Article published online: 2024-10-08

Patient-Centric In Vitro Fertilization Prognostic Counseling Using Machine Learning for the Pragmatist

Mylene W.M. Yao, MD[1] Julian Jenkins, DM[2] Elizabeth T. Nguyen, PhD[1] Trevor Swanson, PhD[1] Marco Menabrito, MD[1]

[1] R&D Department, Univfy, Los Altos, California
[2] Jencap Consulting Ltd., Cardiff, United Kingdom

Address for correspondence Mylene W.M. Yao, MD, Univfy Inc., 171 Main Street, #139, Los Altos, CA 94022 (e-mail: mylene.yao@univfy.com).

Semin Reprod Med 2024;42:112–129

Abstract

Although in vitro fertilization (IVF) has become an extremely effective treatment option for infertility, there is significant underutilization of IVF by patients who could benefit from such treatment. In order for patients to choose to consider IVF treatment when appropriate, it is critical for them to be provided with an accurate, understandable IVF prognosis. Machine learning (ML) can meet the challenge of personalized prognostication based on data available prior to treatment. The development, validation, and deployment of ML prognostic models and related patient counseling report delivery require specialized human and platform expertise. This review article takes a pragmatic approach to review relevant reports of IVF prognostic models and draws from extensive experience meeting patients' and providers' needs with the development of data and model pipelines to implement validated ML models at scale, at the point-of-care. Requirements of using ML-based IVF prognostics at point-of-care will be considered alongside clinical ML implementation factors critical for success. Finally, we discuss health, social, and economic objectives that may be achieved by leveraging combined human expertise and ML prognostics to expand fertility care access and advance health and social good.

Keywords
- prognostic counseling
- live birth probability
- artificial intelligence
- machine learning
- precision medicine

A critical factor for patients to decide whether to proceed with in vitro fertilization (IVF) is understanding their likelihood of achieving a live birth based on their own health data. To help meet this challenge of personalized prognostication, machine learning (ML), a broad discipline within the broader field of artificial intelligence (AI), allows machines to extract relationships from data and learn from it autonomously.[1] Using established ML techniques selected based on dataset attributes and the clinical context of patient counseling, one would develop, validate, deploy, and implement prognostic models for use at point-of-care. Supported by secured cloud computing, a provider–patient counseling report is one way to communicate personalized, validated IVF live birth probabilities (IVF LBP) at scale.[2,3] In this review article, IVF is used broadly and interchangeably with assisted reproductive technology (ART).

Patients considering IVF treatment wish to know their probability of having a live birth from IVF (IVF LBP) and alternative treatments. By showing patients their characteristics such as age, body mass index, ovarian reserve, and clinical diagnosis compared with the whole group used to derive the model, patients may feel reassured the model is considering them as individuals when making predictions.[2,3] Patients also want to know if their prognoses are validated for their particular fertility center's IVF outcomes data.[2,3] The expanding usage of AI data-driven decisions in everyday life encourages patients to trust using technology to support important decisions.

Issue Theme Health Technology and Reproduction; Guest Editor, Shruthi Mahalingaiah, MD, MS

DOI https://doi.org/10.1055/s-0044-1791536.
ISSN 1526-8004.

Thieme Medical Publishers, Inc., 333 Seventh Avenue, 18th Floor, New York, NY 10001, USA

Multiple benefits could arise from personalized prognostication of IVF outcomes using ML. Underutilization of IVF is a major limitation on the use of now highly effective fertility treatments. ML predictive models may address IVF underutilization by providing better quality information to patients to inform their decision-making and by making a course of IVF treatments and achieving an IVF live birth more affordable.[2] Predicting IVF outcome by age alone as frequently used in IVF centers has been shown to underestimate the likelihood of live birth for many individuals, and thus may discourage them from IVF, whereas a more accurate ML prediction model could appropriately encourage patients to embark on an IVF cycle.[2] Similarly, an accurate ML prediction model could encourage patients with a low likelihood of success with their own oocytes to move on to more effective treatment with donor oocytes. Conversely, misconceptions that IVF efficacy is low and unpredictable may compromise an individual's reimbursement by her health insurance plan or, at the population level, discourage health plans from offering IVF insurance coverage or deprioritize state funding in the case of countries with government-based IVF funding. Therefore, efforts to avoid unnecessary underestimation of IVF success, though insufficient on its own, nevertheless are required to achieve parity of IVF funding compared with other areas in medicine.

This review article aims to summarize the use of expert ML-based prognostication of IVF outcomes to support patient counseling. The authors draw from extensive experience meeting patients' and providers' needs with the development of data and model pipelines to produce validated ML models supporting the use of ML center-specific (MLCS) models at the point-of-care and multicenter implementation of MLCS models at scale. This article does not provide an exhaustive review of all IVF prognostic models in the literature but rather prioritizes the most relevant, published models and, as much as possible, models that are in clinical use as examples to support the discussion of IVF prognostic model design, validation considerations, and other requirements for successful clinical usage. Here, we share design and implementation issues we have encountered and the insights we gained from creating and complying with standard operation procedures of our software product life cycle. Our decision-making and execution are guided by ethics, scientific integrity, and compassion. With the expanding capabilities of AI/ML to improve human health and, broadly, humanity, our responsibility to use technology for good is more important than ever before.[4,5]

We will first consider IVF underutilization and the potential role of prognostic counseling in expanding IVF access and utilization. Requirements of using ML-based IVF prognostics at point-of-care will be considered alongside clinical ML implementation factors critical for success. These latter topics are relatively new in the clinical research literature yet are becoming an important part of provider education as ML enters our personal and professional lives. Wherever relevant, reference will be made to how ML implementation has been managed in other areas of medicine including applicable guidelines. We will explore the potential benefits and risks of using ML-based IVF prognostics and ways to evaluate their impact on treatment utilization and outcomes. Finally, we discuss research, social, and economic objectives that may be achieved by leveraging combined human expertise and ML prognostics to advance health and social good.

Underutilization of ART and Challenges in Navigating Fertility Care

Despite its proven safety and efficacy, ART is vastly underutilized and even in patients for whom ART is appropriate and funded through national reimbursement many patients stop treatment prematurely when they would have still had a good chance of success, if they had continued ART.[6–13] One in six people in the reproductive-age group or 100M+ is estimated to have clinical infertility worldwide.[6,7] In the United States, an estimated 10M+ people have infertility, defined by the American Society for Reproductive Medicine (ASRM) as needing medical care to conceive or have a successful pregnancy.[7–9] However, less than 2% of women or couples who could benefit from ART actually used it based on annual reporting by the Society of Assisted Reproductive Technologies (SART).[9] Globally, ~1M babies are born from ~4M ART cycles performed annually.[10,11] Realizing the full family- and society-building potential of IVF requires identifying and solving the barriers in patients' navigation of fertility care for one in six women or couples, nontraditional families such as single women and same-sex couples, and families with hereditary genetic diseases.[6–14]

The causes of the underutilization of ART are complex, but the main barriers cited are emotional stress, uncertainty of treatment success, financial cost, inadequate insurance coverage, and issues arising from limited mechanistic knowledge.[7,14–18] In the United States, while some level of IVF coverage increased from 34% (2015) to 42% (2020) among U.S. employers with 20,000 or more employees, an estimated ~40% of Americans with employer-funded health insurance did not have IVF coverage.[19,20] Warranting special mention are the racial and socioeconomic disparities in fertility care in the United States.[14,20] Despite higher reporting rates of infertility, Black women are less likely to receive fertility diagnostic testing even when care is sought and less likely to receive fertility treatments including IVF. Black and Hispanic women also have lower fertility treatment success for reasons that are not well understood.[5,14,20]

Most countries including Sub-Saharan Africa, low- and middle-income countries in LATAM, South Asia, and East Asia do not have government-funded ART, resulting in low ART utilization.[21] In contrast, many of the industrialized countries already have national support for IVF as a health measure or to attempt to tackle the problem of declining fertility rates or both.[22] Despite such state funding, personal accurate IVF prognostics are typically not available to help patients choose effective treatments and to minimize using ineffective treatments. In fact, ML-based, personalized prognostic counseling may improve efficiency, patient retention, and IVF outcomes while optimizing the use of resources.

Many more people would have a family from IVF if they could afford several IVF treatments. A sustainable family-building program—whether paid by third parties or patients—should consider cumulative live birth probability per IVF cycle because this probability directly impacts the number and cost of IVF cycles needed to have a baby. Once a patient starts IVF treatment, the major limitation of achieving an IVF live birth is the high rate of treatment discontinuation or "drop-out" rate. From first-hand data analysis by our team, historically among self-pay patients (e.g., no government or third-party payer), approximately 80% tend not to return after one failed IVF cycle (aka 80% drop-out rate), while even state-funded or state-mandated covered patients may show a drop-out rate of 30 to 50% after one failed IVF cycle.

Accurate IVF live birth prediction models can support the pricing of IVF treatments based on the outcome of having "a baby or a partial refund." Commonly known as "shared risk" program initially popularized in the 1990s, this pricing method simply charges a discounted fee upfront for performing up to two to three IVF treatments, until a baby or a clinical pregnancy is achieved. If there is no live birth after three IVF treatments, then the patient would be paid a "partial refund." Although patients and fertility centers theoretically benefit from this arrangement, without ML optimization, a substantial percentage of patients may not qualify, or fertility centers set a high upfront fee to protect from financial losses that can be incurred from suboptimal IVF success prediction. In contrast, an ML-driven shared risk program can be offered to the majority of patients and is compliant with the transparency requirements of the Ethics Committee of the American Society of Reproductive Medicine (ASRM).[23,24]

Last but not least, information asymmetry currently exists between the business operations of many fertility centers and their providers and patients in which the business operations may have IVF LBP insights informing qualification for shared risk program; yet, those insights may not be known to providers and patients. We advocate for transparency and scientific literacy which are hand and glove and essential to advancing reproductive care and access to care. To patients and payers, IVF prognostic counseling and the cost of having an IVF baby are one and the same conversation, best supported by locally relevant data and ML.

Solving IVF Prognostic Counseling Challenges

Effective IVF prognostic counseling requires accurate, personalized prognoses and clear, consistent communication of the prognoses.[25,26] The literature has focused on the reporting of IVF live birth rates,[27,28] patients' psychology,[17,25,26] and concerns over patients' overestimates of their personal IVF treatment success probabilities.[29,30] Furthermore, the communication of prognosis to patients tends to be unsupported and relegated to be a matter of personal style, under the label of provider autonomy as commonly seen in other areas of medicine.[31] However, patients' psychology may vary depending on the transparency or information symmetry between patient and provider and it is not possible to measure whether patients under- or overestimated their IVF prognoses in the absence of a validated model and effective communication.

As the efficacy of IVF improves, patients should know their personal prognoses whether poor or excellent, based on their own health profiles. For example, patients with excellent IVF LBP may miss out on an opportunity to have a family if they were to underestimate their prognoses. For patients for whom IVF (with own or donor eggs) and IUI are both possible treatment options, patient counseling is especially important, as there is a wavering consensus on whether IUI or IVF should be offered as first-line treatment for patients with unexplained infertility.[32–34] On the personal front, patients may differ in how they perceive personal tradeoffs such as financial cost, time from work, and side effects versus having a family. Finally, a common complaint from patients is their perception of IVF as a gamble based on the uncertainty and lack of transparency about IVF treatment success on a personal level.

Scaling IVF access and removing health inequities may require IVF prognostic counseling to be delivered by healthcare providers beyond fertility specialists. Motivated by a need to address the clinical and socioeconomic challenges for patients and society at large, we have developed an ML technology platform to support patient counseling, treatment protocol personalization, transparency-driven value-based IVF pricing design, and advancement of precision medicine through the use of accurate, validated clinical prediction models as summarized in the platform schematics.[35] This platform has generated published research, some of which will be further examined later.[2,36–40]

A Pragmatic Overview of Model Design Considerations with Examples from the Literature

Next, we present model design considerations using models reported in the literature to illustrate key points. As much as possible, we focus on pretreatment models designed to support patient counseling prior to starting the first IVF cycle. These model design considerations could be easily extrapolated to address other clinical scenarios—after failing one or more IVF treatments and when considering egg freezing or donor egg IVF, etc.

Dissecting the Literature Based on Model Objectives and the Reporting of Model Validation

A review of predictors of success after IVF by Shingshetty et al following a comprehensive literature search between 1978 and 2023 identified 1,810 publications meeting initial keyword search requirements from which 43 articles were selected for detailed review.[41] However, pragmatically before considering prognostic models it is important to first specify the model objective, which will in turn define relevant clinical variables, outcomes, and other dataset attributes.

►**Table 1** shows the importance of defining the model objectives (e.g., pretreatment counseling, research) and, in the case of pretreatment counseling, the exact clinical contexts (e.g., prior to the first IVF cycle, after a failed IVF cycle). The allowable clinical variables, required outcomes, and data segments to consider for exclusion can then be easily determined. The clinical variables that are commonly tested for predictive value in pretreatment IVF prognostics include age, BMI, ovarian reserve tests (e.g., AMH, AFC, Day 3 FSH), reproductive history, prior IUI or IVF treatment history, and clinical diagnoses such as PCOS, tubal factor, male factor, endometriosis, diminished ovarian reserve, smoking, and the duration of infertility.[42] For example, if the research objective was to understand clinical factors impacting IVF live birth outcomes, it is reasonable to use a dataset comprising clinical variables obtained from pretreatment, ovarian stimulation, and embryology. However, if the objective is to create a prognostic model to counsel patients at the pretreatment stage to consider using IVF for the first time, the variables should be restricted to information that is known and available at the time of patient counseling. The model outcome should be selected to enable the provider to respond to patients' needs. For example, patients wish to know their probability of having a live birth, not a positive pregnancy test. The concept of "reading the ending first" is key, otherwise you may build an excellent model with no clinical utility. This more pragmatic approach diverges from conventional scientific methods requiring the researcher to pose and test a series of hypotheses.

Taking a pragmatic approach we reviewed the 43 articles identified in Shingshetty et al and added further 7 articles—5 published prior to 2024, one published in 2024, and one submitted in 2024.[36–41,43–88] We reviewed all 50 articles and assigned them to subgroups based on clinical context (e.g., applicability of model to current IVF practice), presumed model objective (e.g., research or pretreatment counseling for a particular scenario), modeling method (e.g., logistic regression [LR] vs. other ML techniques), and whether an independent test set was used for validation or testing (see ►**Table 2**).

Although outside of the scope of the IVF prognostic model for patient counseling, we appreciate the 19 articles contributing to the understanding of clinical factors influencing IVF outcomes and/or modeling methods without utility for patient counseling, referred to as "research objective" in ►**Table 2** as they required information only available following treatment.[44,54,60,63,67–73,75–81]

Top-Most Relevant Published IVF Pretreatment Models for Patient Counseling

Now we consider 12 publications to illustrate model design considerations in support of patient counseling prior to starting the first IVF treatment[36–38,45,82–87] (see ►**Table 3**). Accordingly, these publications (a subset highlighted in blue, ►**Table 2**) are selected based on the following: (1) limiting predictors to information available at the time of pretreatment counseling prior to the first IVF cycle; (2) using

Table 1 IVF prediction model objectives should determine model design including dataset attributes, features to be tested, outcomes, and AI/ML techniques

	Model objective and outcome of interest					
	Pretreatment, predict IVF live birth probability			Pretreatment, predict oocyte yield	Research and outcome(s) of interest	
Clinical timing/ context of usage	Prior to 1st IVF cycle using own eggs	After 1 or more failed IVF cycles using own eggs	Prior to IVF cycle using donor egg	Prior to egg freezing	To gain insights, generate hypothesis for testing	
Dataset	1st IVF cycles +/− subsequent IVF cycles labeled with cycle number and linked per patient, linked ETs and outcomes, restricted to IVF cycles using own eggs	Failed IVF cycles and subsequent IVF cycles labeled with cycle number and linked per patient, linked ETs and outcomes, restricted to IVF cycles using own eggs	ETs using donor egg or IVF cycles using known or traditional donors and their subsequent ETs, with donor–recipient linkage and linked per recipient	Egg freezing dataset, IVF dataset, or combined egg freezing and IVF dataset	IVF cycles, linked ETs, linked per patient, and outcome(s) of interest	
Which patients and/or IVF cycles should be additionally excluded?	IVF cycles that address very specific patient population may be included if there is a way to differentiate the labeling of those patients and/or IVF cycles and if there is a way to validate patient subgroups. The ability to include a patient subgroup in the model training and provide subgroup validation will determine if the IVF prognostic model can be appropriately applied to that patient subgroup					
Relevant variables for testing as model predictors	Restrict to variables with known values prior to IVF cycle start	Restrict to variables with known values prior to starting the subsequent IVF cycle; include variables with known values from the prior failed IVF cycle (e.g., oocyte count, blastocyst count)	For both donor and recipient variables, restrict to variables available at the time of counseling patients about donor egg IVF	Restrict to variables with known values prior to starting ovarian stimulation for egg freezing or IVF cycle	Include any variables of interest that are available in the dataset	
Which variables should be additionally excluded?	Consider excluding variables that may not be available at the time of patient counseling. If the model includes variables whose values are logistically challenging to obtain, the model usage will be limited					

Notes: This table is limited to model using structured data and does not attempt to address AI tools aimed to identify blastocysts for transfer. Also, this table serves to illustrate key principles and does not aim to provide an exhaustive list of possible models for prognostic counseling. For example, some scenarios of third-party reproduction are not shown.

Table 2 A review of 50 articles reporting IVF outcomes prediction models, including 43 from Shingshetty et al and 7 additional articles[36–41,43–88]

Relevance for clinical usage in IVF pretreatment counseling	Logistic regression, no test set, no validation	Logistic regression, cross-validation, or independent test set	ML methods, independent test set, or cross-validation	ML methods, no independent test set	Non-LR, Non-ML	Subtotal
Data set not relevant anymore—pre-ICSI, pre-vitrification, day 3 ETs	**8 reports:** Nayudu et al,[46] Hughes et al,[47] Stolwijk et al,[48] Templeton et al,[49] Commenges-Ducos et al,[50] Minaretzis et al,[51] Hunault et al,[52] Ferlitsch et al,[53]	**3 reports:** Bancsi et al,[57] Jones et al (HFEA data 1991–1998),[58] Nelson and Lawlor (HFEA data 2003–2007)[59]			**2 reports:** Stolwijk et al,[61] Lintsen et al[62]	13
Research objective—clinical research and/or technology testing including use of IVF or embryo data	**1 report:** Lehert et al[54]	**3 reports using clinical pregnancies:** Carrera-Rotllan et al,[67] van Loendersloot et al, [68] Zhang et al[69] **4 reports using live birth outcomes:** Vogiatzi et al,[70] Gao et al,[71] Gong et al,[72] Wu et al[73]	**3 reports using +BHCG outcome:** Hassan et al,[74] Barreto et al,[75] Xu et al[44] **5 reports using clinical pregnancies only as outcomes:** Wen et al,[76] Mehrjerd et al,[77] Wang et al,[78] Fu et al,[79] Yang et al[80] **1 report using LB outcomes:** Goyal et al[81]	**1 report:** Vaegter et al. 2017.[60]	**1 report:** Grzegorczyk-Martin et al[63]	19
Pre-treatment counseling, pre-1st IVF cycle only	**2 reports:** Guvenire et al,[55] Metello et al[56]	**1 report:** Dhillon et al[82]	**4 reports:** Qiu et al,[85] Choi et al,[38] Nelson et al,[37] Nguyen et al[86]			7
Pretreatment counseling, pre-1st IVF cycle, after failed IVF cycle or other scenarios	2 reports: Luke et al,[88] McLernon et al[43]	2 reports: McLernon et al,[83] validated separately; Ratna et al[84]	3 reports: Banerjee et al,[6,36] Liu et al,[87] Cai et al[45]			7
Pretreatment counseling, after failed IVF cycle only		**1 report:** La Marca et al[65]				1
eSET counseling		**2 reports:** Ottosen et al,[64] Roberts et al[66]	**1 report:** Lannon et al[39]			3
Subtotal	**13**	**16**	**17**	**1**	**3**	**50**

Table 3 Summary of 12 pretreatment IVF prediction models (or sets of models), dataset, country of origin, size, contemporaneity of test sets, age limits, training method, and model validation metrics

Reference	Data source country of origin	Data source	Data source time period (C1 only)	Data set size	Independent test set	Age limits?	Training method	Model CV or validation in independent test set	Known clinical usage
Group 1. LR center-specific (LR-CS) models									
Dhillon et al[82]	UK	12 sites from one IVF network	2008–2012 training data: 9,915 IVF patients 2013 test data: 2,723 IVF patients			Not mentioned	LR	2013 test set: AUC 0.62 (0.60–0.64)	n/a
Group 2. LR multicenter model									
Luke et al[88]	US	US SART national registry database	Jan 2010–Dec 2016 with FETs to Dec 31, 2017	288,161 IVF patients	Test set was not specified.	18–59 y old	LR update of an earlier model	Model metrics such as AUC were not reported	This model supported a version of the SART online calculator, which is now retired
McLernon et al[43]	US	US SART national registry database	IVF cycles started in 2014–2015, tracked FET outcomes to end of 2016	88,614 IVF patients, 121,561 IVF cycles	Test set was not specified.	18–50 y old	Linear regression	From model training: AUC 0.73 with AMH, AUC 0.71 without AMH	This model supports the current free live SART calculator: https://w3.abdn.ac.uk/clsm/SARTIVF/
McLernon et al[83]	UK	UK HFEA national registry database	1999–2008, FETs and outcomes followed to 2009	113,873 IVF patients, 184,269 IVF cycles	See Ratna et al.[84] See Leijdedkkers et al, 155[b]	Not specified	Discrete time LR	From model training (validation not specified): AUC 0.73 (0.72–0.74)	This model supports the current free live OPIS2 calculator: https://w3.abdn.ac.uk/clsm/opis
Ratna et al[84]	UK	UK HFEA national registry database	Jan 2010- Dec 2016 with FETs to Dec 31, 2017 used as test set	91,035 women, 144,734 IVF cycles	Updated model was not further tested using independent test set.	18–50 y old	LR and recalibration of the McLernon et al model	From model recalibration with no further validation using a separate test set: AUC 0.67 (0.66–0.68)	Same as McLernon et al2016[83]
Group 3. ML center-specific (MLCS) models									
Banerjee et al[36,a]	US	Single center	Training: 2003–2006; test: 2007–2008	Training: 1,676 C1s, test: 643 C1s	Yes: out-of-time, exclusive of training data	Excluded age ≥ 43 from test set	MLCS—GBM	AUC 0.80 versus age control AUC 0.68 (15% improvement); reclassified 83%.	Prototype predating Nguyen et al[86]
Nelson et al[37,a]	UK	Single center	Training: 2006–2010, test: 2011–2012	Training: 2,124 IVF cycles, test: 1,121 IVF cycles	Yes: out-of-time, exclusive of training data	Excluded age > 45 from training or test	MLCS—GBM	AUC 0.716, 6.3% imp over age (0.674), PLORA 29.1 (76.2% improvement), reclassified: 61% higher, 14% lower	Prototype predating Nguyen et al[86]
Nguyen et al[86,a]	US	6 single centers	6 separate datasets, 2013–2022	Dataset sizes range from 200 to 2000 IVF cycles	v1 models: cross-validation (CV) and out-of-time test set exclusive of CV and training data; v2 models: CV	Excluded age ≥ 42 from training and testing; clinical use excluded age ≥ 40 and used a separate model for age ≥ 40	MLCS—GBM and methods as per Banerjee et al., 2010[30] and Nelson et al., 2015.[31]	*Manuscript submitted*	Commercially available to fertility centers (as software-as-a-subscription, SaaS product) as the Univfy PreIVF Report[a]

(*Continued*)

Table 3 (*Continued*)

Reference	Data source country of origin	Data source	Data source time period (C1 only)	Data set size	Independent test set	Age limits?	Training method	Model CV or validation in independent test set	Known clinical usage
Reference	Data source country of origin	Data source	Data source time period (C1 only)	Data set size	Independent test set	Age limits?	Training method	Model CV or validation in independent test set	Known clinical usage
Group 3. ML center-specific (MLCS) models (continued)									
Qiu et al[85]	China	Single center	2014–2018	7,188 first IVF cycles	Training on 70% and testing on 30% of data	Not specified	LR, RF, SVM, XGBoost	Validation on test set (AUC); nested CV x 11 (average accuracy score): LR: AUC 0.71, avg. accuracy 0.68 RF: AUC 0.73, avg. accuracy 0.69 SVM: AUC 0.71, avg. accuracy 0.68 XGBoost: AUC 0.73, avg. accuracy 0.70	n/a
Liu et al[87]	China	Single center	2019–2021	1,857 IVF cycles	2019–2020 data: 80% training, 20% validation; 2021 data: out-of-time testing	20–45 y old	LR, RF, XGBoost, LGBM	2021 test set (similar to validation): LR: AUC 0.645 (0.521, 0.769) RF: AUC 0.641 (0.516, 0.766) XGBoost: AUC 0.644 (0.521, 0.768) LGBM: AUC 0.634 (0.511, 0.758)	n/a
Cai et al[45]	China	Single center	Jan 2013–Dec 2020	26,382 IVF patients	Training 2013–2019; test 2020	20–48 y old, see S9	Training 2013–2019; test 2020		n/a
	Recalibration of McLernon 2016 model (US SART online calculator): All ages: AUC 0.69 (0.68–0.69); <35: AUC 0.58 (0.57, 0.58); ≥35: AUC 0.70 (0.69, 0.71) Recalibration of McLernon 2022 model (UK online calculator): All ages: AUC 0.68 (0.67–0.68); <35: AUC 0.58 (0.57, 0.58); ≥35: AUC 0.67 (0.66, 0.68) Overall, MLCS using endocrinology laboratory values: All ages: AUC 0.74 (0.74, 0.75); <35: AUC 0.68 (0.67, 0.68); ≥35: AUC 0.74 (0.73, 0.75) XGboost: AUC 0.83 (0.83, 0.84); Lasso: AUC 0.75 (0.74, 0.75); GLM: AUC 0.75 (0.74, 0.75)								
Group 4. ML multicenter model validated for each center									
Choi et al[38]	US, Canada, Spain	3 centers, validated for each center	2008–2009	Training 1,061 first IVF cycles; testing: 1,058 first IVF cycles; sampled from a total of 13,076 first IVF cycles		Excluded age ≥ 43 y old	Multicenter ML model trained from blending and weighting model components extracted from 3 center-specific models: AUC 0.634, PLORA = 9.0 Prediction errors ranged from −3.7 to 0.9%		Prototype, commercially available upon request

Abbreviations: AUC, area-under-the-curve of receiver operating characteristic curve; GBM, gradient boosted machine; GLM, generalized linear model; LGBM, light gradient boosted machine; LR, logistic regression; MLCS, machine learning, center-specific model; PLORA, posterior log of odds ratio compared to age control; RF, random forest; SVM, support vector machine.

[a]U.S. Patents including U.S. Patent Number 9,458,495B2; foreign counterparts; and patents issued.

[b]A separate study by Leijdekkers et al, 2018 (155), performed external validation of the McLernon IVF pretreatment model and subsequently updated/recalibrated the model to correct for slight overestimation. External validation of the recalibrated model was not specified.

a dataset with known live birth outcomes; (3) validation of the model using an independent test set or cross-validation. In addition, we include models with known clinical usage even if they do not satisfy all these criteria (a subset highlighted in green, ►**Table 2**).[43,88] However, we did not include models that have not been reported in peer-reviewed research literature.

Data Source, Time Period, Clinical Variables, and Outcomes

The IVF data source may be a single center, group of multiple centers, or national registry database. The choice depends on whether the model usage will be limited to one center, one group, or applied nationally. For optimal results, the training and test data should be representative of the same patient population. Although some researchers advocate testing whether an established model can be adapted and applied elsewhere, if feasible a center-specific model may yield better model performance metrics.[43,45,84]

Since the model applicability may be heavily affected by the time period chosen, data should be from a more recent, shorter time period than to use more years of data. For example, intracytoplasmic sperm injection (ICSI) usage and other innovations improving embryology outcomes (e.g., extended embryo culture, blastocyst vitrification, and elective single embryo transfer [eSET]) became widely adopted in the mid-late 1990s and mid-late 2010s, respectively.[89–93] The center-specific implementation dates of those technologies should also be considered. For example, in recent years, freeze-all followed by serial transfers of single cryopreserved-thawed blastocysts until live birth is reached is increasingly applied when possible. Therefore, the selection of dataset parameters including time periods may affect the appropriate clinical usage of the resulting model.

The relative usefulness of clinical predictors in each model depends on which other clinical predictors are being used, since clinical variables often have overlapping or redundant contributions to LBP models in IVF. Here, we use age and AMH to demonstrate a few practical points, but these concepts can be extrapolated to other clinical predictors as well. For example, if age and AMH are available for most IVF cycles in a dataset, then the relative importance of age and AMH in the resulting model may reflect their true relative contribution to LBP. However, if AMH is available only in a third of the sample, then the resulting model may still perform very well and may best serve that center's own patients even though the model may rely on age more than AMH. Compounding the above is that the relative weighting and scoring of AMH and age are expected to vary depending on the clinical profiles of patients seen locally at each center. Therefore, comments such as "the coefficient/relative importance of AMH is such and such in predicting IVF LBP" should be qualified by the specific patient population, time period, and other practice contexts.

On the topic of clinical outcome to be modeled, model design requirements are distinct from the ongoing controversy in the literature on whether to use live birth or clinical ongoing pregnancy (COP; defined as reaching 12 weeks of gestation with documented fetal cardiac activity on ultrasound) as primary outcomes in clinical trials.[94–96] While we do agree with the goal of using live birth outcome as the outcome to be predicted, we caution that when using binary outcomes (e.g., yes or no, positive or negative), restricting the "positive class" to only live birth necessitates labeling COP as negative class. Treating COP as "no live birth" may compromise model performance and clinical applications because in layman's term, that would be "inaccurate" as only an estimated 5% of COPs do not result in live birth.[94]

Using an example of 200 first IVF cycles in a dataset, let us say that 100 cycles have documented live birth outcomes and 20 cycles have documented COP outcomes. Let us assume that 5% of those 20 cycles (1 cycle) later are found to have ended as second or third-trimester pregnancy losses. By restricting positive outcomes to live births only, the positive class is 50% of the dataset. By applying positive outcomes to both live births and COP outcomes, the positive class is 60%. Later, when the fates of all COPs are confirmed, the positive class is 59%. A model trained using 60% of the data as a positive class is much closer to the truth (59%) than assuming that only 50% of the data has a positive class.

Furthermore, it is not only the positive class rate that would be vastly underestimated but the bigger problem is that the model would be trained based on incorrect relationships between predictors and outcomes. You may ask, "Then why don't we wait until all the COPs have been followed up and we can use a dataset with confirmed final live birth vs. no live birth outcomes confirmed?" The answer to that question is explained in the section "Addressing the Risk of Data Drift**"** later in this article. It may be helpful to note that the objective of a clinical prediction model for patient counseling is vastly different from clinical trials aiming to determine the efficacy and safety of a clinical intervention. Therefore, although the design of clinically practicable prediction models relies on the application of expert clinical research and modeling knowledge, it has requirements and best practices different from conventional clinical trial design.

Last but not least, we should mention that the quality of the data preprocessing and processing steps are critical to successful validation of any prediction models, though they are typically given the least amount of attention in the literature. Establishing and adhering to rules and logic in the data pipeline and frequent code reviews and updates are important tasks in maintaining top-quality data for modeling.

Model Training: Why Machine Learning?

It is important to remain method-agnostic and open-minded to evaluate the benefits and limitations associated with each model training technique. Referring back to ►**Table 3**, Groups 1 and 2 comprise models trained using multivariate LR or simply LR. LR has been in popular usage since circa 1970 for testing and modeling the contribution of several factors in influencing scenarios with binary outcomes.[97–100] However, LR is an early form of ML predating current ML used in

medical research, and it has often been erroneously perceived as the antithesis of ML. While LR is a very important statistical and modeling technique in medical research, it is limited in handling missing data and analyzing continuous and discrete variables, highly correlated variables, nonlinear relationships, and imbalanced data (e.g., low frequency of the positive outcome such as very low live birth rates in older patients).[101–103] Nevertheless, there are established data-processing steps (e.g., imputation, transformation of data) that can modify the data for LR modeling.[45,84]

►**Table 3** (Group 3 publications) reported ML usage. Model training techniques such as LR, generalized linear regression, and various "ML" methods such as Extreme Gradient Boosting (XGB), Lasso, random forest, and gradient boosted machine (GBM) in conjunction with methods to impute missing value, utilize continuous and discrete variables, and perform feature selection as needed have been described.[104–111] See ►**Table 4** for a brief description of commonly used ML techniques and concepts. In contrast to a common misconception, many ML techniques originated decades ago, but medical researchers were not able to take advantage of those methods until the speed and capabilities of personal computers, cloud computing, and data storage became widely available and economical. Although in the earlier years, our attempts to publish research utilizing ML met with significant resistance from reviewers asking us to rationalize the use of ML over conventional LR, the benefits of ML are widely appreciated today.

Many have asked, "what additional benefit is conferred by ML over LR used in conventional medical research?" In the general case, not relating to IVF specifically, the main advantage of ML is its scalability, reproducibility (in terms of repeating the analysis on updated data), and improvements in model performance made possible by its scalability and reproducibility. Due to the prevalence of ML usage across industries, as a discipline, the ML community has established best practices to help make the most meaning from data. Therefore, when using ML and adhering to best practices, we benefit from the collective expertise and knowledge of all experts using data for all applications globally.

For example, for an individual researcher analyzing data sourced from their own center, depending on the dataset attributes, it is possible to add functions such as imputing missing values, adaptations to allow the analysis of both continuous and discrete variables, testing and optimizing tree nodes, to curate an LR model to achieve similar performance as GBM. Also, as a dataset's heterogeneity, features, and sample size decrease, the model performance achieved by LR or other multivariate models and ML may be comparable. However, the above approach would require each center to have its own dedicated researchers and the application of many different unique customizations may also make it more challenging to discern generalizable from center-specific findings.

The scalability and applicability of ML allow the same techniques to be applied rigorously and reproducibly to datasets from many different centers, enabling observation of data and model nuances and establishment of best practice. In other words, with ML, it is possible to re-apply a validated process to different IVF datasets to create center-specific, validated models. This repeatable process levels the playing field for centers varying in size and resources. Other important benefits include the relative ease of updating a model with an updated dataset and the training of a center-specific prediction model or applying center-specific validation of a general, multicenter model.

When discerning the choice of one ML technique over another, the objective measure of model performance using established metrics allows productive scientific discourse, provided researchers understand the advantages, indications for use, and pitfalls of model metrics as discussed later in the section "Model Validation, Training, and Test Sets." Model

Table 4 Commonly used machine learning techniques and concepts

Concept	Description	Example	Reference
Logistic regression	Predicts binary outcomes (yes/no) using input features	Determining the likelihood of pregnancy after IVF	Cox[106]
LASSO	Selects important features and reduces overfitting	Identifying key factors influencing IVF success	Tibshirani[110]
Supervised machine learning	Uses labeled data to predict outcomes	Predicting the success rate of IVF treatments based on patient data	Mitchell[104,150]
Unsupervised machine learning	Finds patterns in unlabeled data	Grouping patients based on ovarian response patterns	Hastie et al[105,150]
Gradient boosting machine (GBM)	Combines multiple weak models to make better predictions	Predicting embryo implantation success	Friedman[107]
Random forests	Uses many decision trees to improve predictions	Predicting patient response to fertility treatments	Breiman[111]
XGBoost	An optimized version of GBM, faster, and more accurate	Advanced models for predicting IVF outcomes	Chen and Guestrin[108]
LightGBM (LGBM)	A faster version of GBM using less memory	Efficiently predicting fertility treatment results	Ke et al[109]

metrics such as receiver operating characteristics (ROC) area under the curve (AUC) should not be compared across publications that differed in their patient populations and other dataset attributes. However, publications that compared performance metrics across modeling techniques applied to the same training and test data are informative. Specifically, Qiu et al, Liu et al, and Cai et al tested four to six ML techniques[45,85,87] (►**Table 3**). Taken together, their results suggested that XGB and Light Gradient Boosting Machine (LGBM) appeared to perform well consistently. Our own research group has found gradient boosting machine (GBM), a comparable implementation to XGB and LGBM, to perform well consistently on datasets from over 50 global and U.S. fertility centers in our client services work and research collaborations as exemplified by Banerjee et al, Choi et al, Nelson et al, and Nguyen et al (submitted).[36–38,86]

Pretreatment models have primarily used structured data, unlike unstructured data such as imaging data used to rank embryos' viability. There may be a misperception among some that the ML techniques used to analyze structured data are not as "advanced" as artificial neural networks (ANN) and deep learning techniques used to analyze unstructured imaging or real-time, data used in other areas of medicine such as diagnosing arrhythmia or screening for breast cancer.[112,113] It is best to be objective and evaluate the ML technique appropriate for each application and corresponding dataset attributes. Consider factors such as economics relating to cloud computing resources, turnaround time, expertise, and potential benefits, pitfalls, or biases when choosing the ML method. Consistent with our experience, Chen et al reported that deep learning does not frequently improve model performance for datasets comprising structured data.[114]

Model Validation, Training, and Test Sets

Over a decade ago, research publications reporting prognostic modeling often omitted model validation.[115] It is now recognized that an unvalidated prediction model lacks credibility. The detailed technical methods of model validation are beyond the scope of this article.[36,37,83,84,116–119] Thus, we highlight a few important points to enable readers to more critically appraise published models.

Clinicians may often use "external validation" when referring to testing a model established using patient data from one location to patients at a different location to determine if the findings are generalizable. In contrast, for ML external validation can also refer to data from a different time period at the same location or mutually exclusive test data sourced from the same location and the same time period.[116,118] Data scientists may specify that allocation of data to training and test sets must be random yet matching for certain clinical attributes, much like matching cases with controls in conventional case–control studies; hence, there is an advantage to use data from the same location. A separate yet important concept is that the overall frequency of the positive call (commonly live birth and/or COP, for IVF prediction) also determines which model validation metrics should be used. For example, if the live birth rate is fairly low in a dataset—approximately 30% or less such as in the case of IVF live birth outcomes for patients 42 years or older using their own eggs—that dataset would typically be considered imbalanced, imposing certain ML techniques and model validation metrics to be used over other methods that may be inapplicable or result in misleading results.

Ideally, ML IVF live birth prediction models are trained and tested using mutually exclusive datasets that are balanced and matching from the same time period (in-time test set), with further testing on an exclusive test set from a time period immediately preceding model deployment (out-of-time relative to the original training and test data), presumed to be most representative of patients being counseled using the deployed model. Further model validation using a test set comprising data of patients being counseled using the deployed model is important to demonstrate that the model does indeed apply to the patients being counseled. The latter model validation could be considered "live model validation (LMV)" demonstrating that the "live" model holds true for current patients. Model validation metrics may be affected by dataset size to different extents. For example, one would ideally want to maximize the size of both the training and test datasets, but in situations where the training and test sets are allocated from the same dataset, increasing the size of one means decreasing the size of the other. Various strategies can be applied to optimize training and test dataset size to maximize model validation results, but attributes and nuances of each type of dataset or even the patient population may determine the choice of training versus test set allocation strategy.

The medical research literature frequently omits a description of the dataset that is ultimately used for model deployment or "production model." Specifically, after model validation confirms that all the steps—data processing, feature testing, training, and validation—have come together to produce a validated model, it is often best practice to deploy a model comprising both training and test data because that model is expected to have the best model performance. However, such a model cannot be further validated until post-deployment data become available for LMV. This ML best practice is conceptually different from the conventional research process.

Some providers are concerned that they should "hold off" from using a model until a test set comprising patients seen currently is available for validation, despite excellent model validation having been demonstrated using historical data from as recent as 1 year ago. In the time that the current patients' IVF treatments can be aggregated into a "current" test set, the originally trained model would actually have become older, even though it has been further tested. In parallel, there may be less hesitation to omit the model validation step altogether because the lack of any validation seems to render the model "evergreen" and timeless. These fallacies are rooted in the well-warranted perception of the risk of data drift. Understanding the risk of data drift helps providers balance theoretical risk and practical benefits.

Addressing the Risk of Data Drift

Models trained and validated on historical data risk not being true for patients treated today or in the future. Such a risk, called data drift, can occur through input data drift, concept drift, and clinical context-related data drift, though these data drift subtypes can affect one another.[120,121] Input data drift includes changes in diagnostic labeling (e.g., documenting polycystic ovarian syndrome as an ovulatory disorder rather than the more specific PCOS diagnosis), demographic changes (e.g., having an increase in the proportion of women younger than 35 years). Data drift related to clinical context of use includes changes in patient management and disease prevalence (e.g., a center had recommended IVF as a first-line treatment to patients with unexplained infertility but now changes to recommend IUI as first-line treatment instead; thus, IVF patients with unexplained infertility will now be patients who already failed at least one to two IUI treatments and the IVF live birth outcomes for this altered patient group may be lower than observed in the historical data). Concept drift, a consequence of input data drift or clinical context-related data drift, refers to a changed relationship between predictor and live birth outcome (e.g., in the previous example, the concept drift is that unexplained infertility as a predictor may now be associated with a lower IVF live birth probability due to different clinical management of those patients). Fortunately, providers tend to be conservative; so, major changes in protocols or treatment recommendations tend not to occur over a short period of time.

Having validated plans for monitoring, reducing, and addressing data drift if needed can help prevent provider data drift concerns. First, providers should understand that healthcare practitioners are only expected to give patients "the best available current prognostic information." In the absence of a validated model, it is more difficult to speak to the quality of the prognostic information. Providers have a simple task of alerting model makers if they make significant treatment changes or if there is a change in the patients treated. In addition, we recommend updating the IVF live birth prediction model using the latest available data every 2 to 3 years or sooner if there are changes to patient demographics or treatment. At the time of model update, we also perform "LMV," validation of the deployed, "live" model using the more recent, and out-of-time data as a test set. Despite data drift deserving caution, fertility centers are typically extremely cautious and avoid making sudden, significant changes to protocols. We have confirmed LMV for fertility centers that have requested it; further, based on a formal reporting of LMV for a sample of six fertility centers, we advocate testing or providing LMV for a larger sample of centers to determine whether this observation can be generalized.[86]

National Registry-Based Online Calculators versus Center-Specific Prediction Models

For providers in countries that do not have a national registry-based online calculator, McLernon et al recommended performing a series of statistical testing, recalibration, and adaptations of the LR models produced using US SART or UK HFEA data by McLernon et al (US SART) and Ratna et al (UK HFEA), respectively.[43,83,84,122,123] However, Cai et al challenged this recommendation by showing that several MLCS models developed de novo using their own center's data, outperformed the US SART and UK HFEA models based on cross-validation results. In addition, Cai et al showed that the US SART- and UK HFEA-adapted models gave poorer validation results for patients younger than 35 years compared with patients 35 years old or older.[43,45,83,84]

When using a multicenter model, it is important to understand whether there are variations in patients across centers and if so to quantify this variability. Using age–AMH–ovulatory disorder diagnosis as a multivariate measure of clinical profiles, Swanson et al reported that inter-center variation of clinical profiles is quantifiable and correlates to live birth outcomes.[124] Specifically, five distinct clinical profiles were demonstrated in 7,742 patients who received IVF treatment from 9 North American centers located in 33 cities across 11 U.S. states and Ontario, Canada. The proportion of patients having each of five distinct clinical profiles varied significantly across centers and the odds of having an IVF live birth varied across these profiles.[124] Also variations in local IVF treatment may contribute toward intercenter variations.

Despite the perception that larger datasets enable more generalizable prediction models, it is important to consider the applicability of models for specific patients in specific centers. In other areas of medicine, there have been varying levels of success in applying ML to clinical registry data, with reports of success in producing clinically applicable models and registry data being inadequate for maximizing the utility of ML.[125–130] National IVF registries were designed for monitoring outcomes and safety, not for supporting individualized prognostication.

National Registry Models and MLCS in the United Kingdom and the United States

In ►**Table 3** (Group 3), the models reported by Banerjee et al and Nelson et al were early prototypes followed by the training and validation by our group of over 50 MLCS models many of which have also been deployed for clinical utilization or to provide operational insights to individual fertility centers. We previously reported that a validated, center-specific ML model computes personalized IVF live birth probabilities with improved discrimination, dynamic range, and posterior log of odds ratio compared to age control models with a significant percentage of patients having higher live birth probabilities than would have been expected based on age alone.[2,36,37]

Even controlling for the data source country (the United States, the United Kingdom), the U.S. and UK national registry based-models (the McLernon models) are different from the MLCS models reported by Banerjee et al (the United States) and Nelson et al (the United Kingdom) in many ways—data source, time period, single versus multiple centers, age limit, and AMH availability in the IVF cycles used, training using LR versus GBM, and validation metrics.[36,37,43,83,84] The

key observation is that MLCS modeling using far fewer IVF cycles achieved comparable or better model performance compared with national registry-based models (►**Table 3**).

Possible Reasons Limiting the Performance of National Registry Models

Exploring possible causes for differences in multicenter versus single-center model performance may inform research and model improvement efforts. First, Ratna et al acknowledged that the UK McLernon model suffered from lack of AMH values in the HFEA data.[84,122] Second, in our experience when patients aged 42+ years are included in the training and test sets, the ROC AUC which reflects whole-model performance may not be representative of model performance for younger women as older women have a disproportionate number of true negatives. Consistent with our experience, Cai et al showed when applying the McLernon models to the Chinese dataset, and the McLernon models performed better in the older age group ($\geq$35 years) than in the younger age group ($<$35 years).[45]

When using multicenter data, one must control contributions to the training and test sets by each center, to avoid having a model that is overrepresented by centers with large volumes. Potential solutions to be tested include sampling and controlling for each center's proportional contribution to the final training and test sets or creating a center label to represent center-specific factors that are not captured by variables in the dataset. For example, Choi et al showed a method for each center to contribute predictive elements or trees rather than IVF cycles to make up the multicenter dataset.[38]

Model Validation

Model review and validation are crucial steps before applying models in clinical practice to demonstrate performance, thereby building trust in model accuracy. Model review involves monitoring descriptive and analytical statistics for various variables, with experts reviewing any irregularities to detect and resolve issues throughout the modeling process.

In this section and in ►**Table 5**, we provide an overview of important model metrics with references for further details. Regardless of the metric used, there should be a control model for comparison.

The AUC of the ROC may be the most widely reported model metric. The ROC AUC measures the ability of the model to discriminate or rank predictions showing the trade-off between the true positive rate (TPR) and false-positive rate (FPR).[116–119] While the ROC AUC measures a model's ability to rank predictions, it has significant drawbacks. For instance, it may not detect clinically meaningful improvements in the model. Moreover, the AUC can be artificially inflated by including specific patient groups, such as those older than 42 years or those with very low live birth rates, giving a false sense of reassurance about the model's performance. This metric is also not suitable for highly imbalanced datasets.

To address limitations of ROC AUC, additional metrics that measure different attributes of the model can be considered (►**Table 5**). In particular, we created the metric, the Posterior Log of Odds Ratio Compared with Age Model (PLORA), for measuring predictive power in the specific context of IVF LBP.[2,36–38,86] PLORA compares the log-likelihood of the IVF prediction model to an age-based control model: that is, "how much more likely will this new model fit the observed data and outcomes compared to the age control model?" This metric is sensitive to dataset size and model improvements and can be communicated in linear scale (e^{PLORA}) for easier understanding by clinicians. Observing a positive PLORA in conjunction with other model metrics provides a comprehensive indication of model performance.

In addition to ROC AUC and PLORA, we employ other important metrics such as IVF LBP distribution, reclassification, and dynamic range to further evaluate models.[131–133] Reclassification examines whether more patients as a group receive higher live birth probabilities better reflecting actual live birth rates, while the dynamic range evaluates the highest and lowest possible live birth probabilities that the model may predict. These metrics provide insight into the strengths and limitations differentiating candidate models.[36–38,45,85–87] Precision, Recall, F1 Score, and PR AUC are also effective at detecting improvements in predicting positive live birth outcomes, especially in imbalanced datasets.[116,117,134] Precision, or positive predictive value, indicates the likelihood that predicted positive outcomes are correct, ensuring patients are not misled about their chances. Recall, or sensitivity, measures the model's ability to identify actual positive outcomes, ensuring that patients with high live birth probabilities are accurately identified. The F1 Score (i.e., the harmonic mean of precision and recall) balances both metrics to provide a comprehensive evaluation of the model's performance. The PR AUC plots precision versus recall without requiring a specific threshold and offers detailed views of the model's predictive capabilities across different thresholds.

These metrics help demonstrate a model's ability to support clinical care and business operations. By leveraging a combination of validation metrics, we can provide more reliable prognostics, ultimately improving clinical decision-making and operational efficiency.

Additional Requirements to Use Validated ML Models in Routine Clinical Care

Although many ML models have been reported in many areas of medicine, there are additional requirements for successful implementation in clinical care as summarized by patient-centric communications; provider collaboration, usability, and explainability; relevance and model performance; ability to handle complex data; scalability including maintenance of quality, user experience, and economic feasibility as usage scales; and adhering to best practice and compliance throughout the product life cycle from raw data processing to production model deployment.[119,135,136]

Table 5 Model performance metrics that are commonly used and/or useful in discerning models varying in performance

Metric	Measurement	Value and interpretation	
Receiver-operating curve area-under-the-curve (ROC AUC)	Measures the ability of the model to discriminate or rank predictions	AUC = 1	The model is a perfect classifier with a maximum true positive rate (TPR) and a minimum false positive rate (FPR)
		AUC $\leq$ 0.5	The model is a poor classifier, no better than random
Posterior log of odds ratio compared with age model (PLORA)	Measures predictive power, comparing the log-likelihood of the IVF prediction model to the age control model	High (positive) PLORA	The IVF prediction model is more effective in predicting successful IVF live birth outcomes than the age-based model
		Low (negative) PLORA	The IVF prediction model is less effective in predicting successful IVF live birth outcomes than the age-based model
Precision/Positive predictive value	Evaluates the model's tendency to overestimate the probability of live birth	High precision	When the model predicts a successful IVF live birth outcome, it is more likely to be correct
		Low precision	A significant proportion of the model's predictions of successful IVF live birth outcomes is likely incorrect (false positives)
Recall/Sensitivity/TPR	Measures the proportion of actual positives (successful IVF live birth outcomes) that are correctly identified by the model	High sensitivity	The model correctly identifies a large proportion of successful IVF live birth outcomes
		Low sensitivity	The model misses a significant number of successful IVF live birth outcomes
F1 score	Measures the harmonic mean of precision and recall	High F1 score	The model has both high precision and high recall given a particular classification threshold
		Low F1 score	The model has low precision, low recall, or low precision and low recall given a particular classification threshold
Precision-recall area-under-the-curve (PR AUC)	Measures the model's overall performance in terms of precision and recall across all thresholds	High PR AUC	The model maintains high precision and high recall across the range of possible classification thresholds
		Low PR AUC	The model struggles to maintain both high precision and high recall across the range of possible classification thresholds

Explainability, Provider Collaboration, and Patient Communications

An IVF prognostic model supports provider–patient relationship at a critical point of the patient's care; therefore, the IVF prognostic information must be clear and easy to understand to both providers and patients. The prognostic information may be presented in an individualized counseling report that includes not only the prognostic information but also the key factors that underpin the prognosis in a graphical format illustrating how the patient compares to other patients treated at the same center.[23]

Scalability, Data Privacy, Compliance, Ethics

In the context of ML-supported provider–patient prognostic counseling, scalability refers to the ability of the ML platform to serve patients through many fertility centers and providers globally with great implementation and model update efficiency at low costs while preserving or improving the quality of the IVF prediction models, the counseling reports, and other supportive services. Scalability is important as it enables the delivery of prognostic information to diverse patient populations. Scalability may be achieved in several ways. For instance, a proprietary, end-to-end platform may

be used to support data model pipelines, no-code implementation of customized model and counseling report specs, deployment and usage of models, report generation, multilingual function, and administrative module. However, the platform and related processes must comply with applicable local data privacy laws such as U.S. Health Insurance Portability and Accountability Act (HIPAA) and the European General Data Protection Regulation (EU GDPR). The regulatory framework and pathways governing medical devices comprising AI/ML and software have continued to evolve with the increasing complexity of the devices and needs of patients and providers.[119] Beyond compliance, it is important to conduct 360-degree review with key stakeholders including collaborators, fertility center, clinical and operational leads and teams, internal team, and patients (directly or indirectly via providers) to consider potential unintended consequences of data strategy, product, and life cycle management decisions including efforts to maximize inclusion of diverse patient groups, ensuring that the data are representative of the people served by the resulting model, and ongoing efforts to maximize affordability of and access to fertility treatments.[4,5]

Adaptability and Operational Efficiency

Successful application of ML at point-of-care requires the merging of clinical care and data/ML workflow in a streamlined way for the clinical team. In view of staff constraints, it is imperative that the streamlining is operationally and efficiently tailored to each center.[137] Many fertility centers have been scaling up services to meet increasing patients' demands for IVF. In the United States, care teams have been augmented by training advanced practice providers (APPs) and/or general obstetricians and gynecologists.[137–139] It is essential that all users are trained on the use of any prognostic tool, but a good tool can support their patient counseling. Additionally, optional integration with fertility centers' electronic health record systems enable largely automated generation of counseling reports.

Real-World Usage, Tracking, and Evaluation

Despite reports of IVF live birth prediction models and online calculators for providers and patients, there are relatively few reports on such real-world use.[43,83,84,140–142] One study compared a group of Australian patients' perceived IVF LBPs against the U.S. SART calculator and another study compared a group of French patients' perceived IVF LBPs against the French registry data.[29,30]

Limited reporting of IVF prognostic tool usage may reflect a variety of challenges such as difficulty in assembling the necessary expertise, limited AI real-world studies, or difficulty in fitting such studies into conventional medical research journals. We have reported clinical implementation of IVF live birth prediction models and our team's multisite implementation experience using the framework recommended by Goldstein et al.[2,86,137,143] A real-world study of the retrospective multicenter experience of 24,238 new IVF patients suggested that usage of a patient-centric, MLCS-based prognostic report was associated with increased IVF conversion among new fertility patients.[143] This study suggested to investigate factors influencing treatment decision-making and real-world optimization of patient-centric workflows incorporating MLCS prognostic reports.[143]

Medical research journals might consider a category dedicated to AI applications encouraging AI publications pertaining to medicine. In addition, it may be helpful to support reviewers with articles on the topic of AI implementation, usage, and guidelines. Indeed, the following guidelines facilitate informed and productive review processes. TRIPOD-AI, STARD-AI, PRISMA-AI, CONSORT-AI, and SPIRIT-AI are extensions of widely used TRIPOD, STARD, PRISM, CONSORT, and SPIRIT guidelines, respectively.[144–149] Since the original, non-AI versions of these guidelines have been widely used by researchers, the AI versions have also become widely recognized by journals.

Unlike other guidelines, the DECIDE-AI guideline was designed for early clinical study stage evaluation of any AI modalities (e.g., diagnostic, prognostic, therapeutic), in live clinical settings and importantly does not require any one study design.[149] The DECIDE-AI guideline prioritizes assessing the risk of data shift and reporting of clinical implementation experience as-is to expedite sharing of the usage experience.[149]

Conclusions

Having described the design, development, validation, and deployment of personalized ML IVF prognostic models, it may be helpful to return to the broader vision of advancing reproductive medicine and increasing fertility care accessibility. Economic modeling using ML IVF prognostic models can inform the allocation of funding to support fertility care with strategies at the local, regional, or national levels. Most importantly, these local strategies are aligned with a global, scientific and ethical approach adaptable to local fertility centers' clinical care and operational needs. A global collaboration of public, private, research, and operational groups developing validated ML IVF prognostic models help to prioritize women's and couples' family-building goals. By helping more people who are proactively seeking fertility care to become parents, we may also help to mitigate the macro level impact of below replacement fertility currently experienced by more than half of all countries.

Authorship Contribution Statement
M.W.M.Y.: writing—original draft, writing—review and editing, conceptualization; J.J.: writing—original draft, writing—review and editing, conceptualization; E.T.N., T.S., M.M.: writing—review and editing.

Attestation Statement
Not applicable.

Data Sharing Statement
Not applicable.

Funding
Each organization funded its own participation.

Conflict of Interests
M.W.M.Y. is employed as CEO by Univfy Inc.; is board director, shareholder, and stock optionee of Univfy; receives payment from patent licensor (Stanford University); is inventor or coinventor on Univfy's issued and pending patents.
E.T.N., T.S., and M.M. are employed by and received stock options from Univfy Inc.
J.J. has no conflicts to declare.

Acknowledgments
The authors thank the following individuals for their assistance, editing, and/or contribution to our research and/or implementation programs: Vincent Kim, B.Sc.; Anjali Wignarajah, M.Sc.; Candice Ortego, RN; Wing H. Wong, PhD; Athena Wu.

References

1 Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023;388(13):1201–1208
2 Jenkins J, van der Poel S, Krüssel J, et al. Empathetic application of machine learning may address appropriate utilization of ART. Reprod Biomed Online 2020;41(04):573–577
3 Sample patient counseling report. Accessed July 18, 2024 at: https://www.univfy.com/research/sample-preivf-report
4 Radanliev P, Santos O, Brandon-Jones A, Joinson A. Ethics and responsible AI deployment. Front Artif Intell 2024;7:1377011
5 Drabiak K, Kyzer S, Nemov V, El Naqa I. AI and machine learning ethics, law, diversity, and global impact. Br J Radiol 2023;96 (1150):20220934
6 World Health Organization (WHO). 2023 1 in 6 people globally affected by infertility: WHO. Accessed on April 29, 2023 at: https://www.who.int/news/item/04-04-2023-1-in-6-people-globally-affected-by-infertility
7 Infertility and IVF Access in the United States. A Human Rights-Based Policy Approach. Center for Reproductive Rights. Accessed June 2, 2024 at: https://reproductiverights.org/our-issues/assisted-reproduction/
8 Practice Committee of the American Society for Reproductive Medicine (ASRM). Definition of Infertility: A Committee Opinion. 2023. Accessed June 2, 2024 at: https://www.asrm.org/practice-guidance/practice-committee-documents/denitions-of-infertility/
9 Society for Assisted Reproductive Technology (SART). 2022More than 73 thousand babies born from assisted reproductive technology cycles done in 2020. Accessed April 29, 2023 at: https://www.sart.org/news-and-publications/news-and-research/press-releases-and-bulletins/more-than-73-thousand-babies-born-from-assisted-reproductive-technology-cycles-done-in-2020/
10 European Society of Human Reproduction and Embryology (ESHRE). ART Factsheet. Accessed July 17, 2024 at: https://www.eshre.eu/Press-Room/Resources
11 International Committee Monitoring Assisted Reproductive Technologies (ICMART). 2018. ICMART Preliminary World Report. 2018. Accessed April 29, 2023 at: https://www.icmartivf.org/reports-publications/ - articles.
12 Family Equality. 2020Facts about LGBTQ+ Families - Fact Sheet. Accessed June 2, 2024 at: https://www.familyequality.org/resources/facts-about-lgbtq-families/
13 National Center for Health Statistics (NCHS). Centers for Disease Control and Prevention (CDC), 2022Key Statistics from the National Survey of Family Growth. Impaired Fecundity and Infertility Sections. Accessed April 29, 2023 at: https://www.cdc.gov/nchs/nsfg/key_statistics/i-keystat.*htm*
14 Ethics Committee of the American Society for Reproductive Medicine. Electronic address: asrm@asrm.org. Disparities in access to effective treatment for infertility in the United States: an Ethics Committee opinion. Fertil Steril 2021;116(01):54–63
15 Klitzman R. How much is a child worth? Providers' and patients' views and responses concerning ethical and policy challenges in paying for ART. PLoS One 2017;12(02):e0171939
16 Rich CW, Domar AD. Addressing the emotional barriers to access to reproductive care. Fertil Steril 2016;105(05):1124–1127
17 Skedgel C, Cubi-Molla P, Mott D, et al. Unmet parenthood goals, health-related quality of life and apparent irrationality: understanding the value of treatments for infertility. PharmacoEconom Open 2023;7(03):337–344
18 Vassena R. Moonshots and last miles: what it may take to treat infertility for all. Reprod Biomed Online 2024;48(02):103642
19 Mercer. New survey finds employers adding fertility benefits to promote DEI, from The Survey on Fertility Benefits. 2021. Accessed June 2, 2024 at: https://www.mercer.us/our-thinking/healthcare/new-survey-finds-employers-adding-fertility-benefits-to-promote-dei.html
20 Weigel G, Ranji U, Long M, Salganicoff A. KFF. Coverage and use of fertility services in the U.S. 2020. Accessed June 2, 2024 at: https://www.kff.org/womens-health-policy/issue-brief/coverage-and-use-of-fertility-services-in-the-u-s/
21 Chiware TM, Vermeulen N, Blondeel K, et al. IVF and other ART in low- and middle-income countries: a systematic landscape analysis. Hum Reprod Update 2021;27(02):213–228
22 The Economist. Can the rich world escape its baby crisis? May 21, 2024. Accessed June 2, 2024 at: https://www.economist.com/finance-and-economics/2024/05/21/can-the-rich-world-escape-its-baby-crisis
23 Sample ML-based IVF refund program cost comparison. Accessed July 17, 2024 at: https://www.univfy.com/research/sample-ivf-refund-program-cost-comparison
24 Ethics Committee of the American Society for Reproductive Medicine. Electronic address: asrm@asrm.org. Financial "risk-sharing" or refund programs in assisted reproduction: an Ethics Committee opinion. Fertil Steril 2024;121(05):783–786
25 van Empel IWH, Aarts JWM, Cohlen BJ, et al. Measuring patient-centredness, the neglected outcome in fertility care: a random multicentre validation study. Hum Reprod 2010;25(10):2516–2526
26 Dancet EA, Van Empel IW, Rober P, Nelen WL, Kremer JA, D'Hooghe TM. Patient-centred infertility care: a qualitative study to listen to the patient's voice. Hum Reprod 2011;26 (04):827–833
27 Moragianni VA, Penzias AS. Cumulative live-birth rates after assisted reproductive technology. Curr Opin Obstet Gynecol 2010;22(03):189–192
28 Cedars MI. Fresh versus frozen: initial transfer or cumulative cycle results: how do we interpret results and design studies? Fertil Steril 2016;106(02):251–256
29 McMahon C, Hammarberg K, Lensen S, Wang R, Mol BW, Vollenhoven BJN. What do women undergoing in vitro fertilization (IVF) understand about their chance of IVF success? Hum Reprod 2024;39(01):130–138
30 Miron-Shatz T, Holzer H, Revel A, et al. 'Luckily, I don't believe in statistics': survey of women's understanding of chance of success with futile fertility treatments. Reprod Biomed Online 2021;42 (02):463–470
31 Thomas JM, Cooney LM Jr, Fried TR. Prognosis reconsidered in light of ancient insights-from Hippocrates to modern medicine. JAMA Intern Med 2019;179(06):820–823
32 Reindollar RH, Regan MM, Neumann PJ, et al. A randomized clinical trial to evaluate optimal treatment for unexplained

infertility: the fast track and standard treatment (FASTT) trial. Fertil Steril 2010;94(03):888–899
33 Man JK-Y, Parker AE, Broughton S, Ikhlaq H, Das M. Should IUI replace IVF as first-line treatment for unexplained infertility? A literature review. BMC Womens Health 2023;23(01):557
34 Osmanlıoğlu Ş, Şükür YE, Tokgöz VY, et al. Intrauterine insemination with ovarian stimulation is a successful step prior to assisted reproductive technology for couples with unexplained infertility. J Obstet Gynaecol 2022;42(03):472–477
35 Artificial intelligence/machine learning (AI/ML) platform for IVF. Accessed July 17, 2024 at: https://www.univfy.com/research/univfy-ml-platform
36 Banerjee P, Choi B, Shahine LK, et al. Deep phenotyping to predict live birth outcomes in in vitro fertilization. Proc Natl Acad Sci U S A 2010;107(31):13570–13575
37 Nelson SM, Fleming R, Gaudoin M, Choi B, Santo-Domingo K, Yao M. Antimüllerian hormone levels and antral follicle count as prognostic indicators in a personalized prediction model of live birth. Fertil Steril 2015;104(02):325–332
38 Choi B, Bosch E, Lannon BM, et al. Personalized prediction of first-cycle in vitro fertilization success. Fertil Steril 2013;99(07):1905–1911
39 Lannon BM, Choi B, Hacker MR, et al. Predicting personalized multiple birth risks after in vitro fertilization-double embryo transfer. Fertil Steril 2012;98(01):69–76
40 Chen SH, Xie YA, Cekleniak NA, Keegan DA, Yao MWM. In search of the crystal ball - how many eggs to a live birth? A 2-step prediction model for egg freezing counseling based on individual patient and center data. Fertil Steril 2019;112(03):83–84
41 Shingshetty L, Cameron NJ, Mclernon DJ, Bhattacharya S. Predictors of success after in vitro fertilization. Fertil Steril 2024;121(05):742–751
42 Examples of clinical variables used in IVF prediction models. Accessed July 17, 2024 at: www.univfy.com/research/sample-ivf-clinical-predictors
43 McLernon DJ, Raja EA, Toner JP, et al. Predicting personalized cumulative live birth following in vitro fertilization. Fertil Steril 2022;117(02):326–338
44 Xu T, de Figueiredo Veiga A, Hammer KC, Paschalidis IC, Mahalingaiah S. Informative predictors of pregnancy after first IVF cycle using eIVF practice highway electronic health records. Sci Rep 2022;12(01):839
45 Cai J, Jiang X, Liu L, et al. Pretreatment prediction for IVF outcomes: generalized applicable model or centre-specific model? Hum Reprod 2024;39(02):364–373
46 Nayudu PL, Gook DA, Hepworth G, Lopata A, Johnston WI. Prediction of outcome in human in vitro fertilization based on follicular and stimulation response variables. Fertil Steril 1989;51(01):117–125
47 Hughes EG, King C, Wood EC. A prospective study of prognostic factors in in vitro fertilization and embryo transfer. Fertil Steril 1989;51(05):838–844
48 Stolwijk AM, Zielhuis GA, Hamilton CJ, et al. Prognostic models for the probability of achieving an ongoing pregnancy after in-vitro fertilization and the importance of testing their predictive value. Hum Reprod 1996;11(10):2298–2303
49 Templeton A, Morris JK. Reducing the risk of multiple births by transfer of two embryos after in vitro fertilization. N Engl J Med 1998;339(09):573–577
50 Commenges-Ducos M, Tricaud S, Papaxanthos-Roche A, Dallay D, Horovitz J, Commenges D. Modelling of the probability of success of the stages of in-vitro fertilization and embryo transfer: stimulation, fertilization and implantation. Hum Reprod 1998;13(01):78–83
51 Minaretzis D, Harris D, Alper MM, Mortola JF, Berger MJ, Power D. Multivariate analysis of factors predictive of successful live births in in vitro fertilization (IVF) suggests strategies to improve IVF outcome. J Assist Reprod Genet 1998;15(06):365–371
52 Hunault CC, Eijkemans MJ, Pieters MH, et al. A prediction model for selecting patients undergoing in vitro fertilization for elective single embryo transfer. Fertil Steril 2002;77(04):725–732
53 Ferlitsch K, Sator MO, Gruber DM, Rücklinger E, Gruber CJ, Huber JC. Body mass index, follicle-stimulating hormone and their predictive value in in vitro fertilization. J Assist Reprod Genet 2004;21(12):431–436
54 Lehert P, Chin W, Schertz J, D'Hooghe T, Alviggi C, Humaidan P. Predicting live birth for poor ovarian responders: the PROsPeR concept. Reprod Biomed Online 2018;37(01):43–52
55 Güvenir HA, Misirli G, Dilbaz S, Ozdegirmenci O, Demir B, Dilbaz B. Estimating the chance of success in IVF treatment using a ranking algorithm. Med Biol Eng Comput 2015;53(09):911–920
56 Metello JL, Tomás C, Ferreira P. Can we predict the IVF/ICSI live birth rate? JBRA Assist Reprod 2019;23(04):402–407
57 Bancsi LF, Huijs AM, den Ouden CT, et al. Basal follicle-stimulating hormone levels are of limited value in predicting ongoing pregnancy rates after in vitro fertilization. Fertil Steril 2000;73(03):552–557
58 Jones CA, Christensen AL, Salihu H, et al. Prediction of individual probabilities of live birth and multiple birth events following in vitro fertilization (IVF): a new outcomes counselling tool for IVF providers and patients using HFEA metrics. J Exp Clin Assist Reprod 2011;8:3
59 Nelson SM, Lawlor DA. Predicting live birth, preterm delivery, and low birth weight in infants born from in vitro fertilisation: a prospective study of 144,018 treatment cycles. PLoS Med 2011;8(01):e1000386
60 Vaegter KK, Lakic TG, Olovsson M, Berglund L, Brodin T, Holte J. Which factors are most predictive for live birth after in vitro fertilization and intracytoplasmic sperm injection (IVF/ICSI) treatments? Analysis of 100 prospectively recorded variables in 8,400 IVF/ICSI single-embryo transfers. Fertil Steril 2017;107(03):641–648.e2
61 Stolwijk AM, Wetzels AM, Braat DD. Cumulative probability of achieving an ongoing pregnancy after in-vitro fertilization and intracytoplasmic sperm injection according to a woman's age, subfertility diagnosis and primary or secondary subfertility. Hum Reprod 2000;15(01):203–209
62 Lintsen AM, Eijkemans MJ, Hunault CC, et al. Predicting ongoing pregnancy chances after IVF and ICSI: a national prospective study. Hum Reprod 2007;22(09):2455–2462
63 Grzegorczyk-Martin V, Roset J, Di Pizio P, et al. Adaptive data-driven models to best predict the likelihood of live birth as the IVF cycle moves on and for each embryo transfer. J Assist Reprod Genet 2022;39(08):1937–1949 [Erratum in: J Assist Reprod Genet. 2023 Feb;40(2):429. PMID: 35767167; PMCID: PMC9428070]
64 Ottosen LD, Kesmodel U, Hindkjaer J, Ingerslev HJ. Pregnancy prediction models and eSET criteria for IVF patients – do we need more information? J Assist Reprod Genet 2007;24(01):29–36
65 La Marca A, Capuzzo M, Donno V, et al. The predicted probability of live birth in in vitro fertilization varies during important stages throughout the treatment: analysis of 114,882 first cycles. J Gynecol Obstet Hum Reprod 2021;50(03):101878
66 Roberts SA, McGowan L, Mark Hirst W, et al; towardSET Collaboration. Reducing the incidence of twins from IVF treatments: predictive modelling from a retrospective cohort. Hum Reprod 2011;26(03):569–575
67 Carrera-Rotllan J, Estrada-García L, Sarquella-Ventura J. Prediction of pregnancy in IVF cycles on the fourth day of ovarian stimulation. J Assist Reprod Genet 2007;24(09):387–394
68 van Loendersloot LL, van Wely M, Repping S, Bossuyt PM, van der Veen F. Individualized decision-making in IVF: calculating the chances of pregnancy. Hum Reprod 2013;28(11):2972–2980
69 Zhang Q, Wang X, Zhang Y, Lu H, Yu Y. Nomogram prediction for the prediction of clinical pregnancy in freeze-thawed embryo transfer. BMC Pregnancy Childbirth 2022;22(01):629

70 Vogiatzi P, Pouliakis A, Siristatidis C. An artificial neural network for the prediction of assisted reproduction outcome. J Assist Reprod Genet 2019;36(07):1441–1448
71 Gao H, Liu DE, Li Y, Wu X, Tan H. Early prediction of live birth for assisted reproductive technology patients: a convenient and practical prediction model. Sci Rep 2021;11(01):331
72 Gong X, Zhang Y, Zhu Y, et al. Development and validation of a live birth prediction model for expected poor ovarian response patients during IVF/ICSI. Front Endocrinol (Lausanne) 2023; 14:1027805
73 Wu Y, Yang R, Lin H, Cao C, Jiao X, Zhang Q. A validated model for individualized prediction of live birth in patients with adenomyosis undergoing frozen-thawed embryo transfer. Front Endocrinol (Lausanne) 2022;13:902083
74 Hassan MR, Al-Insaif S, Hossain MI, Kamruzzaman J. A machine learning approach for prediction of pregnancy outcome following IVF treatment. Neural Comput Appl 2020;32:2283–2297
75 CNBarreto N, Castro GZ, Pereira RG, et al. Predicting in vitro fertilization success in the Brazilian public health system: a machine learning approach. Med Biol Eng Comput 2022;60 (07):1851–1861
76 Wen JY, Liu CF, Chung MT, Tsai YC. Artificial intelligence model to predict pregnancy and multiple pregnancy risk following in vitro fertilization-embryo transfer (IVF-ET). Taiwan J Obstet Gynecol 2022;61(05):837–846
77 Mehrjerd A, Rezaei H, Eslami S, Ratna MB, Khadem Ghaebi N. Internal validation and comparison of predictive models to determine success rate of infertility treatments: a retrospective study of 2485 cycles. Sci Rep 2022;12(01):7216
78 Wang CW, Kuo CY, Chen CH, Hsieh YH, Su EC. Predicting clinical pregnancy using clinical features and machine learning algorithms in in vitro fertilization. PLoS One 2022;17(06):e0267554
79 Fu K, Li Y, Lv H, Wu W, Song J, Xu J. Development of a model predicting the outcome of *in vitro* fertilization cycles by a robust decision tree method. Front Endocrinol (Lausanne) 2022; 13:877518
80 Yang H, Liu F, Ma Y, Di M. Clinical pregnancy outcomes prediction in vitro fertilization women based on random forest prediction model: a nested case-control study. Medicine (Baltimore) 2022; 101(49):e32232
81 Goyal A, Kuchana M, Ayyagari KPR. Machine learning predicts live-birth occurrence before in-vitro fertilization treatment. Sci Rep 2020;10(01):20925
82 Dhillon RK, McLernon DJ, Smith PP, et al. Predicting the chance of live birth for women undergoing IVF: a novel pretreatment counselling tool. Hum Reprod 2016;31(01):84–92
83 McLernon DJ, Maheshwari A, Lee AJ, Bhattacharya S. Cumulative live birth rates after one or more complete cycles of IVF: a population-based study of linked cycle data from 178,898 women. Hum Reprod 2016;31(03):572–581
84 Ratna MB, Bhattacharya S, van Geloven N, McLernon DJ. Predicting cumulative live birth for couples beginning their second complete cycle of in vitro fertilization treatment. Hum Reprod 2022;37(09):2075–2086
85 Qiu J, Li P, Dong M, Xin X, Tan J. Personalized prediction of live birth prior to the first in vitro fertilization treatment: a machine learning method. J Transl Med 2019;17(01):317
86 Nguyen ET, Retzloff MG, Gago LA, et al. Predicting IVF live birth probabilities using machine learning, center-specific and national registry-based models. medRxiv 2024. Doi: 10.1101/2024.06.20. 2430 8970
87 Liu X, Chen Z, Ji Y. Construction of the machine learning-based live birth prediction models for the first in vitro fertilization pregnant women. BMC Pregnancy Childbirth 2023;23(01):476
88 Luke B, Brown MB, Wantman E, et al. A prediction model for live birth and multiple births within the first three cycles of assisted reproductive technology. Fertil Steril 2014;102(03):744–752
89 Palermo G, Joris H, Devroey P, Van Steirteghem AC. Pregnancies after intracytoplasmic injection of single spermatozoon into an oocyte. Lancet 1992;340(8810):17–18
90 Asada Y. Evolution of intracytoplasmic sperm injection: from initial challenges to wider applications. Reprod Med Biol 2024; 23(01):e12582
91 Celada P, Bosch E. Freeze-all, for whom, when, and how. Ups J Med Sci 2020;125(02):104–111
92 Cohen J, Grudzinskas G, Johnson M. Welcome to the '100% club'!. Reprod Biomed Online 2012;24(04):375–376
93 Niederberger C, Pellicer A, Cohen J, et al. Forty years of IVF. Fertil Steril 2018;110(02):185–324.e5
94 Braakhekke M, Kamphuis EI, Dancet EA, Mol F, van der Veen F, Mol BW. Ongoing pregnancy qualifies best as the primary outcome measure of choice in trials in reproductive medicine: an opinion paper. Fertil Steril 2014;101(05):1203–1204
95 Barnhart KT. Live birth is the correct outcome for clinical trials evaluating therapy for the infertile couple. Fertil Steril 2014;101 (05):1205–1208
96 Clarke JF, van Rumste MM, Farquhar CM, Johnson NP, Mol BW, Herbison P. Measuring outcomes in fertility trials: can we rely on clinical pregnancy rates? Fertil Steril 2010;94(05):1647–1651
97 Houwelingen JC, Cessie S. Logistic regression, a review. Stat Neerl 1988;42(04):215–232
98 Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol 2001;54(10):979–985
99 Tolles J, Meurer WJ. Logistic regression: relating patient characteristics to outcomes. JAMA 2016;316(05):533–534
100 JAMA. Logistic Regression—What It Is and How to Use It in Clinical Research. JAMA evidence: Guide to Statistics and Methods. Published online: January 7, 2021. Accessed June 1, 2024 at: https://edhub.ama-assn.org/jn-learning/audio-player/18574348
101 Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: Logistic regression. Perspect Clin Res 2017;8 (03):148–151
102 Advantages and disadvantages of logistic regression. Accessed June 1, 2024 at: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/
103 Kumar S. Assumptions and limitations of logistic regression: navigating the nuances. Accessed June 1, 2024 at: https://medium.com/@skme20417/4-assumptions-and-limitations-of-logistic-regression-navigating-the-nuances-8ef249cc7a01
104 Mitchell TM. Machine Learning. McGraw-Hill; 1997
105 Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer; 2009
106 Cox DR. The regression analysis of binary sequences. J R Stat Soc B 1958;20(02):215–232
107 Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29(05):1189–1232
108 Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785–794
109 Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems. 2017:3149–3157. Available at: https://papers.nips.cc/paper_files/paper/2017. Accessed on September 20, 2024
110 Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc B 1996;58(01):267–288
111 Breiman L. Random forests. Mach Learn 2001;45(01):5–32
112 May M. Eight ways machine learning is assisting medicine. Nat Med 2021;27(01):2–3
113 Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med 2019;25(01):24–29
114 Chen D, Liu S, Kingsbury P, et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. NPJ Digit Med 2019;2:43

115 Leushuis E, van der Steeg JW, Steures P, et al. Prediction models in reproductive medicine: a critical appraisal. Hum Reprod Update 2009;15(05):537–552
116 Srivastava T. 12 Important model evaluation metrics for machine learning everyone should know. (updated 2023). Accessed June 1, 2024 at: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/
117 Cross validation in machine learning. December. 21, 2023. Accessed September 5, 2024 at: https://www.geeksforgeeks.org/cross-validation-machine-learning/
118 Dvorak T. Why isn't out-of-time validation more ubiquitous? Towards Data Science. Feb 11, 2019. Accessed June 1, 2024 at: https://towardsdatascience.com/why-isnt-out-of-time-validation-more-ubiquitous-7397098c4ab6
119 Maleki F, Muthukrishnan N, Ovens K, Reinhold C, Forghani R. Machine learning algorithm validation: from essentials to advanced applications and implications for regulatory certification and deployment. Neuroimaging Clin N Am 2020;30(04):433–445
120 Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics 2020;21(02):345–352
121 Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021;385(03):283–286
122 Human Fertilization & Embryology Authority (HFEA). Multiple births in fertility treatment. 2019. Accessed on April 29, 2024 at: https://www.hfea.gov.uk/about-us/publications/research-and-data/multiple-births-in-fertility-treatment-2019/
123 Curchoe CL, Tarafdar O, Aquilina MC, Seifer DB. SART CORS IVF registry: looking to the past to shape future perspectives. J Assist Reprod Genet 2022;39(11):2607–2616
124 Swanson T, Yao M, Retzloff M, et al. Inter-center variation of patients' clinical profiles is associated with live birth outcomes. Fertil Steril 2023;120(4, Supp):E175
125 Khera R, Haimovich J, Hurley NC, et al. Use of machine learning models to predict death after acute myocardial infarction. JAMA Cardiol 2021;6(06):633–641
126 Kent J. Machine learning limited when applied to clinical data registries. HealthIT Analytics March 11, 2021. Accessed July 17, 2024 at: https://healthitanalytics.com/news/machine-learning-limited-when-applied-to-clinical-data-registries
127 Rodrigues MMS, Barreto-Duarte B, Vinhaes CL, et al. Machine learning algorithms using national registry data to predict loss to follow-up during tuberculosis treatment. BMC Public Health 2024;24(01):1385
128 Jalali A, Lonsdale H, Zamora LV, et al; Pediatric Craniofacial Collaborative Group. Machine learning applied to registry data: development of a patient-specific prediction model for blood transfusion requirements during craniofacial surgery using the pediatric craniofacial perioperative registry dataset. Anesth Analg 2021;132(01):160–171
129 Chen Y, Gue Y, Calvert P, et al; KERALA-AF Registry & APHRS-AF Registry Investigators Joint Senior Authors. Predicting stroke in Asian patients with atrial fibrillation using machine learning: a report from the KERALA-AF registry, with external validation in the APHRS-AF registry. Curr Probl Cardiol 2024;49(04):102456
130 Artemova S, von Schenck U, Fa R, et al. Cohort profile for development of machine learning models to predict healthcare-related adverse events (Demeter): clinical objectives, data requirements for modelling and overview of data set for 2016-2018. BMJ Open 2023;13(08):e070929
131 Kundu S, Aulchenko YS, van Duijn CM, Janssens AC. PredictABEL: an R package for the assessment of risk prediction models. Eur J Epidemiol 2011;26(04):261–264
132 Cook NR, Paynter NP. Performance of reclassification statistics in comparing risk prediction models. Biom J 2011;53(02):237–258
133 Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk prediction instruments: a critical review. Epidemiology 2014;25(01):114–121
134 Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10(03):e0118432
135 Requirements for point-of-care delivery of artificial intelligence/machine learning (AI/ML)-based prognostics at scale. Accessed July 17, 2024 at: https://www.univfy.com/research/point-of-care-requirements
136 The development-to-deployment life cycle of the machine learning-based, center specific (MLCS), prognostic model for use at point-of-care to support patient counseling. Accessed on July 17, 2024 at: https://www.univfy.com/research/mlcs-dev-to-deployment-lifecycle
137 Goldstein J, Weitzman D, Lemerond M, Jones A. Determinants for scalable adoption of autonomous AI in the detection of diabetic eye disease in diverse practice types: key best practices learned through collection of real-world data. Front Digit Health 2023;5:1004130
138 Hariton E, Alvero R, Hill MJ, et al. Meeting the demand for fertility services: the present and future of reproductive endocrinology and infertility in the United States. Fertil Steril 2023;120(04):755–766
139 Adeleye AJ, Kawwass JF, Brauer A, Storment J, Patrizio P, Feinberg E. The mismatch in supply and demand: reproductive endocrinology and infertility workforce challenges and controversies. Fertil Steril 2023;120(3, Pt 1):403–405
140 Society for Assisted Reproductive Technology. Accessed May 10, 2024 at: www.sart.org
141 Society for Assisted Reproductive Technology and University of Aberdeen. Accessed May 10, 2024 at: https://w3.abdn.ac.uk/clsm/SARTIVF/
142 University of Aberdeen. Outcome Prediction in Subfertility, OPIS. Accessed May 10, 2024 at: https://w3.abdn.ac.uk/clsm/opis
143 Yao MWM, Nguyen ET, Retzloff MG, et al. Improving IVF utilization with patient-centric artificial intelligence machine learning (AI/ML): a retrospective multicenter experience. J Clin Med 2024;13(12):3560
144 Salybekov AA, Wolfien M, Hahn W, Hidaka S, Kobayashi S. Artificial intelligence reporting guidelines' adherence in nephrology for improved research and clinical outcomes. Biomedicines 2024;12(03):606
145 EQUATOR Network. What is a Reporting Guideline? Accessed February 20, 2024 at: https://www.equator-network.org/about-us/what-is-a-reporting-guideline/
146 Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. Nat Med 2020;26(06):807–808
147 Tripod Statement. Accessed February 20, 2024 at: https://www.tripod-statement.org/
148 Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AKSPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 2020;370:m3164
149 Vasey B, Nagendran M, Campbell B, et al; DECIDE-AI Expert Group. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ 2022;377:e070904
150 Duda RO, Hart PE. Pattern Classification and Scene Analysis. New York: Wiley; 1973