Problems due to Small Samples and Sparse Data in Conditional Logistic Regression Analysis

Sander Greenland,1 Judith A. Schwartzbaum,2 and William D. Finkle3

Conditional logistic regression was developed to avoid "sparse-data" biases that can arise in ordinary logistic regression analysis. Nonetheless, it is a large-sample method that can exhibit considerable bias when certain types of matched sets are infrequent or when the model contains too many parameters. Sparse-data bias can cause misleading inferences about confounding, effect modification, dose response, and induction periods, and can interact with other biases. In this paper, the authors describe these problems in the context of matched case-control analysis and provide examples from a study of electrical wiring and childhood leukemia and a study of diet and glioma. The same problems can arise in any likelihood-based analysis, including ordinary logistic regression. The problems can be detected by careful inspection of data and by examining the sensitivity of estimates to category boundaries, variables in the model, and transformations of those variables. One can also apply various bias corrections or turn to methods less sensitive to sparse data than conditional likelihood, such as Bayesian and empirical-Bayes (hierarchical regression) methods. Am J Epidemiol 2000;151:531-9.

bias (epidemiology); case-control studies; epidemiologic methods; logistic models; matched-pair analysis; odds ratio; regression analysis; risk assessment

A MATCHED-PAIR STUDY OF CHILDHOOD CANCER

In an analysis of electrical wiring and childhood cancers, Ebi et al. (7) examined data on backyard power lines among case-specular pairs in Denver, Colorado (a specular "case" is a case house and the "control" is a reflection of the case house across the street (8)). The exposure measure, backyard power line, had three levels: three-phase (primary) line, secondary line, and no line. This variable was entered into a conditional logistic regression model as two indicators for line type: $t_1 = 1$ if three-phase, 0 otherwise; and $t_2 = 1$ if secondary, 0 otherwise (9). This model can be expressed as a model for the odds of cancer at the exposure level indicated by $(t_1, t_2)$ relative to the odds when there is no exposure:

$$\text{OR}(t_1, t_2) = \exp(\alpha_1 t_1 + \alpha_2 t_2).$$

The first row of table 1 presents the estimates obtained by conditional maximum likelihood (CML) fitting of model 1 to 259 pairs. Placed in the subject-matter context, the odds ratio estimates of 32 and 14 and most of the points in the confidence intervals are absurd. While backyard power lines might show a stronger association because young children play in backyards near such lines, previous evidence is incompatible with such large odds ratios (10). For example, three-phase power lines produce much stronger fields...
than secondary lines, yet the odds ratio for a threephase line versus a secondary line is only 32/14 or 2.3, compared with 14 for a secondary line versus none.

Various epidemiologic biases could have contributed to these results. Nonetheless, below we argue that the odds ratio estimates were also likely to have been exaggerated by the analytic method.

**Bias in the CML odds ratios**

The CML estimators have infinite bias, because they can be infinite in some samples (6); what we refer to as "statistical bias" is more accurately described as skewing of their distributions toward values far above the true odds ratio. With this skewing, even modest true odds ratios can yield misleadingly large estimates with high probability when certain types of matched sets are infrequent, even if no epidemiologic bias is present.

The second row of table 1 presents estimates from an exact logistic regression program (11). These estimates differ little from the CML estimates, which might seem reassuring. Unfortunately, the point estimates supplied by the program are only slight refinements of the CML estimates (11), and so do not address our concern about the initial results.

Table 2 presents the pairs from which the estimates in table 1 were computed. It appears that the huge CML estimates arise from a single discordant pair with an unexposed case. This possibility can be addressed by examining the influence of small data changes on the estimates (12). The third row of table 1 gives the estimates after reclassifying as unexposed one of the 11 cases with secondary exposure and a three-phase specular. This change puts one pair into the empty cell for unexposed cases with a three-phase specular, and halves the estimates. Given such instability, even slight misclassification or selection bias could drastically affect the results. Even without epidemiologic biases, however, the presence of small counts indicates that large statistical biases may have affected the point estimates.

Many methods have been proposed for coping with small-sample deficiencies of conventional estimates (6, 12-21). The fourth row of table 1 presents estimates obtained by applying a bias correction (13-15) to the CML estimates. This correction results in a 40 percent drop in the odds ratio estimates, but the estimates remain large. This is partly because the correction only accounts for bias on the logarithmic scale, which is less severe than bias on the arithmetic scale (6, 16).

The fifth row of table 1 presents estimates obtained by applying another logarithmic-bias correction, the well-known Haldane correction, which in this example corresponds to adding $1/2$ to each cell in table 2 (6, 18, 21) and then multiplying the counts by a constant so that the total number of informative counts is restored to its original total (21). With six discordant-pair cells and 56 discordant pairs, adding $1/2$ to each cell adds $6(1/2) = 3$ to the total count, so the counts are multiplied by 56/59 to restore the total discordant-pair count to 56. In simple tables, the Haldane correction can remove most of the bias on the logarithmic scale (6, 16, 19, 20); here, it halves the odds ratio estimates.

The sixth row of table 1 presents estimates obtained by applying the Laplace/DeMorgan correction, which in this example corresponds to adding 1 to each cell (18) and then multiplying by 56/62 to restore the discordant-pair total (21). For large positive associations, the resulting estimates are similar in form to estimates obtained from arithmetic-bias corrections (6, 17, 19, 20). The Laplace correction has several advantages over those corrections, however, including absolute logarithmic invariance under exposure recoding. Here, it reduces the estimates by two thirds.

As well as having frequentist justifications (21), both the Haldane and Laplace corrections have Bayesian justifications (21).
derivations from simple types of prior distributions for the cell probabilities \((18, 21)\) and produce estimates of the log odds ratio that are shrunk toward zero \((21)\). The shrinkage decreases as the sample size increases, and can be large only when at least one cell is small (e.g., for a binary exposure, a 20 percent change typically requires a cell of 2 or less for the Haldane correction, 4 or less for the Laplace correction).

Because the bias in maximum likelihood estimates is proportional to the number of parameters in the model \((14, 15, 22)\), another approach to bias reduction is to simplify the fitted model. Although model simplification runs the risk of increasing bias due to model misspecification, evidence suggests that the latter bias would be of lesser magnitude than the sparse-data bias in the present example \((14, 23)\). The only way to simplify model 1 is to combine three-phase exposure and secondary exposure into one variable, so that the fitted model becomes

\[
\text{OR}(x) = \exp(\beta x).
\]

For example, one could set \(x = 0\) if there is no power line, 1 if there is a secondary line, and 2 if there is a three-phase line \((3)\); with this coding, \(\exp(2\beta) = \exp(\beta)^2\) is the three-phase odds ratio and \(\exp(\beta)\) is the secondary odds ratio. The seventh row of table 1 presents the resulting CML estimates. The model fits adequately, and the estimates are one third of those in row 1. If we think model 2 with \(x = 0, 1, 2\) is reasonable, this result is further evidence that the first-row results are statistically biased, because in large samples the estimates from model 1 should vary about the estimates from model 2 when \(x\) is correctly coded.

There are drawbacks in using model 2 as compared with the Laplace correction. Model 2 forces a somewhat arbitrary trend relation among the estimates: With \(x = 0, 1, 2\) in model 2, the three-phase odds ratio must be the square of the secondary odds ratio. In addition, with fewer exposure parameters, the estimates become more sensitive to details of model specification. The eighth row of table 1 presents estimates from model 2 with exposure recoded as \(x = 0, 1, 1.5\). The resulting estimates are approximately double those obtained from the \(0, 1, 2\) coding. We could even get model 2 to duplicate the categorical estimates from model 1 by using \(x = 0, 1, \ln(32/14)\).

The estimates in rows 4–8 of table 1 cover a wide range and do not represent all of the possibilities (e.g., one could compute jackknife estimates \((6)\)). Nonetheless, they all support the impression that the CML odds ratios are severe overestimates. The larger impact of the Laplace correction relative to the maximum likelihood and Haldane corrections occurs because the former is approximating a bias correction on the odds ratio scale, whereas the latter two are approximate bias corrections on the log odds ratio scale \((6, 12–14)\); the logarithmic transformation shrinks large values more than small ones, and hence requires less correction.

Other criteria for estimation

It has been argued that minimization of expected squared error \((\text{ESE})\), also known as mean squared error, is more relevant than minimization of bias \((21, 24, 25)\). This argument is based in part on the fact that \(\text{ESE} = \text{bias}^2 + \text{variance}\), which shows that the ESE incorporates both bias and variance components. Choosing to reduce ESE instead of bias leads naturally to Bayes estimates, as well as to more general procedures known variously as Stein, pseudo-Bayes, empirical Bayes, shrinkage, penalized, ridge, random coefficient, multilevel, or hierarchical estimation \((15, 18, 21, 24–30)\). Many of these estimates are rather complex, but a few simple ones are available for reducing ESE in estimating probabilities. One of the simplest parallels the Haldane and Laplace corrections by adding a constant \(c\) to each cell, but estimates the “best” constant from the data \((18, 21)\). This method is not applicable in the above example, because it requires that the model contain at least three free parameters. However, it is possible to apply Bayesian methods.

The ninth row of table 1 gives Bayes estimates for model 1 derived using penalized likelihood \((5)\) with a bivariate normal prior distribution for the secondary versus none and three-phase versus secondary log odds ratios, which are \(\alpha_2\) and \(\alpha_1 - \alpha_2\) in model 1. The prior mean values are zero, the prior variances are 1, and the prior correlation is 0.5, which implies that \(\alpha_1\) has a prior mean of zero and a prior variance of \(1 + 1 + 2(0.5) = 3\). For the final line, the prior distribution is modified to have mean values of \(\ln(2)/2 = \ln(1.4)\), so that the prior mean of \(\alpha_1\) is \(\ln(2)\). This alteration changes the results only slightly.

Both sets of Bayes estimates resemble the Laplace estimates; this is because all three methods treat the pair of increments \((\alpha_2, \alpha_1 - \alpha_2)\) as having a symmetric and broad prior distribution. For example, the prior used for line 9 assigns 95 percent prior probability to the propositions that the secondary odds ratio lies between \(1/7\) and \(7\) and the three-phase odds ratio lies between \(1/30\) and \(30\), because \(\exp(0 \pm 1.96 \cdot 1) = 1/7, 7\) and \(\exp(0 \pm 1.96 \cdot 3^{1/2}) = 1/30, 30\). Similarly, the prior used for line 10 assigns 95 percent prior probability to the propositions that the secondary odds ratio lies between \(1/15\) and \(15\), because \(\exp[\ln(2)/2 \pm 1.96 \cdot 1] = 1/5, 5\) and \(\exp[\ln(2) \pm 1.96 \cdot 3^{1/2}] = 1/15, 60\). Thus, both priors encompass a much broader range
than was suggested by previous studies (10). Nonetheless, their use has a dramatic impact on the results, because the CML estimates are so large and unstable.

Confidence limits

The pattern of confidence limits in table 1 parallels that of the point estimates, although the variation in the lower limits is much smaller than the variation in the point estimates, even on the logarithmic scale. The CML limits were computed from the estimated standard errors of the CML estimates. Likelihood ratio (not shown) and exact (table 1) limits did not appear more plausible. The failure of exact limits to address the apparent estimate inflation requires that we scrutinize the meaning of and rationale for exact intervals. Exact confidence intervals are guaranteed to contain the true value at least 95 percent of the time, even in small samples (assuming that no epidemiologic bias is present) (1, 5). Nonetheless, as the results in table 1 show, confidence limits (even exact ones) do not capture many sources of uncertainty, such as sensitivity to model choice and to minor data perturbations (12, 31).

Another important problem is that confidence interval theory refers only to an infinite sequence of confidence limits, rather than any pair of limits (32). Because the data in question are just one sample, that theory says nothing about whether the true odds ratios are within the exact limits in table 1. Unfortunately, confidence limits are commonly misinterpreted as probable bounds for parameters. For example, upon examining the exact limits, many if not most people are inclined to think (if not say) that there is a 95 percent chance that the true odds ratio for three-phase lines is between 4.5 and 1,328 if no bias is present, and then ascribe the elevated range to confounding, selection bias, or misclassification. Probable bounds based on the data, however, are Bayesian posterior probability limits (32–34). Although ordinary confidence limits and Bayesian limits are usually similar when the maximum likelihood estimates are much more precise than the background information (32), they can be very different when the data are limited.

Exposing prior distributions

The misinterpretation of confidence limits as probable bounds is particularly misleading for small samples. Contrast the CML limits against the Bayesian limits in table 1. The CML limits approximate the Bayesian limits that would be obtained under an “improper” uniform prior distribution in which every log odds ratio—be it 0, ln(10), ln(1/10), ln(10^30), or ln(10^{-30})—is considered equally probable. In contrast, the Laplace and Haldane limits approximate the posterior limits under priors in which log odds ratios slowly decline in probability as one moves away from the null.

Figure 1 displays these priors as they would appear in a matched-pair study of a dichotomous exposure, along with a standard normal prior (mean 0, variance 1), as used in line 9 of table 1, a normal prior with variance 3, and a normal prior with variance 1/3, all scaled to have a maximum of 1. Which of these priors would be more appropriate in a study of wiring and childhood leukemia? Of approximately 20 studies on the relation between wiring or magnetic fields and childhood can-

![Figure 1](image-url)
cer, few have produced relative risk estimates above 2, so odds ratios above 4 should seem improbable a priori. We may thus view the CML and exact limits in the above example as Bayesianly inappropriate: They give the incorrect impression that a three-phase odds ratio below 4 should seem improbable in light of the data, when in fact the data are so weak that they should not much increase our (low) probability that this odds ratio is above 4. This failing arises precisely because their Bayesian interpretation assumes incorrectly that all possible log odds ratios are a priori equiprobable.

From a frequentist perspective, one potential benefit of using Bayesian methods is that they can yield more precise interval estimates without sacrificing coverage validity, provided that the true parameter value is assigned a relatively high proportion of the maximum prior probability. Figure 1 gives an idea of the situation with a log odds ratio. The standard normal prior appears satisfactory for the example, but it would be a poor choice if the odds ratios could be much larger than 4. In contrast, the Haldane prior appears to be satisfactory beyond the largest odds ratio in the figure.

A MATCHED-SET STUDY OF GliOMA

Bias in CML estimates can become more severe as more covariates are added to the model, raising the following scenario: The investigator adds covariates, sees large increases in the exposure coefficients, interprets these changes as evidence of confounding, and so presents the enlarged coefficients. Another scenario is that, in a dose-response or induction analysis, the investigator moves exposure or timing cutpoints, sees large increases in the exposure coefficients, interprets these increases as evidence of having located the "most relevant" cutpoints, and thus presents the enlarged coefficients. In both scenarios, the enlargement might only be due to enlarged statistical bias. The following example provides data in which either scenario could have occurred.

Table 3 presents results from a case-control study of caloric intake and glioma (a brain tumor) based on 40 matched sets (32 matched pairs plus eight matched triplets with two controls each, matched according to sex, race, and 5-year age categories) (36). The first line in table 3 gives the odds ratio comparing control quartiles from a CML logistic regression with indicators for three calorie categories and body mass index (weight (kg)/height (m)$^2$) as a single continuous variable. As in the electrical wiring example, the odds ratio appears unusually large. Here, however, there are no comparable background studies. For comparison, radiation treatment of the head is considered a strong risk factor for adult brain tumors, with an average odds ratio of approximately 4 being seen in six studies (37). Among other cancers, no calorie-intake odds ratio of 4 or more has been found. Unlike electrical wiring in the first study, caloric intake was assessed by questionnaire, which raises the possibility that the large odds ratio arose from a mix of recall bias and statistical bias. Our further analyses will provide an idea of just how much statistical bias may have contributed to the observed estimate.

**TABLE 3.** Odds ratio estimates for caloric intake from a matched case-control study of glioma*

<table>
<thead>
<tr>
<th>Method</th>
<th>Comparison</th>
<th>OR</th>
<th>95% CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. CML†, control quartiles‡</td>
<td>Top quartile vs. bottom quartile</td>
<td>21</td>
<td>1.9, 233</td>
</tr>
<tr>
<td>2. No. 1 with body mass index deleted</td>
<td>2,500 calories/day vs. 1,000 calories/day</td>
<td>15</td>
<td>2.1, 113</td>
</tr>
<tr>
<td>3. Exact, control quartiles§</td>
<td></td>
<td>16</td>
<td>1.6, 938</td>
</tr>
<tr>
<td>4. No. 3 with body mass index deleted</td>
<td></td>
<td>12</td>
<td>1.8, 170</td>
</tr>
<tr>
<td>5. CML, total-sample quartiles¶</td>
<td></td>
<td>8.1</td>
<td>1.4, 46</td>
</tr>
<tr>
<td>6. No. 5 with body mass index deleted</td>
<td></td>
<td>8.8</td>
<td>1.7, 47</td>
</tr>
<tr>
<td>7. CML, direct calories</td>
<td></td>
<td>6.7</td>
<td>1.5, 31</td>
</tr>
<tr>
<td>8. CML, In(calories)</td>
<td></td>
<td>8.7</td>
<td>1.6, 47</td>
</tr>
<tr>
<td>9. CML, squared calories</td>
<td></td>
<td>4.5</td>
<td>1.2, 17</td>
</tr>
<tr>
<td>10. Null Bayes, direct calories</td>
<td></td>
<td>4.6</td>
<td>1.3, 16</td>
</tr>
<tr>
<td>11. Non-null Bayes, direct calories</td>
<td></td>
<td>5.2</td>
<td>1.5, 18</td>
</tr>
<tr>
<td>12. OML†, control quartiles</td>
<td></td>
<td>31</td>
<td>3.4, 274</td>
</tr>
</tbody>
</table>

* Data were obtained from Schwartzbaum et al. (36). All models included body mass index as a continuous variable unless otherwise indicated.
† OR, odds ratio; CL, confidence limits; CML, conditional maximum likelihood; OML, ordinary (unconditional) maximum likelihood.
‡ Body mass index $p = 0.020$ by likelihood ratio test.
§ Body mass index $p = 0.035$ by exact test; modified CML point estimate and exact limits (10).
¶ Body mass index $p = 0.045$ by likelihood ratio test.
The second line of table 3 gives the odds ratio estimated after deletion of body mass index. By any of the common statistical criteria (change in estimate, statistical significance of body mass index), body mass index would be judged to be a confounder and left in the model; yet it makes the calorie coefficients more implausible. Controlling three diet covariates instead of body mass index only further inflates the calorie odds ratio (36). Lines 3 and 4 provide estimates corresponding to the models in lines 1 and 2 from exact software (11). These estimates are only somewhat lower and still make body mass index appear to be a confounder.

While control of body mass index may have reduced confounding, it is plausible that this control produced more statistical bias. Evidence for the latter hypothesis is provided by examining the estimates in lines 5 and 6, which were obtained using total-sample calorie quartiles rather than control quartiles. This change cuts the estimates in half, suggesting either that the upper control quartile is more relevant biologically or that this quartile is too small to sustain conventional estimation. Note that use of total-sample quartiles rather than control quartiles reverses the direction of apparent confounding.

Table 4 gives data from the glioma study (36) which confirm the small-sample problem. Using control quartiles, there are only four matched sets that have a body mass index confounding. Using control instead of control quartiles reverses the direction of apparent confounding. Note that use of total-sample quartiles rather than control quartiles reverses the direction of apparent body mass index confounding.

Table 4 gives data from the glioma study (36) which confirm the small-sample problem. Using control quartiles, there are only four matched sets that have a control more exposed than the case; using total-sample quartiles, there are nine such sets, although these sets are still spread across six possible cells. In larger studies, the imbalance in numbers produced by using control quartiles rather than total-sample quartiles would be manifested as a loss of power and efficiency (38); with smaller samples it can also increase bias, as apparently happened here.

When the underlying exposure is continuous, categorization can be extremely inefficient (38). Line 7 of table 3 provides the CML estimate of the odds ratio comparing 2,500 calories/day with 1,000 calories/day from models with both calories and body mass index entered as continuous variables; lines 8 and 9 are from models with ln(calories) and squared calories instead of calories (2,500 calories vs. 1,000 calories represents approximately the same difference in intake as the difference between mean intakes in the top and bottom total-sample quartiles). The continuous-calorie estimates are very sensitive to calorie transformation and are still unusually high, suggesting that some bias remains.

One may further check for sample-size problems in the CML analysis by applying likelihood-ratio tests to the conventional (Wald) 95 percent confidence limits given by the program (39). If the sample size is adequate for CML, these tests should both yield a p value close to 0.05; failure to do so provides a warning that the CML estimates may incorporate statistical bias. For the first coefficient in table 3, these tests yield p = 0.017 and p = 0.096 for the lower and upper limits, respectively, which is symptomatic of severe bias. For the coefficient of direct calories (line 7), the tests yield p = 0.057 and p = 0.0009 for the lower and upper limits, indicating that the sample size is inadequate even for this simple CML regression, although the lower limit does not appear too distorted.

Because of the continuous data in this example, the simple bias corrections used above are no longer applicable. An approximate Bayes analysis via penalized likelihood (see Rothman and Greenland (5), chapter 21) is easy, however. Line 10 of table 3 gives the approximate Bayes estimate for 2,500 calories/day versus 1,000 calories/day with continuous calories and body mass index, derived using independent standard normal priors (see figure 1) for the log odds ratios for increments of 1,000 calories/day and 10 units (kg/m²) of body mass index. These priors are weak in subject-matter terms, but they shrink the calorie odds ratio substantially, though the odds ratio remains large. Line 11 repeats the Bayesian analysis, this time with means of ln(2) for both coefficients; this change to moderately positive priors yields an estimate only slightly higher than that obtained from the mean-zero (standard normal) priors.

Finally, to show that inflation can be even more severe in unconditional analysis, we provide on line 12

---

**TABLE 4.** Pair counts from a case-control study of caloric Intake and glioma*

<table>
<thead>
<tr>
<th>Case category</th>
<th>Control category</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Categories using control quartiles (boundaries were 1,164, 1,530, and 1,804 calories)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Mean no. of calories/day</td>
<td>950</td>
</tr>
<tr>
<td>Categories using total-sample quartiles (boundaries were 1,262, 1,642, and 2,045 calories)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Mean no. of calories/day</td>
<td>1,037</td>
</tr>
</tbody>
</table>

* Data were obtained from Schwartzbaum et al. (36); there were 40 cases and 48 controls. For the above tabulation (not the analysis), the eight 1:2 matched sets were incorporated into the pair counts as both possible types of case-control pairings.
of table 3 the results obtained using the control quartiles in an unconditional logistic regression analysis, with the matching factors controlled as indicator variables (one for sex and 12 for age) along with body mass index. The top-quartile estimate is now 50 percent higher than that obtained from CML.

DISCUSSION

The above examples involve very large odds ratios, and hence bias seems an obvious explanation. Nonetheless, CML bias can remain considerable in situations involving modest odds ratios, for which the problem might not be obvious (6, 35). Because of the algebraic equivalence of conditional logistic and Cox-model partial likelihoods, the same biases can afflict proportional hazards analyses. Analysts need to inspect their data in detail in order to alert themselves to the possible danger; likewise, readers should consider the possibility when presented with large maximum likelihood estimates derived from small numbers or from models that contain many covariates.

Several authors have warned about the variance inflation inherent in unnecessary matching or adjustment (40, 41). We have described a companion problem in which unnecessary control can exaggerate associations. The dilemma is how to determine when the problems introduced by further matching or adjustment begin to outweigh the confounding controlled. We know of no simple sufficient criteria. The total sample size or the ratio of sample size to the number of parameters is a common criterion (22), but it must be reformulated in terms of discordant matched sets if applied to conditional logistic regression. For example, in the case-specific data (table 2), the 259 pairs contained too few discordancies with unexposed cases to allow reliable inferences, even from a one-parameter model.

Sensitivity analyses are valuable for detecting sparse-data problems, especially when data displays are too cumbersome to capture multivariate relations. When continuous variables are present, these analyses should involve exploration of dose-response and adjustment models. When one is adding variables, the changes in maximum likelihood estimates from increasing statistical bias can overwhelm changes due to removal of confounding. Therefore, when results appear sensitive to reasonable choices for the analysis, the degree of sensitivity should be reported.

When exposure is the only variable in the model, the bias discussed here applies only to situations in which the underlying association is non-null; thus, it cannot by itself generate a univariate association. Nonetheless, an important possibility is the potential for synergistic interaction of statistical bias with other biases to produce spurious associations. Uncontrolled confounding, misclassification, and selection bias can easily generate odds ratios of 1.2 or even 1.5; statistical bias can then further inflate this odds ratio to a level that might appear too large to be completely explained by the other biases (35).

Remedies for sparse-data bias have various limitations. Exact intervals, while technically valid in the frequentist sense, can be very misleading from a broader scientific perspective, because they lack reasonable Bayesian interpretations when the data are sparse. Direct maximum likelihood bias corrections (13–15) are not available in most statistical packages, remove only logarithmic-scale bias, and are based on approximations that can fail to remove most of that bias (42). Adding a number to each cell is strictly appropriate only when all of the variables are truly discrete (not just arbitrary categorizations of continuous variables); to avoid excessive bias, the numbers need to be chosen carefully to reflect available background information. Bayesian methods can be applied more generally, but they also require thorough attention to background information to avoid excessive bias (although such attention can benefit an analysis).

Large estimates often do signal large effects, especially when found in the context of epidemics and outbreaks, since the latter signal the arrival of a strong cause at their location. In contrast, modern risk factor studies rarely begin with a clear signal in the noise of background incidence, and may slice their data thinly in the search for any signal. Both slicing and searching multiply the opportunities for statistical error and bias, as witnessed by how commonly large estimates arise from exploratory or subgroup analyses of weak effects (43–47).

Even without searching, the slicing implicit in confounder control can contribute to estimate inflation, as apparently happened in the glioma study and may often happen in other studies. For example, Dockerty et al. (46, p. 302) reported an unadjusted odds ratio of 7 (95 percent confidence limits: 0.8, 56) for the association of bedroom mean magnetic field exposure (>0.2 μT vs. <0.1 μT) with childhood leukemia, derived from 113 case-control pairs. On the basis of previous studies, this estimate is already implausibly high, but upon adjustment by conditional logistic regression the odds ratio became 18 (95 percent confidence limits: 1.6, 196); the published article gave neither the number of covariates nor the number of discordant pairs in this analysis. As another example, Abenhaim et al. (48) reported an odds ratio of 23 for the association between appetite-suppressant use over 3 months and risk of primary pulmonary hypertension, based on a conditional logistic regression with nine regressors but only five exposed controls; the unab-
justed odds ratio was not given. In this study, the large estimate does not conflict with background information but nonetheless warrants special scrutiny because of the extreme data sparseness.

Our observations have adverse implications for common confounder-selection strategies. As we noted above, the addition of potential confounders to a model may change estimates to an important degree because of the increase in bias caused by the addition. It may even happen that the control introduces more bias than it removes, as we think occurred in some of the studies cited above. Thus, common strategies can sometimes be detrimental in the analysis of sparse data.

Hierarchical methods are particularly well suited for modeling many variables at once and for checking whether results have been affected by sparse-data problems. Indeed, hierarchical methods can be viewed as techniques for controlling artifacts arising from too many parameters chasing too few data (21, 27, 49), and these methods have performed remarkably well in epidemiologic simulations and examples (24, 27–30, 44, 49–53). A routine for employing simple hierarchical regression in SAS is available on the World Wide Web (54).

If one is reluctant or unable to apply sophisticated techniques, one should still recognize that large estimates may signal statistical bias and investigate this possibility with data tabulations and sensitivity analyses. Researchers and instructors need to recognize that statistical biases may explain large estimates when data are sparse or when many variables are controlled, and can interact synergistically with other biases to inflate null or negligible estimates to apparently important values.

ACKNOWLEDGMENTS

This research was supported by the Wyeth-Ayerst Corporation (Philadelphia, Pennsylvania).

The authors thank Drs. Cande Ananth, Norm Breslow, Ron Brookmeyer, and James Robins for their helpful comments and Drs. Kris Ebi, David Savitz, and Luciano Zaffanella for the use of their data.

REFERENCES


