Human Reproduction, Vol.37, No.3, pp. 565–576, 2022
Advance Access Publication on January 13, 2022 https://doi.org/10.1093/humrep/deab280

human reproduction ORIGINAL ARTICLE *Reproductive epidemiology*

Predictive models of pregnancy based on data from a preconception cohort study

Jennifer J. Yland[1,*,†], Taiyao Wang[2,3,†], Zahra Zad[2,4,†], Sydney K. Willis[1], Tanran R. Wang[1], Amelia K. Wesselink[1], Tammy Jiang[1], Elizabeth E. Hatch[1], Lauren A. Wise[1,‡], and Ioannis Ch. Paschalidis[2,4,5,‡]

[1]Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA [2]Center for Information and Systems Engineering, Boston University, Boston, MA, USA [3]Philips Research North America, Cambridge, MA, USA [4]Division of Systems Engineering, Department of Electrical and Computer Engineering, Boston University, Boston, MA, USA [5]Department of Biomedical Engineering, Boston University, Boston, MA, USA

*Correspondence address. Department of Epidemiology, Boston University School of Public Health, 715 Albany Street, Boston, MA 02118, USA. E-mail: yland@bu.edu https://orcid.org/0000-0001-7870-8971

Submitted on June 25, 2021; resubmitted on November 30, 2021; editorial decision on December 14, 2021

STUDY QUESTION: Can we derive adequate models to predict the probability of conception among couples actively trying to conceive?

SUMMARY ANSWER: Leveraging data collected from female participants in a North American preconception cohort study, we developed models to predict pregnancy with performance of ∼70% in the area under the receiver operating characteristic curve (AUC).

WHAT IS KNOWN ALREADY: Earlier work has focused primarily on identifying individual risk factors for infertility. Several predictive models have been developed in subfertile populations, with relatively low discrimination (AUC: 59–64%).

STUDY DESIGN, SIZE, DURATION: Study participants were female, aged 21–45 years, residents of the USA or Canada, not using fertility treatment, and actively trying to conceive at enrollment (2013–2019). Participants completed a baseline questionnaire at enrollment and follow-up questionnaires every 2 months for up to 12 months or until conception. We used data from 4133 participants with no more than one menstrual cycle of pregnancy attempt at study entry.

PARTICIPANTS/MATERIALS, SETTING, METHODS: On the baseline questionnaire, participants reported data on sociodemographic factors, lifestyle and behavioral factors, diet quality, medical history and selected male partner characteristics. A total of 163 predictors were considered in this study. We implemented regularized logistic regression, support vector machines, neural networks and gradient boosted decision trees to derive models predicting the probability of pregnancy: (i) within fewer than 12 menstrual cycles of pregnancy attempt time (Model I), and (ii) within 6 menstrual cycles of pregnancy attempt time (Model II). Cox models were used to predict the probability of pregnancy within each menstrual cycle for up to 12 cycles of follow-up (Model III). We assessed model performance using the AUC and the weighted-F1 score for Models I and II, and the concordance index for Model III.

MAIN RESULTS AND THE ROLE OF CHANCE: Model I and II AUCs were 70% and 66%, respectively, in parsimonious models, and the concordance index for Model III was 63%. The predictors that were positively associated with pregnancy in all models were: having previously breastfed an infant and using multivitamins or folic acid supplements. The predictors that were inversely associated with pregnancy in all models were: female age, female BMI and history of infertility. Among nulligravid women with no history of infertility, the most important predictors were: female age, female BMI, male BMI, use of a fertility app, attempt time at study entry and perceived stress.

LIMITATIONS, REASONS FOR CAUTION: Reliance on self-reported predictor data could have introduced misclassification, which would likely be non-differential with respect to the pregnancy outcome given the prospective design. In addition, we cannot be certain that all relevant predictor variables were considered. Finally, though we validated the models using split-sample replication techniques, we did not conduct an external validation study.

[†]These authors contributed equally to this work and should be considered joint first authors.

[‡]These last two authors should be considered joint senior authors.

WIDER IMPLICATIONS OF THE FINDINGS: Given a wide range of predictor data, machine learning algorithms can be leveraged to analyze epidemiologic data and predict the probability of conception with discrimination that exceeds earlier work.

STUDY FUNDING/COMPETING INTEREST(S): The research was partially supported by the U.S. National Science Foundation (under grants DMS-1664644, CNS-1645681 and IIS-1914792) and the National Institutes for Health (under grants R01 GM135930 and UL54 TR004130). In the last 3 years, L.A.W. has received in-kind donations for primary data collection in PRESTO from FertilityFriend.com, Kindara.com, Sandstone Diagnostics and Swiss Precision Diagnostics. L.A.W. also serves as a fibroid consultant to AbbVie, Inc. The other authors declare no competing interests.

TRIAL REGISTRATION NUMBER: N/A.

Key words: fertility / fecundability / pregnancy / prospective studies / predictive analytics / machine learning

Introduction

In North America, 10–15% of couples experience infertility, defined as the inability to conceive within 12 months of regular unprotected intercourse (Thoma *et al.*, 2013). In the USA, up to 12% of reproductive aged women and 9.4% of men aged 25–44 years used fertility treatments in 2006–2010 (Chandra *et al.*, 2013). The costs of these services exceed $5 billion in the USA annually (Macaluso *et al.*, 2010) and are expected to increase as couples delay childbearing. Developing better prognostic tools for couples trying to conceive could inform clinical care and mitigate potential costs. For women who are concerned about their fertility potential before they start trying to conceive, an accurate predictive model could facilitate decisions about how long to delay childbearing or how to prioritize other potentially modifiable factors.

Previous research has identified many individual risk factors for infertility and predictors of fecundability (i.e. the per-cycle probability of conception). Female age and BMI, as well as male BMI, have been identified as risk factors for infertility (Homan *et al.*, 2007; Best and Bhattacharya, 2015; Sundaram *et al.*, 2017; Wesselink *et al.*, 2017). In addition, female preconception exposures including alcohol consumption (Fan *et al.*, 2017); sleep quality (Willis *et al.*, 2019); cigarette smoking (Wesselink *et al.*, 2019); use of certain hormonal contraceptives (Yland *et al.*, 2020); dietary factors (Gaskins and Chavarro, 2018); depressive symptoms (Nillni *et al.*, 2016; Evans-Hoeker *et al.*, 2018); stress (Louis *et al.*, 2011; Lynch *et al.*, 2014; Akhter *et al.*, 2016; Wesselink *et al.*, 2018); and environmental exposures such as air pollution (Conforti *et al.*, 2018) and endocrine disrupting chemicals (Kahn *et al.*, 2021) are associated with reduced fecundability. Other male risk factors include exposure to environmental chemicals (Snijder *et al.*, 2012; Buck Louis *et al.*, 2016), cigarette smoking (Soares and Melo, 2008) and short sleep duration (Wise *et al.*, 2018). However, few studies have moved beyond individual risk factors to develop predictive models of pregnancy probability, and the predictive power of these models was modest (Eimers *et al.*, 1994; Collins *et al.*, 1995; Snick *et al.*, 1997; Hunault *et al.*, 2004, 2005; van der Steeg *et al.*, 2007; Coppus *et al.*, 2009).

In this study, we used supervised machine learning methods to predict the cumulative probability of pregnancy over 6 and 12 menstrual cycles and to predict fecundability (the per-cycle probability of conception) in an incident cohort study of pregnancy planners. We considered 163 potential predictors and applied several classification algorithms and variable selection procedures to identify the most accurate models and to evaluate the relative predictive strength of individual risk factors.

Materials and methods

Study population

Pregnancy Study Online (PRESTO) is a web-based preconception cohort study that examines the extent to which environmental and behavioral factors such as diet, exercise and medication use influence fertility and pregnancy outcomes (Wise *et al.*, 2015). The study began in 2013 and is ongoing. Eligible female participants are aged 21–45 years, residing in the USA or Canada, trying to conceive, and not using fertility treatments. We excluded participants with more than one menstrual cycle of pregnancy attempt time at enrollment because these women may have changed their behaviors in response to difficulties conceiving (Wise *et al.*, 2020). We analyzed data from couples who had not yet tried to conceive and those who had tried for one cycle at study entry together. This is consistent with a report by Joffe *et al.* (2005), which indicated that grouping couples with reports of 'zero' and 'one' cycle of pregnancy attempt time does not induce bias. This study included data from 4133 participants enrolled during 2013 through 2019.

Data collection

Female participants completed a baseline questionnaire at enrollment, on which they reported data on sociodemographic factors, behavioral factors, medical and reproductive history, and selected male partner characteristics. Ten days after enrollment, participants were invited to complete the diet history questionnaire II (DHQ II). The DHQ II was designed by the National Cancer Institute and the first version of the DHQ was validated against 24-h dietary recalls in a USA population (Subar *et al.*, 2001; Millen *et al.*, 2006). In validation studies, correlations between energy-adjusted, DHQ-reported food servings and 24-h recall-reported food servings ranged from 0.43 for other starchy vegetables to 0.84 for milk. Based on dietary factors reported via the DHQ II, we assessed overall diet quality using the Healthy Eating Index-2010 (HEI-2010) score (Guenther *et al.*, 2013). Participants completed bimonthly follow-up questionnaires for 12 months, or until reported pregnancy, cessation of pregnancy attempts, study withdrawal or loss to follow-up, whichever occurred first. Data on menstrual cycle dates, pregnancy attempts and pregnancy status were obtained via the baseline questionnaire and updated on each follow-up questionnaire. A complete list of the 163 variables included in this analysis is provided in Table I.

Table I **Complete list of variables included in analysis.**

Category	Variables included in preliminary analysis
Demographic and socioeconomic characteristics	Age, marital status, race,[1] ethnicity, region of residence, urbanization of residential area, year at study entry, highest level of education, parents' education level, household income, employment status, hours/week of work, shift work, night shift frequency in the past month.
Lifestyle, behavioral and wellness factors	Cigarette smoking (if so, number per day); total duration of smoking; history of smoking during pregnancy; use of e-cigarettes (if so, ml/day); frequency of marijuana use; exposure to second-hand smoke; alcohol intake; caffeine consumption; moderate physical activity; vigorous physical activity; sedentary activity; sleep duration; trouble sleeping; perceived stress scale score; major depression inventory score.
Dietary factors and use of supplements	Healthy Eating Index-2010 score; supplemental intake of vitamins A, B1, B2, B3, B4, B5, B6, B7, B12, C, D, E, K; beta-carotene; folic acid; iron; zinc; calcium; magnesium; selenium; omega-3 fatty acids; consumption of whole milk, 2% milk, 1% milk, skim milk, soy milk, other milk, fruit juice, bottled water, tap water, sugar-sweetened soda, diet soda, sugar-sweetened energy drinks, diet energy drinks; use of multivitamins or folic acid supplements.
Early life exposures and family history	Adopted; number of siblings; multiple gestation; born preterm; born with low birthweight; breastfed; delivered via cesarean section; mother's cigarette smoking during pregnancy; mother's age at participant's birth; mother's history of pregnancy complications, miscarriage.
Reproductive characteristics and disorders	Age at menarche; menstrual regularity; menstrual period characteristics (typical length,[2] number of flow days, flow amount, pain); received human papillomavirus vaccine; abnormal pap smear; ever diagnosed with a thyroid condition, fibroids, polycystic ovarian syndrome, endometriosis, a urinary tract infection, pelvic inflammatory disease, chlamydia, herpes, vaginosis, genital warts; recent use of medications for polycystic ovarian syndrome; gravidity; parity; history of cesarean section; years since last pregnancy; history of unplanned pregnancy; history of subfertility or infertility; history of infertility treatment; history of breastfeeding; number of lifetime sexual partners; doing something to improve pregnancy chances; intercourse frequency; using a fertility app; last method of contraception.
Physical characteristics, non-reproductive medical history and medication use	Body mass index; waist measure; Ferriman-Gallwey Hirsutism Score; handedness; number of primary care visits last year; high blood pressure; received influenza vaccine last year; ever diagnosed with migraines (if so, recent migraine frequency), asthma, hay fever, depression, anxiety, gastroesophageal reflux disease, diabetes; use of the following medications in the 4 weeks before baseline: pain medications, antibiotics, asthma medications, diabetes medications; use of psychotropic medications.
Environmental exposures (occupational and personal care product use)	Exposed regularly to agricultural pesticides; metal particulates or fumes; solvents, oil-based paints or cleaning compounds; high temperature environments; chemotherapeutic drugs; engine exhaust; chemicals for hair dyeing, straightening or curing; chemicals for manicure/pedicure. Use of chemical hair relaxer.
Male partner characteristics	Age, body mass index, education, cigarette smoking (if so, number per day), circumcision status.

[1] We conceptualized race as a social construct that serves as a rough proxy for exposure to interpersonal and structural racism.

[2] Menstrual cycle length and regularity were assessed via the following questions on the baseline questionnaire: (i) Did your period become regular on its own without the use of hormonal contraceptives such as the pill, patch, implants or injectables (regular in a way so you can usually predict about when the next period will start)? (ii) Within the past couple of years, has your menstrual period been regular? Please think about those times you were not using hormonal contraceptives. (iii) Thinking about the time(s) when you have not used hormonal contraceptives, what is your typical menstrual cycle length? That is, the number of days from the first day of one menstrual period to the first day of your next menstrual period.

Outcomes

We developed three models to predict (i) pregnancy in fewer than 12 menstrual cycles; (ii) pregnancy within 6 menstrual cycles; and (iii) the average probability of pregnancy per menstrual cycle. We chose these outcome measures to reflect clinically relevant definitions of infertility, subfertility and fecundability (Evers, 2002; Gnoth *et al.*, 2005). We defined the first outcome as fewer than 12 menstrual cycles, rather than ≤12 cycles, because participants who conceive in the twelfth cycle are unlikely to have the opportunity to identify and report their pregnancy before the end of the study period. For the first two models, we used a dataset with one observation per participant and excluded participants who were lost to follow-up before reaching a study endpoint (for the first model, N = 3195; for the second model, N = 3476). For the third model (fecundability), we included all participants under observation regardless of follow-up duration (N = 4133).

Pre-processing and statistical feature selection

We performed several data pre-processing steps to prepare the dataset for feature selection and to avoid model overfitting (Hawkins, 2004). First, we converted categorical variables into indicator variables and standardized each predictor by subtracting its mean and dividing by its SD. Next, for each pair of highly correlated variables (correlation coefficient >0.8), we removed the variable that had a lower correlation with the outcome to avoid issues of collinearity. We then performed statistical feature selection as follows: we evaluated the difference in means or proportions between participants with and without a pregnancy, using the chi-squared test (Cochran, 1952) for binary predictors and the Kolmogorov–Smirnov test for continuous predictors (Massey, 1951). We removed the variables that were not significantly associated with the outcome ($P > 0.05$).

Classification methods

We compared four supervised classification methods to develop predictive models for pregnancy (Hastie et al., 2009; Jiang et al., 2020). Supervised machine learning is an approach in which a dataset is randomly split into a training dataset and a testing dataset. Then, an algorithm (described in greater detail below) is applied to the training data to infer a function that maps a combination of inputs (i.e. predictors) to outputs (i.e. the outcome pregnancy). In a process called feature selection (or elimination), the predictive ability of the model is optimized by selecting variables to improve prediction of the outcome. The model with the final selected set of variables is then trained on the entire training set and its performance evaluated on the testing set.

For Models I (pregnancy in fewer than 12 menstrual cycles) and II (pregnancy within 6 menstrual cycles), we first fit logistic regression models with an added regularization term to penalize an overfit model (Friedman et al., 2010). Models derived using regularization are robust to the presence of outliers in the training data set (Chen et al., 2019; Chen and Paschalidis, 2020). We considered both an ℓ_1-norm (L1LR) and an ℓ_2-norm regularizer (L2LR) (Lee et al., 2006). The former is appropriate if we believe that few variables are predictive of the outcome (sparse model), whereas the latter is appropriate in cases where a dense model is more appropriate. Second, we used support vector machines (SVMs), which find a separating hyperplane in the variable space so that the data points from the two different classes reside on different sides of that hyperplane (Cortes and Vapnik, 1995). We considered both a standard linear SVM with an ℓ_2-norm regularizer (L2SVM) and a linear SVM with an ℓ_1-norm regularizer (L1SVM) designed to induce a sparse solution. Third, we used the Light Gradient Boosting Machine (LightGBM) algorithm, which is an ensemble tree-based model that uses a gradient boosting framework (Mason et al., 1999; Friedman, 2002). Fourth, we used artificial neural networks (ANNs), which attempt to organize data based on structures inspired by mammalian brain functioning (Ripley, 2007). We used Feed Forward Multilayer Perceptron Neural Networks (MLP), with at least three layers of nodes (an input layer, a hidden layer and an output layer) (Salcedo-Sanz, 2016). In a feed-forward ANN, information moves in one direction: from input to output. Because there are intermediate layers of information, the MLP algorithm can model complex non-linear relationships. We tuned several hyperparameters including the number of hidden neurons, the number of layers and the number of iterations. For training, we used a Rectified Linear Unit (ReLU) activation function for the hidden layer and applied the 'Adam' optimizer (Kingma and Ba, 2014). These algorithms were chosen because of their extensive usage and their performance superiority demonstrated in the literature (Brisimi et al., 2018; Hao et al., 2020; Wang et al., 2020).

We present results for full, sparse and parsimonious models. The full models (i.e. least parsimonious) contain all variables selected after statistical feature selection (eliminating variables with no statistically significant relationship with the outcome). The sparse models contain variables selected after both statistical feature selection and recursive feature elimination. Recursive feature elimination is a feature selection algorithm that ranks the predictors selected into the full model by importance and iteratively eliminates the least important variables, ultimately selecting a small set of variables that maximize the area under the receiver operating characteristic curve (AUC) in the testing dataset. The parsimonious models were generated by limiting recursive feature elimination to select a model with up to 15 variables. Specifically, we used L1LR to obtain weights associated with the coefficients of the model and eliminated the variable with the smallest absolute weight. We then performed L1LR to obtain a new model and repeated this process until the final model was selected. The final model maximizes a metric equal to the mean AUC minus the SD of the AUC in the testing dataset (described in more detail below). The parsimonious models are easier to implement and interpret relative to the full models, which have more variables but similar discrimination. To accommodate categorical variables that were recoded as indicator variables in the preprocessing phase, we selected a reference level for each categorical variable and forced every non-reference level to be included in a model if any other (non-reference) level of the categorical variable was selected.

For Model III (fecundability), we fit a discrete-time analog of the Cox proportional hazards model with cycle number as the time scale, allowing for delayed entry into the risk set (i.e. if a participant already had one cycle of pregnancy attempt at enrollment). Participants contributed at-risk cycles to the analysis from enrollment until reported pregnancy or a censoring event, which included initiation of fertility treatment, withdrawal from the study, cessation of pregnancy attempts, loss-to-follow-up or 12 cycles of pregnancy attempt, whichever occurred first. We present results for the full model after statistical feature selection, as described above, and for a parsimonious model. To derive the parsimonious model, we fit separate Cox models with each individual predictor and then sorted the variables based on each model's concordance index. The concordance index is similar to the AUC (described below) but accounts for event time and loss to follow-up (Schmid et al., 2016; Longato et al., 2020). We selected the top fifteen variables and forced non-selected levels of polytomous categorical variables into the final model, as described above.

Sensitivity analysis

We restricted our analyses to nulligravid women with no history of infertility to evaluate the robustness of our results in a population that was presumably naïve to their fertility status.

Performance metrics

For Models I and II, we primarily evaluated model performance using the AUC. The AUC, or C-statistic, quantifies model discrimination, such that a value of 0.5 indicates that discrimination is no better than random, while a value of 1 would indicate perfect prediction. The models were developed and evaluated as follows. First, we split the dataset into five random parts of equal size, where four parts constituted the training dataset, and the fifth part constituted the testing dataset. Second, we used the training dataset to tune the model hyperparameters via 5-fold cross-validation. In 5-fold cross-validation, the training dataset is split into five parts of equal size. A model is trained using four parts as training data, and the resulting model is validated on the fifth part. This procedure is repeated for each of the 5 folds, such that each part of the training dataset is used to validate the model trained on the other four parts of the training dataset. Third, we fit the model with the best cross-validation score (the highest AUC, obtained in Step 2) on the entire training dataset and evaluated its performance metrics on the testing dataset created in Step 1. Fourth, we repeated the first three steps (split the data into five random parts, tune the model hyperparameters with 5-fold cross-validation using the training dataset and evaluate model performance in the testing dataset) five times. Finally, we calculated the mean and SD of the model performance statistics across these five runs.

We also evaluated model performance using the weighted-F1 score. The F1-score is computed as the harmonic mean of positive predictive value and sensitivity, such that the highest value (1.0) indicates both perfect positive predictive value and sensitivity, and the lowest value (0) indicates that either the positive predictive value or the sensitivity is zero. To account for imbalance in the data (i.e. differences in the proportions of participants who did and did not conceive), we computed a weighted F1-score as the average of the F1-scores for participants with and without a pregnancy, weighted by the number of participants in each class. While the AUC is more easily interpretable, the weighted F1-score is more robust to data imbalances (Saito and Rehmsmeier, 2015). Finally, we present weighted-precision and weighted-recall metrics. Precision is equivalent to a positive predictive value, and recall is equivalent to sensitivity. To compute these metrics, we calculate precision and recall for each class (i.e. pregnant versus non-pregnant) and their average weighted by the number of true instances for each class.

For Model III, we evaluated performance using the concordance index, as described above.

All analyses were performed with Python statistical functions. Relevant programs can be accessed here: https://github.com/noc-lab/Predictive-models-of-pregnancy. This repository also contains detailed instructions that can be used by anyone to run the three primary models on their own data. Additional methodological information on how we addressed imbalance in the data and tuning of hyperparameters is provided in the Supplementary File S1.

Results

After excluding participants with incomplete follow-up for Models I and II, we analyzed data from 3195 and 3476 participants for Models I and II, respectively, and 16 876 cycles from 4133 participants for Model III. The study participants were aged 30 years on average and ranged in age from 21 to 44 years. Among the 3195 participants included in Model I, 2747 (86%) became pregnant in 12 menstrual cycles. Among the 3476 participants included in Model II, 2406 (69%) became pregnant within 6 menstrual cycles. The distributions of class (i.e. pregnant versus non-pregnant), overall and by number of menstrual cycles of attempt time at study entry, are presented in Supplementary Tables SI and SII. For each of the three models, the same 163 variables were considered for preprocessing (Table I). After statistical feature selection, 40 variables were selected into the full model predicting pregnancy in 12 menstrual cycles (Model I) and 41 variables were selected into the full model predicting pregnancy within 6 menstrual cycles (Model II). After recursive feature elimination, 30 and 25 variables were selected for the sparse Models I and II, respectively. The final parsimonious models included 14 and 15 variables for Models I and II, respectively. We present performance statistics for the parsimonious models in Table II. The AUC for Model I was 68–70% for all classification algorithms considered (SD: 0.8% to 1.9%). The AUCs for Model II were 65–66% (SD: 1.9% to 2.6%). The L2LR and L2SVM algorithms generally yielded the highest AUC. The weighted-F1 scores were similar across each algorithm, and no algorithm consistently yielded the highest score. The weighted-F1 scores obtained with the L2LR algorithm were 81.8 (SD: 1.0) for Model I and 67.5 (SD: 1.6) for Model II. The parsimonious models performed similarly to the full and sparse models (Supplementary Table SIII). The concordance index for Model III was 63.5% for the full model after statistical feature selection (24 variables) and 62.6% for the final parsimonious model. Supplementary Fig. S1 presents area under the precision-recall curves for Models I, II, IV and V.

Table II Performance metrics for the parsimonious models, PRESTO 2013–2019.

	Performance measure % (SD)							
	Model I				*Model II*			
Algorithm[1]	AUC	Weighted F1 score	Weighted precision	Weighted recall	AUC	Weighted F1 score	Weighted precision	Weighted recall
L2LR	70.2 (1.6)	81.8 (1.0)	80.8 (1.0)	83.3 (1.3)	66.1 (2.1)	67.5 (1.6)	67.2 (1.5)	69.5 (1.4)
L1LR	69.8 (1.8)	81.6 (0.6)	80.6 (0.8)	83.5 (1.1)	66.0 (1.9)	67.4 (1.6)	67.0 (1.6)	69.3 (1.5)
L1SVM	69.8 (1.9)	81.8 (0.8)	80.6 (0.8)	83.6 (0.8)	66.0 (1.9)	67.4 (1.2)	66.9 (1.3)	69.1 (1.3)
L2SVM	70.0 (1.6)	81.5 (1.1)	80.5 (1.2)	83.3 (1.3)	66.2 (2.1)	67.2 (1.0)	66.9 (1.1)	69.6 (0.9)
MLP	69.9 (0.8)	82.1 (0.9)	81.1 (1.2)	83.9 (1.3)	65.1 (2.1)	67.5 (1.5)	67.0 (1.5)	68.5 (1.7)
LightGBM	68.1 (1.4)	81.6 (0.8)	80.8 (0.9)	82.9 (1.2)	64.9 (2.6)	66.9 (1.3)	66.6 (1.4)	67.6 (1.1)

Model I predicts pregnancy in $<$12 menstrual cycles (N = 3195 participants). Model II predicts pregnancy in $<$7 menstrual cycles (N = 3476 participants). The parsimonious models contain variables selected after both statistical feature selection and recursive feature elimination, and limiting recursive feature elimination to select a model with up to 15 variables.

[1] L2LR, ℓ_2-penalized logistic regression; L1LR, ℓ_1-penalized logistic regression; L1SVM, support vector machine (SVM) with an ℓ_1-norm regularizer; L2SVM, SVM with an ℓ_2-norm regularizer; MLP, Feed Forward Multilayer Perceptron Neural Networks; LightGBM, Light Gradient Boosting Machine.

Table III Variables selected by the parsimonious Model I (predicting pregnancy in 12 cycles) using the L2LR algorithm, PRESTO 2013-2019, n = 3195 participants.

Variable	Standardized regression coefficient	Overall		Pregnant		Not pregnant	
		Frequency or mean	SD	Frequency or mean	SD	Frequency or mean	SD
Menstrual cycle length (days)	0.27	29.6	4.0	29.7	4.1	28.7	3.0
Female age at baseline (years)	−0.26	29.8	3.8	29.7	3.6	30.6	4.5
Urbanization of residential area: rural (ref = urbanized area)	0.25	4%	20%	5%	21%	1%	12%
Previously tried to conceive for ≥12 months: 'yes' (ref = 'no, tried for < 12 months')	−0.24	5%	21%	4%	19%	10%	30%
One menstrual cycle of attempt time at study entry (ref = 0)	−0.23	58%	49%	56%	50%	68%	47%
Daily use of multivitamins/folic acid (yes/no)	0.22	84%	37%	85%	35%	73%	44%
Last method of contraception: hormonal IUD (yes/no)[1]	0.19	12%	32%	12%	33%	7%	25%
Female BMI (kg/m^2)	−0.19	26.6	6.5	26.3	6.2	28.4	7.8
Ever breastfed an infant (yes/no)	0.18	31%	46%	32%	47%	22%	41%
Ever been pregnant (yes/no)	0.15	50%	50%	52%	50%	42%	49%
Female education (years)	0.14	16.0	1.2	16.1	1.2	15.8	1.4
Received influenza vaccine in the past year (yes/no)	0.13	53%	50%	54%	50%	44%	50%
Stress (Perceived Stress Scale score)	−0.12	15.5	5.8	15.3	5.8	16.3	5.6
Total number of pregnancies	0.12	1.0	1.4	1.0	1.4	0.8	1.4
Variables forced into the model[2]							
Urbanization of residential area: Canada (ref = urbanized area)	0.01	18%	39%	18%	39%	19%	39%
Urbanization of residential area: urban cluster (ref = urbanized area)	−0.01	8%	27%	8%	27%	8%	27%
Previously tried to conceive for ≥12 months: 'no, never tried before' (ref = 'no, tried for < 12 months')	−0.01	42%	49%	41%	49%	48%	50%

Variables are presented in order of the magnitude of the standardized regression coefficients.

[1] Last methods of contraception were not mutually exclusive and were coded as indicator variables with no reference category. Natural methods included withdrawal, avoiding sex when fertile, calendar methods and monitoring cervical mucus or basal body temperature.

[2] For all models, we selected a reference group for each categorical variable that was recoded as indicator variables in the preprocessing phase and forced every non-reference level to be included in the model if any level of the categorical variable was selected. These variables are listed in addition to the variables selected by the parsimonious model.

In order of decreasing magnitude of the regression coefficients (i.e. strongest to weakest predictor), the variables selected into the parsimonious Model I that were positively associated with pregnancy were menstrual cycle length, living in a rural region, daily use of multivitamins or folic acid, using the hormonal intrauterine device (IUD) as one's most recent method of contraception, having previously breastfed an infant, having ever been pregnant, female education, recent influenza vaccination and gravidity (total number of pregnancies) (Table III). The variables that were inversely associated with pregnancy were female age, having a history of infertility, having completed one menstrual cycle of pregnancy attempt time at study entry (versus zero), female BMI and stress. The distributions of these variables overall, and by pregnancy status, are presented in Table III. Results for parsimonious Models II and III are presented in Tables IV and V, respectively. The variables selected into the parsimonious Model II that were positively associated with pregnancy were daily use of multivitamins or folic acid, having previously breastfed an infant, HEI-2010 score, having a previous unplanned pregnancy, trying to improve one's chances of pregnancy (e.g. charting cycles, ovulation or cervical mucus testing, timing intercourse to the fertile window), and time since the participant's last pregnancy (<1 year). The variables that were inversely associated with pregnancy were female BMI, having a history of infertility, male age, non-use of a fertility app, male BMI, having completed one menstrual cycle of pregnancy attempt time at study entry (versus zero), male partner smoking, female age and having a history of subfertility or infertility. Results were generally similar for Model III. Variables selected into Model III but neither Models I nor II included intercourse frequency and menstrual cycle regularity. The model coefficients and their 95% CIs are presented graphically in Supplementary Fig. S2.

Among 1957 nulligravid women without a history of infertility, we developed models predicting pregnancy in fewer than 12 menstrual

Table IV Variables selected by the parsimonious Model II (predicting pregnancy within 6 cycles) using the L2LR algorithm, PRESTO 2013–2019, n = 3476 participants.

		Overall		Pregnant		Not pregnant	
Variable	**Standardized regression coefficient**	**Frequency or mean**	**SD**	**Frequency or mean**	**SD**	**Frequency or mean**	**SD**
Female BMI (kg/m^2)	−0.11	26.8	6.7	26.1	6.1	28.3	7.7
Daily use of multivitamins/folic acid (yes/no)	0.08	84%	37%	86%	35%	78%	42%
Ever breastfed an infant (yes/no)	0.08	30%	46%	33%	47%	24%	43%
Previously tried to conceive for ≥12 months: 'yes' (ref = 'no, tried for < 12 months')	−0.08	5%	22%	4%	18%	8%	28%
Healthy Eating Index-2010 score (HEI-2010 score)	0.07	66.0	11.2	66.8	10.9	64.3	11.6
Male age (years)	−0.07	31.8	5.0	31.5	4.6	32.4	5.8
Use of fertility app: 'no, but I plan to' (ref = 'yes')	−0.07	8%	27%	6%	24%	11%	31%
History of unplanned pregnancy (yes/no)	0.07	34%	47%	37%	48%	27%	44%
Male BMI (kg/m^2)	−0.07	27.7	5.3	27.3	5.1	28.5	5.6
One menstrual cycle of attempt time at study entry (ref = 0)	−0.06	58%	49%	55%	50%	65%	48%
Male cigarette smoking: 'yes, on a regular basis' (ref = 'no')	−0.06	8%	27%	6%	24%	12%	32%
Female age at baseline (years)	−0.06	29.8	3.8	29.6	3.6	30.3	4.2
Trying to improve chances of pregnancy (yes/no)	0.05	70%	46%	72%	45%	64%	48%
Time since last pregnancy: <1 year (ref = nulliparous)	0.05	22%	41%	24%	42%	18%	38%
History of subfertility or infertility (yes/no)	−0.05	10%	30%	9%	28%	13%	34%
Variables forced into the model[1]							
Previously tried to conceive for ≥12 months: 'no, never tried before' (ref = 'no, tried for < 12 months')	−0.05	42%	49%	40%	49%	46%	50%
Time since last pregnancy: 1-2 years (ref = nulliparous)	0.04	17%	38%	19%	39%	14%	35%
Male cigarette smoking: 'yes, occasionally' (ref = 'no')	−0.02	4%	20%	4%	19%	5%	22%
Time since last pregnancy: ≥5 years (ref = nulliparous)	−0.02	6%	24%	5%	22%	8%	27%
Use of fertility app: 'no' (ref = 'yes')	−0.02	23%	42%	22%	41%	26%	44%
Time since last pregnancy: 3–4 years (ref = nulliparous)	0.02	4%	21%	5%	21%	4%	19%

Variables are presented in order of the magnitude of the standardized regression coefficients.

[1] For all models, we selected a reference group for each categorical variable that was recoded as indicator variables in the preprocessing phase and forced every non-reference level to be included in the model if any level of the categorical variable was selected. These variables are listed in addition to the variables selected by the parsimonious model.

cycles (Model IV), predicting pregnancy within 6 menstrual cycles (Model V) and predicting fecundability (Model VI). We analyzed data from 1571, 1722 and 1957 participants for Models IV, V and VI, respectively. The performance of these models was slightly lower than the analogous models in the full cohort. The performance statistics for the full and sparse Models IV and V are presented in Supplementary Table SIV. Using statistical feature selection, 16 and 12 variables were selected into the full models for Model IV and V, respectively. After recursive feature elimination, 5 and 9 variables were selected for the sparse Models IV and V, respectively. Because fewer than 15 features were selected by each of the sparse models, the sparse models were equivalent to the parsimonious models. Consistent with the main analysis, the L2LR algorithm performed best for the sparse models. The AUCs were 69.5% (SD: 1.4) for Model IV and 65.6% (SD: 2.9) for Model V. The concordance index for Model VI was 60.2%. Variables selected by these models that were positively associated with pregnancy included menstrual cycle length, using a hormonal IUD as one's most recent method of contraception, intercourse frequency, trying to improve one's chances of pregnancy, use of vitamin E supplements and HEI-2010 score. Variables inversely associated with the probability of pregnancy included having completed one menstrual cycle of pregnancy attempt time at study entry (versus zero), female age, male and female BMI, menstrual cycle irregularity, non-use of a fertility app, stress, depressive symptoms, history of vaginosis, male partner smoking, milk consumption and sleep characteristics (Supplementary Tables V, VI, VII). Occupational exposures including exposure to metal particulates or fumes and exposure to high temperature environments were also selected to Model VI, but with very small coefficients (Supplementary Table VII).

Discussion

In this prospective cohort study of 4133 North American pregnancy planners, we applied several supervised learning methods to predict

Table V **Variables selected by the parsimonious Model III (fecundability), PRESTO 2013–2019, n = 4133 participants.**

Variable	Hazard ratio	95% confidence interval
Previously tried to conceive for ≥12 months: 'yes' (ref = 'no, tried for < 12 months')	0.85	(0.80, 0.90)
Ever breastfed an infant (yes/no)	1.16	(1.09, 1.23)
Female BMI (kg/m^2)	0.89	(0.84, 0.93)
Time since last pregnancy: 1–2 years (ref = nulliparous)	1.12	(1.04, 1.21)
Female age at baseline (years)	0.90	(0.85, 0.95)
Trying to improve chances of pregnancy (yes/no)	1.11	(1.06, 1.15)
Female education (years)	1.09	(1.03, 1.15)
Intercourse frequency (times/week)	1.08	(1.03, 1.12)
Male BMI (kg/m^2)	0.93	(0.89, 0.98)
Male cigarette smoking: 'yes, on a regular basis' (ref = 'no')	0.93	(0.89, 0.98)
Has menstrual cycle been regular without hormonal contraception in past 2 years? 'no, irregular' (ref = 'yes, regular')	0.94	(0.89, 0.99)
Daily use of multivitamins/folic acid (yes/no)	1.06	(1.01, 1.11)
Did your period become regular on its own? 'no, irregular' (ref = 'yes, regular')	0.96	(0.92, 1.01)
Male age (years)	0.96	(0.91, 1.01)
Tap water consumption (drinks/week)	1.04	(1.01, 1.07)
Variables forced into the model[1]		
Time since last pregnancy: <1 year (ref = nulliparous)	1.37	(1.14, 1.64)
Time since last pregnancy: 3–4 years (ref = nulliparous)	1.32	(1.01, 1.71)
Male cigarette smoking: 'yes, occasionally' (ref = 'no')	0.87	(0.70, 1.08)
Has menstrual cycle been regular without hormonal contraception in past 2 years? 'unknown, was using hormonal contraception' (ref = 'yes, regular')	1.03	(0.93, 1.14)
Did your period become regular on its own? 'unknown, was using hormonal contraception' (ref = 'yes, regular')	1.02	(0.89, 1.17)
Time since last pregnancy: ≥5 years (ref = nulliparous)	1.01	(0.79, 1.29)

Variables are presented in order of the magnitude of the regression coefficients (i.e. the natural logarithm of the hazard ratio).

[1] For all models, we selected a reference group for each categorical variable that was recoded as indicator variables in the preprocessing phase and forced every non-reference level to be included in the model if any level of the categorical variable was selected. These variables are listed in addition to the variables selected by the parsimonious model.

the probability of pregnancy within three time periods: 12 menstrual cycles, 6 menstrual cycles and on a per-cycle basis. The L2LR and L2SVM algorithms generally yielded the highest AUC, particularly for the parsimonious models. For all models, discrimination (AUC) was close to 70%. The highest AUCs were 71.2% for Model I, 67.1% for Model II, 69.5% for Model IV and 65.6% for Model V. These findings demonstrate that it is possible to develop predictive models with reasonable discrimination using self-reported data in the absence of more detailed medical information such as laboratory or imaging tests.

The discrimination of our models is greater than previously published predictive models for pregnancy independent of fertility treatment, which yielded AUC's between 59% and 64% (Coppus et al., 2009). For example, Eimers et al. (1994) developed a predictive model for pregnancy among 996 couples consulting for infertility care in the Netherlands between 1974 and 1984. The investigators collected data on patient medical history, laboratory tests including semen analysis and postcoital tests (i.e. an examination of the interaction between sperm and the cervical mucus after intercourse), and a gynecologic physical examination. They used forward stepwise Cox regression to produce a model including female age, duration of infertility, primary versus secondary infertility, history of infertility in the male partner's family, sperm motility and the postcoital test results. Similar studies were conducted by Collins et al. (1995), using data from 1061 couples seeking infertility care at eleven Canadian University hospitals, and Snick et al. (1997), using data from 402 couples seeking infertility care at a Dutch general hospital. Hunault et al. (2004) pooled the data from the Eimers, Collins and Snick studies to evaluate the accuracy of these models and to develop two new synthesis models. The synthesis models included female age, duration of subfertility, sperm motility, whether the couple had been referred for infertility care by a general physician or a gynecologist, and the results of a postcoital test. These models were externally validated and found to have AUCs of 59–63% (Hunault et al., 2005; van der Steeg et al., 2007).

Although previous studies predicted the probability of pregnancy independent of fertility treatment, they were exclusively conducted in populations with subfertility using little or no data on lifestyle, environmental and sociodemographic factors (Eimers *et al.*, 1994; Collins *et al.*, 1995; Snick *et al.*, 1997; Hunault *et al.*, 2004, 2005; van der Steeg *et al.*, 2007; Coppus *et al.*, 2009). Our study may be more generalizable to couples across the fertility spectrum, because we included couples with a wide range of reproductive potential. In addition, we considered a range of potential predictors that may be more easily modified than clinical markers such as semen quality or hormone levels. For example, fertility app use, use of multivitamins or folic acid supplements and trying to improve one's chances of pregnancy (e.g. charting cycles, ovulation or cervical mucus testing, timing intercourse to the fertile window) are relatively modifiable behaviors. Lifestyle interventions can also be undertaken to modify individual-level behaviors that may increase a couple's chance of conception, such as promoting a healthy BMI, improving diet and reducing stress. However, many of these behaviors are determined by broader environmental and systemic drivers and thus may be best addressed through macro-level policy interventions that address upstream determinants (e.g. regulation of food supply and marketing). A causal analysis of each risk factor would be worthwhile for future and more targeted work. In this study, there were some variables that appeared to be particularly important predictors of pregnancy. These included female age and BMI, history of infertility, the number of menstrual cycles of pregnancy attempt time at study entry, having previously breastfed an infant and use of multivitamins or folic acid supplements. These findings are generally consistent with previous studies on individual risk factors for infertility that were conducted in other populations (Jensen *et al.*, 1999; Homan *et al.*, 2007; Wise *et al.*, 2011; Cueto *et al.*, 2016). However, having previously breastfed an infant, which was associated with an increased probability of pregnancy in this study, has not been previously studied as a predictor of fecundability. This may reflect underlying fertility, prolonged effects of hormonal changes during breastfeeding or higher socioeconomic status among women who breastfeed their infants (Jones *et al.*, 2011; Odar Stough *et al.*, 2019).

In this study, we developed an additional set of predictive models among nulligravid women with no history of infertility who had been trying to conceive for no more than one menstrual cycle of attempt time at enrollment. The performance of these models was slightly decreased compared with the main analyses. This is likely because having a history of infertility is a strong predictor of future fecundability, and therefore restricting the analytic sample by this variable would limit the predictive ability of the model. This was most obvious in Model V, which predicted pregnancy within six menstrual cycles. In these restricted analyses, the most important predictors of pregnancy across all models were the number of menstrual cycles of pregnancy attempt time at study entry, and female age.

Study limitations include potential misclassification of the predictor variables, given that all data were based on self-reporting. There is limited research on the impact of measurement error on machine learning prediction models (van Doorn *et al.*, 2017; Jiang *et al.*, 2021), and it is unclear how misclassification of the predictors influenced our study results in terms of accuracy and variable selection. There was also the potential for misspecification of the functional form of the predictor variables, which could have influenced the variable selection process. In addition, there may have been some misclassification of our estimate of time to pregnancy, which relied on self-reported menstrual cycle length and date of the last menstrual period. Given the prospective design of the study, such misclassification is likely to be non-differential with respect to the outcome. Bias may also have been introduced if the length of follow-up varied by the predictors under study, as Models I and II did not account for varying lengths of follow-up. However, results were generally consistent with Model III, which accounted for varying lengths of follow-up. Another potential limitation is our lack of inclusion of important predictors of pregnancy, such as hormone levels, which may have reduced the predictive ability of our models. Other potentially important predictors that we did not measure include environmental exposures (Conforti *et al.*, 2018; Hipwell *et al.*, 2019; Kahn *et al.*, 2021), early life adversity (Harville and Boynton-Jarrett, 2013; Jacobs *et al.*, 2015), occupational stress (Barzilai-Pesach *et al.*, 2006; Valsamakis *et al.*, 2019), experiences of discrimination (Krieger, 2000), social disadvantage, neighborhood characteristics (Williams and Collins, 2001) and multigenerational exposures (Eskenazi *et al.*, 2021; Wesselink, 2021). In addition, we lacked comprehensive data on male exposures, which contribute to up to 50% of all subfertility among couples (Irvine, 1998). However, we collected data on several important male characteristics on the female baseline questionnaire, including male age, BMI, education and smoking status. Overall, we considered a diverse range of 163 potential predictors, which is substantially greater than previous studies in this area (Eimers *et al.*, 1994; Collins *et al.*, 1995; Snick *et al.*, 1997; Hunault *et al.*, 2004, 2005; van der Steeg *et al.*, 2007; Coppus *et al.*, 2009). It should be noted that the effect estimates in these models lack causal interpretation, as variables were selected into the final models based on their predictive power, rather than the hypothesized causal structures of the data. Identifying causes of infertility was beyond the scope of this study. Also beyond the scope of this study was the development of models within clinically relevant subgroups (e.g. age >40 years or infertility-related conditions). Finally, though we validated the models using split sample replication techniques, we were unable to conduct an external validation study.

Conclusions

In this large prospective cohort, we used machine learning algorithms to develop predictive models of pregnancy, using three distinct, clinically relevant definitions of infertility, subfertility and fecundability. Comparing results across the three outcomes facilitates robust triangulation of fertility potential; the relative utility of each outcome may depend on a couple's preferences and risk profile. Our methods can predict pregnancy with discrimination as high as 71.2% by properly weighing a small set of predictive variables that include lifestyle and reproductive characteristics. Overall, the most consistent predictors of the probability of conception were female age, female BMI, male age, male BMI, history of infertility, history of breastfeeding, time since the participant's last pregnancy, daily use of multivitamins or folic acid, trying to improve one's chances of pregnancy (e.g. charting cycles, ovulation or cervical mucus testing, timing intercourse to the fertile window), male partner smoking and female education. Among nulligravid women without a history of infertility, the most important predictors were female age, female BMI, male BMI, use of a fertility app and perceived stress. These findings are particularly relevant for couples planning a pregnancy and clinicians providing preconception care to women who are discontinuing contraception in order to conceive. If these models are successfully validated in external populations, they could potentially be implemented as a counseling tool.

Supplementary data

Supplementary data are available at *Human Reproduction* online.

Data availability

The data underlying this article cannot be shared publicly, as PRESTO participants did not provide informed consent to share their data with external entities. The authors have shared their analytic code, along with detailed instructions for using the scripts, at the following location: https://github.com/noc-lab/Predictive-models-of-pregnancy.

Acknowledgments

We acknowledge the contributions of PRESTO participants and staff.

Authors' roles

J.J.Y. was responsible for results interpretation, manuscript writing, revision and finalization. T.W. was responsible for formulation of the study hypotheses and study design, statistical analyses, results interpretation, manuscript writing, revision and finalization. Z.Z. was responsible for statistical analyses, results interpretation, manuscript revision and finalization. S.K.W. was responsible for formulation of the study hypotheses, results interpretation, manuscript revision and finalization. T.R.W. was responsible for data management, manuscript revision and finalization. A.K.W. was responsible for formulation of the study hypotheses, results interpretation, manuscript revision and finalization. T.J. was responsible for results interpretation, manuscript revision and finalization. E.E.H., L.A.W. and I.C.P. were responsible for formulation of the study hypotheses and study design, results interpretation, manuscript writing, revision and finalization.

Funding

The research was partially supported by the U.S. National Science Foundation under grants IIS-1914792, DMS-1664644 and CNS-1645681 and by the National Institutes of Health under grants R01 GM135930, R21 HD072326, R01 HD086742 and grant 1UL1TR001430 to the Clinical & Translational Science Institute at Boston University.

Conflict of interest

L.A.W. received in-kind donations from Sandstone Diagnostics, Swiss Precision Diagnostics, FertilityFriend.com and Kindara.com for primary data collection in PRESTO. L.A.W. serves as a fibroid consultant for AbbVie Inc. The other authors have nothing to disclose.

References

Akhter S, Marcus M, Kerber RA, Kong M, Taylor KC. The impact of periconceptional maternal stress on fecundability. *Ann Epidemiol* 2016;**26**:710–716.e7.

Barzilai-Pesach V, Sheiner EK, Sheiner E, Potashnik G, Shoham-Vardi I. The effect of women's occupational psychologic stress on outcome of fertility treatments. *J Occup Environ Med* 2006;**48**:56–62.

Best D, Bhattacharya S. Obesity and fertility. *Horm Mol Biol Clin Investig* 2015;**24**:5–10.

Brisimi TS, Xu T, Wang T, Dai W, Adams WG, Paschalidis IC. Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach. *Proc IEEE Inst Electr Electron Eng* 2018;**106**:690–707.

Buck Louis GM, Barr DB, Kannan K, Chen Z, Kim S, Sundaram R. Paternal exposures to environmental chemicals and time-to-pregnancy: overview of results from the life study. *Andrology* 2016; **4**:639–647.

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, 1 p following 19.

Chen R, Paschalidis IC. "Distributionally Robust Learning", Foundations and Trends® in Optimization. 2020;**4**:1–243.

Chen R, Paschalidis IC, Hatabu H, Valtchinov VI, Siegelman J. Detection of unwarranted CT radiation exposure from patient and imaging protocol meta-data using regularized regression. *Eur J Radiol Open* 2019;**6**:206–211.

Cochran WG. The chi2 test of goodness of fit. *Ann Math Stat* 1952; **23**:315–345.

Collins JA, Burrows EA, Wilan AR. The prognosis for live birth among untreated infertile couples. *Fertil Steril* 1995;**64**:22–28.

Conforti A, Mascia M, Cioffi G, De Angelis C, Coppola G, De Rosa P, Pivonello R, Alviggi C, De Placido G. Air pollution and female fertility: a systematic review of literature. *Reprod Biol Endocrinol* 2018;**16**:117.

Coppus SF, van der Veen F, Opmeer BC, Mol BW, Bossuyt PM. Evaluating prediction models in reproductive medicine. *Hum Reprod* 2009;**24**:1774–1778.

Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**: 273–297.

Cueto HT, Riis AH, Hatch EE, Wise LA, Rothman KJ, Sørensen HT, Mikkelsen EM. Folic acid supplementation and fecundability: a Danish prospective cohort study. *Eur J Clin Nutr* 2016;**70**:66–71.

Eimers JM, Te Velde ER, Gerritse R, Vogelzang ET, Looman CW, Habbema JD. The prediction of the chance to conceive in subfertile couples. *Fertil Steril* 1994;**61**:44–52.

Eskenazi B, Ames J, Rauch S, Signorini S, Brambilla P, Mocarelli P, Siracusa C, Holland N, Warner M. Dioxin exposure associated with fecundability and infertility in mothers and daughters of Seveso, Italy. *Hum Reprod* 2021;**36**:794–807.

Evans-Hoeker EA, Eisenberg E, Diamond MP, Legro RS, Alvero R, Coutifaris C, Casson PR, Christman GM, Hansen KR, Zhang H *et al.*; Reproductive Medicine Network. Major depression, antidepressant use, and male and female fertility. *Fertil Steril* 2018;**109**:879–887.

Evers JL. Female subfertility. *Lancet* 2002;**360**:151–159.

Fan D, Liu L, Xia Q, Wang W, Wu S, Tian G, Liu Y, Ni J, Wu S, Guo X *et al.* Female alcohol consumption and fecundability: a systematic review and dose-response meta-analysis. *Sci Rep* 2017; **7**:13815.

Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010;**33**: 1–22.

Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal* 2002;**38**:367–378.

Gaskins AJ, Chavarro JE. Diet and fertility: a review. *Am J Obstet Gynecol* 2018;**218**:379–389.

Gnoth C, Godehardt E, Frank-Herrmann P, Friol K, Tigges J, Freundl G. Definition and prevalence of subfertility and infertility. *Hum Reprod* 2005;**20**:1144–1147.

Guenther PM, Casavale KO, Reedy J, Kirkpatrick SI, Hiza HAB, Kuczynski KJ, Kahle LL, Krebs-Smith SM. Update of the healthy eating index: Hei-2010. *J Acad Nutr Diet* 2013;**113**:569–580.

Hao B, Sotudian S, Wang T, Xu T, Hu Y, Gaitanidis A, Breen K, Velmahos GC, Paschalidis IC. Early prediction of level-of-care requirements in patients with covid-19. *Elife* 2020;**9**:e60519.

Harville EW, Boynton-Jarrett R. Childhood social hardships and fertility: a prospective cohort study. *Ann Epidemiol* 2013;**23**:784–790.

Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer Science & Business Media, 2009.

Hawkins DM. The problem of overfitting. *J Chem Inf Comput Sci* 2004;**44**:1–12.

Hipwell AE, Kahn LG, Factor-Litvak P, Porucznik CA, Siegel EL, Fichorova RN, Hamman RF, Klein-Fedyshin M, Harley KG; program orators for Environmental influences on Child Health Outcomes. Exposure to non-persistent chemicals in consumer products and fecundability: a systematic review. *Hum Reprod Update* 2019;**25**:51–71.

Homan GF, Davies M, Norman R. The impact of lifestyle factors on reproductive performance in the general population and those undergoing infertility treatment: a review. *Hum Reprod Update* 2007; **13**:209–223.

Hunault CC, Habbema JD, Eijkemans MJ, Collins JA, Evers JL, Te Velde ER. Two new prediction rules for spontaneous pregnancy leading to live birth among subfertile couples, based on the synthesis of three previous models. *Hum Reprod* 2004;**19**:2019–2026.

Hunault CC, Laven JS, van Rooij IA, Eijkemans MJ, Te Velde ER, Habbema JD. Prospective validation of two models predicting pregnancy leading to live birth among untreated subfertile couples. *Hum Reprod* 2005;**20**:1636–1641.

Irvine DS. Epidemiology and aetiology of male infertility. *Hum Reprod* 1998;**13**:33–44.

Jacobs MB, Boynton-Jarrett RD, Harville EW. Adverse childhood event experiences, fertility difficulties and menstrual cycle characteristics. *J Psychosom Obstet Gynaecol* 2015;**36**:46–57.

Jensen TK, Scheike T, Keiding N, Schaumburg I, Grandjean P. Fecundability in relation to body mass and menstrual cycle patterns. *Epidemiology* 1999;**10**:422–428.

Jiang T, Gradus JL, Lash TL, Fox MP. Addressing measurement error in random forests using quantitative bias analysis. *Am J Epidemiol* 2021;**190**:1830–1840.

Jiang T, Gradus JL, Rosellini AJ. Supervised machine learning: a brief primer. *Behav Ther* 2020;**51**:675–687.

Joffe M, Key J, Best N, Keiding N, Scheike T, Jensen TK. Studying time to pregnancy by use of a retrospective design. *Am J Epidemiol* 2005;**162**:115–124.

Jones JR, Kogan MD, Singh GK, Dee DL, Grummer-Strawn LM. Factors associated with exclusive breastfeeding in the United States. *Pediatrics* 2011;**128**:1117–1125.

Kahn LG, Harley KG, Siegel EL, Zhu Y, Factor-Litvak P, Porucznik CA, Klein-Fedyshin M, Hipwell AE; program orators for Environmental Influences on Child Health Outcomes Program. Persistent organic pollutants and couple fecundability: a systematic review. *Hum Reprod Update* 2021;**27**:339–366.

Kingma DP, Ba J. Adam: a method for stochastic optimization. In: *3rd International Conference for Learning Representations*, San Diego, *arXiv preprint arXiv:1412.6980* 2014.

Krieger N. *Discrimination and Health Social Epidemiology*. Oxford University Press, 2000, 36–75.

Lee S-I, Lee H, Abbeel P, Ng AY. Efficient L1 regularized logistic regression. In: *21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference*, Boston, Massachusetts, USA, 2006, pp.401–408.

Longato E, Vettoretti M, Di Camillo B. A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. *J Biomed Inform* 2020;**108**:103496.

Louis GM, Lum KJ, Sundaram R, Chen Z, Kim S, Lynch CD, Schisterman EF, Pyper C. Stress reduces conception probabilities across the fertile window: evidence in support of relaxation. *Fertil Steril* 2011;**95**:2184–2189.

Lynch CD, Sundaram R, Maisog JM, Sweeney AM, Buck Louis GM. Preconception stress increases the risk of infertility: results from a couple-based prospective cohort study—the life study. *Hum Reprod* 2014;**29**:1067–1075.

Macaluso M, Wright-Schnapp TJ, Chandra A, Johnson R, Satterwhite CL, Pulver A, Berman SM, Wang RY, Farr SL, Pollack LA. A public health focus on infertility prevention, detection, and management. *Fertil Steril* 2010;**93**:16.e1–10.

Mason L, Baxter J, Bartlett P, Frean M. Boosting algorithms as gradient descent in function space. In: *Proceedings of the 12th International Conference on Neural Information Processing Systems*, Denver, CO, 1999, pp.512–518. MIT press, Cambridge, MA, USA.

Massey FJ. The Kolmogorov-Smirnov test for goodness of fit. *J Am Stat Assoc* 1951;**46**:68–78.

Millen AE, Midthune D, Thompson FE, Kipnis V, Subar AF. The National Cancer Institute Diet History Questionnaire: validation of pyramid food servings. *Am J Epidemiol* 2006;**163**:279–288.

Nillni YI, Wesselink AK, Gradus JL, Hatch EE, Rothman KJ, Mikkelsen EM, Wise LA. Depression, anxiety, and psychotropic medication use and fecundability. *Am J Obstet Gynecol* 2016;**215**: 453.e1–8.

Odar Stough C, Khalsa AS, Nabors LA, Merianos AL, Peugh J. Predictors of exclusive breastfeeding for 6 months in a national sample of us children. *Am J Health Promot* 2019;**33**:48–56.

Ripley BD. *Pattern Recognition and Neural Networks*. Cambridge, UK: Cambridge University Press, 2007.

Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;**10**:e0118432.

Salcedo-Sanz S. Modern meta-heuristics based on nonlinear physics processes: a review of models and design procedures. *Phys Rep* 2016;**655**:1–70.

Schmid M, Wright MN, Ziegler A. On the use of Harrell's C for clinical risk prediction via random survival forests. *Expert Syst Appl* 2016;**63**:450–459.

Snick HK, Snick TS, Evers JL, Collins JA. The spontaneous pregnancy prognosis in untreated subfertile couples: the Walcheren primary care study. *Hum Reprod* 1997;**12**:1582–1588.

Snijder CA, Te Velde E, Roeleveld N, Burdorf A. Occupational exposure to chemical substances and time to pregnancy: a systematic review. *Hum Reprod Update* 2012;**18**:284–300.

Soares SR, Melo MA. Cigarette smoking and reproductive function. *Curr Opin Obstet Gynecol* 2008;**20**:281–291.

Subar AF, Thompson FE, Kipnis V, Midthune D, Hurwitz P, McNutt S, McIntosh A, Rosenfeld S. Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: the Eating at America's Table Study. *Am J Epidemiol* 2001; **154**:1089–1099.

Sundaram R, Mumford SL, Buck Louis GM. Couples' body composition and time-to-pregnancy. *Hum Reprod* 2017;**32**:662–668.

Thoma ME, McLain AC, Louis JF, King RB, Trumble AC, Sundaram R, Louis GMB. Prevalence of infertility in the United States as estimated by the current duration approach and a traditional constructed approach. *Fertil Steril* 2013;**99**:1324–1331.e1.

Valsamakis G, Chrousos G, Mastorakos G. Stress, female reproduction and pregnancy. *Psychoneuroendocrinology* 2019;**100**: 48–57.

van der Steeg JW, Steures P, Eijkemans MJ, Habbema JD, Hompes PG, Broekmans FJ, van Dessel HJ, Bossuyt PM, van der Veen F, Mol BW; CECERM study group (Collaborative Effort for Clinical Evaluation in Reproductive Medicine). Pregnancy is predictable: a large-scale prospective external validation of the prediction of spontaneous pregnancy in subfertile couples. *Hum Reprod* 2007; **22**:536–542.

van Doorn S, Brakenhoff TB, Moons KGM, Rutten FH, Hoes AW, Groenwold RHH, Geersing GJ. The effects of misclassification in routine healthcare databases on the accuracy of prognostic prediction models: a case study of the CHA_2DS_2-VASc score in atrial fibrillation. *Diagn Progn Res* 2017;**1**:18.

Wang T, Paschalidis A, Liu Q, Liu Y, Yuan Y, Paschalidis IC. Predictive models of mortality for hospitalized patients with COVID-19: retrospective cohort study. *JMIR Med Inform* 2020;**8**: e21788.

Wesselink AK. Multigenerational effects of environmental exposures. *Hum Reprod* 2021;**36**:539–542.

Wesselink AK, Hatch EE, Rothman KJ, Mikkelsen EM, Aschengrau A, Wise LA. Prospective study of cigarette smoking and fecundability. *Hum Reprod* 2019;**34**:558–567.

Wesselink AK, Hatch EE, Rothman KJ, Weuve JL, Aschengrau A, Song RJ, Wise LA. Perceived stress and fecundability: a preconception cohort study of North American couples. *Am J Epidemiol* 2018;**187**:2662–2671.

Wesselink AK, Rothman KJ, Hatch EE, Mikkelsen EM, Sørensen HT, Wise LA. Age and fecundability in a North American preconception cohort study. *Am J Obstet Gynecol* 2017;**217**:667.e1–667.e8.

Williams DR, Collins C. Racial residential segregation: a fundamental cause of racial disparities in health. *Public Health Rep* 2001;**116**: 404–416.

Willis SK, Hatch EE, Wesselink AK, Rothman KJ, Mikkelsen EM, Wise LA. Female sleep patterns, shift work, and fecundability in a North American preconception cohort study. *Fertil Steril* 2019; **111**:1201–1210.e1.

Wise LA, Mikkelsen EM, Rothman KJ, Riis AH, Sørensen HT, Huybrechts KF, Hatch EE. A prospective cohort study of menstrual characteristics and time to pregnancy. *Am J Epidemiol* 2011;**174**: 701–709.

Wise LA, Rothman KJ, Mikkelsen EM, Stanford JB, Wesselink AK, McKinnon C, Gruschow SM, Horgan CE, Wiley AS, Hahn KA *et al.* Design and conduct of an internet-based preconception cohort study in North America: pregnancy study online. *Paediatr Perinat Epidemiol* 2015;**29**:360–371.

Wise LA, Rothman KJ, Wesselink AK, Mikkelsen EM, Sorensen HT, McKinnon CJ, Hatch EE. Male sleep duration and fecundability in a North American preconception cohort study. *Fertil Steril* 2018; **109**:453–459.

Wise LA, Wesselink AK, Hatch EE, Weuve J, Murray EJ, Wang TR, Mikkelsen EM, T, Sørensen, H, Rothman, KJ. Changes in behavior with increasing pregnancy attempt time: a prospective cohort study. *Epidemiology* 2020;**31**:659–667.

Yland JJ, Bresnick KA, Hatch EE, Wesselink AK, Mikkelsen EM, Rothman KJ, Sørensen HT, Huybrechts KF, Wise LA. Pregravid contraceptive use and fecundability: prospective cohort study. *BMJ* 2020;**371**:m3966.