Six Degrees of Separation No More: Using Data Linkages to Improve the Quality of Cancer Registry and Study Data

David Harris

RTI Health Solutions, Research Triangle Park, NC, United States

ABSTRACT

Background: A data linkage is a process commonly used to determine if persons in one database also reside in a second database. There are two general types of linkages: deterministic (rules-based) and probabilistic (statistical). Specialized linkage software programs such as AutoMatch and Link Plus are used to perform the linkages. For those cancer registries unable to afford a data linkage program, the Centers for Disease Control and Prevention (CDC) offers Link Plus for free on its Web site.

Objective: To explore the variety of reasons to link a database with cancer registry files. The presentation will also illustrate the value of data linkages in increasing the quality of cancer registry and study data.

Methods: The stated objectives will be achieved by offering real-world examples of the value of linking population-based cancer registry databases with other sources. Potential examples include linking a study cohort to a cancer registry database to determine cancer diagnoses and burden among the cohort; using the linkage process to update the vital status and date of last contact for patients in the cancer registry database; evaluating the effectiveness of cancer control and prevention programs; and using linkages for drug safety surveillance studies.

Results: The presentation will include results from data linkages between cancer registry files and other files, including linkages with public use files to update vital status, with cancer control data to evaluate program effectiveness, and with other databases to determine cancer burden in specific populations.

Conclusions: If used properly, data linkages can be effective in increasing the quality of a cancer registry's data, allow researchers to have a better understanding of cancer burden in their cohorts, help to determine if cancer screening efforts are effective, and allow cancer registry data to be used in novel ways.

BACKGROUND

- One goal of population-based cancer registries in the United States (US) is to collect complete, timely, and high-quality data for cancer research and control efforts.
- Data linkages are used to determine if persons in one database also reside in a second database.
- There are two general types of data linkages—deterministic and probabilistic.
  - Table 1 describes three popular data linkage software packages.

Table 1. Description of Data Linkage Software Packages

Linkage Software	Type of Linkage	Advantages	Disadvantages
LinkPlus[1]	Probabilistic	• Free to download • Easy to use • Color-coded manual review page • Accepts NAACCR file format	• Single-pass linkage program • Less control by user for adjusting algorithms • Not able to handle very large files
AutoMatch	Probabilistic or deterministic	• Multi-pass linkage program • Greater control by user to modify algorithms • Ability to perform both probabilistic and deterministic linkages • Name and address standardization capabilities	• Expensive • Older version is DOS-based and no longer supported • Not user-friendly
SAS	Deterministic	• Familiar to many programmers • Able to handle very large files	• Can require extensive programming and macros • Makes review of possible matches difficult

DOS = disk operating system; NAACCR = North American Association of Central Cancer Registries.

OBJECTIVE

- To explore the variety of reasons to link a database with cancer registry files and to illustrate the value of data linkages in increasing the quality of cancer registry and study data.

EXAMPLES FROM STUDIES WHERE DATA LINKAGE WAS USED TO IMPROVE CANCER REGISTRY AND STUDY DATA

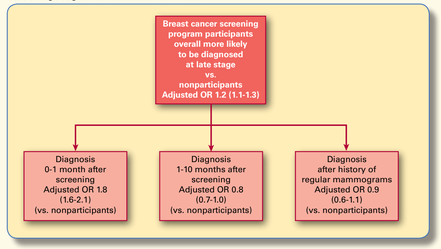
Using Data Linkages to Evaluate the Effectiveness of Cancer Control and Prevention Programs

- Cancer control and prevention programs can be evaluated by linking data from the program to one or more state cancer registries.

Example 1:
Breast Cancer Screening Program Impact on Stage of Diagnosis

- Data from a breast cancer screening program for underserved women was linked to data from the California Cancer Registry to analyze stage at diagnosis for breast cancers versus stage among nonparticipants.[2]
- Overall, women in the program were diagnosed at a later stage than nonparticipants (Figure 1).
- However, by using results from the linkage (data from both databases), when time between mammogram and diagnosis was taken into consideration, the following results were found:
  - Women diagnosed immediately after receiving a mammogram were diagnosed at a statistically significant later stage than nonparticipants, driving the overall results.
  - Women receiving mammograms on a regular basis were diagnosed with late stage tumors at the same rate as nonparticipants, a success for the program.

Figure 1. Stage at Diagnosis Among Women Receiving Mammograms Through a Screening Program for Underserved Women



OR = odds ratio.

Using Data Linkages to Improve the Quality of Cancer Registry Data

- State cancer registries are graded or certified on how complete, accurate, and current their data are.
  - For example, the Surveillance, Epidemiology, and End Results (SEER) program sets criteria for follow-up rates and categorizes the rates as unacceptable, acceptable, or met goal for specific age categories.
  - Table 2 provides an example of SEER's goals for cases diagnosed in 2000-2001 and followed into 2003 from a study performed in October 2004 (Example 2).
- Linkages with other sources can improve date of last contact and provide updated patient contact information.
- Accurate vital status and date of last contact are important for survival analysis and mortality rates.
- Linkages against own database are useful for identifying and removing duplicate cases.

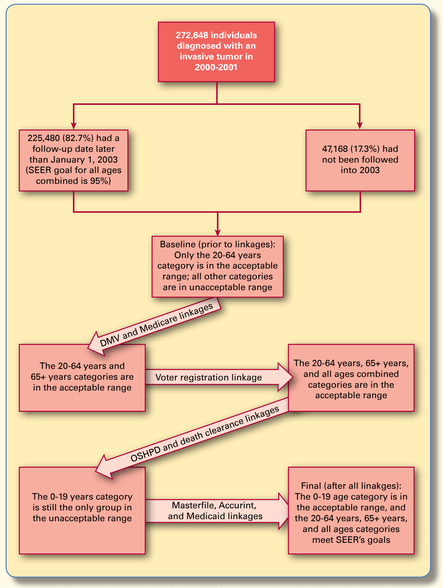
Table 2. SEER Goals for Percentage of Invasive Tumors With Updated Follow-up Data for Cases Diagnosed 2000-2001 and Followed into 2003

Age	Goal	Acceptable	Unacceptable
0-19 years	> 90%	80%-90%	< 80%
20-64 years	> 90%	80%-90%	< 80%
≥ 65 years	> 95%	90%-95%	< 90%
All ages	> 95%	90%-95%	< 90%

Example 2:
Improving Date of Last Contact Using Data Linkages

- A study performed in 2004 analyzed the effect of data linkages in improving date of last contact (follow-up rate) for the Cancer Registry of Greater California (CRGC) SEER region in order to meet SEER's goals (Table 2).[3]
- Data linkages between CRGC and eight separate databases were performed in September and October 2004.
  - Records were updated if the follow-up date from the linkage was later than the date already in the CRGC database, or if the vital status was changed from alive to dead.
- Figure 2 shows the linkages in chronological order and the status of each age group after each linkage.

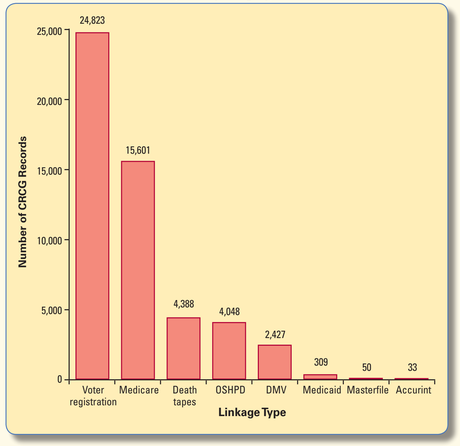
Figure 2. Effect of Linkages in Meeting SEER Goals for Follow-up Rates, by Age Category



DMV = Department of Motor Vehicles; OSHPD = Office of Statewide Health Planning and Development.

- Overall, 51,679 CRGC records among 42,516 individuals were updated. Figure 3 illustrates the number of updates by linkage type .

Figure 3. Number of CRGC Records Updated by Linkage Type, 2000-2001 Cases Followed Into 2003



Using Data Linkages for Tracking Outcomes in Cohorts or Patient Registries

- Investigators can link any type of study cohort or patient registry to one or more state cancer registries to determine the incidence or burden of cancer among the cohort or registry, including changes over time.
- The Food and Drug Administration (FDA) Amendments Act of 2007 gave the FDA additional authority to require pharmaceutical companies to perform postapproval drug safety activities.
  - Linking to state cancer registries is an effective way of tracking the incidence of cancer in a cohort of medication users over an extended period of time.

Example 3:
Linking Two Types of Registries to Investigate Topics of Interest

- A linkage performed between cases from the San Francisco AIDS registry and the California Cancer Registry showed that the use of highly active antiretroviral therapy (HAART) decreased the risk of Kaposi's sarcoma, systemic non-Hodgkin lymphoma (NHL), and central nervous system NHL when comparing pre-HAART and HAART-era AIDS patients.[4]

Example 4:
Linking a Cohort of Medication Users to State Cancer Registries

- A current FDA-mandated study is using data linkage between multiple state cancer registries and a cohort of medication users over an extended period of time to ensure that the incidence of a specific type of cancer is not significantly higher among the cohort.[5]
- Early results from this long-term study demonstrate the feasibility of using a standardized data linkage algorithm to help ensure consistent linkage results among the numerous state cancer registries participating in the study.
- Results can also be used to identify cases where follow-up is required for safety reasons.

CONCLUSION

If used properly, data linkages can be effective in increasing the quality of a cancer registry's data, allow researchers to have a better understanding of cancer burden in their cohorts, help to determine if cancer screening efforts are effective, and allow cancer registry data to be used in novel ways.

REFERENCES

1. Centers for Disease Control and Prevention (CDC). Registry Plus Link Plus, version 2.0. Available at: http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm.
2. Harris DH, Bates JH, Cress R, Tabnak F, Wright WE. Stage of breast cancer diagnosis among medically underserved women in California receiving mammography through a state screening program. Cancer Causes Control. 2004 Sep;15(7):721-9.
3. Harris DH, Sheridan A, Halvorson G. The use of data linkages to increase follow-up rates in California. Poster presented at the North American Association of Central Cancer Registries Annual Conference; June 2005. Cambridge, MA.
4. Pipkin S, Scheer S, Okeigwe I, Schwarcz S, Harris DH, Hessol N. The effect of HAART and calendar period on Kaposi's sarcoma and non-Hodgkin lymphoma: results from a match between an AIDS and cancer registry. AIDS. 2011 Feb 20;25(4):463-71.
5. Harris DH, Midkiff K, Gilsenan A, Andrews E. Public health surveillance collaboration: establishing a linkage algorithm with cancer registries for the Forteo patient registry. Poster presented at the North American Association of Central Cancer Registries Annual Conference; June 2010. Quebec City, Quebec, Canada.

CONTACT INFORMATION

David Harris, MPH
Research Epidemiologist

RTI Health Solutions
200 Park Offices Drive
Research Triangle Park, NC 27709

Phone: +1.919.541.7493
Fax: +1.919.541.7222
E-mail: dharris@rti.org

Presented at: 2011 NAACCR Annual Conference
June 18-24, 2011
Louisville, KY, United States