Confidentiality Protection in Publicly Released Central Cancer Registry Data

Colleen C. McLaughlin, MPH, CTR

Abstract: Public release of data from central cancer registries requires a balance between protecting confidentiality and providing information that is of value for research, public health, and education. Ongoing research in confidentiality protection procedures provides a stepping stone for central cancer registries considering the release of public use data files and detailed, small area incidence data. This article provides a brief review of some established disclosure limitation methodologies and their utility in the context of cancer registry data. Methods available to protect individual level data include recoding variables, limiting the amount of geographic detail, limiting the number of data elements included on the data release, and using data use agreements. Methods for protecting tabular data include suppression and redesigning the tables so that disclosure is minimized.

Key Words: *confidentiality, data release policies, public use files*

Introduction

An important mission of central cancer registries is to initiate and promote data utilization. The primary mechanisms for widespread data dissemination available to central registries are published statistical reports, public use data files, and query capable Internet or PC software, such as the National Cancer Institute's SEER*Stat software. [1,2] All of these dissemination methods are subject to potential breaches in confidentiality, and therefore need to be carefully designed in order to provide data to the fullest extent possible while still realizing the mandate to protect patient confidentiality.

The North American Association of Central Cancer Registries (NAACCR) standard for central cancer registries states "Confidentiality is of paramount concern for all cancer registries. There may be no greater threat to the operation and maintenance of a cancer registry than an actual or perceived breach of confidentiality. In fact, an actual or perceived breach of confidentiality in one registry threatens all registries."[3] The congressional law enacting the National Program of Cancer Registries (NPCR) specifies that all states participating in the program have legislation ensuring confidentiality. [4]

Attentiveness to the disclosure risk inherent in the different forms of public release of data has become increasingly important in light of the introduction of regulations concerning the confidentiality of heath data as part of the Health Insurance Portability and Accountability Act of 1996 (HIPAA).[5] The US Department of Health and Human Services Standards for Privacy of Individually Identifiable Health Information (Privacy Rule), which was released in December 2000 and modified in March 2002, specified that confidential data can be used to create de-identified data that are not subject to regulation, provided the data meet the standards of de-identification laid out in the Rule.[6,7] Under this standard, agencies have a choice of alternative methods to use to de-identify the data, but must assure that there is no reasonable basis to believe that the information can be used to identify an individual. One of the alternatives is to follow the safe harbor method of de-identifying data, which specified in exact terms what data elements needed to be removed or modified on a file prior to public release.[6] A second alternative is to have a person with appropriate knowledge apply accepted statistical methods and document that the risk of disclosure is small.[6] In the March 2002 modification, the Department of Health and Human Services asked for public comment on a third alternative, which is to release more data than allowable by the safe harbor, but only in the context of a data use agreement for the purposes of research, public health, or health care operations. [7]

Definition of Disclosure Limitation

Confidentiality protection, also called disclosure limitation, is the process of minimizing the risk of public identification of a person on whom data are collected and minimizing the risk of disclosing information about that person.[8] Duncan et al outlined 3 distinct types of disclosure.

'Confidentiality Protection in Publicly Released Central Cancer Registry Data'

Address correspondence to Colleen C. McLaughlin, MPH, CTR, Research Scientist, New York State Cancer Registry, New York State Department of Health, Corning Tower, Room 536, Empire State Plaza, Albany, NY 12237. Telephone: 518-474-2255; Fax 518-473-6789; e-mail: ccm01@health.state.ny.us.
Submitted: 07/26/01. Revised: 06/11/02. Accepted: 07/05/02.

"Disclosure occurs when a data subject is identified from a released file (identity disclosure), sensitive information about a data subject is revealed through the released file (attribute disclosure), or the released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible (inferential disclosure)."[8] In terms of cancer registry data, this translates to preventing disclosure that an individual has been diagnosed with cancer as well as preventing disclosure of specific attributes about the cancer patient, such as the type of cancer, stage at diagnosis, types of treatment, etc. The clinical information collected by cancer registries is considered to fall under the same confidentiality conventions that customarily apply to the doctor-patient relationship, and extend indefinitely, even after the patient's death.[3]

It is immediately apparent that direct identifying information, such as first and last name, should not be made available for public release. Less apparent, however, is the need to apply confidentiality protection procedures to data that do not contain identifying information, particularly when those data are released in the form of counts of cancer patients rather than individual level data. In fact, both individual level records (microdata) and frequency counts (tabular data) are subject to potential breaches of confidentiality, particularly when there are only a small number of cancer cases involved and the data are very detailed.[9,10]

Disclosure Risk from Microdata

Microdata are data that contain one record per individual or per tumor.[9,10] The SEER public use database is an example of microdata.[11] (The SEER public use dataset is available only to individuals who sign an agreement stating that they will not attempt to violate the confidentiality of the file.) Because microdata generally contain detailed information about a subject, there is a higher degree of risk associated with identification of a subject on the file as compared to tables of cancer counts.[9,12] For example, identification of a subject on central registry's public use file could disclose the type of cancer the patient was diagnosed with (anatomic site, histology, behavior, grade), how far it had spread at diagnosis, and what treatments were given. The question therefore becomes whether or not a patient can be identified on a file from which direct identifiers such as name and address have been removed. Given the readily available record linkage software and large databases containing personal information, there will be inevitable risk in releasing microdata from cancer registries unless more steps are taken than simply removing names and addresses from the file.[13]

An example of a database readily available for linkage is the Social Security Administration's Death Master File (commonly referred to as the Social Security Death Index, or SSDI).[14] This file can be used to identify individuals in conjunction with cancer registry data containing elements of the dates of birth and death and geographic identifiers. Even when these data elements are not on a cancer registry file directly, they can often be inferred based on other data, such as calculating a year of birth based on the year of diagnosis and age at diagnosis. The uniqueness of an individual in a given population is a risk factor for disclosure.[12] The probability of being unique is highest when a given individual lives in an area with a small population or when the individual is a member of specific subgroup that is very small, such as the extreme elderly. An individual can be located by name on SSDI when one person with a specific, unique combination of year of birth and year of death can be found within the specified geographic area. Using gender, year of birth, year of death and ZIP code of residence, staff at the New York State Cancer Registry were able to identify by name approximately 15% of a sample of cancer cases from one county. It should not be assumed, however, that this is only a problem for small populations. For example, using year of birth and month and year of death in a Web-based SSDI search engine in an attempt to identify a very elderly Queens County, NY resident led to only 3 possible matches out of almost 2 million residents. (Over one third of the states in the US have smaller populations than Queens County, NY). Further sources of information, such as obituaries, can be used to narrow the choices even more.

Minimizing Disclosure Risk in Microdata

There are a number of methods by which microdata could be protected from disclosure risk.[9,15,16] Many of these methods have been developed, tested, and used successfully by the US Census Bureau and other federal agencies on their confidential data sets.[17] The Federal Committee on Statistical Methodology has developed a Checklist on Disclosure Potential of Proposed Data Releases, which can be used as a tool to explore the different options for protecting microdata.[18] Some of the methods, however, are less applicable to central cancer registries. Below is an outline of the methods of protecting data that are most applicable to cancer registries, followed by a brief discussion of other less relevant methods.

1. **Recode variables into intervals and top or bottom coding**
   Recoding variables into intervals is a valuable way of protecting individual data records.[9,15] Examples would include grouping age at diagnosis into 5-year intervals, grouping year of birth into decades, grouping date of diagnosis into yearly intervals and grouping length of follow-up into 3-month intervals. Grouping data in this manner requires the creator of the file to consider the trade-off between the utility of the data and the risk of identification.[9,15] Exact age of diagnosis or year of birth, for example, are needed for analysis less frequently than some other variables such as gender and year of diagnosis. Other variables that are candidates for grouping include detailed race and Hispanic origin categories, place of birth, occupation and industry codes, and cause of death codes.

   Top and bottom coding is a slightly different technique in which the only recoded values are those at the extremes of the distribution.[9,15] In the Queens County example above, the individual could be identified

because persons living beyond their 10th decade are unusual. Persons dying in their 70's are much more difficult to identify, because they are not unique in their communities. To protect the individuals in the extremes of the distribution, the age values for the very elderly are truncated, while the age values for the persons in the middle of the distribution are left unaggregated. The HIPAA Privacy Rule safe harbor specifies that ages for patients over 90 years of age be grouped into a "90+" category.[6] Bottom coding, in which the lower end of the distribution is truncated, could also be applied to the ages of younger patients. Other variables that are candidates for top or bottom coding are year of birth and lifetime number of primary tumors.

2. **Limit geographic detail**
When the US Census Bureau prepared the Public Use Microdata Sample (PUMS) from the 1990 census of the population, the only geographic detail included on the file was identification of areas with at least 100,000 persons.[19] This meant that sparsely populated counties in the rural areas of the nation were grouped together, while data for metropolitan areas were released with grouping of census tracts. Counties with at least 200,000 persons were divided as appropriate. The HIPAA Privacy Rule safe harbor prohibits the release of geographic detail other than state and the first 3 digits of ZIP code, provided there are at least 20,000 persons residing in the 3-digit ZIP code area.[6] Limiting geographic detail is probably the most easily implemented and applicable technique for limiting disclosure risk in cancer registry data. It cannot be the only method applied, however, since, as illustrated in the Queens County example above, unique individuals may be identified even among very large populations when the released file contains enough demographic detail. Limiting the geographic detail also limits the usefulness of the public use data for cancer surveillance of small areas. The inclusion of the data-use agreement alternative for HIPAA Privacy Rule de-identification standard arose, in part, from concerns expressed during the public comment over the lack of geographic detail allowed by the safe harbor.[7]

3. **Limit the number of variables on the file**
The risk of disclosure in microdata is related to the number of variables that can be used to narrow the identification of an individual on the file.[9,15] In the Queens County example, removing month of death on the file would reduce the risk of disclosure by increasing the number of possible matches to the SSDI. SEER does not collect the day portion of date of birth or death in order to reduce the risk of disclosure.[20] Some variables collected by central cancer registries, such as marital status, place of birth, and occupation, are of questionable analytic value due to poor quality coding and incomplete ascertainment.[20] Inclusion of these variables on a public use file would only serve to increase the disclosure risk, with little concomitant benefit.

4. **Restricted use files**
Data use agreements, also called data licenses, provide a method for registries to release data in a more controlled manner than freely available public use files.[21] The SEER Public Use data file and the NAACCR analytic file are released under data licenses.[11,22] This licensing extends the legal responsibilities to protect confidentiality to the data users, thereby providing a means to allow potential users access to more finely detailed data than would otherwise be available.[21] The March 2002 proposed amendment to the HIPAA Privacy Rule de-identification standard includes a provision for data use agreements for the purposes of research, public health and health care operations [7]. Some agencies require Institutional Review Board approval prior to issuing data licenses [21]. Individual registries would also need to determine if and how such arrangements are allowed under their authorizing legislation.

   An alternative method to data licensing is the establishment of data research centers, which allow researchers access to registry data in a controlled setting [23]. More detail can be included in analysis because a secure setting would limit the possibility of linkage to external files. Such data would still be stripped of direct identifying information, such as name. Such centers, however, require physical space and personnel to establish and maintain.[23]

5. **Other Methods**
There are several other methods for limiting disclosure risk in individual level data which have less applicability to cancer registry data. Sampling is one of the most common means of protecting microdata, and is used not only by the US Census Bureau (PUMS data and Current Population Survey), but also in many health related data sets, such as the Behavioral Risk Factor Surveillance System and the National Health Interview Survey.[19,24,25,26] Another more subtle method is to introduce noise (statistical perturbation) in the data file, thereby decreasing the chances that the data associated with any individual is accurate.[9,15,16] Both of these methods are more applicable to data used for research, and run counter to the need for accurate and complete cancer counts for surveillance and monitoring the burden of disease.

Disclosure Risk from Tabular Data

Tabular data or summary statistics are inherently less of a disclosure risk than detailed microdata.[9,12,27] This is because individuals are more difficult to identify in tabular data and less information is available for disclosure if an individual is identified.[12] This does not mean, however, that there is no risk. Disclosure risk arises when a user can associate a specific cell of a table to an individual, thereby revealing more

information to the user than was previously known. If a table of cancer counts stratified by age group, site of cancer and place of residence showed that there was only one child with cancer in a community, for example, then a user in that community who knew of a child with cancer would be able to determine the specific type of cancer in more detail.

Disclosure risk from tabular data is largest for release of data for small geographic areas, such as ZIP codes and census tracts.[12] Particularly vulnerable populations include children and young adults or minorities who reside in predominantly nonminority communities. The majority of childhood cancer cases in New York State occur in ZIP codes with few or no other childhood cases. Of the 1575 residential ZIP codes in New York State, 967 ZIP codes had at least one child with cancer between 1995 and 1999. Of these, 274 ZIP codes had exactly one child diagnosed with cancer during the 5-year period.

Minimizing Disclosure Risk in Tabular Data

As with microdata, there are a number of methods that can be used with tabular data to limit disclosure risk. Many of these are the same as they are for microdata, such as decreasing geographic specificity and grouping data. In fact, using microdata that is itself protected from disclosure risk is the most straightforward way of limiting disclosure risk for tabular data.[9] Protected microdata, however, cannot be the only solution, because it is associated with a large degree of data loss, particularly loss of geographic detail. Being able to release data for small geographic areas is one of the main reasons a cancer registry might choose to release tabular data instead of microdata. In central registries, the most common method of protecting tabular data is by means of employing a threshold rule.[28] To apply a threshold rule, you designate as sensitive any cell that falls below a prespecified minimum number of cases, then protect that cell through one of the following methods.[9,29] One difficult problem with application of a threshold rule is determining what the minimum cell size should be for defining a cell as sensitive. As a rule of thumb, it should be somewhere above 3 and below 20.[9] The size of the threshold would be determined by the probability that any one individual or group of individuals can identify every case in a cell. The larger the threshold, the lower the probability, and the greater the protection. The Federal Committee on Statistical Methodology's Checklist on Disclosure Potential of Proposed Data Releases, also covers methods for protecting tabular data.[18]

1. **Suppression**
   With this method, cells that fall below a predefined threshold are suppressed, so that the exact number of cases in the cell is not disclosed. In order to implement suppression, the designer of the table must assure that the value of the sensitive cell cannot be calculated from the other cells.[9,27,29] For example, if only one cell is suppressed, its value can be determined mathematically by subtracting all the other cells from the total. One method to prevent this is by applying complementary suppression of other nonsensitive cells so that the value of the sensitive cell cannot be calculated.[9,27,29]

   Suppression is an attractive method of limiting disclosure risk, since it results in very little loss of data, but it can be problematic to implement properly.[9,27,29] Correct use of complementary suppression is difficult to do on an ad hoc basis, particularly for tables with many dimensions.[9,27,29] For example, if data were released for the entire state, each county and each ZIP code, with suppression only at the ZIP code level, a user may be able to use the county and state totals to derive the ZIP code total, even when the ZIP code level tables themselves are correctly protected.[29] For this reason, it is recommended that the suppression for all tables to be released should be audited using specially designed software. Several such software packages are available, but they are not widely used.[17,29] If this method were used, further data releases from the registry covering the same geographic areas and same time periods would also need to be audited in light of the previously released data.[29]

   It would be increasingly difficult to use suppression correctly in the context of query based software, since successive queries of the data could undermine the suppression of a given cell.[29] A query for the number of children with brain cancer in a particular county might be suppressed, but the number could be calculated based on the results of 2 separate queries, one for the total number of brain cancers and one for the number of adults with brain cancer. In order to prevent this occurrence, the microdata that underlies the query system must, in and of itself, present a minimum disclosure risk.

2. **Table redesign**
   Table redesign is also based on the concept of a threshold rule, except in this case, sensitive cells are protected by combining them with other cells until the cell totals fall above the threshold.[9] For example, ZIP codes with only a few cancer cases can be combined with other neighboring ZIP codes until the totals reach the minimum cell size. Similarly, the number of dimensions of the table can be reduced so that the cell totals are larger. Exclusion of age from data released for ZIP codes and combining ZIP codes with very few cancer cases are 2 methods used by the New York State Cancer Registry to protect sensitive cells. As with cell suppression, this method limits the ability of the central registry to release additional data about the area, since the design of all the tables released need to be considered concurrently.

3. **Use of rates instead of counts**
   Sensitive cells can also be protected by releasing only rates instead of exact counts, because age-adjusted incidence rates cannot be used to back-track to the number of cancer cases. In such instances, it would be helpful to the data users to also include an indication of which rates were based on relatively few cases. The National Center for Health Statistics has a policy of releasing

death rates only when based on at least 20 deaths.[30] This policy is designed to assure a minimum relative standard error, but also serves to assure that no sensitive cells are included in the released tables.

4. **Other methods**

Other methods for protecting tabular data include sampling and perturbing the table values. Perturbing table values is accomplished by changing some or all cells in a table in some way that maintains the overall message of the table but lessens the accuracy. For example, each cell of a table can be rounded to the nearest tenth. As discussed for microdata, these methods are less applicable to cancer registries.

Conclusions

Cancer registries walk a fine line between protecting confidentiality and providing data needed for cancer surveillance. Cancer registries and other disease registries are unique in the field of statistical disclosure limitation in that the methods acceptable with other datasets, such as perturbation and sampling, would conflict with the mission to provide accurate data for public health surveillance. There is no panacea for protecting confidentiality that will allow one data release product to be used for all purposes with no degree of data loss.[10] Instead, each form of data release will require careful design and review, acknowledging that it may be possible to release certain levels of data only in a restricted manner. Creating public use files that protect confidentiality and fulfill the needs of the users is complicated and time-consuming. Registries must to be willing to devote the necessary time and resources to the process. Users of cancer registry data must be educated that there will always be some degree of data loss associated with confidentiality protection.

References

1. Howe HL. Recommendations for public use data files of national cancer data. Sacramento, Calif: North American Association of Central Cancer Registries, November 1997.
2. National Cancer Institute. SEER*Stat 3.0 CD-ROM, April 2000
3. North American Association of Central Cancer Registries. *Standards for Cancer Registries Vol. III: Standards for Completeness, Quality, Analysis and Management of Data.* Springfield, Ill: North American Association of Central Cancer Registries; 2000: 49-51.
4. Cancer Registries Amendment Act. Public Law 102-515 102d Congress, 1992.
5. Health Insurance Portability and Accountability Act of 1996. Public Law 104-191 104th Congress, 1996.
6. Department of Health and Human Services. *Standards for the Privacy of Individually Identifiable Health Information.* Federal Register. 2002; 65:82818-82819.
7. Department of Health and Human Services. *Standards for the Privacy of Individually Identifiable Health Information.* Federal Register. 2002; 67:14798-14800.
8. Duncan GT, Jabine TB, de Wolf VA, eds. *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics.* Washington, DC: National Academy Press; 1993:22-24.
9. Subcommittee on Disclosure Limitation Methodology, Federal Committee on Statistical Methodology. *Statistical Policy Working Paper 22 - Report on Statistical Disclosure Limitation Methodology.* May 1994.
10. Doyle P, Lane JI, Theeuwes JJ, Zayatz LM. *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies.* New York, NY: Elsevier; 2001:1-6.
11. Surveillance, Epidemiology, and End Results (SEER) Program Public-Use CD-ROM (1973-1999), National Cancer Institute, DCCPS, Cancer Surveillance Research Program, Cancer Statistics Branch; 2002.
12. Elliot M. Disclosure Risk Assessment. In: Doyle P, Lane JI, Theeuwes JJ, Zayatz LM. *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies.* New York, NY: Elsevier; 2001:80-87.
13. Sweeney L. Information Explosion. In: Doyle P, Lane JI, Theeuwes JJ, Zayatz LM. *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies.* New York, NY: Elsevier; 2001:66-73.
14. Social Security Administration Death Master File CD-ROM, US Department of Commerce, National Technical Information Service (NTIS).
15. Domingo-Ferrer J, Torra V. Disclosure Control Methods and Information Loss for Microdata. In: Doyle P, Lane JI, Theeuwes JJ, Zayatz LM. *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies.* New York, NY: Elsevier; 2001:91-108.
16. Domingo-Ferrer J and Torra V. A Quantitative Comparison of Disclosure Control Methods for Microdata. In: Doyle P, Lane JI, Theeuwes JJ, Zayatz LV. *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies.* New York, NY: Elsevier; 2001:111-132.
17. Felso F, Theeuwes JJ, Wagner GG. Disclosure Limitation Methods in Use: Results of a Survey. Domingo-Ferrer J, Torra V. Disclosure Control Methods and Information Loss for Microdata. In: Doyle P, Lane JI, Theeuwes JJ, Zayatz LM. *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies.* New York, NY: Elsevier; 2001:26-36.
18. Interagency Confidentiality and Data Access Group, Federal Committee on Statistical Methodology. Checklist on Disclosure Potential of Proposed Data Releases. Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget. July 1999. http://www.fcsm.gov/docs/checklist_799.doc.
19. Census of Population and Housing, 1990: Public Use Microdata Samples US Technical Documentation / prepared by the Bureau of the Census. – Washington, DC: Bureau of the Census; 1992.
20. Hulstrom D, ed. Standards for Cancer Registries Vol II: *Data Standards and Data Dictornary Version 9.1*, 6th Ed. Springfield, Ill: North American Association of Central Cancer Registries; 2001:11-26.
21. Seastrom M. Licensing. In: Doyle P, Lane JI, Theeuwes JJ, Zayatz LM. *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies.* New York, NY: Elsevier; 2001:279-289.
22. Howe HL. Report of the NAACCR CINA Deluxe Beta Test. Springfield, Ill: North American Association of Central Cancer Registries; 2000.
23. Dunne T. Issues in the Establishment and Management of Secure Research Sites. In: Doyle P, Lane JI, Theeuwes JJ, Zayatz LM. *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies.* New York, NY: Elsevier; 2001:297-302.
24. US Census Bureau and Bureau of Labor Statistics. *Current Population Survey Design and Methodology.* Technical Paper 63. Issued March 2000, TP63.
25. Centers for Disease Control and Prevention. *Behavioral Risk Factor Surveillance System User's Guide.* Atlanta, Ga: US Department of Health and Human Services, Centers for Disease Control and Prevention; 1998.
26. Botman SL, Moore TF, Moriarity CL, Parsons VL. *Design and estimation for the National Health Interview Survey, 1995–2004.* National Center for Health Statistics. *Vital Health Stat.* 2(130). 2000.
27. Duncan Gt, Fienberg SE, Krishnan R, Padman R, Roehrig SF. Disclosure Limitation Methods and Information Loss for Tabular Data. In: Doyle P, Lane JI, Theeuwes JJ, Zayatz LM. *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies.* New York, NY: Elsevier; 2001:142-151.
28. Deapen D, Chen V, et al. *Data Use and Confidentiality Task Force Report.* Springfield, Ill: North American Association of Central Cancer Registries; 2000.
29. Giessing S. Nonperturbative Disclosure Control Methods for Tabular Data. In: Doyle P, Lane JI, Theeuwes JJ, Zayatz LM. *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies.* New York, NY: Elsevier; 2001:198-212.
30. Minino AM, Smith BL. Deaths: Preliminary data for 2000. National vital statistics reports; vol 49 no 12. Hyattsville, Md: National Center for Health Statistics; 2001:37.