Combining Matched and Unmatched Control Groups in Case-Control Studies

Saskia le Cessie, Nico Nagelkerke, Frits R. Rosendaal, Karlijn J. van Stralen, Elisabeth R. Pomp, and Hans C. van Houwelingen

Initially submitted February 8, 2008; accepted for publication July 14, 2008.

Multiple control groups in case-control studies are used to control for different sources of confounding. For example, cases can be contrasted with matched controls to adjust for multiple genetic or unknown lifestyle factors and simultaneously contrasted with an unmatched population-based control group. Inclusion of different control groups for a single exposure analysis yields several estimates of the odds ratio, all using only part of the data. Here the authors introduce an easy way to combine odds ratios from several case-control analyses with the same cases. The approach is based upon methods used for meta-analysis but takes into account the fact that the same cases are used and that the estimated odds ratios are therefore correlated. Two ways of estimating this correlation are discussed: sandwich methodology and the bootstrap. Confidence intervals for the pooled estimates and a test for checking whether the odds ratios in the separate case-control studies differ significantly are derived. The performance of the method is studied by simulation and by applying the methods to a large study on risk factors for thrombosis, the MEGA Study (1999–2004), wherein cases with first venous thrombosis were included with a matched control group of partners and an unmatched population-based control group.

bootstrap; case-control studies; control groups; matching; sandwich estimator; venous thrombosis

Abbreviation: MEGA, Multiple Environmental and Genetic Assessment of Risk Factors for Venous Thrombosis.
who were enrolled consecutively between March 1999 and September 2004. Two control groups were selected. Partners of the patients were used as an individually matched control group. A second population-based control group with the same age and sex distribution as the cases was acquired using random digit dialing.

In the MEGA Study, 2 separate analyses were performed to relate the occurrence of thrombosis to risk factors, like smoking (3). The cases were compared with their partners in a matched case-control analysis using conditional logistic regression. In the second analysis the cases were compared with the population-based controls, using unconditional logistic regression with adjustment for possible confounders, including age and sex.

In both analyses, odds ratios were estimated using only part of the data. Conditional logistic regression ignores the unmatched cases and controls, while the matched partners are not used in unconditional logistic regression. If the 2 sets of odds ratios measure the same quantity—that is, given “infinite” samples, they would be identical—we could combine all data and obtain 1 overall, more efficient estimate.

The need to combine data from different sources arises in many different contexts. Specific settings have been addressed previously—for example, pooling studies with unrelated cases and controls (4), conducting matched case-control studies with missing exposure data (5), and combining matched and unmatched controls to estimate gene-environment interactions (6, 7).

Combining all data and performing 1 overall analysis could be done by writing down the full likelihood of the data. However, this is complicated by the different sampling mechanisms used for the matched and unmatched subjects. Instead, a simple approach is introduced, where the estimates of the odds ratios from the analyses are pooled. This can be seen as performing a small meta-analysis, combining the 2 estimates from the separate analyses. A complicating factor, however, is that the same cases are used twice, which means that the estimated odds ratios in the separate analyses are correlated.

Below, we show how this correlation can be estimated and how an efficient combined estimate of the odds ratio can be obtained.

### POOLING TWO CORRELATED ESTIMATES

We consider the situation with 2 correlated estimates of the same parameter which we want to combine into 1 overall estimate. This is the situation in the MEGA Study, where 1 set of log odds ratios is obtained by conditional logistic regression using the matched pairs, while another set of log odds ratios is obtained from the unconditional logistic regression analysis. Here we assume that the correlation between the estimates is known; in the next section, we discuss how this correlation can be estimated.

The simplest case is 1 risk factor, with both models estimating the same odds ratio, without any confounders—for example, when the effect of a genetic variant is assessed. Here the estimated log odds ratios can be simply averaged.

If one of the log odds ratios is more precisely estimated—for example, because of a larger sample size for one of the control groups—a more accurate pooled estimate is obtained by taking a weighted average. The weighting then should take the standard errors of the 2 estimates into account. This can be done as follows.

Let \( \beta \) be the logarithm of the odds ratio. Let \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) be the 2 estimates of \( \beta \), with standard errors \( SE_1 \) and \( SE_2 \), and with \( \rho \) the correlation between the 2 estimates. The weighted average of the 2 estimates can be written as

\[
\hat{\beta}_{\text{pooled}} = w \hat{\beta}_1 + (1-w) \hat{\beta}_2. \tag{1}
\]

The optimal choice for the weight \( w \) is the one which minimizes the standard error of the pooled estimate \( \hat{\beta}_{\text{pooled}} \). The standard error of this estimate is

\[
SE(\hat{\beta}_{\text{pooled}}) = \sqrt{w^2 SE_1^2 + (1-w)^2 SE_2^2 + 2w(1-w) \rho SE_1 SE_2} \tag{2}
\]

This expression can be minimized with respect to \( w \) by taking the first derivative of equation 2. This yields

\[
w = (SE_2^2 - \rho SE_1 SE_2)/(SE_1^2 + SE_2^2 - 2\rho SE_1 SE_2) \tag{3}
\]

as most efficient choice of \( w \).

Often there are confounding variables. These confounders and their effects can differ in the separate analyses. For example, in the MEGA Study, where the investigators’ interest was in the effect of smoking, partners are more similar regarding lifestyles than unmatched controls, which could result in different effects of lifestyle confounders in the matched analysis. The effect of the risk factor of interest after adjustment for the confounders is assumed to be equal in the 2 analyses.

This situation can be written down formally. To simplify the formulas, we use matrix notation. Let \( \Theta_1 \) be the vector of log odds ratios of the first case-control analysis. This vector can be written as \( \Theta_1 = (\alpha_1, \beta) \), with \( \alpha_1 \) being the vector containing the log odds ratios of the confounding variables and \( \beta \) being the log odds ratio of interest. Similarly, \( \Theta_2 = (\alpha_2, \beta) \) is the vector of log odds ratios in the second analysis. The log odds ratio of interest \( \beta \) is shared by both models, while the log odds ratios for the confounders \( \alpha_1 \) and \( \alpha_2 \) are not shared. Only the estimates of \( \beta \) should be pooled. This pooling and calculation of standard errors can be done in the same way as described without confounders.

So far, we have considered only 1 exposure variable of interest, with 1 pooled odds ratio. We now consider the situation where more odds ratios are pooled simultaneously. This can occur if the exposure is categorical and several dummy variables are used to model its effect, or when the joint effect of 2 or more exposures is studied. Now \( \beta \) is a vector of log odds ratios. Let \( \hat{\beta}_1 \) be the vector of estimates from the first analysis and \( \hat{\beta}_2 \) be the vector of estimates from the second analysis. Suppose for a moment that we know the precision of the estimates, with \( C_{11} = \text{cov}(\hat{\beta}_1) \) and \( C_{22} = \text{cov}(\hat{\beta}_2) \). We also assume that the covariance matrix

\[
C_{ij} = \text{cov}(\hat{\beta}_i, \hat{\beta}_j) \tag{4}
\]
between the 2 estimated vectors of parameters is known: 
\( C_{12} = \text{cov}(\hat{\beta}_1, \hat{\beta}_2) \).

Again the estimates should be pooled in the most efficient way. In the situation with 1 \( \beta \) weights minimizing the standard error of the pooled estimate were used. One can show that this is equivalent to minimizing the weighted squared distance between the observed \( \hat{\beta} \) and the pooled \( \hat{\beta} \), weighting by the inverse of the covariance matrix of \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \). This can be readily extended to more \( \beta \)’s, considering the following regression model:

\[
\begin{pmatrix}
\hat{\beta}_1 \\
\hat{\beta}_2
\end{pmatrix} = N \left( \begin{pmatrix}
I_k & C_{11} & C_{12} \\
I_k & C_{21} & C_{22}
\end{pmatrix}, \begin{pmatrix}
C_{11} & C_{12} \\
C_{21} & C_{22}
\end{pmatrix} \right),
\]

with \( I_k \) being the \( k \)-dimensional identity matrix and \( k \) being the dimension of \( \beta \). The most efficient estimate of \( \beta \) is the weighted least-squares estimate:

\[
\hat{\beta}_{\text{pooled}} = \left[ \begin{pmatrix}
I_k \\
I_k
\end{pmatrix}^T \begin{pmatrix}
C_{11} & C_{12} \\
C_{21} & C_{22}
\end{pmatrix}^{-1} \begin{pmatrix}
I_k \\
I_k
\end{pmatrix} \right]^{-1} \begin{pmatrix}
C_{11} & C_{12} \\
C_{21} & C_{22}
\end{pmatrix}^{-1} \begin{pmatrix}
\hat{\beta}_1 \\
\hat{\beta}_2
\end{pmatrix},
\]

with \( C_{21} = C_{21}' \). If \( \beta \) is 1-dimensional, \( C_{11} = \text{SE}_1^2; C_{22} = \text{SE}_2^2 \), \( \text{SE}_1 = \text{SE}_2 \), and \( I_k = 1 \). Standard matrix calculus shows that this expression for \( \hat{\beta}_{\text{pooled}} \) then equals equation 1.

Standard errors and confidence intervals for the pooled \( \beta \)’s can be derived from the covariance matrix of \( \hat{\beta}_{\text{pooled}} \). Standard calculations show that this matrix equals

\[
\text{cov}(\hat{\beta}_{\text{pooled}}) = \left[ \begin{pmatrix}
I_k \\
I_k
\end{pmatrix}^T \begin{pmatrix}
C_{11} & C_{12} \\
C_{21} & C_{22}
\end{pmatrix}^{-1} \begin{pmatrix}
I_k \\
I_k
\end{pmatrix} \right]^{-1}.
\]

In practice, the matrices \( C_{11}, C_{12}, \) and \( C_{22} \) are unknown. Maximum likelihood estimation automatically yields estimates for \( C_{11} \) and \( C_{22} \), and most software packages calculate them routinely. The estimation of \( C_{12} \) is discussed below.

### ESTIMATING THE CORRELATION BETWEEN THE ESTIMATES

The correlation between the estimated log odds ratios \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) is needed to obtain the optimal pooled estimate with correct standard errors and confidence intervals. Since the joint distribution of the obtained data is unknown, robust methods are needed to estimate the correlation. We consider 2 ways to obtain robust estimates. The first is the “sandwich” estimator, which requires only that the mean response be correctly specified. It is often used in the field of longitudinal data when parameters are estimated with generalized estimating equations (8). An overview is given by Freedman (9).

Bootstrapping is another robust way to estimate the correlation. Here the joint distribution of \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) is estimated by repeatedly drawing samples with replacement from the original data set. Each resampled data set yields new estimates for \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \). The correlation between these bootstrapped estimates is then an (asymptotically unbiased) estimate of the correlation between \( (\hat{\beta}_1, \hat{\beta}_2) \) in the original data set, if the mean response is modeled correctly. Below we provide details on both methods.

### Sandwich estimate

To employ the sandwich method in our situation, we assume that in both analyses maximum likelihood is used to estimate the log odds ratios. Let \( l_i(\theta) \) and \( l_j(\theta) \) be the log-likelihood functions for both analyses, and let \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) be the corresponding maximum likelihood estimates. The log-likelihood functions can be written as the sum of independent parts: \( \sum_i l_i(\theta) \) and \( \sum_j l_j(\theta) \). In the MEGA Study, conditional logistic regression is used in the first case-control analysis, with summation over the matched pairs, while for unconditional logistic regression, summing is done over all cases and unmatched controls. Let \( U_i(\theta) = \partial l_i(\theta)/\partial \theta \) be the score function of the first likelihood—that is, the first derivative of the log likelihood with respect to \( \theta \)—and let \( U_j(\theta) = \partial l_j(\theta)/\partial \theta \) be the score function of the second likelihood function. We can write \( U_i(\theta) = \sum_i U_i(\theta) / \partial \theta \) and similarly for \( U_j(\theta) \). Furthermore, let \( I_i(\theta) \) and \( I_j(\theta) \) be the negative Hessian matrix, with \( I_i(\theta) = -\partial^2 l_i(\theta)/\partial \theta^2 \) and \( I_j(\theta) = -\partial^2 l_j(\theta)/\partial \theta^2 \). Details for the MEGA Study are given in the Web Appendix, which is posted on the Journal’s website (http://aje.oxfordjournals.org/).

By Taylor expansion, one can approximate \( \hat{\theta}_1 - \theta \) by \( \hat{\theta}_1 - \theta = I_1^{-1}(\theta) \sum_i l_i(\theta) \) and \( \hat{\theta}_2 - \theta = I_2^{-1}(\theta) \sum_j U_j(\theta) \). We use this to estimate the covariance between \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \):

\[
E \left[ \left( \hat{\theta}_1 - \theta \right) \left( \hat{\theta}_2 - \theta \right)^T \right] = I_1^{-1}(\theta) \times E \left[ \sum_i U_i(\theta) \sum_j U_j(\theta)^T \right] I_2^{-1}(\theta).
\]

The cross-product \( E[U_i(\theta)U_j(\theta)^T] \) is 0 for any \( i \) and \( j \), unless terms of the same subject are in \( U_i(\theta) \) and \( U_j(\theta) \). In our situation, only the matched cases are used in both the matched and unmatched analysis and occur in both \( U_i(\theta) \) and \( U_j(\theta) \). Hence,

\[
E \left[ \left( \hat{\theta}_1 - \theta \right) \left( \hat{\theta}_2 - \theta \right)^T \right] = I_1^{-1}(\theta) \times E \left[ \sum_{i \in M} U_i(\theta) U_j(\theta)^T \right] I_2^{-1}(\theta),
\]

where \( M \) is the set of matched cases.

We estimate this covariance matrix by means of a sandwich estimator, replacing the expected score products by the observed ones. Hence, \( E \left[ \sum_{i \in M} U_i(\theta) U_j(\theta)^T \right] \) is replaced by the observed cross-product between the score terms, \( \sum_{i \in M} U_i(\hat{\theta}_1) U_j(\hat{\theta}_2)^T \), and for \( I_1(\theta) \) and \( I_2(\theta) \), the observed
negative Hessian matrix in $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively, is used. This yields a sandwich estimate of the covariance matrix between the estimates:

$$\text{cov}(\hat{\theta}_1, \hat{\theta}_2) = I_1^{-1}(\hat{\theta}_1) \left[ \sum_{i \in M} U_{1i}(\hat{\theta}_1) U_{2i}(\hat{\theta}_2)^T \right] I_2^{-1}(\hat{\theta}_2).$$

(4)

From this the covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$ is directly estimated.

Bootstrap

Bootstrapping is an alternative way to estimate the correlation between estimates (10). By repeatedly sampling with replacement from the original data set, a collection of new data sets is obtained. In each of the resampled data sets, both case-control analyses can be repeated, yielding 2 sets of estimates. The correlation between these bootstrapped estimates is then an approximation of the correlation between $(\hat{\beta}_1, \hat{\beta}_2)$ in the original data set.

The manner of resampling should mimic the selection of the controls in the original data set. In the MEGA Study, the total number of cases was fixed, with matched controls being available for part of the cases, and a fixed number of unmatched controls were selected separately. Therefore, separate resamples are drawn from the cases and from the unmatched controls, and a matched control is sampled only when the corresponding case is sampled. This means that the fraction of matched pairs varies among bootstrap samples.

Some technical aspects

The covariance matrix $C_{12}$ between $\hat{\beta}_1$ and $\hat{\beta}_2$ is estimated with sandwich or bootstrap methods, while the matrices $C_{11}$ and $C_{22}$ are estimated using standard maximum likelihood methods. Therefore, the estimate of

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

could occasionally not be a valid covariance matrix, because it is not assured to be positive-definite. In this case, the pooled estimate of $\beta$ can be grossly wrong. A solution is to also estimate $C_{11}$ and $C_{22}$ using the sandwich or bootstrap method. This yields a positive-definite estimate of $C$, but the obtained standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ can differ from those of the separate analyses. An alternative estimate is $USUT$, where $S$ is the sandwich or bootstrap estimate of the total matrix $C$ and $U = \text{blockdiag}(C_{11}^{0.5}, S_{11}^{0.5}, C_{22}^{0.5}, S_{22}^{0.5})$, with $C_{11}$ and $C_{22}$ being the estimated matrices from the separate analyses. This matrix is always positive-definite and yields the same standard errors as the separate analyses. Checking whether the estimate of $C$ is positive-definite can be done by means of its eigenvalues. If any of the eigenvalues are negative, the matrix $USUT$ can be used instead to obtain a positive-definite covariance matrix.

**TESTING WHETHER THE ODDS RATIOS FROM BOTH ANALYSES ARE EQUAL**

In our analysis, we pool $\hat{\beta}_1$ and $\hat{\beta}_2$. This is appropriate if the same odds ratio is estimated in both analyses. In the MEGA Study, the partner controls are matched on several factors, including environmental and lifestyle factors. In the unconditional logistic regression, some of these factors can be included as covariates. When it is possible to include all matching factors, unconditional and conditional logistic regression estimate the same odds ratio (1). Systematic differences indicate that there are unmeasured risk factors that are implicitly controlled in the matched analysis but not in the unconditional logistic regression.

We can calculate confidence intervals for the difference between the odds ratios on the log scale and transform them back to the original scale. The observed difference between the 2 estimates $\Delta = \hat{\beta}_1 - \hat{\beta}_2$ has variance equal to $\text{var}(\Delta) = C_{11} - C_{12} - C_{21} + C_{22}$. This can be used to construct confidence intervals for $\hat{\beta}_1 - \hat{\beta}_2$. To formally test the hypothesis of no systematic difference ($\beta_1 = \beta_2$), the test statistic $T = \Delta^T \left[ \text{var}(\Delta) \right]^{-1} \Delta$ can be used, the distribution of which under the null hypothesis can be approximated by a chi-squared distribution with $k$ df. One could also use this to determine specifically whether one of the components of $\beta_1 - \beta_2$ differed from 0.

---

<table>
<thead>
<tr>
<th>Table 1. Results From a Simulation Study With 100 Matched Pairs and 100 Unmatched Controls$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>Conditional $\hat{\beta}_1$</td>
</tr>
<tr>
<td>Unconditional $\hat{\beta}_2$</td>
</tr>
<tr>
<td>Pooled using sandwich method $\hat{\beta}_{\text{pooled}}$</td>
</tr>
<tr>
<td>Pooled using bootstrap method $\hat{\beta}_{\text{pooled}}$</td>
</tr>
<tr>
<td>$\hat{\rho}$</td>
</tr>
<tr>
<td>$\hat{\rho}$</td>
</tr>
</tbody>
</table>

$^a$ The true $\beta$ was equal to 1. There were 1,000 simulation runs. The observed correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$ was 0.501.

$^b$ The median of the estimates of SE.
SIMULATION

To evaluate the performance of the pooling procedure, we conducted a simulation study. We simulated a large cohort from a logistic model with 2 covariates. The first variable, \( x_1 \), was a categorical covariate indicating whether an individual was exposed (\( x_1 = 1 \)) or not (\( x_1 = 0 \)). The exposure probability was set to 0.5. A second variable, \( x_2 \), was drawn from a uniform distribution and was used as a matching variable. Pairs of observations with the same value of \( x_2 \) were generated.

The outcome \( y \) was generated from a logistic model logit \( \Pr(Y = 1) = \alpha + \beta_1 x_1 + \gamma x_2 \), with \( \alpha = -7 \), \( \beta_1 = 1 \), and \( \gamma = 2 \). This corresponds to logits between \(-7\) and \(-4\) and probabilities on \( Y = 1 \) between \( 9 \times 10^{-4} \) and 0.017. From the large cohort, we drew 100 pairs with discordant outcomes: the matched case-control sample. A random set of 100 controls (observations with \( y = 0 \)) was also drawn.

In each of 1,000 simulations, the cases were compared with the matched controls using conditional logistic regression and with the unmatched controls using unconditional logistic regression. In the matched analysis, only \( x_1 \) was used as a covariate, because \( x_2 \) is constant within each pair and \( \gamma \) is not estimable. The unconditional logistic regression analysis used \( x_1 \) and \( x_2 \) as covariates. The covariance between \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) was estimated using the sandwich estimator and by bootstrapping with 500 bootstrap resamples. Table 1 shows the results.

Both the sandwich and the bootstrap correlation were often close to the observed correlation of 0.501. The sandwich correlation was slightly more variable (SE = 0.084 vs. SE = 0.069), but this did not influence the estimates of the pooled \( \hat{\beta} \), which were nearly always very similar, with means close to the true value 1. The last 2 columns of Table 1 show that the standard error of \( \hat{\beta}_{pooled} \) was estimated well. Calculated 95% confidence intervals contained the true parameter in 95.6% of the cases with the sandwich method and 95.7% of the cases with the bootstrap method. Ignoring the correlation between the estimates yielded standard errors that were too small in almost all simulations (median SE, 0.225; \( p5 = 0.208, p95 = 0.251 \)).

In 3 of the 1,000 simulations, the bootstrap yielded an estimate of

\[
C = \begin{pmatrix}
C_{11} & C_{12} \\
C_{21} & C_{22}
\end{pmatrix}
\]

which was not positive-definite. On these occasions, the pooled \( \hat{\beta} \) was widely off the true value 1, yielding estimates of 2.11, 2.13, and 2.83. All estimates using the sandwich estimator were positive-definite, and the largest estimated pooled \( \hat{\beta} \) was 1.82.

EXAMPLE: THE MEGA STUDY

We now consider data from the MEGA Study (3). In this example, we study the relation between smoking, coded as current or past smoker versus nonsmoker, and the occurrence of thrombosis. Age (continuous), sex, body mass

<p>| Table 2. Estimated Log Odds Ratios (( \hat{\beta} )) and Odds Ratios for the Relation Between Smoking and Thrombosis in Matched, Unmatched, and Pooled Analyses, Using Models Without Any Putative Confoundersa |
|---------------------------------|------|---------|----------|-----------------|----------------|</p>
<table>
<thead>
<tr>
<th></th>
<th>No. of Subjects</th>
<th>( \hat{\beta} )</th>
<th>Standard Error of ( \beta )</th>
<th>Odds Ratio</th>
<th>95% Confidence Interval</th>
<th>Correlation (( \hat{p} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conditional</td>
<td>4,572</td>
<td>0.194</td>
<td>0.069</td>
<td>1.21</td>
<td>1.06, 1.39</td>
<td></td>
</tr>
<tr>
<td>Unconditional</td>
<td>6,598</td>
<td>0.320</td>
<td>0.052</td>
<td>1.38</td>
<td>1.24, 1.52</td>
<td></td>
</tr>
<tr>
<td>Pooled using sandwich method</td>
<td>8,889</td>
<td>0.282</td>
<td>0.047</td>
<td>1.33</td>
<td>1.21, 1.45</td>
<td>0.31</td>
</tr>
<tr>
<td>Pooled using bootstrap method</td>
<td>8,889</td>
<td>0.281</td>
<td>0.047</td>
<td>1.32</td>
<td>1.21, 1.45</td>
<td>0.28</td>
</tr>
</tbody>
</table>

a Test for systematic difference between \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \): \( T = 3.02, 1 \text{ df}; P = 0.08 \).

<p>| Table 3. Estimated Log Odds Ratios (( \hat{\beta} )) and Odds Ratios for the Relation Between Smoking and Thrombosis in Matched, Unmatched, and Pooled Analyses, Using Models With Age, Sex, Body Mass Index, and Pregnancy Included as Confoundersa |
|---------------------------------|------|---------|----------|-----------------|----------------|</p>
<table>
<thead>
<tr>
<th></th>
<th>No. of Subjects</th>
<th>( \hat{\beta} )</th>
<th>Standard Error of ( \beta )</th>
<th>Odds Ratio</th>
<th>95% Confidence Interval</th>
<th>Correlation (( \hat{p} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conditional</td>
<td>4,572</td>
<td>0.238</td>
<td>0.073</td>
<td>1.27</td>
<td>1.10, 1.46</td>
<td></td>
</tr>
<tr>
<td>Unconditional</td>
<td>6,598</td>
<td>0.303</td>
<td>0.054</td>
<td>1.35</td>
<td>1.22, 1.50</td>
<td></td>
</tr>
<tr>
<td>Pooled using sandwich method</td>
<td>8,889</td>
<td>0.284</td>
<td>0.049</td>
<td>1.33</td>
<td>1.21, 1.46</td>
<td>0.30</td>
</tr>
<tr>
<td>Pooled using bootstrap method</td>
<td>8,889</td>
<td>0.284</td>
<td>0.049</td>
<td>1.33</td>
<td>1.21, 1.46</td>
<td>0.28</td>
</tr>
</tbody>
</table>

a Test for systematic difference between \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \): \( T = 0.74, 1 \text{ df}; P = 0.39 \).
DISCUSSION

We have introduced a simple method of combining estimates from several case-control analyses into 1 overall estimate. In ordinary meta-analysis, calculating the pooled mean of 2 studies yields exactly the same estimate as that derived using the individual data (11). Additionally, pooling log odds ratios yields an estimator which is asymptotically equivalent to the maximum likelihood estimate, using individual data. Therefore, we do not expect that much efficiency will be lost with the pooling approach.

We used the sandwich method and the bootstrap method to estimate the covariance matrix between the correlated β’s. Although both estimators are consistent, neither is unbiased. In addition, the sandwich estimator can be quite variable in comparison with parametric estimates (12). In the simulations, the sandwich correlation was slightly more variable, but the pooled estimates were very similar. The sandwich estimate is easy to calculate, and in all simulations it yielded a positive-definite covariance matrix. The bootstrap is more computer-intensive, and it occasionally yielded non-positive-definite covariance matrices. Whether the estimated covariance matrix \( C \) is positive-definite should be checked, because the pooled estimate can otherwise be very wrong. Alternatively, the method described above in “Some technical aspects” could be used.

Matching can be done using measured covariates, like age and sex, in which case it is clear which variables to include in the unconditional regression analysis. If matching is implicitly done on more factors than those measured, like the situation with siblings or partner controls, one should be careful in pooling the results. If there are unmeasured matching factors which are also related to the outcome, the conditional odds ratio, which is a stratum-specific odds ratio, is generally larger than the unconditional odds ratio. The test for equivalence of the odds ratios and the estimated difference with its associated 95% confidence interval can be used to study the magnitude of this problem. If the difference is not too large, pooling could still be done. As in meta-analysis, even if studies do not estimate exactly the same parameter, an overall pooled effect measure with a confidence interval gives an impression of the size of the effect.

Our example was of a large case-control study with a pair-wise matched control group and an unmatched control group. There are many other applications; for example, it is straightforward to apply these methods to control groups with other forms of matching or stratification. Another application is in genetic studies, where matching within families is done to study rare genetic effects and cases are combined with related and unrelated controls in order to estimate gene-environment interaction more efficiently (6, 7). We have used this method to combine family association data with results from an unmatched case-control analysis (13). The methods can also be useful if different confounders are measured in 2 control groups or confounders are measured in a different way. Instead of using complex imputation methods, the pooling methods can be used as an easy alternative.

A SAS macro (SAS Institute Inc., Cary, North Carolina) for pooling odds ratios using the sandwich method is available from the first author upon request.

ACKNOWLEDGMENTS

Author affiliations: Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands (Saskia le Cessie, Hans C. van Houwelingen); Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, the Netherlands (Saskia le Cessie, Frits R. Rosendaal, Karlijn J. van Stralen, Elisabeth R. Pomp); Department of Community Medicine, United Arab Emirates University, Al Ain, United Arab Emirates (Nico Nagelkerke); and Laboratory for Vaccine-Preventable Diseases, National Institute for Public Health and the Environment, Bilthoven, the Netherlands (Hans C. van Houwelingen).

Conflict of interest: none declared.

REFERENCES


