Validity of Methods for Model Selection, Weighting for Model Uncertainty, and Small Sample Adjustment in Capture-Recapture Estimation

Ernest B. Hook\textsuperscript{1,2} and Ronald R. Regal\textsuperscript{3}

In log-linear capture-recapture approaches to population size, the method of model selection may have a major effect upon the estimate. In addition, the estimate may also be very sensitive if certain cells are null or very sparse, even with the use of multiple sources. The authors evaluated 1) various approaches to the issue of model uncertainty and 2) a small sample correction for three or more sources recently proposed by Hook and Regal. The authors compared the estimates derived using 1) three different information criteria that included Akaike's Information Criterion (AIC) and two alternative formulations of the Bayesian Information Criterion (BIC), one proposed by Draper ("two pi") and one by Schwarz ("not two pi"); 2) two related methods of weighting estimates associated with models; 3) the independent model; and 4) the saturated model, with the known totals in 20 different populations studied by five separate groups of investigators. For each method, we also compared the estimate derived with or without the proposed small sample correction. At least in these data sets, the use of AIC appeared on balance to be preferable. The BIC formulation suggested by Draper appeared slightly preferable to that suggested by Schwarz. Adjustment for model uncertainty appears to improve results slightly. The proposed small sample correction appeared to diminish relative log bias but only when sparse cells were present. Otherwise, its use tended to increase relative log bias. Use of the saturated model (with or without the small sample correction) appears to be optimal if the associated interval is not uselessely large, and if one can plausibly exclude an all-source interaction. All other approaches led to an estimate that was too low by about one standard deviation. Am J Epidemiol 1997;145:1138-44.

Bayes theorem; bias (epidemiology); log-linear models; sample size

Capture-recapture methods in epidemiology have been increasingly used in recent years (1-3) since their utility for such application was first demonstrated and popularized by Wittes and coworkers (4-6). One major advance was the introduction of log-linear methods to adjust for source dependencies when data for three or more sources are available (7, 8).

In the application of log-linear methods to data on three or more sources, a major question that has received relatively little attention is the issue of "model selection." With three sources, there are eight different possible capture-recapture log-linear models, each associated with a distinct estimate. With four sources, there are 114 models; with five, there are 6,893; and with six, considerably more (1). Obviously, some models and associated estimates are better than others. If there are \( k \) sources, the best "fit" to the data is always provided by the "saturated" model (i.e., the one with zero degrees of freedom that models a \( k - 1 \) way interaction). This is most complex, however, and is associated usually with the widest confidence interval in relation to the size of the estimate. In this sense, it provides the most conservative estimate, but other simpler models associated with relatively narrower confidence intervals might well be preferable. In this paper, we evaluate five separate methods for model selection and two approaches that weight estimates associated with all models.

Another issue arises, particularly with data sets in which there are sparse or null cells, as to the value of a "small sample" adjustment. Chapman suggested some years ago a small sample adjustment for two sources that is now widely used. If there are two sources, denoted B and C, such that \( a \) is the number in both, \( b \) is the number in source B but not in source C, and \( c \) is the number in source C but not in source B, then he proposed that the unobserved cell total \( x \) be

\[ x = \frac{ab}{b+c} \]
Validity and Capture-Recapture Estimation 1139

estimated by $bcl(a + 1)$ rather than the maximum likelihood estimator, $bcl(a)$. Under a wide variety of conditions, this is optimal (9). Hook and Regal, using large scale simulations for three sources, found that the best adjustment among the many different possibilities considered was derived by adding the value of 1.0 to each cell that is the denominator of the expression for the unobserved cell predicted by the saturated model. (This is one natural extension of the Chapman adjustment from two sources.) From this, Hook and Regal (1) conjectured that, for $k$ or more sources where $k > 3$, the optimal adjustment would be, analogously, the addition of 1.0 to all cells that are in the denominator of the expression associated with the saturated model for $k$ sources. If the investigator uses data from an odd number of sources, then this procedure adds 1.0 to the values of those cells that are the intersection of an even number of sources. If there are an even number of sources, then this adds 1.0 to the values of those cells that are in the intersection of an odd number of sources. We also evaluate here the use of this proposed adjustment.

MATERIALS AND METHODS

Background

There are at least three different information criteria proposed for model selection. One uses "an information criterion" (AIC) proposed by Akaike, more usually denoted as the "Akaike Information Criterion" (10). The formula for this criterion in the context of log-linear methods is as follows:

$$AIC = G^2 - 2(df)$$  

(1)

where $G^2$ is the likelihood ratio statistic associated with the fit of any model to the data, and where df are the degrees of freedom of the model. The model with the lowest AIC is the one selected.

A second criterion, referred to as the "Bayesian Information Criterion" (BIC) (11, 12), was proposed by Schwarz. Recently, Draper (13) suggested a slight alteration. Although some have disputed that this alteration is preferable (14, 15), we consider both of these here, and, to distinguish them, denote them as SIC and DIC, respectively. The formulas for these, respectively, for any model and data set are as follows:

$$SIC = G^2 - (ln N_{obs})(df)$$  

(2)

and

$$DIC = G^2 - (ln (N_{obs}/2\pi))(df)$$  

(3)

where df is the degrees of freedom associated with any model. As with AIC, the model with the lowest associated value is the one selected. These may be denoted as "two pi" (Draper) or "not two pi" (Schwarz).

Such methods, of course, by choosing a single model, ignore the residual issue of "model uncertainty." The latter problem, which also arises in regression and other applications, has been described as "the Achilles' heel" of statistics (16). Its relative neglect has been termed a statistical "quiet scandal" (17). (See also references 18–20.)

Bayesian methods enable such an adjustment but at the price of considerable complexity. Draper (13) has proposed a relatively simple approach to this issue. He suggests that a composite estimate be derived from all models by weighting the estimate from each model $i$ by the amount: $[1/(exp(BIC/2))]$, so that if $\hat{N}_i$ is the estimate associated with model $i$, and BIC, is the Bayesian Information Criterion associated with model $i$ applied to the data in question, then the weighted estimate is as follows:

$$\hat{N}_{WBIC} = \sum (\hat{N}_i \cdot e^{-(BIC/2)}) / \sum e^{-(BIC/2)}$$

(4)

summing over all models.

As there are two competing formulations of BIC, DIC and SIC, we denote here the weighted methods using them as "weighted DIC" and "weighted SIC" and their associated estimates as, respectively, $\hat{N}_{WDIC}$ and $\hat{N}_{WSIC}$. (One might also define analogously a weighted AIC, although there appears no theoretical justification for such, in contrast to that provided for weighted BIC, provided by Draper.) In essence, the weighted BIC is an approximation to a Bayesian method with "flat" priors, that is, with each possible model accorded equal prior probability.

To evaluate the validity of competing methods, we extended a previous approach we used to compare capture-recapture with other methods of estimation (21). We explicate this with the data in table 1. Suppose one has four overlapping sources in a population, A, B, C, and D. These generate 16 nonoverlapping cells. We designate the names and the number observed in each as $a, b, c, d$. The context will make clear the particular meaning. $x$ always denotes the unobserved cell and the number within it. The number observed in each source is the sum of the observations in eight distinct cells as noted in table 1.

Suppose that one misplaced the data on the total in source A but did know the numbers in each cell in source A that was also in some other source (i.e., knew the number observed in cells $a$ through $g$). One could derive an estimate of the total in cell $h$ and consequently the total in source A, $\hat{N}_A$, with a three source capture-recapture analysis using data on the numbers in sources B, C, and D that are also in source A. One
### TABLE 1. Structure of data in four sources and of validity analysis of one source

<table>
<thead>
<tr>
<th>Source A yes</th>
<th>Source B yes</th>
<th>Source B no</th>
<th>Source B no</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source A no</td>
<td>Source A no</td>
<td>Source A no</td>
<td>Source A no</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
N_A &= a + b + c + d + e + f + g + h \\
N_B &= a + b + c + d + e + f + g + h \\
N_C &= a + c + e + f + g + h \\
N_D &= a + b + c + d + e + f + g + h \\
\end{align*}
\]

Capture-recapture analysis of total in source A above:

<table>
<thead>
<tr>
<th>Source B yes</th>
<th>Source B no</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source B no</td>
<td>Source B no</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\hat{N}_A &= a + b + c + d + e + f + g + x \\
\end{align*}
\]

\(x\) in this case is known = \(h\).

Validity analysis compares \(\hat{N}_A\) with \(N_A\).

---

can then compare this estimate, \(\hat{N}_A\), with the known total \(N_A\) to evaluate the accuracy of the particular method used. By comparing the accuracy of estimates associated with different methods of model selection and/or cell adjustment, one can evaluate these various approaches, at least with these data sets. Similarly one could take an analogous approach for sources B, C, and D (21).

---

### Data analyzed

We evaluated five published reports on four or more sources, that is, Bruno et al. (22), Frischer et al. (23), Domingo-Salvany et al. (24), LaPorte et al. (25), and Wittes (26), as cited by Bishop et al. (8). The data of the latter were originally gathered by Jacqueline Fabia for a dissertation on Down's syndrome in Massachusetts (27). She permitted Wittes to present and analyze selected data to illustrate a capture-recapture approach to epidemiologic investigation (J. T. Wittes, Columbia University, personal communication, 1982). Data in these studies were presented on five sources but, because one of them ("schools" or "V" as described by Bishop et al.) captured only 32 cases, we limit analyses only to the totals in the other four sources.

We compared the following seven different methods of model selection or adjustment: the estimates associated with the independent model; the estimate associated with the saturated model; the model with minimum DIC; the model with minimum SIC; the model with minimum AIC; the weighted DIC estimate; and the weighted SIC estimate. (For the reasons noted above, we did not evaluate weighted AIC.) For each type of model selection or method of adjustment we also compared the use of no small sample correction with the use of the correction proposed by Hook and Regal (1). Each of the five references provided data on four or five different overlapping populations of known size, which could be used to evaluate various approaches (three or four source) to capture-recapture estimation, so there were a total of 20 different "samples," not all independent, on each of which we compared 14 different methods of estimation.

We evaluated the mean and standard deviation of relative log bias, \(\ln B_r\), and the absolute value of the relative log bias, \(|\ln B_r|\), of all estimates associated with each particular method. If \(\hat{N}\) is the estimate and \(N\) is...
is the true number in a population, then these expressions are given, respectively, by

\[ \ln B_r = \ln(\hat{N}/N), \tag{5} \]

and

\[ |\ln B_r| = |\ln(\hat{N}/N)|. \tag{6} \]

We use relative log bias (rather than the bias itself) because, for example, for two estimates that are \( x \) and \( 1/x \) times the true population size (e.g., 10 times and 1/10), the magnitudes of the relative log biases are of the same magnitude, albeit in the opposite direction, and thus balance each other. This does not apply to bias itself. The rationale for using the absolute value of relative log bias is illustrated by noting that two values with equal but opposite values of \(|\ln B_r|\) will not only have the same mean value but also a variance and thus a standard deviation of zero. For any set of data such as that evaluated here, evaluation of log \( B_r \) will result in lower means and higher standard deviations than the use of \(|\ln B_r|\). Each provides a slightly different assessment of a particular method.

We also evaluated for each method how often the associated 90 percent confidence interval of any estimate included the known true population value. For estimates derived from the choice of a single model, we used the \( G^2 \)-based method described by Regal and Hook (28). For estimates derived from a weighted method (i.e., weighted DIC or weighted SIC), we combined the method of Regal and Hook with that of the weighted estimates. That is, we calculated the lower and upper limits associated with each model and then weighted these as we did the estimate in formula 4 above to derive weighted lower and upper intervals. We have used this method in previous application (1), although we did not describe the method explicitly there. If there are null entries in one of the cells that make up the denominator associated with the saturated (or complex) model and no small sample correction is used, then one may not be able to derive either a weighted estimate or a confidence limit, as some models, especially the saturated one, may be associated with an undefinably large estimate. For this reason, we limited some of the comparisons to 15 of the 20 data sets.

A reviewer has proposed in essence that we also compare these methods with the method of "internal validity" that we proposed previously (29). Such evaluation requires, however, data sets with not only at least five distinct sources but also every source with at least a moderate number of cases. Those available to us do not meet this criterion.

### RESULTS

Table 2 presents the results for both log \( B_r \) and \(|\ln B_r|\) for the 20 sample populations analyzed. In two data sets, the presence of null cells prevented calculation of a weighted estimate and corresponding interval for some sources, except when a small sample correction was made. Table 3 presents the proportion of 90 percent confidence intervals associated with each approach that yielded the true value. We discuss the trends in these data below.

**TABLE 2.** Means (standard deviations) of log relative bias and of absolute values of log relative bias

<table>
<thead>
<tr>
<th>Adjustment</th>
<th>Independent</th>
<th>Saturated</th>
<th>Minimum DIC*</th>
<th>Minimum SIC*</th>
<th>Minimum AIC*</th>
<th>Weighted DIC</th>
<th>Weighted SIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>No adjustment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subtotal = 15</td>
<td>-0.26 (0.19)</td>
<td>0.04 (0.30)</td>
<td>-0.16 (0.14)</td>
<td>-0.19 (0.16)</td>
<td>-0.12 (0.13)</td>
<td>-0.15 (0.12)</td>
<td>-0.18 (0.14)</td>
</tr>
<tr>
<td>Total = 20</td>
<td>-0.21 (0.19)</td>
<td></td>
<td>-0.12 (0.19)</td>
<td>-0.14 (0.18)</td>
<td>-0.09 (0.17)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjustment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subtotal = 15</td>
<td>-0.26 (0.20)</td>
<td>-0.07 (0.15)</td>
<td>-0.19 (0.19)</td>
<td>-0.20 (0.17)</td>
<td>-0.18 (0.13)</td>
<td>-0.19 (0.12)</td>
<td>-0.20 (0.15)</td>
</tr>
<tr>
<td>Total = 20</td>
<td>-0.22 (0.20)</td>
<td>-0.08 (0.20)</td>
<td>-0.15 (0.16)</td>
<td>-0.16 (0.17)</td>
<td>-0.13 (0.17)</td>
<td>-0.15 (0.15)</td>
<td>-0.16 (0.16)</td>
</tr>
</tbody>
</table>

* DIC, D. Draper's proposed Bayesian Information Criterion; SIC, G. Schwarz's Bayesian Information Criterion for model selection; AIC, an information criterion proposed by H. A. Akaike. (See Appendix regarding weighted AIC results.)
DISCUSSION

One must draw inferences cautiously from this analysis, because it depends critically upon the representative nature of the populations and sources upon which complete data were available. Nevertheless, they appear reasonably typical of the data sets likely to be generally available to epidemiologists for analysis. The small sample adjustment considered here did not show a consistent improvement. In all but one approach to model selection or adjustment, use of the correction did not affect or slightly worsened the relative bias or its absolute value. The correction, however, did improve the performance of the saturated model, not only improving the estimate but also diminishing its standard deviation.

One would, of course, expect the estimate associated with the saturated model to be particularly sensitive to small sample correction, for reasons discussed above. However, the result here with real “data” is counter to the result obtained with large scale simulations. In these, the adjustment analyzed (and all other candidate adjustments considered) was clearly preferable to nonadjustment (30).

There was not a marked difference among the five methods that used an information criterion. The two best were minimum AIC and weighted DIC. Among the two BIC approaches, DIC in all comparisons appears to be slightly better than SIC (a result that also tended to emerge from simulations). Judged by the mean value of the log bias, the weighted approach performed slightly better than did the equivalent minimum model. With log $B_n$, use of the weighted approach resulted in a rise in the mean but a diminishing of the standard deviation compared with $\log B_n$. We interpret these results as implying, at least with these data, a slight preference for use of the weighted methods rather than the associated minimum criteria.

The minimum AIC outperformed both minimum DIC and minimum SIC. An obvious possible approach, therefore, is use of weighted AIC, constructed analogously to weighted DIC and weighted SIC defined above. We had not considered that in our simulations nor in our analyses to date, because there appeared no published theoretical rationale for such an approach. However, in view of the results here, we also plan to evaluate this method in the future, based purely on these empirical observations. (See Appendix).

These analyses show unexpected noteworthy features. With the exception of the saturated model, all methods tend to result in an estimate that is too low by about 1.0 standard deviation. If these data sets are typical of those used by epidemiologists, then the results suggest that capture-recapture estimates derived tend to be biased low. This may indicate that 1) individuals captured by at least two sources are more likely to be caught by multiple sources (i.e., that sources used by epidemiologists tend to have net positive dependence), and/or 2) cases with low catchability by each source tend to have low catchability in general. There are likely to be exceptions, depending on how one constructs and defines the sources. (See Hook and Regal (1) for further discussion of “net dependence” and “source” in capture-recapture methods.)

A second point relates to the nature of the interactions detected by various criteria. The independent model by definition picks always the simplest model. While all three information criteria may in fact select the same model as optimal, if there is any difference among them, then (with more than about 50 cases) minimum SIC will tend to select a simpler model than minimum DIC, which will tend to select a simpler model than minimum AIC. The saturated model by definition always picks the most complex model.

Lastly, on the basis of these observations and previous simulations (30), we make the following recommendations. If there are many small or null cells in the sample (especially, of course, those in the denominator of the expression for the unobserved cell associated with the saturated model), then we suggest use of the correction proposed by Hook and Regal, particularly if
1) the total observed number of cases is relatively small and/or 2) a complex model (e.g., the saturated model) is used to derive an estimate. While use of the correction may make the estimate slightly worse in many cases, this effect is of small significance, whereas use of the correction may make a major improvement in selected but fewer instances. The correction appears likely to protect against extreme or infinite estimates at the cost of nonoptimality when the estimate is not extreme. This implies if the correction makes only a slight diminishment of an estimate, then there is little rationale to use it, and it may well be preferable in fact not to do so.

Use of the saturated model is in general optimal (except under the two conditions discussed further below). Its main disadvantage is the usually greater width of the associated confidence interval. However, the results in table 3 indicate that the "conservative" (i.e., wider) interval associated with this model is more likely to be valid. The investigator should be cautious about use of the saturated model if there are sparse sensitive cells, when the saturated estimate is very unstable, although the width of the confidence interval should in general adjust for that instability in this case. In addition, paradoxically, the investigator should not use the saturated model (or, in fact, any model) if an information criterion suggests that the saturated model is optimal, for if with k sources there is in fact a k way interaction in the data, then the saturated model (which adjusts only for a k - 1 way interaction) will in general appear to be optimal by an information criterion but may well be in serious error. On these grounds, it appears safest to attempt no capture-recapture estimate when the saturated model appears optimal, because of the strong possibility of an undetectable all source interaction in the data.

In results of as yet unpublished simulations, we did find that, if a simple model generated the data, then the saturated model did not perform as well as use of an information criterion (which would tend to choose simpler models). Of course, in that case, we knew the underlying relations in the population because we had specified them. The investigator with any "real" data set in general does not know the underlying relation of the sources and must conjecture their nature. Under these circumstances, any theoretical preferences for "parsimony" in model selection appear to be outweighed by considerations of "validity," at least with the data analyzed here.

Lastly, we emphasize that the conclusions here, though derived from "real" populations and not from simulations, must still be regarded cautiously as preliminary inferences until we gain greater experience with a much larger number of data sets.

ACKNOWLEDGMENTS

Supported in part by the Centers for Disease Control and Prevention (National Center for Health Statistics cooperative agreement U83/CU0908694 –02).

The authors thank Johanna Weber for very helpful assistance, as well as Patricia Podzorski, Janet T. Wittes for comments on the manuscript, and Richard Cormack for comments on limitations of the saturated model.

REFERENCES

21. Hook EB, Regal RR. Validity of Bernoulli census, truncated

APPENDIX

Note added in press (March 20, 1997): While this manuscript was in press, we also evaluated the performance of the weighted AIC approach, defined analogously to the weighted BIC approaches proposed by Draper. For the 15 data sets, analyzed with no adjustment, the mean (and standard deviation) of log relative bias was −0.10 (0.13). For the absolute value of log relative bias, the results were 0.13 (0.10). For the same 15 data sets analyzed with adjustment, these values were, respectively, −0.16 (0.11) and 0.17 (0.11). For all 20 data sets with adjustment the values were, respectively, −0.12 (0.14) and 0.15 (0.10). With regard to the 90 percent confidence intervals associated with each estimate, in seven of 15 cases (unadjusted), this calculated interval included the known estimate. In 12 of 20 cases (adjusted), the calculated 90 percent interval included the known estimate. With these data at least, the weighted AIC approach performs better than either of the alternative weighted BIC approaches evaluated and better than minimum AIC, but not as well as the use of the saturated model.