

Prototyping a Protocol to Identify Duplicate Cancer Reports within New England States' Cancer Registries (2021) Protocols 1 and 2

Project Leads: Judy Rees and Bruce Riddle, NH State Cancer Registry
Contact: crystal.murphy@dartmouth.edu

Summary

This pilot project will test whether pseudonymization techniques are capable of providing estimates of duplicate cases between states.

The project will be carried out using pre-XML, 1995-2018 data from each state's registry. **Please note that the code provided here will need to be revised to accommodate the changes related to XML.**

Pseudonymized files will be created by each registry. New Hampshire State Cancer Registry (NHSCR) will act as the broker state by taking the pseudonymized files from each participating state. As the broker, NHSCR will perform linkages, conduct quality assurance activities, and return files to states. Resolution of duplicates will be performed by states via direct state-to-state communication.

Methods

Phase I

Each state's cancer registry will follow the protocol in the User Documentation to:

1. Create a NAACCR extract of all reportable cases, 1995-2018, with confidential information in the dataset. This extract will be retained by each state and not shared.
2. Run two provided SAS programs on the NAACCR extract to create two CSV files. The resulting CSV files will be retained by each state and not shared.
3. Run the free *OpenPseudonymiser* (OP) program on each of the two files created by the SAS programs. The Internal OP output file, containing the digest and confidential information, will be retained by each registry and not shared. The External OP output file, containing only the digest and state, will be transmitted to NH.
4. Securely transmit the External OP output file to the NHSCR using WebPlus.

NHSCR will:

1. Perform linkages to determine duplicates between states.
2. Return an Excel spreadsheet of potential duplicates to each state. Cases will be identifiable to the submitting state using the crosswalk file retained by the state to re-link the digest to the patient id and other relevant variables.

Phase II: To preserve confidentiality, *each state* will resolve its own duplicates with the other four states. Conversations between states about mutual cases fall under the inter-state data exchange agreement.

Protocol for Phase 1

Create NAACCR Extract

We suggest doing this while still on NAACCR 18, before the change to NAACCR 21 (XML format). You can store the file locally until you are ready to complete the remaining steps.

1. Create a NAACCR extract of all reportable cases, 1995-2018. Variables used in this protocol are:

patient id	N18_20
state	N18_80
sex	N18_220
age	N18_230
date of birth	N18_240
sequence number	N18_380
diagnosis year	4-digit from N18_390
cancer site	2-digit from N18_400
laterality	N18_410
source	N18_500
histology	N18_522
date of last contact	N18_1750
vital status	N18_1760
cause of death	N18_1910
first name	N18_2230
last name	N18_2240
SSN	N18_2320
address at dx	N18_2330

2. Save the file locally, noting the path and filename.

Retrieve Two SAS Programs

Retrieve the SAS programs 'Internal SAS Procedure for OpenPseudonymiser 07-12-2021' and 'External SAS Procedure for OpenPseudonymiser 04-20-2021' from email OR copy them from Appendices A and B of this documentation.

Modify the SAS Programs with Specifics for Your Files

Internal SAS Procedure for OpenPseudonymiser:

Modify the red text in the four '%LET' statements within the SAS program

- Set the two-character abbreviation for your state:
`%let state_ab = \NH ;`
- Set the location of the source file. Please take note of the single quotes:
`%let sourcename = ' Drivename:\folder[s]\filename.dat' ;`
- Set the location of the output files. [Note: These files WILL NOT BE SENT to NHSCR. Only files from the OpenPseudonymiser software will be sent to NHSCR.]
`%let r_outfile = Drivename:\folder[s] ;`
- Set the name of the output file. We recommend that you only change the date.
`%let xfname = INT_bypatcancer_07122021.csv ;`

External SAS Procedure for OpenPseudonymiser:

Modify the red text in the four '%LET' statements within the SAS program

- Set the two-character abbreviation for your state:
`%let state_ab = \NH ;`
- Set the location of the source file. Please take note of the single quotes:
`%let sourcename = ' Drivename:\folder[s]\filename.dat' ;`
- Set the location of the output files. [Note: These files WILL NOT BE SENT to NHSCR. Only files from the OpenPseudonymiser software will be sent to NHSCR.]
`%let r_outfile = Drivename:\folder[s] ;`
- Set the names of the two output files. We recommend that you only change the date.
`%let xfname = _EXTbypatcancer_04202021.csv ;`

Register, Download, and Install the OpenPseudonymiser (OP) Program

Prerequisites:

- You will need Admin permissions on your machine to install the program.
- The program runs on windows computers (XP, Vista, Win7, Win8, Server 2003/8) and uses the .NET framework. It works on both 32 and 64 bit versions of windows.
- You may have to install the .NET framework. If it is not installed, you will get an error message from the OP installer (steps below).
- The .NET framework is available here: <https://www.microsoft.com/en-us/download/details.aspx?id=21>
- Despite the statements that .NET and OP run only on the older versions of Windows listed above, they seem to run fine on our (NH) Windows 10 machines.

OP Steps:

1. Go to <https://www.openpseudonymiser.org>
2. Register by clicking 'Log In' in the upper right then 'Register here' on the login page, or by going to:
<https://www.openpseudonymiser.org/Account/Register.aspx>
3. You must create an account, then activate it. To activate your account, accept the terms by clicking the link in the email you receive from NOREPLY@openpseudonymiser.org. You may need to look in your spam folder for the email.
4. Log in to OpenPseudonymiser. Note that you must check the box to accept the Terms of Service.
5. Click 'The Software' in the navigation bar.
6. Download the software by clicking "Download OpenPseudonymiser Desktop v.2.0.2"
7. Locate and double-click the installer called "OpenPseudonymiser (BatchProcessor) v2.0.2.Setup"
8. Click 'Next' as needed... including accepting the terms of the license agreement.

Retrieve NHSCR SALT

Retrieve the SALT from your email and save it locally. The filename is: NHSCR dupes.EncryptedSalt

The 'SALT' is an extra string of characters appended to each pseudonymized record. This allows data to be shared for a specific project with no risk of the data being cross referenced by another project that uses different salt.

You may read more about SALT in the OP user documentation or here:

<https://www.openpseudonymiser.org/FAQ.aspx>

The SALT we send will be unique to this project. Every state must use the same SALT.

Run OP Program

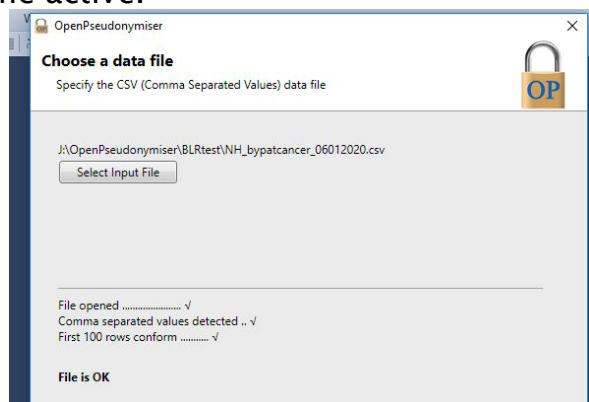
You will run the OP program twice: once on the INTERNAL file and once on the EXTERNAL file created by the SAS programs.

1. Start the OP program.

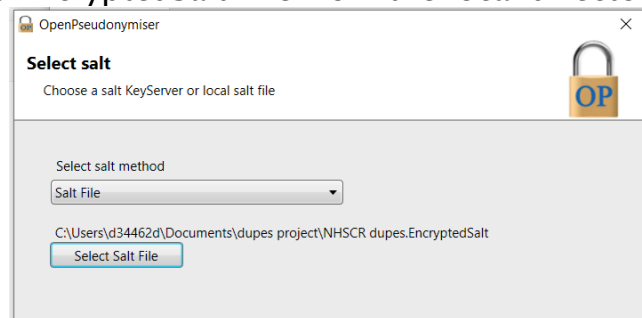
If the program file OpenPseudonymiser does not open correctly, verify you have the .NET framework installed by going to <http://www.microsoft.com/net/> Version 3.5 (or later) of the .NET framework is required.

The first time the program is run, you will be asked to read and accept the license.

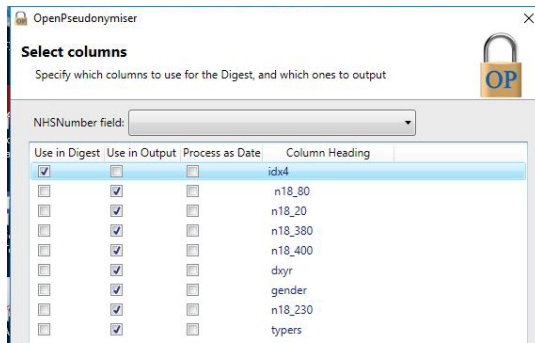
2. Select an input file (one of the two files created by the SAS programs). OP will then perform some basic checks. If these are all okay, the 'Next' button will become active.



3. Select 'Salt File' in the 'Select salt method' dropdown, then select the 'NHSCR dupes.EncryptedSalt' file from the local directory where you saved it.

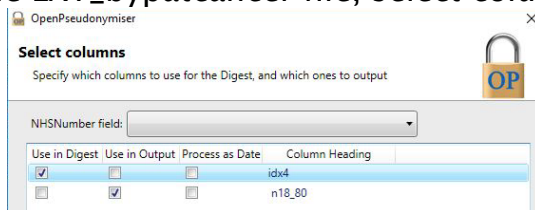


4. For the INT_bypatcancer file, select columns according to this screenshot:



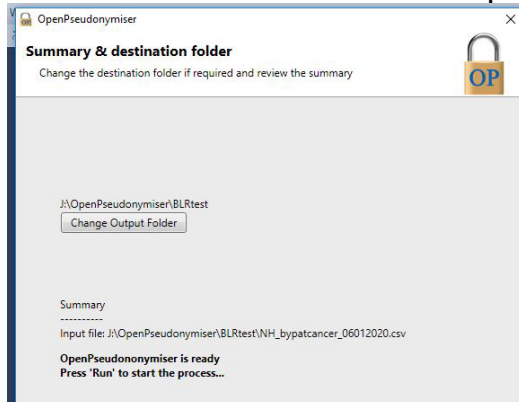
- These selections tell OP to use the column 'idx4' to create the digest. The 'Use in Output' box for column 'idx4' is selected here as you are now creating the crosswalk file that you will use to re-link the potential duplicates returned from NH.
- Note that nothing should be selected in the 'NHSNumber field' drop-down. You do not have to set the drop-down. It clears automatically with the checkbox selections shown above.

For the EXT_byatpcancer file, select columns according to this screenshot:

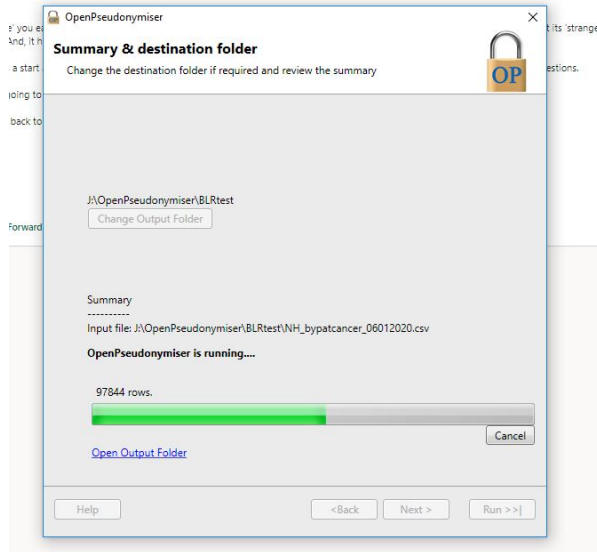


- These selections tell OP to use the column 'idx4' to create the digest. By unselecting the 'Use in Output' box for column 'idx4' we ensure that the resulting CSV file that will be transmitted to NH will not contain this column.
- Note that nothing should be selected in the 'NHSNumber field' drop-down. You do not have to set the drop-down. It clears automatically with the checkbox selections shown above.

5. Designate the destination for the output files.



6. Click 'Run' and wait for the file to be processed.



7. Inspect the output files.

Two files will have been created in the location you specified:

1. The log. This will have the same name as your input file, but will have a file type of “OpenPseudonymiserRunLog” (Note: this file is a record of the settings used in processing. You do not need to submit this file to NHSCR.)
2. The processed CSV file. This will have the same name as your input file but with the word “OpenPseudonymised_” before it. Its contents should look something like this with your state's abbreviation in place of 'NH':

```

== * * * Top of File * * *
|...+....1....+....2....+....3....+....4....+....5....+....6....+....7.]
== Digest, nl8_80
== 4759F0C908B4853A5F3091B5875F1EAC0FC8E15CF6C7B67EBCEFF2305E96CEF3, NH
== F8F6A8A6A17A3FDC43CA1BE81E8D0AA321F6FD51A5B408BEA6FC32FA517EASDC, NH
== E71B66FE1C7C96CD59C5122FC4AB606215164A9B4C867F7C2A2CC23CC2E67D16, NH
== D5207C7824E8623635CB5C1F41A8BB30FE584EDE9BC4A1406B19C75B014FF54A, NH
== 826B93F93D375BB6DE1A8108E9395B2D0A79C0D62F1378F614ACC167E7F3FF5E, NH
== 1EDFB26296293D5158B6FD2D7C340F82404CF673E66354DD6DD9F70359B5DB10, NH
== 2DB49076449AFE5400824BF93A39FC2623C8FED9FB795DEB69CB38818943A8D3, NH
== 8B3795E5509122A5BED869FEA815C6D63E909E747345C5803BAAA21A367B047C, NH

```

8. Repeat steps 2 through 7 on the second file created by the SAS programs.

Retain Crosswalk File

Retain the original INTERNAL OP output file. This is your crosswalk file. In Phase II, you will receive SAS code for re-linking this file to the file of potential duplicates that NH will return to you.

Transfer the Two OP Output Files to NHSCR

Transfer the OP EXTERNAL output file to NHSCR using the same method you use for regular data exchanges with NHSCR:

- Vermont - VT WebPlus
- Maine, Massachusetts, Rhode Island - NHSCR WebPlus

Follow up Activities

We will return the linkages to you in .TXT files with one row for each linked case.

We propose the following:

- Arrange a call with the corresponding state registry and see if a linkage in your state matches the same person and cancer in a neighboring state.
- NH will provide a SAS procedure to re-link returned files to the file generated by the 'External SAS Procedure for OpenPseudonymizer.'
- NH will provide an Excel tool for recording information about each linked case.
- The Excel tool will generate aggregate counts to be returned to NH.

We propose to start these studies with the most recent five years to get an idea of how many linkages are valid, i.e., the same person.

See Phase II Protocol for specifics.

Appendix A to the Phase 1 Protocol

INTERNAL SAS Program to Prepare NAACCR Flat File for OpenPseudonymiser

```
* Internal SAS Procedure for OpenPseudonymiser 07-12-2021 ;
* relink NH SAS Procedure for OP 07-12-2021 ;
options pagesize=35 linesize=122 obs=max errors=20
        pageno=1 nocenter ;

Title Internal SAS Procedure for OpenPseudonymiser 07-12-2021 New Hampshire State
Cancer Registry ;
/* abbreviation of state --change to your state */
%let state_ab = \NH ;
run ;

/* This is the name and location for the NAACCR flat file */
%let sourcename = 'Drivename:\folder[s]\filename.dat' ;
/* This the location for the output files */
%let r_outfile = Drivename:\folder[s] ;
run ;
data source ;
    infile "&sourcename" recfm=v lrecl=24194 missover;
    input @42 N18_20 $8.
           @124 N18_80 $2.
           @222 n18_220 $1.
           @223 n18_230 $3.
           @226 n18_240 $8.
           @542 n18_380 $2.
           @544 dxyr $4.
           @564 n18_522 $4.
           @558 n18_410 $1.
           @577 n18_500 $1.
           @554 n18_400 $4.
           @2775 n18_1750 $8.
           @2785 n18_1760 $1.
           @2940 n18_1910 $4.
           @4049 n18_2230 $40.
           @4089 n18_2240 $40.
           @4328 n18_2320 $9.
           @4348 n18_2330 $60. ;
run ;
proc print data=source (obs=20) ;
run ;
/* setup the strings */
data source ;
    set source ;
    p_name= (trim(n18_2230)||' '||(trim(n18_2240)) ;
    address = trim(n18_2330) ;
    idx3 = n18_220||n18_240||n18_2230||n18_2240 ;
    cancer = substr(n18_400,2,2) ;
    idx4 = n18_220||n18_240||n18_2230||n18_2240||cancer ;
    dupflag = n18_20||cancer ;
    if n18_220 = '1' then gender = 'M' ;
    if n18_220 = '2' then gender = 'F' ;
    if n18_220 ^in ('1','2') then gender = 'O' ;
    if n18_500 = '1' then typers = 'hosp in ' ;
    if n18_500 = '2' then typers = 'rad treat ' ;
    if n18_500 = '3' then typers = 'lab only ' ;
```



```

if n18_500 = '4' then typers = 'phy office ' ;
if n18_500 = '5' then typers = 'nurse-hosp ' ;
if n18_500 = '6' then typers = 'autopsy ' ;
if n18_500 = '7' then typers = 'DCO ' ;
if n18_500 = '8' then typers = 'other ' ;
run ;
/* this step creates file for patient by cancer */
proc sort data=source out=xnhs2 nodupkey ; by dupflag ;
run ;
proc freq data=xnhs2 order=freq ; tables idx4 / out=a ;
title2 data set a IDX4 deduped by dupflag ;
run ;
proc freq data=a ; tables count ;
run ;

%let xfname = INT_bypatcancer_07122021.csv ;
run ;

data _null_ ;
    * file outx ;
    file "&r_outfile&state_ab&xfname" ;
put 'idx4, n28_20, n18_380, n18_80, p_name, address, gender, age, site,
histology, laterality, dxyr,
typers, vitalstatus, causeofdeath, dateoflastcontact' ;
run ;

data _null_ ;
set xnhs2 ;
file "&r_outfile&state_ab&xfname" mod ;
put idx4 ','
n18_20 ','
n18_380 ','
n18_80 ','
p_name ','
address ','
gender ','
n18_230 ',' /* age */
n18_400 ',' /*site */
n18_522 ',' /* histology */
n18_410 ',' /* laterality */
dxyr ','
typers ','
n18_1760 ',' /*vital status */
n18_1910 ',' /* cause of death */
n18_1750 /* date of last contact */ ;
run ;

```

Appendix B to the Phase 1 Protocol

EXTERNAL SAS Program to Prepare NAACCR Flat File for OpenPseudonymiser

```
* External SAS Procedure for OpenPseudonymiser 04-20-2021 ;
options pagesize=35 linesize=122 obs=max errors=20
      pageno=1 nocenter ;

Title External SAS Procedure for OpenPseudonymiser 04-20-2021 New Hampshire State
Cancer Registry ;
/* abbreviation of state --change to your state */
%let state_ab = \NH ;
run ;

/* This is the name and location for the NAACCR flat file */
%let sourcename = 'Drivename:\folder[s]\filename.dat' ;
/* This the location for the output files */
%let r_outfile = Drivename:\folder[s] ;
run ;
data source ;
  infile "&sourcename" recfm=v lrecl=24194 missover;
  input @42 N18_20 $8.
        @124 N18_80 $2.
        @222 n18_220 $1.
        @223 n18_230 $3.
        @226 n18_240 $8.
        @542 n18_380 $2.
        @554 n18_400 $4.
        @4049 n18_2230 $40.
        @4089 n18_2240 $40. ;
run ;
proc print data=source (obs=20) ;
run ;
/* setup the strings */
data source ;
  set source ;
  idx3 = n18_220||n18_240||n18_2230||n18_2240 ;
  cancer = substr(n18_400,2,2) ;
  idx4 = n18_220||n18_240||n18_2230||n18_2240||cancer ;
  dupflag = n18_20||n18_230||cancer ;
run ;
/* this step creates file for patient by cancer */
proc sort data=source out=xnhs2 nodupkey ; by dupflag ;
run ;
proc freq data=xnhs2 order=freq ; tables idx4 / out=a ;
title2 data set a IDX4 deduped by dupflag ;
run ;
proc freq data=a ; tables count ;
run ;

%let xfname = _EXTbypatcancer_04202021.csv ;
run ;

data _null_ ;
  * file outx ;
  file "&r_outfile&state_ab&xfname" ;
  put 'idx4 , n18_80 ' ;
run ;

data _null_ ;
  set xnhs2 ;
```

```
file "&r_outfile&state_ab&xfname" mod ;  
put idx4 ',' n18_80 ;  
run ;
```

Phase II Protocol

Retrieve SAS Program

Retrieve the SAS program 'link matched back to source generic 07-28-2021' from email OR copy it from Appendix A of this documentation.

Modify the SAS Program with Specifics for Your Files

Modify the text in red below within the SAS program

- Set S1 = your state abbreviation and S2 = the state with which you are reconciling (this naming convention will be used in all reconciliation SAS code and filenames):

```
%let s1 = NH ;  
%let s2 = RI ;
```
- Set the path of the file that was returned to your state from NH:

```
data Linked ;  
  infile "Drivename:\folder[s]\&s1._to_&s2._pat_cancer.txt" dlm = ','  
  missover DSD lrecl = 118 ;
```
- Set the path and filename of the INTERNAL file created by OP software from your state Internal File:

```
data source ;  
  infile " Drivename:\folder[s]\s1_INT_bypatcancer_07122021.csv" dlm = ','  
  missover DSD lrecl = 118 ;
```
- Set the path and date in the filename for the output file that will be brought into the Excel spreadsheet for reconciliation:

```
filename relinkx "  
Drivename:\folder[s]\relink2&s1.._to_&s2.._pat_cancer_07132021.txt" ;
```

Populate Reconciliation Tracking Spreadsheet with SAS Output

IMPORTANT:

- **DO NOT MAKE CHANGES TO THE FORMAT OR STRUCTURE OF THE WORKBOOK** (adding/rearranging columns will break linkages required to generate the counts)
- **Work ONLY in the 'data collection' worksheet**

1. Open the SAS output file in Excel: relink_S1_to_S2_pat_cancerMMDDYYYY.txt, where S1 is your state and S2 is the state with which you are reconciling.
2. Click the column header 'A' to select the entire column.
3. Under the 'Data' menu, choose 'Text to Columns...'
4. In the *Text to Columns Wizard*, select 'Delimited'
5. Click 'Next'
6. Select 'Comma' as the delimiter.
7. The contents in the preview window should now be parsed into 16 columns: Digest, PatID, SeqNo, State_ID, PatName, PatAddress, Gender, Age, Site, Histology, Laterality, DX_Yr, Source, VitalStatus, CDeath, DateLastContact (note that you may only see the digest column in the preview window because the contents are so long).
8. Click 'Finish'
9. Select all cells containing values (A1:Pxxx), and only the cells containing values (do not select whole rows or columns).
10. Copy.
11. Open the Reconciliation Tracking Spreadsheet (RTS) you received from NH with the filename 'S1-S2 reconciliation.xlsx' where S1 = your state abbreviation and S2 = the state with which you are reconciling (if necessary, rename the file to reflect the correct S1 and S2).
12. With cursor in RTS cell A2 (directly under 'Digest' heading), from the 'Edit' menu choose 'Paste Special...' and select 'Values' (or use ribbon buttons to Paste Values).
13. All columns with light green or blue headings should now be populated.
14. Sort worksheet by *DCO_Flag*, *Digest*, *DX_Yr* and *PatID*
 - From the 'Data' menu choose 'Sort'
 - In the *Sort* box, check the box in the upper right to indicate 'My list has headers'
 - In the 'Sort by' row, choose '*DCO_Flag*' under 'Column' and 'Z to A' under 'Order'
 - Click the + button in the lower left to add a 'Then by' row
 - In the 'Then by' row, choose '*Digest*' under 'Column' and either 'A to Z' or 'Z to A' under 'Order'
 - Click the + button in the lower left to add a 'Then by' row
 - In the 'Then by' row, choose '*Dx_Yr*' under 'Column' and 'Largest to Smallest' under 'Order'
 - Click the + button in the lower left to add a 'Then by' row
 - In the 'Then by' row, choose '*PtID*' under 'Column' and 'Smallest to Largest' under 'Order'
 - Click OK
15. SAVE!

Reconciliation & Recording in Reconciliation Tracking Spreadsheet (RTS)

The goal of this phase is to reconcile only past 5 years (2018-2014), beginning with the most recent cases (2018). All the linked cases for 1995-2019 have been provided so if the exercise proves useful, all linked cases can be reconciled. We hope that by capturing five years of data, we can evaluate the utility and time investment of this exercise. We hope, too, that what we learn in this process can inform the next iteration.

Each state will use its own Reconciliation Tracking Spreadsheet (RTS) to record the results of the reconciliation of each case. Each state will then submit to NH only aggregated counts. No individual data will leave your state.

IMPORTANT:

- **DO NOT MAKE CHANGES TO THE FORMAT OR STRUCTURE OF THE WORKBOOK** (adding/rearranging columns will break linkages required to generate the counts)
- **Work ONLY in the 'data collection' worksheet**

1. Open the Reconciliation Tracking Spreadsheet (RTS) you created above: 'S1-S2 reconciliation.xlsx' where S1 = your state and S2 = the state with which you are reconciling.
2. If it is not already sorted by *Digest* then *DX_Yr* then *PtID*, do so now (see step 14 above).

For each case:

3. Determine whether the **patient** is a match. If first name, last name, date of birth, and gender match (which they should), look at SSN and middle name to confirm. Record Y/N outcome in Column J *Pt_Match*. You may type or select from the in-cell drop-down menu.
 - If NO, record "N" in Column J *Pt_Match*.
 - a. Note explanation in Column K *PM_Explan* (e.g.: OP fields match but SSN doesn't).
 - b. STOP. Move on to next case; start this step again for that case.
 - If YES, record "Y" in Column J *Pt_Match*.
 - a. Note explanation in Column K *PM_Explan*. Move on to step 4.
4. If *Pt_Match* = Y, determine whether the **tumor** is a match. Use all available information including site, histology, date of diagnosis. Record one of four outcomes in Column L *Tumor_Match*. The in-cell drop-down menu will only populate for cases where *Pt_Match* = Y. Leave *Tumor_Match* if *Pt_Match* = N.
 - D: Duplicate case between two states
 - UD: Uncertain Duplicate
 - DT: Different Tumor under multiple primary rules
 - UDT: Uncertain Different Tumor
5. For all *Tumor_Match* outcomes, note the explanation in Column M *TM_Explan*.
6. If *Tumor_Match* = D: Determine which state (or both or neither) will keep the case in their database. If a case is a full duplicate, indicate which state keeps the case in their database. The assumption is that the other state will delete it from their database. Record the decision in Column N *TM_Action*. The in-cell drop-down menu will only populate for cases where *Tumor_Match* = D. Leave *TM_Action* empty if *Tumor_Match* ≠ D.

7. All records: Determine whether it is a Death Clearance Only (DCO) case. Record Y/N outcome in Column O *DCO_Flag*.
8. SAVE the file periodically!

Prepare & Return Reconciliation Tracking Spreadsheet to NH

When the reconciliation for which you wish to return counts to NH is complete, please follow these steps to run the counts and 'extract' them from their links to the data in the 'data collection' worksheet. You will send only the aggregate counts to NH.

1. Save the Reconciliation Tracking Spreadsheet document locally.
2. Click into the 'pivot tables' sheet.
3. Put your cursor in cell A1. Excel should now show a 'PivotTable Analyze' menu in the ribbon.
4. Under the 'PivotTable Analyze' menu in the ribbon, click 'Refresh' to pop up a contextual menu. Choose 'Refresh All'.
5. Click the box above the row headings to the left of the column headings to select the entire contents of the worksheet.
6. Copy.
7. Click into the 'values from pivot table' tab.
8. With cursor in cell A1, from the 'Edit' menu choose 'Paste Special...' and select 'Values' (or use ribbon buttons to Paste Values).
9. Right-click the 'values from pivot table' tab to pop up a contextual menu. Choose 'Move or Copy...'
10. In the Move or Copy box, choose '(new book)' under 'To book:' and check the 'Create a copy' box. Click OK.
11. The contents of the 'values from pivot table' tab should now be open in a new workbook.
12. Save the new workbook with filename 'reconciliation_S1_S2_yyyy.mm.dd' where S1 = your state and S2 = the state with which you are reconciling (eg 'reconciliation_NH_RI_2021.04.09').
13. Send 'reconciliation_S1_S2_yyyy.mm.dd' to New Hampshire.

Appendix C to the Phase 2 Protocol

SAS Program to Link Matched File back to Source

```
* Link matched back to source 07-28-2021 ;
/* this program links the file returned to your state by New Hampshire that
contains
the linked cases between your state and another state the file you created using
the 'Internal
SAS Program containing PHI. The output from this file will be imported to an
Excel spreadsheet for
reconciliation.
Your state is labelled 's1' and the other state is labeled 's2'. */

* this is the file returned your state by New Hampshire that contains the
linkages between
two states Be sure to change the file location ;
Options symbolgen mlogic mprint mfile;
%let s1 = NH ;
%let s2 = RI ;
data Linked ;
  infile "Drivename:\folder[s]\&s1_to_&s2_pat_cancer_05042021.txt" dlm = ','
missover DSD lrecl = 100 ;
  format digest $64. ;
  format &s1.stateid $2. ;
  format &s2.stateid $2. ;
input digest $
      &s1.stateid $
      &s2.stateid $ ;
run ;
/* this file was created by the 'Internal SAS program' and contains the PHI for
the
reconciliation. The infile will be the name and location of the file created. */
data source ;
  infile "Drivename:\folder[s]\s1_INT_bypatcancer_07122021.csv" dlm = ','
  missover DSD lrecl = 200 firstobs =2 ;
  /* Digest, n18_80, n18_20, n18_380, n18_400, dxyr, gender, n18_240, typers */
  format      digest $64. ;
format      n18_20 8. ;
format      n18_380 $2.;
format      n18_80 $2. ;
format      p_name  $80. ;
format      address $60. ;
format      gender  $2. ;
format      n18_230 $3. /* age */ ;
format      n18_400 $4. /*site */ ;
format      n18_522 $4. /* histology */ ;
format      n18_410 $1. /* laterality */ ;
format      dxyr    4. ;
format      typers  $12. ;
format      n18_1760 $1. /*vital status */ ;
format      n18_1910 $4. /* cause of death */ ;
format      n18_1750 $8. /* date of last contact */ ;
input      digest $
          n18_20
          n18_380 $
          n18_80 $
          p_name  $
```



```

        address $
        gender $
        n18_230 $
        n18_400 $
        n18_522 $
        n18_410 $
        dxyr
        typers $
        n18_1760 $
        n18_1910 $
        n18_1750 $ ;

run ;
proc print data=source (obs=25) ;
run ;
proc sort data=linked ; by digest ;
run ;
proc sort data=source ; by digest ;
run ;
data result1 ;
    merge linked (in=q1)
           source (in=q2) ; by digest ;
if q1 & q2 ;
run ;
filename relinkx
"Drivename:\folder[s]\relink2&s1.._to_&s2.._pat_cancer_07132021.txt" ;
data _null_ ;
    set result1 ;
    file relinkx ;
put digest ','
    n18_20 ','
    n18_380 ','
    n18_80 ','
    p_name ','
    address ','
    gender ','
    n18_230 ',' /* age */
    n18_400 ',' /*site */
    n18_522 ',' /* histology */
    n18_410 ',' /* laterality */
    dxyr ','
    typers ','
    n18_1760 ',' /*vital status */
    n18_1910 ',' /* cause of death */
    n18_1750 /* date of last contact */ ;
run ;
run ;

```