

Using Occupancy Models to Estimate the Number of Duplicate Cases in a Data System without Unique Identifiers

Ruiguang Song, Timothy Green, Matthew McKenna, and M. Kathleen Glynn
Centers for Disease Control and Prevention

Abstract: Data systems collecting information from different sources or over long periods of time can receive multiple reports from the same individual. An important example is public health surveillance systems that monitor conditions with long natural histories. Several state-level systems for surveillance of one such condition, the human immunodeficiency virus (HIV), use codes composed of combinations of non-unique personal characteristics such as birth date, soundex (a code based on last name), and sex as patient identifiers. As a result, these systems cannot distinguish between several different individuals having identical codes and a unique individual erroneously represented several times. We applied results for occupancy models to estimate the potential magnitude of duplicate case counting for AIDS cases reported to the Centers for Disease Control and Prevention with only non-unique partial personal identifiers. Occupancy models with equal and unequal occupancy probabilities are considered. Unbiased estimators for the numbers of true duplicates within and between case reporting areas are provided. Formulas to calculate estimators' variances are also provided. These results can be applied to evaluating duplicate reporting in other data systems that have no unique identifier for each individual.

Key words: Case surveillance, duplicate reporting, occupancy model, soundex.

1. Introduction

Public health surveillance systems that monitor conditions with long natural histories can receive multiple reports from different sources regarding the same affected individual. For example, an individual may change his/her place of residence and seek care for the disease under surveillance, likely resulting in case reports from both places. If there is a unique identifier for each individual submitted with surveillance reports, then duplicate reports can be easily identified and removed from the surveillance system. However, because of confidentiality concerns, national surveillance systems do not collect information on variables that can uniquely identify a person. For example, name and social security number may be reported to a state surveillance system as a part of routine reporting,

but are not reported to CDC (Centers for Disease Control and Prevention) for national HIV/AIDS surveillance purposes. Instead, an identifier is often created based on several descriptors. This identifier will not be unique. When information submitted to a surveillance system cannot uniquely identify an individual, and the potential for duplicate reports being submitted to the system exists, the system must use additional information to determine if cases with the same non-unique identifiers represent the same person. For this discussion, we call reports with the same partial personal identifiers “potential duplicates”. Among these, we classify reports representing the same person as “true duplicates” and those representing different persons as “non-duplicates”.

National AIDS surveillance data in the United States have the potential for duplicate reporting and do not have unique identifiers to identify and remove duplicate reports. Using the data available at the national level (cases reported to CDC), one cannot determine whether cases with the same partial personal identifiers represent the same person and therefore are true duplicates. However, it is possible to estimate the expected number of non-duplicates from the potential duplicates in a surveillance system based on the probability of matching on these partial personal identifiers. Larsen (1994) considered this problem in a register of HIV infected persons, using a method to estimate the number of distinct individuals in the register based on the date of birth of each entry and classical occupancy theory where each ball has the same chance of falling into any one of the cells. While this method may be applicable as applied to the date of birth in a given year, it cannot be applied to identifiers where individuals have an unequal chance to take each possible value of the identifier, e.g., the soundex (a code based on last name using a method of encryption, see Fenna, 1984).

Under the classical occupancy model where each ball has the same chance of falling into each cell, the explicit formula for the expected number of empty cells is available. However, the explicit formula for the variance associated with the observed number of empty cells is not available. A similar situation occurs under the model with unequal occupancy probabilities. Only approximate formulas for the variance are available, see Chistyakov (1967), Holst(1971), and Sevastyanov (1972). In this paper, we provide exact variance formulas for the observed number of empty cells under the two occupancy models. They are presented in Sections 2 and 3, respectively. In Section 4, we consider a model with cells filled by colored balls. All of these results can be applied to evaluating duplicates in a data system. As an example, we use occupancy models to evaluate duplication in AIDS case reporting. Results are presented in Section 5. Finally, some concerns and recommendations are presented in the discussion section.

2. Occupancy Model with Equal Occupancy Probabilities

Suppose that r balls are randomly distributed to n cells. Assume that each ball has an equal chance of being distributed to each cell. Let $M_{r,n}$ be the number of cells remaining empty. According to occupancy theory (see Feller, 1968, page 102), the probability distribution of $M_{r,n}$ is given by

$$Pr(M_{r,n} = m) = \binom{n}{m} \sum_{i=1}^{n-m} (-1)^i \binom{n-m}{i} (1 - (m+i)/n)^r \quad (2.1)$$

where $\binom{x}{k}$ is the binomial coefficient equal to the number of combinations of k items selected from x items. Note that this formula is difficult to handle because of the potential for rounding error. A useful recursive formula is available (see Feller, 1968, page 60):

$$Pr(M_{r+1,n} = m) = Pr(M_{r,n} = m) \frac{n-m}{n} + Pr(M_{r,n} = m+1) \frac{m+1}{n} \quad (2.2)$$

Based on this recursive formula, one can derive the mean and variance of $M_{r,n}$. An alternative but simpler way to derive the mean and variance is presented in Section 3. As a special case of equations (3.1) and (3.6), the mean and variance of $M_{r,n}$ defined in (2.1) are:

$$E(M_{r,n}) = n \left(1 - \frac{1}{n}\right)^r \quad (2.3)$$

and

$$Var(M_{r,n}) = n^2 \left\{ \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right)^r + \left(1 - \frac{1}{n}\right)^r \left[\frac{1}{n} - \left(1 - \frac{1}{n}\right)^r \right] \right\} \quad (2.4)$$

Since we are interested in the number of occupied cells and the number of balls that exceed the minimum necessary to fill the occupied cells, we consider variables $K_{r,n}$, the number of occupied cells, and $D_{r,n}$, the number of balls r , minus the number of cells occupied by the r balls, $D_{r,n} = r - K_{r,n} = r - (n - M_{r,n})$. Therefore, we have

$$E(D_{r,n}) = r - E(K_{r,n}) = r - [n - E(M_{r,n})] \quad (2.5)$$

and, given n and r ,

$$Var(M_{r,n}) = Var(K_{r,n}) = Var(D_{r,n}) \quad (2.6)$$

If $K_{r,n}$ is observed but r is unknown, then we can estimate r by solving equation (2.3) for r :

$$\hat{r} = \frac{\log(1 - K/n)}{\log(1 - 1/n)} \quad (2.7)$$

Using the delta method, the variance of the above estimator is

$$\text{Var}(\hat{r}) \approx \left(\frac{1}{(n - K) \log(1 - 1/n)} \right)^2 \text{Var}(K_{r,n}) \quad (2.8)$$

The above results can be applied to situations where cells do not all have the same probability of being occupied, but can be divided into subgroups such that within each subgroup each cell has an equal probability of being occupied by a ball.

3. Occupancy Model with Unequal Occupancy Probabilities

In this section, we assume that the occupancy probabilities differ from cell to cell. Let $M_{r,n}$ denote the number of empty cells after r balls have been placed into n cells with occupancy probabilities p_1, \dots, p_n , and $\sum_{i=1}^n p_i = 1$. Then, the expected number of empty cells can be expressed as

$$\begin{aligned} E(M_{r,n}) &= \sum_{i=1}^n \text{Pr}(\text{the } i\text{-th cell is empty}) \\ &= \sum_{i=1}^n (1 - p_i)^r \end{aligned} \quad (3.1)$$

Given m , the number of empty cells, we can estimate r , the number of balls, by solving the above equation for r with $E(M_{r,n}) = m$.

Using the binomial expansion followed by interchanging the order of summation in (3.1), we have:

$$E(M_{r,n}) = n - r + \sum_{t=2}^r (-1)^t \binom{r}{t} q_t \quad (3.2)$$

where

$$q_t = \sum_{i=1}^n p_i^t \quad (3.3)$$

Therefore, the expected number of excess balls is given by

$$E(D_{r,n}) = r - [n - E(M_{r,n})] = \sum_{t=2}^r (-1)^t \binom{r}{t} q_t \quad (3.4)$$

Given s ($2 \leq s < r$), the above formula can be approximated by

$$E(D_{r,n}) \approx \sum_{t=2}^s (-1)^t \binom{r}{t} q_t \quad (3.5)$$

The difference between (3.4) and (3.5) is the sum of smaller terms $\sum_{t=s+1}^r (-1)^t \binom{r}{t} q_t$. Because of the cancelation of positive and negative terms, the difference can be small. If $\binom{r}{t} q_t$ decreases with t when $t \geq s$, then the above approximation has a maximum error less than $\binom{r}{s+1} q_{s+1}$. Since $q_{t+1} \leq p_{\max} q_t$ where $p_{\max} = \max\{p_1, \dots, p_n\}$, it follows that $\binom{r}{t} q_t \leq \binom{r}{t+1} q_{t+1}$ if $t \geq (rp_{\max} - 1)/(1 + p_{\max})$.

Similar to the occupancy model with equal occupancy probabilities, the variances of $M_{r,n}$, $K_{r,n}$ and $D_{r,n}$ are all the same for given n and r . The common variance is:

$$\begin{aligned} \text{Var}(M_{r,n}) &= \sum_{i=1}^n (1-p_i)^r [1 - (1-p_i)^r] \\ &\quad + 2 \sum_{i < j} \{ [1 - (p_i + p_j)]^r - (1-p_i)^r (1-p_j)^r \} \end{aligned} \quad (3.6)$$

This can be easily proved by considering the number of empty cells as a sum of binary variables: $M_{r,n} = \sum_{i=1}^n X_i$, where X_i is the indicator variable for the i -th cell after r balls have been distributed into the n cells. The probability that the i -th cell is empty is given by $Pr(X_i = 1) = (1-p_i)^r$ and the probability that two cells, say the i -th and j -th cells, are empty is $Pr(X_i = 1, X_j = 1) = [1 - (p_i + p_j)]^r$. Therefore, the variance and covariance of these binary variables are

$$\text{Var}(X_i) = E(X_i^2) - [E(X_i)]^2 = (1-p_i)^r - (1-p_i)^{2r} \quad (3.7)$$

and

$$\text{Cov}(X_i, X_j) = E(X_i X_j) E(X_i) E(X_j) = [1 - (p_i + p_j)]^r - (1-p_i)^r (1-p_j)^r \quad (3.8)$$

Combining (3.7) and (3.8) gives (3.6).

Three approximate formulas for the variance can be found in the literature. The first approximation is given by Chistyakov (1967): if $\log(r/n)$ is bounded, then $M_{r,n}$ has an asymptotic normal distribution with variance

$$\sigma_1^2 = \sum_{i=1}^n e^{-rp_i} (1 - e^{-rp_i}) - r \left(\sum_{i=1}^n p_i e^{-rp_i} \right)^2 \quad (3.9)$$

This approximation is quite accurate. It is slightly greater than the true variance and its maximum relative error is small:

$$\text{Var}(M_{r,n}) \leq \sigma_1^2 \leq \frac{r}{r-1} \text{Var}(M_{r,n}) \quad (3.10)$$

Under the condition of equal occupancy probabilities, it can be shown that for fixed r ,

$$\sigma_1^2 \xrightarrow{n \rightarrow \infty} \frac{r}{r-1} \text{Var}(M_{r,n}) \quad (3.11)$$

If n is large but r/n is small, Holst (1971) gave a simpler approximation for the variance of $M_{r,n}$:

$$\sigma_2^2 = \frac{r^2}{2} \sum_{i=1}^n p_i^2 \quad (3.12)$$

If n is large, but r/n is not small or the expected number of empty cells is small, Sevastyanov (1972) proved that $M_{r,n}$ has an asymptotic Poisson distribution with a variance equal to its mean:

$$\sigma_3^2 = \sum_{i=1}^n (1 - p_i)^r \quad (3.13)$$

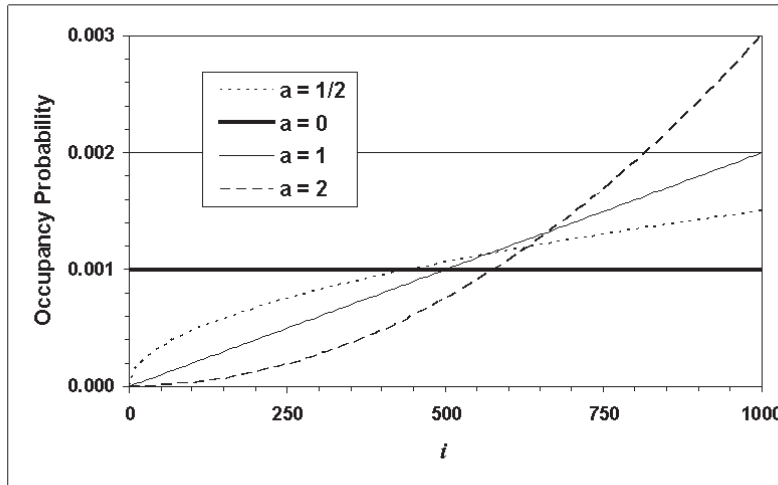


Figure 1: Distributions of occupancy probabilities that are proportional to i^a for $n = 1000$.

Table 1: True and approximate values of standard deviation of the number of occupied cells for $n = 1000$.

r	$E(K_{r,n})$	σ	σ_1	σ_2	σ_3	σ_1/σ	$\min(\sigma_2, \sigma_3)/\sigma$
$a = 0$							
2	2.0	0.03	0.04	0.04	31.59	1.413	1.415
5	5.0	0.10	0.11	0.11	31.54	1.117	1.121
10	10.0	0.21	0.22	0.22	31.46	1.053	1.062
50	48.8	1.06	1.07	1.12	30.84	1.009	1.052
100	95.2	2.05	2.06	2.24	30.08	1.004	1.091
500	393.6	7.40	7.40	11.18	24.62	1.000	1.512
1000	632.3	9.86	9.86	22.36	19.18	1.000	1.945
3000	950.3	6.31	6.31	67.08	7.05	1.001	1.117
5000	993.3	2.54	2.54	111.80	2.59	1.000	1.021
$a = 1/2$							
2	2.0	0.03	0.05	0.05	31.59	1.413	1.415
5	5.0	0.11	0.12	0.12	31.54	1.117	1.122
10	10.0	0.22	0.23	0.24	31.47	1.053	1.063
50	48.6	1.12	1.13	1.19	30.84	1.008	1.056
100	94.6	2.15	2.16	2.37	30.09	1.004	1.100
500	383.8	7.58	7.58	11.86	24.82	1.000	1.565
1000	607.3	9.99	9.98	23.71	19.82	1.000	1.985
3000	907.7	7.69	7.70	71.13	9.61	1.000	1.249
5000	965.0	5.05	5.05	118.56	5.91	1.000	1.171
$a = 1$							
2	2.0	0.04	0.05	0.05	31.59	1.412	1.415
5	5.0	0.11	0.13	0.13	31.54	1.116	1.122
10	9.9	0.24	0.26	0.26	31.47	1.053	1.064
50	48.4	1.21	1.22	1.29	30.85	1.008	1.064
100	93.7	2.31	2.32	2.58	30.10	1.004	1.117
500	368.1	7.75	7.76	12.91	25.14	1.000	1.664
1000	568.0	9.94	9.94	25.81	20.79	1.000	2.092
3000	834.1	8.60	8.60	77.44	12.88	1.000	1.497
5000	900.4	6.93	6.93	129.07	9.98	1.000	1.440
$a = 2$							
2	2.0	0.04	0.06	0.06	31.59	1.411	1.415
5	5.0	0.13	0.15	0.15	31.54	1.116	1.124
10	9.9	0.28	0.30	0.30	31.47	1.052	1.067
50	47.9	1.38	1.40	1.50	30.86	1.008	1.083
100	91.7	2.60	2.60	3.00	30.14	1.003	1.155
500	336.9	7.88	7.88	15.00	25.75	1.000	1.903
1000	495.9	9.55	9.55	29.99	22.45	1.000	2.352
3000	704.9	8.90	8.91	89.98	17.18	1.000	1.929
5000	771.5	8.03	8.03	149.96	15.12	1.000	1.883

To see how accurate these approximations are, we compared the approximate variances with the true value for occupancy probability distributions $p_i = i^a/c$, where $c = \sum_{i=1}^n i^a$. When $a = 0$, occupancy probabilities are equal. The distributions for $a = 0, 1/2, 1$, and 2 are shown in Figure 1. True and approximate values of the standard deviation are provided in Table 1. From Table 1, we can see that

$$\sigma \leq \sigma_1 \leq \min(\sigma_2, \sigma_3) \quad (3.14)$$

where $\sigma = \sqrt{\text{Var}(M_{r,n})}$.

4. Occupancy Problem When Cells Are Filled with Balls of Different Colors

In this section, we consider an occupancy model with cells filled by different colored balls. For simplicity, suppose that balls are colored either black or white and there are r_1 black and r_2 white balls. We are interested in the expected number of black balls falling in cells that have white balls. Suppose that balls of both colors are distributed into n cells with the same distribution probabilities p_1, \dots, p_n . Suppose that the r_2 white balls are distributed in k_2 cells labeled i_1, \dots, i_{k_2} . The probability that a ball falls in these cells is

$$p_{white} = \sum_{j=1}^{k_2} p_{i_j} \quad (4.1)$$

Let R_{12} be the number of black balls occupying cells i_1, \dots, i_{k_2} . Then R_{12} has a binomial distribution $\text{Bin}(r_1, p_{white})$. Therefore, we have

$$E(R_{12}) = r_1 p_{white} \quad \text{and} \quad \text{Var}(R_{12}) = r_1 p_{white} (1 - p_{white}) \quad (4.2)$$

In the equal occupancy probability situation, all $p_i = 1/n$ and $p_{white} = k_2/n$. The mean and variance are

$$E(R_{12}) = r_1 k_2 / n \quad (4.3)$$

and

$$\text{Var}(R_{12}) = r_1 k_2 (n - k_2) / n^2. \quad (4.4)$$

5. Application to Analysis of Duplicates in AIDS Case Reporting

As we mentioned earlier, in AIDS surveillance, the partial personal identifiers reported to CDC cannot uniquely identify an individual. Among data elements

reported to CDC, sex, date of birth, and soundex (a code based on last name using a method of encryption, see Fenna, 1984) are used for duplicate evaluation. Since the frequency of letters in last names is not uniform, the soundex does not have a uniform distribution over its possible values in the general population. This is also true among persons with AIDS. Although sex is fairly uniformly distributed in the general population, persons with AIDS are more likely to be male than female. Dates of birth for persons with AIDS do not have a uniform distribution over calendar dates, so we cannot directly apply the occupancy model with equal occupancy probabilities to evaluate the duplicate reporting in AIDS surveillance. Note, however, that the date of birth can be considered a combination of birth year and birth day in a year. Among persons diagnosed with AIDS, birth year does not have a uniform distribution over the calendar years. However, the birth day is quite uniformly distributed within a calendar year. Therefore, if we stratify the reported AIDS cases by sex, soundex, and birth year, then we can apply the occupancy results developed in section 2 to evaluate duplicate reporting based on matched birth days within each stratum.

Table 2: Estimates related to duplicate reporting of AIDS cases reported to CDC by June of 2004.

	Male	Female	Total
Number of reported cases (n)	748,950	173,312	922,262
Number of distinct codes* (k)	675,132	165,284	840,416
Estimated number of AIDS cases (r)	716,230	168,308	884,538
standard deviation of r	233	58	240
95% confidence interval: low	715,773	168,195	884,068
high	716,686	168,421	885,008
Number of potential duplicates ($n - k$)	73,818	8,028	81,846
Expected number of non-duplicates ($r - k$)	41,098	3,024	44,122
Estimated number of true duplicates ($n - r$)	32,720	5,004	37,724
Percent of true duplicates ($(n - r)/r$)	4.37	2.89	4.09
95% confidence interval: low	4.31	2.82	4.04
high	4.43	2.95	4.14

*combination of sex, soundex, and date of birth.

If there were no true duplicates (multiple cases reported for the same person) in the AIDS surveillance system, then given the number of cases reported to the system (which would equal the number of persons with AIDS reported to the system), the number of distinct combinations of sex, soundex, and date of birth would satisfy the equations provided in the previous sections. In this application,

persons or cases are considered as balls, and cells comprise the various combinations of sex, soundex, and date of birth. Since the number of sex, soundex, and date of birth combinations is observable and not affected by true duplicate reporting, we can work backwards to estimate the number of reported persons with AIDS. If the actual number of reported cases is greater than this estimated number of distinct persons reported, then there exists evidence of duplicate reporting (true duplicates) in the AIDS case reporting system.

As of June 2004, there were 922,835 AIDS cases reported to CDC. After eliminating 573 cases with unknown sex or soundex, or incomplete date of birth, $n = 922,262$ cases are used in our duplicate analysis; they “occupy” 840,416 distinct combinations of sex, soundex, and date of birth, thus revealing 81,846 potential duplicates. By following the procedure described in Section 2, we estimate that 37,724 or 4.09% of total reported cases are true duplicates with a 95% confidence interval (4.04%, 4.14%). This percentage varied by sex (see Table 2).

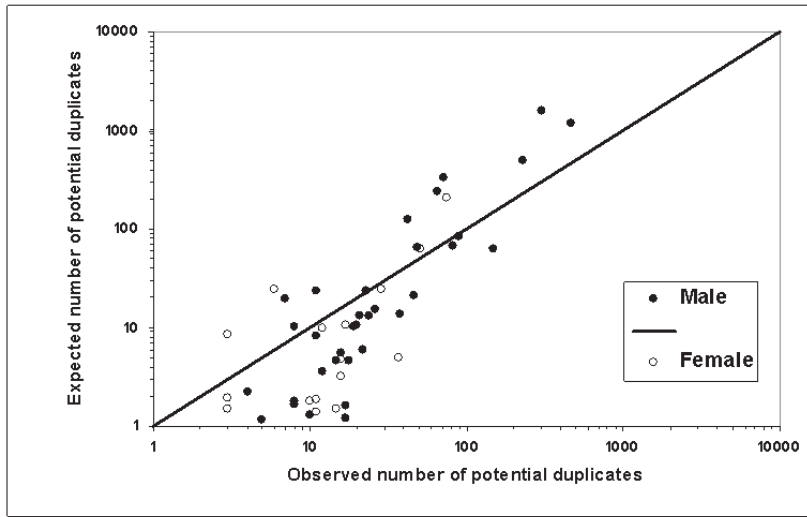


Figure 2: Observed number of potential duplicates vs. expected number of potential duplicates (without true duplicates) within each diagnosis state by sex. Results are based on cases reported to CDC by June of 2004.

We next analyzed the potential duplicates for cases diagnosed *within* each state. If there are no true duplicates in each state, then the number of potential duplicates can be estimated using the formulas provided in section 2. Based on AIDS cases reported to CDC by June of 2004, we compare the observed numbers of potential duplicates in each state to the expected number of potential

duplicates (Figure 2). States with a small number of AIDS cases tend to have more true duplicates (the observed number of potential duplicates greater than the expected number of potential duplicates), while in states with a large number of AIDS cases, the observed numbers of potential duplicates are consistently less than the expected numbers of potential duplicates.

We also evaluated the problem of duplicate reporting of AIDS cases *between* states. Cases diagnosed in one state may have the same partial personal identifier as cases diagnosed in other states. The number or proportion of true inter-state duplicates can be estimated using the method described in section 4. Results based on AIDS cases reported to CDC by June of 2004 are shown in Figure 3. States with a smaller number of potential duplicates tend to have a higher proportion of true duplicates.

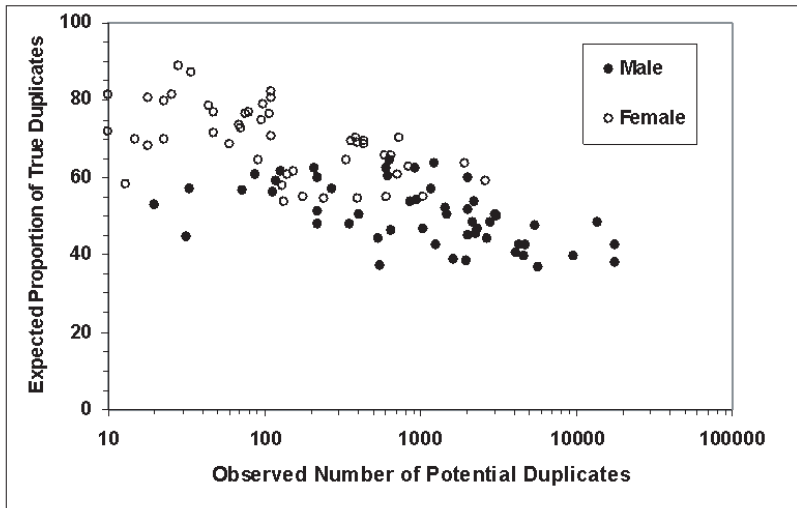


Figure 3: Expected proportion of true duplicates among potential duplicates in a state with other states. Results are based on cases reported to CDC by June of 2004.

6. Summary and Discussion

In this paper, we provided formulas to calculate the exact variance of the number of empty cells in occupancy problems with equal or unequal occupancy

probabilities. We also considered a generalized occupancy problem with balls of different colors. These results are useful for evaluating duplicate reporting in case surveillance of a disease in which a unique identifier is not available but for which duplicate case reporting is likely to occur.

Although the duplication analysis cannot tell whether a particular pair of potential duplicates is a pair of true duplicates or non-duplicates, it does help identify problems and can be used to estimate the magnitude of true duplicates in a surveillance system. The precision of the duplicate analysis depends on the number n of cells and the number r of cases in each stratum. More precisely, it depends on r and the probability that two distinct cases have the same code composed of a combination of personal identification variables. This probability takes on a minimum value of $1/n$ when all possible codes are equally likely. On the other hand, if, for example, soundex (with more than 8000 possible outcomes) is the only partial personal identifier considered, then two persons have approximately a $1/450$ probability of having the same code.

An assumption underlying our work is that the data elements used as a partial personal identifier, or stratification variables, have values that do not vary over time for each individual. If for some reason, a person's partial personal identifiers used for duplication assessment change, then the method will underestimate the number of true duplicates. For example, a woman's soundex could be changed due to marriage. Also, data entry errors of partial personal identifiers could result in underestimating the number of true duplicates unless the same error occurs consistently in reporting.

Using this occupancy model, we estimate that approximately 4% of AIDS cases reported in the national AIDS surveillance system up to June 2004 represent duplicate reports. This is consistent with previous findings (page 40, Centers for Disease Control and Prevention, 2004) that less than 5% of HIV and AIDS cases in the national surveillance database are duplicates, and is in compliance with recommended performance standards for well functioning surveillance systems that set a minimum performance standard for duplicate reports of $\leq 5\%$. These methods could be used to monitor national and state surveillance systems as a performance indicator — should estimated duplication rates exceed 5%, evaluation of surveillance practices for intrastate duplication efforts and communication between states for interstate duplication assessment can be undertaken.

Acknowledgement

We thank the editor and the referee for their constructive comments which led to a substantial improvement of this paper.

Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control

and Prevention.

References

- Centers for Disease Control and Prevention (2004). *HIV/AIDS Surveillance Report 2003*, (Vol. 15). US Department of Health and Human Services, Centers for Disease Control and Prevention.
- Chistyakov, V. P. (1967). Discrete time distribution in the problem of balls falling in cells with arbitrary probabilities. *Math. Notes Acad. Sc. USSR* **1**, 6-11.
- Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. I, third edition. John Wiley and Sons.
- Fenna, D. (1984). Phonetic reduction of names. *Computer Programs Biomedicine* **19**, 31-36.
- Holst, L. (1971). Limit theorems for some occupancy and sequential occupancy problems. *Annals of Mathematical Statistics* **42**, 1671-1680.
- Larsen, S. O. (1994). Estimation of the number of people in register from the number of birthdates. *Statistics in Medicine* **13**, 177-183.
- Sevastyanov, B. A. (1972). Poisson limit law for a scheme of sums of dependent random variables. *Theoretical Probability Application* **17**, 695-699.

Received September 8, 2005; accepted December 6, 2005.

Ruiguang Song
CDC Mail Stop E-48
1600 Clifton Road
Atlanta, GA 30333, USA
RSong@cdc.gov

Timothy Green
CDC Mail Stop E-48
1600 Clifton Road
Atlanta, GA 30333, USA
TGreen@cdc.gov

Matthew McKenna
CDC Mail Stop E-47
1600 Clifton Road
Atlanta, GA 30333, USA
MMcKenna@cdc.gov

M. Kathleen Glynn
CDC Mail Stop E-47
1600 Clifton Road
Atlanta, GA 30333, USA
KGlynn@cdc.gov