Procedures for Conducting Data Linkages with the California Cancer Registry

Chronic Disease Surveillance and Research Branch
California Department of Public Health
California Cancer Registry
1631 Alhambra Blvd., Suite 200
Sacramento, CA 95816
(916) 731-2500 Phone
(916) 454-1538 FAX

California Cancer Registry Data Linkage Experience

The California Cancer Registry (CCR) has routinely performed probabilistic data linkages with cancer registry data since 1992. Currently, CCR uses the National Cancer Institute's Match*Pro software to perform most data linkages. For some projects, CCR has also used the software Integrity (previously named AUTOMATCH), and Link Plus, developed by the Centers for Disease Control and Prevention (CDC).

CCR data has been linked to a variety of state and federal databases as well as various research cohorts. The results of these linkages have enhanced CCR data by providing updated patient follow-up information, identifying major comorbid illnesses among cancer patients, and providing valuable information for research and policy analyses. The CCR database is routinely linked to other statewide databases such as the California Office of Statewide Health Planning and Development (OSHPD) hospital discharge data, the California Department of Motor Vehicles driver license files, California Voter Registration files, and the California Department of Public Health (CDPH) Center for Health Statistics death certificate files. Results from these linkages provide updated follow-up information for a substantial number of CCR records.

The CCR database has also been linked to many diverse research cohorts including:

- Using Cancer Registries to Assess Quality of Cancer Care (Harvard Medical School)
- Cancer Genetics Research Information System (University of California, Irvine)
- Cancer Risk in in Solid Organ Transplant Recipients (National Cancer Institute)
- Assisted Reproductive Technology and Risk of Childhood Cancer (Michigan State University)
- Costs of Treating Breast Cancer in the Medi-Cal (Medicaid) Population in California (Department of Health Care Services)
- Childhood Cancer record Linkage Project (UC Berkeley)
- Continued Follow-up of PLCO Screening Trial Participants (National Cancer Institute)"
- Multiethnic/Minority Study of Diet and Cancer (University of Southern California)
- Follow-up of Cancer Prevention Study (CPS)-II Participants through Linkage with State Cancer Registries (American Cancer Society)
- Cancer Incidence in the United Farm Workers (United Farm Workers Health Collaborative)
- California Teachers' Study (University of Southern California)
- The Child Health and Development Study (California Department of Public Health)
- Every Woman Counts (Department of Health Care Services)

Procedures for Linkage

Data record linkage is a process to determine whether a record in one file matches a record, or several records, in another file. 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 in 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 in the two files. Anyone wanting to perform data linkage must have an understanding of the variables in 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 in the two data sets, both data sets are prepared for linkage. The file format for CCR linkages is found on page 4. The codes for all categorical data must match those of the CCR. For example, to follow CCR conventions, sex must be coded as "1" or "2" and not "M" or "F." Data values that are missing must be distinguished from zeroes in numeric fields. Similar information in 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 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 fields for the linkage. If necessary, a phonetic spelling of the street name and city may be done. After data preparation is complete, dictionaries of the data file are created for the match.

Next, 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, a procedure called "blocking" is used. Blocking 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. The 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 all the weights is used to help determine cutoffs for the weights on pairs of records. CCR will discuss these histograms with the researcher to jointly determine the cutoff thresholds. An upper cutoff is established above which all pairs are considered matches. A lower cutoff is also established below which all pairs are considered non-matches. Those pairs of records with a weight in between the upper and lower 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 performed 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. Finally, the CCR produces an encrypted data file that is uploaded to a secure server. Researchers are granted temporary log-in credentials to the secure server to retrieve their data. All passwords are provided over the phone.

Required File Format for CCR Data Linkage

Data files should be ASCII fixed field length files 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 (CCYYMMDD)	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) should also be included if available.

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

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 character fields are spaces and missing values for numeric fields are 9's.

2. All letters should be in capitals.

3. Names should not have embedded blanks, commas, apostrophes, or periods. Hyphens are allowed.

4. No “Jr.”, “II”, “III”, etc. in names.