

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

Data solidarity for machine learning for embryo selection; a call for the creation of an open access repository of embryo data

Abstract

The last decade has seen an explosion of machine learning (ML) applications in healthcare with mixed and sometimes harmful results, despite much promise and associated hype (Heaven, 2020). A significant reason for these reverses is the premature implementation of machine learning algorithms into clinical practice. In this paper we argue the critical need for “data solidarity” for machine learning for embryo selection. A recent Lancet and Financial Times (FT) commission defined data solidarity as “an approach to the collection, use, and sharing of health data and data for health that safeguards individual human rights while building a culture of data justice and equity, and ensuring that the value of data is harnessed for public good” (Kickbusch et al., 2021).

Introduction

Transparency and reproducibility are key features of the scientific process. Without them it is not possible to validate or build on the work of others. There are two necessary pre-requisites, access to the actual, if not the type of data used in the study and, in the context of machine learning, computational reproducibility (Haibe-Kains et al., 2020). In this paper, we focus on the critical need for access to data.

Developing an ML algorithm for embryo selection

Clinical data from fertility treatment cycles, including outcome data (usually, success or failure of implantation, pregnancy, livebirth or healthy liveborn), are combined with measurements of the embryo, usually static or video embryo images that are labelled to identify key points or features. Key steps are accessing the data, training, testing and then validating the ML algorithm. The goal of the ML process is to construct a predictive model (a formula) which turns the embryo images or measurements into a useful prediction of the outcome.

Embryo data and known confounders

The unique situation in IVF studies is that endpoints of interest are contributed by a wide range of factors, both known and unknown. For example, maternal age, the most known confounding factor, has been ignored in some studies investigating embryonic contribution to treatment outcomes such as implantation or live birth. It is paramount that datasets used for model development include at least the known confounding factors so that bias is minimized. Indeed, there are increasing data showing embryo morphokinetics are altered by patient, treatment and laboratory related factors (Liu et al., 2019). Therefore, embryo selection models that include embryonic parameters alone could potentially lead to biased prediction and this demands further emphasis amongst researchers, not just computer scientists who develop prediction models, but also IVF professionals.

Access to data

The key issue is whether ML algorithms, specifically those developed using deep learning, developed in one population are generalizable to other groups. ML algorithms are known to be subject to bias, through overfitting to the training data, data shifts over time, and bias against unrepresented patient populations (Kaushal et al., 2020; Wu et al., 2021). They are

further subject to other biases, such as cognitive, automation, technical and other biases (Challen et al., 2019; Sujan et al., 2019). Examples abound of deep learning algorithms that showed promise in the laboratory, but have failed to generalize in clinical practice (Heaven, 2020; Wang et al., 2020; Zech et al., 2018).

An absolutely crucial step to reduce the risk of bias and consequent failure in the real-life clinical setting is to validate the ML algorithm on an external dataset, independent of the dataset that it was trained on. As well as external validation, access to a large, good quality, diverse database is beneficial to all. It allows researchers to construct new algorithms and to replicate an algorithm constructed using this dataset, an essential part of scientific rigour. The vast majority of deep learning studies in medicine are not reproducible due to lack of an open database. Breast cancer is one example, where the only publicly available dataset is DDSM, which is created from decade old equipment and has low quality images that are not labeled well (<http://www.eng.usf.edu/cvpgr/Mammography/Database.html>). The more recent databases on breast radiology, such as the InBreast database (Moreira et al., 2012), asks users to hand over intellectual property rights before even accessing the database. Even for non-deep learning studies it can be difficult to track down datasets from medical papers – most will not share data due to fear of data privacy leaks. Consider, for instance, the retracted papers on hydroxy chloroquine treatment for Covid-19. Not even the authors could access the database, and the whole study appears to be fake (Hellman, 2020). There are ethical issues relating to consent, privacy and confidentiality with allowing access to patient data, even when it is truly non-identifying, which can be difficult and need to be considered before allowing such access (Adibuzzaman et al., 2018). However, clinical and laboratory data are essential for the ethical development of ML-based tools, to ensure patients are benefitted and not harmed, as well as and for audit and research. More importantly though, good quality data is life-saving (or life-creating for patients). Failure to access all available data (including unpublished commercial data) leads to ineffective or harmful interventions continuing to be employed (Savulescu et al., 1996). In the case of IVF, it can lead to women having unnecessary repeated cycles of IVF, with its attendant morbidity, not to mention financial expense. There is a moral obligation to make data available.

Sharing data requires effort, time and expense. It must be “cleaned up”, made anonymous, formatted, and corrected as necessary. Furthermore, the data owners (clinics) may be unsure how the data will be used, or worse misused. They might also be concerned that they are undervaluing the data and might wish to hold out for a (higher) price. The late Hans Rösling coined the phrase “database hugging disorder” to describe this reticence to share data (Khokhar, 2017).

Development of an open access, comprehensive repository of embryo images and data

New paradigms for obtaining, storing and sharing data generally, as advocated and soon to be required by the National Institutes of Health (Jorgenson et al., 2021), and specifically for the use of digital technologies (Kickbusch et al., 2021) are needed.

An open access comprehensive repository would allow researchers to train and evaluate ML algorithms for embryo selection on all kinds of real-world data, data that these models would typically be exposed to, in a wide variety of clinical environments. Therefore, this database should be as inclusive as possible, accepting data from all sources so long as they provide a minimum of information per case: for example, the embryo image, the age of the mother at egg collection, and whether the embryo developed into a live-born baby.

95 Data repositories could be achieved under the auspices of government (e.g. HFEA,
1 though
2 96 this would be UK only); professional bodies, such as Academy of Clinical Embryologists;
3 or
4 97 academic institutions, who would oversee the repository, evaluate proposals to access
5 the
6 98 data, allow access to researchers under license, and ensure that the data is used
7 ethically,
8
9 99 and studies ultimately published. Such a repository will have the added benefit of
10
11 100 permitting a comparison of different ML algorithms, as well as considerably speeding up the
12 101 implementation of new, effective algorithms into clinical practice (Kamran et al., 2022).
13 102 Professional bodies, as well as the reproductive medicine community itself, should urge the
14 103 contribution of embryo images and relevant clinical and laboratory information (such as the
15 104 results of genetic analyses) to this central repository. Since the stakes are so high in the
16 105 creation of a new life, and the possibility of compromising the next generation exists, it is
17 106 urgent and essential that such a central repository be established.
18
19 107 This presents formidable ethical challenges (particularly around anonymization and consent)
20 108 which can and must be met if AI is to be fit for the clinic. Many of these issues have already
21 109 been considered (Jorgenson et al., 2021). Crucial is at the outset of collecting and storing
22 110 the data, to consider the moral imperative for data sharing, and to build in systems and
23 111 processes to make data solidarity an inherent attribute of the process.
24
25 112 The reproductive community can benefit from the experience of other clinical databases
26 113 that have been established for the purposes of sharing data with researchers, such as
27 114 Nightingale Science (<https://www.nightingalescience.org/>), a database of CT scans released
28 115 by Duke University ([https://cvit.duke.edu/resource/database-for-benchmarking-organ-](https://cvit.duke.edu/resource/database-for-benchmarking-organ-dose-estimates-in-ct/)
29 116 [dose-estimates-in-ct/](https://cvit.duke.edu/resource/database-for-benchmarking-organ-dose-estimates-in-ct/)), the MIMIC database (<https://lcp.mit.edu/mimic>); neuroimages in
30 117 the Human Brain Project (<https://www.humanbrainproject.eu/en/>); non-clinical databases
31 118 – such as IBM’s “diversity in faces” dataset
32 119 (<https://www.ibm.com/blogs/research/2019/01/diversity-in-faces/>), and of course the
33 120 original “ImageNet” which spurred a whole host of research in computer vision
34 121 (<https://www.image-net.org/>).
35
36 122 Key characteristics of these accessible datasets appear to be that the project is driven by
37 123 visionaries, who cross medical and computer science boundaries, and whose primary goal
38 124 appears to be to contribute to the betterment of medical care; and that there is sufficient
39 125 funding – for example, the Nightingale Science project is funded by a \$2Million grant from
40 126 Eric Schmidt, the former Google chief executive.

127 128 **Premature clinical implementation**

129 The output of an ML algorithm which has been developed and validated internally are
130 performance statistics for that algorithm trained and tested on particular datasets
131 (representing particular populations) for predicting a specified outcome. Unfortunately, it is
132 at this point that the algorithm is submitted for clinical-use regulatory approval, and
133 companies start to aggressively market and sell their products (Wu et al., 2021).
134 There are two significant problems with premature implementation. First, biases are
135 rampant in data, and disguised, especially when black-box models are used (O’Connor,
136 2021). Second, there are large gaps in the evidence base on the interface of digital
137 technologies and health (Kickbusch et al., 2021).

138 139 **Generating evidence of clinical effectiveness**

140 As noted, each algorithm provides measures of performance with respect to its own limited
141 dataset. An absolutely crucial step is validating the ML algorithm on an external dataset. The

dataset should be large and diverse, with as varied and comprehensive data inputs as available, and ideally from different timepoints than the datasets that the ML algorithm was trained on. If the external dataset includes the type of patients and embryos on whom it is intended to use the ML model, its validation on this dataset will give powerful reassurance that it is ready to test in clinical practice.

After external validation, the product is ready for clinical testing in the field before we can confidently, and ethically, apply ML tools into clinical practice (Afnan et al., 2021b). In view of the uncertainties, it would be wise to adopt a “precautionary, mission-orientated and value-based approach” (Kickbusch et al., 2021).

First, clinicians need to be trained on how to use the algorithm, and interpret the output, including understanding levels of uncertainty, so that the whole system (AI + human) performs well (Kickbusch et al., 2021).

Then the product needs to be tested in the way it is envisioned that it will be applied in clinical practice, with the patient-desired outcome pre-specified. The gold standard for testing any intervention in the clinic setting is the randomized controlled trial. ML algorithms are no exception, yet they are rarely tested (Nagendran et al., 2020), despite medicine being replete with examples of promising interventions which have failed to live up to the hype, including deep learning based AI software for healthcare (Heaven, 2020).

Safety concerns

We have elsewhere argued that although initially requiring more work, interpretable ML, and not black-box ML is safer, and in the long run likely to be more accurate as interpretable models will allow problems to be detected and solved (Afnan et al., 2021a). This is of particular importance when it comes to embryo selection, since these algorithms may determine which person will be born.

Once clinical effectiveness has been determined, within a minimum pre-determined safety standard, then post-implementation surveillance is essential to ensure safety of this intervention for embryo selection, for both the individual and society, as even small biases can be amplified over generations. This will require the establishment of a specific registry, which already exists in most countries, and ongoing surveillance and the contribution of this data to the central repository.

Case study

A good example of how a black-box constructed algorithm might have benefitted from being tested in a large external dataset is the Virtus health device registered as a medical device in Australia ([https://www.ebs.tga.gov.au/servlet/xmlmillr6?dbid=ebs/PublicHTML/pdfStore.nsf&docid=7F71610B157CC231CA258687003CBCD3&agid=\(PrintDetailsPublic\)&actionid=1](https://www.ebs.tga.gov.au/servlet/xmlmillr6?dbid=ebs/PublicHTML/pdfStore.nsf&docid=7F71610B157CC231CA258687003CBCD3&agid=(PrintDetailsPublic)&actionid=1)) “... to provide clinical decision support for embryo assessment. The device evaluates early embryo development through acquired embryo time lapse images/videos to assist embryo selection during assisted reproduction. The device is intended as an adjunct to clinical decisions. The final assessment and decision shall be made by the embryologist.” This device predicted with 93% accuracy a fetal heart pregnancy (Tran et al., 2019). There is currently a non-inferiority RCT registered (<https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=379161&isReview=true>) for this product. However, as noted in our previous paper (Afnan et al., 2021b), the embryos that the algorithm was trained on was a comparison between good quality embryos and

embryos that were of such poor quality that they would have been discarded and not considered for transfer. Had a large and diverse independent external embryo repository been available, the algorithm developed by Tran and colleagues could have been readily and easily validated not just for performance statistics between good- and poor-quality embryos in their own dataset, but also in a large and diverse dataset in more clinically relevant situations, giving more confidence, or otherwise, to the researchers, funders, and trial participants to investigate this algorithm's effectiveness in an RCT.

A call to action

This paper is primarily a call to professional embryology societies, such as the Academy of Clinical Embryologists which is served by this journal, to develop and oversee access to a large comprehensive database, the Embryo Repository, independent of commercial interests, and monitor the ethical use of the data. Access to such a database could be licensed and paid for, recouping the costs to the society and to the individual clinics. We call on computer scientists to share source code, allowing others to replicate their work, a key element of scientific rigor, and to allow others to advance the field by building on their work (Buda et al 2021).

We also call on clinicians to become knowledgeable about ML, and not be taken in by the hype; to regulators to insist on reliable evidence of clinical effectiveness, and not just performance statistics; and to governments or professional societies to add ML intervention post-implementation surveillance.

A change in culture is required to make data sharing the rule rather than exception. This change in culture can be encouraged by funders, such as NIH (Jorgenson et al., 2021), and journals, such as Nature Portfolio journals, which state that "...authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications." (<https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards>). We note that this journal (RBMO) "encourages" sharing of data and source code (<https://www.rbmojournal.com/content/authorinfo>).

Conclusion

A comprehensive, diverse, accessible database is essential for robust clinical development of effective ML and thus provide benefit and prevent harm to patients. There is a moral obligation to contribute and share data. Self-regulation requires taking responsibility. The Reproductive Medicine and Computer Science communities have the opportunity to lead the way for the ethical implementation of ML algorithms in embryo selection, by sharing data and code, thereby safely and rapidly improving patient care. Let's not succumb to passivity or trade this priceless opportunity for short-term narrow considerations, such as commercial gain.

References

- Adibuzzaman, M., DeLaurentis, P., Hill, J., Benneyworth, B.D., 2018. Big data in healthcare – the promises, challenges and opportunities from a research perspective: A case study with a model database. AMIA Annu. Symp. Proc. April 16, 384–392.
- Afnan, M.A.M., Liu, Y., Conitzer, V., Rudin, C., Mishra, A., Savulescu, J., Afnan, M., 2021a. Interpretable, Not Black-Box, Artificial Intelligence Should be Used for Embryo Selection. Hum. Reprod. Open. <https://doi.org/https://doi.org/10.1093/hropen/hoab040>

- 236 Afnan, M.A.M., Rudin, C., Conitzer, V., Savulescu, J., Mishra, A., Liu, Y., Afnan, M., 2021b.
- 237 Ethical Implementation of Artificial Intelligence to Select Embryos in in Vitro
- 238 Fertilization. AIES 2021 - Proc. 2021 AAAI/ACM Conf. AI, Ethics, Soc. 316–326.
- 239 <https://doi.org/10.1145/3461702.3462589>
- 240 Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., Tsaneva-Atanasova, K., 2019.
- 241 Artificial intelligence, bias and clinical safety. BMJ Qual. Saf. 28, 231–237.
- 242 <https://doi.org/10.1136/bmjqs-2018-008370>
- 243 Haibe-Kains, B., Adam, G.A., Hosny, A., Khodakarami, F., Shraddha, T., Kusko, R., Sansone,
- 244 S.A., Tong, W., Wolfinger, R.D., Mason, C.E., Jones, W., Dopazo, J., Furlanello, C.,
- 245 Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., Greene, C.S.,
- 246 Broderick, T., Hoffman, M.M., Leek, J.T., Korthauer, K., Huber, W., Brazma, A., Pineau,
- 247 J., Tibshirani, R., Hastie, T., Ioannidis, J.P.A., Quackenbush, J., Aerts, H.J.W.L., 2020.
- 248 Transparency and reproducibility in artificial intelligence. Nature 586, E14–E16.
- 249 <https://doi.org/10.1038/s41586-020-2766-y>
- 250 Heaven, W.D., 2020. Google’s medical AI was super accurate in a lab. Real life was a
- 251 different story. [WWW Document]. MIT Technol. Rev.
- 252 Hellman, J., 2020. Authors retract major COVID-19 paper on effects of hydroxychloroquine.
- 253 Hill.
- 254 Jorgenson, L.A., Wolinetz, C.D., Collins, F.S., 2021. Incentivizing a New Culture of Data
- 255 Stewardship. Jama 20814, 1–2. <https://doi.org/10.1001/jama.2021.20489>
- 256 Kamran, F., Tang, S., Otles, E., McEvoy, D.S., Saleh, S.N., Gong, J., Li, B.Y., Dutta, S., Liu, X.,
- 257 Medford, R.J., Valley, T.S., West, L.R., Singh, K., Blumberg, S., Donnelly, J.P., Shenoy,
- 258 E.S., Ayanian, J.Z., Nallamothu, B.K., Sjoding, M.W., Wiens, J., 2022. Early identification
- 259 of patients admitted to hospital for covid-19 at risk of clinical deterioration: model
- 260 development and multisite external validation study. Bmj e068576.
- 261 <https://doi.org/10.1136/bmj-2021-068576>
- 262 Kaushal, A., Altman, R., Langlotz, C., 2020. Geographic distribution of US cohorts used to
- 263 train deep learning algorithms. JAMA - J. Am. Med. Assoc. 324, 1212–1213.
- 264 <https://doi.org/10.1001/jama.2020.12067>
- 265 Khokhar, T., 2017. Hugs and databases: in memory of Hans Rosling [WWW Document]. URL
- 266 <https://blogs.worldbank.org/opendata/hugs-and-databases-memory-hans-rosling>
- 267 (accessed 2.20.22).
- 268 Kickbusch, I., Piselli, D., Agrawal, A., Balicer, R., Banner, O., Adelhardt, M., Capobianco, E.,
- 269 Fabian, C., Singh Gill, A., Lupton, D., Medhora, R.P., Ndili, N., Ryś, A., Sambuli, N., Settle,
- 270 D., Swaminathan, S., Morales, J.V., Wolpert, M., Wyckoff, A.W., Xue, L., Bytyqi, A.,
- 271 Franz, C., Gray, W., Holly, L., Neumann, M., Panda, L., Smith, R.D., Georges Stevens,
- 272 E.A., Wong, B.L.H., 2021. The Lancet and Financial Times Commission on governing
- 273 health futures 2030: growing up in a digital world. Lancet 398, 1727–1776.
- 274 [https://doi.org/10.1016/s0140-6736\(21\)01824-9](https://doi.org/10.1016/s0140-6736(21)01824-9)
- 275 Liu, Y., Feenan, K., V, C., Matson, P., 2019. Assessing efficacy of Day 3 embryo time-lapse
- 276 algorithms retrospectively: impacts of dataset type and confounding factors. Hum.
- 277 Fertil. 22, 182–190.
- 278 Moreira, I., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M., Cardoso, J., 2012. INbreast:
- 279 toward a full-field digital mammographic database. Acad Radiol. 19, 236–48.
- 280 <https://doi.org/10.1016/j.acra.2011.09.014>
- 281 Nagendran, M., Chen, Y., Lovejoy, C.A., Gordon, A.C., Komorowski, M., Harvey, H., Topol,
- 282 E.J., Ioannidis, J.P.A., Collins, G.S., Maruthappu, M., 2020. Artificial intelligence versus

clinicians: Systematic review of design, reporting standards, and claims of deep
 learning studies in medical imaging. *BMJ* 368, 1–12. <https://doi.org/10.1136/bmj.m689>
 O'Connor, M., 2021. Algorithm's 'unexpected' weakness raises larger concerns about AI's
 potential in broader populations. [WWW Document]. Healthimaging. URL
[https://www.healthimaging.com/topics/artificial-intelligence/weakness-ai-broader-](https://www.healthimaging.com/topics/artificial-intelligence/weakness-ai-broader-patient-)
 patient- (accessed 11.12.21).
 Savulescu, J., Chalmers, I., Blunt, J., 1996. Are research ethics committees behaving
 unethically? Some suggestions for improving performance and accountability. *BMJ* 313,
 1390–3.
 Sujan, M., Furniss, D., Grundy, K., Grundy, H., Nelson, D., Elliott, M., White, S., Habli, I.,
 Reynolds, N., 2019. Human factors challenges for the safe use of artificial intelligence in
 patient care. *BMJ Heal. Care Informatics* 26, 1–5. [https://doi.org/10.1136/bmjhci-2019-](https://doi.org/10.1136/bmjhci-2019-100081)
 100081
 Tran, D., Cooke, S., Illingworth, P.J., Gardner, D.K., 2019. Deep learning as a predictive tool
 for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. *Hum.*
Reprod., [Comment in: *Hum Reprod.* 2020 Feb 29;35(2):482; PMID: 32053171
[\[https://www.ncbi.nlm.nih.gov/pubmed/32053171\]](https://www.ncbi.nlm.nih.gov/pubmed/32053171)][Comment in: *Hum Reprod.* 2020
 Feb 29;35(2):483; PMID: 32053191
[\[https://www.ncbi.nlm.nih.gov/pubmed/32053191\]](https://www.ncbi.nlm.nih.gov/pubmed/32053191) 34, 1011–1018.
<https://doi.org/https://dx.doi.org/10.1093/humrep/dez064>
 Wang, X., Liang, G., Zhang, Y., Blanton, H., Bessinger, Z., 2020. Inconsistent Performance of
 Deep Learning Models on Mammogram Classification. *J. Am. Coll. Radiol.* 17(6), 796–
 803. <https://doi.org/10.1016/j.jacr.2020.01.006>
 Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D.E., Zou, J., 2021. How medical AI devices
 are evaluated: limitations and recommendations from an analysis of FDA approvals.
Nat. Med. 27, 582–584. <https://doi.org/https://doi.org/10.1038/s41591-021-01312-x>
 Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K., 2018. Variable
 generalization performance of a deep learning model to detect pneumonia in chest
 radiographs : A cross-sectional study. *PLoS Med.* 15, 1–17.
<https://doi.org/10.1371/journal.pmed.1002683> N