Personalized prediction of first-cycle in vitro fertilization success

Bokyung Choi, Ph.D.,[a] Ernesto Bosch, M.D.,[b] Benjamin M. Lannon, M.D.,[c,d,e] Marie-Claude Leveille, Ph.D.,[f] Wing H. Wong, Ph.D.,[a,g] Arthur Leader, M.D.,[f] Antonio Pellicer, M.D.,[b] Alan S. Penzias, M.D.,[c,d,e] and Mylene W. M. Yao, M.D.[a]

[a] Univfy, Los Altos, California; [b] Instituto Valenciano de Infertilidad, Valencia, Spain; [c] Boston IVF, Waltham, Massachusetts; [d] Department of Obstetrics and Gynecology, Beth Israel Deaconess Medical Center, Boston, Massachusetts; [e] Department of Obstetrics, Gynecology, and Reproductive Biology, Harvard Medical School, Boston, Massachusetts; [f] Ottawa Fertility Centre, Ottawa, Ontario, Canada; and [g] Department of Statistics, School of Humanities and Sciences, Stanford University, Stanford, California

Objective: To test whether the probability of having a live birth (LB) with the first IVF cycle (C1) can be predicted and personalized for patients in diverse environments.
Design: Retrospective validation of multicenter prediction model.
Setting: Three university-affiliated outpatient IVF clinics located in different countries.
Patient(s): Using primary models aggregated from >13,000 C1s, we applied the boosted tree method to train a preIVF-diversity model (PreIVF-D) with 1,061 C1s from 2008 to 2009, and validated predicted LB probabilities with an independent dataset comprising 1,058 C1s from 2008 to 2009.
Intervention(s): None.
Main Outcome Measure(s): Predictive power, reclassification, receiver operator characteristic analysis, calibration, dynamic range.
Result(s): Overall, with PreIVF-D, 86% of cases had significantly different LB probabilities compared with age control, and more than one-half had higher LB probabilities. Specifically, 42% of patients could have been identified by PreIVF-D to have a personalized predicted success rate >45%, whereas an age-control model could not differentiate them from others. Furthermore, PreIVF-D showed improved predictive power, with 36% improved log-likelihood (or 9.0-fold by log-scale; >1,000-fold linear scale), and prediction errors for subgroups ranged from 0.9% to 3.7%.
Conclusion(s): Validated prediction of personalized LB probabilities from diverse multiple sources identify excellent prognoses in more than one-half of patients. (Fertil Steril® 2013;99:1905–11. ©2013 by American Society for Reproductive Medicine.)
Key Words: Personalized medicine, IVF prediction, IVF outcomes, infertility, IVF success rate

Discuss: You can discuss this article with its authors and with other ASRM members at http://fertstertforum.com/choib-personalized-medicine-ivf-success-rate/

Use your smartphone to scan this QR code and connect to the discussion forum for this article now.*

* Download a free QR code scanner by searching for "QR scanner" in your smartphone's app store or app marketplace.

In vitro fertilization (IVF) offers the highest per-treatment success rate for most patients diagnosed with infertility. However, it has remained underutilized; fewer than 3% of the estimated 7 million women/couples suffering from infertility access IVF. Although there are likely many reasons for this underutilization, such as high cost, limited insurance reimbursement, and success with other treatments, unclear benefits and unrealistic expectations compromise a patient's level of confidence in pursuing IVF treatment.

Some women erroneously believe that IVF secures their chances to have a baby despite delayed family building, whereas others may not realize that their personalized success rates are much higher than typically quoted with the use of age-based or filtered reporting. In the former scenario, a delay in trying to conceive, or in pursuing IVF, not only compromises a woman's personalized chances of success, but contributes to the lowering of overall and age-based success rates and the perceived limited success of IVF. On the other hand, potential age-based underestimation of success rates may discourage women >35 years of age from pursuing a treatment that in reality offers them excellent chances of having a baby.

Although informative guides (e.g., Fertistat, American Society for Reproductive Medicine patient resources,

Received October 22, 2012; revised and accepted February 8, 2013; published online March 21, 2013.
B.C. is a full time employee, and has been granted stock options at Univfy Inc. E.B. is a consultant for and receives payment for lectures by MSD, Merck-Serono, and Ferring Pharmaceuticals. B.M.L. has nothing to disclose. M.-C.L. has nothing to disclose. W.H.W. is a cofounder, stock holder, and board member of Univfy. A.L. is a medical advisor of Univfy. A.P. has nothing to disclose. A.S.P. is collaborating with Univfy to develop prediction tools via data sharing. M.W.M.Y. is a co-founder, full time employee, stock holder, and board member of Univfy.
Reprint requests: Mylene W. M. Yao, M.D., Dept. of R&D, Univfy, 5150 El Camino Real, Suite B-23, Los Altos, California 94022 (E-mail: mylene.yao@univfy.com).

Fertility and Sterility® Vol. 99, No. 7, June 2013 0015-0282/$36.00

http://dx.doi.org/10.1016/j.fertnstert.2013.02.016

public reporting by the Society for Assisted Reproductive Technologies) have been available to raise awareness of the risks of infertility and when to seek infertility medical care, there was previously no rigorously validated online tool providing quantitative personalized IVF success rates to infertility patients based on prediction modeling of multicenter North American and European IVF outcomes data (1–4). Personalized quantitative prognostics convey an important message to patients and society at large that IVF success is largely predictable, based on science and evidence, rather than a roll of the dice. Having this resource alone may minimize uncertainty and confusion among patients and enhance their confidence in infertility treatment options.

Previously, we reported the first validated clinic-specific models predicting live birth and multiple birth probabilities and their performance as quantitatively measured by predictive power, discrimination, calibration, dynamic range, and reclassification (5, 6). We received enthusiastic requests from many IVF providers and infertility patients asking our research team to develop a prediction model that is applicable to diverse patient populations despite cross-center differences in clinical protocols. In addition, we were asked to make such advanced IVF prediction modeling accessible to patients. Other research groups have attempted to develop multicenter or cross-center prediction models for IVF outcomes, but have not been successful in their validation (7).

We aimed to push the performance and utility of IVF prediction modeling beyond single-center validation to develop a more diverse multicenter validated prediction model, PreIVF-Diversity (PreIVF-D), to predict the probability of having a live birth in a patient's first IVF treatment. Having diverse representation of patient population and treatment protocols makes this predictive technology relevant to patients and doctors beyond the centers involved in this research. In parallel to the statistical research, enterprise-grade engineering implementation has made the personalized prediction of IVF success accessible in real time, via a paid online web-based tool. Reporting of both predicted probability and percentile ranking may also support IVF providers in the refining clinical protocols based on prognostics.

Here, we report the methods and validation results of the multicenter prediction model, PreIVF-D, which predicts the probability of having a live birth with a patient's first IVF treatment. We discuss the strengths, limitations, and applications of PreIVF-D, as well as the potential impact of personalized prognostics on the overall IVF success rates and access to IVF at the population level.

METHODS

Patients, IVF Treatments, and Clinical Outcomes

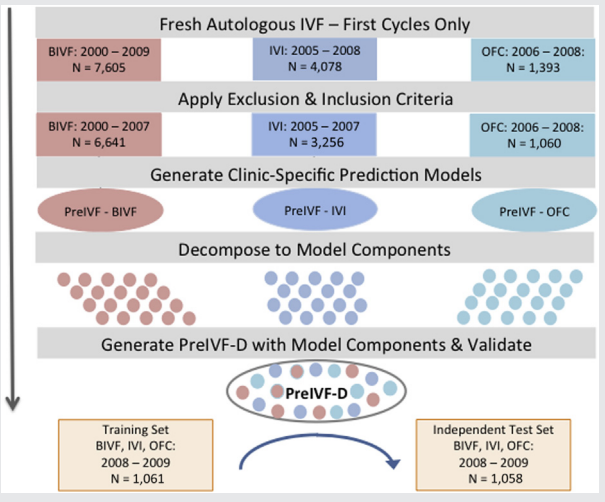
The retrospective cohort comprised 13,076 first IVF treatment cycles using fresh autologous eggs and fresh embryo transfers (ETs) that were performed at three academically affiliated private IVF clinics, located in three different countries: 7,605 cases from Boston IVF (BIVF), Waltham, Massachusetts, from January 1, 2000, to December 31, 2009; 4,078 cases from IVI-Valencia (IVIV), Spain, from January 1, 2005, to December 31, 2008; and 1,393 cases from Ottawa Fertility Centre (OFC), Ottawa, Canada, from January 1, 2006, to December 31, 2009. Inclusion and exclusion criteria were applied to define cases for generating clinic-specific prediction models for each clinic (Fig. 1). Each center obtained its own Institutional Review Board approval to conduct retrospective IVF prediction modeling research with researchers from Univfy.

Patients underwent ovarian stimulation protocols according to a combination of physician's recommendation and clinical protocol as described previously (8–10). Embryos were cultured according to each clinic's standard protocols. Ultrasound-guided ET was performed 3–5 days after oocyte retrieval according to clinic protocol. The number of embryos transferred was based on national and clinic guidelines as well as individual patient needs. Patients were followed for at least 1 year from the start of their IVF cycles to confirm IVF and pregnancy outcomes.

Data Collection, Exclusion Criteria, Definition of Live Birth, and Variables

Baseline demographic, clinical and laboratory data were collected according to standard clinic practices as described previously (8–10). Medical record review was used to

FIGURE 1



Clinical data sources, datasets, and prediction model design used to generate and validate PreIVF-Diversity (PreIVF-D), a prediction model that predicts the probability of live birth with a patient's first IVF treatment. Outcomes data on first IVF cycles were extracted from deidentified data sets provided by three clinics: Boston IVF (BIVF; *red*), Instituto Valenciano de Infertilidad (IVI; *blue*), and Ottawa Fertility Centre (OFC; *green*). N indicates the number of IVF treatments. After applying inclusion and exclusion criteria, outcomes data from eligible cycles were used to generate clinic-specific prediction models (*ovals*), which were each decomposed to model components (*circles*) (patent pending). Using a training set, PreIVF-D (*oval*) was generated by selecting model components from all three clinics. PreIVF-D was validated by testing on the independent data set comprising data from all three clinics: training and independent test datasets (*pale orange*).

Choi. Predicting first IVF cycle outcomes. Fertil Steril 2013.

confirm pregnancy outcome. Each clinic obtained approval from its Institutional Review Board to perform retrospective data collection, aggregation, and deidentification for statistical analysis and prediction modeling.

We excluded first IVF cycles that met any of the following criteria: A prior IVF cycle had been performed at another fertility clinic which could be ascertained from the database alone; the IVF cycle was canceled before oocyte retrieval; clinical outcome was not known; the patient's age was ≥43 years; ET occurred on days other than day 2–6; and IVF cycles that did not use gonadotropin injection.

The outcome of an IVF cycle was defined as "live birth" if: 1) the fresh ET directly resulted in a live birth; or 2) the fresh ET did not result in a live birth, but the transfer of cryopreserved-thawed embryos that were produced by that fresh IVF cycle resulted in a live birth. Linking the outcomes of fresh and cryopreserved ET in this way provided information on the total reproductive potential of the embryos produced by a fresh cycle (5, 6, 11, 12).

We analyzed the following variables, each of which was available from at least one clinic. Patient's demographics and reproductive history included age at the time of first IVF treatment, body mass index, smoking status, gravidity, parity, pregnancy losses before 20 weeks' gestation, number of ectopic pregnancies, antral follicle count, day 3 serum FSH, and year. Clinical diagnoses included polycystic ovarian syndrome or disease, diminished ovarian reserve, tubal disease, endometriosis, recurrent miscarriage, unexplained infertility, uterine causes, other causes, male factor. Male partner's reproductive health included age, total motile sperm count, use of sperm extraction method, and use of donor sperm.

Statistical Analysis

We generated the PreIVF-D model with the use of a multistep procedure. First, we used baseline clinical variables and data that were available before starting IVF, to develop clinic-specific PreIVF models. Each clinic-specific PreIVF model was trained with variables and eligible outcomes data obtained from that clinic alone: PreIVF-BIVF: 6,641 cases from January 1, 2000, to December 31, 2007; PreIVF-IVI: 3,256 cases from January 1, 2005, to December 31, 2007; and PreIVF-OFC: 1,060 cases from January 1, 2006, to December 31, 2008. Briefly, for each clinic-specific model, we computed the log-likelihood based on the Bernoulli distribution and applied generalized boosted models (GBM), a free software implementation of stochastic gradient boosting algorithm, to build a boosted tree model using a maximum of 70,000 trees and tenfold cross-validation (5, 6).

We built PreIVF-D by blending and weighting the individual components from all three clinic-specific models to form a resulting model that was adjusted for the different numbers of cases available from each clinic. Training of PreIVF-D was performed with a training dataset comprising an aggregate of 1,061 independent cases that were not used to generate the original clinic-specific PreIVF models: BIVF: 483 cases from January 1, 2009, to December 31, 2009; IVI: 411 cases from January 1, 2008, to December 31, 2008; and OFC: 167 cases from January 1, 2009, to December 31, 2009 (Fig. 1).

PreIVF-D was compared with an age-based control model (Age model) that was generated from 10,957 cases by applying GBM to patient's age alone based on age categories (<35, 35–37, 38–40, 41–42) that are used by the Society for Assisted Reproductive Technologies and Centers for Disease Control and Prevention (1, 2, 4). We used all available cases, including those used to develop each clinic-specific model, to generate this Age model, to allow the most stringent conditions for comparing the PreIVF-D and Age models. The performance of the PreIVF-D model was measured and validated with the use of a further independent test set that comprised 1,058 test cases: BIVF: 481 cases from January 1, 2008, to December 31, 2009; IVI: 411 cases from January 1, 2008, to December 31, 2008; and OFC: 166 cases from January 1, 2009, to December 31, 2009 (Fig. 1). All of the results reported are validated test results of PreIVF-D, not merely a description of the training set or results pertaining to clinic-specific PreIVF models.

We determined the posterior probability of having a live birth in the first IVF cycle based on the collective phenotype profile of the patient and her male partner, or the patient's phenotype profile alone if donor sperm is used. Predictive power is described as the improvement in the log-likelihood of predicting the probability of having a live birth in the first IVF cycle with the PreIVF-D relative to the Age model, using Baseline-Diversity (Baseline-D), a control model in which no predictors are used. In other words, Baseline-D is the mean probability of having a live birth in the first IVF cycle if not a single predictor, not even age, is used. Log-likelihoods (LL) were computed with the use of GBM.

$$\%\ \text{improvement} = \left(\left[\left(\text{LL}_{\text{PreIVF-D}} - \text{LL}_{\text{Baseline}}\right) - \left(\text{LL}_{\text{Age}} - \text{LL}_{\text{Baseline}}\right)\right] / \left[\text{LL}_{\text{Age}} - \text{LL}_{\text{Baseline}}\right]\right) \times 100\%$$

To compare the clinical utility of PreIVF-D, frequency distributions of predicted live birth probabilities for Baseline control, Age-D, and PreIVF-D were compared, and the validity of this comparison was determined by receiver operating characteristic (ROC) analysis. Dynamic range describes the probabilities of live birth that can be predicted with the use of PreIVF-D compared with Age-D. Calibration of PreIVF-D was tested by comparing the predicted versus observed probabilities of live birth of test cases with the those of Age-D based on the Holsmer-Lemeshow goodness-of-fit test. Those test cases were defined by six groups with successive probabilities of live birth: <10.0%, 10.0%–19.9%, 20.0%–29.9%, 30.0%–39.9%, 40.0%–45.0%, and >45.0%.

RESULTS

Training and Test Sets for PreIVF-D

Together the training and test sets comprised 2,119 first fresh IVF cycles that used the patients' own eggs. Separately, the training set that was used to train PreIVF-D showed a live birth rate of 38.1% (95% confidence interval [CI] 0.35–0.41), and the test set that was used to test

validation of PreIVF-D showed a live birth rate of 38.4% (95% CI 0.35–0.41). The mean values for each variable did not differ significantly between training and test sets (Table 1).

Predictors and Their Relative Importance

The PreIVF-D model assigned relative importance to each prognostic factor, with the total relative importance set arbitrarily at 100%. Variables with the highest nonredundant prognostic contribution to the PreIVF-D model, and their relative importance, were age of patient (60.1%), total motile sperm count (9.6%), body mass index (9.5%), day 3 serum FSH (5.0%), and antral follicle count (4.5%); other factors, each with <3.0% relevance, made up the remaining 11.4% of relative importance. Therefore, before a patient's first IVF treatment, 60% of her personalized prognosis is predicted by her age and 40% by other clinical factors. More importantly, the relative influence of various factors is not the same for each patient, and these measures serve as an overview only.

Predictive Power and Prediction Error

The ability to predict the probability of live birth is improved by 35.7% with PreIVF-D compared with Age-D. This improvement represented an increase in predictive power by 9.0-fold on the log scale (>1,000-fold in linear scale).

Sometimes, in predictive modeling, increased predictive power may be achieved at the expense of accuracy, by increasing prediction error (prediction error = 1 − accuracy), which is a separate measure from predictive power. Prediction error is determined for a subgroup of patients rather than specific individual patients. For each of the models, PreIVF-D and Age-D, we divided the test set into subgroups of patients according to their predicted probabilities and measured the prediction error for each subgroup (Table 2). Most remarkably, PreIVF-D has prediction errors of only 0.9% and −1.2% for patients who had predicted live birth probabilities of >45.0% and 40.0–45.0%, respectively. Therefore, we confirmed that calibration or accuracy has not been sacrificed in achieving the predictive power shown in the PreIVF-D model.

Measure of Utility (Usefulness) by Ranking, Dynamic Range, Reclassification, and Discrimination

When determining whether a prediction model is clinically useful, predictive power must first be confirmed to be superior to currently available methods. However, for a predictive model to be useful, its ability to rank patients correctly according to prognosis (e.g., discrimination) must be equivalent to or better than the control model. The ranking provided by PreIVF-D is more reliable than the control, because ROC analysis showed that the ability of PreIVF-D to discriminate patients with differential probabilities of live birth showed an improvement of 3.2% over Age-D (the areas under the ROC curves [AUCs] for PreIVF-D and Age-D measured 0.634 and 0.614, respectively.) Furthermore, the dynamic range of predicted probabilities is extended from

TABLE 1

Comparison of variables between the PreIVF-D training and test datasets.

	Training set (n = 1,061)		Independent test set (validation) (n = 1,058)	
Variable	**Mean[a]**	**SD**	**Mean[a]**	**SD**
Age, y	34.5	4.2	34.5	4.1
Gravidity	0.8	1.2	0.9	1.3
Parity	0.2	0.5	0.2	0.6
No. of ectopic pregnancies	0.03	0.2	0.07	0.3
No. of pregnancy losses (<20 wk)	0.4	0.9	0.5	1.0
Smoking[b]	0.3	0.5	0.3	0.5
Year	2008.6	0.5	2008.6	0.5
Body mass index	25.5	6.1	26.5	6.8
Serum day 3 FSH, mIU/mL	7.3	2.6	7.7	3.5
Antral follicle count	29.4	25.9	32.5	29.2
Male partner's age, y	37.2	6.0	37.4	5.6
Total motile sperm count (million/mL)	54.3	81.0	52.4	79.8
Use of frozen sperm[b]	0.1	0.3	0.1	0.3
Use of donor sperm[b]	0.05	0.21	0.05	0.22
Clinical diagnoses				
Tubal factor[b]	0.1	0.3	0.1	0.4
Diminished ovarian reserve[b]	0.08	0.27	0.07	0.26
Endometriosis[b]	0.08	0.27	0.09	0.29
Male factor[b]	0.6	0.5	0.6	0.5
Recurrent miscarriage[b]	0.03	0.17	0.04	0.19
Polycystic ovaries/polycystic ovarian syndrome[b]	0.09	0.29	0.09	0.28
Unexplained infertility[b]	0.05	0.23	0.05	0.21

[a] For continuous variables, the mean indicates the mean value of each variable. For categoric variables, the mean indicates the average number of positive occurrences. There was no significant difference in the mean values between training and test sets for all variables ($P>.5$).
[b] Categoric variables have values "true" or "false."

Choi. Predicting first IVF cycle outcomes. Fertil Steril 2013.

TABLE 2

Percentage of patients and their probabilities of live birth predicted by each prediction model: PreIVF-D, Age Control, and Baseline Control (no predictor).

Predicted probability of live birth	Baseline (no predictor)	Baseline prediction error	Age control	Age prediction error	PreIVF-D	PreIVF-D prediction error
>45.0%			0%		41.6%	0.9%
40.0%–45.0%			49.6%	−4.9%	13.5%	−1.2%
30.0%–39.9%	100%	−3.0%	24.2%	−1.5%	19.3%	−0.1%
20.0%–29.9%			19.1%	0.2%	13.2%	−3.7%
10.0%–19.9%			0%		7.3%	−1.5%
<10.0%			7.1%	−1.1%	5.1%	−1.7%

Note: The use of PreIVF-D identified the top 18.3%, 41.5%, 55.1%, and 63.9% percentiles of patients who have the highest probabilities of live birth at >50.0%, >45.0%, >40.0%, and >35.0%, respectively. A negative prediction error indicates that the mean predicted probability is lower than the mean observed probability in that group.

Choi. Predicting first IVF cycle outcomes. Fertil Steril 2013.

having four discrete probabilities (age <35 y, 49.6%; age 35–37 y, 24.2%; age 38–40 y, 19.1%; age 41–42 y, 7.1%) to predicted live birth probabilities, ranging from 3.9% to 57.0%.

Specifically, reclassification analysis showed that with PreIVF-D, 86% of cases had significantly different live birth probabilities compared with age control ($P<.05$), with 57% and 28% showing higher and lower live birth probabilities, respectively; 42% of patients were found to have predicted live birth probabilities >45.0% and 18.3% to have predicted live birth probabilities of >50.0%. In contrast, no patients could be identified by Age-D to have >45% live birth probabilities.

Note that the difference in ranking between PreIVF-D and Age-D is not merely a shift of the same patients being told that they have probabilities of 40%–45% when their probabilities are >45%. The same patients are not necessarily at the top of each model's percentile ranking. For example, a patient predicted to have 38% live birth probability by age may be reclassified to have 46% live birth probability by PreIVF-D, and because PreIVF-D has improved predictive power, the probability of 45% as predicted by PreIVF-D is closer to the truth by 9 times on the log-scale (>1,000 times on a linear scale); this patient is also part of a group whose PreIVF-D prediction differs from observed live birth rate by 0.9%.

We also made our best attempt to refine the age control model. We applied the boosted tree method to analyze age with the use of more than 10,000 first IVF cycles, without restricting age to conventionally defined categories, to test whether the age model—overall or for the high-prognosis group—could perform better. However, the predictive power and AUC were not improved (data not shown). Therefore, PreIVF-D still has the best performance overall and in each of the measured parameters.

DISCUSSION

First, and foremost, this research work was performed in response to requests from many patients and reproductive endocrinology-infertility specialists (REIs). Patients indicated that they would like to receive predicted probabilities of treatment success that are personalized to their clinical data beyond age-based statistics, and REIs indicated that they wished to embrace personalized medicine and deliver transparent prognoses that are tailored to each patient's clinical situation, but that they lacked an accurate and validated user-friendly tool to support this goal. Therefore, we responded to these requests by testing whether a prediction model could be validated to provide personalized predicted probabilities for a patient's first IVF treatment. This information can be used to raise awareness of the benefits of IVF and to support a patient's decision to pursue IVF.

Second, a key strength of our study design is the use of data from multiple centers with diverse patient populations without requiring each individual clinic's data set to comply with a specific mandatory format or set of clinical protocols. For example, cases may have data on day 3 FSH levels or antral follicle count, but for the purpose of model building, neither was mandatory, because our goal was to develop not only a prediction model but also an approach to building a prediction model that can be easily replicated or extended to include centers that may have different clinical protocols or have collected different variables in their databases. To our knowledge, multicenter models that are developed via decomposition and recompilation of individual model components without merging and normalization of the centers' actual data sets have not been reported in the medical literature. Not only is this novel method critical to maximize the utility of information from diverse clinics and data infrastructures to provide high-quality personalized prognoses to patients contemplating IVF, but it also has far-reaching implications beyond reproductive medicine. This method of accessing diverse data infrastructures without standardization overcomes a major obstacle in personalized medicine and health care technology: the inability to analyze retrospective multicenter data because of nonuniformity of data infrastructure.

Third, we applied stringent methods, including: the use of distinct training and test sets to develop and validate the prediction model, respectively; defining the outcome measure to reflect the total reproductive potential of the first IVF cycle by linking the fresh cycle with its frozen ETs; and evaluating PreIVF-D based on objective and quantitative measures that are required for clinical utility. We demonstrate that these criteria—predictive power, the ability to rank cases correctly based on predicted probabilities, dynamic range, and reclassification rate—are met to support clinical utility (5, 6). Although most papers report AUC based on ROC analysis as a way to determine discrimination, AUC measures depend heavily on the cases that are included in the analysis. For

example, had we included cases with age >43 years, the AUC of the model would have increased dramatically, because that is an age group for which outcomes prediction is "easy." In addition, the measures of a prediction model and its utility are based on the data available at the time that a decision is made. For example, we have shown previously that based on data available by the time of ET, modeling can provide personalized predictions of live birth outcomes with an AUC of 0.8 (5). However, because the IVF treatment would be complete at that point with pending serum pregnancy test results, the utility of that model is low, despite extremely high predictive power and discrimination.

When interpreting the results, it is important to note the meaning of various statistical terms. We use log-likelihood to measure predictive power, which means "how likely the data will fit the model," or for nonstatisticians, "how much more likely the test data are represented by one prediction model over the control model." Owing to discrepancy between the common English language and statistical definitions, the words "predictability" and "accuracy" are often used interchangeably in the English language, but in the context of reporting research findings in an original research article, these terms have distinct meanings. When measuring predicted live birth probabilities, the prediction error for a group of patients is the difference between the expected live birth probability and the observed live birth probability as a percentage of the observed live birth probability. Prediction error may also be expressed in terms of accuracy, where accuracy = 1 − prediction error. The overall accuracy of all subgroups is measured by calibration. Therefore, the terms calibration, prediction error, and accuracy all measure the proximity between observed and expected live birth probabilities, but they do not inform us of a model's predictive power (13–17).

On a practical level, the most important reference is the table that shows how a particular predicted probability correlates with percentile rankings based on the PreIVF-D and age control models (Table 2). This information is expected to help patients understand their prognoses in a context that is meaningful to them. In areas outside of reproductive medicine, most people relate excellent chances of an outcome with 90%–100%, good chances with 80%–90%, and poor chances with <50%. A major challenge in counseling patients about the benefits of IVF is that without a percentile scale, a patient may not realize that having >40% chance to have a baby with IVF is very good. This reference can also support physicians in refining or developing prognosis-driven clinical protocols. For example, a clinic may establish clinical protocols for patients who have unexplained infertility based on varying levels of predicted probability of live birth. Over time, collection of outcomes data can be fed back into the prediction model to support a data-driven method to further refine protocols.

Our research design meets with a couple of constraints. First, the availability of IVF live birth outcomes lag behind treatment start dates for more than a year because of the duration of pregnancy, follow-up collection of live birth outcomes, and time required to develop and validate prediction models, so it is not possible to have a prediction model that is validated for patients who are going through IVF in real time. However, because PreIVF-D predicts live birth probability in terms of a patient's full reproductive potential (e.g., live birth with the transfer of fresh and/or frozen embryos), rather than the live birth probability per fresh ET, the prediction is not likely to be affected by changes in ET policies. Second, as outcomes data comprising other variables, such as serum antimüllerian hormone levels or preimplantation genetic screening data, become available, their use can also be incorporated into prediction models.

Although it is not possible for us to prove within the scope of the present study that PreIVF-D is definitively valid for a clinic outside of this study, the current representation of diverse patient populations and clinical protocols from three collaborating clinics in three different countries should allow clinics that share similar demographics and practice patterns to use PreIVF-D with confidence. Furthermore, we invite research collaboration to help a clinic to test whether PreIVF-D is validated for its clinic-specific data. This validation work can be performed easily by analyzing ~100 cycles of deidentified data, which makes it feasible for most clinics to participate in personalization prognosis. In our ongoing research, we aim to collaborate with other clinics to perform this type of validation study. This approach also allows uncommon clinical profiles to be aggregated across centers, which improves the quality of prognostic information. In fact, the use of a multicenter-derived prediction model may be the most cost-effective and practical approach, because we have found that for the prediction of live births in first IVF cycles, the improvement offered by a clinic-specific model (i.e., each of PreIVF-BIVF, PreIVF-IVIV, or PreIVF-OFC applied to a clinic-specific test set) over PreIVF-D is minimal (data not shown). Nonetheless, if a clinic determines that its demographics or protocols are sufficiently different from cases represented by PreIVF-D, there is always the option to develop and validate a prediction model based on its own data.

Another potential application is to use personalized outcomes prediction to support patient selection for research protocols, especially protocols that are being evaluated for patients with poor prognosis. A general problem with prospective interventional research trials is that large trials may fail due to the inability to identify patients who may benefit most from a certain new intervention, and subgroup analyses that are not determined a priori may be criticized. Adding an objective method to either define subgroups a priori or to define criteria for patient recruitment is very likely to improve the success and cost-effectiveness of research trials.

This multicenter prediction model, together with our previous work pertaining to clinic-specific predictive modeling, meets an urgent need to bring personalized prognosis to the forefront for our patients. We are in the midst of an emergence of statistical learning across medical disciplines in which prediction models in prostate cancer, cardiovascular disease, and chronic kidney disease have enriched medicine with new ways of applying personalized prognostics. However, unlike those disease areas, where complex prediction models are still clinic specific and in the research and development (R&D) stage, thus not yet accessible, we have gone beyond the R&D stage and established online access by patients and health care

providers. Furthermore, the role of personalized prognostics in reproductive medicine is much clearer. Infertility patients would have only diminishing success rates with time whether they choose natural, non-IVF, or IVF treatments; the threshold probability of success needed to proceed with IVF is subject to each patient's own value system rather than governed by guidelines.

Finally, our findings reveal that age-based estimates of live birth probabilities in IVF currently provide a suboptimal basis to guide clinical decisions by patients and providers, because live birth probabilities are vastly underestimated for many patients considering their first IVF cycles. Here, we describe two brief hypothetical examples. Patient A is 38.1 years old, has a body mass index of 20.0 kg/m^2, day 3 FSH of 9.8 mIU/mL, and normal semen analysis. Patient B is 34.4 years old, has a body mass index of 24.6, day 3 FSH of 7.4 mIU/mL, and normal semen analysis. According to PreIVF-D, the predicted probabilities of having successful first-IVF for patients A and B are 34.1% and 49.4%, respectively, which differ from age-based predicted probabilities of 26.5% and 42.2% for age groups 38–40 and <35 years, respectively (1, 4).

We do not know how many patients currently hold back from or delay pursuing their first IVF cycle because they perceive their success rates to be unacceptable. However, even if a small percentage of patients who would normally be hesitant to proceed would now feel more confident to pursue IVF based on personalized prognosis, that represents an increase in IVF utilization by good-to-excellent prognosis patients, which in turn would improve the overall success rates and utilization of IVF. Therefore, we propose that the use of rigorously developed and validated personalized prediction tool in the infertility community will improve access to and utilization of ART care to help a greater number of patients to build healthy families.

Acknowledgments: The authors thank Beth Malizia, M.D., Michele R. Hacker, Sc.D., Laura Dodge, M.P.H., and Brent Barrett, Ph.D., for database design and data processing at BIVF; Denny Sakkas, Ph.D., for discussion; Rocio Gandia Franco for database design and data extraction at IVI; Alina Tartia, Ph.D., for database work at OFC; and Kenneth Santo-Domingo, M.Sc., at Univfy for his support in data analysis and manuscript preparation.

REFERENCES

1. Society for Reproductive Technologies. SART CORS Online. All SART Member Clinics: Clinic Summary Report. Available at: https://www.sartcorsonline.com/rptCSR_PublicMultYear.aspx?ClinicPKID=0. Accessed February 5, 2013.
2. American Society for Reproductive Medicine. ASRM resources for patients. Available at: http://www.reproductivefacts.org/. Accessd February 5, 2013.
3. Bunting L, Boivin J. Development and preliminary validation of the fertility status awareness tool: Fertistat. Hum Reprod 2010;25:1722–33.
4. Sunderam S, Kissin DM, Flowers L, Anderson JE, Folger SG, Jamieson DJ, et al. Assisted reproductive technology surveillance—United States, 2009. MMWR Surveill Summ 2012;61:1–23.
5. Banerjee P, Choi B, Shahine LK, Jun SH, O'Leary K, Lathi RB, et al. Deep phenotyping to predict live birth outcomes in in vitro fertilization. Proc Natl Acad Sci U S A 2010;107:13570–5.
6. Lannon BM, Choi B, Hacker MR, Dodge LE, Malizia BA, Barrett CB, et al. Predicting personalized multiple birth risks after in vitro fertilization-double embryo transfer. Fertil Steril 2012;98:69–76.
7. Leushuis E, van der Steeg JW, Steures P, Bossuyt PM, Eijkemans MJ, van der Veen F, et al. Prediction models in reproductive medicine: a critical appraisal. Hum Reprod Update 2009;15:537–52.
8. Eaton JL, Hacker MR, Harris D, Thornton KL, Penzias AS. Assessment of day-3 morphology and euploidy for individual chromosomes in embryos that develop to the blastocyst stage. Fertil Steril 2009;91:2432–6.
9. Labarta E, Bosch E, Alama P, Rubio C, Rodrigo L, Pellicer A. Moderate ovarian stimulation does not increase the incidence of human embryo chromosomal abnormalities in in vitro fertilization cycles. J Clin Endocrinol Metab 2012;97:E1987–94.
10. Wen SW, Leader A, White RR, Leveille MC, Wilkie V, Zhou J, et al. A comprehensive assessment of outcomes in pregnancies conceived by in vitro fertilization/intracytoplasmic sperm injection. Eur J Obstet Gynecol Reprod Biol 2010;150:160–5.
11. Garrido N, Bellver J, Remohi J, Simon C, Pellicer A. Cumulative live-birth rates per total number of embryos needed to reach newborn in consecutive in vitro fertilization (IVF) cycles: a new approach to measuring the likelihood of IVF success. Fertil Steril 2011;96:40–6.
12. Malizia BA, Hacker MR, Penzias AS. Cumulative live-birth rates after in vitro fertilization. N Engl J Med 2009;360:236–43.
13. Yao M. Part 1: Complex IVF data and machine learning. In: Fertility chronicles: predicting IVF success 101. October 5, 2012. Available at: https://www.univfy.com/fertilitychronicles/personalizedprognosticsblog. Accessed February 5, 2013.
14. Yao M. Part 2: Rethinking our assumptions in fertility. In: fertility chronicles: predicting IVF success 101. October 12, 2012. Available at: https://www.univfy.com/fertilitychronicles/rethinking-IVF-assumptions. Accessed February 5, 2013.
15. Yao M. Part 3: Applying boosted tree to build IVF prediction models. In: Fertility chronicles: predicting IVF success 101. October 22, 2012. Available at: https://www.univfy.com/fertilitychronicles/Boosted-Tree. Accessed February 5, 2013.
16. Yao M. Part 4: Testing whether an IVF prediction model "works." In: Fertility chronicles: predicting IVF success 101. November 21, 2012. Available at: https://www.univfy.com/fertilitychronicles/Prediction-Modeling. Accessed February 5, 2013.
17. Yao M. Part 5: How accurate is your IVF prediction model? In: Fertility chronicles: predicting IVF success 101. December 13, 2012. Available at: https://www.univfy.com/fertilitychronicles/Accuracy-MY. Accessed February 5, 2013.