

Probabilistic population estimation of the size and overlap of data sets based on date of birth

Steven M. Banks^{*,†} and John A. Pandiani

The Bristol Observatory, 521 Hewitt Road, Bristol, VT 05443, U.S.A.

SUMMARY

Probabilistic population estimation is a statistical procedure for deriving unduplicated counts of the number of people represented in data sets that do not include unique person identifiers and the number of people shared by data sets that do not share personal identifiers. Because the procedure relies on anonymous data sets, the personal privacy of individuals and the confidentiality of medical records is protected. This paper describes the mathematics of probabilistic population estimation, and applies the procedure to an important contemporary public policy issue. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

In recent years, the proliferation of electronic databases in conjunction with dramatic advances in data processing technology have led to increasing concern about the confidentiality of medical records [1–4]. The nation is engaged in a major public policy debate. This debate involves a classic confrontation of public and private goods [5]. Personal privacy is an important value in contemporary American society. Regulation in support of personal privacy and the confidentiality of medical records would certainly receive widespread support if it did not have the potential to undermine another set of important values. These values favour rational administration of public programmes and the right of people to be able to make informed choices regarding their health care. Lawmakers and the public health community are faced with a major public policy dilemma that is the result of the apparent tension between two contemporary American values. This problem is particularly important in areas where stigmatizing conditions, such as mental illness, raise the level of sensitivity of people in the stigmatized group and among members of the general population.

This paper will describe probabilistic population estimation [6]. Probabilistic population estimation is a statistical procedure for determining the unduplicated number of people represented in data sets that include multiple records per person but do not include unique person identifiers. Section 2 of this paper will provide the mathematical derivation of the statistic. Section 3 will describe a method for using these unduplicated person counts to measure the

* Correspondence to: Steven M. Banks, The Bristol Observatory, 521 Hewitt Road, Bristol, VT 05443, U.S.A.

† E-mail: bristob@together.net

unduplicated number of people who are shared by data sets that do not include unique person identifiers. Section 4 will present the results of a series of simulation studies that provide the statistical properties of the estimator. Section 5 will illustrate the use of the estimator in policy research. In conclusion, Section 6 will discuss the utility of these estimates with regard to a wide range of health policy issues.

2. MATHEMATIC DERIVATION

Probabilistic population estimation uses a solution to the classical mathematical occupancy problem (coupon collector problem) to determine the unduplicated number of people represented in data sets that do not include unique person identifiers. In this paper, the techniques and results from the study of urn models [7] will be used to provide a theoretical justification for the technique. This technique provides an alternative to the work of Larsen [8], but differs in both the theoretical underpinning and in its results. A comparison of the two methods and their results will be provided.

The solution to the coupon collector problem used in probabilistic population estimation begins with a decomposition of the problem. This decomposition involves breaking down the larger question (how many unique individuals are represented in a data set) into a series of smaller questions for which the mathematical solution is known. In this case, a data set is first divided into discrete segments that describe individuals who share a year of birth and gender. Then, using a decomposition argument, the expected total number of individuals needed to fill a prespecified number of dates of birth is equivalent to the expected number of individuals needed to fill one date of birth, plus the expected number needed to fill a second date of birth once the first is filled, plus the expected number necessary to fill a third once the second is filled etc., until the prespecified number of dates of birth is filled. For birth dates, when a uniform distribution is assumed, the expected number of individuals is determined by

$$P_j(d) = \sum_{i=1}^d \frac{365}{365 - i}$$

where P_j is the estimate of the population in a distinct gender/year of birth cohort, and d (which depends on the specific cohort j) is the number of observed birth dates within that cohort.

The variance of the number of people is determined by

$$\sigma^2(P_j(d)) = \sum_{i=1}^d \frac{(i \times 365)}{(365 - i)^2}$$

The total number of people represented in a data set is obtained by summing the population parameters over all gender/year of birth cohort subsets. These formulae are the basis for the probabilistic population estimation procedure.

To establish the theoretical derivation of probabilistic population estimation, we begin by examining the problem of ‘sequential occupancy’ as described by Johnson and Kotz [7] in their book *Urn Models and their Applications*. In the language of urn models, let there be (m) urns and let the probability of a ball landing in an urn be $(1/m)$. Fix in advance a required number of empty urns (k) , and count the number of balls, which we will call $N(k)$, necessary

to achieve the desired number of empty urns. In this setting, $N(k)$ can be regarded as the sum of $(m - k)$ random variables, Y_1, Y_2, \dots, Y_{m-k} , where Y_j is equal to the number of balls needed to place the first ball in the j th new urn once $(j - 1)$ urns already have at least one ball in each of them. The Y 's are random variables, mutually independent, but not identically distributed. In this setting, $Y_1 = 1$, and Y_j , for each $j \geq 2$ has a geometric distribution, with

$$E(Y_j) = \frac{m}{(m - j + 1)}$$

and the variance estimator

$$\text{var}(Y_j) = \frac{m(j - 1)}{(m - j + 1)^2}$$

Probabilistic population estimation does not fix in advance the number of birthdays that will be seen (or left empty). Instead the estimate is conditioned on the number of observed birthdays. In addition, and more importantly, the original intent of the sequential occupancy problem was to determine the number of balls that were necessary to first place a ball in the k th urn. In our setting, we know that k birthdays (urns) have been observed (filled), however, we do not know that the last record (ball) was the first to fill the k th birthday. Thus, when we have observed k birthdays, we know that, at a minimum, we are at least one person shy of seeing the $k + 1$ th birthday. Therefore, when we observe k birthdays, we define $P(k) = N(k + 1) - 1$, where $N(k + 1)$ is defined as in the sequential occupancy problem above. Algebraic simplification and a mathematical change of variables yields the two formulae that are used for probabilistic population estimation.

As noted earlier, Larsen [8] provided a solution to the problem of the number of unique individuals in a data set that used a maximum likelihood approach. He applied his solution to the estimation of the number of people represented in an anonymous chlamydia registry in one county in Denmark. When the same data are analysed using probabilistic population estimation, the estimate of the expected number of people was almost identical to Larsen's estimate (probabilistic population estimation indicated 1128 people, compared to 1127 in Larsen's paper). The similarity in the estimates using the two procedures is not a coincidence. Larsen's estimate of the number of individuals in a specific gender/year of birth cohort may be rewritten as an integral. It may be shown that probabilistic population estimation is the upper Riemann sum of that integral.

While the population estimates provided by the two methods are very similar, the variances are very different. The variance estimate provided by Larsen is seven times as large as the estimate provided by probabilistic population estimation (174 for probabilistic population estimation compared to Larsen's 1298). The 95 per cent confidence interval constructed using Larsen's method is 2.7 times larger than the confidence interval provided by probabilistic population estimation.

The smaller confidence intervals associated with probabilistic population estimation are of value, of course, only if these confidence intervals include the true value. In order to compare the validity of the confidence intervals provided by probabilistic population estimation with the confidence intervals provided by Larsen's probabilistic computations, a data set that describes all people discharged from New York State Office of Mental Health facilities between 1993 and 1995 was analysed. This data set includes multiple records for some people, but it does contain a unique person identifier, so the actual number of people represented is known.

Table I. Comparison of the coverage percentage of confidence intervals probabilistic population estimation (PPE) and Larsen's probabilistic technique.

Technique	Level of confidence		
	95%	90%	80%
PPE	94.8%	89.1%	79.8%
Larsen	100%	99.4%	98.4%

Probabilistic population estimation and the method used by Larsen were applied to this data set for each of the 193 gender/year-of-birth cohorts and confidence intervals associated with 95 per cent, 90 per cent and 80 per cent confidence were constructed. Table I shows the proportion of the 193 cohorts in which the specified confidence interval included the actual number of people.

The results of this analysis clearly demonstrate that Larsen's procedure produces confidence intervals that substantially overstate the uncertainty of the estimate. Larsen's confidence intervals included the actual population size much more frequently than specified, indicating that the confidence intervals are not accurate. The confidence intervals provided by probabilistic population estimation included the actual population size at almost exactly the specified level of confidence.

3. CASELOAD OVERLAP

To probabilistically determine the number of people shared across data sets that do not include a common person identifier, the sizes of three populations are determined, and the results are compared. First, the number of people represented in each of the original data sets is determined. Second, the two data sets are combined (concatenated), and the number of unique individuals represented in the combined data set is determined.

The number of people who are shared by the two data sets is the difference between the sum of the numbers of people represented in the two original data sets and the number of people represented in the combined data set. In terms of mathematical set theory [9], the size of the intersection of two sets ($A \cap B$) is the difference between the sum of the sizes of the two sets ($A + B$) and the size of the union of the two sets ($A \cup B$):

$$n(A \cap B) = n(A) + n(B) - n(A \cup B)$$

where $n(\cdot)$ is the number of unique individuals included in a data set. The size of the two original data sets and the size of the union of the two may be obtained using the probabilistic procedure described in Section 2. Because some of the terms in the union of A and B are also included in the size of the larger data set, the variance estimator will be substantially smaller than would otherwise be the case. (For a more detailed discussion of the mathematical derivation of this statistical procedure, see Banks and Pandiani [10].)

One verification study included a comparison of the probabilistically predicted population size and overlap of the caseloads of the Vermont State Hospital and Vermont Correctional Facilities to actual parameters for 1989 to 1995 [11]. The confidence intervals for both populations were less than two per cent (plus or minus) of the point estimate for every year.

Table II. Vermont Correctional Facilities and Vermont State Hospital caseload size and overlap: 1989–1995.

Year	Correctional Facilities		State Hospital		Both	
	(Estimated)	(Actual)	(Estimated)	(Actual)	(Estimated)	(Actual)
1995	5240 ± 70	5282	324 ± 3	323	73 ± 14	65
1994	4947 ± 68	4931	335 ± 3	333	71 ± 15	62
1993	4971 ± 69	4959	313 ± 3	311	57 ± 15	58
1992	4878 ± 69	4866	390 ± 4	389	65 ± 16	61
1991	4819 ± 70	4806	408 ± 4	407	74 ± 16	78
1990	4744 ± 69	4716	454 ± 4	455	91 ± 19	90
1989	4416 ± 67	4395	521 ± 5	524	121 ± 17	100

All 14 of the actual population size parameters were included by the 95 per cent confidence intervals. The confidence intervals for the overlap estimates were somewhat larger (between 15 per cent and 20 per cent) and six of the seven actual population overlap parameters were included by the confidence intervals. The results of this verification are presented in Table II.

4. RESULTS OF A SERIES OF STUDIES

A large number of simulation studies have been conducted to provide information on the statistical properties of the estimator. Simulation studies have focused on the impact of three factors on the accuracy of the estimates. These factors include population size, the amount of overlap between population, and the uniformity of dates of birth in the population.

The formal derivation of probabilistic population estimation assumes that dates of birth are uniformly distributed across a year. The distribution of dates of birth has been investigated by Nunnikhoven [12] using 10 years of vital records data for the United States. He found very small departures from the uniformity in dates of birth. This small departure from uniformity actually tends to decrease the bias associated with probabilistic population estimation as will be seen below.

Numerous simulation studies of the effect of variation in uniformity on the bias associated with the probabilistic estimates have been performed. Simulations have been conducted under conditions in which the distribution of dates of birth are uniform and where they depart from uniformity. The non-uniform condition that was modelled had the distribution of dates of birth exceed the uniform by 5 per cent for half of the year and set the distribution to be 5 per cent less than the uniform for the other half of the year. Each point on the graph of the results of these simulation studies, shown in Figure 1, represents a minimum of 9000 simulations, and some represent more than 60 000 simulations.

When the dates of birth are uniform, probabilistic population estimation always has a small positive bias (approximately 0.5 per cent) across most of the range of the 365 potential birthdays. The bias is larger when more than 90 per cent of the dates of birth are filled. However, when dates of birth deviate from uniformity, probabilistic population estimation becomes less biased, and includes examples of both negative and positive bias over the range of observed dates of birth.

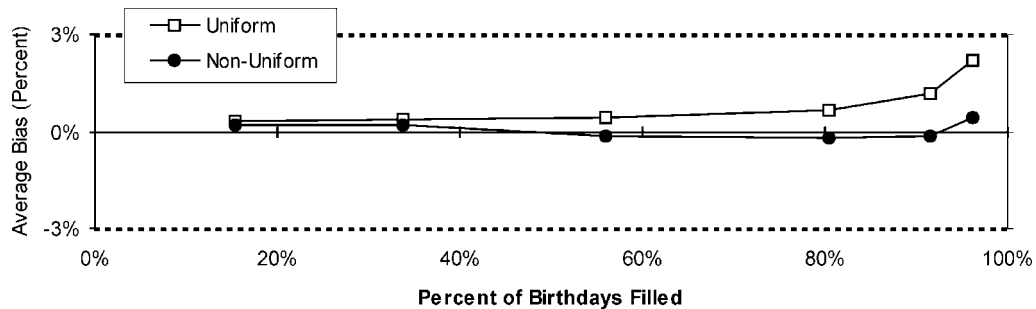
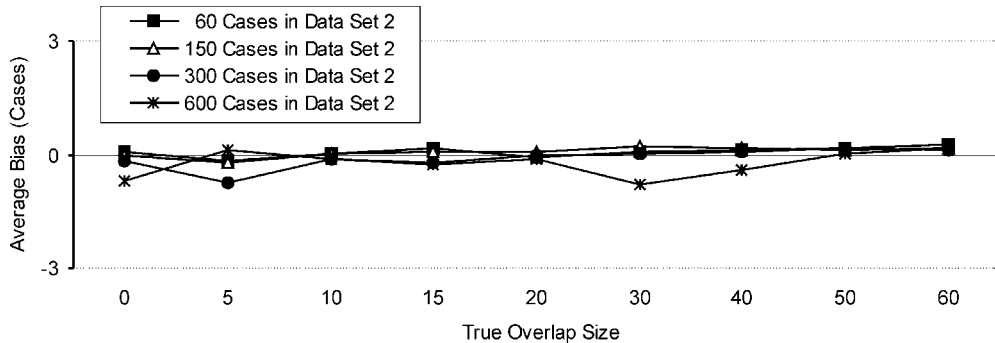


Figure 1. Validity of population size estimates for uniform and non-uniform distributions of birthdays.

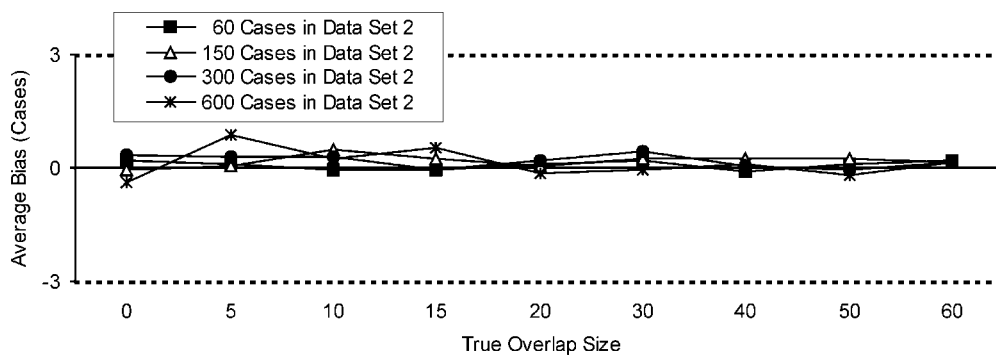


60 Cases in Data Set 1

Figure 2. Validity of population overlap estimate for uniform distribution.

The simulation studies for population overlap are somewhat more complex, as they may be influenced by the size of both data sets and by the amount of overlap. For the simulation studies of population overlap described here, one data set was fixed at 60 individuals. The size of the second data set was varied from 60 to 150, 300 and 600. The size of the overlap in the simulations was varied from 0 to 5, 10, 15, 20, 30, 40, 50 and all 60 individuals. Figure 2 presents the results of the simulations for uniformly distributed dates of birth, and Figure 3 presents the results of the simulations for non-uniformly distributed dates of birth. Little bias is observed across the range of potential overlap for either the uniform or the non-uniform condition. In this demonstration and in numerous other simulation studies, probabilistic population estimation has produced reliable and valid estimates of population overlap, and these estimates improve when there is a slight departure from uniformity in dates of birth.

These simulations indicate that probabilistic population estimation yields little or no systematic bias in the production of point estimates for either population size or population overlap. These simulations also demonstrate that the 95 per cent confidence intervals have appropriate coverage.



60 Cases in Data Set 1

Figure 3. Validity of population overlap estimate for non-uniform distribution.

5. THE ESTIMATOR IN POLICY RESEARCH

For purposes of illustration the process and results of applying these analytic tools to a current public health policy concern will be described. For more than 30 years, public policy in the United States has favoured the reduction of state psychiatric hospital populations and the development of community alternatives. This policy has raised concern about the care of the people who would have been served in state hospitals. This concern is frequently based on the expectation that the care and custody of these people is being transferred to other institutions that are less able to provide appropriate services. The two institutions most often mentioned as possible destinations for discharged or diverted patients are correctional facilities [13] and local general hospitals [14]. This idea that one institutional setting may serve as a substitute for another may be referred to as the hypothesis of 'transinstitutionalization'.

This application of the method of probabilistic population estimation will focus on the degree to which discharged state hospital patients utilized alternative institutions by measuring rates of incarceration and hospitalization after state hospitalization in 17 up-state New York counties. The subjects of this study are 638 adults (446 men and 192 women) who had an episode of state hospitalization in one of 17 up-state New York counties during 1995. Data sets that included the date of birth and gender and a service system specific person identifier for all state hospital patients in these counties were obtained from the New York Office of Mental Health.

In order to determine the rates of incarceration of members of this state hospital cohort, data sets that included the date of birth and gender of all adults who had been incarcerated in local jails in these counties during 1996 were obtained from the county correctional authorities. In order to determine rates of hospitalization for psychiatric treatment in local general hospitals, similar data sets were obtained from the New York State Department of Health.

Because of the substantial differences between men and women in rates of incarceration and rates of psychiatric treatment in general hospitals, all analyses were conducted separately for the two genders. For purposes of comparison, the incarceration rate and the hospitalization rate for the general population of these 17 counties were determined by dividing the number of people served in each of these sectors by the total adult population of the regions. These rates

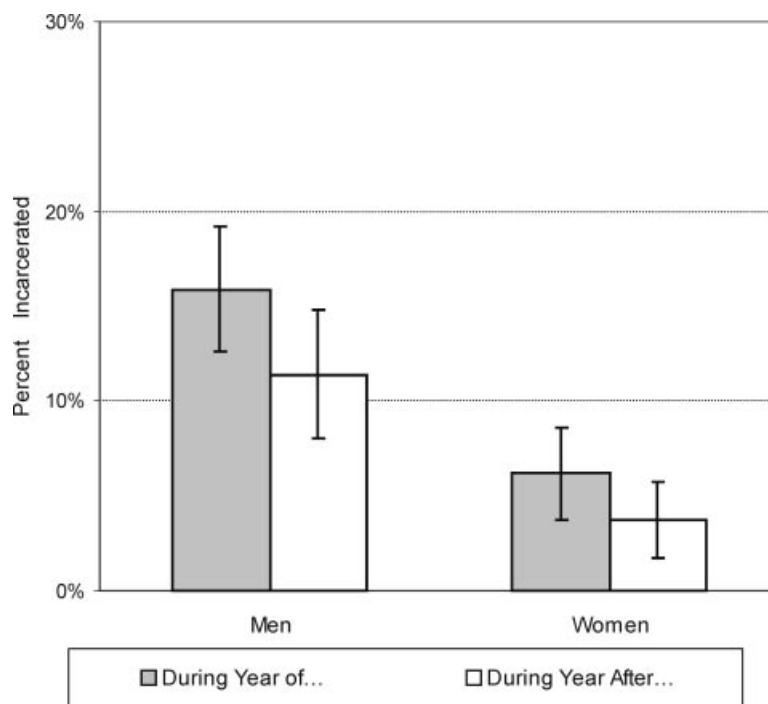


Figure 4. Rates of incarceration (w/95 per cent CI) during year of state hospitalization and during year after state hospitalization.

are used for comparison with the rates of incarceration and hospitalization of state hospital patients.

People who spent time in state hospitals in New York State were substantially more likely than members of the general population to have spent time in local jails and general hospitals (all $p < 0.001$). Men who had spent time in a state hospital were more than seven times as likely as other men to be incarcerated and more than 21 times as likely to be hospitalized. Women who had spent time in a state hospital were more than 12 times as likely as other women to be incarcerated and more than 80 times as likely to be hospitalized.

For people who had spent time in a state psychiatric hospital during the study period, rates of incarceration were significantly lower (both $p < 0.05$) during the year after state hospitalization than during the year in which an episode of state hospitalization had occurred. This pattern was evident for both men and women. Men were more likely than women to be incarcerated ($p < 0.01$) during both time periods (Figure 4).

Rates of hospitalization in alternative inpatient settings for people who had spent time in a state psychiatric hospital were also significantly lower during the year after state hospitalization than during the year in which an episode of state hospitalization had occurred ($p < 0.001$). This pattern was evident for both men and women. Women were more likely than men to receive psychiatric services in local general hospitals ($p < 0.01$) during both time periods (Figure 5).

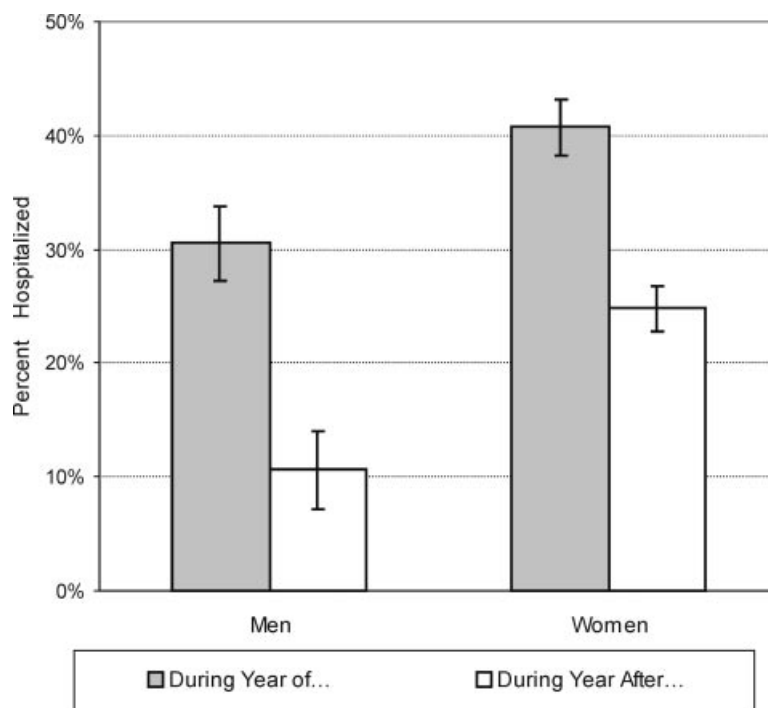


Figure 5. Hospitalization in other settings (w/95 per cent CI) during year of state hospitalization and during year after state hospitalization.

6. CONCLUSION

This paper has provided a summary description of the mathematical derivation of probabilistic population estimation, a useful tool for addressing many contemporary public health issues. The simulation studies demonstrate that the method produces valid and reliable estimates across a wide range of conditions using only two variables (date of birth and gender). This tool will be increasingly useful, as comprehensive data sets become more widely available in both public and private sector settings.

This tool will be particularly valuable as increasing concerns about personal privacy and the confidentiality of medical records lead to increasing limitation on access to data sets that include personally identifiable information. In the application described in this paper, an important public policy issue that involved a particularly sensitive population was addressed without any threat to personal privacy. This methodology has also been used to evaluate the efficacy of substance abuse treatment programs [5], compare clinical outcomes of groups who were treated under differing protocols [15] and to measure the treated prevalence of a disorder in a statewide population [6]. In the future this methodology should prove to be useful for research on the distribution and outcomes of disorders where confidentiality and privacy are considered to be especially important, such as is the case with HIV and AIDS.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the contributions of Peter Landi, M.S. and Peter Lambert, Ph.D., to the data management and analysis for this project, and Lucille M. Schacht, M.S., for her assistance with data presentation.

REFERENCES

1. Secretary's Advisory Committee on Automated Personal Data Systems. Records, Computers, and Rights of Citizens: Report of the Advisory Committee on Automated Personal Data Systems, US Department of Health, Education, and Welfare. U.S. Government Printing Office: Washington, DC, 1973.
2. Hendricks E, Hayden T, Novik JD. *Your Right to Privacy: A Basic Guide to Legal Rights in an Information Society*. Southern Illinois University Press: Carbondale, IL, 1990.
3. Donaldson MS, Lohr KN. *Health Data in the Information Age: Use Disclosure and Privacy*. Institute of Medicine, National Academy Press: Washington, DC, 1994.
4. Shalala DE. Confidentiality of Individually-Identifiable Health Information; Recommendations of the Secretary of Health and Human Services, pursuant to section 264 of the Health Insurance Portability and Accountability Act of 1996. 11 September 1997. <http://aspe.os.dhhs.gov/admnsimp/PVCREC2.HTM#COVERAGE>.
5. Pandiani JA, Banks SM, Schacht LM. Personal privacy vs. public accountability: a technological solution to an ethical dilemma. *Journal of Behavioral Health Services and Research* 1998; **25**(4):456–463.
6. Banks SM, Pandiani JA. The use of state and general hospitals for inpatient psychiatric care. *American Journal of Public Health* 1998; **88**:448–451.
7. Johnson NL, Kotz S. *Urn Models and their Applications. An Approach to Modern Discrete Probability Theory*. Wiley: New York, 1977.
8. Larsen OL. Estimation of the number of people in a register from the number of birthdates. *Statistics in Medicine* 1994; **13**:177–183.
9. Whitehead AN, Russell B. *Principia Mathematica*. Vol. 1, 2nd edn. Cambridge University Press: Cambridge, 1927.
10. Banks SM, Pandiani JA. *A Methodology for Probabilistically Estimating Caseload Size and Overlap*. The Evaluation Center, HSRI, 1999.
11. Pandiani JA, Banks SM. Service Systems Research in the Information Age. Proceedings of the Sixth Annual Conference on State Mental Health Agency Services Research and Program Evaluation. National Association of State Mental Health Program Directors, Arlington, VA, 1996.
12. Nunnikhoven TS. A birthday problem solution for non-uniform birth frequencies. *American Statistician* 1992; **46**:270–274.
13. Torrey EF, Stieber J, Ezekiel J, Wolfe SM, Scharfstein J, Noble JH, Flynn LM. *Criminalizing the Seriously Mentally Ill: The Abuse of Jails as Mental Hospitals*. Public Citizen's Health Research Group and the National Alliance for the Mentally Ill: Washington, DC, 1992.
14. Regier A, Goldberg D, Taube A. The de facto US Mental Health Services Delivery system: a public health perspective. *Archives of General Psychiatry* 1978; **35**:685–693.
15. Banks SM, Pandiani JA, Gauvin L, Reardon E, Schacht LM, Zovistoski A. Practice patterns and hospitalization rates: a statewide program evaluation. *Administration and Policy in Mental Health* 1998; **26**:33–44.