Running Title: Predictive models of miscarriage

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

Jennifer J Yland PhD, MS[a,*]; Zahra Zad[b,c,*]; Tanran R Wang MPH[a]; Amelia K Wesselink PhD[a]; Tammy Jiang PhD, MPH [a]; Elizabeth E Hatch PhD[a]; Ioannis Ch. Paschalidis PhD[b,c,d] ; Lauren A Wise ScD[a]

[a] Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA;
[b] Hariri Institute for Computing and Computational Science & Engineering, Boston University, Boston, MA, USA;
[c] Department of Electrical and Computer Engineering, Division of Systems Engineering, Boston University, Boston, MA, USA;
[d] Department of Biomedical Engineering, Boston University, Boston, MA, USA
[*] These authors contributed equally to this work and should be considered joint first authors.

Corresponding Author:
Jennifer J Yland
yland@bu.edu
(ORCID 0000-0001-7870-8971)
Department of Epidemiology
Boston University School of Public Health
715 Albany Street
Boston, Massachusetts, 02118, USA

Article type: Cohort study

Funding Statement: The research was partially supported by the NSF under grants CCF-2200052, IIS-1914792, and DMS-1664644, and by the NIH under grants R01 GM135930, R21 HD072326, R01 HD086742 and grant 1UL1TR001430 to the Clinical & Translational Science Institute at Boston University.

Disclosure Statement: LAW received in-kind donations from Swiss Precision Diagnostics and Kindara.com for primary data collection in PRESTO. LAW serves as a consultant for AbbVie Inc. and the Gates Foundation. The other authors have nothing to disclose.

Attestation Statement: Data regarding any of the subjects in the study has not been previously published unless specified.

Data Sharing Statement: 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-miscarriage/

Abstract Word Count: 317
Text Word Count: 3,199

Capsule: We used supervised machine learning methods to generate predictive models of miscarriage based on self-reported preconception data.

Structured Abstract

Objective: To use self-reported preconception data to derive models that predict risk of miscarriage.

Design: Prospective preconception cohort study.

Subjects: Study participants were female, aged 21-45 years, residents of the United States or Canada, and attempting spontaneous pregnancy at enrollment during 2013-2022. Participants were followed for up to 12 months of pregnancy attempts; those who conceived were followed through pregnancy and postpartum. We restricted analyses to participants who conceived during the study period.

Exposure: On baseline and follow-up questionnaires completed every 8 weeks until pregnancy, we collected self-reported data on sociodemographic factors, reproductive history, lifestyle, anthropometrics, diet, medical history, and male partner characteristics. We included 160 potential predictor variables in our models.

Main Outcome Measures: The primary outcome was miscarriage, defined as pregnancy loss before 20 weeks' gestation. We followed participants from their first positive pregnancy test until miscarriage or a censoring event (induced abortion, ectopic pregnancy, loss to follow-up, or 20 weeks' gestation), whichever occurred first. We fit both survival and static models, using Cox proportional hazards models, logistic regression, support vector machines, Gradient Boosted Trees, and Random Forest algorithms. We evaluated model performance using the concordance index (survival models) and the weighted-F1 score (static models).

Results: Among 8,720 participants who conceived, 20.4% reported miscarriage. In multivariable models, the strongest predictors of miscarriage were female age, history of miscarriage, and male partner age. The weighted-F1 score ranged from 73-89% for static models and the concordance

index ranged from 53-56% for survival models, indicating better discrimination for the static models compared with the survival models (*i.e.*, ability of the model to discriminate between individuals with and without miscarriage). No appreciable differences were observed across strata of miscarriage history or among models restricted to ≥8 weeks' gestation.

Conclusion: Our findings suggest that miscarriage is not easily predicted based on preconception lifestyle characteristics, and that advancing age and history of miscarriage are the most important predictors of incident miscarriage.

Keywords: miscarriage; spontaneous abortion; machine learning; predictive modeling; pregnancy

Introduction

Miscarriage, or pregnancy loss before 20 completed weeks of gestation, affects approximately 20% of recognized pregnancies (1). The strongest identified predictors of miscarriage are older parental age and history of miscarriage (2). Other reported risk factors include low and high body mass index (BMI) (3,4), caffeine consumption (5–7), alcohol intake (8–11), and smoking (12–14), though the etiology of miscarriage remains poorly understood.

Several studies have developed predictive models of miscarriage among individuals receiving treatment with assisted reproduction technology (ART) (15–17), individuals with recurrent miscarriage (18–21), and individuals with threatened miscarriage (22). Most of these studies have relied on clinical assessments such as early pregnancy ultrasound measurements and laboratory values. Other studies have attempted to predict miscarriage based on early pregnancy characteristics (*e.g.*, parental age, ultrasound measurements, and laboratory values) in general populations (23,24). However, no study has derived a predictive model of miscarriage using prospectively collected data on the couple during the preconception period. Predicting primary (i.e., first-time) miscarriage is of great importance, given the high rate of miscarriage and the impact of miscarriage on mental health and fertility outcomes. Moreover, primary miscarriage likely shares many risk factors with recurrent miscarriage (25).

In a North American prospective preconception cohort study, we predicted risk of miscarriage using 160 self-reported variables describing a variety of preconception sociodemographic, lifestyle, dietary, and anthropometric factors. We used supervised machine learning methods with several classification algorithms and variable selection procedures.

Materials and Methods

Study population

Pregnancy Study Online (PRESTO) is an ongoing web-based preconception cohort study that collects data on a variety of environmental and behavioral factors in addition to pregnancy outcomes (26). At enrollment, eligible participants were female, aged 21-45 years, residents of the United States (US) or Canada, and trying to conceive without the use of fertility treatment. Participants were followed for up to 12 months of pregnancy attempts, during which time they could have initiated fertility treatment. Participants who conceived were followed through pregnancy and postpartum.

During 2013 through 2022, 16,631 female participants enrolled in PRESTO and completed a baseline questionnaire. We excluded 37 participants who were not from the US or Canada, 120 who were already pregnant at study entry, 203 who completed the baseline questionnaire <11 weeks before analysis (and therefore had no opportunity for follow-up), and 41 who completed the baseline questionnaire >2 months after the screening questionnaire. Approximately 36% of participants were lost to follow-up. Among those who were lost to follow-up, we successfully collected information on pregnancy for 25% of participants via email or phone contact, or by searching for baby registries and birth announcements online; for 5% by linking to birth registries in selected states (CA, FL, MA, MI, NY, OH, PA, TX); and for 5% by linking to FertilityFriend.com data (a mobile computing fertility-tracking app).

In total, 8,739 participants became pregnant during follow-up (we included only the first observed pregnancy per participant in these analyses). We excluded 19 participants with missing data on categorical variables (handling of missing data is described in the Supplementary Material), retaining a total of 8,720 participants in the dataset used for our analysis. The institutional review board at Boston University Medical Campus approved the study protocol.

Data collection

Female participants completed a baseline questionnaire and follow-up questionnaires every eight weeks until pregnancy. Those who conceived completed an early pregnancy questionnaire at a median of 9 weeks' gestation and a late pregnancy questionnaire at approximately 32 weeks' gestation. On baseline, follow-up, and pregnancy questionnaires, we collected data on pregnancy status, sociodemographic factors, lifestyle and behavioral factors, anthropometrics, medical and reproductive history, and selected male partner characteristics. Reproductive history included gravidity, parity, and history of miscarriage (i.e., miscarriages that occurred prior to enrolling in PRESTO), among other variables. Participants were also invited to complete the web-based Diet History Questionnaire (DHQ II: 2013-2019; DHQ III: 2020-2022) ten days after enrollment. The DHQ was designed by the National Cancer Institute and the first version of the DHQ was validated against 24-hour dietary recalls in a US population (27,28). We used DHQ data to calculate the Healthy Eating Index-2010 (HEI-2010) score, a measure of diet quality (29). For time-varying characteristics, we prioritized data collected most recently before conception to avoid bias due to conditioning on future information (30). Table 1 provides a complete list of the 160 variables included in this analysis and when they were ascertained. Ninety variables were binary, 58 were continuous, and 12 were categorical. Table S1 describes the percentage of

missingness for each predictor variable and the Methods Supplement provides an overview of how missing data were handled.

Assessment of miscarriage

We defined miscarriage as pregnancy loss before 20 completed weeks of gestation (including blighted ovum and chemical pregnancy but excluding ectopic pregnancy and induced abortion). On follow-up questionnaires, participants reported the date of their last menstrual period, whether they were currently pregnant, and whether they had experienced a miscarriage since completing their previous questionnaire. Participants who reported a miscarriage were asked how many weeks the pregnancy lasted and on what date the pregnancy ended. Pregnant participants reported the due date of their current pregnancy and the date of their first positive pregnancy test. Pregnant participants were asked to report the method(s) used to confirm their pregnancy (*e.g.*, home pregnancy test, urine or blood test in doctor's office, ultrasound). More than 95% of participants reported using a home pregnancy test to identify their pregnancy.

For participants who reported a miscarriage, we used the participant's reported gestational weeks at loss when available (defined as weeks since the last menstrual period). Among participants who did not report their gestational week at loss but who reported a due date (11%), we estimated gestational age as: (pregnancy end date – (pregnancy due date – 280 days))/7 (31). Among participants who reported neither their gestational week at loss nor their due date (21%), we estimated week at loss as: (pregnancy end date – last menstrual period date)/7. Approximately 97% of miscarriages were identified via study questionnaires; the remaining 3%

were identified via the study withdrawal form, via email or phone contact, by linking to birth registries, or by linking to FertilityFriend.com data.

Statistical analysis

We used supervised machine learning methods to generate predictive models of miscarriage. We generated both static and survival models. Static models predict the risk or odds of miscarriage without consideration of time at loss, while survival models predict the rate of miscarriage (conceptualized as time to miscarriage). For all analyses, we first performed several pre-processing steps including statistical feature selection. For static models, we used a variety of supervised classification methods including linear (*e.g.*, logistic regression) and non-linear (*e.g.*, Gradient Boosted Trees) algorithms. For survival models, we fit Cox proportional hazards models. For both static and survival models, we generated full and sparse models. The full models contain all variables selected by statistical feature selection, whereas the sparse models contain all variables selected by both statistical feature selection and univariate feature selection for survival models or recursive feature elimination for static models. We evaluated model performance via the area under the receiver operating characteristic curve (AUC), precision and recall metrics, and the weighted-F1 score for static models, and via the concordance index for survival models. These methods are described in greater detail in the Supplementary Material.

Sensitivity analyses

We repeated all analyses among primigravid participants to generate models predictive of primary miscarriage, which may have different predictors from secondary or recurrent miscarriage. We also restricted the dataset to ≥8 gestational weeks to assess the extent to which

predictors differed for later losses, which are less likely to be attributable to random chromosomal aberrations (32). All analyses were performed with Python packages. Relevant programs can be accessed here: https://github.com/noc-lab/Predictive-models-of-miscarriage/

Results

We analyzed data from 8,720 pregnant participants, among whom 1,775 (20.4%) experienced miscarriage during the 12-month study period. Miscarriages were reported as early as 3 gestational weeks (median=6; interquartile range: 5-8 gestational weeks). We observed 567 late miscarriages (32% occurring ≥8 gestational weeks). The distribution of gestational weeks at miscarriage is presented in Table S2. Mean age was 30 years for female participants and 32 years for male partners. Mean BMI of female participants was 27 kg/m^2 and 28 kg/m^2 for male partners. Approximately one quarter of couples resided in the Northeast US, while 22% resided in the South, 22% in the Midwest, 16% in the West, and 16% in Canada. Approximately one quarter of participants had a previous miscarriage, 35% had had an unplanned pregnancy before enrolling in PRESTO, and about half were parous. Almost 14% of female participants reported any history of subfertility or infertility, and 7% of study pregnancies were conceived via fertility treatment.

Survival models

After statistical feature selection, 17 variables remained in the dataset. The variables selected into the full survival model are presented in Table S3. The variables selected into the sparse survival model are presented in Table 2. The strongest two predictors in the sparse survival model were female age at conception (HR=1.19; 95% CI: 1.11, 1.27) and history of miscarriage

(HR=1.10; 95% CI: 1.03, 1.17), which were both positively associated with miscarriage (Table 2). All other variables selected into the sparse model had very small or null associations with miscarriage. Variables that were very slightly positively associated with miscarriage were use of omega-3 or fish oil supplements (HR=1.04; 95% CI: 0.99, 1.10), number of prior pregnancies (HR=1.04; 95% CI: 0.99, 1.10), history of subfertility or infertility (HR=1.04; 95% CI: 0.97, 1.11), male partner age at conception (HR=1.03; 95% CI: 0.97, 1.10), and having a history of unplanned pregnancy (HR=1.01; 95% CI: 0.94, 1.09). Variables that were very slightly inversely associated with miscarriage included having been pregnant before (HR=0.95; 95% CI: 0.87, 1.05) and being vaccinated against human papillomavirus (HPV) (HR=0.98; 95% CI: 0.93, 1.04). The concordance index of the final sparse survival model, applied to the testing dataset, was 55.4%, indicating poor-to-moderate discrimination (*i.e.*, ability of the model to discriminate between individuals with and without miscarriage).

When we restricted the incident period to ≥8 gestational weeks (n=6,993; 32% of all miscarriages), 4 variables remained after statistical feature selection. The strongest predictors of miscarriage were female age at conception, male partner age at conception, and history of unplanned pregnancy, each of which was positively associated with miscarriage (Table S4). The Healthy Eating Index-2010 score was also selected into this model and was inversely associated with miscarriage. The concordance index for this model was 55.6%.

When we restricted to primigravid participants (n=4,267), 9 variables remained after statistical feature selection. In this model, variables that were positively associated with miscarriage included female age at conception, male age at conception, use of omega-3 or fish oil

supplements, recent use of psychotropic medications, and female BMI; variables that were inversely associated with miscarriage included being married, use of oral contraceptives as the most recent contraceptive method, residence in the Northeast US, and the Healthy Eating Index-2010 score (Table S5). The concordance index for this model was 57.4%. Among primigravid participants who contributed ≥8 gestational weeks to the analysis (n=3,488), only female and male partner age remained after statistical feature selection, and the concordance index was 53.3% (Table S6).

Static models

Variables selected into the full static models are presented in Table S3. After recursive feature elimination, there were 9 variables in the sparse model (Table 3). Performance metrics for all static models are presented in Table 4. The weighted-F1 score ranged from 72.6% for the LR-L1 model to 73.5% for the RF model. The two most important variables selected into the sparse static model were female age at conception and history of miscarriage, which were both positively associated with miscarriage.

When we restricted the incident period to ≥8 gestational weeks (6,993 pregnancies), 4 features remained after statistical feature selection, and 2 remained after recursive feature elimination (Table S7). Female and male age at conception were the final two variables selected into the sparse model, with a weighted-F1 score of 88.0%. Among primigravid participants (n=4,267), 9 features remained after statistical feature selection, and all of these remained after recursive feature elimination. The weighted-F1 score of the sparse model was 73.8%, and the two most important variables selected into the model were residing in the Northeast US (negatively

associated with miscarriage) and female age at conception (positively associated with miscarriage) (Table S8). Among primigravid participants with pregnancies lasting ≥8 gestational weeks (n=3,488), 2 features remained after statistical feature selection and only 1 remained in the final sparse model: male age at conception (Table S9). The weighted-F1 score for this model was 88.5%.

Discussion

In this prospective cohort study of North American pregnancy planners, we developed predictive models for miscarriage based on self-reported preconception data. Previous studies have identified few confirmed causes of miscarriage, and the strongest identified risk factors in these studies were age and history of miscarriage (2). In the present study, we generated models with moderate predictive power: the weighted-F1 score ranged from 73-89% for static models and the concordance index ranged from 53-56% for survival models. However, the AUC was <60% for all static models. Consistent with previous studies, our findings indicate that advancing female and male partner age are the most important predictors of miscarriage, and that female age is generally more predictive than male age. After age, history of miscarriage appeared to be the strongest predictor of miscarriage. These factors were consistently predictive of miscarriage across a variety of models and settings.

Our study identified several preconception dietary factors as predictors of miscarriage, albeit most associations were very small and consistent with the null. Specifically, a healthier diet as measured by the Healthy Eating Index-2010 score (*e.g.*, greater intake of fruits and vegetables, whole grains, dairy, seafood & plant proteins, and unsaturated fats) was associated with a

slightly lower rate of miscarriage. In addition, use of omega-3 or fish oil supplements was associated with a slightly increased rate of miscarriage and several B-vitamins were selected with inconsistent associations. Several studies have investigated the relation between dietary factors and miscarriage (33–39). One study – with a similar design to PRESTO – reported an inverse association between adherence to Nordic dietary guidelines (which emphasize fish consumption) and risk of miscarriage (35). Another study evaluated the association between pre-pregnancy adherence to three dietary patterns – the Healthy Eating Index 2010, the Alternative Mediterranean Diet, and the Fertility Diet (FD) – and risk of miscarriage among 15,950 pregnancies in the Nurses' Health Study II (34). The authors reported no association between these dietary patterns and miscarriage. The role of dietary factors remains debated, and the predictive ability of these variables in our study was small.

An unexpected finding in our study was the selection of smoking status into the sparse static model and the full survival model in the full study population (i.e., not restricted by gravidity or gestational week). However, the overall prevalence of smoking was quite low in this study population (4%), and this variable was not consistently selected into all models. Moreover, the detrimental health effects of smoking tobacco are well documented, and several studies have identified a positive association between current smoking and miscarriage risk (13,14).

The following variables were selected into models developed among primigravid participants but not among those who were previously pregnant: marital status, pregravid use of oral contraceptives, recent use of psychotropic medications, and female BMI. Being married was associated with a lower rate of miscarriage, which could be related to higher socioeconomic

position, greater social and emotional support, and lower stress levels. However, factors such as perceived stress scores and household income were not selected as important predictors of miscarriage during the statistical feature selection process. Some (40–42) but not all (43,44) studies reported that pregravid use of oral contraceptives was associated with a lower risk of miscarriage compared with non-use, in agreement with the present study. However, a recently published paper conducted in PRESTO reported that pregravid use of oral contraceptives was not associated with miscarriage (45). This contrast may be due to differences in model selection, as the previous publication aimed to estimate potential causal effects of contraceptive use. The potential association between use of psychotropic medications and miscarriage has been debated. However, a recent study reported that use of antidepressants was not associated with miscarriage after controlling for depression diagnosis (46). High BMI has previously been associated with an increased risk of miscarriage (3,4). Among 5,132 couples who conceived in a Danish preconception cohort study, the adjusted HR for miscarriage among women with BMI ≥30 kg/m^2 relative to those with BMI 20-24 kg/m^2 was 1.23 (95% CI: 0.98, 1.54) (4).

We attempted to isolate predictors of later miscarriage, as earlier miscarriages (<8 weeks' gestation) are more likely to be due to chromosomal abnormalities than later losses (47). However, the predictive ability of models restricted to ≥8 gestational weeks was no better than those generated in the entire dataset, and the list of variables selected for these models was similar to those based on full spectrum of gestational ages (all miscarriages).

Previous studies have developed models to predict miscarriage in special populations, such as couples with recurrent miscarriage (18–21) or those using ART (15–17). These studies largely

relied upon ultrasound measurements (e.g., gestational sac size, crown-rump length, fetal heart rate) or laboratory values (e.g., beta-human chorionic gonadotropin, progesterone levels) during early pregnancy. One study in the Netherlands attempted to predict pregnancy outcome among 526 couples with unexplained recurrent miscarriage (21). Data on previous miscarriages and fertility treatment; and male and female age, BMI, and smoking status were included, and all were identified as potential predictors of miscarriage, with an AUC of 0.66. The present study greatly expands on the breadth of potential predictors assessed. Moreover, our findings might be useful for couples who wish to understand their risk for miscarriage before trying to conceive spontaneously.

Study limitations include bias due to missingness or misclassification of predictor variables. All data were self-reported, and certain variables such as dietary factors or medication use may be more vulnerable to misclassification than others. The impact of misclassification on our findings is challenging to quantify, as there is little research on the impact of measurement error on machine learning prediction models (48,49). Outcome misclassification is also possible but unlikely: more than 95% of participants reported using at-home-pregnancy tests and we ascertained miscarriages as early as 3 weeks' gestation. In addition, although we evaluated a wide range of variables, we were unable to include environmental exposures (e.g., phthalates, phenols, pesticides, etc.) as potential predictors. Moreover, we did not evaluate interactions between the independent variables, such as depressive symptoms and use of psychotropic medications. Finally, though we validated the models using split-sample replication techniques, we were unable to conduct an external validation study. Given that more than 93% of PRESTO

participants had spontaneous conceptions, our results may not generalize to ART-conceived conceptions.

Conclusions

In this study, we used a variety of supervised machine learning methods to generate predictive models of miscarriage based on self-reported preconception data. We considered 160 potential predictors of miscarriage and analyzed data from nearly 9,000 pregnancies. Female age, male age, and history of miscarriage were the most important predictors of miscarriage, consistent with existing knowledge. The overall performance of our models was moderate. Our findings suggest that miscarriage is not easily predicted based on preconception lifestyle characteristics, including reproductive and medical factors.

Authors' Roles: All authors were responsible for formulation of the study hypotheses and study design, statistical analyses, results interpretation, manuscript writing, revision, and finalization.

Acknowledgments: We acknowledge the contributions of PRESTO participants and staff.

References

1. Rossen LM, Ahrens KA, Branum AM. Trends in Risk of Pregnancy Loss Among US Women, 1990-2011. Paediatric and perinatal epidemiology 2018;32(1):19–29.

2. Wilcox AJ, Weinberg CR, O'Connor JF, Baird DD, Schlatterer JP, Canfield RE, et al. Incidence of early loss of pregnancy. New England Journal of Medicine 1988;319(4):189–94.

3. Arck PC, Rücke M, Rose M, Szekeres-Bartho J, Douglas AJ, Pritsch M, et al. Early risk factors for miscarriage: a prospective cohort study in pregnant women. Reprod Biomed Online 2008;17(1):101–13.

4. Hahn KA, Hatch EE, Rothman KJ, Mikkelsen EM, Brogly SB, Sørensen HT, et al. Body size and risk of spontaneous abortion among danish pregnancy planners. Paediatr Perinat Epidemiol 2014;28(5):412–23.

5. Savitz DA, Chan RL, Herring AH, Howards PP, Hartmann KE. Caffeine and miscarriage risk. Epidemiology 2008;19(1):55–62.

6. Weng X, Odouli R, Li D-K. Maternal caffeine consumption during pregnancy and the risk of miscarriage: a prospective cohort study. Am J Obstet Gynecol 2008;198(3):279.e1-8.

7. Hahn KA, Wise LA, Rothman KJ, Mikkelsen EM, Brogly SB, Sørensen HT, et al. Caffeine and caffeinated beverage consumption and risk of spontaneous abortion. Hum Reprod 2015;30(5):1246–55.

8. Klonoff-Cohen H, Lam-Kruglick P, Gonzalez C. Effects of maternal and paternal alcohol consumption on the success rates of in vitro fertilization and gamete intrafallopian transfer. Fertil Steril 2003;79(2):330–9.

9. Henriksen TB, Hjollund NH, Jensen TK, Bonde JP, Andersson A-M, Kolstad H, et al. Alcohol consumption at the time of conception and spontaneous abortion. Am J Epidemiol 2004;160(7):661–7.

10. Andersen A-MN, Andersen PK, Olsen J, Grønbæk M, Strandberg-Larsen K. Moderate alcohol intake during pregnancy and risk of fetal death. Int J Epidemiol 2012;41(2):405–13.

11. Feodor Nilsson S, Andersen PK, Strandberg-Larsen K, Nybo Andersen A-M. Risk factors for miscarriage from a prevention perspective: a nationwide follow-up study. BJOG 2014;121(11):1375–84.

12. Venners SA, Wang X, Chen C, Wang L, Chen D, Guang W, et al. Paternal smoking and pregnancy loss: a prospective study using a biomarker of pregnancy. Am J Epidemiol 2004;159(10):993–1001.

13. George L, Granath F, Johansson ALV, Annerén G, Cnattingius S. Environmental tobacco smoke and risk of spontaneous abortion. Epidemiology 2006;17(5):500–5.

14. Nielsen A, Hannibal CG, Lindekilde BE, Tolstrup J, Frederiksen K, Munk C, et al. Maternal smoking predicts the risk of spontaneous abortion. Acta Obstet Gynecol Scand 2006;85(9):1057–65.

15. Choong S, Rombauts L, Ugoni A, Meagher S. Ultrasound prediction of risk of spontaneous miscarriage in live embryos from assisted conceptions. Ultrasound Obstet Gynecol 2003;22(6):571–7.

16. Yi Y, Lu G, Ouyang Y, lin G, Gong F, Li X. A logistic model to predict early pregnancy loss following in vitro fertilization based on 2601 infertility patients. Reprod Biol Endocrinol 2016;14:15.

17. Liu L, Jiao Y, Li X, Ouyang Y, Shi D. Machine learning algorithms to predict early pregnancy loss after in vitro fertilization-embryo transfer with fetal heart rate as a strong predictor. Comput Methods Programs Biomed 2020;196:105624.

18. Quenby SM, Farquharson RG. Predicting recurring miscarriage: what is important? Obstet Gynecol 1993;82(1):132–8.

19. Caetano MR, Couto E, Passini R, Simoni RZ, Barini R. Gestational prognostic factors in women with recurrent spontaneous abortion. Sao Paulo Med J 2006;124(4):181–5.

20. Dai Y-F, Lin L-Z, Lin N, He D-Q, Guo D-H, Xue H-L, et al. APA scoring system: a novel predictive model based on risk factors of pregnancy loss for recurrent spontaneous abortion patients. J Obstet Gynaecol 2022;42(6):2069–74.

21. du Fossé NA, van der Hoorn M-LP, de Koning R, Mulders AGMGJ, van Lith JMM, le Cessie S, et al. Toward more accurate prediction of future pregnancy outcome in couples with unexplained recurrent pregnancy loss: taking both partners into account. Fertil Steril 2022;117(1):144–52.

22. Huang J, Lv P, Lian Y, Zhang M, Ge X, Li S, et al. Construction of machine learning tools to predict threatened miscarriage in the first trimester based on AEA, progesterone and β-hCG in China: a multicentre, observational, case-control study. BMC Pregnancy Childbirth 2022;22(1):697.

23. DeVilbiss EA, Mumford SL, Sjaarda LA, Connell MT, Plowden TC, Andriessen VC, et al. Prediction of pregnancy loss by early first trimester ultrasound characteristics. Am J Obstet Gynecol 2020;223(2):242.e1-242.e22.

24. Li S, Li D, Ma Y. A mathematical model to predict the probability of a successful pregnancy. J Obstet Gynaecol Res 2022;48(7):1632–40.

25. Cramer DW, Wise LA. The epidemiology of recurrent pregnancy loss. Semin Reprod Med 2000;18(4):331–9.

26. Wise LA, Rothman KJ, Mikkelsen EM, Stanford JB, Wesselink AK, McKinnon C, et al. Design and Conduct of an Internet-Based Preconception Cohort Study in North America: Pregnancy Study Online. Paediatric and perinatal epidemiology 2015;29(4):360–71.

27. Subar AF, Thompson FE, Kipnis V, Midthune D, Hurwitz P, McNutt S, et al. Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires : the Eating at America's Table Study. American journal of epidemiology 2001;154(12):1089–99.

28. Millen AE, Midthune D, Thompson FE, Kipnis V, Subar AF. The National Cancer Institute diet history questionnaire: validation of pyramid food servings. American journal of epidemiology 2006;163(3):279–88.

29. Guenther PM, Casavale KO, Reedy J, Kirkpatrick SI, Hiza HAB, Kuczynski KJ, et al. Update of the Healthy Eating Index: HEI-2010. J Acad Nutr Diet 2013;113(4):569–80.

30. Suissa S, Dell'Aniello S. Time-related biases in pharmacoepidemiology. Pharmacoepidemiol Drug Saf 2020;29(9):1101–10.

31. ACOG Committee Opinion No 579: Definition of term pregnancy. Obstetrics and gynecology 2013;122(5):1139–40.

32. Savitz DA, Hertz-Picciotto I, Poole C, Olshan AF. Epidemiologic measures of the course and outcome of pregnancy. Epidemiologic reviews 2002;24(2):91–101.

33. Hsiao PY, Fung JL, Mitchell DC, Hartman TJ, Goldman MB. Dietary quality, as measured by the Alternative Healthy Eating Index for Pregnancy (AHEI-P), in couples planning their first pregnancy. Public Health Nutrition 2019;22(18):3385–94.

34. Gaskins AJ, Rich-Edwards JW, Hauser R, Williams PL, Gillman MW, Penzias A, et al. Prepregnancy dietary patterns and risk of pregnancy loss1,2,3. The American Journal of Clinical Nutrition 2014;100(4):1166–72.

35. Laursen ASD, Johannesen BR, Willis SK, Hatch EE, Wise LA, Wesselink AK, et al. Adherence to Nordic dietary patterns and risk of first-trimester spontaneous abortion. Eur J Nutr 2022;61(6):3255–65.

36. Gaskins AJ, Nassan FL, Chiu Y-H, Arvizu M, Williams PL, Keller MG, et al. Dietary patterns and outcomes of assisted reproduction. American Journal of Obstetrics & Gynecology 2019;220(6):567.e1-567.e18.

37. Karayiannis D, Kontogianni MD, Mendorou C, Mastrominas M, Yiannakouris N. Adherence to the Mediterranean diet and IVF success rate among non-obese women attempting fertility. Human Reproduction 2018;33(3):494–502.

38. Twigt JM, Bolhuis MEC, Steegers EAP, Hammiche F, van Inzen WG, Laven JSE, et al. The preconception diet is associated with the chance of ongoing pregnancy in women undergoing IVF/ICSI treatment. Human Reproduction 2012;27(8):2526–31.

39. Wesselink AK, Willis SK, Laursen ASD, Mikkelsen EM, Wang TR, Trolle E, et al. Protein-rich food intake and risk of spontaneous abortion: a prospective cohort study. Eur J Nutr 2022;61(5):2737–48.

40. Hahn KA, Hatch EE, Rothman KJ, Mikkelsen EM, Brogly SB, Sørensen HT, et al. History of oral contraceptive use and risk of spontaneous abortion. Ann Epidemiol 2015;25(12):936-941.e1.

41. Sackoff J, Kline J, Susser M. Previous use of oral contraceptives and spontaneous abortion. Epidemiology 1994;5(4):422–8.

42. Rothman KJ. Fetal loss, twinning and birth weight after oral-contraceptive use. N Engl J Med 1977;297(9):468–71.

43. Risch HA, Weiss NS, Clarke EA, Miller AB. Risk factors for spontaneous abortion and its recurrence. Am J Epidemiol 1988;128(2):420–30.

44. Jellesen R, Strandberg-Larsen K, Jørgensen T, Olsen J, Thulstrup AM, Andersen A-MN. Maternal use of oral contraceptives and risk of fetal death. Paediatr Perinat Epidemiol 2008;22(4):334–40.

45. Yland JJ, Bresnick KA, Hatch EE, Wesselink AK, Mikkelsen EM, Rothman KJ, et al. Pregravid contraceptive use and fecundability: prospective cohort study. BMJ 2020;371:m3966.

46. Kjaersgaard MIS, Parner ET, Vestergaard M, Sørensen MJ, Olsen J, Christensen J, et al. Prenatal antidepressant exposure and risk of spontaneous abortion - a population-based study. PLoS One 2013;8(8):e72095.

47. Pflueger SMV. Cytogenetics of Spontaneous Abortion [Internet]. In: Gersen SL, Keagle MB, editors. The Principles of Clinical Cytogenetics. Totowa, NJ: Humana Press; 2005 [cited 2022 Sep 26]. p. 323–45.Available from: https://doi.org/10.1385/1-59259-833-1:323

48. Jiang T, Gradus JL, Lash TL, Fox MP. Addressing Measurement Error in Random Forests using Quantitative Bias Analysis. American journal of epidemiology 2021;

49. van Doorn S, Brakenhoff TB, Moons KGM, Rutten FH, Hoes AW, Groenwold RHH, et al. The effects of misclassification in routine healthcare databases on the accuracy of prognostic prediction models: a case study of the CHA2DS2-VASc score in atrial fibrillation. Diagnostic and Prognostic Research 2017;1(1):18.

50. Cochran WG. The chi2 Test of Goodness of Fit. The Annals of Mathematical Statistics 1952;23(3):315–45.

51. Massey FJ. The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association 1951;46(253):68–78.

52. 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. Journal of Biomedical Informatics 2020;108:103496.

53. Schmid M, Wright MN, Ziegler A. On the use of Harrell's C for clinical risk prediction via random survival forests. Expert Systems with Applications 2016;63:450–9.

54. Cortes C, Vapnik V. Support-vector networks. Machine learning 1995;20(3):273–97.

55. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33(1):1–22.

56. Friedman JH. Stochastic gradient boosting. Computational statistics & data analysis 2002;38(4):367–78.

57. Mason L, Baxter J, Bartlett P, Frean M. Boosting algorithms as gradient descent in function space. Nips; 1999.

58. Breiman L. Random Forests. Machine Learning 2001;45(1):5–32.

59. Hastie T, Tibshirani R, Friedman J. Ridge Regression. In: The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media; 2009. p. 61–5.

60. 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(3):e0118432.

Table 1. Complete list of variables included in analysis to generate predictive models of miscarriage in PRESTO, 2013-2022.

Category	Variables
Demographic and socioeconomic characteristics	Age*, marital status, region of residence, urbanization of residential area, 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	Years in a steady relationship, 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, B5, B6, B7, B12, C, 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, 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; mother's history of miscarriage.
Reproductive characteristics and disorders	Use of fertility treatment to conceive the study pregnancy (if yes, type of treatment); history of miscarriage; age at menarche; menstrual regularity; menstrual period characteristics (typical length, 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; Ferriman-Gallwey Hirsutism Score; 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; last method of contraception; number of menstrual cycles to conceive the study pregnancy.
Physical characteristics, non-reproductive medical history, and medication use	Body mass index; waist circumference; 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.

*These variables were considered time-varying characteristics and were updated on follow-up questionnaires completed after the baseline questionnaire but before conception.

Table 2. Variables selected by sparse survival model to predict miscarriage in PRESTO, 2013-2022.

Variable	Hazard Ratio[1] (95% CI)
Female age at conception (years)	1.19 (1.11, 1.27)
History of miscarriage (yes/no)	1.10 (1.03, 1.17)
Ever pregnant before (yes/no)	0.95 (0.87, 1.05)
Use of omega-3 or fish oil supplements (yes/no)	1.04 (0.99, 1.10)
Number of prior pregnancies	1.04 (0.99, 1.10)
History of subfertility or infertility[2] (yes/no)	1.04 (0.97, 1.11)
Male age at conception (years)	1.03 (0.97, 1.10)
Ever received HPV vaccine	0.98 (0.93, 1.04)
History of unplanned pregnancy (yes/no)	1.01 (0.94, 1.09)
Previously tried to conceive for ≥12 months[2]: "no, never tried before" (ref = "no")	1.00 (0.93, 1.08)
Variables forced into the model[3]	
Previously tried to conceive for ≥12 months: "yes" (ref = "no")	0.99 (0.77, 1.26)

Abbreviations: CI, confidence interval; HPV, human papillomavirus.

[1] Continuous variables were standardized; the effect estimate is the hazard ratio for a one-unit increase in the z-score for that variable.

[2] History of subfertility or infertility is derived from participants' responses to questions about their reproductive history and was defined as having previously tried to conceive for ≥6 months for any prior pregnancy; previously tried to conceive for ≥12 months was participants' response to the question, "have you ever tried for ≥12 months without conceiving?"

[3] For all models, we selected a reference group for each categorical variable that was recoded as an indicator variable 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 sparse model.

Table 3. Variables selected by sparse static model (logistic regression with an ℓ2-norm regularization term) to predict miscarriage in PRESTO, 2013-2022.

Variable	OR (95% CI)	β	Correlation with outcome	Overall frequency (std.) or mean (std.)	Frequency or mean, by outcome status[1]	
					Miscarriage	*No miscarriage*
Female age at conception (years)	1.23 (1.20, 1.27)	0.21	0.09	30.2 (3.9)	30.9	30.0
History of miscarriage (yes/no)	1.16 (1.13, 1.20)	0.15	0.07	26% (44%)	32%	24%
Female smoking: current regular smoker (ref = never smoker)	0.89 (0.86, 0.91)	-0.12	-0.04	4% (19%)	3%	4%
Geographic region of residence: Northeast US (ref = South US)	0.90 (0.87, 0.92)	-0.11	-0.04	24% (43%)	21%	25%
Healthy Eating Index-2010 score (HEI-2010 score)	0.92 (0.90, 0.95)	-0.08	-0.02	66.8 (9.2)	66.4	66.8
Use of omega-3 or fish oil supplements (yes/no)	1.06 (1.04, 1.09)	0.06	0.04	19% (39%)	22%	18%
Use of vitamin B6 (yes/no)	1.05 (1.02, 1.08)	0.05	0.04	5% (21%)	6%	4%
Ever pregnant before (yes/no)	0.96 (0.93, 0.99)	-0.04	0.04	51% (50%)	55%	50%
Use of vitamin C (yes/no)	1.04 (1.01, 1.07)	0.04	0.03	7% (25%)	8%	6%
Variables forced into the model[2]						
Geographic region of residence: Canada (ref = South US)	0.97 (0.94, 1.00)	-0.03	0.00	16% (37%)	15%	16%
Female smoking: former smoker (ref = never smoker)	0.97 (0.95, 0.99)	-0.03	0.00	12% (33%)	13%	12%
Geographic region of residence: Midwest US (ref = South US)	0.99 (0.96, 1.02)	-0.01	0.01	22% (41%)	22%	22%
Female smoking: current occasional smoker (ref = never smoker)	0.99 (0.97, 1.01)	-0.01	0.00	3% (16%)	3%	3%
Geographic region of residence: West US (ref = South US)	1.00 (0.97, 1.03)	0.00	0.02	16% (37%)	18%	16%

Abbreviations: β, regression coefficient; CI, confidence interval; LR-L2, logistic regression model with an L2 penalty; OR, odds ratio (exp[β]); std, standard deviation; US, United States.

[1] These cells should be interpreted as the mean or percentage for each variable among individuals with or without miscarriage. For example, the average age of female participants who experienced a miscarriage was 30.9 years.

[2] For all models, we selected a reference group for each categorical variable that was recoded as an indicator variable 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 sparse model.

Table 4. Performance metrics for the static models predicting miscarriage in PRESTO, 2013-2022.

	Performance Measure (%) (Standard Deviation)				
Algorithm	**AUC**	**Accuracy**	**Weighted-F1 Score**	**Weighted Precision Score**	**Weighted Recall Score**
Full population					
LR-L1	56.8 (1.0)	75.8 (0.6)	72.6 (0.3)	70.9 (0.3)	75.8 (0.6)
SVM-L1	56.9 (1.0)	76.3 (0.9)	72.8 (0.2)	71.2 (0.4)	76.3 (0.9)
GBT	60.5 (0.7)	77.1 (0.9)	73.0 (0.3)	71.5 (0.5)	77.1 (0.9)
RF	59.3 (1.1)	77.5 (1.0)	73.5 (0.7)	72.2 (1.2)	77.5 (1.0)
LR-L2_RFE	57.6 (0.6)	76.3 (0.9)	72.7 (0.5)	71.0 (0.6)	76.3 (0.9)
Subset: ≥8 Gestational Weeks					
LR-L1	55.6 (2.8)	91.4 (0.2)	88.2 (0.2)	86.4 (0.6)	91.4 (0.2)
SVM-L1	55.6 (2.8)	91.4 (0.2)	88.2 (0.2)	86.4 (0.6)	91.4 (0.2)
GBT	58.5 (2.8)	91.4 (0.4)	88.2 (0.1)	86.7 (0.8)	91.4 (0.4)
RF	57.0 (3.0)	90.4 (0.6)	87.8 (0.4)	86.0 (0.7)	90.4 (0.6)
LR-L2_RFE	56.4 (2.5)	91.2 (0.6)	88.0 (0.2)	85.9 (0.8)	91.2 (0.6)
Subset: Primigravid					
LR-L1	57.3 (1.1)	78.6 (0.6)	73.8 (0.7)	71.7 (1.2)	78.6 (0.6)
SVM-L1	57.2 (1.1)	78.5 (0.7)	73.8 (0.6)	71.7 (1.2)	78.5 (0.7)
GBT	57.6 (1.9)	77.7 (1.7)	74.0 (1.3)	72.2 (1.6)	77.7 (1.7)
RF	56.0 (2.1)	75.2 (2.0)	73.1 (1.3)	71.7 (1.3)	75.2 (2.0)
LR-L2_RFE	57.3 (1.1)	78.6 (0.6)	73.8 (0.6)	71.7 (1.0)	78.6 (0.6)
Subset: Primigravid ≥8 Gestational Weeks					
LR-L1	53.4 (4.1)	92.0 (0.2)	88.7 (0.3)	86.0 (1.5)	92.0 (0.2)
SVM-L1	53.4 (4.1)	92.0 (0.2)	88.7 (0.3)	86.0 (1.5)	92.0 (0.2)
GBT	51.2 (4.8)	91.9 (0.3)	88.6 (0.2)	85.6 (0.5)	91.9 (0.3)
RF	51.5 (5.0)	91.6 (0.4)	88.5 (0.2)	86.0 (0.7)	91.6 (0.4)
LR-L2_RFE	55.5 (3.5)	91.7 (0.5)	88.5 (0.2)	85.5 (0.4)	91.7 (0.5)

Abbreviations: LR-L1=logistic regression with an ℓ1-norm regularization term; SVM-L1=support vector machines with an ℓ1-norm regularization term; GBT=Gradient Boosted Trees; RF=Random Forest; LR-L2 RFE=logistic regression with an ℓ2-norm regularization term.

Supplementary Material

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

Jennifer J Yland, Zahra Zad, Tanran R Wang, Amelia K Wesselink, Tammy Jiang, Elizabeth E Hatch, Lauren A Wise, Ioannis Ch. Paschalidis

---

Methods Supplement

Glossary

We define key variable selection methods and performance metrics below:

1. **Statistical feature selection:** a variable selection process used during the pre-processing phase for all models. We tested the association between each variable and the outcome and removed variables that were not independently associated with the outcome based on $p > 0.05$. We used the chi-squared test (50) for binary predictors and the Kolmogorov-Smirnov test for continuous predictors (51).
2. **Univariate feature selection:** a variable selection process, applied after statistical feature selection for all survival models. Univariate feature selection evaluates each feature independently based on its relationship with the outcome. We fit individual Cox proportional hazards models for each variable, such that each model contained only one independent variable, and we recorded the concordance index for each model. We ranked variables based on the associated concordance index and selected the top 10 variables with highest concordance index.
3. **Recursive feature elimination (RFE):** a variable selection process applied after statistical feature selection for all static models. RFE ranks the predictors selected into the full model (*i.e.,* by statistical feature selection) by importance and iteratively eliminates the least important variables. Importance is assessed by the absolute value of the variable coefficient in a logistic regression model derived using an $\ell 1$-norm regularization term. RFE ultimately selects a small set of variables that maximize the AUC in the training dataset.

4. **Five-fold cross validation (see *performance evaluation* below):** a process used to tune model parameters. First, we split the training dataset (80% of the full dataset) into five equal parts, or folds. Second, we trained the model using four parts. Third, we validated the model on the fifth part. We repeated these three steps for each of the five folds, such that each part of the full training dataset was used to validate the model trained on the other four parts of the training dataset. Finally, we selected the subset of values for the model parameters that led to the model with the best performance (*i.e.*, the highest AUC).
5. **AUC:** a performance metric used for static models. The Receiver Operating Characteristic (ROC) curve was created by plotting the true positive rate against the false positive rate at various thresholds. The c-statistic, or the Area Under the ROC Curve (AUC), is used to evaluate prediction performance. The AUC 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.
6. **Weighted-F1 Score:** a performance metric used for static models. The F1 score is computed as the harmonic mean of positive predictive value (i.e., precision) and sensitivity (i.e., recall), and ranges from 0 to 1. A score of 1 indicates both perfect positive predictive value and sensitivity, while a score of 0 indicates that either the positive predictive value or the sensitivity is zero. We calculated a weighted-F1 score to account for imbalance in the proportion of participants who had a miscarriage. The weighted-F1 score is the average of the scores across participants with and without a miscarriage, weighted by the number of participants in each class.
7. **Concordance index:** a performance metric used for survival models. The concordance index is the fraction or percent of the pair of observations which are concordant and show

a goodness-of-fit statistic for survival analysis. The concordance index is a generalization of the AUC that accounts for event time and loss to follow-up (52,53). Like the AUC, a value of 0.5 indicates that discrimination is no better than random, while a value of 1 would indicate perfect prediction.

Pre-processing and statistical feature selection

We performed several data pre-processing steps:

1) First, we converted each categorical variable into a set of indicator variables (reference groups were selected case-by-case for each categorical variable).
2) Second, we handled missing data as follows: we excluded individuals with missing data on categorical variables. For binary variables, we replaced missing values with zero. For continuous variables, we replaced missing values with the median value of available data. Six categorical variables had missing values, ranging from 5 missing values for having previously tried to conceive for ≥12 months to 14 missing values for handedness. Forty-nine binary variables had missing values, ranging from 7 missing values for having ever been pregnant to 1,998 for maternal history of miscarriage. All continuous variables had missing values, from 1 missing value menstrual cycle at study entry to 4,334 missing values for Ferriman-Gallwey Hirsutism Score.
3) Third, we addressed potential collinearity issues as follows: for each pair of highly correlated variables (correlation coefficient >0.9), we removed one variable.
4) Fourth, we performed statistical feature selection (described above).
5) Last, we standardized each continuous variable by subtracting its mean and dividing by its standard deviation.

Deriving and testing static models

In static (non-survival) models, the outcome was defined as miscarriage (yes/no). Individuals with an ongoing pregnancy at 20 weeks' gestation and those with an earlier censoring event (loss to follow-up, ectopic pregnancy, or induced abortion) were classified as 'no.'

After pre-processing and statistical feature selection, we randomly split the dataset into a training dataset (80% of the total dataset) and a testing dataset (20% of the total dataset). We then applied a variety of supervised classification methods to the training dataset, including linear and non-linear algorithms. These algorithms infer a function that maps a combination of inputs (*i.e.*, predictors) to outputs (*i.e.*, the outcome miscarriage). Linear models included logistic regression (LR) and support vector machines (SVM) (54), to which we added an $\ell 1$-norm regularization term to penalize an overfit model (LR-L1 and SVM-L1) (55). Non-linear models included Gradient Boosted Trees (GBT) (56,57) and Random Forest (RF) algorithms, which are tree-based learning algorithms (58). Linear models may be more interpretable because the magnitude of the coefficients is directly related to the importance of the predictor. However, non-linear models are more complex and typically yield better classification performance.

We generated full and sparse models. The full models were generated with the algorithms described above (LR-L1, SVM-L1, GBT, and RF) and contain all variables selected by statistical feature selection. To generate the sparse models, we applied Recursive Feature Elimination (RFE, defined above) after statistical feature selection. Then, we fit a logistic regression model with an L2 penalty (LR-L2 RFE) to derive the prediction strength of the features selected via

RFE. The L2 penalty improves stability of the estimations and controls for high correlation between variables (59).

Performance evaluation

We evaluated performance of the static models as follows:

1) First, we randomly split the entire dataset into five equally sized parts. Four parts comprised the training dataset (80%) and one part comprised the testing dataset (20%).
2) Second, we tuned the model hyperparameters on the training dataset using five-fold cross validation (defined above).
3) Third, we evaluated the performance of the model obtained in Step 2 on the testing dataset (20% of the full dataset; this portion was not used in five-fold cross validation).
4) We repeated the first three steps (split the data into five random parts, tune the model hyperparameters with five-fold cross validation using the training dataset, evaluate model performance in the testing dataset) five times.
5) Finally, we calculated the mean and standard deviation of each performance metric over these five runs.

In addition to the AUC, we evaluated model performance using the weighted-F1 score (defined above), which is more robust to imbalanced data than the AUC (60). We also calculated weighted-precision (*i.e.*, positive predictive value) and weighted-recall (*i.e.*, sensitivity) metrics as follows: we calculated precision and recall among participants with and without a miscarriage, and calculated the average scores across groups, weighted by the number individuals in each class.

Deriving and testing survival models

In survival models, participants entered the risk set at the week of their first positive pregnancy test and were followed until miscarriage or a censoring event (ectopic pregnancy, induced abortion, loss to follow up, or 20 weeks' gestation), whichever occurred first. We fit Cox proportional hazards models with gestational week as the time scale. We accounted for ties using Efron's method and applied an L2 penalty to all models (59). Full models contain all variables selected by statistical feature selection. To generate the sparse models, we applied Univariate Feature Selection (defined above) after statistical feature selection. For survival models, we evaluated performance with the concordance index (defined above) using five-fold cross validation.

Table S1. Missing data among predictor variables in PRESTO.

Variable Name	Participants with missing data	
	N	% (out of 8,739)
Categorical Variables		
Handedness	14	<1%
Female smoking status	9	<1%
Male smoking status	9	<1%
Menstrual cycle regularity (initial)	7	<1%
Tried to get pregnant for 12 months or more	5	<1%
Menstrual cycle regularity (recent)	5	<1%
Binary variables		
Mother's history of miscarriage	1998	22.9
Mother's history of pregnancy problems	1394	16.0
Male partner circumcision status	1055	12.1
Secondhand smoking status (current, at home)	820	9.4
Last method of contraception, barrier methods	819	9.4
Last method of contraception, natural methods	819	9.4
Participant was born preterm	803	9.2
Secondhand smoking status (current, at work)	742	8.5
Conceived through fertility treatment	729	8.3
Participant was born with low birth weight	680	7.8
History of an abnormal Pap smear	566	6.5
Secondhand smoking status (age 0-10, at home)	550	6.3
Ever visited a physician for difficulty getting pregnant	527	6.0
Mother's history of C-section for participant's birth	341	3.9
Participant was a twin/triplet	161	1.8
Working rotating shifts	157	1.8
Continuous		
Ferriman-Gallwey score	4334	49.6
Waist measure	3304	37.8
Current e-cigarettes (ml/day)	3107	35.6
Duration participant was breastfed	2926	33.5
Number of lifetime sexual partners	2376	27.2
Total HEI-2010 score	2178	24.9
Mother's age at participant's birth	2010	23.0
Number of older siblings	1994	22.8
Mother's smoking history while pregnant (number of cigarettes)	1331	15.2
Male BMI	1012	11.6
Father's education	439	5.0
Household income	230	2.6
Mother's education	205	2.3
Night shift frequency in past month	198	2.3
Male education	170	1.9
Job hours/week	126	1.4

Note: All categorical variables are presented in this table; however, we only present continuous and binary variables with >1% missingness here.

Table S2. Distribution of gestational age at miscarriage in PRESTO, 2013-2022.

Gestational week at miscarriage	N(%)
Total	N=1,775
3	53 (3.0)
4	358 (20.2)
5	346 (19.5)
6	305 (17.2)
7	146 (8.2)
8	143 (8.1)
9	137 (7.7)
10	123 (6.9)
11	67 (3.8)
12	49 (2.8)
13	15 (0.8)
14	9 (0.5)
15	8 (0.5)
16	5 (0.3)
17	6 (0.3)
18	3 (0.2)
19	2 (0.1)

Table S3. Variables selected by the full survival model predicting miscarriage in PRESTO, 2013-2022.

Variable	Hazard Ratio[1] (95% CI)
Female age at conception (years)	1.20 (1.12, 1.29)
Female smoking: current regular smoker (ref = never smoker)	0.90 (0.84, 0.96)
History of miscarriage (yes/no)	1.11 (1.04, 1.18)
Geographic region of residence: Northeast US (ref = South US)	0.93 (0.87, 0.99)
Use of vitamin B7 (yes/no)	1.07 (0.99, 1.14)
Healthy Eating Index-2010 score (HEI-2010 score)	0.94 (0.89, 0.99)
Use of vitamin B6 (yes/no)	1.04 (0.99, 1.10)
Ever pregnant before (yes/no)	0.96 (0.87, 1.05)
Number of prior pregnancies	1.04 (0.98, 1.10)
Use of vitamin B1 (yes/no)	0.96 (0.89, 1.04)
Use of omega-3 or fish oil supplements (yes/no)	1.04 (0.99, 1.09)
Male age at conception (years)	1.04 (0.97, 1.10)
History of subfertility or infertility (yes/no)	1.03 (0.97, 1.11)
Ever received HPV vaccine	0.99 (0.94, 1.04)
Use of vitamin C (yes/no)	1.01 (0.96, 1.06)
History of unplanned pregnancy (yes/no)	0.99 (0.92, 1.07)
Previously tried to conceive for ≥12 months: “no, never tried before” (ref = “no”)	1.00 (0.92, 1.08)
Variables forced into the model[2]	
Female smoking: former smoker (ref = never smoker)	0.98 (0.93, 1.03)
Geographic region of residence: Canada (ref = South US)	0.98 (0.93, 1.04)
Geographic region of residence: West US US (ref = South US)	1.01 (0.95, 1.07)
Female smoking: current occasional smoker (ref = never smoker)	0.99 (0.94, 1.05)
Geographic region of residence: Midwest US (ref = South US)	1.00 (0.94, 1.07)
Previously tried to conceive for ≥12 months: “yes” (ref = “no”)	1.00 (0.94, 1.07)

Abbreviations: CI, confidence interval; HPV, human papillomavirus; US, United States.

[1] Continuous variables were standardized; the effect estimate is the hazard ratio for a one-unit increase in the z-score for that variable.

[2] For all models, we selected a reference group for each categorical variable that was recoded as an indicator variable 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 sparse model.

Table S4. Variables selected by the sparse survival model predicting miscarriage after restricting to ≥8 gestational weeks in PRESTO, 2013-2022.

Variable	Hazard Ratio[1] (95% CI)
Female age at conception (years)	1.17 (1.05, 1.30)
Male age at conception (years)	1.09 (0.98, 1.20)
History of unplanned pregnancy (yes/no)	1.07 (0.98, 1.16)
Healthy Eating Index-2010 score (HEI-2010 score)	0.96 (0.88, 1.04)

Abbreviations: CI, confidence interval.
Note: The Sparse and Full models were equivalent.
[1] Continuous variables were standardized; the effect estimate is the hazard ratio for a one-unit increase in the z-score for that variable.

Table S5. Variables selected by the sparse survival model predicting miscarriage among primigravid participants in PRESTO, 2013-2022.

Variable	Hazard Ratio[1] (95% CI)
Married (yes/no)	0.94 (0.88, 0.99)
Female age at conception (years)	1.07 (1.00, 1.14)
Last method of contraception was oral contraceptives (yes/no)	0.94 (0.88, 1.00)
Geographic region of residence: Northeast US (ref = South US)	0.94 (0.88, 1.01)
Male age at conception (years)	1.05 (0.98, 1.13)
Use of omega-3 or fish oil supplements (yes/no)	1.05 (0.99, 1.11)
Recent use of psychotropic medications (yes/no)	1.04 (0.98, 1.10)
Female BMI (kg/m^2)	1.04 (0.97, 1.10)
Healthy Eating Index-2010 score (HEI-2010 score)	0.97 (0.91, 1.03)
Variables forced into the model[2]	
Geographic region of residence: West US (ref = South US)	1.05 (0.99, 1.12)
Geographic region of residence: Canada (ref = South US)	0.99 (0.92, 1.05)
Geographic region of residence: Midwest US (ref = South US)	1.01 (0.95, 1.08)

Abbreviations: BMI, body mass index; CI, confidence interval; US, United States.

[1] Continuous variables were standardized; the effect estimate is the hazard ratio for a one-unit increase in the z-score for that variable.

[2] For all models, we selected a reference group for each categorical variable that was recoded as an indicator variable 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 sparse model.

Table S6. Variables selected by the sparse survival model predicting miscarriage after restricting to ≥8 gestational weeks among primigravid participants in PRESTO, 2013-2022.

Variable	Hazard Ratio[1] (95% CI)
Female age at conception (years)	1.07 (0.98, 1.18)
Male age at conception (years)	1.06 (0.96, 1.16)

Abbreviations: CI, confidence interval.
[1] Continuous variables were standardized; the effect estimate is the hazard ratio for a one-unit increase in the z-score for that variable.

Table S7. Variables selected by the sparse static model (logistic regression with an ℓ2-norm regularization term) predicting miscarriage after restricting to ≥8 gestational weeks in PRESTO, 2013-2022.

					Mean, by outcome status[1]	
Variable	**OR (95% CI)**	**β**	**Correlation with outcome**	**Overall mean (std.)**	***Miscarriage***	***No miscarriage***
Female age at conception (years)	1.19 (1.14, 1.23)	0.17	0.07	30.1 (3.8)	30.9	30.0
Male age at conception (years)	1.09 (1.06, 1.13)	0.09	0.06	31.9 (4.9)	32.9	31.9

Abbreviations: β, regression coefficient; CI, confidence interval; LR-L2, logistic regression model with an L2 penalty; OR, odds ratio (exp[β]); RFE, recursive feature elimination; std, standard deviation.

[1] These cells should be interpreted as the mean for each variable among individuals with or without miscarriage. For example, the average age of female participants who experienced a miscarriage was 30.9 years.

Table S8. Variables selected by the sparse static model (logistic regression with an ℓ2-norm regularization term) predicting miscarriage among primigravid participants in PRESTO, 2013-2022.

Variable	OR (95% CI)	β	Correlation with outcome	Overall frequency (std.) or mean (std.)	Frequency or mean, by outcome status[1]	
					Miscarriage	*No miscarriage*
Geographic region of residence: Northeast US (ref = South US)	0.88 (0.84, 0.92)	-0.13	-0.05	26% (44%)	22%	27%
Female age at conception (years)	1.12 (1.07, 1.17)	0.11	0.05	29.5 (3.5)	29.9	29.4
Married (yes/no)	0.90 (0.87, 0.94)	-0.10	-0.05	94% (24%)	91%	95%
Last method of contraception was oral contraceptives (yes/no)	0.90 (0.87, 0.94)	-0.10	-0.04	29% (45%)	25%	30%
Use of omega-3 or fish oil supplements (yes/no)	1.09 (1.06, 1.13)	0.09	0.04	18% (39%)	22%	18%
Recent use of psychotropic medications (yes/no)	1.09 (1.06, 1.13)	0.09	0.05	13% (33%)	16%	12%
Female BMI (kg/m2)	1.05 (1.01, 1.09)	0.05	0.03	26.3 (6.2)	26.7	26.2
Male age at conception (years)	1.04 (0.99, 1.09)	0.04	0.05	31.3 (4.6)	31.8	31.2
Healthy Eating Index-2010 score (HEI-2010 score)	0.96 (0.93, 1.00)	-0.04	-0.02	67.5 (9.0)	67.2	67.5
Variables forced into the model[2]						
Geographic region of residence: West US (ref = South US)	1.05 (1.01, 1.10)	0.05	0.04	15% (36%)	19%	15%
Geographic region of residence: Canada (ref = South US)	0.95 (0.91, 0.99)	-0.05	-0.01	17% (38%)	16%	18%
Geographic region of residence: Midwest US (ref = South US)	0.98 (0.94, 1.02)	-0.02	0.01	20% (40%)	21%	20%

Abbreviations: β, regression coefficient; BMI, body mass index; CI, confidence interval; LR-L2, logistic regression model with an L2 penalty; OR, odds ratio (exp[β]); RFE, recursive feature elimination; std, standard deviation; US, United States.

[1] These cells should be interpreted as the mean or percentage for each variable among individuals with or without miscarriage. For example, the average age of female participants who experienced a miscarriage was 29.9 years.

[2] For all models, we selected a reference group for each categorical variable that was recoded as an indicator variable 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 sparse model.

Table S9. Variables selected by the sparse static model (logistic regression with an ℓ2-norm regularization term) predicting miscarriage after restricting to ≥8 gestational weeks among primigravid participants in PRESTO, 2013-2022.

					Mean, by outcome status	
Variable	**OR (95% CI)**	**β**	**Correlation with outcome**	**Overall mean (std.)**	***Miscarriage***	***No miscarriage***
Male age at conception (years)	1.25 (1.20, 1.30)	0.22	0.06	31.3 (4.6)	32.3	31.2

Abbreviations: β, regression coefficient; CI, confidence interval; LR-L2, logistic regression model with an L2 penalty; OR, odds ratio (exp[β]); RFE, recursive feature elimination; std, standard deviation.