Running title:
IVF live birth prediction for the pragmatist

Title:
Predicting IVF live birth probabilities using machine learning, center-specific and national registry-based models.

Elizabeth T. Nguyen PhD[a], Matthew G. Retzloff MD[b], L. April Gago MD[c], John E. Nichols MD[d], John F. Payne MD[d], Barry A. Ripps MD[e], Michael Opsahl MD[f], Jeremy Groll MD[g], Ronald Beesley MD[f], Lorie Nowak PhD[g], Gregory Neal MD[b], Jaye Adams MD[b], Trevor Swanson PhD[a], Xiaocong Chen MSc[a], Mylene W. M. Yao MD[a].

[a]R&D Department, Univfy, Los Altos, CA, US.
[b]Fertility Center of San Antonio, San Antonio, TX, US.
[c]Gago Center for Fertility, Brighton, MI, US.
[d]Piedmont Reproductive Endocrinology Group, Greenville, SC, US.
[e]NewLIFE Fertility, Pensacola, FL, US.
[f]Poma Fertility, Kirkland, WA, US.
[g]SpringCreek Fertility, Dayton, OH, US.

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

Article type: Observational Study - Study of prognostics model

Funding statement: Each organization funded its own participation.

CRediT Authorship Contribution Statement
Elizabeth T. Nguyen: Writing - original draft, Writing - review & editing, Methodology, Investigation, Conceptualization, Formal analysis, Data curation, Validation, Software, Visualization. **Matthew G. Retzloff, L. April Gago, John E. Nichols, John F. Payne, Barry A. Ripps, Michael Opsahl, Jeremy Groll, Ronald Beesley, Lorie Nowak, Gregory Neal, Jaye Adams:** Writing - review, Data curation, Conceptualization. **Trevor Swanson and Xiaocong Chen:** Validation, Software, Methodology, Data curation. **Mylene Yao:** Writing - original draft, Writing - review & editing, Methodology, Investigation, Conceptualization, Project administration, Visualization, Resources.

Disclosure Statement / Declaration of Interests:
M Yao 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 co-inventor on Univfy's issued and pending patents.
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 SART.

Attestation Statement:
Data pertaining to aggregate analyses could be made available to the editors of the journal for review or query upon request.

Data Sharing Statement:
Raw data are not available for sharing with other researchers. However, we can support or collaborate with other researchers to apply similar methods to test other researchers' data.

Word Count: 350 words for abstract and 3247 words in the text

Capsule:
Machine learning-based, center-specific models predicted higher IVF live birth probabilities and improved validation metrics compared to a national registry-based multicenter model for 6 geographically distributed fertility centers in the US.

Structured Abstract

Objective:
To compare the performance of machine learning based, center-specific (MLCS) models and the US national registry-based, multicenter model (SART model) in predicting IVF live birth probabilities (LBPs) for 6 unrelated, geographically diverse US fertility centers.

Design:
Retrospective observational design.

Subjects:
Test sets comprised first IVF cycle data (2013-2022) extracted from a retrospective cohort of 4,645 patients at 6 fertility centers.

Intervention or Exposure:
The initial (MLCS1) and updated (MLCS2) models were compared against age control. MLSC2 and SART models were compared.

Main Outcome Measures:
Model validation metrics, reported in median and interquartile range (IQR), were compared using Wilcoxon signed-rank test: ROC AUC, posterior log-likelihood of odds ratio compared to age (PLORA), Precision-Recall (PR) AUC, F1 score and continuous net reclassification improvement (NRI).

Results:
MLCS1 and MLCS2 models showed improved AUC and PLORA compared to age control; MLCS1 models were validated using out-of-time test data. MLCS2 models showed improved PLORA 23.9 (IQR 10.2, 39.4) compared to 7.2 (IQR 3.6, 11.8) for MLCS1, $p<0.05$. MLCS2 showed higher median PR AUC at 0.75 (IQR 0.73, 0.77) compared to 0.69 (IQR 0.68, 0.71) for SART, $p<0.05$. In addition, the median F1 Score was higher for MLCS2 compared to SART model across predicted live birth probability (LBP) thresholds sampled at deciles 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.

At these six centers, using 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%. That means for every 100 patients who will have a first IVF cycle live birth, using LBR ≥ 50% as threshold, the MLCS2 model can identify 9 more such patients without overcalling or overestimating LBPs compared to the SART model.

Conclusion:
MLCS models accurately assign higher IVF LBPs to more patients compared to the SART model at 6 US fertility centers. We recommend testing a larger sample of fertility centers to evaluate generalizability of MLCS model benefits.

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

Introduction

Despite the proven safety and efficacy of assisted reproductive technology (ART), patients' navigation of fertility care continues to be met with barriers limiting ART's family-building potential for millions of people worldwide. Providers have an important responsibility in providing accurate and meaningful 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. (1-3) (IVF is used broadly to mean ART, including the use of ovarian stimulation, ICSI, freeze-all, and fresh or frozen ETs.)

For over a decade, we have reported the development and clinical usage of ML-based, center-specific (CS) or MLCS IVF prognostic models to support provider-patient counseling. (1, 4-6) This MLCS approach has been successfully applied to fertility centers in diverse geographies with and without IVF insurance coverage mandates (e.g. US) or a mix of self-pay and government-paid IVF (e.g. Ontario, Canada; UK and EU). MLCS models have supported providers with validated, localized and personalized pre-treatment counseling regarding first IVF treatment, repeat IVF after one or more failed IVF treatments (also called post-treatment), egg freezing, donor egg IVF treatment and elective single embryo transfer (eSET). (1, 4-8) We have reported methods and validation to demonstrate improved model performance of MLCS over control models including the ability to reclassify more patients to having higher live birth probabilities. Overall, the MLCS approach is expected to provide more locally relevant prognostic information as it is unaffected by inter-center variations in patients' attributes and clinical or embryology laboratory protocols. (1, 4-5, 9-12) Further, conventional ML methods have remained comparable or even superior to deep learning methods when applied to train structured healthcare data. (13)

Nonetheless, there is a perception that multicenter, registry-based IVF prognostics models as exemplified by the "McLernon models" -- US Society for Assisted Reproductive Technology (SART) pretreatment model (aka SART calculator or SART model) and the "UK McLernon 2022 model" -- are

"sufficient". Both the SART calculator and UK McLernon 2022 model were developed using large data sets and are accessible to the public via online calculator websites. (14-18)

The limited meaning of ROC-AUC notwithstanding, we have reported MLCS models with AUCs comparable to those of registry-based models. For example, we reported an external validation of pretreatment MLCS models showing AUCs of 0.80 (US center, 2010) and 0.72 (UK center, 2015), which compared favorably to the AUCs of 0.73 and 0.71 reported for the SART 2021 model training (non-external validation) with and without AMH as predictor, respectively and AUC of 0.67 reported for the external validation of the UK McLernon 2022 model. (4-5, 14-16)

Using a single center's dataset comprising ~26K+ IVF cycles, Cai et al reported improved and more locally relevant model performance using the MLCS approach, refuting the recommendation by McLernon et al to develop center-specific models by recalibrating from the SART or UK McLernon models. (9) Many US providers have asked us to show the differential prognostic information provided by the SART calculator and an MLCS model and whether MLCS is applicable to small-to-midsize US fertility centers with much lower IVF volumes compared to the report by Cai et al. However, a head-to-head comparison between the MLCS and McLernon pretreatment models has not been performed for centers reporting to the US or UK registries. Addressing these questions will help us to develop best practices for IVF prognostic counseling, which is critical for advancing and expanding 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, individual small-to-midsize fertility centers operating in 22 locations across 9 states in 4 US regions (West, Southeast, Southwest and Midwest) with processed datasets comprising 4,645 IVF cycles in aggregate available for model evaluation. MLCS and SART models (with and without AMH as predictor, based on AMH availability) were evaluated using metrics including AUC-ROC, AUC improvement over age control, predictive power, precision, recall, F1 score, and precision-recall AUC. (19) We also addressed data drift, a scenario in which changes in the distribution of clinical attributes, relationship between predictors and treatment outcomes and/or the relevant importance of predictors occurring after a model is deployed causes a previously validated model to have decreased clinical relevance. (20) Our goal was to focus on objective comparisons that would directly and practically inform clinical practice and patient experience.

Materials and 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 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. (21) Univfy Inc. received an exempt status from institutional review board (IRB) to conduct this research.

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

compared to age control model (PLORA) were substantially as previously reported. (4-5) The training data were limited to IVF cycles using autologous oocytes and embryos with the female's age under 42.

The MLCS model life cycle and evaluation steps are detailed in SI Methods and SI Figure 1. Consecutive years of data within the 2013-2022 period were used for training and testing varied slightly across centers. 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. (20, 22). Using a larger, more recent, historical data set, each center's first model (MLCS1) was replaced by an updated model (MLCS2) in clinical usage at the start of this study.

Adapting test sets to enable comparison of MLSC2 and SART models.
The SART model responses were obtained by using the pre-treatment model formulae, with and without AMH predictor, reported by McLernon et al., 2021. (16) De novo model validation (DNMV) was performed for each center using model responses from applying each center's own MLCS2 model and the SART model to DNMV1. DNMV1, a test set modified to enable testing by both MLSC2 and SART models, was limited to first IVF cycles with age under 40, BMI value, and male factor diagnosis value (true or false). The SART model with or without AMH predictor was used according to AMH availability.

Intentional design difference between MLCS and SART methods resulted in clinical ongoing pregnancy (COP) being assigned live birth by MLCS and no live birth by SART, affecting ~4.8% of aggregate data. (16) The DNMV2 test set was finalized after removing those differentially labeled cycles from DNMV1 test set.

Comparing MLCS2 and SART models
In addition to AUC and PLORA, we computed Precision, Recall and F1 scores for the MLCS2 and SART models using each center's DNMV1 and DNMV2. PR AUC was calculated for the MLCS2 vs SART LBP for all 6 centers in aggregate. (19) We also tested the MLCS2 and SART models for reclassification, which measured the percentage of cases having different live birth probabilities from the two models. The age-based live birth rates stated in the finalized 2020 SART National Summary were used as age control because practically, that is the number that providers and patients could see if they were not using any prediction models. (23)

Statistical analyses
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. Continuous net reclassification improvement (NRI) was used to measure the likelihood of correctly re-assigning a higher or lower IVF live birth probability with MLCS2 compared to SART models. (24, 25)

The EQUATOR Reporting Guidelines including "TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods" were followed. (26, 27)

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 MLCS1 and MLCS2 model cross validation results showed positive AUC Improvement and PLORA, indicating they were superior to their respective age control models. (SI Table 1.)

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. (SI Table 1.)

To test for the risk of data drift, we performed LMV testing on each center's MLCS1 model using a center-specific, out-of-time data set. There was no significant difference in the AUC and PLORA between the LMV and CV results for MLCS1 models; therefore, there was no detectable data drift based on LMV. (SI Table 1.)

MLCS2 and SART models were evaluated for each center using the modified, center-specific test sets, 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 AMH value and they were tested by the SART model without 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 PR AUC, which are considered to be more sensitive 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 deciles 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 which showed median F1 Scores of 0.74 (IQR=0.71, 0.75)and 0.69 (IQR=0.67, 0.72)for MLCS2 and SART, respectively, using the DNMV2 test set.

PR AUC was 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 2). While overall precision was comparable between MLCS2 and SART, MLCS2 models showed higher rates of recall across most precision rates and more cycles with higher IVF live birth probabilities for both DNMV1 and DNMV2 across six centers (Figure 1).

We constructed a 4x4 reclassification table to give the practical, clinical context of the improvement conferred by MLCS2. Table 3 showed the number of patients falling into one of 16

spots based on their LBPs as computed by MLCS2 and SART using these 4 LBP categories: <25%, 25-49.9%, 50-74.9%, and ≥75%. For example, of 249 patients who would have received SART-LBPs < 25%, 168 (67%) had LBPs of 25-50% as validated by the MLCS2 model. This pattern of higher LBPs from MLCS2 is consistently seen over the entire LBP range. The continuous net reclassification index (NRI) showed a net 18.3% (95% CI 13.3%, 23.2%) of patients were more correctly assigned a higher or lower probability with MLCS2 compared to SART when tested using the DNMV1 test set ($p<0.001$). Similar findings were obtained using the DNMV2 test set. (Table 3.)

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 report to the SART registry. Here, the retrospective study design was appropriate because the prognostic models were previously trained, tested, already in clinical usage and evaluation of the models' technical performance were not biased by the retrospective design.

We took the pragmatic realist approach to address "how are the predictions different for the patients seen in the centers today?" The MLCS2 models performed better than the SART model in terms of metrics considered more sensitive to improvements in predicting the positive class (i.e. live birth prediction in this study) -- the PR AUC, F1 score and Recall. (19) Taken together, providers at those six centers can expect the following practical difference at LBP ≥ 50%: the MLCS2 model identifies ~84% of patients who would go on to have an IVF live birth, whereas the SART model identifies only ~75%. In other words, for every 100 patients who will have an IVF live birth, with LBR ≥ 50%, the MLCS2 model can identify 9 more such patients than the SART model without overcalling or overestimating LBPs compared to the SART model. This example reflects MLCS2 models' improved Recall over SART model.

To provide a thorough comparison, we used PR AUC, a metric sensitive in detecting improvements in predicting the positive class (live birth) and can tolerate imbalanced dataset. Consistent with the above example, the MLSC2 models showed improved ability to make live birth prediction calls, with a higher PR AUC compared to SART model PR AUC for the DNMV1 and DNMV2 test data used in this study ($p< 0.05$).

Although scientifically, the F1 score and PR AUC results were robust and definitive in showing the improved ability of MLCS2 models to provide appropriately higher LBPs, we also used several visuals -- a 4 x 4 reclassification table to illustrate differential LBPs and histograms showing differential frequency distribution of LBPs between the MLCS2 and SART models. In the clinical context of IVF LBP counseling, the continuous NRI is useful in testing whether differential LBPs correlate with likelihoods of more or fewer actual live births. In contrast, using continuous or categorical NRI as a metric may not be as beneficial for clinical contexts where moving up or down more than one risk category can have very different clinical meaning (e.g. oncology). (28)

Compared to the SART model, we found MLCS2 models to have improved PR AUC and F1 score yet comparable ROC AUC despite the differential dataset sizes used for model creation. (The SART model used 121K+ IVF cycles whereas the MLCS models used a median dataset size of 547 IVF cycles.) We believe it is not a constructive use of time to dissect each aspect of model design for its impact on model performance. Instead, we take the view that many factors -- including ML, CS approach, data cleaning and modeling pipeline quality assurance, the use of expert human

supervision, software automation and close collaboration with providers -- all contributed to improved MLCS2 model performance. One other difference that warrants mention is the greater number of consecutive years covered by the MLCS data sets, as that allowed for more freeze-all cycles to generate outcomes to reflect their more realistic and higher live birth probabilities. In other words, we accepted the comparison of two live models in toto, viewing any differences or constraints in the model design or data set construction to be intentional.

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.

ROC AUC was not different between the MLCS2 and SART models, presumably because it was rather insensitive to improvements in positive class prediction. However, ROC AUC was lower for the SART model in this study compared to ROC AUC measured using a national dataset. It is possible that the inclusion of patients up to age 50 in the SART model training data provided a larger proportion of true negative cases (such as having a high rate of IVF failure in older patients), resulting in a higher model-wide ROC AUC. (19) Cai et al eluded to the adapted McLernon models having better performance for patients over 35 compared to those under 35. (9) In our client services work, we typically train pretreatment models for under 40 and 40+ separately and we further perform subgroup validation for incremental age groups as the live birth rates for ages 41-42 vary significantly from those for 43+, for example. The ability to discern patients with different IVF prognoses support the appropriate delivery of compassionate care (such as using prognosis-driven empathy in the words and tone) and validated optimism according to the validated IVF LBP, efforts that are important and not mutually exclusive. (29)

The initial and ongoing motivation for our research group is to improve access to fertility care and IVF treatment to help more people succeed in building a family. The SART model, in the format of a free online calculator, is a valuable patient education tool that encourages patients to seek care. However, at the point when patients have completed their diagnostic workup and are being counseled by providers to start IVF treatment, patients are interested to know their center-specific IVF live birth probabilities.

As with the adoption of any new technology especially in the post-pandemic era, quantifying care navigation and improved workflow efficiency are critical. We have reported results from a retrospective study measuring patients' treatment utilization after receiving MLCS-counseling, the results of which informed the design of a prospective trial (in progress). (30) In the context of clinical workflow, MLCS models can support an evolving, more diverse range of healthcare providers -- such as advanced practice providers (APPs), nurse practitioners and general obstetrician-gynecologists -- to perform patient counseling, further improving scalability and accessibility of IVF treatments. (31, 32)

With IVF access as our north star, we take the view that improving model metrics such as F1 Score and PR AUC is urgently needed for several reasons. First, IVF success predictability directly affects the extent to which IVF can be financially de-risked for payers including patients as consumers and enterprise payers such as health plans and employers. Further, we should correct the erroneous

perceptions that IVF treatments are prone to failure or not predictable, which continue to discourage patients and payers.

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. (33). We have established a globally applicable framework for analyzing IVF data to inform locally relevant, practical clinical decisions. We welcome collaboration to scale research tackling crucial questions related to race/ethnicity, other social determinants of health, molecular mechanisms of clinical infertility, IVF usage and IVF outcomes for research. (34) We hope this study will help to advance reproductive medicine beyond dichotomies of multicenter versus center-specific or ML versus non-ML. Ultimately, the multicenter-scaling of the MLCS approach is expected to maximize benefit to individuals and society by addressing health inequities, supporting provider-patients prognostics counseling, de-risking financial support for IVF care and advancing precision medicine in reproductive health.

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.).

References

1. 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.
2. Moragianni VA, Penzias AS. Cumulative live-birth rates after assisted reproductive technology. Curr Opin Obstet Gynecol. 2010 Jun;22(3):189-92. doi: 10.1097/GCO.0b013e328338493f. PMID: 20216414.
3. Cedars MI. Fresh versus frozen: initial transfer or cumulative cycle results: how do we interpret results and design studies? Fertil Steril. 2016 Aug;106(2):251-6. doi: 10.1016/j.fertnstert.2016.06.001. Epub 2016 Jun 17. PMID: 27322878.
4. 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.
5. 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.
6. 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.
7. 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.

8. 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.
9. 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.
10. Wang N, Yin X, Tao Y, Wang Y, Zhu Q. Cumulative live birth rates over multiple complete cycles of in vitro fertilisation cycles: 10-year cohort study of 20,687 women following freeze-all strategy from one single centre. Arch Gynecol Obstet. 2022 Jan;305(1):251-259. doi: 10.1007/s00404-021-06063-1. Epub 2021 Aug 4. PMID: 34350510.
11. Zhu H, Zhao C, Xiao P, Zhang S. Predicting the Likelihood of Live Birth in Assisted Reproductive Technology According to the Number of Oocytes Retrieved and Female Age Using a Generalized Additive Model: A Retrospective Cohort Analysis of 17,948 Cycles. Front Endocrinol (Lausanne). 2021 Apr 30;12:606231. doi: 10.3389/fendo.2021.606231. PMID: 33995268; PMCID: PMC8120808.
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. Society for Assisted Reproductive Technology and University of Aberdeen. URL: https://w3.abdn.ac.uk/clsm/SARTIVF/ (last accessed May 10, 2024)
18. University of Aberdeen. Outcome Prediction in Subfertility, OPIS. URL: https://w3.abdn.ac.uk/clsm/opis (last accessed May 10, 2024)
19. 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.

20. 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.
21. 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.
22. https://towardsdatascience.com/why-isnt-out-of-time-validation-more-ubiquitous-7397098c4ab6
23. Society for Assisted Reproductive Technology. URL: www.sart.org (last accessed May 10, 2024)
24. 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.
25. 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.
26. EQUATOR Network. Enahance the QUAlity and Transparency Of health Research. URL: equator-network.org (last accessed May 10, 2024)
27. 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)
28. 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.
29. 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
30. Yao MWM, Nguyen ET, Retzloff MG, Gago LA, Copland S, Nichols JE, Payne JF, Opsahl M, Cadesky K, Meriano J, Donesky BW, Bird III J, Peavey M, Beesley R, Neal G, Bird Jr. JS, Swanson T, Chen X, Walmer DK. Improving IVF utilization with patient-centric artificial intelligence-machine learning (AI/ML): A retrospective multicenter experience. JCM 2024;13(12):3560; https://doi.org/10.3390/jcm13123560
31. Hariton E, Alvero R, Hill MJ, Mersereau JE, Perman S, Sable D, Wang F, Adamson GD, Coutifaris C, Craig LB, Hosseinzadeh P, Imudia AN, Johnstone EB, Lathi RB, Lin PC, Marsh EE, Munch M, Richard-Davis G, Roth LW, Schutt AK, Thornton K, Verrilli L, Weinerman RS, Young SL, Devine K. 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.
32. 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.

33. 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.
34. 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.

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 IVF outcomes 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

Table 2. This table shows the median and IQR for model cross validation metrics -- AUC and PLORA -- measured by testing each center's MLCS2 and SART models using each center's modified test sets (DNMV1 and DNMV2) and using the SART 2020 age group-based live birth rate 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	PLORA	F1 at 50% LBP threshold*	PR AUC*
DNMV1 test set includes IVF cycles with COP outcomes (N=4,645)	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 excluding IVF cycles with COP outcomes (N=4,421)	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)

Table 3. Reclassification table comparing IVF live birth predicted probability models MLCS2 and SART across centers for (A) DNMV1 and (B) DNMV2. The continuous NRI was 18.3% (95% CI 13%, 23%) for DNMV1 and 15.8% (95% CI 10.7%, 20.8%) for DNMV2; $p<0.001$.

A.

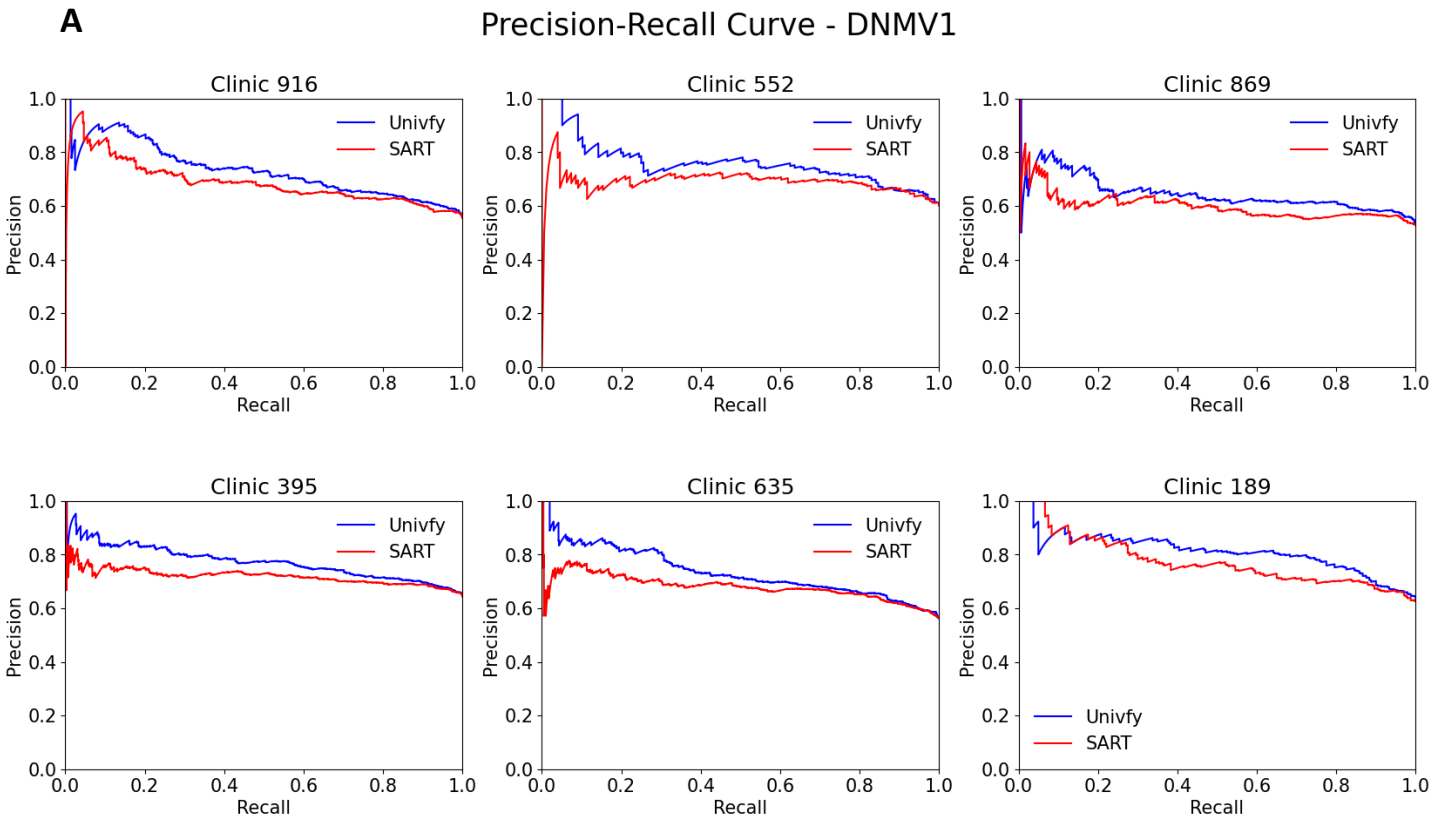
DNMV1	**MLCS2**				
SART Model	< 25%	25% ≥ x > 50%	50% ≥ x > 75%	≥ 75%	Total
< 25%	79	168	2	0	249
25% ≥ x > 50%	12	735	566	7	1320
50% ≥ x > 75%	0	144	2445	487	3076
≥ 75%	0	0	0	0	0
Total	91	1047	3013	494	4645

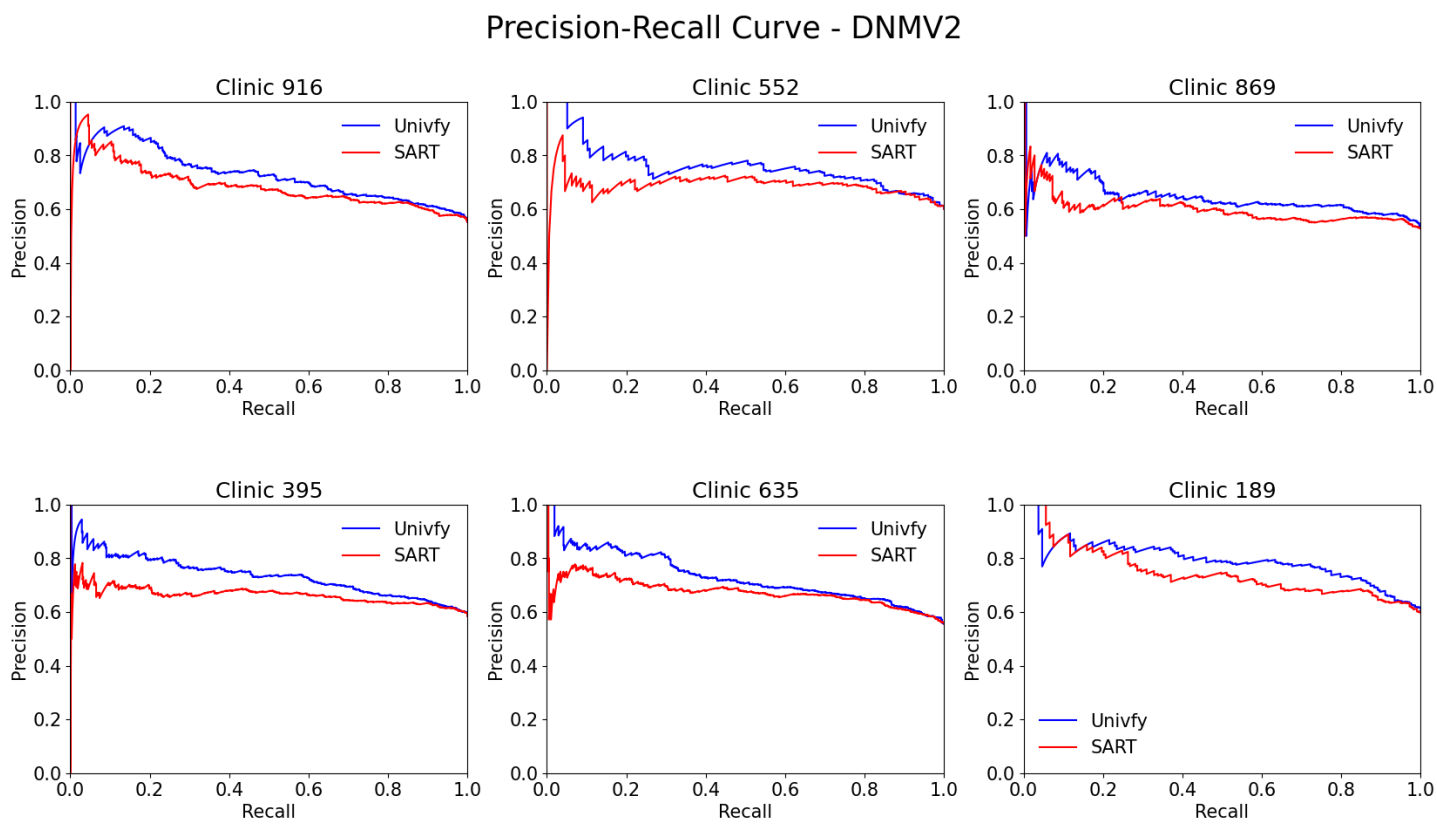
B.

DNMV2	**MLCS2**				
SART Model	< 25%	25% ≥ x > 50%	50% ≥ x > 75%	≥ 75%	Total
< 25%	78	163	1	0	242
25% ≥ x > 50%	12	726	525	7	1270
50% ≥ x > 75%	0	143	2336	430	2909
≥ 75%	0	0	0	0	0
Total	90	1032	2862	437	4421

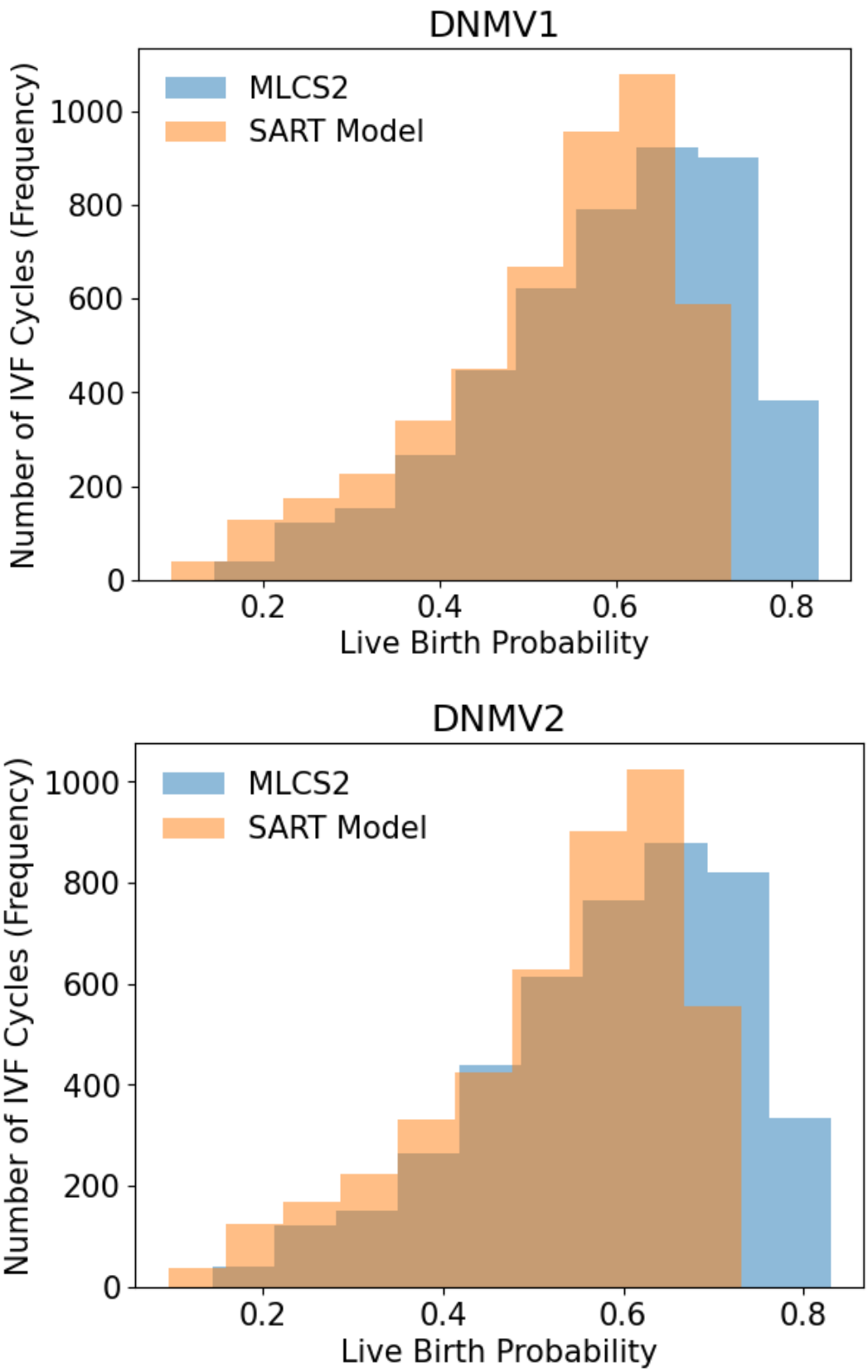
Figure 1. Comparison of MLCS2 and SART models using (A) Precision-Recall curves for each of the 6 clinics using each center's modified test sets, DNMV1 and DNMV2, for MLCS2 (labeled Univfy) and SART models; (B) frequency distributions of live birth probabilities using MLCS2 and the SART models for modified test sets DNMV1 and DNMV2.

A





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

Supplementary Information (SI)

Predicting IVF live birth probabilities using machine learning, center-specific and national registry-based models. Nguyen et al.

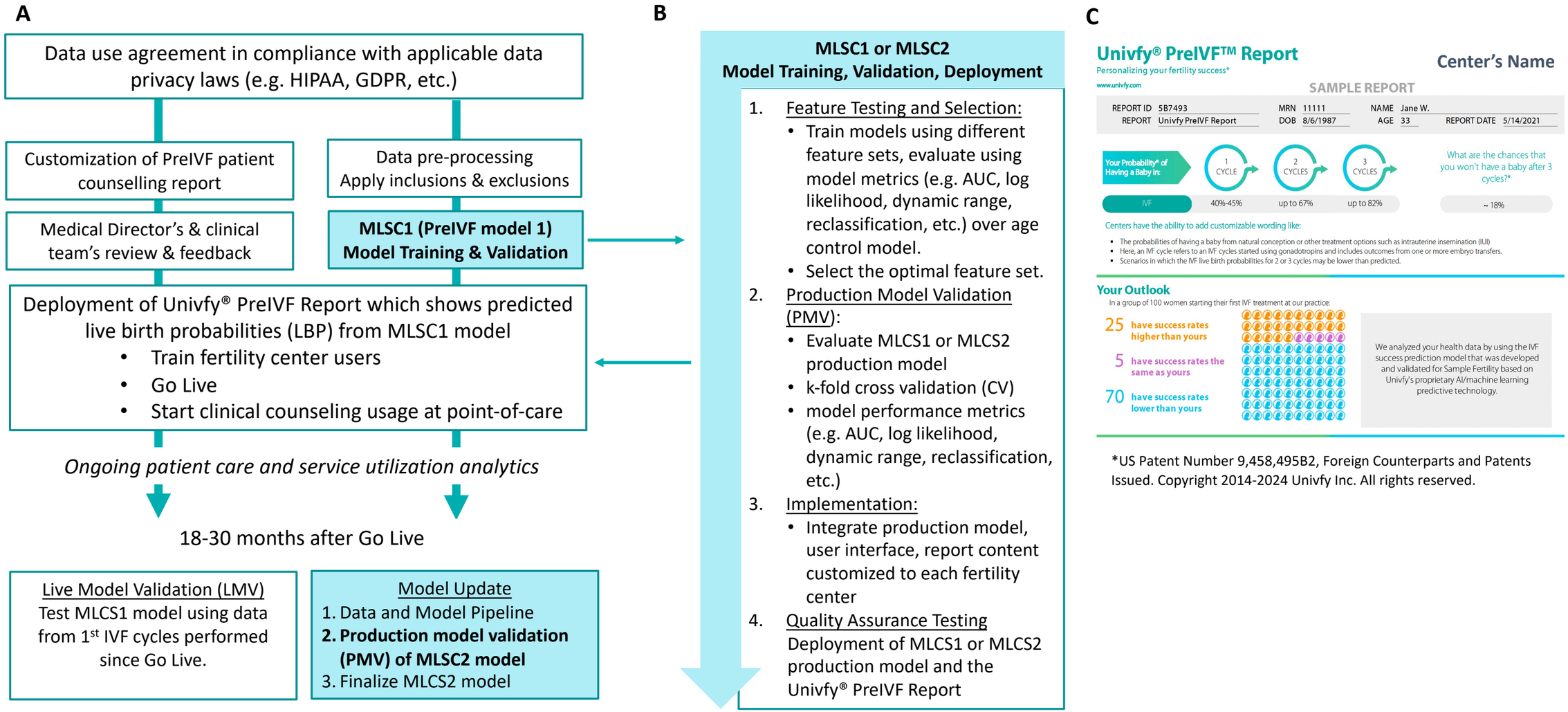
SI Methods

Cross validation (CV). 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, can also be translated to the linear scale (e^{PLORA}) to facilitate communications with non-statisticians. (1, 4-5) CV of both MLCS1and MLCS2 was reported using median and interquartile range (IQR) across 6 centers for ROC-AUC, ROC-AUC improvement and PLORA.

SI Table 1. This table shows the median and interquartile range (IQR) for cross validation and live model validation metrics -- AUC, AUC Improvement over Age model and PLORA -- for MLCS1 and MLCS2 models across 6 centers.

		Model validation results: median and IQR	
Model	Validation	AUC	PLORA
MLCS 1	CV, in-time	0.66 (IQR = 0.61, 0.68)	7.2 (IQR = 3.6, 11.8)
MLCS 2	CV, in-time	0.67 (IQR = 0.66, 0.68)	23.9 (IQR = 10.2, 39.4)
MLCS 1	LMV, out-of-time	0.65 (IQR = 0.63, 0.66)	6.7 (IQR = 2.2, 12.0)

SI Figure 1. The development-to-deployment life cycle of the machine learning-based, center-specific, 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). (B) Model pipeline supports feature testing, model training, validation analysis, deployment to production and quality testing. (C) Point-of-care deliverable as illustrated by the first page of the sample provider-patient counseling report showing how the MLSC model's predicted IVF live birth probabilities are communicated. (The identifiers including Report ID, MRN, DOB, Name, Age and Report Date, on the report are fictitious and are generated using a "demo" clinic that uses only dummy, fictitious data for illustration only.)



*US Patent Number 9,458,495B2, Foreign Counterparts and Patents Issued.

*US Patent Number 9,458,495B2, Foreign Counterparts and Patents Issued. Copyright 2014-2024 Univfy Inc. All rights reserved.