American Journal of Epidemiology
Vol. 149, No. 4
Printed in U.S.A.
Copyright © 1999 by The Johns Hopkins University School of Hygiene and Public Health
All rights reserved

Capture-Recapture Models Including Covariate Effects

Kate Tilling and Jonathan A. C. Sterne

Capture-recapture methods are used to estimate the incidence of a disease, using a multiple-source registry. Usually, log-linear methods are used to estimate population size, assuming that not all sources of notification are dependent. Where there are categorical covariates, a stratified analysis can be performed. The multinomial logit model has occasionally been used. In this paper, the authors compare log-linear and logit models with and without covariates, and use simulated data to compare estimates from different models. The crude estimate of population size is biased when the sources are not independent. Analyses adjusting for covariates produce less biased estimates. In the absence of covariates, or where all covariates are categorical, the log-linear model and the logit model are equivalent. The log-linear model cannot include continuous variables. To minimize potential bias in estimating incidence, covariates should be included in the design and analysis of multiple-source disease registries. Am J Epidemiol 1999;149:392–400.

capture-recapture methods; log-linear models; logistic models; models, statistical; registries

Accurate information on the prevalence or incidence of a disease may be required by epidemiologists, health services researchers, or health care planners and providers. One way to obtain such data is to set up a specialist registry, with the aim of registering all cases in a population. However, even when the methodology and resources for such a registry are well-established (such as with regional cancer registries in the United Kingdom and around the developed world), it is unlikely that every case will be registered. Thus, some estimate of the coverage of the registry, and hence of the true number of affected individuals, is needed (1).

The capture-recapture method was initially developed to estimate the size of wildlife populations. Animals are trapped, marked, and released on a number of occasions, and the individual trapping histories are then used to estimate the size of the population. The method has been increasingly used to estimate the size of human populations (such as the number of people with a given disease), using overlapping multiple sources of notification as the "captures." A major epidemiologic application has been to the estimation of the completeness of cancer registries (2, 3), but the method has been used to estimate the sizes of many populations, including homeless people (4) and children born with genetic disorders (5).

Received for publication October 20, 1997, and accepted for publication June 15, 1998.

Abbreviation: CI, confidence interval.

From the Department of Public Health Sciences, Guy's, King's and St Thomas's School of Medicine, London SE1 3QD, United Kingdom.

Two recent reviews of capture-recapture methods in epidemiology have emphasized the use of log-linear models (using Poisson regression) to investigate the relations between the sources used and the number of people "captured" (6, 7). This method makes two major and related assumptions about the probability of capture. Firstly, it is assumed that the capture probabilities for different notification sources are not all dependent. Thus, if there are only two sources, they are assumed to be independent. Applying standard capture-recapture methods to two dependent sources has been shown algebraically to underestimate the population size if the two sources are positively dependent, and to overestimate the population size in the case of negative dependence (8). Secondly, the probability of capture by a given source is assumed to be the same for each individual in the population.

These assumptions are violated when the probabilities of capture depend on covariates, such as severity of disease or age. One proposed solution using log-linear models is to stratify the data according to the covariate(s), estimate the total number of cases for each stratum, and then combine these estimates (6). Where covariate information is not available and there are more than two sources, multiple different two-source estimates can be made, considering each source against all of the others pooled (9).

An alternative approach, one taken by Alho (10), is to use a multinomial logit model to relate the characteristics of persons captured to their probability of being captured by each source. The probabilities of capture by the different sources are obtained by conditioning on

Downloaded from academic.oup.com/aje/article/149/4/392/59124 by Michigan State University user on 14 November 2026

being captured at least once. The resulting parameter estimates can be used to estimate the proportion of the population missed, assuming that the population has the same covariate distribution as the sample. Huggins (11) independently suggested a similar approach to the problem. Sanathanan (12) showed that this conditional estimate is asymptotically equivalent to the maximum likelihood estimator (based on the complete data, including the unobserved individuals) of the true population size.

METHODS

In this paper, we consider the problem of estimating the size of a closed population using two sources of notification (a single capture and recapture). We describe the use of logit models (based on the multinomial sampling scheme) and log-linear models (based on the Poisson sampling scheme) to estimate the population size. Each of these may be implemented using standard statistical software. We then extend the multinomial logit model to include the effect of individual-level covariates, and show that an equivalent log-linear model cannot be fitted when one of the covariates is continuous. Simulated data are used to compare estimates of population size derived from the different modeling approaches under varying conditions. We use bootstrap methods to obtain the 95 percent confidence intervals for the population size using each method, and compare their coverage. All computations were performed using Stata software (13).

Two sources, no covariates

The capture-recapture problem with two sources can be set out as in table 1, where x is the unknown number of missed cases.

The probabilities of being captured by source 1 (Π_1) and source 2 (Π_2) are assumed to be independent and identical for each individual. We write l_1 and l_2 for the logit of the probabilities of capture by source 1 and source 2, respectively. Assuming that the probabilities of capture by each source are independent, the probability of being missed by both sources is given by

$$\Pi_0 = (1 - \Pi_1)(1 - \Pi_2) = [(1 + e^{l_1})(1 + e^{l_2})]^{-1}.$$

TABLE 1. Notation for the number of cases captured by each source in a capture-recapture problem

Source 1	Source 2		
	No	Yes	Total
No	x	N_1	$x + N_1$
Yes	N_2	N_3	$N_2 + N_3$
Total	$x + N_2$	$N_1 + N_3$	

We use the indicator variable q_i to define whether the individual i is captured by no sources ($q_i = 0$), by source 1 only ($q_i = 1$), by source 2 only ($q_i = 2$), or by both sources ($q_i = 3$). Following the approaches of Alho (10) and Huggins (11), we condition on the observed individuals ($q \neq 0$) to estimate the model parameters, and then use the model parameters to estimate the population size. We use p_{ij} to represent the probability that $q_i = j$, conditional on the individual's being observed. We assume that the p_{ij}'s are identical for each individual i, and so we represent p_{ij} by p_j. Thus,

$$p_1 = \frac{\text{probability (in source 1 only)}}{\text{probability (observed)}}$$

$$= \frac{\Pi_1(1 - \Pi_2)}{\Pi_1(1 - \Pi_2) + \Pi_2(1 - \Pi_1) + \Pi_1\Pi_2}$$

$$= \frac{e^{l_1}}{e^{l_1} + e^{l_2} + e^{l_1 + l_2}}.$$

Similarly,

$$p_2 = \frac{e^{l_2}}{e^{l_1} + e^{l_2} + e^{l_1 + l_2}};$$

$$p_3 = \frac{e^{l_1 + l_2}}{e^{l_1} + e^{l_2} + e^{l_1 + l_2}}.$$

For a given contingency table, the likelihood following from the Poisson sampling scheme is proportional to the likelihood following from the multinomial sampling scheme (14). The log-likelihood for the multinomial model can be expressed as

$$\sum N_i \log(\pi_i)$$

and that for the Poisson model as

$$\sum N_i \log(\mu_i) - \mu_i,$$

where

$$\mu_i = E(N_i) = \pi_i(N_1 + N_2 + N_3) = \pi_i N.$$

The log-likelihood for the Poisson model can be expressed (15) as

$$\sum N_i \log(\pi_i) + \text{Poisson log-likelihood for } N.$$

Downloaded from academic.oup.com/aje/article/149/4/392/59124 by Michigan State University user on 14 November 2026

Thus, any differences between the estimates from a multinomial logit model and those from a Poisson log-linear model for a given table will arise not from differences in the likelihoods but from the way the models are parameterized and estimated. The equivalence of the two likelihoods also means that conditioning on the probability of being observed, which is implicit in the usual log-linear approach, yields an estimate which is asymptotically equivalent to the unconditional maximum likelihood estimate (12).

Estimation using the logit model. The systematic part of a logit model for the cell probabilities (p_{ij}) of the observed part of the table ($q_i \neq 0$) can be represented using parameters β_j defined by

$$p_j = \frac{e^{\beta_j}}{e^{\beta_1} + e^{\beta_2} + e^{\beta_3}}.$$

To identify the model, one of the parameters is set to zero. Thus, the systematic part of the model can be defined by

$$\beta_1 = 0;$$

$$\log\left(\frac{p_j}{p_1}\right) = \beta_j,\ j = 2, 3.$$

Relating the coefficients to the logits of the original capture probabilities, we find that

$$\beta_2 = l_2 - l_1;$$

$$\beta_3 = l_2.$$

Since $p_1 + p_2 + p_3 = 1$, we find

$$p_1 = (1 + e^{\beta_2} + e^{\beta_3})^{-1}.$$

We then assume independence of the original capture probabilities Π_1 and Π_2, to estimate the proportion of cases missed (Π_0). Using the subscript *lo* to denote the logit model without covariates, the proportion of cases missed is estimated by

$$p_{lo} = \frac{e^{\beta_2}}{(e^{\beta_3} + 1)(e^{\beta_3} + e^{\beta_2})}.$$

The total population size is estimated by

$$N_{lo} = \sum_{q_i \neq 0} \frac{1}{(1 - p_{lo})} = \frac{N_1 + N_2 + N_3}{1 - p_{lo}}.$$

The log-likelihood for this model is equivalent to

$$N_1 \log\left(\frac{1}{1 + e^{\beta_2} + e^{\beta_3}}\right) + N_2 \log\left(\frac{e^{\beta_2}}{1 + e^{\beta_2} + e^{\beta_3}}\right) +$$

$$N_3 \log\left(\frac{e^{\beta_3}}{1 + e^{\beta_2} + e^{\beta_3}}\right).$$

Estimation using the log-linear model. The log-linear model for the cell probabilities (p_j) of the observed part of the table ($q_i \neq 0$) can be defined using parameters α_j by:

$$\log(p_1) = \alpha_1. \quad (1)$$

$$\log(p_j) = \alpha_1 + \alpha_j,\ j = 2, 3. \quad (2)$$

Such a log-linear model, with the constraint that $p_1 + p_2 + p_3 = 1$, is equivalent (15) to the logit model described above, with

$$\alpha_1 = -\log[1 + e^{l_2 - l_1} + e^{l_2}];$$

$$\alpha_j = \beta_j,\ j = 2, 3.$$

Thus, the estimate of the number of cases missed obtained from fitting this log-linear model is equivalent to that obtained from the logit model.

The log-likelihood from this model can be expressed as

$$N_1\alpha_1 - e^{\alpha_1} + N_2(\alpha_1 + \alpha_2) - e^{(\alpha_1 + \alpha_2)} +$$

$$N_3(\alpha_1 + \alpha_3) - e^{(\alpha_1 + \alpha_3)},$$

which is equivalent to the multinomial log-likelihood plus the Poisson log-likelihood for the total number of people captured.

Because the logit model estimates only the ratio of the probabilities, whereas the log-linear model estimates the actual probabilities, there is no analogue for α_1 in the logit model. The constraint that $p_1 + p_2 + p_3 = 1$ means that α_1 must satisfy $\alpha_1 = -\log[1 + e^{\alpha_2} + e^{\alpha_3}]$.

An alternative parameterization of the log-linear model would be to fit the model with no constant term, by estimating

$$\phi_2 = \log(p_2) = \alpha_1 + \alpha_2;$$

$$\phi_3 = \log(p_3) = \alpha_1 + \alpha_3.$$

Downloaded from academic.oup.com/aje/article/149/4/392/59124 by Michigan State University user on 14 November 2026

The parameter $\alpha_1 = \log(p_1)$ can then be estimated using the constraint $p_1 + p_2 + p_3 = 1$. This alternative parameterization of the model leads to the same estimates of p_1, p_2, and p_3, and thus to the same estimate of the number of missing cases. However, the parameterization described in equations 1 and 2 emphasizes the fact that the logit model and the log-linear model lead to identical parameter estimates and thus to identical estimates of the number of cases missed.

Comparison of the logit and log-linear models. The log-likelihood from the Poisson log-linear model is equivalent to that obtained from the multinomial logit model plus the Poisson log-likelihood for the total number of cases captured. Thus, as shown above, the two models can be parameterized to give the same parameter estimates and standard errors. The Poisson model gives rise to an extra parameter. Both models can be fitted using standard statistical software.

Example 1. A data set consisting of 5,000 cases was generated, with capture probabilities (Π_1 and Π_2) given by

$$\text{logit}\,\Pi_1 = 0.5, \Pi_1 = 0.622;$$

$$\text{logit}\,\Pi_2 = 1, \Pi_2 = 0.731.$$

The true parameter values for the log-linear and multinomial logit models are therefore 0.5 (α_2 and β_2) and 1 (α_3 and β_3). For the log-linear model, the true value of α_1 is -1.68.

For these simulated data, the number of cases captured by each combination of sources is shown in table 2. Fitting a log-linear model to the observed cell counts of table 2 gives the following parameter estimates and 95 percent confidence intervals (95% CIs):

$$\alpha_1 = -1.6425 \text{ (95\% CI } -1.7089 \text{ to } -1.5760);$$

$$\alpha_2 = 0.4408 \text{ (95\% CI } 0.3557 \text{ to } 0.5260);$$

$$\alpha_3 = 0.9608 \text{ (95\% CI } 0.8827 \text{ to } 1.0389).$$

TABLE 2. Number of cases captured by each source in example 1, where the two sources are independent

Source 1	Source 2		
	No	Yes	Total
No	504	1,352	1,856
Yes	870	2,274	3,144
Total	1,374	3,626	5,000

The estimated number of cases missed is 517. The parameter estimates for each model are close to their true values.

Two dependent sources

We now assume that we have a covariate vector x_i for individual i. The probability of capture of individual i by source 1 and source 2 is given by Π_{1i} and Π_{2i}, respectively. We assume that these probabilities are independent. The capture probabilities are related to the covariate vector by the parameters l_i and parameter vectors k_i, such that

$$\text{logit}(\Pi_{1i}) = l_1 + k_1^T x_i;$$

$$\text{logit}(\Pi_{2i}) = l_2 + k_2^T x_i.$$

The probability (Π_{0i}) of individual i being missed is then given by

$$\Pi_{0i} = (1-\Pi_{1i})(1-\Pi_{2i}) =$$

$$[(1 + e^{l_1 + k_1^T x_i})(1 + e^{l_2 + k_2^T x_i})]^{-1}.$$

We define the indicator variable q_i as before. The probabilities p_{ij} of individual i falling into category j are related to the parameters l_i and parameter vectors k_i by

$$p_{i1} = \frac{e^{l_1 + k_1^T x_i}}{e^{l_1 + k_1^T x_i} + e^{l_2 + k_2^T x_i} + e^{(l_1 + l_2) + (k_1^T x_i + k_2^T x_i)}}; \quad (3)$$

$$p_{i2} = \frac{e^{l_2 + k_2^T x_i}}{e^{l_1 + k_1^T x_i} + e^{l_2 + k_2^T x_i} + e^{(l_1 + l_2) + (k_1^T x_i + k_2^T x_i)}};$$

$$p_{i3} = \frac{e^{(l_1 + l_2) + (k_1^T x_i + k_2^T x_i)}}{e^{l_1 + k_1^T x_i} + e^{l_2 + k_2^T x_i} + e^{(l_1 + l_2) + (k_1^T x_i + k_2^T x_i)}}.$$

Logit model. The multinomial logit model extends naturally to include covariate effects. The systematic part can be represented using the parameters β_j and the vectors of parameters δ_j, such that

$$p_{ij} = \frac{e^{\beta_j + \delta_j^T x_i}}{e^{\beta_1 + \delta_1^T x_i} + e^{\beta_2 + \delta_2^T x_i} + e^{\beta_3 + \delta_3^T x_i}}.$$

As before, this model is identified by setting one set of parameters to zero. We reparameterize the model such that

$$\beta_1 = \delta_1 = 0;$$

$$(p_{ij}/p_{i1}) = e^{\beta_j + \delta_j^T x_i}, \; j = 2, 3.$$

Downloaded from academic.oup.com/aje/article/149/4/392/59124 by Michigan State University user on 14 November 2026

Relating the coefficients to the logits of the original capture probabilities,

$$\beta_2 = l_2 - l_1;$$

$$\beta_3 = l_2;$$

$$\delta_2 = k_2 - k_1;$$

$$\delta_3 = k_2.$$

Since $p_{i1} + p_{i2} + p_{i3} = 1$, we find

$$p_{i1} = \left[1 + e^{l_2 - l_1 + (k_2 - k_1)^T x_i} + e^{l_2 + k_2^T x_i}\right]^{-1}.$$

We then assume that the probabilities of capture by each source are independent for each i, and thus estimate the probability of being missed (p_{0i}) for individual i with covariate x_i. Using the subscript lc to denote the logit model including the covariate effect, we estimate the total population size by

$$N_{lc} = \sum_{q_i \neq 0} \frac{1}{(1 - p_{0i})}.$$

The multinomial logit model can be parameterized and used in this way to estimate the population size regardless of whether the covariates are continuous or categorical.

Log-linear model. The systematic part of the log-linear model can be represented using parameters α_j and the vector of parameters γ_j, such that

$$\log(p_{i1}) = \alpha_1 + \gamma_1 x_i; \quad (4)$$

$$\log(p_{ij}) = \alpha_j + \alpha_1 + (\gamma_j + \gamma_1)x_i, \; j = 2, 3. \quad (5)$$

Where the covariates are categorical, a log-linear model equivalent to the logit model described above can be fitted (15). This is done by including the interaction between the variable (q_i) and each categorical covariate, together with the fullest interaction term among the covariates. This is equivalent to augmenting the covariate vector x_i with dummy variables. For each individual, one component of this vector is equal to 1 (corresponding to the overall category within which the individual falls), with the rest being zero. The parameters α_j and the vector of parameters γ_j then correspond to the parameters β_j and the vector of parameters δ_j, such that

$$\beta_j = \alpha_j, \; j = 2, 3;$$

$$\delta_j = \gamma_j, \; j = 2, 3.$$

This model is equivalent to the usual approach of stratifying by the categorical covariate. The estimate of population size derived from this model is identical to that derived from the logit model. As in the logit model, the assumption used is that the sources are independent given the level of the covariate.

Again, there is no analogue for α_1 or γ_1 in the logit model. As before, the constraint that $p_{i1} + p_{i2} + p_{i3} = 1$ means that α_1 and γ_1 can be calculated from the other parameters. As in the simple case, an alternative parameterization of the model omitting α_1 and γ_1 is possible and leads to the same estimates as the parameterization given by equations 4 and 5. Including α_1 and γ_1 shows more clearly the relation between the logit model and the log-linear model.

When there are continuous covariates, a log-linear model equivalent to the logit model cannot be directly fitted (16). Consider the log-linear model described in terms of coefficients α_j and γ_j above. It is not possible to express the probability p_{i1} (equation 3) in the log-linear systematic form (equation 4). This is because the denominator of p_{i1} contains the sum of terms in the exponential of x_i (equation 3). The log of p_{i1} cannot be decomposed into a term which is linear in x_i, as required by the parameterization of the log-linear model (equation 4).

Because

$$\log(p_{ij}) = \beta_j + \delta_j^T x_i + \log(p_{i1}), \quad j = 2, 3,$$

the probabilities p_{i2} and p_{i3} can be estimated by using p_{i1} as an offset in a log-linear model, once p_{i1} is known. Thus, the estimates of the parameters β_j and the vector of parameters δ_j from the log-linear model with offset p_{i1} and the logit model will be identical. However, the standard errors for the parameter estimates from the log-linear model with offset p_{i1} will be underestimated because the model treats p_{i1} as known.

Comparison of the logit and log-linear models. Where there are categorical covariates, these can be included in either a Poisson log-linear model or a multinomial logit model. The two models can be parameterized to give the same parameter estimates and standard errors, with an extra parameter for each level of the categorical covariate in the Poisson model. Where there are several categorical covariates, this can lead to a large number of extra parameters' being estimated when using the log-linear model. In this case, fitting the logit model directly would be computationally preferable.

Continuous covariates can be included in a multinomial logit model but not in an equivalent Poisson log-linear model. The only way to include such variables in a Poisson log-linear model would be to treat each level as a separate category, which would lead to the estimation of a large number of parameters and so to a

Downloaded from academic.oup.com/aje/article/149/4/392/59124 by Michigan State University user on 14 November 2026

maximum likelihood estimate which may be inconsistent or biased (16).

Example 2. As in the previous example, a data set of 5,000 cases was generated, but in addition, for each case, a single continuous covariate x_i was randomly drawn from the standard normal distribution. The probabilities of being captured by source 1 (Π_{1i}) and source 2 (Π_{2i}) were

$$\text{logit } \Pi_{1i} = 0.5 + 0.4x_i;$$

$$\text{logit } \Pi_{2i} = 1 - 0.3x_i.$$

The true values of the parameters are

$$\beta_2 = l_2 - l_1 = 0.5;$$

$$\beta_3 = l_2 = 1;$$

$$\delta_2 = k_2 - k_1 = -0.7;$$

$$\delta_3 = k_2 = -0.3.$$

The number of cases registered by each combination of sources, together with the mean and standard deviation of the covariate, are shown in table 3. The mean of the covariate is higher for those not captured by source 1 than for those captured by source 1, with the opposite relation holding between the covariate mean and capture by source 2. When the logit model including the covariate effect was fitted to the observed cell probabilities, the parameter estimates and 95 percent confidence intervals were

$\beta_2 = 0.4526$ (95% CI 0.3624 to 0.5427);

$\beta_3 = 0.9898$ (95% CI 0.9078 to 1.0717);

$\delta_2 = -0.7814$ (95% CI −0.8738 to −0.6891);

$\delta_3 = -0.3609$ (95% CI −0.4426 to −0.2793).

Each of these parameter estimates is close to its true value. Using this multinomial logit model, it is estimated that 482 cases were missed by the registry.

As discussed above, a log-linear model could not be fitted directly.

Simulations

The effect of dependence of capture probabilities on a covariate under each of the models was examined using simulated data sets. For each data set, 500 cases were generated, and for each case a single covariate x_i was randomly drawn from the standard normal distribution. Five hundred cases were chosen as an approximation of the size of the sample likely in many epidemiologic situations—large enough so that random variation will not completely dominate the results, but small enough for variation in the estimators to have an important effect on the estimate of the population size. The probabilities of being captured by source 1 (Π_{1i}) and source 2 (Π_{2i}) were generated according to

$$\text{logit } \Pi_{1i} = 0.5 + Ax_i;$$

$$\text{logit } \Pi_{2i} = 1 + x_i.$$

Five different situations were simulated, with A taking the values −1, −0.5, 0, 0.5, and 1. Each situation was simulated 100 times. For each data set, the population size was estimated using the crude logit model without including the covariate and the logit model adjusting for the continuous covariate. A categorical covariate was then derived by dividing the continuous covariate into quartiles. The population size was then estimated using a logit model including this categorical covariate. Bootstrap methods (with 500 replications) were used to derive 95 percent confidence intervals for each of the estimates of population size. We drew a sample of size N (with replacement) from the data set of N subjects and estimated the population size using each of the three methods for that replication. This was repeated for each of the 500 replications, and the standard deviation was calculated as the

TABLE 3. Number of cases captured by each source in example 2, and the associated covariate mean, where the probability of capture by each source depends on the continuous covariate

Source 1	Source 2								
	No			Yes			Total		
	No.	Mean	SD*	No.	Mean	SD	No.	Mean	SD
No	498	0.03	0.925	1,389	−0.36	0.995	1,887	−0.26	0.992
Yes	891	0.38	0.985	2,222	0.04	0.961	3,113	0.14	0.980
Total	1,389	0.26	0.978	3,611	−0.11	0.994	5,000	−0.01	1.00

* SD, standard deviation.

Downloaded from academic.oup.com/aje/article/149/4/392/59124 by Michigan State University user on 14 November 2026

standard deviation of the 500 estimates. Ninety-five percent confidence intervals were derived by taking the 2.5th and 97.5th percentiles of the values from the 500 replications.

RESULTS

Table 4 shows the mean estimates of population size, the mean estimates of the standard deviation, and the coverage of the nominal 95 percent confidence intervals (i.e., the proportion of the calculated confidence intervals including the true population size) obtained from the crude model (with no covariates). Where the two sources were independent ($A = 0$), the mean estimate from this crude model (500.1) was very close to the true population size (500). In this case, the coverage of the nominal 95 percent confidence intervals was 85 percent. However, once dependence between the sources was introduced, the crude model produced biased estimates.

The mean estimates of population size, the mean estimates of the standard deviation, and the coverage of the nominal 95 percent confidence intervals obtained from the logit model including the continuous covariate and the logit model including the categorical covariate are also shown in table 4. Each of these models produced less biased estimates of population size than the crude model, except where the sources were independent. In the independent case, the bias of the estimates obtained from either model including the covariate was small. The estimates obtained from these two models were similar, although the estimates obtained from the model including the continuous covariate seemed to be marginally less biased than those obtained from the stratified analysis when there was negative dependence between the sources, and more biased when there was positive dependence.

The overall coverage proportions of the bootstrap confidence intervals for each model (i.e., the proportion of confidence intervals containing the true number of cases missed) were consistently lower than the nominal value of 95 percent, even in the case of independence ($A = 0$). The coverage of confidence intervals from the logit model including the continuous covariate was higher than that of confidence intervals from the stratified analysis in four of the five situations simulated, even when the estimate of the population size was more biased.

DISCUSSION

For capture-recapture data without covariates, the usual log-linear model is equivalent to the logit model. Thus, either approach can be used in this case. The crude estimate of the population size can be seriously biased when there is dependence between the two sources. It is therefore necessary to examine the association of capture probability with measured covariates, and to include such covariates in capture-recapture analyses.

Both the log-linear model and the logit model can be extended to include the effects of categorical covariates. However, the need to include interaction terms and baseline probabilities means that this is more complicated for the log-linear model. For instance, with two covariates with four categories each, the logit model has 14 parameters, whereas the equivalent log-linear model has 30. While the logit model can include the effect of continuous covariates (10), there is no equivalent log-linear model.

In simulations, including covariate effects in capture-recapture models reduced the bias in the estimates of population size in comparison with the crude model. It appeared that the estimate from the model with the continuous covariate increased as the sources became more positively dependent, whereas that from the stratified model decreased, although the differences were relatively small. There was no significant trend in coverage of the confidence intervals with increasing dependence of the sources for either model. These results taken alone would seem to imply that the continuous method is least biased for sources with negative dependence and the stratified method is least biased for sources with positive dependence, although this needs further study. An idea of the direction of dependence on a given covariate can be obtained from a table showing the mean of the covariate for each combination of sources (table 3).

Bootstrap methods can be used to derive confidence intervals for the population size. However, in our simulations, the coverage of such confidence intervals was consistently lower than the nominal coverage of 95 percent. Such confidence intervals must therefore be interpreted with caution.

If there is a relation between covariates and the probability of capture, the distribution of the covariates will not be the same in the persons observed as in the entire population. This can be clearly seen in table 3, where the covariate distribution varies between those captured by source 1 only, those captured by source 2 only, those captured by both sources, and those not observed. Analyses (for instance, calculation of age-standardized incidence rates) based on the distribution of these covariates in the observed population might be biased. There is potential for using models including the covariates to estimate the distribution of the covariates in the population. Information about the relation between the capture probabilities and the

Downloaded from academic.oup.com/aje/article/149/4/392/59124 by Michigan State University user on 14 November 2026

TABLE 4. Estimates of population size and coverage* of 95% confidence intervals from three different models, for varying degrees of dependence of capture by each source on a continuous covariate

Dependence of logit of capture by first source on continuous covariate (*A*)	Mean estimate	Minimum	Maximum	Mean standard deviation	Coverage of 95% confidence interval (%)
Crude model					
−1	548.4	519.6	592.8	12.51	0
−0.5	524.6	501.7	561.0	10.28	30
0	500.1	471.3	535.7	8.30	85
0.5	478.7	453.1	499.6	7.08	31
1	461.5	430.1	489.6	6.21	4
Model including continuous covariate					
−1	502.0	478.0	538.8	7.96	89
−0.5	503.5	474.0	532.8	8.67	81
0	502.6	473.2	532.8	10.43	89
0.5	504.1	461.7	552.2	15.80	92
1	510.3	455.3	581.5	22.61	91
Stratified model					
−1	508.8	484.4	541.6	9.67	78
−0.5	505.2	483.5	539.5	9.59	81
0	503.2	472.5	558.1	11.14	86
0.5	499.2	464.7	531.1	13.35	88
1	495.1	451.7	557.4	16.67	84

* The percentage of the confidence intervals for the number of cases missed containing the true value.

covariates can also be used to target future registries. For example, if the probability of capture were inversely related to disease severity, an additional source specific to severely diseased people could be sought.

Multinomial logit models have not been used frequently in capture-recapture problems. One such application was estimation of the population size from the 1990 Post-Enumeration Survey in the United States (17). In addition to estimating the size of the population, the estimated probabilities of enumeration from the final logit model were used to identify "hard-to-enumerate" individuals. This showed the feasibility of applying the logit regression method including covariates to large samples, and that further information was available from the estimated probabilities of capture.

Models including covariates can only account for observable variability in capture probabilities. If there is dependence between sources which is not explained by observed covariates, the estimates resulting from the logit model will still be biased. A practical example concerns breast cancer notification by clinicians and by death certificates (8). Positive dependence between the two sources was suspected, due to more extensive contacts between clinicians and patients with more serious disease, who then went on to die. Where there are three or more sources, dependence between lists can be estimated. In such cases, one suggested approach is to combine those lists which show dependence (6) and then apply capture-recapture methods to the resulting aggregated sources.

Where there are three or more sources of notification and the covariates relating to the capture probabilities are not available, alternative models are available. These include a version of the Rasch model for the three-source problem, which has also been applied to the Post-Enumeration Survey in the United States (18). The model used has been shown to be equivalent to a quasi-symmetric log-linear model (19). It has been suggested (18) that when three or more sources are available, the analyst should stratify on the observable covariates associated with the capture probabilities and then use such a model within each stratum, to accommodate the unexplainable heterogeneity of capture probabilities. Thus, the use of the logit model including observable covariates is particularly important when there are only two sources of notification.

The performance of the population size estimates based on the logit model and the stratified model in practical situations has been little explored. In particular, the behavior of the estimates when the conditions shown by Alho (10) to imply the consistency and asymptotic normality of the maximum likelihood estimate (including the condition that both very small and very large probabilities of capture should be rare) are not met has not been examined. Practical examination of capture-recapture models including covariate

Downloaded from academic.oup.com/aje/article/149/4/392/59124 by Michigan State University user on 14 November 2026

effects, particularly where an independent estimate of the population size is available, is needed.

Failure to control for the effect of covariates associated with capture in the analysis can lead to biased estimates of the population size. Thus, potential covariates should be included in both the design of and the analysis of multiple-source disease registries. This will enable capture-recapture methods to be used more effectively in providing accurate estimates of registry coverage and incidence and prevalence of disease.

ACKNOWLEDGMENTS

Dr. Jonathan Sterne was supported financially by the UK Department of Health. Kate Tilling was supported by the Stroke Research Fund, the United Medical and Dental Schools of Guy's and St Thomas' Hospitals, and Glaxo Wellcome (Uxbridge, England).

The authors thank Dr. Charles Wolfe for co-supervision of this work.

REFERENCES

1. Benn RT, Leck I, Nwene UP. Estimation of completeness of cancer registration. Int J Epidemiol 1982;11:362–7.
2. Robles SC, Marrett LD, Clarke EA, et al. An application of capture-recapture methods to the estimation of completeness of cancer registration. J Clin Epidemiol 1988;41: 495–501.
3. Schouten LJ, Straatman H, Kiemeney LA, et al. The capture-recapture method for estimation of cancer registry completeness: a useful tool? Int J Epidemiol 1994;23:1111–16.
4. Fisher N, Turner SW, Pugh R, et al. Estimating numbers of homeless and homeless mentally ill people in north east Westminster by using capture-recapture analysis. BMJ 1994; 308:27–30.
5. Lie RT, Heuch I, Irgens LM. Maximum likelihood estimation of the proportion of congenital malformations using double registration systems. Biometrics 1994;50:433–44.
6. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. Epidemiol Rev 1995;17: 243–64.
7. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation. I. History and theoretical development. Am J Epidemiol 1995;142:1047–58.
8. Brenner H. Uses and limitations of the capture-recapture method in disease monitoring with two dependent sources. Epidemiology 1995;6:42–8.
9. Hook EB, Regal RR. Effect of variation in probability of ascertainment by sources ("variable catchability") upon "capture-recapture" estimates of prevalence. Am J Epidemiol 1993; 137:1148–66.
10. Alho JM. Logistic regression in capture-recapture models. Biometrics 1990;46:623–35.
11. Huggins RM. On the statistical analysis of capture experiments. Biometrika 1989;76:133–40.
12. Sanathanan L. Estimating the size of a multinomial population. Ann Math Stat 1972;43:142–52.
13. Stata Corporation. Stata statistical software: release 5.0. College Station, TX: Stata Corporation, 1997.
14. Birch MW. Maximum likelihood in three-way contingency tables. J R Stat Soc B 1963;25:220–33.
15. Agresti A. Categorical data analysis. New York, NY: John Wiley and Sons, Inc, 1990.
16. Baker SG. The multinomial-Poisson transformation. Statistician 1994;43:495–504.
17. Alho JM, Mulry MH, Wurdeman K, et al. Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. J Am Stat Assoc 1993;88:1130–6.
18. Darroch IN, Fienberg SE, Glonek GF, et al. A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. J Am Stat Assoc 1993;88: 1137–48.
19. Agresti A. Simple capture-recapture models permitting unequal catchability and variable sampling effort. Biometrics 1994;50:494–500.

Downloaded from academic.oup.com/aje/article/149/4/392/59124 by Michigan State University user on 14 November 2026