Deep phenotyping to predict live birth outcomes in in vitro fertilization

Prajna Banerjee[a,1], Bokyung Choi[b,1], Lora K. Shahine[a,c], Sunny H. Jun[a,d], Kathleen O'Leary[a], Ruth B. Lathi[a], Lynn M. Westphal[a], Wing H. Wong[e], and Mylene W. M. Yao[a,2]

[a]Department of Obstetrics and Gynecology, Stanford University School of Medicine, Stanford, CA 94305; [b]Department of Applied Physics, School of Humanities and Sciences, Stanford University, Stanford, CA 94305; [c]Pacific Northwest Fertility and In Vitro Fertilization Specialists, Seattle, WA 98104; [d]Department of Obstetrics and Gynecology, Palo Alto Medical Foundation, Fremont, CA 94538; and [e]Department of Statistics, School of Humanities and Sciences, Stanford University, Stanford, CA 94305

Edited* by Grace Wahba, University of Wisconsin, Madison, WI, and approved June 1, 2010 (received for review February 24, 2010)

Nearly 75% of in vitro fertilization (IVF) treatments do not result in live births and patients are largely guided by a generalized age-based prognostic stratification. We sought to provide personalized and validated prognosis by using available clinical and embryo data from prior, failed treatments to predict live birth probabilities in the subsequent treatment. We generated a boosted tree model, IVF_{BT}, by training it with IVF outcomes data from 1,676 first cycles (C1s) from 2003–2006, followed by external validation with 634 cycles from 2007–2008, respectively. We tested whether this model could predict the probability of having a live birth in the subsequent treatment (C2). By using nondeterministic methods to identify prognostic factors and their relative nonredundant contribution, we generated a prediction model, IVF_{BT}, that was superior to the age-based control by providing over 1,000-fold improvement to fit new data ($p < 0.05$), and increased discrimination by receiver–operative characteristic analysis (area-under-the-curve, 0.80 vs. 0.68 for C1, 0.68 vs. 0.58 for C2). IVF_{BT} provided predictions that were more accurate for ~83% of C1 and ~60% of C2 cycles that were out of the range predicted by age. Over half of those patients were reclassified to have higher live birth probabilities. We showed that data from a prior cycle could be used effectively to provide personalized and validated live birth probabilities in a subsequent cycle. Our approach may be replicated and further validated in other IVF clinics.

regression tree model | in vitro fertilization prediction | live birth prediction | in vitro fertilization prognostics | personalized medicine

In vitro fertilization (IVF) has enabled the conception of 1% of newborns in the United States per year, and 1 million babies worldwide since its inception (1, 2). For most subfertile couples, IVF treatment offers the highest live birth rate per treatment cycle. However, the decision to pursue IVF treatment after a failed attempt is challenging due to the high cost and uncertain prognosis. This lack of information about how to modify treatments to improve a couple's chance of a live birth may contribute to risks of multiple gestations or further futile treatments (3, 4).

Numerous factors, such as patient's age and embryo parameters, are associated with IVF outcomes (5–7), but their relative influence on live birth outcomes is not understood. Thus, prognostic counseling in IVF has largely been guided by age-based data with minor adjustments based on other clinical factors (2, 8–10). In addition, IVF prediction models described previously have limited utility. Some models predicted pregnancy status rather than live births; others were developed before current clinical and laboratory protocols; and importantly, most prediction models did not link outcomes of cryopreserved–thawed embryo transfers (11–13).

Previous IVF prediction models have also not been evaluated by all key criteria—likelihood to fit (i.e., predict), calibration, discrimination, and reclassification—that are specific and essential to prognostic models. For example, prediction of a future clinical state is most meaningfully expressed as probability, rather than "yes" or "no" prediction often used in diagnostics (13–15). In addition, using sensitivity and specificity to measure "accuracy" may not be as meaningful for a prognostic test as it is for diagnostics. Instead, the power and utility of a prognostic test is measured by its likelihood to fit new data, concordance between predicted probabilities to observed outcomes, range of probabilities that can be predicted, and its ability to discriminate patients by prognoses. Most importantly, these criteria must be externally validated by an independent dataset (13).

We aimed to develop a prognostic tool that would provide patients with an evidence-based and personalized prediction of their live birth outcome. We propose that the probability of live birth outcome per cycle is determined by a highly predictable component, in addition to random effects. We previously proved that pregnancy status could be predicted by boosted tree analysis that stratified patients according to clinical profiles (16). In this study, we used boosted tree to perform "deep phenotyping" (17) (e.g., the sorting of patients into subsets defined by similar clinical characteristics) by using data that are known prior to, and during, the first IVF cycle, to generate a model for predicting live birth probabilities in IVF (see *Methods*).

Rigorous evaluation of this prediction model, IVF_{BT}, showed that it is superior to age-based control models according to emerging statistical criteria for prognostics (15). Many variables were found to have unique and complex relationships. By *not* making assumptions about these relationships or the relative influence of variables, we were able to use available clinical and embryo data to better predict live birth outcomes in the current/first cycle (C1). However, because embryo data are critical to the establishment of the IVF_{BT} model, this prediction model will not be able to provide live birth probabilities prior to starting the first IVF cycle. Nevertheless, importantly, we have shown that this clinical data pertaining to a failed IVF cycle can predict the couple's probability of a live birth in a subsequent cycle (C2). Further, our strategy is fluid and may be replicated in other IVF clinics. This model may be further validated for its direct applications in other clinics to significantly improve prognostic counseling after

Author contributions: P.B., B.C., L.M.W., W.H.W., and M.W.M.Y. designed research; P.B., B.C., L.K.S., S.H.J., and M.W.M.Y. performed research; P.B., B.C., L.K.S., S.H.J., K.O., R.B.L., L.M.W., W.H.W., and M.W.M.Y. analyzed data; and P.B., B.C., L.K.S., S.H.J., K.O., R.B.L., L.M.W., W.H.W., and M.W.M.Y. wrote the paper.

Conflict of interest statement: M.W.M.Y and W.H.W. have cofounded a new company to make personalized prognostics accessible to IVF patients. The company is in the start-up, prefunding, precommercialization stage at the time of manuscript submission. L.K.S. is an employee at Pacific Northwest Fertility and IVF Specialists, and S.H.J. is an employee at Palo Alto Medical Foundation.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

See Commentary on page 13559.

[1]P.B. and B.C. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: mylene.yao@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1002296107/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1002296107

Downloaded at Palestinian Territory, occupied on December 21, 2021

www.manaraa.com

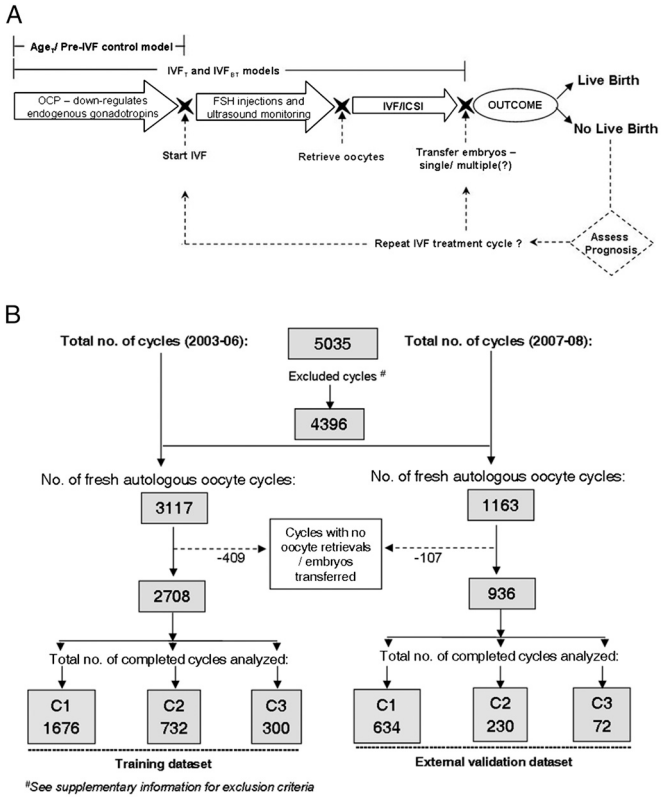

Fig. 1. IVF treatment cycle and data collection from the SHC. (*A*) Sequence of events in a typical IVF treatment cycle. This schematic describes the factors analyzed during model generation with respect to an IVF treatment cycle. If no live birth results and there are no cryopreserved embryos available, then the ability to predict subsequent cycle outcomes a priori would support decision-making. OCP, oral birth control pill, is commonly used to down-regulate endogenous gonadotropin hormone secretion; ICSI, intracytoplasmic sperm injection, is commonly used instead or in addition to in vitro fertilization to fertilize eggs in vitro; FSH, follicle stimulating hormone. (*B*) Source of clinical data. The numbers in the gray boxes indicate the number of fresh cycles performed at SHC from 2003–2008. See *SI Methods* for detailed explanation of exclusion criteria. The fresh cycles analyzed were further classified into C1, C2, and C3, which tracked data and outcomes to the corresponding fresh IVF cycle. The red boxes represent the three datasets used for model generation and validation.

failure of an IVF cycle (Fig. 1*A*). Finally, this approach may also be applied in future investigations to better identify patients who are at risk for multiple births.

Results

Cycles, Variables, and Their Association with Live Birth Outcomes. Data and outcomes for 5,035 IVF treatments performed in 2003–2008 were retrieved. Inclusion and exclusion criteria were fulfilled by 3,117 fresh, autologous oocyte IVF cycles from 2003–2006. The training set was formed by 2,708 completed cycles, which comprised 1,676; 732; and 300 completed C1, C2, and C3 cycles, respectively, for analysis (Fig. 1*B*). Overall, the live birth rates for C1, C2, and C3, were 29%, 18%, and 14%, respectively. Of 1,196 C1 patients (71% of 1,676) who did not have a live birth, 732 (61%) returned for C2 treatment and 464 (39%) of patients dropped out (see *SI Text*).

Thirty of the 52 variables were confirmed to be significantly associated with live birth outcomes by univariate analysis (p value < 0.05, Table 1). For example, normal fertilization and the rate of blastocyst development were positively associated, whereas clinical diagnosis of diminished ovarian reserve and the number of unfertilized eggs were negatively associated with live birth outcomes.

We previously found that stepwise logistic regression was not appropriate (*SI Text*). In addition, many pairs of continuous variables were highly correlated by Pearson correlation coefficient (Table S1). However, the potential for significant interactions among variables was not known. For these reasons and others discussed above, we applied a boosted tree method to 1,676 completed C1 and their outcomes, including outcomes of 440 linked cryopreserved–thawed embryo transfer procedures from the 2003–2006 training set, to generate prediction models for live birth outcomes (Fig. 1*B* and Table 1).

Personalized Prognosis Without Stratification. Traditional classification tree models have been used in infertility research to establish predictive criteria by training prediction models to recognize complex relationships among variables (16, 18). Using the dataset comprising 1,676 C1 patients from 2003–2006 (hereafter, training dataset) (Fig. 1*B*), we generated an example of a tree model by classification and regression trees (CART) (19), and demonstrated that it offered improved live birth prediction compared to an age-control model in C1 (see *SI Methods* and Fig. S1). Most importantly, CART analysis revealed unique and complex interactions at least among the top five prognostic variables (*SI Text* and Fig. S1). However, the boosted tree model, which aggregates a collection of simple trees, is known to produce significantly superior results (20–23). Hence, we applied the boosted tree methodology to generate a model (IVF_{BT}), to predict live birth outcomes.

Data analysis by Generalized Boosted Models (GBM®) (23), a free software implementation of stochastic gradient boosting algorithm (20), revealed the relative and nonredundant influence of all 52 variables, which was used to generate the IVF_{BT} model to predict a continuous range of live birth outcome probabilities, without partitioning the population into discrete and discontinuous groups (see *Methods*). Relative influence is the independent contribution to outcomes that is made by each variable independent of the other 51 variables and is scaled to 100 as the maximum relative influence made collectively by all variables. The top influential factors were rate of blastocyst development (relative influence, 26%), the total amount of gonadotropins administered (10%), the number of eight-cell embryos (9%), embryo cryopreservation (7%), the age of female patient (6%), endometrial thickness (6%), and total number of embryos (6%) (Fig. 2).

We externally validated the IVF_{BT} model using an independent dataset comprising 634 C1 and 230 C2 from 2007–2008 (hereafter, external validation set), that are unrelated to the patients and cycles of the training set used to generate the model (Fig. 1*B* and Table S2). Of note, 21 of 52 clinical variables (40%) were significantly different between the training and validation datasets, including age, amount of gonadotropins required, and number of eight-cell embryos (Table S2). Thus, the independent dataset truly served as external validation. To evaluate the predictive power of IVF_{BT}, we tested the null hypothesis that the like-

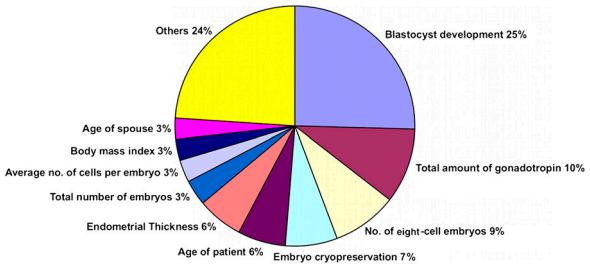

Fig. 2. Prognostic factors and their relative influence in the IVF_{BT} model. Of 52 variables analyzed, the top 10 nonredundant variables and their "relative influence" are shown here. The model was established by setting the sum of all relative influences to 100%. Variables with an individual influence of <2.5% were placed under "Others."

Downloaded at Palestinian Territory, occupied on December 21, 2021
www.manaraa.com

Table 1. Variables analyzed and their association with live birth outcome by univariate analysis of 1,676 C1 cycles from 2003–2006 (training dataset)

Variables	Estimate*	SEM	p value†	Mean‡	SD	% missing§
Pre-IVF factors						
Continuous variables						
Age of patient, yr	−3.50E−04	3.60E−05	$p < 0.001$	40.99	4.56	0
Age of male partner, yr	−8.30E − 05	1.80E−05	$p < 0.001$	42.95	5.82	0
Body mass index, kg/m^2	−5.60E−03	7.30E−03	0.53	24.9	9.75	9.4
No. of previous pregnancies¶	−4.40E−02	4.70E−02	0.39	1.04	1.33	0.1
No. of previous term deliveries‖	−2.00E−01	8.50E−02	0.04	0.27	0.6	2.8
Spontaneous miscarriages**	−1.80E−01	7.30E−02	0.04	0.41	0.82	1.7
Serum day 3 FSH, IU/L	−4.00E−02	1.60E−02	0.02	8.05	4.92	5.1
Year	+2.60E−02	4.80E−02	0.66	4.32	1.12	0
Categorical variables						
Diminished ovarian reserve	−1.10E+00	1.20E−01	$p < 0.001$	0.39	0.49	0
Polycystic ovarian syndrome	+6.20E−01	1.70E−01	$p < 0.001$	0.09	0.29	0
Unexplained female infertility	+4.50E−01	1.80E−01	0.02	0.09	0.28	0
Other causes for infertility	−1.30E−01	1.20E−01	0.37	0.34	0.47	0
Tubal disease	−9.50E−02	1.70E−01	0.65	0.12	0.32	0
Uterine fibroids	−2.00E−01	2.00E−01	0.40	0.09	0.28	0
Endometriosis	−7.30E−02	1.60E−01	0.67	0.14	0.35	0
Male infertility	+4.50E−02	1.10E−01	0.72	0.42	0.49	0
Male-only infertility causes	+4.60E−01	1.60E−01	0.009	0.11	0.32	0
Tubal ligation	+1.20E−01	4.10E−01	0.79	0.01	0.12	0
Hydrosalpinx	+1.80E−01	3.70E−01	0.67	0.02	0.14	0
Season	+2.10E−01	1.50E−01	0.32	0.22	0.42	0
Protocol and treatment factors						
Continuous variables						
Total amount of gonadotropin, units	−3.60E−04	2.70E−05	$p < 0.001$	4,769.14	1990.4	0
Endometrial thickness, mm	+1.90E−01	2.80E−02	$p < 0.001$	10	2.1	0.3
No. of sperm motile after wash, million/mL	−3.60E−04	2.90E−04	0.32	128.41	244.6	2.2
No. of sperm motile before wash, million/mL	−2.60E−04	3.30E−04	0.53	106.46	198.93	2.3
Total no. of oocytes	+9.9.0E−02	8.70E−03	$p < 0.001$	10.25	6.63	0
Normal and mature oocytes, %	+1.30E+00	1.10E+00	0.35	0.97	0.08	0.8
Normal fertilization, %	+9.20E−01	2.60E−01	$p < 0.001$	0.65	0.24	0
Unfertilized eggs, %	−7.90E−01	2.70E−01	0.007	0.27	0.23	0
Abnormally fertilized eggs, %	−6.50E−01	4.90E−01	0.30	0.08	0.14	0
Total number of embryos	+1.50E−01	1.20E−02	$p < 0.001$	5.92	4.94	0
Compaction on day 3	+2.80E−01	2.70E−01	0.40	0.09	0.2	0
Average no. of cells per embryos	+2.90E−01	4.50E−02	$p < 0.001$	6.75	1.32	2.6
No. embryos arrested at ≥4 cells	−1.30E−02	2.70E−03	$p < 0.001$	16.74	23.52	2.6
No. of eight-cell embryos	+2.70E−01	2.20E−02	$p < 0.001$	2.81	2.84	2.6
Percentage of eight-cell embryos, %	+9.50E−03	1.90E−03	$p < 0.001$	40.46	29.08	2.6
Blastocyst development, %	+3.20E+00	2.40E−01	$p < 0.001$	0.15	0.23	0
Embryo cryopreservation, %	+2.30E+00	2.50E−01	$p < 0.001$	0.11	0.19	0
Average grade of embryos	−1.50E−01	1.10E−01	0.29	1.83	0.61	35.5
Total no. of transferred embryos	−9.00E−02	4.10E−02	0.06	2.57	1.32	0
Average nos. of cells per transferred embryos	+4.20E−01	6.60E−02	$p < 0.001$	7.17	1.35	0.5
Average grade of transferred embryos	−2.70E−01	1.40E−01	0.10	1.74	0.63	33.3
No. of transferred embryos with ≥4 cells	−2.10E−02	5.00E−03	$p < 0.001$	9.9	24.1	0.5
No. of transferred embryos with ≥8 cells	+3.40E−01	6.10E−02	$p < 0.001$	1.38	1.16	0.5
Percentage of eight-cell embryos transferred	+1.20E−02	2.00E−03	$p < 0.001$	48.71	37.61	0.5
Categorical variables						
Oral contraception for down-regulation of endogenous gonadotropins	+6.60E−01	1.90E−01	0.001	0.87	0.34	0.1
Stimulation protocol	−1.20E+00	1.20E−01	$p < 0.001$	0.47	0.5	0.7
Sperm from male partner	+2.20E−01	2.60E−01	0.52	0.04	0.2	0
Sperm from donor	−4.00E−01	4.60E−01	0.51	0.02	0.13	0
Performance of intracytoplasmic sperm injection	+3.30E−01	1.10E−01	0.003	0.41	0.49	0
Assisted hatching	−9.50E−01	1.90E−01	$p < 0.001$	0.32	0.47	0
Day 5 embryo transfer	+1.40E+00	1.20E−01	$p < 0.001$	0.27	0.44	5.0
Preimplantation genetic diagnosis/screening	+3.00E−02	2.80E−01	0.91	0.04	0.19	0

The top half of Table 1 (Pre-IVF factors) represents 20 variables that are available before the start of the IVF treatment cycle. The bottom half of Table 1 (Protocol and treatment factors) represents 32 variables that are based on the treatment cycle and protocol. The variables have also been classified as continuous or categorical based on their numeric/nonnumeric values. FSH, follicle stimulating hormone.

*Positive and negative estimates indicate positive and negative association with live birth outcomes, respectively.

†The p value represents significance of association with live birth outcome.

‡Mean for continuous variables indicates the mean value of each variable among the entire dataset analyzed; mean for categorical variables indicates the average number of positive occurrences in the entire dataset analyzed.

§Percentage of data entries missing. Note: Data entries for "Average grade of embryos" was not complete for ~30% of cycles, and "Compaction rate" was not entered routinely in the 2007–2008 dataset.

¶Number of previous positive clinical pregnancies

‖Number of previous deliveries carried to term, 37 weeks gestation.

**Miscarriages refer to development stopped at or after 5 weeks gestation.

Downloaded at Palestinian Territory, occupied on December 21, 2021
www.manaraa.com

lihood of IVF_{BT} to fit independent data was equal to that of control models, and found that IVF_{BT} was more than 1,000 times more likely to fit new data compared to a control model that used age alone (Age_T) or variables that are known prior to starting IVF (pre-IVF factors) ($p < 0.0001$; see *SI Methods*). Hence, the use of pre-IVF factors to predict C2 outcomes was not further pursued.

Thus, we have shown that live birth probabilities in IVF can be predicted using clinical variables known prior to starting IVF, response to hormonal stimulation, and embryo parameters. When used alone, age may be misleading as a prognostic factor, due to the complex and as yet uncharacterized relationships among age, hormonal response, and embryo parameters such as blastocyst formation rate and total number of embryos (Fig. 2*A*).

Predicting Live Birth Outcomes for the Next Cycle. The IVF_{BT}-model-required embryo data, which would not be known prior to starting an IVF cycle, are critical to development of the IVF_{BT} model. To determine its clinical utility, we tested whether data from the index cycle (C1) could serve as proxy for C2 data to predict live birth outcomes for C2, without using clinical data from C2. Our goal was to simulate the scenario in which patients do not have a live birth after C1 and wish to know their personalized probabilities to have a live birth if they were to repeat the same treatment (e.g., C2).

Although cycle number itself was not found to be an independent predictor by GBM analysis, we found that the clinical/embryo phenotypes were typically worse in C2 than C1 (Table S2). Thus, using a prediction model that is based on C1 data alone would overestimate the live birth probabilities for C2. Using the IVF_{BT} model, we generated C1–C2 predicted probability pairs from the training data, to determine a linear model in the logit scale, to measure the difference between C1 and C2 predictions. This linear model was then used to calculate IVF_{BT}-C2, by adjusting the IVF_{BT} prediction for C2. The Age_T model was modified by using patient's age at C2 (rather than C1) in the training set to generate Age_T-C2, which controls for C2 prediction that was attributed to age alone, without the use of other pre-IVF, treatment, or embryo factors. All of the variables used C1 values as proxy to predict C2 outcomes in the IVF_{BT}-C2 model.

The use of IVF_{BT}, which was generated using C1 data only, would not have introduced bias from self-selection of C2 population (e.g., patients' decision to take C2 treatment or to drop out of treatment), but the use of C2 outcomes in the adjustment procedure to generate IVF_{BT}-C2 had the potential to introduce bias. However, we ascertained that introduction of this type of bias was unlikely because the probabilities of live birth outcome were not differentially distributed between patients who returned for C2 treatment and patients who dropped out of treatment (see *SI Text* and Table S3). Finally, we evaluated the predictive power of IVF_{BT}-C2 as described for IVF_{BT}-C1 above. We tested the null hypothesis that the likelihood of IVF_{BT}-C2 to fit independent data was equal to that of the Age_T model, and found that IVF_{BT}-C2 was 1,000 times more likely to fit new data compared to the Age_T model ($p < 0.05$). Therefore, IVF_{BT}-C2 improved the fit significantly to the new data compared to control (see *SI Methods*).

Testing Calibration and Discrimination of Prognostic Model. The applicability of a prognostic tool also depends on calibration, which measures the concordance between predicted probabilities and rates of observed outcomes (13, 14). Therefore, calibration could be thought of as a measure of accuracy that is meaningful for prognostic tests. Using the external validation set, we tested calibration by stratifying patients into groups based on their predicted probabilities of live birth by IVF_{BT} and Age_T for C1 prediction, and $IVF_{BT}C2$ and Age_TC2 for C2 prediction (Fig. 3 *A* and *B*). Overall, both sets of models were very well calibrated at a 95% confidence interval ($p > 0.1$, Holsmer–Lemeshow goodness-of-fit test, Fig. 3 *A* and *B*, and *SI Methods*).

Nevertheless, calibration graphs could not fully illustrate the dynamic range of predicted probabilities for all patients. The significantly wider dynamic range of IVF_{BT} compared to Age_T is demonstrated by a scatter plot analysis of predicted probabilities for all patients in the external validation sets for C1 and C2 predictions. For example, for C1 prediction, IVF_{BT} predicted live birth probabilities ranging from near zero to 80%, whereas conventional age categories predicted discrete live birth probabilities ranging from 5% to 41% (Fig. 3*C*). Similarly, for C2 prediction, IVF_{BT} predicted live birth probabilities ranging from near zero to ~50%, whereas adjusted age categories predicted probabilities ranging from near zero to 33% (Fig. 3*D*).

Interestingly, although cycle number did not have independent effects on live birth outcomes by GBM analysis, age-based prediction of live birth outcomes for C1 and C2 were significantly different ($p < 0.005$, two-way ANOVA test, Fig. 3 *C* and *D*). These results highlight the overestimation of live birth rates per cycle, using the current age-based paradigm, which is typically presented without adjustments for repeat treatments (9, 24).

We further assessed the ability of IVF_{BT} to discriminate patients with different probabilities of live birth by computing the area under the receiver–operator curve (14). The area under the curve (AUC) measures discrimination based on the true and false positive rates at a series of arbitrarily defined thresholds. Although the AUC may be less meaningful for prognostic than diagnostic tests (15), it allowed a direct comparison with the Templeton model (25), one of few reputed live birth prediction models in IVF (AUC = 0.63). Using the external validation set to test C1 prediction, we found that IVF_{BT} (AUC = 0.8) was superior to Age_T (AUC = 0.68) (Fig. 3*E*). For C2 prediction, $IVF_{BT}C2$ (AUC = 0.68) was similarly superior to Age_TC2 (AUC = 0.58) (Fig. 3*F*). Thus, both prediction models were ~17% more discriminatory than respective age controls and the Templeton model.

Finally, to understand the full utility of our prediction model, we determined the percentage of patients whose prognosis would be reclassified—be given a different probability of achieving live birth—by IVF_{BT} compared to Age_T. Overall, IVF_{BT} and $IVF_{BT}C2$ models predicted live birth probabilities that were significantly different and out of the range predicted by their respective age controls in ~83% and ~60% of patients, respectively (Fig. 3 *C* and *D*). Of patients that were reclassified, 50% and 60% were assigned higher predicted live birth probabilities by IVF_{BT} and $IVF_{BT}C2$ models, respectively, compared to the probabilities predicted by age models. The reclassification rate for $IVF_{BT}C2$ would have been even higher if the validation set had comprised a larger number of cycles.

Collectively, given the superior fit and calibration of IVF_{BT}, its high rate of reclassification indicated that the current age-based paradigm may provide misleading live birth outcome probabilities for a large portion of patients. Overall, we have demonstrated that the IVF_{BT} model, which utilized a large composite of commonly recorded clinical variables, is robust and superior to the age-control model. Most importantly, we provided proof that live birth outcomes can be predicted by applying the IVF_{BT} model to analyze clinical data obtained from a previous IVF cycle at the same clinic.

Discussion

Both medical and personal decisions regarding IVF treatment may hinge on the probability of live birth outcome for many patients. Given the financial, physical, and emotional costs of IVF, high-quality personalized prognostic information should assist patients' decisions to continue treatment, pursue alternative options such as egg or embryo donation, or drop out of treatment (26). Patients have been counseled to focus on cumulative live birth rates rather than success rate per cycle (8, 27). However,

SEE COMMENTARY
APPLIED MATHEMATICS
Downloaded at Palestinian Territory, occupied on December 21, 2021
www.manaraa.com

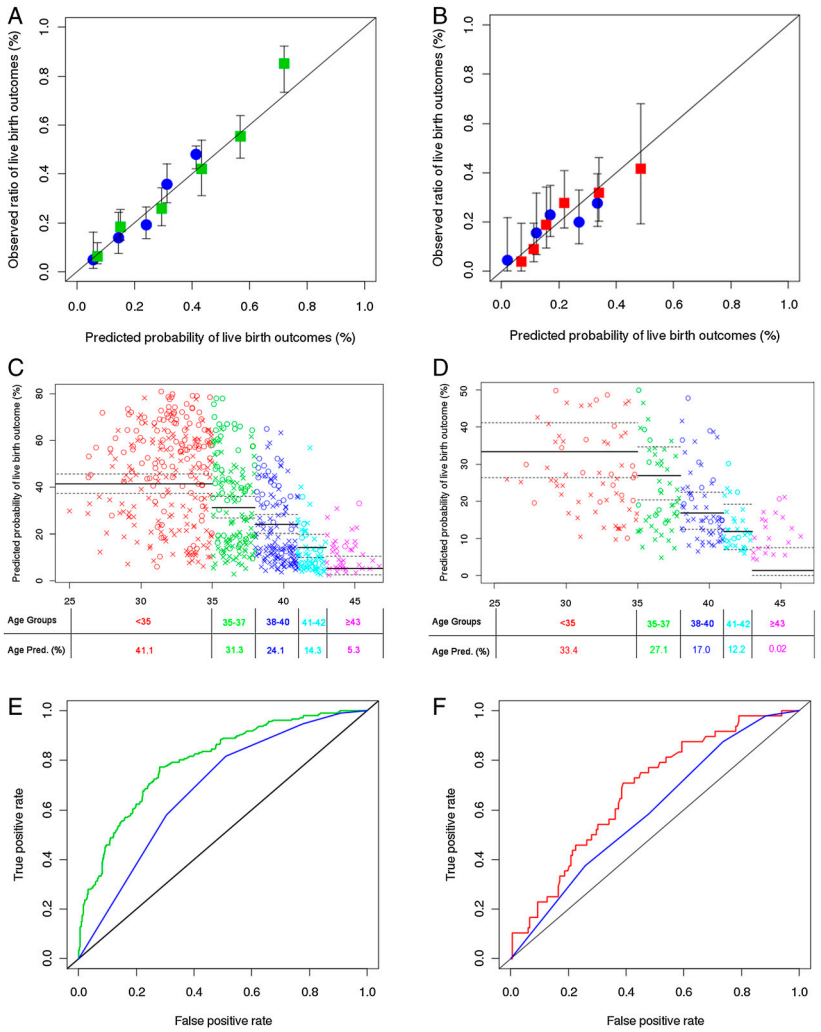

Fig. 3. Evaluation of IVF_{BT} and IVF_{BT} C2 in predicting live birth outcomes in index (C1) and subsequent cycle (C2). (*A*) Model calibration for C1 predictions. The five groups in the Age_T (blue circles) correspond to mean live birth probabilities as predicted by Age_T for age groups <35, 35–37, 38–40, 41–42, and ≥43, whereas the six groups in the IVF_{BT} model (*A*, green squares; *B*, red squares) indicate those predicted by IVF_{BT} model. Error bars indicated 95% confidence interval. Calibration was measured by comparing the predicted and observed live birth rates of 634 index (C1) cycles (*A*, green squares) from the 2007–2008 validation set. Note that the grouping was only performed here to facilitate comparison between IVF_{BT} and the Age_T models as grouping itself was not required for model development. The diagonal line in the graph refers to an ideal calibration if the predicted and observed live birth probabilities were identical. (*B*) Model calibration for C2 predictions. The mean live birth probabilities predicted by Age_TC2 (blue circles) and IVF_{BT}C2 (red squares) were compared to mean observed outcomes for 230 C2 in the external validation dataset from 2007–2008. Both Age_TC2 and IVF_{BT}C2 were found to be well calibrated ($p > 0.1$). (*C* and *D*) Patient-specific predictions and reclassification. To compute reclassification of predicted outcomes by IVF_{BT}, we plotted each patient from 634 C1 (*C*) and 230 C2 (*D*) from the 2007–2008 validation set, based on age and predicted live birth probabilities. The observed outcome of each patient was plotted as live birth (O) and no live birth (X). The bold lines indicate probabilities of live birth predicted by Age_T (*C*) or Age_TC2 (*D*) with their confidence intervals (upper and lower dotted lines). Note the narrow range of probabilities by age-control models compared to the wide dynamic range predicted by IVF_{BT} or IVF_{BT}C2. (*E* and *F*) Evaluation of discrimination by receiver-operative characteristic (ROC). ROC curves of the IVT_{BT} model (*E*, green line for C1 prediction; *F*, red line for C2 prediction) and Age_T or Age_TC2 (blue line) are shown for C1 prediction (*E*) and C2 prediction (*F*) of 2007–2008 validation sets. For both C1 and C2 predictions, the area under the ROC curve is significantly higher for the IVF_{BT} and IVF_{BT}C2, respectively, compared to age controls, which indicates that the IVF_{BT} model is more discriminatory.

live birth probabilities have traditionally been based on age, and determined for large groups of patients, despite limited applicability of age-based probabilities for the individual patient. Despite efforts to generate an IVF prediction model to address these needs, even the most rigorously tested published prediction models could not be validated (13).

We have established an externally validated, highly discriminatory, well-calibrated, and robust prediction model that can use available clinical data from a previous cycle to predict live birth rates in a subsequent cycle, without additional clinical or laboratory testing. The IVF treatment itself often has been proposed as a diagnostic/prognostic tool in addition to serving as a therapy. However, that concept has applied to a small subset of patients in whom very serious defects of sperm, oocyte, embryo, or fertilization would only be revealed by IVF (28, 29). Our findings show that the first IVF cycle can be both prognostic and therapeutic for *all* patients, because it would provide quantitative, customized prediction of the live birth probability in subsequent cycles. This concept is radically different from the current paradigm, in which age is a major predictor, and other factors may be used to adjust the outcomes according to various semiquantitative scoring methods.

This study of IVF outcomes analysis has comprehensively and simultaneously addressed challenges that have previously contributed to the "black-box" nature of IVF "success rates." Mechanistically, the most significant finding is the complex relationships among key prognostic factors in influencing live birth outcomes. For translational and clinical applications, patient-specific prediction of live birth outcome can be applied to support physicians in counseling patients and to empower patients in decision-making after a failed cycle. Further, these patient-specific live birth probabilities may also be used to determine patient-specific cumulative live birth rates and the number of treatments that may be required to reach a live birth (8, 27). Other applications of these customized prognostics may include improved candidate identification for elective single-embryo transfer to decrease the rate of multiple gestations (30), enhanced clinical research trial design by refining patient selection, and ultimately, improved maternal and neonatal health by determining risk factors. Finally, the power of boosted tree analysis in predicting IVF outcomes suggests that clinical questions in other areas of medicine may also benefit from such deep phenotyping.

The strengths of our study include the analysis of many variables without the need to assume their interactions a priori. Our model analyzed live birth rates pragmatically, by including outcomes of linked cryopreserved embryos. The application of our model was not affected by patients' self-selection for subsequent treatment, which might be a concern in other studies (8). Specifically, we found that the IVF_{BT} profiles of patients who dropped out after failed C1 treatments were similar to patients who returned for a repeat treatment (Table S3). These findings suggested that some patients might have been influenced by self-perceived probabilities of live birth outcomes that had little correlation with their true prognosis. Alternatively, some patients might have dropped out due to factors unrelated to prognosis, such as finances, or a combination of these and other factors.

Downloaded at Palestinian Territory, occupied on December 21, 2021
www.manaraa.com

Our work is clinically applicable because data collection methods met high standards of quality IVF clinics. Our approach does not require fixed coefficients or variables that are typical of logistic regression models. Because treatment regimens, embryology laboratory protocols, and data collection vary among clinics, a key strength is that our approach may be replicated for validation in other IVF clinics to establish clinic-specific prediction models that are meaningful to patients in the context of their clinic. This study is limited by constraints related to data collection present in most IVF clinics. First, some factors that may have prognostic value, such as ethnicity (31), antral follicle counts, and serum anti-Mullerian hormone levels, were not available in our dataset; hence, they could not be tested in our models. Second, many indications for IVF treatment were not defined quantitatively, and their use might have varied among physicians, which might explain why baseline diagnoses did not contribute to the prediction model. The prognostic value of variables that are not well defined in our dataset may be evaluated by analyzing datasets from clinics that use and record those variables routinely. Hence, live birth prediction models generated at different clinics may comprise different predictors, depending on the available variables, clinical volume, and any clinic- or demographics-specific nuances that are not currently understood. Although most of the top prognostic factors are expected to be consistently significant among live birth prediction models derived from different clinics, we anticipate that their relative importance in predicting live birth outcomes may be different among clinics.

Alternatively, we propose that IVF overrides many pathophysiological factors of infertility, such as anovulation, tubal disease, and male factor. Thus, conventional clinical diagnoses may be less relevant in predicting IVF outcomes, as failure likely results from molecular defects in oocyte, sperm, or embryos that are currently not overcome by IVF. Despite being influenced by multiple maternal, paternal, and embryo variables, and their potential interactions, we have established that live birth outcomes in IVF can be subjected to rigorous scientific investigation, and they can be predicted.

Methods

Data on clinical diagnoses, IVF treatment protocol and response, embryology data, and treatment outcomes for all IVF cycles performed between January 1, 2003 and December 31, 2008 at Stanford Hospital and Clinics (SHC) were retrieved from our clinical database (BabySentryPro, BabySentry, Ltd.) or medical records as necessary. Retrospective data collection, deidentification, and analysis were performed according to a human subjects protocol approved by the SHC Institutional Review Board.

The inclusion criteria for data analysis were freshly stimulated, nondonor oocyte IVF cycles performed at SHC. For each patient, the first IVF cycle performed at SHC was defined as C1, whereas subsequent cycles following failed IVF treatments were designated C2, C3, etc. We applied exclusion criteria and restricted our analysis to cycles that had complete embryo data to generate a model to predict live birth probabilities (see Fig. 1*B* and *SI Text* for exclusion criteria).

Fifty-two variables, selected for high data quality and completeness, were analyzed in an unbiased fashion without specific ranking a priori (Table 1). Over 90% of data fields were completed for 50 of the 52 variables (see Table 1 for details). Twenty of the 52 variables are factors known to physicians prior to starting an IVF treatment cycle (pre-IVF factors) and 32 variables become available only during or at the conclusion of an IVF treatment (protocol and treatment factors, Table 1 and Fig. 1*A*).

Here, the outcome of a cycle was determined by outcomes of transfers using fresh or cryopreserved–thawed embryos that resulted from fresh, stimulated cycles. Live birth was defined as the delivery of a live baby beyond 24 weeks gestation after the transfer of fresh or cryopreserved embryos that had resulted from eligible IVF cycles. The outcome, "no live birth," encompassed all other outcomes such as negative or declining serum β-human chorionic gonadotropin, clinical pregnancy loss, intrauterine fetal death, and ectopic pregnancy.

ACKNOWLEDGMENTS. We thank the Embryology staff, L. Morcom and M. Madrid for data entry; and B. Behr, V. Baker, P. Donnelly, C. Mak, A. Milki, D. Owen, S. Quake, R. Tibshirani, and L. Wu for valuable discussion. M.W.M.Y. received support from National Institutes of Health (NIH) Grants R01HD57970 and K12HD01249; W.H.W. received support from NIH 1R01HG004634. This project was supported by the Coulter Foundation Translational Research Program at Stanford University.

1. Bonduelle M, et al. (2005) A multi-centre cohort study of the physical health of 5-year-old children conceived after intracytoplasmic sperm injection, in vitro fertilization and natural conception. *Hum Reprod* 20:413–419.
2. Sunderam S, et al. (2009) Assisted reproductive technology surveillance—United States, 2006. *MMWR Surveill Summ* 58:1–25.
3. Jansen RP (2005) Benefits and challenges brought by improved results from in vitro fertilization. *Intern Med J* 35:108–117.
4. Verhaak CM, et al. (2007) Women's emotional adjustment to IVF: A systematic review of 25 years of research. *Hum Reprod Update* 13:27–36.
5. Giorgetti C, et al. (1995) Embryo score to predict implantation after in-vitro fertilization: Based on 957 single embryo transfers. *Hum Reprod* 10:2427–2431.
6. Volpes A, et al. (2004) Number of good quality embryos on day 3 is predictive for both pregnancy and implantation rates in in vitro fertilization/intracytoplasmic sperm injection cycles. *Fertil Steril* 82:1330–1336.
7. Srouji SS, et al. (2005) Predicting in vitro fertilization live birth using stimulation day 6 estradiol, age, and follicle-stimulating hormone. *Fertil Steril* 84:795–797.
8. Malizia BA, Hacker MR, Penzias AS (2009) Cumulative live-birth rates after in vitro fertilization. *N Engl J Med* 360:236–243.
9. *All SART Member Clinics: Clinic Summary Report* (Society for Assisted Reproductive Technologies, Birmingham, AL), This report summarizes information from all IVF cycles performed in 2006 in the United States that were submitted to SART by member clinics https://www.sartcorsonline.com/rptCSR_PublicMultYear.aspx?ClinicPKID=0.
10. Collins J (2001) Cost-effectiveness of in vitro fertilization. *Semin Reprod Med* 19:279–289.
11. Hunault CC, et al. (2004) Two new prediction rules for spontaneous pregnancy leading to live birth among subfertile couples, based on the synthesis of three previous models. *Hum Reprod* 19:2019–2026.
12. Stolwijk AM, et al. (1996) Prognostic models for the probability of achieving an ongoing pregnancy after in-vitro fertilization and the importance of testing their predictive value. *Hum Reprod* 11:2298–2303.
13. Leushuis E, et al. (2009) Prediction models in reproductive medicine: A critical appraisal. *Hum Reprod Update* 15:537–52.
14. Cook NR (2007) Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 115:928–935.
15. Cook NR (2008) Statistical evaluation of prognostic versus diagnostic models: Beyond the ROC curve. *Clin Chem* 54:17–23.
16. Jun S, et al. (2008) Defining human embryo phenotypes with cohort-specific prognostic factors in in vitro fertilization. *PLoS ONE* 3:e2562.
17. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. *Science* 322:881–888.
18. Guzick DS, et al. (2001) Sperm morphology, motility, and concentration in fertile and infertile men. *N Engl J Med* 345:1388–1393.
19. Breiman L, Friedman J, Olshen R, Stone C (1984) *Classification and Regression Trees* (Wadsworth, Belmont, CA).
20. Friedman J (1999) *Stochastic Gradient Boosting. Technical Report* (Department of Statistics, Stanford University, Stanford, CA) http://www-stat.stanford.edu/~jhf/ftp/stobst.pdf.
21. Friedman J (2002) *Tutorial: Getting Started with MART in R* (Department of Statistics, Stanford University, Stanford, CA) http://www-stat.stanford.edu/~jhf/r-mart/tutorial/tutorial.pdf.
22. Friedman J (1999) *Greedy Function Approximation: A Stochastic Boosting Machine* (Department of Statistics, Stanford University, Stanford, CA) http://www-stat.stanford.edu/~jhf/ftp/trebst.pdf.
23. Ridgeway G (2007) gbm: Generalized Boosted Regression Models. R Package Ver. 1.6-3. http://cran.r-project.org/web/packages/gbm/.
24. *Assisted Reproductive Technology Success Rates: Preliminary Data* (Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta).
25. Templeton A, Morris JK, Parslow W (1996) Factors that affect outcome of in-vitro fertilisation treatment. *Lancet* 348:1402–1406.
26. Domar AD, Smith K, Conboy L, Iannone M, Alper M (2009) A prospective investigation into the reasons why insured United States patients drop out of in vitro fertilization treatment. *Fertil Steril* doi:10.1016/j.fertnstert.2009.06.020.
27. Guzick DS, Wilkes C, Jones HW, Jr (1986) Cumulative pregnancy rates for in vitro fertilization. *Fertil Steril* 46:663–667.
28. De Sutter P (2006) Rational diagnosis and treatment in infertility. *Best Pract Res Cl Ob* 20:647–664.
29. Randolph JF, Jr (2000) Unexplained infertility. *Clin Obstet Gynecol* 43:897–901.
30. Khalaf Y, et al. (2008) Selective single blastocyst transfer reduces the multiple pregnancy rate and increases pregnancy rates: A pre- and postintervention study. *Bjog—Int J Obstet Gy* 115:385–390.
31. Shahine LK, et al. (2009) Poor prognosis with in vitro fertilization in Indian women compared to Caucasian women despite similar embryo quality. *PLoS One* 4:e7599.

Downloaded at Palestinian Territory, occupied on December 21, 2021

www.manaraa.com