Supplementary Materials and methods

Multiple imputation

We performed multiple imputation with a fully conditional specification (FCS) to impute missing day of ovulation and covariates in 20 datasets (Liu and De, 2015). The FCS model was chosen because it does not assume a joint multivariate normal distribution and provides additional flexibility in handling continuous and categorical variables (Liu and De, 2015). Missing values were imputed using predicted mean matching because this method imputes values very similar to actual values in the dataset in terms of distribution and variable type (discrete, continuous or categorical).

We initially performed multiple imputation using the smaller sample size (N = 362). All variables that might predict missingness or would potentially be associated with miscarriage were included in the final model used to impute missing values. The multiple imputation model only included women who had a 25-hydroxyvitamin D [25(OH)D] measurement and a reported pregnancy outcome (these variables were included in the final model only to help predict the missing day of ovulation and were not imputed). Women who conceived may have contributed multiple menstrual cycles of observation prior to conceiving. We calculated several woman-specific variables across all of her observed cycles during the preconception period: the mean day of ovulation as measured by an ovulation prediction kit (including the conception cycle, if available), the mean day of ovulation as measured by cervical mucus and basal body temperature (categorized), mean cycle length in the daily diary (excluding the conception cycle which does not have a 'length'), mean alcohol intake, mean caffeine intake and mean duration of exercise. The imputation model also included time since estrogen use (categorized into groups: >3 months, 3 months, 2 months or 1 month). Variables specific to the conception cycle were alcohol and caffeine intake. Other variables included were age at the start of each menstrual cycle, time from the day of ovulation to a positive pregnancy test, race, BMI, education, gravidity, parity, smoking status at baseline, alcohol and caffeine intake reported at baseline. Participant characteristics included in the model were age, race, education, BMI and smoking status. The imputed values for the day of ovulation were used to determine gestational ages at the time of miscarriage.

We performed another multiple imputation using the entire dataset (N = 681) to impute missing 25(OH)D (exposure variable), gestational age at time of miscarriage (outcome) and covariates. We included all the variables included in the previous imputation except the pregnancy outcome variable. In addition, we included a variable which defined the gestational age from the time of a positive pregnancy test until a pregnancy outcome (live birth, miscarriage, induced abortion), prenatal supplement use at the time of blood draw and season of blood draw. We use the imputed gestational age from the pregnancy outcome variable to calculate the gestational age at miscarriage defined using the day of ovulation.

Diagnostics

The resulting 20 imputations were examined using trace plots and diagnostics were performed on five randomly selected datasets. We used kernel density plots to check the differences in distribution between the observed imputed datasets (Liu and De, 2015). The Kolmogorov-Smirnov test and t-tests were used to quantify the differences between observed and imputed data. We used P < 0.05 to detect a difference in distribution between the observed and imputed datasets.

Reference