

Toward an Operational Definition of “Personally Identifiable Information”

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

and

John A. Pandiani, Ph.D.
The Bristol Observatory

The Bristol Observatory
521 Hewitt Road, Bristol, VT 05443
802-453-7070

bristob@together.net

Summary

This paper provides an operational definition of “personally identifiable information” that will allow health researchers and data archive administrators to make reasonable decisions about access to data. Statistical procedures derived from probability theory are used to determine the population size necessary to achieve specified levels of identifiability. These statistical procedures are applied to the problem of determining the size of a community in which the combination of date of birth and gender may be considered personally identifying information. Date of birth and gender tend to be personally identifiable in communities of less than 2,600 people. When populations reach 72,000, false identification is more likely than correct identification of individual people. Data bases that include the complete date of birth and the gender of individuals in conjunction with county of residence should rarely be considered personally identifying. The statistical procedures used here can be applied to any set of personal characteristics when the distribution of these characteristics in the population is known. These procedures can provide a rational basis for making decisions about the public release of information.

In recent years, the proliferation of electronic data bases in conjunction with dramatic advances in data processing technology have led to increasing concern about threats to personal privacy and, more specifically, about the confidentiality of medical records.^{1,2,3} In 1996, the United States Congress passed the Health Insurance Portability and Accountability Act (PL 104-191) that requires the Secretary of Health and Human Services to promulgate federal regulations that protect the privacy of health information if congress had not enacted legislation in this area by August of 1999. In December 2000, Standards for Privacy of Individually Identifiable Health Information; Final Rule (45 CFR 164) were promulgated. As of this writing, the regulations are undergoing extended congressional review. The nation is engaged in a major public policy debate. The results of this debate will shape health information policy and have a significant impact on public health research for years to come.

The current debate regarding medical records privacy involves a classic confrontation of public and private goods.⁴ Personal privacy is an important value in contemporary American society. Legislation in support of personal privacy and the confidentiality of medical records would be likely to receive widespread support if it did not have the potential to undermine another set of important national values. These values favor rational administration of public programs and the right of people to be able to make informed choices regarding their health care. These values support the evaluation and

public accountability of health care programs and research in the areas of health care and public health. 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 tension led former Secretary of Health and Human Services Donna Shalala to specifically identify exemptions for public health, health research, and oversight of the health care system in her recommendations to the congress regarding the confidentiality of individually identifiable health information.⁵

This paper contributes to this important public policy discussion by approaching the issue of personally identifiable information from a quantitative empirical perspective. Until now the discussion has been dominated by value statements and imprecisely defined terms. We will propose a rational/empirical approach to the definition of personally identifiable information and will demonstrate the application of this approach to the determination of the point at which widely available demographic data elements (date of birth, gender, and place of residence) can be considered personally identifying. Finally, we will discuss the implications of this approach for public health researchers and for managers of data bases that include data elements that may, in combination, become personally identifying information.

DEFINITIONS OF PERSONALLY IDENTIFIABLE INFORMATION

This paper will focus on two components of the definition of personally identifiable information that permeate discussions of medical records privacy. The first issue we will address is the degree of certainty regarding the identification of individual people that is

intended or required. The second issue regards the type(s) of information that should be considered to be personally identifiable. To date these issues have not been adequately discussed or defined in a useful way.

In her recommendations to the Congress regarding medial records privacy, Secretary Shalala defined protected information to include :

“... any information ... including demographic information ...that identifies the individual, or with respect to which there is a reasonable basis to believe that the information can be used to identify the patient” ... “Reasonableness may depend on a judgment based on what other information is known to be available to a recipient and the amount of effort and time that would be needed to achieve positive identification.” ...“The standard we recommend should not be read to mean that information is identifiable if there is a remote chance that somebody might possibly be able to identify a patient from a general description.”⁵

This definition of personally identifiable information includes reference to degrees of identifiability that range from “positive identification” through “reasonable accuracy” to “a remote chance that somebody might be able to identify a patient.” Unfortunately, It provides little guidance to people responsible for controlling access to potentially personally identifiable information. This lack of a clear operational definition has the potential to leave public health researchers and data archivists facing one of two equally undesirable situations. First, there is the very real possibility that fear of liability would lead to an unnecessarily strict limitation on access to information. Such limitation could compromise the ability of researchers and evaluators to protect the public good. On the

other hand, inadequate sensitivity to the ability of current and emerging data processing technologies to combine and sort through huge data archives in search of improbable combinations of characteristics could lead to inadequate protection of sensitive and potentially damaging information.

The second important issue regards the types of information that are at issue. Such information certainly includes all of our traditional personal identifiers. Legislation protecting the identity of recipients of drug and alcohol treatment, for instance, defines patient identifying information to include:

“...the name, address, social security number, fingerprints, photograph, or similar information by which the identity of a patient can be determined with reasonable accuracy and speed either directly or by reference to other publicly available information.”⁶

The Institute for Medicine offers a substantially broader definition of personally identifiable information when it includes:

“...items of information (e.g. the fact of a physician visit on a given day) that will allow identification of an individual when combined with other facts (e.g. zip code, date of birth, and gender).”²

The Public Health Service endorsed an even stricter definition of personally identifiable information when it promulgated guidelines that prohibited the release of data tapes that contain:

“...any detailed information about the subject that could facilitate identification... (e.g., exact date of the subject’s birth).”⁷

While the growing concern regarding the potential identifiability of combinations of data elements is certainly well placed, there has been little discussion of the degree to which and the conditions under which this concern is justified. In the pages that follow, we will introduce a mathematical procedure for determining the potential for personal identification that is provided by combinations of otherwise innocuous data elements. We will apply this procedure to the evaluation of the degree of personal identifiability that is provided by the combination of five widely available data elements: year of birth, month of birth, day of birth, gender, and place of residence. Finally, we will discuss the relationship between levels of mathematical certainty/uncertainty regarding personal identification and the concept of personally identifiable information.

The analysis is designed to answer two questions. The first question pertains to the minimum size of the group, community, or population that is likely to produce at least one duplication of potentially identifying information. This question relates to likelihood of “false positive” identification at the group or community level: the named characteristics of at least one member of a population have a perfect match in a sensitive data set, but the match is spurious. The second question pertains to the size of a group, community, or population that is likely to produce duplication of potentially identifying information for an individual person. This question relates to likelihood of “false positive” identification at the individual person level: the named characteristics of a particular member of a population have a perfect match in a sensitive data set but the match is spurious.

While the analysis that follows will focus specifically on the potential for personal identification based on date of birth and gender, the procedures may be applied to any set

of personal characteristics for which the distribution in the general population is known.

METHOD

In order to provide examples of computational procedures that are broadly applicable, each of the two questions introduced above will be answered for varying numbers of data elements and levels of certainty. The determination of the size of a group in which two or more individuals are likely to share specified personal characteristics will be demonstrated for two data elements (month and day of birth), for three data elements (year, month, and day of birth), and for four data elements (year, month, and day of birth, and gender).

In addition, the determination of the size of a group in a specified individual is likely to share specified personal characteristics will be demonstrated for two data elements (month and day of birth), and for four data elements (year, month, and day of birth, and gender). Since the probabilities vary with the size of individual date of birth and gender cohorts in this second example, results will be provided for both the average cohort and for the smallest cohort. (Population distributions are modeled on the distribution of the population of the United States in 1990. Population distributions for specified populations will vary.)

The determination of the probability that more than one person in the population will share a birthday (month and day only) is provided by the solution to the classic birthday problem.⁸ When we assume that birthdays are uniformly distributed in the population, the probability of the sharing of a birthday is given by:

$$1 - \prod_{i=1}^n \frac{365-i}{365} \quad [1]$$

where n is the number of people in the population. If n is greater than 365 the probability that at least two people will share a birthday is equal to 1. When a group includes more than 365 people there will always be at least two people who share a birthday.

Similar logic provides the probabilities related to the sharing date of birth (month, day, and year) and gender. The determination of the probability that more than one person in the population will share a date of birth and gender is provided by:

$$1 - \left(\prod_{k=1}^j \left(\prod_{i=1}^{n_k} \frac{365-i}{365} \right) \right) \quad [2]$$

where j is the number of year of birth cohorts (or the number of year of birth and gender cohorts), and n_k is the number of individuals in the k^{th} cohort. If any n_k is greater than 365, the probability that at least two people in the group, community, or population will share a birthday is equal to 1.

The probability that a specified person will share his or her birthday with another person in the same population is provided by:

$$1 - \left(\frac{364}{365} \right)^n \quad [3]$$

where n is the number of people in the total population.⁷ The resulting probability approaches 1 but never reaches 1. No matter how large the population, it is possible that an individual is the only person with the specified birthday.

The probability that a specified person will share his or her date of birth and gender with another person in a population can be derived in two ways. The first provides the average probability that the specified date of birth and gender will be shared. This probability is given by:

$$\frac{\sum_{k=1}^j \left(1 - \left(\frac{364}{365} \right)^{n_k} \right)}{j} \quad [4]$$

where j is the number of year of birth and gender cohorts and n_k is the number of individuals in the k^{th} cohort. The probability resulting from this calculation will approach 1 but never reach 1. The sharing of a date of birth and gender is never guaranteed.

The second probability associated with a person sharing his or her date of birth and gender is the minimum probability. This will produce a different result than the above calculation because there is substantial variation in the size of date of birth and gender cohorts in the general population. In 1998 there were more males living in the United

States who were born in 1990 than males who were born in 1940, and there are fewer males born in 1940 than females born in 1940. The probability that a person in the smallest year of birth and gender cohort in a population will share his or her birthday with another person in the same cohort is given by:

$$\min_{n_k} \left(1 - \left(\frac{364}{365} \right)^{n_k} \right) \quad [5]$$

The minimum probability provided by formula [5] is always smaller than the average probability provided by formula [4].

RESULTS

Table One presents the population sizes associated with specified levels of certainty that no two people will share their birthday (month and day), that no two people will share their date of birth (month, day, and year), and that no two people will share their date of birth and gender. For purposes of this exercise, populations are assumed to have a distribution similar to the age and gender distribution of people under 70 years of age residing in the United States in 1996.⁹

Levels of certainty are grouped into three broad categories. For purpose of this exercise, a probability of less than .05 that two people share specified identifiers is described as indicating a fair degree of certainty that no two people share the identifier(s). In this situation, the potential identifiers could be considered to be personally identifying. A

probability of between .05 and .50 that two people share specified identifiers is described as uncertainty as to whether any two people share the identifier(s). The situation in which the probability that at least two people share specified identifiers is greater than .50 is described as being more likely than not that at least two people share the specified identifiers. In this situation, the identifiers should not be considered to be personally identifying.

In groups of fewer than eight people, it is fairly certain that birthday is unique to every person. One out of twenty groups of 7 randomly selected people will include at least two people with the same birthday. In this situation, birthday could be considered to be personally identifying. As soon as there are more than seven people in a group, however, we would argue that birthday should no longer be considered to be personally identifying. In a group of 24 or more people, it is more likely than not that two or more people will share their birthday. One half of all randomly selected groups of 24 people will include two people who have the same birthday.

When year of birth is added to birthday, the size of the group in which date of birth could be considered to be personally identifying increase to 142, and when gender is considered in conjunction with date of birth, the size of the group in which this combination of identifiers can be considered to be personally identifying increases to 285. As soon as there are 421 or more people in a group or community, it is more likely than not that two or more people will share their date of birth and gender. One half of all randomly selected groups of 421 people will include two people who have the exact same date of birth and gender.

Table Two presents the population sizes associated with levels of certainty that a specified person will share his or her birthday, and that a specified person will share his or her date of birth and gender.

In a group of less than 20 people, it is fairly certain that birthday is unique to every person: it may be considered to be personally identifying. One out of twenty groups of 20 randomly selected people will include at least one person with the same birthday as the specified person. In a group of more than 253 people, it is more likely than not that someone will share this person's birthday. In this situation, birthday cannot be said to be personally identifying.

The average probability and the minimum probability that a specified person will share his or her date of birth and gender with another member of groups of specified size are also presented in Table Two. The average probability of the specified person sharing his or her date of birth and gender exceeds .05 when the population size reaches 2,593. One out of twenty communities of 2,593 people will include a person with the same date of birth and gender as the specified person. This probability exceeds .50 when the size of a community reaches 37,039. One half of all communities with a population of this size will include at least one person who shares the date of birth and gender of a pre-specified person.

The minimum probability that a specified person will share his or her date of birth and gender with another member of the community is the probability associated with the smallest date of birth and gender cohort in the population. In 1997, among residents of the United States under 70 years of age, the smallest date of birth gender cohort was

comprised of men born in 1930. The probability of the specified person in this age and gender category sharing date of birth and gender exceeds .05 when the population size reaches 5,328. In a community of more than 5,327 people, the observation of a precise match of date of birth and gender should not be considered the basis for personal identification of an individual with even fair certainty. The probability of the specified person in the smallest age and gender category sharing date of birth and gender exceeds .50 when the size of a community reaches 71,985. In a community of more than 71,985 people, it is more likely than not that personal identification based on matching date of birth and gender will be erroneous.

DISCUSSION

The mathematical procedures described above provide a methodology for determining the uniqueness of specified personal identifiers in populations of specified sizes. This methodology was applied to the problem of determining the size of communities in which information about a person's date of birth and gender could be considered to be personally identifiable. This methodology, however, can be applied to information about a wide range of personal characteristics that are recorded in electronic data bases. Other examples of potentially identifying information that could be evaluated in this way include ethnicity, the first letter(s) of a person's name, and marital status.

A focus on the distribution of the named characteristic(s) in the general population rather than in specified data bases is a key component of this procedure. This focus is appropriate to the issue of personal privacy because a violation of personal privacy occurs

when a record in an electronic data base is linked to a specified member of the general population. The identification of the electronic record of a person who is known to be HIV positive is not the issue. The public policy issue relates to the ability to determine whether a member of the general population whose HIV status was not known is represented in an electronic HIV registry. The ability to identify such an individual with a fair degree of certainty is a violation of personal privacy. The ability to identify such a person is a function of the size of the larger community of which the person is a member, and the distribution of the specified personal characteristics in that community.

Two concepts related to personal identifiability have been associated with specific statistical criteria. The concept, "a fair degree of certainty" of individual identifiability, was associated with the level of statistical significance ($p < .05$) that is widely accepted in the scientific community.¹⁰ The concept "more than likely not identifiable" was associated with a greater than even chance ($p > .50$) that the named characteristics are shared. The mathematical procedures used above provide a clear set of operations by which these concepts may be defined. We believe that personal privacy is threatened when there is "a fair degree of certainty" of individual identifiability ($p < .05$). We also believe that the criterion of "more than likely not identifiable" ($p > .50$) far exceeds a reasonable definition of personally identifiable information.

The selection of specific quantitative criteria that will be used to determine the appropriateness of releasing specific of data items should be the result of political discussion in the congress and in the scientific community. For purposes of discussion, we will operationally define personally identifiable information by reference to the mid-point

between these two extremes. A probability of duplication of potentially identifying information of $<.275$ will be treated as potentially personally identifying. (The authors do not advocate the adoption of this or any universal criterion. Rather, we favor the use of different criteria for different users of information.)

A probabilistically based operational definition of personally identifiable information can be applied at both the community level (as in Table One) and at the individual person level (as in Table Two). At the community level, application of our working operational definition ($p<.275$) would indicate that date of birth and gender should not be considered to be personally identifying in communities of more than 350 people. In a community of more than 350 people, it is sufficiently likely that at least two people of the same gender share their date of birth to exceed our operational definition. At the personal level, application of our operational definition would indicate that date of birth and gender should not be considered to be personally identifying in communities of more than 20,000 (for the average person) or 40,000 (for the person in the smallest year of birth and gender cohort). In a community of more than 20,000 people, the likelihood that a specified person will share his or her date of birth and gender with another person in the same community is sufficiently large that it should not be considered to provide “positive identification” or even “reasonable accuracy.”

Based on 1997 population estimates, information regarding date of birth, gender, and county of residence should not be considered to be personally identifying of the vast majority of residents of the United States. Only three counties in the United States have a population of less than 350 people. Almost 95% of the residents of the United States live

in counties with populations greater than 20,000, and more than 85% live in counties with more than 40,000 residents. In counties (or other social or geographical groupings) that exceed these thresholds, there is little danger that date of birth and gender would be personally identifying. In places or groups where calculations of the type demonstrated above indicate that personal identifiability is an issue, it is unlikely that public health research would be harmed by the grouping of neighboring or similar entities so as to reduce the likelihood of personal identification.

The selection of an appropriate threshold for the release of potentially identifying information has important implications for both the personal privacy and for the public good that is served by health research. While this paper has focused on issues of personal privacy, it is important to note that the data elements discussed here have significant utility for health research and program evaluation. Emerging statistical methodologies for probabilistic determination of population size^{11,12} and population overlap,¹² even when unique personal identifiers are not available, have demonstrated utility for measuring the treated prevalence of disorders,^{13,14} variation in levels of access to care,¹⁵ and treatment outcomes^{16,17,18} without threatening personal privacy⁴.

The ability of public health and health care researchers, program evaluators, and others to make use of existing data resources to provide information that will allow public policy makers, the professional community, and the general public to make rational decisions is an important public good. The current public debate on the privacy of medical records has the potential to result in a significant limitation of that capability. This potential is in large part the result of the lack of clear operational definitions of terms and careful consideration

of the implications of the adoption of different operational definitions. This paper has attempted to contribute to the current public policy debate by proposing a rational/empirical framework in which the ongoing public policy debate can proceed in a realistic and reasonable manner. We hope that the tools provided in this paper will prove to be useful to the people responsible for making decisions about the release of data sets, and that these decisions will be made within guidelines that provide reasonable protection to personal privacy at the same time as they allow for the advance of knowledge in public health, health care, and health related program evaluation.

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. Washington DC. U.S. Government Printing Office. 1973.
2. Hendricks E, Hayden T, Novik JD. Your Right to Privacy: A Basic Guide to Legal Rights in an Information Society. Carbondale, IL: Southern Illinois University Press;1990.
3. Donaldson MS, Lohr KN. Health Data in the Information Age: Use Disclosure and Privacy. Institute of Medicine. Washington DC: National Academy Press; 1994. *p 51*.
4. 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 (Forthcoming).
5. 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. September 11, 1997. <http://aspe.os.dhhs.gov/admnsimp/PVCREC2.HTM#COVERAGE>
6. United States Code 290dd-2; 42 CFR 2.11(b)
7. National Center for Health Statistics. NCHS Staff Manual on Confidentiality. Hyattsville MD: Department of Health and Human Services; 1984.
8. Feller W. An Introduction to Probability Theory and Its Applications (2nd ed). New York, NY: John Wiley; 1957.

9. U.S. Bureau of the Census. Statistical Abstract of the United States: 1997. Washington, DC; 1997.
10. Fisher RA. The Design of Experiments, 6th edition. Hafner Publishing Co. New York; 1951.
11. Larsen OL. Estimation of the number of people in a register from the number of birth-dates. *Statistics in Medicine*. 1994;13:177-183.
12. Banks SM, Pandiani JA. Probabilistic Population Estimation of the Size and Overlap of Data Sets Based on Date of Birth. *Statistics in Medicine*, In Press.
13. Banks SM, Pandiani JA. The Use of State and General Hospitals for Inpatient Psychiatric Care. *American Journal of Public Health*. 1998;88:448-451
14. Banks SM and Pandiani JA. Enhancing the value of cancer registries and other data sources. *American Journal of Public Health*, 2001: 91 (1):157.
15. Pandiani JA, Banks SM, Gauvin L. A Global Measure of Access to Mental Health Services for a Managed Care Environment. *Journal of Mental Health Administration*. 1997;23:268-277.
16. Banks SM, Pandiani JA, Gauvin L, Reardon E, Schacht LM, and Zovistoski A. Practice Patterns and Hospitalization Rates: A Statewide Program Evaluation. *Administration and Policy in Mental Health*. 1998: 26(1):33-44.
17. Pandiani JA, Banks SM, Schacht LM. Using Incarceration Rates to Measure Mental Health Program Performance. *Journal of Behavioral Health Services and Research*, 1998;25:300-311.
18. Banks SM, Pandiani JA, Schacht LM, and Gauvin, L. Age and Mortality Among

Problem Drinkers. *Addiction*. 2000 ; 95(8): 1249-1254.

TABLE 1 - Population Sizes That Produce Duplication of Potentially Identifying Information

Level of Identifiability	<u>Basis of Potential Identification</u>		
	Birthday (Month/day)	Date of Birth (month/day/year)	Date of Birth (month/day/year) and Gender
Fairly certain that it is ($p < .05$)	2 - 7	2 - 142	2 - 285
Not certain that it is... ($.05 < p < .50$)	8 - 23	143 - 209	286 - 420
More than likely it is not... ($p > .50$)	24 +	210 +	421 +

TABLE 2 - Population Sizes For Levels of Identifiability of a Specified Person

Level of Identifiability	<u>Basis of Potential Identification</u>		
	Birthday (Month/day)	Date of Birth and Gender Person in Average Cohort	Person in Smallest Cohort ¹
Fairly certain that it is ... (p < .05)	2 - 19	2 - 2,593	2 - 5,327
Not certain that it is... (.05 < p < .50)	20 - 252	2,594 - 37,038	5,328 - 71,984
More than likely it is not... (p > .50)	253 +	37,039 +	71,985+

¹ Smallest cohort is the year of birth and gender cohort with the fewest number of members. In 1996, 69 year old males were the smallest cohort.

ACKNOWLEDGEMENTS

The authors acknowledge the contribution of Lucille M. Schacht to the data analysis and the preparation of this paper. We also acknowledge the contribution of John Petrilla to our understanding of the legal and policy issues associated with concept of “personally identifiable information.”