A Methodology
For Probabilistically Estimating
Caseload Size and Overlap

Applied to Inpatient Behavioral Health Care Services Provided to Residents of Vermont During 1990-1994

Steven M. Banks, Ph.D.
The Bristol Observatory, Ltd.

John Pandiani, Ph.D.
Vermont Department of Developmental and Mental Health Services

The analysis of inpatient data set reported here was supported by The Evaluation Center @ HSRI, and the Vermont Mental Health Statistics Improvement Project (HR1 SM46203). The method of probabilistically population estimation was the result of independent collaboration of the authors. The authors gratefully acknowledge the assistance of Lisa Gauvin and Lucille Schacht. The authors also thank the project by the Vermont State Hospital, the Vermont Health Care Authority, and the Brattleboro Retreat for making data available to the project.

For more information contact Steven Banks
The Bristol Observatory
21 Hewitt Road
Bristol, VT 05443
802.453.3353 bristob@together.net

Introduction To Probabilistic Population Estimation

Probabilistic Population Estimation, a new statistical procedure, provides information on the unduplicated number of people who are involved in more than one institution, program, group, or activity, either concurrently or in sequence. This information can help to answer important questions about access to services and treatment outcomes that have previously remained unanswered due to the lack of unique person identifiers within and across data sets.

Access to care is one of the most important issues regarding behavioral health care in the emerging managed care environment. The complexity of the behavioral health care service sector and the lack of common person identifiers across service providers have severely limited the development of measures of access to services. With Probabilistic Population Estimation, the lack of common identifiers is no longer a problem and the complexity of the problem is greatly simplified (Banks and Pandiani, Forthcoming). Combined with emerging methodologies for estimating the prevalence of disorders in general populations, Probabilistic Population Estimation can provide valid and reliable measures in one of the most important and most difficult issues facing program administrators and payors today (Pandiani, Banks, and Gauvin, 1997).

Probabilistic Population Estimation is also a powerful tool for evaluating treatment outcomes across geographical regions, provider organizations, treatment programs, or managed care plans (Pandiani and Banks, 1997). Rates of hospitalization subsequent to community mental health treatment, for instance, may be determined using Probabilistic Population Estimation to measure the amount of overlap between community program caseload during one year, and inpatient population during subsequent years (Banks and Pandiani, 1997). When probabilistic estimates of incarceration subsequent to treatment (outcomes) are combined with probabilistic estimates of incarceration before treatment (access), a very powerful risk adjusted measure of program performance is the result (Pandiani, Banks, Schacht, Forthcoming).

Probabilistic Population Estimation has three important advantages. First, the personal privacy of individuals and the confidentiality of medical records are assured because Probabilistic Population Estimation does not depend on information that identifies specific individuals. Second, because the methodology relies on existing data bases, it does not require the commitment of substantial amounts of staff time or financial resources. Finally, Probabilistic Population Estimation can support retrospective evaluation of changes in systems of care that have occurred in the past, and provide longitudinal baseline data for evaluating current or anticipated changes in systems of care wherever basic client information resides in electronic data bases.

This working paper presents two applications of Probabilistic Population Estimation (PPE). The first is an estimation of the unduplicated number of residents of an entire state who were hospitalized for psychiatric or substance abuse services during a five year period. The second applies the PPE methodology to construction of a global measure of access in a managed mental health care system.

Methodology

The innovative analytical technique used in this project allows researchers, policy analysts, and evaluators to answer two basic questions that have frequently remained unanswered because existing data sets lack unique person identifiers across organizations and service sectors. “How many people have contact with a service system?” and “How many people are served by more than one organization, service sector, or service system?” The ability to answer these questions is based on *Probabilistic Population Estimation*, a statistical procedure derived from the solution to the classic mathematical “coupon collector” problem (Feller, 1957). The ability of this statistic to provide probabilistic estimates (with known confidence intervals) of these basic parameters of service systems is particularly valuable where issues of confidentiality or organizational complexity limit the availability of unique identifiers, or the lack of adequate financial resources inhibit the development of comprehensive integrated data warehouses. PPE combines information on the distribution of dates of birth in data sets with information on the distribution of dates of birth in the general population to produce valid and reliable estimates of the number of people represented. These estimates include unduplicated counts of the people represented in individual data sets that do not include unique person identifiers, and the number of people shared across data sets that do not include common person identifiers

In the classical coupon collector problem, the solution to the problem answers the question "How many baseball cards must a collector collect to obtain a complete set of cards, when the probability of every card being in a given bubble gum package is known?". In the current application, the same logic is used to answer the questions, “How many unique individuals are represented in a data set that does not include a unique person identifier?”, and “How many unique individuals are shared by data sets that do not include common person identifiers?”

Determining Population Size

The solution to the coupon collector problem used in Probabilistic Population Estimation begins with a decomposition of the problem which does not involve mathematical approximation. This decomposition involves breaking down the larger question into a series of smaller questions for which the mathematical solution is known. In this case, a data set is divided into discreet segments that describe individuals who share a year of birth and gender. Using decomposition, the total number of individuals needed to fill a prespecified number of dates of birth is equivalent to the number of individuals needed to fill one date of birth, plus the number needed to fill a second date of birth once the first is full, plus the 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 number of individuals is determined by:

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

where *j* represents a distinct gender/year of birth cohort, and *l* is the number of observed birth dates within that cohort.

The variance of the number of people is determined by:

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

The total number of people represented in a data set (P_{Total}) is obtained by summing the population parameters over all gender/year of birth cohort subsets:

$$P_{Total} = \sum_{j=1}^{k} P_j$$

where *k* is the total number of gender/year of birth cohort subsets.

The construction of the 95% confidence interval for the point estimate derived above involves a two step process. First, the total variance σ^2_{Total} is obtained by summing the variance for each gender/year of birth cohort:

$$\sigma^2_{Total} = \sum_{j=1}^{k} \sigma^2(P_j)$$

where *k* is the total number of gender/year of birth cohort subsets. The estimate for the

$$P_{Total} \pm 1.96\ \sigma(P_{Total})$$

95% confidence interval is then constructed:

Determining Population Overlap

Probabilistic Population Estimation also determines the amount of overlap between two data sets that do not include a common individual identifier. To find the population overlap, Probabilistic Population Estimation employs the same decomposition approach that is used to determine population size. First, the procedure described above for determining population size is applied to the first and the second original data sets. Then, the number of individuals needed to increase the number of unique birth dates from the larger of the two data sets to the total number of unique birth dates in a combined data set is calculated. This calculation is similar to the approach used in determining population size and can be represented by:

$$P_j(b,c) = \sum_{k=b+1}^{c} \frac{365}{365 - k}$$

where $P_j(b,c)$ is the population estimate for a gender/year of birth cohort "j" necessary for the larger number of unique birth dates from either the first or second data sets ("b") to increase to a total number of unique birth dates in the combined data set ("c").

When the number of birth dates in the larger of the original data sets is equal to the number in the combined data set, the determination is provided by:

$$P_j(b,c) = \frac{1}{2}\left(\frac{365}{365 - c}\right)$$

For each year of birth/gender cohort, the variance $\sigma^2(P_j(b,c))$ of the number of people necessary to increase the number of observed birth dates from the number of birth dates observed in the larger of the two original data sets to the number of birth dates observed in the combined data set is determined by:

$$\sigma^2(P_j(b,c)) = [\begin{array}{ll} \sum_{k=b+1}^{c} \frac{(k * 365)}{(365 - k)^2} & when\ \ b < c \\ \frac{1}{4} * \frac{(k * 365)}{(365 - k)} & when\ \ b = c \end{array}$$

The population overlap is the difference between the number of people represented in the smaller of the two original data sets and the number of people needed for the increase in the number of birth dates observed from the larger of the two original data sets to the combined data set. The method of calculating population overlap depends on whether the smaller number of individuals represented in either of the original data sets is known (based on a data set specific unique person identifier) or has been statistically determined by the steps identified above.

The calculation of population overlap is determined by:

$$P(\text{overlap}) = P(a) - P(b,c)$$

where P(a) is the number of people represented (population size) in the smaller data set and P(b,c) is the population size necessary for increasing the number of birth dates observed in the larger (i.e., having more observed birth dates) of the two original data sets to the number of birth dates observed in the combined data set. P(a) is the actual number of people when this number is known. When this number is not known, P(a) is the number determined by the calculations above.

When the sum of the number of birth dates in the two original data sets is equal to the number of birth dates in the combined data set, the overlap is set to zero. When the population overlap derived by the above step is less than zero, the population overlap parameter is set to zero which is the smallest overlap possible.

Calculating the variance of the population overlap $\sigma_P^2(\text{overlap})$ is determined by:

$$\sigma_P^2(\text{overlap}) = \sigma^2(P(a)) + \sigma^2(P(b,c))$$

where $\sigma^2(P(a))$ is the variance of the number of people represented (population size) in the smaller data set and $\sigma^2(P(b,c))$ is the variance of the population size necessary for increasing the number of birth dates observed in the larger (i.e., having more observed birth dates) of the two original data sets to the number of birth dates observed in the combined data set. The variance of the population size in the smaller data set $\sigma^2(P(a))$ is equal to zero when the number of people is known, and is calculated as indicated previously when the number of people is not known. When the sum of the number of birth dates in the two original data sets is equal to the number of birth dates in the combined data set, the variance of the overlap $\sigma_P^2(\text{overlap})$ is set to zero.

Application #1: Unduplicated Count of Vermont Residents Hospitalized for Behavioral Health Care Services During 1990-1994

This report presents previously unavailable information on the unduplicated number of residents of an entire state who were hospitalized for psychiatric or substance abuse services during a five year period. This unduplicated count includes episodes of hospitalization that occurred in general hospitals, state and private psychiatric hospitals and the Veterans Hospital in one state, and general hospitals in its neighboring states are included in this overview. The report is based on the analysis of existing administrative and operational data bases. No new data collection was required. The unduplicated counts of people served across service sectors where no common person identifiers are available is made possible by the application of *Probabilistic Population Estimation* (Patent Pending), a new statistical procedure that combines knowledge of the distribution of dates of birth in the general population with information of the distribution of dates of birth in data sets to determine the number of people who are represented in those data sets.

The need to determine the number of people who are treated for psychiatric and substance abuse disorders across treatment settings and service sectors has been recognized for at least two decades. In 1977, The President's Commission on Mental Health specifically requested information on the number of people were being treated in which settings. Regier et. al.(1978) observed that existing data resources could provide reasonable estimates of the amount of service being provided, but it would be extremely difficult, if not impossible, to answer important questions about the number of people who were being served. More recently, the NIMH National Plan (1991) argued that there is an urgent need to better understand the relationship of state and general hospital caseloads. "Although large numbers of patients are now seen in both psychiatric hospitals and psychiatric units of general hospitals, very little is known about how these facilities fit into the rest of the mental health system." (p. 38)

Data Sources

The results presented in this report are based on analysis of data sets from three sources: a Hospital Discharge Data Set maintained by the Vermont Department of Health, the Vermont State Hospital database maintained by that facility, and an extract from the database of the Brattleboro Retreat, the state's only private psychiatric hospital. There are no unique person identifiers that are shared by the three data sets.

The statewide Hospital Discharge Data Set provided computer records that describe every episode of hospitalization of a Vermont resident for mental health or substance abuse treatment in general hospitals in Vermont and neighboring New Hampshire, or in the Veterans Administration Hospital in White River Junction, VT. Records that describe episodes of hospitalization coded under the Mental Health and Substance Abuse Major Diagnostic Categories (MDCs) were abstracted by Health Department staff and provided to the research team. Each record included the date of birth, gender, and zip code of the patient, the beginning and end dates that define each episode of hospitalization, a facility identifier, and diagnostic codes associated with each episode. The records do not include any patient identifiers. The Hospital Discharge Data Set provided information on 14,460 episodes of hospitalization in 15 general hospitals in Vermont, 17 general hospitals in New Hampshire, as well as Vermont's Veterans hospital.

Data on episodes of hospitalization that occurred at the Vermont State Hospital (the state's only public psychiatric facility) were obtained from a standard State Hospital Inpatient Episode Data Set (Leginski, et. al., 1989). Records that describe every hospitalization of an adult at the state hospital during 1990 through 1994 were abstracted by the research team. This data set contains information on the date of birth, gender, and zip code of each patient, the beginning and end dates of each episode, and diagnostic codes. The Vermont State Hospital data set includes a unique patient identifier that was used to validate the estimation procedures described below (See Appendix 2). The Vermont State Hospital data set provided information on 2,177 episodes of hospitalization that involved 1,182 unique individuals.

Data on episodes of hospitalization of Vermont Residents that occurred at the Brattleboro Retreat were provided to the project by that institution. Records included information on the date of birth, gender and zip code of each patient, the beginning and end dates of each episode, and diagnostic codes. The Brattleboro Retreat data set also included a facility specific patient identifier that was used to validate the estimation procedures. The Brattleboro Retreat data set provided information on 1,447 episodes of hospitalization involving 1,030 Vermonters.

Together, these data sets provide a comprehensive accounting of the amount of inpatient behavioral health care provided to Vermont residents in Hospitals in Vermont and New Hampshire during 1990 through 1994. The major shortcoming of this aggregation of data sets is the lack of unique client identifiers that would allow the research team to determine the number of people who were hospitalized for behavioral health care, and to determine the degree to which individuals are served in more than one type of facility. The method of probabilistic population estimation described below is designed to allow researchers and evaluators to address these and other questions about the number of people served and the number of people shared by different organizations and service sectors when these organizations and service sectors do not share a common client identifier.

Results

Approximately 9,871± Vermont residents received inpatient care for mental health or substance abuse problems in Vermont or New Hampshire during 1990-1994. The Vermont State Hospital served 12%± of these people. General hospitals in Vermont served 66%±, and general hospitals in New Hampshire served 19%±. Ninety-three percent (±) of the people who received inpatient behavioral health care during this period were hospitalized in only one type of institution. Seven percent (±) received inpatient behavioral health care in more than one type of hospital.

(Detailed results are provided in Tables 1-4. In these tables, all person counts that are based on Probabilistic Population Estimation are accompanied by precise confidence intervals. In this discussion, confidence intervals of 2% or less are indicated by the symbol "±", for larger confidence intervals the actual value is specified.)

THE HOSPITALS

General Hospitals in Vermont

During 1990 through 1994, more Vermonters received inpatient behavioral health care from general hospitals than in any other setting. Sixty-six percent (±) of all Vermont residents who received inpatient behavioral health care services received services from general hospital in Vermont. Local general hospitals accounted for 57% of all episodes of hospitalization but only 29% of all patient days.

Most people who received behavioral health care in their local general hospitals during 1990-1994 did not receive similar services in another type of facility during the same time period. Only 8%± were also hospitalized in another type of facility for behavioral health care (8%± in the Vermont State Hospital, 7%± in New Hampshire hospitals, 4%± in the Brattleboro Retreat, and 1%± in the Vermont Veterans Administration Hospital).

Vermont State Hospital

During 1990 through 1994 the Vermont State Hospital provided more inpatient behavioral health care to Vermonters than any other type of facility. The Vermont State Hospital accounted for 46% of all patient days, but only 12% of all episodes of care and 12%± of all people served.

Forty-six percent (±) of Vermonters served by the Vermont State Hospital also received services from another hospital. Forty-two percent (±) were also served by a general hospital. Eight percent (±) received services at New Hampshire hospitals. Six percent (±) were hospitalized at the Brattleboro Retreat. Only 3% (±) were served by the Veterans Administration Hospital in White River Junction.

The Brattleboro Retreat

Between 1990 through 1994, a growing number of Vermonters received inpatient behavioral health care from the Brattleboro Retreat. Ten percent (±) of all Vermonters who received inpatient behavioral health care during this period were served by the Brattleboro Retreat. The number of Vermonters served by the Brattleboro Retreat increased by 45%, from 204 to 296 during the study period. The Brattleboro Retreat accounted for 9% of all patient days and 8% of all episodes of hospitalization between 1990 and 1994.

Thirty-four percent (±) of all Vermonters served by the Brattleboro Retreat also received services from another hospital. Twenty-six (±) percent were also served by general hospitals in Vermont. Thirteen percent (±) were also served by New Hampshire hospitals. Six percent (±) were also treated in the Vermont State Hospital. Two percent (±) were also served by the Veterans Administration Hospital in White River Junction.

The Veterans Hospital in White River Junction, VT

The Veterans Hospital in White River Junction accounted for a relatively small portion of the behavioral health care services provided to Vermonters overall. Only 7%± of all Vermonters who received inpatient behavioral health care services were served by the Veterans Hospital. The Veterans Hospital accounted for 5% of all patient days and 7% of all episodes of hospitalization between 1990 and 1994.

Nineteen percent (±3%) of Vermonters served by the Veterans Hospital also received services from another hospital. Thirteen percent (±3%) were also served by Vermont general hospitals. Five percent (±3%) were also treated in the Vermont State Hospital, and 3% (±3%) were also treated at the Brattleboro Retreat. Four percent (±3%) were also served by New Hampshire Hospitals.

New Hampshire Hospitals

Between 1990 and 1994, a growing number of Vermonters received inpatient behavioral health care from general hospitals in New Hampshire. General hospitals in New Hampshire served 19%± of all Vermonters who received inpatient behavioral health care services. These hospitals accounted for 11% of the patient days and 16% of the episodes of hospitalization for these people.

The number of Vermonters who received behavioral health care in New Hampshire general hospitals increased from 374± to 558± between 1990 and 1994. Twenty-nine percent (±) of the Vermonters served by New Hampshire general hospitals also received services from a hospital in Vermont. Twenty-four percent (±) were also served by a Vermont general hospital. Seven percent (±) were also treated at the Brattleboro Retreat, and 5%± were also treated in the Vermont State Hospital. Two percent (±) were also served by the Veterans Administration Hospital in White River Junction.

REGIONAL VARIATION

There were substantial differences among Vermont's regions in the number of people who received inpatient behavioral health care during 1990 through 1994. During that period, 9,871± Vermonters received behavioral health care services on an inpatient basis at least once. This translates to a statewide rate of 1,729± patients per 100,000 population.

Windsor and Windham counties in the southeast had the highest utilization rates (2,448 and 2,250 ± respectively) for the period as a whole and both increased substantially during the five years. Windsor County's utilization rate increased by 16%± and Windham County's rate increased by 42%± between 1990-91 and 1993-94.

The Addison County and Franklin-Grand Isle regions in Vermont's northwestern Champlain Valley had the lowest rates in the state (1,106± and 1,234± per 100,000) for the period as a whole. Addison County experienced a 12%± increase between 1990-91 and 1993-94 while utilization in the Franklin-Grand Isle area increased by 13%±.

Washington County, the home of the Vermont State Hospital, ranked third in utilization for the period as a whole (2,065±), and decreased by 13%± between 1990-91 and 1993-94.

The utilization rate for Chittenden County, the state's most populace region, was substantially below the statewide average (1,629 ±) and decreased by 6%± between 1990-91 and 1993-94.

APPENDIX A

Population Estimates
for Specified Numbers of Birth Dates Within a Given Year
(Based on Probabilistic Population Estimation)

Birth dates	Number of People	Birth dates	Number of People
1	1.003 ± .103	190	269 ± 22
10	10.15 ± .776	200	290 ± 24
20	20.6 ± 1.54	210	313 ± 27
30	31 ± 2.3	220	338 ± 29
40	42 ± 3.2	230	364 ± 32
50	54 ± 4	240	392 ± 35
60	66 ± 5	250	423 ± 38
70	78 ± 6	260	456 ± 42
80	90 ± 7	270	493 ± 46
90	104 ± 8	280	534 ± 51
100	117 ± 9	290	580 ± 57
110	131 ± 10	300	632 ± 64
120	146 ± 11	310	694 ± 73
130	161 ± 13	320	768 ± 85
140	177 ± 14	330	860 ± 101
150	194 ± 15	340	985 ± 125
160	211 ± 17	350	1177 ± 171
170	229 ± 19	360	1603 ± 325
180	249 ± 20	364	2364 ± 912

Probabilistic Population Estimation (Patent Pending) is a statistical procedure which combines knowledge of the distribution of dates of birth in the general population with information of the distribution of dates of birth in data sets to derive an unduplicated count of people who are represented in those data sets. (Banks and Pandiani, 1996)

APPENDIX B

Vermont Correctional Facilities And Vermont State Hospital Caseload Size and Overlap: 1989- 1995

Year(s)	Correctional Facilities			State Hospital			Both			% OF VSH		
	(Estimated)		(Actual)	(Estimated)		(Actual)	(Estimated)		(Actual)	(Estimated)		(Actual)
1989-95	21,804	±377	21, 954	1,815	±16	1,818	606	±100	569	33%	±6%	31%
1995	5,240	±70	5,282	324	±3	323	73	±14	65	23%	±4%	20%
1994	4,947	±68	4,931	335	±3	333	71	±15	62	21%	±5%	19%
1993	4,971	±69	4,959	313	±3	311	57	±15	58	18%	±5%	19%
1992	4,878	±69	4,866	390	±4	389	65	±16	61	17%	±4%	16%
1991	4,819	±70	4,806	408	±4	407	74	±16	78	18%	±4%	19%
1990	4,744	±69	4,716	454	±4	455	91	±19	90	20%	±4%	20%
1989	4,416	±67	4,395	521	±5	524	121	±17	100	23%	±3%	19%

The accuracy of probabilistic estimates of population size and overlap may be determined whenever unique person identifiers are available. This table provides an example of the verification of size of two populations and their overlap. Actual population parameters are based on unique person identifiers in each data base. Actual population overlap parameters are based on manual comparison the names of individuals in the two data sets. (Pandiani and Banks, 1996)

REFERENCES

Banks, S.M., and Pandiani, J.A. (1996). Mathematical Derivation and Properties of The Probabilistic Population Estimator. Albany, NY: Paradigm Associates.

Banks, S.M. and Pandiani, J.A. (1997). An Evaluation of Practice Guidelines for Major Depression in Non-research Settings. NIMH Conference on Improving the Condition of People with Mental Illness: The Role of Services Research.

Banks, S.M., and Pandiani, J.A. (Forthcoming). The Utilization of State and General Hospitals for Inpatient Psychiatric Care. American Journal of Public Health.

Caring for People with Severe Mental Disorders: A National Plan of Research to Improve Services National Advisory Mental Health Council. (1991). Rockville, MD; National Institute of Mental Health; DHHS Publication ADM 91-1762.

Feller, W. (1957). An Introduction to Probability Theory and Its Applications (2nd ed.). New York: John Wiley.

Leginski, W.A., Croze, C., Driggers, J., et. al. (1989). Data Standards for Mental Health Decisions Support Systems. NIMH. Series FN 10. DHHS Pub. No. (ADM)89-1589. Washington, DC: Supt. Of Docs., U.S. Govt. Print. Off.

Regier, A., Goldberg, D., and Taube, A. (1978). The de facto US mental health services system: A public health perspective. Archives of General Psychiatry, 35: 685-93.

Pandiani J.A. and Banks S.M. (1996) “Service Systems Research in the Information Age.”

Sixth Annual National Conference on State Mental Health Agency Services Research and Program Evaluation.

Pandiani J.A. and Banks S.M, Gauvin L (1997) A Global Measure of Access to Mental Health Services for a Managed Care Environment. The Journal of Mental Health Administration.

Pandiani, J.A. and Banks, S.M. and Schacht, L.M. (Forthcoming) Using Incarceration Rates to Measure Program Performance. Journal of Behavioral Health Services and Research.

The White House: Executive Order No. 11973 - President's Commission on Mental Health. Office of the White House Press Secretary, Feb. 17, 1977.

Application #2: A Global Measure of Access to Mental Health Services for a Managed Care Environment