

Procedures for Conducting Data Linkages with the California Cancer Registry

Chronic Disease and Surveillance Branch
California Department of Public Health
California Cancer Registry
1825 Bell Street, Suite 102
Sacramento, CA 95825
(916) 779-0300 Phone
(916) 779-0264 FAX

Table of Contents

Table of Contents	2
Check List for Researchers	3
Overview of the California Cancer Registry	4
Goals	4
Background	4
How are CCR data used?	5
How much research is conducted using CCR data?	5
What about patient confidentiality?	5
Data Quality and Completeness	5
CCR Data Linkage Experience	6
Procedures for Linkage	7
File Format for Data Sent to the CCR	9
Costs for CCR to Perform Linkages:	10
Laws Governing CCR.....	11

Check List for Researchers

Proposing to Conduct Data Linkages with the California Cancer Registry

- Contact Ann Brunson, or designee at (916) 779-2673, abrunson@ccr.ca.gov to initiate process and discuss procedures.
- Review and complete the application requirements for *Policies and Procedures for Access to and Disclosure of Confidential Data from the California Cancer Registry* (http://www.ccrca.org/PDF/CCRDataAccessDisclo_v04.4.pdf).
- Submit to the following to Chronic Disease and Surveillance Branch (CDSRB):
 - Study protocol (without Appendices)
 - List of requested data items from the California Cancer Registry (CCR) including brief justification
 - Current Institutional Review Board (IRB) Approval (Institution's IRB and State of California's Committee for the Protection of Human Subjects (CPHS) IRB (if needed))
 - Appendix 3: Agreement for Disclosure of CCR Data* signed by the principal investigator and responsible institution official (included in *Policies and Procedures for Access to and Disclosure of Confidential Data from the CCR*).
 - Appendix 5: Procedures to Maintain the Confidentiality of CCR Data* signed by the principal investigator (included in *Policies and Procedures for Access to and Disclosure of Confidential Data from the CCR*)
 - Notice of Grant Award if applicable
- Once CDSRB approves the project and signs the *Agreement for Disclosure of CCR Data*, Ann Brunson (916-779-2673, abrunson@ccr.ca.gov<mailto:mallen@ccr.ca.gov>) will send the approval and make arrangements for the transfer of your data to the CCR and subsequent linkage.

Overview of the California Cancer Registry

The mission of the California Cancer Registry is to eliminate the effects of cancer by developing tools and using them for research into the prevention, causes, detection, and cures of cancer.

Goals

- ◆ Conduct research into the causes and cures of cancer.
- ◆ Identify unequal burdens in cancer risk factors and incidence, and gaps in treatment and outcomes among the State's diverse population subgroups.
- ◆ Build and maintain an easily accessible data system that will support policy analysis and program decisions.
- ◆ Provide up-to-date information to the public to address concerns about cancer incidence.

Background

The California Cancer Registry (CCR) was created by the legislature in response to public demands that more be done to find the prevention, causes, and cures for cancer. Since 1988 every cancer, except basal and squamous cell carcinoma, diagnosed in California is required by law to be reported to the CCR, which is operated by the Cancer Surveillance and Research Branch of the California Department of Public Health (CDPH). The Public Health Institute (PHI) has a contract with the CDPH to assist in the operation of the CCR, including facilitating externally funded projects utilizing CCR data.

The CCR is one of the largest registries in the world and is internationally recognized for its high quality cancer data. Because of California's racial diversity, the CCR provides information on cancer in racial/ethnic groups that cannot be obtained from any other source.

Approximately 130,000 new cancer cases and about 52,000 cancer deaths are reported to the CCR each year. The CCR is a three-tiered system:

- Medical treatment facilities collect and report cancer data from their medical records. Physicians report information of cancer patients who are not referred to a medical treatment facility.
- A network of eight regional registries receives these data and checks for accuracy, performs analyses, and conducts studies specific to the local area.
- The Cancer Surveillance and Research Branch in Sacramento collates these data, performs additional quality control and analyzes the data on a statewide basis.

How are CCR data used?

CCR data are used to:

- Study cancer causes and risk factors,
- Conduct epidemiological and clinical research
- Evaluate patterns of treatment and stage of diagnosis
- Disseminate information for planning and early detection programs
- Respond to state and local questions and concerns about cancer
- Provide information to citizens, legislators, and health professionals.

How much research is conducted using the CCR?

- Since 1988, approximately 450 research projects using CCR data have been initiated.
- Researchers have published over 2,084 articles in scientific journals and books.

What about patient confidentiality?

All data collected by the CCR reporting system are subject to the confidentiality provision of the Health and Safety Code and the Government Code. Confidential information can only be released for research purposes to qualified investigators whose study protocols have been approved by a federally-designated committee for the protection of human subjects, and who comply with additional conditions specified by the CCR.

Data Quality and Completeness

It requires the CCR about eighteen months after the close of a calendar year to collect, quality control, consolidate and produce analysis files for greater than 95% of the cases for a given year. For example, in August 2004, it was estimated that case reporting for 2002 was 97.5% complete for invasive cancers other than prostate gland. Given the volatility of prostate cancer incidence since 1989 due to PSA screening increases, it is difficult to assess completeness of case reporting for this cancer. The entire state of California is now a part of the national SEER program and meets their stringent requirements for completeness. The CCR also has historical data from the San Francisco Bay area dating back to 1973. The registry now contains about 2.2 million records. Information is collected about the patient, the tumor, the treating physician, the treating facility, and the first course of treatment - over 300 data elements for each tumor. There is a unique record for each tumor and about 10% of the tumors are additional primary site tumors on individuals already in the registry. We have out-of-state case sharing agreements with 22 states, including all of the bordering states.

CCR Data Linkage Experience

The CCR has routinely performed probabilistic data linkages with the cancer registry data since 1992. Starting in 1995, the frequency of linkages significantly increased and the registry began using the software program AUTOMATCH to perform the probabilistic data record linkages. AUTOMATCH was bought by another company and the software changed names to Integrity. This company was recently bought by Ascential Software, Inc. Linkages have been done on a variety of other state and federal databases as well as various researcher databases. The results of the linkages have enhanced the CCR database, updated patient follow-up, identified major comorbid illnesses and provided valuable information for research and policy analyses. The CCR database is annually linked with the Health Care Financing Administration/Medicare (HCFA) enrollment file of 6.0 million records. Linkages with the 3.7 million records each year from hospital discharge files of the Office of Statewide Health Planning and Development (OSHPD) have been done frequently. The CCR has linked with the California Medicaid enrollment file and the driver license files of the California Department of Motor Vehicles. Annually the CCR database is linked with CDPH Center for Health Statistics Death Master Files. For follow-up, the results of each of these linkages have been able to update about 8%-10% of CCR records. The CCR has also linked the CCR database with many diverse cohorts, some of which are:

1. "Using Cancer Registries to Assess Quality of Cancer Care" in collaboration with Harvard Medical School.
2. "Cancer diagnoses and Health Plan Switching Among University of California Employees" in collaboration with UCI.
3. "Fertility Drugs and Ovarian Cancer", UCSF.
4. "NIH/AARP Diet and Health Study", Westat.
5. "Costs of Treating Breast Cancer in the Medi-Cal(Medicaid) Population in California", DHS
6. "Prostate Cancer in High, Medium and Low Risk Populations", Stanford
7. "Multiethnic/Minority Study of Diet and Cancer", USC
8. "Follow-up of CPS-II Participants through Linkage with State Cancer Registries", American Cancer Society
9. "Cancer Incidence in the United Farm Workers, United Farm Workers Health Collaborative
10. California Teachers' Study, USC
11. The Child Health and Development Study, CDPH
12. Human Population Laboratory, CDPH
13. The Breast and Cervical Cancer Program, CDPH
14. The Breast Cancer Early Detection Program, CDPH

Procedures for Linkage

Data record linkage is the process where we determine if a record in one file matches to one or several records in another file. Currently, the CCR uses software named “Integrity” that enables us to perform probabilistic linkages. Both data files must have common variables such as name, social security number, date of birth, residence, race, and place of birth. Some of these variables should not be different between the files other than for some mistakes in the coding of information or missing information on one of the files. Social security number, date of birth, race, and place of birth all fall into this category. On the other hand, people commonly change their name and their residence, and these variables may be different on the two files. Anyone wanting to perform data linkage must have an understanding of the variables on each file and an exact understanding of when the data were collected that is likely to change or be modified over time.

Once there is a basic understanding of the information available on the two data sets, both data sets are prepared for the linkage. The file format for CCR linkages is found on page 9. The codes for all categorical data must match the CCR database. For example, sex must be coded as “1” and “2” and not “M” and “F” to follow the CCR conventions. Data values that are missing must be distinguished from zeroes of numeric fields. Similar information on the two files may need a flag variable created for the match rather than comparing complex character values. Individual’s names should be standardized on both files. CCR staff first separate names into individual surname and given name fields. Then nicknames are changed into standardized names and names that are the same but are spelled differently are changed into a standard spelling of the name. Usually, the original name is preserved and a new variable with the standardization of the original name is created. Lastly, a phonetic spelling or coding of the name is created. Address information also needs to be standardized and separated into individual parts for the linkage. If necessary, a phonetic spelling of the street name and city may be done. After all of the preparation of the data, dictionaries of the data file are created for the match.

Next the CCR staff discusses matching specifications with the researcher requesting the linkage and CCR prepares the specification document. It is not feasible to compare one record in the first file to every record in the other file. Consequently, we do a procedure called “blocking” that provides a means of looking at only those pairs of records with a high probability of matching and limiting the number of pairs examined. Blocking requires that a specific variable or combination of variables be an exact match and then the remaining variables are compared to determine if the pair of records is a likely or unlikely match. To allow for minor differences between variables that were required to be exact, a second linkage is done requiring different variables to be exact. The matching specifications delineate how many linkages are done, which variables must be exact during each linkage, and which variables are matched.

After the matching specifications are agreed upon, both files are indexed by the blocking fields rather than sorted, and a frequency analysis is done on all variables.

The frequency analysis is used in constructing tables that estimate the likelihood of a chance agreement in a field for a pair of records. This likelihood for each variable is used in calculating an overall agreement weight for each pair of records. The higher the weight, the more likely the pair of records is a match.

The matching algorithm is now run for the first blocking variables. A histogram of the all of the weights is used to help determine cutoffs for the weights on pairs of records. CCR will discuss these histograms to jointly determine the cutoff thresholds. An upper cutoff is established above which all pairs are considered matches. A lower cutoff is also determined below which all pairs are considered non-matches. Those pairs of records with a weight in between the lower cutoff and the upper cutoff are then reviewed manually to determine if they are a match or not. After this is complete, the matching algorithm is repeated for the next blocking variables. More possible matches are identified and reviewed. This process continues until all blocking runs are complete.

Once the linkage is complete, data are extracted from the client file and the CCR file for all of the matched records. Additional checks are usually made to determine if one record from the client file matched to two or more records in the CCR file. This process will sometimes identify duplicate records in one or both of the original data files. Once the duplicate records are removed, the necessary cancer registry variables are placed in a file for analysis. The CCR generates a data dictionary of all coding variables and copies of coding schemes, eg. County codes, hospital numbers, etc. Finally the CCR produces an encrypted data file that is emailed through the CCR Certified Email system or if the file is too large, put on a CD and express mails all materials to the researcher. The password is provided by e-mail or phone.

File Format for Data Sent to the CCR

Data files should be ASCII fixed field length files on CD or arranged to be submitted through the CCR Certified Email system with the following variables:

Variable	Columns
First Name	1-15
Middle Name	16-30
Last Name	31-50
Social Security Number (no hyphens)	51-59
Date of Birth (MMDDCCYY)	60-67
Sex (1=Male, 2=Female)	68
Street Address	69-93
City	94-108
Zip code	109-113

Additional personal identifying information (e.g. maiden name, hospital record number) if available should also be included.

Data files should be checked to ensure that no duplicate records are included.

If personal identifiers WILL be returned from the CCR, an additional sequential identification variable must be included to facilitate merging the data set of linked records with the original data set. This variable MUST be unique for each record in the original data set and in the data set provided for linkage.

If personal identifiers WILL NOT be returned from the CCR, any grouping variables (e.g. occupation code, individual status code, type of position, etc.) that will be needed for analysis after the linkage must be included. For these variables, include the name of the variable and columns where the variable may be found.

Restrictions on variables:

1. Missing values for all names, street address and city are spaces and for all numeric variables are 9's.
2. All letters should be in capitals.
3. All names should have no embedded blanks, commas, apostrophes, or periods. Hyphens are allowed.
4. No "Jr.", "II", "III", etc. in names.

Costs for CCR to Perform Linkages:

Basic linkage costs:

Less than 10,000 records:	\$ 3,000
10,000 - 49,999 records:	\$ 4,500
50,000 - 99,999 records:	\$ 6,000
100,000 - 249,999 records:	\$ 7,000
250,000 – 499,999 records:	\$ 8,500
500,000 – 999,999 records:	\$10,000
More than 1,000,000 records:	\$15,000

Multi-year linkage (i.e. OSHPD linkage) costs:

Less than 10,000 records:

\$3,000 + 20% (\$600) for 1-5 years of data
\$3,000 + 40% (\$1,200) for 6-10 years of data
\$3,000 + 60% (\$1,800) for 11+ years of data

10,000 – 49,999 records:

\$4,500 + 20% (\$900) for 1-5 years of data
\$4,500 + 40% (\$1,800) for 6-10 years of data
\$4,500 + 60% (\$2,700) for 11+ years of data

50,000 – 99,999 records:

\$6,000 + 20% (\$1,200) for 1-5 years of data
\$6,000 + 40% (\$2,400) for 6-10 years of data
\$6,000 + 60% (\$3,600) for 11+ years of data

100,000 – 249,999 records:

\$7,000 + 20% (\$1,400) for 1-5 years of data
\$7,000 + 40% (\$2,800) for 6-10 years of data
\$7,000 + 60% (\$4,200) for 11+ years of data

250,000 – 499,999 records:

\$8,500 + 20% (\$1,700) for 1-5 years of data
\$8,500 + 40% (\$3,400) for 6-10 years of data
\$8,500 + 60% (\$5,100) for 11+ years of data

500,000 – 999,999 records:

\$10,000 + 20% (\$2,000) for 1-5 years of data
\$10,000 + 40% (\$4,000) for 6-10 years of data
\$10,000 + 60% (\$6,000) for 11+ years of data

More than 1,000,000 records:

\$15,000 + 20% (\$3,000) for 1-5 years of data
\$15,000 + 40% (\$6,000) for 6-10 years of data
\$15,000 + 60% (\$9,000) for 11+ years of data

Adjustments to costs:

Repeated linkage = reduce basic cost by 20%.

Please contact Gretchen Agha at the Public Health Institute (916) 779-2672, gcagha@ccr.ca.gov regarding payment processing.

Laws Governing CCR

California Law consists of 29 codes, covering various subject areas, the State Constitution and Statutes. The California Cancer Registry was established in accordance with the California Health and Safety Code sections 100330 and 103875-103885.

- Law establishing the California Cancer Registry: California Health and Safety Code, sections 103875-103885.
- Law concerning confidentiality of information collected by the California Cancer Registry: California Health and Safety Code, section 103885(g).

These sections of the Health and Safety Code are contained in *Policies and Procedures for Access to and Disclosure of Confidential Data from the California Cancer Registry*.

Confidentiality provisions are also addressed in:

- Government Code commencing with Section 6250, and specifically 6250 (c) and 6250 (k) (California Public Records Act).
- CA Civil Code Section 1798.24 and CA Welfare and Institutions Code Section 10850 (California Information Practices Act).