Link Plus - Self-Training Manual for Linkage

Developed by Northwest Tribal Registry Project staff,
Northwest Portland Area Indian Health Board

** DISCLAIMER: This un-official manual was developed for internal training, and is **not** sanctioned by CDC or Link Plus. **

Please direct comments/suggestions to ideanw@npaihb.org. Thanks!

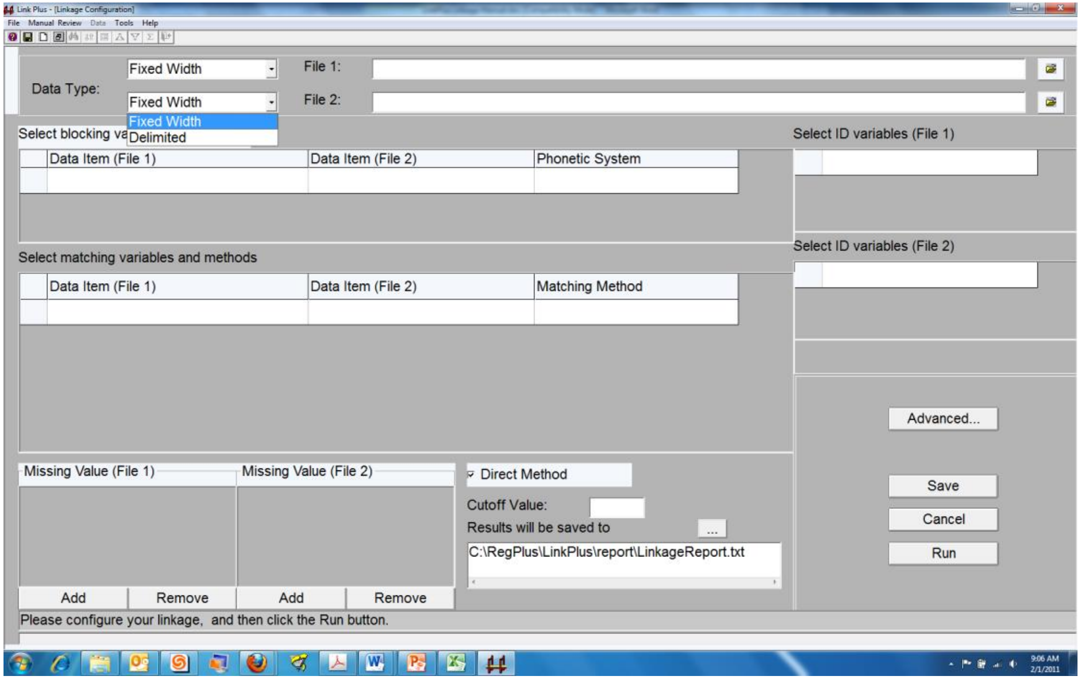
These instructions are written using the example of a record linkage between the Northwest Tribal Registry (NTR) and a state health registry (HR) such as a state cancer registry. The NTR is a list of AI/AN enrollees at NW IHS and tribal clinics. Some details and suggestions are specific to our project protocols and may not apply to all linkage studies.

1 – Getting Started

1. Download and install Link Plus software (www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm)
2. Create a folder where all linkage files will be saved (for this example "Linkage Folder")
3. Save NTR and HR in *Linkage Folder* as .txt files (fixed-width, tab- or comma-delimited)
4. Open Link Plus software
5. Select **File > New Linkage.** The linkage configuration screen opens. (These instructions can also be used to identify duplicates within a single file; select **File > New Deduplication** at this point instead.)

2 – Specifying Files

1. Choose Data Type (fixed width or delimited) of linkage files, using drop-down boxes.



Data Type: Fixed Width

Fixed Width

Results will be saved to C:\RegPlus\LinkPlus\report\LinkageReport.txt

Please configure your linkage, and then click the Run button.

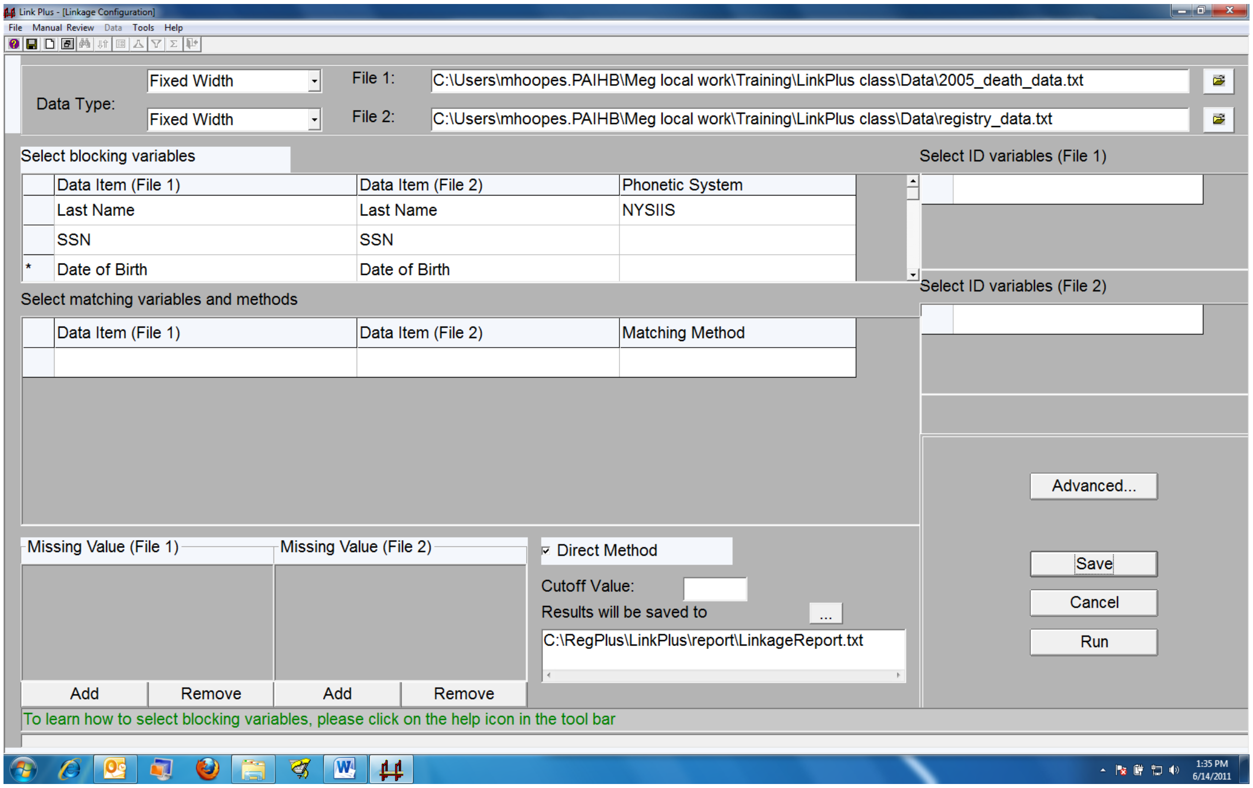
2. Identify linkage files. File 1 and File 2 have a one-to-many relationship. This means that a single record in File 1 can match multiple records in File 2. A general rule of thumb is to set the file you want to improve as File 1. E.g., we want each record for Tom Becker in the state health registry to have as many chances as possible to match a Tom Becker record in the NTR file, thus we will specify the health registry as File 1.
    a. Click on the folder icon to the right of the input box for File 1.
    b. The ***Open Data File*** dialog box opens. Navigate to the Linkage Folder. Select HR file and click **open.**
        ii. For fixed width files: ***Data Import Form For Fixed Width File*** dialog box opens.
            1. Specify record layout file (if a text layout file has already been created) by clicking on the folder icon. Otherwise, enter Data Item name, Start Position (column number), and Length for each data field. You will then be prompted to save the layout file if you haven't done so already.
            2. Click **View Data** and verify that all data fields are displaying correctly. If the data are not displaying correctly, look for errors in Start Position and Length.
            3. Click **OK.**
        iii. For delimited files: ***Data Import Form For Delimited File*** dialog box opens.
            1. Select delimiter. Select "Data item names in first row" option if appropriate.
            2. Click **View Data** and verify that all data fields are displaying correctly.
            3. Click **OK.**
    c. Repeat 2.a-b for File 2.

3 – Blocking Variables

Blocking variables are the fields that must match exactly (or phonetically if specified) for Link Plus to consider a record pair as a possible match and be sent for comparison on Matching Variables. Common blocking variables include Last Name, First Name, SSN, and DOB. If all four of these are used as Blocking Variables, then if any one of them matches identically in the two files, Link Plus will go on to compare match variables and assign a score.

1. In the **Select blocking variables** section click on the first empty cell in the **Data Item (File 1)** column of the grid. The drop down box menu contains all of the variable names that will be imported from File 1 (similarly for File 2). Use the drop down boxes to choose your blocking variables. Each line should represent the variables from the two files to be compared (i.e. the Social Security Number field for each file should be on the same line).
    a. Use the drop down box for **Phonetic System** to choose Soundex or NYSIIS where appropriate[1]. Numeric fields such as SSN and DOB do not use a phonetic system.
    b. Repeat until all blocking variables have been specified.

[1] NYSIIS is more specific but Soundex may bring more pairs for comparison when used for blocking. Studies have reported an accuracy increase of 2.7% over Soundex, and NYSIIS may perform better when Spanish names are used. If you're unsure, NYSIIS is recommended.


Link Plus - [Linkage Configuration]
File Manual Review Data Tools Help
Advanced...
Save
Cancel
Run
Add
Remove
Add
Remove
Select ID variables (File 1)
Select ID variables (File 2)
Missing Value (File 1)
Missing Value (File 2)
1:35 PM
6/14/2011

Data Type: Fixed Width / Fixed Width

File 1: C:\Users\mhoopes.PAIHB\Meg local work\Training\LinkPlus class\Data\2005_death_data.txt

File 2: C:\Users\mhoopes.PAIHB\Meg local work\Training\LinkPlus class\Data\registry_data.txt

Select blocking variables

Data Item (File 1)	Data Item (File 2)	Phonetic System
Last Name	Last Name	NYSIIS
SSN	SSN	
Date of Birth	Date of Birth	

Select matching variables and methods

Data Item (File 1)	Data Item (File 2)	Matching Method

☑ Direct Method

Cutoff Value:

Results will be saved to C:\RegPlus\LinkPlus\report\LinkageReport.txt

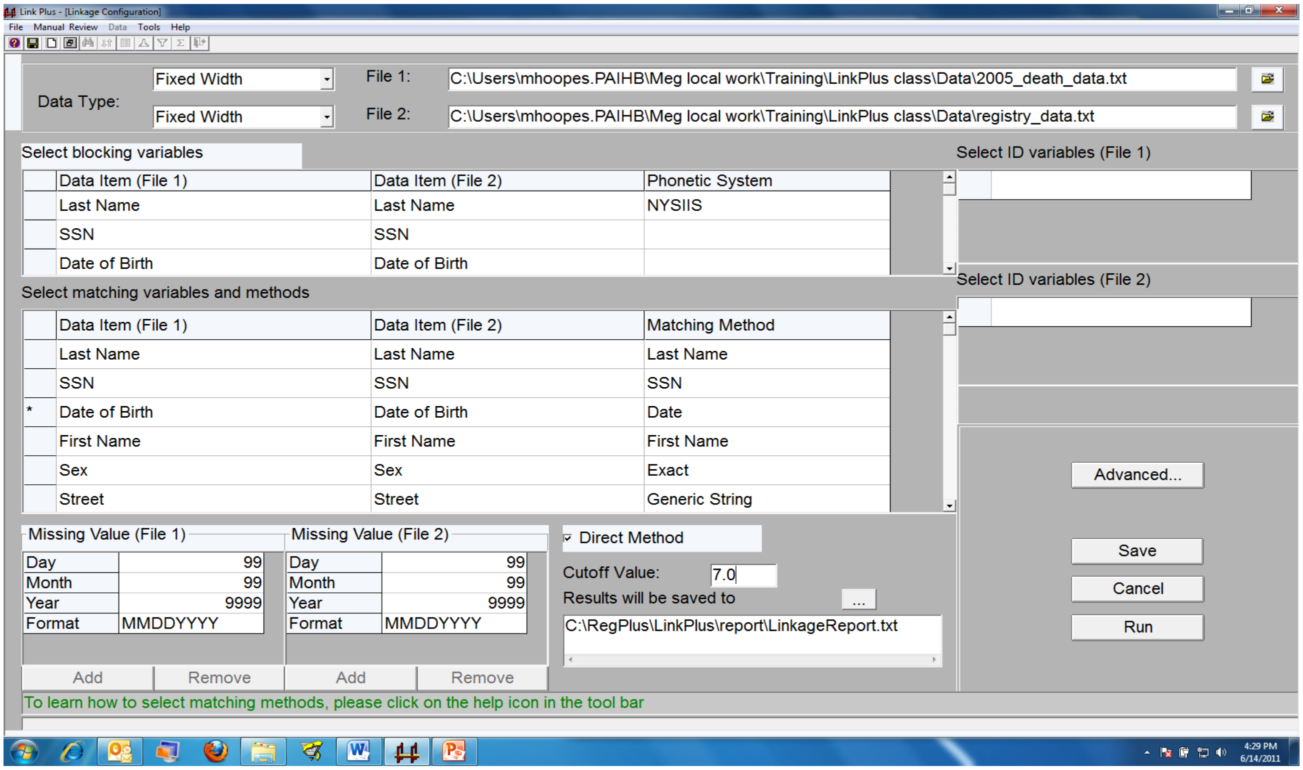
To learn how to select blocking variables, please click on the help icon in the tool bar

4 – Matching Variables & Missing Values

1. Select matching variables and methods. This is similar to selecting the blocking variables except that you should **choose all of the blocking variables** in addition to any other variables that will be used for comparison. Common matching variables include: Last Name, Middle Name, First Name, Sex, DOB, SSN, Race, Address.
    b. In the **Select Matching variables and methods** section click on the first empty cell in the **Data Item (File 1)** column of the grid. The drop down box menu contains all of the variable names that will be imported from File 1 (similarly for File 2). Use the drop down boxes to choose your matching variables just as you did for the blocking variables. Each line should represent the variables from the two files to be compared.
    c. Use the drop down box for **Matching Method** to choose the specific matching method associated with each pair of matching variables. The variables Last Name, Middle Name, First Name, SSN, and DOB have a specific matching method associated with them. The variable Sex will use the **Exact** matching method. Street Address should use the **Generic String** method.
    d. Repeat this step until all matching variables have been specified.
2. Define Missing Values. Link Plus automatically considers null or empty values as missing data for Matching Variables and allows the user to indicate additional values which are to be treated as missing data by the program. Most databases have a method for dealing with missing values for

each variable.[2] Specifying the convention used in each file produces better match scores and simplifies clerical review.

a. Select the Matching Variable that a missing value should be assigned to by clicking on the gray box to the left of the ***Data Item (File 1)***. The Matching Variable line should now be highlighted and an asterisk should appear in the box.
b. Click the **Add** button in the **Missing Values** grid for the appropriate file[3].
c. A cell opens where you can enter a missing value for the specified variable. Multiple missing values may be entered for a single variable by clicking **Add** again.
d. File 1 and File 2 may have different assigned missing values for the same comparison variable; be sure to enter them correctly for each.
e. Complete this step for each variable that has an assigned missing variable.



Data Type: Fixed Width (File 1), Fixed Width (File 2)

File 1: C:\Users\mhoopes.PAIHB\Meg local work\Training\LinkPlus class\Data\2005_death_data.txt

File 2: C:\Users\mhoopes.PAIHB\Meg local work\Training\LinkPlus class\Data\registry_data.txt

Select blocking variables

Data Item (File 1)	Data Item (File 2)	Phonetic System
Last Name	Last Name	NYSIIS
SSN	SSN	
Date of Birth	Date of Birth	

Select ID variables (File 1)

Select ID variables (File 2)

Select matching variables and methods

	Data Item (File 1)	Data Item (File 2)	Matching Method
	Last Name	Last Name	Last Name
	SSN	SSN	SSN
*	Date of Birth	Date of Birth	Date
	First Name	First Name	First Name
	Sex	Sex	Exact
	Street	Street	Generic String

Missing Value (File 1)

Day	99
Month	99
Year	9999
Format	MMDDYYYY

Missing Value (File 2)

Day	99
Month	99
Year	9999
Format	MMDDYYYY

☑ Direct Method

Cutoff Value: 7.0

Results will be saved to

C:\RegPlus\LinkPlus\report\LinkageReport.txt

To learn how to select matching methods, please click on the help icon in the tool bar

5 – Other Options

1. Select ID Variables if desired. Although Link Plus reports will automatically include **Matching** Variables, you may want to specify other variables to be included in the reports, called **ID Variables**. Alternately, you can select ID variables later when setting up the clerical review screen, and/or select additional variables after completing clerical review to export with the

[2] For instance, a common code of missing values for the DOB variable is 9999999. Other common values are Unknown or UNK.

[3] Entering missing values for date variables works a little differently from other variables. The **Day, Month,** and **Year** components appear automatically. Each of the components must be specified separately and the variable format must also be specified.

results. Including ID Variables at this time will affect the performance of the software, so it is recommended to wait.

2. Select or de-select **Direct Method**. The m-probability is the probability that matching variables agree given that a comparison pair is a match. Selecting the Direct Method will use the software's default m-probabilities. You can view the default probabilities and/or define your own by clicking on the **Advanced…** button. De-selecting Direct Method will cause Link Plus to use the Expected-Maximization (EM) algorithm to **compute** the m-Probabilities **based on your data** (probability estimates will reflect the characteristics of the data dynamically). The Direct Method is more efficient, but the EM method may improve results. If you're unsure, the Direct Method is recommended.
3. Enter **Cut-off Value**. The cut-off value is the score above which comparison pairs are accepted as potential matches and presented for manual review. The cut-off value recommended by Link Plus is 7.0.

6 – Specifying Output Files & Saving Configuration

1. Specify linkage report name and location. At the completion of a linkage run, Link Plus will generate a linkage report, named **LinkageReport.txt**, and store it by default in the Report folder of the Link Plus directory. The report is a tab delimited text file, and can be opened with a text editor or a spreadsheet program. (Link Plus also generates a non-match report, named **Non_MatchReport.txt**, which contains records from File 2 not matched to records in File 1.) The report header lists the files that were used and all of the parameters that were set for the run. Below the header is the report detail. Records are presented in comparison pairs sorted by their linkage scores in descending order. Pairs with scores above the selected cut-off value are listed. The detail row for each record contains all of the Matching and ID Variables used in the linkage.
    a. By default Link Plus will write over the report file each time Link Plus is run. To rename the report and/or change the location **(recommended)**, click on the ellipses (...) to the right of **Results will be saved to**. The **Save linkage results to the file below** dialog box opens.
    b. Use the **Save in:** drop down box to navigate to the folder where you would like to save the reports. For our example, the Linkage Folder that we created earlier.
    c. Type the name you would like to save the match report in the **File name:** box, and click **Save**. The linkage report folder and name should now be changed to your specifications. Your configuration screen should now look similar to the image on p. 4.
4. Save your work. This way you can use the configuration that you have set up in the future.
    a. Click **Save**. The **Save Configuration** dialog box opens.
    b. Use the **Save in:** drop down box to navigate to the folder where you would like to save the configuration file. For our example, the Linkage Folder.
    c. Enter the name you would like to use for the configuration file in the **File name:** box, and click **Save**.

7 – Running the Linkage

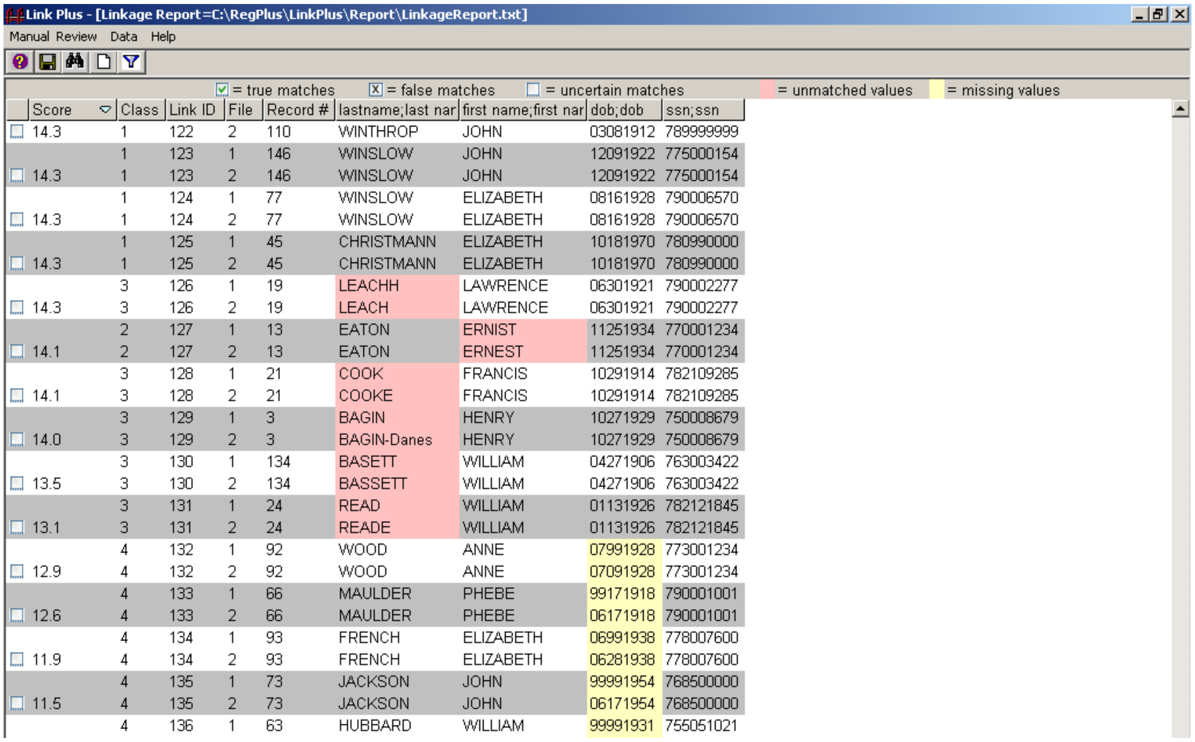
1. Click **Run** to run the linkage. Go get a cup of coffee while you're waiting patiently.
2. A box in the Link Plus software informs you that the "Link process is done" and displays some information including the time it took to run and the locations of the saved linkage report and the non-match report. Close this box or select **Conduct manual review now** if you are ready to continue.

8 – Manual Review

1. Create the layout for the manual review. Click **Manual Review** > **New View** in the toolbar.
   a. Select the linkage report (you may need to navigate to the file) and click **Open**.
   b. **Select Additional Fields For Manual Review** dialog box opens. If you would like to view additional fields during the manual review, you have two methods to choose from (this is where you can add those ID Variables that you may have skipped over earlier). Either click the box to the left of Field Names in the **Add fields in File 1** (or File 2) and the variables will be shown on separate columns on the manual review screen; or use the drop down boxes for the **Add fields in File 1 and corresponding fields in File 2** section, then click the large up arrow button (located in the center of the panel) to add the fields you have selected.
   c. Verify that all desired fields are now in the **Fields for manual review** window.
   d. Change the order of fields on the manual review list if desired. Those fields that are most important for comparison should be listed first and those that are least important should be listed last. For example: SSN, DOB, Last Name, First Name, Middle Name, Sex, Address
      i. Select the variable that should be moved in the list.
      ii. Click the small up arrow (to the left of the large up arrow button) or the small down arrow (to the right of the large up arrow button) to move the variable accordingly.
      iii. Click **OK** when finished.
2. The Manual Review screen opens for the Linkage Report with the variables ordered as you just specified. The Manual Review screen allows you to identify true matches and export files for matches and non-matches (and uncertain matches if desired). Higher scores are more likely to be true matches based on the matching variables and methods specified in the linkage configuration. Link Plus offers many options and features to assist in the manual review process.

   Here are some options and tips:

   - To improve readability, you can right-click on the SSN and DOB headers to add dashes. Columns can also be sorted, hidden/unhidden, and re-sized.
   - Class categories are determined by which of the matching variables match exactly and the definition of each class can be viewed by clicking on **Data** > **Definition of "Class"**; match status of blocking and matching variables are color coded.
   - At any time you can display only uncertain matches by clicking on the icon that looks like a funnel or all pairs by clicking on the icon that looks like an upside down funnel. Or click **Data** > **Display Uncertain Matches Only...** and **Data** > **Display All** to accomplish the same task.
   - Users can switch between two viewing modes: the Datasheet View (shown below) and the Pair View. Additional options can be explored in the **Manual Review** and **Data** pull-down menus.
   - The Manual Review process can be saved and resumed at a later time. Select **Manual Review > Save View As…** to save a .view file. To re-open, you will need the .view file **and** the original linkage report created by Link Plus.


Link Plus - [Linkage Report=C:\RegPlus\LinkPlus\Report\LinkageReport.txt]
Manual Review
Data
Help

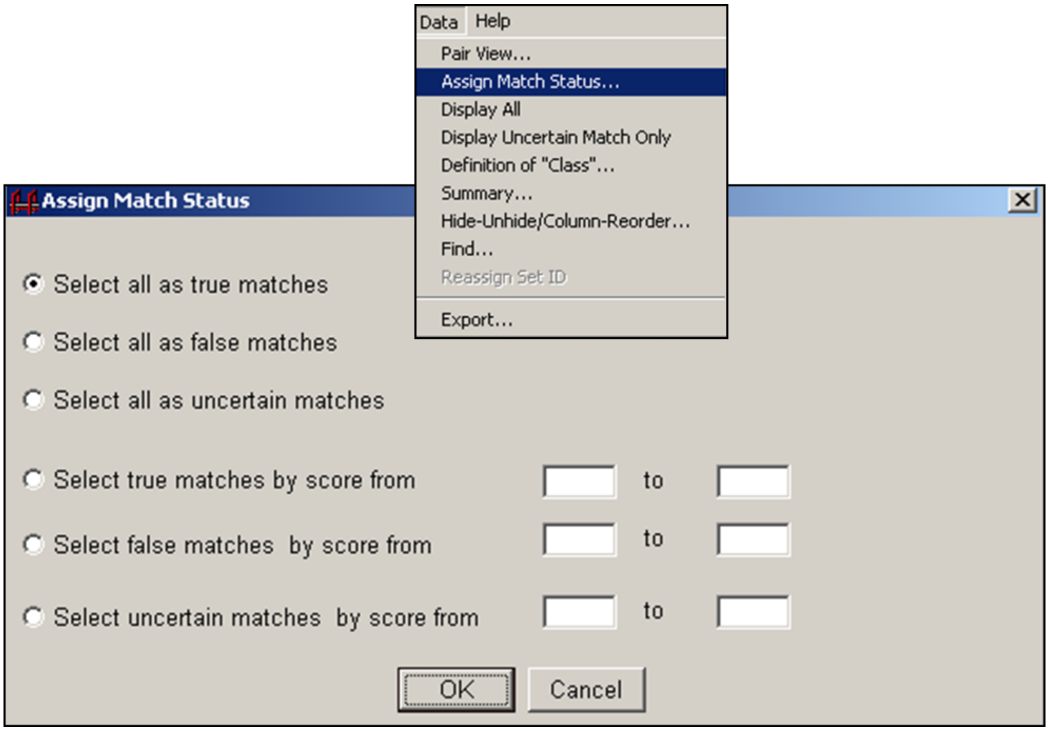
☑ = true matches ☒ = false matches □ = uncertain matches 🟥 = unmatched values 🟨 = missing values

	Score	Class	Link ID	File	Record #	lastname;last nar	first name;first nar	dob;dob	ssn;ssn
□	14.3	1	122	2	110	WINTHROP	JOHN	03081912	789999999
		1	123	1	146	WINSLOW	JOHN	12091922	775000154
□	14.3	1	123	2	146	WINSLOW	JOHN	12091922	775000154
		1	124	1	77	WINSLOW	ELIZABETH	08161928	790006570
□	14.3	1	124	2	77	WINSLOW	ELIZABETH	08161928	790006570
		1	125	1	45	CHRISTMANN	ELIZABETH	10181970	780990000
□	14.3	1	125	2	45	CHRISTMANN	ELIZABETH	10181970	780990000
		3	126	1	19	LEACHH	LAWRENCE	06301921	790002277
□	14.3	3	126	2	19	LEACH	LAWRENCE	06301921	790002277
		2	127	1	13	EATON	ERNIST	11251934	770001234
□	14.1	2	127	2	13	EATON	ERNEST	11251934	770001234
		3	128	1	21	COOK	FRANCIS	10291914	782109285
□	14.1	3	128	2	21	COOKE	FRANCIS	10291914	782109285
		3	129	1	3	BAGIN	HENRY	10271929	750008679
□	14.0	3	129	2	3	BAGIN-Danes	HENRY	10271929	750008679
		3	130	1	134	BASETT	WILLIAM	04271906	763003422
□	13.5	3	130	2	134	BASSETT	WILLIAM	04271906	763003422
		3	131	1	24	READ	WILLIAM	01131926	782121845
□	13.1	3	131	2	24	READE	WILLIAM	01131926	782121845
		4	132	1	92	WOOD	ANNE	07991928	773001234
□	12.9	4	132	2	92	WOOD	ANNE	07091928	773001234
		4	133	1	66	MAULDER	PHEBE	99171918	790001001
□	12.6	4	133	2	66	MAULDER	PHEBE	06171918	790001001
		4	134	1	93	FRENCH	ELIZABETH	06991938	778007600
□	11.9	4	134	2	93	FRENCH	ELIZABETH	06281938	778007600
		4	135	1	73	JACKSON	JOHN	99991954	768500000
□	11.5	4	135	2	73	JACKSON	JOHN	06171954	768500000
		4	136	1	63	HUBBARD	WILLIAM	99991931	755051021

3. Manually assigning Match Status.
   a. To assign a pair as a true match click in the Match Status check box once, a check will appear. This designates the pair as a true match. A pair can also be assigned as a true match by pressing the “M” key when the pair’s Match Status check box is highlighted.
   b. To assign a pair as a non-match click in the Match Status check box twice, an “x” will appear. This designates the pair as a false or non-match. A pair can also be assigned as a false match by pressing the “N” key when the pair’s Match Status check box is highlighted.
   c. To assign a pair as an uncertain match click in the Match Status check box until it is blank. A pair can also be assigned as an uncertain match by pressing the “B” key when the pair’s Match Status check box is highlighted.
4. Assigning Match Status by Score. After an initial review of the results it is often possible (and preferable) to identify a score above which all proposed matches are true matches (upper cut-off) and a score below which all proposed matches are false matches (lower cut-off). The remaining possible matches will need to be clerically reviewed and individually identified as true or false matches.
   a. **Data** > **Assign Match Status...**. The Assign Match Status dialogue box opens.
   b. After choosing one of the options and inserting scores where required, click **OK**.
   c. Link Plus returns to the Manual Review screen to continue reviewing the proposed matches.
   d. Assigning upper and lower cut-off scores can be repeated several times, and is often an iterative process.

      Example:
      - Solid matches somewhere between 20-30
      - Non-matches between 10-20
      - Upper cut-off set at 27 and higher; Lower cut-off set at 17 and lower

- Gray Area = Greater than 17 and less than 27



Assign Match Status

● Select all as true matches

○ Select all as false matches

○ Select all as uncertain matches

○ Select true matches by score from ___ to ___

○ Select false matches by score from ___ to ___

○ Select uncertain matches by score from ___ to ___

e. Match status can be changed at any time during the manual review process. Keep in mind, if you re-assign match status by score, it will over-write any match determinations you may have already made within that score range.

f. When are done, select **View Uncertain Matches** to make sure that all match pairs have been assigned a match status.

5. **Tips for manual review**
   - Focus initially on SSN and DOB
     - Names have a lot of issues (spelling and spacing)
   - Once matches on SSN go away, pay attention to DOB, name, and sex
   - First and middle name switches are common
   - Use race & address variables if available
   - First time you start doubting if a pair is a match
     - Score will be upper cut-off score
     - Anything above is considered a match
   - When you start to see junk
     - Score will be lower cut-off score
     - Anything below is considered a non-match
   - Keep an eye out for
     - Husbands and wives matching (SSN's match/sex different))
     - Brothers, sisters, and twins (LN match, SSN off by 1)
   - Two people should review so that results can be combined and resolved
   - These are suggestions - need to know your own data

9 – Exporting Files

1. Export Results. You have formatting options available for the export file.
    a. Format export file. Click **Tools** > **Options....** The options dialogue box opens. Click on the **Export** tab. Link Plus exports all files in delimited format but offers delimiter options of Tab, Comma, Semi Colon, or you can specify a different delimiter in Other.
        i. Match Status of export file. More than likely you will want to export true matches, and possibly false matches. Link Plus will not export them to the same file so you must identify which you want exported. Click in the appropriate box.
        ii. Click **OK**.
    b. Identify fields that will be included in the export file. Click **Data** > **Export....**
        i. The **Export Setting** dialogue box opens. Fields can be selected by clicking in the check box to the left of File 1 and File 2. As you click in the box the field is automatically added to the Exported Fields box. At a minimum, you will probably want to export the unique ID number from each file.
        ii. If you decide that you want to reorder the fields after adding them to the Exported Fields box you can do so by highlighting the field to be repositioned and using the up and down arrows. When you have added all of the fields to be exported and they are ordered correctly click **Export**.
    c. The **Create a delimited file** dialogue box opens.
    d. Use the **Save in:** box to identify where you want the export file to be saved.
    e. Use the **File name:** box to identify the name of the file. If you are exporting matches you should identify it as such, similarly for non-matches. Click **OK.**
    f. The report opens in your preferred text editor.
    g. The Link Plus system remains on the **Export Setting** dialogue box. Close the dialogue box.
    h. If you want to export another file (i.e. Non-matches or Matches) repeat this step.

10 – Finishing Up

1. Inspect the exported reports and the non-match (residual) file. It may be necessary to delete sensitive information such as SSN, name, street address, etc. before leaving the state registry premises. Import the file into your favorite data processing software.
2. Any files that are sensitive or restricted should be deleted from the laptop according to project protocols prior to leaving the state registry premises.
3. It is recommended that you create dummy files and practice this process a few times before attempting your first “live” linkage.

Additional tips

- With real data Link Plus may take a while to read files
- Be patient – linkage times vary. But don't wait days – your computer can run out of virtual memory.
- Keep a 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.
- 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
- 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.
- 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.

Comments? Suggestions? Corrections? Please send to ideanw@npaihb.org*. Thank you!*