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

William Howe
Information Management Services, Inc.

August 26, 2024

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
- 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
    - 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

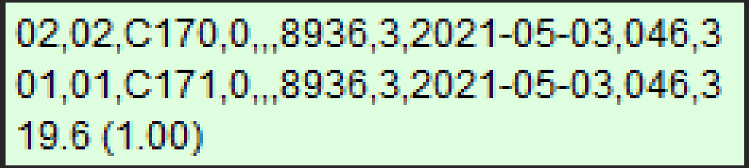
- New configuration file for Match*Pro 2.5.
- EM Algorithm has been enabled. This will increase the runtime but will result in less manual review.

Tumor Deduplication Changes

- New configuration file for Match*Pro 2.5.
- Hash code calculation has been removed. Status archive will now store site, laterality, histology, behavior, and date of dx.
- Sequence number, tumor record number, ICD-O-2 histology and behavior, age at diagnosis, and type of reporting source are no longer included in the list of items Match*Pro generates to find duplicates and are no longer displayed by default (you can add columns if you want to show them).

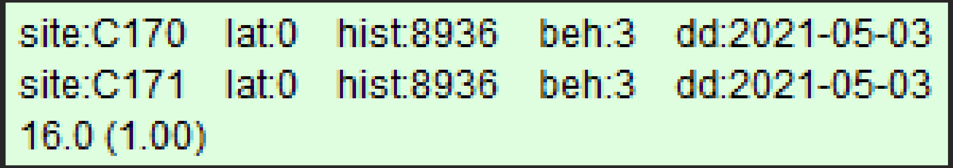
Tumor Deduplication Changes

- The way tumors are presented on the manual review screen has been updated to make things easier to read (fewer items, labels added).



```
02,02,C170,0,,,8936,3,2021-05-03,046,3
01,01,C171,0,,,8936,3,2021-05-03,046,3
19.6 (1.00)
```

→



```
site:C170   lat:0   hist:8936   beh:3   dd:2021-05-03
site:C171   lat:0   hist:8936   beh:3   dd:2021-05-03
16.0 (1.00)
```

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

- 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

- 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 need to submit the tumor status archive to NAACCR so that your duplication rate can be calculated accurately, especially if your duplicate percentage is hovering around 1% for 2007-2022 (GOLD) or 1% for 2018-2022 (SILVER).

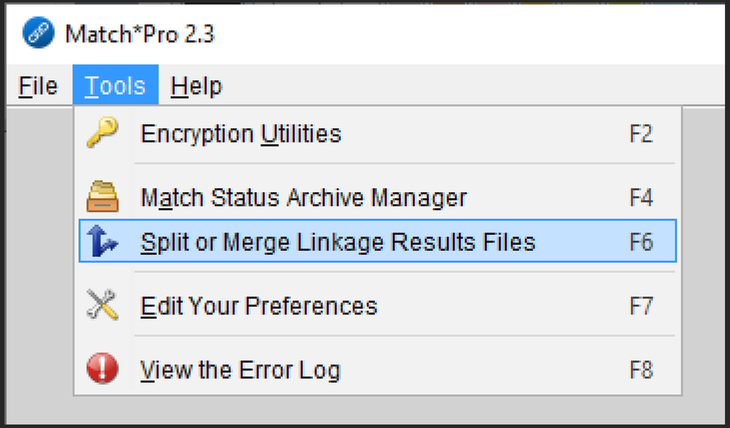
  - Consolidate the duplicate tumors in your registry's database.

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!

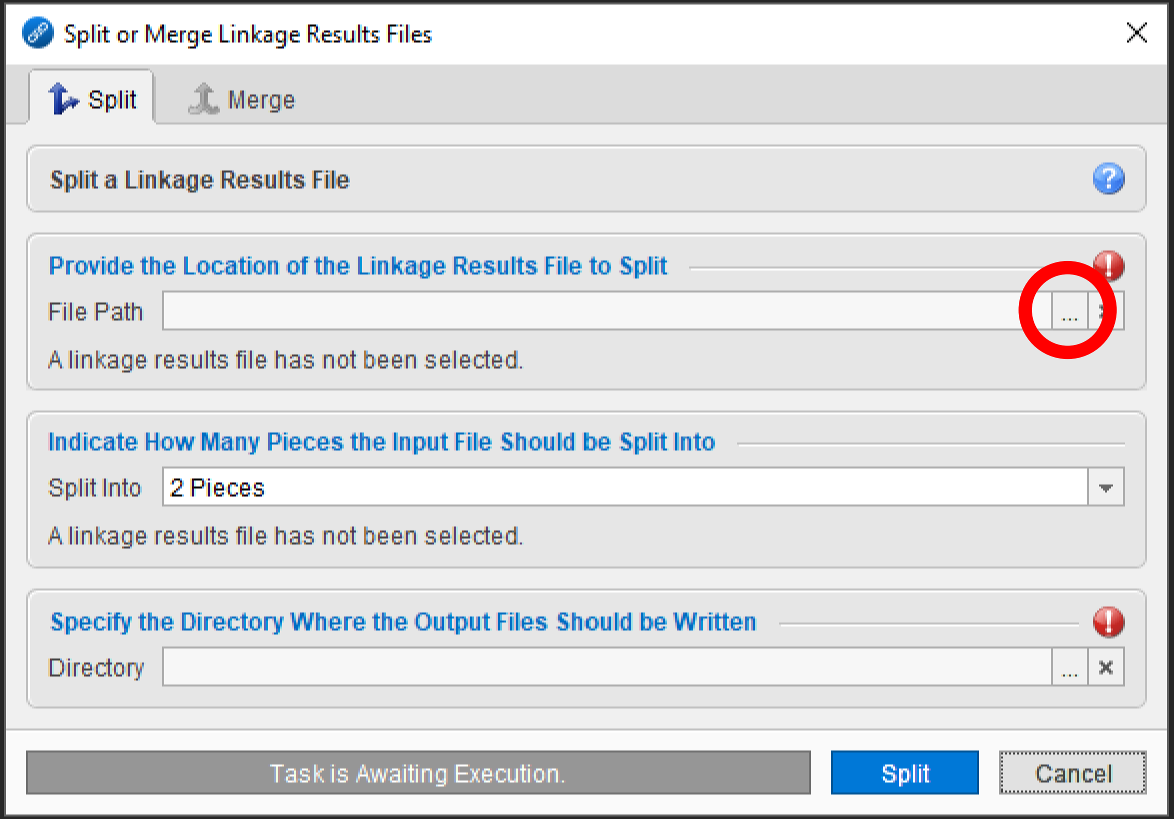
Tips & Strategies for Manual Review

- 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.



Split or Merge Linkage Results Files

Split a Linkage Results File

Provide the Location of the Linkage Results File to Split

File Path

A linkage results file has not been selected.

Indicate How Many Pieces the Input File Should be Split Into

Split Into: 2 Pieces

A linkage results file has not been selected.

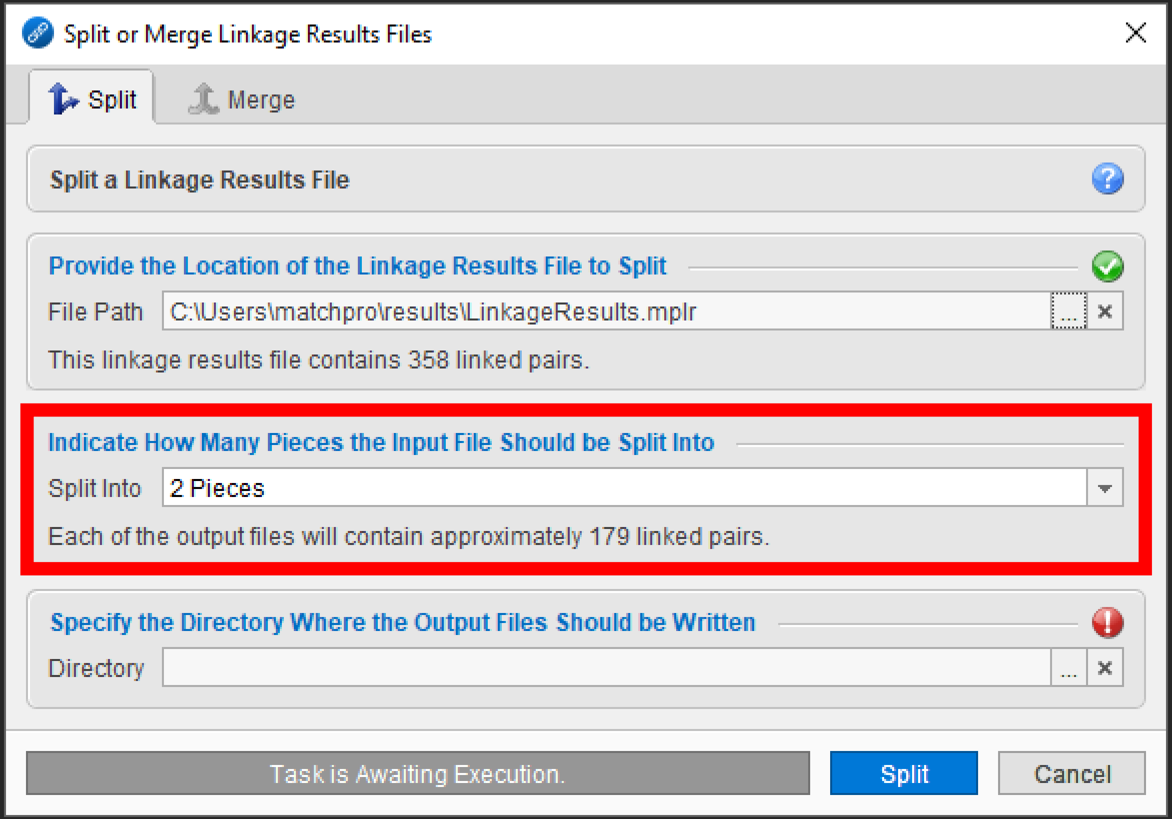
Specify the Directory Where the Output Files Should be Written

Directory

Task is Awaiting Execution.

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.



Provide the Location of the Linkage Results File to Split

File Path: C:\Users\matchpro\results\LinkageResults.mplr

This linkage results file contains 358 linked pairs.

Indicate How Many Pieces the Input File Should be Split Into

Split Into: 2 Pieces

Each of the output files will contain approximately 179 linked pairs.

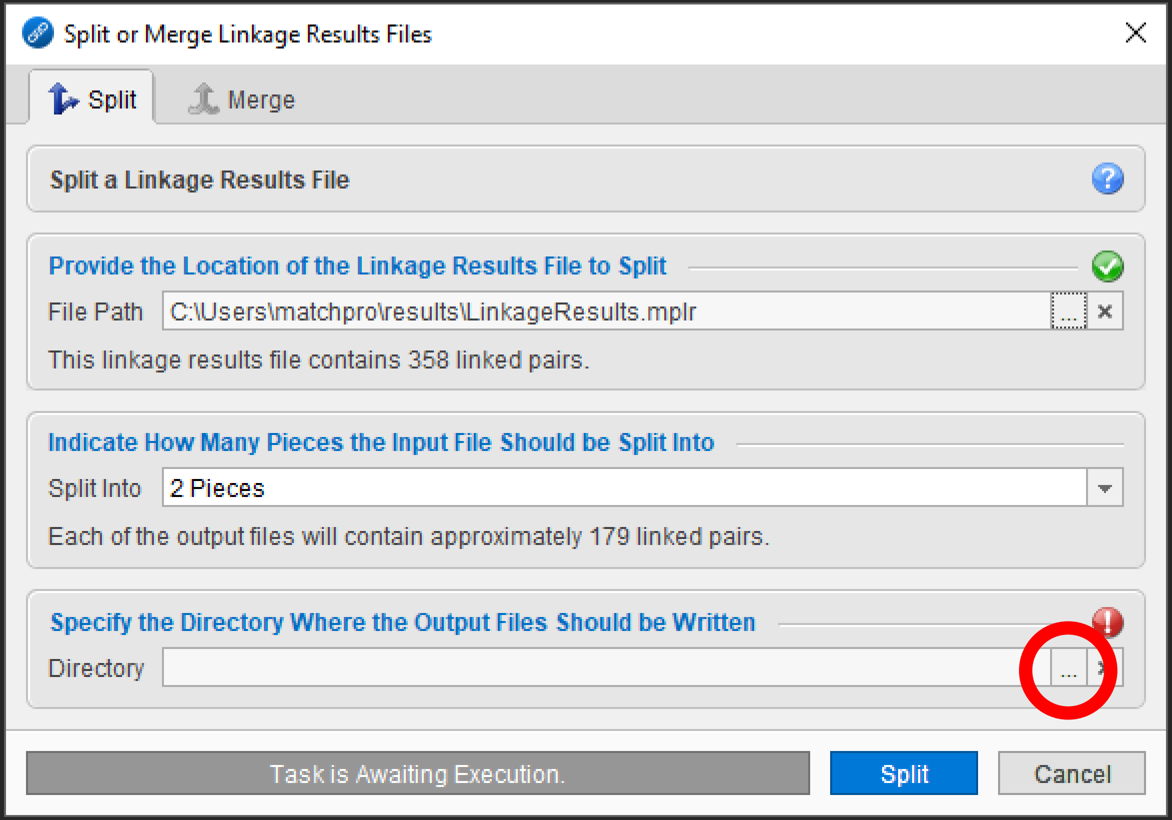
Specify the Directory Where the Output Files Should be Written

Directory:

Task is Awaiting Execution.

Tips & Strategies for Manual Review

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



Split or Merge Linkage Results Files

Split a Linkage Results File

Provide the Location of the Linkage Results File to Split

File Path: C:\Users\matchpro\results\LinkageResults.mplr

This linkage results file contains 358 linked pairs.

Indicate How Many Pieces the Input File Should be Split Into

Split Into: 2 Pieces

Each of the output files will contain approximately 179 linked pairs.

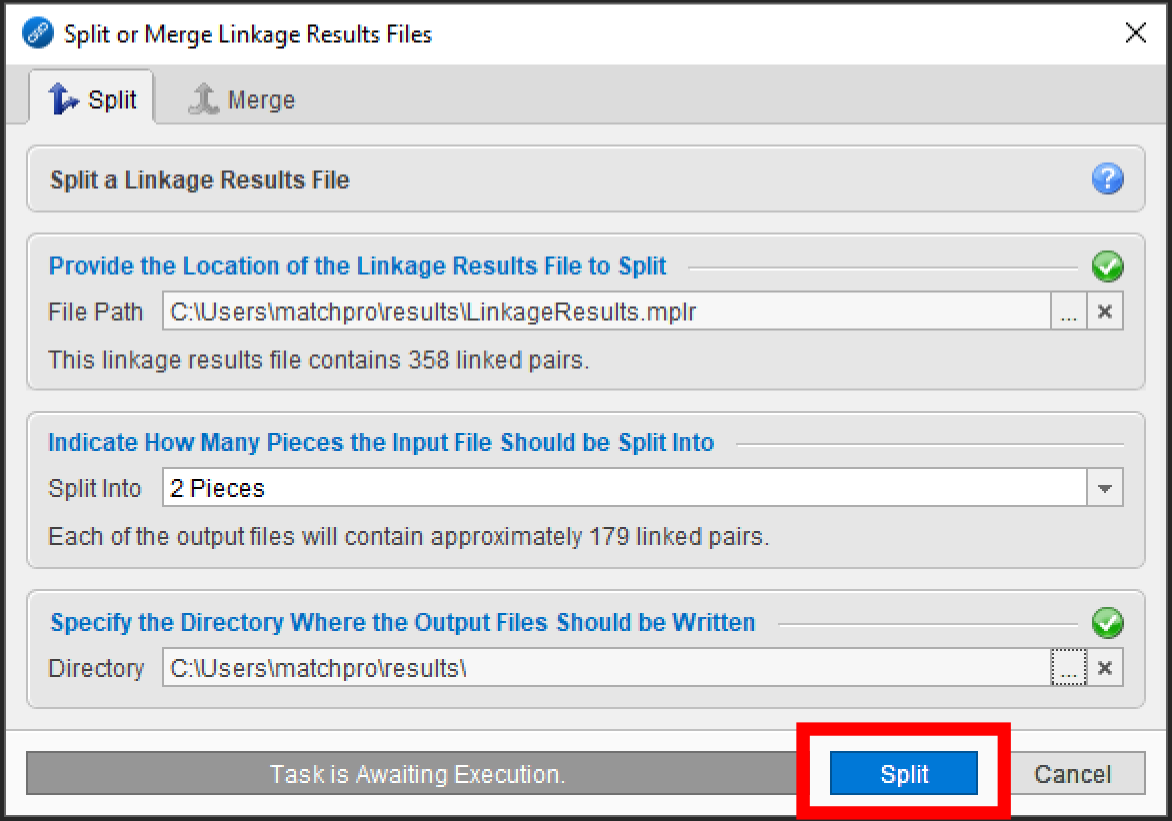
Specify the Directory Where the Output Files Should be Written

Directory:

Task is Awaiting Execution.

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



Provide the Location of the Linkage Results File to Split

File Path: C:\Users\matchpro\results\LinkageResults.mplr

This linkage results file contains 358 linked pairs.

Indicate How Many Pieces the Input File Should be Split Into

Split Into: 2 Pieces

Each of the output files will contain approximately 179 linked pairs.

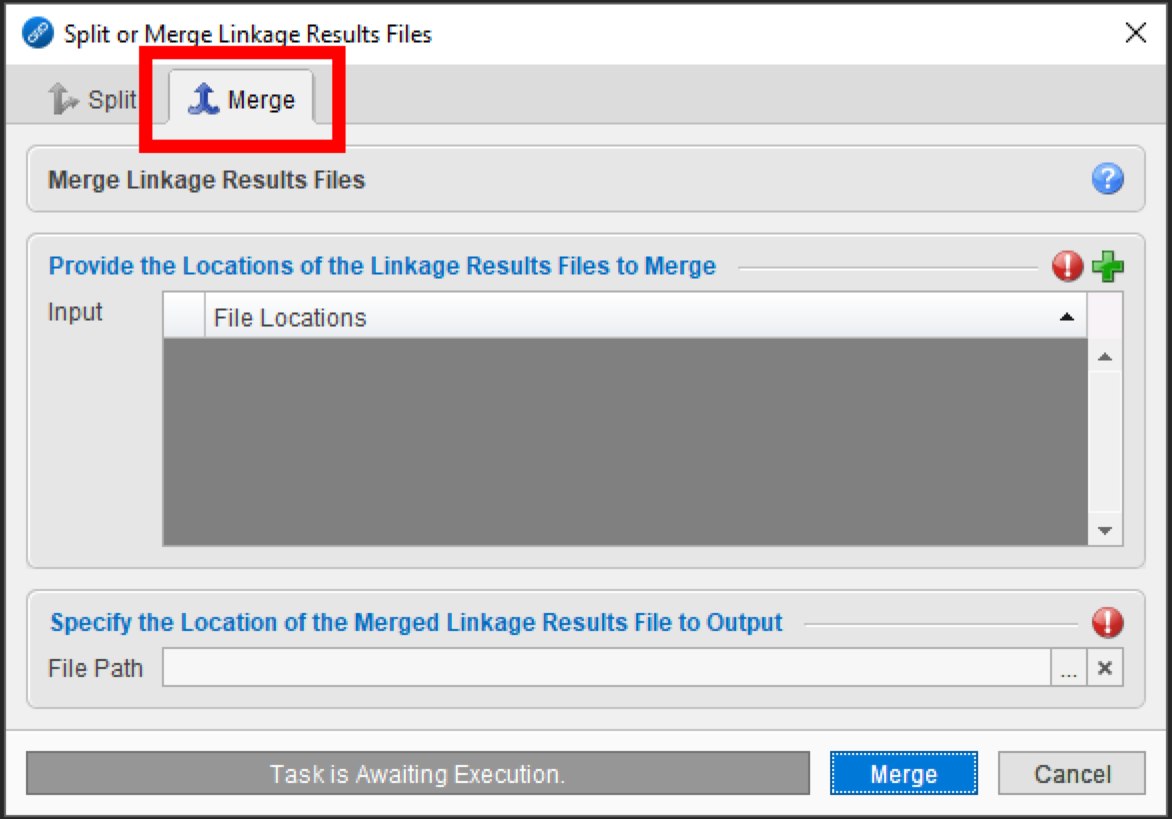
Specify the Directory Where the Output Files Should be Written

Directory: C:\Users\matchpro\results\

Task is Awaiting Execution.

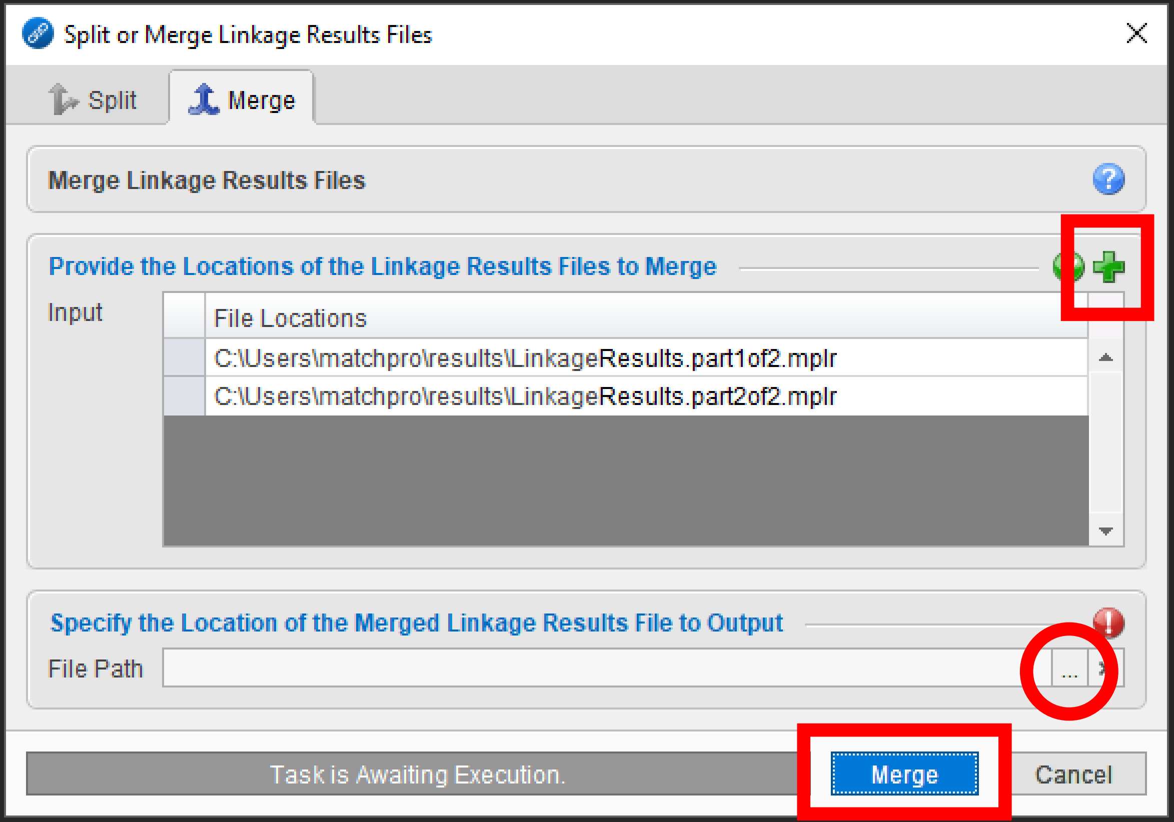
Tips & Strategies for Manual Review

- 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.



Split or Merge Linkage Results Files

Merge Linkage Results Files

Provide the Locations of the Linkage Results Files to Merge

Input

File Locations
C:\Users\matchpro\results\LinkageResults.part1of2.mplr
C:\Users\matchpro\results\LinkageResults.part2of2.mplr

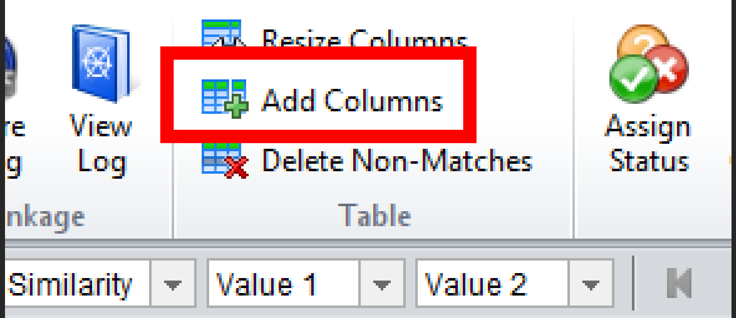
Specify the Location of the Merged Linkage Results File to Output

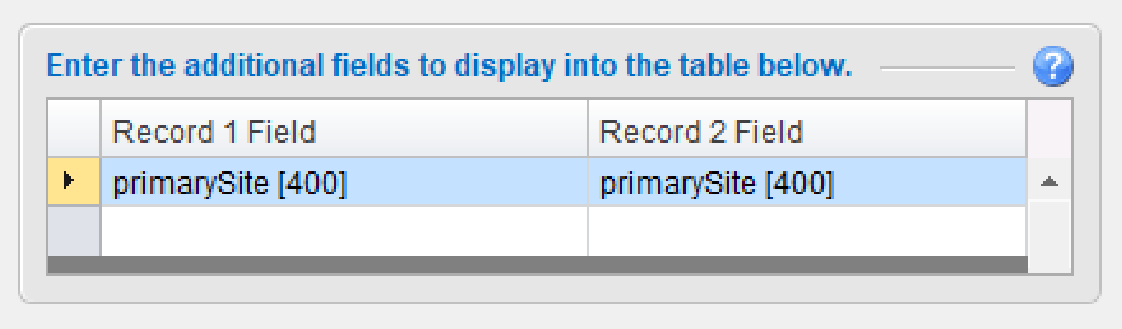
File Path

Task is Awaiting Execution.

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 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.




Enter the additional fields to display into the table below.

Record 1 Field	Record 2 Field
primarySite [400]	primarySite [400]