STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:2041–2051
Published online 7 April 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2081

Methods for capture–recapture analysis when cases lack personal identifiers‡

Betsy L. Cadwell[1,*,†] Philip J. Smith[2] and Andrew L. Baughman[3]

[1]*Division of Diabetes Translation, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, U.S.A.*
[2]*Immunization Services Division, National Immunization Program, Centers for Disease Control and Prevention, U.S.A.*
[3]*Epidemiology and Surveillance Division, National Immunization Program, Centers for Disease Control and Prevention, U.S.A.*

SUMMARY

Methods for estimating the size of a closed population from a capture–recapture study require the availability of unique identifiers on each of two lists. These identifiers are used to identify the number of individuals appearing on both lists. When the number of individuals appearing on both lists cannot be determined with certainty from the data, matching between the lists is problematic. In this paper, we develop a weighted estimator to account for all possible matches between two lists. A bootstrap procedure is proposed for estimation. To illustrate the methods, we used two lists that recorded New York State (NYS) hospitalizations due to pertussis in 1996 to estimate the number of persons hospitalized for pertussis in NYS that year. Published in 2005 by John Wiley & Sons, Ltd.

KEY WORDS: capture–recapture; matching algorithm; bootstrap; percentile interval

1. INTRODUCTION

Capture–recapture methods are used to estimate the size of a closed population from overlapping lists of cases [1–3]. After the cases on each list are enumerated and cases recorded on both lists are identified, an estimate is obtained for the number of cases that are not on either list. This estimate of the number of unknown cases is combined with the number of matched cases (recorded on both lists) and unmatched cases (appearing on a single list) to yield an estimate of the size of the population. To estimate the number of unknown cases, Lincoln and Petersen developed the maximum likelihood estimator which was first applied

*Correspondence to: Betsy L. Cadwell, CDC/NCCDPHP, MS K-10, 3470 Buford Highway, Atlanta, GA 30341, U.S.A. and CDC/NCCDPHP, 2858 Woodcock Blvd., Atlanta, GA 30341, U.S.A.
†E-mail: bcadwell@cdc.gov

Received October 2003
Published in 2005 by John Wiley & Sons, Ltd. *Accepted September 2004*

to estimate the size of human populations by Sekar and Deming [4–6]. Later the maximum likelihood estimator was adjusted for use with small samples [7, 8]. Bayesian methods for capture–recapture studies have been described by Smith [9–11].

To obtain a valid population estimate using capture–recapture methods with two lists, three conditions are assumed: for each list, the probability a case is recorded on the list is equal for all cases; the probability of being recorded on one list is independent of the probability of being recorded on the other; and there is no immigration or emigration to or from the population during the study period. Even if these assumptions are violated, Hook and Regal suggest the estimates may still be valid [1]. Regardless of whether these assumptions are met or not, many methods assume cases appearing on each list to be uniquely identifiable (e.g. discernable from all other cases on the list) such that a case recorded on both lists can be identified with certainty. Often at times a variable is available which is different for each individual recorded on a list. We refer to this variable as a unique identifier. If both lists utilize the same unique identifier then finding cases recorded on both lists with certainty can be done with relative ease by matching the unique identifiers from both lists. Cases that can be identified as existing on both lists are referred to as unique matches.

Epidemiological applications of capture–recapture methods utilize surveillance data or administrative lists [12–14]. Administrative lists often do not have sufficient amounts of specific information on individual characteristics to ensure cases are uniquely identifiable. When surveillance or administrative lists do not contain unique identifiers, there exist methods for matching cases to obtain capture–recapture estimates. Methods for matching cases include, empirically deciding upon a set of characteristics that define matches or conducting a record linkage project that uses probabilistic matching algorithm [15, 16]. Using a probabilistic matching algorithm involves choosing the one most likely match from many possible matches and does not explicitly account for variation attributable to other possible matches. Other limitations of record linkage projects include multiple passes through the data, exclusion of putative duplicate records, arbitrary decision rules to define matches and potential difficulty in replicating the results. Laska *et al.* describe methods for estimating population size in capture–recapture studies when records cannot be matched uniquely [17]. Regardless of which matching methods are used, discussion of the sensitivity of capture–recapture estimates to assumptions about the accuracy of matching is recommended [18].

In this paper we modify commonly used capture–recapture methods to overcome the limitations of capture–recapture analysis when cases may not be identified uniquely. In our methods we create profiles, or collections of characteristics, that describe one or more individuals on a list. For example, if a profile were defined by gender, birth month, and birth year, then a possible profile is males born in May 1976. For profiles appearing on both lists, we account for potential matches of cases between the two lists. Two methods for estimating population size are developed to account for potential matches: (1) a weighted estimator, which considers all potential matches, and (2) a bootstrap estimator, which simulates the weighted estimator by resampling profiles with replacement from all potential matches. We illustrate the use of the bootstrap estimator by estimating the number of pertussis hospitalizations reported during 1996 in New York State, using two data lists that do not share a unique identifier but both lists contain information on the gender, birth month, birth year, year of illness and month of illness for each hospitalization that can be used to create profiles.

2. METHODS

2.1. The Chapman estimator: unique matches

A unique match is defined as a comparison of a case on one list with a case from the second list that is for the same person. For two lists, A and B, the observed counts are: the number of uniquely matched cases on both lists A and B, X_{AB}, the number of cases on list A but not on list B, $X_{A\bar{B}}$, and the number of cases on list B but not on list A, $X_{\bar{A}B}$. The unknown number of cases on neither list A nor list B is estimated by

$$\hat{X}_{\bar{A}\bar{B}} = (X_{A\bar{B}} X_{\bar{A}B}/(X_{AB} + 1))$$

The modified Chapman Lincoln–Petersen estimate of the population total is

$$\hat{N} = X_{AB} + X_{A\bar{B}} + X_{\bar{A}B} + \hat{X}_{\bar{A}\bar{B}} \tag{1}$$

and its estimated variance is

$$\hat{V}(\hat{N}) \doteq \frac{(X_{AB} + X_{A\bar{B}} + 1)(X_{AB} + X_{\bar{A}B} + 1)\ X_{A\bar{B}} X_{\bar{A}B}}{(X_{AB} + 1)^2\ (X_{AB} + 2)}$$

2.2. The weighted estimator: non-unique matches

When characteristics on the two administrative lists do not allow individuals to be matched uniquely, potential matches can be identified by merging two or more characteristics to create a profile. Table I provides an example of a profile that could be created from variables (gender, year of illness, birth month and birth year) in two lists (A and B). When more than one person is defined by the same profile, there are many ways to match cases between the two lists with respect to the profile. Consider the last profile in Table I. There is one record in A and two records in B that are defined by the same profile. One possibility is that there are zero matches on this profile implying there are three unique individuals (Table II). In other words, the individual appearing on list A does not appear on list B. The other possibility is that there is one match on the profile implying only two unique individuals on both A and B for profile 3. One match would indicate the individual appearing on list A was also listed on list B. For our example profile, there is only one way to produce zero matches. Yet, there are two possible ways to produce one match for this profile (Table II). When there are many profiles, and each may be associated with more than one person, individuals in lists A and B could match in many ways. For a given profile, we refer to each possible way individuals on lists A and B could match as a profile match configuration. Let A_i denote the number of individuals on list A in profile i, B_i denote the number of individuals on list B in profile i, $S_i = \min\{A_i, B_i\}$, P denote the number of unique profiles, and j_i denote the number of matches between lists A and B for cases in profile i. The number of ways a profile match configuration with j_i matches can occur for profile i, $i = 1, \ldots, P$, is

$$q_{ji} = \binom{A_i}{j_i} \binom{B_i}{j_i} \tag{2}$$

Table I. An example of profiles for data with non-unique identifiers.

Profile number	Gender	Year of illness	Birth month/birth year	Number of individuals, list *A*	Number of individuals, list *B*
1	Female	1993	04/1992	2	6
2	Male	1995	07/1995	0	1
3	Female	1994	04/1994	1	2

Table II. Possible data tables for profile 3 in Table I.

Profile Number 3	List *A*	List *B*
Data table resulting from zero matches on profile 3		
Female, 1994, 04/1994	1 (subject 1)	0
Female, 1994, 04/1994	0	1 (subject 2)
Female, 1994, 04/1994	0	1 (subject 3)
Data table resulting from one match on profile 3		
Female, 1994, 04/1994	1 (subject 1)	1 (subject 1)
Female, 1994, 04/1994	0	1 (subject 2)
Alternative data table resulting from one match on profile 3		
Female, 1994, 04/1994	0	1 (subject 1)
Female, 1994, 04/1994	1 (subject 2)	1 (subject 2)

$j_i = 0, \ldots, S_i$. Letting $j_1, j_2, \ldots, j_p$ denote a combination of profile match configurations in which there are j_1 matches in profile 1, j_2 matches in profile 2,..., and j_P matches in profile P, the number of ways $j_1, j_2, \ldots, j_p$ can occur is

$$q_{j_1, j_2, \ldots, j_p} = \prod_{i=1}^{P} q_{ji}$$

and the probability of $j_1, j_2, \ldots, j_p$ is

$$\pi_{j_1, j_2, \ldots, j_p} = \frac{q_{j_1, j_2, \ldots, j_p}}{\sum_{j_1=0}^{Sj_1} \sum_{j_2=0}^{Sj_2} \cdots \sum_{j_P=0}^{Sj_P} q_{j_1, j_2, \ldots, j_p}} \tag{3}$$

The modified Chapman Lincoln–Petersen estimator corresponding to $j_1, j_2, \ldots, j_p$ is

$$N_{j_1, j_2, \ldots, j_P} = \sum_{i=1}^{P} j_i + \left(\sum_{i=1}^{P} A_i - \sum_{i=1}^{P} j_i \right) + \left(\sum_{i=1}^{P} B_i - \sum_{i=1}^{P} j_i \right) + \frac{(\sum_{i=1}^{P} A_i - \sum_{i=1}^{P} j_i)(\sum_{i=1}^{P} B_i - \sum_{i=1}^{P} j_i)}{1 + \sum_{i=1}^{P} j_i} \tag{4}$$

Table III. Weighted estimate for example data (Table I).

Profile match configuration: $j_1, j_2, \ldots, j_p$	Number of ways profile match configuration can occur $q_{j_1,j_2,\ldots,j_p}$	Chapman estimate $N_{j_1,j_2,\ldots,j_p}$
0,0,0	1	39
1,0,0	12	19
2,0,0	15	12.3
0,0,1	2	19
1,0,1	24	12.3
2,0,1	30	9
Total	84	12.5

where, $\sum_{i=1}^{P} j_i$ is the number of matches, $(\sum_{i=1}^{P} A_i - \sum_{i=1}^{P} j_i)$ is the number of cases on list A without a match, $(\sum_{i=1}^{P} B_i - \sum_{i=1}^{P} j_i)$ is the number of cases on list B without a match, and $(\sum_{i=1}^{P} A_i - \sum_{i=1}^{P} j_i)(\sum_{i=1}^{P} B_i - \sum_{i=1}^{P} j_i)/(1 + \sum_{i=1}^{P} j_i)$ is the modified Chapman estimator for the number of cases appearing on neither list. The weighted estimator of the population total is

$$\hat{N} = \sum_{j_1=0}^{Sj_1} \sum_{j_2=0}^{Sj_2} \cdots \sum_{j_P=0}^{Sj_P} \pi_{j_1,j_2,\ldots,j_P} N_{j_1,j_2,\ldots,j_P} \tag{5}$$

Table III provides estimates for the example when our weighted estimator is implemented. In this example, the total number of unique matches ranges from 0 to 3 and the point estimate for the population size ranged from 9 to 39 with a weighted average of 12.5. It can be shown that the weighted estimator is nearly unbiased for the population size (Appendix A.1).

2.3. The bootstrap estimator: non-unique matches

The weighted estimator (5) may be useful when the number of profiles (P) is small or when relatively few records exist in each profile; i.e. A_i and B_i are small. However, $q_{j_1,j_2,\ldots,j_p}$ becomes very large when there are either a large number of profiles (P) or when A_i and/or B_i are moderate or large for many profiles. In this case, we use a bootstrap procedure to obtain a point estimate for the closed population size, N, and to derive percentile confidence intervals for the estimate. The bootstrap procedure is carried out in two steps. The first step accounts for variation attributable to sampling and the second step accounts for variation attributable to uncertainty of matches.

Step 1: In the rth bootstrap replicate sample, $r = 1, \ldots, R_1$, we obtain a with-replacement sample of size $\sum_{j=1}^{P} A_i$ from the cases on list A and of size $\sum_{j=1}^{P} B_i$ from the cases on list B. From these samples we note the number of times each P profile from each list was selected. For the rth bootstrap replicate, let A_i' denote the number of times profile i is sampled from list A, B_i' denote the number of times profile i is sampled from list B, $S_i' = \min\{A_i', B_i'\}$.

$$q_{ji}' = \binom{A_i'}{j_i}\binom{B_i'}{j_i} \quad \text{and} \quad q_{j_1,j_2,\ldots,j_P}' = \prod_{i=1}^{P} q_{ji}'$$

Then we compute the probability of a combination of profile match configurations $j_1, j_2, \ldots, j_P$,

$$\pi'_{j_1,j_2,\ldots,j_p} = \frac{q'_{j_1,j_2,\ldots,j_p}}{\sum_{j_1=0}^{S'j_1} \sum_{j_2=0}^{S'j_2} \cdots \sum_{j_P=0}^{S'j_P} q'_{j_1,j_2,\ldots,j_p}}$$

Step 2: Next we obtain R_2 with-replacement samples from all combinations of profile match configurations from the rth bootstrap replicate sample. In this step, a combination of profile match configurations $j_1, j_2, \ldots, j_P$ is sampled with probability $\pi'_{j_1,j_2,\ldots,j_p}$. Each combination of profile match configurations resampled in this step represents one of the ways individuals on list A could correctly match individuals on list B. For the sth resample from the data in the rth replicate, let $X_{AB}^{(r,s)}$ denote the number of uniquely matched cases on both list A and list B, $X_{A\bar{B}}^{(r,s)}$ denote the number of cases on list A but not on list B, $X_{\bar{A}B}^{(r,s)}$ denote the number of cases on list B but not on list A, and

$$\hat{X}_{\bar{A}\bar{B}}^{(r,s)} = (X_{A\bar{B}}^{(r,s)} X_{\bar{A}B}^{(r,s)} / (X_{AB}^{(r,s)} + 1))$$

denote the estimated number of individuals in the population on neither list A nor list B, and

$$\hat{N}^{(r,s)} = X_{AB}^{(r,s)} + X_{A\bar{B}}^{(r,s)} + X_{\bar{A}B}^{(r,s)} + \hat{X}_{\bar{A}\bar{B}}^{(r,s)}$$

denote the modified Chapman Lincoln–Petersen estimator. The estimator for replicate r is $\hat{N}^{(r)} = \sum_{s=1}^{R_2} \hat{N}^{(r,s)}/R_2$ and the mean bootstrap replicate estimator is

$$\hat{N}_{\text{boot}} = \sum_{r=1}^{R_1} \hat{N}^{(r)}/R_1$$

Bootstrap percentile confidence intervals for the closed population size can be obtained using quantiles of the replicate estimates, $\hat{N}^{(r)}$, $r = 1, \ldots, R_1$ [19]. The bootstrap estimator is approximately unbiased for the closed population size (Appendix A.2).

3. EXAMPLE: PERTUSSIS HOSPITALIZATIONS DURING 1996 IN NEW YORK STATE

Reliable estimates of the burden of disease are needed to evaluate disease control policies. Similar to other nationally reportable diseases, evidence suggests the numbers of pertussis cases and hospitalizations in the United States are underestimated [20]. The combination of increasing disease and possible underestimation of reporting motivated a study to estimate the magnitude of pertussis hospitalizations by state and year in the United States using a two-source capture–recapture analysis. We report here only the results for the state of New York during 1996.

3.1. Data sources

Two surveillance lists were identified, both of which captured hospitalizations from the population of interest during 1996. The first list is from the National Electronic Telecommunications System for Surveillance (NETSS) the second list from the Health Care Information Association (HCIA).

NETSS is a concatenation of weekly reports submitted electronically from state health departments to the Centers for Disease Control and Prevention. Included in these reports are cases of diseases determined to be nationally notifiable by the Council of State and Territorial Epidemiologists [21]. The data are primarily used to rapidly identify disease epidemics.

The HCIA database contains acute-care hospital discharge records from both public and proprietary state data during 1992–1996. The database contains reports from more than 2500 non-federal, self-selected, acute-care hospitals in the United States. These reports represent approximately 40 per cent of the total U.S. hospital discharges.

3.2. Application of bootstrap estimator

The five variables common to both surveillance lists were gender, birth month, birth year, year of illness, and month of illness (hospitalization month from HCIA matched with cough-onset month from NETSS). These common variables do not identify individuals uniquely within a list and thus we cannot be certain that individuals across lists are unique. Therefore, we used this set of variables to define profiles. Because of the lack of unique identifiers and because the number of profiles was large, we implemented the bootstrap estimator to estimate the total number of pertussis hospitalizations in New York State in 1996.

The HCIA database listed 200 records of hospitalizations due to pertussis, and the NETSS database listed 123 records of pertussis hospitalizations in New York State, 1996. Each record was defined by one of 157 profiles determined by the above-mentioned five variables. The HCIA database contained 88 unique profiles and the average number of cases described by each profile was 1.40 (range 1–5). The NETSS database contained 113 unique profiles and the average number of cases described by each profile was 1.77 (range 1–6). However, 51 (41 per cent) cases listed in NETSS matched more than one individual in HCIA. In HCIA, 85 (43 per cent) cases matched more than one case in NETSS. The number of combinations of profile match configurations exceeded 4×10^{15}. Thus, we used the bootstrap method with $R_1 = 500$ and $R_2 = 250$. Figure 1 shows a histogram of the bootstrap replicate estimates, $\hat{N}^{(r)}$, $r = 1, \ldots, R_1$. The estimate for the number of pertussis hospitalizations in New York State in 1996 was the mean of the replicate estimates, $\hat{N}_{\text{boot}} = 894$. The 2.5 per cent and 97.5 per cent quantiles of $\hat{N}^{(r)}$ were used to give the 95 percentile confidence interval (CI) (737, 1102).

3.3. Evaluation of bootstrap estimator

To examine the impact of profile definition on our results, we implemented various profile definitions for the 1996 New York State data. The first, presented above, defined profiles by gender, year of illness, month of illness, birth month, and birth year. In each additional profile definition, one of the variables used to define the initial profile is removed thereby creating a less specific definition then the first. With the less specific matching criterion, the estimate decreases and the confidence intervals narrow (Table IV). This result is not surprising because with a less specific definition the number of potential matches decreases and therefore the total populations estimate must decrease.

4. DISCUSSION

A fundamental requirement of capture–recapture methods for estimating population totals is that cases can be matched using unique identifiers. However, the methods we describe

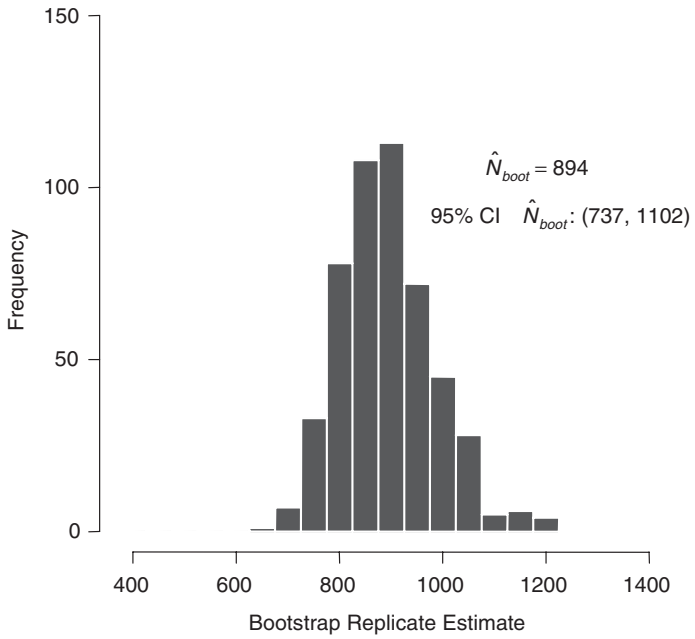

Figure 1. New York 1996: distribution of bootstrap replicate estimates ($R_1 = 500$; $R_2 = 250$).

Table IV. Impact of profile definition on population estimates for NY 1996.

Profile definition	Number of profiles	Population estimate (95% CI)
Gender, year of illness, month of illness, birth month, birth year	157	894 (737, 1102)
Gender, year of illness, birth month, birth year	58	468 (421, 523)
Gender, year of illness, birth year	18	362 (346, 380)
Gender, year of illness	2	324 (321, 329)

*Using the bootstrap estimator ($R_1 = 500$; $R_2 = 250$).

estimate the total population of cases from two lists using non-unique identifiers to form profiles defined by characteristics common to each list. We account for the uncertainty of uniqueness by developing the weighted estimator, which incorporates all possible matching configurations when non-uniqueness exists. In addition, we developed the bootstrap estimator for use when there are large numbers of profiles or large numbers of individuals within profiles.

Both the weighted and bootstrap estimators assume cases within lists are not duplicated. That is, any two cases defined by the same profile within a list are two different cases. It is therefore important to carefully evaluate lists to ensure that there is, at least theoretically, no duplication of cases on any one list providing information for the capture–recapture analysis.

However, if duplicates do exist within one or both lists the resulting population estimates will be inflated.

The methods presented here are not recommended as a substitute for the use of unique identifiers when they do exist within and between lists. The weighted estimator is a feasible alternative when assumptions of uniqueness cannot be met, but implementation may be limited to situations with few profiles or few cases per profile. The bootstrap estimator may allow for broader applications of capture–recapture methods in the absence of unique identifiers, and overcomes the previously mentioned limitation of the weighted estimator. Although this work demonstrates that capture–recapture estimates can be obtained when cases may not be uniquely identifiable, the resulting population estimates are impacted by the specificity used to define profiles. Additionally, evaluations of the bias and variance of the estimators should be performed to develop guidelines for use of the estimators.

Epidemiologic and demographic applications of the capture–recapture method have devoted considerable resources to ascertaining matches using characteristics of cases available in administrative lists. These matches are often uncertain and do not account for the variation attributable to the uncertainty of the matching. Moreover, the capture–recapture estimate may be biased because of reliance on only one of many possible matches among comparison pairs of records. Results from our study show that the matching procedures that rely on only one match may not be necessary. Reasonable capture–recapture estimates may be obtained that account for uncertainty resulting from non-unique identifiers as well as uncertainty attributable to sampling variation.

APPENDIX A

A.1. Justification for the weighted estimator as an approximately unbiased estimator for the unknown size of the closed population

Conditional on matching configuration $j_1, j_2, \ldots, j_P$ being the true configuration, the modified Chapman Lincoln–Petersen estimate is approximately unbiased [8]; i.e.

$$E[\hat{N}_{j_1,j_2,\ldots,j_P|j_1,j_2,\ldots,j_P}] \doteq N$$

As a consequence, the weighted estimator is approximately unbiased for the true but unknown size of the closed population

$$E_{j_1,j_2,\ldots,j_P}\{E[\hat{N}_{j_1,j_2,\ldots,j_P|j_1,j_2,\ldots,j_P}]\} \doteq N$$

A.2. Justification for the bootstrap estimator as an unbiased estimator for the weighted estimator

The weighted estimator is sensible in that it is an average over all estimates that correspond to each possible profile match configuration that could have been observed. The bootstrap estimator has the same expectation as the weighted estimator. To show this, let $\hat{N}^{(r)} = \sum_{s=1}^{R_2} \hat{N}^{(r,s)}/R_2$ denote the rth replicate estimator obtained from the rth replicate data table, where $\hat{N}^{(r,s)}$ denotes the modified Chapman Lincoln–Petersen estimator from the sth resampled combination of profile match configurations from the rth replicate data table.

The replicate estimator for the rth data table may be re-expressed as

$$\sum_{j_1=0}^{Sj_1}\sum_{j_2=0}^{Sj_2}\cdots\sum_{j_P=0}^{Sj_P}\frac{t^{(r)}_{(j_1,j_2,\ldots,j_P)}\hat{N}^{(r)}_{j_1,j_2,\ldots,j_P}}{R_2},$$

where $t^{(r)}_{(j_1,j_2,\ldots,j_P)}$ denotes the number of times the combination of profile match configurations $j_1, j_2, \ldots, j_P$ is selected out of the R_2 subreplicates drawn with replacement from the rth replicate data table. $\hat{N}^{(r)}_{j_1,j_2,\ldots,j_P}$ is the Chapman Lincoln–Petersen estimator for the rth replicate data table corresponding to the combination of profile match configurations $j_1, j_2, \ldots, j_P$. Conditional on the rth replicate table, $E[t^{(r)}_{(j_1,j_2,\ldots,j_P)}|r] = R_2 \times \pi^{(r)}_{j_1,j_2,\ldots,j_P}$. Therefore,

$$E[\hat{N}^{(r)}] = \sum_{j_1=0}^{Sj_1}\sum_{j_2=0}^{Sj_2}\cdots\sum_{j_p=0}^{Sj_P}\pi^{(r)}_{j_1,j_2\ldots,j_p}\hat{N}^{(r)}_{j_1,j_2,\ldots,j_p}$$

i.e. the weighted estimator for the rth replicate table. Averaging

$$\sum_{j_1=0}^{Sj_1}\sum_{j_2=0}^{Sj_2}\cdots\sum_{j_p=0}^{Sj_P}\pi^{(r)}_{j_1,j_2,\ldots,j_p}\hat{N}^{(r)}_{j_1,j_2,\ldots,j_p}$$

over all replicate tables yields a simulated value for the weighted estimator.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the useful comments provided by Chima Ohuabunwo, Kristine Bisgard, Chuck Vitek, and Ted Thompson. In addition, we thank Mary McCauley for her editorial assistance.

REFERENCES

1. Hook EB, Regal RR. Capture–recapture methods in epidemiology: methods and limitations. *Epidemiologic Reviews* 1995; **17**(2):243–264 (correction appears in *Epidemiologic Reviews* **148**:1219).
2. International Working Group for Disease Monitoring and Forecasting. Capture–recapture and multiple-record systems estimation I: history and theoretical development. *American Journal of Epidemiology* 1995; **142**: 1047–1058.
3. International Working Group for Disease Monitoring and Forecasting. Capture–recapture and multiple-record systems estimation II: applications in human diseases. *American Journal of Epidemiology* 1995; **142**: 1059–1068.
4. Lincoln FC. Calculating waterfowl abundance on the basis of banding returns. *Circular No. 118*. US Department of Agriculture, Washington, DC, 1930; 1–4.
5. Petersen CGJ. The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station* 1896; **6**:1–48.
6. Sekar C, Deming WE. On a method of estimating birth and death rates and extent of registration. *Journal of the American Statistical Association* 1949; **44**:101–115.
7. Chapman CJ. Some properties of the hypergeometric distribution with applications to zoological censuses. *University of California Publication in Statistics* 1951; **1**:131–160.
8. Wittes JT. On the bias and estimated variance of Chapman's two-sample capture–recapture population estimate. *Biometrics* 1972; **28**:592–597.
9. Smith PJ. Bayesian methods for multiple capture–recapture surveys. *Biometrics* 1988; **44**:1177–1190.
10. Smith PJ. Bayesian methods for prevalence estimation form incomplete administrative lists. *Statistics in Medicine* 1990; **10**:113–118.
11. Smith PJ. Bayesian analyses for a capture–recapture model. *Biometrika* 1991; **2**:399–407.

12. Verstraeten T, Baughman AL, Cadwell B, Zanardi L, Haber P, Chen RT. Vaccine Adverse Event Reporting System Team. Enhancing vaccine safety surveillance: a capture–recapture analysis of intussusception after rotavirus vaccination. *American Journal of Epidemiology* 2001; **154**:1006–1012.
13. Galil K, Pletcher MJ, Wallace BJ, Seward J, Meyer PA, Baughman AL, Wharton M. Tracking varicella deaths: accuracy and completeness of death certificates and hospital discharge records, New York state, 1989–1995. *American Journal of Public Health* 2002; **92**:1248–1250.
14. Vitek CR, Pascual FB, Baughman AL, Murphy TV. Increase in deaths from pertussis among young infants in the United States in the 1990s. *The Pediatric Infectious Disease Journal* 2003; **22**:624–634.
15. Jaro MA. Probabilistic linkage of large public health data files. *Statistics in Medicine* 1995; **14**:491–498.
16. Blakely T, Salmond C. Probabilistic linkage and a method to calculate the positive predictive value. *International Journal of Epidemiology* 2002; **31**:1246–1252.
17. Laska EM, Meisner M, Wanderling J, Seigel C. Estimating population size and duplication rates when records cannot be linked. *Statistics in Medicine* 2003; **22**:3403–3417.
18. Hook EB, Regal RR. Recommendation for presentation and evaluation of capture–recapture estimates in epidemiology. *Journal of Clinical Epidemiology* 1999; **52**:917–926.
19. Efron B, Tibishriani RJ. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Press: Boca Raton, FL, 1993.
20. Sutter RS, Cochi SL. Pertussis hospitalizations and mortality in the United States, 1985–1988. *Journal of the American Medical Association* 1992; **267**(3):386–391.
21. National Electronic Telecommunications System for Surveillance, CDC Epidemiology Program Office. http://www.cdc.gov/epo/dphsi/netss.htm (accessed 19 September, 2003).