Title:

Predicting IVF live birth probabilities using machine learning, center-specific models: validation results and potential benefits over national registry-based models.

Elizabeth T. Nguyen PhD[1], Matthew G. Retzloff MD[2], L. April Gago MD[3], John E. Nichols MD[4], John F. Payne MD[4], Barry A. Ripps MD[5], Michael Opsahl MD[6], Jeremy Groll MD[7], Ronald Beesley MD[6], Gregory Neal MD[2], Jaye Adams MD[2], Lorie Nowak PhD[7], Trevor Swanson PhD[1], Xiaocong Chen MSc[1], Mylene W. M. Yao MD[1].

[1]R&D Department, Univfy, Los Altos, CA, US.
[2]Fertility Center of San Antonio, San Antonio, TX, US.
[3]Gago Center for Fertility, Brighton, MI, US.
[4]Piedmont Reproductive Endocrinology Group, Greenville, SC, US.
[5]NewLIFE Fertility, Pensacola, FL, US.
[6]Poma Fertility, Kirkland, WA, US.
[7]SpringCreek Fertility, Dayton, OH, US.

Corresponding Author:
Author Name: Mylene W. M. Yao, MD
Affiliation: Univfy Inc.
Mailing Address: 171 Main Street, #139, Los Altos, CA 94022
Phone: 1-650-799-8003
Email: mylene.yao@univfy.com

Ongoing improvement of pretreatment live birth prognostication for in vitro fertilization (IVF) is critical for informing fertility patients' treatment decisions, advocating for IVF coverage and supporting value-based IVF care. The US national registry Society for Assisted Reproductive Technology (SART) IVF live birth prediction (LBP) model (SART model) has been widely adopted for its prognostic support without external validation or utilization studies. We conducted a retrospective model validation study to compare the IVF LBP performance of machine learning, center-specific (MLCS) models versus the SART model in 6 unrelated US fertility centers using their respective center-specific test sets comprising an aggregate of 4,635 patients' first-IVF cycle data. Compared to the SART model, MLCS2 showed higher median Precision Recall AUC at 0.75 (IQR 0.73, 0.77) vs. 0.69 (IQR 0.68, 0.71), $p<0.05$ and higher median F1 Score across LBP thresholds. Further, MLCS1 showed no evidence of data drift when validated using out-of-time test data from a later period. Reclassification analysis showed that MLCS2 models assigned more appropriate and higher IVF LBPs compared to the SART model, which underestimated patient prognoses (continuous net reclassification index: 18.3%, $p<0.0001$). Overall, MLCS2 and SART models assigned 30% of patients to differential prognostic groups, with MLCS2 assigning 26% of patients to a higher LBP category compared to the SART model. Importantly, MLCS2 models identified 11% of patients to have LBP ≥ 75%, whereas the SART model detected none. This group had a live birth rate of 81%. We recommend testing a larger sample of fertility centers to further evaluate MLCS model benefits and limitations.

Introduction

Infertility, declared a medical disease and global health issue by the World Health Organization (WHO), is estimated to affect 107M-172M women or couples globally and over 10M in the US (1-14). Despite the proven safety and efficacy of assisted reproductive technology (ART), patients' access to and utilization of ART continue to be met with barriers. Of the couples who could benefit from ART to have a family, only 5% was estimated to use it (4). Further, an estimated 10-50% of patients discontinued fertility care before achieving a live birth (15, *personal communications*). Gaps at different levels of fertility care access included cost of treatment, stigma, emotional stress and uncertain probability of success (5-7, 14-15). (As in vitro fertilization (IVF) is the most common type of ART, the rest of the article will refer to IVF instead of ART.)

Our research has focused on improving IVF access and utilization for patients who have proactively and voluntarily sought fertility care at IVF centers. The potential contribution of artificial intelligence (AI)/machine learning (ML) to improve IVF pretreatment counseling warrants investigation on a broader scale, based on a reported 2-3 fold increase in IVF utilization among new patients across seven independent fertility centers in US and Canada (16). Providers have an important responsibility to offer validated, personalized and relevant prognostic counseling to educate patients about the potential benefits and limitations of IVF and to consider a course of IVF treatments to maximize the probability of having a baby (7, 17, 18). In addition, the cost of having an IVF baby is elusive for an individual patient because it depends heavily on the women's or couple's IVF live birth prognosis. IVF cost-to-live birth transparency is urgently needed for both patients and providers to support a very personal decision, whether for a medical or social/family need.

We previously reported the development and clinical usage of machine learning, center-specific (MLCS) IVF prognostic models to support provider-patient counseling across geographies (e.g. US, Canada, UK and EU) which varied in IVF payers including patients themselves, employers, health insurance plans and government (7, 16, 19-24). Geographically distributed, local delivery of fertility care helps to remove barriers in patient-centric care. However, patients' clinical characteristics varied significantly across fertility centers and those inter-center variations were associated with differential IVF live birth outcomes (25). Therefore, patient-centric care includes providing patients with prognostic counseling that is relevant and reflects clinical characteristics of the local patient population.

Nonetheless, there is a widely held perception among providers that the national registry-based, center-agnostic IVF pretreatment prognostics model developed by the US Society for Assisted Reproductive Technology (SART model) "suffices" based on its usage of a large national dataset from 121,561 IVF cycles started in a two-year period (2014-2015) and its free online access, even though model validation using an external test set and clinical utilization studies have not been reported (26, 27). Further, recommendations were made to fertility centers in countries outside of US and UK to adapt local prognostic models by recalibrating the US SART and UK Human Fertilisation and Embryology Authority (HFEA) data-based models using local data, but Cai et al. showed that de novo MLCS model gave much improved model validation metrics (26, 28-30). Validation of MLCS IVF pretreatment models using separate test sets have been reported and the design, validation and clinical usage considerations for IVF pretreatment prognostic model has been reviewed in-depth(7, 19, 20, 24, 28, 31, 32).

The above body of work led providers to ask 1) whether and how MLCS and SART model predictions differ for patients seeking fertility care in the US today, and 2) whether MLCS models are applicable to small-to-midsize US fertility centers. Indeed, a head-to-head comparison between the MLCS and SART pretreatment models for US centers has not been reported. Providers also want to know if "...the [MLCS] model has been validated for patients who received IVF after MLCS-based counseling?" Translated to ML and clinical research lingo, providers are asking "is there data drift causing previously validated models not to be applicable to patients doing IVF in a later time period?" Addressing these questions will help us to develop best practices for IVF live birth prognostic counseling, which is critical for advancing fertility care in the US and globally.

This retrospective cohort study aimed to compare the performance of the MLCS and SART pretreatment models for six unrelated small-to-midsize US fertility centers operating in 22 locations across 9 states in 4 US regions (West, Southeast, Southwest and Midwest) using their respective center-specific test sets comprising an aggregate of 4,635 patients' first-IVF cycle data. This study's primary outcome is a set of model performance metrics -- area-under-the-curve of the receiver operating characteristic curve (AUC-ROC), AUC improvement over age control, predictive power, precision, recall, F1 score, and precision-recall AUC -- used to evaluate MLCS and SART models (33). (Depending on the AMH availability of each cycle, the SART model with and without AMH as predictor would be used (26).) Live birth outcomes are primary outcomes for the prediction models being evaluated and are not primary outcomes of this study. We also addressed data drift, a scenario in which changes in patients, their characteristics, or treatment protocols cause a previously validated model not to be clinically applicable anymore (33, 34). To clarify, the AI techniques used and the problems addressed by this study are distinct from AI usage to improve embryo selection and its efficiency (35-37). Also, we did not use deep learning, foundational models or generative AI, so risks of "hallucination" do not apply (38-40).

Results

Six centers participated in this study. Table 1 shows for each model validation, the MLCS model tested, time period of each data set, IVF volume range represented by the data set, data set usage (e.g. used for both training and testing or testing only), and the validation type (in-time or out-of-time).

For each center, the initial, version 1 MLCS model (MLCS1) and the updated, version 2 MLCS model (MLCS2) cross validation results showed improved AUC and positive posterior log odds ratio compared to age (PLORA), indicating they were superior to their respective age control models. The median and interquartile range (IQR) of AUC, AUC improvement, and PLORA are summarized in Table 2. Next, we tested whether the AUC and PLORA of MLCS2 were improved over those of MLCS1. Across 6 centers, AUC was similar between MLCS1 & MLCS2, but PLORA of MLCS2 (23.9, IQR 10.2, 39.4) was improved over those of MLCS1 (7.2, IQR 3.6, 11.8), $p<0.05$. Therefore, the model update process (i.e. using a larger data set including more recent years of data) resulted in improved model performance (Table 2).

To test for the risk of data drift, we performed live model validation (LMV) testing on each center's MLCS1 model using a center-specific, out-of-time data set. There was no detectable data drift based on observing comparable AUC and PLORA values between the LMV and cross validation (CV) results for MLCS1 model (Table 1).

MLCS2 and SART models were evaluated for each center using the modified, center-specific de novo model validation test sets, de novo model validation test sets 1 and 2 (DNMV1 and DNMV2), each comprising an aggregate of 4,645 and 4,421 unique patient-cycles across 6 centers. The overall rates of live birth labeling were 58.5% and 56.4% for DNMV1 and DNMV2, respectively. Further, only ~5% of patient-cycles did not have an AMH value and they were tested by the SART model formulae that did not require AMH predictor.

AUC and PLORA were not significantly different between the MLCS2 and SART models for either DNMV1 or DNMV2 (Table 2). Further evaluation was performed using the model metrics F1 Score (the harmonic mean of precision and recall) and Precision-Recall (PR) AUC, which are more sensitive than the AUC ROC in detecting improvements in predicting the positive class which is live birth prediction in the context of this study.

The median F1 Score was higher for MLCS2 compared to SART model across predicted live birth probability (LBP) thresholds sampled at ≥40%, ≥50%, ≥60%, ≥70%. For example, at the 50% LBP threshold, MLCS2 had a median F1 Score of 0.74 (IQR=0.72, 0.78) compared to 0.71 (IQR=0.68, 0.73) for SART using the DNMV1 test set. Similar findings were observed using the DNMV2 test set (Table 3).

PR AUC was also significantly higher for MLCS2 using DNMV1. The median PR AUC was 0.75 (IQR=0.73, 0.77) for MLCS2 and 0.69 (IQR=0.68, 0.71) for SART across the 6 centers, $p<0.05$ (Table 2). The findings were similar when tested using DNMV2 test set, $p<0.05$ (Table 3). At these six centers, assessing recall at the LBP threshold of ≥ 50%, MLCS2 models can identify ~84% of patients who would go on to have IVF live births, while the SART model can only identify ~75%. In other words, the use of SART model would have resulted in 9% of patients missing a good prognosis of ≥50% LBP. While overall precision was comparable between MLCS2 and SART, across all precision rates, MLCS2 models showed higher rates of recall and more cycles with higher IVF live birth probabilities across six centers (Figure 1A). The higher recall rates of MLCS2 are consistent with MLCS2 model generating an LBP distribution that is shifted to the right (i.e. having more patients with higher LBPs) compared to SART model (Figure 1B). Similar findings were observed for DNMV2.

Although the PR AUC, F1 Score and recall rates sufficed to show that the MLCS2 models provide more appropriate LBPs compared to the SART model, those model metrics were not intuitive for clinicians and patients. To facilitate communications with clinicians, we created a 4x4 reclassification table for the DNMV1 test set to show the concordance and discordance of LBPs made by MLCS2 and SART models in a practical, clinical context (see Methods and Table 4). (Note: Analyses related to model validation, LBPs and live birth rates were all necessarily performed at the group level, not at the individual patient level because it would not be possible or scientifically feasible to evaluate those measures on individual patients.)

Overall, 70% (3259 of 4645) of patients had concordant LBPs between SART and MLCS2 models (Table 4, green cells), while 30% (1386 of 4645) of the patients had discordant LBPs (Table 4, peach or blue cells) -- meaning, MLCS2 and SART models placed 1386 patients into different prognostic categories. Importantly, for the 30% of patients with discordant LBP prognostic categories, each group's live birth rate aligned with MLCS2 model predictions whether MLCS2 assigned a higher or lower prognostic category than SART. Of the patients with discordant LBPs, 89% (1230 of 1386)

were given higher LBPs by MLCS2 (Table 4, blue cells) and 11% (156 of 1386) were given lower LBPs by MLCS2 (Table 4, peach cells) compared to SART model.

Patient groups (in blue) that got moved up to a higher prognostic category by MLCS2 showed live birth rates that were in the range predicted by MLCS2 models. For example, of the patients that the SART model assigned to Low LBP (N=249), Medium LBP (N=1320) and High LBP (N=3076), 67%, 43% and 16%, respectively, were upgraded to a higher prognostic group by MLCS2 models. Further, these "upgraded" patient groups had live birth rates (LBR) that matched with the expected LBP range: patients upgraded from Low LBP to Medium LBP had 33% LBR; patients upgraded from Medium LBP to High LBP had 61% LBR; and patients upgraded from High LBP to Very High LBP had LBR 80%. (Table 4.)

Importantly, no patients received LBP ≥ 75% from the SART model whereas 11% of patients received LBP ≥ 75% from the MLCS2 models and the live birth rate for patients in this Very High LBP prognostic group was 81%. In summary, MLCS2 assigned 26% (1230 of 4645) of patients to a higher LBP category compared to the SART model. SART-LBPs underestimated live birth rates for each of those discordant groups across the spectrum of prognostic categories.

The converse was also observed: 12 of 1320 (1%) patients assigned to Medium LBP and 144 of 3076 (5%) patients assigned to High LBP by SART were downgraded to Low LBP and Medium LBP categories by MLCS2 models, respectively, yielding 0% LBR in the Low LBP group and 37% LBR in the Medium LBP group. Therefore, across all discordant groups, MLCS2 gave more appropriate LBPs compared to the SART model based on alignment of each group's LBR to the expected LBP range for the MLCS2-associated prognostic category. Similar results were obtained from the DNMV2 test set.

Using continuous net reclassification index (NRI), compared with unsuccessful patients, patients with a live birth outcome were 18.3% (95% CI 13.3%, 23.2%) more likely to be given a higher LBP with MLCS2 compared to SART when tested using the DNMV1 ($p<0.001$). Similar findings were obtained using the DNMV2. Although, the differential prognoses affecting 30% of patients were not a measure of the models, they help to contextualize the improved model metrics in PR AUC, F1 Score and Recall observed for the MLCS2 over the SART models.

Discussion

This study compared individual MLCS models and the SART model for pretreatment IVF live birth prognostics for six unrelated, geographically distributed US fertility centers that reported to the SART registry. The retrospective study design was appropriate because the prognostic models were previously trained, tested, and already in clinical usage, and evaluation of the models' technical performance were not biased by the retrospective design. MLCS model validation was performed prior to clinical usage and the deployed MLCS models were tested using an out-of-time test set, coined live model validation, LMV. In addition to validating the models, those results also indicated that there was no detectable data drift.

We took the pragmatic realist approach to address "how do the MLCS and SART model predictions differ for the patients seeking care today?" The MLCS2 models performed better than the SART model in predicting the positive class (i.e. live birth prediction), as indicated by the PR AUC, F1 score and Recall (33). Further, using the 4 x 4 reclassification table to define concordance and

discordance between MLCS2 and SART model LBPs provided clinically intuitive interpretation of the results. Overall, 30% of all patients in the DMNV1 test set showed discordance between MLCS2 and SART model LBPs, and of those 89% of patients were placed in lower prognostic groups by SART whereas 11% of patients were placed in higher prognostic groups by SART. All discordant groups received more appropriate LBPs from MLCS2 models than from the SART model, whether the discordance stemmed from under- or overestimation of LBPs by the SART model. Further, the under- and overestimation by SART affected patients with LBPs across the spectrum from Low LBP to Very High LBP were affected.

Why are these results important? First, patients and providers should be aware of the best available source of IVF live birth prognosis to inform patients' decision-making. No additional justification should be needed. Patients who have very poor prognosis deserve to know their specific treatment limitations to inform funding IVF and subjecting themselves to physical intervention, whereas patients with good or excellent prognosis should not be discouraged by an underestimation of prognoses, which may deter or delay IVF, potentially resulting in not having a family. Further, in countries that allow the use of donor egg IVF, patients with poor prognosis may choose to use donor eggs for a higher probability of success.

The improved F1 Score and PR AUC model metrics were reflected by a high percentage of patients placed into groups with more appropriate and higher LBPs given by MLCS2 and underestimation of LBPs by SART. These findings can directly translate to support IVF pricing to the lower cost per IVF baby, as explained further in Yao et al. (24). This type of IVF pricing is already offered to self-pay patients or consumers, but it can be expanded and adapted for enterprise payers such as health plans and employers. For context, the cost per IVF baby is a significant barrier to IVF access, and in most US states, there is no mandated IVF coverage by health insurance plans (41, 42). Many fertility centers offer shared risk or refund programs to self-pay patients, with the goal of increasing the feasibility of doing a course of several IVF treatments. However, if the actuarial-like models backing these shared risk programs have suboptimal F1 Score, PR AUC and Recall, more patients may be deemed ineligible to enter the program. Therefore, achieving high precision and recall would alleviate the financial barrier while maximizing live birth outcomes for many patients (24). The MLCS modeling framework enables providers to offer true value-based IVF care at scale, through transparent IVF pricing and live birth outcomes specific to each fertility center.

Prediction models that drive value-based IVF care can also support provider-patient counseling, an important part of patient-centric care. Most importantly, transparency can be achieved by using the same IVF live birth prediction model to support both value-based care and provider-patient counseling. The SART model, in the format of a free online calculator, presumably serves as an educational tool that encourages patients to seek care. However, at the point when patients have completed their diagnostic workup and are being counseled by providers about the benefits and limitations of IVF, patients are interested to know their IVF live birth probabilities at that particular center.

The MLCS design aimed to strengthen the patient-provider relationship by prioritizing patients' top concern: "Is this IVF success prediction based on your center's own data?" Whereas the MLCS live birth prediction directly responds to that question, the SART online calculator was not designed to address that concern. The SART online calculator's disclaimer states, "The estimates are based on the data we have available and may not be representative of your specific experience...Please speak with your doctor about your specific treatment plan and potential for success."

Like many innovative products used in healthcare, the MLCS-based counseling report has been developed and is sold by Univfy Inc., for-profit company. Currently, Univfy charges a fee to fertility centers for the center-specific implementation of the MLCS-based counseling model. Fertility centers using the MLCS-based report have independently chosen to fund the cost of these reports as a complimentary service to their patients. Therefore, patients are currently not bearing the cost of this technology and it is also equitably priced based on local IVF pricing and IVF volume. The authors of this study believe it is a public service to share this research study's findings as we expect providers, researchers and patients would appreciate knowing about the capabilities of MLCS models. We leave it to stakeholders of the free market -- comprising patients, providers, private equity investors, biopharma, employers, health insurance plans and benefits companies -- to choose performance over freemium based on their patients' and providers' needs.

MLCS-based counseling reports may also improve patient-centric care by streamlining provider-patient flow and clinical workflow. In the context of clinical workflow, MLCS models can support diverse healthcare providers -- such as advanced practice providers (APPs), nurse practitioners and general obstetrician-gynecologists -- in performing patient counseling, further improving scalability and accessibility of IVF treatments (43, 44). MLCS models and their associated counseling reports have also been developed to support other ART clinical counseling scenarios -- e.g. after one or more failed IVF cycles, prior to egg freezing, whether to use their own eggs or eggs from a donor -- which included addressing patients with a poor prognosis and delivering prognosis with compassion (7, 19, 21-24, 45).

Having discussed the application and scope of real-world benefits of MLCS models -- including patient-centric care and lowering the cost per IVF baby -- we now turn to the technical aspects affecting the differential performance of MLCS and SART models. Considering the SART model used 121,561 IVF cycles whereas the MLCS2 models used a median dataset size of 1163 IVF cycles (IQR 658-1662 IVF cycles) to achieve comparable ROC AUC and improved PR AUC and F1 score, the MLCS approach is 200x more data efficient (26). Here, we hypothesized several factors to be driving the improved metrics: 1) The greater number of consecutive years covered by the MLCS data sets allowed for more freeze-all cycles to generate outcomes that reflect more realistic and higher live birth probabilities. 2) The lower-than-expected SART model ROC AUCs for the DNMV datasets may have resulted from the under 40 age limit of this study, whereas the original report of SART model training included age up to 50 years of age (24, 33). Inclusion of older patients can artificially increase the ROC AUC and that issue is discussed in-depth in a separate review (33). 3) MLCS enables greater flexibility in the number and scope of clinical predictors that can be tested for use in the prediction model to capture greater inter-patient differences and dynamic range of model prediction (7, 24-25, SI Figure 1). 4) Protocolization, maintenance of data processing and modeling pipelines, quality assurance, expert human supervision and most importantly, close collaboration with providers and centers' operational teams likely contributed to the quality and validation of MLCS models (7, 24). As the use of ML gains maturity in healthcare, the emphasis shifts to delivering highly scalable, secured pipelines for model pre-processing, model training and model deployment (46). The quality control and considerations used from design to deployment specific to the production of IVF pretreatment live birth prediction models for use at the point-of-care are reviewed by Yao et al, 2024 (24).

Having outlined the likely reasons for improvement, we recommend viewing the differences between the MLCS2 and SART models in total. Instead of dissecting older models and datasets, we

recommend to focus collaborative research efforts to study model metrics on a wider range of fertility centers -- larger centers, academic centers, centers in IVF coverage-mandated states and publicly funded IVF -- to learn the potential benefits and limitations of the use of center agnostic and center-specific multicenter models in comparison with MLCS models, for the prediction of IVF live birth probability in different clinical settings.

More broadly, the highly scaled MLCS framework -- creating MLCS models for many centers with collaboration with providers and quality control for the modeling -- can be used to advance reproductive research. As local specificity is crucial in solving health inequities in fertility care and IVF, which have remained significant and largely unsolved (47, 48), the MLCS framework can be applied to dissect the contribution of social determinants of health to fertility care utilization, access to care and IVF live birth outcomes. In addition, the MLCS framework can be applied to support the evaluation of add-ons, which aimed to improve IVF live birth outcomes but may have inconclusive or controversial clinical results. The acceptance of add-on procedures and the evaluation of each type of add-on have been challenging due to a combination of factors including the heterogeneity of the studies and patient populations (49). Those challenges relate to evaluating IVF patients as a whole and not having locally validated prognostic groups. Ultimately, amassing insights from many MLCS models is expected to advance precision medicine in IVF with cost-efficiency and capabilities to evaluate new diagnostic and therapeutic interventions.

To summarize, we have established a globally applicable framework for MLCS modeling of IVF live birth prediction to inform locally relevant patient-provider counseling, clinical workflow and value-based IVF care to lower the cost per IVF baby. Collectively, improvements in these areas are expected to contribute towards making IVF treatment accessible to more patients with a resulting increase in the families built and number of singleton babies born. We believe multicenter collaboration and collaboration between public and private sectors will accelerate and scale research tackling crucial questions related to racial disparities in IVF, molecular mechanisms of clinical infertility and IVF success, and ways to expand access to IVF care. We hope this study will help to advance reproductive medicine beyond dichotomies of multicenter versus center-specific or ML versus non-ML prediction models. Ultimately, the multicenter scaling of a machine learning, localized approach is expected to maximize benefit to people wanting a family while addressing health inequities and patient needs, de-risking IVF costs and advancing precision medicine in reproductive health.

Methods

Research data sources, de-identified data sets and prior reporting of methods

De-identified IVF treatment clinical variables and outcomes data previously linked and processed as part of Univfy client services were anonymized and entered into Univfy research database as per research protocol. The original data sources included electronic medical record (EMR) and SART CORS, the US national registry database managed by SART (50). Univfy Inc. submitted research protocol to institutional review board (IRB) which designated the exempt status for our research protocol, which used anonymized data.

Briefly, definitions of IVF treatments, live birth and methods used for data collection, exclusion criteria, use of center-specific variables and MLCS model life cycle and evaluation steps including model training and testing, gradient boosted machines (GBM) on the Bernoulli distribution, and the use of ROC-AUC, AUC improvement over age control model ("AUC improvement") and posterior

log of odds ratio (PLORA) were substantially as previously reported and are summarized here for ease of reference (7, 19, 20, 24, 51, Figure 2). Advantages of machine learning in general and GBM specifically over conventional methods such as logistic regression were reviewed previously (7, 24, 51). The first page of a sample provider-patient IVF pretreatment counselling report based on MLCS, live birth probability prediction model, commercially known as the Univfy® PreIVF Report, is available at https://www.univfy.com/research (52). In the MLCS models and throughout this article, an IVF cycle is defined as gonadotropin ovarian stimulation cycle with the intention of retrieving oocytes for IVF or ICSI for embryo culture and blastocyst transfer with or without PGT-A or freeze-all.

Cross validation (CV) and model metrics

Our standard model evaluation procedure required k-fold cross validation on an in-time test set (the test and training data sources were contemporaneous) to compute the ROC-AUC, AUC improvement over age control model ("AUC improvement") and the posterior log odds ratio compared to age control model (PLORA) as previously reported. In layman terms, PLORA describes "given a certain LBP prediction, how much more likely will the MLCS model be correct compared to age control?" PLORA, expressed in the log scale with log base e; non-statisticians may prefer to translate to linear scale (e^{PLORA}) for more intuitive understanding (7, 19, 20, 24, 51, 53, Figure 2). CV of both MLCS1 and MLCS2 was reported using median and interquartile range (IQR) across 6 centers for ROC-AUC, ROC-AUC improvement and PLORA.

Briefly, a model with higher Precision (aka positive predictive value, PPV) is more likely to be correct when predicting a successful IVF live birth outcome; a model with higher Recall (aka sensitivity or true positive rate, TPR) would correctly identify a larger proportion of successful IVF live birth outcomes. F1 score measures the harmonic mean of Precision and Recall such that a high F1 score describes a model with both high Precision and high Recall at a particular live birth probability threshold (24, 33, 54). See *" Creating center-specific, de novo model validation test sets (DNMV1 and DNMV2) to enable statistical analyses of metrics of the MLCS2 and SART models"* below for detailed methods used to create and analyze test sets for model comparison.

Data used for model training and testing, inclusion and exclusion criteria

Consecutive years of data within the 2013-2022 period were used for model training and testing performed for each center independently of the others. Each center's Univfy report usage period started in 2016-2019 with data collection ending in 2020-2022 (Table 1). To assess the risk of data drift, we performed post-deployment, live model validation (LMV) per center, using an out-of-time test set from a time period following and exclusive from the MLCS1 training and test data (24, 34, 55, SI Figure 1). Using a larger, more recent, historical data set, each center's first model (MLCS1) was replaced by an updated model (MLCS2) using the same MLCS model life cycle and evaluation (SI Figure 1). The MLCS2 models were in clinical use at the time of writing this article.

Inclusion criteria for model training and testing were: 1) IVF cycles started, 2) IVF cycles using/ intending to use the patient's own eggs and uterus. Exclusion criteria were: FETs not linked to an original IVF ovarian stimulation cycle within that dataset; cancellations for reasons unrelated to IVF (e.g. covid, personal reasons, etc.); the use of donor egg, gestational carrier, embryo donation, egg freezing or fertility preservation for any reason; cycles without the female patient's age or outcome; cycles with age 42 and up; freeze-all cycles that have not had any FETs and batched cycles; PGT-M usage. IVF cycles that resulted in cancellation of egg retrieval, no blastocyst or euploid blastocyst available for transfer or PGT-A usage were not excluded.

IVF labeling criteria in the data used for MLCS training and testing

IVF cycles were labeled as having "no live birth outcome" if they resulted in cancellation at any point due to reasons related to the IVF cycle (e.g. poor ovarian response, thin endometrium, no viable oocytes, no fertilization, no blastocyst for transfer or no euploid blastocyst, etc.). In the context of MLCS models, an IVF cycle has achieved a live birth outcome if at least one live birth or clinical ongoing pregnancy were documented from one or more fresh and/or frozen embryo transfer(s) using the number of blastocyst(s) according to ASRM guidelines (56, 57).

To be clear, while MLCS and SART pretreatment models had live birth outcomes as primary outcomes for model prediction, live birth outcome itself is not the primary outcome in this study. Rather, this study's primary outcome is to evaluate MLCS and SART models using model performance metrics. However, comparison of models' metrics requires providing methods used to process data, train and test MLCS pretreatment models.

Model predictors

The model predictors used by each center's MLCS2 model and their relative importance vary across centers, despite drawing from a similar set of clinical variables (27-29, SI Figure 1). The SART pretreatment model predictors were either also used by MLCS2 models or were determined to have no non-redundant predictive impact based on other predictors used by the MLCS2 models. The MLCS2 models used additional predictors not used by the SART model even though they were recorded in SART-CORS (24, 26, SI Figure 1).

MLCS enables greater flexibility in the number and scope of clinical predictors that can be tested for use in the prediction model to capture greater inter-patient differences. Although the MLCS and SART model training sets were comparable in including female patient's age, BMI, clinical diagnoses, and reproductive history, the MLCS models used one or more ovarian reserve tests (e.g. AMH, D3 FSH, or AFC) reflecting each center's practice without being affected by inter-center laboratory differences irrelevant to each center. However, AMH value was available in ~95% of cycles. (SI Figure 1).

Creating center-specific, de novo model validation test sets (DNMV1 and DNMV2) to enable comparison of MLCS2 and SART models

We adapted each of the 6 center-specific MLCS2 test sets to a center-specific, de novo model validation test set (DNMV1) that allows validation of MLCS2 and SART models for each center. This adaptation was performed by limiting the test set to first IVF cycles started in 2013-2022; age under 40 (required by MLCS model); IVF cycles having BMI value and yes/no for male factor, ovulatory disorder, PCOS, uterine factor and unexplained infertility diagnoses (required by SART model); yes/no for full term birth(s). The 6 DNMV1 test sets together comprised 4,645 first IVF cycles.

For the purpose of MLCS2 and SART model comparison, for each center, de novo MLCS2 model validation was performed by obtaining MLCS2 model responses for each center's DNMV1 test set. Similarly, each center's own SART model responses (aka predictions) were obtained by using pre-treatment model formulae, with and without AMH predictor, reportedly used to support the online SART calculator, as provided in the supplement of McLernon et al., 2022 (26). Other than to confirm the accurate implementation of those formulae using a few test cases specifically for this study, we did not interact with or use the online SART calculator website for this research study.

In the context of IVF live birth prediction models, model training required labeling each IVF cycle has having the binary outcomes "live birth" versus "no live birth". One intentional design difference between MLCS and SART models is that in MLCS2, the models are trained with live births and clinical ongoing pregnancies labeled as "live birth", whereas the SART model labeled live births as "live birth" and clinical ongoing pregnancies as "no live birth" (26).

Since the impact of this intentional design difference was not known, we reasoned that we should create test sets void of IVF cycles with clinical ongoing pregnancies, so that we could be sure any model comparison results are not simply due to those cycles having differential outcome labeling between MLCS and SART models. Therefore, we created a second set of 6 independent test sets (DNMV2) by removing the ~4.8% of IVF cycles with clinical ongoing pregnancies from the DNMV1 test sets.

Statistical analyses involving model metrics used to compare MLCS2 and SART models

The MLCS2 and SART models were compared using six independent, center-specific DNMV1 test sets for Precision, Recall, Precision Recall AUC (PR AUC), and F1 scores to allow for reporting of median and interquartile range of the model metrics, whereas one aggregate DNMV1 test set was used to compare MLCS and SART models for continuous net reclassification improvement (NRI). The entire process of model comparisons -- six independent center-specific DNMV test sets and one aggregate DNMV test set -- were repeated using DNMV2 test set (24, 33, 53-55, 58). Model metrics were reported using median and interquartile range (IQR) across 6 centers. Wilcoxon signed-rank test, allowing for non-parametric paired-testing, was used to compare MLCS2 and SART model metrics paired by center (58).

Reclassification measured the percentage of cases having different live birth predictions from the two models (59, 60). To provide clinical context, we created a 4x4 reclassification table for the DNMV1 test set to show the concordance and discordance of LBPs made by MLCS2 and SART models. Patients are placed into one of 16 groups based on their LBPs as computed by MLCS2 and SART (MLCS2-LBPs and SART-LBPs, respectively), using 4 arbitrarily defined, yet clinically intuitive prognostic categories based on LBPs: Low LBP (LBP <25%), Medium LBP (LBP 25-49.9%), High LBP (LBP 50-74.9%) and Very High LBP (LBP ≥75%), (Table 4). Patient groups are concordant if SART and MLCS2 models placed the patients into the same prognostic category (colored green in Table 4); patient groups are discordant if SART and MLCS2 models placed patients into different prognostic categories (colored blue or peach in Table 4). The same procedure was applied to the DNMV2 test set.

Further, an alternative method, continuous net reclassification improvement (NRI) was used to measure the likelihood of re-assigning a higher or lower IVF live birth probability with MLCS2 compared to SART models (60). The age-based live birth rates for each clinic stated in the finalized 2020 SART National Summary were used as age control models because practically, that is the number that providers and patients would use if they were not using any prediction models (61).

The EQUATOR Reporting Guidelines including "TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods" were followed (62, 63).

Data Availability

The anonymized data sets containing clinical data are not available for sharing with other researchers because they are the property of each collaborating center.

Code Availability

The code for computing model metrics and comparisons of model metrics can be shared under a confidentiality agreement for research purposes only. Code from the data pre-processing, processing and model pipelines are proprietary and can be shared via licensing. However, we would be open to supporting or collaborating with other researchers to accelerate your research.

References

1. World Health Organization. Infertility Prevalence Estimates, 1990-2021. Global Report, 3 Apr. 2023, https://www.who.int/publications/iiitem/978920068315.
2. U.S. Department of Health and Human Services. (2024, March 13). Fact sheet: In vitro fertilization (IVF) use across the United States. Retrieved from https://www.hhs.gov/about/news/2024/03/13/fact-sheet-in-vitro-fertilization-M-use-across-united-states.html
3. Definition of Infertility: A Committee Opinion. (2023). *American Society for Reproductive Medicine.* https://www.asrm.org/practice- guidance/ practice-committee-documents/ denitions-of-infertility /
4. Yao M. Sept. 3, 2024. We are sharing infographics to bring together fertility care needs and ART stats reported by many researchers, epidemiologists, public health experts, demographers, and policy makers to provide an estimate of the current gap in ART usage. [Post]. LinkedIn. https://www.linkedin.com/posts/mylene-yao-m-d-049a2915_global-assisted-reproductive-technology-gap-activity-7234585961043505153-erKg/
5. Adamson G.D., Zegers-Hochschild F., Dyer S. (2023). Global fertility care with assisted reproductive technology, *Fertility and Sterility,* 120(3 Pt 1), 473-482. doi: 10.1016/j.fertnstert.2023.01.01
6. Fauser B.C.J.M., Adamson G.D., Boivin J., et al. (2024). Declining global fertility rates and the implications for family planning and family building: an IFFS consensus document based on a narrative review of the literature, *Human Reproduction Update,* 30(2), 153-173. doi: 10.1093/humupd/dmad028
7. Jenkins J, van der Poel S, Krüssel J, Bosch E, Nelson SM, Pinborg A, Yao MW. Empathetic application of machine learning may address appropriate utilization of ART. Reproductive BioMedicine Online 2020; 41:573-577.
8. Cox C.M., Thoma M.E., Tchangalova N., et al. (2022). Infertility prevalence and the methods of estimation from 1990 to 2021: a systematic review and meta-analysis, *Human Reproduction Open,* 2022(4). doi: 10.1093/hropen/hoac051
9. Boivin J., Bunting L, Collins J.A, Nygren KG. (2007). International estimates of infertility prevalence and treatment-seeking: potential need and demand for infertility medical care, *Human Reproduction,* 22(6), 1506-1512. doi: 10.1093/humrep/dem046
10. ESHRE. ART Fact Sheet. Nov. 2023. https://www.eshre.eu/Europe/Factsheets-and-infographics
11. Adamson G.D., Dyer S., Chambers G., et al. (2022) International Committee for Monitoring Assisted Reproductive Technology: World Report on Assisted Reproductive Technology, 2018. https://www.icmartivf.org/reports-publications/presentations/

12. Adamson G.D., de Mouzon J., Chambers G., et al. (2022) International Committee for Monitoring Assisted Reproductive Technology: World Report on Assisted Reproductive Technology, 2019. https://www.icmartivf.org/reports-publications/presentations/
13. Zegers-Hochschild F.,Adamson G.D., Dyer S., et al. (2017). The International Glossary on Infertility and Fertility Care, 2017, Fertility and Sterility, 108(3), 393-406. http:/ /dx.doi.org/10.1016/j.fertnstert.2017.06.005.
14. Bunting L.,Tsibulsky I., Boivin J. (2013). Fertility knowledge and beliefs about fertility treatment: findings from the International Fertility Decision-making Study, *Human Reproduction,* 28(2), 385-397. doi: 10.1093/humrep/des402
15. Collura B, Hayward B, Modrzejewski KA, Mottla GL, Richter KS, Catherino AB. Identifying Factors Associated with Discontinuation of Infertility Treatment Prior to Achieving Pregnancy: Results of a Nationwide Survey. Journal of Patient Experience. 2024;11. doi:10.1177/23743735241229380
16. Yao MWM, Nguyen ET, Retzloff MG, Cadesky K, Gago LA, Copland S 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. https://doi.org/10.3390/jcm13123560.
17. Malizia BA, Hacker MR, Penzias AS. Cumulative live-birth rates after in vitro fertilization. N Engl J Med. 2009 Jan 15;360(3):236-43. doi: 10.1056/NEJMoa0803072. PMID: 19144939.
18. Smith ADAC, Tilling K, Nelson SM, Lawlor DA. Live-Birth Rate Associated With Repeat In Vitro Fertilization Treatment Cycles. JAMA. 2015 Dec 22-29;314(24):2654-2662. doi: 10.1001/jama.2015.17296. PMID: 26717030; PMCID: PMC4934614.
19. Banerjee P, Choi B, Shahine LK, Jun SH, O'Leary K, Lathi RB, Westphal LM, Wong WH, Yao MW. Deep phenotyping to predict live birth outcomes in in vitro fertilization. Proc Natl Acad Sci U S A. 2010 Aug 3;107(31):13570-5. doi: 10.1073/pnas.1002296107. Epub 2010 Jul 19. PMID: 20643955; PMCID: PMC2922227.
20. 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 Aug;104(2):325-32. doi: 10.1016/j.fertnstert.2015.04.032. Epub 2015 May 21. PMID: 26003269.
21. Choi B, Bosch E, Lannon BM, Leveille MC, Wong WH, Leader A, Pellicer A, Penzias AS, Yao MW. Personalized prediction of first-cycle in vitro fertilization success. Fertil Steril. 2013 Jun;99(7):1905-11. doi: 10.1016/j.fertnstert.2013.02.016. Epub 2013 Mar 21. PMID: 23522806.
22. Lannon BM, Choi B, Hacker MR, Dodge LE, Malizia BA, Barrett CB, Wong WH, Yao MW, Penzias AS. Predicting personalized multiple birth risks after in vitro fertilization-double embryo transfer. Fertil Steril. 2012 Jul;98(1):69-76. doi: 10.1016/j.fertnstert.2012.04.011. Epub 2012 Jun 4. PMID: 22673597.
23. Chen SH, Xie YA, Cekleniak NA, Keegan DA, Yao MWM. In search of the crystall 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(3):E83-E84 Supp. doi: 10.1016/j.fertnstert.2019.07.339.
24. Yao MWM, J Jenkins, Nguyen ET, Swanson T, Menabrito M. Patient-centric IVF prognostics counseling using machine learning for the pragmatist. Semin Repro Med 2024, *accepted*.
25. Swanson T, Yao MWM, Retzloff M, Gago LA, Copland S, Nichols JE et al. Inter-center variation of patients' clinical profiles is associated with IVF live birth outcomes. Fertil Steril 2023;120(4):E175. doi: 10.1016/j.fertnstert.2023.08.517.
26. McLernon DJ, Raja EA, Toner JP, Baker VL, Doody KJ, Seifer DB, Sparks AE, Wantman E, Lin PC, Bhattacharya S, Van Voorhis BJ. Predicting personalized cumulative live birth following in vitro

fertilization. Fertil Steril. 2022 Feb;117(2):326-338. doi: 10.1016/j.fertnstert.2021.09.015. Epub 2021 Oct 19. PMID: 34674824.

27. Society for Assisted Reproductive Technology and University of Aberdeen. URL: https://w3.abdn.ac.uk/clsm/SARTIVF/ (last accessed May 10, 2024)
28. Cai J, Jiang X, Liu L, Liu Z, Chen J, Chen K, Yang X, Ren J. Pretreatment prediction for IVF outcomes: generalized applicable model or centre-specific model? Hum Reprod. 2024 Feb 1;39(2):364-373. doi: 10.1093/humrep/dead242. PMID: 37995380; PMCID: PMC10833083.
29. McLernon DJ, Steyerberg EW, Te Velde ER, Lee AJ, Bhattacharya S. Predicting the chances of a live birth after one or more complete cycles of in vitro fertilisation: population based study of linked cycle data from 113 873 women. BMJ. 2016 Nov 16;355:i5735. doi: 10.1136/bmj.i5735. PMID: 27852632; PMCID: PMC5112178.
30. Ratna MB, Bhattacharya S, McLernon DJ. External validation of models for predicting cumulative live birth over multiple complete cycles of IVF treatment. Hum Reprod. 2023 Oct 3;38(10):1998-2010. doi: 10.1093/humrep/dead165. PMID: 37632223; PMCID: PMC10546080.
31. 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(1):317. doi: 10.1186/s12967-019-2062-5. PMID: 31547822; PMCID: PMC6757430.
32. 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(1):476. doi: 10.1186/s12884-023-05775-3. PMID: 37370040; PMCID: PMC10294395.
33. 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 Mar 4;10(3):e0118432. doi: 10.1371/journal.pone.0118432. PMID: 25738806; PMCID: PMC4349800.
34. Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications and potential remedies. Br J Radiol. 2023 Oct;96(1150):20220878. doi: 10.1259/bjr.20220878. Epub 2023 Mar 27. PMID: 36971405; PMCID: PMC10546450.
35. Rajendran S, Brendel M, Barnes J, Zhan Q, Malmsten JE, Zisimopoulos P, Sigaras A, Ofori-Atta K, Meseguer M, Miller KA, Hoffman D, Rosenwaks Z, Elemento O, Zaninovic N, Hajirasouliha I. Automatic ploidy prediction and quality assessment of human blastocysts using time-lapse imaging. Nat Commun. 2024 Sep 5;15(1):7756. doi: 10.1038/s41467-024-51823-7. PMID: 39237547; PMCID: PMC11377764.
36. Illingworth PJ, Venetis C, Gardner DK, Nelson SM, Berntsen J, Larman MG, Agresta F, Ahitan S, Ahlström A, Cattrall F, Cooke S, Demmers K, Gabrielsen A, Hindkjær J, Kelley RL, Knight C, Lee L, Lahoud R, Mangat M, Park H, Price A, Trew G, Troest B, Vincent A, Wennerström S, Zujovic L, Hardarson T. Deep learning versus manual morphology-based embryo selection in IVF: a randomized, double-blind noninferiority trial. Nat Med. 2024 Aug 9. doi: 10.1038/s41591-024-03166-5. Epub ahead of print. PMID: 39122964.
37. Diakiw SM, Hall JMM, VerMilyea MD, Amin J, Aizpurua J, Giardini L, Briones YG, Lim AYX, Dakka MA, Nguyen TV, Perugini D, Perugini M. Development of an artificial intelligence model for predicting the likelihood of human embryo euploidy based on blastocyst images from multiple imaging systems during IVF. Hum Reprod. 2022 Jul 30;37(8):1746-1759. doi: 10.1093/humrep/deac131. PMID: 35674312; PMCID: PMC9340116.
38. Chen D, Liu S, Kingsbury P, Sohn S, Storlie CB, Habermann EB, Naessens JM, Larson DW, Liu H. Deep learning and alternative learning strategies for retrospective real-world clinical data. NPJ Digit Med. 2019 May 30;2:43. doi: 10.1038/s41746-019-0122-0. PMID: 31304389; PMCID: PMC6550223.

39. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, Pfeffer MA, Fries J, Shah NH. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med. 2023 Jul 29;6(1):135. doi: 10.1038/s41746-023-00879-8. PMID: 37516790; PMCID: PMC10387101.
40. Isabelle Bousquette. A Clamor for generative AI (even if something else works better). The Wall Street Journal. July 22, 2024. Available at: https://www.wsj.com/articles/a-clamor-for-generative-ai-even-if-something-else-works-better-d9bd0257 (Accessed on July 24, 2024)
41. RESOLVE - The National Infertility Association. Insurance coverage by state. June 17, 2024. https://resolve.org/learn/financial-resources-for-family-building/insurance-coverage/insurance-coverage-by-state/ (Accessed Sept. 9, 2024)
42. Ekechi C. Addressing inequality in fertility treatment. Lancet 2021;398(10301):645-646. doi: 10.1016/S0140-6736(21)01743-8
43. Hariton E, Alvero R, Hill MJ, Mersereau JE, Perman S, Sable D et al. Meeting the demand for fertility services: the present and future of reproductive endocrinology and infertility in the United States. Fertil Steril 2023 Oct;120(4):755-766. doi: 10.1016/j.fertnstert.2023.08.019. Epub 2023 Sep 4. PMID: 37665313.
44. 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):P403-405. doi: 10.1016/j.fertnstert.2023.01.007.
45. Klipstein S. The role of compassionate reproductive care and counseling in the face of futility. Fertil Steril 2023;120(3):P409-411. doi: 10.1016/j.fertnstert.2023.01.012
46. Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. Nat Biomed Eng. 2022 Dec;6(12):1330-1345. doi: 10.1038/s41551-022-00898-y. Epub 2022 Jul 4. PMID: 35788685.
47. Richard-Davis G, Morris J. No longer separate but not close to equal: navigating inclusivity in a burgeoning field built on injustice. Fertil Steril 2023;120(3):P400-402. doi: 10.1016/j.fertnstert.2022.11.013.
48. Ekechi C. Addressing inequality in fertility treatment. Lancet. 2021 Aug 21;398(10301):645-646. doi: 10.1016/S0140-6736(21)01743-8. Epub 2021 Aug 2. PMID: 34352200.
49. ESHRE Add-ons working group; Lundin K, Bentzen JG, Bozdag G, Ebner T, Harper J, Le Clef N, Moffett A, Norcross S, Polyzos NP, Rautakallio-Hokkanen S, Sfontouris I, Sermon K, Vermeulen N, Pinborg A. Good practice recommendations on add-ons in reproductive medicine†. Hum Reprod. 2023 Nov 2;38(11):2062-2104. doi: 10.1093/humrep/dead184. PMID: 37747409; PMCID: PMC10628516.
50. 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 Nov;39(11):2607-2616. doi: 10.1007/s10815-022-02634-6. Epub 2022 Oct 21. PMID: 36269502; PMCID: PMC9722991.
51. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of Statistics 2001;29(5):1189-1232.
52. Univfy Inc. 2024. Research Support - Sample Patient Counseling Report. https://www.univfy.com/research (Accessed on Sept. 9, 2023)
53. Cross validation in machine learning. Dec. 21, 2023. https://www.geeksforgeeks.org/cross-validation-machine-learning/ (Accessed 2024).
54. Srivastava T. 12 Important model evaluation metrics for machine learning everyone should know (updated 2023). https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/ (Accessed Jun 1, 2024).

55. https://towardsdatascience.com/why-isnt-out-of-time-validation-more-ubiquitous-7397098c4ab6
56. Practice Committee of the American Society for Reproductive Medicine and the Practice Committee for the Society for Assisted Reproductive Technologies. Guidance on the limits to the number of embryos to transfer: a committee opinion. Fertil Steril 2021;116:651-54.
57. Practice Committee of the American Society for Reproductive Medicine and the Practice Committee for the Society for Assisted Reproductive Technologies. Guidance on the limits to the number of embryos to transfer: a committee opinion. Fertil Steril 2017;107:901-3.
58. Kundu S, Aulchenko YS, van Duijn CM, Janssens AC. PredictABEL: an R package for the assessment of risk prediction models. Eur J Epidemiol. 2011 Apr;26(4):261-4. doi: 10.1007/s10654-011-9567-4. Epub 2011 Mar 24. PMID: 21431839; PMCID: PMC3088798.
59. Cook NR, Paynter NP. Performance of reclassification statistics in comparing risk prediction models. Biom J. 2011 Mar;53(2):237-58. doi: 10.1002/bimj.201000078. Epub 2011 Feb 3. PMID: 21294152; PMCID: PMC3395053.
60. 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 Jan;25(1):114-21. doi: 10.1097/EDE.0000000000000018. PMID: 24240655; PMCID: PMC3918180.
61. Society for Assisted Reproductive Technology. URL: www.sart.org (last accessed May 10, 2024)
62. EQUATOR Network. Enahance the QUAlity and Transparency Of health Research. URL: equator-network.org (last accessed May 10, 2024)
63. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, Boulesteix AL, Camaradou JC, Celi LA, Denaxas S, Denniston AK, Glocker B, Golub RM, Harvey H, Heinze G, Hoffman MM, Kengne AP, Lam E, Lee N, Loder EW, Maier-Hein L, Mateen BA, McCradden MD, Oakden-Rayner L, Ordish J, Parnell R, Rose S, Singh K, Wynants L, Logullo P. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. URL: https://www.equator-network.org/reporting-guidelines/tripod-statement/ (last accessed May 10, 2024)

Keywords
live birth probability, IVF live birth prediction, artificial intelligence, machine learning, SART, fertility prognosis

Acknowledgements
The authors thank the following individuals for their assistance, editing, advisory, insightful comments and contributions to the present research: Faith Ripley, BS, CPC (PREG); Patrick McCarthy, MBA (Poma Fertility); Amanda McCarthy, MBA (Poma Fertility); Brijinder S. Minhas, PhD, HCLD, MBA (NewLIFE); Wing H. Wong, PhD (Advisor); Vincent Kim, B.Sc. (Univfy Inc.); Marco Menabrito, MD (Univfy Inc.); Anjali Wignarajah, M.Sc. (Univfy Inc.); Candice Ortego (Univfy Inc.); Athena T. R. Wu (editing).

Funding
Each organization funded its own participation.

Authorship Contribution
ETN and MWMY contributed to the original study conceptualization, design, methodology, preparation of data subsets for analyses, statistical analyses, data interpretation and visualization;

writing of the original manuscript drafts and revisions. ETN, TS, XC contributed to code development, data processing, modeling, and methodologies related to those processes. ETN, MWMY, TS, XC contributed to curation of processed data. MGR, LAG, JEN, JFP, BAR, MO, JG, RB, LN, GN, JA contributed to conceptualization of study, data collection, curation of data, interpretation of data processing and modeling results, review and revision of manuscript. All authors approved submission of the manuscript for publication.

Ethics Declaration

M Yao is employed as CEO by Univfy Inc. and is board director, shareholder and stock optionee of Univfy; she is inventor or co-inventor on Univfy's issued and pending patents and receives payment from patent licensor (Stanford University). ET Nguyen, T Swanson, X Chen are employed by and received stock options from Univfy Inc. M Retzloff performs paid consulting work as Nexplanon trainer for Organon and is Treasurer for the Society for Reproductive Technology (SART).

Supplementary Information

Supplementary Figure 1.

Tables and Figures

Table 1. This table shows the time period, number of years and IVF volume represented by each data set matched against the MLCS model being tested. We also indicated whether (i) the dataset was used for both training and testing or testing only and (ii) the model validation was cross validation (CV) using in-time data or live model validation (LMV) using out-of-time data.

Clinic	MLCS-based PreIVF model tested	Attributes of dataset used			Data use: both (train and test) or test only	Model Validation Type	
		Time period	Number of years	Number of IVF cycles (range)		CV or LMV	In-time or out-of-time
916	MLCS 1	2014-2016	3	501-1000	both	CV	in-time
	MLCS 2	2014-2020	7	1001-2000	both	CV	in-time
	MLCS 1	2017-2020	4	501-1000	test only	LMV	out-of-time
552	MLCS 1	2014-2016	2.5	101-200	both	CV	in-time
	MLCS 2	2014-2020	7	301-500	both	CV	in-time
	MLCS 1	2016-2020	4.5	101-200	test only	LMV	out-of-time
635	MLCS 1	2013-2016	4	501-1000	both	CV	in-time
	MLCS 2	2013-2020	8	1001-2000	both	CV	in-time
	MLCS 1	2017-2020	4	501-1000	test only	LMV	out-of-time
189	MLCS 1	2014-2018	5	301-500	both	CV	in-time
	MLCS 2	2014-2020	7	501-1000	both	CV	in-time
	MLCS 1	2019-2020	2	201-300	test only	LMV	out-of-time
869	MLCS 1	2014-2018	5	501-1000	both	CV	in-time
	MLCS 2	2016-2020	5	501-1000	both	CV	in-time
	MLCS 1	2019-2020	2	201-300	test only	LMV	out-of-time
395	MLCS 1	2013-2018	6	501-1000	both	CV	in-time
	MLCS 2	2013-2021	9	1001-2000	both	CV	in-time
	MLCS 1	2019-2021	2	501-1000	test only	LMV	out-of-time

Abbreviations: MLCS1 = machine learning, center-specific model version 1; MLCS2 = machine learning, center-specific model version 2; CV = cross validation using in-time data; LMV = live model validation, a term we coined for external validation or the use of out-of-time data in model validation.

Table 2. This table shows the median and interquartile range (IQR) for cross validation and live model validation metrics -- AUC, AUC Improvement over Age-only model and PLORA -- for MLCS1 and MLCS2 models across 6 centers. See Methods for source of the age-only control model (61).

		Model validation results: median and IQR		
Model	**Validation**	**AUC**	**AUC Improvement compared to age-only control**	**PLORA**
MLCS 1	CV, in-time	0.66 (IQR = 0.61, 0.68)	41.0% (IQR = 32.8%, 49.6%)	7.2 (IQR = 3.6, 11.8)
MLCS 2	CV, in-time	0.67 (IQR = 0.66, 0.68)	44.1% (IQR = 36.4%, 50.2%)	23.9 (IQR = 10.2, 39.4)
MLCS 1	LMV, out-of-time	0.65 (IQR = 0.63, 0.66)	7.6% (IQR = 5.2%, 8.7%)	6.7 (IQR = 2.2, 12.0)

Abbreviations: MLCS1 = machine learning, center-specific model version 1; MLCS2 = machine learning, center-specific model version 2; PLORA = posterior log odds ratio compared to age.

Table 3. This table shows the median and interquartile range (IQR) for model cross validation metrics -- AUC ROC and PLORA -- measured by testing each center's MLCS2 and SART models using each center's de novo model validation test sets (DNMV1 and DNMV2) and using the SART 2020 age group-based live birth rate for each clinic as the age control "model". *MLCS2 models showed significantly higher PR AUC score than SART, $p<0.05$.

Test set for cross validation			**Model validation results: median and IQR (in parenthesis)**			
Test set	**In-time or out-of-time data**	**Model**	**AUC ROC**	**PLORA**	**F1 score at 50% LBP threshold**	**PR AUC***
DNMV1 test set (N=4,645) includes IVF cycles with clinical pregnancy outcomes	in-time	MLCS2	0.64 (0.62, 0.66)	28.1 (15.2, 49.4)	0.74 (0.72, 0.78)	0.75 (0.73, 0.77)
	out-of-time	SART	0.65 (0.63, 0.66)	22.5 (15.8, 46,5)	0.71 (0.68, 0.73)	0.69 (0.68, 0.71)
DNMV2 test set (N=4,421) excludes IVF cycles with clinical pregnancy outcomes	in-time	MLCS2	0.64 (0.62, 0.66)	23.5 (13.3, 33.9)	0.74 (0.71, 0.75)	0.73 (0.73, 0.75)
	out-of-time	SART	0.65 (0.62, 0.66)	20.2 (14.9, 40.9)	0.69 (0.67, 0.72)	0.68 (0.66, 0.69)

Abbreviations: AUC = area under the curve; ROC = receiver-operator curve analysis; PLORA = posterior log odds ratio compared to age control model; MLCS2 = machine learning, center-specific model version 2; DNMV = de novo model validation test set adapted to enable testing of MLCS2 and SART models; DNMV1 = the DNMV test set which included IVF cycles with clinical pregnancy outcomes; DNMV2 = the DNMV test set which excluded IVF cycles with clinical pregnancy outcomes; PR AUC = Precision-Recall AUC

Table 4. Reclassification table comparing IVF live birth prediction (LBP) models MLCS2 and SART across centers for (A) DNMV1 and (B) DNMV2.
A. Reclassification table using the DNMV1 test set (aggregate of 6 center-specific test sets) showing the number of IVF cycles receiving predicted IVF live birth prediction (LBP) in each range by the SART model vs. MLCS2 model.

Abbreviations: MLCS2 = machine learning, center-specific model version 2; DNMV = de novo model validation test set adapted to enable testing of MLCS2 and SART models ; DNMV1 = the DNMV test set which included IVF cycles with clinical pregnancy outcomes; DNMV2 = the DNMV test set which excluded IVF cycles with clinical pregnancy outcomes; N=the number of patients being placed into each of the 16 groups. LBR = live birth rate of each group. B. The same procedure was applied to the DNMV2 test set.

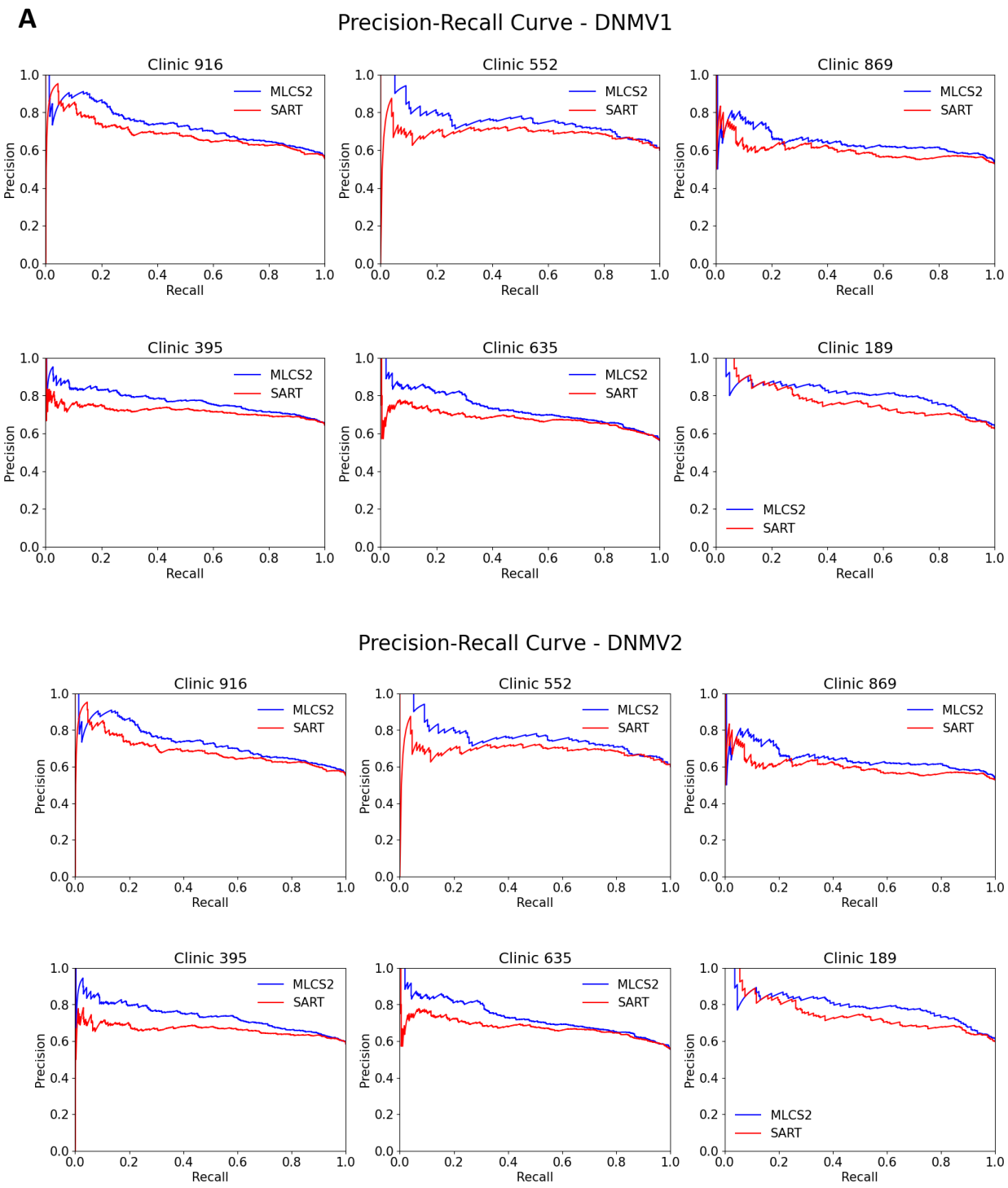
		MLCS2 Model Prognostic Categories (LBP range)				
		Low LBP < 25%	**Medium LBP 25% to 49.9%**	**High LBP 50% to 74.9%**	**Very High LBP ≥ 75%**	**Subtotal**
SART Model Prognostic Categories (LBP range)	**Low LBP < 25%**	N = 79 LBR = 9%	N = 168 LBR = 33%	N = 2 LBR = 50%	N = 0	N = 249 LBR = 26%
	Medium LBP 25% to 49.9%	N = 12 LBR = 0%	N = 735 LBR = 38%	N=566 LBR = 61%	N = 7 LBR = 100%	N = 1320 LBR = 48%
	High LBP 50% to 74.9%	N = 0	N = 144 LBR = 37%	N=2445 LBR = 64%	N = 487 LBR = 80%	N = 3076 LBR = 66%
	Very High LBP ≥ 75%	N = 0	N = 0	N = 0	N = 0	N = 0
	Subtotal	N = 91 LBR = 8%	N = 1047 LBR = 37%	N = 3013 LBR = 64%	N = 494 LBR = 81%	N = 4645 LBR = 59%

Table 4 (Cont'd.)

B. The same reclassification procedure (above) was applied to the DNMV2 test set.

		MLCS2 Model Prognostic Categories (LBP range)				
		Low LBP < 25%	**Medium LBP 25% to 49.9%**	**High LBP 50% to 74.9%**	**Very High LBP ≥ 75%**	**Subtotal**
SART Model Prognostic Categories (LBP range)	**Low LBP < 25%**	N = 78 LBR = 8%	N = 163 LBR = 31%	N = 1 LBR = 0%	N = 0	N = 242 LBR = 24%
	Medium LBP 25% to 49.9%	N = 12 LBR = 0%	N = 726 LBR = 37%	N=525 LBR = 58%	N = 7 LBR = 100%	N = 1270 LBR = 46%
	High LBP 50% to 74.9%	N = 0	N = 143 LBR = 36%	N=2336 LBR = 63%	N = 430 LBR = 78%	N = 2909 LBR = 64%
	Very High LBP ≥ 75%	N = 0	N = 0	N = 0	N = 0	N = 0
	Subtotal	N = 90 LBR = 7%	N = 1032 LBR = 36%	N = 2862 LBR = 62%	N = 437 LBR = 78%	N = 4421 LBR = 56%

Figure 1. Comparison of MLCS2 and SART models using (A) Precision-Recall curves for each of the 6 clinics using each center's de novo model validation test sets, DNMV1 and DNMV2; (B) frequency distributions of live birth probabilities using MLCS2 and the SART models for DNMV1 and DNMV2 test sets each comprising data from 6 centers' test sets in aggregate.



MLCS2 = machine learning, center-specific model version 2
DNMV1 = de novo model validation test set 1 (includes IVF with clinical pregnancy outcomes)
DNMV2 = de novo model validation test set 1 (excludes IVF with clinical pregnancy outcomes)
Precision, Recall and Precision-Recall Curve – see glossary and SI Methods for explainer

B

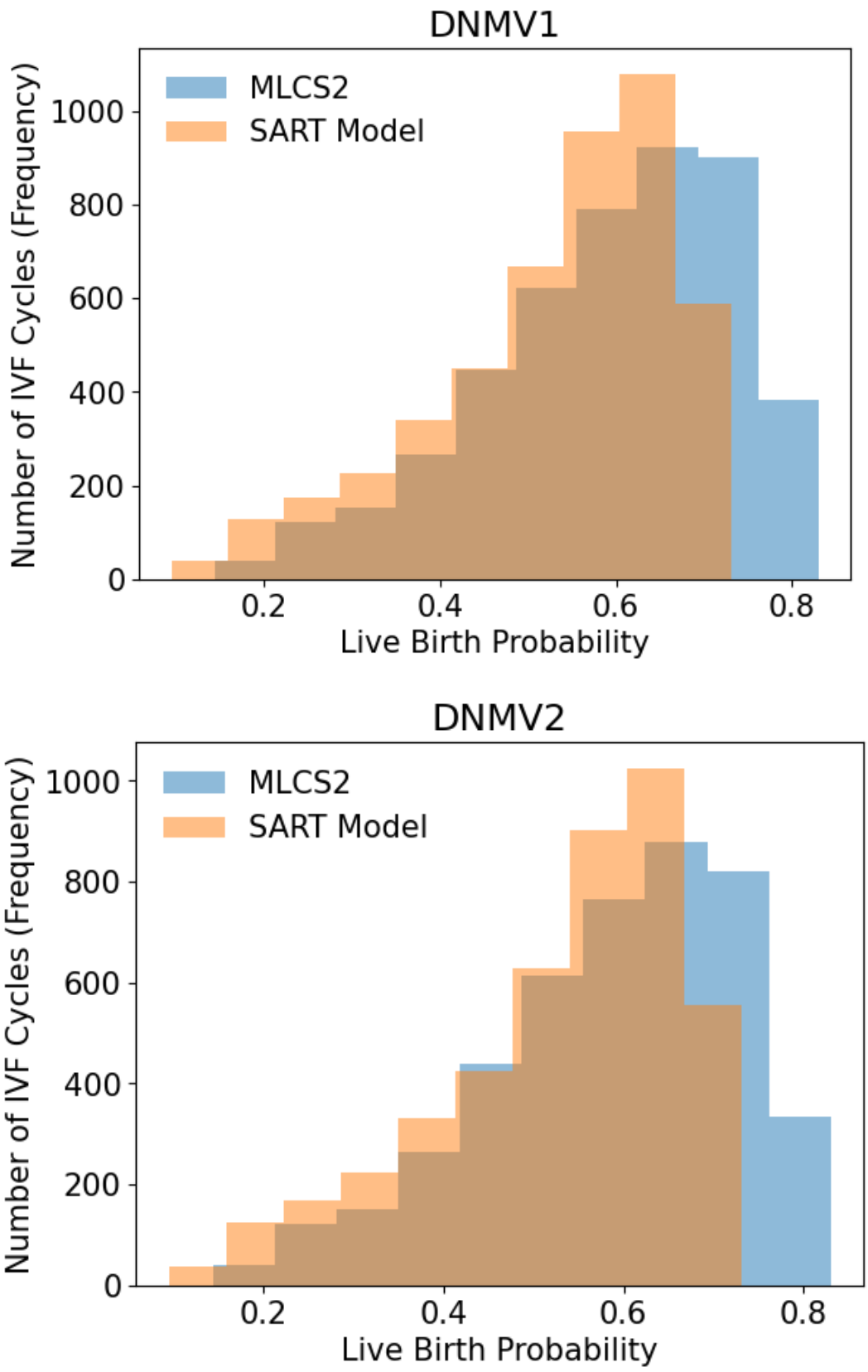
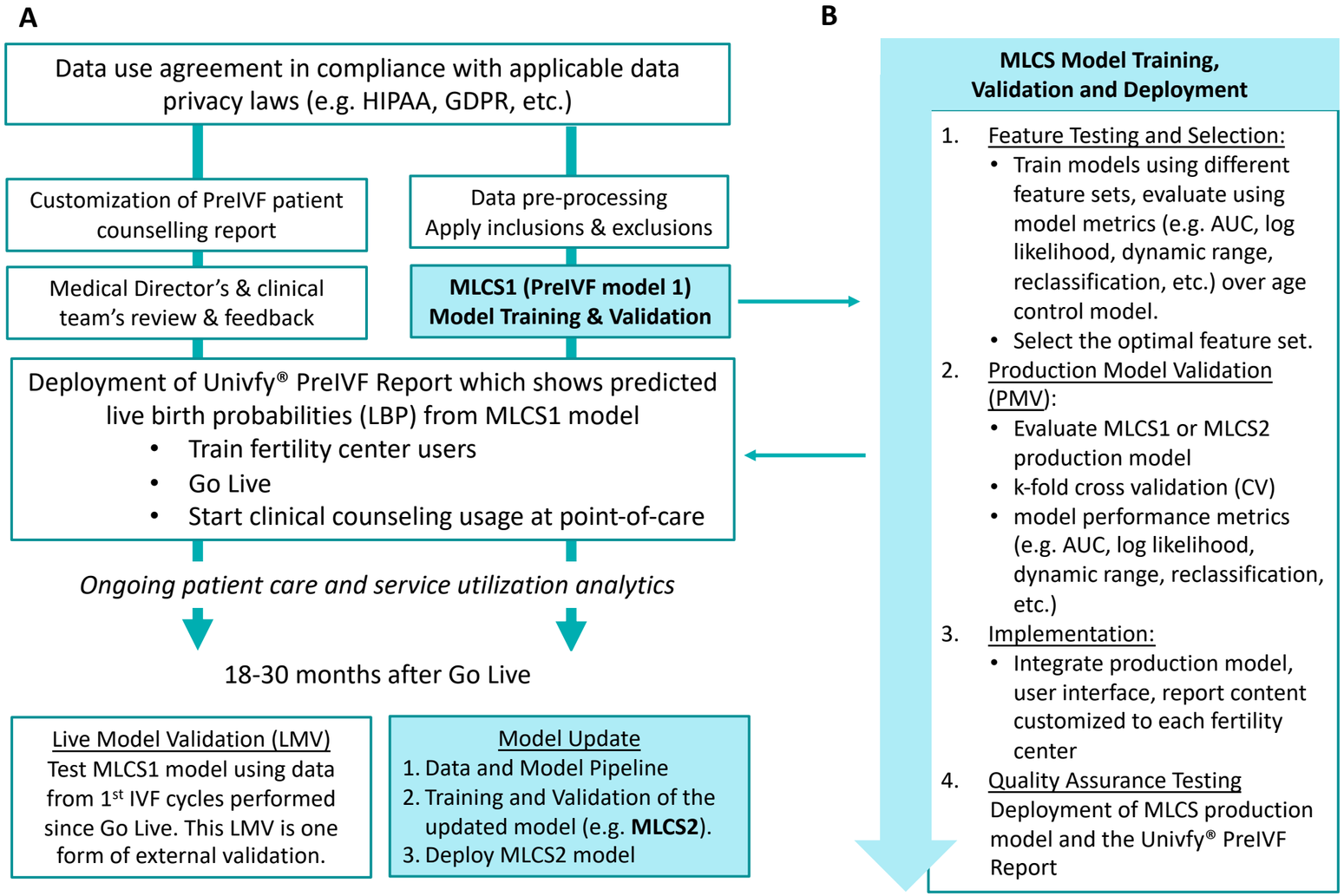

DNMV1
MLCS2
SART Model
Number of IVF Cycles (Frequency)
0
200
400
600
800
1000
0.2
0.4
0.6
0.8
Live Birth Probability
DNMV2
MLCS2
SART Model
Number of IVF Cycles (Frequency)
0
200
400
600
800
1000
0.2
0.4
0.6
0.8
Live Birth Probability

Figure 2. 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*. (A) The MLCS-based, PreIVF model (MLCS model) product life cycle comprises the steps of data pre-processing, model training and validation, deployment and post-deployment validation (or live model validation). MLCS1, MLCS2, etc. indicates that each MLCS model will be replaced by an updated MLCS model trained and tested with a more recent data set which may also become cumulatively larger. (B) Model pipeline supports feature testing, model training, validation analysis, deployment to production and quality testing. "MLCS" is used generically to indicate the steps used for MLCS1, MLCS2 or any subsequent updates of MLCS model for a particular fertility center.



SI Figure 1. MLCS2 model predictors and their relative importance, cross-referenced with SART pretreatment model predictors*. (1-3) The model predictors used by each center's MLCS2 model and their relative importance vary across centers, despite drawing from a similar set of clinical variables. The SART pretreatment model predictors were either also used by MLCS2 models or were determined to have no non-redundant predictive impact based on other predictors used by the MLCS2 models. The MLCS2 models used additional predictors not used by the SART model even though they were recorded in SART-CORS. See Yao et al, 2024 for more in-depth discussion of model predictors and relative importance (1).

Abbreviations.
MLCS2 = machine learning, center-specific model version 2
SART = Society of Assisted Reproductive Technologies
AMH = serum anti-mullerian hormone level
BMI = body mass index
Day 3 FSH = serum day 3 follicular stimulating hormone level
PCOS = polycystic ovarian syndrome
PCO = polycystic ovaries
IVF = in vitro fertilization

References.

1. Yao MWM, J Jenkins, Nguyen ET, Swanson T, Menabrito M. Patient-centric IVF prognostics counseling using machine learning for the pragmatist. Semin Repro Med 2024, *accepted*.

2. McLernon DJ, Raja EA, Toner JP, Baker VL, Doody KJ, Seifer DB, Sparks AE, Wantman E, Lin PC, Bhattacharya S, Van Voorhis BJ. Predicting personalized cumulative live birth following in vitro fertilization. Fertil Steril. 2022 Feb;117(2):326-338. doi: 10.1016/j.fertnstert.2021.09.015. Epub 2021 Oct 19. PMID: 34674824.

3. 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 Nov;39(11):2607-2616. doi: 10.1007/s10815-022-02634-6. Epub 2022 Oct 21. PMID: 36269502; PMCID: PMC9722991.