|  |
| --- |
|  |

|  |
| --- |
| **Updating the Past or Present Hepatitis C Registry - Standard Operating Procedures** |
|  |
|  |
|  |
|  |
| **California Department of Public Health (CDPH)** |
|  |
|  |

Table of Contents

[I. Purpose 1](#_Toc389921331)

[II. CDPH Contacts 1](#_Toc389921332)

[III. Overview 2](#_Toc389921333)

[Data Dictionaries 2](#_Toc389921334)

[Code Repositories in Github 2](#_Toc389921335)

[IV. Standard Operating Procedures 3](#_Toc389921336)

[Step 1: Move Quest data files from the S: (Richmond) drive to the R: (Sacramento) drive 3](#_Toc389921337)

[Step 2: Download and prepare new CalREDIE data 4](#_Toc389921338)

[Step 3: Obtain new AVSS and other local system data 4](#_Toc389921339)

[Step 4: Download SF eFTP Dataset 5](#_Toc389921340)

[Step 5: Update EpiInfo Dataset (Optional) 6](#_Toc389921341)

[Step 6: Run Prepartion Codes 01 Standard\_Header SAS code 6](#_Toc389921342)

[Step 7: Run 02 Quest Lab Dataset Creation SAS code 6](#_Toc389921343)

[Step 8: Run 02b\_CalREDIE Dataset Creation SAS code 7](#_Toc389921344)

[Step 9: Run 02c\_AVSS Dataset Creation SAS code 7](#_Toc389921345)

[Step 10: Run 02d\_SFeFTP Dataset Creation SAS code 7](#_Toc389921346)

[Step 11: Run 02e\_EpiInfo Datasource Creation SAS code (Optional) 8](#_Toc389921347)

[Step 12: Run Data Cleaning and Compilation SAS code 8](#_Toc389921348)

[Step 13: Review the scores in the “trough” in SAS Graph output 8](#_Toc389921349)

[Step 14: Run 07 Deduplication and 08 Final Steps SAS code 9](#_Toc389921350)

[V. References 9](#_Toc389921351)

[I. Appendix A. Steps to Prep and Combine Source Data to Make Main01 and Main02 11](#_Toc389921352)

[II. Appendix B. Definition of Registry SAS Programs 12](#_Toc389921353)

[III. Appendix C. Definition of Permanent Registry Datasets 14](#_Toc389921354)

# 

# Purpose

The purpose of this document is to provide direction and assistance to California Department of Public Health (CDPH) staff who conduct viral hepatitis surveillance. This document will describe the procedures for updating the past or present viral hepatitis C registry on an annual basis.

# CDPH Contacts

The following individuals at CDPH play key roles in viral hepatitis surveillance:

STD Control Branch

* Heidi Bauer, STD Control Branch, Chief
* Michael Samuel, STD Control Branch, Epidemiology and Surveillance Section, Chief
* Joan Chow, STD Control Branch, Epidemiology Unit, Chief
* Rachel McLean, STD Control Branch, Office of Adult Viral Hepatitis Prevention, Chief

Immunization Branch

* Kathleen Harriman, Immunization Branch, Vaccine Preventable Disease Epidemiology Section, Chief
* Erin Murray, Immunization Branch, Epidemiologist Supervisor

# Overview

In 2012, CDPH created a past or present hepatitis C case registry that the STD Control Branch uses to characterize the epidemiology of viral hepatitis in California. The past or present hepatitis C case registry requires a deduplication and matching methodology described in detail in the publication by Glenn Wright, *“Probabilistic Record Linkage in SAS”*[*1*](#_ENREF_1).

## Data Dictionaries

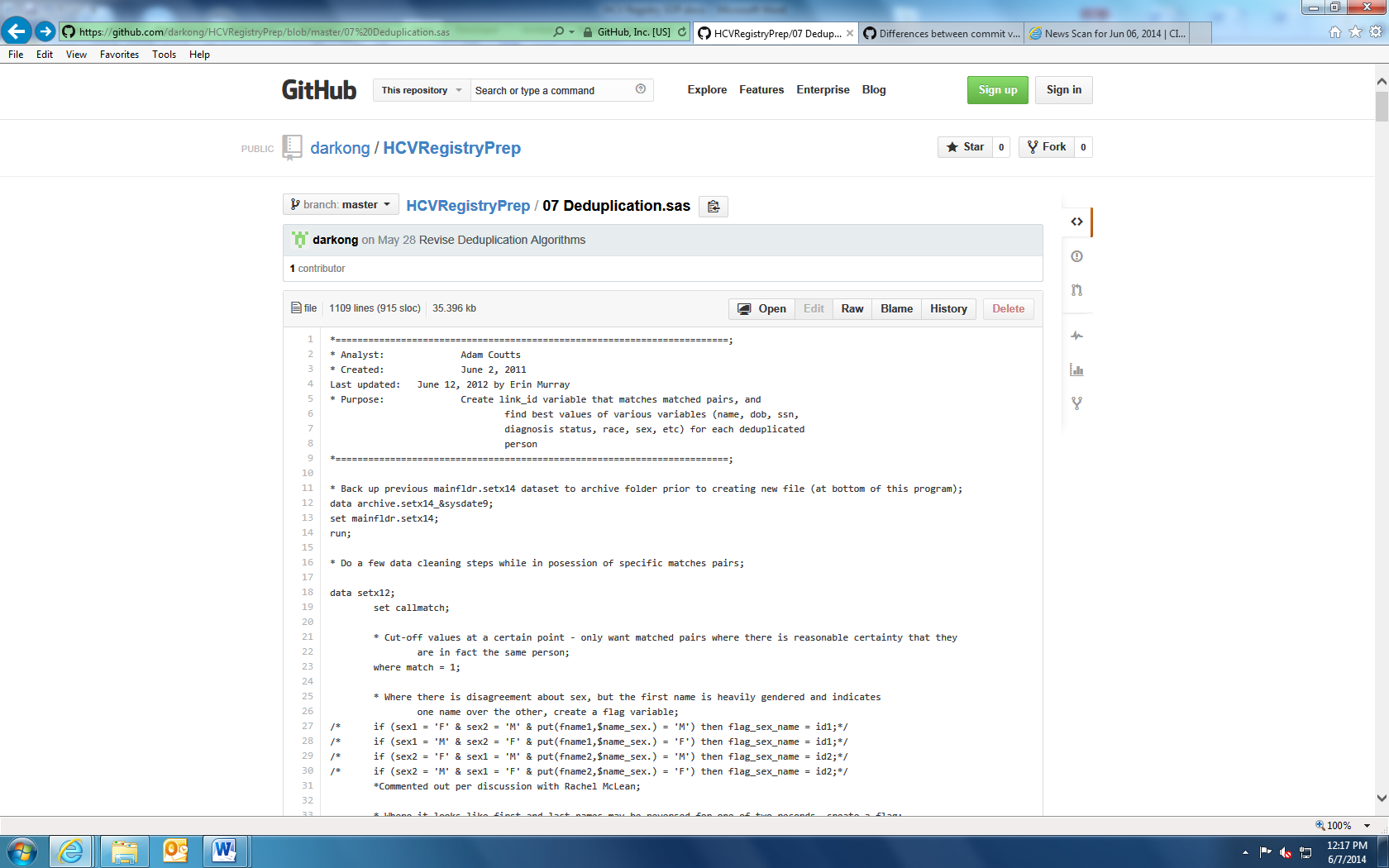
The data dictionary for the past or present hepatitis C case registry can be found here (*R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep)*; variables in the registry will not be explained in detail in this document.

## Code Repositories in Github

The Registry Code (programs 01A-08), the data dictionary, and this SOP are contained in the folder *R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep.* This folder is connected to GitHub which means that any past changes made to the contents are linked to the cloud and you can see past changes and the documented rationale for those changes. You can review the past changes to this folder at the following website:

<https://github.com/darkong/HCVRegistryPrep>

To check the history of the files, click the file you are interested in, and click history:



The site will then show you the dates when the code was changed, will highlight the changes made, and display comments regarding the reason for the changes.

# Standard Operating Procedures

**Step 1:** Move Quest data files from the S: (Richmond) drive to the R: (Sacramento) drive

**Step 2:** Download and prepare CalREDIE data

**Step 3:** Obtain new AVSS and other local system data

**Step 4:** Download SF eFTP dataset

**Step 5:** Update EpiInfo dataset (Optional)

**Step 6:** Run Preparation Codes 01 Standard Header and 01A Macros

**Step 7:** Run 02 Quest Lab Dataset Creation SAS code

**Step 8:** Run 02b\_CalREDIE Dataset Creation SAS code

**Step 9:** Run 02c\_AVSS Dataset Creation SAS code

**Step 10:** Run 02d\_SFeFTP Dataset Creation SAS code

**Step 11:** Run 02e\_EpiInfo Datasource Creation SAS code (optional)

**Step 12:** Run Data Cleaning and Compilation SAS codes

**Step 13:** Review the scores in the “trough” in SAS Graph

**Step 14:** Run 07 Deduplication and 08 Final Steps SAS code

Note: Unlike the HBV registry, new records from the HCV source files are brought in during Steps 1-11 in the actual HCV registry code. In Step 12 these new records are compiled into set 101, which is then appended to the old registry and Cartesian Matching is performed. These new records are matched to new records and to old records but old records are not matched to old records. Because matching old records to old records has been previously done to create the old registry, it was determined to be unnecessary and would only increase processing time. Steps 13-14 identify likely duplicates among the matched pairs, resolve discrepant data among matches, and outputs the final datasets main01 and main02.

Because the HCV registry code keeps the new records separate from the old, it therefore allows Cartesian matching to be restricted so that old data is not matched again to old data. This means that link\_ids in the HCV registry will have more consistency from year to year. in contrast, the link\_ids in the HBV registry will not be consistent from year to year.

Lastly, one other major difference between the HCV and HBV registries is that the HBV registry includes acute cases from 2010 to the present. The HCV registry only includes chronic cases.

## Step 1: Move Quest data files from the S: (Richmond) drive to the R: (Sacramento) drive

***Inputs:*** S:\State Surveillance\Labs\Quest and Foundation Lab Reports\Lab Data\Quest Data,

2010 – Present\Final\_CADOH\_HCV\_Signalmmmyyyy.xls

S:\State Surveillance\Labs\Quest and Foundation Lab Reports\Lab Data\Quest Data,

2010 – Present\FINAL\_CADOH\_HCVmmmyyyy.xls

***Outputs:*** R:\State Surveillance\Quest and Foundation Lab Reports\Lab Data\Quest Data,

2010 - Present\yyyy\Final\_CADOH\_HCV\_Signalmmmyyyy.xls

R:\State Surveillance\Quest and Foundation Lab Reports\Lab Data\Quest Data,

2010 - Present\yyyy\FINAL\_CADOH\_HCVmmmyyyy.xls

***Run:***

1. Rachel McLean receives a notification from Quest whenever new data files are available from Quest. She downloads the files from their secure server to the following location: S:\State Surveillance\Labs\Quest and Foundation Lab Reports\Lab Data\Quest Data, 2010 – Present.
2. Copy all new monthly data files received from the S: drive to the following location on the R: drive: R:\ State Surveillance\Quest and Foundation Lab Reports\Lab Data\Quest Data, 2010 - Present\yyyy
3. There are two Quest data files for each month, one containing all Hepatitis C tests and a second containing the signal to cut-off ratios for reactive HCV antibody tests. Be sure to move both files for each month.
4. Do not change the file names.

Note: mmm = first three letters of month; yyyy = four digit year

## Step 2: Download and prepare new CalREDIE data

***Inputs:*** None

***Outputs:*** R:\State Surveillance\CalREDIE\2012-2013 Hep C Data\HCVJulxxxxJunyyyy.tsv

**Run:**

Download from the Data Distribution Portal all the system tab export for all Chronic HCV Incidents using the start date of Jul 1 of the previous year to June 30 of the current year. This will give the system tab info for all chronic HCV incidents. The file will download as a tsv file.

Note: xxxx = previous year, yyyy = current year

## Step 3: Obtain new AVSS and other local system data

***Inputs:*** None

***Outputs:*** R:\State Surveillance\HBV Data Analysis\HBV Match 2012\Datasets\Source Data\Morb\NonCalREDIE Cumulative\XXXX

***Run:***

This file is created by completing Steps 1 and 2 of the HBV SOP found here: R:\State Surveillance\HBV Registry\HBVRegistryPrep\HBV Registry SOP.docx

Note: xxxx = previous year

## Step 4: Download SF eFTP Dataset

***Input:*** None

***Output:*** R:\State Surveillance\SF County eFTP

Data\cdhs\_sf\_chronic\_c\_dlmmddyyyy.sas7bdat

Note: mm = two digit month; dd = two digit day; yyyy = four digit year

***Run:***

1. If you do not already have access, request permission to the SF eFTP site from Rachel McLean ([Rachel.McLean@cdph.ca.gov](mailto:Rachel.McLean@cdph.ca.gov)) or Kathleen Harriman at [Kathleen.Harriman@cdph.ca.gov](mailto:Kathleen.Harriman@cdph.ca.gov), depending on who your supervisor is.
2. Log onto the SF eFTP site (<https://eft.dhcs.ca.gov/EFTClient/Account/Login.htm>) using your username and password.
3. Download the SF SAS Dataset cdhs\_sf\_chronic\_c.sas7bdat to the following location: R:\State Surveillance\SF County eFTP Data, using the following naming convention: cdhs\_sf\_chronic\_c\_dlmmddyyyy.sas7bdat, where mm = two digit month; dd = two digit day; yyyy = four digit year
   * It is possible the a file with the name cdhs\_sf\_chronic\_c.sas7bdat will not be on the eFTP site
   * If this is the case, download the file called sf\_cases\_chronicc.sas7bdat to the following location: R:\State Surveillance\SF County eFTP Data, using the following naming convention: sf\_cases\_chronicc\_dlmmddyyyy.sas7bdat, where mm = two digit month; dd = two digit day; yyyy = four digit year
4. Make sure to check that names are included in the downloaded dataset
   * The sf\_cases\_chronicc.sas7bdat has not contained patient names in the past
5. If the names are not included, contact Amy Nishimura at SFDPH at: [amy.nishimura@sfdph.org](mailto:amy.nishimura@sfdph.org) to request a new dataset with both first and last names
   * Once SF has indicated that the dataset is available, start over at step 2.

Note: If your username and password for the eFTP site does not work and the online “forgot password” reset also does not work, then email Van Randon ([Van.Randon@dhcs.ca.gov](mailto:Van.Randon@dhcs.ca.gov)) and cc: Walter Shinn ([Walter.Shinn@cdph.ca.gov](mailto:Walter.Shinn@cdph.ca.gov)) and request that your password be reset.

## Step 5: Update EpiInfo Dataset (Optional)

***Input*:** None

***Output:***

G:\CDPH\DCDC\STD\Groups\Office of Adult Viral Hepatitis Prevention\Surveillance\State Surveillance\CMRs\HEP\_C.DBF

***Run:***

Add any data from loose CMRs to the EpiInfo file above. However, the EpiInfo dataset is a one time addition that should be included in the 2013 data as it was not added in the 2012 update. But after 2013 update it should not be needed any longer

## Step 6: Run Preparation Codes 01 Standard\_Header SAS code

***File Locations:***

R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep\01 Standard Header.sas

R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep\01A Macros.sas

**Run:**

Run in the following order:

1. 01 Standard Header
2. 01A Macros

## Step 7: Run 02 Quest Lab Dataset Creation SAS code

***File Locations:***

R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep\02 Quest Lab Dataset Creation.sas

***Run:***

1. Update the highlighted macro variables at the very beginning of the program to reflect the appropriate years

%let n=2012; \*year for July-December data files;

%let m=2013; \*year for January-June data files;

n should reflect the previous year and m should reflect the current year.

1. Save and run the revised program.

## Step 8: Run 02b\_CalREDIE Dataset Creation SAS code

***File Location:***

R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep\02b\_CalREDIE Dataset Creation.sas

***Run:***

1. Update the highlight macro variables at the very beginning of the program to reflect the appropriate years

\*Assign year to macro variable for easy changing from year to year;

%let y1=2012; \*year for July-December data files;

%let m1=Jul;\*months for July-December data files;

%let y2=2013; \*year for January-June data files;

%let m2=Jun; \*months for January-June data files;

n should reflect the previous year and m should reflect the current year.

1. Save and run the revised program.

## Step 9: Run 02c\_AVSS Dataset Creation SAS code

***File Location:***

R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep\02c\_AVSS Dataset Creation.sas

***Run:***

1. Update the highlighted macro variable at the very beginning of the program to be the **previous** year

%let year = 2012;

1. Save and run the revised program.

## Step 10: Run 02d\_SFeFTP Dataset Creation SAS code

***File Location:***

R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep\02d\_SFeFTP Dataset Creation.sas

***Run:***

1. Update the highlighted macro variables at the very beginning of the program to reflect the appropriate files

\* Enter the dataset that was used during the previous run;

%let ds1=sfeftp.Cdhs\_sf\_chronic\_c\_dl04022012;

\* Enter the new dataset;

%let ds2=sfeftp.Cdhs\_sf\_chronic\_c\_dl06072012;

In theory, the file listed as ds2 should become ds1 and the newly downloaded file from Step 4 should become ds2

1. Save and run the revised program.

## Step 11: Run 02e\_EpiInfo Datasource Creation SAS code (Optional)

***File Location:***

R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep\02e\_EpiInfo Dataset Creation.sas

***Run:***

Run the above program. If Step 5 was needed, then run this code.

## Step 12: Run Data Cleaning and Compilation SAS code

***File Location:***

R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep\03 Datasource Preparation.sas

R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep\04 Prison Macro.sas

R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep\05 Merging.sas

R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep\06 Matched Pairs.sas

***Run:***

Run the above programs in the following order:

1. 03 Datasource Preparation

If Step 5 was needed then create set007

1. 04 Prison Macro
2. 05 Merging
3. 06 Matched Pairs – See **Step 13** below for details on reviewing the output

## Step 13: Review the scores in the “trough” in SAS Graph output

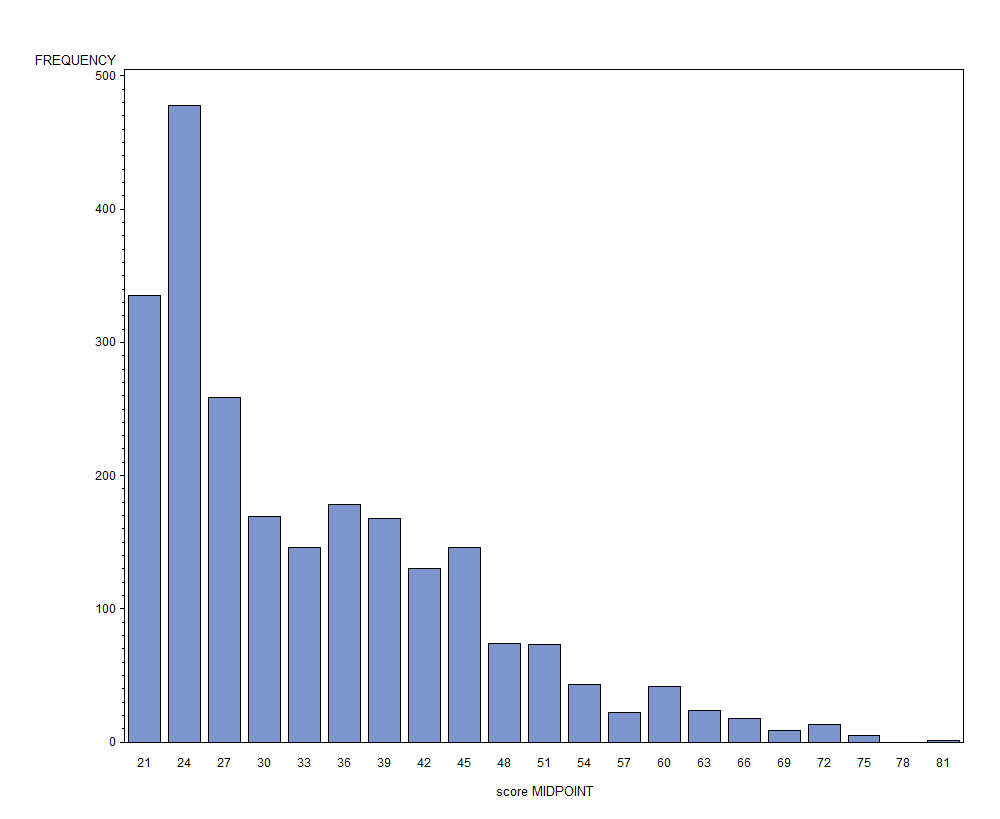
***File Location:***

R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep\06 Matched Pairs.sas

***Run:***

1. In program 06 Matched Pairs, SAS Graph will produce a distribution curve of the scores from the linked pairs dataset (Chart 1)
2. Identify the scores that reflect the dip before the bell curve (trough)
3. Confirm that the scores that contain the trough are close to 28.5 – 33. The scores do not have to be identical to the trough, but should be roughly the same.
4. If the trough scores are dramatically different, this is most likely due to error in the newly uploaded datasets. The Epidemiologist will need to review the updated data to look for data entry errors, make corrections, and start the process over.

**Chart 1. Printout of SAS Graph is Program, “06 Matched Pairs”**



The bars to the right of the trough represent cases that matched

The arrows indicate the scores of the “trough” before the rise where cases are deemed a match. These scores identify the range of scores with potential matches that are further analyzed in the SAS program

The bars to the left of the trough represent cases that did not match

## Step 14: Run 07 Deduplication and 08 Final Steps SAS code

***File Location:***

R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep\07 Deduplication.sas

R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep\08 Final Steps.sas

***Run:***

Run the above programs in the following order:

1. 07 Deduplication
2. 08 Final Steps

# References

1. Wright G. Probabilistic Record Linkage in SAS. *SAS Users Group*. San Francisco: Western Users of SAS Software (WUSS); 2011.

A copy of the paper can be found here: *S:\State Surveillance\Surveillance Analyses References*

# Appendix A. Steps to Prep and Combine Source Data to Make Main01 and Main02

old main02

AVSS

CalREDIE

SFeFTP

Quest

EpiInfo\*

set101

Main01

Main02

= datasets with 1 year’s worth of data

= datasets with multiple years of data

Dataset locations\*\*:

|  |  |
| --- | --- |
| Quest | R:\State Surveillance\Quest and Foundation Lab Reports\Lab Data\Quest Data, 2010 - Present\yyyy |
| CalREDIE | R:\State Surveillance\CalREDIE\yyyy-(yyyy-1) Hep C Data |
| AVSS | R:\State Surveillance\HBV Data Analysis\HBV Match 2012\Datasets\Source Data\Morb\NonCalREDIE Cumulative\yyyy\nonCalREDIEcumul.sas7bdat |
| SFeFTP | R:\State Surveillance\SF County eFTP Data |
| EpiInfo\* | G:\Surveillance\State Surveillance\CMRs\HEP\_C.dbf |
| Set101 | Created in Step 12 |
| Old Main02 | Main02 from prior update |
| Main01 | R:\State Surveillance\Adam Coutts Code Folder\Datasets\main01.sas7bdat |
| Main02 | R:\State Surveillance\Adam Coutts Code Folder\Datasets\main02.sas7bdat |

\* EpiInfo dataset is a one time addition that should be included in the 2013 data as it was not added in the 2012 update. But after 2013 update it should not be needed any longer

\*\*yyyy = current year, yyyy-1 = prior year

# Appendix B. Definition of Registry SAS Programs

* **01 Standard\_Header:** This program assigns the directories and runs all of the formats that will be used in later programs.
* **01A Macros:** This program installs macros used in later programs.
* **02 Quest Dataset Creation:** This program reads in and formats the Quest laboratory data.
* **02b CalREDIE Dataset Creation:** This program reads in and formats the CalREDIE data.
* **02c AVSS Dataset Creation:** This program reads in and formats the morbfile and AVSS laboratory (HCV and PCRHCV variables) data.
* **02d SFeFTP Dataset Creation:** This program reads in and formats the SF eFTP data.
* **02e EpiInfo Dataset Creation:** This program reads in and formats the CMRs that were entered into EpiInfo. As no more CMRs are expected to be entered, this program is may not need to be run.
* **03 Datasource Preparation:** This program revises and standardizes the code to merge all of the source datasets together (Quest, CalREDIE, morbfile, and SF eFTP).
* **04 Prison Macro:** This program calls in the prison macro which determines whether a given case was reported from a state prison.
* **05 Merging:** This program merges source datasets into one main dataset, processes the data, and recodes variables in preparation for the match.
* **06 Matched Pairs:** This program creates matched pairs and “Linked Pairs” SAS datasets of cases to deduplicate in the next program. This program applies scores to potential matched variables; Lastname, Firstname, Date of Birth, Social Security Number, Prison status, and geography.
* **07 Deduplication:** This program creates a “link\_id” variable that links matched pairs and finds the best value for the following variables: name, date of birth, date of death, social security number, diagnosis status, race (if available), sex, prison, and location for each individual case. This methodology is described in more detail in the paper, *“Probabilistic Record Linkage in SAS”*[*1*](#_ENREF_1).
* **08 Final Steps:** This program creates the final datasets: main01 with reports aggregated into person-level data and main02 with event level data and the original demographic information associated with each report.

# Appendix C. Definition of Permanent Registry Datasets

The permanent registries are as follows:

* **Main01:** This is the permanent past or present HCV Registry and is a SAS dataset of cumulative past or present HCV deduplicated person-level data. Each record represents an individual reported with a past or present HCV and each individual is represented once in the dataset.

* **Main02:** This is the permanent past or present HCV Registry and is a SAS dataset of cumulative past or present HCV duplicated event-level data. Each record represents a reported case of past or present HCV. Multiple records corresponding to the same individual indicate that one individual was reported with past or present HCV multiple times.

Data dictionaries for these registries are located here: R:\State Surveillance\Chronic HCV Registry\HCVRegistryPrep\Hep C\_Data Dictionary.docx