**NAACCR 2023 Call for Data – 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. **The steps outlined in this document should only be performed AFTER you've successfully deduplicated the patients in your database using Match*Pro and consolidated all of the duplicate patients in your database.**

1. To get started, you will need to download and install Match*Pro version 2.4.2. The software can be downloaded from https://seer.cancer.gov/tools/matchpro/. **You should use this specific version of the software (or a later version if it is available) to ensure that the library being used by the software to match the tumors is up to date.**

2. After you have 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 23) 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)

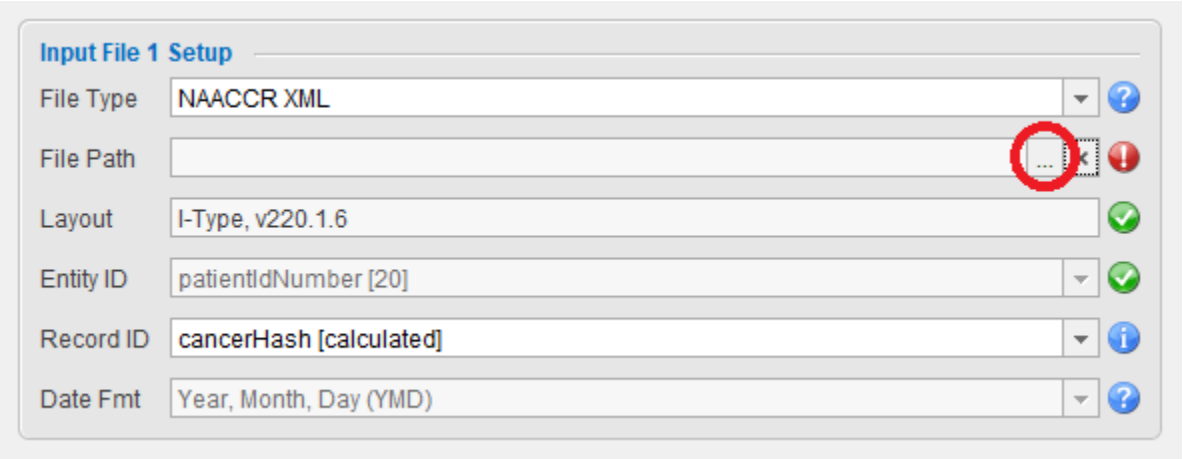
3. Once you have created the extract you are ready to begin the process of deduplicating the tumors in your database. A linkage configuration file named **naaccr-tumor-dedup-20230303.mplc** was included with these instructions for this purpose.

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

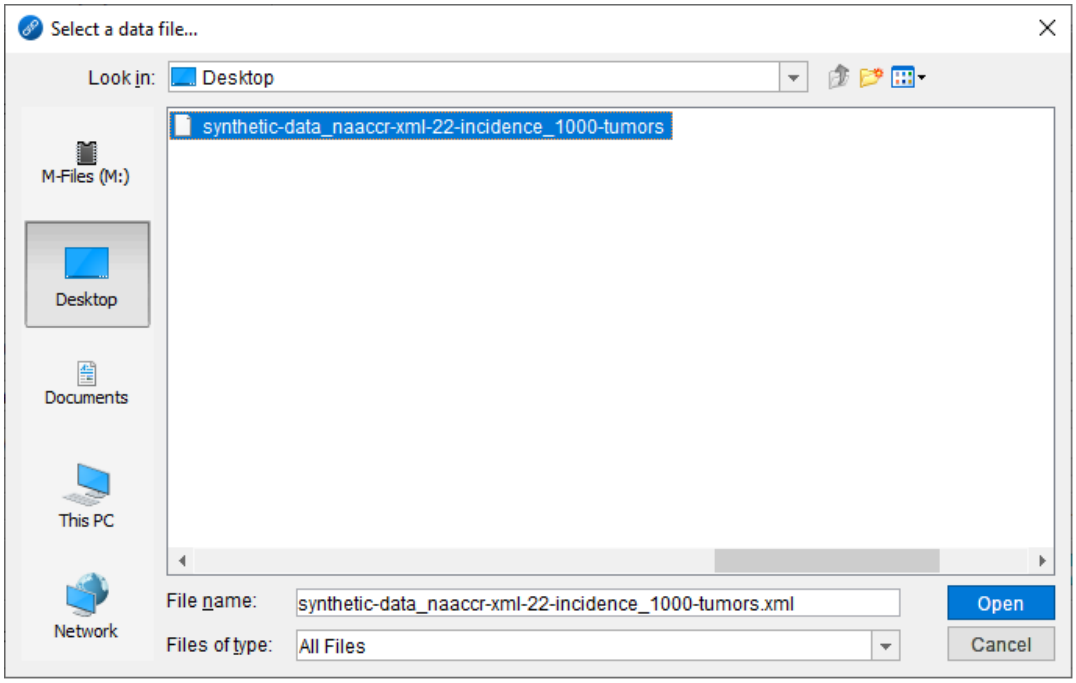
    a. If, for some reason, the file does not open automatically then you will need to manually open the file. To do this you will need to start the Match*Pro software (a shortcut for which should have been created on your desktop during the installation process). Once the software is running, click on the File menu and select "Open Linkage Configuration …" from the list of options (this is the second 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 will need to provide Match*Pro with the location of the extract you created in step 2. There are five tabs on the linkage configuration screen. The first tab, 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.




    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.

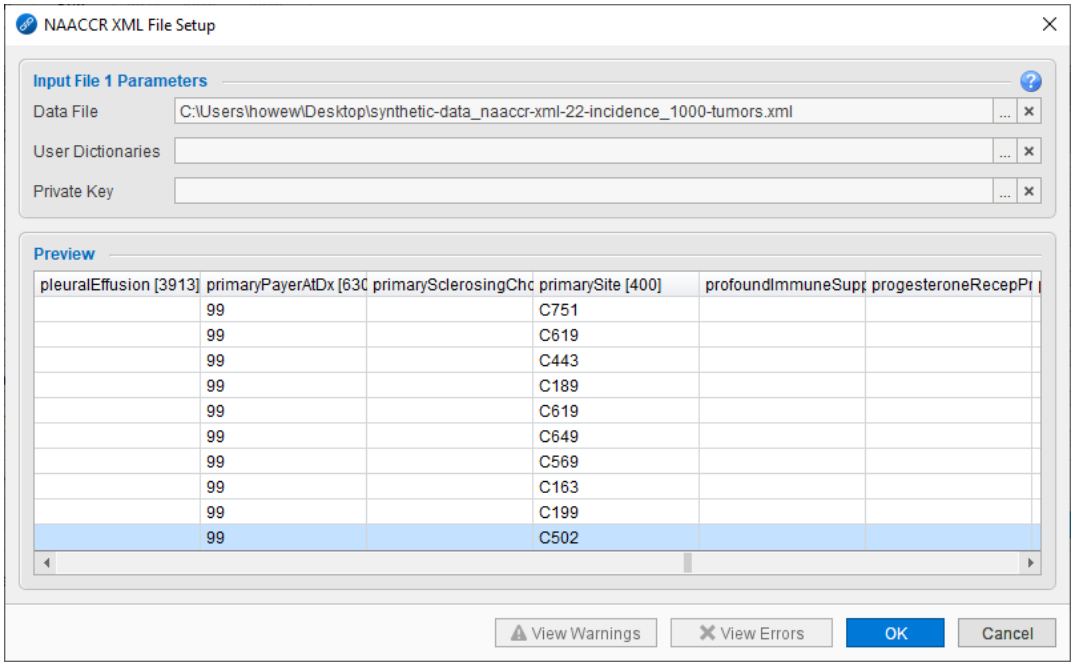


c. The NAACCR XML File Setup dialog will be displayed.

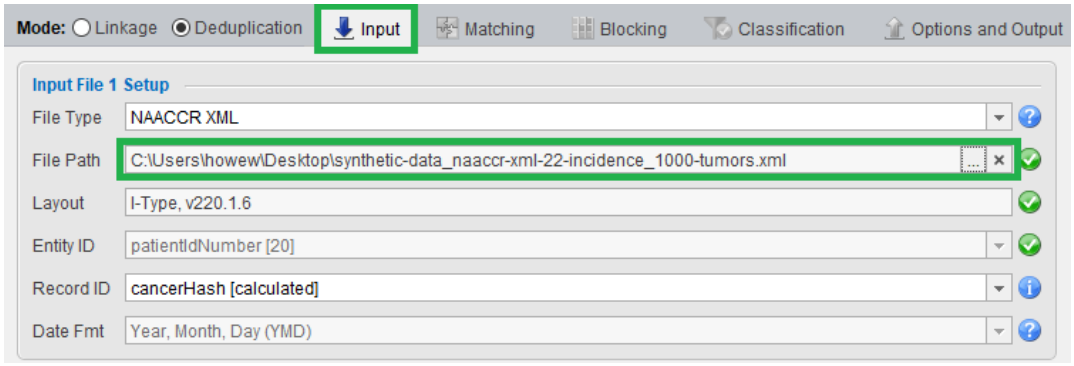
If the file you selected references a user-dictionary then you will receive a warning that a user-dictionary has not been provided (because you have not provided it yet).

**YOU CAN IGNORE THIS WARNING** since there are not any user-defined fields being used in this process or you can provide the dictionary in question (it is easier to just ignore the warning).

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

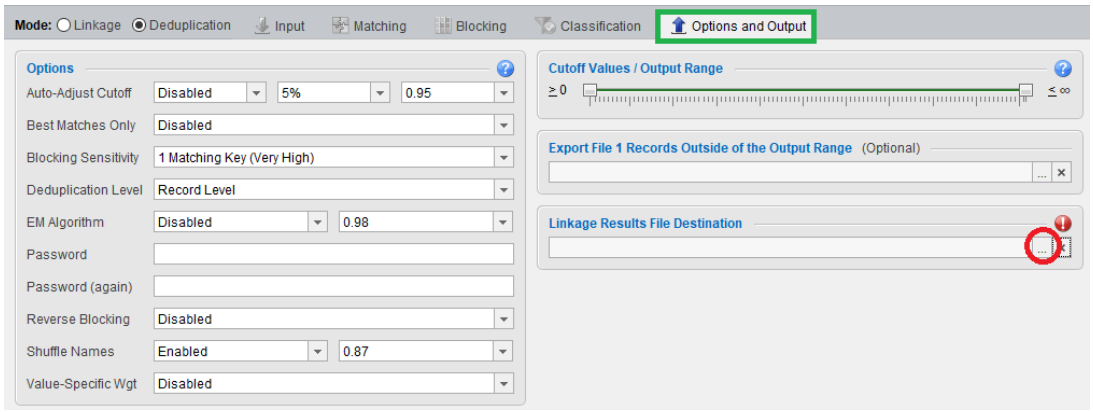



d. The name and location of the extract will be displayed in the text box.

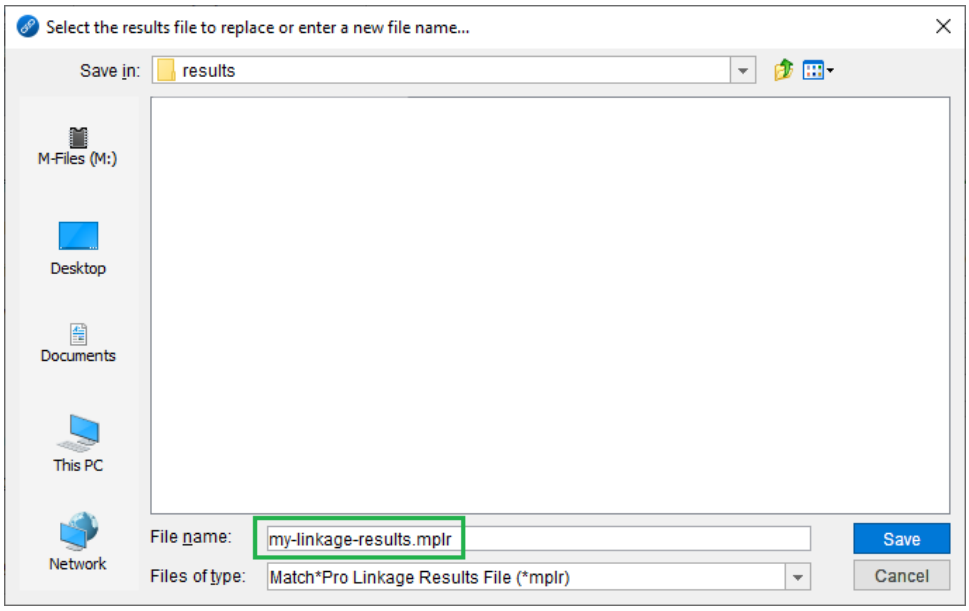


5. You are now finished with the input tab. Switch to the **Options and Output** tab. This is the fifth and final tab that is displayed on the linkage configuration screen. Here you will need to provide Match*Pro with the location of where you would like the linkage results file to be created.

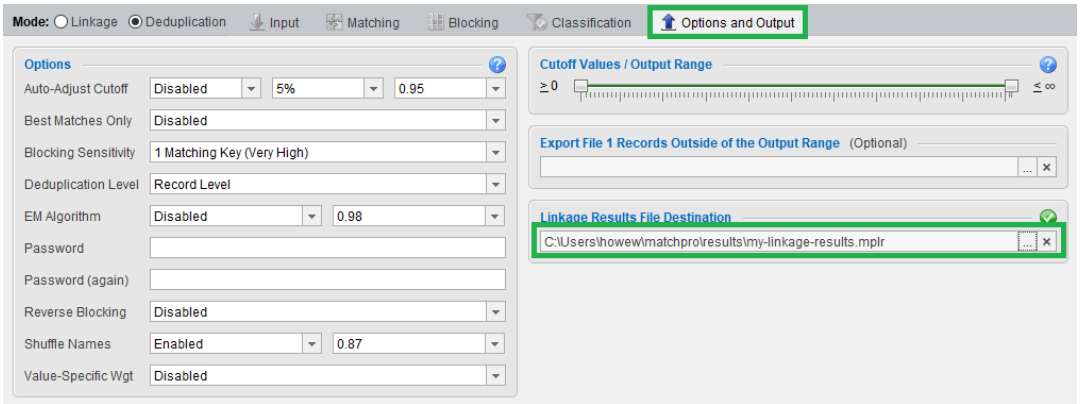
    a. Press the browse button associated with the **Linkage Results File Destination**, which has been circled in **RED** in the image below.



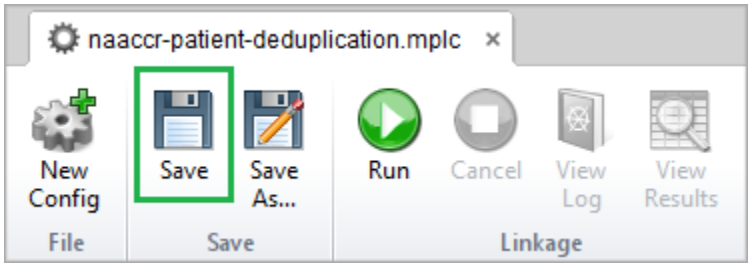

    b. A save dialog will appear. Browse to the location of where you would like the results file to reside, then enter a filename and press the **SAVE** button. **PLEASE BE ADVISED THAT THE LINKAGE RESULTS FILE SHOULD BE CREATED IN A FOLDER ON YOUR C:/ DRIVE AS OPPOSED TO A NETWORK DRIVE AS SLOW AND/OR DROPPED NETWORK CONNECTIONS CAN CORRUPT THE FILE (PARTICULARLY IF IT IS LARGE). THE RESULTS FILE SHOULD REMAIN IN THE FOLDER ON YOUR C:/ DRIVE UNTIL ALL OF THE WORK OUTLINED IN THIS DOCUMENT HAS BEEN COMPLETED. DO NOT SAVE DIRECTLY TO C:/ - YOU WILL LIKELY GET A PERMISSIONS ERROR IF YOU DO. YOU MUST SPECIFY A FOLDER.**



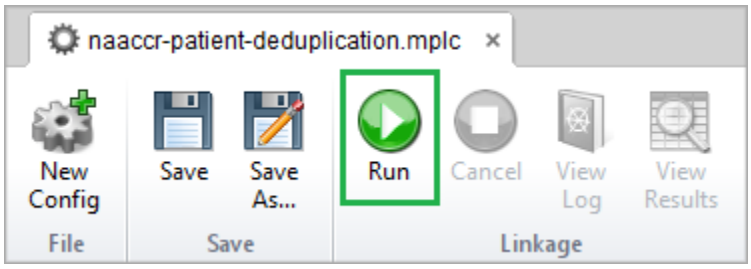
c. The name and future location of the results file will be displayed in the text box.



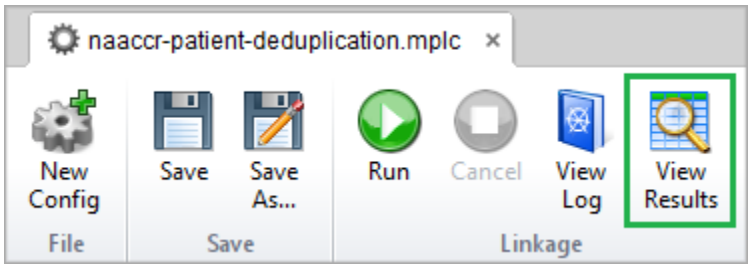

6. Press the **SAVE** button, which is located at the top of the linkage configuration screen, to save the changes that you have made to the configuration file.




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. 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:

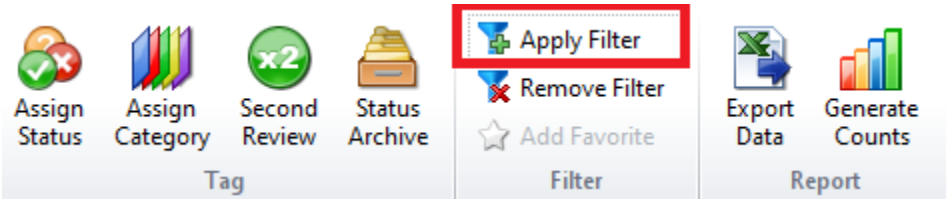
    a. https://github.com/imsweb/mph
    b. https://seer.cancer.gov/tools/solidtumor/
    c. https://seer.cancer.gov/tools/mphrules/

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

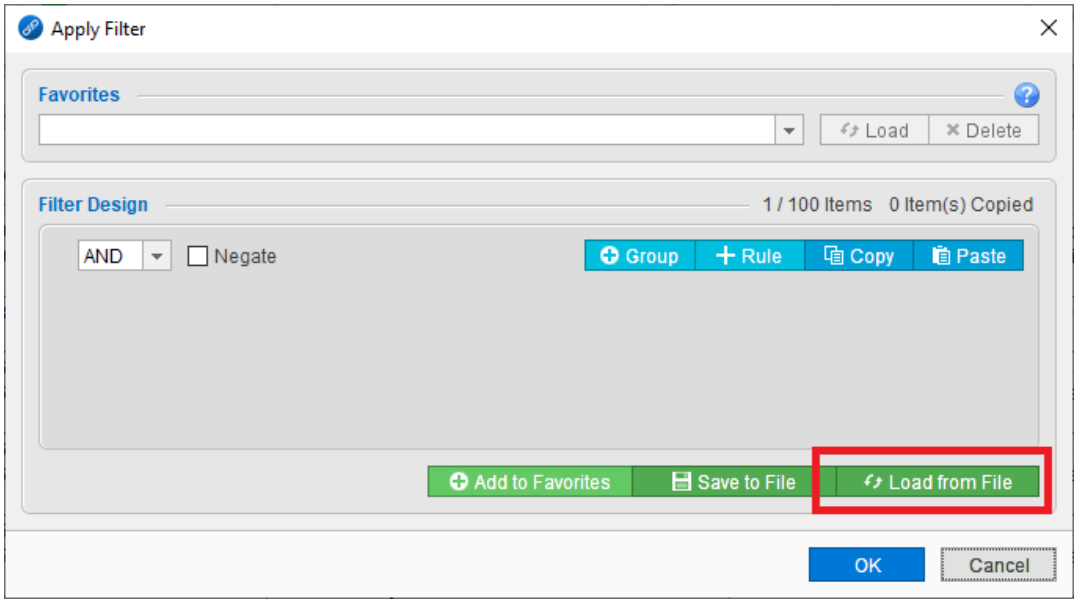
    If there are pairs on the result screen, continue to step 11.

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.

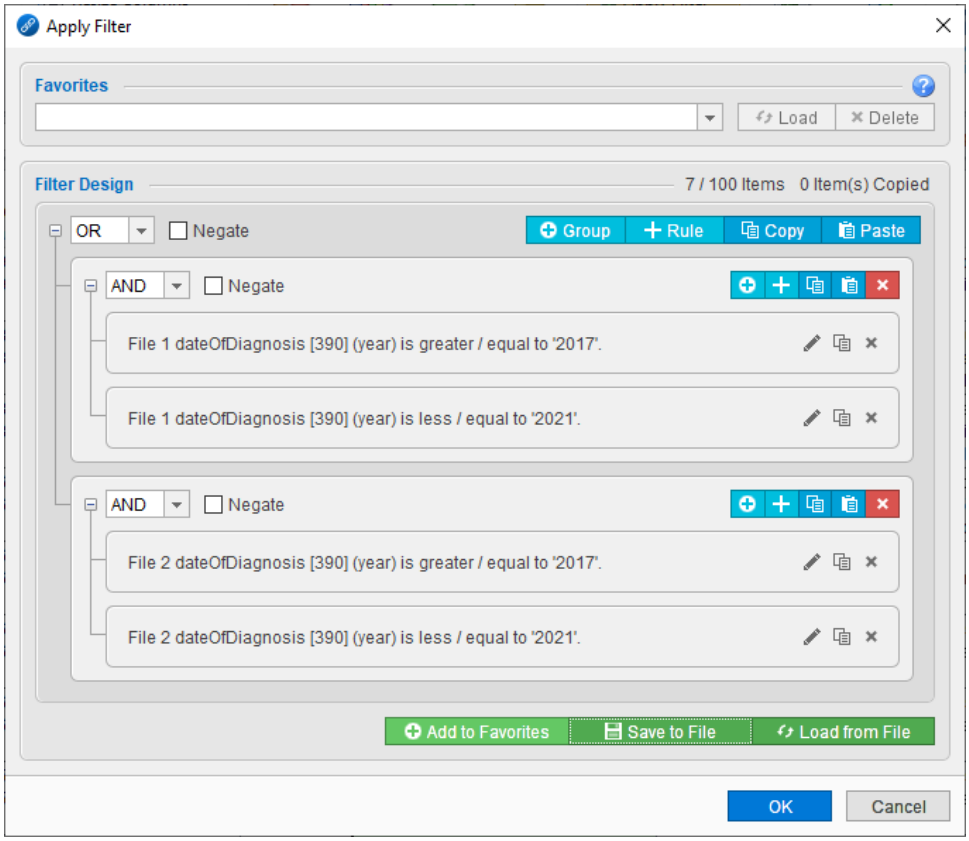
    a. Press the **APPLY FILTER** button. The Apply Filter dialog will be displayed.



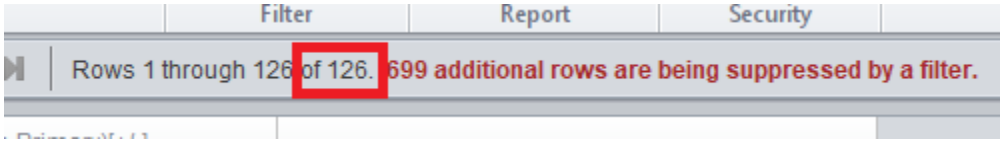
b. Press the Load From File button. This can be found in the lower-right corner of the dialog.



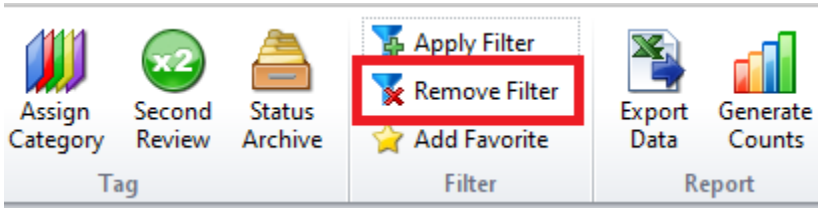

c. A file selection dialog will be displayed. Select the file named **filter-on-2017-2021-dx.mplf** then press the **OPEN** button. The filter will be displayed.



d. Press the **OK** button. The dialog will close, and you will 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 three counts:

    a. a count for the total number of pairs (obtained in step 10)
    b. a count for the number of pairs that reference cases diagnosed between 2017-2021 (obtained in step 11)
    c. a count for the number of pairs that reference cases diagnosed in 2021 (obtained in step 12)

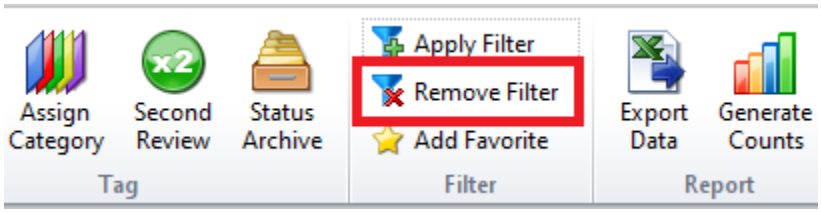
14. Please be advised that some of the cases that are flagged as single primaries may NOT be true duplicates. This could be due to issues with the Solid Tumor / MPH algorithms or other facts regarding to the two cases. We estimate that less than 5% of the results fall into this category. You will need to use any/all information at your disposal (abstracts, etc.) to determine whether you agree with the results.

    If you disagree with a result, you can submit a question using the Ask a SEER Registrar submission form. A SEER Registrar will review and determine if there is a problem with the algorithms or will provide an explanation for the result. Any problems with the Solid Tumor / MPH algorithms will be reported to the appropriate development teams by the SEER registrar, so there is no need to contact Match*Pro Support regarding these issues. The Ask a SEER Registrar submission page can be found here: https://seer.cancer.gov/registrars/contact.html

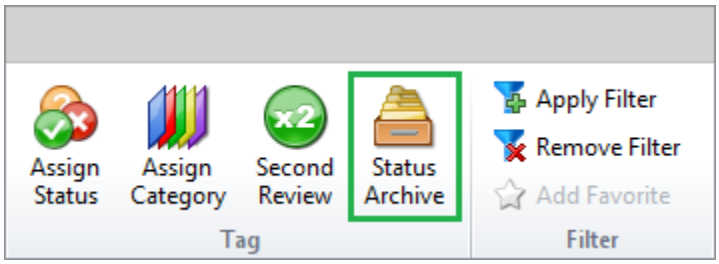
15. If you have never used Match*Pro to deduplicate tumors before or if you have, but you are no longer in possession of the match status archive from the previous run, SKIP to step 16. Otherwise…

    a. Locate your status archive file from last year. When you have found it, **MAKE A COPY** of the file for safe keeping just in case something goes wrong during this process.

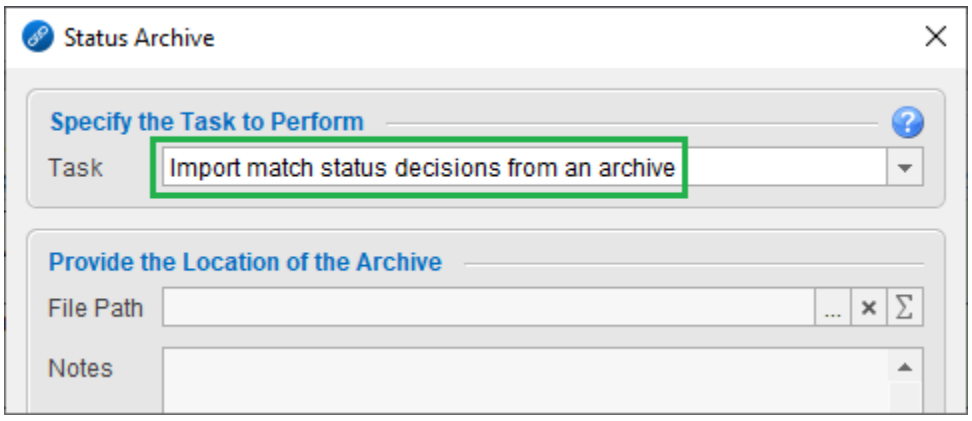
    b. Press the **REMOVE FILTER** button. **This step is very important for what we are about to do, so make sure that you press it at least once to be sure.**



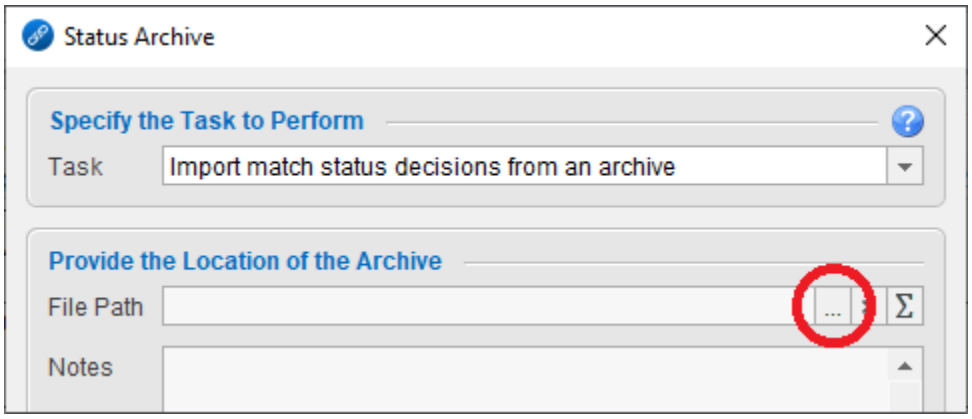

    c. Press the **STATUS ARCHIVE** button.



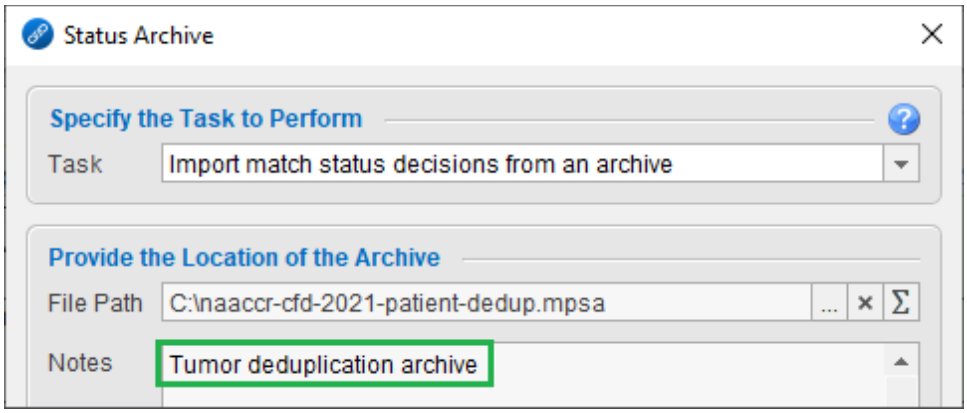

    d. Select "**Import match status decisions from an archive**" from the drop down.



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




f. After the file is selected the notes will be displayed. **Check the notes to confirm you selected the correct file before you proceed.**




g. 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 reviewed by staff from your registry at some point in the past. They are not matches and they do not need to be reviewed a second time. Any yellow/uncertain pairs that remain will need to be reviewed.

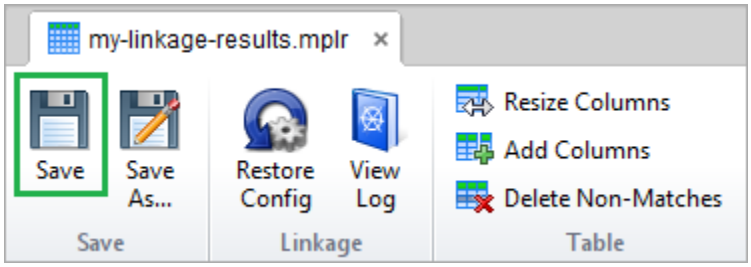
16. Starting with the 2021 cases and working backwards through 2017 and earlier, determine whether you agree with Match*Pro regarding whether the two cases in each pair are single primaries.

    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. **Any pairs you mark as a match should be consolidated in your database.**

17. When you are finished with your review, make note of the following:

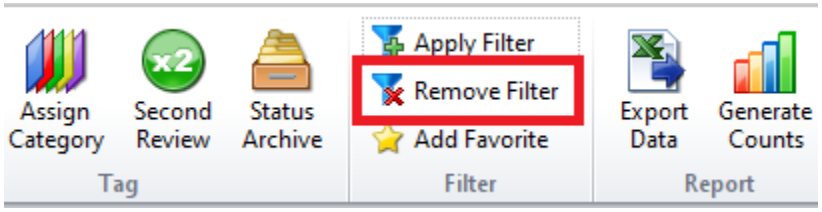
    a. the number of pairs you reviewed.
    b. the number of pairs you accepted as single primaries and resolved in your database.
    c. how far back you went with your review (i.e., did you review 2021 cases, cases going back to 2017, or cases going back to before 2017?)

18. **PRESS THE SAVE BUTTON** at the top of the screen to lock in all the decisions you made during the manual review process.

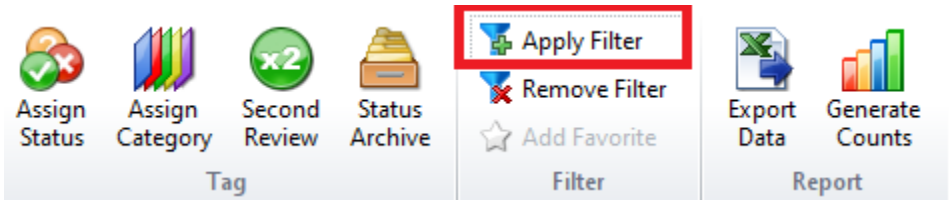


**The remaining steps explain how to create or update the match status archive file that contains the list of non-matches. You can use this file in future linkages (such as next year's tumor deduplication) to save time on the manual review.**

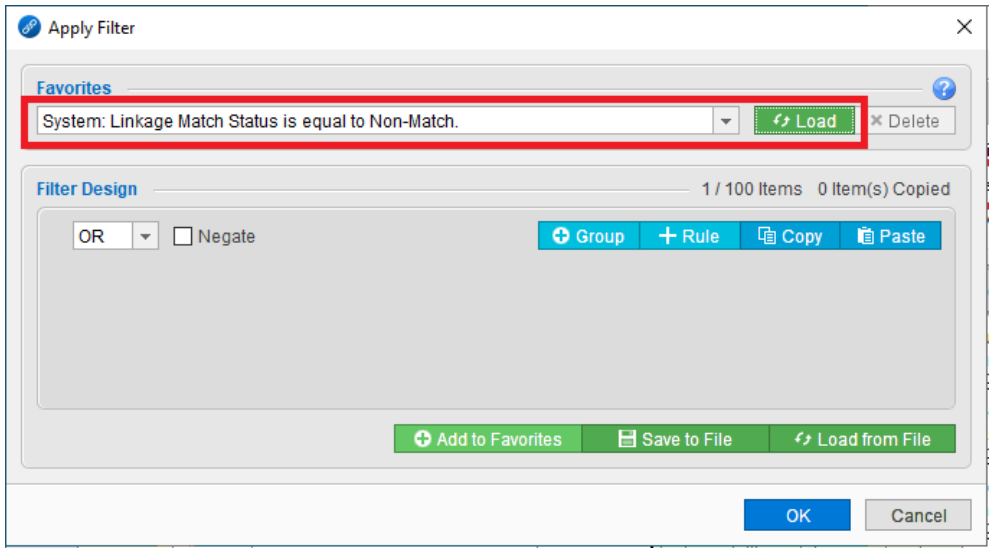
19. Press the **REMOVE FILTER** button to clear any filters that might be in place. **This step is very important for what we are about to do, so make sure that you press it at least once to be sure.**



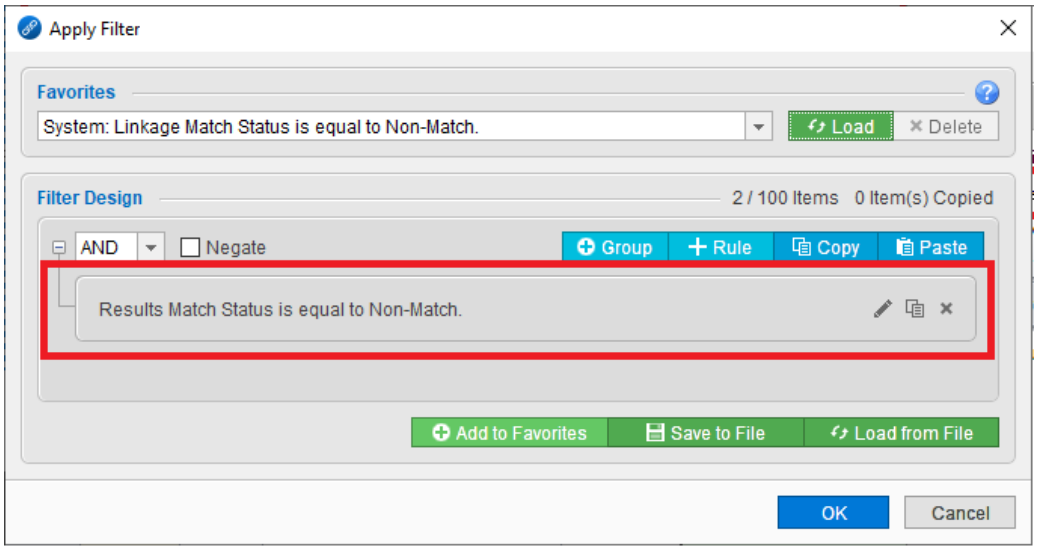

20. Press the **APPLY FILTER** button. The Apply Filter dialog will appear.



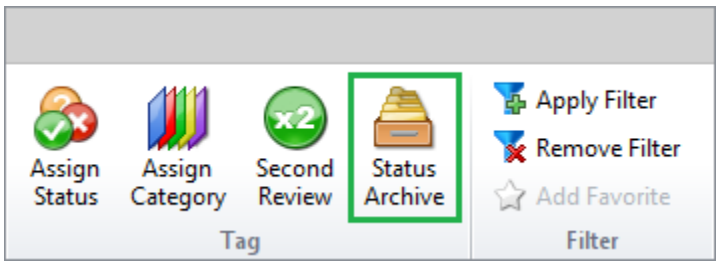

21. Select "**System: Linkage Match Status is equal to Non-Match**" from the drop down at the top of the dialog, then press the **LOAD** button.



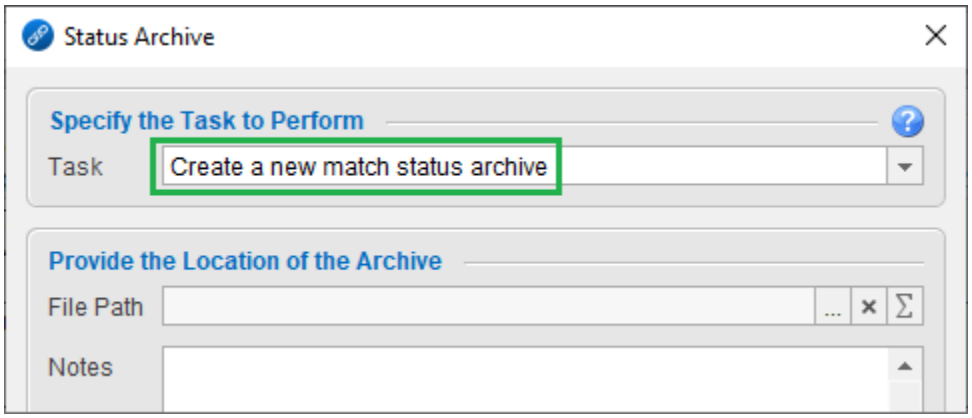
22. The filter criteria will be displayed. Press the **OK** button. The Apply Filter dialog will close. When you return to the results screen you will only see non-matches.



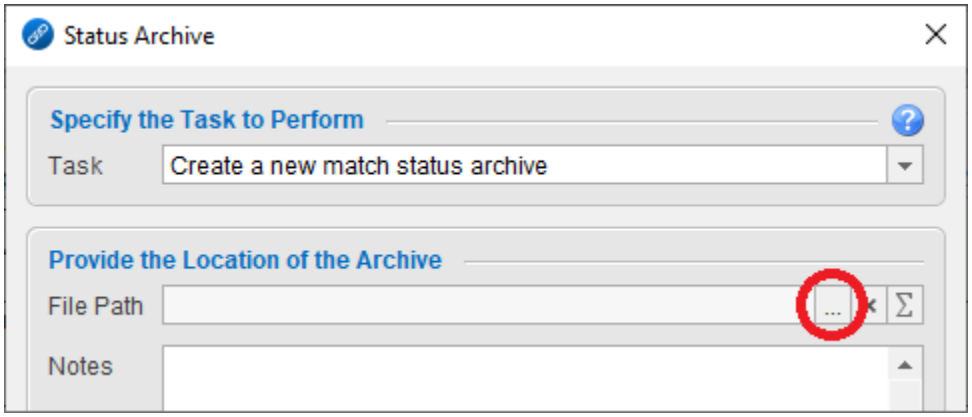

23. Press the **STATUS ARCHIVE** button. The Status Archive dialog will be displayed.



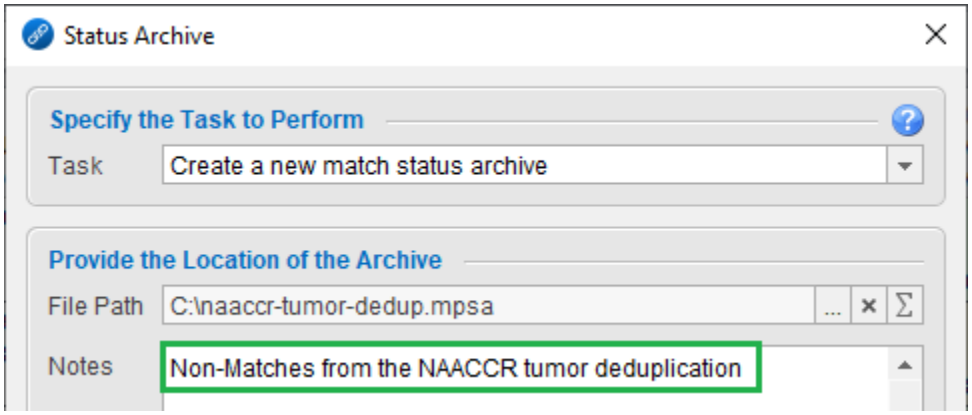

24. If you are creating a new archive, select "**Create a new match status archive**" from the drop down. If you are updating an existing archive, select "**Append data to an existing match status archive**."



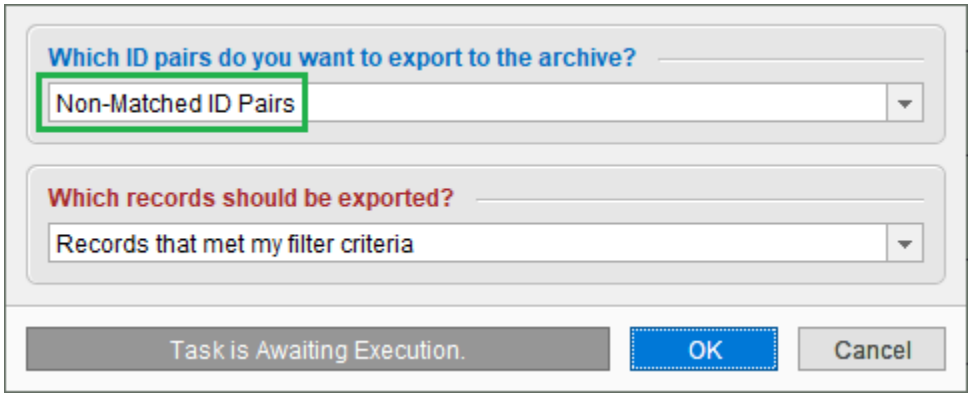
25. If you are creating a new archive, provide the location of where you would like to save the archive. If you are updating an existing archive, provide the location of the archive you want to update. The button you need to press to perform either action is circled in **RED** in the image below.




26. If you are creating a new archive, enter a description of the archive in the notes section (*e.g.,* "Non-Matches from the NAACCR tumor deduplication"). **If you are updating an existing archive, check the notes to make sure you selected the correct file** and, if so, make any changes to the notes that you think are necessary.




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



28. Press the **OK** button. The status archive will be created or updated.

    **Make sure to save this file for next year.**

    Close the dialog to return to the manual review screen.

29. **CONGRATULATIONS!!!** You are finished.