COLLECTION AND VERIFICATION OF DATA FOR MATCHED RECORDS FROM US CANCER AND HIV/AIDS REGISTRIES

Janice Watkins, Oak Ridge Associated Universities, T. Borges, Robert Stafford, Oak Ridge National Laboratory
Robert Biggar, James Goedert, National Cancer Institute
Janice Watkins, ORAU, MS 45, P.O. Box 117, Oak Ridge, TN 37830

Key Words: AIDS, HIV/AIDS Registry, Cancer Registry, data verification, record matching

BACKGROUND

Data for investigating cancer rates, cofactors, and disease progression in HIV positive individuals may help to promote an understanding of the relationship between tumor formation and immunity. HIV/AIDS and Cancer registries at selected U.S. sites were linked to find records of individuals common to both registries. These records were matched and the data collected into files containing variables to be utilized in later analyses. Systematic data verification was done on all records in the matched files to ensure maximum quality and consistency of data across site locations. These are the locations where matched files were collected:

- Atlanta
- Colorado
- Connecticut
- Florida
- Illinois
- Los Angeles
- Massachusetts
- New York
- New Jersey
- San Diego
- San Francisco
- Seattle

METHODS

Matching

The Cancer and HIV/AIDS records were linked using the commercially available AutoMatch Software (MatchWare Technologies, Burtonsville, MD), which employs a probabilistic methodology. The matches were done using last name, first name, and soundex codes for name, gender, race, birth, death, address, and, when available, Social Security number and middle name or initial. Each linked record received a score dependent upon the strength of the match and was automatically accepted or rejected when the weight was compared to a selected cut-off limit. Questionable matches, identified by low scores, were verified manually. Keys were built to link the match records and were maintained only by the two registries to assure confidentiality.

Data Scrubbing

A new database, devoid of any identifying information to ensure registrant confidentiality, was created from linked records for the site location by assembling data from both registries into one record per individual. In addition to demographic data and diagnosis dates, the database contained generally complete information on the mode of exposure to HIV and cancer grades and morphologies. Some data were also included from the HIV/AIDS registry on the results and dates of various medical tests given, but these data were missing for the majority of records.

Data were verified using a computer program developed to standardize date formats, cancer diagnosis codes, race codes, and gender codes. A dual purpose of the program is to identify errors or inconsistent values between fields. This program makes minor changes automatically and documents each change in a comment field. In addition, adjustments or corrections to fields of major importance are captured in a flag containing a cumulative score that has a unique value for each combination of changes. Particular attention is given to comparing values from the two registries for dates, ages at diagnoses, gender, race, and AIDS-associated tumor types. Also, all gender-specific cancer cases (cervical, uterine, testicular, and penile) are checked against gender listed in the record.

An additional field was created for each record to verify the last date when the individual was known to be alive. The value of this field corresponded to the death date when available, or it was the most recent date found in any field from either the AIDS/HIV or the cancer registry.

To increase compatibility of data across site locations, the matched file from each site was investigated manually for data not in standard form due to site-specific practices. Major discrepancies between values in the cancer and the AIDS/HIV registries were also rectified manually. When an inconsistency could not be resolved, the cancer registry value was used. These changes were also noted in the comment field and documented in the flag.

RESULTS

Approximately 7.1 million cancer records were matched against nearly 400,000 HIV/AIDS registry records. Of these, 44,700 records belonged to individuals common to both registries. Among accepted matches, nearly 40% had errors or discrepancies between common data from the two registries, mainly associated with dates. However, most date discrepancies involved differences within a few days, less than three months, or less than two years. AIDS-defining tumors occurring after an initial diagnosis tended to be found only in the cancer registries.

Figure 1 presents the number of matches made at each site location for individuals having records in both the cancer and the HIV/AIDS registries. As seen in Figure 1, the number of individuals listed in both the AIDS and cancer registries varied greatly by location. Since New York has two HIV/AIDS registries (one for New York City and one for the remainder of the state), separate matches with the New York State Cancer Registry and between the HIV/AIDS registries were done. When these two matched files were combined into one, duplicate records were removed for any individuals matched in both the New York City registries and the New York State registries. The largest number of matches from a single registry occurred with New York City. There were two distinct concentration centers for matches, California and the New York/New Jersey area. Other areas of the country contributed substantially fewer matches.

Figure 2 shows the percent of total records in which changes were made in each field of major importance, as documented in the flag. Over 60% of the matches had clean data for which no changes were necessary. Fewer than 5% of the records had changes in any fields of major importance with the exception of Karposi's sarcoma, which often appeared only in the HIV/AIDS registry.

CONCLUSIONS

A systematic linkage process that protects the confidentiality of matched individuals while efficiently uncovering matches was applied to identify records common to HIV/AIDS and Cancer registries throughout the United States. Data verification/validation was demonstrated to be an essential step in the data collection process in order to provide a database suitable for analysis in research studies. Great care must be taken to ensure consistent quality and data compatibility when dealing with large databases from multiple registry site locations.

This report concerns work conducted by the Oak Ridge National Laboratory, Oak Ridge, Tennessee under contract No. Y1-CP-8040-2 from the National Cancer Institute. ORNL is managed by Lockheed Martin Energy Research Corporation, for the U.S. Department of Energy under Contract No. DEAC0596OR22464.

Figure 1

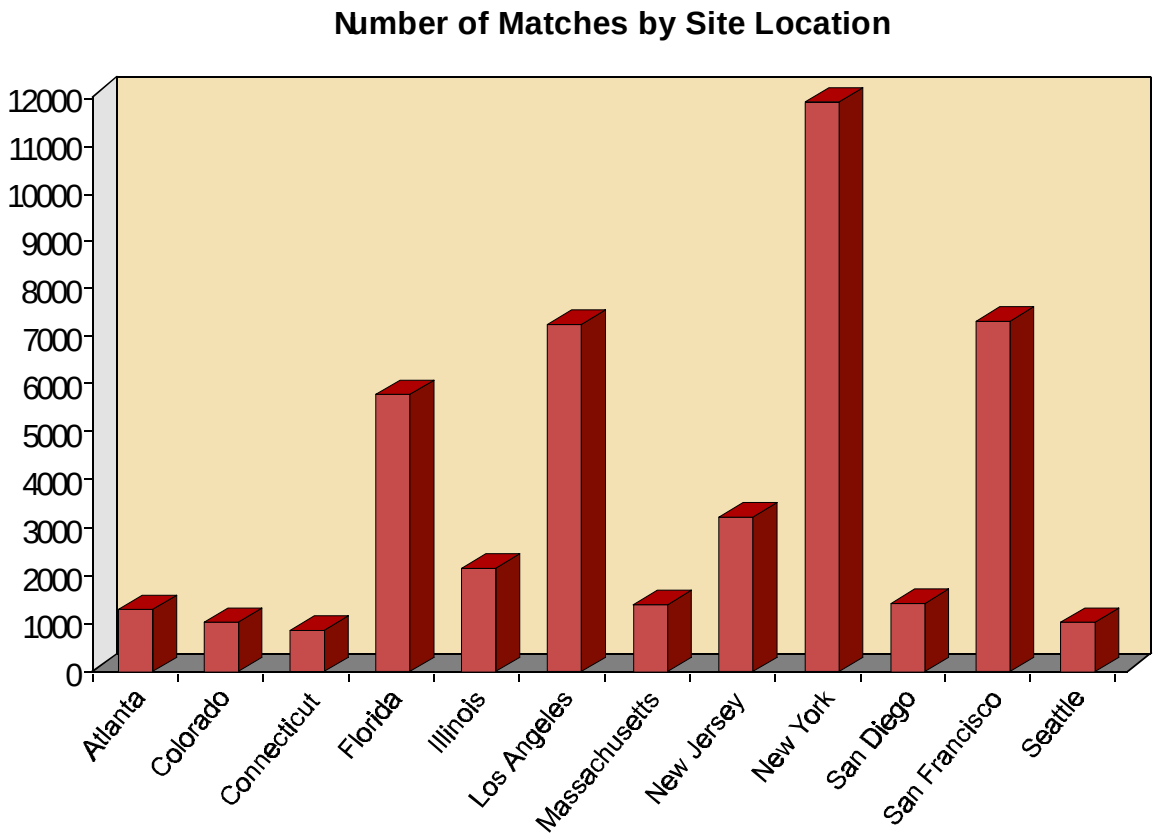
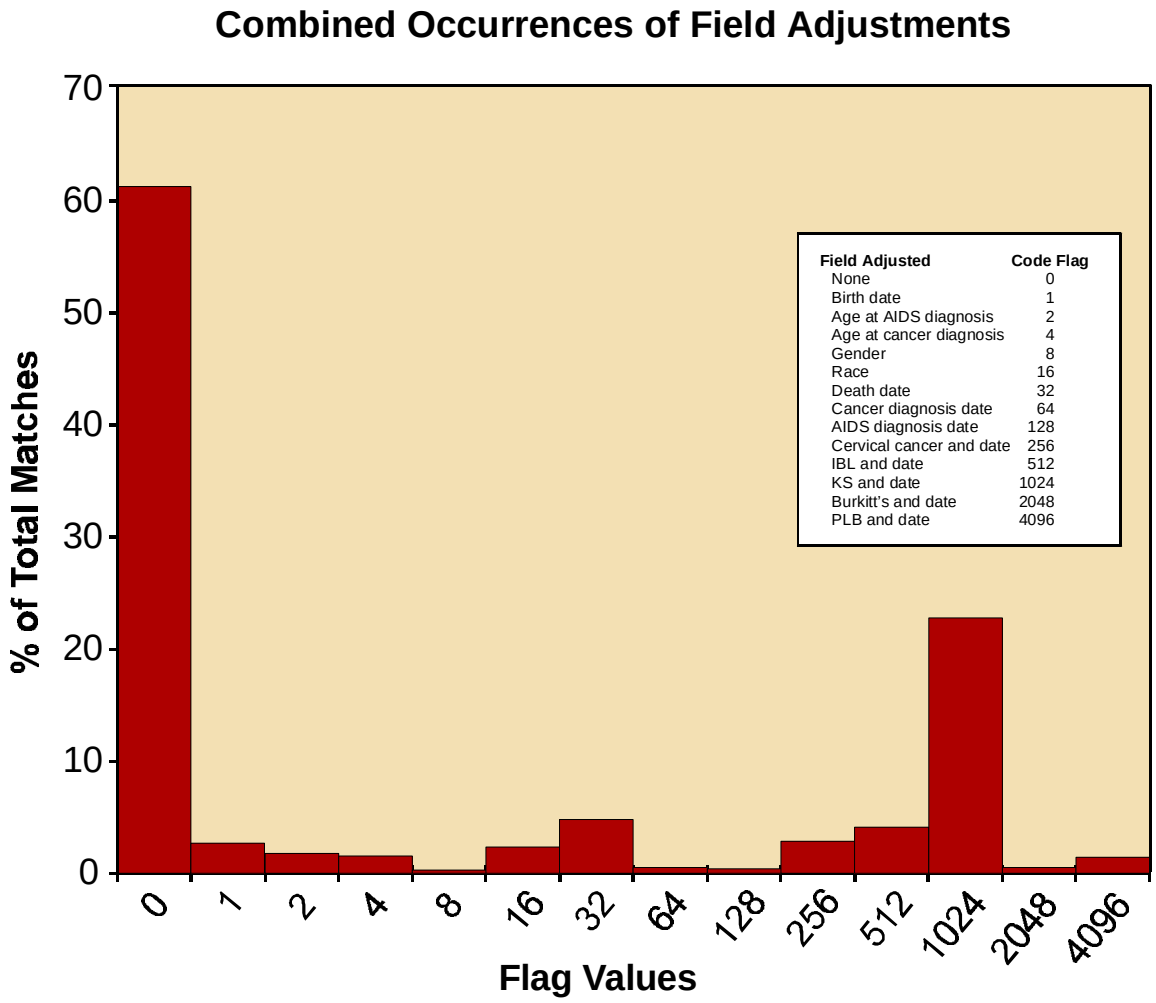

Number of Matches by Site Location
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Atlanta
Colorado
Connecticut
Florida
Illinois
Los Angeles
Massachusetts
New Jersey
New York
San Diego
San Francisco
Seattle

Figure 2


Combined Occurrences of Field Adjustments
% of Total Matches
70
60
50
40
30
20
10
0
0
1
2
4
8
16
32
64
128
256
512
1024
2048
4096
Flag Values
Field Adjusted
Code Flag
None 0
Birth date 1
Age at AIDS diagnosis 2
Age at cancer diagnosis 4
Gender 8
Race 16
Death date 32
Cancer diagnosis date 64
AIDS diagnosis date 128
Cervical cancer and date 256
IBL and date 512
KS and date 1024
Burkitt's and date 2048
PLB and date 4096