## STDCB Hepatitis B Registry Data Dictionary

## General note

\* For many variables, basic standardization was performed, including but not limited to

\* Punctuation characters stripped out where not needed

\* Double spaces compressed to single spaces

\* Trailing and leading blanks cut out

\* Capitalization standardized (usually, but not always, to all capital letters)

========================================================================================

========================================================================================

## Datasets: main01\_chronichbv

- MasterHBV\_Chronic: Deduplicated, person-level records of Chronic Hepatitis B

**(1) link\_id**

- 8 numbers

- Unduplicated, person-level id variable - derived from the in-house created ID variable - Used to link main01 and main02 records

- Please note that as the main01 and main02 files get updated, the link\_id may change.

**(2) acute**

- 1 number

- Indicates whether patient had an acute hepatitis B report

- Patients can have both chronic and acute reports

**(3) chronic**

- 1 number

- Indicates whether patient had a chronic hepatitis B report

- Patients can have both chronic and acute reports

**(4) best\_first\_name**

- 20 characters

- First name of patient

\* Variable is an amalgamation of:

\* San Francisco County public health data system variable FNAME

\* The CalREDIE variable FIRSTNAME

\* The Morbfile variable FNA

\* 'SR','JR','II','III','IV' stripped off of end

\* If missing and middle name is not missing, or if first name is a substring of middle name (i.e. "M" and "Mary"), then what was middle name becomes first name

\* If a person clearly has a first name for a last name and a last name for a first name (something extreme like "Peterson Jennifer"), then names are reversed

\* If a case has the same name for first and last name, and a matched case (found in deduplicating process) indicates that first name is the one that should be changed, then first name is changed to be the same as that of the matched case

\* If comparison with a matched case (found in deduplicating process) indicates that first and last names are reversed, then they are un-reversed

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), if there is disagreement as to FIRST\_NAME among various episode-level records, take the most common FIRST\_NAME - \*unless\*,

- A different value starts with the current best name choice but is longer

- The current best name choice is four or more char and is contained within the comparison value

- The current best name choice is short, and the comparison value is considerably longer

- If percentage is tied and no other rules apply, take the more common spelling

**(5) best\_last\_name**

- 35 characters

- Last name of patient

\* Variable is an amalgamation of:

\* San Francisco County public health data system variable LNAME

\* The CalREDIE variable LASTNAME

\* The Morbfile variable LNA

\* Obvious misspellings of common names were corrected

\* 'SR','JR','II','III','IV' stripped off of end

\* If first name has two "words" and last name is missing, then the second word becomes last name

\* If a person clearly has a first name for a last name and a last name for a first name (something extreme like "Peterson Jennifer"), then names are reversed

\* If last name is a substring of middle name (i.e. "R" and "Rogers"), then what was middle name becomes last name

\* If a cases has the same name for first and last name, and a matched case (found in deduplicating process) indicates that last name is the one that should be changed, then last name is changed to be the same as that of the matched case

\* If comparison with a matched case (found in deduplicating process) indicates that first and last names are reversed, then they are un-reversed

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), if there is disagreement as to LAST\_NAME among various episode-level records, take the most common LAST\_NAME - \*unless\*,

- A different value starts with the current best name choice but is longer

- The current best name choice is four or more char and is contained within the comparison value

- The current best name choice is short, and the comparison value is considerably longer

- If percentage is tied and no other rules apply, take the more common spelling

**(6) best\_middle\_name**

- 20 characters

- Middle name (or, more often, initial) of patient

\* Variable is an amalgamation of:

\* The CalREDIE variable MIDDLENAME

\* The Morbfile variable MNA

\* If first name has two "words" and middle name is missing, then the second word becomes middle name (if last name and middle name are missing, then the second name becomes the last name)

\* If middle name is same or a subset of first or last name, then middle name deleted

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), if there is disagreement as to MIDDLE\_NAME among various episode-level records, take the most common MIDDLE\_NAME - \*unless\*,

- A different value starts with the current best name choice and then is longer

- The current best name choice is four or more char and is contained within the comparison value

- The current best name choice is short, and the comparison value is considerably longer

- If percentage is tied and no other rules apply, take the more common spelling

\* For middle names, if there is disagreement between possibilities and one value matches the first letter of the chosen best first name, then delete that value;

\* In situations where (1) middle name is missing or it is an initial substring (usually, first initial), of an alternate first name, and (2) there is a value for first name that is clearly different (not a nickname, not linguistically similar, not an initial substring) than the first name which was taken), then save the first name that is an initial, or less common, as the middle name

**(7) best\_ssn**

- 9 numbers

- Social security number

\* Variable is an amalgamation of:

\* The Morbfile variable SSN, XSSN

\* The CalREDIE variable SSN

\* Values that are clearly missing (i.e., "000000000","000000001","999999999") set to missing

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), then any ssn that is uncontested (i.e., the only non-missing value) or that is present for at least seventy-five percent of episodes is chosen as the final value.

**(8) best\_sex**

- 1 character

- Sex/gender of patient

\* Variable is an amalgamation of:

\* The San Francisco County public health data system variable SEX

\* The CalREDIE variable SEX

\* The Morbfile variable SEX

\* Value of "O" ("Other") recoded to "T" ("Transgender")

\* All values besides "M", "F", and "T" set to "U" ("Unknown")

\* If sex is missing and a record shows incarceration in a specifically male-only or female-only state prison, then sex is set to that value

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), then whichever gender appears more than half the frequency is chosen as the final value. If there is a tie, then assign SEX as missing.

\* If any episode-level value is "T", then person-level value is set to "T"

**(9) best\_race\_ethnicity**

- 35 characters

- Race/ethnicity of patient

\* Variable is an amalgamation of:

\* Calculated from San Francisco County public health data system variables ETHNIC\_GRP, RACE\_AMIND, RACE\_ASIAN, RACE\_BLACK, RACE\_OTHER, RACE\_PACISLAND, and RACE\_WHITE

\* Calculated from the CalREDIE and Morbfile variables ETHNICITY and RACE

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), if there was disagreement between race/ethnicity values, then a value was taken in ascending order of this ranking of specificity :

**(10) best\_date\_of\_birth**

- Date variable

- Date of birth

\* Variable is an amalgamation of:

\* San Francisco County public health data system variable birth\_dt

\* The CalREDIE variable DOB

\* The Morbfile variable DOB

\* If DATE\_OF\_BIRTH is over 110 years before the date that program "05 Merging.sas" was run, then 100 years is added. After that, if DATE\_OF\_BIRTH is before January 1, 1900, or after the date that program "05 Merging.sas" was run, then DATE\_OF\_BIRTH is set to missing.

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), then the most common day-of-the-month of birth, the most common month of birth, and the most common year of birth are combined to create final person-level date of birth.

\* If the patient is incarcerated in a state prison, but their calculated age is <14 then date\_of\_birth is set to missing

**(11) age**

- Numeric variable

- Calculated age of patient at first report [(FIRSTDATE-DATE\_OF\_BIRTH)/365.25]

\* The 365.25 accounts for leap years

**(12) best\_date\_of\_death**

- Date variable

- Date of death

\* Variable is an amalgamation of:

\* The CalREDIE variable DTDEATH

\* The Morbfile variable DTH

\* If DATE\_OF\_DEATH is same as DATE\_OF\_BIRTH, then DATE\_OF\_DEATH set to missing

\* If DATE\_OF\_DEATH is over 100 years before the date that program "05 Merging.sas" was run, then 100 years is added. After that, if DATE\_OF\_DEATH is before January 1, 1900, or after the date that program "05 Merging.sas" was run, then DATE\_OF\_DEATH is set to missing.

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), take any DATE\_OF\_DEATH this exists for seventy percent or more of non-missing episode-level records. If such a value does not exist, then set DATE\_OF\_DEATH to missing.

**(13) firstyear**

- Numeric variable

- Year of first entry into system, excluding diagnosis date - calculated during data process [year(firstdate)]

**(14) firstdate**

- Date variable

- Date of first entry into system for either chronic or acute hepatitis B, excluding diagnosis date - calculated during data process

\* Variable is calculated first for each episode as the earliest/first/smallest out of [EPISODE\_DATE\_1, EPISODE\_DATE\_2, EPISODE\_DATE\_3, EPISODE\_DATE\_4, COLLECTION\_DATE, RESULT\_DATE].

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), then the earliest FIRSTDATE value is taken for the person-level record.

\* This variable represented the earliest date associated with the case being first reported to public health for Chronic HBV cases.

**(15) firstdate\_ch**

- Date variable

- Date of first entry into system for chronic hepatitis B, excluding diagnosis date - calculated during data process

\* Variable is calculated first for each episode as the earliest/first/smallest out of [EPISODE\_DATE\_1, EPISODE\_DATE\_2, EPISODE\_DATE\_3, EPISODE\_DATE\_4, COLLECTION\_DATE, RESULT\_DATE].

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), then the earliest FIRSTDATE value is taken for the person-level record.

\* This variable represented the earliest date associated with the case being first reported to public health as a chronic hepatitis B case.

**(16) firstdate\_ac**

- Date variable

- Date of first entry into system for acute hepatitis B, excluding diagnosis date - calculated during data process

\* Variable is calculated first for each episode as the earliest/first/smallest out of [EPISODE\_DATE\_1, EPISODE\_DATE\_2, EPISODE\_DATE\_3, EPISODE\_DATE\_4, COLLECTION\_DATE, RESULT\_DATE].

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), then the earliest FIRSTDATE value is taken for the person-level record.

\* This variable represented the earliest date associated with the case being first reported to public health as a chronic hepatitis B case.

**(17) dxdate**

- Date variable

- Date of first entry into system - calculated during data process

\* Variable is calculated first for each episode as the earliest/first/smallest out of [EPISODE\_DATE\_1, EPISODE\_DATE\_2, EPISODE\_DATE\_3, EPISODE\_DATE\_4, COLLECTION\_DATE, RESULT\_DATE, DIAGNOSIS\_DATE].

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), then the earliest FIRSTDATE value is taken for the person-level record.

\* This variable was created because we don’t particularly trust the diagnosis date variable and also because the date of diagnosis could legitimately precede the date of entry into our system.

**(18) dxdate\_ch**

- Date variable

- Date of first entry into system for chronic hepatitis B - calculated during data process

\* Variable is calculated first for each episode as the earliest/first/smallest out of [EPISODE\_DATE\_1, EPISODE\_DATE\_2, EPISODE\_DATE\_3, EPISODE\_DATE\_4, COLLECTION\_DATE, RESULT\_DATE, DIAGNOSIS\_DATE].

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), then the earliest FIRSTDATE value is taken for the person-level record.

\* This variable was created because we don’t particularly trust the diagnosis date variable and also because the date of diagnosis could legitimately precede the date of entry into our system.

**(19) dxyear**

- Numeric variable

- Date of first entry into system - calculated during data process [year(dxyear)]

**(20) agedx**

- Numeric variable

- Calculated age of patient at first diagnosis [(DXDATE-DATE\_OF\_BIRTH)/365.25]

\* The 365.25 accounts for leap years

**(21) prison\_ever**

- 1 character

- Has this person ever been in prison

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), then if any value of variable PRISON is "M" (i.e., male state prison) or "F" (i.e., female state prison), then PRISON\_EVER is "S" (has served time in state prison). If this is not the case, but any value of "O" (i.e., has been in prison/jail, but cannot say for certain if it is a state prison) exists in any record for this person, then PRISON\_EVER is set to "O". Otherwise, PRISON\_EVER defaults to the value "-".

**(22) prison\_firstrpt**

- 1 character

- Was this person in prison at the time they were first reported to public health

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), then if the value of variable PRISON for the patient’s record with the earliest FIRSTDATE is "M" (i.e., male state prison) or "F" (i.e., female state prison), then PRISON\_FIRSTRPT is "Y" (was serving time in state prison at the time they were first reported to public health). Otherwise, PRISON\_FIRSTRPT defaults to the value "-".

**(23) main\_diagnosis**

- 1 number

- The most comprehensive laboratory-based diagnosis we have received for this person - derived from taking the lowest number for a person of the DIAGNOSIS2 variable from main02 dataset

- This variable is not being use for Hepatitis B – empty until we use lab data from CalREDIE

\* If a person has more than one episode with a positive band on RIBA test (i.e. RESULT\_NAME is in ["BAND PATTERN 5-1-1 (p)/cl00 (p)","C22P","C33C","NS5"], and RESULT\_VALUE is "REACTIVE"), then that person's MAIN\_DIAGNOSIS becomes either "1" or "2" (depending on whether the accession number was reported to the state or not, respectively).

**(24) overall\_diagnosis**

- 1 number

- The most comprehensive diagnosis we have ever received for this person - derived from taking the lowest number for a person of the DIAGNOSIS2 variable from main02 dataset or if the variable DIAGNOSIS from the main02 dataset is “CONFIRMED”, then the value of OVERALL\_DIAGNOSIS is set to “1”.

- This variable is not being use for Hepatitis B – empty until we use lab data from CalREDIE

\* If a person has more than one episode with a positive band on RIBA test (i.e. RESULT\_NAME is in ["BAND PATTERN 5-1-1 (p)/cl00 (p)","C22P","C33C","NS5"], and RESULT\_VALUE is "REACTIVE"), then that person's OVERALL\_DIAGNOSIS becomes either "1" or "2" (depending on whether the accession number was reported to the state or not, respectively).

**(25) records\_per**

- 4 numbers

- The number of event-level records that are affiliated with this person

**(26) first\_lhj**

- 20 characters

- The local health jurisdiction at the time of first contact

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), this variable is created as the value of the LOCAL\_HEALTH\_JURISDICTION at the time of the record with the earliest FIRSTDATE value.

**(27) first\_city**

- 25 characters

- The city at the time of first contact

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), this variable is created as the value of the LOCAL\_HEALTH\_JURISDICTION at the time of the record with the earliest FIRSTDATE value.

**(28) common\_lhj**

- 20 characters

- The local health jurisdiction the patient is most commonly associated with

\* If 2 lhjs are equally most common (ie: 50% from lhj 1 and 50% from lhj 2) then this field is blank

**(29) common\_city**

- 20 characters

- The city the patient is most commonly associated with

\* If 2 cities are equally most common (ie: 50% from city 1 and 50% from city 2) then this field is blank

**(30) country**

- 13 characters

- Country of origin of patient; US or Other

\* Variable is an amalgamation of:

\* Calculated from San Francisco County public health data system variable birth\_country

\* Calculated from the CalREDIE and Morbfile variables COUNTRY

========================================================================================

========================================================================================

## Datasets: main02\_chronichbv

- Event-level records (often more than one per person)

- Main02\_chronicHBV: event level records of chronic HBV cases

**(1) link\_id**

- 8 numbers

- Unduplicated, person-level id variable - derived from the in-house created ID variable

- Used to link main01 and main02 records - often the same as the ID (for people we have only a single record for, or for the root/base record of a person we have multiple records for) but sometimes different (for non-lead linked records for people for multiple records)

**(2) first\_name**

- 20 characters

- First name of patient as in the original record

\* 'SR','JR','II','III','IV' stripped off of end

\* If missing and middle name is not missing, or if first name is a substring of middle name (i.e. "M" and "Mary"), then what was middle name becomes first name

\* If a person clearly has a first name for a last name and a last name for a first name (something extreme like "Peterson Jennifer"), then names are reversed

\* If a case has the same name for first and last name, and a matched case (found in deduplicating process) indicates that first name is the one that should be changed, then first name is changed to be the same as that of the matched case

\* If comparison with a matched case (found in deduplicating process) indicates that first and last names are reversed, then they are un-reversed

**(3) last\_name**

- 35 characters

- Last name of patient in the original record

\* 'SR','JR','II','III','IV' stripped off of end

\* If missing and middle name is not missing, or if first name is a substring of middle name (i.e. "M" and "Mary"), then what was middle name becomes first name

\* If a person clearly has a first name for a last name and a last name for a first name (something extreme like "Peterson Jennifer"), then names are reversed

\* If a case has the same name for first and last name, and a matched case (found in deduplicating process) indicates that first name is the one that should be changed, then first name is changed to be the same as that of the matched case

\* If comparison with a matched case (found in deduplicating process) indicates that first and last names are reversed, then they are un-reversed

**(4) middle\_name**

- 20 characters

- Middle name (or, more often, initial) of patient in the original record

\* 'SR','JR','II','III','IV' stripped off of end

\* If missing and middle name is not missing, or if first name is a substring of middle name (i.e. "M" and "Mary"), then what was middle name becomes first name

\* If a person clearly has a first name for a last name and a last name for a first name (something extreme like "Peterson Jennifer"), then names are reversed

\* If a case has the same name for first and last name, and a matched case (found in deduplicating process) indicates that first name is the one that should be changed, then first name is changed to be the same as that of the matched case

\* If comparison with a matched case (found in deduplicating process) indicates that first and last names are reversed, then they are un-reversed

**(2) id**

- 8 numbers

- A variable (created during our data processing) that is unique to an individual record. It may be the same as link\_id (for people we have only a single record for, or for the root/base record of a person we have multiple records for)

**(3) data\_source**

- 22 characters

- Explains which data provider/system an individual record came from

**(4) occupation**

- 25 characters

- Occupation of patient (practically, most of this information when present is "inmate", and is mostly missing for other occupations)

\* Variable is an amalgamation of:

\* Calculated from the CalREDIE variables OCCLOCATION, OCCSETTINGSPEC, OCCSETTINGTYPE, OCCSPECIFY, and OCCUPATION (all contain minimal data)

\* The Morbfile variable OCCUPATION

\* If OCCUPATION is missing and PRISON is "M", "F", or "O", then OCCUPATION is set to "Inmate"

**(5) date\_of\_onset**

- Date variable

- Date of onset of hepatitis C

\* Variable is an amalgamation of:

\* The CalREDIE variable DTONSET

\* The Morbfile variable DON

\* If DATE\_OF\_ONSET is over 100 years before the date that program "05 Merging.sas" was run, then 100 years is added. After that, if DATE\_OF\_ONSET is before January 1, 1900, or after the date that program "05 Merging.sas" was run, then DATE\_OF\_ONSET is set to missing.

**(6) date\_of\_diagnosis**

- Date variable

- Date of onset of hepatitis C

\* Variable is an amalgamation of:

\* The CalREDIE variable DTDIAGNOSIS

\* The Morbfile variable DDX

\* If DATE\_OF\_DIAGNOSIS is over 100 years before the date that program "05 Merging.sas" was run, then 100 years is added. After that, if DATE\_OF\_DIAGNOSIS is before January 1, 1900, or after the date that program "05 Merging.sas" was run, then DATE\_OF\_DIAGNOSIS is set to missing.

**(7) mmwr\_year**   
- 8 numbers

- MMWR year for reporting (Believe that this has something to do with year of diagnosis)

- Currently only applicable to HEP C registry

\* Variable consists of the AVSS/legacy data system variable YEAR (recoded from a two digit year value to a four digit (if the value is over 11, then 1900 got added (i.e., "77" to "1977"); if the value was under 12, then 2000 was added (i.e. "4" to "2004"))

**(8) date\_of\_death**

- Date variable

- Date of death

\* Variable is an amalgamation of:

\* The CalREDIE variable DTDEATH

\* The Morbfile variable DTH

\* If DATE\_OF\_DEATH is same as DATE\_OF\_BIRTH, then DATE\_OF\_DEATH set to missing

\* If DATE\_OF\_DEATH is over 100 years before the date that program "05 Merging.sas" was run, then 100 years is added. After that, if DATE\_OF\_DEATH is before January 1, 1900, or after the date that program "05 Merging.sas" was run, then DATE\_OF\_DEATH is set to missing.

**(9) patient\_address**

- 50 characters

- Patient home street address - usually, either this or account address exists - if the two are the same, then usually the address is a prison

\* Variable is an amalgamation of:

\* The CalREDIE variable ADDRESS

\* The Morbfile variable ADDR

\* Spelling of "PO BOX is standardized (from "P.O. BX", etc.)

\* Various forms of "Homeless", "Transient", "No Address", "Not Given", etc. set to blank

\* Addresses under three characters set to missing

\* If address appears to indicate a prison address but it slightly misspelled, spelling is standardized

\* If variable is missing, but other variables indicate residence in a specific prison/jail, then variable is filled in with information for that prison/jail

**(10) patient\_city**

- 25 characters

- City of patient home address

\* Variable is an amalgamation of:

\* For records from the San Francisco County public health data system, if the country of residence (CNTY\_RES variable) is "San Francisco", then PATIENT\_CITY is set to "SAN FRANCISCO"

\* The CalREDIE variable CITY

\* The Morbfile variable CITY

\* Misspelled city names corrected

\* If length of patient city name is less than three, then try to derive a city name from patient zip code, account city, or account zip code. If all these fail, set to missing.

\* If variable is missing, but other variables indicate residence in a specific prison/jail, then variable is filled in with information for that prison/jail

**(11) patient\_zip\_code**

- 8 numbers

- Zip code of patient home address

\* Variable is an amalgamation of:

\* The CalREDIE variable ZIP (occasionally taken from variable STATE, when data misformatted)

\* The Morbfile variable ZIP

\* Zip code 99999 set to missing (there is no zip code '99999' in the USA - it is apparently a missing code)

\* Occasionally extracted out of last five characters of PATIENT\_ADDRESS variable

\* If variable is missing, but other variables indicate residence in a specific prison/jail, then variable is filled in with information for that prison/jail

**(12) census\_tract**

- 10 characters

- Census tract at time of record collection

\* Variable is an amalgamation of:

\* The CalREDIE variable CTRACT

\* The Morbfile variable CTRACT

**(13) account\_name**

- 40 characters

- The name of the doctor or institution that ordered the test - usually, the name of a prison, a clinic, a hospital, or a doctor

\* Variable is an amalgamation of:

\* Calculated from the CalREDIE variables RSLOCATION and ACCOUNT\_NAME (and occasionally from RSNAME and SUBMITER)

\* The Morbfile variable SOURCE

\* If address appears to indicate a prison address but it slightly misspelled, spelling is standardized

\* If variable is missing, but other variables indicate residence in a specific prison/jail, then variable is filled in with information for that prison/jail

\* If, for all cases, a given ORDERING\_DOCTOR name always correlates with a given ACCOUNT\_NAME, then, for cases where ACCOUNT\_NAME is missing and ORDERING\_DOCTOR is not, then assign the appropriate ACCOUNT\_NAME

**(14) account\_address**

- 80 characters

- Street address of ordering account - usually, either this or patient home address exists - if it is the same, then usually the address is a prison

\* Variable is an amalgamation of:

\* The CalREDIE variable RSADDRESS (occasionally the variable RSLOCATION)

\* The Morbfile variable RSADDRESS (occasionally the variable RSLOCATION)

\* If address appears to indicate a prison address but it slightly misspelled, spelling is standardized

\* If variable is missing, but other variables indicate residence in a specific prison/jail, then variable is filled in with information for that prison/jail

\* If, for all cases, a given ORDERING\_DOCTOR name always correlates with a given ACCOUNT\_NAME, then, for cases where ACCOUNT\_NAME and ACCOUNT\_ADDRESS are missing and ORDERING\_DOCTOR is not, then assign the ACCOUNT\_ADDRESS correlated with the appropriate ACCOUNT\_NAME

**(15) account\_city**

- 24 Characters

- City of ordering account address

\* Variable is an amalgamation of:

\* The CalREDIE variable RSCITY

\* The Morbfile variable RSCITY

\* Misspelled city names corrected

\* If variable is missing, but other variables indicate residence in a specific prison/jail, then variable is filled in with information for that prison/jail

\* If, for all cases, a given ORDERING\_DOCTOR name always correlates with a given ACCOUNT\_NAME, then, for cases where ACCOUNT\_NAME and ACCOUNT\_CITY are missing and ORDERING\_DOCTOR is not, then assign the ACCOUNT\_CITY correlated with the appropriate ACCOUNT\_NAME

**(16) account\_zip\_code**

- 8 numbers

- Zip code of ordering account address

\* Variable is an amalgamation of:

\* The CalREDIE variable RSZIPCODE

\* The Morbfile variable RSZIPCODE

\* Zip code 99999 set to missing (there is no zip code '99999' in the USA - it is apparently a missing code)

\* If variable is missing, but other variables indicate residence in a specific prison/jail, then variable is filled in with information for that prison/jail

\* If, for all cases, a given ORDERING\_DOCTOR name always correlates with a given ACCOUNT\_NAME, then, for cases where ACCOUNT\_NAME and ACCOUNT\_ZIP\_CODE are missing and ORDERING\_DOCTOR is not, then assign the ACCOUNT\_ZIP\_CODE correlated with the appropriate ACCOUNT\_NAME

**(17) local\_health\_juris**

- 20 characters

- The local health jurisdiction that this episode is correlated with

\* Variable is an amalgamation of:

\* A value of "San Francisco" for records from the San Francisco County public health data system

\* The CalREDIE variable LHJ

\* The Morbfile variable LHJ

\* If variable is missing, but other variables indicate residence in a specific prison/jail, then variable is filled in with information for that prison/jail

\* If variable is missing, and PATIENT\_CITY or ACCOUNT\_CITY are ['Berkeley', 'Long Beach', 'Pasadena'], then LOCAL\_HEALTH\_JURISDICTION is set to the name of that city. If LHJ variable is still missing, then assign the value of whatever California county that PATIENT\_ZIP\_CODE or ACCOUNT\_ZIP\_CODE map to. If LHJ is still missing, and PATIENT\_ZIP\_CODE or ACCOUNT\_ZIP\_CODE indicate out-of-state address, then assign value of "OUT OF STATE".

**(18) Laboratory**

- 50 characters

\* Variable is an amalgamation of:

\* The CalREDIE variable LABORATORY

\* The Morbfile variable LABORATORY

**(19) diagnosis**

- 20 characters

- Specific diagnosis information, in word form, that corresponds to the local health jurisdictions resolution status of the case. The status might or might not be accompanied by laboratory information.

\* Variable is an amalgamation of:

\* The CalREDIE variable RSTATUS (resolution status)

\* All records reported by SF through the eFTP site were considered confirmed

\* All records reported to the Morbfile were considered confirmed

**(20) diagnosis2**

- 8 numbers

- A single number, calculated in-house, that expresses whether a record can be considered a confirmed chronic hepatitis B case or not based solely on laboratory data.

- Currently only applicable to HEP C registry

**(21) ordering\_doctor**

- 30 characters

- The name of the medical doctor that ordered the hep c test done

- Currently only applicable to HEP C registry

\* Variable is an amalgamation of:

\* Calculated from the CalREDIE variables RSNAME and SUBMITER

\* If clearly missing ("NOT GIVEN", "NO ORDERING DOCTOR", "UNKNOWN") then set to missing

\* If in form of [Last, First], reordered to [First Last]

**(22) episode\_date\_1**

- Date variable

- First of a heterogeneous mixture of different date variables that correlate roughly with the time the record was collected

\* Variable is an amalgamation of:

\* San Francisco County public health data system variable REP\_DT

\* The CalREDIE variable DTEPISODE

\* The Morbfile variable DEP

\* If EPISODE\_DATE\_1 is before January 1, 1900, or after the date that program "05 Merging.sas" was run, then EPISODE\_DATE\_1 is set to missing.

Note: see algorithm for how Scott determined which date to enter as “date received”

**(23) episode\_date\_2**

- Date variable

- Second of a heterogeneous mixture of different date variables that correlate roughly with the time the record was collected

\* Variable is an amalgamation of:

\* The CalREDIE variable DTCREATE

\* The Morbfile variable DTCREATE

\* If EPISODE\_DATE\_2 is before January 1, 1900, or after the date that program "05 Merging.sas" was run, then EPISODE\_DATE\_2 is set to missing.

**(24) episode\_date\_3**

- Date variable

- Third of a heterogeneous mixture of different date variables that correlate roughly with the time the record was collected

\* Variable is an amalgamation of:

\* The CalREDIE variable DTRECEIVED

\* The Morbfile variable DAT

\* If EPISODE\_DATE\_3 is before January 1, 1900, or after the date that program "05 Merging.sas" was run, then EPISODE\_DATE\_3 is set to missing.

**(25) episode\_date\_4**

- Date variable

- Fourth of a heterogeneous mixture of different date variables that correlate roughly with the time the record was collected

\* Variable consists of the CalREDIE variable DTSUBMIT

\* The Morbfile variable DTSUBMIT

**(26) collection\_date**

- Date variable

- The date that a biological sample was obtained from the client

- Currently only applicable to HEP C registry

\* Variable is an amalgamation of:

\* The Quest Labs variable COLLECTION\_DATE

\* The Foundation Labs variable DRAW\_DATE

\* If COLLECTION\_DATE is before January 1, 1900, or after the date that program "05 Merging.sas" was run, then COLLECTION\_DATE is set to missing.

**(27) result\_date**

- Date variable

- The date a sample was logged into the Quest lab system, or results received by Foundation lab system

- Currently only applicable to HEP C registry

\* Variable is an amalgamation of:

\* The Quest Labs variable DATE\_OF\_SERVICE

\* The Foundation Labs variable RESULT\_DATE

\* If RESULT\_DATE is before January 1, 1900, or after the date that program "05 Merging.sas" was run, then RESULT\_DATE is set to missing.

**(28) firstdate**

- Date variable

- Date of first entry into system, excluding diagnosis date - calculated during data process

\* Variable is calculated first for each episode as the earliest/first/smallest out of [EPISODE\_DATE\_1, EPISODE\_DATE\_2, EPISODE\_DATE\_3, EPISODE\_DATE\_4, COLLECTION\_DATE, RESULT\_DATE].

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), then the earliest FIRSTDATE value is taken for the person-level record.

**(29) reporter\_type**

- 9 characters

- Who made the report of Hep C

- Currently only applicable to HEP C registry

\* Variable consists of the AVSS/legacy data system variable RTYP

**(30) prison**

- 1 character

- A variable, derived from a wide selection of other variables, that expresses incarceration status at the time of record collection

Possible values are:

- "M" for incarcerated in a male state prison ("AVENAL STATE PRISON", "CA CORRECTION CENTER", "CA CORRECTION INSTITUTE", "CA INSTITUTE FOR MEN", "CA MED FACILITY", "CA MEN'S COLONY", "CA STATE PRISON CENTINELA", "CA STATE PRISON CORCORAN", "CA STATE PRISON LOS ANGELES", "CA STATE PRISON SACRAMENTO", "CA STATE PRISON SOLANO", "CAL SUBST ABUSE TREATMENT FCLTY", "CALIPATRIA STATE PRISON", "CHUCKAWALLA VALLEY STATE PRISON", "CORRECTIONAL TRAINING FACILITY", "DEUEL VOCATIONAL INST", "FOLSOM STATE PRISON", "HIGH DESERT STATE PRISON", "IRONWOOD STATE PRISON", "KERN VALLEY STATE PRISON", "MULE CREEK STATE PRISON", "NORTH KERN STATE PRISON", "PELICAN BAY STATE PRISON", "PLEASANT VALLEY STATE PRISON", "RJ DONOVAN CORRCTNL FACLTY", "SALINAS VALLEY STATE PRISON", "SAN QUENTIN STATE PRISON", "SIERRA CONSERVATION CENTER", "WASCO STATE PRISON", and men in "CA REHABILITATION CENTER")

- "F" for incarcerated in a female state prison ("CA INSTITUTION FOR WOMEN", "CENTRAL CA WOMENS FACILITY", "NORTHERN CA WOMENS FACILITY", "VALLEY STATE PRISON FOR WOMEN", and women in "CA REHABILITATION CENTER")

- "O" for incarcerated in another prison or jail (most are unknown/unspecified (and in some cases may be a "M" or "F" facility, although it can't be known for sure), but common specific non-CDCR facilities include "ATASCADERO STATE HOSPITAL", "FEDERAL PENITENTIARY LOMPOC", "LA COUNTY JAIL TWIN TOWERS", "LAS COLINAS DETENTION CENTER", "LERDO JAIL FACILITY", "SACRAMENTO COUNTY MAIN JAIL", and "SAN DIEGO CENTRAL DETENTION FCLTY")

- "-" for not incarcerated

If specific institution indicates a specific state prison, then "M" or "F" status are coded. Otherwise, "O" status is given.

\* Incarceration status is determined in the affirmative if

- ACCOUNT\_NAME, ACCOUNT\_ADDRESS, PATIENT\_ADDRESS, code at the end of ORDERING\_DOCTOR (i.e., "Sally Jones (NKSP)"), or patient/address zip code (in a couple cases, where prisons have their own zip code) indicate residence in a specific jail

- If "CDC" or "INMATE" is part of FIRST\_NAME

- If ACCOUNT\_NAME, ACCOUNT\_ADDRESS, or PATIENT\_ADDRESS include one of the text strings ["INMATE","JAIL","CORRECT,"PRISON","DETENTION","CUSTODY","INCARCER"]

- If OCCUPATION is "Inmate"

**(31) test\_name**

- 45 characters

- An overview variable that captures the name of a biological test done for this care episode

- Currently only applicable to HEP C registry

\* Variable is an amalgamation of:

\* For records from Quest Lab, this variable is copied from ORDER\_NAME [if ORDER\_NAME is the same as RESULT\_NAME, or if ORDER\_NAME is equal to certain values]. But the RESULT\_NAME takes precedence and may be copied over into TEST\_NAME when the result value correlates more closely with a single test value and/or when the order name could indicate more than one different type of result (such as when a reflex is present).

\* The Foundation Labs variable TEST\_NAME

\* The hand-entered lab-reports/accession numbers variable TEST

**(32) result\_name**

- 36 characters

- The name of the result of a biological test (there often are multiple result names for a single test, for example when there is a reflex attached to a base test or when a hepatitis panel was ordered- this variable specifies which result is being listed)

- Currently only applicable to HEP C registry

\* Variable is an amalgamation of:

\* The Quest Labs variable RESULT\_NAME (standardized down to a smaller number of possible values)

\* The Foundation Labs variable TEST\_NAME

\* The hand-entered lab-reports/accession numbers variable TEST

**(33) order\_name**

- 41 characters

- The name of the test that was ordered (there often are multiple result names for a single test name, for example when there is a reflex attached to a base test or when a hepatitis panel was ordered)

- Currently only applicable to HEP C registry

\* Variable is an amalgamation of:

\* The Quest Labs variable ORDER\_NAME

\* The Foundation Labs variable TEST\_NAME

\* The hand-entered lab-reports/accession numbers variable TEST

**(34) result\_value**

- 20 characters

- The value of a the result of a biological test

- Currently only applicable to HEP C registry

\* Variable is an amalgamation of:

\* The Quest Labs variable RESULT\_VALUE

\* The Foundation Labs variable TEST\_RESULT

\* The hand-entered lab-reports/accession numbers variable INDEX

**(35) result\_comment**

- 200 characters

- Comments submitted by Quest Lab concerning a biological test result

- Currently only applicable to HEP C registry

\* Variable consists of the Quest Labs variable RESULT\_COMMENT

**(36) patient\_id**

- 22 characters

- A heterogeneous collection of data-system-specific unique (or unique-ish) identification variables. Some were, in their original datasets, purely numeric, some were alphanumeric.

\* Variable is an amalgamation of:

\* San Francisco County public health data system variable local\_id

\* The CalREDIE variable INCIDENTID

\* The Morbfile variable ID, INCIDENTID

**(37) firstdxdate**

- Date variable

- Date of first entry into system - calculated during data process

\* Variable is calculated first for each episode as the earliest/first/smallest out of [EPISODE\_DATE\_1, EPISODE\_DATE\_2, EPISODE\_DATE\_3, EPISODE\_DATE\_4, COLLECTION\_DATE, RESULT\_DATE, DIAGNOSIS\_DATE].

\* After deduplicating processes pairs up all episode-level records into matches (i.e., groups by persons), then the earliest FIRSTDATE value is taken for the person-level record.

\* This variable was created because we don’t particularly trust the diagnosis date variable and also because the date of diagnosis could legitimately precede the date of entry into our system.

==================================================================================================================================================================================