

NAACCR Tumor Deduplication Instructions

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

1. To get started, you'll need to download and install Match*Pro version 2.3. 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 the fields listed below for ALL records of eligible primary tumors diagnosed from 2007-2021. 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. Tumor Record Number (#60)
 - c. Sequence Number (#380)
 - d. Primary Site (#400)
 - e. Laterality (#410)
 - f. Histology ICD-O-2 (#420)
 - g. Behavior Code ICD-O-2 (#430)
 - h. Histology ICD-O-3 (#522)
 - i. Behavior Code ICD-O-3 (#523)
 - j. Date of Diagnosis (#390)
 - k. Age at Diagnosis (#230)
 - l. Type of Reporting Source (#500)
 - m. Override Site/Lat/Seq No (#2010, AKA the Inter-record Edit 09 review flag)
3. Once you've created the extract you are ready to begin the process of deduplicating the tumors in your database. A linkage configuration file named [naaccr-tumor-dedup-20230303.mplc](#) was included with these instructions for this purpose.

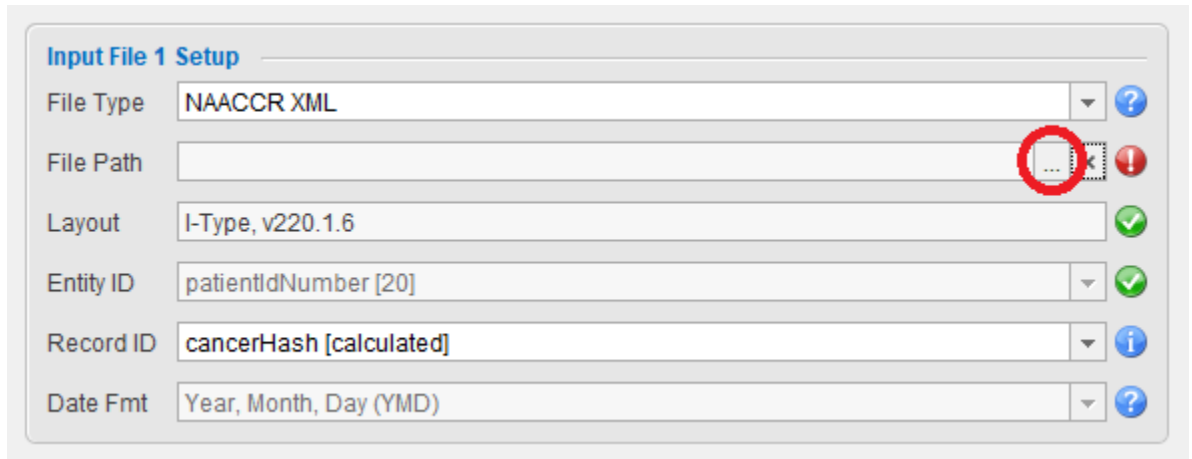
Extract the configuration file from the zip folder, then double-click on it. It should open automatically with Match*Pro.

- 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

NAACCR Tumor Deduplication Instructions

default.

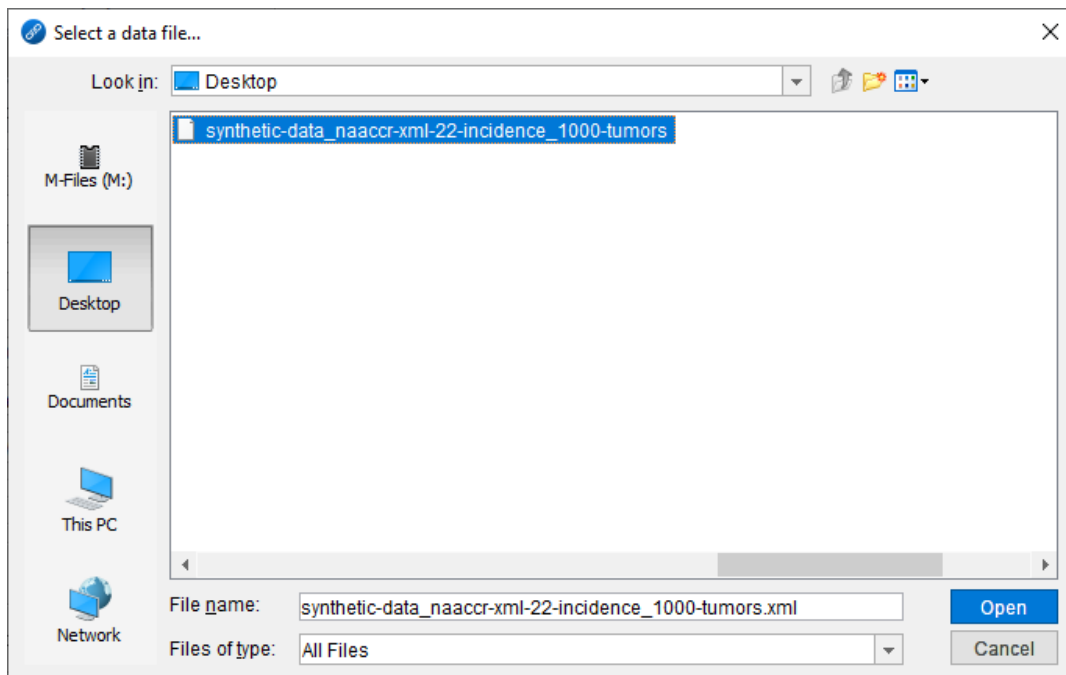
- a. Press the browse button associated with the **File Path** for **File 1**, which has been circled in **RED** in the image below.



The 'Input File 1 Setup' dialog box contains the following fields and controls:

- File Type:** NAACCR XML (dropdown menu with a help icon)
- File Path:** An empty text field with a browse button (three dots in a square) circled in red and a warning icon.
- Layout:** I-Type, v220.1.6 (dropdown menu with a checkmark icon)
- Entity ID:** patientIdNumber [20] (dropdown menu with a checkmark icon)
- Record ID:** cancerHash [calculated] (dropdown menu with an information icon)
- Date Fmt:** Year, Month, Day (YMD) (dropdown menu with a help icon)

- 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.



- c. The NAACCR XML File Setup dialog will be displayed.

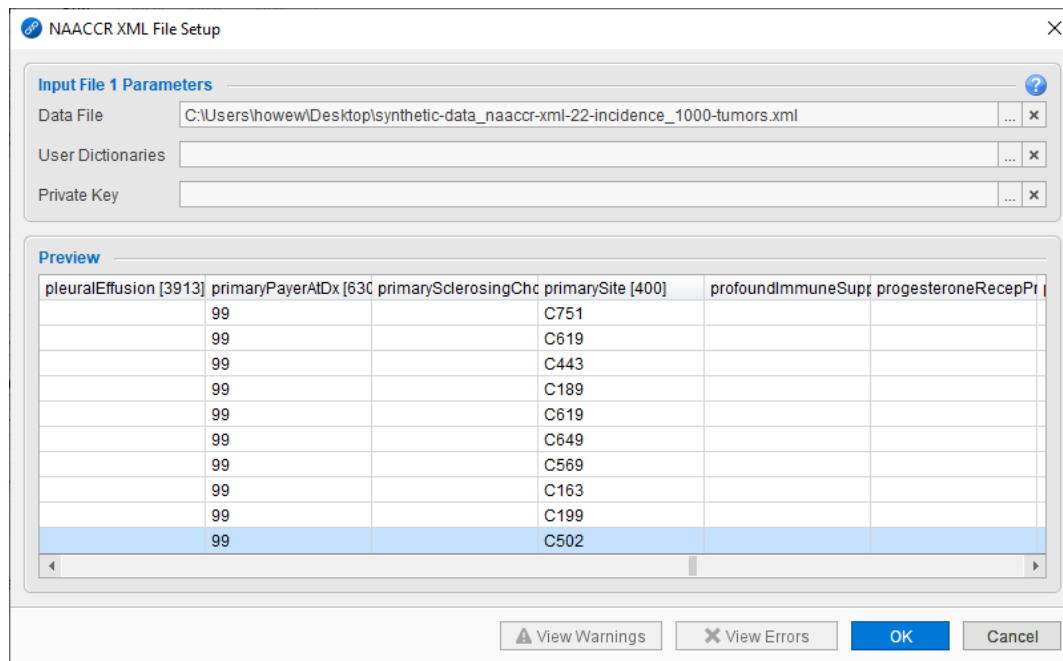
If the file you selected references a user-dictionary then you will receive a warning that a

NAACCR Tumor Deduplication Instructions

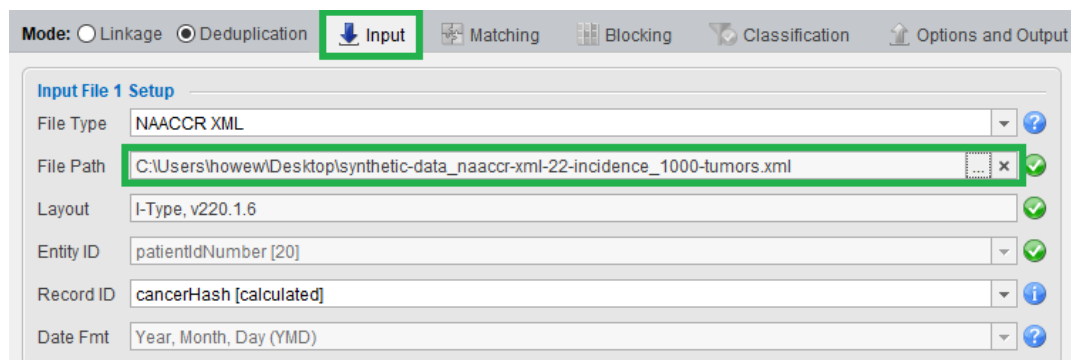
user-dictionary has not been provided (because you haven't provided it yet).

YOU CAN IGNORE THIS WARNING since there aren't any user-defined fields being used in this process or you can provide the dictionary in question (it's probably easier to just ignore the warning).

You can use the preview window to verify that all the fields have been populated. Once you're convinced that all 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.



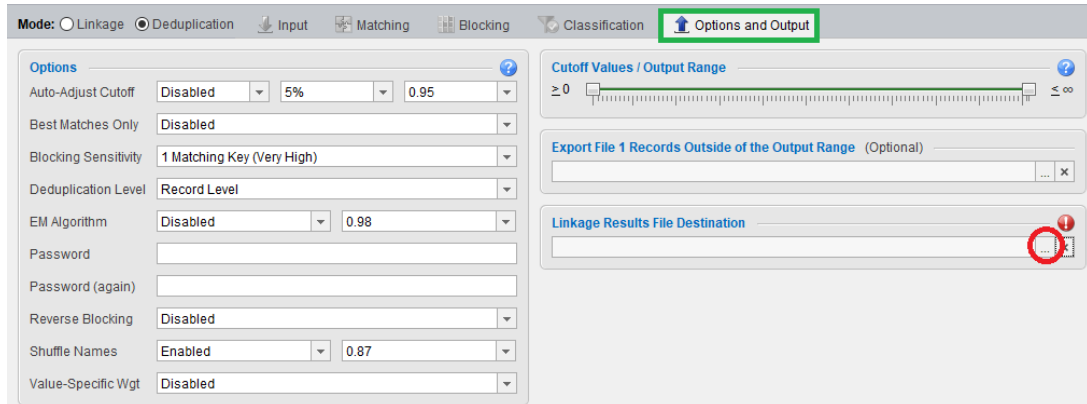
- d. The name and location of the extract will be displayed in the text box.



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.

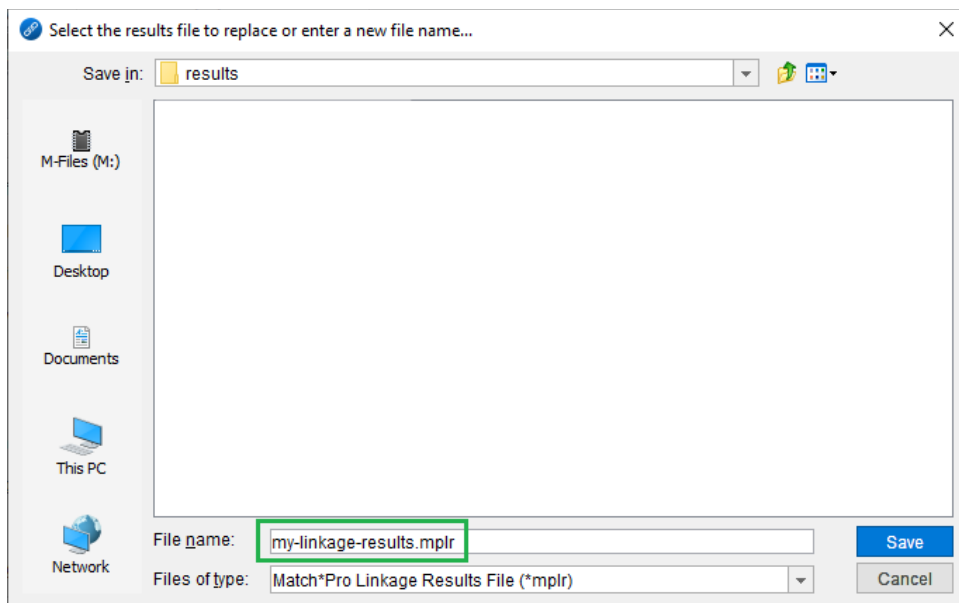
NAACCR Tumor Deduplication Instructions

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



The screenshot shows the 'Options and Output' tab of the NAACCR Tumor Deduplication software. The 'Linkage Results File Destination' field is visible, and a red circle highlights the browse button (folder icon) next to it. Other options like 'Auto-Adjust Cutoff', 'Best Matches Only', and 'Blocking Sensitivity' are also visible.

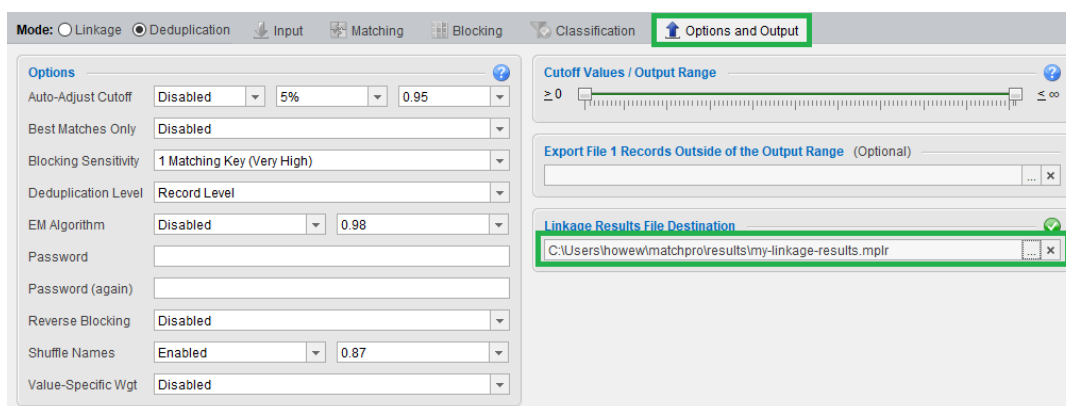
- b. A save dialog will appear. Browse to the location where you'd like the results file to reside, then enter a filename and press the **SAVE** button. **PLEASE BE ADVISED THAT THE LINKAGE RESULTS FILE SHOULD BE CREATED ON YOUR C:/ DRIVE AS OPPOSED TO A NETWORK DRIVE AS SLOW AND/OR DROPPED NETWORK CONNECTIONS 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.**



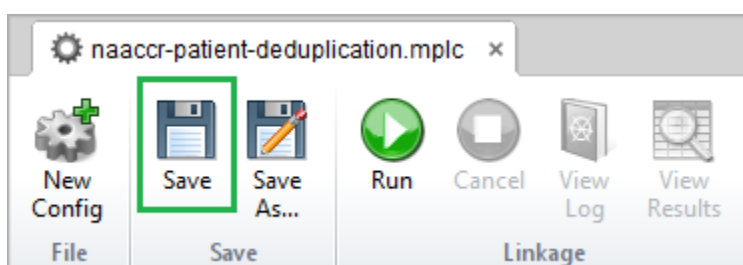
The screenshot shows a Windows 'Save As' dialog box. The 'Save in' field is set to 'results'. The 'File name' field contains 'my-linkage-results.mplr', which is highlighted with a green box. The 'Files of type' dropdown is set to 'Match*Pro Linkage Results File (*.mplr)'. The 'Save' button is visible.

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

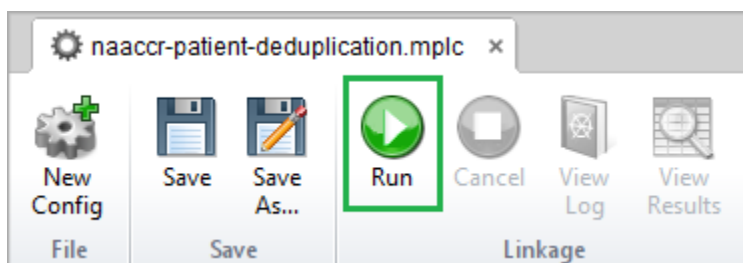
NAACCR Tumor Deduplication Instructions



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

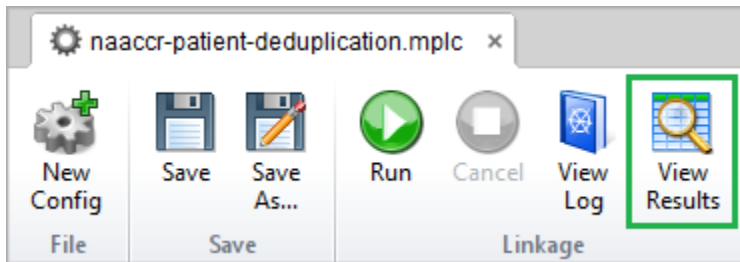


7. Press the **RUN** button. The linkage process will begin. The run time will vary depending on the number of records that are in the extract.



NAACCR Tumor Deduplication Instructions

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. Match*Pro uses an external library to determine if two tumors represent a single primary. If the library determines that two tumors represent the same primary, then those two tumors will appear on the results screen. For more information about the Java library and the Solid Tumor and MPH rulesets it implements, see:

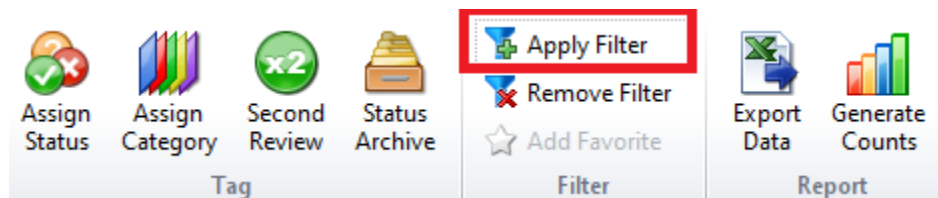
- a. <https://github.com/imsweb/mph>
- b. <https://seer.cancer.gov/tools/solidtumor/>
- c. <https://seer.cancer.gov/tools/mphrules/>

10. Write down the number of pairs that appear on the results screen. If there aren't any pairs on the results screen (unlikely, but possible) then the number is ZERO and you are finished.

If there are pairs on the result screen, continue to step 11.

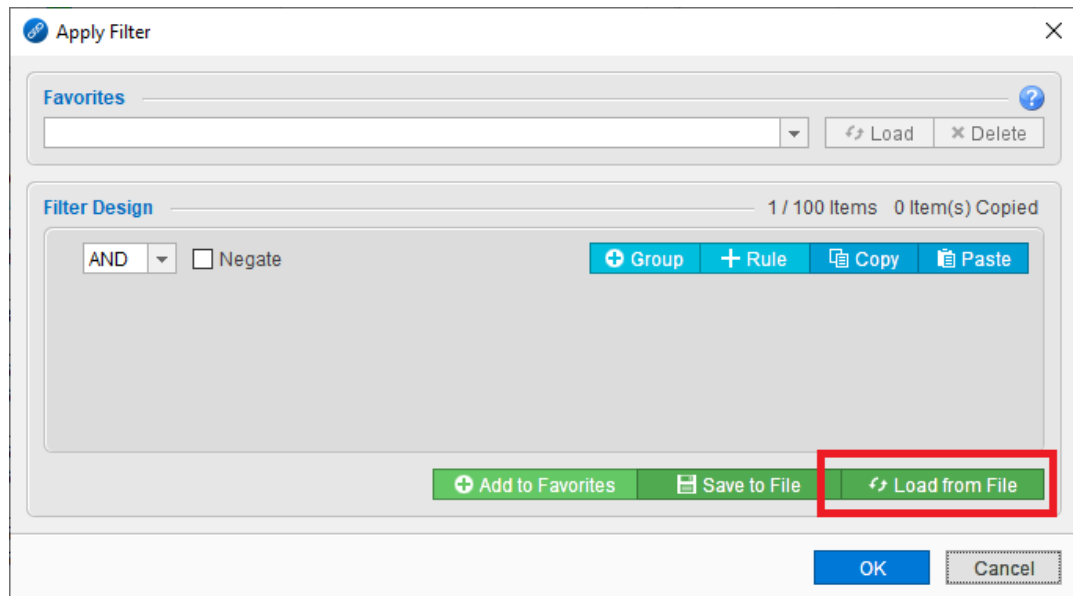
11. Next, you will need to determine how many of the pairs reference a case diagnosed between 2017 and 2021. A filter definition file named **filter-on-2017-2021-dx.mplf** was included in the zip file for this purpose.

- a. Press the **APPLY FILTER** button. The Apply Filter dialog will be displayed.

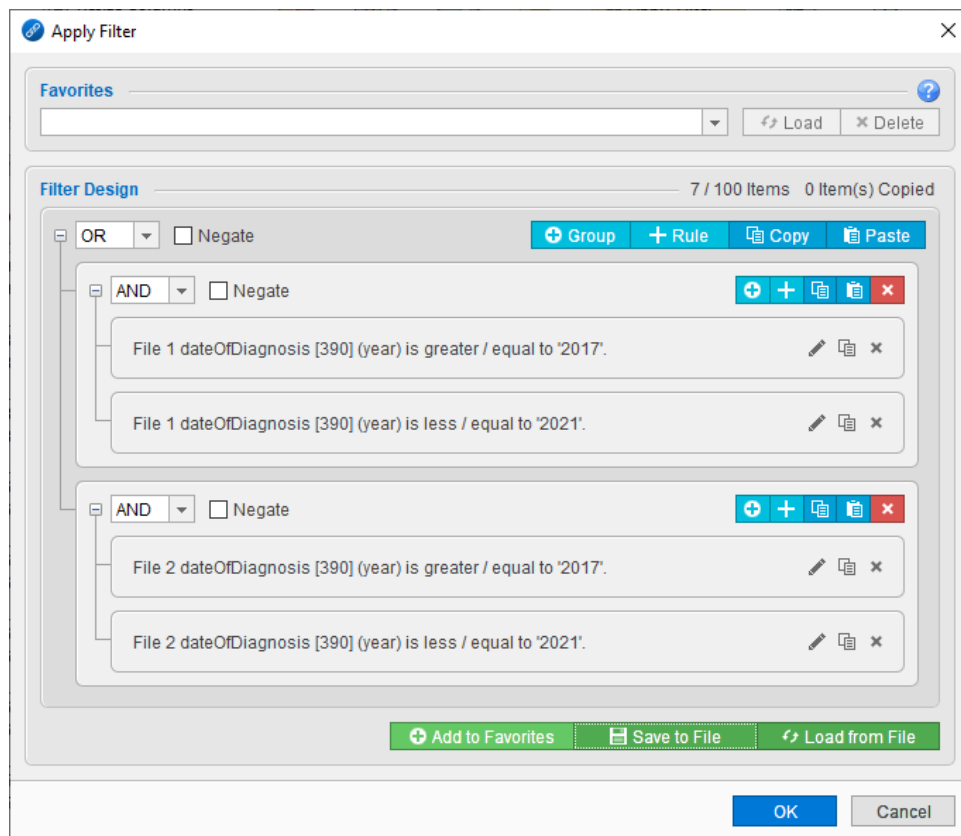


NAACCR Tumor Deduplication Instructions

- b. Press the Load From File button. This can be found in the lower-right corner of the dialog.

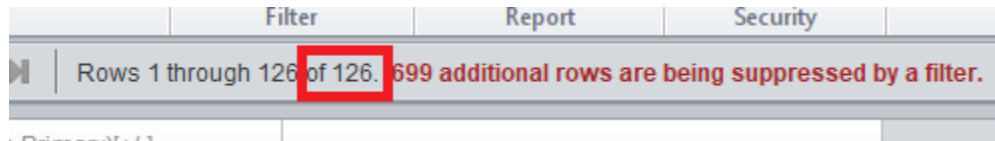


- c. A file selection dialog will be displayed. Select the file named **filter-on-2017-2021-dx.mplf** then press the **OPEN** button. The filter will be displayed.

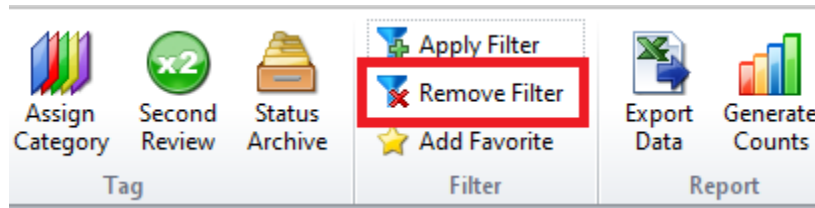


NAACCR Tumor Deduplication Instructions

- d. Press the **OK** button. The dialog will close, and you'll be taken back to the results screen. Write down the number of records that remain after applying the filter. This information appears above the table.



- e. Press the **REMOVE FILTER** button.



12. Repeat step 11, but this time select the file named **filter-on-2021-dx.mplf** to determine the number of pairs that reference a case diagnosed in 2021.
13. At this point you should have 3 counts:
- a count for the total number of pairs (obtained in step 10)
 - a count for the number of pairs that reference cases diagnosed between 2017-2021 (obtained in step 11)
 - a count for the number of pairs that reference cases diagnosed in 2021 (obtained in step 12)
14. Please be advised that some of the cases that are flagged as single primaries by the library will NOT be true duplicates. This could be due to issues with the library or other facts regarding to the two cases. We estimate that roughly 5-10% of the results fall under this category. You will need to use any/all information at your disposal (abstracts, etc.) to determine whether you agree with the library or not.

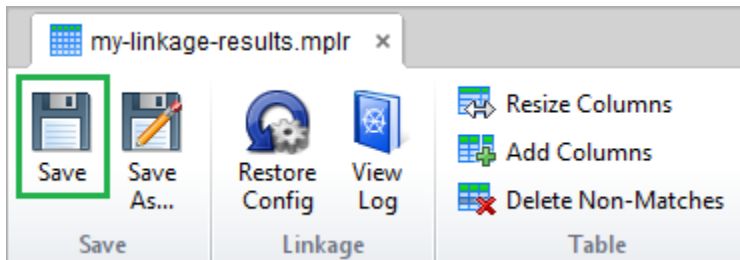
Starting with the 2021 cases and working backwards through 2017 and earlier, determine whether you agree with Match*Pro regarding whether the two cases in each pair are single primaries.

If you agree with Match*Pro, mark the pair a match. If you disagree, mark the pair a non-match. You can do this by clicking on the check box or the 'X' in the margin of each row.



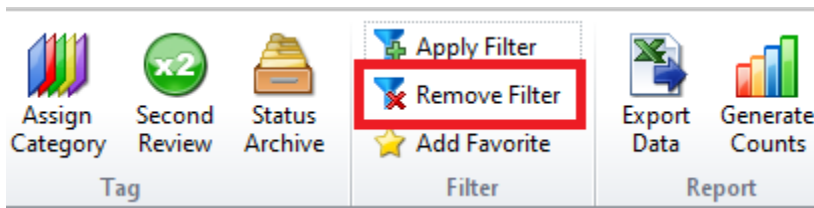
NAACCR Tumor Deduplication Instructions

15. When you are finished with your review, make note of the following:
- the number of pairs you reviewed
 - the number of pairs you accepted as single primaries and resolved in your database
 - how far back you went with your review (i.e., did you review 2021 cases, cases going back to 2017, or cases going back to before 2017?)
16. **PRESS THE SAVE BUTTON** at the top of the screen to lock in all the decisions you made during the manual review.

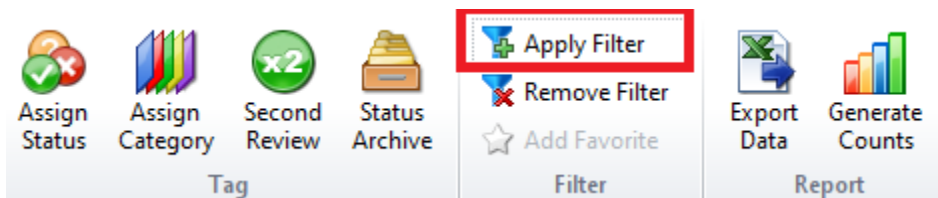


The remaining steps explain how to create a match status archive file that contains the list of non-matches. You can use this file in future linkages (such as next year's tumor deduplication) to save time on the manual review.

17. Press the **REMOVE FILTER** button to clear any filters that might be in place.

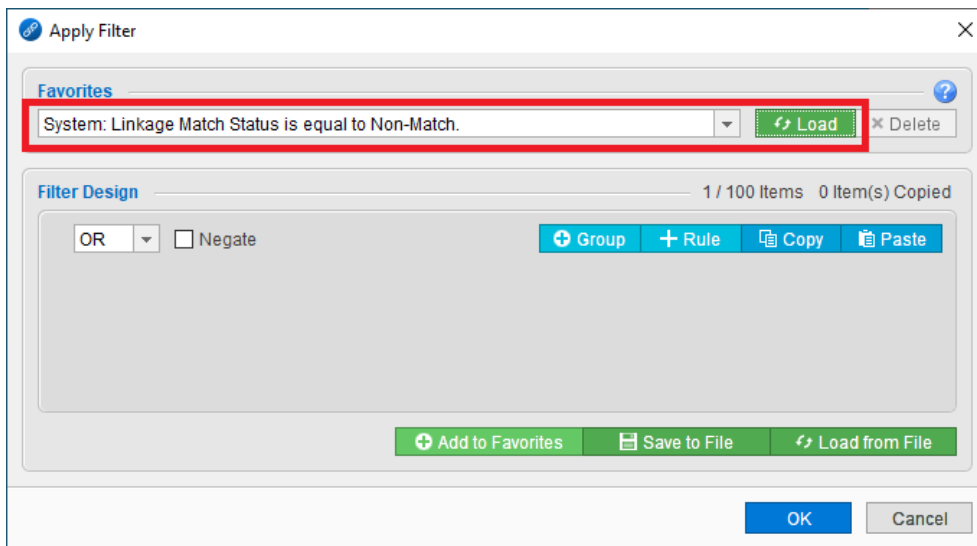


18. Press the **APPLY FILTER** button. The Apply Filter dialog will appear.

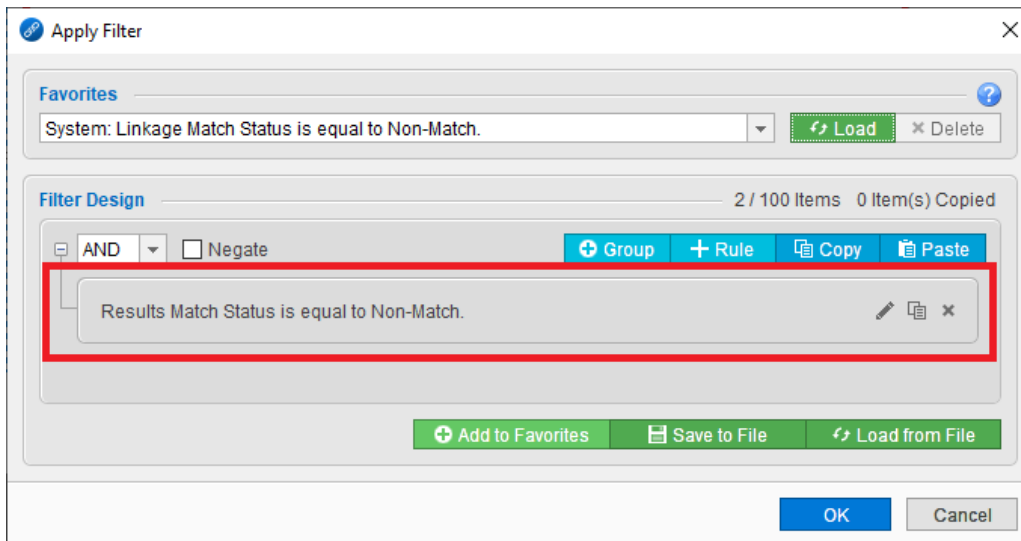


NAACCR Tumor Deduplication Instructions

19. Select “**System: Linkage Match Status is equal to Non-Match**” from the drop down at the top of the dialog, then press the **LOAD** button.

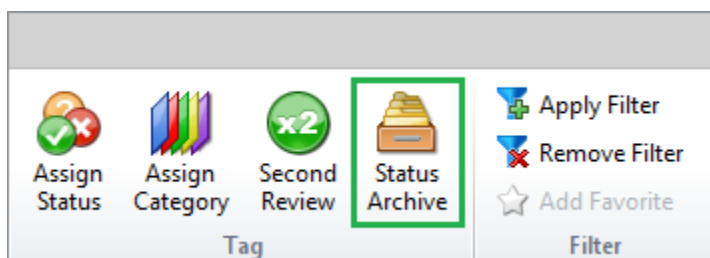


20. The filter criteria will be displayed. Press the **OK** button. The Apply Filter dialog will close. When you return to the results screen you will only see non-matches.

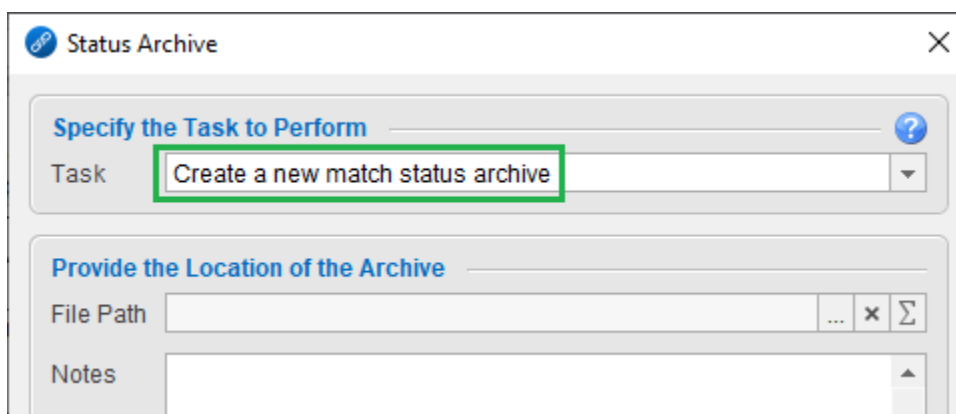


NAACCR Tumor Deduplication Instructions

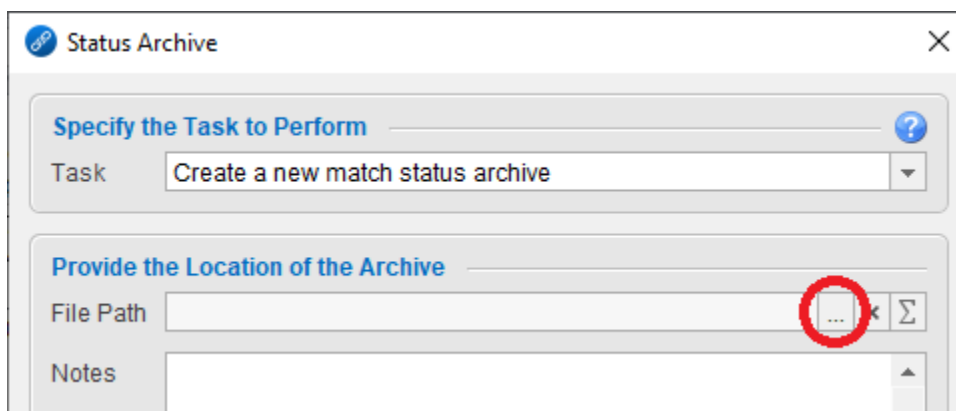
21. Press the **STATUS ARCHIVE** button. The Status Archive dialog will be displayed.



22. Select “Create a new match status archive” from the drop down.

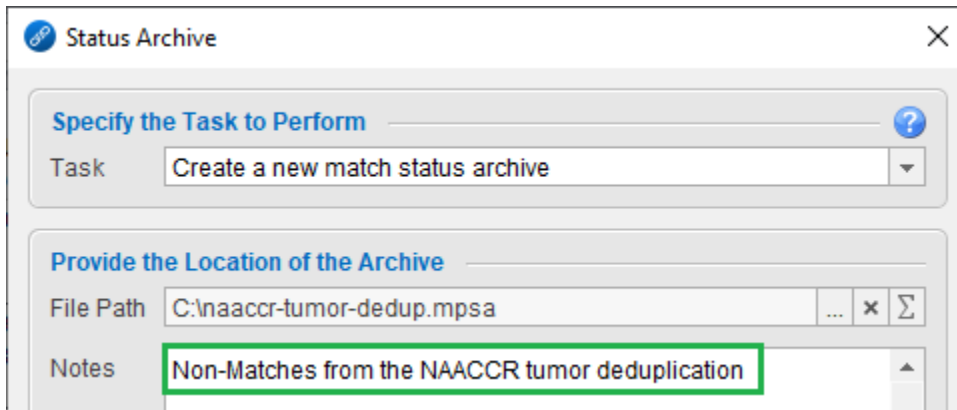


23. Provide the location of where you’d like to save the archive. The button you need to press to do this is circled in **RED** in the image below.



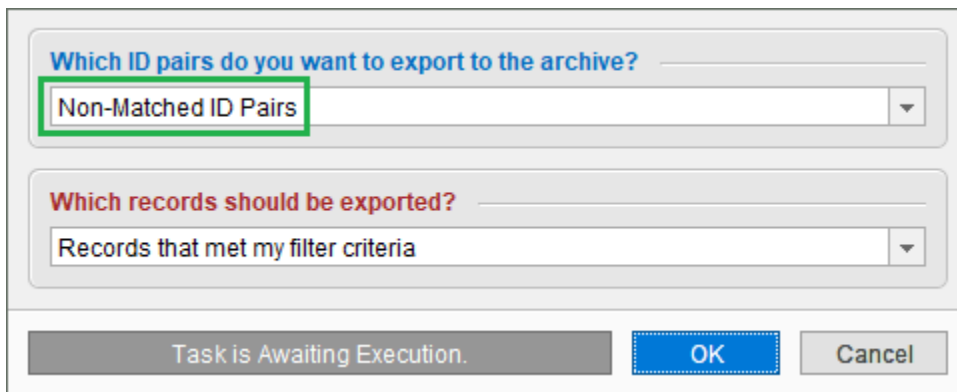
NAACCR Tumor Deduplication Instructions

24. Enter a description of the archive in the notes section (e.g., “Non-Matches from the NAACCR tumor deduplication”).



The screenshot shows a dialog box titled "Status Archive". It has two main sections. The first section, "Specify the Task to Perform", contains a dropdown menu labeled "Task" with the selected option "Create a new match status archive". The second section, "Provide the Location of the Archive", contains a text field for "File Path" with the value "C:\naaccr-tumor-dedup.mpsa" and a "Notes" text area containing the text "Non-Matches from the NAACCR tumor deduplication". The "Notes" text area is highlighted with a green rectangular box.

25. Select “Non-Matched Pairs” from the drop down towards the lower half of the dialog.



The screenshot shows the lower half of the "Status Archive" dialog box. It contains two dropdown menus. The first dropdown, titled "Which ID pairs do you want to export to the archive?", has the option "Non-Matched ID Pairs" selected and is highlighted with a green rectangular box. The second dropdown, titled "Which records should be exported?", has the option "Records that met my filter criteria" selected. At the bottom of the dialog, there is a status bar that says "Task is Awaiting Execution.", followed by "OK" and "Cancel" buttons.

26. Press the **OK** button. The status archive will be created in the location you specified in step 23.
Make sure to save this file for next year.

Close the dialog to return to the manual review screen.

27. **CONGRATULATIONS!!!** You are finished.