**Sample IRB protocol excerpt describing linkage methods with Link Plus software**

Data Sources

Description of data sources to be linked. Describe source and ownership of data, date ranges, approximate number of records, and data fields available for matching.

Record Linkage

1. Description of staff members & roles.
2. Prepare data sets to be linked (with a description of who will prepare each data set, how and where they will be stored in preparation for linkage).

To allow for increased efficiency and maximum probability scores, we will attempt to make blocking and matching fields as comparable as possible. Data in each file will be cleaned to minimize data errors (i.e., nonsense records) and to standardize fields, such as date formats and middle initial/middle name fields. Data preparation steps will be completed in SAS software, before each data set being exported as a <fixed width/delimited> text file for use by Link Plus.

1. Linkage

We will use the linkage software Link Plus, a program developed by CDC for use by the National Program of Cancer Registries. Link Plus runs in a Microsoft Windows environment and can be used with any data type, provided they are in a delimited or fixed-width text format. Link Plus employs probabilistic record linkage algorithms to identify matching records across two files and/or detect duplicate records in a single data file (deduplication). Link Plus has ability to identify not only deterministic matches in which comparison fields are exactly identical between the two files, but also possible matches in situations where minor differences exist between the two records (e.g., transposed digits in a social security number, missing data, or misspellings of a name). Match scores are assigned for each potential match pair based on the defined linkage algorithm; higher scores indicate greater probability of a match. Table 1 defines basic terms and concepts used in the linkage process, and summarizes our use of them for the current project.

The linkage will be conducted by trained project staff and occur on-site at <location>. In this way, no confidential data elements contained in the partnering data system will need to leave the premises. This will also allow for transparency and collaboration between Party1 and Party2 throughout the process. We will conduct each record linkage the project’s stand-alone laptop computer, which will not be connected to any network. The data file to be linked with will be transferred to a designated folder on the laptop for the linkage, and permanently deleted (i.e., securely “wiped”) from the computer’s memory prior to leaving the location.

One pass will be performed:

1. Both data files will be copied onto the designated computer loaded with Link Plus software.
2. Both files will be imported into Link Plus, and the first several rows of data will be checked to ensure that all fields were read in and formatted properly.
3. Blocking and matching parameters will be specified within the software:
4. Blocking variables:

* Last name (NYSIIS phonetic system)
* First name (NYSIIS phonetic system)
* Date of birth
* Social Security Number

1. Probabilistic matching variables (with Link Plus matching method in parentheses):

* First name (first)
* Last name (last)
* Date of birth (date)
* Social security number (SSN)
* Sex (exact)
* Address (generic string)

1. Missing values and date formats will be defined where appropriate
2. The Direct Method will be used, which applies the software’s default M-probability levels. The Link Plus-recommended cut-off value of 7.0 will be used to cut off non-matches.
3. The location of output files will be specified to a folder on the computer, and the linkage configuration saved to this same location prior to running the linkage.

After the linkage has run, we will assign match status to all matched record pairs using the Link Plus Manual Review screen and user-preferred options. Additional ID variables (such as suffix, street address) may be selected at this time to assist in the clerical review process. The reviewer will review the results and assign upper and lower cut-off scores. Clerical review of all matched pairs between these values will be conducted independently by two reviewers. Any discrepant results will be discussed and decided upon by consensus.

After all record pairs have been assigned a match status (match or non-match) we will export a text file containing the unique record ID of all matches. This ID number can then be used to merge the match results back into the original/analytic data set(s).

Staff will keep notes of the process throughout linkage set-up and clerical review. These are likely to include: the length of time the linkage took to run, total number of matched pairs above cut-off value, user-defined upper and lower cut-off values, and number of true matches.

Prior to leaving the linkage site, all data files with Protected Health Information (PHI) will be encrypted, and all linkage-associated files provided by Party2 will be deleted by securely wiping the files from the computer. A file containing previously agreed-upon health data, but no confidential demographic variables, may be provided to Party1 at this time. Any such files will also be encrypted and saved on the Project laptop.

|  |  |  |
| --- | --- | --- |
| **Table 1*.* Probabilistic linkage concepts** | | |
|  | **Description** | **Use for proposed project\*** |
| **Blocking** | An initial step to reduce the number of record comparisons and increase efficiency of linkage. At least one blocking variable must match *exactly* between the two records being compared; subsequent comparisons are made only within blocks. | Blocking variables:   * Last name * First name * Social security number * Date of birth |
| **Matching** | Within a block, matching variables are compared to generate a match score for each record pair.  Match scores for each variable are:   * Field-specific (matching DOB is scored higher than matching sex) * Value-specific (last name of ‘Hoopes’ is scored higher than ‘Smith’ due to frequency of occurrence) | Matching variables:   * Last name * First name * Middle name * Social security number * Date of birth * Sex * Address   The user may designate a specific matching algorithm for each variable. |
| **Match score** | The total probability weight assigned to each record pair; equal to the sum of scores generated by comparing each match field. Based on software-calculated M probability (sensitivity) and U probability (specificity). | The range of match scores is examined to determine upper and lower cut-off values. High match scores are designated true matchesand scores below 7.0 are automatically designated false matches. Record pairs between cut-off values are clerically reviewed. |
| **Clerical review** | Case-by-case review of uncertain matches that fall between the upper and lower cut-off values. Additional fields (“ID variables”) may be considered to assist in the designation of match status. This process can be completed independently by two or more reviewers to increase reliability. | ID variables:   * Street address * City, state, zip code * Suffix   Double clerical review will be conducted whenever possible |
| \* specifics may vary based on characteristics of matching data sets | | |

Data Security

Describe the protocol for handling PHI throughout the process, including who will have access to data sets, how access will be restricted, how they will be transferred between computers, and how resulting files will be stored. Reference confidentiality statements and data sharing agreements if relevant.