CALIFORNIA CANCER REGISTRY

Procedures for Conducting Data Linkages with the California Cancer Registry

Chronic Disease Surveillance and Research Branch

California Department of Public Health

April 2020

CALIFORNIA CANCER REGISTRY DATA LINKAGE EXPERIENCE

BACKGROUND

Since 1992, the California Cancer Registry (CCR) has routinely performed probabilistic data linkages with cancer registry data. Currently, CCR uses the National Cancer Institute's Match*Pro software to perform most data linkages.

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, address, race, and place of birth. Some of these variables should not differ 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 understanding of when the data were collected and what variables are 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 3. 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. Address information needs to be standardized and separated into individual fields for the linkage, such as street address, city, zip code, and state.

Next, CCR staff discusses matching specifications, such as which variables are available in the cohort file, with the researcher requesting the linkage. Then, CCR prepares a linkage configuration file in the software used for the linkage. 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. The matching specifications delineate which variables must be exact during each linkage, and which variables are matched.

The matching algorithm runs on every pair of records that match on any two blocking variables and a score is calculated for them. Those pairs of records with a weight between the upper and lower cutoff are classified as a match, uncertain, or non-match, depending on the configuration of the linkage. Finally, manual review of uncertain and non-match records is performed.

Once the linkage is complete, data are extracted from the client file and the CCR file for all 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 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

Microsoft Excel (.csv, .xlsx) or SAS (.sas7bdat) data files are preferred with the following variables:

VARIABLE
FIRST NAME
MIDDLE NAME
LAST NAME
SOCIAL SECURITY NUMBER (NO HYPHENS)
DATE OF BIRTH (CCYYMMDD)
SEX (1=MALE, 2=FEMALE)
STREET ADDRESS
CITY
ZIP CODE

Additional personal identifying information (e.g. maiden name, hospital record number) may 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. Street addresses may have spaces.
4. No “Jr.”, “II”, “III”, etc. in names

Note: To obtain any death-related data fields (e.g., vital status, cause of death and survival time), researchers need additional approval from the Center for Health Statistics and Informatics. Please contact the research coordinator at research@ccr.ca.gov to start an application.