*Using Match*Pro to Deduplicate Patients & Tumors for the NAACCR Call for Data*

William Howe

Information Management Services, Inc.

August 28, 2023

NAACCR

IMS

Information Management Services, Inc.

Agenda

- About Match*Pro
- Where to find the Match*Pro software and the other files that are needed to perform the patient and tumor deduplication linkages
- Where to find training if you aren't familiar with Match*Pro
- Changes since last submission (patient deduplication) and earlier this year (tumor deduplication)
- Overview of the deduplication process
- Tips & Strategies for Manual Review

*About Match*Pro*

- Funded by the National Cancer Institute (NCI)
- Developed by Information Management Services, Inc. (IMS)
- Probabilistic record linkage software based on framework developed by Fellegi & Sunter (1969)

Links to Software & Configuration Files

- Links are on the NAACCR Call for Data "Tools" page
- https://www.naaccr.org/call-for-data/#datatools
- Scroll down to the "Deduplication" Heading
  - You will find a link to download a zip file containing…
    - Detailed, step-by-step, instructions for how to run the patient and tumor deduplication linkages
    - The Match*Pro linkage configuration files that are needed to perform those linkages
    - Match*Pro filter definition files to subset cases by DX year (used in tumor deduplication)
  - There is also a link to the landing page for the latest release of the Match*Pro software (version 2.4.2)
    - https://seer.cancer.gov/tools/matchpro/

*Match*Pro Training*

- Links are on the NAACCR Call for Data "Tools" page
- https://www.naaccr.org/call-for-data/#datatools
- Scroll down to the "Deduplication" Heading
  - You will find links to download recordings of two previous webinars that go over the processes for using Match*Pro to deduplicate patients and tumors in more detail than what will be provided today.
- For more general training regarding the usage of Match*Pro, NAACCR also provides recordings from an educational workshop conducted in June 2021. These recordings are based on an earlier version of Match*Pro, but most of the information presented during each of the two 4-hour sessions is still relevant.
- https://education.naaccr.org/products/matchpro-record-linkage-software

Patient Deduplication Changes

- Last year there were two versions of the configuration file: one for registries who have used Match*Pro to deduplicate their patients before and who are still in possession of their match status archive from last year and one for registries who have never used Match*Pro before or that have misplaced the archive. This year there is just one configuration file for both scenarios.

- Last year the linkage configuration file for registries with a status archive did not implement any match classification logic. This year, registries with a status archive will be able to use a configuration with match classification logic built into it.

- The match classification logic that is used to identify high-quality matches, was updated to always require at least a semi-good match on the date of birth regardless of how well everything else matches. It now also requires a very good first name match if SSN isn't an exact match, regardless of how well everything else matches.

Tumor Deduplication Changes

- The linkage configuration file and the filter definition files are the same as they were earlier in the year but changes have been made to the Solid Tumor / MPH library and to the way Match*Pro uses the library.
- The annotations in Match*Pro now identify the module that was used to evaluate the two tumors and the applicable rule where the module stopped. This change was made to make the review process a little easier and to provide more transparency regarding the algorithms used.
- The Solid Tumor / MPH library has been updated to fix bugs and to implement 2023 updates to the solid tumor rules.

Overview of the Deduplication Process

- Download the Match*Pro software and resources mentioned on slide 4.
- Create a NAACCR-XML extract for the purposes of patient deduplication. Specific information regarding how the file should be made (fields required, years to include, etc.) can be found in the instructions.
- Use the Match*Pro software, the linkage configuration file, the XML file, and the instructions to identify the duplicate patients in your database and to create/update the match status archive containing the non-matches.

Overview of the Deduplication Process

- NAACCR will be looking to collect the following information in the fall submission, so be sure to write it down as you review the results from the patient deduplication linkage. Details on how to obtain this information are provided in the instructions.
  - How many matches (duplicates) were found?
  - How many non-matches were there?
  - How many uncertain pairs were there (if more than zero)?

Overview of the Deduplication Process

- When you are finished deduplicating your patients:
  - Put the match status archive containing the non-matches from the patient deduplication process in a safe location so that it can be used next year.
  - Consolidate the duplicate patients in your registry's database.

Overview of the Deduplication Process

- After the duplicate patients have been consolidated in your registry's database, create a new NAACCR-XML extract for the purposes of tumor deduplication. Specific information regarding how the file should be made (fields required, years to include, etc.) can be found in the instructions.
- Use the Match*Pro software, the linkage configuration file, the XML file, and the instructions to identify the duplicate tumors in your database and to create/update the match status archive containing the non-matches.
- Note that the linkage is configured to only look at cases diagnosed in/after 2007.

Overview of the Deduplication Process

- NAACCR will be looking to collect the following information in the fall submission, so be sure to write it down as you review the results from the tumor deduplication linkage. Details on how to obtain this information are provided in the instructions.
    - How many potential duplicates were found for cases diagnosed after 2007?
    - How many potential duplicates were found for cases diagnosed between 2017 and 2021?
    - How many potential duplicates were found for cases diagnosed in 2021?
    - How many potential duplicates did you resolve?
        - If you resolved cases, did you fully resolve the 2021 cases?
        - If you resolved cases, did you fully resolve the 2017-2021 cases?
        - If you resolved cases, did you resolve cases before 2017? If the answer is yes, how far back did you go?

Overview of the Deduplication Process

- When you are finished deduplicating your tumors:
  - Put the match status archive containing the non-matches from the tumor deduplication process in a safe location so that it can be used next year.

    You will also need to submit the tumor status archive to NAACCR

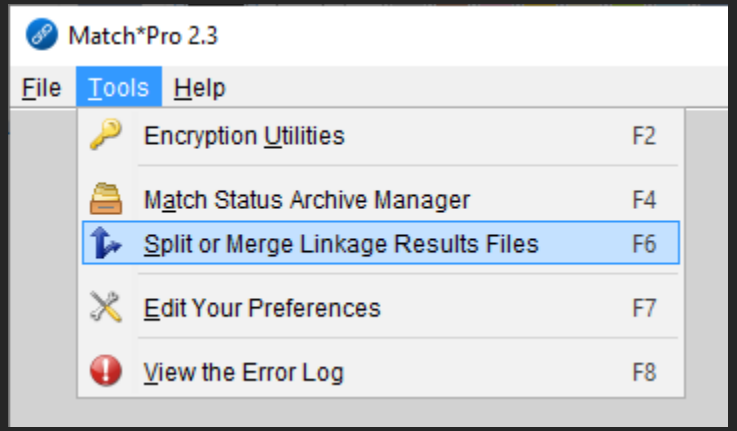
  - Consolidate the duplicate tumors in your registry's database.

Tips & Strategies for Manual Review

- Save often.
- Make backup copies.
- Use Match*Pro's color-coded categories feature to keep track of the pairs you've looked at (right click on a row or rows, select categorize, then choose a color).
- Registries may need to spend 15-30 minutes reviewing each pair of duplicate tumors – so don't wait to get started!

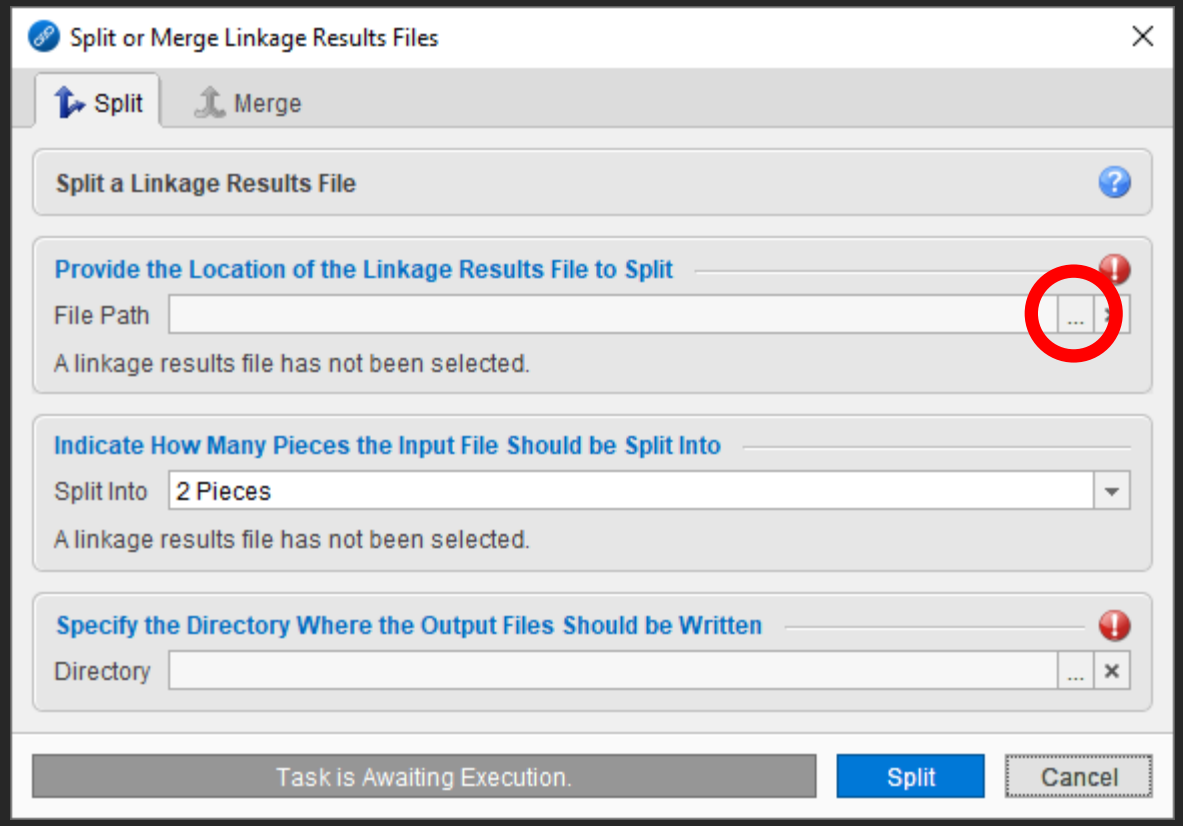
Tips & Strategies for Manual Review

- If your registry has a lot of cases, you may want to employ a divide-and-conquer strategy for the manual review so that you can split the workload up between several CTRs.
- Note: it is not possible for multiple people to edit a Match*Pro results file simultaneously, but the software does provide a tool that can be used to split a results file up into smaller pieces.
- To access the tool, select Split or Merge Linkage Results Files from the Tools menu.



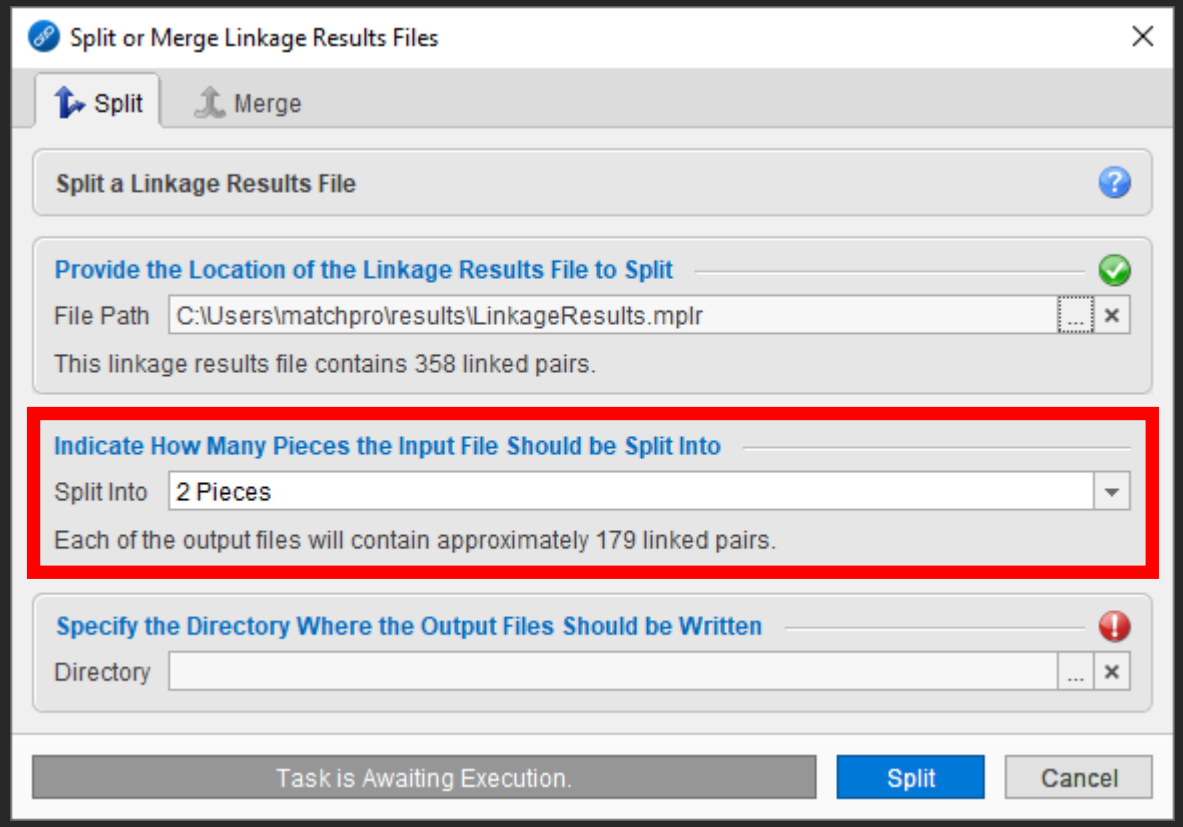
Tips & Strategies for Manual Review

- A dialog will be displayed.
- Provide the location of the results file you want to split up.
- After you select the file, you will be told how many pairs it contains.



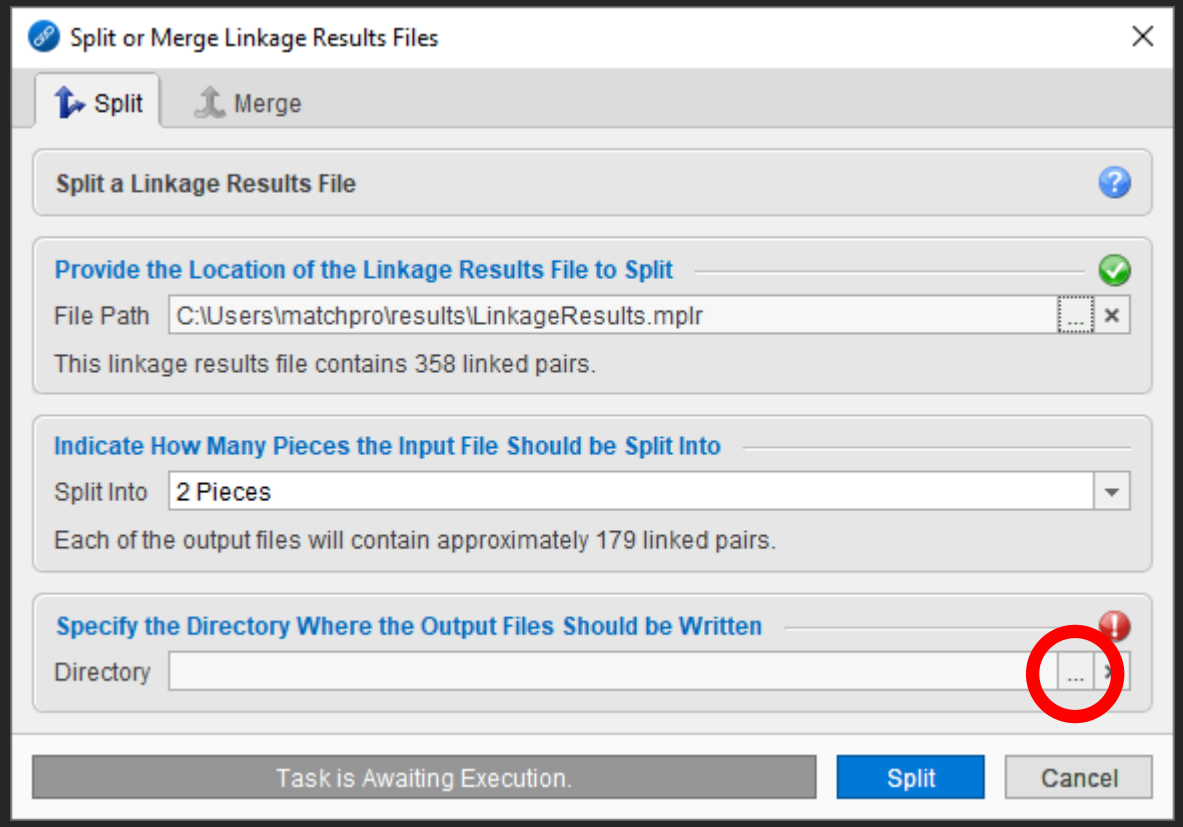
Tips & Strategies for Manual Review

- Next, indicate how many pieces the results file should be split into using the dropdown.
- You can select a value between 2 and 9.



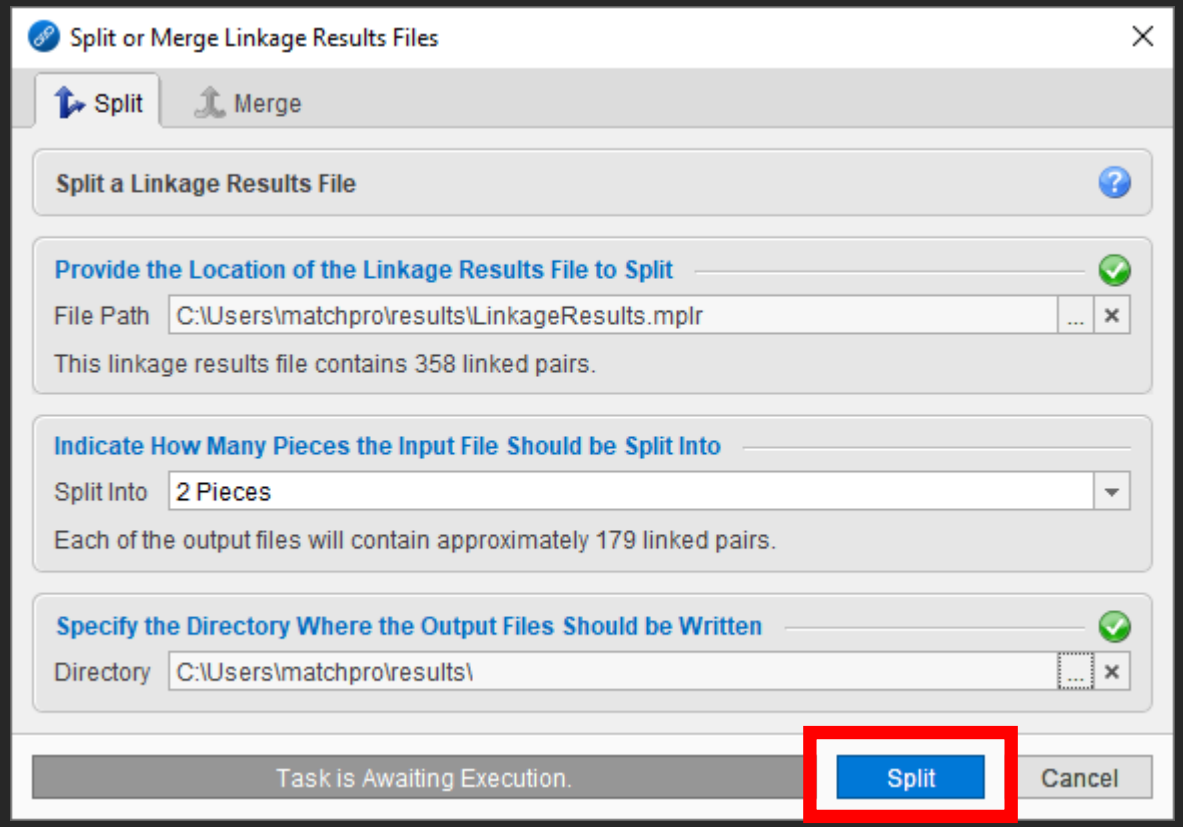
Tips & Strategies for Manual Review

- Next, specify where the smaller results files should be written.



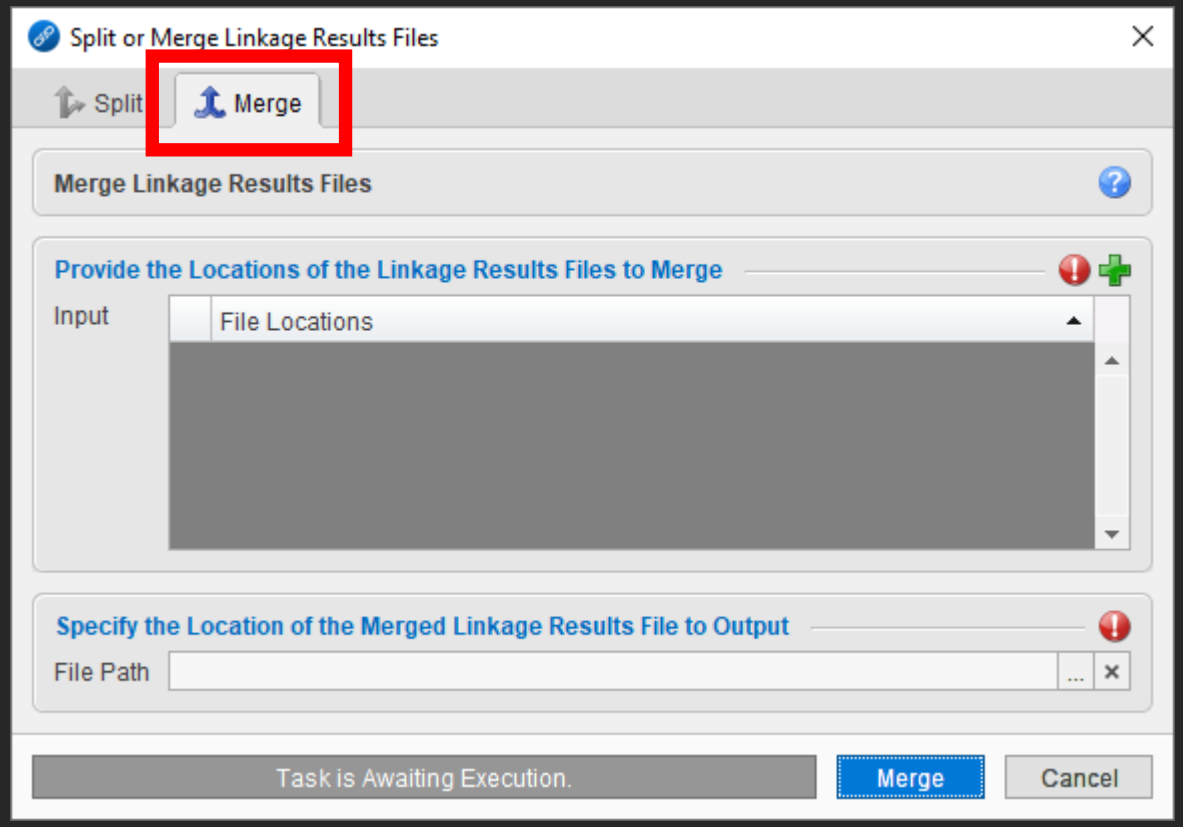
Tips & Strategies for Manual Review

- Press the Split button.
- The process may take a few moments to run.
- When its finished you will find the newly created results files in the folder you specified on the previous slide.
- They will have names like:
  - LinkageResults.part1of2
  - LinkageResults.part2of2



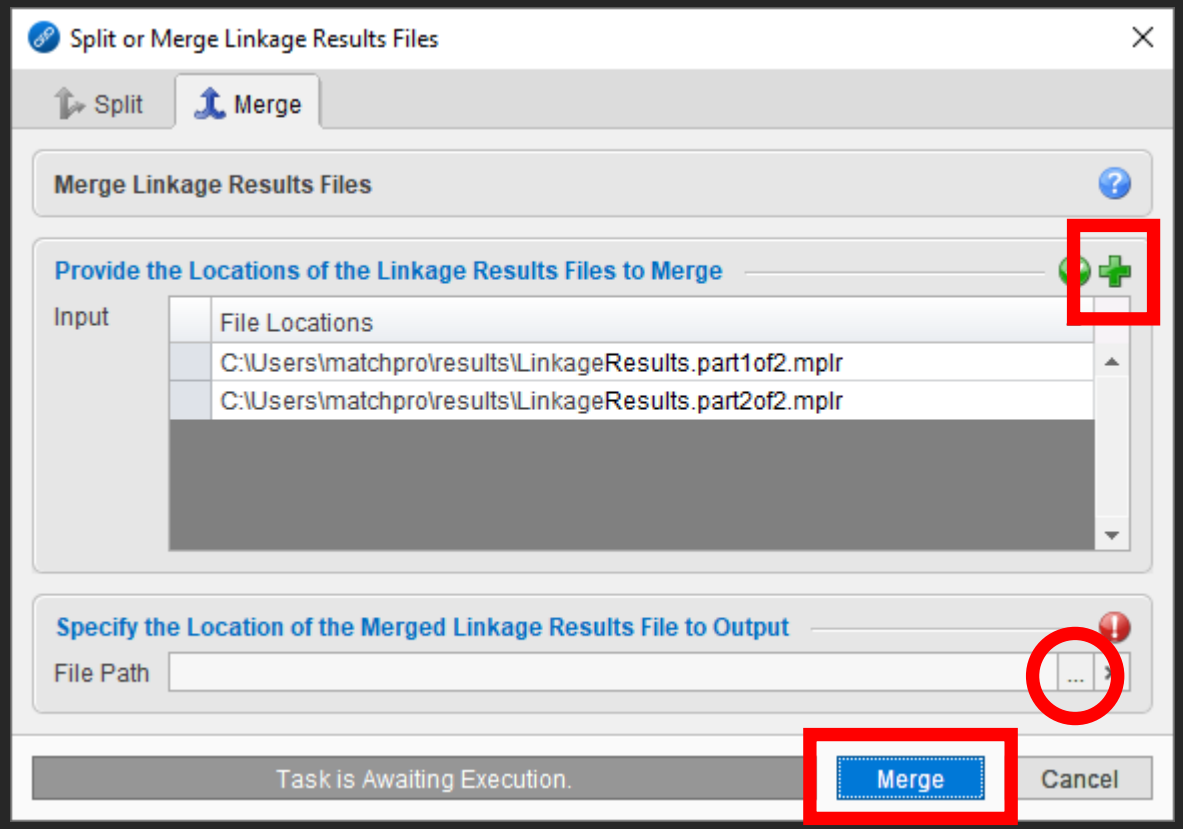
Tips & Strategies for Manual Review

- Note: the cases are split randomly between the files. It is not possible to export specific subsets of cases to one file or another.
- You can recombine all the pieces after they have been individually reviewed using the Merge tab.



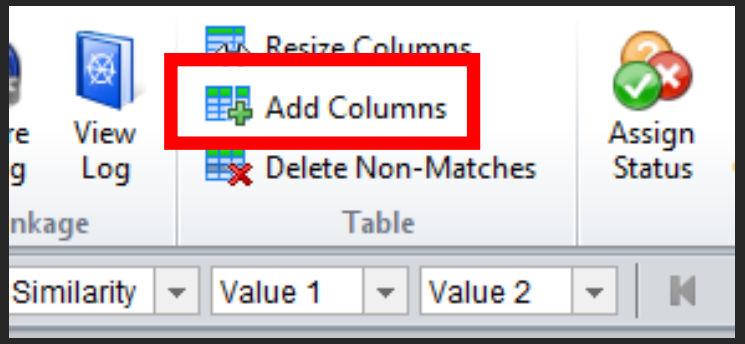
Tips & Strategies for Manual Review

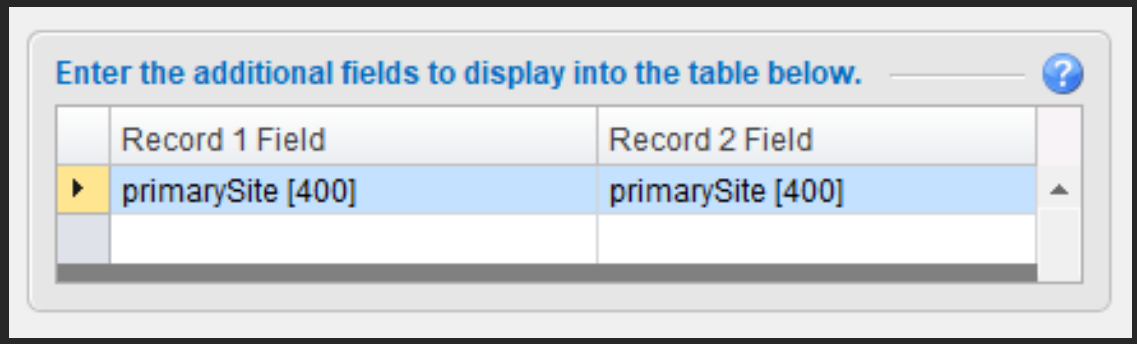
- To merge files, press the Plus sign to add each file you want to merge to the list.
- Specify where you'd like the combined file to be written.
- Press the Merge button to combine the files.



Tips & Strategies for Manual Review

- You may also find it helpful to subset (filter) and/or sort the records by primary site and histology. This way you can work through all the cases of a specific type at the same time to avoid switching between manuals between every other row.
- You can add columns (*e.g.*, a primary site column) by pressing the Add Columns button on the manual review screen and selecting the fields in the table. Each row corresponds to a new column.





QUESTIONS ?