Methods for Capture- Recapture Analysis When Cases May Not Be Identified Uniquely

Betsy L. Cadwell[†], Philip J. Smith[‡], Andrew L. Baughman[‡]

[†] Center for Disease Control and Prevention, Division of Diabetes Translation, 3470 Buford Highway, NE, Atlanta, GA 30341; [‡] Center for Disease Control and Prevention, National Immunization Program, 1600 Clifton Road, NE, Atlanta, GA 30333.

1. Introduction

Capture-recapture methods are used for estimating the unknown size of a closed population from overlapping lists of cases [1, 2, 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 and unmatched cases 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 [4, 5]. Chapman later adjusted the estimator for use with small samples [6, 7].

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 a list is independent of the probability of being recorded on the other list; and there is no immigration or emigration to 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 the validity of these assumptions, current methods require cases appearing on each list to be uniquely identifiable such that unique matches between lists can be determined.

Epidemiological applications of capture-recapture methods utilize surveillance data or administrative lists. Administrative lists often do not have a sufficient amount of specific information on individual characteristics to insure cases are uniquely identifiable. In this paper we modify commonly used capture-recapture methods to overcome this problem. In our methods we create profiles, or a collection of characteristics, that describe one or more individuals on a list. For example, if profiles were defined by gender, birth month and birth year then a possible profile is males born in May of 1976. For profiles appearing on both lists, we account for potential matches of cases between the two lists. Two methods are developed to account for potential matches: (1) a weighted estimator, which considers all potential matches and (2) a bootstrap estimator, which considers potential matches by resampling profiles with replacement. 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 without unique identifiers.

2. Methods

2.1 The Chapman Estimator: Unique Matches

When two lists, A and B, the observed counts are: the number of uniquely matched cases on both list A and list 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

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

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

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

and its estimated variance is approximately

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

2.2 The Weighted Estimator: Non-unique Matches

When the identifiers 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" (e.g. males born in May of 1976). When more than one person on a list has the same profile, there are many ways to match cases between the two lists with respect to the profile. Further, when there are many profiles, each of which may be associated with more than one person, there are many different ways that individuals in list A could match individuals in list B. We refer to each of the possible ways in which individuals on list A could match individuals on list B as a "profile match configuration."

Letting A_i denote the number of individuals on list A with profile i, B_i denote the number of individuals on list B with profile i, $S_i = \min\{A_i, B_i\}$, P denote the number of unique profiles, and j_i the number of ways profile match configuration can occur for profile i, $i = 1, \ldots, P$ is

$$q_{ji} = \binom{A_i}{j_i}\binom{B_i}{j_i}, \qquad (2)$$

$j_i = 0,\ldots,S_i$. Letting $j_1, j_2, \ldots, j_p$ denote a combination of profile match configurations in which there are j_1 matches on profile 1, j_2 matches on profile 2, …, and j_P matches on 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

$$\pi_{j_1, j_2, \ldots, j_p} = \frac{q_{j_1, j_2, \ldots j_p}}{\sum_{j_1=0}^{S_{j_1}} \sum_{j_2=0}^{S_{j_2}} \cdots \sum_{j_P=0}^{S_{j_P}} q_{j_1, j_2, \ldots j_p}} \; . \qquad (3)$$

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

$$\hat{N}_{j_1 j_2 \ldots j_P} = \sum_{i=1}^{P} j_i + \left(\sum_{i=1}^{P} A_i - \sum_{i=1}^{P} j_i\right) + \left(\sum_{i=1}^{P} B_i - \sum_{i=1}^{P} j_i\right) + \frac{\left(\sum_{i=1}^{P} A_i - \sum_{i=1}^{P} j_i\right)\left(\sum_{i=1}^{P} B_i - \sum_{i=1}^{P} j_i\right)}{1 + \sum_{i=1}^{P} j_i} \qquad (4)$$

where, $\sum_{i=1}^{P} j_i$ is the number of matches, $\left(\sum_{i=1}^{P} A_i - \sum_{i=1}^{P} j_i\right)$ is the number of cases on list A without a match, $\left(\sum_{i=1}^{P} B_i - \sum_{i=1}^{P} j_i\right)$ is the number of cases on list B without a match, and $\frac{\left(\sum_{i=1}^{P} A_i - \sum_{i=1}^{P} j_i\right)\left(\sum_{i=1}^{P} B_i - \sum_{i=1}^{P} j_i\right)}{1 + \sum_{i=1}^{P} j_i}$ is the modified Chapman estimator for the number of cases appearing on neither list. When there is only one profile match configuration, cases are uniquely identifiable and equations (4) and (1) are identical. The weighted estimator of the population total is

$$\hat{N} = \sum_{j_1=0}^{S_{j_1}} \sum_{j_2=0}^{S_{j_2}} \cdots \sum_{j_P=0}^{S_{j_P}} \pi_{j_1, j_2, \ldots, j_P} \, \hat{N}_{j_1, j_2, \ldots, j_P} \; . \qquad (5)$$

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 there are relatively few records existing in each profile, i.e. A_i and B_i are small. However, $q_{j_1, j_2, \ldots, j_p}$ approaches infinity 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.

In the r-th bootstrap replicate sample, $r = 1,\ldots,R_1$, we obtain a with-replacement sample of size $\sum_{j=1}^{P} A_i$ from list A and of size $\sum_{j=1}^{P} B_i$ from list B. From these samples we note the number of individuals belonging to each of the *P* profiles on each list. For the r-th bootstrap replicate let A_i' denote the number of individuals sampled from list A that have profile *i*, B_i' denote the number of individuals on list B with profile *i*, $S'_i = \min\{A_i', B_i'\}$, $q'_{ji} = \binom{A_i'}{j_i}\binom{B_i'}{j_i}$, and $q'_{j_1,j_2,\ldots,j_P} = \prod_{i=1}^{P} q'_{ji}$.

Then, we compute the probability of every profile match configuration $j_1, j_2, \ldots, j_P$,

$$\pi'_{j_1,j_2,\ldots,j_p} = \frac{q'_{j_1,j_2,\ldots j_p}}{\sum_{j_1=0}^{S'_{j1}}\sum_{j_2=0}^{S'_{j2}}\cdots\sum_{j_P=0}^{S'_{jP}} q'_{j_1,j_2,\ldots j_p}} .$$

Next, we obtain R_2 with-replacement samples from the profile match configurations $j_1, j_2, \ldots, j_P$ from the *r*-th bootstrap replicate sample. In this step, profile match configuration $j_1, j_2, \ldots, j_P$ is sampled with probability $\pi'_{j_1,j_2,\ldots,j_p}$. Each of the profile match configurations resampled in this step represent one of the ways in which individuals on list A could correctly match individuals on list B. For the s-th, resample from the data in the r-th replicate, $X_{AA}^{(r,s)}$ 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

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

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 $\hat{N}^{(r)} = \sum_{s=1}^{R_2} \hat{N}^{(r,s)} \Big/ R_2$ and the mean bootstrap replicate estimator is

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

Bootstrap percentile confidence intervals for the close population size can be obtained using quantiles of the replicate estimates, $\hat{N}^{(r)}$, $r = 1,\ldots,R_1$. It can be shown that the mean bootstrap replicate estimator is unbiased for the weighted estimator, and that the weighted estimator minimizes squared error loss.

3. Example: Pertussis Hospitalizations During 1996 in the State of New York

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 [9]. 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-sources 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 which both captured hospitalizations from the population of interest during 1996. The first list comes from the National Electronic Telecommunications System for Surveillance (NETSS); the second list was obtained 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 [10]. The data are primarily used to rapidly identify disease epidemics.

The HCIA database contains acute-care hospital discharge records form 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% of the total US hospital discharges.

3.2 Application of Bootstrap Estimator

The attributes 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 attributes do not identify individuals uniquely within a list and thus we cannot be certain that the individuals across lists are unique. Therefore, we used this set of attributes 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 during the year 1996.

The HCIA database listed 200 records of hospitalizations due to pertussis while 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 variables. For HCIA there were 88 unique profiles and the average number of cases described by each profile was 1.40 (Range: 1-5). For NETSS there were 113 unique profile; the average number of cases described by each profile was 1.77 (Range: 1-6). However, 51 (41%) cases listed in NETSS matched more than one individual in HCIA. In HCIA, 85 (43%) cases matched more than one case in NETSS. The number of matching configurations exceeds 4×10^{15}. Thus, we used the bootstrap method with $R_1 = 500$ times 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 during the year 1996 is the mean of the replicate estimates, $\hat{N}_{boot} = 894$. The 2.5% and 97.5% quantiles of $\hat{N}^{(r)}$ were used to give the 95% percentile confidence interval, (737, 1102).

To examine the impact of profile definition on our results, we implemented two profile definitions for the 1996 New York State data. The first, presented above, defined the profile by gender, year of illness, month of illness, birth month and birth year. The second profile definition used only gender, year of illness, birth month, and birth year. Thus the second profile definition can be considered less specific then the first. As expected, with the less specific matching criterion the estimate decreased to 468 (95% CI: 421, 523).

4. Discussion

A fundamental requirement of capture-recapture methods for estimating population totals is cases can be matched using unique identifiers. In our research 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 a large number of profiles are present or when large numbers of individuals exist within profiles. The bootstrap estimator is approximately unbiased for the closed population size.

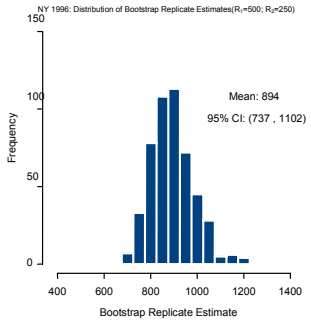
Both the weighted and bootstrap estimators assume individuals within lists are not duplicated. That is, two or more cases defined by the same profile within a list are in fact unique. It is therefore important to carefully evaluate the administrative lists used to ensure that there is, at least theoretically, no duplication of individuals on any one list providing information for the capture recapture analysis. However, if duplicates do exist within one or both of the lists, the resulting estimates will be inflated.

The methods presented here are not recommended as a substitute for use of unique identifiers when they do exist within and between lists. Yet, using the weighted or bootstrap estimator is a feasible alternative when assumptions of uniqueness cannot be met and allows for broader applications of capture-recapture methods. Although this work demonstrates that capture-recapture estimates can be obtained when cases are not uniquely identifiable, the numerical results suggest that estimates can depend sensitively on the specificity of the profile definition.

5. 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.

Figure 1: Distribution of Replicate Estimates



7. References

1. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiologic Reviews* 1995; **17**(2):243-64 [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.* Washington DC: US Department of Agriculture, 1930: 1-4.
5. Petersen CGJ. The yearly immigration of young plaice into the Limfjord from the German sea. *Rep Dan Biol Stat* 1896; **6**:1-48.
6. Chapman CJ. Some properties of the hypergeometric distribution with applications to zoological censuses. *U California Public Stat* 1951; **1**:131-60.
7. Wittes JT. On the bias and estimated variance of Chapman's two-sample capture-recapture population estimate. *Biometrics* 1972; **28**: 592-7.
8. Efron B, Tibishriani RJ. *An Introduction to the Bootstrap*. Chapman & Hall/CRC: Boca Raton, 1993.
9. Sutter RS, Cochi SL. Pertussis hospitalizations and mortality in the United States, 1985-1988. *JAMA* 1992; **267**(3): 386-91.
10. http://www.cdc.gov/epo/dphsi/netss.htm; Internet; accessed October 7, 2002.