Link Plus Tip Sheet and Resources

Improving Data & Enhancing Access (IDEA-NW) Project
Northwest Tribal Epidemiology Center

Link Plus is a free probabilistic record linkage and de-duplication program developed at CDC's Division of Cancer Prevention and Control. Originally designed for use by CDC's National Program of Cancer Registries, the program can be used with any type of data in fixed width or delimited format. It can be downloaded here: http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm

Linkage Resources

Linkage toolkit including a Link Plus manual developed by IDEA-NW project staff:
http://www.npaihb.org/epicenter/project/linkage_resources/

Tribal data linkage toolkit (most of the same materials as on NPAIHB site):
http://www.cste.org/dnn/TribalDataLinkageToolkit/tabid/489/Agg1907_SelectTab/2/Default.aspx

HRSA linkage training materials:
http://webcast.hrsa.gov/conferences/mchb/mchepi_2009/data_linkage.htm

SAS macro programs developed to assist in the cleaning and linking of data using Link Plus (developed by Washington Education Research and Data Center):
http://www.erdc.wa.gov/briefs/technical/default.asp

Link Plus tips

With thanks to John Sabel (State of Washington) and Melissa Jim (IHS)

Setting up – Link Plus has its own set of default paths to folders (e.g., Configuration, Export, Report) that are based off the C drive (e.g., `C:\RegPlus\LinkPlus\Export`). It really likes these paths. You will quickly find that you are forever copying directory paths to the actual location of these folders from Windows Explorer to Link Plus. To more easily navigate to your project's folders and make Link Plus projects portable:

- Map the root directory of every Link Plus project to a single drive letter
    - Choose a drive letter, such as "L", to associate with all Link Plus projects. Then place shortcuts in each default Link Plus directory on your C drive to the analogous directory immediately off the chosen drive letter.
        - For example, if your chosen drive letter is "L", in the `C:\RegPlus\LinkPlus\Export` folder put a shortcut to `L:\Export`
    - Then before beginning any Link Plus project, map the location of the root directory of the project to that drive letter

Data cleaning

- Clean all linking data fields in a consistent manner
    - e.g., make all characters upper case, remove all dashes, sex and dates in same format
    - Name elements should be in separate fields (e.g., LAST, FIRST, MIDDLE, SUFFIX)

Setting up & running the linkage

- File 1 and File 2 have a one-to-many relationship: each record in File 1 can match 0 to many records in File 2.
    - If you are looking to "improve" one file (such as correct the race field), set it as File 1; otherwise,
    - If only one file is rigorously deduplicated, set that as File 1
- With real data Link Plus may take a while to read files
- When running the linkage, be patient – linkage times vary. But don't wait days; your computer can run out of virtual memory.
- Takes a lot of CPU
    - Shutdown and restart computer right before linkage to clear up as much space as possible
    - Turn off screen saver and close all other programs
- Make note of your file sizes (number of records) and the linkage time so you can compare performance and know what to expect for future linkages with your file(s). Some real-life examples:
    - 200,000 records matched to 10,000 recs: < 2 min.
    - 200,000 recs. matched to 800,000 recs: 47 min.
    - 2.4 million recs. matched to 400,000 recs: 2 hrs.
    - 2.4 million recs. matched to 1.2 million recs: 6.5 hrs.

Clerical Review

- If the clerical review screen appears to have no matching fields within record pairs, you may be missing a carriage return in one of your input files.
- Determining upper and lower cut-off values is often an iterative process
    - Focus initially on SSN and DOB
    - Once matches on SSN go away, pay attention to DOB, name, and sex
    - First time you start doubting a match, that score will be the upper cut-off: anything above it is considered a true match
    - When you start to see junk, score will be lower cut-off: anything below is considered a false match
- Names become more common in gray area. Keep an eye out for
    - Husbands and wives matching (SSNs match/sex different)
    - Brothers, sisters, and twins (last name match, SSN off by 1)
- Other fields ("ID variables") may help: address, race, tribe
- In the current version of Link Plus (v2.0), assigning match status by score will overwrite any matches you've already assigned in that score range, even if they are hidden from view
    - Keep note of the score ranges you've assigned previously
- Sorting by class may help in lengthy clerical reviews so you can more efficiently focus on specific differences. Be careful about assigning match status by score when doing this, though (see previous point).
- When deciding on match status for uncertain matches, consider the implications and use of the results.
    - Will results be used to "correct", update, or supplement the state HR? Call matches more conservatively.
    - Will results be used to merge two patient registries and you just want to assess the overlap? Can probably call matches more liberally.
    - Consider evidence **for** and **against** calling a match. E.g., missing SSN provides no evidence **for** a match, but different SSNs provide evidence **against** a match.
- Two people should conduct clerical review if possible (together or independently) and resolve discrepancies together