Prediction modeling for live birth in *in vitro* fertilization

Kristen E. Gray

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Melissa A. Schiff, Chair

Jacqueline R. Starr

Kathleen Lin

Yingye Zheng

Program Authorized to Offer Degree:

Public Health-Epidemiology

University of Washington

Abstract

Prediction modeling for live birth in *in vitro* fertilization

Kristen E. Gray

Chair of the Supervisory Committee:

Melissa A. Schiff

Department of Epidemiology

Background: Approximately 150,000 women undergo *in vitro* fertilization (IVF) each year to treat infertility. The success of IVF is limited, and the procedure is costly, time-consuming, and poses physical and emotional health risks to the patient. Therefore, generating personalized probabilities of live birth may assist patients and clinicians in decision-making. We sought to examine the ability of individual biomarkers, including anti-Müllerian hormone (AMH, a biomarker of ovarian reserve), and multivariable models to predict the probability of live birth prior to initiating stimulation for IVF.

Methods: We included fresh, autologous IVF cycles initiated between 2005 and 2011 from five U.S. infertility clinics. We developed and validated multivariable models predicting probabilities of live birth in 23,154 first IVF cycles, as well as in 8,184 second IVF cycles after a single prior failed cycle using varying levels of model complexity: (a) backwards stepwise logistic regression (p>0.2) with only parameter main effects, (b) with main effects and interactions, and (c) boosted regression trees. For first cycles, eligible predictors included those obtained at the baseline infertility evaluation (e.g., demographics, anthropometrics, pregnancy history, infertility diagnosis, stimulation protocol); which were also examined in second cycles in addition to the treatment response in the previous failed cycle (e.g., dose of gonadotropins, egg and embryo characteristics, cycle outcome, etc.). For comparison, we fit age category and linear age models. Due to missing data we imputed 15 datatsets using multiple imputation by chained equations. In the 20% of data reserved for validation, we estimated the receiver operating characteristic curve (ROC), area under the ROC curve (AUC), and the difference in AUCs between all models, along with bootstrapped 95% confidence intervals (CIs). In a subsample of data from a single clinic, we evaluated the ability of AMH to predict live birth in all fresh, autologous IVF cycles from 2010-2011 (N=834) and compared to widely collected biomarkers of ovarian reserve, including age, antral follicle count (AFC), and follicle stimulating hormone (FSH). We estimated the ROC curves, AUCs, and difference in AUCs between biomarkers, along with bootstrapped 95% CIs. We also evaluated the performance of AMH within subgroups based on age, body mass index (BMI), polycystic ovary syndrome (PCOS) status, and infertility diagnosis.

Results: In first IVF cycles, all predictors were included in the main effects and interactions model. All multivariable models performed similarly (AUCs=0.67, 95% CIs=0.66, 0.69) and only slightly better than age-based models (age category AUC=0.64, 95% CI=0.63, 0.65; linear age AUC=0.65, 95% CI=0.64, 0.67). In second IVF cycles, many variables from the failed first cycle were included as predictors in addition to most baseline variables. Multivariable models performed only slightly better than age-based models (AUCs=0.63), with AUCs ranging from 0.67 (main effects, 95% CI=0.65, 0.70) to 0.72 (boosted regression, 95% CI=0.68, 0.77). When we examined individual biomarkers of ovarian reserve, AMH, age and FSH had similar performance with AUCs ranging from 0.63 (95% CI=0.59, 0.67) to 0.67 (95% CI=0.64, 0.71); FSH had the poorest performance (AUC=0.55, 95% CI=0.51, 0.59). Only FSH had a significantly different AUC from AMH (difference=0.08, 95% CI=0.04, 0.13). No substantial differences in AMH performance were observed within subgroups.

Conclusion: Multivariable models performed only slightly better than simple age-based models or models based on other single biomarkers of ovarian reserve. There was very little improvement in accuracy with increasing model complexity, with small or no differences when using boosted regression compared to stepwise techniques. All models/individual predictors had only modest performance with AUCs below 0.72. The minimal improvements in model performance for multivariable models are likely not substantial enough to warrant widespread clinical application, which would necessitate software development for calculating individualized probabilities. Despite the modest performance overall, there may be subgroups of women in whom the predictors and chance of live birth differ. Future investigations should examine whether models developed within relevant subgroups, such as those based on age, race, and diagnosis, have better performance.

Table of Contents

List of Figures

Chapter 1. Prediction modeling for live birth in *in vitro* fertilization: personalized predictions in first and second cycles

Chapter 2. Prediction modeling for live birth in *in vitro* fertilization: anti-Müllerian hormone and other biomarkers of ovarian reserve

List of Tables

Chapter 1. Prediction modeling for live birth in *in vitro* fertilization: personalized predictions in first and second cycles

Chapter 2. Prediction modeling for live birth in *in vitro* fertilization: anti-Müllerian hormone and other biomarkers of ovarian reserve

Acknowledgements

We would like to thank Dr. Paul Lin and Laura Waibel from Seattle Reproductive Medicine for their assistance with this project.

Chapter 1. Prediction modeling for live birth in *in vitro* fertilization: personalized predictions in first and second cycles

INTRODUCTION

Infertility affects at least 6% of married or cohabitating couples [1], and the use of *in vitro* fertilization (IVF) as a treatment for infertility has become a mainstay since its first successful implementation over 40 years ago [2]. In 2011 approximately 152,000 IVF cycles were reported to the Society for Assisted Reproductive Technology (SART) in the United States [3], a 40% increase since 2001 [2]. Infants conceived using assisted reproductive technologies (including IVF) now account for approximately 1% of all births in the US [4]. Despite improvements in IVF techniques, the success of this procedure is still limited, with only 29% of cycles resulting in a live birth [3]. Further, IVF can be costly (average cost of $12,400 per cycle [5]) and time-consuming, and it also poses risks to the patient. Therefore identifying a patient's individual probability of having a successful cycle before its initiation could improve decision-making for both patients and physicians.

Currently, clinicians often use cumulative information from similarly aged patients to estimate a woman's probability of IVF treatment success, such as pregnancy or live birth [6]. In particular, U.S. clinicians use the publicly available SART data [3], which present probabilities of live birth stratified by predominantly age categories, to provide women with estimates of their chance of success. In other circumstances, physicians may use clinic-specific data to generate an individual center's algorithm to estimate the likelihood of live birth for any given patient, but in which age would be a key determinant. Although chronological age can strongly influence fertility, the age-related rate of decline in fertility may vary dramatically between women. Indeed, the success of IVF can also be predicted by additional markers of ovarian reserve, as well as demographic, anthropometric, and infertility characteristics [7,8], as has been investigated in previous prediction models [9].

Among the reported models that predict individualized probabilities of IVF success, most have included pregnancy as the measure of IVF success, whereas the ultimate goal of treatment for both patients and clinicians is a live birth. Even among those studies predicting live birth, many have included predictors ascertained after ovarian stimulation. Inclusion of these predictors precludes model application before treatment initiation, the time when personalized predicted probabilities are most useful to patients. Furthermore, very few studies have been based on U.S. data. Across Europe eligibility criteria for IVF vary and may include, depending on the country: age limits, medical indications, marital or cohabitation status, and sexual orientation; and all EU countries have either partial or full insurance coverage of IVF [10]. Therefore, findings from European studies may not be applicable to U.S. populations that are not subject to these eligibility criteria. The few U.S.-based studies have had limited sample sizes and often used data from a single clinic, which may also limit applicability to the more than 400 infertility clinics nationwide. Therefore we sought to develop a model to predict the probability of live birth in first IVF cycles using variables obtained prior to ovarian stimulation in a multi-center U.S. study population.

Almost all models developed using U.S. or non-U.S. data have predicted IVF's success probability (live birth or pregnancy) in either first cycles or in all cycles regardless of prior attempts. However, these models may not be applicable to women with a failed first cycle since patient characteristics and responses to treatment may differ between those with and without a live birth. Furthermore, because responses across cycles are likely correlated within a woman, predictors from the first failed cycle obtained *after* stimulation may be used to better predict live birth in second cycles compared to predictors obtained prior to stimulation in the first cycle. For these reasons, we also aimed to develop a model to predict the probability of live birth in second IVF cycles contingent on a single previous failed cycle using data obtained before stimulation in the second cycle.

METHODS

We conducted a retrospective cohort study to develop multivariable models predicting live birth in women (1) who had never previously undergone IVF, and (2) who had a single prior failed cycle. We split the dataset, using 80% of cycles for model development, and the remaining 20% for model validation. We used backwards stepwise logistic regression and, as a sensitivity analysis, boosted regression trees, to build the prediction models. Within cycle, we compared these models to (1) the expected probabilities based on age-stratified SART data, (2) a model with linear age only, and (3) to each other to estimate relative improvements in prediction.

Data Source

We used electronic medical records data from five private infertility clinics located in (1) Washington State; (2) California; (3) Florida; (4) South Carolina; and (5) Washington, D.C., Maryland, Pennsylvania, Virginia, which are part of a larger national network of fertility centers.

The clinics included in the national network share a common electronic medical records system, along with a mechanism for creating queries accessible across clinic sites. However, the clinics are networked only administratively; each clinic remains autonomous in its clinical practice. Thus physician management and clinic-specific protocols guide the course of treatment for each patient, and these likely varied across the five clinics. To address the aims of this project, we created a standardized query that was applied to the electronic medical records database in each clinic.

This project was approved by the Institutional Review Board at the University of Washington.

Study Subjects

For all analyses, female patients undergoing autologous fresh IVF cycles with or without intra-cytoplasmic sperm injection (ICSI) between January 2005 and August 2010 were eligible for inclusion; those using donor eggs or cryopreserved eggs or embryos were excluded. All analyses included women who received prior infertility services other than IVF or ICSI, such as intrauterine insemination (IUI). For first cycle analyses, eligible for inclusion were all female patients who had never previously undergone IVF treatment. We excluded women with a prior cycle at any clinic (ascertained by self-report or via medical records if prior cycles occurred at the same clinic). For second cycle analyses, female patients who had undergone a single prior cycle that did not result in a live birth, and who had a second cycle at the same clinic as the first were included.

Outcome

The primary outcome of interest was live birth (yes/no) in the cycle of interest, as reported by the patient. The outcome was obtained by the clinics via telephone follow-up with patients who self-reported the outcome. Because clinics are required to report to the Society for Assisted Reproductive Technology (SART), outcome data are nearly complete and are validated in compliance with SART guidelines [11]. Fewer than 2% of cycles have been reported to be discrepant between the SART and medical records data for cycle outcomes [2].

Predictors

First cycles. All potential predictors of live birth were determined *a priori* and selected from data recorded in the electronic medical record prior to ovarian stimulation. These variables included demographics (year of cycle initiation, age, race/ethnicity [White, Asian, Black, Hispanic, other]); anthropometrics (body mass index [BMI], height, weight); reproductive history (never pregnant/never live birth, ever pregnant/never live birth, ever pregnant/ever live birth); infertility diagnosis (diminished ovarian reserve [DOR], endometriosis, tubal factor, male infertility, ovulation disorders/polycystic ovaries [PCO], uterine factor, unexplained, other, or multiple factors); and the stimulation protocol to be used (antagonist, normal responder protocol, high responder protocol, and low responder protocol). For cycles using a GnRH agonist (Lupron) for pituitary down-regulation, we defined suspected normal responders as those using GnRH agonist administered in long luteal protocol with standard agonist dosing [12]; and with higher agonist dosing as suspected high responders. Because they are usually reserved for suspected poor responders, we categorized Lupron microdose, flare, and stop protocols as low responder protocols [13]. We categorized cycles with a GnRH antagonist separately (antagonist protocol). We did not include clinic and year as eligible predictors, as their inclusion would preclude application to clinics outside of the sample and to future cycles.

Due to changes in required reporting to SART, height and weight became standard fields in clinic databases in 2007 and, as a result, were almost entirely complete as of 2008. However, this information was missing in earlier years, particularly for the Washington State clinic. Therefore, we abstracted these data on-site from the electronic medical records at the Washington State clinic.

Second cycles. Variables included in the first cycle analyses were also eligible in second cycle analyses, with information updated from the second cycle (e.g., age reflected values at the time of the second, not the first, cycle). Variables obtained after stimulation in the prior failed cycle were also eligible as potential predictors. These included the total dose (IUs) of gonadotropins administered; number of days of stimulation; maximum estradiol level during stimulation; number of follicles >14 mm observed on ultrasound during stimulation; number of oocytes aspirated; number of germinal vesicle, metaphase I, and metaphase II oocytes; source of semen (partner/any donor); method of semen collection (ejaculation/other); number of oocytes fertilized via IVF and via ICSI; total number of oocytes fertilized; whether assisted hatching was performed in oocytes (no/at least some); number of embryos transferred; day of embryo transfer (<3, 3-4, 5, 6+, no transfer); whether preimplantation genetic diagnosis was performed (yes/no); human chorionic gonadotropin (hCG) level measured after embryo transfer; number of gestational sacs observed on ultrasound; fetal cardiac activity (FCA, number of distinct heartbeats identified); first cycle outcome (not pregnant; biochemical pregnancy; ectopic pregnancy; clinical intrauterine pregnancy resulting in spontaneous abortion, therapeutic abortion, or stillbirth); and whether any complications occurred in the cycle (yes/no)**.** These complications included anesthetic complications, hemorrhage, infection, side effects of medication, moderate hyperstimulation, severe hyperstimulation, complications from procedures, psychological stress, and other complications.

Multiple imputation. For first cycles, almost all predictors had <3.5% missing data, except for race (10.9%) and biochemical pregnancies (40.1%); however, exclusion of women with *any* missing data would have reduced the sample size by half. Therefore, we used multiple imputation by chained equations to generate many complete datasets in which missing data were replaced with plausible values for both first and second cycles simultaneously [14]. Although 3-5 imputations were formerly considered sufficient for imputation [15], more recent data suggests that additional imputations are required particularly when the fraction of missing information is large [16]. Therefore, we created 15 imputed datasets, in which any observations with missing values were replaced with plausible values.

Data analysis

Model development. Clinicians often estimate patients' probability of IVF success by applying age group-specific observed probabilities of live birth reported in SART. Therefore, before fitting more complicated prediction models, for comparison purposes we assigned each subject a predicted probability based on her age group and the age group-specific live birth percentages reported in SART. We call this the "age category model." Based on information reported to SART in 2011 for all autologous fresh cycles [3], we assigned live birth probabilities as follows: <35 years of age, 40.1%; 35-37 years, 31.8%; 38-40 years, 21.5%; 41-42 years, 12.2%; and 43+ years, assigned 4.2%. Furthermore, we fit a model with age (continuous) as the only predictor, the "linear age model;" we compared the performance of these age-based predictions to the more complex multivariable prediction models described below.

Before performing stepwise regression, we examined whether including higher-order (non-linear) age terms (age^2 and age^3) improved model fit over a linear term alone. Using the likelihood ratio (LR) test, we compared a model with age and age^2 to an age-only model, and a model with age, age^2, and age^3 to a model with age and age^2. If inclusion of the higher order age term resulted in a p-value<0.05 for the LR test, the higher order term was included in the list of potential predictors.

For the primary analyses, we employed backwards stepwise logistic regression to develop models predicting the probability of live birth versus no live birth. We initially included all potential predictors. After performing the regression with all variables, the predictor with the highest p-value was removed and the remaining model rerun. We repeated this process until all variables in the model were associated with live birth at a p-value of <0.2. We added the previously excluded variables back one at a time and included any that predicted live birth with a p-value of <0.2 to generate the final "main effects model".

To generate the "interaction model" we added interaction terms for some of the variables that could plausibly modify associations. For both first and second cycle analyses, we included an interaction between each predictor in the main effects model with age (continuous), weight, (continuous), BMI (continuous), reproductive history (never pregnant/never live birth, ever pregnant/never live birth, ever pregnant, ever live birth), and race (White, Asian, Black, and Hispanic; the other category was excluded because of limited sample size). We included interactions with age because at sufficiently advanced ages, other variables may not add any information to the predicted probability. Weight/BMI and pregnancy history were examined for interactions because despite other similar indicators of success, heavier women may have a lower probability of pregnancy or live birth compared to normal weight women [17], and women with a demonstrated capacity to become pregnant may have a greater probability than those without. Finally, we included interactions with race because known differences in reproductive potential across races [18] may affect the ability of other variables to predict IVF success. Interactions were added one at a time to each of the imputed datasets, and those with a Wald test p-value<0.05 in at least 11 of the 15 imputed datasets were further examined. Finally, all interactions meeting these criteria were entered together and retained in the final model only if the Wald test p-values were <0.05 in at least 11 datasets.

For purposes of comparison to work by others, we employed boosted regression trees [19], which attempt to improve the predictive performance of a single model by fitting and combining many models [20]. Regression trees, which belong to the classification and regression tree (CART) family of models, are built by creating binary splits at various predictor values chosen to minimize prediction errors [20]. Boosted regression trees use boosting, a machine learning algorithm, to build and combine trees, leading to a linear combination of many trees [20]. With boosting, the first regression tree is fit to the available data; thereafter, each regression tree is fit to the residuals of the previous tree, modeling the variation not explained by the previous model [20]. We allowed up to 10,000 trees and up to 6-way interactions to generate the “boosted regression model.” To minimize overfitting, we shrunk estimates by a factor of 0.005 and employed bagging, whereby only 60% of observations in the development sample were used to fit a new tree to the residuals of the last tree. We fit boosted regression trees to each imputed dataset.

Model evaluation. We performed internal model validation of the five models (age category, linear age, main effects, interaction, and boosted regression models) for both first and second cycles using the remaining 20% of data, which was independent of the model development sample. Evaluating model development and validation in the same data can result in overfitting, leading to overly optimistic estimates of model performance [21]. Splitting the data reduces the degree of overfitting.

We first assessed the discrimination of each model, the ability to distinguish women with a live birth from those without, using nonparametric estimates of the receiver operating characteristic (ROC) curve and area under the ROC curve (AUC). The ROC curve is a plot of the sensitivity versus 1-specificity across all possible thresholds of predicted probability. The AUC is the area under this curve, with a value of 0.50 indicating no discrimination and 1.00 for perfect discrimination. To determine which model had the best discrimination, we estimated the difference in AUCs between each of the models (within cycle number). To explore the potential generalizability, we examined the AUC within each clinic and year of initiation. We also estimated the AUC within age categories (<35, 35-37, 38-40, 41-42, 43+ years of age) for each model to determine if they have better or worse performance in specific age groups that are relevant to clinicians.

We calculated each model's sensitivity and specificity, which required defining a threshold of predicted probability above which subjects would be classified as having "high probability" of having a live birth (hereafter referred to as "live birth likely") and below which they would be classified as having "low probability" (referred to as "live birth unlikely"). Sensitivity measures the proportion of subjects with the outcome of interest (live birth) correctly classified as live birth likely, and specificity the proportion of subjects without the outcome that were classified correctly as live birth unlikely. There is very little research investigating how estimated probabilities of success affect patients' decision-making; however, decisions at a given predicted probability are likely influenced by many factors that vary among women, such as socioeconomic status and age. Therefore, we examined a number of different thresholds that may be relevant to defining live birth likely or unlikely, including 5%, 10%, 15%, 20%, and 25%.

For each threshold we also calculated each model's positive and negative predictive values (PPV and NPV, respectively), which are influenced by the prevalence of live birth in the population. The PPV reflects the proportion of participants who had a live birth among those classified as live birth likely; the NPV is the proportion who did not have a live birth among those identified as live birth unlikely. These measures of accuracy provide estimates of how often the results based on the predictive model reflect the underlying condition (i.e., live birth).

We calculated standard errors for all measures of model performance by using a bootstrap procedure with 300 sampled datasets in order to obtain 95% confidence intervals. For the multiply imputed datasets, model performance (and performance measures' bootstrapped standard errors) was calculated within each imputation, and Rubin's Rules were applied to pool estimates across datasets [22].

To assess model calibration, i.e., agreement between predicted and observed outcomes, we generated calibration plots in which the predicted probability was averaged within deciles of predicted probability, as was the observed probability within the decile. We plotted the predicted vs. observed probabilities; perfect prediction occurs when observed and predicted probabilities are equal (e.g., on the 45 degree line). Furthermore, to examine the extent to which predicted probabilities varied within a given age and provided more individualized estimates than age alone, we plotted the predicted probabilities from the linear age, main effects, and interactions models by age for first and second cycles.

For the boosted regression trees, we calculated the percent influence of each predictor (separately for each imputed dataset and then, for each predictor, averaged across the imputed datasets). Percent influence provides information about the relative importance of each predictor in the boosted regression tree [19]. Each split on a variable in a tree results in an increase in the log likelihood; the influence is the sum of the increase in log likelihood across all trees due to a particular variable, scaled to total 100 across all variables. The ROC curve and AUC, as well as the sensitivity, specificity, PPV, and NPV at the cutpoints, were calculated as for the other models.

The second cycle models had many more potential predictors available and were developed in a selected, specific group of women. To determine if addition of information from the first cycle improved prediction in this group, we compared the AUC of the second cycle main effects and interactions models to the first cycle models applied to the second cycles, as described above. We also examined whether there were systematic differences in the predictions generated from the first and second cycle models by comparing the means of these predicted probabilities within each imputation and averaging.

RESULTS

First cycles

Across clinics, there were 23,247 first cycles. We excluded 93 first cycles with missing information on outcome, source of semen, and method of semen collection, as collinearity in variables used for imputation precluded their inclusion, resulting in 23,154 first cycles for analysis. The majority of women were 35 years of age or older and white, with BMI in the normal range (Table 1). Half the women had never been pregnant. The most common infertility diagnoses were unexplained (19.8%), male infertility (19.7%), and multiple factors (19.4%); the most common stimulation protocol was for normal responders. At 34.6%, the percentage of cycles resulting in live birth was slightly higher than the national average of 29.3% [3], and 27.2% of women with a live birth had a multiple gestation.

Model development.

Age only models. When age alone was included in a model, for every 1-year increase in age, the odds of live birth decreased by 0.11 (OR=0.892, 95% CI=0.896, 0.898).

Backwards stepwise regression. Using the LR test, age^2 and age^3 both improved model fit at the $p<0.05$ level, and were included as eligible predictors. In the 18,555 first cycles reserved for model development, age, age^2, age^3, weight, height, stimulation protocol, race, diagnosis, and pregnancy history were predictors in the main effects model (Table 2); only BMI was eliminated.

When we added interaction terms to the main effects model, interactions between age (linear) and (1) weight, (2) race (with exclusion of “other” race due to small cell sizes), and (3) infertility diagnosis were retained (Table 3), meaning that the relationship between age and the predicted probability of live birth differed by these variables.

Boosted regression. In the boosted regression model, age had the largest influence (52.2%, Figure 1). The next largest contributor was the type of stimulation protocol (10.4%), BMI (10.2%), infertility diagnosis (8.5%), weight (6.9%), race (5.3%), height (5.1%), and pregnancy history (1.4%).

Model performance. In general, all first cycle models performed similarly with AUCs ranging from 0.64 (95% CI=0.63, 0.65) for the age category model to 0.67 (95% CI=0.66, 0.69) for all multivariable models (Table 4). Differences between the models were minimal, ranging from -0.003 to 0.04. This similarity in models also was reflected in the ROC curves (Figure 2).

Across clinics, the AUCs for the multivariable models varied from approximately 0.65 to 0.70, with the smallest AUC consistently observed in the clinic with the fewest cycles (Table 5). AUCs also varied by year of cycle initiation across multivariable models from 0.64 to 0.70; however there was no consistent trend in these differences. Within age categories, multivariable models had the best performance in the women of advanced age (≥43 years, Table 6), with AUCs ranging from 0.64 (95% CI=0.50, 0.78) to 0.69 (95% CI=0.56, 0.82). The smallest AUCs were in the women 41-42 years of age across multivariable models, with AUCs ranging from 0.51 (95% CI=0.42, 0.60) to 0.53 (95% CI=0.43, 0.62).

Across all thresholds and all models, sensitivity was high (≥0.82) and specificity was low (≤0.39), with moderate PPV (0.34-0.42) and high NPV (0.80-1.00, Table 7). The PPV and NPV are a function of the prevalence of live birth, which was 34.1% in the validation sample. With greater prevalence, the PPV would be higher and the NPV lower and vice versa with lower prevalence. At the lowest investigated threshold of 5% for defining live birth likely and unlikely, sensitivity was close to 1.00 for all models and specificity was approximately zero. At the 5% threshold, the PPV was 0.34 to 0.36 for all models, and the NPV was 0.93 for the main effects, 0.94 for the age category and interactions, and approached 1.00 for the linear age and boosted regression tree models. At the highest threshold defining live birth likely and unlikely of 25%, sensitivities were 0.82 for the age category model and 0.88 for all other models; specificities ranged from 0.30 to 0.39. Across models, PPVs ranged from 0.39 to 0.42 and NPV from 0.80 to 0.84. Although multivariable models performed better than the age category or the linear age model, no multivariable model emerged as superior in terms of AUC, sensitivity, specificity, PPV, and NPV.

In the calibration plot, the predicted and observed probabilities within deciles were not far off from perfect calibration (equal predicted and observed probabilities) (Figure 3). There were no obvious systematic differences within models (e.g., all observations too high, too low, or too extreme). When we examined the predicted probabilities from the main effects and interactions model by age, there was substantial variability within a given age, as compared to the linear age model with only a single predicted probability for each year of age (Figure 4).

Second cycles

Of the 15,133 women who did not have a live birth in their first IVF cycle, 8,283 elected to proceed with a second cycle at the same clinic (54.7%). Because of limited sample sizes at the extremes of age, we eliminated cycles in which the woman was less than 25 years or greater than 44 years of age (n=96); we also excluded 3 cycles for which the day of embryo transfer was missing. The remaining 8,184 were similar to the women in the first cycle analysis, except they were older, more likely to have been pregnant, more likely to have previously had spontaneous abortions and biochemical pregnancies, and were less likely to have normal or high response stimulation protocols (Table 1). The live birth rate was 29.0% in second cycles, slightly lower than first cycles.

Model development.

Age only model. When age alone was included in a model, for every 1-year increase in age, the odds of live birth decreased by 0.11 (OR=0.894, 95% CI=0.882, 0.906), as in the first cycles.

Backwards stepwise regression. Using the LR test, age^2 and age^3 improved model fit and were included as eligible predictors in second cycle analyses. In the 6,534 first cycles reserved for model development, age, age^2, age^3, BMI, weight, stimulation protocol, race, and pregnancy history were selected from the second cycle predictors (Table 8). From the failed first cycle, total amount of gonadotropins, maximum estradiol, number of metaphase II (mature) oocytes retrieved, whether assisted hatching was performed, the number of embryos transferred, whether PGD was performed, the day of embryo transfer, and whether any complications occurred in the cycle were included as predictors.

When we added interaction terms to the second cycle main effects model, interactions between (1) maximum estradiol (linear) and race (with exclusion of “other” race due to small cell sizes), and (2) BMI and day of embryo transfer (collapsing the <3 days and 4 days categories) were retained (Table 9).

Boosted regression. Age had the largest relative influence (21.9%) in the boosted regression model, although it was smaller than in first cycles (Figure 5). The next largest contributors were maximum estradiol level (13.0%), total dose of gonadotropins administered (10.4%), number of metaphase II oocytes (7.6%), number of oocytes aspirated (6.7%), the number of follicles>14mm observed on ultrasound (5.7%), and BMI (5.5 %).

Model performance. Once again, the age category and linear age models had the lowest AUCs at 0.63 (95% CI=0.60, 0.66 and 0.61, 0.66, respectively, Table 4). The main effects and interaction models performed slightly better with AUCs of 0.68 and 0.67 (both 95% CI=0.65, 0.70), respectively, with differences from the age category and linear age models ranging from 0.04 to 0.05. The boosted regression tree model had the largest AUC at 0.72 (95% CI=0.68, 0.77), which was an improvement of 0.09 over the age category and linear age models, and 0.05 compared to the main effects and interactions models (Figure 6).

Across clinics, the AUCs were fairly similar for the multivariable models except for in the smallest clinic, where the AUCs were much lower between 0.34 and 0.47 (Table 5). In this clinic, increasing predicted probability of live birth was associated with decreased odds of live birth,resulting in AUCs below 0.50. AUCs also varied across multivariable models by year of cycle initiation, but without any consistent pattern. For all multivariable models, the AUCs were largest in women ≥43 years of age and smallest in those <35 years of age (Table 6). The improved AUC among the oldest age group was particularly pronounced in the boosted regression model, with an AUC of 0.76 (95% CI=0.53, 1.00).

Similar to first cycle analyses, sensitivity was high (≥0.74) and specificity was low (≤0.49), with low to moderate PPV (0.28 to 0.38) and high NPV (0.81 to apporaching 1.00, Table 7) for all models at a live birth rate of 27.9%. At the 5% threshold for defining live birth likely and unlikely, sensitivity was close to 1.00 for all models and specificity was near zero; the PPV was 0.28 to 0.30 and the NPV approached 1.00. At the 25% threshold, sensitivities ranged from 0.74 to 0.81 and specificities from 0.46 to 0.49; PPV ranged from 0.35 to 0.38 and NPV from 0.81 to 0.87. At this threshold, the boosted regression tree model had the best performance on all measures, although these improvements were quite small.

In the calibration plot, the predicted and observed probabilities within deciles were similar. The predictions were not systematically different from the observed probabilities, although the greatest disparities between observed and predicted probabilities occurred in the largest decile for all models (Figure 7). There was substantial variability in the predicted probabilities for the main effects and interactions models within age, similar to first cycle results (Figure 4).

When we applied the first cycle models to second cycles, predicted probabilities were larger than the second cycle predictions by about 0.0047 (0.47%) for the main effects and 0.00049 for the interactions models (data not shown). The AUC of the first cycle main effects model applied to second cycles was 0.05 (95% CI=-0.05, -0.009) smaller than the second cycle main effects model, and 0.02 (95% CI=-0.034 -0.001) smaller for the interactions models.

DISCUSSION

In this study we investigated a number of models for predicting live birth. These models could assist patients and clinicians in decision-making about initiating IVF and repeating IVF after a prior failed cycle (second cycles). Despite the increase in use of IVF in the U.S., there are very few reports of models to predict live birth or ongoing pregnancy in U.S. populations among women undergoing IVF. Prediction models specific to U.S. populations using factors known prior to cycle initiation would be very useful because IVF is expensive [5], is rarely covered by insurance [23], and can have emotional [24] and health sequelae [25, 26]. Clinicians currently use population-based age-stratified estimates of live birth from SART data; however age alone may not provide the most accurate prediction of success. We found that multivariable models, produced through stepwise regression with or without interactions or boosted regression trees, performed slightly better than models using age categories or linear age. We observed that the prediction models performed best in the oldest group of women (≥43 years of age) who have the lowest probability of live birth (<5%), which suggests that the models may prove to be particularly helpful in this group.

Only three publications to date report on models to predict live birth, the ultimate outcome of interest, using data from U.S. populations [27-29]. Among these studies, two included both pre- and post-stimulation variables [27-28] to predict in first cycles; therefore, they are of little use to patients in deciding whether or not to proceed with treatment. The only U.S. study of pre-stimulation factors predicting live birth in first IVF cycles included data from Spain, Canada, and a single Boston, Massachusetts clinic and employed a boosted regression tree approach to model development [29]. The study included ~7,600 U.S. cycles, whereas we included data from five U.S. clinics, leading to the largest U.S. sample by far of almost 19,000 cycles for model development. Choi et al reported that age of the patient was the most influential predictor (60.1%), followed by sperm count (9.6%), BMI (9.5%), day 3 serum follicle stimulating hormone (FSH, 5.0%), and antral follicle count (AFC, 4.5%), resulting in an AUC of

0.64 in the validation sample, slightly better than a boosted regression tree model with age alone (AUC=0.61). [29]. We obtained a similar but slightly larger AUC of 0.67 in the boosted regression tree model, in which age was the most influential predictor at 52.2%. However, this boosted regression tree model had no improvement in model performance over the various stepwise models, despite increased model flexibility. Although boosted regression trees initially held promise for improving predictions, more recent investigations suggest that logistic regression may perform similarly to boosting [30].

Unlike Choi et al, we did not have complete information on sperm count, or the biomarkers of ovarian reserve, including FSH and AFC in the dataset, which may be important predictors of IVF success. These predictors are typically available to clinicians and have been associated with IVF outcomes [31-33]. However, we included stimulation protocol, which the physician determines *a priori* based on a number of patient characteristics including age, ovarian reserve markers, and infertility diagnosis, which ultimately may have served as a more comprehensive proxy for FSH and AFC alone. Indeed, stimulation protocol had the second highest influence in this sample (10.4%). Furthermore, a diagnosis of male infertility would likely reflect sperm parameters, such as sperm count, and the infertility diagnosis had the fourth highest influence (8.5%). We also did not have other potentially important predictors of live birth, such as smoking status [34] or duration of infertility [35]. Despite this missing information, because an extremely strong association between a predictor(s) and the outcome of interest is needed to improve classification, it is unlikely that inclusion of any of these missing predictors would lead to substantially better model performance [36]. Even associations that would typically be considered strong, such as an odds ratio of 2 or 3, do not lead to improvements in classification when added to existing prediction models [36,37]; none of the excluded variables have associations sufficiently strong to enhance classification.

We also generated models for live birth prediction in a second cycle, as women with a prior failed cycle are a more specific group with possibly different treatment responses than all women initiating IVF for the first time. In addition, information on the response to treatment in the prior cycle can presumably be leveraged to predict the probability of live birth in second cycles because of correlations within a woman. Therefore the opportunity exists to generate a more accurate prediction model within this subgroup. Indeed, we observed that when the first cycle models were applied to women with a second cycle after a prior failed cycle, the predictions were larger and the AUCs were smaller than the second cycle-specific models, although these differences were small.

There has been only one report on a model predicting live birth specifically in second IVF cycles, using data from Stanford Hospital and Clinics [28]. Banerjee et al generated a boosted regression tree model predicting live birth in first IVF cycles using predictors obtained following stimulation. They subsequently adjusted this model for second cycle data to account for the typically worse treatment response in second cycles, without which predicted probabilities would be overestimated. The most influential factors were rate of blastocyst development (26%), the total amount of gonadotropins administered (10%), the number of eight-cell embryos available (9%), embryo cryopreservation (7%), age (6%), endometrial thickness (6%), and total number of embryos obtained (6%). Only age (21.9%) and the total amount of gonadotropins administered (10.4%) overlapped with the most influential predictors we observed in the boosted regression model. Other important contributors were maximum estradiol level (13.0%), number of metaphase II oocytes (7.6%), number of oocytes aspirated (6.7%), the number of follicles>14mm observed on ultrasound (5.7%), and BMI (5.5%). With the exception of BMI, these factors all reflect the response to stimulation with larger values indicating a better response, which at least partially contributes to an increased chance of live birth.

We did not have information on embryo development, such as the rate of blastocytst development or the number of eight-cell embryos, nor did we have data on endometrial thickness as Banerjee et al did. However, oocyte characteristics such as the number of metaphase II oocytes, number of oocytes aspirated, and the number of follicles>14mm during stimulation were all highly influential, and oocyte characteristics may be related to the embryo characteristics [38], and thus serve as proxies for these variables in the data. Despite this lack of information, in the validation sample of 230 second cycles the AUC of the Banerjee model was slightly lower than the current boosted regression tree model (0.68 vs. 0.72), with both models performing better than those with age alone. Similar to first cycles, the boosted regression tree model led to only slight improvements in AUC over the other multivariable models that had AUCs of ~0.68.

We generated the second cycle model only in those women with an additional cycle after the first cycle failed, thus the sample was more specific and selected. In contrast, the Banerjee et al model was generated initially in first cycles, which included women who did and did not proceed with a second IVF cycle. However, their adjustment for differences between first and second cycle predictions likely would have mitigated these effects, as adjustment was based on those with a second cycle. Both models therefore may be biased in that they exclude information from women who opted out of treatment after a failed cycle, a group that may differ from those proceeding with treatment. In the current sample, first cycle treatment responses were similar between women who did and did not proceed with treatment, as were the probabilities of pregnancy estimated from the various first cycle models. Therefore use of this more selected sample likely did not bias the second cycle models appreciably.

The multivariable models we generated had only moderate performance, all with AUCs of 0.67 to 0.72. For all models across varying thresholds defining live birth likely and unlikely (5% to 25%), sensitivities were high and specificities were low; most women who had a live birth were classified as live birth likely but few women without a live birth were classified as live birth unlikely. Even at the largest investigated threshold of 25% predicted probability, the PPV was only 0.35 to 0.40 across first and second cycle models, and at the lowest threshold of 5% the PPV was 0.28 to 0.34. However, in the infertility setting it may be more grievous to inappropriately deter women from treatment (a false negative result) as compared to proceeding with treatment and having an unsuccessful cycle (a false positive result). Therefore, high sensitivity is desirable, even at the expense of low specificity.

Because there is no universal probability at which all women would elect to proceed with treatment, we investigated a range of plausible values. With a national live birth rate of ~30%, many women must find this predicted probability acceptable to proceed with treatment, supporting the largest threshold of 25%. Women >42 years of age have a probability of live birth of <5%, yet almost 6,000 women each year undergo IVF at these ages, indicating a willingness to proceed with treatment despite very limited success [3]. Furthermore, Dutch women reported a willingness to pay €1000 (1350 U.S. dollars) out-of-pocket when the predicted live birth rate was 6% or more [39]. Therefore, even very low predicted probabilities may have sufficient value to warrant paying for treatment traditionally covered by insurance. In the U.S., where insurance coverage for infertility is uncommon, acceptable probabilities of live birth may vary based on socioeconomic status, among other factors. Indeed, income appears to be the most important predictor of pursuing infertility treatment among women who had an infertility evaluation [40].

Despite the suboptimal performance of these models, personalized predictions may still be helpful in making treatment decisions. Almost 1/3 of women retrospectively reported being dissatisfied or only somewhat satisfied with the information provided on their chance of a live birth in IVF treatment, which was particularly true for women without a live birth (49%) [41]. Personalized predictions from one of the models reported here or in other publications may provide the additional information patients desire.

However, a remaining question is whether use of more complex models leads to sufficiently improved performance to encourage widespread deployment in the clinical setting. For age-based models, like the linear age and age category models, simple normograms can be utilized to provide patients with individualized probabilities of live birth. However, for more complex multivariable models, software is required to estimate probabilities. With such small differences between models in sensitivities and specificities across all cutoffs examined herein, improvements in the number of women appropriately classified would be small, even across all U.S. cycles within a year.

Although these models had similar performance when evaluated in the entire sample, we did observe improved accuracy in subgroups, particularly among older women. We also observed interactions between age and a number of baseline characteristics (race, diagnosis, and BMI), suggesting that the relationships between some predictors and live birth vary within subgroups. Therefore, future investigations should consider developing models *within* particular subgroups of women in whom predictors could vary. These tailored prediction models may prove to perform better than models applicable to all women, and could be targeted to groups for whom the clinical utility would be high, such as those with the lowest probabilities or the most variability in outcomes.

REFERENCES

1. Chandra A, Copen CE, Stephen EH. Infertility and impaired fecundity in the United States, 1982–2010: Data from the National Survey of Family Growth. *Natl Health Stat Report*. 2013;67:1-18.
2. Centers for Disease Control and Prevention, American Society for Reproductive Medicine, Society for Assisted Reproductive Technology. *2010 Assisted Reproductive Technology National Summary Report*. Atlanta: U.S. Department of Health and Human Services; 2012.
3. Centers for Disease Control and Prevention, American Society for Reproductive Medicine, Society for Assisted Reproductive Technology. *2011 Assisted Reproductive Technology Fertility Clinic Success Rates Report*. Atlanta: U.S. Department of Health and Human Services; 2013.
4. Sunderam S, Chang J, Flowers L, et al. Assisted Reproductive Technology Surveillance-United States, 2006. *MMWR Surveillance Summary.* 2009;58:1-25.
5. ASRM Frequently Asked Questions About Infertility, Question 6. Is In Vitro Fertilization Expensive? http://www.asrm.org/detail.aspx?id=3023. Accessed October 8, 2013.
6. Jones CA, Christensen AL, Salihu H, et al. Prediction of individual probabilities of livebirth and multiple birth events following in vitro fertilization (IVF): a new outcomes counseling tool for IVF providers and patients using HFEA metrics. *J Exp Clin Assist Reprod*. 2011;8:3. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3183499/
7. van Loendersloot LL, van Wely M, Limpens J, Bossuyt PM, Repping S, van der Veen F. Predictive factors in in vitro fertilization (IVF): a systematic review and meta-analysis. *Hum Reprod Update*. 2010;16:577-89.
8. Rittenberg V, Seshadri S, Sunkara SK, Sobaleva S, Oteng-Ntim E, El-Toukhy T. Effect of body mass index on IVF treatment outcome: an updated systematic review and meta-analysis. *Reprod Biomed Online.* 2011;23:421-439.
9. Leushuis E, van der Steeg JW, Steures P, et al. Prediction models in reproductive medicine: a critical appraisal. *Hum Reprod.* 2009;15:537-552.
10. Brigham KB, Cadier B, Chevreul K. The diversity of regulation and public financing of IVF in Europe and its impact on utilization. *Hum Reprod.* 2013;28:666-675.
11. Adashi EY, Wyden R. Public reporting of clinical outcomes of assisted reproductive technology programs: implications for other medical and surgical procedures. *JAMA.* 2011;306:1135-1136.
12. Johnston-MacAnanny EB, DiLuigi AJ, Engmann LL, Maier DB, Benadiva CA, Nulsen JC. Selection of first in vitro fertilization cycle stimulation protocol for good prognosis patients: gonadotropin releasing hormone antagonist versus agonist protocols. *J Reprod Med*. 2011;56:12-6.
13. Karande V, Gleicher N. A rational approach to the management of low responders in in-vitro fertilization. *Hum Reprod.*1999;14:1744-1748.
14. Royston P. Multiple imputation of missing values. *Stata J.* 2004;4:227–241.
15. Schafer JL, Olsen MK. Multiple imputation for multivariate missing data problems: A data analyst’s perspective. *Multivariate Behav Res.* 1998;33:545–571.
16. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prec Sci.* 2007;8:206-213.
17. Bellver J, Ayllón Y, Ferrando M, et al. Female obesity impairs in vitro fertilization outcome without affecting embryo quality. *Fertil Steril.* 2010;93:447-454.
18. Butts SF, Seifer DB. Racial and ethnic differences in reproductive potential across the life cycle. *Fertil Steril.* 2010;93:681-690.
19. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat.* 2001;29:1189-1232.

20. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. *J Anim Ecol.* 2008;77:802-813.
21. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* New York: Springer; 2009.
22. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley & Sons; 1987.
23. Resolve: The National Infertility Association. Insurance Coverage in Your State. http://www.resolve.org/family-building-options/insurance_coverage/state-coverage.html. Accessed November 27, 2013.
24. Verhaak CM, Smeenk JM, Evers AW, Kremer JA, Kraaimaat FW, Braat DD. Women's emotional adjustment to IVF: a systematic review of 25 years of research. *Hum Reprod Update*. 2007;13:27-36.
25. Delvigne A, Rozenberg S. Epidemiology and prevention of ovarian hyperstimulation syndrome (OHSS): a review. *Hum Reprod Update.* 2002;8:559-577.
26. Reddy UM, Wapner RJ, Rebar RW, Tasca RJ. Infertility, assisted reproductive technology, and adverse pregnancy outcomes: executive summary of a National Institute of Child Health and Human Development workshop. *Obstet Gynecol.* 2007;109:967-77.
27. Minaretzis D, Harris D, Alper MM, Mortola JF, Berger MJ, Power D. Multivariate analysis of factors predictive of successful live births in in vitro fertilization (IVF) suggests strategies to improve IVF outcome. *J Assist Reprod Genet.* 1998;15:365-371.
28. Banerjee P, Choi B, Shahine LK, et al. Deep phenotyping to predict live birth outcomes in in vitro fertilization. *Proc Natl Acad Sci USA.* 2010;107:13570-13575.
29. 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;99:1905-1911.
30. Austin PC, Lee DS, Steyerberg EW, Tu JV. Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods? *Biomed J.* 2012;54:657-673.
31. Chan SY, Wang C, Chan ST, et al. Predictive value of sperm morphology and movement characteristics in the outcome of in vitro fertilization of human oocytes. *J In Vitro Fert Embryo Transf*. 1989;6:142-148.
32. Creus M, Peñarrubia J, Fábregues F, et al. Day 3 serum inhibin B and FSH and age as predictors of assisted reproduction treatment outcome. *Hum Reprod*. 2000;15:2341-2346.
33. Holte J, Brodin T, Berglund L, Hadziosmanovic N, Olovsson M, Bergh T. Antral follicle counts are strongly associated with live-birth rates after assisted reproduction, with superior treatment outcome in women with polycystic ovaries. *Fertil Steril.* 2011;96:594-599.
34. Klonoff-Cohen H, Nataraian L, Marrs R, Yee B. Effects of female and male smoking on success rates of IVF and gamete intra-Fallopian transfer. *Hum Reprod*. 2001;16:1382-1390.
35. van Loendersloot LL, van Wely M, Limpens J, Bossuyt PM, Repping S, van der Veen F. Predictive factors in in vitro fertilization (IVF): a systematic review and meta-analysis. *Hum Reprod Update.* 2010;16:577-589.
36. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004;159:882-890.
37. Kattan MW. Judging new markers by their ability to improve predictive accuracy. *J Natl Cancer Inst.* 2003;95:634–635.

38. Loutradis D, Drakakis P, Kallianidis K, Milingos S, Dendrinos S, Michalas S. Oocyte morphology correlates with embryo quality and pregnancy rate after intracytoplasmic sperm injection. *Fertil Steril.* 1999;72:240-244.
39. Musters AM, de Bekker-Grob EW, Mochtar MH, van der Veen F, van Mello NM. Women's perspectives regarding subcutaneous injections, costs, and live birth rates in IVF. *Hum Reprod.* 2011;26:2425-2431.
40. Kessler LM, Craig BM, Plosker SM, Reed DR, Quinn GP. Infertility evaluation and treatment among women in the United States. *Fertil Stereil.* 2013;100:1025-1032.
41. Hammarberg K, Astbury J, Baker HWG. Women's experience of IVF: a follow-up study. *Hum Reprod.* 2001;16:374-383.

TABLES

Table 1.1. Characteristics of women undergoing their first *in vitro* fertilization cycle and women undergoing a second cycle after a prior failed cycle, 2005-2011

		First cycles		*Second cycles after prior failed cycle*	
		N	**%**	**N**	**%**
Total		23154	100.0	8184	100.0
Year	*2005*	3011	13.0	752	9.2
	2006	2953	12.8	1118	13.7
	2007	3278	14.2	1267	15.5
	2008	3434	14.8	1365	16.7
	2009	3685	15.9	1343	16.4
	2010	3806	16.4	1301	15.9
	2011	2987	12.9	1038	12.7
Age (years)	*<25*	229	1.0	0	0.0
	25-29	2722	11.8	586	7.2
	30-34	7399	32.0	2250	27.5
	35-39	8799	38.0	3499	42.8
	40+	4005	17.3	1849	22.6
Race/Ethnicity	*Asian*	3368	14.5	1228	15.0
	Black	2573	11.1	900	11.0
	Hispanic	1383	6.0	469	5.7
	Other	130	0.6	41	0.5
	White	13186	56.9	4869	59.5
	missing	2514	10.9	677	8.3
Body mass index (kg/m^2)	*<18.5 (underweight)*	664	2.9	224	2.7
	18.5-<25 (normal)	12797	55.3	4364	53.3
	25-<30 (overweight)	5232	22.6	1779	21.7
	30+ (obese)	3695	16.0	1342	16.4
	missing	766	3.3	475	5.8
Height (feet)	*0 to <5*	449	1.9	147	1.8
	5 to <5.5	13524	58.4	4614	56.4
	5.5 to <6	8303	35.9	2907	35.5
	6+	238	1.0	88	1.1
	missing	640	2.8	428	5.2
Weight (lbs)	*<100*	207	0.9	67	0.8
	100 to <150	13005	56.2	4457	54.5
	150 to <200	7159	30.9	2432	29.7
	200 to <250	1838	7.9	676	8.3
	250+	227	1.0	90	1.1
	missing	718	3.1	462	5.6

Table 1.1, cont.

		First cycles		*Second cycles after prior failed cycle*	
		N	**%**	**N**	**%**
Gravidity	*0*	11611	50.1	3569	43.6
	1	5845	25.2	2331	28.5
	2	2914	12.6	1199	14.7
	3+	2783	12.0	1084	13.2
	missing	1	0.0	1	0.0
Term births	*0*	18553	80.1	6659	81.4
	1	3510	15.2	1190	14.5
	2+	1089	4.7	334	4.1
	missing	2	0.0	1	0.0
Premature births	*0*	22484	97.1	7915	96.7
	1	561	2.4	237	2.9
	2+	105	0.5	31	0.4
	missing	4	0.0	1	0.0
Spontaneous abortions	*0*	17703	76.5	5569	68.0
	1	3642	15.7	1814	22.2
	2+	1804	7.8	799	9.8
	missing	5	0.0	2	0.0
Biochemical pregnancies	*0*	12842	55.5	1767	21.6
	1	867	3.7	4889	59.7
	2+	161	0.7	177	2.2
	missing	9284	40.1	1351	16.5
Diagnosis	*Diminished ovarian reserve*	1961	8.5	844	10.3
	Endometriosis	1120	4.8	408	5.0
	Tubal factor	2256	9.7	730	8.9
	Male infertility	4552	19.7	1585	19.4
	Other	1549	6.7	514	6.3
	Ovulation disorders/polycystic ovaries	1722	7.4	476	5.8
	Unexplained	4594	19.8	1652	20.2
	Uterine factor	317	1.4	142	1.7
	Multiple factors	4482	19.4	1722	21.0
	missing	601	2.6	111	1.4
Stimulation protocol	*GnRH Antagonist*	6409	27.7	3468	42.4
	GnRH Agonist				
	Normal response	10556	45.6	2074	25.3
	High response	2082	9.0	334	4.1
	Low response	3953	17.1	2287	27.9
	missing	154	0.7	21	0.3

Table 1.1, cont.

		First cycles		*Second cycles after prior failed cycle*	
		N	**%**	**N**	**%**
Cycle canceled	*No*	20260	87.5	6316	77.2
	Yes	2894	12.5	1868	22.8
Outcome	*Biochemical*	9068	39.2	3714	45.4
	Clinical intrauterine pregnancy+ live birth	8021	34.6	2374	29.0
	Clinical intrauterine pregnancy+spontaneous abortion/stillbirth/therapeutic abortion	1700	7.3	581	7.1
	Ectopic	178	0.8	66	0.8
	Not pregnant	4187	18.1	1449	17.7
Number live born	*0*	15133	65.4	5810	71.0
	1	5842	25.2	1743	21.3
	2	2109	9.1	611	7.5
	3	70	0.3	20	0.2

Table 1.2. Parameter estimates for a model predicting live birth in first *in vitro* fertilization cycles, main effects only, 2005-2011[a] (n=18555)

		95% CI		
	Odds Ratio	**LB**	**UB**	**p-value**
Age (years)	0.19	0.09	0.42	<0.001
Age^2 (years^2)	1.06	1.03	1.08	<0.001
Age^3 (years^3)	1.00	1.00	1.00	<0.001
Weight (per 10 lbs)	0.98	0.97	0.99	<0.001
Height (per 6 inches)	1.17	1.09	1.26	<0.001
Stimulation protocol				
Antagonist	0.74	0.69	0.80	<0.001
Normal response	1.00	reference		-
High response	1.33	1.19	1.49	<0.001
Low response	0.63	0.57	0.70	<0.001
Race				
Asian	0.74	0.67	0.82	<0.001
Black	0.72	0.64	0.80	<0.001
Hispanic/Latino	0.77	0.67	0.88	<0.001
Other	0.68	0.44	1.05	0.08
White	1.00	reference		-
Diagnosis				
Diminished ovarian reserve	0.64	0.54	0.76	<0.001
Endometriosis	0.95	0.81	1.11	0.52
Tubal factor	0.96	0.85	1.09	0.56
Male factor	1.00	reference		-
Other	0.89	0.77	1.03	0.11
Ovulation disorders/polycystic ovaries	1.28	1.13	1.46	<0.001
Unexplained	1.07	0.97	1.18	0.21
Uterine factor	0.76	0.57	1.02	0.07
Multiple factors	0.82	0.74	0.91	<0.001
Pregnancy history				
Never pregnant/never live birth	1.00	reference		-
Ever pregnant/never live birth	0.93	0.86	1.00	0.47
Ever pregnant/ever live birth	1.11	1.02	1.21	0.02
Constant	4.15E+06	784.40	2.30E+10	<0.001

CI-confidence interval, LB-lower bound, UB-upper bound

[a] Model built using backwards stepwise logistic regression with p-value≥0.1 for exclusion

Table 1.3. Parameter estimates for a model predicting live birth in first *in vitro* fertilization cycles, main effects and interactions, 2005-2011[a] (n=18431)

		95% CI		
	Odds Ratio	**LB**	**UB**	**p-value**
Age (years)	0.15	0.06	0.33	<0.001
Age^2 (years^2)	1.06	1.04	1.09	<0.001
Age^3 (years^3)	1.00	1.00	1.00	<0.001
Weight (per 10 lbs)	0.87	0.80	0.94	<0.001
Height (per 6 inches)	1.18	1.10	1.28	<0.001
Stimulation protocol				
Antagonist	0.74	0.69	0.80	<0.001
Normal response	1.00	reference		-
High response	1.33	1.19	1.47	<0.001
Low response	0.63	0.57	0.70	<0.001
Race				
Asian	0.18	0.08	0.41	<0.001
Black	0.29	0.12	0.73	0.01
Hispanic/Latino	0.41	0.14	1.23	0.12
White	1.00	reference		-
Diagnosis				
Diminished ovarian reserve	0.21	0.04	1.13	0.07
Endometriosis	0.85	0.21	3.37	0.82
Tubal factor	2.10	0.73	6.03	0.17
Male factor	1.00	reference		-
Other	0.15	0.05	0.47	0.001
Ovulation disorders/polycystic ovaries	0.50	0.17	1.41	0.20
Unexplained	1.20	0.51	2.86	0.67
Uterine factor	0.05	0.003	0.83	0.04
Multiple factors	0.99	0.44	2.24	0.99
Pregnancy history				
Never pregnant/never live birth	1.00	reference		-
Ever pregnant/never live birth	0.92	0.85	0.99	0.04
Ever pregnant/ever live birth	1.12	1.03	1.22	0.01
Age*Weight	1.003	1.001	1.006	0.005
Age*Race				0.004
Age*Asian	1.04	1.02	1.07	<0.001
Age*Black	1.03	1.00	1.05	0.05
Age*Hispanic	1.02	0.99	1.05	0.27
Age*White	1.00	reference		-

Table 1.3, cont.

		95% CI		
	Odds Ratio	**LB**	**UB**	**p-value**
Age*Diagnosis				0.002
Age*Diminished ovarian reserve	1.03	0.99	1.08	0.19
Age*Endometriosis	1.00	0.96	1.05	0.88
Age*Tubal factor	0.98	0.95	1.01	0.15
Age*Male factor	1.00	reference		-
Age*Other	1.05	1.02	1.09	0.002
Age*Ovulation disorders/polycystic ovaries	1.03	1.00	1.06	0.08
Age*Unexplained	1.00	0.97	1.02	0.80
Age*Uterine factor	1.08	1.00	1.17	0.06
Age*Multiple factors	0.99	0.97	1.02	0.66
Constant	2.84E+08	3.46E+04	2.33E+12	<0.001

CI-confidence interval, LB-lower bound, UB-upper bound

[a] Model built using backwards stepwise logistic regression with p-values≥0.1 for exclusion; interaction terms included when p-values<0.05

Table 1.4. Area under the receiver operating characteristic curve (AUC) and differences between AUCs in models predicting live birth in first *in vitro* fertilization cycles and second cycles after a prior failed cycle, 2005-2011

	First cycles				*Second cycles after a prior failed cycle*			
			95% CI				*95% CI*	
Model	**N**	**AUC**	**LB**	**UB**	**N**	**AUC**	**LB**	**UB**
Age category[a]	4599	0.64	0.63	0.65	1650	0.63	0.60	0.66
Linear age[b]	4599	0.65	0.64	0.67	1650	0.63	0.61	0.66
Main effects[c]	4599	0.67	0.66	0.69	1650	0.68	0.65	0.70
Interactions[d]	4572	0.67	0.66	0.69	1644	0.67	0.65	0.70
Boosted regression[e]	4572	0.67	0.66	0.69	1644	0.72	0.68	0.77
Differences								
Linear age vs. age category	4599	0.01	0.009	0.02	1650	0.004	-0.004	0.01
Main effects vs. age category	4599	0.04	0.02	0.05	1650	0.05	0.02	0.07
Interactions vs. age category	4572	0.03	0.02	0.05	1644	0.04	0.02	0.07
Boosted vs. age category	4572	0.03	0.02	0.04	1644	0.09	0.05	0.14
Main effects vs. linear age	4599	0.02	0.01	0.03	1650	0.04	0.02	0.06
Interactions vs. linear age	4572	0.02	0.01	0.03	1644	0.04	0.02	0.06
Boosted vs. linear age	4572	0.02	0.01	0.03	1644	0.09	0.05	0.13
Interactions vs. main effects	4572	-0.001	-0.004	0.002	1644	-0.003	-0.009	0.003
Boosted vs. main effects	4572	-0.003	-0.008	0.001	1644	0.05	0.009	0.09
Boosted vs. interactions	4572	-0.002	-0.007	0.003	1644	0.05	0.01	0.09

CI-confidence interval, AUC-area under the receiver operating characteristic curve, LB-lower bound, UB, upper bound

[a] Assigned probabilities of live birth based on age as follows: <35 years: 40.1%, 35-37 years: 31.8%, 38-40 years: 21.5%, 41-42 years: 12.2%, ≥43 years: 4.2%

[b] Linear term for age

[c] Backwards stepwise logistic regression using a p-value≥0.1 for exclusion, with main effects only (no interactions)

Table 1.4, cont.

[d] Backwards stepwise logistic regression using a p-value$\geq$0.1 for exclusion, with main effects and interactions between variables

[e] Boosted regression tree with up to 10000 regression trees, 6-way interactions, shrunk by a factor of 0.005, and bagging of 60% of data

Table 1.5. Area under the receiver operating characteristic curve of models predicting live birth in first in vitro fertilization cycle and second cycles after a prior failed cycle by clinic and year of cycle initiation, 2005-2011

			Linear Age[a]			Main Effects[b]			Interactions[c]			Boosted Regression[d]		
				95% CI			*95% CI*			*95% CI*			*95% CI*	
Clinic		**N**	**AUC**	**LB**	**UB**	**AUC**	**LB**	**UB**	**AUC**	**LB**	**UB**	**AUC**	**LB**	**UB**
First cycles	1	641	0.66	0.62	0.70	0.68	0.64	0.73	0.68	0.64	0.73	0.68	0.64	0.72
	2	2757	0.65	0.63	0.67	0.67	0.64	0.69	0.67	0.65	0.69	0.66	0.64	0.68
	3	96	0.58	0.45	0.70	0.65	0.51	0.79	0.65	0.51	0.78	0.65	0.52	0.78
	4	558	0.67	0.63	0.72	0.69	0.65	0.74	0.70	0.65	0.74	0.69	0.64	0.74
	5	547	0.65	0.61	0.70	0.69	0.65	0.74	0.69	0.64	0.73	0.69	0.64	0.74
Second cycles	1	152	0.64	0.55	0.74	0.73	0.64	0.81	0.73	0.64	0.82	0.77	0.68	0.87
	2	1132	0.63	0.60	0.67	0.68	0.64	0.71	0.68	0.64	0.71	0.73	0.68	0.77
	3[e]	20	0.57	0.24	0.90	0.34	0.04	0.63	0.37	0.07	0.66	0.47	0.16	0.78
	4	189	0.68	0.57	0.78	0.70	0.59	0.81	0.69	0.58	0.80	0.75	0.66	0.85
	5	157	0.58	0.47	0.68	0.60	0.50	0.71	0.59	0.48	0.69	0.64	0.52	0.75
Year														
First cycles	2005	573	0.66	0.61	0.70	0.66	0.62	0.62	0.67	0.62	0.62	0.66	0.62	0.71
	2006	615	0.68	0.64	0.73	0.68	0.64	0.73	0.68	0.64	0.73	0.69	0.64	0.73
	2007	629	0.67	0.63	0.71	0.70	0.66	0.74	0.70	0.66	0.75	0.69	0.65	0.73
	2008	705	0.65	0.60	0.69	0.67	0.63	0.71	0.67	0.63	0.71	0.67	0.63	0.71
	2009	749	0.67	0.63	0.71	0.69	0.64	0.73	0.69	0.65	0.73	0.68	0.64	0.72
	2010	767	0.63	0.59	0.67	0.68	0.64	0.72	0.68	0.64	0.72	0.67	0.63	0.71
	2011	561	0.63	0.58	0.68	0.64	0.60	0.69	0.64	0.59	0.69	0.65	0.60	0.70
Second cycles	2005	247	0.59	0.49	0.69	0.60	0.52	0.52	0.59	0.52	0.52	0.66	0.57	0.75
	2006	235	0.62	0.54	0.70	0.72	0.52	0.68	0.71	0.52	0.67	0.78	0.70	0.86
	2007	259	0.61	0.54	0.69	0.68	0.61	0.74	0.68	0.62	0.75	0.72	0.65	0.79
	2008	277	0.66	0.59	0.73	0.66	0.60	0.73	0.66	0.59	0.72	0.70	0.63	0.77
	2009	269	0.60	0.52	0.67	0.69	0.62	0.75	0.70	0.63	0.76	0.73	0.66	0.81
	2010	241	0.68	0.62	0.75	0.71	0.65	0.78	0.71	0.65	0.78	0.76	0.69	0.83
	2011	122	0.65	0.58	0.73	0.68	0.57	0.79	0.64	0.54	0.75	0.74	0.64	0.85

Table 5, cont.

CI-confidence interval, AUC-Area under the receiver operating characteristic curve, LB-lower bound, UB-upper bound

[a] Linear term for age

[b] Backwards stepwise logistic regression using a p-value$\geq$0.1 for exclusion, with main effects only (no interactions)

[c] Backwards stepwise logistic regression using a p-value$\geq$0.1 for exclusion, with main effects and interactions between variables

[d] Boosted regression tree with up to 10,000 regression trees, 6-way interactions, shrunk by a factor of 0.005, and bagging of 60% of data

[e] AUC<0.50 because increasing predicted probability of live birth associated with decreased odds of live birth in this subset

Table 1.6. Performance of models predicting live birth in first *in vitro* fertilization cycles and second cycles after a prior failed cycle by age category, 2005-2011

			Linear Age[a]			Main Effects[b]			Interactions[c]			Boosted Regression[d]		
				95% CI			*95% CI*			*95% CI*			*95% CI*	
	Age category	**N**	**AUC**	**LB**	**UB**	**AUC**	**LB**	**UB**	**AUC**	**LB**	**UB**	**AUC**	**LB**	**UB**
First cycles	<35	2067	0.56	0.53	0.58	0.60	0.58	0.63	0.60	0.58	0.63	0.60	0.58	0.63
	35-37	1108	0.52	0.48	0.55	0.60	0.56	0.63	0.59	0.55	0.63	0.59	0.55	0.63
	38-40	913	0.55	0.51	0.59	0.58	0.54	0.62	0.59	0.54	0.63	0.59	0.54	0.63
	41-42	355	0.46[e]	0.38	0.54	0.51	0.42	0.60	0.53	0.43	0.62	0.52	0.43	0.62
	≥43	156	0.51	0.41	0.60	0.69	0.56	0.82	0.64	0.50	0.78	0.64	0.51	0.77
Second cycles	<35	547	0.49	0.45	0.54	0.57	0.52	0.62	0.57	0.52	0.62	0.65	0.58	0.73
	35-37	423	0.53	0.47	0.58	0.63	0.58	0.69	0.63	0.57	0.69	0.69	0.62	0.76
	38-40	430	0.53	0.47	0.59	0.66	0.60	0.72	0.63	0.56	0.69	0.71	0.63	0.78
	41-42	188	0.61	0.53	0.70	0.67	0.55	0.80	0.68	0.56	0.80	0.77	0.68	0.87
	≥43	62	0.50	0.22	0.79	0.68	0.31	1.05	0.69	0.30	1.08	0.76	0.53	1.00

CI-confidence interval, AUC-area under the receiver operating characteristic curve, LB-lower bound, UB-upper bound

[a] Linear term for age

[b] Backwards stepwise logistic regression using a p-value≥0.1 for exclusion, with main effects only (no interactions)

[c] Backwards stepwise logistic regression using a p-value≥0.1 for exclusion, with main effects and interactions between variables

[d] Boosted regression tree with up to 10,000 regression trees, 6-way interactions, shrunk by a factor of 0.005, and bagging of 60% of data

[e] AUC<0.50 because increasing predicted probability of live birth associated with decreased odds of live birth in this subset

Table 1.7. Performance of models predicting live birth in first *in vitro* fertilization cycles and second cycles after a prior failed cycle at various predicted probability thresholds for defining high versus low probability of live birth, 2005-2011

Threshold defining high vs. low risk	Performance measures[a]	*First cycles*					*Second cycles after a prior failed cycle*				
		Age category[b]	Linear age[c]	Main effects[d]	Inter-actions[e]	Boosted regression[f]	Age category[b]	Linear age[c]	Main effects[d]	Inter-actions[e]	Boosted regression[f]
5%	Sens	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00
	Spec	0.05	0.00	0.02	0.01	0.00	0.05	0.00	0.01	0.01	0.00
	PPV	0.36	0.34	0.34	0.34	0.34	0.30	0.28	0.28	0.28	0.28
	NPV	0.94	-	0.93	0.94	1.00	0.93	-	0.97	0.97	-
10%	Sens	0.99	1.00	0.98	0.98	0.98	0.99	1.00	0.99	0.99	1.00
	Spec	0.05	0.00	0.08	0.07	0.07	0.05	0.00	0.08	0.09	0.02
	PPV	0.36	0.34	0.35	0.35	0.35	0.30	0.28	0.29	0.30	0.28
	NPV	0.94	-	0.89	0.90	0.88	0.93	-	0.94	0.97	1.00
15%	Sens	0.96	1.00	0.96	0.96	0.96	0.94	1.00	0.95	0.95	0.97
	Spec	0.16	0.00	0.16	0.16	0.14	0.19	0.01	0.19	0.19	0.18
	PPV	0.38	0.34	0.37	0.37	0.37	0.32	0.28	0.31	0.31	0.31
	NPV	0.89	1.00	0.88	0.87	0.88	0.88	0.94	0.92	0.91	0.95
20%	Sens	0.96	0.97	0.92	0.92	0.93	0.94	0.93	0.89	0.87	0.91
	Spec	0.16	0.09	0.23	0.23	0.23	0.19	0.18	0.34	0.35	0.34
	PPV	0.38	0.36	0.38	0.38	0.38	0.32	0.31	0.34	0.34	0.35
	NPV	0.89	0.86	0.85	0.86	0.86	0.88	0.88	0.89	0.88	0.91
25%	Sens	0.82	0.88	0.88	0.88	0.88	0.74	0.74	0.77	0.75	0.81
	Spec	0.39	0.30	0.32	0.33	0.32	0.46	0.47	0.49	0.49	0.49
	PPV	0.42	0.39	0.40	0.40	0.40	0.36	0.35	0.37	0.36	0.38
	NPV	0.80	0.83	0.83	0.84	0.84	0.81	0.82	0.84	0.84	0.87

Sens – sensitivity, Spec – specificity, PPV – positive predictive value, NPV – negative predictive value

[a] PPV & NPV calculated using probability of live birth observed in the study population

[b] Assigned probabilities of live birth based on age as follows: <35 years: 40.1%, 35-37 years: 31.8%, 38-40 years: 21.5%, 41-42 years: 12.2%, ≥43 years: 4.2%

[c] Linear term for age

[d] Backwards stepwise logistic regression using a p-value≥0.1 for exclusion, with main effects only (no interactions)

[e] Backwards stepwise logistic regression using a p-value≥0.1 for exclusion, with main effects and interactions between variables

[f] Boosted regression tree with up to 10000 regression trees, 6-way interactions, shrunk by a factor of 0.005, and bagging of 60% of data

Table 1.8. Parameter estimates for a model predicting live birth in second *in vitro* fertilization cycles after a prior failed cycle, main effects only, 2005-2011[a] (n=6534)

		95% CI		
	Odds Ratio	**LB**	**UB**	**p-value**
Variables from second cycle				
Age (years)	0.09	0.01	1.02	0.05
Age^2 ($years^2$)	1.08	1.01	1.16	0.03
Age^3 ($years^3$)	1.00	1.00	1.00	0.02
Body mass index (kg/m^2)	0.97	0.95	1.00	0.05
Weight (per 10 lbs)	1.03	0.99	1.08	0.14
Stimulation protocol				
Antagonist	0.82	0.71	0.96	0.01
Normal response	1.00	reference		-
High response	1.02	0.77	1.35	0.89
Low response	0.95	0.80	1.13	0.58
Race				
Asian	0.84	0.71	0.99	0.04
Black	0.70	0.57	0.85	<0.001
Hispanic/Latino	0.85	0.66	1.10	0.22
Other	0.51	0.21	1.24	0.14
White	1.00	reference		-
Pregnancy history				
Never pregnant/never live birth	1.00	reference		-
Ever pregnant/never live birth	1.15	1.02	1.31	0.03
Ever pregnant/ever live birth	1.16	0.99	1.36	0.07
Variables from prior failed cycle				
Total amount of gonadotropins (per 500 IUs)	0.97	0.95	0.99	<0.001
Maximum estradiol (per 100 IUs)	1.005	1.001	1.010	0.02
Metaphase II oocytes retrieved	1.02	1.01	1.04	<0.001
Assisted hatching performed				
None	1.00	reference		-
At least some	0.88	0.75	1.02	0.10
Embryos transferred	0.92	0.83	1.03	0.14
Preimplantation genetic diagnosis				
No	1.00	reference		-
Yes	0.60	0.41	0.88	0.01
Day of embryo transfer				
<3 days	0.32	0.04	2.60	0.29
3 days	1.00	reference		-
4-5 days	1.05	0.90	1.24	0.53
6+ days	0.97	0.76	1.23	0.79
No transfer	0.64	0.48	0.84	0.001

Table 1.8, cont.

		95% CI		
	Odds Ratio	**LB**	**UB**	**p-value**
Complications				
No	1.00	reference		-
Yes	1.56	1.03	2.38	0.04
Constant	1.69E+11	0.10	2.97E+23	0.07

CI-confidence interval, LB-lower bound, UB-upper bound, IU-international units

[a]Model built using backwards stepwise logistic regression with p-values$\geq$0.1 for exclusion

Table 1.9. Parameter estimates for a model predicting live birth in second *in vitro* fertilization cycles after a prior failed cycle, main effects and interactions, 2005-2011[a] (n=6491)

		95% CI		
	Odds Ratio	**LB**	**UB**	**p-value**
Variables from second cycle				
Age (years)	0.09	0.01	1.02	0.05
Age^2 (years^2)	1.08	1.01	1.16	0.03
Age^3 (years^3)	0.999	0.998	0.9999	0.02
BMI (kg/m^2)	0.99	0.96	1.02	0.33
Weight (per 10lbs)	1.03	0.99	1.08	0.15
Stimulation protocol				
Antagonist	0.83	0.71	0.96	0.01
Normal response	1.00	reference		-
High response	1.03	0.78	1.37	0.81
Low response	0.95	0.80	1.13	0.57
Race				
Asian	0.75	0.56	1.02	0.07
Black	0.82	0.60	1.13	0.23
Hispanic/Latino	0.51	0.32	0.82	0.01
White	1.00	reference		-
Pregnancy history				
Never pregnant/never live birth	1.00	reference		-
Ever pregnant/never live birth	1.14	1.00	1.30	0.05
Ever pregnant/ever live birth	1.16	0.99	1.36	0.07
Variables from prior failed cycle				
Total amount of gonadotropins (per 500 IUs)	0.97	0.95	0.99	0.001
Maximum estradiol (per 100 IUs)	1.004	0.998	1.009	0.19
Metaphase II oocytes retrieved	1.03	1.01	1.04	<0.001
Assisted hatching performed				
None	1.00	reference		-
At least some	0.88	0.75	1.02	0.09
Embryos transferred	0.93	0.84	1.04	0.19
Preimplantation genetic diagnosis				
No	1.00	reference		-
Yes	0.59	0.41	0.87	0.01
Day of embryo transfer				
<4 days	1.00	reference		-
4-5 days	1.81	0.89	3.67	0.10
6+ days	5.09	1.52	17.09	0.01
No transfer	0.95	0.44	2.02	0.89

Table 1.9, cont.

		95% CI		
	Odds Ratio	**LB**	**UB**	**p-value**
Complications				
No	1.00	reference		-
Yes	1.59	1.04	2.42	0.03
Race*Maximum estradiol				0.01
Asian*Maximum estradiol	1.00	0.99	1.02	0.38
Black* Maximum estradiol	0.99	0.98	1.00	0.22
Hispanic/Latino*Maximum estradiol	1.02	1.01	1.04	0.01
White*Maximum estradiol	1.00	reference		-
BMI*Day of embryo transfer				0.04
BMI*<4 days	1.00	reference		-
BMI*4-5 days	0.98	0.95	1.01	0.14
BMI*6+ days	0.94	0.89	0.98	0.01
BMI*no transfer	0.99	0.96	1.01	0.31
Constant	1.43E+11	0.07	2.91E+23	0.08

CI-confidence interval, LB-lower bound, UB-upper bound, BMI-body mass index, IU-international units

[a]Model built using backwards stepwise logistic regression with p-values$\geq$0.1 for exclusion; interaction terms included when p-values<0.05

FIGURES

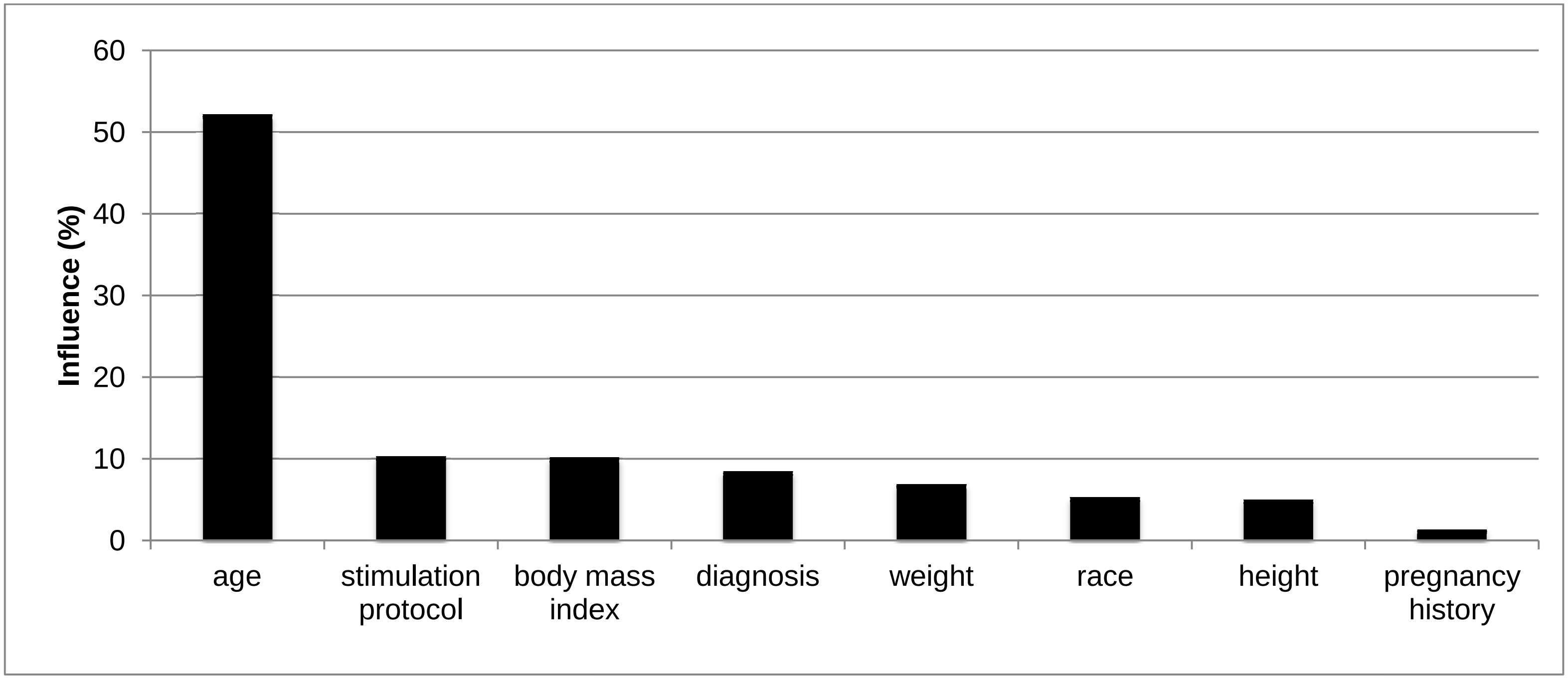

Figure 1.1. **Influence of variables predicting live birth in second *in vitro* fertilization cycles after a prior failed cycle in boosted regression model, 2005-2011**. The boosted regression model was built allowing up to 10,000 regression trees and 6-way interactions, shrunk by a factor of 0.005 and bagging 60% of data. Each split on a variable in a regression tree results in an increase in the log likelihood; the influence is the sum of the increase in log likelihood across all trees due to a particular variable, scaled to total 100% across all variables. Age had the largest influence, followed by stimulation protocol, body mass index, and diagnosis.

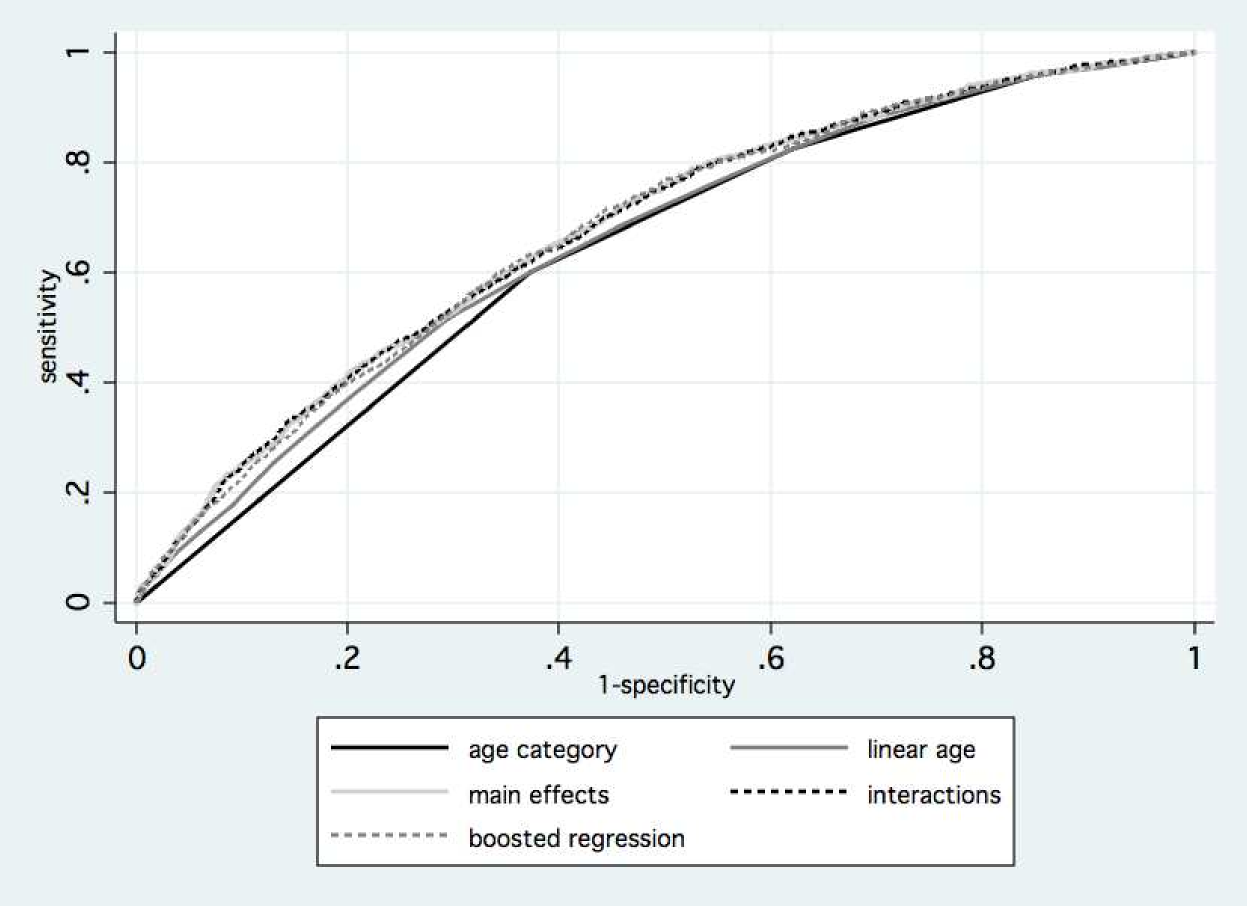

Figure 1.2. Receiver operating characteristic curves (ROC) for models predicting live birth in first *in vitro* fertilization cycles, 2005-2011. The age category model assigned probabilities of live birth based on age as follows: <35 years: 40.1%, 35-37 years: 31.8%, 38-40 years: 21.5%, 41-42 years: 12.2%; >43 years: 4.2%. The linear age model was based on a single linear age term. The main effects and interactions models were developed using backwards stepwise logistic regression with a p-value>0.1 for exclusion. The boosted regression model was built allowing up to 10,000 regression trees and 6-way interactions, shrunk by a factor of 0.005 and bagging 60% of data. All of the curves are very close together, reflecting similar model performance.

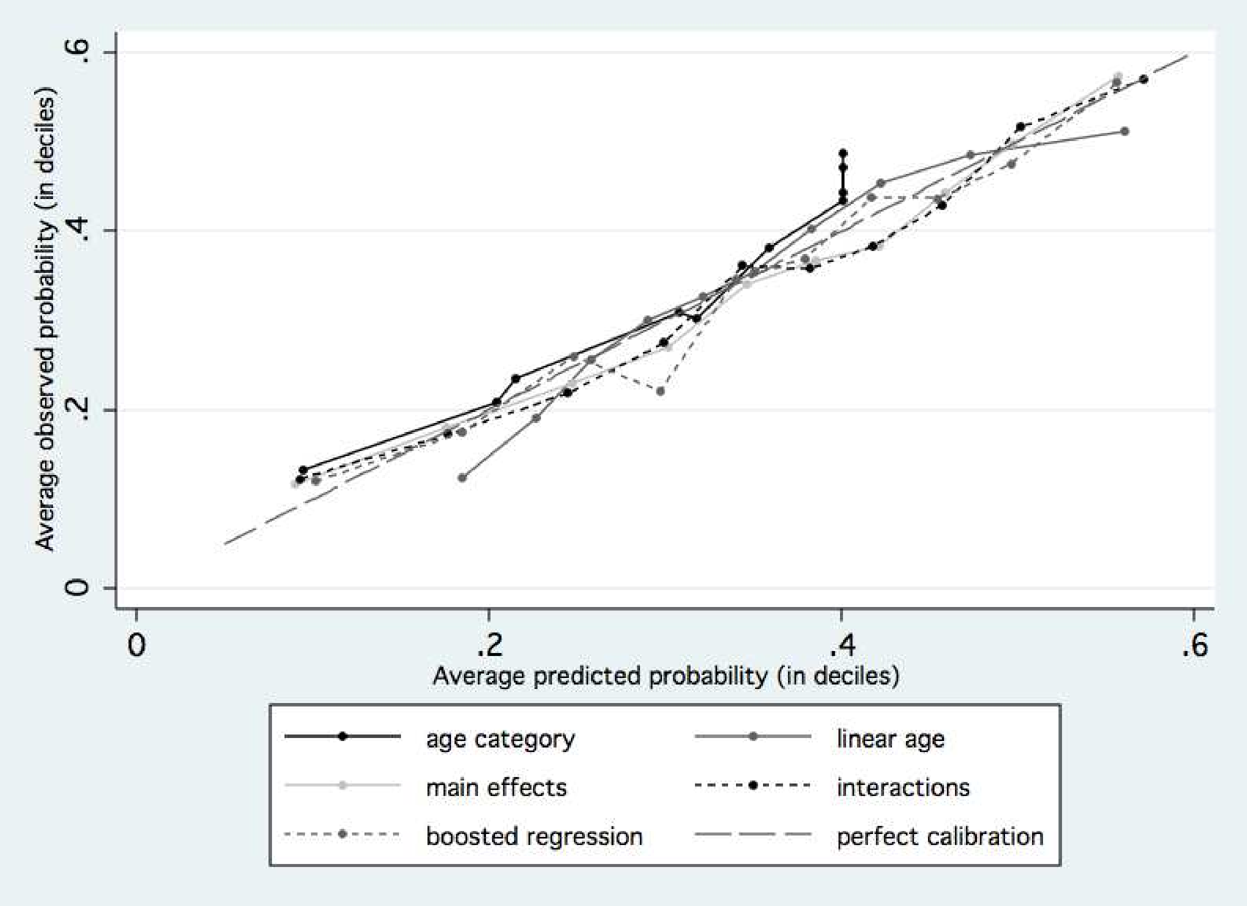

Figure 1.3. Calibration plot for models predicting live birth in first *in vitro* fertilization cycles, 2005-2011. Calibration plots display the average predicted probability versus the average observed probability within deciles of predicted probability. Perfect calibration occurs when observed and predicted probabilities are equivalent. The age category model assigned probabilities of live birth based on age as follows: <35 years: 40.1%, 35-37 years: 31.8%, 38-40 years: 21.5%, 41-42 years: 12.2%, >43 years: 4.2%. The linear age model was based on a single linear age term. The main effects and interactions models were developed using backwards stepwise logistic regression with a p-value>0.1 for exclusion. The boosted regression model was built allowing up to 10,000 regression trees and 6-way interactions, shrunk by a factor of 0.005 and bagging 60% of data. All models have good calibration with observed and predicted probabilities not far from perfect calibration.

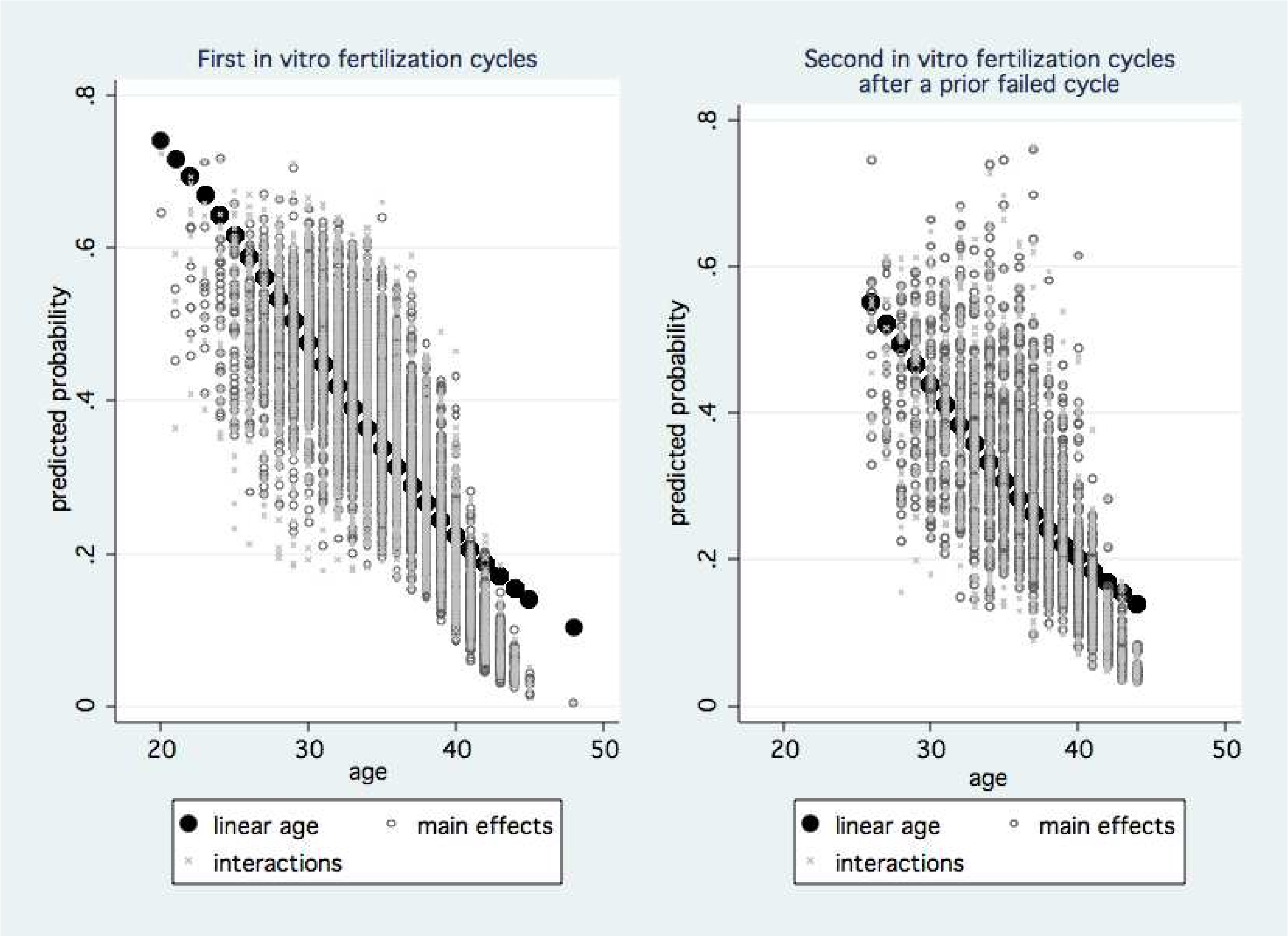

Figure 1.4. Predicted probabilities of live birth in first *in vitro* fertilization cycles and second cycles after a prior failed cycle by age in linear age, main effects, and interactions models, 2005-2011. The linear age model was based on a single linear age term. The main effects and interactions models were developed using backwards stepwise logistic regression with a p-value≥0.1 for exclusion. In both first and second cycles, the predicted probabilities among women of the same age vary quite dramatically for the main effects and interactions model, compared to the linear age models.

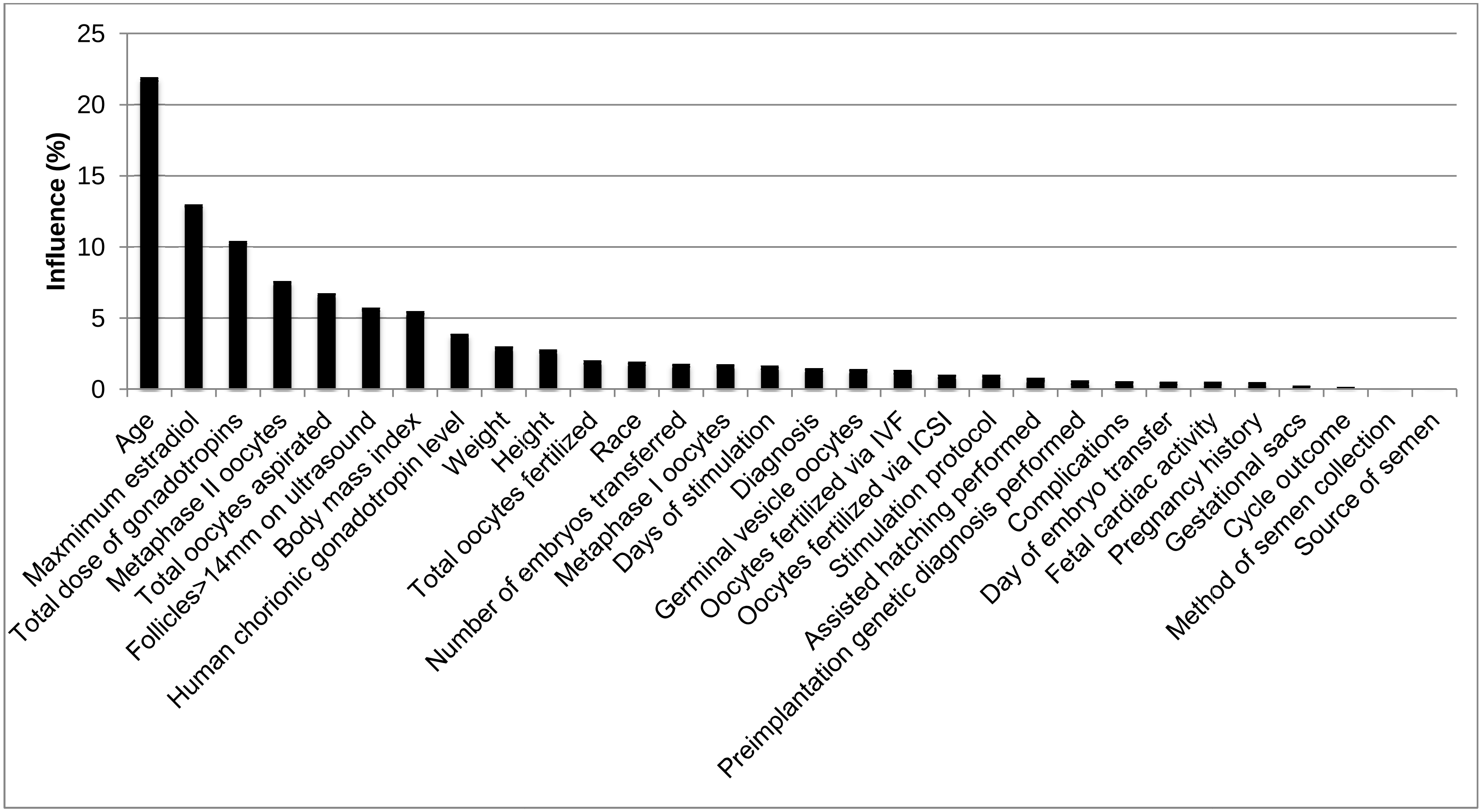

Figure 1.5. Influence of variables predicting live birth in second *in vitro* fertilization cycles after a prior failed cycle in boosted regression tree model, 2005-2011. The boosted regression model was built allowing up to 10,000 regression trees and 6-way interactions, shrunk by a factor of 0.005 and bagging 60% of data. Each split on a variable in a regression tree results in an increase in the log likelihood; the influence is the sum of the increase in log likelihood across all trees due to a particular variable, scaled to total 100% across all variables. Age had the strongest influence of all variables, followed by the maximum estradiol level, total dose of gonadotropins administered, the number of metaphase II oocytes retrieved, the number of follicles>14mm observed on ultrasound, and BMI (body mass index). All other variables had influences of <5%.

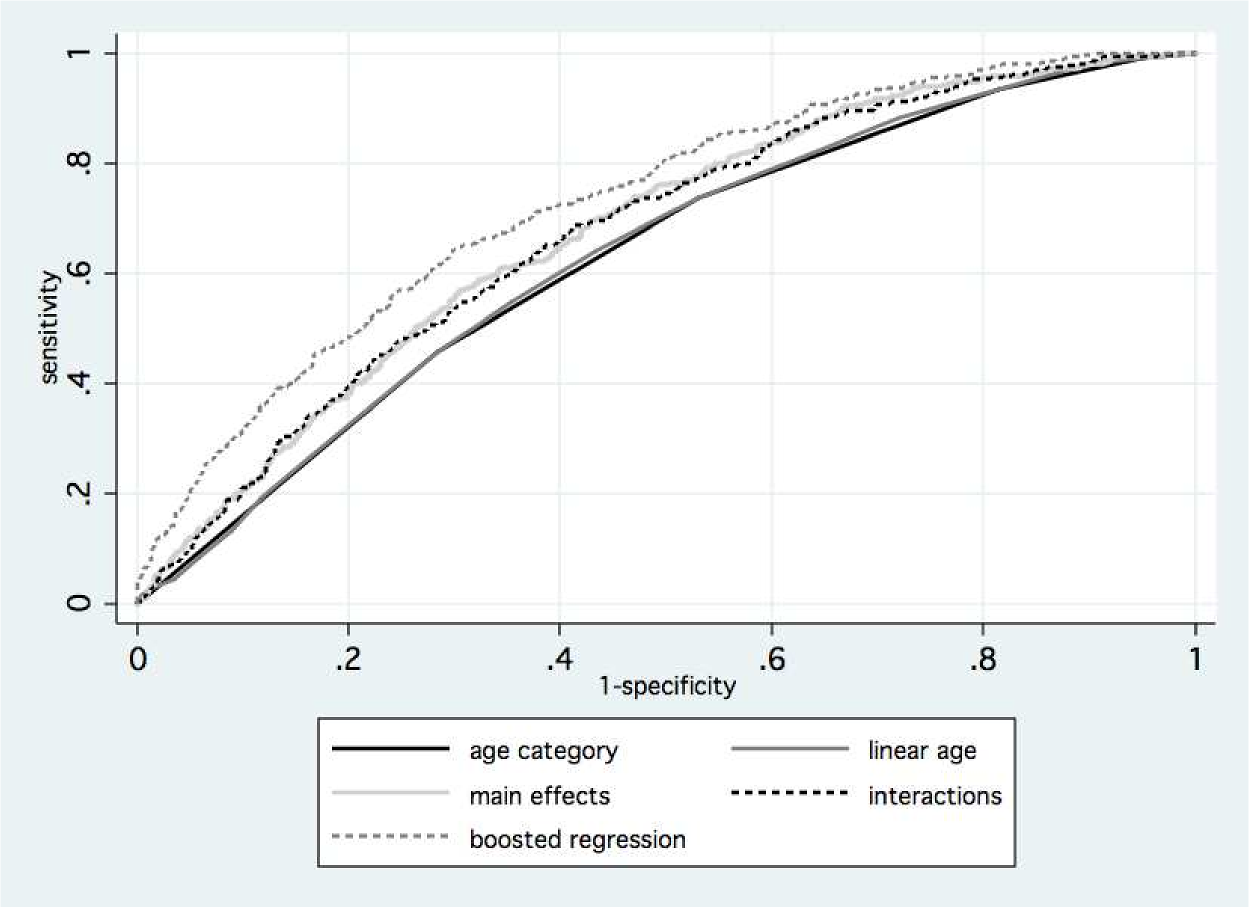

Figure 1.6. Receiver operating characteristic curves (ROC) for models predicting live birth in second *in vitro* fertilization cycles after a prior failed cycle, 2005-2011. The age category model assigned probabilities of live birth based on age as follows: <35 years: 40.1%, 35-37 years: 31.8%, 38-40 years: 21.5%, 41-42 years: 12.2%, >43 years: 4.2%. The linear age model was based on a single linear age term. The main effects and interactions models were developed using backwards stepwise logistic regression with a p-value>0.1 for exclusion. The boosted regression model was built allowing up to 10,000 regression trees and 6-way interactions, shrunk by a factor of 0.005 and bagging 60% of data. The boosted regression model has the best performance as it comes closer to the upper left corner (maximum sensitivity and specificity). The main effects and interactions models have overlapping curves, demonstrating similar performance. The age category and linear age models had the poorest performance with curves below the other models.

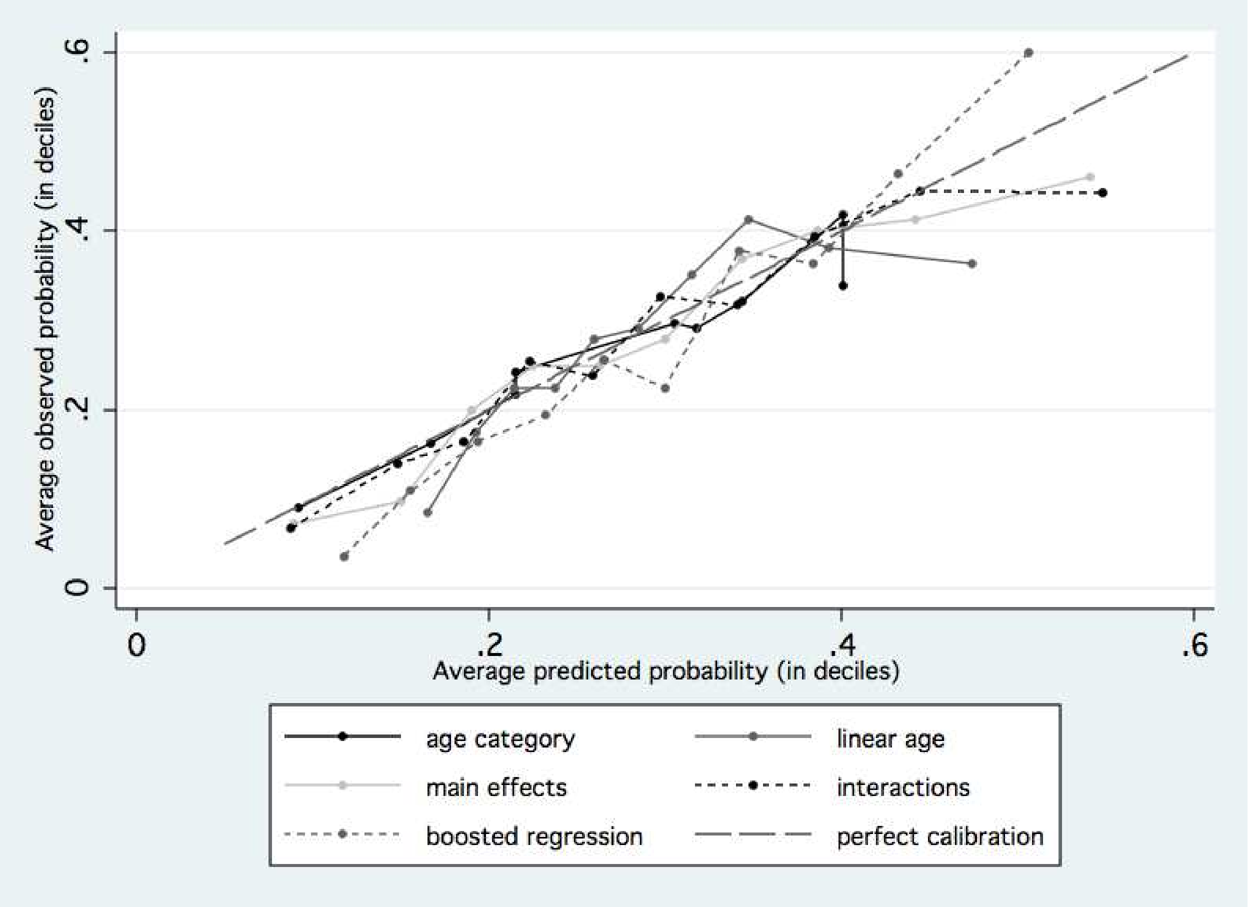

Figure 1.7. Calibration plot for models predicting live birth in second *in vitro* fertilization cycles after a prior failed cycle, 2005-2011. Calibration plots display the average predicted probability versus the average observed probability within deciles of predicted probability. Perfect calibration occurs when observed and predicted probabilities are equivalent. The age category model assigned probabilities of live birth based on age as follows: <35 years: 40.1%, 35-37 years: 31.8%, 38-40 years: 21.5%, 41-42 years: 12.2%, >43 years: 4.2%. The linear age model is based on a single linear age term. The main effects and interactions models were developed using backwards stepwise logistic regression with a p-value>0.1 for exclusion. The boosted regression model was built allowing up to 10,000 regression trees and 6-way interactions, shrunk by a factor of 0.005 and bagging 60% of data. All models have good calibration with observed and predicted probabilities not far off from perfect calibration. Deviations from perfect calibration are greatest for all models in the 10th decile.

Chapter 2. Prediction modeling for live birth in *in vitro* fertilization: anti-Müllerian hormone and other biomarkers of ovarian reserve

INTRODUCTION

Since the introduction of *in vitro* fertilization (IVF) to treat infertility, clinicians have been interested in identifying a single test to reliably quantify ovarian reserve and ovarian aging. Ovarian reserve refers to the quantity and quality of follicles in the ovary at a given time [1], which declines with age ultimately leading to menopause. It is one measure of a woman's fertility and can be used to predict the response to IVF, among other reproductive outcomes. A common measure of ovarian reserve is female age, because the number of oocytes decreases exponentially with age [2] in concert with a decrease in oocyte quality [3-6]. Other individual biomarkers of ovarian reserve, such as follicle stimulating hormone (FSH) and antral follicle count (AFC), are commonly measured as predictors of IVF success. Recently, anti-müllerian hormone (AMH) has emerged as a potential candidate to improve prediction of IVF success compared with other established markers of ovarian reserve.

AMH is a dimeric glycoprotein of the transforming growth factor beta (TGFβ) family of growth and differentiation factors [7]. AMH is expressed in granulosa cells of primary follicles, is maximal in preantral and small antral follicles, and is not expressed in larger antral follicles [8]. Therefore, AMH is expressed in follicles that have undergone recruitment from the primordial follicle pool but have not been selected for dominance [9]. The release of AMH from ovarian granulosa cells is proportional to the number of developing follicles in the ovaries [8], thus reflecting ovarian reserve. Women with higher AMH levels have more developing follicles and greater ovarian reserve compared with women who have lower AMH levels. Indeed, higher AMH levels [10] and a slower rate of change in AMH [11] have been associated with longer time to menopause (i.e., greater ovarian reserve) after adjustment for age. Although AMH is not expressed in primordial follicles, the true measure of ovarian reserve, the number of recruited follicles correlates with the size of the primordial follicle pool [12] and the more clinically relevant outcome of response to ovarian stimulation.

AMH may be a better marker of ovarian reserve than FSH and AFC because it is independent of the menstrual cycle [13-15], which means it can be measured on any day of the cycle. Even endocrine influences such as hormonal contraceptive use [16], gonadotropin releasing hormone agonists [17], and pregnancy [18] do not significantly affect AMH measurement. In addition, while most other markers of ovarian reserve and aging (age, FSH, AFC) are distantly related to the initial development of primordial follicles, AMH production in follicles first appears in primary follicles. Substances related to earlier follicle development, like AMH, may correlate better with the size of the primordial follicle pool.

Despite a number of studies demonstrating the ability of AMH to predict response to ovarian stimulation or pregnancy, few have assessed the role of AMH in predicting live birth [19-26], the outcome of primary concern for patients. Furthermore, none of these studies have used data from the U.S., which may limit the generalizability of findings to U.S. populations. Across Europe, eligibility criteria for IVF vary and may include: age limits, medical indications, marital or cohabitation status, and sexual orientation; and all EU countries have either partial or full insurance coverage of IVF [27]. Therefore we sought to examine the ability of AMH to predict live birth in IVF cycles in the U.S. and to compare its predictive ability to other common predictors, such as age, FSH, and AFC. We also evaluated whether the predictive ability of AMH varied by age, body mass index (BMI), polycystic ovary syndrome (PCOS) status, and infertility diagnosis, since these factors may influence how well AMH predicts live birth. Lastly, we evaluated whether addition of AMH improved prediction of live birth in IVF in a previously developed multivariable model.

METHODS

Study Subjects

We used electronic medical records data from a single infertility clinic located in Washington State. For all analyses, female patients undergoing autologous fresh IVF cycles with or without intra-cytoplasmic sperm injection (ICSI) between January 2010 and August 2011 were eligible for inclusion; those using donor eggs or cryopreserved eggs or embryos were excluded. All cycles were eligible such that more than one cycle per woman could be included in analyses.

This project was approved by the Institutional Review Board at the University of Washington.

Outcome

The primary outcome of interest was live birth (yes/no) in the cycle of interest. The outcome was obtained by the clinic via telephone follow-up with patients who self-reported the outcome. Because this clinic is required to report to the Society for Assisted Reproductive Technology (SART), outcome data are nearly complete and are validated in compliance with SART guidelines [28]. Fewer than 2% of cycles had discrepancies between the SART and medical records data for cycle outcomes [29]. Only subjects with outcome data were included in analyses.

Predictors

Predictors of live birth included age, FSH, AMH, and AFC. Subject's date of birth was obtained at the initial clinic visit, from which age was calculated and recorded in the electronic medical record. As part of the initial infertility evaluation, blood draw was performed by venipuncture, typically before 9:30am. Samples were collected in a primary tube and allowed to clot for 15-20 minutes, after which they were centrifuged at 3600 rpm for 10 minutes to separate serum. For FSH measurement, serum samples were sent to the clinic's central laboratory, where they were measured the same day as collection. Samples that could not be sent to the lab the same day of blood draw were stored between 0 and +4°C and sent the next day. After analysis, serum was stored for one month between -25°C and -15°C. Basal FSH was measured in mIU/mL using the Immulite 2000 (Diagnostic Product Corporation, Los Angeles, CA, USA) assay according to the manufacturer's protocol. The limit of detection for FSH was 0.1 mIU/mL. A third party laboratory, LabCorp, measured AMH for the clinic. After clotting and centrifugation, samples were frozen between -25°C and -15°C and sent to LabCorp for processing, which occurred within one day. AMH measurement was performed with the Diagnostic Systems Laboratory ELISA kit (Webster, TX, USA) following manufacturer protocols, with a limit of detection 0.1 ng/mL. For AFC, clinicians counted the number of follicles observed on vaginal ultrasound in each ovary; the total from both ovaries was the AFC. AFC was measured for the majority of subjects within seven days before the cycle start in order to obtain basal levels, but was not consistent with respect to day of the menstrual cycle.

Not all women had AFC, FSH, and AMH measured in each cycle, therefore values from other cycles were included if they were measured within one year prior to the start of stimulation in the index cycle ("cycle start" hereafter), as there appears to be little inter-cycle variation in this timeframe [30]. To obtain basal and not stimulated AFC, we linked multiple cycles within a woman and excluded any values measured between the cycle start and the day of cancellation, retrieval, or transfer. We only included women in the analyses who had age, FSH, AFC, and AMH measured within one year prior to the cycle start. Women missing one or more of these values or who had had measurements outside of the year window were excluded.

Values below the limit of detection. Both AMH and FSH have a lower limit of detection (LOD), below which the measurement cannot be obtained. Exclusion of these observations can lead to bias in the estimates of model performance; however, inclusion requires defining a replacement value. Because <3% of AMH and <1% of FSH values were below the LOD, we replaced AMH and FSH values below the LOD with LOD/2. With such a small proportion of subjects affected by the LOD, simple replacement is unlikely to introduce substantial bias [31-32].

Data Analysis

Study sample characteristics. We first examined the characteristics of women included and excluded from analyses to determine if there were any differences in these two groups that could potentially introduce bias. Characteristics included demographics, anthropometry, reproductive history, and infertility diagnosis, as well as the IVF cycle characteristics and response. We also examined the distribution of the various individual predictors of IVF success (age, FSH, AFC, and AMH), as well the timing of their measurement.

Individual predictors. We fit a logistic regression model with live birth as the outcome and a linear term for each of the predictors separately. We did not examine non-linear terms or transformations because measures for evaluating performance of single biomarkers are rank-based and therefore would be equivalent regardless of model parameterization. We took into account repeated cycles per woman by using clustered sandwich estimators of the standard error for the logistic regression models, which was used to calculate all 95% confidence intervals (CI). We plotted the nonparametric receiver operating characteristic (ROC) curve for each predictor, which demonstrates the discrimination of the model -- the ability to distinguish women with and without a live birth. We also estimated the area under the ROC curve (AUC) and 95% CI. AUCs range from 0.50 to 1.00, with a value of 0.50 indicating no ability to discriminate and perfect discrimination at a value of 1.00. To determine if AMH performed better than the other predictors, we estimated the difference in AUCs between AMH and each of age, FSH, and AFC and 95% CIs. For the AUCs and differences in AUCs, 95% CIs were calculated using a bootstrap procedure with 1000 samples. Samples were drawn such that *all* cycles from a particular woman were included if the woman was selected into the bootstrap sample, accounting for repeated cycles.

Assessment of the predictors' sensitivity and specificity required defining a cutoff probability above which women would be classified as high probability of live birth (hereafter referred to as “live birth likely”) and below as low probability (“live birth unlikely”). Sensitivity is the proportion of women who had a live birth who were correctly classified as live birth likely; the specificity is the proportion correctly classified as live birth unlikely among those who did not have a live birth. Without perfect sensitivity and specificity (almost impossible in practice), some women will be misclassified (inappropriately assigned live birth likely or unlikely). The relative “costs” of these two types of misclassification may be unequal, but ascribing a specific or relative cost to each is often fraught with guesswork. However, in the infertility setting the goal is typically to minimize the number of women inappropriately classified as live birth unlikely (false negative rate, 1-sensitivity) who may be excluded from or opt out of treatment unnecessarily, which means maximizing the sensitivity. Therefore, across models we examined the specificity and 95% CI at various levels of sensitivity ranging from 0.80 to 0.99, with a corresponding false negative rate of 0.20 to 0.01. We also examined the positive and negative predictive values (PPV and NPV, respectively) at the various sensitivity levels, which are a function of the probability of live birth in the study population. PPV is the proportion of women who had a live birth among those classified as live birth likely; NPV is the proportion of women who did not have a live birth among those classified as live birth unlikely. A more accurate model would have higher specificity, PPV, and NPV at the selected sensitivities.

Sensitivity analyses. Bias may have been introduced due to the exclusion of women who were missing a value for one or more predictors. One concern about missing data is that at sufficiently high (or low) values of a particular biomarker, a clinician may have decided that measurement of other biomarkers would add no further value to assessing a patient’s prognosis.

Therefore women at the extremes of predictor values may have been excluded. To examine this possibility, we performed sensitivity analyses in which we estimated the AUC and 95% CI for each predictor among all women who had that specific predictor measured within one year of the cycle start, leading to different sample sizes for each predictor. We also examined the sensitivity of results to the window prior to the cycle start for retrieving biomarker values from the electronic medical record. In these analyses, we estimated the AUC and 95% CI using cutoffs of within 90 days and within 2 years prior to the cycle start.

Performance of AMH within subgroups. Because the association between AMH and IVF outcomes (pregnancy and live birth) has been reported to vary with women's age [33], we examined wither the AUC varied across age categories used by SART (<35, 35-37, 38-40, 41-42, and $\geq$43 years) [34]. We also examined the AUC within categories of BMI (underweight/normal: <25kg/m^2, overweight: 25 to <30 kg/m^2, obese: $\geq$30kg/m^2). On average, obese women have lower AMH levels than their normal weight peers even after adjustment for age and race [35], and they also have lower rates of live birth following IVF [36]; however it is unclear if these differences lead to variation in predictive ability of AMH across BMI categories.

Furthermore, we evaluated whether polycystic ovary syndrome (PCOS) modified the AUC of AMH. PCOS is diagnosed when at least two of the following three features are present: (1) oligo- and/or anovulation, (2) clinical and/or biochemical signs of hyperandrogenism, and (3) polycystic ovaries on transvaginal ultrasonography [37]. AMH is 2-3 times higher in serum from women with PCOS than in women with normal ovaries [38], and the production of AMH within granulosa cells is also increased [39]. Despite these higher levels of AMH, which in general are associated with better IVF outcomes, women with PCOS have worse or similar outcomes as compared to non-PCOS women [40]. Therefore, the performance of AMH may vary between women with and without PCOS. Because we did not have complete information on PCOS status in the dataset, we used two proxy definitions for PCOS: (1) a diagnosis of ovulation disorder/polycystic ovaries (PCO), and (2) a diagnosis of ovulation disorder/PCO and BMI$\geq$25kg/m^2 because 1/3 to 1/2 of women with PCOS are overweight or obese [41].

Lastly, because AMH reflects ovarian reserve, rather than other components of infertility, we examined whether the AUC of AMH varied by infertility diagnosis (diminished ovarian reserve, tubal factor, or male factor). A woman was classified into a particular diagnosis category if she had at least one indication for the diagnosis across the three available diagnosis variables and did not have an indication for the remaining two diagnoses of interest (e.g., women with an indication for two or more of the diagnoses of interest were excluded from analyses).

Incremental improvement over age alone. We also added AMH to age in a prediction model to examine if the two predictors together could improve performance over age alone. For these analyses we split the sample, using 80% for model development and 20% for model validation. In the development sample we examined various parameterizations of AMH (linear and log-transformed) in logistic regression models with a linear term for age. For the parameterization with the lowest p-value, in the validation sample we estimated the AUC and difference in AUC compared to an age-only model, along with 95% CIs, as described above. Furthermore, we estimated the AUC and 95% CI of this AMH and age model within the SART age categories described previously.

Application to multivariable prediction model. In previous work, we developed multivariable models to predict the probability of live birth in women undergoing their first IVF cycle using predictors obtained prior to stimulation. A model built using backward stepwise logistic regression with a p-value$\geq$0.2 for removal included age, age^2, age^3, height (inches), weight (pounds), stimulation protocol to be used (antagonist, normal responder protocol, high responder protocol, and low responder protocol), race/ethnicity (White, Asian, Black, Hispanic, other), infertility diagnosis (diminished ovarian reserve [DOR], endometriosis, tubal factor, male infertility, ovulation disorders/polycystic ovaries [PCO], uterine factor, unexplained, other, or multiple factors), and pregnancy history (never pregnant/never live birth, ever pregnant/never live birth, ever pregnant/ever live birth).

We assessed whether addition of AMH to this multivariable model improved prediction accuracy. The group of women included in this analysis differed from the group in which we developed and tested the multivariable prediction model and from the analyses described above. Here, we include only women with (1) a first autologous fresh IVF cycle, (2) initiated between 2010 and 2011 in the Washington State clinic, and (3) an AMH value within one year of the cycle start. Because a large proportion (~50%) of women were missing at least one of the predictors, we used multiple imputation by chained equations to create 15 complete datasets in which missing data were replaced with plausible values, excluding AMH [42]. We split the dataset prior to imputation and used 80% to re-fit the multivariable model with and without AMH, removing the other category from race/ethnicity and the low responder protocol from the stimulation protocol due to small cell sizes.

We used the remaining 20% of observations to assess model performance. We plotted the ROC curve for the models with and without AMH, and estimated the AUC and corresponding 95% CI. To determine if AMH improved model discrimination, we obtained the difference in AUCs and 95% CIs associated with models that did and did not include AMH. We also calculated each model’s sensitivity, specificity, PPV, and NPV. We examined these measures of performance across a number of different cutoffs that may be relevant to defining live birth likely versus unlikely, including 5%, 10%, 15%, 20%, 25%.

We calculated standard errors for all measures of model performance by using a bootstrap procedure with 300 sampled datasets in order to obtain 95% CIs. For the multiply imputed datasets, model performance (and performance measures' bootstrapped standard errors) was calculated within each imputation, and Rubin’s Rules were applied to pool estimates across datasets [43].

RESULTS

Study sample characteristics

There were a total of 1542 fresh autologous IVF cycles between 2010 and 2011; 834 (54.1%) cycles had measurement of all biomarkers within the appropriate window and complete data on the outcome of the cycle (Table 1). Across the 834 cycles, the majority of women were 35 years of age or older (59.0%), white (56.5%), of normal BMI (57.1%), and nulligravid (50.7%). The most common infertility diagnoses were male infertility (27.5%), multiple factors (24.8%), and DOR (12.5%). The live birth rate was 39.6%, and 22.4% of those with a live birth delivered two or more infants. Five hundred forty-five women had only one cycle included in analyses, 112 had two, 29 had three, and two had four. Among excluded women, 212 only were missing one or more predictors, 202 only had measurements outside the 1-year window, and 1 only was missing the cycle outcome; 293 women were excluded for multiple reasons (data not shown). Cycles excluded from analyses were fairly similar, except they were less likely to have a diagnosis of male infertility, to have embryos transferred, to become pregnant, and to have a live birth. Women with included cycles had slightly lower values of AMH, larger values of AFC, and similar values of FSH to those women excluded from analyses (Table 2).

Individual predictors

Logistic regression. When fitting a logistic regression model with each of the predictors and the dichotomous live birth outcome, we observed positive relationships with live for AMH and AFC, and negative relationships for FSH and age (Table 3).

Model performance. Across predictors, AUC values were similarly modest with the exception of FSH (Table 4). The AUC for AMH was 0.63 (95% CI=0.59, 0.67), 0.64 (95% CI=0.60, 0.67) for AFC, and 0.67 (95% CI=0.64, 0.71) for age. The AUC for FSH was smaller at 0.55 (95% CI=0.51, 0.59). When comparing AMH to the other predictors, there was a substantial difference only from FSH (difference=0.08, 95% CI=0.04, 0.13); all other differences were negligible (absolute value of differences≤0.04). These differences were also reflected in the ROC curves (Figure 1), with FSH having a curve below that of the other predictors.

Across all sensitivities, specificities were smallest for FSH, ranging from 0.02 to 0.25 (Table 5). Age had the largest specificities across all sensitivities, ranging from 0.06 to 0.49. Specificities for AMH were the second smallest across all sensitivities. These results reiterate the findings from the AUC and ROC curves that FSH has the poorest performance among the predictors and age had slightly better performance, but similar to AMH and FSH. Overall the specificity was quite low for all predictors even at the lowest sensitivity examined (0.80), with ≤50% of the women who did not have a live birth correctly classified as live birth unlikely. This percentage got substantially smaller with increasing sensitivity. Across all sensitivities and predictors, the PPV was moderate and the NPV was large, ranging from 0.40 to 0.50 and 0.66 to 0.88, respectively.

Sensitivity analyses. Across all sensitivity analyses, AUCs were similar to the primary analysis for all predictors (Table 4).

Performance of AMH within subgroups. Across age groups, there was variability in the performance of AMH in predicting live birth (Table 6). Women younger than 35 years of age had lower performance than older age groups, with an AUC of 0.53 compared to a range of 0.57 (35-37 years old) to 0.65 (38-40 years old).

Across categories of BMI, AUCs were similar, ranging from 0.62 to 0.63.

Performance was similar in women with and without PCOS when defined as a diagnosis of ovulation disorder/PCO (AUC of 0.63 vs. 0.62) and slightly better for women with PCOS when defined as a diagnosis of ovulation disorder/PCO in addition to being overweight or obese (AUC of 0.68 vs. 0.62).

Within diagnoses, the performance of AMH in women with DOR and male infertility was similar with AUCs of 0.56 and 0.53, respectively. The AUC for tubal factor was slightly higher (AUC=0.65); however, confidence intervals were wide for all estimates.

Incremental improvement over age alone

The logistic regression model with log-transformed AMH had a smaller p-value than linear AMH in the development sample. When we compared the performance of a model with log-transformed AMH and linear age (AUC=0.69, 95% CI=0.49, 0.66, data not shown) to age alone, there was very small improvement in AUC (difference in AUC=0.02, 95% CI=-0.0005, 0.04).

Within age groups, the predictive ability of the AMH and age model was slightly better with increasing age with an AUC of 0.55 (95% CI=0.49, 0.61) and 0.58 (95% CI=0.50, 0.67) in women <35 and 35-37 years of age, respectively, and 0.62 (95% CI=0.47, 0.76) and 0.64 (95% CI=0.24, 1.05) in women 41-42 and ≥43 years of age, respectively (data not shown). However, CIs were overlapping for all age groups and were very wide for the oldest age groups.

Application to multivariable prediction model

There were 718 women who underwent their first IVF cycle between 2010 and 2011 with complete outcome data, as well as an eligible AMH measurement. When we updated the multivariable model to include AMH, the AUC in the validation sample of 141 women was 0.69 (95% CI=0.60, 0.79), almost identical to that in the model without AMH (Table 7). There was no difference in the AUC estimates (difference=-0.003, 95% CI=-0.03, 0.02) between these two models, which was also reflected in the ROC curves (Figure 2).

Sensitivities and specificities were similar for the model with and without AMH across all cutoffs defining live birth likely and unlikely, as were the PPVs and NPVs.

DISCUSSION

In this study we investigated the ability of AMH to predict live birth in IVF cycles, and compared its performance to other established predictors, including AFC, FSH, and age. We were particularly interested in AMH because it displays less intra-cycle variability than AFC and FSH [44, 45] and is not altered by exogenous or endogenous endocrine influences [16-18]. Although a number of previous studies have examined AMH as a predictor of live birth in the infertility setting, this work was warranted because all other studies have utilized data from clinics outside of the U.S. where inclusion criteria for IVF, insurance coverage, and governmental regulations differ from the U.S. We hypothesized that AMH would outperform the other predictors since it is more directly related to the underlying ovarian reserve and may be more sensitive to its underlying changes [46]. Contrary to this hypothesis, models with AMH performed similarly to those with AFC or age, although slightly better than models with FSH: models with AMH had an AUC of 0.63, compared to 0.64 for AFC, 0.67 for age, and 0.55 for FSH.

The predictive ability of AMH is consistent with that observed in other studies, in which AUCs ranged from 0.57 to 0.66 [20-24, 26]. The small variations may be attributable to sampling variability or to actual differences in study populations, including differences in sample size, geographic location, and eligible patient ages and infertility diagnoses. In addition, parameterizations of AMH have differed across studies (e.g., continuous vs. categorical). We were concerned that exclusion of subjects missing any biomarker values or those measured outside of the one-year window may have led to bias in estimated performance. However, AUC estimates in sensitivity analyses where the inclusion criteria were modified were almost identical to those in primary analyses, further supporting the stability of the findings within this study and across others.

In almost all of the comparable studies, AMH performed similarly to other single predictors of live birth, including AFC, FSH, and age, as we also observed [21, 23, 26]. Across these studies, different predictors emerged as superior based on statistical comparisons (AMH [21, 26], age [23]); however, these increases were quite small and likely would not translate to major improvements in classification. Consistent with previous investigations, all single predictors had low to moderate discrimination in the current study, with AUCs ranging from 0.55 to 0.67, suggesting that none adequately distinguished women who went on to have a live birth from those who did not. The particularly low AUC for FSH may be attributable to its greater intra-cycle variability [45], as this biomarker was not measured on a particular day of the menstrual cycle, blunting its performance.

Across acceptable sensitivities (0.80 to 0.99), all predictors had low to moderate specificities: among women who did not have a live birth ≤50% were correctly classified as live birth unlikely. We evaluated model performance at selected sensitivities because this allows comparison of specificities, PPVs, and NPVs across models without having to select an optimal cutoff for defining live birth likely versus unlikely, and criteria for optimizing may differ across clinicians and patients (e.g., different weighting of false positives and false negatives). Furthermore, as incorrectly refusing or deterring someone from treatment is a potentially more costly error than inappropriately proceeding with treatment, we examined model performance with fixed, small false negative rates (1-sensitivity).

Across all predictors, PPVs were moderate and NPVs were high, indicating that a negative test result accurately reflected the true outcome of the cycle (no live birth), but a positive test did not. Therefore, for women classified as live birth unlikely, the majority would not go on to have a live birth. However, less than 50% of women classified as live birth likely would actually have a live birth. With a high NPV a clinician may be more comfortable excluding a woman from treatment if classified as live birth unlikely, since the classification accurately reflects the underlying outcome (no live birth). Because PPV and NPV are influenced by the prevalence of the outcome, their low and high values, respectively, are partially attributable to the low prevalence of live birth in IVF. With a greater prevalence of live birth, the PPV increases and NPV decreases.

The limited ability of these markers to predict live birth is likely at least in part a function of the more distant relationship between measures of ovarian reserve and a live birth outcome. Ovarian reserve measures the quantity and quality of primordial follicles in the ovaries, but other factors affecting conception, implantation, and pregnancy loss are not necessarily captured by these single measures. In particular, factors that affect uterine receptivity and spontaneous abortion (aside from oocyte quality) are likely not reflected in predictors examined here. Indeed, adjustment for oocyte yield has been reported to nullify the association between AMH and live birth [26], suggesting AMH may only reflect quantity and quality of primordial follicles. Furthermore, the predictive ability of these ovarian reserve measures was substantially larger in some studies examining poor response or excessive response, typically defined by the number of oocytes retrieved, which is more closely related to ovarian reserve [19, 26]. However, live birth is ultimately the outcome of interest for both patients and physicians, and use of other outcomes can be misleading. Inclusion of women with male factor infertility may also have blunted the performance of the markers, as has been reported in one prior study [22], because ovarian reserve may be less informative for this infertility diagnosis. Indeed, the AUC was slightly lower for male factor than tubal factor infertility.

We also examined whether addition of AMH to age could improve model performance and observed only a very small improvement in the AUC compared to an age-only model, suggesting little benefit to including AMH with age to determine individualized probabilities of live birth. Contrary to these findings, one study observed an AUC of 0.66 for a model with categories of age and AMH, with substantial improvements over age (AUC=0.55) alone [24], which was confirmed in external validation [25].

Furthermore, we evaluated whether addition of AMH to an extant multivariable prediction model improved prediction accuracy. In a model that included age, age^2, age^3, height, weight, stimulation protocol to be used, race/ethnicity, infertility diagnosis, and pregnancy history, addition of AMH did not improve model AUC, which was ~0.70 for models with and without AMH. AMH did not lead to better performance as measured by the sensitivity, specificity, PPV, and NPV either. The lack of improvement may be because AMH is not associated strongly enough with live birth to improve the AUC in an existing model that already includes many variables associated with live birth. It is underappreciated that associations typically considered moderate, such as an odds ratio of 2 or 3, do not lead to improvements in classification when added to existing prediction models [47,48], and each unit increase in AMH was associated with an odds ratio of only 1.13.

We also examined the AUCs with various groups of women defined by age, BMI, and PCOS status. AUCs increased slightly with more advanced age. In women <35 years of age, the AUC was 0.53 and increased to 0.65 in women 38-40 years of age. These findings are consistent with other studies, including one that reported an association between AMH tertile and live birth rate in women 34-37 years of age, but no association in those <34 and 38-41 years of age [33]. In another study, the AUC of log-transformed AMH was 0.51 in women <35 years of age, which was smaller than the AUC of 0.65 in women ≥35 years of age [22]. Therefore, AMH may be a more useful predictor of live birth in older women, in whom the probability of live birth is already diminished due to advanced ovarian aging.

In obese women, AMH levels [35] and live birth rates [36] are reportedly lower than those of normal weight women; therefore, AMH may not be equally predictive of live birth across BMI. However, we observed that the AUC for AMH was almost identical within all groups of BMI. We also examined whether the predictive ability of AMH varied by PCOS status, as women with PCOS have reportedly higher levels of AMH [38] and lower or equivalent live birth rates following IVF [40]. However, when we defined PCOS as a diagnosis of ovulation disorder/PCO, the AUCs were almost identical. Because this diagnosis may include many women without PCOS who have an unrelated ovulation disorder, we examined a more specific definition that included being overweight/obese, since ~50% of women with PCOS are obese [41]. With this definition the differences in AUCs were more apparent, with an AUC of 0.68 in women with PCOS compared to 0.62 in underweight or normal weight women without an ovulation disorder/PCO diagnosis. There were only 48 women who met this definition for PCOS, resulting in a wide confidence interval and an unstable estimate of the model performance. Therefore, these analyses should be repeated in another larger sample with an actual PCOS diagnosis based on Rotterdam criteria [37]. Furthermore, although live birth rates have been found to be similar in women undergoing IVF with and without PCOS, we observed higher rates in women with PCOS than without, which may have influenced these findings.

We also examined whether the performance of AMH varied by the infertility diagnosis. Because measures of ovarian reserve, such as AMH, may be used to define a DOR diagnosis, they may have limited predictive ability within this subgroup. Indeed, mean and variability of AMH were smaller among women with DOR than male and tubal factor infertility. We observed that the AUC was similar among women with DOR and male infertility, and slightly higher for tubal factor. Because confidence intervals were wide, it is unclear if there are true differences in AMH performance by infertility diagnosis.

In conclusion, models with AMH had similar accuracy to other single markers of ovarian reserve for predicting live birth following IVF, and all single predictors investigated had only modest prediction accuracy. Despite the lack of differences in accuracy across predictors, age may be the best predictor of live birth in practice as it is recorded for all patients, costs nothing to obtain, and does not require any invasive tests, such as vaginal ultrasound or blood draw. For practical reasons, AMH may be a better predictor than FSH and AFC, as the latter measures of ovarian reserve (and particularly FSH) display more marked intra-cycle variability than AMH [44, 45], thus necessitating measurement within a specific timeframe for comparability. We observed that AMH was a better predictor of live birth in some subgroups of women, particularly older women and women with PCOS. Future studies should aim to replicate these findings and to determine the underlying mechanisms for differences within these subgroups. Furthermore, because almost no studies have included AMH as an eligible predictor in multivariable prediction models, future investigations should incorporate this variable into model development steps. This work will further elucidate if AMH in combination with other known predictors facilitates identification of women likely to have or not have a live birth.

REFERENCES

1. Domingues TS, Rocha AM, Serafini PC. Tests for ovarian reserve: reliability and utility. *Curr Opin Obstet Gynecol.* 2010;22:271–276.
2. Broekmans FJ, Soules MR, Fauser BC. Ovarian aging: mechanisms and clinical Consequences. *Endocr Rev*. 2009;30:465–493.
3. Cuckle HS, Wald NJ, Thompson SG. Estimating a woman's risk of having a pregnancy associate with Down's syndrome using her age and alpha-fetoprotein level. *Br J Obstet Gynaecol*. 1987;94:387-402.
4. Wilcox AJ, Weinberg CR, O'Connor JF, et al. Incidence of early loss of pregnancy. N Engl J Med. 1988;319:189-94.
5. Whittaker PG, Taylor A, Lind T. Unsuspected pregnancy loss in healthy women. Lancet 1983;321:1126-1127.
6. Hakim RB, Gray RH, Zacur H. Infertility and early pregnancy loss. *Am J Obstet Gynecol.* 1995;172:1510-7.
7. Cate RL, Mattaliano RJ, Hession C, et al. Isolation of the bovine and human genes for Mullerian inhibiting substance and expression of the human gene in animal cells. *Cell* 1986;45:685–698.
8. Broer SL, Mol B, Dolleman M, Fauser BC, Broekmans FJ. The role of anti-Mullerian hormone assessment in assisted reproductive technology outcome. *Curr Opin Obstet Gynecol.* 2010;22:193–201.
9. La Marca A, Broekmans FJ, Volpe A, Fauser BC, Macklon NS. Anti-Mullerian hormone (AMH): what do we still need to know? *Hum Reprod.* 2009;24:2264–2275.
10. Freeman EW, Sammel MD, Lin H, Gracia CR. Anti-mullerian hormone as a predictor of time to menopause in late reproductive age women. *J Clin Endocrinol Metab.* 2012;97:1673-80.
11. Freeman EW, Sammel MD, Lin H, Boorman DW, Gracia CR. Contribution of the rate of change of antimüllerian hormone in estimating time to menopause for late reproductive-age women. *Fertil Steril.* 2012;98:1254–1259.e2.
12. Scheffer GJ, Broekmans FJ, Dorland M, Habbema JD, Looman CW, te Velde ER. Antral follicle counts by transvaginal ultrasonography are related to age in women with proven natural fertility. *Fertil Steril*. 1999;72:845–851.
13. La Marca A, Stabile G, Artenisio AC, Volpe A. Serum anti-Müllerian hormone throughout the human menstrual cycle. *Hum Reprod.* 2006; 21:3103–3107.
14. Tsepelidis S, Devreker F, Demeestere I, Flahaut A, Gervy CH, Englert Y. Stable serum levels of anti-Mullerian hormone during the menstrual cycle: a prospective study in normoovulatory women. *Hum Reprod.* 2007;22:1837–1840.
15. Hehenkamp WJ, Looman CW, Themmen AP, de Jong FH, Te Velde ER, Broekmans FJ. Anti-Mullerian hormone levels in the spontaneous menstrual cycle do not show substantial fluctuation. *J Clin Endocrinol Metab.* 2006;91:4057-4063.
16. Somunkiran A, Yavuz T, Yucel O, Ozdemir I. Anti-Mullerian hormone levels during hormonal contraception in women with polycystic ovary syndrome. *Eur J Obstet Gynecol Reprod Biol.* 2007;134:196–201.
17. Arbo E, Vetori DV, Jimenez MF, Freitas FM, Lemos N, Cunha-Filho JS. Serum antimullerian hormone levels and follicular cohort characteristics after pituitary suppression in the late luteal phase with oral contraceptive pills. *Hum Reprod.* 2007;22:3192–3196.
18. La Marca A, Giulini S, Orvieto R, De Leo V, Volpe A. Anti-Müllerian hormone concentrations in maternal serum during pregnancy. *Hum Reprod.* 2005;20:1569–1572.
19. Mutlu MF, Erdem M, Erdem A, et al. Antral follicle count determines poor ovarian

response better than anti-Müllerian hormone but age is the only predictor for live birth in in vitro fertilization cycles. *J Assist Reprod Genet.* 2013;30:657-665.

20. Mjumder K, Gelbaya TA, Laing I. Nardo LG. The use of anti-Müllerian hormone and antral follicle count to predict the potential of oocytes and embryos. *Eur J Obstet Gynecol Reprod Biol.* 2010;150:166-170.
21. Lukaszuk K. Kunicki M, Liss J, Lukaszuk M, Jakiel G. Use of ovarian reserve parameters for predicting live births in women undergoing *in vitro* fertilization. *Eur J Obstet Gynecol Reprod Biol.* 2013;168:173-177.
22. Lee TH, Liu CH, Huang CC, Hsieh KC, Lin PM, Lee MS. Impact of female age and male infertility on ovarian reserve markers to predict outcome of assisted reproduction technology cycles. *Reprod Biol Endocrinol.* 2009;7:100. doi:10.1186/1477-7827-7-100.
23. Li HWR, Lee VCY, Lau EYL, Yeung WSB, Ho PC, Ng EHY. Role of baseline antral follicle count and anti-Mullerian hormone in prediction of cumulative live birth in the first in vitro fertilisation cycle: a retrospective cohort analysis. *PLoS ONE.* 2013;8:e61095.
24. La Marca A, Nelson SM, Sighinolfi G, et al. Anti-Müllerian hormone-based prediction model for a live birth in assisted reproduction. *Reprod Biomed Online.* 2011;22:341-349.
25. Khader A, Lloyd SM, McConnachie A, et al. External validation of anti-Müllerian hormone based prediction of live birth in assisted conception. *J Ovarian Res.* 2013;6:3. doi:10.1186/1757-2215-6-3
26. Nelson SM, Yates RW, Fleming R. Serum anti-Müllerian hormone and FSH: prediction of live birth and extremes of response in stimulated cycles--implications for individualization of therapy. *Hum Reprod.* 2007;22:2414-2421.
27. Brigham KB, Cadier B, Chevreul K. The diversity of regulation and public financing of IVF in Europe and its impact on utilization. *Hum Reprod.* 2013;28:666-675.
28. Adashi EY, Wyden R. Public reporting of clinical outcomes of assisted reproductive technology programs: implications for other medical and surgical procedures. *JAMA*. 2011;306:1135-1136.
29. Centers for Disease Control and Prevention, American Society for Reproductive Medicine, Society for Assisted Reproductive Technology. *2010 Assisted Reproductive Technology National Summary Report*. Atlanta: U.S. Department of Health and Human Services; 2012.
30. McIlveen M, Skull JD, Ledger WL. Evaluation of the utility of multiple endocrine and ultrasound measures of ovarian reserve in the prediction of cycle cancellation in a high-risk IVF population. *Hum Reprod.* 2007;22:778-785.
31. Perkins NJ, Schisterman EF, Vexler A. Receiver operating characteristic curve Inference from a sample with a limit of detection. *Am J Epidemiol*. 2007;165:325-333.
32. Lynn HS. Maximum likelihood inference for left-censored HIV RNA data. *Stat Med*. 2001;20:35-45.
33. Wang JG, Douglas NC, Nakhuda GS, et al. The association between anti-Müllerian hormone and IVF pregnancy outcomes is influenced by age. *Reprod Biomed Online.* 2010;21:757-761.
34. Centers for Disease Control and Prevention, American Society for Reproductive Medicine, Society for Assisted Reproductive Technology. *2011 Assisted Reproductive Technology Fertility Clinic Success Rates Report*. Atlanta: U.S. Department of Health and Human Services; 2013.
35. Freeman EW, Gracia CR, Sammel MD, Lin H, Lim LC, Strauss JF 3rd. Association of anti-mullerian hormone levels with obesity in late reproductive-age women. *Fertil Steril.* 2007;87:101-106.
36. Rittenberg V, Seshadri S, Sunkara SK, Sobaleva S, Oteng-Ntim E, El-Toukhy T. Effect of body mass index on IVF treatment outcome: an updated systematic review and meta-

analysis. *Reprod Biomed Online.* 2011;23:421-439.

37. Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Fertil Steril.* 2004;81:19-25.
38. Fallat ME, Siow Y, Marra M, Cook C, Carrillo A. Müllerian-inhibiting substance in follicular fluid and serum: a comparison of patients with tubal factor infertility, polycystic ovary syndrome, and endometriosis. *Fertil Steril.* 1997;67:962-965.
39. Pellatt L, Hanna L, Brincat M, et al. Granulosa cell production of anti-Müllerian hormone is increased in polycystic ovaries. *J Clin Endocrinol Metab.* 2007;92:240–24*5.*
40. Heijnen EM, Eijkemans MJ, Hughes EG, Laven JS, Macklon NS, Fauser BC. A meta-analysis of outcomes of conventional IVF in women with polycystic ovary syndrome. *Hum Reprod Update.* 2006;12:13-21.
41. Balen AH, Conway GS, Kaltsas G et al. Polycystic ovarian syndrome: the spectrum of the disorder in 1741 patients. *Hum Reprod.* 1995;10:2107-2111.
42. Royston P. Multiple imputation of missing values. *Stata J*. 2004;4:227–241.
43. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley & Sons; 1987.
44. Deb S, Campbell BK, Clewes JS, Pincott-Allen C, Raine-Fenning NJ. Intracycle variation in number of antral follicles stratified by size and in endocrine markers of ovarian reserve in women with normal ovulatory menstrual cycles. *Ultrasound Obstet Gynecol.* 2013;41:216-222.
45. van Disseldorp J, Lambalk CB, Kwee J, et al. Comparison of inter- and intra-cycle variability of anti-Mullerian hormone and antral follicle counts. *Hum Reprod.* 2010;25:221-227.
46. de Vet A, Laven JS, de Jong FH, Themmen AP, Fauser BC. Antimüllerian hormone serum levels: a putative marker for ovarian aging. *Fertil Steril.* 2002;77:357-362.
47. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol.* 2004;159:882-890.
48. Kattan MW. Judging new markers by their ability to improve predictive accuracy. *J Natl Cancer Inst*. 2003;95:634–635.

TABLES

Table 2.1. Distribution of patient and cycle characteristics in included and excluded *in vitro* fertilization cycles, 2010-2011

		Excluded[a]		Included[b]	
		N	**%**	**N**	**%**
Total		708	100.0	834	100.0
Year	*2010*	309	43.6	460	55.2
	2011	399	56.4	374	44.8
Age	*<25*	7	1.0	2	0.2
	25-29	65	9.2	80	9.6
	30-34	205	29.0	260	31.2
	35-39	254	35.9	318	38.1
	40+	177	25.0	174	20.9
Prior *in vitro* fertilization cycles	*0*	480	67.8	563	67.5
	1	162	22.9	195	23.4
	2+	64	9.0	76	9.1
	missing	2	0.3	0	0.0
Race	*Asian*	104	14.7	136	16.3
	Black	20	2.8	14	1.7
	Hispanic	23	3.2	11	1.3
	Other	3	0.4	4	0.5
	White	383	54.1	471	56.5
	missing	175	24.7	198	23.7
Body mass index (kg/m^2)	*Underweight*	13	1.8	22	2.6
	Normal	377	53.2	476	57.1
	Overweight	164	23.2	205	24.6
	Obese	154	21.8	131	15.7
Height (feet)	*<5*	14	2.0	20	2.4
	5 to <5.5	400	56.5	490	58.8
	5.5 to <6	286	40.4	308	36.9
	6+	8	1.1	16	1.9
Weight (lbs)	*<100*	5	0.7	9	1.1
	100 to <150	383	54.1	468	56.1
	150 to <200	223	31.5	295	35.4
	200 to <250	82	11.6	55	6.6
	250+	14	2.0	7	0.8
	missing	1	0.1	0	0.0
Gravidity	*0*	346	48.9	423	50.7
	1	170	24.0	221	26.5
	2	92	13.0	97	11.6
	3+	100	14.1	93	11.2

Table 2.1, cont.

		Excluded[a]		Included[b]	
		N	**%**	**N**	**%**
Diagnosis	*Diminished ovarian reserve*	71	10.0	104	12.5
	Endometriosis	45	6.4	40	4.8
	Tubal factor	47	6.6	67	8.0
	Male infertility	123	17.4	229	27.5
	Other	42	5.9	41	4.9
	Ovulation disorders/polycystic ovaries	40	5.6	36	4.3
	Unexplained	72	10.2	75	9.0
	Uterine factor	8	1.1	11	1.3
	Multiple factors	169	23.9	207	24.8
	missing	91	12.9	24	2.9
Stimulation protocol	*Antagonist*	248	35.0	312	37.4
	Agonist				
	Normal responder	89	12.6	105	12.6
	High responder	351	49.6	416	49.9
	Low responder	3	0.4	1	0.1
	missing	17	2.4	0	0.0
Cycle canceled	*No*	625	88.3	775	92.9
	Yes	83	11.7	59	7.1
Days of stimulation	*1 to 6*	17	2.4	14	1.7
	7 to 13	675	95.3	796	95.4
	14+	15	2.1	23	2.8
	missing	1	0.1	1	0.1
Source of semen	*Donor*	25	3.5	39	4.7
	Partner	683	96.5	795	95.3
Method of semen collection	*Aspiration*	23	3.2	24	2.9
	Biopsy	14	2.0	15	1.8
	Ejaculation	671	94.8	795	95.3
Follicles >14mm observed on ultrasound prior to retrieval	*1 to 4*	71	10.0	80	9.6
	5 to 9	220	31.1	306	36.7
	10 to 14	147	20.8	249	29.9
	15+	100	14.1	140	16.8
	missing	170	24.0	59	7.1

Table 2.1, cont.

		Excluded[a]		Included[b]	
		N	**%**	**N**	**%**
Follicles aspirated	*0*	12	1.7	4	0.5
	1 to 4	39	5.5	57	6.8
	5 to 9	155	21.9	187	22.4
	10 to 14	145	20.5	202	24.2
	15+	282	39.8	329	39.4
	missing	75	10.6	55	6.6
Oocytes fertilized	*0*	102	14.4	75	9.0
	1 to 4	189	26.7	219	26.3
	5 to 9	214	30.2	322	38.6
	10 to 14	107	15.1	143	17.1
	15+	96	13.6	75	9.0
Embryos transferred	*0*	132	18.6	93	11.2
	1	152	21.5	227	27.2
	2	271	38.3	385	46.2
	3+	144	20.3	128	15.3
	missing	9	1.3	1	0.1
Day of embryo transfer	*<4*	284	40.1	299	35.9
	4+	283	40.0	441	52.9
	No transfer	141	19.9	94	11.3
Outcome	*Biochemical pregnancy*	250	35.3	332	39.8
	Clinical intrauterine pregnancy + live birth	224	31.6	330	39.6
	Clinical intrauterine pregnancy + spontaneous abortion/stillbirth/ therapeutic abortion	49	6.9	55	6.6
	Ectopic	3	0.4	4	0.5
	Not pregnant	172	24.3	113	13.5
	missing	10	1.4	0	0.0
Number live born	*0*	474	66.9	504	60.4
	1	177	25.0	256	30.7
	2	46	6.5	73	8.8
	3+	1	0.1	1	0.1
	missing	10	1.4	0	0.0

[a] Cycles with one or more predictors missing or measured more ≥1 year prior to the cycle start or after the cycle start, or with missing cycle outcome data

[b] Cycles with all predictors available and measured within one year prior to the cycle start and complete cycle outcome data

Table 2.2. Distribution of predictors of live birth and timing of their measurement in included and excluded in vitro fertilization cycles, 2010-2011

					Percentile			
Predictor	**n**	**mean**	**sd**	**min**	**25%**	**50%**	**75%**	**max**
Included cycles								
Anti-Müllerian hormone (ng/mL)	814[a]	2.53	2.68	0.10	0.86	1.80	3.20	33.00
Days from cycle start to measurement	834	-117.90	89.14	-364	-175	-90	-44	0
Antral follicle count	834	16.20	10	0	9	14	21	70
Days from cycle start to measurement	834	-7.95	24.08	-360	-5	-4	-2	0
Follicle stimulating hormone (mIU/mL)	832[b]	7.79	3.54	0.91	5.78	7.19	8.82	36.00
Days from cycle start to measurement	834	-100.11	82.00	-360	-139	-70	-36	0
Age (years)	834	35.52	4.41	23	32	36	39	45
Excluded cycles								
Anti-Müllerian hormone (ng/mL)	300	2.85	3.60	0.20	1.00	1.70	3.40	30.00
Antral follicle count	708	10.36	11.17	0	0	9	16	58
Follicle stimulating hormone (mIU/mL)	300	7.75	3.73	0.24	5.87	6.90	8.73	35.30
Age (years)	708	35.73	4.58	21	32	36	40	47

sd-standard deviation

[a] 20 cycles have anti-Müllerian hormone measurements below the limit of detection

[b] 2 cycles have follicle stimulating hormone measurements below the limit of detection

Table 2.3. Association between predictors and live birth in in vitro fertilization cycles, 2010-2011

		95% CI		
Predictor	**Odds Ratio**	**LB**	**UB**	**p-value**
Anti-Müllerian hormone (ng/mL)	1.13	1.06	1.19	<0.001
Antral follicle count	1.04	1.03	1.06	<0.001
Follicle stimulating hormone (mIU/mL)	0.94	0.90	0.99	0.01
Age (years)	0.87	0.84	0.90	<0.001

CI-confidence interval, LB-lower bound, UB-upper bound

Table 2.4. Area under the receiver operating characteristic curve for predictors of live birth in in vitro fertilization cycles, 2010-2011

				Sensitivity analyses											
				Availability of each individual predictor				Measurement within 90 days of cycle start				Measurement within 2 years of cycle start			
		95% CI				*95% CI*				*95% CI*				*95% CI*	
Predictor	**AUC**	**LB**	**UB**	**n**	**AUC**	**LB**	**UB**	**n**	**AUC**	**LB**	**UB**	**n**	**AUC**	**LB**	**UB**
Anti-Müllerian hormone (ng/mL)	0.63	0.59	0.67	1049	0.63	0.59	0.67	377	0.63	0.59	0.67	896	0.63	0.59	0.67
Antral follicle count	0.64	0.60	0.67	1218	0.64	0.60	0.67	377	0.64	0.60	0.67	896	0.64	0.60	0.67
Follicle stimulating hormone (mIU/mL)	0.55	0.51	0.59	996	0.55	0.51	0.59	377	0.55	0.51	0.59	896	0.55	0.51	0.59
Age (years)	0.67	0.64	0.71	834	0.67	0.64	0.71	377	0.67	0.64	0.71	896	0.67	0.64	0.71
Difference in AUC compared to anti-Müllerian hormone															
Antral follicle count	-0.004	-0.04	0.03												
Follicle stimulating hormone (mIU/mL)	0.08	0.04	0.13												
Age (years)	-0.04	-0.08	0.001												

CI-confidence interval, AUC-area under the receiver operating characteristic curve, LB-lower bound, UB-upper bound

Table 2.5. Performance of predictors of live birth *in in vitro* fertilization cycles, 2010-2011

		Predictor			
Sensitivity	**Performance measure**	**Anti-Müllerian hormone (ng/mL)**	**Antral follicle count**	**Follicle stimulating hormone (mIU/mL)**	**Age (years)**
0.80	*Specificity*	0.37	0.39	0.25	0.49
	Positive PV	0.46	0.47	0.41	0.50
	Negative PV	0.75	0.76	0.66	0.77
0.85	*Specificity*	0.30	0.32	0.21	0.37
	Positive PV	0.45	0.45	0.41	0.46
	Negative PV	0.76	0.77	0.68	0.78
0.90	*Specificity*	0.22	0.27	0.17	0.28
	Positive PV	0.43	0.44	0.41	0.46
	Negative PV	0.78	0.80	0.72	0.78
0.95	*Specificity*	0.17	0.13	0.08	0.20
	Positive PV	0.43	0.43	0.40	0.45
	Negative PV	0.83	0.80	0.71	0.81
0.99	*Specificity*	0.05	0.04	0.02	0.06
	Positive PV	0.41	0.41	0.40	0.41
	Negative PV	0.87	0.84	0.80	0.88

CI-confidence interval, PV-predictive value

Table 2.6. Area under the receiver operating characteristic curve for anti-Müllerian hormone in predicting live birth in *in vitro* fertilization by age, body mass index, polycystic ovary syndrome, and infertility diagnosis 2010-2011

				95% CI	
	N	**Live birth rate (%)**	**AUC**	**LB**	**UB**
Age					
<35	342	54.1	0.53	0.47	0.59
35-37	170	42.4	0.57	0.48	0.66
38-40	201	26.9	0.65	0.56	0.74
41-42	87	17.2	0.61	0.45	0.76
≥43	34	11.8	0.64	0.32	0.96
Body mass index					
Underweight/normal	498	38.0	0.62	0.57	0.68
Overweight	205	47.3	0.62	0.54	0.70
Obese	131	33.6	0.63	0.52	0.74
Diagnosis of ovulation disorder/polycystic ovaries					
No	723	38.9	0.62	0.57	0.67
Yes	87	46.0	0.63	0.50	0.76
Diagnosis of ovulation disorder/polycystic ovaries and overweight/obese					
No	446	37.2	0.62	0.57	0.68
Yes	48	45.8	0.68	0.52	0.85
Infertility diagnosis					
Diminished ovarian reserve	127	20.5	0.56	0.42	0.71
Tubal factor	93	49.5	0.65	0.53	0.78
Male factor	284	46.8	0.53	0.46	0.60

CI-confidence interval, AUC-area under the receiver operating characteristic curve, LB-lower bound, UB-upper bound

Table 2.7. Performance of multivariable models predicting live birth in first *in vitro* fertilization cycles with and without anti-Müllerian hormone, 2010-2011

			No AMH	**AMH**	**Difference**
		AUC	0.70	0.69	-0.003
		95% CI	0.60, 079	0.60, 0.79	-0.03, 0.02
Cutoff of predicted probability defining live birth likely versus unlikely	0.05	*Sensitivity*	0.97	0.97	
		Specificity	0.02	0.01	
		Positive PV	0.37	0.36	
		Negative PV	0.63	0.39	
	0.10	*Sensitivity*	0.96	0.96	
		Specificity	0.06	0.07	
		Positive PV	0.37	0.38	
		Negative PV	0.72	0.76	
	0.15	*Sensitivity*	0.94	0.94	
		Specificity	0.12	0.15	
		Positive PV	0.38	0.39	
		Negative PV	0.76	0.82	
	0.20	*Sensitivity*	0.92	0.93	
		Specificity	0.22	0.21	
		Positive PV	0.41	0.41	
		Negative PV	0.82	0.83	
	0.25	*Sensitivity*	0.88	0.88	
		Specificity	0.31	0.30	
		Positive PV	0.43	0.42	
		Negative PV	0.82	0.81	

AMH-anti-Müllerian hormone, AUC-area under the receiver operating characteristic curve, CI-confidence interval, PV-predictive value

FIGURES

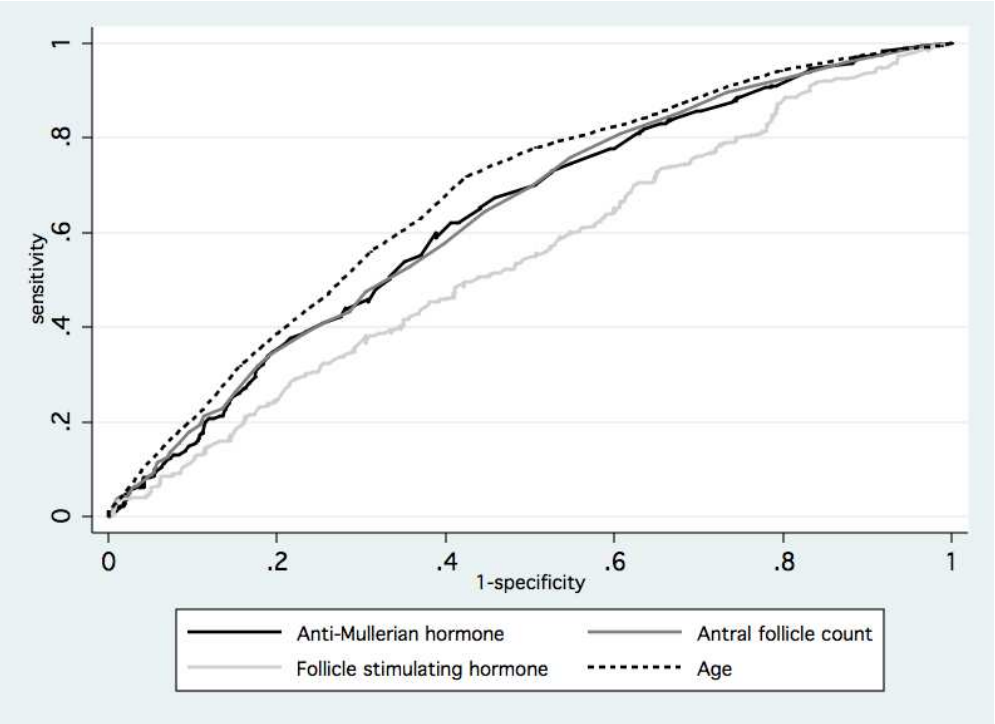

Figure 2.1. Receiver operating characteristic curves for predictors of live birth in *in vitro* fertilization cycles, 2010-2011. The curves display the sensitivity and 1-specificity across all thresholds for defining live birth likely and unlikely, with more accurate models closer to the upper left-hand corner. Follicle stimulating hormone had the poorest performance with a curve below that of the other predictors. Antral follicle count, age, and anti-Müllerian hormone have similar performance with overlapping curves.

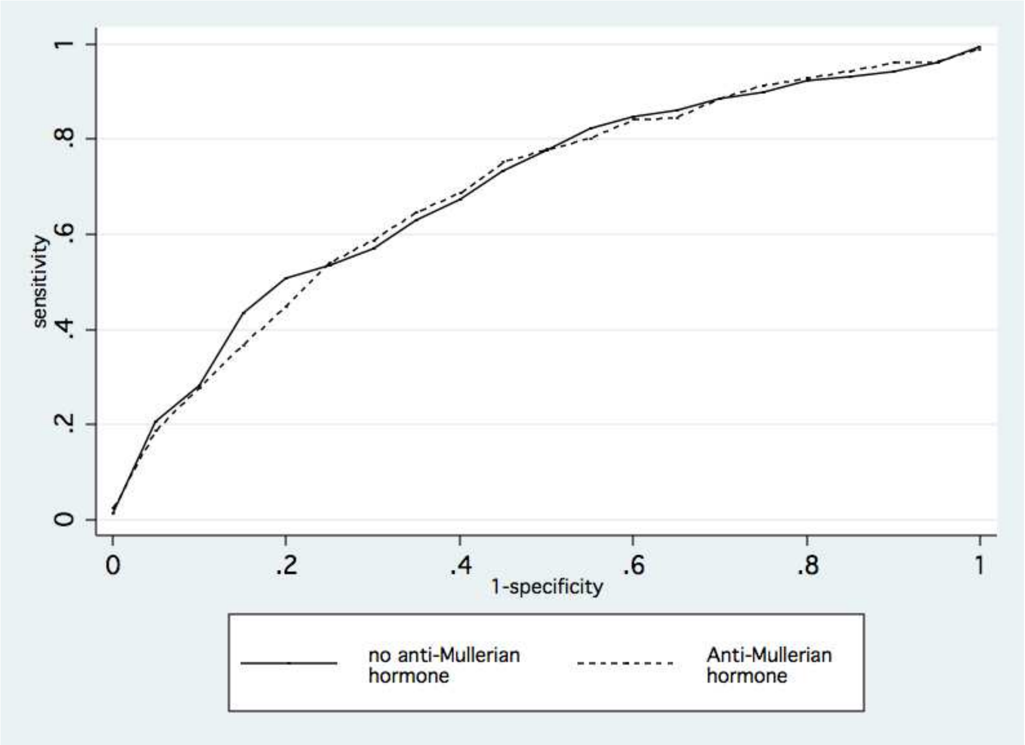

Figure 2.2. Receiver operating characteristic curves for multivariable models predicting live birth in *in vitro* fertilization cycles with and without anti-Müllerian hormone, 2010-2011. Multivariable models were developed in previous work and included age, age^2, age^3, number of prior intrauterine insemination cycles, height, weight, stimulation protocol to be used, race/ethnicity, infertility diagnosis, and pregnancy history, with and without anti-Müllerian hormone. The curves display the sensitivity and 1-specificity across all thresholds for defining live birth likely and unlikely, with more accurate models closer to the upper left-hand corner. There was no difference in the models with and without anti-Müllerian hormone, indicating no improvement in model performance with the addition of anti-Müllerian hormone.