SUPPLEMENTARY ONLINE MATERIAL

Folate supplementation and twin pregnancies

Stein Emil Vollset, Håkon K Gjessing, Anne Tandberg, Thorbjørn Rønning, Lorentz M Irgens, Valborg Baste, Roy M Nilsen and Anne Kjersti Daltveit.
Epidemiology, 2005.

This supplement provides a more detailed description of the statistical misclassification models used in the paper to account for underreporting of IVF and folate use, and to estimate the effect of folate intake on mono- and dizygotic twins separately when zygosity is unknown. We refer to the original paper for a fuller discussion of design, data collection, variable description and results.

Misclassification model

Variables

To obtain a more correct estimate of the effect of folate intake on the risk of having a twin birth it is necessary to account for the assumed 45% underreporting of folate use. Let F be a dummy variable indicating the true folate status (non-user/user). Since F is not observed directly, let F_o be the observed folate status. We then use $P(F_o|F)$ to denote the probability of observing a specific status conditional on the underlying true status. This conditional probability thus incorporates the 45% probability that a folate user should be reported as a non-user.

Since IVF is a strong confounder it is also important to include a correction for the assumed 12.7% underreporting of IVF. This is included in the model in the same manner as the folate misclassification, with I and I_o denoting the true and observed IVF status of the pregnancy, and $P(I_o|I)$ the misclassification probabilities.

While F and I are covariates with measurement error, it is also useful to consider measurement error in the outcome variable. Let Y denote the full outcome of a birth, which has four possible categories: 1 = Singleton, 2 = Monozygotic twins, 3 = Like-sex dizygotic twins, 4 = Unlike-sex dizygotic twins. If the full value of Y were observed the effect of folate on twinning rate could be studied directly from these categories. However, zygosity is not known, and we only observe Y_o with three categories: 1 = Singleton, 2 = Like-sex twins, 3 = Unlike-sex twins. In contrast to the full outcome, the monozygotic and dizygotic like-sex twins cannot be separated from the observed data. The misclassification probabilities $P(Y_o|Y)$ are trivially zero or one, determined by the contraction of the two groups of twins.

In addition to the above variables the model includes maternal age A as a categorical covariate.

Likelihood

We use maximum likelihood to estimate all parameters. The observed data consist of cell counts for all combinations of Y_o, F_o, I_o and A; we use a multinomial likelihood with cell probability

$$P(Y_o, F_o, I_o, A) = \sum_{Y,F,I} P(Y_o, F_o, I_o, Y, A, F, I),$$

where the right hand side is a sum of the full likelihood over all unobserved variables. The full likelihood can be factored as follows:

$$\begin{aligned} P(Y_o, F_o, I_o, Y, A, F, I) &= P(F_o|Y_o, I_o, Y, A, F, I)P(Y_o, I_o, Y, A, F, I) \\ &= P(F_o|F)P(Y_o, I_o, Y, A, F, I) \\ &= P(F_o|F)P(I_o|I)P(Y_o|Y)P(Y, A, F, I), \end{aligned}$$

by assuming non-differential misclassification for folate and IVF status (we assume the misclassification to be independent of both the outcome and of other covariates). When separating out the effects of the misclassifications, the remaining part of the likelihood is $P(Y, A, F, I) = P(Y|A, F, I)P(A, F, I)$. The cell counts for the unobserved covariates, $P(A, F, I)$, are estimated as separate parameters, without any assumption on their joint distribution. The core of the likelihood is the conditional probability $P(Y|A, F, I)$, which is the likelihood contribution in a model without misclassification. This is the part that contains information about effect parameters.

Parametrization
The conditional probability $P(Y|A, F, I)$ is modelled in two different ways:

Not accounting for zygosity:
The outcome Y is reduced to two values, $1 =$ Singleton, $2 =$ Twin, and $P(Y|A, F, I)$ is modelled as a logistic regression with independent variables F, I and A, with effects of F and A nested within I, i.e. we obtain one set of effect estimates for folate and age for each IVF status.

Accounting for zygosity:
Using the fact that like-sex and unlike-sex dizygotic twins have (practically) the same probability (Weinberg's rule), $P(Y|A, F, I)$ is modelled as a multinomial logit model, with effects of F and A nested within I, and with separate parameters for mono- and dizygotic outcomes.