

Probabilistic Record Linkages for Generating a Comprehensive Epidemiological Research File on Maternal and Infant Health

Beate Danielsen, Health Information Solutions

Date Last Printed: 11/4/2000

1. Table of Contents

1. Table of Contents	2
2. Introduction	6
3. Probabilistic Record Linkage	7
<i>Method</i>	7
<i>Similarity Measures</i>	9
<i>The SOUNDEX Transformation</i>	10
4. Autoexec.sas	11
5. Main Navigation Screen	13
6. Linkage of Vital Statistics Birth and Vital Statistics Death Data	14
<i>Step 1: Setting Linkage Parameters and Reviewing Linkage Status</i>	15
<i>Step 2: Reading Vital Statistics Death Data</i>	16
<i>Step 3: Extracting Subfile from Vital Statistics Death File to be Used for the Linkage</i>	17
<i>Step 4: Extracting Subfile from the Vital Statistics Birth File to be Used for Linkage</i>	19
<i>Step 5: Generating Value-Specific Frequency Information and the Set of u-Probabilities</i>	20
<i>Step 6: Setting of m- and u-Probabilities for Linkage Run</i>	21
<i>Step 7: Running a Linkage Step</i>	22
<i>Step 8: Clerical Review of Match Results</i>	25
<i>Step 9: Recalculate m-Probabilities and Check Convergence</i>	26
<i>Step 10: Generating Results File and Result Summary</i>	27
<i>Step 11: Bias Check</i>	27
7. Linkage of Vital Statistics Birth and Infant/Maternal Hospital Discharge Data	28
<i>Step 1: Setting Linkage Parameters and Reviewing Linkage Status</i>	28
<i>Step 2: Run readdat1: Extract Infant and Maternal Discharge Record from Master Patient Discharge Data File</i>	29
<i>Step 3: Run readdat2: Combine this Year's and Next Year's Infant and Maternal Records into one File</i>	30

<i>Step 4: Run extpddi: Generate Minimum Infant Discharge File for Vital Statistics Birth and Infant Discharge Record Data Linkage</i>	32
<i>Step 5: Run extpddm: Generate Minimum Maternal Discharge File for Vital Statistics Birth and Maternal Discharge Record Data Linkage</i>	37
<i>Step 6: Run readvs: Read Vital Statistics Data</i>	38
<i>Step 7: Run extvs: Extract Minimum Vital Statistics Data File for Linkage to Maternal and Infant Discharge Data</i>	39
<i>Step 8: Run getdat: Prepare Hospital Level Record Comparison</i>	41
<i>Step 9: Run gettable: Generate Hospital Level Comparison of Record Numbers to Identify Hospital Mergers, etc.</i>	41
<i>Step 10: Update Hospital Crosswalk File with Results of Previous Step</i>	43
<i>Step 11: Run Final Preparation Step to Add DHS Hospital ID to Patient Discharge Databases</i>	44
<i>Step 12: Generating Value-Specific Frequency Information and the Set of u-Probabilities</i>	44
<i>Step 13: Setting of m- and u-Probabilities for Linkage Run</i>	45
<i>Step 14: Adding Unlinked Infant Discharge Records to the Vital Statistics Birth Subfile (Only Needed for Maternal Linkage!)</i>	47
<i>Step 15: Linkage of Newborn Discharge and Vital Statistics Birth Data</i>	47
<i>Step 16: Linkage of Maternal Discharge and Vital Statistics Birth Data</i>	51
<i>Step 17: Clerical Review of Match Results</i>	52
<i>Step 18: Generating Results Files</i>	54
<i>Step 19: Recalculate m-Probabilities and Check Convergence</i>	58
<i>Step 20: Generate Results Summary</i>	59
<i>Step 21: Bias Check</i>	60

8. Add-On Linkage of Unlinked Vital Statistics Records and Linked Infant/Maternal Discharge Records----- 61

<i>Step 1: Setting Linkage Parameters and Reviewing Linkage Status</i>	61
<i>Step 2: Run extsubdat: Extract Unlinked Births and Infant/Maternal Record Matches</i>	62
<i>Step 3: Generating Value-Specific Frequency Information and the Set of u-Probabilities</i>	63
<i>Step 4: Setting of m- and u-Probabilities for Linkage Run</i>	64
<i>Step 5: Linkage of Unlinked Births and Infant/Maternal Matches</i>	65
<i>Step 6: Clerical Review of Match Results</i>	66
<i>Step 7: Generate Results File</i>	68
<i>Step 8: Recalculate m-Probabilities and Check Convergence</i>	68

Step 9: Update sastmp.vsbvsdIM to Include Additional Matches -----68

9. Linkage of Delivery and Maternal Prenatal/Postnatal Admission Data ----- 72

Step 1: Setting Linkage Parameters and Reviewing Linkage Status-----72

Step 2: Run extrlnbase: Extract All Records Pertaining to Women Aged 11 to 70 that are Not Deliveries -----73

Step 3: Run getPPMom1: Extract All Delivery Records -----74

Step 4: Run getPPMom2: Generate File of Possible Prenatal/Postnatal Admissions -----75

Step 5: Generating Value-Specific Frequency Information and the Set of u-Probabilities -----75

Step 6: Setting of m- and u-Probabilities for Linkage Run -----76

Step 7: Linkage of Maternal Delivery Record and Prenatal/Postnatal Hospitalizations -----78

Step 8: Clerical Review of Match Results-----80

Step 9: Recalculate m-Probabilities and Check Convergence -----81

Step 10: Generate Results File-----82

Step 11: Results Summary-----82

10. Linkage of Birth Record and Transfers/Re-Admissions within the first year of life ----- 83

Step 1: Setting Linkage Parameters and Reviewing Linkage Status-----83

Step 2: Run prepmin: Extract All Possible Birth Records-----84

Step 3: Generating Value-Specific Frequency Information and the Set of u-Probabilities -----88

Step 4: Setting of m- and u-Probabilities for Linkage Run -----90

Step 5: Linkage of Transfers and Births-----91

Step 6: Linkage of Re-Admissions and Births -----93

Step 7: Clerical Review of Match Results-----93

Step 8: Recalculate m-Probabilities and Check Convergence -----95

Step 10: Generate Results File-----96

Step 10: Generate Results Summary-----97

11. Generation of Final Summary File ----- 98

Location of Input Files -----99

Step 1: Remove from vsbvsdIM Newborn Records Linked as Transfers; Link VS Births -----99

Step 2: Obtain Full Input File for Linked Prenatal/Postnatal Admissions ----- 101

Step 3: Merge full Maternal Discharge Record to Cohort----- 101

<i>Step 4: Append Transfers/Re-Admissions to Cohort and Link to Infant Discharge Records -----</i>	<i>102</i>
<i>Step 5: Append Prenatal/Postnatal Maternal Admissions -----</i>	<i>102</i>
<i>Step 6: Garble Up Record Identifiers to Disallow Link Back to Published OSHPD or DHS Records ----</i>	<i>103</i>
<i>Step 7: Sort Data Set by Reference Birth Identifier and Within Reference Birth Identifier by Order of Events -----</i>	<i>103</i>
<i>Step 8: Beautify Cohort File and Produce Result Summaries -----</i>	<i>104</i>

2. Introduction

The purpose of this document is to summarize the use of the Graphical User Interface for record linkages.

The following record linkages can be performed using the GUI:

- Vital Statistics Birth and Vital Statistics Death Data: In California, the Department of Health Services published two files that summarize the birth experience in a year, the vital statistics birth file and the birth cohort file. The vital statistics birth file includes information collected at time of birth. For each baby for which a birth certificate was filed, a record is included in this file. The birth cohort file includes all information collected at time of birth, and information on any deaths that occurred to infants in this file during the first year of life is linked as well. As a cohort of infants must be followed over the span of one year, the birth cohort file can be constructed at the earliest one year after collection of the vital statistics birth file. In addition, for babies who die out of state, records have to be retrieved from other U.S. States. Therefore, there is usually a lag of three to four years between the current year and the most recent year for which the birth cohort file is available. By linking California death certificates to the vital statistics birth file, the time lag in the availability of death information can be reduced.
- Vital Statistics Birth and Infant Hospital Discharge Data. While the vital statistics birth file is rich in the description of risk factors, it lacks in the description of outcomes. Usually, only birth weight and gestational age are the primary outcomes used from this file. The discharge record complements the set of outcomes in that it provides detailed diagnosis, procedure, and resource use (length of stay, charges) information.
- Vital Statistics Birth and Maternal Hospital Discharge Data. The only maternal outcome collected on the birth certificate is maternal death which is very rare. The availability of the maternal discharge record does not only allow the study of additional outcomes, but it also allows the assessment of additional risk factors (hypertension, diabetes, etc.) that are not reliably available from the birth certificate.
- Maternal Delivery Record and Maternal Prenatal and Postnatal Hospitalization Records: The availability of antepartum and postpartum complications allows an improved assessment of risk factors and longer term outcomes.
- Infant discharge and transfer records: For infants who were transferred prior to being discharged home for the first time after birth, the hospitalization record is incomplete. Resource information is truncated, diagnosis and procedure information is not completely recorded. Linkage of transfers enables the complete study of outcomes.
- Infant discharge and re-admission records: Any re-admissions during the neonatal period and/or first year of life provide additional outcomes that can be studied.

Note that it is possible to use this GUI with non-California data as long as the same variable names and types are used.

3. Probabilistic Record Linkage

The methods of record linkage used in this GUI are based on probabilistic linkage techniques described in the book "Handbook of Record Linkage. Methods for Health and Statistical Studies, Administration, and Business" by Howard B. Newcombe and several publications by Matthew Jaro.

It is important to keep in mind that probabilistic linkage theory provides the best possible match between records given the structure of the data. However, it is not a precise science, in that it must be taken into account that linkage errors occur. The output files produced contain several markers that will ascertain the likely correctness of a match. Some of the matches produced will be false positives, while some records that should be matched will not be matched (false negatives).

Method

A preparatory step for the data linkage is the calculation of a set of u-probabilities. A **u-probability** is defined as the probability that a variable matches for two records even though the records do not constitute a match. For instance, the u-probability for gender is about 0.5 since we can expect gender to be the same if two records are selected at random and compared for gender. Note that the lower the u-probability for a variable, the better it distinguishes between records. The u-probabilities are determined based on a data set of "unlinkable" pairs that was derived for each linkage situation.¹ More details on the estimation of the u-probabilities is included in the description of relevant parts of the GUI below.

Besides a set of u-probabilities, a set of m-probabilities needs to be determined. An **m-probability** is defined as the probability that a variable matches for two records if the records do constitute a match. The m-probability of a variable reflects how reliably it is coded. The higher the m-probability for a variable, the more reliable it is. As there is no easy way to determine the m-probability of a variable beforehand, the initial values are usually estimated from experience. Data linkage runs are then repeated with a set of m-probabilities that is estimated from the set of matches. The set of m-probabilities is re-estimated until convergence is achieved, i.e., the estimated m-probabilities do not differ by more than a predetermined amount alpha. For our purposes, alpha is set to 0.001.

In order to link two data sets, each record in the first data set has to be compared with each record in the second data set. As an example, the number of records in the vital statistics birth data for 1997 was 525,455. The number of infant discharge records for 1997 was 516,151. For matching these data sets, if we were to carry out a comparison of each record on the vital statistics birth file with each record on the infant discharge file, we would need to study 271,208,869,155 comparisons, an inefficient route to accomplish the task. Therefore, comparisons are carried out within blocks of observations. In general, these blocks of observations should be designed such that:

1. The variables they are based on are very reliable, i.e., have high m-probabilities.

¹ U-probabilities are derived from a file of "unlinkable" pairs. A file of unlinkable pairs is created by matching randomly selected records from the first data set to randomly selected records from the second data set. To ensure that the resulting file does not include any true matches, all records which might be true matches are removed. The u-probability is determined as the ratio of records in agreement across all unlinkable pairs.

2. The mean number of observations per block should be small (about 2).
3. Different blocks should be based on variables that are unrelated to each other. For instance, a block based on ZIP code and a second block based on the street address do not make much sense since ZIP codes and street addresses are related to each other and nothing would be gained from matching pairs within blocks induced by ZIP codes and then blocks induced by the street address.

The data linkages proceed in these steps:

Determination of an appropriate blocking structure.

1. Determination of the total number of simple agreements for each possible pair within each block. A simple agreement for a variable on two files is 1 if the two variables agree, 0 otherwise.
2. Determination of the general frequency weight for each pair exceeding a critical value of simple agreements. The general frequency weight for a variable was defined as follows:



The overall general frequency weight is determined as the sum of the logarithms to base 2 of the individual variable's general frequency weights.

3. Determination of the value-specific frequency weight for each pair exceeding a critical value for the general frequency weight. In contrast to the general frequency weight, the value-specific not only takes into account the variable, but also takes into account the value of the variable. In other words, the general frequency weight for a variable is scaled up if the values observed for the variable are relatively scarcer and they are scaled down if the values observed for the variables are relatively more common. The value-specific frequency weight was obtained as the sum of the logarithms to base 2 of the individual variable's value-specific frequency weights.
4. Determination of cutoff threshold. We determined a threshold for the value-specific frequency weight beyond which we considered pairs of records as matches. As the size of the linkage task at hand is enormous, clerical review was primarily carried out to obtain conservative match cutoffs.



Figure 1 demonstrates how the general frequency weight changes dependent upon values for the m- and u-probabilities. For instance, a frequency weight of 10 for a variable can be judged as very good evidence in favor of a match based on that variable.

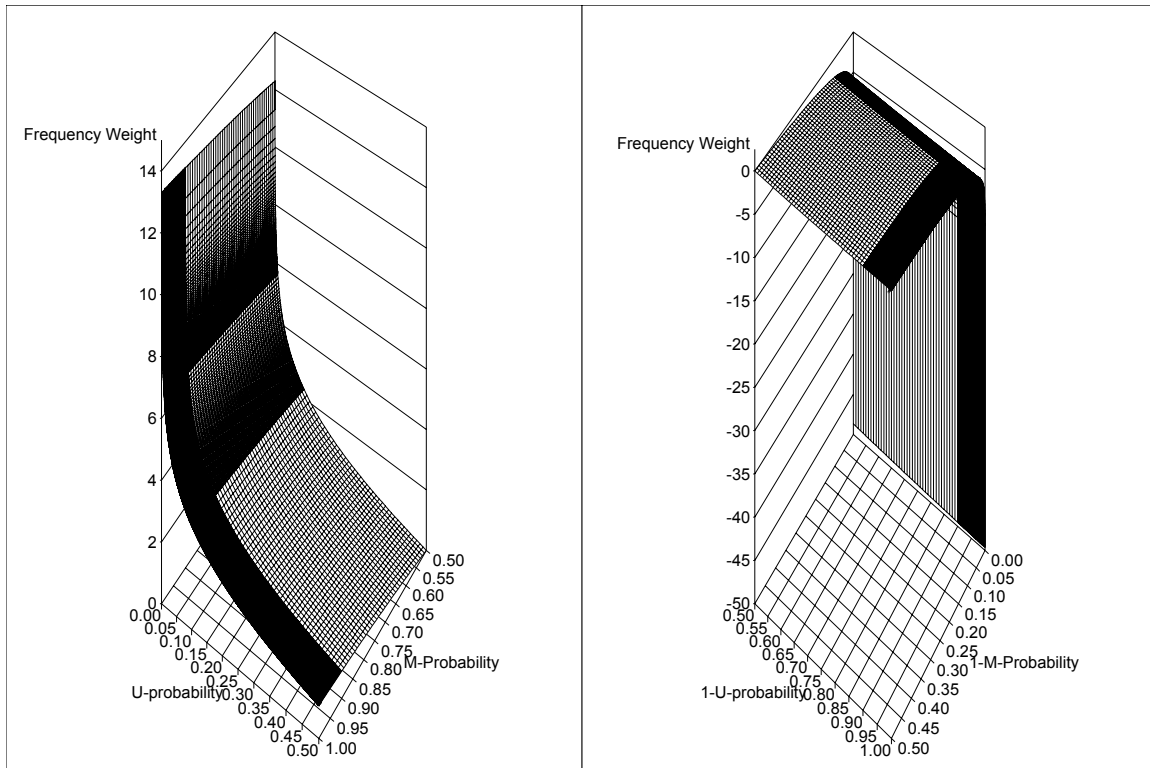


Figure 1: Agreement and Disagreement Weights by m- and u-probability Values

Similarity Measures

For some variables, the requirement of perfect agreement for assignment of the full match weight is very strict. The perfect agreement criterion can be softened by obtaining similarity measures for two strings that need to be compared. A similarity measure reflects to which extent two strings have the same source. A value of 1 indicates perfect agreement and a value of 0 indicates complete disagreement. Similarity measures are particularly useful for the comparison of Social Security Numbers for which simple transpositions or incorrect reporting of only 1 or 2 digits might occur.

We use the Jaro comparator to carry out string comparisons.

1. Compute string lengths.
2. Find number of common characters in the two strings. The definition of common is that the agreeing character must be within $\frac{1}{2}$ the length of the shorter string.
3. Find number of transposition. The definition of a transposition is that the character from one string is out of order with the corresponding character from the other string.

$$4. \quad \text{jaro}(\text{string1}, \text{string2}) = \frac{1}{3} \left(\frac{\# \text{ common}}{\text{length of string 1}} + \frac{\# \text{ common}}{\text{length of string 2}} + 0.5 \times \frac{\# \text{ transpositions}}{\# \text{ common}} \right)$$

The Jaro comparator is incorporated in the calculations of the value-specific frequency weights for the record linkage number, ZIP code, and name comparisons in the following fashion. If the Jaro comparator exceeds 0.7, the agreement weight is multiplied by the Jaro comparator. Otherwise, the full disagreement

weight is applied. Note that if the Jaro comparator is 1 reflecting full agreement, the full agreement weight is applied.

Note whenever similarity measures are used to prorate the probabilistic match weight, for the calculation of the value-specific weight, the larger of `prob` and `prob_` is used. This is necessary since a comparison of a very common and a similar, but very rare string, might lead to a match weight that exceeds an agreement match on the very common string.

A more thorough discussion of similarity measures can be found in Porter, 1999.²

The SOUNDEx Transformation

The SOUNDEx function encodes a character string according to an algorithm originally developed by Margaret K. Odell and Robert C. Russel (US Patents 1261167 (1918) and 1435663 (1922)). The algorithm is described in Knuth, The Art of Computer Programming, Volume 3. See also, SAS on-line documentation.

To obtain the SOUNDEx transformation of a character string, the following rules are used:

1. Retain the first letter of the character string and discard the following letters: A E H I O U W Y
2. Assign the following numbers to these classes of letters:

1: B F P V

2: C G J K Q S X Z

3: D T

4: L

5: M N

6: R

3. If two or more adjacent letters have the same letter class, discard all but the first.
4. Add trailing zeroes and truncate the result to a length of 4.

Note that for our purposes, rather to a length of 4, we truncated to a length of 3.

² Porter EH, Winkler WE. Approximate string comparisons and its effect on an advanced record linkage system. Bureau of the Census. Technical report RR97/92. 1997.

4. Autoexec.sas

The following file should be used as `autoexec.sas` file for the linkage application, or these lines should be run before any calls to the GUI are made:

```

OPTIONS FORMCHAR ="|----|+|----+=|-\<>";
OPTIONS COMPRESS=YES NOCENTER ERRORS=1;
LIBNAME auxdata V8 " f:\pdd-vs linkages\auxiliary data";
LIBNAME gui      V8 " f:\pdd-vs linkages\guis";
LIBNAME saslib   V8 " f:\pdd-vs linkages\data";
LIBNAME tmp      V8 " f:\pdd-vs linkages\tmp";
%LET tmpprogs= f:\pdd-vs linkages\tmp;
%LET path0=f:\pdd-vs linkages\programs;
%LET path1=f:\pdd-vs linkages\formats;
%LET path3=f:\pdd-vs linkages\temporary files;
%LET vstype=BCF;
%LET next=Y;
%LET year=1997;
LIBNAME sastmp  V8 "&path3";

DM "AF CAT=gui.linkages.main.frame";

```

The `FORMCHAR` option is used since the designer preferred the Lucida font in the output and log windows. The compression option should be edited if the SAS platform used does not yet provide a stable compression mechanism.

The first path for a permanent library named `auxdata` is the location of the crosswalk files providing a connection between the 4-digit maternity hospital identifier and the 6-digit OSHPD hospital identifier. Note that these crosswalk files differ from year to year and that they need to be updated for each year for which data are linked. We will discuss the updating mechanism later as it is built into the GUI. The libname `gui` is essential in that it provides the location of all the GUIs constructed for this application. The libname `saslib` is necessary and points to the path of permanent input data files, e.g., the birth cohort SAS data sets. The library `tmp` is used if for the maternal prenatal and postpartum linkages which creates a number of large subfiles. If disk space is tight, it is possible to choose a different disk location for these temporary files. Note that all files that are created in this directory are also removed after the GUI has run.

The macro variable `tmpprogs` is the path to a temporary location. The GUI constructs several ASCII input files that are then included in the SAS programs that it runs. These files do not need to be kept around and can be removed after linkages are completed. The macro variables `path0`, `path1`, and `path3` establish a default path for the program files (macros), format files, and temporary data files. These paths can be overwritten by the GUI. The macro variables `vstype`, `next`, and `year` are convenient to set at this point, especially the first two since the default values here are the ones that will be most commonly used.

Finally, the `DM` instruction starts the GUI.

Note that the above lines are using the Windows syntax for the formulation of the data path. The GUI is already prepared to also run on the Linux platform, and the paths can be easily adapted to conform to Linux conventions.

5. Main Navigation Screen

The main navigation screen is displayed in Figure 2. It allows to jump to the next linkage to be undertaken. If a linkage has been previously started and if work on it is to continue, it is also possible to obtain a list of projects by clicking on the down arrow of the combination box in the bottom of the screen. In this case, all subsequent screens will try to retrieve the most current information for the project specified.

Data Linkages
Health Information Solutions, October 2000

- 1 Vital Statistics Birth and Death Data
- 2 Vital Statistics Birth and Infant/Maternal Discharge Record
- 3 ADD-QN: Unlinked Births and PDDI-PDDM Linkage
- 4 Delivery and Prenatal/Postnatal Records
- 5 Infant Discharge and Transfers/Re-Admissions
- 6 Generate Summary File from all Linkages for Cohort

Enter project name:

Figure 2: Start-up Screen of Graphical User Interface (GUI) for Linkages

6. Linkage of Vital Statistics Birth and Vital Statistics Death Data

Figure 3 shows the main navigation GUI screen for accomplishing the match between the vital statistics birth and death files.

Figure 3: VSBVSD Linkage, Main Screen

The linkage is guided by the sequence of push buttons on the left of the screen. The checkboxes allow to check off the completion of the step of the linkage. The combo box on the right (data entry box with the down arrow next to it) can be used to access information on projects as they were previously stored. Clicking on the down arrow will bring up a list of all current and/or completed projects along with their status and parameter settings.

Note that if a project name was selected on the first GUI screen (see Figure 2), the GUI will attempt to load the information for this project.

If a project has never been stored before, a name should be typed over the text entry field 'Project Name'. Using the 'Save' push button will save the current status of the project. Saving is fast and it is recommendable to save project information after each step.

The remainder of this section discusses the function of each push button. It also discusses the role of any macros that are executed in the background of the GUI as a result.

Step 1: Setting Linkage Parameters and Reviewing Linkage Status

The first push button 'Enter Linkage Parameters' leads to a screen that displays parameters for the current linkage environment (Figure 4).

Path to linkage macros: h:\pdd-vs linkages\programs

Path to formats: h:\pdd-vs linkages\formats

Path to output data: h:\pdd-vs linkages\temporary files

Year of linkage: 1997

VS/BCF file? ☒ VS ☐ BCF

Next year of death data available? ☒ Yes ☐ No

Additional weight information? ☐ Yes ☒ No

M-Probabilities Version: 0

Linkage Block: 1

VSPW cutoffs: 50 60 60 60

Done Cancel Reset

Figure 4: VSBVSD Linkage, Linkage Parameters Entry

This screen can be used to adapt the path to the SAS Macro files that are used throughout the GUI, the path to formats needed to read the data, and the path to a SAS data library that holds temporary output files specific for the year of linkage. **If it is necessary to save these temporary output files for one year's linkages, it is necessary to choose a different path for each year.**

Besides paths, the linkage parameter/status screen also allows the setting of:

- The year for which data are read and/or linked.
- The type of input file: Vital Statistics Birth File (VS) and Birth Cohort File (BCF). For the death linkage VS is the default and at this point only reasonable choice.
- Indicator flag for availability of next year of data: Usually for the most recent year of vital statistics birth data at most the death data for the same year are available, not for the subsequent year. If data for the subsequent year are available, this flag should be set to 'Yes,' otherwise it should be set to 'No.'
- Status indicator flag for amount of detail information that was most currently retained as part of the linkage result files. Before a linkage run is started, the user is prompted to indicate whether or not detail information on each variable's contribution to the overall match weight should be

- retained. This flag is 'Yes' if the user decided to retain this information; it is 'No' if the user decided not to retain this information. More details about the effect of retaining information are included in the step discussing the starting of linkage runs.
- Status on the version of m-probabilities used for the most current linkage run: Linkage runs are repeated until convergence of m-probabilities is achieved. The m-probabilities version indicator shows what the most recent linkage run was for which m-probabilities were calculated. Note that this information is saved as part of the projects file, i.e., it is possible to quit the GUI and restart it to continue linking and still retain all the most current status information on linkage progress. However, read the notes on stopping and restarting linkages to make sure any necessary temporary information stored in the SAS work library is also saved.
 - Status of most current block that was linked: This entry keeps track of the most recent linkage block that was completed.
 - Block-specific thresholds for the value-specific frequency weights. The first box corresponds to the first block, the second box corresponds to the second block, etc. Only match pairs that exceed the value-specific frequency weight will be included in the results file.

The 'Done' pushbutton executes any changes made to the parameters. The 'Cancel' pushbutton allows to leave this screen without making any changes to the current parameter settings. The 'Reset' button will reset all parameter values to their defaults, observing correct path syntax depending upon the OS environment.

Step 2: Reading Vital Statistics Death Data

The vital statistics death data are read with the macro **%readvdsd** stored in *macros for getting data ready for linkages.sas*. The macro has two parameters: the year for which data is read, and the record length (LRECL) for the file. The user is prompted for the year for which death data should be read. For instance, for the death linkage of the 1997 birth data, the 1997 and 1998 death data sets should be read. Of course, if the following year of data is not yet available, this is not possible.

Figure 5 shows the pop-menu that is displayed to allow the user to read the correct year of data for the 1997 data linkage. Figure 6 shows the screen that prompts the user for the input path of the raw data.

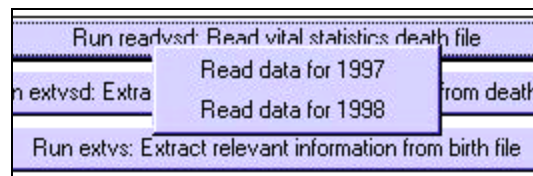


Figure 5: VSBVSD Linkage, Pop-Menu for Setting the Year for Which Data are Read

Enter path to external file:

Enter LRECL:

Figure 6: VSBVSD Linkage, Setting Input Path for Raw Vital Statistics Death Data

Besides reading all the variables included in the file, a unique record identifier is created:

```
dlinkid = &year.000000 + _N_;
```

This identifier is retained in all subfiles that are extracted from this master vital statistics death file as it will allow linkage back to the original information for a record.

The code for reading the vital statistics death data does not change any of the raw data as they are stored in the file except for one element. For the ICD-9 diagnosis code, any dashes are removed from the string.

A note on reading the vital statistics birth data:

As for the majority of years, the linkage of vital statistics death data to vital statistics birth data is not necessary, the vital statistics birth data are not read as part of the portion of the GUI that accomplishes the vital statistics birth/death file linkage. The vital statistics birth file is read as part of the VSPDD GUI, i.e., the routine that links the vital statistics birth and infant/maternal hospital discharge data. For the purpose of reading the raw vital statistics birth data, it is therefore necessary to exit this screen and enter the portion of the GUI that deals with the vital statistics birth and infant/maternal discharge data linkage.

Step 3: Extracting Subfile from Vital Statistics Death File to be Used for the Linkage

It makes sense to only retain the minimally necessary information from the vital statistics death file for the purpose of the linkage. This step is accomplished by running the **%extvdsd** macro stored in *macros for getting data ready for linkages.sas*. The macro has only one parameter, an indicator flag showing whether or not the next year of death data is available. The macro creates a file called `sastmp.subvdsd`.

Important note: The death linkage input subfile is generated based on one or two input files. At a minimum the current year's death records are searched for any persons with birth date in the current year. If the next year of death data is also available, this file is searched, too. Note that both files – if both are available and the macro parameter `next` is set equal to 'Y' in the parameter entry screen – should be read using the **%readvdsd** macro prior to attempting to run the **%extvdsd** macro. Otherwise the **%extvdsd** macro will fail.

Table 1 shows the variables extracted. It also shows the type (character or numeric) and the length of the variables.

Table 1: VSBVSD Linkage, Variables Extracted from the Vital Statistics Death File for Linkage

Variable Name	Description	Type	Length
bmomn	Mom's birth name	Char	15
bmomnsdx	SOUNDEX transformation of Mom's birth name	Char	3
bthdate	Birth date	Num	4
county	County of birth	Char	2
detrace	Race detail	Char	2
dobd	Day of birth	Num	3
dobm	Month of birth	Num	3
doby	Year of birth	Num	3
dthdate	Date of death	Num	4
fn	First name	Char	12
fnsdx	SOUNDEX transformation of first name	Char	3
hisp	Hispanic ethnicity indicator	Char	1
ldadn	Father's last name	Char	15
ldadnsdx	SOUNDEX transformation of father's last name	Char	3
linkid	Unique record identifier	Num	8
ln	Last name	Char	20
lnsdx	SOUNDEX transformation of last name	Char	3
mn	Middle name	Char	12
mnsdx	SOUNDEX transformation of middle name	Char	3
race	Race - abridged coding	Char	1
sex	Gender	Char	1
sporig	Hispanic origin detail	Char	1
zip	5-digit ZIP code of residence at time of death	Char	5

Note that variables are added after the extraction, the SOUNDEX version of all of the name variables and individual variables representing day, month, and year of birth. Race and Hispanic origin variables are generated to obtain a uniform coding across databases. The following two tabulations list the detail coding in the left column and the abridged coding in the right column.

Recoding for Race:	
'10' (White)	White ('1')
'20' (Black)	Black ('2')
'30' (American Indian), '57' (Eskimo), '58' (Aleut)	Native American/Eskimo ('3')
'98' (Unknown), '99' (Not Stated)	Unknown ('6')
All others	Asian/Pacific Islander ('4')

Recoding for Ethnicity:

`1' (Not Hispanic)	Non-Hispanic (`2')
`9' (Unknown)	Unknown (`3')
All others	Hispanic (`1')

The record identifier `dlinkid` is renamed to `linkid`.

Step 4: Extracting Subfile from the Vital Statistics Birth File to be Used for Linkage

It makes sense to only retain the minimally necessary information from the vital statistics birth file for the purpose of the linkages. This step is accomplished by running the **%extvs** macro stored in *macros for getting data ready for linkages.sas*. The macro has no parameters, but assumes that the variables `vstype`, `year`, and `nextyear` are globally available. The macro creates a file called `sastmp.subvs`.

The macro **%extvs** is also used later to generate the vital statistics linkage input data for the infant linkage and maternal linkage. It is essential for the global macro variable to be set to VS for the macro to do the extraction appropriate for the linkage to the death data.

Table 2 shows the variables extracted as well as their type and their length. The grayed variables are not necessary for the actual linkage. They are retained to provide additional information on linkage quality.

Table 2: VSBVSD Linkage, Variables Extracted from the Vital Statistics Birth Data for Linkage

Variable name	Description	Type	Length
<code>bmomn</code>	Mom's birth last name	Char	15
<code>bmomnsdx</code>	SOUNDEX transformation of Mom's birth last name	Char	3
<code>bthdate</code>	Date of birth	Num	8
<code>bthwght</code>	Uncorrected birth weight	Num	8
<code>county</code>	County of occurrence of birth	Char	2
<code>detrace</code>	Race of mother	Char	2
<code>dobd</code>	Day of birth	Num	3
<code>dobm</code>	Month of birth	Num	3
<code>doby</code>	Year of birth	Num	3
<code>fmomn</code>	Mom's first name	Char	8
<code>fmomnsdx</code>	SOUNDEX transformation of Mother's first name	Char	3
<code>fn</code>	First name	Char	12
<code>fnsdx</code>	SOUNDEX transformation of first name	Char	3
<code>hisp</code>	Hispanic ethnicity indicator	Char	1
<code>ldadn</code>	Father's last name	Char	15
<code>ldadnsdx</code>	SOUNDEX transformation of father's last name	Char	3
<code>linkid</code>	Unique record identifier	Num	8
<code>ln</code>	Last name	Char	20
<code>lnsdx</code>	SOUNDEX transformation of last name	Char	3
<code>lrn</code>	Local Registrar's Number (USELESS!)	Char	6
<code>mathosp</code>	CA maternity hospital	Char	4

mn	Middle name	Char	12
mnsdx	SOUNDEX transformation of middle name	Char	3
race	Race - abridged coding	Char	1
sex	Gender of child	Char	1
sfn	State File Number (USELESS!)	Char	6
sporig	Spanish origin of mother	Char	1
twin	Twin indicator	Char	1
zip	Zip code of mother's residence	Char	5

Naming convention:

The names of variables that are used in the linkages are the **same** for the two data sets involved in the match. For instance, if sex is used to establish a link between two data sets, a variable named sex has to exist on each of the data sets to be linked. All variables that are ancillary, i.e., not used to establish a link, must have different names.

Step 5: Generating Value-Specific Frequency Information and the Set of u-Probabilities

All macros used in this section are stored in the file *macros for probabilistic record linkage preparations.sas*.

The generation of the value-specific frequency information is driven by three macros: **%freqsD**, **%reduceD**, and **%vsfreqsD**.

The purpose of the **%freqsD** macro is to generate for each variable participating in the linkage procedure a table with each possible value of the variable and its percent frequency. These tables are used to generate the value-specific frequency weights for each variable in the probabilistic linkage. The macro has three parameters: the name of the variable to be analyzed, whether or not missing values should be considered a valid category for this variable, and the variable type (numeric or character). The macro writes an output data set in the `sastmp` directory that is named after the variable with a `D` appended to it. For instance, for the variable `sex`, an output data set by the name of `sastmp.sexD` would be created. The output data set includes three variables: the variable value, the probability of occurrence of this value in the first data set (birth data), `probB`, the probability of occurrence of this value in the second data set (death data), `probD`.

Besides this file, the macro appends an observation to the file `sastmp.genfreqsD`. The file `sastmp.genfreqsD` consists of a description of the variable, `des`, which is the variable name and the general frequency (uncorrected sums of squares (USS)) for the variable for the birth and death file, `genfreqB` and `genfreqD`.

For some variables, the same probability value occurs for more than one variable value. The macro **%reduceD** sorts the values of a variable in descending order of their frequency, determines how often a probability level occurs, and groups values with the same probability level together.

Another issue is that for some variables, a large number of values occur rarely. In this situation, it is inefficient to use the complete table of value-specific frequencies. Rather we have chosen a threshold

probability, usually 0.0001, below which the value-specific probability is set to the same value. The macro **%vsfreqsD** writes an ASCII file that assigns for each variable value the value-specific probability level using IF ... ELSE ... constructs. The ASCII files are named using the variable name, appending a 'D', and adding the extension .sas. Note that the macro describes probabilities for the variable values from the birth file with the variable `prob`, the probabilities for the variable values from the death file are described with `prob_`.

The generation of the set of u-probabilities is based on the construction of a file of unlinkable pairs. The macro **%uprobsD** is used to create this file. Randomly records are merged from the birth and death file. Any records that match on date of birth, gender, and SOUNDEX transformed first name are eliminated to make sure that the file really only includes records that are not linkable. The resulting file `ulpairs` is evaluated for agreement on variable values for all the variables that are used in the linkage (**%getuprob**); the macro **%uprob** is then used to calculate the probability of agreement on a variable value for each variable. The result of these evaluations is stored in file `sastmp.uprobD`. The file consists of the name of the variable and its u-probability, i.e., the estimated probability of agreement based on the file `ulpairs`. Note that the macro calls to the **%uprob** macros happen via the SCL of the `gui.vsbvds.main.frame`.

The file `_dobyD.sas` needs to be edited. Otherwise, the linkage macros will crash.

Step 6: Setting of m- and u-Probabilities for Linkage Run

Prior to running a probabilistic linkage step, it is necessary to set the m- and u-probabilities for each variable. These probabilities are set in the form of global macro variables. For instance, for the variable `sex`, global macro variables name `&msex` and `&usex` are created to represent this variable's m- and u-probability respectively. The screen `gui.vsbvds.probs.frame` is used to set the m- and u-probabilities. The screen capture in Figure 7 shows the screen, as it appears when the screen is first entered.

The m-probabilities are shown in the left column, the u-probabilities in the right column. The box in the top right indicates which version of m-probabilities is shown. As we have not completed an iteration through the linkage procedure, the current iteration is zero. The u-probabilities are derived from [Step 5: Generating Value-Specific Frequency Information and the Set of u-Probabilities](#) described above. The m-probabilities are set to initial values that are based on experience. These initial values can be changed using this screen if desired. The 'Set' button in the right bottom corner is used to actually set the m- and u-probability for each variable. The 'End' button below it is used to exit the screen.

This screen also pops up after [Step 9: Recalculate m-Probabilities and Check Convergence](#) further below. The m-probabilities will then be updated to reflect the estimate from the linkage run most recently completed.

Note that by changing the iteration number, a different starting set of m-probabilities can be loaded. For instance to use the last set of m-probabilities from the 1997 linkage run as starting values for the 1996 linkage, copy the file `sastmp.mprobDX` (where X corresponds to the last completed linkage run) for 1997 in the `sastmp` directory for 1996 (they might be the same), and enter the number X as the iteration number on this screen.

doby	0.99	1	Iteration: <input type="text" value="0"/>
dobd	0.99	0.031159838430	
dobm	0.99	0.079438353529	
ln	0.95	0.000769378726	
lnsdx	0.95	0.003654548951	
fn	0.95	0.002115791498	
fnsdx	0.95	0.005577995768	
mn	0.95	0.001538757453	
mnsdx	0.95	0.005770340450	
ldadh	0.95	0.000577034045	
ldadnsdx	0.95	0.002115791498	
bmomn	0.95	0.000769378726	
bmomnsdx	0.95	0.003269859588	
sex	0.99	0.493748797845	
race	0.9	0.663781496441	<input type="button" value="Set"/> <input type="button" value="End"/>
detrace	0.9	0.656472398538	
hisp	0.9	0.505289478745	
sporig	0.9	0.474129640315	
zip	0.95	0.001154068090	

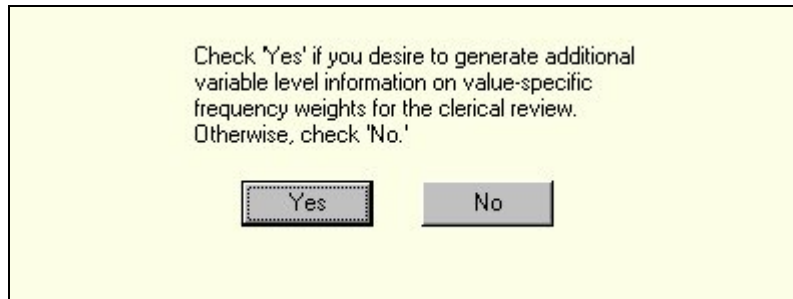
Figure 7: VSBVSD Linkage, Setting of M- and U-Probabilities

Step 7: Running a Linkage Step

Clicking on 'Linkage of vital statistics death and birth data' (Figure 3) leads first to a prompt asking for the linkage block to run. Block 1 of the linkage is based on date of birth and gender. Block 2 of the linkage is based on month of birth, year of birth, and the SOUNDEX transformed last name. Block 3 of the linkage is based on day of birth, year of birth, and the SOUNDEX transformed last name. *To complete a linkage step, all blocks must be run in this order.* This is necessary as the block 2 linkage run will only include deaths and births that did not match in block 1, etc. Note that it is recommendable to run a clerical review of the match results after each block has completed (see [Step 8: Clerical Review of Match Results](#) discussed next).

A second box (Figure 8) asks the user whether any detail information on the individual variables' contributions to the overall match weight should be retained in the data set of matches. Note that retaining this information will lead to higher storage requirements as for each variable two types of match weights are retained. It will also lead to slower execution times as SAS will have to work with larger data

sets in the background. However, the information can be useful in understanding the composition of the overall general and value-specific frequency weight especially for the first iteration.



Check 'Yes' if you desire to generate additional variable level information on value-specific frequency weights for the clerical review. Otherwise, check 'No.'

Yes No

Figure 8: VSBVSD Linkage, Pop-Up Box on the Inclusion on Detailed Variable-Specific Contributions to the overall Match Weight

It is important to keep in mind that the application does not prompt for a threshold for the value-specific frequency weight here. Rather, thresholds/cutoffs for the value-specific frequency weight are stored in the parameter settings as part of [Step 1: Setting Linkage Parameters and Reviewing Linkage Status](#). Matches with a value-specific frequency weight below the acceptable threshold minimum as set in Step 1 are eliminated before the matches are evaluated. Matches are evaluated by checking the match data sets for ties (e.g., two or more birth records link to the death record; two or more death records link to the same birth record, or both). Some ties can be resolved by retaining only the record with the higher match weight. Some ties can be randomized, especially those for which the two birth records and/or the two death records pertain to multiples.

Two macros stored in *macros for probabilistic record linkages.sas* carry out the linkage tasks. The first macro is **%linkmacD**, the second macro is **%mrkmtchs**.

The macro **%linkmacD** has the following parameters:

Input1	The first data set that includes information on deaths. Note that in the course of the macro all match and block variables will be renamed to include an _ as the last character of their name.
Input2	The second data set that includes information on births. The name of all block and match variables remains unchanged.
Lnkd	The output file of matched pairs.
Path	The path to the location of the files that were created by the %vsfreqsD macro discussed in Step 5: Generating Value-Specific Frequency Information and the Set of u-Probabilities . Usually, these files are stored in the same library that the SAS libname <code>sastmp</code> refers to.
Compcrit	A critical value for all simple agreements: The macro forms all possible matched pairs within a block. As this can result in a large number of matches, as a first cut the macro evaluates all simple agreements within a block, in other words, a variable <code>comp</code> is created that counts the number of times two variables agree. If <code>compcrit</code> is set to a value larger than zero, all matched pairs with <code>compcrit</code> agreements or fewer are removed from the file of matched pairs prior to the calculation of general and value-specific frequency weights.
Crit	A critical value for the general frequency weight. In contrast to the value-specific frequency

	<p>weight, the general frequency weight is based on m- and u-probabilities only. In case of agreement on a variable, the general frequency weight for the variable is the ratio of m- and u-probability; in case of disagreement on a variable, the general frequency weight is the ratio of $(1 - m\text{-probability})$ and $(1 - u\text{-probability})$. Note that in case of a variable with a high m-probability and low u-probability (the most desirable situation for a match variable), this leads to a large positive number in case of agreement and number less than 1 in case of disagreement. As logs are taken prior to obtaining the general frequency weight by summing the contributions of all individual variables, agreement results in a large positive weight, while disagreement results in a large negative weight. Note that the general frequency weight is the <u>same</u> for different values of the variable. (For more details on probabilistic record linkages, see also 3 above).</p> <p>If <code>crit</code> is larger than 0, its value is used to further exclude matched pairs with general frequency weights equal to or below the <code>crit</code> value.</p>
Block	Indicates the block number, in case of the death linkage, this is 1, 2, or 3.
Srtvr1	The first numeric block variable.
Srtvr2	The second numeric block variable.
Srtvr3	The third numeric block variable.
Srtvr4	The fourth numeric block variable.
Srtvr5	The fifth numeric block variable.
Csrtvr1	The first character block variable.
Csrtvr2	The second character block variable.
Csrtvr3	The third character block variable.
Csrtvr4	The fourth character block variable.
Csrtvr5	The fifth character block variable.
Lastvar	The name of the last block variable. If <u>any</u> numeric block variables are present, this is the last numeric block variable. If <u>only</u> character block variables are present, this is the last character block variable.
Special	Available parameter, not yet used. Can be used to force specific conditions on the file of matched pairs.
Type	Version of linkage: <code>D</code> for death, <code>T</code> for transfers, <code>R</code> for re-admissions, <code>P</code> for prenatal/postpartum records, <code>I</code> for infant PDD, <code>M</code> for maternal PDD, <code>A</code> for add-on.
Develop	Indicator that is <code>Y</code> or <code>N</code> . If <code>Y</code> , variable-specific contributions to the value-specific frequency weight are included in the file of matched pairs. If <code>N</code> , these weights will not be included.

Furthermore the macro needs the following global parameters to be set:

- All m- and u- probabilities for the match variables.
- The year of the linkage.
- The vital statistics input file type (VS, BCF, VSBVSD).

The macro **%mrkmtchs** executes immediately after the **%linkmacD** macro. The macro has three parameters, the version of linkage (`D` for death), the current block number, and the vital statistics input file type (VS, BCF, VSBVSD). It resolves any ties in the matched file by declaring this matched pair the final match that had the largest value-specific frequency weight. If there is a tie for the value-specific frequency weight, the match is accomplished by randomization. Note that there were no ties for the value-

specific frequency weight in the 1997 vital statistics birth/death linkage. The issue of randomized matching is of more importance for the vital statistics birth and infant PDD match as well as the transfer and re-admissions matches. A more elaborate discussion of randomized matching can therefore be found in [Section 7, Step 15: Linkage of Newborn Discharge and Vital Statistics Birth Data](#).

Step 8: Clerical Review of Match Results

The following options are available for the clerical review:

- Review the first 1000 virtual IDs (A virtual ID is defined by a unique combination of block variables).
- Review the first 1000 twins.
- Review the first 1000 virtual IDs with at least three records.
- Review the first 1000 matches in ascending order of the value-specific frequency weight.
- Review the first 1000 matches in descending order of value-specific frequency weight.

All these options are applied to the most recent completed linkage block, i.e., a clerical review should happen immediately after the linkage for one block has been finished.

The screen capture in Figure 9 displays the clerical review screen.³ The screen is primarily used to establish match cutoffs. In some circumstances, specific `linkids` are marked for removal from the matched file or marked for inclusion even though the match weight is below what is considered the minimum match weight. These circumstances are rare as just the size of the linkage tasks at hand does not allow time for an elaborate manual clerical review.

³ Confidential information is blacked out to not violate the privacy of persons in the data set.

Virtual ID: 553 Block size: 5 Birth Date: 20/1997 Gender: 2

Vital Statistics Birth Record(s):

	linkid	Sex of child	doby	dobm	dobd	ln	fn	mn	l
1	307355	Female	1997		20				
2	307356	Female	1997		20				
3	274042	Female	1997		20				L
4	320487	Female	1997		20				L
5	357930	Female	1997		20				

Vital Statistics Death Record(s):

	linkid_	sex_	doby_	dobm_	dobd_	ln_	fn_	mn_	ldadm_
1	1997176074	2	1997		20				
2	1997176074	2	1997		20				
3	1997176074	2	1997		20				
4	1997176074	2	1997		20				
5	1997121213	2	1997		20				

Linkage information:

	comp	gfw	vsfw	zipjarol	keep
1	9	20.41	50.50	0.613	N
2	9	20.41	50.50	0.613	N
3	10	17.73	53.60	0.275	N
4	13	54.32	127.83	1.000	Y
5	14	63.64	149.44	1.000	Y

Navigation buttons: [Previous], [Next], [First], [Last], [End]

Figure 9: VSBVSD Linkage, Clerical Review Screen

Step 9: Recalculate m-Probabilities and Check Convergence

After all three blocks of a linkage run are completed, the m-probabilities have to be re-calculated to check convergence. This step can be accomplished by clicking on the pushbutton 'Recalculate m-probabilities; check convergence' (Figure 3). First a temporary file is created via the macro **%inputmprob** stored in *macros for probabilistic record linkage.sas*. This temporary file retains information for each match on which variables matched and which variables did not match. The macro takes three parameters, the version of linkage currently run (e.g., **D** for deaths, **P** for prenatal/postpartum maternal records, etc.), the year of the linkage, and the vital statistics input file type (VS, BCF, VSBVSD). Note that for the death linkages the input file type is always VS.

The macro **%getmprob** stored in *macros for probabilistic record linkage.sas* is then used to obtain an updated file of m-probabilities, `sastmp.mprobDX`. Where **x** corresponds to the current linkage run iteration. If **x** is larger than 1, the macro also compares the current iteration's m-probabilities to the previous iteration and prints the result in the SAS output window. The result can be used to determine whether or not convergence has occurred. We considered a linkage run to have converged if all differences in m-probabilities were less than 0.01.

After the recalculation of the m-probabilities is completed, the GUI automatically calls the entry `gui.vsbvds.probs.frame` as discussed in [Step 6: Setting of m- and u-Probabilities for Linkage Run](#). If convergence has not yet happened, it can be used to set the most current version of m-probabilities that will be used in the next iteration of the linkage procedure.

Step 10: Generating Results File and Result Summary

Clicking on 'Generate results file and summarize results' (Figure 3) leads to a pop menu that presents three options:

a) Should the results file and summary be generated from two years of data input? b) Should the results file and summary be generated from one year of data input? c) Is the results file already created and should only a results summary be generated? For instance, if the vital statistics death file for the subsequent year was used in the linkage and a results file has not been previously created, a) is the appropriate choice.

The results file is generated as part of the SCL of `gui.vsbvds.main.frame` in the `pb_summary` section. The results file is written permanently as `saslib.vsbvdsXXXX` where `XXXX` corresponds to the year for which the linkage is carried out.

The file includes all the information on the vital statistics birth file and where linked all the information on the vital statistics death file for those infants that were linked. It includes only information from the vital statistics death file for all records that were not linked to a birth record.

As the variables that were used in the linkage have the same name on both data sets, all variables from the death file are renamed by having a `D` added to their original name, e.g., `sex` is renamed to `sexD`. For the unlinked death records, the variables `sex`, `bthdate`, `hisp`, `race`, and `zip` are generated. These records can still be eligible for linkage in the PDD-VS linkage, therefore they need to have variables that the link can be based upon.

Two additional variables are generated: The variable `death` with values 'Y' and 'N,' which indicates whether or not an infant died, and the variable `ndeath` with values 'Y' and 'N,' which indicates whether or not an infant died in the neonatal period.

The linkage is summarized with a tabulation of the linked/unlinked infant deaths and linked/unlinked neonatal deaths. Furthermore, a univariate descriptive statistical summary for the value-specific frequency weight is generated. Relevant portions of the summaries should be inserted into the results tables for the current linkage year.⁴

Step 11: Bias Check

The bias check verifies the distribution of birth weight among infant deaths, neonatal deaths, postneonatal deaths and survivors. Relevant portions of the bias check should be inserted into the results tables for the 1997.

⁴ The results spreadsheet is usually stored as `XXXX linkage results.xls` (where `XXXX` should be replaced by the year of the linkage) in `\pdd-vs linkages\results`.

7. Linkage of Vital Statistics Birth and Infant/Maternal Hospital Discharge Data

Figure 10 shows the main navigation GUI screen for accomplishing the match between the vital statistics birth and infant/maternal hospital discharge records.

Enter Linkage Parameters:	<input type="checkbox"/>	Project Name	Save
Run readdat1: Extract infant and maternal discharge from main PDD file	<input type="checkbox"/>		
Run readdat2: Combine this year's and next year's infant/maternal records	<input type="checkbox"/>		
Run extpddi: Generate minimum files for data linkages	<input type="checkbox"/>	Check Log	Check Output
Run extpddm: Generate minimum files for data linkages	<input type="checkbox"/>		
Run readvs: Generate vital statistics data	<input type="checkbox"/>	End	
Run extvs: Extract minimum VS data set for linkage	<input type="checkbox"/>		
Run getdat: Prepare hospital level record comparison	<input type="checkbox"/>		
Run gettable: Generate hospital level comparison of record numbers	<input type="checkbox"/>		
Update hospital cross walk file	<input type="checkbox"/>		
Run final preparation step to add DHS ID to PDD databases	<input type="checkbox"/>		
Value-specific frequencies and estimation of u-probabilities	<input type="checkbox"/>		
Setting of m- and u-probabilities for linkage run	<input type="checkbox"/>		
Linkage of newborn discharge and vital statistics birth data	<input type="checkbox"/>	Run extvs2: Add unlinked PDDI records to sastmp.subvs	<input type="checkbox"/>
Linkage of maternal discharge and vital statistics birth data	<input type="checkbox"/>		
Review newborn/maternal data linkage	<input type="checkbox"/>		
Generate results file	<input type="checkbox"/>		
Recalculate m-probabilities; check convergence	<input type="checkbox"/>		
Results Summary	<input type="checkbox"/>		
Check results for bias	<input type="checkbox"/>		

PDD-VS Linkage
Health Information Solutions, October 2000

Figure 10: VSPDD Linkage, Main Screen

As for the death linkage, the push buttons in the left portion of the screen guide through the linkage steps. As explained under the death linkage the current status and any match parameters can be saved and restored using the top right portion of the screen (see beginning of Section 6).

The actions executed by each push button are explained in detail below.

Step 1: Setting Linkage Parameters and Reviewing Linkage Status

The first push button 'Enter Linkage Parameters' in Figure 10 leads to a screen that displays parameters for the current linkage environment.

Path to linkage macros:	d:\pdd-vs linkages\programs	
Path to formats:	d:\pdd-vs linkages\formats	
Path to output data:	d:\pdd-vs linkages\temporary files	
Year of linkage:	1997	
VS input file?	<input type="radio"/> VS <input type="radio"/> BCF <input checked="" type="radio"/> VSBVSD	
Next year of PDD data available?	<input checked="" type="radio"/> Yes <input type="radio"/> No	
Additional weight information?	<input type="radio"/> Yes <input checked="" type="radio"/> No	
M-Probabilities Version:	0	0
Linkage Block:	1	1
VSPFW cutoffs:	25	10
<div>Done</div> <div>Cancel</div> <div>Reset</div>		

Figure 11: VSPDD Linkage: Linkage Parameter Settings

This screen is very similar to the one displayed in Figure 4 for the vital statistics birth and death linkage. There are two important differences though:

1. The Vital Statistics input file radio box gives one more choice, VSBVSD. This choice should be checked if the input data is the linked vital statistics birth infant death file that was created previously (see [Section 6, Linkage of Vital Statistics Birth and Vital Statistics Death Data](#). Specifically, [Step 10: Generating Results File and Result Summary](#)). The vital statistics birth/death resulted in the generation of an output file named `saslib.vsbvsdXXXX` where `XXXX` refers to the year of the linkage. It is this file that is used as linkage input if VSBVSD is checked.
2. The VSPFW cutoffs are the same for all blocks of the newborn discharge record linkage (there is only one, though) and all blocks of the maternal discharge record linkage. The entry in the first entry box under VSPFW cutoffs is used as the cutoff for the linkage to the infant discharge data; the entry in the second entry box is used as the cutoff for the linkage to the maternal discharge data.

Step 2: Run `readdat1`: Extract Infant and Maternal Discharge Record from Master Patient Discharge Data File

It is assumed that the master patient discharge data are stored in a directory whose libname in SAS is `hdf`. Furthermore, it is assumed that the master patient discharge data has previously been created, i.e., the GUI does not create the master patient discharge file itself.

The macro **%readdat1** is used to read all discharge records that:

- Pertain to a newborn discharge.
- Pertain to a discharge of an infant that was born in the current year of linkage.

- Pertain to a discharge of an infant that was born in the year prior to the year for which the linkage is carried out.
- Pertain to a delivery discharge.

The reason not only infants with birth date in the current year but also birth date in the prior year are extracted is that records to infants born in the prior year are needed in the re-admission and transfer linkage for that year. Extracting records for both linkages at the same time saves resources. Note that the delivery discharges will also include discharges to women who delivered their baby in the previous year.

The macro **%readdat1** stored in *macros for getting data ready for linkages.sas* assumes the existence of two global macro variables: `year` which refers to the current year of the linkage and `prevyear` which refers to the prior year.

A record is considered a newborn record if any of these conditions hold:

- The principal or any of the 24 other diagnoses starts with the two character sequence 'v3.'
- The Diagnosis Related Group (DRG) is one of 385 to 391.
- 1994 or earlier: The admission type is '04' (newborn) or the admission source is '18' (newborn).⁵
1995 or later: The admission source code for level of care is '7' (newborn).

All other records are considered to be infant transfer or re-admission records. Note that starting with year 1995 and for later years, records with admission type '3' (infant less than 24 hours old) are also included in the extracted data set.

A record is considered a maternal delivery record if any of these conditions hold:

- The principal or any of the 24 other diagnoses is one of 'v270', 'v271', 'v272', 'v273', 'v274', 'v275', 'v276', or 'v277'.
- The Diagnosis Related Group (DRG) is one of 370 to 375.
- 1994 or earlier: The admission type is '05' (delivery).

Note that as a record is added to the patient discharge data depending upon the date of discharge, this extraction will include some maternal delivery records to women who actually delivered their baby in the previous year.

The infant data extracted are stored in the file `saslib.hdfiXXXX` where `XXXX` corresponds to the year of discharge data from which records are extracted; the maternal data extracted are stored in the file `saslib.hdfmXXXX`.

Step 3: Run readdat2: Combine this Year's and Next Year's Infant and Maternal Records into one File

The macro **%readdat2** stored in *macros for getting data ready for linkages.sas* combines the data extracted in the previous step for different years into one master data set with all newborn, infant, and maternal discharge records that were recorded for or were in relationship to a birth event in the year of the linkage.

⁵ For 1983, records with admission type '05' (newborn) are also included. Note that this coding was only used for this one year.

Specifically, the file `saslib.hdfiXXXXYYYY` is created by:

1. Extracting all infant discharges with birth year equal to the year of the linkage from file `saslib.hdfiXXXX` where `XXXX` refers to the year of the linkage.
2. Extracting all infant discharges with birth year equal to the year of the linkage from file `saslib.hdfiYYYY` where `YYYY` refers to the year following the linkage.

The file `saslib.hdfmXXXXYYYY` is created by:

1. Extracting all maternal delivery records for the year of the linkage from file `saslib.hdfmXXXX` where `XXXX` refers to the year of the linkage.
2. Extracting all maternal delivery records with admission date in the year of the linkage from file `saslib.hdfmYYYY` where `YYYY` refers to the year following the linkage.

Note that in the first step, we do not exclude deliveries to women who were admitted in the previous year. We proceed this way since some women get admitted in the previous year, however, the birth might still occur in the subsequent year.

If the patient discharge data are not yet available for the year following the current year of linkage, only the data extracted in the first step are used to participate in any matching. Note that the parameter `next` that is set in step 1 on the parameter entry screen (on page 28), indicates whether or not the next year of discharge data is available or not.

The macro **%readdat2** needs the following global macro variables: `year`, the current year of the linkage, `nextyear`, the year following the current year, and `next`, an Y/N indicator flag that indicates whether next year's discharge data are available or not.

After 1994, OSHPD changed the definitions of some of the variables recorded in the patient discharge record. Several of the variables that are used in the linkage are affected by these changes.

- Admission source
- Admission type
- Discharge status
- Hispanic origin
- Race
- Payer source

For the generation of the 1994 file, the new version of each variable is retained. Furthermore, for the 1995 data where possible, the 1994 coding is generated from the 1995 coding and added to the file. Race as reported in 1995 records is recoded according to the following table:

Race	Ethnicity	1995 Race	1994 Race
	Any	'2'	'02' Black
	'1'	Any	'03' Hispanic
	Not '1'	'1'	'01' Non-Hispanic White

	Not '1'	'3'	'04' Native American
	Not '1'	'4'	'05' Asian
	Not '1'	'5'	'06' Other
	Not '1'	'6'	'07' Unknown

Step 4: Run extpddi: Generate Minimum Infant Discharge File for Vital Statistics Birth and Infant Discharge Record Data Linkage

It makes sense to only retain the minimally necessary information from the vital statistics birth and infant discharge file for the purpose of the linkage. This step is accomplished by running the **%extpddi** macro stored in *macros for getting data ready for linkages.sas*. The macro assumes the existence of two global parameters: `year` referring to the current year of linkage and `nextyear` referring to the year following the current linkage year.

Besides retaining only minimal information needed for the linkage, the macro **%extpddi** also generates several new variables that represent versions of the variables needed for the linkage that are better suitable for the linkage task.

1. The macro generates a variable named `newborn`. This variable indicates whether any of the 25 diagnoses in the record started with 'v3' indicating a newborn admission.
2. For 1995 or later years, the variable `race` is generated as a copy of the OSHPD patient race variable, however, category '5' (other) is recoded as '6' (unknown).
3. The macro generates a variable called `bwgrp`. This variable refers to the birth weight group to which an infant was assigned at time of birth. Birth weight groups are specified in 250 gram intervals starting from 500 grams to 2,500 grams. Assignment is made according to the following rules:

ICD-9 codes starting with 764.0, 764.1, 764.2, 764.9, 765.0, 765.1 ...	Birth weight category	Category
ending in 1	Under 500 grams	1
ending in 2	500 to under 750 grams	2
ending in 3	750 to under 1,000 grams	3
ending in 4	1,000 to under 1,250 grams	4
ending in 5	1,250 to under 1,500 grams	5
ending in 6	1,500 to under 1,750 grams	6
ending in 7	1,750 to under 2,000 grams	7
ending in 8	2,000 to under 2,500 grams	8
ending in 9	2,500 grams or over	9
No fifth digit given	Categorized as after any of the previous categories	10

All other kids are categorized as category 11.

4. The variable `csr` is generated based on whether or not the newborn discharge record gives evidence of that this baby was delivered vaginally or via cesarean section. If the fifth digit of the principal or any of the other 24 diagnoses that start with 'v30', 'v31', 'v32', 'v33', 'v34', 'v35',

'V36', or 'V39' is a 1, the baby is considered to be delivered via cesarean, otherwise vaginally. The variable `csr` is a '0'/'1' variable.

5. The variable `bornin` is generated: '1' indicates that this baby was born in the hospital; '2' indicates that the baby was born prior to admission to the hospital; '3' indicates that the baby was born prior to admission to the hospital and never hospitalized after birth. The variable is based on the 25 diagnosis codes. For any diagnosis code starting with 'V3' the fourth digit is checked. If it is '1' the baby was born in the hospital; if it is '2' the baby was born prior to admission to the hospital; if it is '3' the baby was born prior to admission to the hospital and never admitted to the hospital after birth. For simplicity a '0'/'1' variable named `bornout` is created that is '0' if baby is born in the hospital, and '1' if baby is born prior to admission to the hospital.
6. The variable `twin` is generated to indicate whether or not the record refers to the birth of a multiple. If any of the 25 diagnoses is one of 'V31', 'V32', 'V33', 'V34', 'V35', 'V36', 'V37', or 761.5, it is assumed to be a multiple birth, otherwise a singleton. The variable `twin` is a 'Y'/'N' variable.
7. For years after 1994:

A record is not considered a re-admission (`rehosp='N'`) if any of the following conditions are true:

- a. The admission source level of care is equal to '7' (newborn).
- b. The baby is a newborn born in the hospital or born outside the hospital and then admitted to the hospital and its admission source level of care is one of the following codes: '1' (Home), '3' (Ambulatory Surgery), '7' (Newborn), '8' (Prison/Jail), '9' (Other).
- c. The baby is a newborn, the admission date is equal to its birth date, it is assigned to one of the neonatal DRGs, and the admission source level of care is one of the following codes: '1' (Home), '3' (Ambulatory Surgery), '7' (Newborn), '8' (Prison/Jail), '9' (Other).

For years prior to 1994:

A record is not considered a re-admission (`rehosp='N'`) if any of the following conditions are true:

- a. The admission source is equal to '18' (newborn).⁶
- b. The baby is a newborn born in the hospital or born outside the hospital and then admitted to the hospital and its admission source level of care is one of the following codes: '01' (1983: Routine), '02' (1983: Emergency Room), '05' (1983: Newborn), '06' (1983: Other), '11' (Routine), '12' (Emergency Room), '17' (Home Health Service), '18' (Newborn), '19' (Other).
- c. The baby is a newborn, the admission date is equal to its birth date, it is assigned to one of the neonatal DRGs, and the admission source level of care was one of the following codes: '01' (1983: Routine), '02' (1983: Emergency Room), '05' (1983: Newborn), '06' (1983: Other), '11' (Routine), '12' (Emergency Room), '17' (Home Health Service), '18' (Newborn), '19' (Other).

For 1994:

For those records that originate from the 1995 discharge data, the rules are used that are listed under for years after 1994; for those records that originate from the 1994 discharge data, the rules are used that are listed under for years prior to 1994.

⁶ For 1983, admission source '05' is also used.

8. An indicator is generated that reflects whether or not an infant died in the hospital.
- For years after 1994:
- If the discharge status is '11' (died), this record pertains to a death (`death = 'Y'`).
- For years prior to 1994:
- If the discharge status is '08' (died), this record pertains to a death (`death = 'Y'`).
- For 1994:
- If the record comes from the 1995 file and the discharge status is '11' (died), this record pertains to a death (`death = 'Y'`). If the record comes from the 1994 file and the discharge status is '08' (died), this record pertains to a death (`death = 'Y'`).
9. Assign each newborn discharge record a discharge status code that can be used to correspond to admission codes in a transfer:
- The following table is used for patient discharge records from files after 1994:

Code	Code2	Discharge Status for Newborn Records	Admission Source for Transfers
'1'	'1'	'01' Home '10' Left Against Medical Advice '12' Home Health Service	'1' Home '7' Newborn
'2'	'2'	'09' Prison	'8' Prison
'3'	'2'	'02', '05' Acute Care Facility	'5' Acute Care Facility
'4'	'2'	'03', '06' Other Care Facility	'6' Other Care Facility
'5'	'2'	'04', '07' Long Term Care Facility	'4' Long Term Care Facility
'6'	'2'	'08' Residential Care Facility	'2' Residential Care Facility
'7'	'1'	'13' Other	'3' Ambulatory Surgery '9' Other
'8'	'8'	'11' Died	NA

The variables `code` and `code2` are created according to column 3 for all records that do not pertain to re-admissions or transfers. For re-admissions and transfers, they are created according to column 4.

The following table is used for patient discharge records from 1994 or earlier years:

Code	Code2	Discharge Status for Newborn Records	Admission Source for Transfers
'1'	'1'	'01' Routine Discharge '06' Left Against Medical Advice '07' Home Health Service	'01' 1983: Routine '02' 1983: Emergency Room '05' 1983: Newborn '11' Routine Discharge '12' Emergency Room '17' Home Health Service '18' Newborn

'3'	'2'	'02' Short-Term Acute Care Facility	'13' Short-Term Acute Care Facility
'4'	'2'	'03' Intermediate Care Facility	'14' Intermediate Care Facility
'5'	'2'	'04' Skilled Nursing Facility	'15' Skilled Nursing Facility
'6'	'2'	'05' Other Facility	'16' Other Facility
'7'	'1'	All other codes	'19' Other and all other codes
'8'	'8'	'08' Died	NA

10. For years 1995 and later, a variable called `smhosp` is created. The variable indicates for a newborn record whether the discharge status indicated a transfer within the same hospital, to a different hospital, or whether it did not indicate any transfer. For a transfer/re-admissions, the variable indicates whether the current admission comes from the same hospital, a different hospital, or not from a hospital.

Smhosp	Discharge Status for Newborn Records	Admission Source Care Licensed Under for Transfers/Readmissions
'1'	'02' To acute care in this hospital '03' To other care in this hospital '04' To long term care in this hospital	'1' This Hospital
'2'	'05' To acute care in other hospital '06' To other care in other hospital '07' To long term care in other hospital	'2' Another Hospital
'3'	All others	'3' Not a Hospital

Finally, three output data sets are created that are all used in the subsequent linkages:

- `Sastmp.subhdfI`: Includes all newborn discharge records, i.e., all records for which `rehosp='N'`. This file is used in the linkage of vital statistics birth data to infant discharge data.
- `Sastmp.subhdfT`: Includes all transfer records to infants one year or younger, i.e., all records for which `rehosp = 'Y'` with admission source '03' (1983: Other hospital), '04' (1983: Other Facility), '13' (Short-Term Acute Care Facility), '14' (Intermediate Care Facility), '15' (Skilled Nursing Facility), or '16' (Other Facility) for years prior to 1995, and '4' (Long Term Care Facility), '5' (Acute Inpatient Hospital Care), or '6' (Other inpatient hospital care) for 1995 or later.
- `Sastmp.subhdfR`: Includes all re-admission records to infants one year or younger, i.e., all remaining records for which `rehosp = 'Y'`.

Table 3 shows the variables that are retained for the linkages in each of the three files.

Table 3: VSPDD Linkage, Variables Retained in Newborn, Transfer, and Re-Admission Linkage Input Files

Variable Name	Description	Type	Length
ADMDATE	Admission Date	Numeric	5
BTHDATE	Birth Date	Numeric	5
COUNTY	County of occurrence	Character	2

DIAG1	Diagnosis 1	Character	5
DIAG2	Diagnosis 2	Character	5
DIAG3	Diagnosis 3	Character	5
DIAG4	Diagnosis 4	Character	5
DIAG5	Diagnosis 5	Character	5
DISDATE	Discharge Date	Numeric	5
DRG	Diagnosis Related Group	Character	3
LINKID	Unique record ID	Character	13
PAYMSO	Payer Source	Character	2
PRDIAG	Principal Diagnosis	Character	5
ZIPHOSP	Hospital Zip	Character	5
bornin	Born in/outside the hospital	Character	1
bwgrp	Birth weight group	Character	2
Code	Admission/Discharge Code	Character	1
Code2	Admission/Discharge Abridged Code	Character	1
Csr	Delivery mode (c/s vs. vaginal)	Character	1
death	Death in hospital	Character	1
dobd	Day of birth	Numeric	3
dobm	Month of Birth	Numeric	3
doby	Year of Birth	Numeric	3
Hisp	Patient Ethnicity	Character	1
hospid	Hospital identifier	Character	6
Race	Patient race	Character	1
rehosp	Re-admission indicator	Character	1
Sex	Patient Sex	Character	1
smhsp	Patient admitted/discharge to same hospital	Character	1
Twin	Multiple birth	Character	1
Zip	HDF Zip	Character	5

Highlighted variables are added to the file. Note that the variables `hisp` and `smhsp` are only used for years 1995 and later.

As mentioned earlier, starting from 1995, OSHPD made several changes to some of the variables needed for the data linkage. Payer source, race, and ethnicity are among the variables affected by this change. For 1995 onwards, the macro brings the coding used for payer source in concordance with the one used by DHS as shown in the following table.

Payer Source	'02'	'01' Medicare '02' MediCal
	'03'	'03' Worker's Compensation

	'05'	'05' CHAMPUS/CHAMPVA/VA '06' Other Governmental
	'06'	'10' Blue Cross/Blue Shield (not HMO/PPO)
	'07'	'09' Private Insurance Company (not HMO/PPO)
	'08'	'07' HMO '08' PPO
	'09'	'11' Self-Pay
	'10'	'12' Charity Care '13' No Charge
	'11'	'14' Other Non-Governmental

Starting from 1995, the linkage uses race and ethnicity as two separate variables.

Step 5: Run extpddm: Generate Minimum Maternal Discharge File for Vital Statistics Birth and Maternal Discharge Record Data Linkage

The macro **%extpddm** stored in *macros for getting data ready for data linkages.sas* generates the file `sastmp.subhdfM`. As in the previous step, only variables are retained that are necessary for the data linkage of vital statistics birth file and maternal discharge record. The macro assumes the existence of two global variables, `year`, referring to the year of the linkage, and `nextyear`, referring to the year following the linkage.

Besides retaining only minimal information needed for the linkage, the macro **%extpddm** also generates several new variables that represent versions of the variables needed for the linkage that are better suitable for the linkage task.

1. The macro generates a variable named `bthdate`. This variable represents the estimated birth date of the baby. If the principal procedure date is not missing, the birth date is the principal procedure date, otherwise it is the admission date.
2. The macro generates a variable named `twin`. This variable is 'Y' for all multiple deliveries, otherwise, it is set to 'N'. A record is considered to pertain to a multiple delivery if one of the 25 diagnosis codes in the record is one of the following: '761.5', 'v27.2', 'v27.3', 'v27.4', 'v27.5', 'v27.6', 'v27.7', or '651'.
3. The macro generates a variable named `fdeath`. This variable indicates whether the current record pertains to a fetal death ('Y'). A record is considered to pertain to a fetal death if one of the 25 diagnosis codes in the record is one of the following: 'v27.1', 'v27.3', 'v27.4', 'v27.6', or 'v27.7'.
4. The macro generates a variable named `csr`. This variable indicates whether the current record pertains to a cesarean section ('Y'). A record is considered to pertain to a cesarean section if one of the 21 procedure codes in the record is one of the following: '740', '741', '742', '744', '745', '749', or if one of the 25 diagnosis codes in the record is: '6697' or if the Diagnosis Related Group (DRG) is '370' or '371'.

As mentioned earlier, starting from 1995, OSHPD made several changes to some of the variables needed for the data linkage. Payer source, race, and ethnicity are among the variables affected by this change. The macro brings the coding used for payer source in concordance with the one used by DHS (see above). Starting from 1995, the linkage uses race and ethnicity as two separate variables.

The following variables are included in the minimum maternal discharge record file used for the data linkage:

Table 4: VSPDD Linkage, Variables Retained in Maternal Delivery Record Input Linkage Files

Variable name	Description	Type	Length
Linkid	Unique record identifier	Character	13
Paymso	Payer Source	Character	2
bthdate	Estimated birth date	Numeric	8
Csr	Delivery mode (c/s vs. vaginal)	Character	1
Dobd	Day of birth	Numeric	3
Dobm	Month of birth	Numeric	3
Doby	Year of birth	Numeric	3
Fdeath	Fetal death	Character	1
Hisp	Patient Ethnicity	Character	1
Hospid	OSHPD hospital ID	Character	6
madmdate	Admission Date	Numeric	5
mbthdate	Birth Date	Numeric	5
Mdobd	Maternal day of birth	Numeric	3
Mdobm	Maternal month of birth	Numeric	3
Mdoby	Maternal year of birth	Numeric	3
Mdrg	Diagnosis Related Group	Character	3
mppdate	Principal Procedure Date	Numeric	5
mziphosp	Hospital Zip	Character	5
race	Race	Character	1
twin	Multiple birth	Character	1
zip	HDF Zip	Character	5

Highlighted variables are added to the file. Note that the variable `hisp` is only used for years 1995 and later.

Step 6: Run readvs: Read Vital Statistics Data

The use of this step is a bit tricky since we added the ability to add infant deaths to the GUI.

For years for which only the vital statistics birth data are available, the sequence of events is a little confusing.

1. Data are read by running **%readvs** from `gui.vspdd.main.frame`.
2. Deaths are linked by running the steps of `gui.vsbvds.main.frame`.
3. Return to `gui.vspdd.main.frame`, however **%readvs** does not have to be re-run after these steps as the file that is used to link infant/maternal discharge records to is `saslib.vsbvdsXXXX` where `XXXX` refers to the year for which the linkage is carried out.

There is no such sequencing for the birth cohort file.

The macro **%readvs** stored in *macros for probabilistic record linkage preparations.sas* reads the vital statistics birth data and birth cohort file data from the ASCII input files provided by DHS. After clicking on the appropriate push button, the user is prompted to input the path to the external file in which the data are stored as well as its record length (Figure 12).

Figure 12: VSPDD Linkage, Input Parameters for Reading Raw Vital Statistics Birth Data

Note that the macro **%readvs** has to be updated for years for which we do not yet have data to reflect any changes in the record layout that might have happened. Our experience though is that the column position of variables in this file remains constant over time. However, it is possible that variables get added in previously empty space, or that variables get dropped.

Step 7: Run extvs: Extract Minimum Vital Statistics Data File for Linkage to Maternal and Infant Discharge Data

The macro **%extvs** stored in *macros for probabilistic record linkage preparations.sas* was already discussed in [Step 4: Extracting Subfile from the Vital Statistics Birth File to be Used for Linkage in Section 6](#). The macro is also used to extract the vital statistics data needed for the linkage to the infant and maternal discharge data. The difference is that the `vstype` parameter of the macro is set to either BCF (if the birth cohort file is available for the year of the linkage) or VSBVSD (if the linkage to the death file is produced by HIS).

IF VSBVSD is used, i.e., the linkage is to the HIS produced link between vital statistics birth and death data, for unlinked death records, the `linkid` is set to `dlinkid`. As `dlinkid` was generated as:

```
dlinkid = &year.000000 + _N_;
```

it is possible to proceed in this fashion.

The file generates/edits several variables for the linkage:

- Maternal day of birth, month of birth, and year of birth.
- Missing payer sources are recoded to '99'. Payer source '01' is recoded as '02'.
- Birth weight groups are created according to the following table:

Under 500 grams	'01'
500 to under 750 grams	'02'

750 to under 1,000 grams	'03'
1,000 to under 1,250 grams	'04'
1,250 to under 1,500 grams	'05'
1,500 to under 1,750 grams	'06'
1,750 to under 2,000 grams	'07'
2,000 to under 2,500 grams	'08'
All others with birth weight not equal to 9,998 or 9,999	'11'
9,998 and 9,999	'99'

- For years prior to 1995, maternal race and Spanish origin are used to code the variable `race` used for the data linkage according to the following table:

'10' (White) and not of Hispanic Origin	Non-Hispanic White ('1')
'20' (Black) and not of Hispanic Origin	Black ('2')
Any race and of Hispanic Origin	Hispanic ('3')
'30' (American Indian), '57' (Eskimo), '58' (Aleut)	Native American/Eskimo ('4')
'4x' (Asian), '52' (Asian Indian), '53' (Filipino)	Asian ('5')
'98' (Unknown), '99' (Not Stated)	Unknown ('7')
All others	Other ('6')

- For years 1995 and later, maternal race is used to code the variable `race` used for the data linkage according to the following table:

'10' (White)	White ('1')
'20' (Black)	Black ('2')
'30' (American Indian), '57' (Eskimo), '58' (Aleut)	Native American/Eskimo ('3')
'98' (Unknown), '99' (Not Stated)	Unknown ('6')
All others	Asian/Pacific Islander ('4')

- For years 1995 and later, Spanish origin of mother is used to code the variable `hisp` used for the data linkage according to the following table:

'1' (Not Hispanic)	Non-Hispanic ('2')
'9' (Unknown)	Unknown ('3')
All others	Hispanic ('1')

Table 5 shows the vital statistics dataset for the data linkage.

Table 5: VSPDD Linkage, Variables Retained from Vital Statistics Data for Linkage Input File

Variable Name	Description	Type	Length
<code>bthdate</code>	Birth date	Numeric	8
<code>bthwght</code>	Birth weight (grams)	Numeric	8
<code>bwgrp</code>	Birth weight group	Character	2
<code>county</code>	County of occurrence of birth	Character	2
<code>csr</code>	C-section indicator	Character	1
<code>death</code>	Death indicator	Character	1
<code>dobd</code>	Day of birth	Numeric	3
<code>dobm</code>	Month of birth	Numeric	3

doby	Year of birth	Numeric	3
Hisp	Hispanic origin	Character	1
Linked	Unique record identifier	Numeric	8
mathosp	California DHS hospital ID	Character	4
mbthdate	Birth date of mother	Numeric	8
Mdobd	Day of birth of mother	Numeric	3
mdobm	Month of birth of mother	Numeric	3
mdoby	Year of birth of mother	Numeric	3
paymso	Payer source	Character	2
race	Race	Character	1
sex	Sex of child	Character	1
twin	Multiple birth indicator	Character	1
zip	ZIP	Character	5

Step 8: Run `getdat`: Prepare Hospital Level Record Comparison

Macro **%getdat** stored in *macros for probabilistic record linkage preparations.sas* is used to obtain for each OSHPD and/or DHS hospital a frequency count of

- Single live births and multiple live births
- Fetal deaths for singletons and multiples
- Number of re-admissions

The output generated by the macro **%getdat** is used by the macro **%gettable** to tabulate these numbers for each hospital and print a summary for all hospitals.

In order to function the macro **%getdat** requires the global macro variable `vstype` to be set to VS, VSBVSD, or BCF.

Step 9: Run `gettable`: Generate Hospital Level Comparison of Record Numbers to Identify Hospital Mergers, etc.

The hospital level comparison of record numbers is generated by the macro **%gettable** stored in *macros for probabilistic record linkage preparations.sas*. The goal is to identify:

1. Hospitals that came into existence in the year of the linkage or started to report to OSHPD. These hospitals have to be added to the crosswalk file `auxdata.namesXXXX` where XXXX corresponds to the year for which the crosswalk file is valid.
2. Hospitals that merged or changed their ID number because of some other event. For these hospitals the crosswalk `auxdata.namesXXXX` where XXXX corresponds to the year for which the crosswalk file is valid has to be updated.
3. To identify two or more OSHPD/DHS hospitals that report together even though the hospitals are in physically different locations. Most of the time, the DHS/OSHPD maternity hospital identifier will be different for each of the hospitals involved in such reporting under the same ID number.

The hospitals can be identified by carefully studying the printed output generated by this macro. An excerpt of the printout is shown below.

```

INFORMATION ON LINKAGE ELIGIBLE RECORDS FROM INFANT AND MATERNAL
HOSPITAL DISCHARGE RECORD AS WELL AS VITAL STATISTICS RECORD

cnty=

mathosp  HOSPID  name                BCFZIP  bcfelig  hdfelig  mhdfelig  bcfw  hdfw  mhdfw  mhdfidh  hdfrehsp

```

0000	1991	0	0	25	0	0	0	0
0998	495	0	0	9	0	0	0	0
0999	1718	0	0	37	0	0	0	0
9999	188	0	0	15	0	0	0	0

cnty=Alameda

mathosp	HOSPID	name	BCFZIP	bcfelig	hdfelig	mhdfelig	bcftw	hdftw	mhdfw	mhdfdth	hdfrehsp
0001	010735	ALAMEDA HOSPITAL	94501	477	478	478	8	8	4	2	4
0003	010739	ALTA BATES MEDICAL CENTER	94705	4750	4798	4770	179	165	87	30	149
0005	010776	CHILDREN'S HOSPITAL	94609	0	1	0	0	0	0	0	3338
0007	013619	SAN LEANDRO HOSPITAL	94578	1	0	0	0	0	0	0	0
0008	010805	EDEN HOSPITAL	94546	1116	1121	1120	20	20	10	5	15
0013	010846	HIGHLAND GENERAL HOSPITAL	94602	1062	1084	1071	25	25	13	21	22
0014	010858	KAISER HOSPITAL: HAYWARD	94545	3069	3096	3079	80	69	39	22	417
0015	010856	KAISER HOSPITAL: OAKLAND	94611	1876	1888	1875	51	54	26	22	282
0016	010937	SUMMIT MEDICAL CENTER - HAWTHOR	94609	3092	3086	3096	80	79	41	32	62
0022	010967	ST. ROSE HOSPITAL	94540	1522	1551	1542	38	31	19	15	13
0024	010983	VALLEY CARE MEDICAL CENTER	94588	1318	1330	1321	42	39	21	2	101
0026	010987	WASHINGTON TOWNSHIP HOSPITAL	94538	2319	2337	2326	61	55	28	10	171

cnty=Amador

mathosp	HOSPID	name	BCFZIP	bcfelig	hdfelig	mhdfelig	bcftw	hdftw	mhdfw	mhdfdth	hdfrehsp
0028	030786	SUTTER AMADOR HOSPITAL	95642	217	217	216	2	2	1	0	14

The printout shows separate sections for each California county. For the first four entries, the county is missing. These entries refer to births at home ('0000'), other not-in-hospital births ('0998'), hospitals with no maternity hospital code or new hospital for this year ('0999') and unknown place of birth or out-of-state birth ('9999').

The tabulation shows the following variables:

Mathosp	California DHS hospital identifier
Hospid	OSHPD hospital identifier
Name	Hospital name
Bcfzip	5-digit ZIP code of hospital
Bcfelig	Number of singleton live birth records eligible for linkage that were found in the vital statistics birth file
Hdfelig	Number of singleton live birth records eligible for linkage that were found in the infant hospital discharge file
Mhdfelig	Number of singleton live birth records eligible for linkage that were found in the maternal hospital discharge file
Bcftw	Number of multiple live birth records eligible for linkage that were found in the vital statistics file
Hdftw	Number of multiple live birth records eligible for linkage that were found in the infant hospital discharge file
Mhdfw	Number of multiple live birth records eligible for linkage that were found in the maternal hospital discharge file
Mhdfdth	Number of fetal deaths recorded in the maternal discharge data
Hdfrhsp	Number of infant re-admissions that occurred at this hospital
Fdeath	Number of fetal deaths recorded in the birth cohort file

The variable `fdeath` is only included if the listing pertains to a birth cohort file.

The number of multiple live births records found in the maternal discharge data needs to be at least doubled in order to be comparable to the number of multiple live births records found in the vital statistics birth file or infant discharge file since one delivery record will need to be linked to more two or more birth records.

Note that the macro **%gettable** prints an overview that includes all the hospitals, however, it also prints an overview that only pertains to those hospitals for which the difference between the number of records eligible from the vital statistics birth file is less than or more than 10% of the number of records eligible from the infant discharge data and/or maternal discharge data. An example for this portion of the printout is shown below. Note that some DHS hospitals will never have a matching OSHPD hospital: Military hospitals, birthing centers, and other hospitals that do not report to OSHPD.

cnty=Kings

diffi	difffm	mathosp	HOSPID	name	BCFZIP	bcfelig	hdfelig	mhdfeleg	bcftw	hdftw	mhdfw	mhdfddth	hdfrehsp
0.000	100.000	0108	160702	CORCORAN DISTRICT HOSPITAL	93212	1	1	2	0	0	0	0	29
100.000	100.000	0112		NAVAL HOSPITAL - LEMOORE	93246	309	0	0	10	0	0	0	0

cnty=Lassen

diffi	difffm	mathosp	HOSPID	name	BCFZIP	bcfelig	hdfelig	mhdfeleg	bcftw	hdftw	mhdfw	mhdfddth	hdfrehsp
1.361	21.769	0119	180919	LASSEN COMMUNITY HOSPITAL	96130	294	290	230	4	4	1	0	15

cnty=Los Angeles

diffi	difffm	mathosp	HOSPID	name	BCFZIP	bcfelig	hdfelig	mhdfeleg	bcftw	hdftw	mhdfw	mhdfddth	hdfrehsp
33.333	33.333	0131	190045	AVALON MUNICIPAL HOSPITAL	90704	3	2	2	0	0	0	0	0
100.000	100.000	0294		SUBURBAN HOSPITAL	90280	1	0	0	0	0	0	0	0
100.000	100.000	0667		NATURAL CHILDBIRTH INSTITUTE/	90230	51	0	0	0	0	0	0	0
100.000	100.000	0726		DOCTORS OFFICE	90039	19	0	0	0	0	0	0	0
100.000	100.000	0731		BEVERLY HEALTH & BIRTHING CENTE	90004	2	0	0	0	0	0	0	0
100.000	100.000	0773		NATURAL CHOICE BIRTH CENTER	91105	65	0	0	0	0	0	0	0

If a new pair of hospital identifiers is identified it can be added to the crosswalk file using **Step 10: Update Hospital Crosswalk File with Results of Previous Step**.

However, as the crosswalk file only allows an injective connection between the two sets of identifiers, OSHPD or DHS hospitals reporting under the same identifier in one database, but under two or more identifiers in the other database need to be added in the code for the following macros directly:

1. **%gettable**
2. **%mtchpddI**
3. **%mtchpddM**
4. **%preps**
5. **%hspsmmry** stored in *macros for summarizing all linkages in one file*.

The changes are obvious when the macro code is looked at in detail.

Care must be taken to always code those hospitals that have several different identifiers to the same master identifier that is used in the crosswalk file!

Step 10: Update Hospital Crosswalk File with Results of Previous Step

Write this section when I actually have to do it.

Step 11: Run Final Preparation Step to Add DHS Hospital ID to Patient Discharge Databases

Clicking on the pushbutton for this step leads to three choices: To update all files, to update only the discharge data, to update only the vital statistics data.

Using the crosswalk file `auxdata.namesXXXX` where `XXXX` corresponds to the year for which the crosswalk is valid, the macros **%mtchpddI** and **%mtchpddM** add a variable `mathosp` to the files `sastmp.subhdfi` and `sastmp.subhdfm`, i.e., they update the discharge data. The variable `hspmtch` is set to `Y` for all records for which the hospital was matched; if the hospital was not matched, the variable `hspmtch` is set to `N`.

The macro **%prepvs** finishes the preparations for the vital statistics birth data. For all DHS hospital identifiers for which no matching OSHPD identifier was found, the DHS identifier is set to 'ZZZZ.' Note that – as with all data edits undertaken for the purpose of the data linkage – this edit will only be used for the data linkage, the final linked product will, of course, have the original data value.

Step 12: Generating Value-Specific Frequency Information and the Set of u-Probabilities

All macros used in this section are stored in the file *macros for probabilistic record linkage preparations.sas*. Note that this section is very similar to [Step 5: Generating Value-Specific Frequency Information and the Set of u-Probabilities in Section 6 above](#).

The generation of the value-specific frequency information is driven by three macros: **%freqsB**, **%reduceB**, and **%vsfreqsB**.

The purpose of the **%freqsB** macro is to generate for each variable to be used in the linkage procedure a table with each possible value of the variable and its percent frequency. These tables are used to generate the value-specific frequency weights for each variable for the probabilistic linkage. The macro has three parameters: the name of the variable to be analyzed, whether or not missing values should be considered a valid category for this variable, and the variable type (numeric or character). The macro writes an output data set in the `sastmp` directory that is named as the variable with a `B` appended. For instance, for the variable `sex`, an output data set by the name of `sastmp.sexB` would be created. The output data set includes three variables: the variable value, the probability of occurrence of this value in the first data set (birth data), `probB`, the probability of occurrence of this value in the second and third data set (infant/maternal discharge data), `probI/probM`.

Besides this file, the macro appends an observation to the file `sastmp.genfreqsB`. The file `sastmp.genfreqsB` consists of a description of the variable, `des`, which is just the variable name and the general frequencies (uncorrected sums of squares (USS)) for the variable on the birth, infant discharge, and maternal discharge file, `genfreqB`, `genfreqI` and `genfreqM`.

For some variables, the same probability value occurs for more than one variable value. The macro **%reduceB** sorts the values of a variable in descending order of frequency, determines how often a probability level occurs, and groups values with the same probability level together.

Another issue is that for some variables, a large number of values occur rarely. In this situation, it is inefficient to use the complete table of value-specific frequencies. Rather we have chosen a threshold probability, usually 0.0001, below which the value-specific probability is set to the same value. The macro **%vsfreqsB** writes an ASCII file that assigns for each variable value the value-specific probability level using IF ... ELSE ... constructs. The ASCII files are named using the variable name, appending an 'I' or 'M' corresponding to the linkage for which they apply, and adding the extension .sas. Note that the macro describes probabilities for the variable values from the birth file with the variable `prob`, the probabilities for the variable values from the infant or maternal discharge file are described with `prob_`.

The generation of the set of u-probabilities is based on the construction of a file of unlinkable pairs. The macro **%uprobs** is used to create this file. Randomly records are merged from the birth, infant discharge, and maternal discharge file. For the infant discharge data, any records that match on hospital, date of birth, and gender are eliminated to make sure that the file really only includes records that are not linkable. For the maternal discharge data, any records that match on hospital, date of birth and maternal date of birth are eliminated. The resulting files `uprobI` and `uprobM` are evaluated for agreement on variable values for all the variables that are used in the linkage. For the birth-discharge data linkage, the screen SCL of `gui.vspdd.main.frame` carries out this step. The macro **%uprob** is used to calculate the probability of agreement on a variable value for each variable. The result of these evaluations is stored in the file `sastmp.uprobI` and `sastmp.uprobM`. These files consist of the name of the variable and its u-probability, i.e., the estimated probability of agreement based on the files `uprobI` and `uprobM`. The calls to the **%uprob** macros happen via the SCL of `gui.vspdd.main.frame`.

Step 13: Setting of m- and u-Probabilities for Linkage Run

Prior to running a probabilistic linkage step, it is necessary to set the m- and u-probabilities for each variable. These probabilities are set in the form of global macro variables. For instance, for the variable `sex`, global macro variables named `msex` and `usex` are created to represent this variable's m- and u-probability respectively. The screen `gui.vspdd.probs.frame` is used to set the m- and u-probabilities. The screen capture in Figure 13 shows the screen, as it appears when the screen is first entered.

Variable	m-probability	u-probability
mathosp	0.9999	0.004935884677
zip	0.95	0.001398879758
doby	0.9999	0.978340109749
dobd	0.9999	0.031941087828
dobm	0.9999	0.080962535422
race	0.9	0.557796669604
sex	0.9999	0.488813700681
twin	0.9999	0.926917062352
csr	0.9999	0.653426591983
bwgrp	0.9999	0.848757972951
paymso	0.9	0.339948631922
hisp	0.9	0.477452067517
death	0.9999	0.965721759404
...		
...		

Figure 13: VSPDD Linkage, Setting M- and U-Probabilities for Linkage

The m-probabilities are shown in the left column, the u-probabilities in the right column. The radio box in the top left corner needs to be checked to reference the linkage to be carried out, newborns or mothers. Only after a box is checked, the fields on the screen will populate. The box in the top right corner indicates which version of m-probabilities is shown. As we have not completed any iterations through the linkage procedure, the current iteration is zero or empty. The u-probabilities are derived from the [Step 12: Generating Value-Specific Frequency Information and the Set of u-Probabilities](#). The m-probabilities are set to initial values. These initial values can be changed using this screen if desired. The 'Set' button in the right bottom corner is used to actually set the m- and u-probability for each variable. The 'End' button below it is used to exit the screen.

Note that this screen also pops up after [Step 19: Recalculate m-Probabilities and Check Convergence](#). The m-probabilities will then be updated to reflect the estimate from the linkage run most recently completed.

Note that by changing the iteration number, a different starting set of m-probabilities can be loaded. For instance to use the last set of m-probabilities from the 1997 vital statistics birth-infant discharge record linkage as starting values for the 1996 linkage, copy the file `sastmp.mprobIX` (where X corresponds to the last completed linkage run) for 1997 in the `sastmp` directory for 1996 (they might be the same), and enter the number X as the iteration number on this screen.

Step 14: Adding Unlinked Infant Discharge Records to the Vital Statistics Birth Subfile (Only Needed for Maternal Linkage!)

The macro `%extvs2` adds unlinked infant discharge records to the pool of vital statistics birth records that the maternal records are matched against. In order for this macro to function properly, the file `resultI` has to exist. Note that this file is created in [Step 18: Generating Results Files](#) after the infant linkage has converged. If the file `resultI` does not exist, the user is automatically routed to this step. Immediately after the creation of `resultI`, `sastmp.subvs` is updated. Should the user cancel out of the generation of `resultI`, `sastmp.subvs` is not updated.

For the unlinked infant discharge records a `linkid` that is used when running the maternal linkage has to be created. As the `linkid` from the patient discharge data is a character variable and as `linkid` from the vital statistics data is a numeric variable, we cannot simply substitute the former for the latter. The principle for constructing `linkids` for the patient discharge data `linkid` is the following:

Year of PD Data	CD input:	_N_, i.e., the sequence number corresponding to the way records are stored in the original PD data, formatted as eight characters with leading zeroes.
	'A' first CD	
	'B' second CD	
	'C' third CD	
	'D' fourth CD	

For instance, the first record read for the 1997 PD data would be assigned the `linkid` '1997A00000001.' In order to obtain an all numeric `linkid` for the unlinked infant discharge records, we replaced 'A' by '6', 'B' by '7', 'C' by '8', and 'D' by '9', and converted the result to numeric. In the example the `linkid` 1997600000001 would be used.

Step 15: Linkage of Newborn Discharge and Vital Statistics Birth Data

Clicking on 'Linkage of newborn discharge and vital statistics birth data' (Figure 10) leads to the box familiar from the linkage to the death data asking the user whether any detail information on the individual variables' contributions to the overall match weight should be retained in the data set of matches (Figure 8). Note that retaining this information will lead to higher storage requirements as for each variable two types of match weights are retained. It will also lead to slower run-times of the linkage procedure as the data sets used are larger in size. However, the information can be useful in understanding the composition of the overall general and value-specific frequency weight.

It is important to keep in mind that the application does not at this point prompt for a threshold for the value-specific frequency weight. Rather, thresholds/cutoffs for the value-specific frequencies are stored in the parameter settings as part of [Step 1: Setting Linkage Parameters and Reviewing Linkage Status](#). Matches with a value-specific frequency weight below the acceptable threshold minimum are eliminated before the matches are evaluated. Matches are evaluated by checking the match data sets for ties (e.g., two or more birth records link to the same newborn discharge record; two or more newborn discharge records link to the same birth record, or both). Some ties can be resolved by retaining only the record with the higher match weight. Some ties can be randomized, especially those for which the two birth records and/or the two newborn discharge records pertain to multiples.

Two macros stored in *macros for probabilistic record linkages.sas* carry out the linkage tasks. The first macro is **%linkmac**. It carries out the linkage tasks. The second macro is **%mrkmtchs**. It resolves ties.

The macro **%linkmac** has the following parameters:

Input1	The first data set that includes information on infant or maternal discharges. Note that in the course of the macro all match and block variables are renamed to include an '_' as the last character of their name.
Input2	The second data set that includes information on births. The name of all block and match variables remains unchanged.
Lnkd	The output file of matched pairs.
Path	The path where the files that were created by the %vsfreqs macro discussed in Step 12: Generating Value-Specific Frequency Information and the Set of u-Probabilities . Usually, these files are stored in the same library that the SAS libname <code>sastmp</code> refers to.
Compcrit	A critical value for all simple agreements: The macro forms all possible matched pairs within a block. As this can result in a large number of matches, as a first cut the macro evaluates all simple agreements within a block, in other words, a variable <code>comp</code> is created that counts the number of times the linkage variables agree. If <code>compcrit</code> is set to a value larger than zero, all matched pairs with fewer than or equal to <code>compcrit</code> agreements are removed from the file of matched pairs prior to the calculation of general and value-specific frequency weights.
Crit	A critical value for the general frequency weight. In contrast to the value-specific frequency weight, the general frequency weight is based on m- and u-probabilities only. In case of agreement on a variable, the general frequency weight for the variable is the ratio of m- and u-probability; in case of disagreement on a variable, the general frequency weight is the ratio of $(1 - m\text{-probability})$ and $(1 - u\text{-probability})$. Note that in case of a variable with a high m-probability and low u-probability (the most desirable situation for a match variable), this leads to a large positive number in case of agreement and number less than 1 in case of disagreement. As logs are taken prior to summing the frequency weights for all variables, agreement results in a large positive weight; disagreement results in a large negative weight. Note that the general frequency weight is the <u>same</u> for different values of the variable. (See also Section 3) If <code>crit</code> is larger than 0, its value is used to further exclude matched pairs with general frequency weights below or equal to <code>crit</code> value.
Block	Indicates the block number, in case of the infant linkage, this can only be 1. In case of the maternal linkage, this can be a number as high as 10.
Srtvr1	The first numeric block variable.
Srtvr2	The second numeric block variable.
Srtvr3	The third numeric block variable.
Srtvr4	The fourth numeric block variable.
Srtvr5	The fifth numeric block variable.
Csrtvr1	The first character block variable.
Csrtvr2	The second character block variable.

Csrtvr3	The third character block variable.
Csrtvr4	The fourth character block variable.
Csrtvr5	The fifth character block variable.
Lastvar	The name of the last block variable. If <u>any</u> numeric block variables are present, this is the last numeric block variable. If <u>only</u> character block variables are present, this is the last character block variable.
Special	Available parameter, not yet used. Can be used to force specific conditions on the file of matched pairs.
Type	Version of linkage: D for deaths, T for transfers, R for re-admissions, P for prenatal/postpartum records, I for infant PDD, M for maternal PDD, A for add-on.
Develop	Indicator that is Y or N . If Y , variable-specific contributions to the value-specific frequency weight are included in the file of matched pairs. If N , these weights will not be included.

Furthermore the macro needs the following global parameters to be set:

- All m- and u- probabilities for the match variables.
- The year of the linkage.
- The vital statistics input file type (VS, BCF, VSBVSD).

For the infant linkage, only one linkage block is used. This linkage block is based on day of birth, month of birth, and hospital of birth.

The macro **%mrkmtchs** executes immediately after the **%linkmac** macro. The macro has three parameters, the version of linkage (**I** for infant discharge; **M** for maternal discharge file), the current block number, and the vital statistics input file type (VS, BCF, VSBVSD). It resolves any ties in the matched file by declaring this matched pair the final match that has the largest value-specific frequency weight. If there is a tie for the value-specific frequency weight, the match is accomplished by randomization.

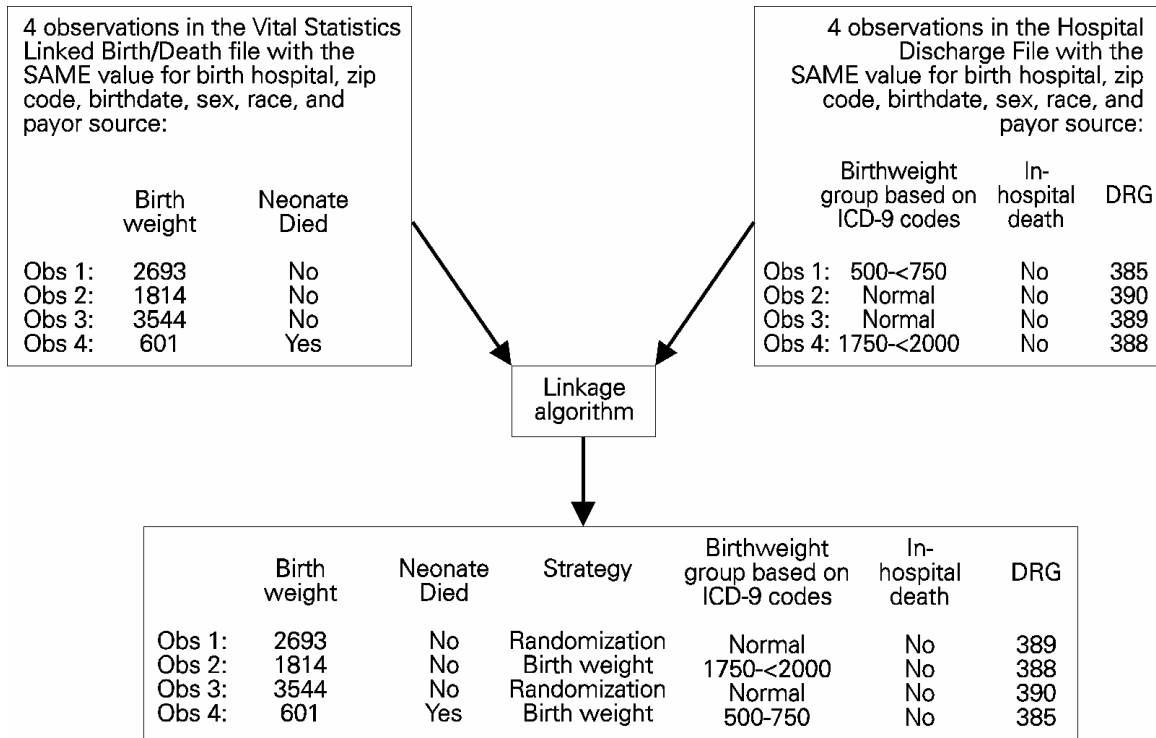


Figure 14: Example for Randomized Matching

Randomized matching was explored in detail in Herrchen, Gould, Nesbitt "Vital Statistics Linked Birth/Infant Death and Hospital Discharge Record Linkage for Epidemiological Studies," *Computers and Biomedical Research*, 30, 290-305, 1997. For the linkage of birth data to infant discharge data, it is of importance to link ill babies - specifically low birth weight babies or infants who passed away during the hospital stay - well. Figure 14 shows an example how our linkage procedure will use information available through the ICD-9-CM code and discharge status for this group of infants. For two of the four records additional birth weight information is available and used to accomplish the correct match. For the other two records, the match is randomized. The vast majority of records randomized pertain to healthy babies for whom the information in the discharge data that is linked to the vital statistics birth file is similar: no complicating diagnoses, length of stay of between 0 and 2 days for vaginal deliveries and 0 and 4 days for cesarean sections, similar charges. False positive matches with different information on for instance, diagnoses, length of stay, charges, will not make the linked data useless for epidemiological, population-based studies. However, a conservative bias will be introduced in findings that are based on analyses including such records.

Of course, one might ask why randomized matching should be done at all; why not leave out tied records? The problem is that the bias that would be introduced by leaving out tied records would compromise the quality of many epidemiological studies, the reason being that most of the randomized matching happens for the 10 hospitals with the largest numbers of births in California. Results for these hospitals and the regions of California they are in would be affected by leaving out tied matches.

Step 16: Linkage of Maternal Discharge and Vital Statistics Birth Data

Prior to running the maternal linkage, if any unlinked infant discharge records from the vital statistics birth/infant discharge linkage are to be included as possible matches, those need to be added to the file `sastmp.subvs` using the `%extvs2` macro (see [Step 14: Adding Unlinked Infant Discharge Records to the Vital Statistics Birth Subfile \(Only Needed for Maternal Linkage!\)](#) above).

Clicking on 'Linkage of maternal discharge and vital statistics birth data' (Figure 10) leads to a box asking the user whether any detail information on the individual variables' contributions to the overall match weight should be retained in the data set of matches. Note that retaining this information will lead to higher storage requirements as for each variable two types of match weights are retained. Speed of data processing is also affected. However, the information can be useful in understanding the composition of the overall general and value-specific frequency weight.

For the maternal linkage, the user of the linkage GUI is furthermore prompted for the linkage block that should be run.

A problem of linking multiple births to maternal records is that the match is no longer a bijective match. Rather, it is possible that multiple births records match to the same maternal record. In order to account for this situation, the program flow is changed. The following chart clarifies the program flow for the maternal linkage:

1. Run block 1 of vital statistics birth/maternal hospital discharge record linkage.
↓
2. Write all matches to the file of "vital statistics birth matches." Use all remaining vital statistics birth records for further matching for block 1.
↓
3. Write all matches involving singleton birth to the file of "maternal matches." Write all matches involving a twin birth in either the vital statistics birth file or maternal discharge file to the file of "maternal matches" and "maternal repeats." Use the file of "maternal repeats" for further matching in block 1.
↓
4. Run block 1 of vital statistics birth/maternal hospital discharge record linkage with "maternal repeats" and unmatched vital statistics birth records.
↓
5. Add all matches to the file of "vital statistics birth matches." Use all remaining vital statistics birth records for further matching for this block.
↓
6. Add all matches involving a twin birth in either the vital statistics birth file or maternal discharge file to the file of "maternal matches". Replace the "maternal repeats" with matches obtained in this step. Note that this number is going to be smaller compared to the previous step. At this point, twins are linked while triplets and higher order multiples are still left over.
↓

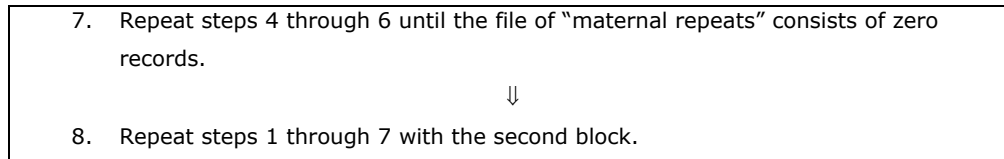


Figure 15: VS-PDD Linkage, Flow of Control for Linkage of Vital Statistics Birth Data to Maternal Discharge Data

The pop menu in Figure 16 generates this program flow.



Figure 16: VS-PDDM, Pop Menu that Generates Program Flow for Linkage of Vital Statistics and Maternal Delivery Discharge Records

For the maternal linkage, the first block is based on mother’s birth date and hospital of birth; the second block is based on estimated delivery day and month and hospital of birth.

The macro **%linkmac** is used to produce linked records. The macro **%mrkmtchs** is used to resolve ties. See [Step 15: Linkage of Newborn Discharge and Vital Statistics Birth Data](#) for a detailed description of both.

Step 17: Clerical Review of Match Results

Figure 17 shows a screen capture with the options available to select records to be clerically reviewed. Note that it makes sense to do a clerical review of the infant linkage immediately after the infant linkage was run, and prior to running the maternal linkage.

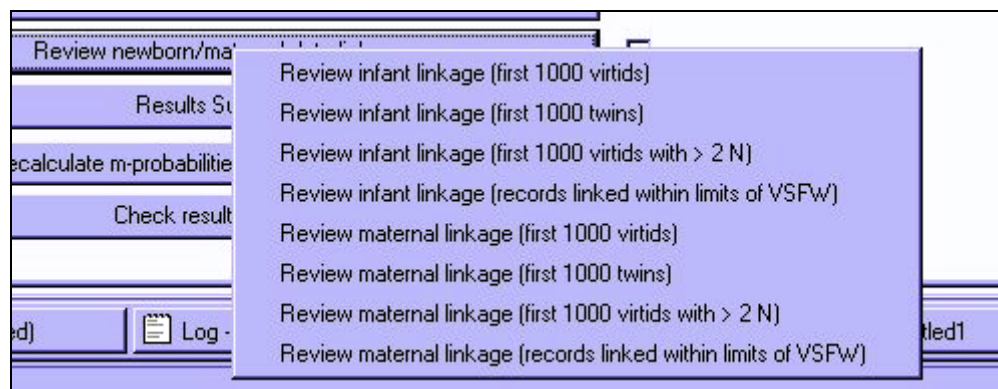


Figure 17: VS-PDD, Clerical Review Options Pop Menu

Figure 18 displays the clerical review screen. This screen is primarily used to establish match cutoffs. In some circumstances, specific linkids are marked for removal from the matched file or marked for

inclusion even though the match weight is below what is considered the minimum match weight. These circumstances are rare, as just the size of the linkage tasks at hand does not allow time for an elaborate manual clerical review.

The screen in Figure 18 demonstrates the second option 'Review infant linkage (first 1000 twins).' The virtual ID or block within which matches are searched is defined by the hospital, birth date, and gender. There were six possible matches in this block. The two multiples are correctly matched, as the first was delivered vaginally, while the second twin was delivered via cesarean section. The remaining four matches consist of two records in the discharge data matched to two records in the vital statistics birth date. The probabilistic match weight identifies the two of the four possible matches that are correct. Note that Figure 17 also displays the clerical review screen when detailed information on each variable's contribution to the match weight is retained: For instance, *w2_paymso* is the contribution – prior to taking logs – of the payer source variable to the overall match weight.

Virtual ID: 44 Hospital Name: 0001 Block size: 6 Birth Date: [redacted] Gender: 2

Vital Statistics Records:

	ID	C/S	Twin	Race	Payer	Hispanic	Death	ZIP	BW	BWGT	Status
1	50517	0	Y	2	08	1	N	94619	11	2605	Y
2	50518	1	Y	2	08	1	N	94619	08	2479	Y
3	50432	0	N	1	05	1	N	94501	11	3756	Y
4	50434	0	N	4	08	2	N	94502	11	3964	Y
5	50434	0	N	4	08	2	N	94502	11	3964	N
6	50432	0	N	1	05	1	N	94501	11	3756	N

OSHPD Discharge Records:

	ID	C/S	Twin	Race	ZIP	BW	DRG	Status	Payer	Hispanic	Death
1	1997A00000835	0	Y	2	94619	11	391	Y	08	2	N
2	1997A00000836	1	Y	2	94619	11	391	Y	08	2	N
3	1997A00000819	0	N	1	94501	11	391	Y	05	2	N
4	1997A00000818	0	N	1	94501	11	391	Y	08	2	N
5	1997A00000819	0	N	1	94501	11	391	N	05	2	N
6	1997A00000818	0	N	1	94501	11	391	N	08	2	N

Linkage information:

	w2_csr	w2_twin	w2_race	w2_hisp	w2_paymso	zipjarol	w2_zip	w2_bwgrp	vsfw	keep
1	1.60	1600.08	193.30	0.19	5.75	1.000	2138633.33	1.15	79.94	Y
2	22.78	1600.08	193.30	0.19	5.75	1.000	2138633.33	0.00	72.89	Y
3	1.60	1.05	1.62	0.19	5154.39	1.000	472320.06	1.15	70.10	Y
4	1.60	1.05	0.23	3.19	5.75	0.800	302284.84	1.15	60.76	Y
5	1.60	1.05	0.23	3.19	0.15	0.800	302284.84	1.15	41.24	N
6	1.60	1.05	1.62	0.19	0.15	1.000	472320.06	1.15	40.67	N

Navigation buttons: [Previous], [Next], [First], [Last], [End]

Figure 18: VS-PDD, Clerical Review Screen

For the fourth and eighth option on the pop menu in Figure 18, the following screen prompts the user to enter a minimum and maximum linkage weight. The clerical review screens will then show all virtual Ids for which at least one match occurred with a linkage weight within the given boundaries.

Figure 19: VSPDD Linkage, Setting Boundaries for the Value-Specific Frequency Weight within which a Clerical Review is Carried Out

Step 18: Generating Results Files

Figure 20 displays three options for generating results files: 1) should a results be generated for the infant linkage? 2) Should a results file be generated for the maternal linkage? 3) Should a results file be generated for both linkages?

Figure 20: VS-PDD, Pop Menu for the Generation of Results Files

For the calculation of the m-probabilities of the infant linkage, the results file for the infant linkage, `resultI`, has to be generated. For the calculation of the m-probabilities of the maternal linkage, the results file for the maternal linkage, `resultM`, has to be generated. For the creation of the results file summarizing both linkages, the results file for the infant and the maternal linkage need **not** be generated. If this option is chosen, the results files for the infant and maternal linkages will be generated prior to generating the linked file consisting of both linkages.

The macro **%resultI** generates the results file for the infant linkage. The macro **%resultM** generates the results file for the maternal linkage. Both macros require three parameters: 1) the threshold of the value-specific frequency weight; 2) the number of linkage blocks (note multiple runs of the same linkage block as displayed in Figure 15 are counted multiple times!); 3) the year of the linkage.

In addition, the macro **%resultI** requires a fourth parameter, the vital statistics input file. This parameter is necessary as the construction of the result file is based on `sastmp.subvs`. However, if the unlinked infant records are added to `sastmp.subvs` in order to perform the maternal linkage including those records as well - see [Step 14: Adding Unlinked Infant Discharge Records to the Vital Statistics Birth Subfile \(Only Needed for Maternal Linkage!\)](#), this version of the file would lead to an incorrect summary file. In this situation, the GUI will use the file `sastmp.subvs2` **only if the checkbox for [Step 14: Adding Unlinked](#)**

Infant Discharge Records to the Vital Statistics Birth Subfile (Only Needed for Maternal Linkage!) is checked!

The user is prompted for the threshold to use via a screen that looks like Figure 21. The default entry on the screen corresponds to the one entered in [Step 1: Setting Linkage Parameters and Reviewing Linkage Status](#).



Figure 21: VS-PDD, Screen for Entry of Match Cutoff for Maternal Linkage

The macro **%resultI** creates the file `resultI`. The file `resultI` includes all variables of relevance that were added by the linkage procedure as well as all the variables used in the linkage from the vital statistics and infant discharge data. The following variables are added by the linkage procedure:

Block	Block in which this match occurred (should always be '1').
Keep	Flag that indicates whether or not this record was kept after imposing the linkage threshold and resolution of ties (should always be 'Y')
LinkedI	Indicator variable that is 'Y' if the current record has the vital statistics birth record linked to the infant discharge record. <code>LinkedI</code> is equal to 'I' if the current record is an infant discharge record that was NOT linked to a vital statistics record. <code>LinkedI</code> is equal to 'B' if the current record is a vital statistics birth record that was NOT linked to an infant discharge record.
Nomtchs	Total number of matches that occurred in the block within which the current record was linked. Note that each unique block is described by the added variable <code>virtid1</code> .
Tie	Indicates that this record was the result of resolving a tie. The tie might have been resolved since this record had the highest linkage weight. The tie might have also been resolved by randomization. Note that virtually all multiple birth records have the tie flag set to 'Y'.
Virtid1	A unique number assigned to each different block imposed by the blocking variables.
vsfw	Value-specific frequency weight or match weight that indicates the confidence in this match. The larger <code>vsfw</code> , the better the match.

As the variable `linkid` only exists for the vital statistics birth record, the macro also generates a `linkid` for the unlinked infant discharge records (`linkedI EQ 'I'`). This variable is generated as described in [Step 14: Adding Unlinked Infant Discharge Records to the Vital Statistics Birth Subfile \(Only Needed for Maternal Linkage!\)](#).

The macro **%resultM** creates the output file `resultM`. The file `resultM` includes all variables of relevance that were added by the linkage procedure as well as all the variables used in the linkage from the vital statistics and infant discharge data. The following variables are added by the linkage procedure:

Block	Block in which this match occurred.
Keep	Flag that indicates whether or not this record was kept after imposing the linkage threshold and resolution of ties (should always be 'Y')
LinkedM	Indicator variable that is 'Y' if the current record has the vital statistics birth record linked to the maternal discharge record. LinkedM is equal to 'M' if the current record is a maternal discharge record that was NOT linked to a vital statistics record. LinkedM is equal to 'B' if the current record is a vital statistics birth record that was NOT linked to a maternal discharge record.
Nomtchs	Total number of matches that occurred in the block within which the current record was linked. Note that each unique block is described by the added variable <code>virtid1</code> .
Tie	This record was the result of resolving a tie. The tie might have been resolved since this record had the highest linkage weight. The tie might have also been resolved by randomization. Note that virtually all multiple birth records have the tie flag set to 'Y'.
Virtid1	A unique number assigned to each different block imposed by the blocking variables.
vsfw	Value-specific frequency weight or match weight that indicates the confidence in this match. The larger <code>vsfw</code> , the better the match.

The macro **%resultB** summarizes both linkages and creates the output file `sastmp.vsbvsdIM`. The variables included in this file are shown below.

Table 6: VS-PDD Linkage, Contents of Final Results File `sastmp.vsbvsdIM`

Variable Name	Description	Variable Type	Variable Length
DRG	Diagnosis Related Group	Character	3
BlockI	Block of linkage (PDDI)	Character	1
BlockM	Block of linkage (PDDM)	Character	1
bthwght	Birth weight (grams)	Numeric	8
death	Death indicator (VS)	Character	1
death_	Death indicator (PDD)	Character	1
dobd	Day of birth	Numeric	3
dobm	Month of birth	Numeric	3
doby	Year of birth	Numeric	3
dobyI	Year of birth (PDDI)	Numeric	3
dobyM	Estimated year of birth (PDDM)	Numeric	3
fdeath_	Fetal death (PDDM)	Character	1
hisp	Hispanic ethnicity (VS)	Character	1
linkedB	Infant and maternal PDD linked to VS	Character	1
linkedI	Infant PDD linked to VS	Character	1
linkedM	Maternal PDD linked to VS	Character	1
Linkid	Unique record identifier (VSB, PDDI, VSD)	Numeric	8

linkidI	Unique record identifier (PDDI)	Character	13
linkidM	Unique record identifier (PDDM)	Character	13
mathosp	CA maternity hospital code	Character	4
mdobd	Mother's day of birth (VS)	Numeric	3
mdobdM	Mother's day of birth (PDDM)	Numeric	3
mdobm	Mother's month of birth (VS)	Numeric	3
mdobmM	Mother's month of birth (PDDM)	Numeric	3
mdoby	Mother's year of birth (VS)	Numeric	3
mdobyM	Mother's year of birth (PDDM)	Numeric	3
NomtchsI	Number of matches within block of this match (PDDI)	Numeric	8
NomtchsM	Number of matches within block of this match (PDDM)	Numeric	8
paymso	Payer source (VS)	Character	2
Race	Race (VS)	Character	1
tieI	Record was tied (PDDI)	Character	1
tieM	Record was tied (PDDM)	Character	1
twin	Multiple birth (VS)	Character	1
twinB	Multiple birth (VS) (used to check that resultI and resultM have the same twin indicator for the same VS record)	Character	1
twinI	Multiple birth (PDDI)	Character	1
twinM	Multiple birth (PDDM)	Character	1
VirtidI	Virtual ID number (VS for PDDI linkage)	Numeric	5
VirtidM	Virtual ID number (VS for PDDM linkage)	Numeric	5
vsfwI	Value-specific frequency weight (PDDI)	Numeric	8
vsfwM	Value-specific frequency weight (PDDM)	Numeric	8
zip	Zipcode of mother's residence	Character	5

Several variables are added to the original list of variables that were input to the linkage. These variables are highlighted. The variables `blockI`, `blockM`, `nomtchsI`, `nomtchsM`, `tieI`, `tieM`, `virtidI`, `virtidM`, `vsfwI`, and `vsfwM` are all renamed versions of the variables `block`, `nomtchs`, `tie`, `virtidI`, and `vsfw`. They are explained in the discussion of the generation of the `resultM` and `resultI` files. The variables `linkedI` and `linkedM` were also discussed at that point. The only variable added by the macro `%resultB` is `linkedB`. This variable is generated according to the following rules:⁷

⁷

```

IF linkedI EQ 'Y' AND linkedM EQ 'Y' THEN linkedB = 'Y';          /* linked to maternal and infant record */
ELSE IF linkedI EQ 'Y' AND linkedM EQ 'B' THEN linkedB = 'I';    /* VSB linked to infant record only */
ELSE IF linkedI EQ 'B' AND linkedM EQ 'Y' THEN linkedB = 'M';    /* VSB linked to maternal record only */
ELSE IF linkedI EQ ' ' AND linkedM EQ 'Y' THEN linkedB = 'N';    /* infant discharge records linked ONLY to
                                                                    maternal discharge record */
ELSE IF linkedI EQ ' ' AND linkedM IN ('M') THEN linkedB = 'C';  /* unlinked maternal discharge records */
ELSE IF linkedI IN (' ' 'B') AND linkedM EQ 'B' THEN linkedB = 'B'; /* unlinked VSB records */
ELSE IF linkedI EQ 'I' AND linkedM EQ ' ' THEN linkedB = 'A';    /* unlinked infant discharge records */

```

'A'	Unlinked infant record
'B'	Unlinked birth record
'C'	Unlinked maternal record
'N'	Infant and maternal record linked, not VSB
'I'	VSB and infant record linked, not maternal record
'M'	VSB and maternal record linked, not infant record
'Y'	VSB, infant, and maternal record linked

Step 19: Recalculate m-Probabilities and Check Convergence

After the infant or maternal linkage runs are completed, the m-probabilities have to be re-calculated to check convergence. The user is prompted to indicate whether to re-calculate m-probabilities for the maternal or infant linkage (see Figure 22). **Note that the files `resultI` and `resultM` have to be generated via Step 18: Generating Results Files** Step 18: Generating Results File **in order to obtain the updated set of m-probabilities for the infant and maternal linkage respectively.** If these results files are not generated, an error message will be issued and the system will not proceed with this step.

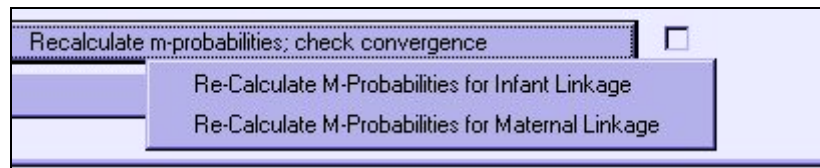


Figure 22: VS-PDD, Pop Menu for Recalculation of M-Probabilities

First a temporary file is created via the macro `%inputmprob` stored in *macros for probabilistic record linkage.sas*. This temporary file retains information for each match on whether a variable agreed or disagreed. The macro takes three parameters, the version of linkage currently run (e.g., `D` for deaths, `P` for prenatal/postpartum maternal records, `I` for infant discharge, `M` for maternal discharge), the year of the linkage, and the vital statistics input file type (VS, BCF, VSBVSD).

The macro `%getmprob` stored in *macros for probabilistic record linkage.sas* is then used to obtain an updated file of m-probabilities, `sastmp.mprobIX` and `sastmp.mprobMX`. Where `X` corresponds to the current linkage run iteration. If `x` is larger than 1, the macro also compares the current iteration's m-probabilities to the previous iteration and prints the result in the SAS output window. The result can be used to determine whether or not convergence has occurred. We considered a linkage run to have converged if all differences in m-probabilities were less than 0.01.

After the recalculation of the m-probabilities is completed, the GUI automatically calls the entry `gui.vspdd.probs.frame` as discussed in [Step 13: Setting of m- and u-Probabilities for Linkage Run](#). If convergence has not yet happened, it can be used to set the most current version of m-probabilities that will be used in the next iteration of the linkage procedure.

Note that at this point the iteration number always has to be entered in the upper right hand text entry box to obtain the most current set of m-probabilities.

Step 20: Generate Results Summary

The user is presented with three options to obtain a summary of the linkage results. These options are displayed in Figure 23.

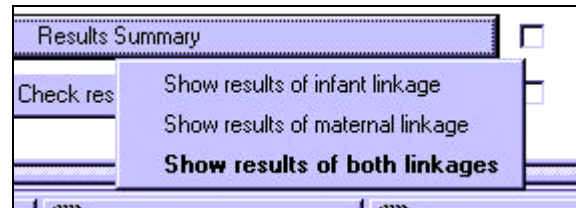


Figure 23: VS-PDD, Pop Menu for Generation of a Results Summary

The first item listed on the pop-menu leads to a summary of the linkage result (linked/unlinked records) for the infant linkage by plurality. The second item leads to a summary of the linkage results for the maternal linkage by plurality. The third item leads to a tabulation of the linkage status of each observation as well as a number of more detailed results of the linkage at the hospital level.

The results for the infant and maternal linkage, and the first portion of the results for both linkages are generated by the screen SCL of `gui.vspdd.main.frame` in the `pb_summary` section. The additional detailed results are generated by the macro `%rsldtl` stored in *macros for probabilistic record linkage.sas*

The tabulation of linkage status of each observation should be used to populate the worksheet 'VSPDD Summary' in the spreadsheet named `XXX linkage results.xls` where `XXXX` refers to the year of the linkage. The detailed results at the hospital level should be used to populate the worksheet 'VS' for the vital statistics birth file and 'BCF' for the birth cohort file in the same spreadsheet.

An excerpt of the 1997 linkage results is shown below in Table 7.

Table 7: Example for Linkage Results

Note that the vital statistics data include records in hospitals that do not report to OSHPD.

Multiple VS (Y/N)	Multiple PDDM (Y/N)	Multiple PDDI (Y/N)	linked to both records	linked to mom's record only	linked to infant record only	linked	unlinked VS record	unlinked mom's record	unlinked infant record
						infant and maternal record, not VSB			
No	No	No	493,158	-	-	-	-	-	-
No	No	Yes	85	-	-	-	-	-	-
No	Yes	No	162	-	-	-	-	-	-
No	Yes	Yes	34	-	-	-	-	-	-
Yes	No	No	41	-	-	-	-	-	-
Yes	No	Yes	192	-	-	-	-	-	-

Yes	Yes	No	690	-	-	-	-	-	-
Yes	Yes	Yes	12,028	-	-	-	-	-	-
No	No	NA	-	3,837	-	-	-	-	-
No	Yes	NA	-	4	-	-	-	-	-
Yes	No	NA	-	-	-	-	-	-	-
Yes	Yes	NA	-	73	-	-	-	-	-
No	NA	No	-	-	2,147	-	-	-	-
No	NA	Yes	-	-	-	-	-	-	-
Yes	NA	No	-	-	11	-	-	-	-
Yes	NA	Yes	-	-	112	-	-	-	-
NA	No	No	-	-	-	3,084	-	-	-
NA	No	Yes	-	-	-	11	-	-	-
NA	Yes	No	-	-	-	27	-	-	-
NA	Yes	Yes	-	-	-	77	-	-	-
No	NA	NA	-	-	-	-	12,590	-	-
Yes	NA	NA	-	-	-	-	291	-	-
NA	NA	NA	-	-	-	-	203	-	-
NA	No	NA	-	-	-	-	-	4,439	-
NA	Yes	NA	-	-	-	-	-	164	-
NA	NA	No	-	-	-	-	-	-	4,209
NA	NA	Yes	-	-	-	-	-	-	73
			506,390	3,914	2,270	3,199	13,084	4,603	4,282
Total Number of VS Records:						525,658			
Total Number of PDDM Records (double counting multiples):						518,106			
Total Number of PDDI Records:						516,141			
Percent VS records linked to both records:						96.33%			
Percent VS records linked to PDDM record:						97.08%			
Percent VS records linked to PDDI record:						96.77%			
Percent PDDI records linked to both records:						98.11%			
Percent PDDI records linked to VS record:						98.87%			
Percent PDDI records linked to PDDM record:						98.73%			
Percent PDDM records linked to both records:						97.74%			
Percent PDDM records linked to VS record:						98.49%			
Percent PDDM records linked to PDDI record:						98.36%			

Step 21: Bias Check

The bias check is driven by the macro **%bias** stored in *macros for probabilistic record linkage.sas*. It verifies the distribution of birth weight among linked and unlinked records.

8. Add-On Linkage of Unlinked Vital Statistics Records and Linked Infant/Maternal Discharge Records

As we have pointed out in the previous section under Step 14: Adding Unlinked Infant Discharge Records to the Vital Statistics Birth Subfile (Only Needed for Maternal Linkage!), we added unlinked infant discharge records to the file against which we searched for matches to the maternal discharge record. These matches are reflected in column 7 of Table 7. This column shows the number of matches of an infant and maternal discharge record, yet the infant discharge record had not been linked to a vital statistics record.

As adding the information in the maternal discharge record to the infant discharge record adds information on the mother's date of birth and fetal death to the file, this additional information can be used to attempt to link the subset of infant/maternal matches to the vital statistics birth file. This attempt is made in this add-on step (Figure 24).

PDD-VS Linkage - ADD-ON
Health Information Solutions, October 2000

Enter Linkage Parameters:	<input checked="" type="checkbox"/>	VS1997	Save
Run extsubdat: Obtain unlinked birth records and linked PDDI-PDDM records	<input checked="" type="checkbox"/>		
Value-specific frequencies and estimation of u-probabilities	<input checked="" type="checkbox"/>		
Setting of m- and u-probabilities for linkage run	<input type="checkbox"/>	Check Log	Check Output
Linkage of unlinked births to PDDI-PDDM matches	<input type="checkbox"/>	End	
Review matches	<input type="checkbox"/>		
Generate results file	<input type="checkbox"/>		
Recalculate m-probabilities; check convergence	<input type="checkbox"/>		
Update sastmp.vsbvsdLM with additional matches	<input type="checkbox"/>		

Figure 24: PDD-VS ADD-ON: Main Screen

As for the death linkage, the push buttons in the left portion of the screen guide through the linkage steps. As explained under the death linkage the current status and any match parameters can be saved and restored using the top right portion of the screen.

The actions executed with each push button are explained in detail below.

Step 1: Setting Linkage Parameters and Reviewing Linkage Status

The first push button 'Enter Linkage Parameters' (Figure 22) leads to a screen that displays parameters for the current linkage environment.

The parameter screen (Figure 25) has the same functionality as similar screens previously described. Note that the VSFW cutoffs refer to the cutoffs for the value-specific frequency weight for the first block and second block respectively.

The screenshot shows a dialog box titled "VSPDD ADD-ON Linkage, Linkage Parameter Screen". It contains the following fields and options:

- Path to linkage macros:**
- Path to formats:**
- Path to output data:**
- Year of linkage:**
- VS input file?** ☐ VS ☐ BCF ☒ VSBVSD
- Next year of PDD data available?** ☒ Yes ☐ No
- Additional weight information?** ☐ Yes ☒ No
- M-Probabilities Version:**
- Linkage Block:**
- VSFW cutoffs:**

At the bottom are three buttons: **Done**, **Cancel**, and **Reset**.

Figure 25: VSPDD ADD-ON Linkage, Linkage Parameter Screen

Step 2: Run *extsubdat*: Extract Unlinked Births and Infant/Maternal Record Matches

Macro **%extsubdat** stored in *macros for getting data ready for linkages.sas* is used to extract the unlinked births and infant/maternal record matches. The basis of the extraction is the file `sastmp.vsbvsdIM` that was created in Step 18: Generating Results Files, Section 7. As was explained in this section, the variable `linkedB` was used to keep track of the match status of each observations. **%extsubdat** utilizes all observations with match status 'B' for the unlinked births file; it utilizes all observations with match status 'N' for the infant/maternal matches. If the file `sastmp.vsbvsdIM` does not exist, the step does not generate any results, and a warning is issued for the user.

The macro **%extsubdat** extracts information from `sastmp.subvs` on each unlinked birth. It extracts information from `sastmp.subhdfI` and `sastmp.subhdfM` for each linked infant/maternal match. The macro creates the file `sastmp.subvsA` and `sastmp.subpddA`. The file `sastmp.subvsA` has the same variables as `sastmp.subvs` has. The file `sastmp.subpddA` has the same variables that `sastmp.subhdfI` has plus variables for maternal age (`mdobd`, `mdobm`, `mdoby`) and fetal death (`fdeath`) from `sastmp.subhdfM`. Furthermore, `sastmp.subpddA` includes the record identifier for the infant discharge record (`linkid`), and it includes the record identifier for the maternal discharge record (`linkidM`).

Step 3: Generating Value-Specific Frequency Information and the Set of u-Probabilities

All macros used in this section are stored in the file *macros for probabilistic record linkage preparations.sas*. Note that this section is very similar to [Step 5: Generating Value-Specific Frequency Information and the Set of u-Probabilities in Section 6 above](#).

The generation of the value-specific frequency information is driven by three macros: **%freqsA**, **%reduceA**, and **%vsfreqsA**.

The purpose of the **%freqsA** macro is to generate for each variable to be used in the linkage procedure a table with each possible value of the variable and its percent frequency. These tables will be used to generate the value-specific frequency weights for each variable for the probabilistic linkage. The macro has three parameters: the name of the variable to be analyzed, whether or not missing values should be considered a valid category for this variable, and the variable type (numeric or character). The macro writes an output data set in the `sastmp` directory that is named as the variable with a `P` appended. For instance, for the variable `sex`, an output data set by the name of `sastmp.sexA` would be created. The output data set includes three variables: the variable value, the probability of occurrence of this value in the first data set (unlinked births), `probX`, the probability of occurrence of this value in the second data set (infant/maternal matches), `probY`.

Besides this file, the macro appends an observation to the file `sastmp.genfreqsA`. The file `sastmp.genfreqsA` consists of a description of the variable, `des`, which is the variable name and the general frequency (uncorrected sums of squares (USS)) for the variable on the unlinked births file and the file of infant/maternal matches.

For some variables, the same probability value occurs for more than one variable value. The macro **%reduceA** sorts the values of a variable in descending order, determines how often a probability level occurs, and groups values with the same probability level together.

Another issue is that for some variables, a large number of values occur rarely. In this situation, it is inefficient to use the complete table of value-specific frequencies. Rather we have chosen a threshold probability, usually 0.01, below which the value-specific probability is set to the same value. The macro **%vsfreqsA** writes an ASCII file that assigns for each variable value the value-specific probability level using IF ... ELSE ... constructs. The ASCII files are named using the variable name, appending a 'A' and adding the extension `.sas`. Note that the macro describes the probabilities for the variable values from the unlinked births file with the variable `prob`, the probabilities for the variable values from the file of infant/maternal matches are described with `prob_`.

The generation of the set of u-probabilities is based on the construction of a file of unlinkable pairs. The macro **%probsA** is used to create this file. Randomly records are merged from the file of unlinked births records and from the file of infant/maternal matches. Any records that agree on birth date, maternal birth date, and gender are removed. The file of unlinkable pairs is called `ulpairs`.

The screen SCL of `gui.vspdd2.main.frame` takes care of evaluating the file of unlinkable pairs for the calculation of u-probabilities. After constructing the file `uprobA` with indicator variables for agreement and disagreement of the variables to be used in the linkage, the macro **%uprob** is used to generate the file of u-probabilities `sastmp.uprobA`.

Step 4: Setting of m- and u-Probabilities for Linkage Run

Prior to running a probabilistic linkage step, it is necessary to set the m- and u-probabilities for each variable. These probabilities are set in the form of global macro variables. For instance, for the variable `sex`, global macro variables name **msex** and **usex** are created to represent this variable's m- and u-probability respectively. The screen `gui.vspdd2.probs.frame` is used to set the m- and u-probabilities. The screen capture in Figure 26 shows the screen, as it appears when the screen is first entered.

The m-probabilities are shown in the left column, the u-probabilities in the right column. The box in the top right indicates which version of m-probabilities is shown. As we have not completed any iterations through the linkage procedure, the current iteration is zero. The u-probabilities are derived from the previous step described above. The m-probabilities are set to initial values. These initial values can be changed using this screen if desired. The 'Set' button in the right bottom corner is used to actually set the m- and u-probability for each variable. The 'End' button below it is used to exit the screen.

mathosp	<input type="text" value="0.9"/>	<input type="text" value="1E-8"/>	Iteration: <input type="text" value="0"/> <input type="button" value="Set"/> <input type="button" value="End"/>
zip	<input type="text" value="0.9"/>	<input type="text" value="0.000625195373"/>	
doby	<input type="text" value="0.99"/>	<input type="text" value="1"/>	
dobd	<input type="text" value="0.99"/>	<input type="text" value="0.029384182557"/>	
dobm	<input type="text" value="0.99"/>	<input type="text" value="0.079087214754"/>	
race	<input type="text" value="0.9"/>	<input type="text" value="0.534542044388"/>	
sex	<input type="text" value="0.99"/>	<input type="text" value="0.494529540481"/>	
twin	<input type="text" value="0.98"/>	<input type="text" value="0.933729290403"/>	
csr	<input type="text" value="0.98"/>	<input type="text" value="0.725539231009"/>	
bwgrp	<input type="text" value="0.95"/>	<input type="text" value="0.809628008752"/>	
mdoby	<input type="text" value="0.99"/>	<input type="text" value="0.042200687714"/>	
mdobd	<input type="text" value="0.99"/>	<input type="text" value="0.028758987183"/>	
mdobm	<input type="text" value="0.99"/>	<input type="text" value="0.082213191622"/>	
paymso	<input type="text" value="0.9"/>	<input type="text" value="0.120350109409"/>	
hisp	<input type="text" value="0.9"/>	<input type="text" value="0.521412941544"/>	
death	<input type="text" value="0.98"/>	<input type="text" value="0.961237886839"/>	
...	<input type="text"/>	<input type="text"/>	
...	<input type="text"/>	<input type="text"/>	
...	<input type="text"/>	<input type="text"/>	

Figure 26: ADD-ON Linkage, Setting of M- and U-Probabilities for Linkage Run

Note that this screen also pops up after [Step 8: Recalculate m-Probabilities and Check Convergence](#). The m-probabilities will then be updated to reflect the estimate from the linkage run most recently completed.

By changing the iteration number, a different starting set of m-probabilities can be loaded. For instance to use the last set of m-probabilities from the 1997 linkage as starting values for the 1996 linkage, copy the file `sastmp.mprobAX` (where x corresponds to the last completed linkage run) for 1997 in the `sastmp` directory for 1996 (they might be the same), and enter the number x as the iteration number on this screen.

Step 5: Linkage of Unlinked Births and Infant/Maternal Matches

Clicking on 'Linkage of newborn discharge and vital statistics birth data' leads to a box asking the user whether any detail information on the individual variables' contributions to the overall match weight should be retained in the data set of matches (Figure 8). Note that retaining this information will lead to higher storage requirements as for each variable two types of match weights are retained. However, the

information can be useful in understanding the composition of the overall general and value-specific frequency weight.

It is important to keep in mind that the application does not at this point prompt for a threshold for the value-specific frequency weight. Rather, thresholds/cutoffs for the value-specific frequencies are stored in the parameter settings as part of [Step 1: Setting Linkage Parameters and Reviewing Linkage Status](#). Matches with a value-specific frequency weight below the acceptable threshold minimum are eliminated before the matches are evaluated. Matches are evaluated by checking the match data sets for ties (e.g., two or more birth records link to the same newborn discharge record; two or more newborn discharge records link to the same birth record, or both). Some ties can be resolved by retaining only the record with the higher match weight. Some ties can be randomized, especially those for which the two birth records and/or the two newborn/maternal discharge record matches pertain to multiples.

Two macros stored in *macros for probabilistic record linkages.sas* carry out the linkage tasks. The first macro is **%linkmac**. It carries out the linkage tasks. The second macro is **%mrkmtchs**. It resolves ties.

The macro **%linkmac** was described in detail in [Step 15: Linkage of Newborn Discharge and Vital Statistics Birth Data, Section 1 above](#). Note that for the purpose of linking unlinked births and infant/maternal matches, the macro is called with the parameter `version set to 'A'`.

Two linkage blocks are used for years 1989 and later. The first linkage block is based on day of birth, month of birth, and year of mother's birth. The second linkage block is based on month of mother's birth, day of mother's birth, and gender. For years prior to 1989, one linkage block is used. This linkage block is based on day of birth, month of birth, and age of mother at time of birth.

The macro **%mrkmtchs** executes immediately after the **%linkmac** macro. It resolves any ties in the matched file by declaring this matched pair the final match that has the largest value-specific frequency weight. If there is a tie for the value-specific frequency weight, the match is accomplished by randomization.

Step 6: Clerical Review of Match Results

Figure 27 shows a screen capture with the options available to select records to be clerically reviewed.

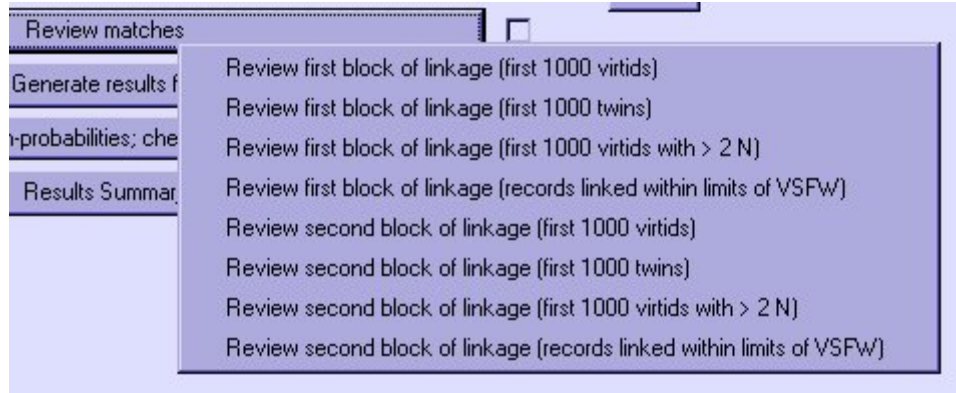


Figure 27: VSPDD-Add-on Linkage, Pop-Menu for the Clerical Review of the Linkage of Unlinked Births and Infant/Maternal Matches

The screen capture in Figure 28 displays the clerical review screen. The screen is primarily used to establish match cutoffs. In some circumstances, specific linkids are marked for removal from the matched file or marked for inclusion even though the match weight is below what is considered the minimum match weight. These circumstances are rare, as just the size of the linkage tasks at hand does not allow time for an elaborate manual clerical review.

Vital ID: 31 Block size: 2 Mom's Birth Date: Birth Date: 2 Gender: 2

Vital Statistics Records:

	ID	disID	dobd	dobm	doby	C/S	Twin	Race	mdobd	mdobm	mdob
1	39635	ZZZZ			1	1997	0	N	1		

OSWPD Discharge Records:

	ID	mathosp	dobd	dobm	doby	C/S	Twin	Race	ZIP	BW	DRG	Status	md
1	1997D00488543	0747			1	1997	0	N	1	95020	11	391	Y

Linkage information:

	zipjarol	vswr	keep
1		29.76	Y

Navigation buttons: < > <> End

Figure 28: VSPDD-Add-on Linkage, Clerical Review of Linkage of Unlinked Births and Infant/Maternal Matches

For the fourth and eighth option on the pop-menu in Figure 27, the user is prompted to enter a minimum and maximum linkage weight (see Figure 19). The clerical review screens will then show all virtual Ids for which at least one match occurred with a linkage weight within the given boundaries.

Step 7: Generate Results File

The macro **%resultA** stored in *macros for probabilistic record linkages.sas* is used to generate a results file for this linkage. The macro requires three parameters: the year of the linkage, the cutoff for the value-specific frequency weight for the first block, and the cutoff for the value-specific frequency weight for the second block (for years 1989 and later only). The macro creates a file named `resultA`. This file includes only the matches that were found. In addition, it includes all the information necessary to re-calculate the m-probabilities in the next step. In other words, it is necessary to create the file `resultA` in order to be able to calculate the m-probabilities.

Step 8: Recalculate m-Probabilities and Check Convergence

After the linkage run is completed, the m-probabilities have to be re-calculated to check convergence. The file `resultA` created in [Step 7: Generate Results File](#) needs to exist for this step to execute. The user is alerted if the file does not exist.

First a temporary file is created via the macro **%inputmprob** stored in *macros for probabilistic record linkage.sas*. This temporary file retains information for each match on whether a variable agreed or disagreed. The macro takes three parameters, the version of linkage currently run (e.g., `D` for deaths, `P` for prenatal/postpartum maternal records, `I` for infant discharge, `M` for maternal discharge, `A` for add-on), the year of the linkage, and the vital statistics input file type (VS, BCF, VSBVSD).

The macro **%getmprob** stored in *macros for probabilistic record linkage.sas* is then used to obtain an updated file of m-probabilities, `sastmp.mprobAX`, where `x` corresponds to the current linkage run iteration. If `x` is larger than 1, the macro also compares the current iteration's m-probabilities to the previous iteration and prints the result in the SAS output window. The result can be used to determine whether or not convergence has occurred. We considered a linkage run to have converged if all differences in m-probabilities were less than 0.01.

After the recalculation of the m-probabilities is completed, the GUI automatically calls the entry `gui.vspdd2.probs.frame` as discussed in [Step 4: Setting of m- and u-Probabilities for Linkage Run](#). If convergence has not yet happened, it can be used to set the most current version of m-probabilities that will be used in the next iteration of the linkage procedure.

Step 9: Update `sastmp.vsbvsdIM` to Include Additional Matches

Finally, the matches stored in `resultA` (see [Step 7: Generate Results File](#)) are used to update the linkage information stored in `sastmp.vsbvsdIM`. The variables in this file were explained in [Step 18: Generating Results Files, Section 7](#). Of these variables, the variables that are highlighted in Table 8 are updated.

Table 8: VS-PDD-Add-on Linkage, Contents of File `sastmp.vsbvsdIM`

Variable	Description	Variable	Variable
----------	-------------	----------	----------

Name		Type	Length
DRG	Diagnosis Related Group	Character	3
BlockI	Block of linkage (PDDI)	Character	1
BlockM	Block of linkage (PDDM)	Character	1
btwght	Birth weight (grams)	Numeric	8
death	Death indicator (VS)	Character	1
Death_	Death indicator (PDD)	Character	1
dobd	Day of birth	Numeric	3
dobm	Month of birth	Numeric	3
doby	Year of birth	Numeric	3
dobyI	Year of birth (PDDI)	Numeric	3
dobyM	Estimated year of birth (PDDM)	Numeric	3
fdeath_	Fetal death (PDDM)	Character	1
hisp	Hispanic ethnicity (VS)	Character	1
linkedB	Infant and maternal PDD linked to VS	Character	1
linkedI	Infant PDD linked to VS	Character	1
linkedM	Maternal PDD linked to VS	Character	1
Linkid		Numeric	
	Unique record identifier (VSB, PDDI, VSD)		8
linkidI	Unique record identifier (PDDI)	Character	13
linkidM	Unique record identifier (PDDM)	Character	13
mathosp	CA maternity hospital code	Character	4
mdobd	Mother's day of birth (VS)	Numeric	3
mdobdM	Mother's day of birth (PDDM)	Numeric	3
mdobm	Mother's month of birth (VS)	Numeric	3
mdobmM	Mother's month of birth (PDDM)	Numeric	3
mdoby	Mother's year of birth (VS)	Numeric	3
mdobyM	Mother's year of birth (PDDM)	Numeric	3
NomtchsI	Number of matches within block of this match (PDDI)	Numeric	8
NomtchsM	Number of matches within block of this match (PDDM)	Numeric	8
paymso	Payer source (VS)	Character	2
Race	Race (VS)	Character	1
tieI	Record was tied (PDDI)	Character	1
tieM	Record was tied (PDDM)	Character	1
twin	Multiple birth (VS)	Character	1
	Multiple birth (VS) (used to check that resultI and resultM have the same twin indicator for the same VS record)		
twinB		Character	1
twinI	Multiple birth (PDDI)	Character	1
twinM	Multiple birth (PDDM)	Character	1
VirtidI		Numeric	
	Virtual ID number (VS for PDDI linkage)		5
VirtidM		Numeric	
	Virtual ID number (VS for PDDM linkage)		5
vsfwI	Value-specific frequency weight (PDDI)	Numeric	8

vsfwM	Value-specific frequency weight (PDDM)	Numeric	8
Zip	Zipcode of mother's residence	Character	5

The variable `linkid` is updated with the found unique birth record identifier. The variables `linkedB`, `linkedI`, and `linkedM` are updated to 'Y' for all additional linked records.

The following variables are added to the file to help recognize records that were linked in this step as well as help evaluate the match:

Variable Name	Description	Variable Type	Variable Length
BlockA	Block of linkage	Character	1
NomtchsA	Number of matches within block of this match	Numeric	8
tieA	Record was tied	Character	1
VirtidA	Virtual ID number	Numeric	5
vsfwA	Value-specific frequency weight	Numeric	8

Note that the difference in the number of records between the old `sastmp.vsbvdsIM` and the new `sastmp.vsbvdsIM` should be equal to the number of matches found in this linkage. Prior to updating `sastmp.vsbvdsIM`, the user is asked whether the previous version of the file should be stored. If the user responds 'Yes' in the message box, the previous version will be stored as `sastmp.vsbvdsIMbackup`.

After `sastmp.vsbvdsIM` has been updated, in order to re-generate the results summaries and the check for bias, it is necessary to exit out of the add-on linkage screen and re-enter the screen for the linkage of vital statistics and infant/maternal discharge data. The last two steps of this linkage should be re-run with the updated results. For instance, after re-running the result summary for 1997, the excerpt shown in Table 7 is updated to what is shown in Table 9.

Table 9: Example for Updated Linkage Results

Multiple VS (Y/N)	Multiple PDDM (Y/N)	Multiple PDDI (Y/N)	linked to both records	linked to mom's record only	linked to infant record only	linked infant and maternal record, VS record			
						maternal record, not VSB	unlinked VS record	unlinked mom's record	unlinked infant record
No	No	No	494,740	-	-	-	-	-	-
No	No	Yes	86	-	-	-	-	-	-
No	Yes	No	166	-	-	-	-	-	-
No	Yes	Yes	36	-	-	-	-	-	-
Yes	No	No	41	-	-	-	-	-	-
Yes	No	Yes	193	-	-	-	-	-	-
Yes	Yes	No	693	-	-	-	-	-	-
Yes	Yes	Yes	12,056	-	-	-	-	-	-
No	No	Missing	-	3,837	-	-	-	-	-
No	Yes	Missing	-	4	-	-	-	-	-

Yes	No	Missing	-	-	-	-	-	-	-
Yes	Yes	Missing	-	73	-	-	-	-	-
No	Missing	No	-	-	2,147	-	-	-	-
No	Missing	Yes	-	-	-	-	-	-	-
Yes	Missing	No	-	-	11	-	-	-	-
Yes	Missing	Yes	-	-	112	-	-	-	-
Missing	No	No	-	-	-	1,502	-	-	-
Missing	No	Yes	-	-	-	9	-	-	-
Missing	Yes	No	-	-	-	20	-	-	-
Missing	Yes	Yes	-	-	-	47	-	-	-
No	Missing	Missing	-	-	-	-	11,001	-	-
Yes	Missing	Missing	-	-	-	-	259	-	-
Missing	Missing	Missing	-	-	-	-	203	-	-
Missing	No	Missing	-	-	-	-	-	4,439	-
Missing	Yes	Missing	-	-	-	-	-	164	-
Missing	Missing	No	-	-	-	-	-	-	4,209
Missing	Missing	Yes	-	-	-	-	-	-	73
			508,011	3,914	2,270	1,578	11,463	4,603	4,282
					1,621				
Total Number of VS Records:					525,658				
Total Number of PDDM Records (double counting multiples):					518,106				
Total Number of PDDI Records:					516,141				
Percent VS records linked to both records:					96.64%				
Percent VS records linked to PDDM record:					97.39%				
Percent VS records linked to PDDI record:					97.07%				
Percent PDDI records linked to both records:					98.42%				
Percent PDDI records linked to VS record:					99.18%				
Percent PDDI records linked to PDDM record:					98.73%				
Percent PDDM records linked to both records:					98.05%				
Percent PDDM records linked to VS record:					98.81%				
Percent PDDM records linked to PDDI record:					98.36%				

9. Linkage of Delivery and Maternal Prenatal/Postnatal Admission Data

Figure 29 shows the main navigation GUI screen for accomplishing the match between the delivery record and any maternal prenatal/postnatal admissions.

Figure 29: PPMOM Linkage, Main Screen

As for the death linkage, the push buttons in the left portion of the screen guide through the linkage steps. As explained under the death linkage the current status and any match parameters can be saved and restored using the top right portion of the screen.

The actions executed with each push button are explained in detail below.

Step 1: Setting Linkage Parameters and Reviewing Linkage Status

The first push button 'Enter Linkage Parameters' (Figure 29) leads to a screen that displays parameters for the current linkage environment.

The parameter screen has the same functionality as similar screens previously described. Note that the VSWF cutoffs refer to the cutoffs for the value-specific frequency weight for the first block and second block respectively.

Path to linkage macros: f:\pdd-vs linkages\programs

Path to formats: f:\pdd-vs linkages\formats

Path to output data: f:\pdd-vs linkages\temporary files

Year of linkage: 1997

VS/BCF file? ☐ VS ☐ BCF ☒ VSBVSD

Next year of PDD data available? ☒ Yes ☐ No

Additional weight information? ☒ Yes ☐ No

M-Probabilities Version: 0

Linkage Block: 2

VSPW cutoffs: 25 65 0

Done Cancel Reset

Figure 30: PPMOM Linkage, Setting Parameters

Step 2: Run extrlnbase: Extract All Records Pertaining to Women Aged 11 to 70 that are Not Deliveries

When clicking on the second pushbutton in Figure 29, a window appears prompting the user to enter the year for which data should be extracted. It also prompts the user for information on the path to the hospital discharge data and which SAS version engine was used to create the discharge data (Figure 31).

Enter year of discharge data from which data should be extracted: 1997

Enter path to discharge data: k:\

Indicate SAS database engine: ☒ V6 ☐ V8

Done Cancel

Figure 31: PPMOM Linkage, Setting Input Specifications for Extraction of Potential Prenatal/Postpartum Records from Discharge Data

The macro **%extrlnbase** stored in *macros for getting data ready for linkages.sas* is used to read all discharge records that pertain to women aged 11 to 70 that are not deliveries, i.e., is assigned to one of the DRGs '370', '371', '372', '373', '374', or '375.' The data are stored in `sastmp.rlnXXXX` where `XXXX` corresponds to the year of discharge data for which data were extracted.

In order to link all deliveries in the year of the linkage to possible prenatal or postnatal hospitalizations, it is necessary to obtain the files `sastmp.rlnXXXX` for the year of the linkage, for the previous year, and for the following year. Note that some of these files might already have been constructed when the linkage for another year was carried out. It might not be necessary to always run this step for the current, previous, and following year.

The next tabulation shows the variables that are included in the file `sastmp.rlnXXXX`.

The coding of all variables remains unchanged except for payer source which is recoded according to the following scheme:

Payer Source	'02'	'01' Medicare '02' MediCal
	'03'	'03' Worker's Compensation
	'05'	'05' CHAMPUS/CHAMPVA/VA '06' Other Governmental
	'06'	'10' Blue Cross/Blue Shield (not HMO/PPO)
	'07'	'09' Private Insurance Company (not HMO/PPO)
	'08'	'07' HMO '08' PPO
	'09'	'11' Self-Pay
	'10'	'12' Charity Care '13' No Charge
	'11'	'14' Other Non-Governmental

Step 3: Run `getPPMom1`: Extract All Delivery Records

The macro `%getPPMom1` stored in *macros for getting data ready for linkages.sas* extracts all possible delivery records against which any hospitalization to women aged 11 to 80 with a non-delivery diagnosis will be matched. It is based on the file `sastmp.subhdfM`. We considered using the file `sastmp.vsbvsdIM` that is generated as a result of the infant and maternal discharge record linkage.⁸

⁸ We actually tried to run the procedure with a file constructed in the following way:

All of the following records are included as delivery records in the linkage:

- Linked to vital statistics birth, infant, and maternal record.
- Linked to vital statistics birth and infant record.
- Linked to vital statistics birth and maternal record.
- Linked to infant and maternal record.
- Unlinked maternal records.
- Unlinked vital statistics birth records.

Only those delivery records were extracted for which the principal procedure date or the year of admission was the year of the linkage. The problem was that for women who gave birth to multiple babies, multiple record identifiers would have been introduced in the file. We dropped the thought for this reason. Note

Prenatal and postpartum records are only matched to delivery records that were recorded in OSHPD hospitals. Women who delivered in non-OSHPD hospitals are not matched by this procedure. Note that this differs from the re-admission and transfer linkage for the children discussed later. For the linkage of transfers/re-admissions any birth is considered, even those that happened in a military hospital, birthing center, or other facility that does not report to OSHPD.

Step 4: Run getPPMom2: Generate File of Possible Prenatal/Postnatal Admissions

Macro **%getPPMom2** stored in *macros for getting data ready for linkages.sas* combines the data extracted under Step 2: Run extrlnbase: Extract All Records Pertaining to Women Aged 11 to 70 that are Not Deliveries for the year of the linkage, the previous year, and the subsequent year into one file, `sastmp.subrln`.

Step 5: Generating Value-Specific Frequency Information and the Set of u-Probabilities

All macros used in this section are stored in the file *macros for probabilistic record linkage preparations.sas*. Note that this section is very similar to Step 5: Generating Value-Specific Frequency Information and the Set of u-Probabilities in Section 6 above.

The generation of the value-specific frequency information is driven by three macros: **%freqsP**, **%reduceP**, and **%vsfreqsP**.

The purpose of the **%freqsP** macro is to generate for each variable to be used in the linkage procedure a table with each possible value of the variable and its percent frequency. These tables will be used to generate the value-specific frequency weights for each variable for the probabilistic linkage. The macro has three parameters: the name of the variable to be analyzed, whether or not missing values should be considered a valid category for this variable, and the variable type (numeric or character). The macro writes an output data set in the `sastmp` directory that is named as the variable with a `P` appended. For instance, for the variable `sex`, an output data set by the name of `sastmp.sexP` would be created. The output data set includes three variables: the variable value, the probability of occurrence of this value in the first data set (birth data), `probP`, the probability of occurrence of this value in the second data set (infant/maternal discharge data), `probP`.

Besides this file, the macro appends an observation to the file `sastmp.genfreqsP`. The file `sastmp.genfreqsP` consists of a description of the variable, `des`, which is the variable name and the general frequency (uncorrected sums of squares (USS)) for the variable on the delivery file and the file of possible prenatal/postnatal admissions.

that the linkage to the maternal discharge data hinges on the record linkage number, in other words, the loss is not big.

For some variables, the same probability value occurs for more than one variable value. The macro **%reduceP** sorts the values of a variable in descending order, determines how often a probability level occurs, and groups values with the same probability level together.

Another issue is that for some variables, a large number of values occur rarely. In this situation, it is inefficient to use the complete table of value-specific frequencies. Rather we have chosen a threshold probability, usually 0.0001, below which the value-specific probability is set to the same value. The macro **%vsfreqsP** writes an ASCII file that assigns for each variable value the value-specific probability level using IF ... ELSE ... constructs. The ASCII files are named using the variable name, appending a 'P' and adding the extension .sas. Note that the macro describes the probabilities for the variable values from the delivery file with the variable `prob`, the probabilities for the variable values from the possible prenatal/postnatal admissions file are described with `prob_`.

The generation of the set of u-probabilities is based on the construction of a file of unlinkable pairs. The macro **%probsP** is used to create this file. Randomly records are merged from the file of delivery records and from the file of possible prenatal/postnatal admissions. Any records that agree on record linkage number (RLN), maternal date of birth and that referred to a hospitalization within the prenatal/postnatal period are removed. The file of unlinkable pairs is called `ulpairs`.

The screen SCL of `gui.ppmom.main.frame` takes care of evaluating the file of unlinkable pairs for the calculation of u-probabilities. After constructing the file `uprobP` with indicator variables for agreement and disagreement of the variables to be used in the linkage, the macro **%uprob** is used to generate the file of u-probabilities `sastmp.uprobP`.

Step 6: Setting of m- and u-Probabilities for Linkage Run

Prior to running a probabilistic linkage step, it is necessary to set the m- and u-probabilities for each variable. These probabilities are set in the form of global macro variables. For instance, for the variable `sex`, global macro variables name **msex** and **usex** are created to represent this variable's m- and u-probability respectively. The screen `gui.ppmom.probs.frame` is used to set the m- and u-probabilities. The screen capture in Figure x shows the screen, as it appears when the screen is first entered.

hfpapat	<input type="text" value="3045872157424"/>	<input type="text" value="0.015644578937"/>	Iteration: <input type="text" value="7"/>
hisp	<input type="text" value="0.943262526878"/>	<input type="text" value="0.504516656208"/>	
hospid	<input type="text" value="0.761272561412"/>	<input type="text" value="0.004054336713"/>	
mdobd	<input type="text" value="0.991985404313"/>	<input type="text" value="0.032426874348"/>	
mdobm	<input type="text" value="0.996758324102"/>	<input type="text" value="0.083956438360"/>	
mdoby	<input type="text" value="0.995031602267"/>	<input type="text" value="0.013801165475"/>	
paymso	<input type="text" value="0.877028083664"/>	<input type="text" value="0.367914440593"/>	
race	<input type="text" value="0.930784518146"/>	<input type="text" value="0.517168376173"/>	
rlin	<input type="text" value="0.879113181729"/>	<input type="text" value="1E-8"/>	
zip	<input type="text" value="0.856258552160"/>	<input type="text" value="0.001317561689"/>	
...	<input type="text"/>	<input type="text"/>	<input type="button" value="Set"/> <input type="button" value="End"/>
...	<input type="text"/>	<input type="text"/>	
...	<input type="text"/>	<input type="text"/>	
...	<input type="text"/>	<input type="text"/>	
...	<input type="text"/>	<input type="text"/>	
...	<input type="text"/>	<input type="text"/>	
...	<input type="text"/>	<input type="text"/>	

Figure 32: PPMOM Linkage, Setting M- and U-Probabilities for Linkage

The m-probabilities are shown in the left column, the u-probabilities in the right column. The box in the top right indicates which version of m-probabilities is shown. As we have not completed any iterations through the linkage procedure, the current iteration is zero. The u-probabilities are derived from the previous step described above. The m-probabilities are set to initial values. These initial values can be changed using this screen if desired. The 'Set' button in the right bottom corner is used to actually set the m- and u-probability for each variable. The 'End' button below it is used to exit the screen.

Note that this screen also pops up after . The m-probabilities will then be updated to reflect the estimate from the linkage run most recently completed.

Note that by changing the iteration number, a different starting set of m-probabilities can be loaded. For instance to use the last set of m-probabilities from the 1997 linkage as starting values for the 1996 linkage, copy the file `sastmp.mprobPX` (where X corresponds to the last completed linkage run) for 1997 in the `sastmp` directory for 1996 (they might be the same), and enter the number X as the iteration number on this screen.

Step 7: Linkage of Maternal Delivery Record and Prenatal/Postnatal Hospitalizations

Clicking on 'Linkage of prenatal and postpartum maternal records' (Figure 29) leads to a box asking the user whether any detail information on the individual variables' contributions to the overall match weight should be retained in the data set of matches (Figure 8). Note that retaining this information will lead to higher storage requirements as for each variable two types of match weights are retained. Processing time will also be increased. However, the information can be useful in understanding the composition of the overall general and value-specific frequency weight.

It is important to keep in mind that the application does not at this point prompt for a threshold for the value-specific frequency weight. Rather, thresholds/cutoffs for the value-specific frequencies are stored in the parameter settings as part of [Step 1: Setting Linkage Parameters and Reviewing Linkage Status](#). Matches with a value-specific frequency weight below the acceptable threshold minimum are eliminated before the matches are evaluated. Matches are evaluated by checking the match data sets for ties (e.g., two or more birth records link to the same prenatal/postpartum admission record; two or more prenatal/postpartum admissions link to the same delivery record, but have conflicting admission and discharge dates; or both). Most ties can be resolved by retaining only the record with the higher match weight. Some ties can be randomized.

Two macros stored in *macros for probabilistic record linkages.sas* carry out the linkage tasks. The first macro is **%linkmacP**. It carries out the linkage tasks. The second macro is **%mrkmtchsP**. It resolves ties.

The macro **%linkmacP** has the following parameters:

Input1	The first data set that includes information on prenatal/postnatal admissions. Note that in the course of the macro all match and block variables will be renamed to include an '_' as the last character of their name.
Input2	The second data set that includes information on deliveries. The name of all block and match variables remains unchanged.
Lnkd	The output file of matched pairs.
Path	The path where the files that were created by the %vsfreqsP macro discussed in Step 5: Generating Value-Specific Frequency Information and the Set of u-Probabilities was created. Usually, these files are stored in the same library that the SAS libname <code>sastmp</code> refers to.
Compcrit	A critical value for all simple agreements: The macro forms all possible matched pairs within a block. As this can result in a large number of matches, as a first cut the macro evaluates all simple agreements within a block, in other words, a variable <code>comp</code> is created that counts the number of times two variables agree. If <code>compcrit</code> is set to a value larger than zero, all matched pairs with less than or equal to <code>compcrit</code> agreements are removed from the file of matched pairs prior to the calculation of general and value-specific frequency weights.
Crit	A critical value for the general frequency weight. In contrast to the value-specific frequency weight, the general frequency weight is based on m- and u-probabilities only. In case of agreement on a variable, the general frequency weight for the variable is the ratio of m- and

	u-probability; in case of disagreement on a variable, the general frequency weight is the ratio of $(1 - m\text{-probability})$ and $(1 - u\text{-probability})$. Note that in case of a variable with a high m-probability and low u-probability (the most desirable situation for a match variable), this leads to a large positive number in case of agreement and number less than 1 in case of disagreement. As logs are taken prior to summing the frequency weights for all variables, agreement results in a large positive weight; disagreement results in a large negative weight. Note that the general frequency weight is the <u>same</u> for different values of the variable. If <code>crit</code> is larger than 0, its value is used to further exclude matched pairs with general frequency weights less than or equal to the <code>crit</code> value.
Block	Indicates the block number.
Srtvr1	The first numeric block variable.
Srtvr2	The second numeric block variable.
Srtvr3	The third numeric block variable.
Srtvr4	The fourth numeric block variable.
Srtvr5	The fifth numeric block variable.
Csrtvr1	The first character block variable.
Csrtvr2	The second character block variable.
Csrtvr3	The third character block variable.
Csrtvr4	The fourth character block variable.
Csrtvr5	The fifth character block variable.
Lastvar	The name of the last block variable. If <u>any</u> numeric block variables are present, this is the last numeric block variable. If <u>only</u> character block variables are present, this is the last character block variable.
Special	Available parameter, not yet used. Can be used to force specific conditions on the file of matched pairs.
Type	Version of linkage: P for prenatal/postpartum records.
Develop	Indicator that is Y or N. If Y, variable-specific contributions to the value-specific frequency weight are included in the file of matched pairs. If N, these weights will not be included.

Furthermore the macro needs the following global parameters to be set:

- All m- and u- probabilities for the match variables.
- The year of the linkage.
- The vital statistics input file type (VS, BCF, VSBVSD).

The first block of the maternal prenatal/postnatal admissions linkage is based on the record linkage number, the encrypted social security number. The second block of the maternal prenatal/postpartum admissions linkage is based on the mother's date of birth.

The macro **%mrkmtchsP** executes immediately after the **%linkmacP** macro. The macro has five parameters, the current block number, the threshold value for the probabilistic match weight, the year of the linkage, the prenatal/postnatal admissions data set that was used in the linkage, and the version of the linkage ('P'). The prenatal/postnatal admissions data set used in the linkage is the file `sastmp.subrln` for the first block, for the second block, it consists of those admissions only that were not linked in the first block.

The macro **%mrkmtchsP** resolves any ties in the matched file by declaring this matched pair the final match that has the largest value-specific frequency weight. If there is a tie for the value-specific frequency weight, the match is accomplished by randomization.

Step 8: Clerical Review of Match Results

Figure 33 shows a screen capture with the options available to select records to be clerically reviewed. Figure 34 shows the second pop-menu asking the user to identify which linkage block is to be clerically reviewed.

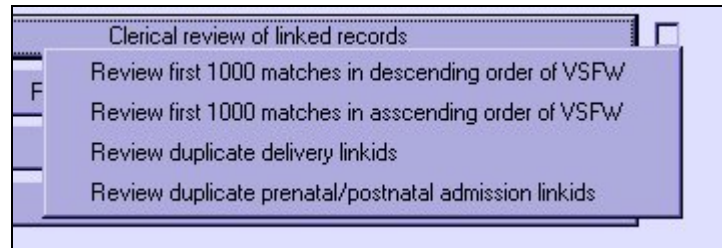


Figure 33: PPMOM Linkage, First Pop-Menu prior to Clerical Review



Figure 34: PPMOM Linkage, Second Pop-Menu prior to Clerical Review

Virtual ID: 612 Block size: 10 Birth Date: RLN:

Birth Record:

	ID	RLN	Mnth	Day	Yr	Admit	Dischg	HFA	ZIP	Race	HospId	Payer	Hispanic
1	1997C00252738							1103		1	331216	02	1
2	1997B00881044							1012		1	301175	02	1

Readmission/Transfer Records:

	ID	RLN	Mnth	Day	Yr	Admit	Dischg	HFA	ZIP	Race	HospId	Payer	Hispanic	DRG
1	1996C00246135							1103		1	331216	02	1	381
2	1997B00880800							1012		1	301175	02	1	384

Linkage information:

	comp	gfw	vsfw	rhjarol	zipjarol	tie_pp	tie_d	keep
1	9	33.67	65.00	.	1.000	1	1	Y
2	9	33.67	65.33	.	1.000	1	1	Y

Navigation buttons: < > <> End

Figure 35: PPMOM Linkage, Clerical Review Screen

The clerical review is carried out in the setting displayed in Figure 35. The screen is primarily used to establish match cutoffs. In some circumstances, specific linkids are marked for removal from the matched file or marked for inclusion even though the match weight is below what is considered the minimum match weight. These circumstances are rare, as just the size of the linkage tasks at hand does not allow time for an elaborate manual clerical review.

The screen below demonstrates the second option 'Review first one thousand matches in ascending order ...' for block 2.

Step 9: Recalculate m-Probabilities and Check Convergence

After the linkage runs are completed, the m-probabilities have to be re-calculated to check convergence.

First a temporary file is created via the macro **%inputmprob** stored in *macros for probabilistic record linkage.sas*. This temporary file retains information for each match on whether the variables from the two files agreed. The macro takes three parameters, the version of linkage currently run (e.g., **D** for deaths, **P** for prenatal/postpartum maternal records, **I** for infant discharge, **M** for maternal discharge), the year of the linkage, and the vital statistics input file type (VS, BCF, VSBVSD).

The macro **%getmprob** stored in *macros for probabilistic record linkage.sas* is then used to obtain an updated file of m-probabilities, `sastmp.mprobPX`, where `x` corresponds to the current linkage run iteration. If `x` is larger than 1, the macro also compares the current iteration's m-probabilities to the previous iteration and prints the result in the SAS output window. The result can be used to determine whether or not convergence has occurred. We considered a linkage run to have converged if all differences in m-probabilities were less than 0.01.

After the recalculation of the m-probabilities is completed, the GUI automatically calls the entry `gui.pppm.probs.frame` as discussed in [Step 6: Setting of m- and u-Probabilities for Linkage Run](#). If convergence has not yet happened, it can be used to set the most current version of m-probabilities that will be used in the next iteration of the linkage procedure.

Step 10: Generate Results File

The results file for the prenatal/postnatal admissions linkage is generated with the **%resultP** macro that is stored in *macros for probabilistic record linkages.sas*. The macro requires a four parameters: the probabilistic match threshold for block 1, the probabilistic match threshold for block 2, the year of the linkage, and the number of blocks.

The macro creates a file called `resultsP` that includes all valid matches with weights larger than the match thresholds specified. The file retains the following information on each match.

Block	Block in which this match occurred.
Keep	This flag indicates whether or not this record was kept after imposing the linkage threshold and resolution of ties (should always be 'Y')
Nomtchs	Total number of matches that occurred in the block within which the current record was linked. Note that each unique block is described by the added variable <code>virtid1</code> .
Tie	This record was the result of resolving a tie. The tie might have been resolved since this record had the highest linkage weight. The tie might have also been resolved by randomization.
Virtid1	A unique number assigned to each different block imposed by the blocking variables.
vsfw	Value-specific frequency weight or match weight that indicates the confidence in this match. The larger VSFW, the better the match.

Step 11: Results Summary

The results summary lists the number of matches obtained per block. It also shows the distribution of the value-specific frequency weight or probabilistic match weight in each block.

10.Linkage of Birth Record and Transfers/Re-Admissions within the first year of life

Figure 36 shows the main navigation GUI screen for navigating through the linkage of transfers and re-admissions to the birth record.

Figure 36: RT Linkage, Main Screen

As for the death linkage, the push buttons in the left portion of the screen guide through the linkage steps. As explained under the death linkage the current status and any match parameters can be saved and restored using the top right portion of the screen.

The actions executed with each push button are explained in detail below.

Step 1: Setting Linkage Parameters and Reviewing Linkage Status

The first push button 'Enter Linkage Parameters' leads to a screen that displays parameters for the current linkage environment.

This screen is very similar to the one displayed in Figure 4 for the vital statistics birth and death linkage. Note that in contrast to the death linkage, the additional choice 'VSBVSD' is available as vital statistics input file. In other words, the file against transfers and re-admissions are matched is the linked vital statistics birth/death file as generated by the steps described in Section 6. The choice 'VS' should be selected if data should be linked to the vital statistics birth file. The choice 'BCF' should be selected if transfers and re-admissions should be linked to the birth cohort file.

Three VSW cutoffs need to be specified. The first box corresponds to the cutoff for the transfer linkage. The second box corresponds to the cutoff for the re-admission linkage. The third box corresponds to the cutoff that was used for the linkage of vital statistics birth and infant discharge data. The latter is needed for successfully running Step 2: Run prepmin: Extract All Possible Birth Records.

Path to linkage macros: h:\pdd-vs linkages\programs

Path to formats: h:\pdd-vs linkages\formats

Path to output data: h:\pdd-vs linkages\temporary files

Year of linkage: 1997

VS/BCF file? ☐ VS ☐ BCF ☒ VSBVSD

Next year of PDD data available? ☒ Yes ☐ No

Additional weight information? ☐ Yes ☒ No

M-Probabilities Version: 0 0

Linkage Block: 1 1

VSW cutoffs: 33.5 20 30

Done Cancel Reset

Figure 37: RT Linkage, Linkage Parameters Entry

Step 2: Run prepmin: Extract All Possible Birth Records

The macro **%prepmin** stored in *macros for probabilistic record linkage preparations.sas* is used to generate the file of births against which transfers and re-admissions are matched. It also generates the set of transfers as well as the set of re-admissions.

Births Input File:

The file of births against which transfers and re-admissions are matched consists of:

1. all successful matches of birth certificate to infant discharge records excluding infants who died in the hospital during their newborn stay; note that if the linked birth/death certificate is used as input, this will include any infant deaths that were linked to the birth certificate.
2. all birth certificate records that could not be matched to an infant discharge record excluding fetal deaths; note that if the linked birth/death certificate is used as input, this will include any infant deaths that were not linked to the birth certificate.
3. all infant discharge records that could be matched to a maternal discharge records, but not to a birth certificate record.

The successful matches are extracted from `vspddi1`. For this reason in the previous step the specification of the final probabilistic match weight is needed. Note that any unmatched infant discharge records are not used in the births input file, rather these records are appended to the file of transfers under the assumption that these cases might have been coded incorrectly. Proceeding in this fashion has the advantage that we actually might still be able to find a match for these cases among the unlinked birth records.

In order to easily identify birth records, a `linkid` is constructed that is:

1. for infant discharge records equal to the record identifier `linkidI`.
2. for birth certificate and linked death certificate records formed by preceding `linkidB` with `BXXXX` where `XXXX` refers to the year of the linkage.
3. for death certificate records that were not linked to a birth certificate formed by preceding `linkidB` with `D`. Note that this is necessary as the `linkidB` for these records already includes the year as its first 4 characters.

The births file is generated by first linking the file `sastmp.subvs2` to the file of successful matches, `sastmp.vsbvdsIM`. Remember that `sastmp.subvs2` consists of all birth and unlinked death certificate records; it does not include unlinked newborn discharge records. At this point, the file `sastmp.vsbvdsIM` was updated with any additional matches between vital statistics birth record and infant/maternal discharge records that were achieved through the add-on steps described in Section 8.

Newborn discharge records that were neither linked to a birth record nor linked to a maternal delivery record are at this point assumed to be transfers that were not coded correctly, in other words, rather than allocating these records to the pool of births that transfers and re-admissions are matched against, these records are allocated to the pool of transfers.

The following table shows the variables included in the births file. Note that some of the variables that appear in this file were generated in [Step 4: Run extpddi: Generate Minimum Infant Discharge File for Vital Statistics Birth and Infant Discharge Record Data Linkage](#) and [Step 7: Run extvs: Extract Minimum Vital Statistics Data File for Linkage to Maternal and Infant Discharge Data in Section 7](#). These variables are marked below.

For the vital statistics birth/death file and the birth cohort file, the variable `death` is added. It indicates whether or not a baby died during its first year of life.

The discharge date is retained in the birth file since it helps to assure that a transfer is linked that does not have dates that overlap with the initial hospitalization.

Table 10: Variables in Births File Against Which Transfers/Re-Admissions are Matched

Variable			
Name	Description	Type	Length
ADMDATE	Admission Date	Numeric	4
DISDATE	Discharge Date	Numeric	4
DRG	Diagnosis Related Group	Character	3

HOSPID	OSHPD Hospital ID	Character	6
PRDIAG	Principal Diagnosis	Character	5
DIAG1	Diagnosis 1	Character	5
DIAG2	Diagnosis 2	Character	5
DIAG3	Diagnosis 3	Character	5
DIAG4	Diagnosis 4	Character	5
DIAG5	Diagnosis 5	Character	5
ZIPHOSP	Hospital Zip	Character	5
bornin	Born in hospital	Character	1
bthdate	Birth date	Numeric	4
bthwght	Birth weight	Numeric	8
bwgrp	Birth weight group	Character	2
code	Discharge Status Recode 1	Character	1
code2	Discharge Status Recode 2	Character	1
county	County of occurrence of birth	Character	2
Csr	C-section Delivery Indicator	Character	1
death	Death Indicator	Character	1
dobd	Day of Birth	Character	3
dobm	Month of Birth	Character	3
doby	Year of Birth	Character	3
hfpapat	Patient's Health Facility Planning Area	Character	4
Hisp	Patient Ethnicity	Character	1
Ilinkid		Numeric	8
	Unique record identifier for birth certificate record		
inhspdth	In-hospital Death	Character	1
Linkid	Unique record identifier for this birth record (see text)	Character	13
mathosp	DHS Hospital Identifier	Character	4
paymso	Payer Source	Character	2
Race	Race	Character	1
rehosp	Re-admission Indicator	Character	1
Sex	Sex of child	Character	1
smhsp	Discharged to Same Hospital	Character	1
twinB	Multiple Birth According to Birth Certificate	Character	1
twinI	Multiple Birth According to Infant Discharge Record	Character	1
Zip	Zipcode of Residence	Character	5

Transfers Input File:

The file of transfers that is matched to the births file consists of all records that were recognized as transfers in Step 4: Run extpddi: Generate Minimum Infant Discharge File for Vital Statistics Birth and Infant Discharge Record Data Linkage in Section 7. These records were stored as `sastmp.subhdfT`. In addition, any newborn discharge records that were not linked to a birth record or a maternal record, are at this point assumed to be miscoded transfers. These records are added to the set of transfers.

Table 11: RT Linkage, Variables Retained in Transfer File

Variable			
Name	Description	Type	Length
Health Information Solutions		7/19/05	

ADMDATE	Admission Date	Numeric	5
BTHDATE	Birth Date	Numeric	5
COUNTY	County of Occurrence of Birth	Character	2
DIAG1	Diagnosis 1	Character	5
DIAG2	Diagnosis 2	Character	5
DIAG3	Diagnosis 3	Character	5
DIAG4	Diagnosis 4	Character	5
DIAG5	Diagnosis 5	Character	5
DISDATE	Discharge Date	Numeric	5
DRG	Diagnosis Related Group	Character	3
LINKID	Unique record identifier	Character	13
PAYMSO	Payer Source	Character	2
PRDIAG	Principal Diagnosis	Character	5
ZIPHOSP	Hospital Zip	Character	5
bornin	Born in hospital	Character	1
bwgrp	Birth weight group	Character	1
code	Admission code	Character	1
code2	Admission code (abridged)	Character	1
death	In-hospital death	Character	1
dobd	Day of birth	Numeric	3
dobm	Month of birth	Numeric	3
doby	Year of birth	Numeric	3
hfpapat	Patient HFPA	Character	4
hisp	Patient Ethnicity	Character	1
hospid	OSHPD Hospital ID	Character	6
hspmtch	Hospital matched to DHS hospital	Character	1
inhspdth	In-hospital death	Character	1
linkedB	Linkage status	Character	1
mathosp	CA maternity hospital code	Character	4
qu_brt	Unlinked newborn record assumed to be transfer indicator	Character	1
race	Race	Character	1
rehosp	Re-admission indicator - always Y	Character	1
sex	Patient Sex	Character	1
smhsp	Admitted from unit in same hospital	Character	1
twin	Twin indicator	Character	1
zip	HDF Zip	Character	5

Re-Admissions Input File:

The file of re-admissions that is matched to the births file consists of all records that were recognized as transfers in Step 4: Run extpddi: Generate Minimum Infant Discharge File for Vital Statistics Birth and Infant Discharge Record Data Linkage in Section 7. These records were stored as `sastmp.subhdfR`.

Table 12: RT Linkage, Variables Retained in Re-Admissions File

Variable			
Name	Description	Type	Length
ADMDATE	Admission Date	Numeric	5
BTHDATE	Birth Date	Numeric	5
COUNTY	County of Occurrence of Birth	Character	2
DIAG1	Diagnosis 1	Character	5
DIAG2	Diagnosis 2	Character	5
DIAG3	Diagnosis 3	Character	5
DIAG4	Diagnosis 4	Character	5
DIAG5	Diagnosis 5	Character	5
DISDATE	Discharge Date	Numeric	5
DRG	Diagnosis Related Group	Character	3
LINKID	Unique record identifier	Character	13
PAYMSO	Payer Source	Character	2
PRDIAG	Principal Diagnosis	Character	5
ZIPHOSP	Hospital Zip	Character	5
bornin	Born in hospital	Character	1
bwgrp	Birth weight group	Character	1
Code	Admission code	Character	1
Code2	Admission code (abridged)	Character	1
death	In-hospital death	Character	1
Dobd	Day of birth	Numeric	3
Dobm	Month of birth	Numeric	3
Doby	Year of birth	Numeric	3
hfpapat	Patient HFP A	Character	4
Hisp	Patient Ethnicity	Character	1
hospid	OSHPD Hospital ID	Character	6
inhspdth	In-hospital death	Character	1
linkedB	Linkage status	Character	1
race	Race	Character	1
rehosp	Re-admission indicator - always Y	Character	1
sex	Patient Sex	Character	1
smhsp	Admitted from unit in same hospital	Character	1
twin	Twin indicator	Character	1
Zip	HDF Zip	Character	5

Step 3: Generating Value-Specific Frequency Information and the Set of u-Probabilities

All macros used in this section are stored in the file *macros for probabilistic record linkage preparations.sas*. Note that this section is very similar to Step 5: Generating Value-Specific Frequency Information and the Set of u-Probabilities in Section 6 above.

The generation of the value-specific frequency information is driven by three macros: **%freqsTR**, **%reduceTR**, and **%vsfreqsTR**.

The purpose of the **%freqsTR** macro is to generate for each variable to be used in the linkage procedure a table with each possible value of the variable and its percent frequency. These tables will be used to generate the value-specific frequency weights for each variable for the probabilistic linkage. The macro has three parameters: the name of the variable to be analyzed, whether or not missing values should be considered a valid category for this variable, and the variable type (numeric or character). The macro writes an output data set in the `sastmp` directory that is named as the variable with a `TR` appended. For instance, for the variable `sex`, an output data set by the name of `sastmp.sexTR` would be created. The output data set includes three variables: the variable value, the probability of occurrence of this value in the births data set, `probB`, the probability of occurrence of this value in the transfers data set, `probT`, and the probability of occurrence of this value in the re-admissions data set, `probR`.

Besides this file, the macro appends an observation to the file `sastmp.genfreqsTR`. The file `sastmp.genfreqsTR` consists of a description of the variable, `des`, which is the variable name and the general frequency (uncorrected sums of squares (USS)) for the variable on the births, transfers, and re-admissions files respectively.

For some variables, the same probability value occurs for more than one variable value. The macro **%reduceTR** sorts the values of a variable in descending order, determines how often a probability level occurs, and groups values with the same probability level together.

Another issue is that for some variables, a large number of values occur rarely. In this situation, it is inefficient to use the complete table of value-specific frequencies. Rather we have chosen a threshold probability, usually 0.0001, below which the value-specific probability is set to the same value. The macro **%vsfreqsTR** writes an ASCII file that assigns for each variable value the value-specific probability level using IF ... ELSE ... constructs. The ASCII files are named using the variable name, appending a 'T' or an 'R' for the transfer and re-admissions linkage respectively, and adding the extension `.sas`. Note that the macro describes the probabilities for the variable values from the births file with the variable `prob`, the probabilities for the variable values from the possible transfers/re-admissions file are described with `prob_`.

The generation of the set of u-probabilities is based on the construction of a file of unlinkable pairs. The macro **%probsTR** is used to create this file. Randomly records are merged from the file of births records to the file of transfers. Any records that agree on record linkage number birth date and gender are removed. The file of unlinkable pairs is called `ulpairsT`. The macro repeats the same steps for the re-admissions.

The macro **%getuprob** evaluates the file of unlinkable pairs for agreement on the variables that were identified as linkage variables. The macro has one parameter, `type`, which is `R` or `T` depending upon whether the evaluation should occur for `ulpairsT` or `ulpairsR`. For the comparison of county of birth, past linkages were used to identify those counties within which transfers of kids were observed. In other words,

the variables county was not only considered in agreement if a transfer occurred within the county, but also within a subset of the remaining California counties. In contrast to previous evaluations of agreement of variables on the file of unlinkable pairs, for transfers and re-admissions, the evaluation also includes the diagnosis codes. All diagnoses except for V-codes are compared.

After constructing the files `uprobT` and `probR` with indicator variables for agreement and disagreement of the variables to be used in the linkage, the macro `%uprob` is used to generate the file of u-probabilities `sastmp.uprobTR`.

Step 4: Setting of m- and u-Probabilities for Linkage Run

<input checked="" type="radio"/> Transfers <input type="radio"/> Re-admissions		Iteration: <input type="text" value="0"/>
county	<input type="text" value="0.9999"/>	<input type="text" value="0.531067108856"/>
paymso	<input type="text" value="0.9"/>	<input type="text" value="0.395776433866"/>
race	<input type="text" value="0.9"/>	<input type="text" value="0.486441669918"/>
hospid	<input type="text" value="0.9"/>	<input type="text" value="0.004243074522"/>
zip	<input type="text" value="0.95"/>	<input type="text" value="0.000975419430"/>
dobd	<input type="text" value="0.9999"/>	<input type="text" value="0.031847444401"/>
dobm	<input type="text" value="0.9999"/>	<input type="text" value="0.079984393289"/>
doby	<input type="text" value="0.9999"/>	<input type="text" value="1"/>
sex	<input type="text" value="0.9999"/>	<input type="text" value="0.506291455325"/>
hfpapat	<input type="text" value="0.95"/>	<input type="text" value="0.015801794771"/>
code	<input type="text" value="0.9"/>	<input type="text" value="0.340226297307"/>
code2	<input type="text" value="0.9"/>	<input type="text" value="0.342079594225"/>
bwgrp	<input type="text" value="0.9"/>	<input type="text" value="0.675087787748"/>
smdy	<input type="text" value="0.9"/>	<input type="text" value="0.002097151775"/>
dx4	<input type="text" value="0.9"/>	<input type="text" value="0.031018337885"/>
dx3	<input type="text" value="0.9"/>	<input type="text" value="0.056574326960"/>
hisp	<input type="text" value="0.9"/>	<input type="text" value="0.475907140070"/>
smhsp	<input type="text" value="0.9"/>	<input type="text" value="0.347346859149"/>
death	<input type="text" value="0.9"/>	<input type="text" value="0.000195083886"/>
...	<input type="text"/>	<input type="text"/>

Figure 38: RT Linkage, Setting of M- and U-Probabilities for Linkage

Prior to running a probabilistic linkage step, it is necessary to set the m- and u-probabilities for each variable. These probabilities are set in the form of global macro variables. For instance, for the variable sex, global macro variables name **ms_{sex}** and **us_{sex}** are created to represent this variable's m- and u-probability respectively. The screen `gui.rt.probs.frame` is used to set the m- and u-probabilities. The screen capture in Figure 38 shows the screen, as it appears when is first entered and the user has checked 'Transfers' in the radio box in the top left corner of the screen.

The m-probabilities are shown in the left column, the u-probabilities in the right column. The box in the top right indicates which version of m-probabilities is shown. As we have not completed any iterations through the linkage procedure, the current iteration is zero. The u-probabilities are derived from the previous step described above. The m-probabilities are set to initial values. These initial values can be changed using this screen if desired. The 'Set' button in the right bottom corner is used to actually set the m- and u-probability for each variable. The 'End' button below it is used to exit the screen.

The radio box in the top left corner of the screen can be used to set for which linkage m- and u-probabilities are set.

Note that this screen also pops up after . The m-probabilities will then be updated to reflect the estimate from the linkage run most recently completed.

Note that by changing the iteration number, a different starting set of m-probabilities can be loaded. For instance to use the last set of m-probabilities from the 1997 transfer linkage as starting values for the 1996 transfer linkage, copy the file `sastmp.mprobTX` (where X corresponds to the last completed linkage run) for 1997 in the `sastmp` directory for 1996 (they might be the same), and enter the number X as the iteration number on this screen.

Step 5: Linkage of Transfers and Births

Clicking on 'Linkage of transfers and births' (Figure 36) leads to a box asking the user whether any detail information on the individual variables' contributions to the overall match weight should be retained in the data set of matches (Figure 8). Note that retaining this information will lead to higher storage requirements as for each variable two types of match weights are retained. However, the information can be useful in understanding the composition of the overall general and value-specific frequency weight.

It is important to keep in mind that the application does not at this point prompt for a threshold for the value-specific frequency weight. Rather, thresholds/cutoffs for the value-specific frequencies are stored in the parameter settings as part of [Step 1: Setting Linkage Parameters and Reviewing Linkage Status](#). Matches with a value-specific frequency weight below the acceptable threshold minimum are eliminated before the matches are evaluated. Matches are evaluated by checking the match data sets for ties (e.g., two or more birth records link to the same transfer record; two or more transfer records link to the same birth record and have a conflict in admission and discharge dates). Most ties can be resolved by retaining only the record with the higher match weight. Some ties can be randomized, especially those for which the two birth records and/or the two newborn discharge records pertain to multiples.

Two macros stored in *macros for probabilistic record linkages.sas* carry out the linkage tasks. The first macro is **%linkmacTR**. It carries out the linkage tasks. The second macro is **%mrkmtchsTR**. It resolves ties.

The macro **%linkmacTR** has the following parameters:

Input1	The first data set that includes information transfers or re-admissions. Note that in the course of the macro all match and block variables will be renamed to include an '_' as the last character of their name.
Input2	The second data set that includes information on births. The name of all block and match variables remains unchanged.
Lnkd	The output file of matched pairs.
Path	The path where the files that were created by the %vsfreqsTR macro discussed in Step 3: Generating Value-Specific Frequency Information and the Set of u-Probabilities was created. Usually, these files are stored in the same library that the SAS libname <code>sastmp</code> refers to.
Compcrit	A critical value for all simple agreements: The macro forms all possible matched pairs within a block. As this can result in a large number of matches, as a first cut the macro evaluates all simple agreements within a block, in other words, a variable <code>comp</code> is created that counts the number of times two variables agree. If <code>compcrit</code> is set to a value larger than zero, all matched pairs with fewer than or equal to <code>compcrit</code> agreements are removed from the file of matched pairs prior to the calculation of general and value-specific frequency weights.
Crit	A critical value for the general frequency weight. In contrast to the value-specific frequency weight, the general frequency weight is based on m- and u-probabilities only. In case of agreement on a variable, the general frequency weight for the variable is the ratio of m- and u-probability; in case of disagreement on a variable, the general frequency weight is the ratio of (1 – m-probability) and (1 – u-probability). Note that in case of a variable with a high m-probability and low u-probability (the most desirable situation for a match variable), this leads to a large positive number in case of agreement and number less than 1 in case of disagreement. As logs are taken prior to summing the frequency weights for all variables, agreement results in a large positive weight; disagreement results in a large negative weight. Note that the general frequency weight is the <u>same</u> for different values of the variable. If <code>crit</code> is larger than 0, its value is used to further exclude matched pairs with general frequency weights less or equal to <code>crit</code> value.
Block	Indicates the block number, in case of the infant linkage, this can only be 1. In case of the maternal linkage, this can be a number as high as 10.
Srtvr1	The first numeric block variable.
Srtvr2	The second numeric block variable.
Srtvr3	The third numeric block variable.
Srtvr4	The fourth numeric block variable.
Srtvr5	The fifth numeric block variable.
Csrtvr1	The first character block variable.
Csrtvr2	The second character block variable.
Csrtvr3	The third character block variable.

Csrtvr4	The fourth character block variable.
Csrtvr5	The fifth character block variable.
Lastvar	The name of the last block variable. If <u>any</u> numeric block variables are present, this is the last numeric block variable. If <u>only</u> character block variables are present, this is the last character block variable.
Special	Available parameter, not yet used. Can be used to force specific conditions on the file of matched pairs.
Type	Version of linkage: T for transfers, R for re-admissions.
Develop	Indicator that is Y or N . If Y , variable-specific contributions to the value-specific frequency weight are included in the file of matched pairs. If N , these weights will not be included.

Furthermore the macro needs the following global parameters to be set:

- All m- and u- probabilities for the match variables.
- The year of the linkage.
- The vital statistics input file type (VS, BCF, VSBVSD).

Two blocks are used to accomplish the transfer linkage. The first block is based on date of birth and infant's health facility planning area. The second block is based on date of birth and gender.

The macro **%mrkmtchsTR** executes immediately after the **%linkmacTR** macro. The macro has five parameters, the type of linkage (T for transfers; R for re-admissions), the current block number, the threshold probabilistic match weight, the year of the linkage, and the input data set of transfers. The input data set consists of all transfers when running **%mrkmtchsTR** for the first block; for the second block, it consists of all transfers that were not linked in the first block. The macro is used to resolve ties for the birth record and ties that come to exist because dates of hospitalizations for two or more matching transfers overlap. Most ties can be resolved by declaring this matched pair the final match that has the largest value-specific frequency weight. If there is a tie for the value-specific frequency weight, the match is accomplished by randomization.

Step 6: Linkage of Re-Admissions and Births

For the re-admission linkage the procedure is the same as for the transfer linkage described in the previous paragraph. The blocking structure is also the same as for the transfer linkage.

Step 7: Clerical Review of Match Results

Figure 39 shows a screen capture with the options available to select records to be clerically reviewed.

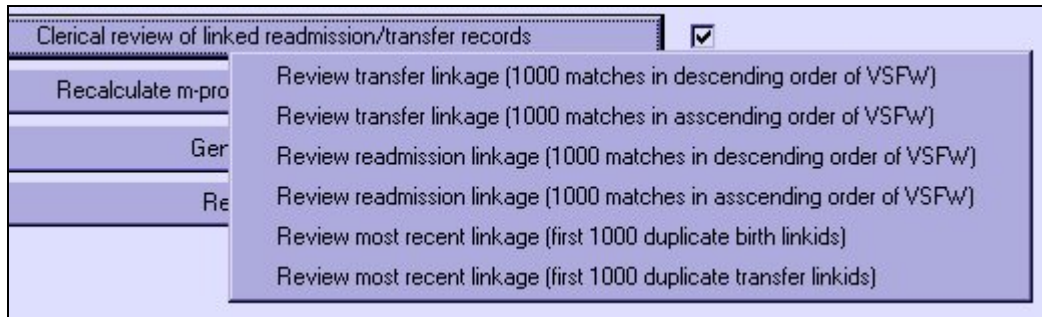


Figure 39: RT Linkage, Pop-Menu for Clerical Review

After this menu, a second menu pops up asking the user to identify the block that should be reviewed (Figure 40).

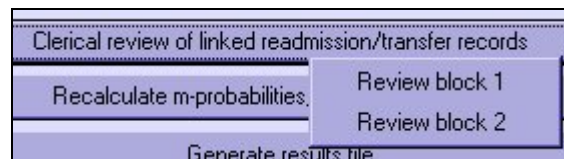


Figure 40: RT Linkage, Pop-Menu for Clerical Review Asking for Linkage Block

Figure 41 displays the clerical review screen. The screen is primarily used to establish match cutoffs. In some circumstances, specific linkids are marked for removal from the matched file or marked for inclusion even though the match weight is below what is considered the minimum match weight. These circumstances are rare, as just the size of the linkage tasks at hand does not allow time for an elaborate manual clerical review.

Virtual ID: 9826 Block size: 1 Birth Date: Gender: 1

Birth Record:

	ID	twm1	twm2	ZIP	Race	EW	HFA	smhsp	Payer	Hispanic	County	code	code2	Discharge	HspId	DX1	DX2
1	1997C00662936	N	N		1	11	1211	3	02	2	San Bernardino	1	1		361343	V3000	

Readmission/Transfer Records:

	ID	ZIP	Race	EW	HFA	smhsp	Payer	Hispanic	County	code	code2	Admit	Discharge	HspId	DX1	DX2
1	1997C00560344		1	11	1211	2	08	2	San Bernardino	3	2			361246	769	7706

Linkage information:

	comp	gfw	vfw	mpjrol	tie_rt	tie_b	keep
1	9	21.251870126	46.12	0.600	1	1	Y

Navigation buttons: < > <> End

Figure 41: RT Linkage, Clerical Review Screen

Step 8: Recalculate m-Probabilities and Check Convergence

After the transfer or re-admission linkage runs are completed, the m-probabilities have to be re-calculated to check convergence. The user is prompted to indicate whether to re-calculate m-probabilities for the transfer or re-admission linkage (Figure 42).

Recalculate m-probabilities: check convergence

Re-Calculate M-Probabilities for transfer linkage

Re-Calculate M-Probabilities for readmission linkage

Results Summary

Figure 42: RT Linkage, Pop-Menu for Recalculation of M-Probabilities

For the re-calculation of m-probabilities, first a temporary file is created via the macro **%inputmprob** stored in *macros for probabilistic record linkage.sas*. This temporary file retains information for each match on which variables matched and which variables did not match. The macro takes three parameters, the version of linkage currently run (e.g., **D** for deaths, **P** for prenatal/postpartum maternal records, **I** for infant discharge, **M** for maternal discharge, **R** for re-admissions, **T** for transfers, **A** for add-on), the year of the linkage, and the vital statistics input file type (VS, BCF, VSBVSD).

The macro **%getmprob** stored in *macros for probabilistic record linkage.sas* is then used to obtain an updated file of m-probabilities, `sastmp.mprobTX` and `sastmp.mprobRX`. Where *x* corresponds to the current linkage run iteration. If *x* is larger than 1, the macro also compares the current iteration's m-probabilities to the previous iteration and prints the result in the SAS output window. The result can be used to determine whether or not convergence has occurred. We considered a linkage run to have converged if all differences in m-probabilities were less than 0.01.

After the recalculation of the m-probabilities is completed, the GUI automatically calls the entry `gui.rt.probs.frame` as discussed in [Step 4: Setting of m- and u-Probabilities for Linkage Run](#). If convergence has not yet happened, it can be used to set the most current version of m-probabilities that will be used in the next iteration of the linkage procedure.

Step 10: Generate Results File

The final results file is created by the macro **%_mrkmtchsTR** which is stored in *macros for probabilistic record linkages.sas*. While the macro **%mrkmtchsTR** ran immediately after a linkage run was completed to identify the set of final matches and remove ties, it is necessary that a similar check is run for all linked transfers and re-admissions combined. Prior to running **%_mrkmtchsTR** the file `vspddTR` is created as part of the SCL of `gui.rt.main.frame`. It consists of the files `mtchsT1`, `mtchsT2`, `mtchsR1`, and `mtchsR2`, the results files of the individual matches for each block produced by the macro **%mrkmtchsTR**. Added is a variable `type` that indicates whether the match was found in the transfer or re-admission run, and a variable `block` that indicates in which block the match was found.

The macro **%mrkmtchsTR** will check for any violations of the re-admission history, i.e., conflicts in the date of admission and discharge for each baby. If it finds a violation, the record with the highest match weight will be retained as a match. For 1997, we found about 100 violations.

The following variables were added to by the linkage procedure to allow assessment of the quality of the match.

Block	Block in which this match occurred
Keep	This flag indicates whether or not this record was kept after imposing the linkage threshold and resolution of ties (this indicator is 'Y' or 'N'; need to make sure that for the final linkage master file, only records with <code>keep EQ 'Y'</code> are retained)
Nomtchs	Total number of matches that occurred in the block within which the current record was linked. Note that each unique block is described by the added variable <code>virtid1</code> .
Tie_b	Two or more birth records were linked to this transfer/re-admission record; ties was resolved by using the match with the highest match weight; if match weight was tied, match was randomized
Tie_rt	Two or more transfers/re-admissions has overlapping admission and discharge dates; those records were retained as matches that had the highest match weight; if match weight was tied, match was randomized.
Virtid1	A unique number assigned to each different block imposed by the blocking variables.
vsfw	Value-specific frequency weight or match weight that indicates the confidence in this

match. The larger VSFW, the better the match.

Step 10: Generate Results Summary

The options for the results summary are shown in Figure 43.

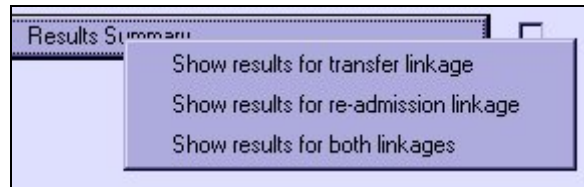


Figure 43: RT Linkage, Pop-Menu for Results Summary

The following results are displayed for the user selection.

- The number of matches per block.
- The number of matches by plurality and ties.
- The distribution of the value-specific frequency weight.

11. Generation of Final Summary File

All macros mentioned in this section are stored in the SAS program file *macros for storing all linkages in one file.sas*.

Health Information Solutions: PDD-VS Linkage

Generation of Summary File from All Linkages for Cohort

Location of input files	<input type="checkbox"/>	Project Name	Save
Remove from VSBVSDIM newborn records linked as transfers; link VS births	<input type="checkbox"/>		
Obtain full input file for linked prenatal/postnatal admissions	<input type="checkbox"/>		
Merge full maternal discharge record to cohort	<input type="checkbox"/>	Check Log	Check Output
Append transfer/re-admissions to VSBVSDIM and merge to full discharge record	<input type="checkbox"/>		
Append prenatal/postnatal admissions to VSBVSDIM	<input type="checkbox"/>	End	
Garble up record identifiers	<input type="checkbox"/>		
Sort data set by BRTHID and order of events	<input type="checkbox"/>		
Add BTHDATE of birth record to all records	<input type="checkbox"/>		
Run all steps at once			

Figure 44: Generation of Summary File, Main Screen

Figure 44 displays the main screen guiding through the generation of the summary file that includes all linked records in a format that will allow the use of the data in epidemiological studies. Each of the steps involved are explained in more detail in the sections below.

Location of Input Files

Year of cohort/linkage:

Data library of VSBVSDIM:

Data library of non-delivery/newborn records:

Data library of initial data inputs:

Is next year of discharge data available? ☐ Yes ☐ No

Figure 45: Generation of Summary File, Location of Input Files

Figure 45 shows that the user can specify the SAS data library for:

- the location of the file `vsbvsdIM`. This file was generated after the vital statistics birth, infant discharge, and maternal discharge records were linked. It allows the construction of the final summary file as it includes for each birth, newborn, or maternal delivery record information on its linkage status as well as ancillary variables that allow assessment of the quality of the match. The location can be picked from a drop-down list of active data libraries.
- The location of the result files for the maternal prenatal/postnatal admission linkage and the infant transfer/re-admission linkage. The location can be picked from a drop-down list of active data libraries.
- The location of the master input files. These files are essentially copies of the vital statistics or discharge records as they were read from the raw data..

Besides these data libraries, the user also has to specify the year for which the files should be created, and the user has to indicate whether the next year of discharge data is available or not. The latter is mostly of importance for the creation of the correct file of reference raw data for the prenatal/postnatal admissions linkage.

Step 1: Remove from `vsbvsdIM` Newborn Records Linked as Transfers; Link VS Births

The file `vsbvsdIM` that was created as the result of the vital statistics birth, infant and maternal discharge record linkage (Step 18: Generating Results Files in Section 7 and Step 9: Update `sastmp.vsbvsdIM` to Include Additional Matches in Section 8) still considers all unlinked infant discharge records as newborn records. However, in the transfer linkage, we assumed that for some reason these records had been coded as newborns, but did in fact pertain to a transfer. Any records that were linked as a transfer record need to be removed from `vsbvsdIM` to obtain a correct reference set of birth records, newborn, and delivery records. This is the first task that is accomplished by the macro **%step1**.

The second task accomplished by this macro is the matching of the linked vital statistics birth/infant death file to `vsbvsdIM`. Note that `vsbvsdIM` only retained some of the information found in the birth or death

certificate. Matching `vsbvsdIM` to the original linked birth/death file adds all the information in the birth record to the summary file. From now on the file `vsbvsdIM` is named `cohXXXX` where `XXXX` is the year of the linkage.

This step also performs a third important task. It generates the variable `brthID` according to the following rules:

- If a vital statistics birth record was not linked, linked to an infant discharge record only, linked to a maternal discharge record only, or linked to both discharge records, the `brthID` is equal to the unique record identifier of the birth record as stored in `vsbvsdXXXX`. For birth records, the identifier is preceded by `BXXXX`, for death records, the identifier is preceded by `DXXXX`. The numeric linked is concatenated to this string with leading zeroes.
- If an infant discharge record was not linked or only linked to the maternal discharge record, the `brthID` is equal to the unique record identifier of the infant discharge record.
- If a maternal discharge record was not linked, the `brthID` is equal to the unique record identifier of the maternal discharge record.

In the following, this `brthID` is also referred to as **reference birth identifier**.

Prior to doing any work, the GUI will check whether all necessary input files are available for the step to be carried out successfully. If input files are missing or paths undeclared, the message box displayed in Figure 46 appears.

Similar error boxes will pop up if the step finds an unexpected number of observations or if there is a problem with linking based on the record identifiers.

If the step is completed without errors, a message box as displayed in Figure 47 should appear.

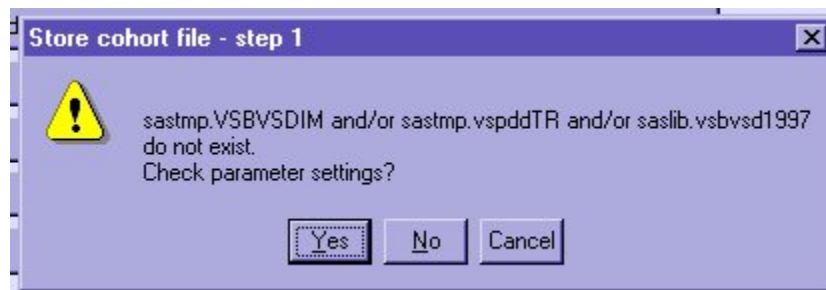


Figure 46: Generation of Summary File, Message Box Indicating Inability to Locate Data Files in Step 1



Figure 47: Generation of Summary File, Message Box Step 1

Step 2: Obtain Full Input File for Linked Prenatal/Postnatal Admissions

The file `resultP` generated as part of the prenatal/postnatal maternal admissions linkage contains information on the record identifier of the delivery record as well as the record identifier of the prenatal/postnatal re-admission. In this step, complete information for all prenatal/postnatal maternal admissions is added and a temporary file `rln` is created.

Note that we have a particular problem here due to the use of multiple births in the linkages: One maternal delivery record links to more than one birth record in the case of twins. Therefore, for each maternal delivery record, all birth record identifiers are obtained, and the prenatal/postnatal admission history is merged to each of these birth record identifiers. In addition, a variable is generated `_twinwght` that is 1 for the first birth record matched to a maternal delivery record, and 0 otherwise. The variable `_twinwght` can be used to obtain a unique count of all mothers in the database.



Figure 48: Generation of Summary File, Message Box Step 2

This step also takes care of renaming all variables that come from the maternal discharge record by adding a capital `M` to their name.

A message box such as the one displayed in Figure 48 appears if the step completed successfully.

Step 3: Merge full Maternal Discharge Record to Cohort

In this step, the maternal discharge record is added to the summary file `cohXXXX`. In other words, the complete information stored in the discharge record is added to the file. The step checks to make sure that

all maternal records that were linked to a birth or newborn discharge record are indeed represented in the master maternal delivery discharge file.

If the maternal discharge data are merged successfully, a message box as displayed in Figure 49 should appear.



Figure 49: Generation of Summary File, Message Box Step 3

Step 4: Append Transfers/Re-Admissions to Cohort and Link to Infant Discharge Records

The file `vspddTR` contains all transfers/re-admissions that were successfully linked to a birth/newborn record. Note that only records with `keep EQ 'Y'` are included in the summary file. The records in `vspddTR` are appended to the cohort file, and subsequently the cohort file is matched to the complete infant discharge record. The GUI checks whether any record identifiers in the file of matches could not be found in the reference file of complete infant discharge records.

A message box such as the one shown in Figure 50 pops up after the step completes successfully.



Figure 50: Generation of Summary File, Message Box Step 4

Step 5: Append Prenatal/Postnatal Maternal Admissions

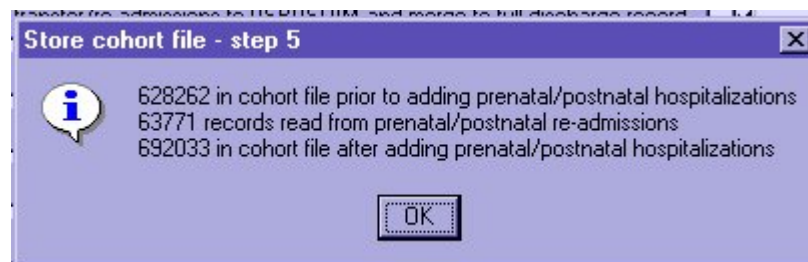


Figure 51: Generation of Summary File, Message Box Step 5

All prenatal/postnatal admissions with complete discharge information as generated in [Step 2: Obtain Full Input File for Linked Prenatal/Postnatal Admissions](#), `rln`, is appended to the summary file `cohXXXX`.

A message box such as the one shown in Figure 51 pops up after the step completes successfully.

Step 6: Garble Up Record Identifiers to Disallow Link Back to Published OSHPD or DHS Records

As the record identifiers used in the linkages are based on the sequence in which the data appear in the raw data files as they were received from OSHPD and DHS, it is necessary to garble up the record identifiers to disallow any linkage back to information published by OSHPD and DHS in other files.

This step scrambles up the record identifier for the reference birth record, birth record, infant discharge record, and maternal discharge record.

A message box such as the one shown in Figure 52 pops up after the step completes successfully.



Figure 52: Generation of Summary File: Message Box Step 6

Step 7: Sort Data Set by Reference Birth Identifier and Within Reference Birth Identifier by Order of Events

In order to enable a database user to connect a birth with the mother's prenatal/postnatal hospitalization history and the infant's transfers/re-admissions, the cohort file is sorted by the reference birth record identifier, `brthID`. Within each `brthID`, the file is sorted by the admission dates in the maternal and infant records, the discharge date in the maternal and infant discharge record and the `_input` value, the source data set. For this purpose, two variables, `order1` and `order2`, are created in the following fashion:

Variable	<code>_input EQ 'B'</code>	<code>_input EQ 'I'</code>	<code>_input EQ 'M'</code>
<code>Order1</code>	<code>Bthdate</code>	<code>AdmdateI</code>	<code>AdmdateM</code>
<code>Order2</code>	<code>Bthdate</code>	<code>DisdateI</code>	<code>DisdateM</code>

In case that a mother was admitted and discharged on the same day of the delivery, the data set is also sorted by a `_input`, however, for the purpose of this step, contributions from the maternal prenatal/postnatal admission linkage are given the value 'A.' This guarantees that in the above situation, this prenatal hospitalization will come before the birth hospitalization.

After this step completes, a screen similar to Figure 53 appears.



Figure 53: Generation of Summary File, Message Box Step 7

Step 8: Beautify Cohort File and Produce Result Summaries

The final steps for completion of the cohort's summary file are performed by the macro **%step8**. This macro generates several additional variables that have proved to be a useful addition to the summary file in the past:

_dob	The date of birth is added to all prenatal/postnatal admissions and transfer/re-admission records. For unlinked maternal delivery admissions, this variable is set to the principal procedure date, if the principal procedure date is missing, it is set to the admission date.
_diffI	The difference in days between this record's admission date and birth date. This variable is useful to define the follow-up period during which transfers and re-admissions are included in an analysis.
_diffM	The difference in days between this record's admission date and the newborn's birth date/delivery date. This variable is useful in identifying the timing of the maternal admission relative to the time of birth. Note that this number is negative for prenatal admissions.
_losI	A copy of the <code>lenstayI</code> variable, however, this variable is of type numeric.
_losM	A copy of the <code>lenstayM</code> variable, however, this variable is of type numeric.
_chargesI	A copy of the <code>chargesI</code> variable, however, this variable is of type numeric.
_chargesM	A copy of the <code>chargesM</code> variable, however, this variable is of type numeric.
_dthind	A second death indicator variable. Takes the value '1' for a neonatal death, takes the value '2' for a postneonatal death, takes the value '0' if this infant survived the first year.
AgedaysI	Defined according to OSHPD conventions that were established for years 1995 and later, i.e., it is the baby's age in days of under 1-year olds <i>at time of admission</i> . For under 4-year olds, it is the number of days alive since the last birthday at time of admission. For all other ages, it is missing. See OSHPD manual for 1995 or later for a more detailed explanation.
AgeyrsI	Defined according to OSHPD conventions that were established for years 1995 and later, i.e., it is the baby's age in years <i>at time of admission</i> . See

	OSHDP manual for 1995 or later for a more detailed explanation.
agedaysM	Defined according to OSHDP conventions that were established for years 1995 and later, i.e., it is the mother's age in days of under 1-year olds <i>at time of admission</i> . For under 4-year olds, it is the number of days alive since the last birthday at time of admission. For all other ages, it is missing. See OSHDP manual for 1995 or later for a more detailed explanation.
ageyrsM	Defined according to OSHDP conventions that were established for years 1995 and later, i.e., it is the mother's age in years <i>at time of admission</i> . See OSHDP manual for 1995 or later for a more detailed explanation.

Furthermore, in this step, all variables are assigned labels. The definition of all of the variables can be found in the spreadsheet *list of variables in linked cohort file.xls*.

Finally, this step runs some more summary statistics that should be added to the spreadsheet XXXX *linkage results.xls*. Specifically, the following items are generated:

- A tabulation of _input by _linkedB.
- A tabulation of low birth weight by linkage status for hospitals that could be matched and for hospitals that could not be matched.
- A tabulation of more refined birth weight categories by linkage status for hospitals that could be matched and for hospitals that could not be matched.
- A tabulation of the death outcome by linkage status for hospitals that could be matched and for hospitals that could not be matched.
- A tabulation of DRG by linkage status for hospitals that could be matched and for hospitals that could not be matched.
- A hospital level tabulation of all births of records eligible for linkage, records linked to both files, records linked only to one file, and records that were not linked at all.
- A hospital level tabulation of singleton births of records eligible for linkage, records linked to both files, records linked only to one file, and records that were not linked at all.
- A hospital level tabulation of multiple births of records eligible for linkage, records linked to both files, records linked only to one file, and records that were not linked at all.

After this step completes, a screen such as the one displayed in Figure 54 appears.



Figure 54: Generation of Results File, Message Box Step 8