HERC
Health Economics Resource Center
VA Palo Alto Health Care System
795 Willow Road (152 MPD)
Menlo Park, CA 94025

MEMORANDUM

Date: August 2018

To: QUERI Researchers

From: HERC

Subject: Creating a Finder File

OBJECTIVE

This document provides a guide to creating a finder file. Think of the finder file as "key" that enables researcher to link participants in their study to VA administrative and utilization files. Sometimes the finder file is called a crosswalk file.

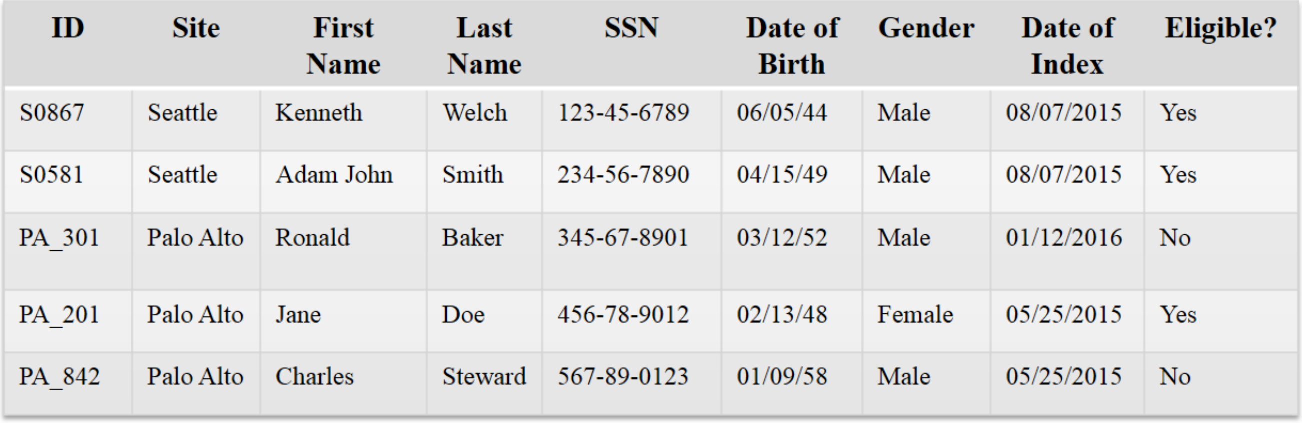
METHODS

Creating a finder file

1. **Finder file structure**
   - The finder file should have 1 record per participant (see Example 1). This file structure can be easily merged to other data tables.
   - The name of the primary linking variable (SSN) must be the same in the finder file as it is in the VA administrative file. The variable type (i.e., numeric, text) must also match.
   - The secondary variables in the finder file should not match the names in the VA administrative data.
2. **Primary linking variable**: You will need to collect the Social Security Numbers (SSNs) to identify study participants. VA administrative data use a range of identifiers including SSN, and scrambled SSN. However, participants only know their SSN. If you are going to collect this information from participants, then you need to ask for their SSN. You will need to have approvals to collect SSNs.[1]
3. **Corroborating variables**: Include other variables to verify that you have identified the correct study participants in the administrative data. Common secondary variables include:
   - Date of birth
   - Gender
4. **Create a unique study ID**. This can be alpha numeric—anything you want—but it needs to be unique to the study and not derived from other unique identifiers.
5. **Verify that the finder file is complete.**
   - There can be no missing data on the primary linking variable.
   - Missing data on corroborating variables is fine.
   - Ensure that there are no duplicate records.
   - Example 2 shows a finder file with errors.

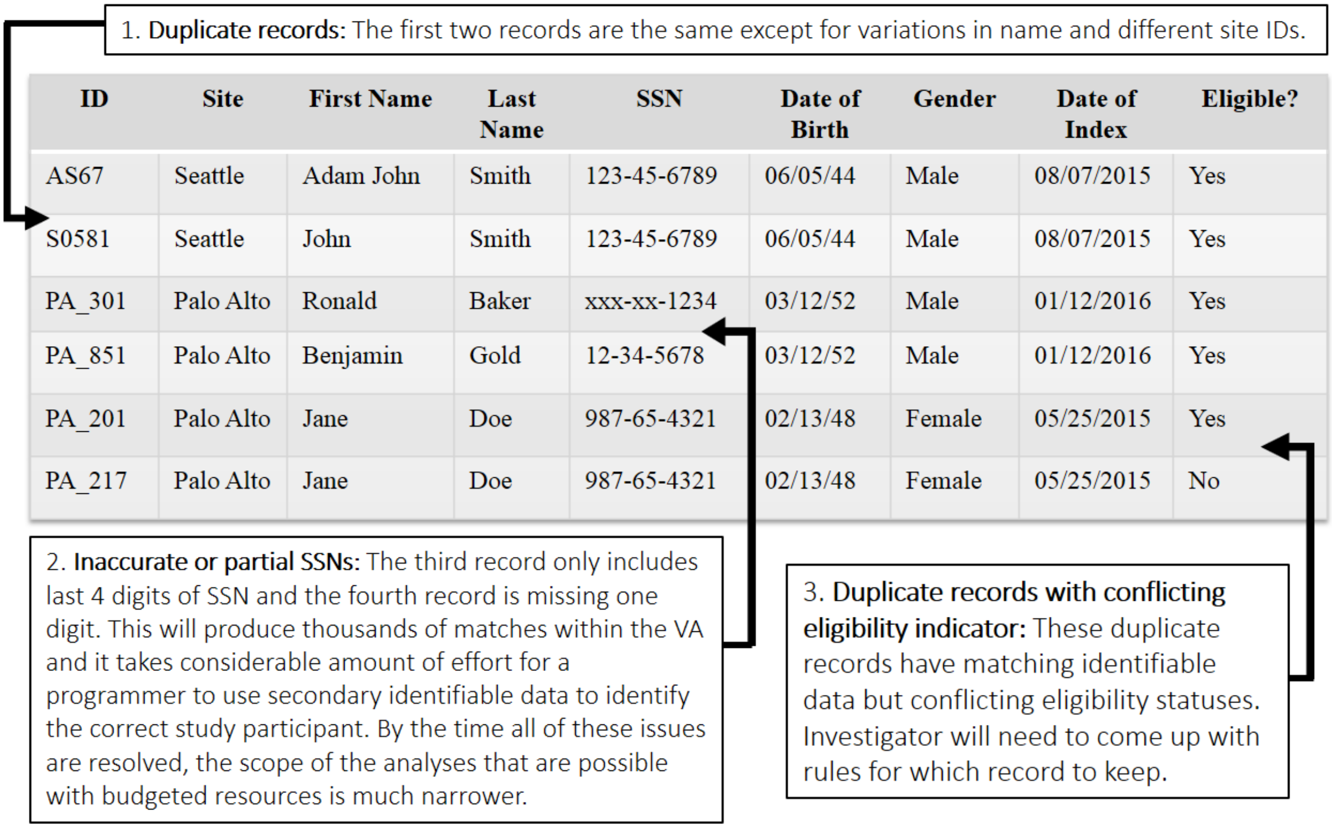
[1] If you don't have the SSN, it is possible to collect a lot of other information, such as date of birth, gender, height, race, ICD-9 diagnoses, and procedure dates, that would allow you to identify a likely match in the administrative data. But this is bad practice, so don't do it.

Example 1: Sample Finder File WITH NO ERRORS



ID	Site	First Name	Last Name	SSN	Date of Birth	Gender	Date of Index	Eligible?
S0867	Seattle	Kenneth	Welch	123-45-6789	06/05/44	Male	08/07/2015	Yes
S0581	Seattle	Adam John	Smith	234-56-7890	04/15/49	Male	08/07/2015	Yes
PA_301	Palo Alto	Ronald	Baker	345-67-8901	03/12/52	Male	01/12/2016	No
PA_201	Palo Alto	Jane	Doe	456-78-9012	02/13/48	Female	05/25/2015	Yes
PA_842	Palo Alto	Charles	Steward	567-89-0123	01/09/58	Male	05/25/2015	No

Example 2: Sample Finder File WITH ERRORS



1. **Duplicate records:** The first two records are the same except for variations in name and different site IDs.

ID	Site	First Name	Last Name	SSN	Date of Birth	Gender	Date of Index	Eligible?
AS67	Seattle	Adam John	Smith	123-45-6789	06/05/44	Male	08/07/2015	Yes
S0581	Seattle	John	Smith	123-45-6789	06/05/44	Male	08/07/2015	Yes
PA_301	Palo Alto	Ronald	Baker	xxx-xx-1234	03/12/52	Male	01/12/2016	Yes
PA_851	Palo Alto	Benjamin	Gold	12-34-5678	03/12/52	Male	01/12/2016	Yes
PA_201	Palo Alto	Jane	Doe	987-65-4321	02/13/48	Female	05/25/2015	Yes
PA_217	Palo Alto	Jane	Doe	987-65-4321	02/13/48	Female	05/25/2015	No

2. **Inaccurate or partial SSNs:** The third record only includes last 4 digits of SSN and the fourth record is missing one digit. This will produce thousands of matches within the VA and it takes considerable amount of effort for a programmer to use secondary identifiable data to identify the correct study participant. By the time all of these issues are resolved, the scope of the analyses that are possible with budgeted resources is much narrower.

3. **Duplicate records with conflicting eligibility indicator:** These duplicate records have matching identifiable data but conflicting eligibility statuses. Investigator will need to come up with rules for which record to keep.

Using a Finder File

A. Match the finder file to the VA administrative data.

- Many VA datasets use encrypted identifiers, such as scrambled SSN.
- It is good practice to create a finder file that has the real SSN and the encrypted identifiers.
- To do this, you will need to merge the finder file to the VA SSN / scrambled SSN crosswalk file.
  a. This merge should be a 1:1 merge
  b. If data errors are found, the records that don't match should be reviewed. Common data errors to look for include missing data (i.e., missing fields) and inaccurate data (i.e., wrong or partial SSN)
  c. When the data merge, it is helpful to corroborate the merge with the corroborating variables (date of birth and gender).
- Once this merge is complete, you will have a finder file with real SSN and the encrypted identifiers, such as scrambled SSN. This is a highly sensitive file. This dataset should be saved – it is the master key. This dataset is considered highly sensitive and so it should be encrypted.
- Use the master key dataset to create a new linking dataset that includes the encrypted identifiers (such as scrambled SSN), but excludes the real SSN.

B. Match the data.

- Merge the finder file with the VA administrative files. The merge should happen using the encrypted identifiers (e.g., scrambled SSN).
- Again, it is good practice to confirm the merge and reconcile any problems. If problems are found in the finder file, then it may be necessary to fix the master key finder file as well.