

# National Program of Cancer Registries Data Linkages with Public and Private Data Sources

SM Van Heest, JD Rogers, Centers for Disease Control and Prevention, Division of Cancer Prevention and Control, Cancer Surveillance Branch

## Background—

Central cancer registries in CDC's National Program of Cancer Registries (NPCR) commonly conduct data linkages to improve data accuracy, quality, and completeness. Data linkages identify missing cancer cases, cross-validate data, and help with studies about comorbidities and survival. They also provide opportunities for registries to enhance relationships with other cancer control programs while potentially decreasing the cost of data collection.

The NPCR Interim Progress Report Guidance (year 2) outlines the following schedule for data linkages—

- Central cancer registries link with state death files at least every year, and incorporate the results on vital status and cause of death into the registry database.
- Central cancer registries link with the National Death Index at least every other year, and incorporate the results on vital status and cause of death into the registry database.
- Central cancer registries link with the state breast and cervical cancer early detection program and the state colorectal cancer screening program at least every year to identify potentially missed cases, reconcile differences between the two systems, and update appropriate data fields to capture post-linkage information.
- Central cancer registries link with the Indian Health Service (IHS) at least every five years. Central cancer registries with IHS Contract Health Service Delivery Area counties link their records with patient registration records from IHS every year

The central cancer registry uses linkages to address gaps identified in data quality and completeness reviews or to improve the usefulness of the data. Other potential sources of information include—

- Statewide electronic health and discharge files.
- Claims data for casefinding and completeness of required data items.
- Census data (or similar) for socio-demographic variables.
- Birth records for demographic information.
- Department of Motor Vehicle records for demographic information.
- Voter registration files for demographic information.
- The Social Security Administration's Death Master File.

- Medicare and Medicaid claim files.
- Other data sources, including pathology reports, the American Hospital Directory, the Medicare Physician Identification and Eligibility Registry, the Area Resource File, geocoding, and the American Hospital Directory.

Data linkages populate or enhance data such as patient identification number, date of birth, sex, race, address including ZIP Code, Social Security number, type of health plan and coverage, primary site and number of sites, date and method of diagnosis, cancer stage, date and cause of death, Medicare and Medicaid eligibility, and cause of death and comorbidities/complications.

## Methods—

Registry success stories provide examples of data linkages that take advantage of publicly available datasets.—

- The Arizona Cancer Registry linked with Hopi and Navajo datasets. <http://www.cdc.gov/cancer/npcr/success/west/az/az1.htm>
- The Arkansas Central Cancer Registry linked cases from the Enhanced HIV/AIDS Reporting System (eHARS). <http://www.cdc.gov/cancer/npcr/success/south/ar/ar1.htm>
- The New York State Cancer Registry linked to the state hospital discharge database. <http://www.cdc.gov/cancer/npcr/success/northeast/ny/ny1.htm>
- The Oregon State Cancer Registry linked with the Social Security Death Index and the National Death Index, providing missing Social Security numbers and correcting errors in the dates of birth and death. <http://www.cdc.gov/cancer/npcr/success/west/or/or2.htm>
- The Michigan Cancer Surveillance Program linked to two tribal registries. <http://www.cdc.gov/cancer/npcr/success/northeast/>

## Discussion—

Investigators may link databases to examine changes in patterns of care, the use of cancer tests and procedures, and the cost of cancer treatment, as well as to enhance casefinding and follow-up.

In the future, biomarker data may be incorporated into the bilateral linked datasets and the registry datasets by using the Biorepository Alliance (in Georgia) and the SEER Residual Tissue Repository Program (captures discarded tissue, linked to SEER data). These linkages could lead to the creation of linked, de-identified analytical datasets tailored for specific analyses.

Linking cancer registry data with state vital records, the Social Security Death Index, and the National Death Index will provide follow-up information needed for survivorship studies. NPCR is addressing data collection and methodologic issues related to measuring survival time, including short- and long-term survival projections for recent patient cohorts, to estimate prevalence.

Cancer registry data linkage has limitations. Nationally, not all data is included in files, and coverage varies by geographic region. State data limitations include differing data layouts and coding schemes, data sets that have layouts that sometimes vary by year, and missing data.

## Conclusion—

Investigators in central cancer registries are encouraged to link databases to examine changes in patterns, especially those that enhance casefinding and follow-up. Linked data from multiple databases are a valuable resource in cancer surveillance and in the evaluation of factors influencing cancer trends and quality of care. These linkages expand the richness of cancer registry data by facilitating examination of a variety of health-related issues. Data linkages with private and publicly available databases may decrease the cost of data collection and improve the timeliness of reporting.

