All Models Are Wrong, but Some Are Useful

Carol Lynn Curchoe¹

Received: 13 July 2020 / Accepted: 14 July 2020 / Published online: 7 October 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Statistician George E.P. Box famously said, “All models are wrong, but some are useful”. Reproduction researchers and clinicians are grappling to critically evaluate the recent deluge of artificial intelligence (AI) studies to determine if they are “useful” for prediction, or at the very least, can automate the manual, routine, and subjective drudgery of day-to-day clinical practice.

Predictive modeling has evolved into a standalone subdiscipline of reproductive medicine, and the literature have been analyzed and evaluated [1] using formal systems, such as; the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) and prediction model risk of bias assessment tool (PROBST). Predictive modeling’s place as a support tool is growing in clinical practice, but gains in patient outcomes and improved workflows have yet to be fully realized.

Will artificial intelligence also fail to live up to the hype? And why is AI so breathtakingly exciting?

Despite early promise, global approaches in transcriptomics, epigenomics, and proteomics [2] have largely failed to unambiguously identify statistically relevant clinical signatures for complex problems in reproduction, while simultaneously providing a wealth of basic knowledge that improves our understanding of embryo development. Successful human reproduction is a complex problem with so many variables that even the best-planned studies can quickly become confounded, yielding inconclusive results.

Inherent variation and subjectivity are the enemy of consistency and objectivity. AI can address these challenges (presumably with greater accuracy) to assemble solutions that cannot be resolved by the human senses. AI is perfectly suited for the seemingly intractable questions of reproductive medicine, for example, embryo selection [3], the complex dialogue between endometrium and embryo and recurrent miscarriage [4], the physiological function of the uterus and disease states

like endometriosis and adenomyosis [5], therapeutic targets for biological and chronological ovarian ageing [6], preimplantation genetics to improve pregnancy outcomes [7], and recurrent implantation failure [8].

If you are struggling to understand what artificial intelligence is let alone how it works, consider this great analogy. Imagine the “gold standard” outcome, i.e., a healthy, live-born infant, conceived quickly through reproductive medicine, as a ball of crumpled paper. The ball has many sheets of paper and each sheet of paper is a different complex problem in and of itself: patient demographics, gamete quality, disease etiology, embryo quality, uterine lining and receptivity, and more. Artificial intelligence “uncrumples” this ball, working backward through each step, slowly unraveling the mystery of how the ball came to be crumpled.

AI finds neat representations for complex, highly folded data, and machine learning is the main instrument to build artificial intelligence. It works superbly to make forecasts on incoming data, using historical data for training. To date, a wealth of embryo, oocyte, and semen videos and images have been retrospectively analyzed, and Fernandez et al. review the resulting AI model types, accuracy, strengths and weaknesses in this timely work in the current edition of *Journal of Assisted Reproduction and Genetics*.

The performance obtained from artificial intelligence (AI) models and their possible clinical utility, hinges on the quality and size of databases (BIG is better), the types and distribution of data, and the choice of AI methods applied. However, there is currently little consensus on publication and evaluation standards [9], and it begs the question—do AI models and databases differ enough from other types of statistical modeling to need their own checklist, similar to TRIPOD, but specific for AI? Indeed, the TRIPOD steering group has proposed the development of specific reporting guidance for prediction model studies that are based on artificial intelligence or machine learning methods for model development, validation, or updating.

Another challenge for researchers is the need for “Big Data” and collaboration for prospective, cohort studies. These data are highly sensitive, strictly regulated, and patient

✉ Carol Lynn Curchoe
cburton@fertilitylabsciences.com

¹ CCRM IVF Network, Lone Tree, CO, USA

consent and ethical approval are mandatory. “Big Data” is a lot like a teenage love affair. Everyone wants it, everyone thinks they have it, and everyone thinks they are doing it right. But “Big Data” can present even bigger problems—data collection, preprocessing, and sharing practices ethics, as well as tagging and annotation, are as much an art as they are a science.

A critical step in modeling is preprocessing and the treatment of images, before feature extraction and model training. Clumsy preprocessing and federated subsets of data generated from various geographic locations (with different cropping and contrast, taken on a wide variety of cameras and microscopes) may yield an inconsistent database that leads AI models to poor accuracy. A world-wide common Reproduction Git image and video embryo repository (similar to genome sequence databases) and automated methods to crop, remove patient information, rotate, and size images or video is a worthwhile goal that will take immense international collaboration. Commercial block chain solutions are seeking to provide distributed frameworks to enable privacy-preserving, collaborative machine learning, and guarantee traceability and authenticity of data using a distributed ledger [10].

There is another flaw in all studies that use embryo images and video (analyzed by both AI and statistical models) to predict pregnancy outcome. The most common type of images used are blastocyst images taken before biopsy and cryopreservation, thawing, and transfer. The culture conditions of the lab, technical competency of the operator (both clinical [11] and laboratory [12, 13]), and embryo quality and/or expansion postthaw are not analyzed in models that use clinical pregnancy as an end-point, despite being dependent on them.

There are significant challenges to implementing any AI system in a meaningful way into a clinical IVF laboratory workflow. Implementation of an AI at the actual point of care in a routine, easy, and automated fashion is nearly impossible for the majority of IVF labs that still paper charts and do not take a single image of their patient’s embryos, much less video, or augment patient demographics, clinical and lab KPIs, and other relevant data streams (ultrasound images, and PGT results, competency assessments) into a single dashboard.

Certainly, most embryo cryostorage inventory “systems” are rows of neatly labeled 3 ring binders, backed up by a spreadsheet. The original, hand-written embryo cryorecord has a sacrosanct quality to it that I do not see as easily replaceable in the hearts and minds of clinical embryologists.

Now that the low hanging (retrospective, 2D, and morphokinetic) fruit have been plucked, research scientists, and clinicians are tasked to transcend simply packaging established pregnancy predictors into fancy new machine learning algorithms. Prospective design, BIG data, standardized outcome measures, and external validation with integration of clinical and laboratory KPIs plus patient demographics

will help identify novel variables and hidden relationships that allow for superior predictive capabilities.

Current controversies in the field illuminate the difficulty we have with subjective grading [14], false negatives, mosaic embryos, and the near-certainty that we are “discarding” viable embryos. Bias and black-box problems mean that AI systems that seek to diagnose, rather than rank, embryos for transfer have the potential to worsen this problem.

Perhaps, predictive modeling has not been readily embraced because much of the culture of reproductive medicine, its beliefs, and aspirations are rooted in making “everyday miracles”. We have all been overjoyed when a shabby embryo leads to a positive beta [14]. At the end of the day, every embryo should be given a chance to defy the odds, no matter what the model says.

Compliance with ethical standards

Conflict of interest CC is the founder of ART Compass, an AI-based software platform for IVF Laboratory quality management.

References

1. Ratna MB, Bhattacharya S, Abdulrahim B, McLernon DJ. A systematic review of the quality of clinical prediction models in in vitro fertilisation. *Hum Reprod*. 2020;35(1):100–16.
2. Krisher RL, Schoolcraft WB, Katz-Jaffe MG. Omics as a window to view embryo viability. *Fertil Steril*. 2015;103(2):333–41.
3. Bormann CL, Thirumalaraju P, Kanakasabapathy MK, Kandula H, Souter I, Dimitriadis I, et al. Consistency and objectivity of automated embryo assessments using deep neural networks. *Fertil Steril*. 2020;113(4):781–7 e1.
4. Macklon NS, Brosens JJ. The human endometrium as a sensor of embryo quality. *Biol Reprod*. 2014;91(4):98.
5. Kunz G, Leyendecker G. Uterine peristaltic activity during the menstrual cycle: characterization, regulation, function and dysfunction. *Reprod BioMed Online*. 2002;4(Suppl 3):5–9.
6. Alviggi C, Humaidan P, Howles CM, Tredway D, Hillier SG. Biological versus chronological ovarian age: implications for assisted reproductive technology. *Reprod Biol Endocrinol*. 2009;7:101.
7. Munne S, et al. Preimplantation genetic testing for aneuploidy versus morphology as selection criteria for single frozen-thawed embryo transfer in good-prognosis patients: a multicenter randomized clinical trial. *Fertil Steril*. 2019;112(6):1071–9 e7.
8. Bashiri A, Halper KI, Orvieto R. Recurrent implantation failure—update overview on etiology, diagnosis, treatment and future directions. *Reprod Biol Endocrinol*. 2018;16(1):121.
9. 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(4):591–600.
10. Galtier MNMC (2019) Substra: a framework for privacy-preserving, traceable and collaborative Machine Learning. *ArVix*
11. Cirillo F, Patrizio P, Baccini M, Morengi E, Ronchetti C, Cafaro L, et al. The human factor: does the operator performing the embryo transfer significantly impact the cycle outcome? *Hum Reprod*. 2020;35(2):275–82.

12. Franco JG Jr, et al. Key performance indicators score (KPIs-score) based on clinical and laboratorial parameters can establish benchmarks for internal quality control in an ART program. *JBRA Assist Reprod*. 2017;21(2):61–6.
13. Swain JE. Controversies in ART: can the IVF laboratory influence preimplantation embryo aneuploidy? *Reprod BioMed Online*. 2019;39(4):599–607.
14. Morbeck DE. Blastocyst culture in the era of PGS and FreezeAlls: is a 'C' a failing grade? *Hum Reprod Open*. 2017;2017(3):hox017.

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