Appendix 1: Statistical methods

This appendix describes the details of the statistical model used to estimate the association between birth defect outcomes and exposure to SSRIs, the meta-analysis approach used to summarize available information for development of prior distributions for the pertinent model parameters and the Markov Chain Monte Carlo (MCMC) algorithm used to derive estimates of model parameters. In addition, we provide a description of the approach used to assess the potential impact of missing information on analysis results.

Statistical model

We assumed that the log odds of participant *i* in the National Birth Defects Prevention Study (NDBPS) having a child with a specific birth defect, which we will call $\ln(odds_i)$, can be estimated using the model

$$\ln(odds_i) = \beta_0 + \beta_1 * E_i + \beta_2 * MatEd_i + \beta_3 * Race_i + \beta_4 * Obs_i + \beta_5 * S_i \quad .$$

In the above equation, the variable E_i = 1 if the woman was exposed to the SSRI of interest, $MatEd_i$ = 1 if the woman had 12 or less years of education, $Race_i$ = 1 if the woman reported race ethnicity other than non-Hispanic white, Obs_i = 1 if the mother has body mass index > 30 and S_i = 1 if the mother reported smoking from one month before to the end of the first trimester. If the woman reported values of these variables other than those assigned values of 1 as described, the variables were set to zero. The parameters, $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ and β_5 represent the log odds relating the defect outcome to the variables with β_1 being the parameter of primary interest with the exponentiated form corresponding to the odds ratio (OR) relating maternal exposure to the SSRI and the risk of a subsequent birth defect. Estimates for the parameters in the above model were developed using a Bayesian approach that requires specification of prior distributions for all model parameters. For the parameter relating the log odds of the birth defect outcome and exposure to the SSRI, β_1, these priors were developed based on available published results, when such results were available, using the meta-analysis approach described below. Priors for all other model parameters were assumed to be non-informative and were estimated using a Normal distribution with mean zero and large variance.

Meta-Analysis for Development of Prior Distributions for Model Parameters

Use of the Bayesian modeling approach required specification of prior distributions for all model parameters. Prior distribution estimates for the β_1 parameter were based on published estimated odds ratios (OR) and associated confidence intervals (CI) relating a birth defect and maternal exposure to a specific SSRI early in pregnancy. When two or more publications were available that presented information on the same exposure/birth defect association, a summary prior distribution was developed using a Bayesian meta-analysis. In the meta-analysis, the published natural logarithm of the OR estimate was assumed to be drawn from a Normal distribution with mean equal to the estimated log OR and variance derived from the estimated CI. To describe the model used in the meta-analysis, let $ln(OR_i)$ be the observed log odds ratio in study *i*, with associated standard error σ_i^2 which is assumed to be known and equal to that reported in the study. At the first level, we assume

$$ln(OR_i) \sim N(\mu_i, \sigma_i^2)$$

where

$$\mu_i = \mu + \gamma_i .$$

Here, μ represents the true underlying odds ratio relating the SSRI and the birth defect of interest and γ_i is a study-specific random effect. The model is completed by assuming the next level of priors such that

$$\mu \sim N(0, 0.67)$$

and

$$\gamma_i \sim N(0, \delta)$$

with a hyper-prior for the variance of the random effects given by

$$ln(\delta) \sim N(-2, 0.37).$$

Note that the assumed prior distribution for the underlying true OR implies a 95% credible interval for this parameter of approximately zero to 5. Assumptions on the variance of the study-specific random effects can be very influential and potentially bias analysis result[1]. To address this potential for bias, we used the prior for *ln(δ)* given above as suggested by Turner et. al.[2].

Prior distributions for the β_1 parameter in the model above were assumed to be Normal with mean and variance corresponding to the posterior mean and variance of the estimated true underlying log OR developed in the meta- analysis. For cases in which only one estimated OR relating the SSRI and the outcome of interest was available in the published literature, we assumed a Normal prior for β_1 with mean given by the log of the presented OR estimate and variance derived from its associated CI. In cases for which no appropriate OR estimate was available in the literature, we assumed a non-informative Normal prior for β_1.

Markov Chain Monte Carlo Estimates

Posterior estimates for the model parameters were developed using MCMC methods in which iterative sampling from a series of conditional distributions eventually produces samples from the desired posterior distribution[3]. In our application of the MCMC approach, we derived 100,000 iterative samples for each of the model parameters, discarded the first 10,000 samples to increase the likelihood of convergence to the posterior distribution and retained every fourth remaining sample up to 10,000 samples to reduce autocorrelation among the posterior samples. As a result, the posterior estimates of the model parameters presented in this paper correspond to 10,000 samples from the posterior distribution and are summarized using the median of the sampled values and a 95% equal tailed credible interval (CI) defined by the 2.5^{th} and 97.5^{th} percentiles of the 10,000 samples.

Missing data sensitivity analysis

The primary results presented here were derived using a complete case analysis that included only those NBDPS participants who reported values for all the variables used in the logistic regression model. While birth defect outcome was known for all study participants, women missing information on exposure to the medication of interest, race/ethnicity, maternal education, obesity and smoking were excluded from the analyses. Approximately 6-7% of the study population was missing information for at least one of these variables.

A complete case analysis provides unbiased estimates of the parameters of interest as long as the unobserved information is missing completely at random (MCAR)[4] . Under the MCAR assumption, women with missing information are a random sample drawn from all study participants. If the MCAR assumption is violated, that is, if the value of the missing information that would have been observed is associated with some collection of known or unknown attributes of the study participants, then the complete case analysis conducted here could result in biased estimates of the ORs. To assess this possibility, we conducted two sensitivity analyses focused on plausible violations of the assumption that information is MCAR among NBDPS participants. In the first assessment, we assumed that missing information on the predictor variables could be imputed based on the values of the other predictor variables that were observed for that participant. To do this, we fit a logistic regression model in which the 0 or 1 value for the missing predictor variable was estimated using available information on all observed values of the other predictor variables and on the birth defect outcome, that is we assumed the missing information was missing at random MAR[4]. For example, if smoking status was missing for a study participant, a value for her smoking status was estimated using available information on her race/ethnicity, education level, obesity status and if her child had the defect.

Using this imputation approach under the Bayesian model, we developed not only posterior estimates of the missing information but also estimates of the ORs of interest reflecting that imputation. This approach produces unbiased estimates as long as the missing information can be modeled using observed information available on that individual.

As an additional sensitivity assessment, we considered the possibility that the missingness can be related to what the value of that variable would have been had it been observed, beyond the level that can be accounted for by other observed information available. This situation is referred to as missing not at random (NMAR)[4] or informative missingness. To assess the potential for such informative missingness, we considered a scenario in which the probability that a missing variable has a given value depends on the case/control status of the subject missing that data. Under this scenario, the probability that a control participant has a value of one for any missing covariate is given by the proportion of controls with observed information for that variable having a value of one. For cases with missing information, however, we assume that the probability of a value of one for the missing variable is 0.5 times that proportion used for controls with missing data. For example, in this sensitivity assessment, we assume that controls with a missing value for exposure to the SSRI of interest were two times more likely than a case missing the same information to have actually been exposed to the SSRI. We focused on this model because it would tend to decrease the estimate of the ORs relating the predictor variables and to the birth defect outcomes. Therefore, if a relatively large OR is estimated in the complete case analysis but the corresponding OR estimate is substantially smaller when we assume this scenario for informative missingness, then it could be that the complete case estimate is primarily an artifact of the pattern of missing data among subjects as opposed to an unbiased estimate of the true level of association. It should be noted that these sensitivity assessments are based on unverifiable assumptions, that is, we cannot observe the true association between missingness probability and what the reported value might have been had it been observed. Our goal in conducting these assessments was to evaluate the impact of plausible assumptions on how information might be missing and what the potential implications are of these assumptions of the complete case analysis results[5].

References

1. Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian anal 2006:515-34 doi: 10.1214/06-BA117A[published Online First: Epub Date]|.
2. Turner RM, Jackson D, Wei Y, et al. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. Statistics in medicine 2015;**34**(6):984-98 doi: 10.1002/sim.6381[published Online First: Epub Date]|.
3. Lunn D, Jackson C, Best N, et al. *The BUGS book: a practical introduction to Bayesian analysis*: CRC Press, 2012.
4. Little RA, Rubin DD. *Statistical Analysis with Missing Data*. New York, NY: John Wiley and Sons, 1986.
5. Li T, Hutfless S, Scharfstein DO, et al. Standards should be applied in the prevention and handling of missing data for patient-centered outcomes research: a systematic review and expert consensus. Journal of clinical epidemiology 2014;**67**(1):15-32 doi: 10.1016/j.jclinepi.2013.08.013[published Online First: Epub Date]|.