# N3C Clinical: Cancer Enclave

## Kickoff Meeting

1. Initiate project kickoff meeting with technical contacts at our 6 pilot sites (Duke, Johns Hopkins, University of Chicago (SEER), UNC-Chapel Hill, Washington University, & University of Nebraska)
   1. Go through the proposed project plan.
   2. Homework for sites; need answers to the questions below:
      1. Do you know your local cancer registries contact person?
      2. Do you know which of the NAACCR items are populated by your local registry?
      3. Do you know what version of NAACCR you can export your cancer registry data in? (*UNC is XML v24*)
      4. Do you know what PHI is available? (*we will help you identify PHI if you don’t)*
      5. Does the data have a person identifier that you can link to your EHR data and local cancer registry? (we have MRN's populated well)
         1. If yes, is it well populated/ high match rate?
      6. Does your local cancer registry contain data only for the patients of your academic medical center/ EHR or for more hospitals in the area? this will determine the order of when to run your phenotype? do you need to only find cancer patients from within that EHR/healthcare system? Does your cancer registry have any issues for reuse of the data for research or sharing of the data?
      7. What is the time lag in your registry? (*UNC is 6 months behind*)
      8. Does your registry have a fee to extract data and how much?
      9. Are the regulatory documents (DUA, DTA, Linkage Honest Broker Agreements, or IRB Reliance agreements) signed?
      10. Are you submitting as a PCORnet or OMOP site?
      11. Can you support a payload frequency of 2-4 times a year, with quarterly preferred?
      12. Does your local registry have data from 2018 onward?
      13. Any other questions?
2. After we receive answers from the sites, we will schedule another follow-up call.

## EHR CDM Project Plan:

1. Find [longitudinal phenotype](https://github.com/National-COVID-Cohort-Collaborative/tenant-pilot/tree/main/PhenotypeScripts) on Git and pull your site’s clinical cohort.
2. Find the [extracts scripts](https://github.com/National-COVID-Cohort-Collaborative/tenant-pilot/tree/main/ExtractScripts) on Git and extract EHR CDM tables for that cohort
3. Walk through [export](https://github.com/National-COVID-Cohort-Collaborative/tenant-pilot/tree/main/Exporters) steps
4. Send zipped payload of EHR data (PCORnet or OMOP) through N3C SFTP and use the following naming convention: *SITE\_CLINICAL\_CDM\_ YYYYMMDD (UNC\_CLINICAL\_PCORNET\_20230516)*
5. Do not drop your @resultsDatabaseSchema.N3C\_CLINICAL\_COHORT (we will need the patient ID’s for a future patient matching step)

## Cancer Registry/ NAACCR Project Plan:

### Exporting cancer registry data:

1. Identify and contact your local registry (or registries) for cancer cases.  
   Ask them how you can extract or get an extract of all cancer cases (no time bound). XML v24 is preferred but any XML version will do (v23 or less).
   1. Save XML export locally
2. These were the steps at UNC:
   1. Logged into local cancer registry system: <https://crstarnew.ers-can.com/login>
   2. We selected ‘enhanced reporting’ and then filled out ‘select population’ A screenshot of a computer

      Description automatically generated
   3. We selected abstract cases only, specified our hospital list, and then added selection criteria. We discovered that we couldn’t export the entire population in one go (out of memory error), so we split the population into two by diagnosis date. Your site might not have this issue and you might be able to export the entire population in one report. However, if you run into the same issue that we did, then you can split the population using the diagnosis data. *NOTE: We had a lower bound for our date range because we were told that data before then was unreliable. Your site might not have this issue in which case, pull from as far back as you can. Upper bound will be till present day.*
   4. Clicked ‘select’ to create the population and it will return the # of records selected and you can then save the population label.
   5. Repeat for second population if you had one
   6. Then we were ready to export the registry data in the NAACCR XML v24 (our site is on v24 but your site might be on an earlier version). We selected ‘Data Exports’, then ‘State Export’, then found our first population label saved in the previous step, then select ‘Zip Data File’, and hit Run.A screenshot of a computer

      Description automatically generated
   7. When the export is complete, it saved the zipped folder in our local downloads folder. We retrieved it from there and saved it in our N3C Clinical Cancer Tenant Folder. *Note that our registry system had the default export file type to be the latest XML version but your system might ask you to specify that.*

### Parsing XML to CSV

1. Go here and follow the instructions to download SEER’s File\*Pro (v3.46): [Download the File\*Pro Software (cancer.gov)](https://seer.cancer.gov/tools/filepro/download)
   1. Submit registration form
   2. You’ll receive a download link in your email which will be valid for 24 hrs.
   3. Download and install File\*Pro
2. Open File\*Pro A screenshot of a computer

   Description automatically generated
3. On the Input tab, click Select File to View/Process and find your previously saved XML export.
4. Once you select the XML export, a pop-up window called Data Warnings might show up which displays Data Quality issues with certain rows of data from certain fields. Please ignore and close the pop-up window.
5. File\*Pro will automatically load the XML export and display the Data Overview tab A screenshot of a computer

   Description automatically generated
6. Go to the Filter tab and make sure that ‘No Filter (all data will be processed)’ is selected A screenshot of a computer

   Description automatically generated
7. Go to Recode tab and make sure that ‘No data changes’ is selectedA screenshot of a computer

   Description automatically generated
8. Go to Output tab and select ‘Create a copy of the input file’A screenshot of a computer

   Description automatically generated
9. Go to Output Options and edit the following fields: A screenshot of a computer

   Description automatically generated
   1. Edit Target File to indicate destination folder and file name
   2. Select the compression of the created file as ‘Same as Input File’
   3. Select the format of the created file as ‘CSV’
   4. Under Format Options, the Extracted Fields, click Change. In the pop-up window below, Load the file ‘FilePro Output Column List’ which we have provided separately. This has the full list of columns that we want in the output. Click ‘ok’ A screenshot of a computer

      Description automatically generated
   5. Under Format Options, the Extracted Fields will auto-populate and show the number of extracted columns (this will depend on the XML version but if your site is on v24 then it should say 783 fields and if your site is on v23 then it should say 782 fields)
   6. Select the style of the column headers as ‘Field Numbers (NAACCR Item Numbers) - N999’
   7. Select the value separator as ‘Comma’
   8. Select how to handle new-lines in values as ‘Escape’
   9. Select how to handle quoting values as ‘Use when needed’
   10. Under Advanced Options, select the end-of-line type for the created file as ‘DOS/Windows (CRLF)’
   11. No other selections need to be made below this
   12. Select ‘Process Data File Now’ (this might take a bit- it took us 3 mins.)
10. If extract is successful then the window will look like this with some basic counts/stats A screenshot of a computer

    Description automatically generated
11. Go ahead and check csv by clicking on Open Folder and it will lead to target folder where you can then open the CSV file.

### Patient Matching

1. Take the CSV file (783 columns) and load it into your RDBMS where you can do your patient matching.
2. At UNC, our cancer registry export already has a MRN field which we will use to match the patient to their Epic MRN. This captures a little more than 50% of our cohort.
3. The other 40% is captured by having an exact match on FNAME, LNAME, DOB, & SSN.
4. The last step is to look for MRN merges and that captures a handful more. We are not looking for any non-exact matches.
5. After this, we then crosswalk our matched cohort to their OMOP person\_id (about 83% of our matched cohort were in OMOP).
6. Now your cancer registry data should have 784 columns because of the OMOP person\_id. Please rename to N3C\_PATIENT\_ID
7. Export and save as a CSV since this will be used in the next step.

### Removing PHI Columns

1. Next step is to remove columns that contain PHI from the matched cohort CSV. Columns that contain PHI were identified using the Incidence level record type (from the NAACCR data dictionary for v23 & v24) as well as using the PCORnet Tumor Table specs for private fields. We have done this for NAACCR versions 23 and 24. If you are using the earlier versions then we will help you double check to make sure that you have dropped the PHI columns correctly.
2. Open up FilePro and load the 784 columns CSV. Click process in the ‘Unknown Format’ pop-up window.A screenshot of a computer

   Description automatically generated
3. Skip Data Overview, Filter, and Recode tabs (leave as default) and click on Output. Select ‘Create a copy of the input file’A screenshot of a computer

   Description automatically generated
4. Go to Output Options and edit the following fields: A screenshot of a computer

   Description automatically generated
   1. Edit Target File to indicate destination folder and file name
   2. Select the compression of the created file as ‘Same as Input File’
   3. Select the format of the created file as ‘CSV’
   4. Under Format Options, the Extracted Fields, click Change. In the pop-up window below, Load the file ‘FINAL OUTPUT NO PHI’ which we have provided separately. This has the full list of columns that we want remaining in the output after we have dropped the PHI containing columns. Click ‘ok’A screenshot of a computer

      Description automatically generated
   5. Now under Format Options, the Extracted Fields will auto-populate and show the number of extracted columns (651 or 650 depending on the XML version) A screenshot of a computer

      Description automatically generated
   6. Select the style of the column headers as ‘Same as Input File’
   7. Select the value separator as ‘Comma’
   8. Select how to handle new-lines in values as ‘Escape’
   9. Select how to handle quoting values as ‘Use when needed’
   10. Under Advanced Options, select the end-of-line type for the created file as ‘DOS/Windows (CRLF)’
   11. No other selections need to be made below this
   12. Select ‘Process Data File Now’ (this might take a bit)
5. If extract is successful then the window will look like this with some basic counts/stats A screenshot of a computer

   Description automatically generated
6. Go ahead and check csv by clicking on Open Folder and it will lead to target folder where you can then open the CSV file. Verify that the columns containing PHI have indeed dropped.

### SFTP NAACCR Data

1. To submit your NAACCR csv file via SFTP:
   1. Take the output file from the previous step, rename to SiteAbbrev\_NAACCR\_YYYYMMDD.csv
   2. Then manually submit that with (or after) your EHR CDM payload.
   3. For more detailed submission instructions, please check the [github page](https://github.com/National-Clinical-Cohort-Collaborative/tenant-pilot/wiki/Cancer-Enclave-Specific-Instructions).