NAACCR 2022 Call for Data – Patient and Tumor Deduplication Instructions

This document will explain how to use the Match*Pro record linkage software to deduplicate patients and cancer cases that may exist in your registry's database.

1. To get started, you'll need to download and install Match*Pro version 2.2.3. The software can be downloaded from https://seer.cancer.gov/tools/matchpro/.

2. After you've downloaded and installed the software the next step will be to create an extract containing patient names, dates of birth, social security numbers, sex, telephone number, and addresses (both current address and address at DX) for ALL records of eligible primary tumors diagnosed from 1995-2021. If your registry's inception year is not on/before 1995 (*i.e., your registry does not have complete data until 1996 or later*) then the start date for the extract should coincide with your registry's inception year. The extract should include cases obtained through data exchange agreements with other central cancer registries, federal facilities like the Veteran's Administration, and other non-hospital data sources. The extract should be created in the NAACCR-XML (version 22) format. To minimize the linkage runtime, create an extract containing ONLY these fields:

   a. Patient Id Number (#20)
   b. Name—First (#2240)
   c. Name—Last (#2230)
   d. Name—Maiden (#2390)
   e. Name—Middle (#2250)
   f. Name—Birth Surname (#2232)
   g. Date of Birth (#240)
   h. Social Security Number (#2320)
   i. Telephone (#2360)
   j. Sex (#220)
   k. Addr Current—No & Street (#2350)
   l. Addr Current—City (#1810)
   m. Addr Current—Postal Code (#1830)
   n. Addr Current—State (#1820)
   o. Addr at DX—No & Street (#2330)
   p. Addr at DX—City (#70)
   q. Addr at DX—Postal Code (#100)
   r. Addr at DX—State (#80)

3. Once you've created the extract you are ready to begin the process of deduplicating the patients in your database. Two linkage configuration files were included with these instructions for this purpose.

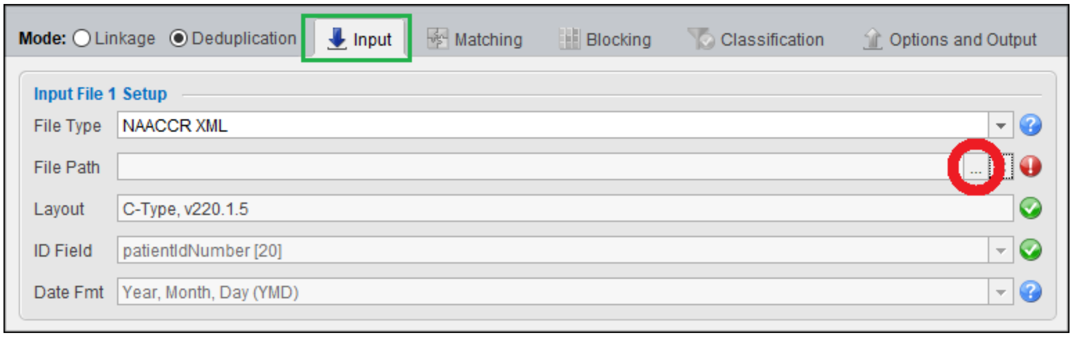
   If you used Match*Pro to deduplicate your database last submission AND you still have the Status Archive you created last year, then should use the configuration file named **naaccr-patient-deduplication-with-status-archive.mplc.** Otherwise, you should use the configuration file named **naaccr-patient-deduplication.mplc**.

Extract the configuration file from the zip folder, then double-click on it. It should open automatically with Match*Pro.

a. If, for some reason, the file does not open automatically then you will need to manually open the file. To do this you'll need to start the Match*Pro software (a shortcut for which should have been created on your desktop during the installation process). Once the software is running, click on the File menu and select "Open Linkage Configuration …" from the list of options (this is the 2nd option in the list). A file selection dialog will appear. Browse to the location of the linkage configuration file, select it, and press the open button. The linkage configuration will be opened.

4. Now that the linkage configuration file is open, you'll need to provide Match*Pro with the location of the extract you created in step 2. There are 5 tabs on the linkage configuration screen. The first tab, which is labeled "**Input**", is where you will perform this step. This tab is shown to you by default.

   a. Press the browse button associated with the **File Path** for **File 1**, which has been circled in **RED** in the image below.



Mode: ○ Linkage ● Deduplication

Input File 1 Setup

File Type: NAACCR XML

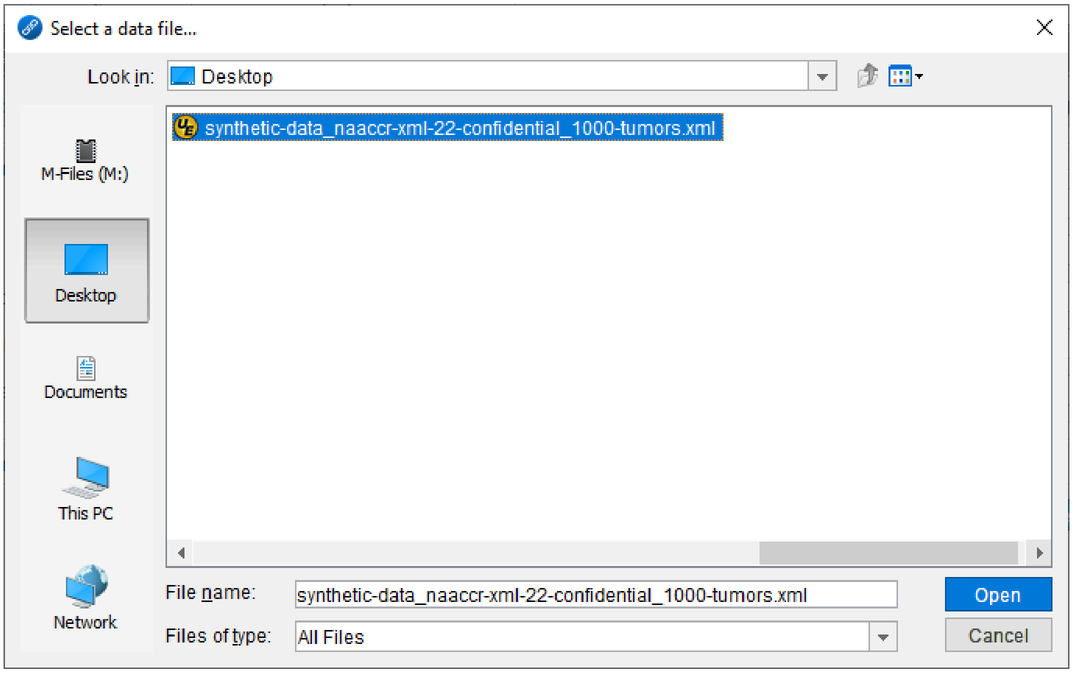
File Path:

Layout: C-Type, v220.1.5

ID Field: patientIdNumber [20]

Date Fmt: Year, Month, Day (YMD)

b. A file selection dialog will appear. Browse to the location of the extract you created in step 2, select the file, and then press the **OPEN** button.



Select a data file...

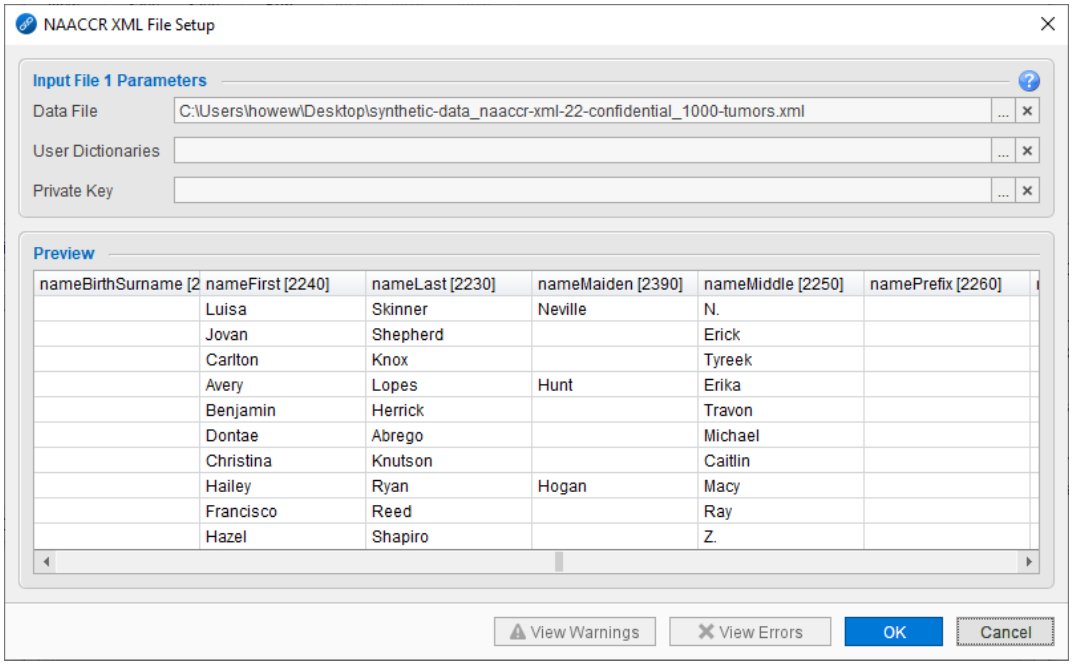
Look in: Desktop

synthetic-data_naaccr-xml-22-confidential_1000-tumors.xml

File name: synthetic-data_naaccr-xml-22-confidential_1000-tumors.xml

Files of type: All Files

c. The NAACCR XML File Setup dialog will be displayed. You can use the preview window to verify that all of the fields have been populated. Once you're convinced that all of the fields are being read in correctly, press the **OK** button. The dialog will close, and you'll be returned to the Input tab on the linkage configuration screen.



NAACCR XML File Setup

Input File 1 Parameters

Data File: C:\Users\howew\Desktop\synthetic-data_naaccr-xml-22-confidential_1000-tumors.xml

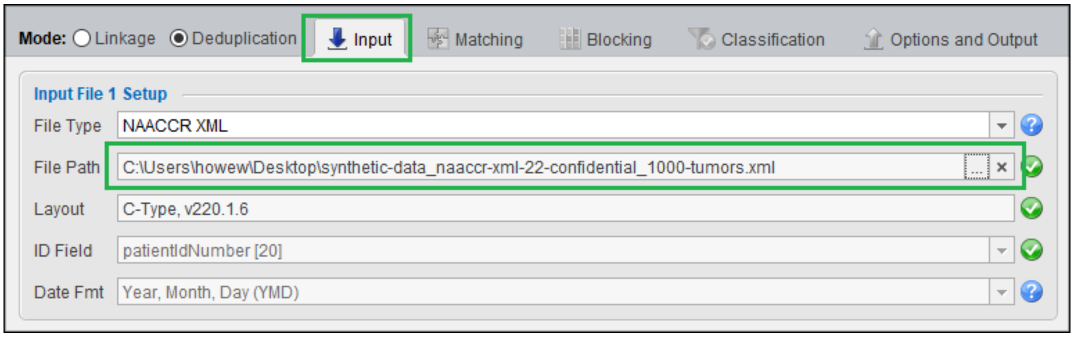
User Dictionaries:

Private Key:

Preview

nameBirthSurname [2	nameFirst [2240]	nameLast [2230]	nameMaiden [2390]	nameMiddle [2250]	namePrefix [2260]
	Luisa	Skinner	Neville	N.	
	Jovan	Shepherd		Erick	
	Carlton	Knox		Tyreek	
	Avery	Lopes	Hunt	Erika	
	Benjamin	Herrick		Travon	
	Dontae	Abrego		Michael	
	Christina	Knutson		Caitlin	
	Hailey	Ryan	Hogan	Macy	
	Francisco	Reed		Ray	
	Hazel	Shapiro		Z.	

d. The name and location of the extract will be displayed in the text box.



Input File 1 Setup

File Type: NAACCR XML

File Path: C:\Users\howew\Desktop\synthetic-data_naaccr-xml-22-confidential_1000-tumors.xml

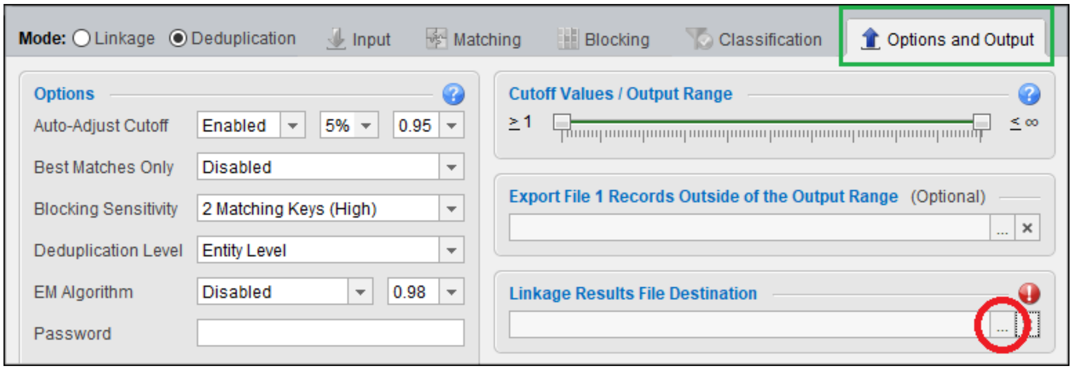
Layout: C-Type, v220.1.6

ID Field: patientIdNumber [20]

Date Fmt: Year, Month, Day (YMD)

5. You are now finished with the input tab. Switch to the **Options and Output** tab. This is the 5th and final tab that's displayed on the linkage configuration screen. Here you'll need to provide Match*Pro with the location of where you'd like the linkage results file to be created.

   a. Press the browse button associated with the **Linkage Results File Destination**, which has been circled in **RED** in the image below.



Options

Auto-Adjust Cutoff: Enabled 5% 0.95

Best Matches Only: Disabled

Blocking Sensitivity: 2 Matching Keys (High)

Deduplication Level: Entity Level

EM Algorithm: Disabled 0.98

Password:

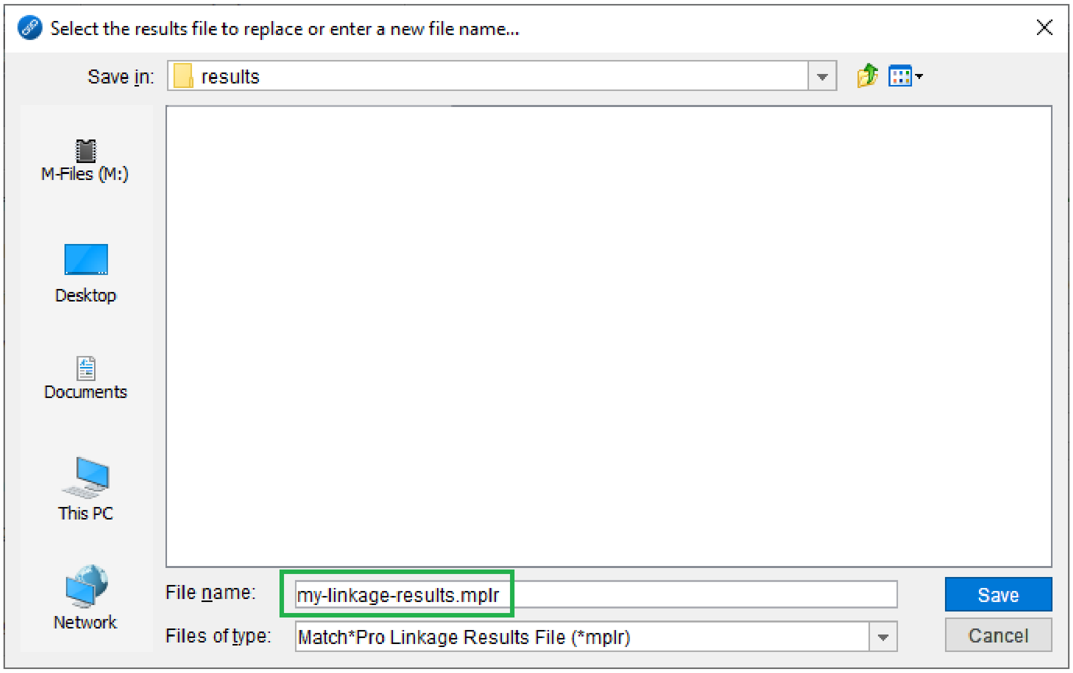
Cutoff Values / Output Range

≥ 1 ≤ ∞

Export File 1 Records Outside of the Output Range (Optional)

Linkage Results File Destination

b. A save dialog will appear. Browse to the location of where you'd like the results file to reside, then enter a filename and press the **SAVE** button. **PLEASE BE ADVISED THAT THE LINKAGE RESULTS FILE SHOULD BE CREATED ON YOUR C:/ DRIVE AS OPPOSED TO A NETWORK DRIVE AS SLOW AND/OR DROPPED NETWORK CONNECTIONS CAN CORRUPT THE FILE (PARTICULARLY IF IT IS LARGE). THE RESULTS FILE SHOULD REMAIN ON YOUR C:/ DRIVE UNTIL ALL OF THE WORK OUTLINED IN THIS DOCUMENT HAS BEEN COMPLETED.**



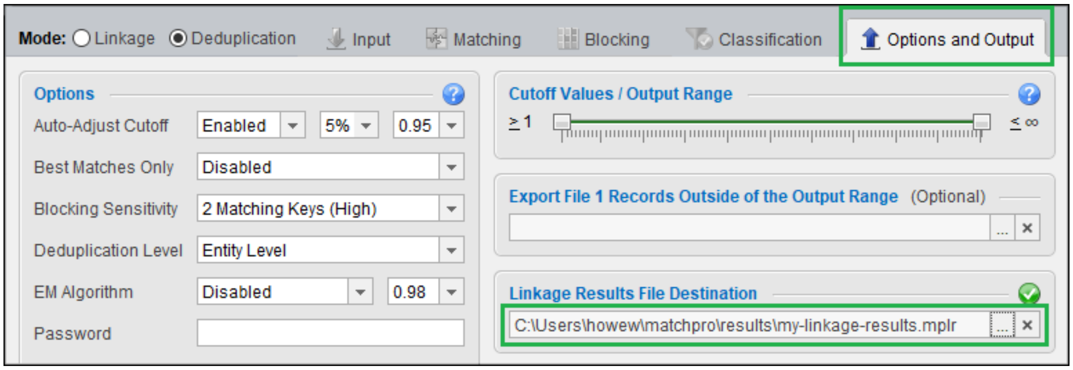
Select the results file to replace or enter a new file name...

Save in: results

File name: my-linkage-results.mplr

Files of type: Match*Pro Linkage Results File (*mplr)

c. The name and future location of the results file will be displayed in the text box.



Mode: ○ Linkage ● Deduplication

Options

Auto-Adjust Cutoff	Enabled	5%	0.95
Best Matches Only	Disabled		
Blocking Sensitivity	2 Matching Keys (High)		
Deduplication Level	Entity Level		
EM Algorithm	Disabled	0.98	
Password			

Cutoff Values / Output Range

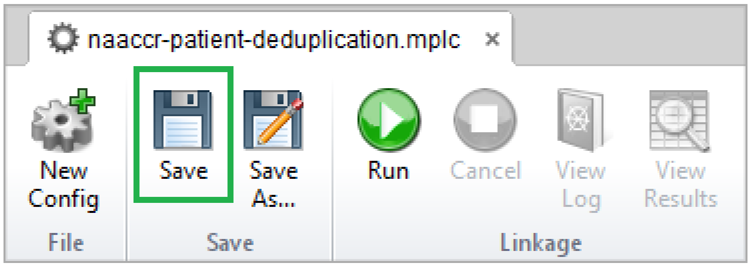
≥ 1 ≤ ∞

Export File 1 Records Outside of the Output Range (Optional)

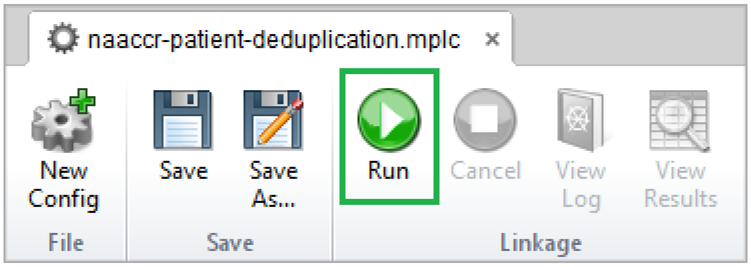
Linkage Results File Destination

C:\Users\howew\matchpro\results\my-linkage-results.mplr

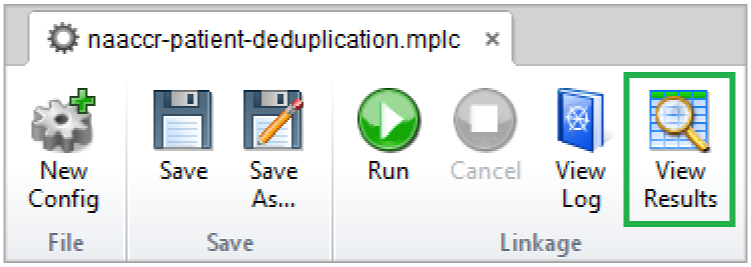
6. Press the **SAVE** button, which is located at the top of the linkage configuration screen, to save all of the changes that you've made to the configuration file.



7. Press the **RUN** button. The linkage process will begin. The run time will vary depending on the number of records that are in the extract.

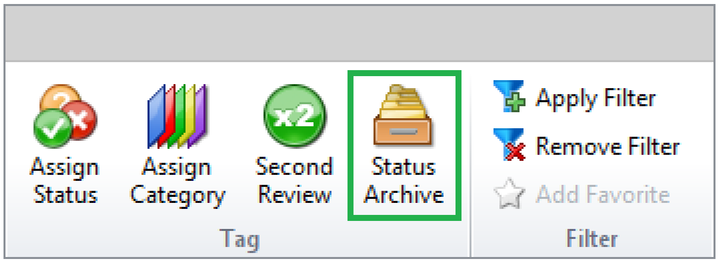


8. Once the linkage has finished running, press the **VIEW RESULTS** button to open the linkage results file. The linkage results screen will be displayed.

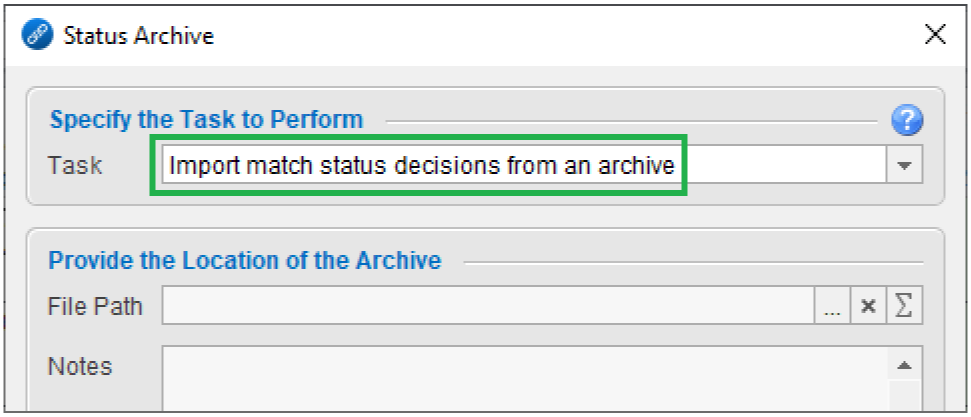


9. If you did not use Match*Pro to deduplicate your patients last year or you did not create a status archive last year, skip this step and proceed to step 10. You should only perform this step if you are using the naaccr-patient-deduplication-with-status-archive.mplc configuration file.

   a. Press the **STATUS ARCHIVE** button, which is located at the top of the linkage results screen. The Status Archive dialog will appear.



   b. Select "**Import match status decisions from an archive**" from the drop down.


Status Archive

Specify the Task to Perform

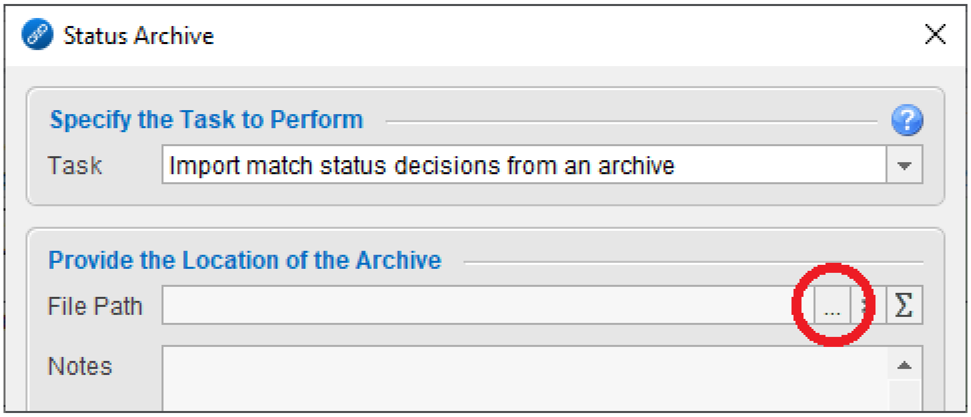
Task: Import match status decisions from an archive

Provide the Location of the Archive

File Path:

Notes:

   c. Provide the location of the status archive you created following last year's patient deduplication linkage. The button you need to press to do this is circled in **RED** in the image below.


Status Archive

Specify the Task to Perform

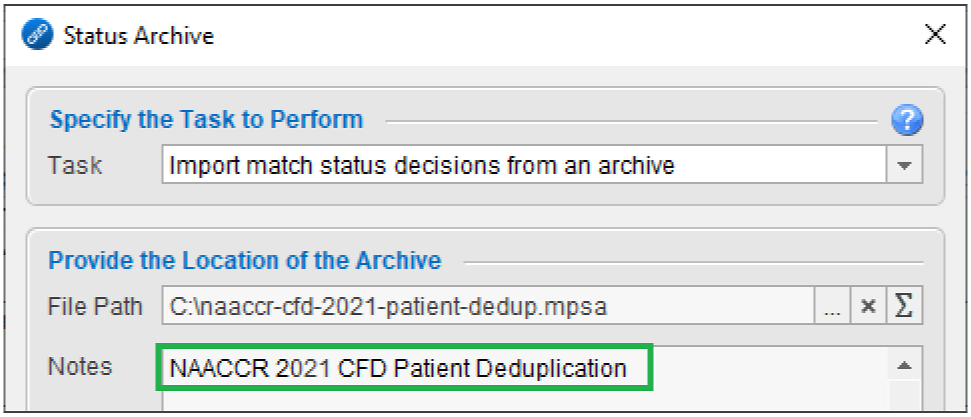
Task: Import match status decisions from an archive

Provide the Location of the Archive

File Path:

Notes:

d. After the file is selected the notes will be updated. Check the notes to confirm you selected the correct file.


Status Archive
Specify the Task to Perform

Task: Import match status decisions from an archive

Provide the Location of the Archive

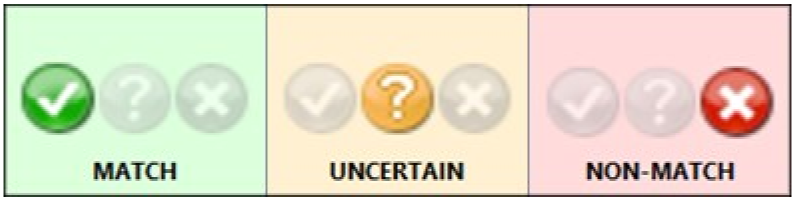
File Path: C:\naaccr-cfd-2021-patient-dedup.mpsa

Notes: NAACCR 2021 CFD Patient Deduplication

e. Press the **OK** button. The dialog will close and the match statuses on the linkage results screen will update. Any pairs that change from uncertain/yellow to non-match/red were previously reviewed by staff from your registry last year. They are not matches and they do not need to be reviewed a second time.

f. **SKIP TO STEP 13.**

10. Depending on the data, Match*Pro may have already classified some of these potential duplicates as matches or non-matches. **Matches** will have a **green check mark** next to them and **non-matches** will have a **red "X"** next to them. The remaining potential duplicates will have a **yellow question mark** next to them. These are the **uncertain** pairs.

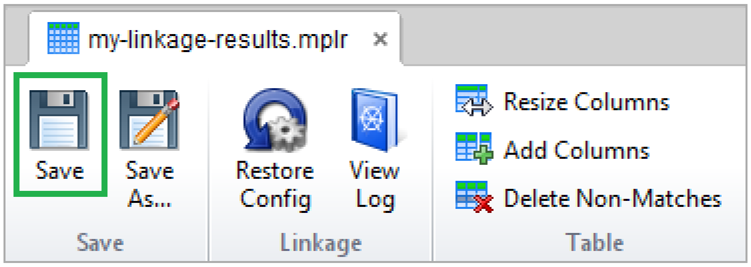


11. Take a moment to review all of the pairs with a **green check mark** next to them to see if any of them are **false positives** (pairs that were declared a match that you don't believe are actually matches). If you see them, change the match status of those pairs to 'non-match' by clicking on the red "X".

12. Next, take a moment to review all of the pairs with a **red "X"** next to them to see if any of them are **false negatives** (pairs that were declared a non-match that you believe are actually a match). If you see them, change the match status of those pairs to 'match' by clicking on the green check mark.

13. You'll then need to review the remaining yellow/uncertain pairs and assign them either a 'match' or 'non-match' status. This is where the bulk of your time performing the manual review will likely be spent.

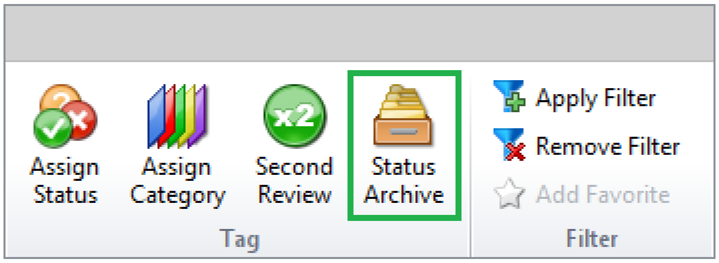
14. When you've finished reviewing all of the pairs **PRESS THE SAVE BUTTON** at the top of the screen to lock in all of the decisions you made during the manual review.



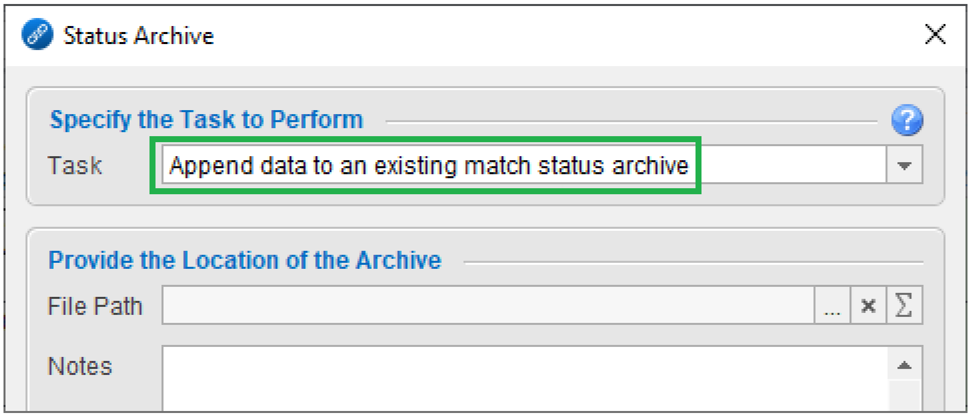
15. If you did not use Match*Pro to deduplicate your patients last year or you did not create a status archive last year, skip this step and proceed to step 16. You should only perform this step if you are using the naaccr-patient-deduplication-with-status-archive.mplc configuration file.

    Next, take a moment to **UPDATE** the status archive for next year.

    a. Press the **STATUS ARCHIVE** button, which is located at the top of the linkage results screen. The Status Archive dialog will appear.

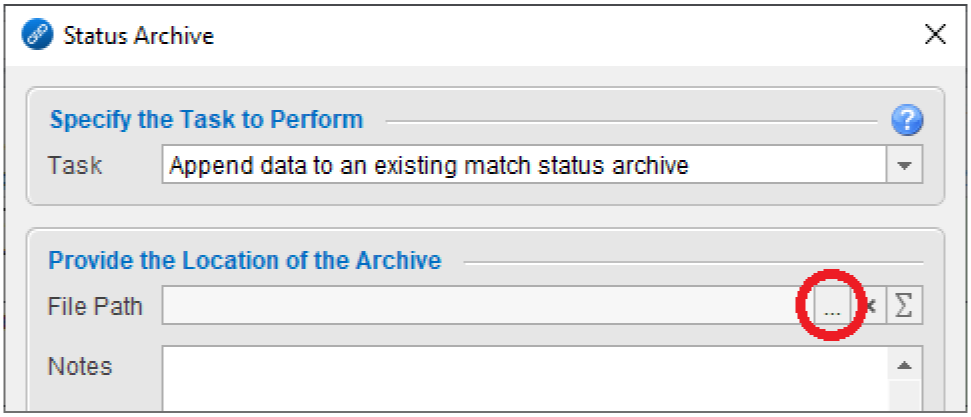


b. Select "**Append data to an existing match status archive**" from the drop down.


Status Archive
Specify the Task to Perform
Provide the Location of the Archive
File Path
Notes

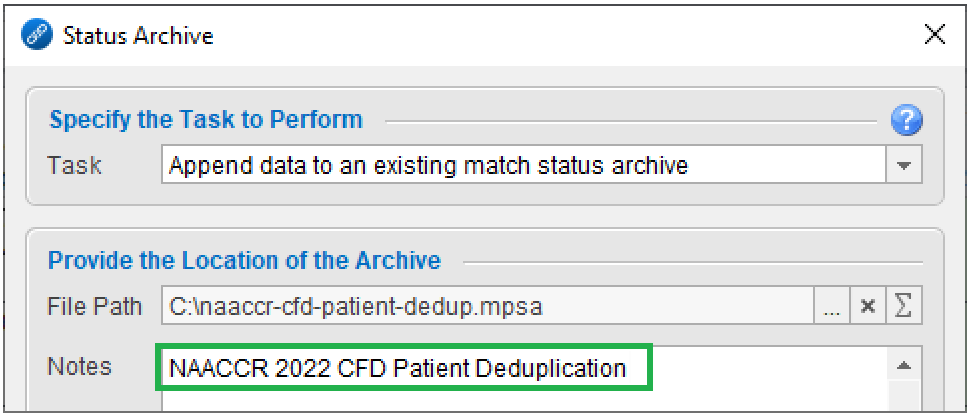
Task: Append data to an existing match status archive

c. Provide the location of the existing match status archive. This would be the one you selected in step 9. The button you need to press in order to do this is circled in **RED** in the image below.


Status Archive
Specify the Task to Perform
Provide the Location of the Archive
File Path
Notes
...

Task: Append data to an existing match status archive

d. Update the description in the notes field

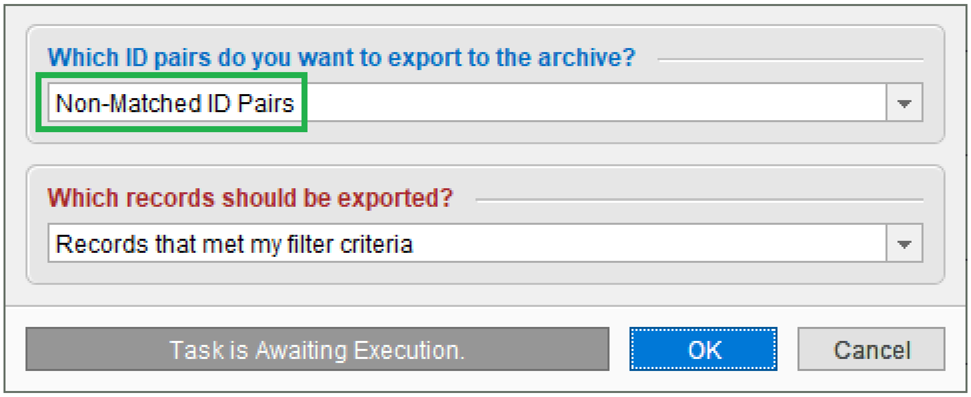

Status Archive
Specify the Task to Perform
Provide the Location of the Archive

Task: Append data to an existing match status archive

File Path: C:\naaccr-cfd-patient-dedup.mpsa

Notes: NAACCR 2022 CFD Patient Deduplication

e. Select "Non-Matched Pairs" from the drop down towards the lower half of the dialog.



Which ID pairs do you want to export to the archive?

Non-Matched ID Pairs

Which records should be exported?

Records that met my filter criteria

Task is Awaiting Execution.

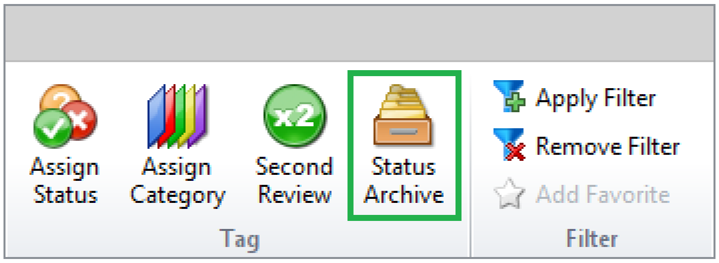
f. Press the OK button. The status archive will be updated with the new information from this year. **Make sure to save this file for next year.**

g. Close the dialog and return to the manual review screen

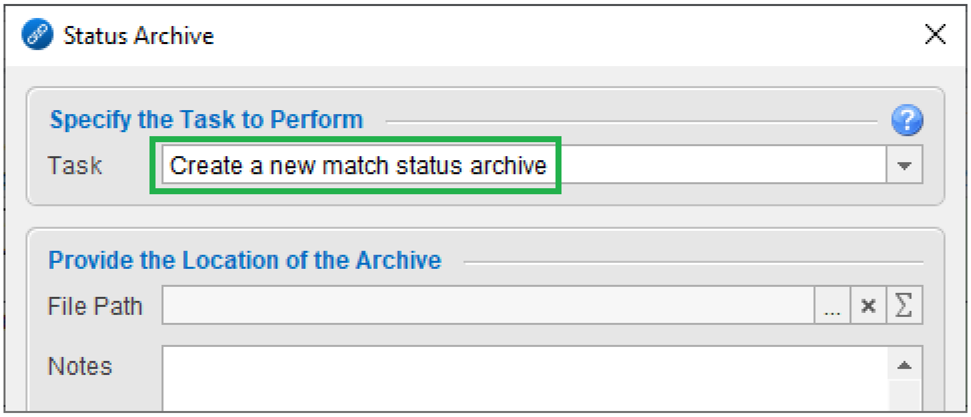
h. SKIP TO STEP 17

16. Next, take a moment to create a **NEW** status archive for next year.

a. Press the **STATUS ARCHIVE** button, which is located at the top of the linkage results screen. The Status Archive dialog will appear.



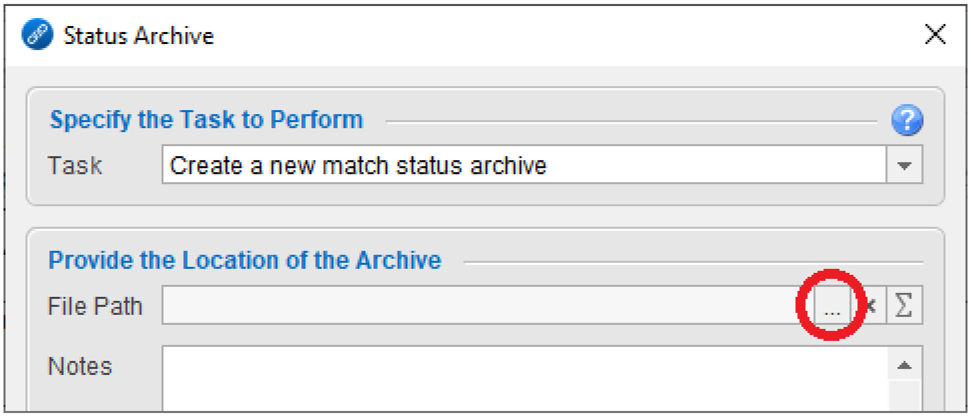
b. Select "**Create a new match status archive**" from the drop down.


Status Archive
Specify the Task to Perform
Provide the Location of the Archive
Notes

Task: Create a new match status archive

File Path:

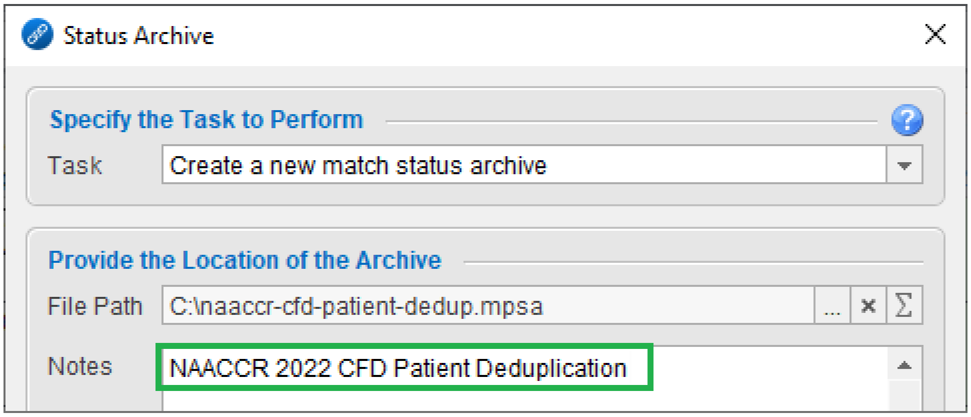
c. Provide the location of where you'd like to save the archive. The button you need to press in order to do this is circled in **RED** in the image below.


Status Archive
Specify the Task to Perform
Provide the Location of the Archive
...
Notes

Task: Create a new match status archive

File Path:

d. Enter a description of the archive in the notes section (*e.g.,* "NAACCR 2022 CFD patient deduplication").

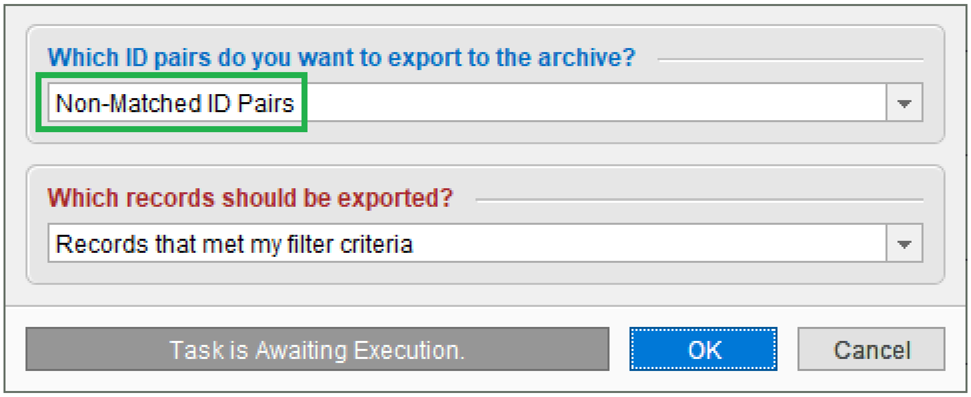

Status Archive
Specify the Task to Perform
Provide the Location of the Archive

Task: Create a new match status archive

File Path: C:\naaccr-cfd-patient-dedup.mpsa

Notes: NAACCR 2022 CFD Patient Deduplication

e. Select "**Non-Matched Pairs**" from the drop down towards the lower half of the dialog.



Which ID pairs do you want to export to the archive?

Non-Matched ID Pairs

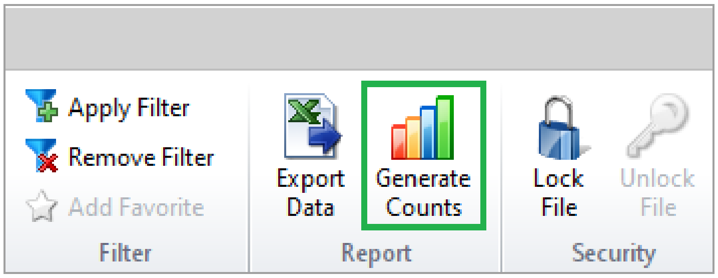
Which records should be exported?

Records that met my filter criteria

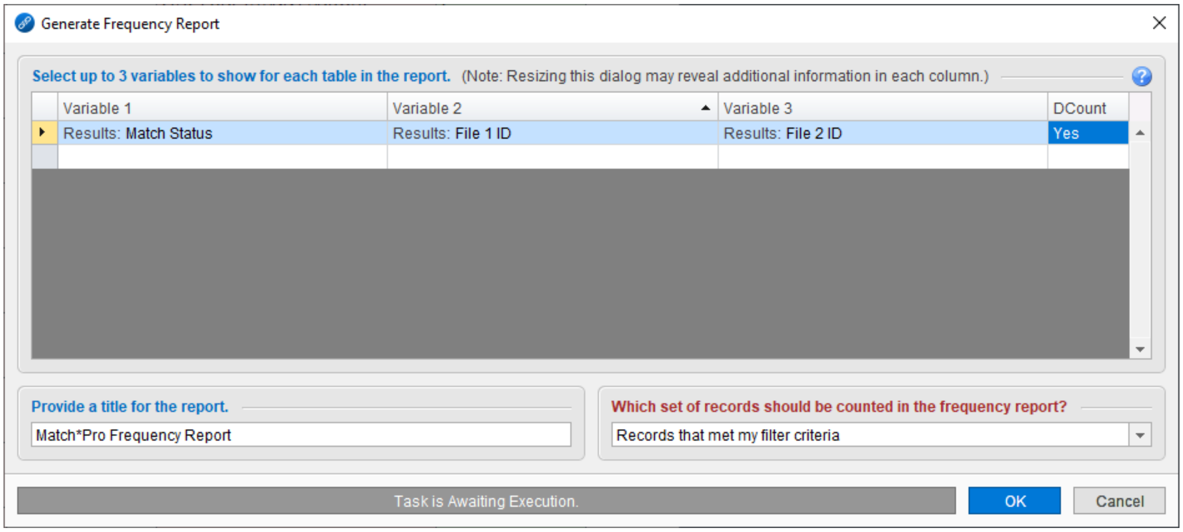
f. Press the **OK** button. The status archive will be created in the location you specified in step 15c, above. **Make sure to save this file for next year.**

g. Close the dialog to return to the manual review screen.

17. Press the **GENERATE COUNTS** button, which is located at the top of the linkage results screen. The Generate Counts dialog will appear.



18. Select "**Results: Match Status**", "**Results: File 1 ID**", "**Results: File 2 ID**", and "**Yes**" from the dropdowns in the first row in the table.



Generate Frequency Report

Select up to 3 variables to show for each table in the report. (Note: Resizing this dialog may reveal additional information in each column.)

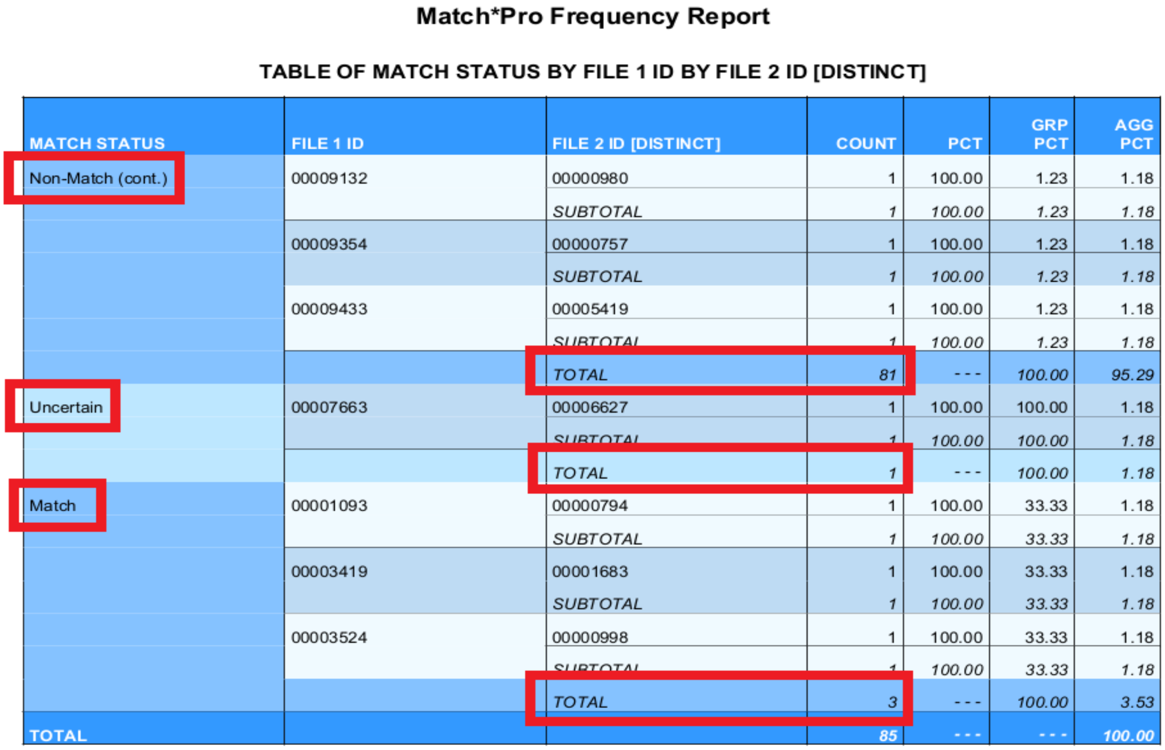
Variable 1	Variable 2	Variable 3	DCount
Results: Match Status	Results: File 1 ID	Results: File 2 ID	Yes

Provide a title for the report.
Match*Pro Frequency Report

Which set of records should be counted in the frequency report?
Records that met my filter criteria

Task is Awaiting Execution.

19. Press the **OK** button. The dialog will close, and a frequency report will be displayed showing you the number of matches, non-matches, and uncertain cases. If you performed a full manual review and the number of uncertain cases is zero, then you will only see the number of matches and non-matches. **Write down the sub-totals for the number of patient pairs that are matches, non-matches, or uncertain, as NAACCR will be asking you for them later in the submission process.**

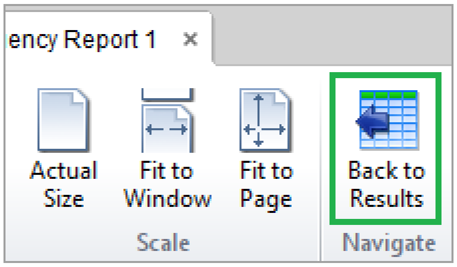


Match*Pro Frequency Report

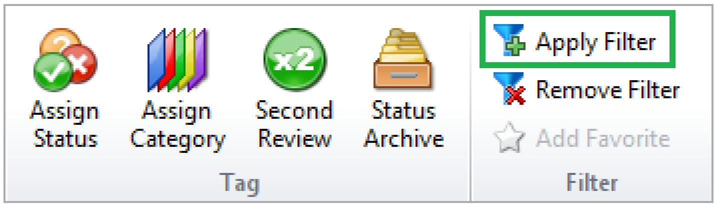
TABLE OF MATCH STATUS BY FILE 1 ID BY FILE 2 ID [DISTINCT]

MATCH STATUS	FILE 1 ID	FILE 2 ID [DISTINCT]	COUNT	PCT	GRP PCT	AGG PCT
Non-Match (cont.)	00009132	00000980	1	100.00	1.23	1.18
		SUBTOTAL	*1*	*100.00*	*1.23*	*1.18*
	00009354	00000757	1	100.00	1.23	1.18
		SUBTOTAL	*1*	*100.00*	*1.23*	*1.18*
	00009433	00005419	1	100.00	1.23	1.18
		SUBTOTAL	*1*	*100.00*	*1.23*	*1.18*
		TOTAL	*81*	- - -	*100.00*	*95.29*
Uncertain	00007663	00006627	1	100.00	100.00	1.18
		SUBTOTAL	*1*	*100.00*	*100.00*	*1.18*
		TOTAL	*1*	- - -	*100.00*	*1.18*
Match	00001093	00000794	1	100.00	33.33	1.18
		SUBTOTAL	*1*	*100.00*	*33.33*	*1.18*
	00003419	00001683	1	100.00	33.33	1.18
		SUBTOTAL	*1*	*100.00*	*33.33*	*1.18*
	00003524	00000998	1	100.00	33.33	1.18
		SUBTOTAL	*1*	*100.00*	*33.33*	*1.18*
		TOTAL	*3*	- - -	*100.00*	*3.53*
TOTAL			*85*	- - -	- - -	*100.00*

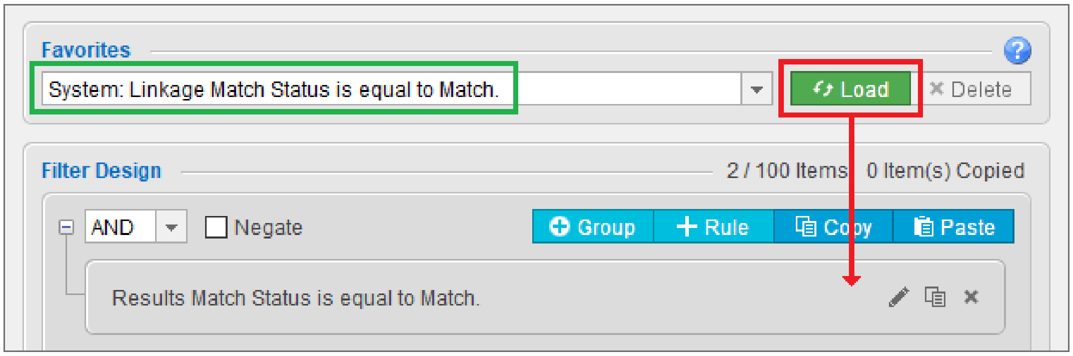
20. Press the **BACK TO RESULTS** button, which is located at the top of the frequency report screen, to return to the linkage results screen.



21. Press the **APPLY FILTER** button, which is located at the top of the linkage results screen. The Apply Filter dialog will appear.



a. From the drop-down containing the list of **Favorites**, select the filter labeled "**System: Linkage Match Status is equal to Match**", then press the **LOAD** button. The filter criteria will be displayed.


Favorites

System: Linkage Match Status is equal to Match.

Filter Design

2 / 100 Items 0 Item(s) Copied

AND ☐ Negate

Results Match Status is equal to Match.

b. Press the **OK** button, which is located in the bottom-right corner of the dialog. The Apply Filter dialog will close. At this point you will only be looking at the confirmed duplicate patients.

22. Return to your database and resolve/consolidate all of the duplicate patients that were identified during the linkage process.

23. **CONGRATULATIONS!!!** You've finished deduplicating your patients for the NAACCR submission.