**DATA PROCESSING PROCEDURES 04:**

**PREPARING AND VALIDATING THE STANDARD RECODE**

Table of Contents

[1. Introduction 4](#_Toc52552344)

[2. Overview and purpose of the Standard Recode Data File 4](#_Toc52552345)

[2.1 Introduction 4](#_Toc52552346)

[2.2 Definitions 4](#_Toc52552347)

[2.3 Rationale for having a standard recode data set 5](#_Toc52552352)

[2.4 Selective standardization of response categories 5](#_Toc52552353)

[2.5 Simplification of variables where appropriate and addition of summary variables 6](#_Toc52552354)

[2.6 Calculated variables to facilitate analysis 8](#_Toc52552355)

[2.7 Treatment of country specific variables 9](#_Toc52552356)

[3. Required Steps for Creating and Validating the Standard Recode File 9](#_Toc52552357)

[3.1 Overview of Process 9](#_Toc52552358)

[4. Recode Dictionaries 11](#_Toc52552359)

[4.1 Treatment of Non-De-Facto Cases 11](#_Toc52552360)

[4.2 Sample Design Variables 11](#_Toc52552361)

[4.3 Initial Steps in Modifying Dictionary 11](#_Toc52552362)

[4.4 Key Points for Handling Standard Variables 12](#_Toc52552363)

[4.5 Country Specific Variables 12](#_Toc52552364)

[4.6 Treatment of Unused Standard Variables 13](#_Toc52552365)

[4.7 Switching Between Relative and Fixed Positions in the Recode Dictionary 14](#_Toc52552366)

[4.8 Finalizing the Dictionary 15](#_Toc52552367)

[5. Recode Applications 15](#_Toc52552368)

[5.1 Modifying the Recode Application 15](#_Toc52552369)

[5.2 Treatment of Multiple Response Questions 17](#_Toc52552370)

[5.3 Modifying the Consistency Applications 18](#_Toc52552371)

[5.3.1 Introduction 18](#_Toc52552372)

[5.3.2 Preparing the A2Q File for the CONSISx Application 18](#_Toc52552373)

[5.3.3 Running the CONSIS Application 19](#_Toc52552374)

[6. Recode Documentation 19](#_Toc52552375)

[7. Frequency Checking 20](#_Toc52552376)

[7.1 Introduction and Overview 20](#_Toc52552377)

[7.2 Guide to Production of Frequencies for Checking 21](#_Toc52552378)

[7.3 Procedures for Checking Frequencies 22](#_Toc52552379)

[8. Preparation of Data Files and Documentation for Distribution and Archiving 23](#_Toc52552380)

[9. Archive All Available Country Documentation 25](#_Toc52552381)

[10. Recode Specifications for Malaria Indicator Surveys (DHS7) 26](#_Toc52552382)

[Appendix A. Correspondence of Raw Data Dictionary Sections to Recode Data File Sections 29](#_Toc52552383)

[Appendix B. Variable Naming Conventions 31](#_Toc52552384)

# Introduction

After all the data finalization steps have been completed the raw data file must be converted to the standard recode file structure (“Recode file”) as a prerequisite for all further steps in processing the survey. The recode file is used to generate tables for the main report and is also the format used for public distribution of data in the [DHS Data Archive](https://dhsprogram.com/data/). In addition, the recode data file is used to populate the [Statcompiler](https://www.statcompiler.com/en/) database with indicators from the survey.

In the recode file variables are combined and created in a form that is easy to use for analysis. Many summary variables and computed variables used in the DHS reports are included in the recode file. Most importantly, the data in the recode file are in a standardized format allowing easy comparison of data among surveys within a country or among countries.

The three components in preparing the recode file: recode dictionaries, recode applications and recode documentation will be discussed. Each of these components are best prepared simultaneously, since modifications to the standard dictionary or standard application should generally be reflected in either or both of the other two components. It is thus suggested to use multiple windows: Word for the documentation, and CSPro or a text editor such as [Notepad ++](https://notepad-plus-plus.org/) or [Gvim](https://www.vim.org/) for the applications.

All files used during data entry and data processing, as well as all available country documentation that have to be archived and prepared for distribution are listed in the last two sections of this chapter.

# Overview and purpose of the Standard Recode Data File

## Introduction

After data collection, DHS data are put through a standardization process where the original raw data is transformed into a common data format termed the “DHS Standard Recode”. This Standard Recode data file is then used for generating tabulations for the DHS Final Report, and also forms the basis of the datasets that are distributed to the public on the DHS Website.

The purpose of this document is to instructions for creating and checking, the standard recode file and procedures for the recoding and treatment of country specific variables. It draws extensively on information in the [DHS Standard Recode Manual](https://www.dhsprogram.com/pubs/pdf/DHSG4/Recode7_DHS_10Sep2018_DHSG4.pdf).

## Definitions

“Raw data”: this term is used to refer both to data collected in the field, and associated CSPro data dictionaries that were used to capture that data. This applies both to surveys where data was collected electronically or where paper questionnaires were used in the field and then entered later.

The structure of the raw data mirrors the structure of the questionnaire very closely. Almost all fields in the questionnaire will have a corresponding variable in the data dictionary. The names of variables in most cases will correspond directly to question numbers or names in the questionnaire.

The “recode data” structure by contrast does not mirror the country questionnaire directly. Instead, it consists of all variables in the raw data, converted into a standardized structure. This standardized structure is the same across all surveys in a given phase of the DHS. For certain variables, responses to multiple questions in the raw data are combined to facilitate analysis.



## Rationale for having a standard recode data set

The main rationale for having a standardized data set is that it facilitates analysis across different DHS surveys, both over different countries, and over different phases of the DHS project. The DHS project has been conducted over a 30 year period, during which time the content and structure of the questionnaires has changed considerably. For example, during the early 2000’s, many questions were added to address sexual behavior and knowledge and attitudes to HIV/AIDS. As sections have been added, section numbers have changed. For example, in DHS phase 2 the section on marriage was section 5 of the woman’s questionnaire, in the most recent version of the questionnaire it is section 7. However, using the standard recode data set, an analyst can be sure that variable names will not change across project phases. In the example of the marriage section, for all DHS datasets, the standard recode variable V501 always contains marital status, regardless of the phase of DHS that the survey took place in.

Similarly, having a standardized data structure also minimizes the differences in questionnaire design and content across different countries. Due to questions being added and deleted to meet the needs of different countries, a given question may be assigned a different question number, and hence variable name in the raw data, across different countries. However, in the standard recode data set, the same standard variable names are always used, regardless of the country or phase of DHS.

## Selective standardization of response categories

For most standard recode variables, response codes are standardized and will not change regardless of country or project phase. For example, variable V501 current marital status always has the same response codes.

**V501 Current marital status**

0 Never in union

1 Married

2 Living with partner

3 Widowed

4 Divorced

5 No longer living together/separated

(m) 9 Missing

However, for certain variables, mostly ones which have a high degree of variation between countries and regions, country-specific response codes are used that may differ across countries, but which are nevertheless organized into standardized groups based on the first digit of the response code. An example of such a variable is HV215, main roof material of the household building. In the table below, the response categories for surveys in Benin and Jordan are shown. Because of the regional and economic differences between Benin and Jordan, different types of construction materials are used in each country for the roofs of houses. Even with these different response categories however, the analyst can assume that all codes between 11 and 19 represent natural materials, those between 21-29 rudimentary materials and those between 30-39 finished materials.

|  |  |  |
| --- | --- | --- |
| **HV215 Main roof material response categories** | | |
|  | **Benin 2017 DHS** | **Jordan 2017-18 DHS** |
| **10** | **NATURAL** | **NATURAL** |
| 11 | No roof |  |
| 12 | Thatch / palm leaf |  |
| 13 | Sod |  |
| **20** | **RUDIMENTARY** | **RUDIMENTARY** |
| 21 | Rustic mat | Mud bricks |
| 22 | Palm / bamboo | Mud bricks with stones |
| 23 | Wood planks |  |
| 24 | Cardboard |  |
| **30** | **FINISHED** | **FINISHED** |
| 31 | Metal | Concrete |
| 32 | Wood | Zinc/metal |
| 33 | Zinc / cement fiber |  |
| 34 | Ceramic tiles |  |
| 35 | Cement |  |
| 36 | Roofing shingles |  |
| 96 | Other | Other |

Other examples of these types of variables include:

HV201: Source of drinking water

HV214: Main wall material

V3A07: First source for current method

M15: Place of delivery

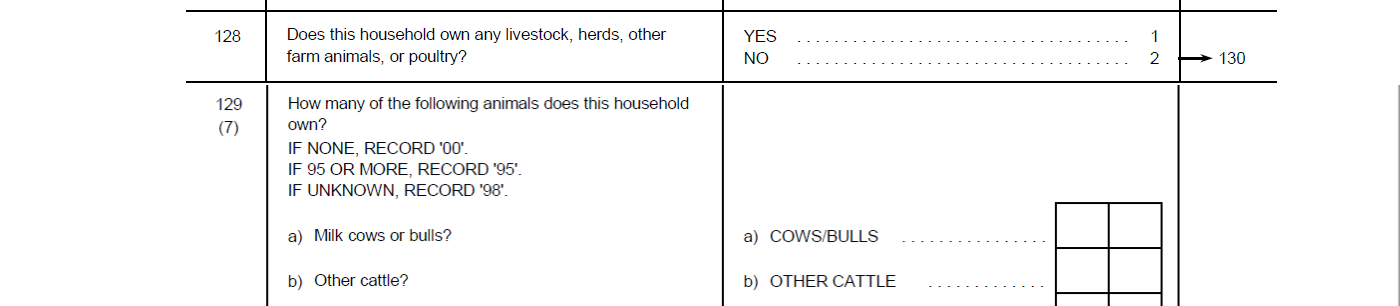
V829: Place where last HIV test was taken

## Simplification of variables where appropriate and addition of summary variables

Summary variables like number of living children, number of children in the last 5 years preceding the survey, number of de-facto household members etc. are often used in analysis. To facilitate analysis, the standard recode data structure contains a large number of such summary variables, including those that are used in the DHS reports.

In addition to summary variables, wherever possible, data items are simplified and adapted in the recode structure as compared with their raw data equivalents to ease analysis. Three principal ways this is done are:

For many questions requiring a numeric response in the original questionnaire, there will be an initial filter question to determine if the numeric question applies to the respondent. For example, Q128 in the household questionnaire asks if the household possesses any livestock or farm animals such as cows. If the response is “Yes”, the respondent is then asked how many animals of various types they possess in questions Q129A-F. For households with no farm animals, all responses Q129-AF will be set to system missing in the raw data, which can cause difficulties when using analysis software, particularly if arithmetic operations are being used (for example, to sum up the total number of animals to calculate a mean). To address this problem, the recode data variables for farm animals, HV246-AF, are set to a value of zero if QH128 has a “No” response and questions QH129A-F were skipped. This approach is used for all similar numeric variables in the recode data set



All variables with yes/no response categories are dichotomized. This is done by converting the “No” response in the raw data, which by convention is always coded with a value of 2 in the raw data, to a zero value in the recode data.

|  |  |  |  |
| --- | --- | --- | --- |
| **Household has electricity** | | | |
| **Raw Data: QH132A** | | **Recode Data: HV206** | |
| **Code** | **Value** | **Code** | **Value** |
| 1 | Yes | 1 | Yes |
| 2 | No | 0 | No |

The DHS questionnaire contains many questions that allow for more than one response. For example, Q415 in the woman’s questionnaire asks for source of antenatal care. Since the respondent may receive antenatal care from more than one source, this question allows multiple responses. In the raw data, this information is collected as one single alphanumeric variable with alphabetical codes to indicate the various response categories. However, for purposes of analysis, particularly with statistics packages such as SPSS, alphanumeric variables are difficult to work with. Consequently, in the recode dataset, multiple response variables are split into disjoint, dichotomized variables, one for each possible response.

The example below shows the mapping between the raw data responses for Q410, source of antenatal care in the 2017-18 Jordan DHS, and the recode variables M57A-Z. Note that this is also an example of a question where there is a mixture of standard categories such as “respondent’s home” and “other home” that are always coded into the same variables, and country specific categories such as “Royal medical services”.

|  |  |  |
| --- | --- | --- |
| **Response for Q410** | **Recode variable** | **Label for M57A-Z** |
| A" | M57A | Antenatal care: respondent's home |
| B" | M57B | Antenatal care: other home |
|  | M57C | NA - Antenatal care: country specific home |
|  | M57D | NA - Antenatal care: country specific home |
| C" | M57E | Antenatal care: government hospital |
| D" | M57F | Antenatal care: Government health center |
| E" | M57G | Antenatal care: University hospital |
| F" | M57H | Antenatal care: Royal Medical Services |
| G" | M57I | Antenatal care: Other public |
|  | M57J | NA - Antenatal care: country specific public health |
|  | M57K | NA - Antenatal care: country specific public health |
|  | M57L | NA - Antenatal care: country specific public health |
| H" | M57M | Antenatal care: private hospital/clinic |
| "I" | M57N | Antenatal care: UNRWA health center |
| "J" | M57O | Antenatal care: UNHCR/other NGO |
| "K" | M57P | Antenatal care: Other private |
|  | M57Q | NA - Antenatal care: country specific private medical |
|  | M57R | NA - Antenatal care: country specific private medical |
|  | M57S | NA - Antenatal care: country specific other |
|  | M57T | NA - Antenatal care: country specific other |
|  | M57U | NA - Antenatal care: country specific other |
|  | M57V | NA - Antenatal care: country specific other |
| "X" | M57X | Antenatal care: other |

## Calculated variables to facilitate analysis

The recode data set contains a number of calculated variables to facilitate analysis. These variables will combine two more existing variables in the recode dataset. For example, recode variable V602: fertility preference, or desire for more children. Two questions about the desire to have another child are asked in the woman questionnaire: one is asked of women who are currently pregnant, and another one asked of women who are not currently pregnant. If working with the raw data, an analyst would need to write logic to include both questions. In the recode file these two variables are combined and in a form that is easier to use for analysis.

## Treatment of country specific variables

The standard recode data set is intended to be fully complete and contain all data collected during the survey, including country specific questions that are not part of the DHS standard questionnaire. All country-specific questions are included in specific sections in the data dictionary as shown in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Questionnaire** | **Type of variable** | **Raw** | **Recode section** |
| **Household** | Non-repeated variables from cover sheet, household characteristics or other single sections | AHSECOVER, AHSEC01, AHSEC02, AHSEC03, AHSEC03, AHSEC04, ABSECOVER | RECH3 Survey specific Household variables |
| Repeated variables from household roster | AHSEC01 | RECH4 Survey specific Household Schedule variables |
| **Woman** | Variables from any single section in the woman’s questionnaire |  | REC91 |
| Birth history | AWSEC2B | REC92 Country specific Birth History variables |
| Pregnancy and Postnatal Care | AWSEC04 | REC94 Country specific Maternity variables |
| Vaccinations and child health | AWSEC05, AWSEC06A, AWSEC06B | REC95 Country specific Health and Vaccination variables |

# Required Steps for Creating and Validating the Standard Recode File

## Overview of Process

The overall process for creating and validating the Standard recode data set for a given survey is as follows:

1. Data Finalization is completed prior to running the Key Indicator Report (KIR) tables
2. The standard recode dictionary and applications are modified and a preliminary draft recode data set is produced
3. The data processor responsible for the survey generates and checks the raw data frequencies against the recode frequencies to ensure that the recode data is correct, and makes any necessary corrections.
4. An application named CONSIS is adapted and run on the recode data which carries out a number of consistency checks on the recoded data. All consistency messages are reviewed and followed up. Follow up typically involves either corrections to the data or documentation of identified issues.
5. A second data processor conducts another frequency review once the CONSIS step has been finalized
6. A formal review is carried out on the recode and raw data file to ensure that no personal identifying information (PII) is contained in the data sets. This review is carried out by the survey data processor and a second review is carried out by a senior member of the data processing team.
7. Once any corrections have been implemented from the second frequency review and PII checks, the data file is passed to the Data Archive for export and distribution on the DHS Program data archive website.

Work should begin on the standard recode dictionary and applications well in advance of the data finalization activity phase. If possible, an initial check of the recode file should be done using data from the field. All recode applications **must be based** on the most recent, up to date versions of the standard recode RECODEx applications on Sharepoint. It is never acceptable to re-use old applications from previous surveys.

The following activities **must be been completed** prior to running any tables for the DHS Final Report:

* 1. Applications and dictionaries for generating the Recode files completed and checked
  2. Recode data files generated and checked, including male recode (if applicable)
  3. All country specific variables and sections added
  4. All Not Applicable variables identified and labeled
  5. 1st Frequency check by data processor responsible for survey completed and any issues identified during the check resolved

Prior to final report tables being sent to the publications group for final formatting, the following activities must have been completed:

1. CONSIS application run and all issues identified corrected in the data
2. 2nd frequency check by another member of the data processing team completed and all issues identified by the check resolved

At least one week prior to the release of the Final report the following activities must be completed:

1. All outstanding activities described in steps 3 and 4 above
2. Completion of the country recode documentation
3. PII check completed and PII document signed by 1st Data processor, and PII review completed and document signed by 2nd Data Processor
4. Distrib checklist completed
5. Files uploaded to the .\Distrib network location ready for exporting and distribution

# Recode Dictionaries

The recode data file is structured in two parts, standard sections and country-specific sections. The standard sections contain the same variables in the same positions for all countries. The country-specific sections contain variables specific to the country.

The standard recode dictionary for DHS VIII type questionnaires (RECODE8.DCF), containing the household and individual questionnaires, can be found in [SharePoint under the Standard DHS8 Project](https://icfonline.sharepoint.com/sites/ihd-dhs/Standard8/SitePages/DHS-Standard-Materials.aspx). Four-digit fields for each variable pertaining to a year are used in RECODEx.DCF, whereas recodes up to DHS4 had two digits for these variables.

A key difference between RECODE8.DCF and recodes up to DHS6 is the inclusion of dates calculated in Century Day Code. In addition, the ages of children and therefore the base for many sections pertaining to children in the DHS7 and DHS8 recode are also computed based on the Century Day Code, whereas in recodes up to DHS6 these were calculated from the Century Month Code. Adjusting MRECODEx.DCF in the same directory creates the husband/male recode file.

In the first rounds of the DHS, the records in the recode files followed the structure of the questionnaires closely. In subsequent rounds of the DHS many questions and whole sections were added to the questionnaires and some questions moved to different parts of the questionnaires. Therefore, the structure of the recode files follows the questionnaires less closely, but there is still a close correspondence between questionnaires and recode structure.

## Treatment of Non-De-Facto Cases

Incomplete interviews are included in the recodes, but non-de-facto women and men are excluded from the recodes (unless the sample selected was a de-jure sample in which case, non-de-jure women and men are excluded from the recodes).

## Sample Design Variables

Ensure that the appropriate labels are included for variables concerned with sample design. These are:

HV022/V022/MV022: Sample strata

HV023/V023/MV023: Stratification used in sample design

Usually, the strata comprise the codes for regions broken down by urban and rural, but in some surveys other stratification methods may be used. Verify with the sampler, the correct label for each value.

## Initial Steps in Modifying Dictionary

1. Set the survey name and year in the Dictionary Label section in CSProDesigner.
2. Change the maximum number of occurrences for multiple records in the household data to their true maximum. For RECH1 and, if applicable, RECHMH, RECHDIS and RECH4, this number is the maximum number of household members and can be found in the frequencies of variable HVIDX. For biomarker sections RECH5 (children), RECH6 (women) and RECH6 (men), this number corresponds to the highest number of occurrences for each section in the survey data. Set household level country specific multiple records and groups to true maximum.
3. In the woman’s recode, do NOT change the maximum occurrences for any section The calendar record, if it exists, must have at least 2 occurrences.
4. Remove records that are not used and that are not required to be present. Records that should be removed if unused are shown below. **No other records** **should be removed even if all variables are unused.**

Household – all after REECH6

Woman: all after REC81

Men: all after MREC80

## Key Points for Handling Standard Variables

1. With regard to dates, the DHS Standard Recode does not retain the originally reported dates of birth (except for day of birth).
2. Variable names, locations and lengths in standard sections cannot be changed. Country specific variables are added only in the country specific sections. If a variable exists in the raw data that would normally be recoded into a standard variable, but has longer length than can be accommodated by the standard variable, it should be stored as a country specific variable.

## Country Specific Variables

In general, every variable from the raw data file should be included in the recode data file. **No data should be lost**. If a standard variable cannot be recoded without collapsing categories, then a country-specific variable must be added with the original categories. If non-standard groupings of the variable are used in the country report, a variable containing these groupings must be added as country-specific.

Exceptions to the above rule include raw data variables containing originally reported dates of birth, filter variables, variables with repeated data, and variables that are used by the CAPI system for controlling program flow.

To identify country specific variables, run the *findcs.exe* utility once all country specific variables have been added to the recode dictionary and associated logic included in the recode applications. This utility has three file parameters as shown below.

|  |  |
| --- | --- |
| **DCF Filename:** this is the raw data dictionary filename  **APP Filename**: the name of the recode application  **LST Filename**: name of the output file containing |  |

The *findcs* utility looks for the name of each variable in the raw data dictionary and checks to see if that variable name occurs in the recode application. At the end of the run, a list of all variables found in the raw data dictionary but not in the recode application is output in the LST filename.

This list should be reviewed and checked to identify which variables should be included in the recode data as country specific.

Country-specific variables should be named Sxxx where xxx is the original question number. Household country specific variables should be named SHxxx. Male questionnaire variables should be named SMxxx, irrespective of whether the data is from a male or husband questionnaire.

Country-specific variables should be in ascending numerical order based on the question number of the variable, in the appropriate section.

When creating country specific variables, they should be coded using the same principles as standard variables, as described in section 2 of this manual. For example:

* Yes/No variables must be recoded to codes 1:Yes, 0:No
* Source variables should be coded using codes 11-19, 21-29 etc to represent higher level categories such as public and private sector
* Multiple response variables must be broken into a series of dichotomized yes/no variables in the same way as is done for standard variables (see section 4.2 below for more information)
* Country specific variables must not contain any decimal points or alphabetical characters, that is, they should all be integers. If a country specific raw data variable has decimal places it should be converted into a country specific variable with an implied decimal place, and this should be clearly documented in the variable label.
* Wherever possible country specific variables should be merged where it makes sense to do so

Cover variables that are country-specific must be present for all households, women, and men even if their specific questionnaire is incomplete.

If the survey data contains country specific ever-married factors, they should be clearly labeled with an indication of what specific background variable they are to be used with. For example, if there is an ever-married factor for education attainment based on V149, then this information should be included that in the label of the country specific factor variable.

## Treatment of Unused Standard Variables

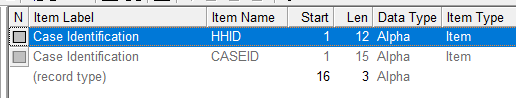
For any unused standard variables, the characters ‘NA-‘ must be added to the first 3 positions of the variable label in the dictionary, and the variable name documented in the country recode documentation.

To identify unused variables, use the *nasearch*.*exe* utility with the irfreq.lst file as input. The *nasearch* utility should not be run until after the initial frequency check for the recode file has been completed. This utility checks the frequency output file for all variables that have only not applicable values and outputs a list of variables as a text file. This list should be reviewed carefully to ensure that all variables identified are really unused, rather than being set to not applicable due to a programming logic error. If necessary, the recode logic should be corrected to include any variables wrongly set to not applicable.

Once a final list of unused variables has been made, use the *naassign.exe* utility to add the ‘NA’ labels to the dictionary.

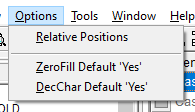
## Switching Between Relative and Fixed Positions in the Recode Dictionary

The dictionary containing the household and women’s data is designed in such a way that the two identification fields HHID and CASEID overlap, as is shown in the image below. The standard recode dictionary has household ID beginning in position 1 for 12 characters, individual ID beginning in position 1 for 15 characters and record type beginning in 16 for 3 characters. **Data values must begin in location 19 on all records.**

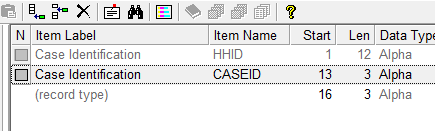


In order for variables to be permitted to overlap when using the CSPro designer, the default setting is for all variable positions to be fixed in the dictionary. However, this setting makes it difficult to re-order country specific variables due to the fact that it prevents variables being inserted in a list of existing variables in a record.

The default dictionary setting can be changed to “relative positions” to facilitate addition and re-ordering of country specific variables. This is done by clicking on the Options menu item and checking “Relative”.

****

When switching to relative position, change the individual ID, CASEID,to begin in position 13 with a length of 3; that will keep all variables in the correct location.



When switching back to fixed positions, remember to change the individual ID to begin in 1 and have a width of 15.

Note that this restriction does not apply to the male data dictionary.

## Finalizing the Dictionary

The following steps should be completed:

1. Ensure that RELATIVE position is not checked in the dictionary – it must be in FIXED position
2. Ensure that the HHID is in position 1 for 12, the CASEID is in position 1 for 15, and the REC\_ID is in position 16 for 3.
3. All unused standard variables are marked with the ‘ -NA’ label.
4. Multiple records are set for their maximum (if allowed to be changed)
5. All unnecessary records have been removed.
6. Spelling errors have been corrected
7. The dictionary name has been updated with the country name, year, and phase of DHS.

# Recode Applications

The recode application uses as input the final raw data file

Two programs are used to generate and check the standard recode data file. The recode application (RECODEx.APP) to generate the data file, and the consistency application (CONSISx.APP) to check the consistency among variables in the recode data file. For the men’s recode, MRECODEx.APP generates the recode file. To check the consistency, it is necessary to create two small files containing household IDs and women’s line numbers and basic information. These files are used with the men’s recode when running MCONSISx.APP

## Modifying the Recode Application

The standard recode applications for DHS type household and individual questionnaires can be found on Sharepoint in the \RECODE folder. RECODEx.APP should be used with dictionaries with four-digit year fields. The standard recode application for a male/husband type survey (MRECODEx) is located in the same directory.

Below are the steps and general guidelines in constructing a country specific recode application from the standard recode application:

1. Run the A2Q utility on the RECODEx and MRECODEx applications.
2. Remove logic that is related to standard variables and sections that are not used in the survey data. In the recode application, look for the X or Z variables from the A2Q that identify unused standard questions or variables that are in any way different from the standard questionnaire.
3. Modify the constants found in the ‘dictionary’ proc. Check the ‘calendar’ variable creations with care.

* Change CALBEGY to the first year of the calendar in the questionnaire.
* Set METHOTH to the occurrence of the first “other” category in the contraceptive table.
* Adjust xprm and/or xsec when there are more or fewer than 6 grades in primary and secondary levels respectively.
* VMOTHER and VMARKED need only be changed when different codes are used to indicate in the vaccination table that the mother said that the child received the vaccination or that it was evident from the vaccination card that the child received the vaccination but no exact date was given.
* HSEcountry specific is used to indicate whether a 3 or 5 year period is used in the health section.
* Check to see whether LASTCH should be set to 1 (only last child asked A408-A423 and A437-A444) or if additional children were asked these questions.
* Do not change ACALLEN and VCALLEN.
* Adjust ACALSTRS when the codes in the columns used in the calendar are different from the standard. These codes should correspond to their equivalents in VCALSTRS. All codes in VCALSTRS are standard. If it is not possible to fit the country specific codes in the country-specific letter codes allotted in VCALSTRS then use column 5. This column is generally used to record additional events in the calendar, specific to a country.
* VCALMETH, VCALDISC are generally not changed.

For the male recode set the following:

* Set METHOTH to the occurrence of the first other category in the contraceptive table.
* Adjust xprm and/or xsec when there are more or fewer than 6 grades in primary and secondary levels respectively.

1. Review the logic in the education functions LevEduc, GraEduc, SingEduc and AttnEduc to ensure that the logic matches the levels and grades in the country education system. If necessary, modify the logic.
2. Do a search for ‘!!!’ characters in the logic which identify places where logic may need to be changed to address particular characteristics of the country survey data. Carefully review each instance to determine if modifications are needed.
3. Review the logic and response category labels for standard variables which nevertheless have country specific response categories. Variables which most commonly require adjustment are listed below:

|  |  |
| --- | --- |
| Source of drinking water | HV201 |
| Source of non-drinking water | HV202 |
| Type of toilet facility | HV205 |
| Main floor material | HV213 |
| Main wall material | HV214 |
| Main roof material | HV215 |
| Source of drinking water | V113 |
| Source known for any method | V379 |
| Source known for any method | V380 |
| Source of condoms used for last sex | V762 |
| Source of antimalarial during pregnancy | ML2 |
| Who checked respondent health before discharge | M64 |
| Who checked respondent health after discharge/ | M68 |
| Who checked child health before discharge | M76 |
| Who performed daughter's circumcision | G124 |

1. Check the logic for V301-V307 and V312 to ensure that the contraceptive methods in the raw data are correctly mapped to the standard codes in the recode data.
2. Be sure to assign the ‘cover’ country specific variables outside of the logic restricting assignment to cases with a complete result code (QHRESULT/QRESULT = 1) so that all households, women and men have data for those country specific variables, including cases without a complete interview.
3. Pay attention to the logic handling assignments for all multiple response variables, particularly those related to source of health care, tests or drugs. In most countries the standard logic will need to be modified to ensure that the responses are correctly mapped from the raw data to the standard recode variables. More detail on this is provided in section 4.2 below.
4. If messages regarding ‘no occurrence’ of certain records are present in the recode application listing, all such messages should be followed up and corrected. There should be no run time messages present for any recode application.

Once the recode dictionary and application are complete, run the application and check the LST file for warnings. If warnings exist then the recode dictionary and/or the application should be adjusted. Also check the recode data file for stars (e.g. use *find.exe* in a DOS window or load the data file into Multi-Edit or Notepad++ to find them). The stars should be eliminated by adjusting the dictionary and/or application. Repeat running the application until no more warnings are generated or all are considered acceptable and no more stars appear in the output data file. It is also helpful to run the HR, IR and MRFREQ applications and look for “@” or “default” to identify out of range values or stars.

## Treatment of Multiple Response Questions

The standard DHS questionnaire contains a number of questions that can have multiple responses. All multiple response variables are split into disjoint, dichotomized variables, one for each possible response. A number of these variables list sources of various types of health care or medications which will usually require some modification to the standard logic, and to the recode dictionary labels, to fit the specific sources used in the country. A list of these variables (current as of DHS7) is shown below.

|  |  |
| --- | --- |
| V3A00A-X | Source of family planning for non-users |
| M2A-M3N | Source of Prenatal care |
| M57A-M57X | Source of Antenatal care |
| H12A-H12X | Source of treatment for diarrhea |
| H32A-H32X | Source of treatment for fever/cough |
| V770A-V784X | Source of STI advice/treatment |

For these types of variables, there are some response categories such as government hospital or pharmacy that must be recoded to specific standard recode variables. In the example shown below for H12A-Z, source of treatment for diarrhea, these categories are shown in gray. All other response categories should be recoded in variables with the label prefixed with “CS”, making sure to map to the appropriate type of facility or source of treatment such as Public Sector, Private medical, and other sector. The label for each variable should be replaced with the label indicating the specific source. Note that all source labels should be translated into English. When mapping multiple response country specific categories to standard variables, they should be assigned the same order as the sources given in the country questionnaire. An example of this mapping is given in section 2.5 above using the example of variables M57A-Z in the 2016 Jordan DHS.

In some surveys there may be too many response categories within a given sector or group to fit within the standard variables available. In this situation, do not put the extra response categories in a standard variable that is labeled for another sector. Instead, use one of the available variables in the correct sector to combine several of the additional response categories into one category, and then preserve the separate response categories as country-specific variables. Document this both in the label for the variable used for combining several categories and in the country recode document.

|  |  |  |  |
| --- | --- | --- | --- |
| **H12A** | Diarrhea: government hospital | **H12N** | Diarrhea: CS private medical |
| **H12B** | Diarrhea: CS public sector | **H12O** | Diarrhea: CS private medical |
| **H12C** | Diarrhea: CS public sector | **H12P** | Diarrhea: CS private medical |
| **H12D** | Diarrhea: CS public sector | **H12Q** | Diarrhea: CS private medical |
| **H12E** | Diarrhea: CS public sector | **H12R** | Diarrhea: CS private medical |
| **H12F** | Diarrhea: CS public sector | **H12S** | Diarrhea: CS other sector |
| **H12G** | Diarrhea: CS public sector | **H12T** | Diarrhea: CS other sector |
| **H12H** | Diarrhea: CS public sector | **H12U** | Diarrhea: CS other sector |
| **H12I** | Diarrhea: CS public sector | **H12V** | Diarrhea: CS other sector |
| **H12J** | Diarrhea: private hospital/clinic | **H12W** | Diarrhea: CS other sector |
| **H12K** | Diarrhea: private pharmacy | **H12X** | Diarrhea: Other |
| **H12L** | Diarrhea: private doctor | **H12Y** | Diarrhea: no treatment |
| **H12M** | Diarrhea: CS private medical | **H12Z** | Diarrhea: medical treatment |

## Modifying the Consistency Applications

### Introduction

The CONSISx application for the woman questionnaire and the MCONSISx application for the male questionnaire can also be found on Sharepoint in the \RECODE folder. The consistency applications are used to check the consistency between variables in the recode data files. Run for the first time, the error-message listing can be very large, so it should only be run for 1% of the file at the beginning. Each message should be looked at and be addressed in one of four ways:

1. The message signals a genuine mistake in the recode data set and thus the recode application has to be adjusted and run again.
2. The message indicates a problem in the raw data which needs to be adjusted (for example a skip error in the data entry application), and run again. Note that all issues identified in the raw data set must be corrected first in the raw data before running the recode application again.
3. The message indicates a mistake in the CONSISx application. Variables might not be available in the recode data set and should thus be excluded from the CONSISx application or country specific variables are not included in the CONSISx and should thus be added. Adjust the CONSISx application and run it again.
4. The message signals exceptional or remarkable categories for variables or answers to a combination of variables. These should be documented in the recode documentation under ‘Notes on the Recode File’. All messages in the Level 0 postproc, with message numbers higher than 20000, belong to this category. If the message indicates a genuine inconsistency in the data file, not due to the recode application, then this needs to be documented under ‘Inconsistencies in the data file’. After the recode documentation is adjusted the corresponding error-message can be braced out.

It is thus an iterative, and sometimes time consuming, process to “clean” the recode and CONSISx applications and to add explanatory text to the recode documentation.

It is suggested to start at the listing by looking at the message summary at the end of the listing. It can be helpful to use a text editor or Excel to sort the messages in reverse order of frequency. Initially, focus on those messages that are generated for just about all of the cases – these are almost certainly problems due to not applicable variables or other aspects of the CONSIS logic not fitting the design of the specific survey,. After modifying the CONSISx logic for those messages, rerun the application for a slightly larger percentage of the cases.

### Preparing the A2Q File for the CONSISx Application

To simplify the work of modifying the CONSISx application to remove or adapt logic to address unused standard variables, the A2Q utility should be used to highlight unused variables in the CONSISx file.

To do this, take the list of unused variables from the final output produced by the *nasearch.exe* utility. Use a text editor to create a second column as shown below, renaming each variable with the suffix \_NA.

HV026 -> HV026\_NA //NA- Place of residence

HV028 -> HV028\_NA //NA- Household weight for male subsample (6 decimals)

HV030 -> HV030\_NA //NA- Field supervisor

HV031 -> HV031\_NA //NA- Field editor

HV027 -> HV027\_NA //NA- Office editor

Run the A2Q on the CONSISx application, then, go through the application searching for the \_NA string to identify unused variables. Modify or remove logic as appropriate.

### Running the CONSIS Application

The messages produced that are numbered over 20000 are due to a ‘kount’ array that is initialized at 0 in the preproc of the dictionary and incremented each time an error is identified, without producing an individual error message. The processor can create a second array “kount2” and handle it the same way as the kount array for ‘errors’ to count errors without producing individual messages. Number the messages 22000 + the index for the message to be produced from the kount2 array.

Following is an overview of the steps and general guidelines in preparing and running the CONSISx application:

1. Check the CONSISx application and modify any functions or constants that were modified in the RECODEx app, such as any educational logic or type of methods logic.
2. Create a second ‘KOUNT’ (KOUNT2) array to count errors that have been flagged as for documentation.
3. Remove all CONSISx logic for records not in the country recode application.
4. Do not remove CONSISx logic just because it is difficult to determine why a given message is being generated. It should be possible to identify all problems and either resolve or document them.
5. Do NOT spend time reading through the entire app. Just run it for about 2 seconds. The listing will be long –and the number of households, women and/or men in the listings will point to logic that needs to be modifed. As the logic becomes more finely tuned, the CONSIS application can be run for longer periods of time, and the the ‘KOUNT2’ counter can be used to eliminate error listings case by case.
6. Each message has a descriptive comment beside it in the CONSISx logic describing the reason for the message. It is important to carefully read these comments when investigating a given message.

# Recode Documentation

A country recode document must be produced for each survey. A template for the recode documentation can be found on Sharepoint in the \Recode folder. Rename this word document file ccIRvv.DOC, where cc is country code, and vv is recode version.

When this file is loaded the following template information is requested: country, country code and initials of DP person producing the documentation. This information is used in the title and the last update field at the end of the file.

Notes on completing the country recode document:

1. On the cover page substitute the questionmarks (????) with the relevant information and add additional information (e.g. other data files used in this survey). Include unweighted numbers for all households, completed households, all de-facto women, completed de-facto women, births in last 60 months (including month of interview), all de-facto men, completed de-facto men, all women in domestic violence module (if applicable). Also included the weighted ns for completed households, completed de-facto women, births in last 60 months, completed de-facto men, completed women in domestic violence (if applicable).

If there is a country specific population in the survey, it is wise to put the weighted and unweighted number of those populations also. It gives users a number with which to check their results. If the survey is an MIS, include the number of ‘last’ children found in the maternity and child health records in place of the births in last 60 months and be sure to label it correctly.

1. Under **“Sections and Variables Unused”**, List the NA variables in 5-6 columns (neatly) in order.
2. Likewise, print a listing of the country specific sections of the dictionary and insert them under the **“Sections Added”** portion of the documentation. List the country specific variables in ascending order by Record with complete labels (try to avoid abbreviations). The labels should be clarified and there should be no abbreviations in this part of the documentation.
3. The **“Notes on the Recode File”** and **“Inconsistencies in the Data File”** can be completed when the CONSIS applications are considered complete. Be sure the items listed in these 2 sections are listed in the order in which they are found in the dictionary. Review the CONSISx output and organize the comments so that they are orderly – that is, the HH comments begin with HV0….variables and end with the last variable at that level. The women’s comments should begin with V0…., followed by comments for V100s, V200s, etc. The same applies to the men’s comments or notes.
4. To save time, the text of notes and inconsistencies can be adapted from a pre-existing consistency document, with just the number of cases affected being changed.
5. If the value set for for any standard variables was changed, be sure to put a note in the document about this.
6. It is helpful to users to include information regarding the recoding of education and occupation variables. The **“Coding Additions”** should contain all standard variables with additions to the standard codes and all variables where the recode dictionary does not list all categories used. Since the dictionary now holds up to 1000 value labels, this might no longer be needed.
7. Documentation for the male recode should generally have the same contents and should follow the documentation for the household and individual recode.

When changes are made to the recode data file after earlier versions of this data set have been distributed then a data alert should be produced.

# Frequency Checking

## Introduction and Overview

Checking the recode is manually intensive and requires hand checking of frequencies from the recode data file with frequencies from the raw data file. It is thus necessary to produce unweighted frequencies of all variables in the raw data file, with frequencies of individual methods in the contraceptive section and frequencies of only the children born in the last 3 or 5 years before the survey in the health section. Furthermore, the frequency must be restricted to completed interviews for de-facto respondents. Standard applications exist both for the raw and recoded data sets and are described in more detail in section 6.3 below.

All DHS and MIS standard recode data sets **must** undergo two frequency checks prior to final tables being produced. The first frequency check is done by the programmer responsible for the survey and must be completed and all outstanding issues resolved prior to running tabulations for the Final Report. The second frequency check is to be performed by another data processing person and must be completed prior to tables being finalized, that is, sent to Publications for formatting.

Note that when the recode is provided to a second data processing person for review, the recode should be considered in its final state by the country data processor, with documentation completed, all unused variables marked ‘-NA’, all country specific variables included, and all outstanding issues identified by the CONSISx application run resolved.

## Guide to Production of Frequencies for Checking

The standard applications for checking the recode and raw frequencies have been reorganized so that women’s variables, including the birth history, sibling history, and domestic violence are produced before the variables limited to births of children < 60 months ago. In addition, the IQFREQx application in the \Prelim directory has been adjusted so that the anthropometry data collected at the household will appear for the children in the woman’s questionnaires along with all the other children variables. Household characteristic variables that are standard are also included in the woman’s frequencies. The data processor should add any country-specific household characteristics to the IQFREQ application at the woman’s level. Be sure to use the *breakdown* keyword with any alpha variable frequencies added to the application. Use the *disjoint* keyword only for sub-items when it is necessary to distinguish between the occurrences (for example, Q301/V301).

The applications HRFREQ and IRFREQ to produce frequencies for the household and individual recode data files can be found on Sharepoint in the \RECODE directory. If REC83 and REC84 (maternal mortality) and REC85 (AIDS module) are used in IRFREQ then delete the braces surrounding these records. Do the same for the country specific sections in REC97, REC98 and REC99. When there are additional country specific sections then these must be added to the frequency application as well. Use the application MRFREQ in the same directory to produce frequencies for the male recode file. Likewise, if MREC83 and MREC84 (maternal mortality) and MREC85 (AIDS module) are used in MRFREQ then delete the braces surrounding these records. Do the same for the country specific sections in MREC97, MREC98 and MREC99.

The following guidelines should be followed when producing frequencies for checking:

1. Frequencies must be parallel and for the same populations. They must include the cover page variables (raw) for all households and cover page of individuals along with any HH cover page variables being brought to individual level for all defacto women and men.
2. Variables run with nofreq stat in the raw must be run the same way in the recode.
3. Alpha variables in the raw must be run ‘breakdown’ to facilitate the checking of the VnnnA-Z variables.
4. Be sure to include a NOFREQ STAT frequency for any variables excluded from the normal frequency command for each record. Be sure to be consistent in the use of NOFREQ STAT between the raw and recode so all variables can be checked.
5. There should be a frequency for all households, all defacto women, and all defacto men for all variables that should exist (questionnaire cover) whether the result is complete or not. Check that incomplete households, women and men have weights of 0 and that they have data for country specific variables that come from the cover page of each questionnaire.
6. Be sure to include frequencies of the 1-digit wealth index, heights, weights and sample weights for HH, men, women and DV weights or other weights that will be in the recode. Use ‘nofreq stat’ for the anthropometry statistics and any variables (not weights) that are 5-12 digits in length.

## Procedures for Checking Frequencies

When the raw and recode frequencies are passed on to a reviewer, the documentation must be complete and given to the reviewer, together with the questionnaires. The reviewer must perform the following checks:

1. Check that no standard variables are missing from the recode.
2. Check that all country specific variables are in the standard recode and are documented.
3. Check frequencies of constructed variables as well as simple variables; if necessary use cross-tabulations or other applications (e.g. original recode applications) to do this.
4. Check that spell checking of documentation is performed.
5. Check grammar of documentation.
6. Check that the documentation follows the standard format.

The programmer should perform these same checks before all materials are handed to the reviewer. In addition, the programmer should make sure that:

* There will be no stars in the exported data, caused by NotAppl not being defined. To make sure that this is not the case, run a simple EXPORT on the recode data files and check the exported data file for stars (“\*”).
* There are no default values in the frequencies, caused by NotAppl not being defined.
* There are no ampersands (@) in the frequencies, caused by out of range categories in the raw and recode dictionaries.

**Detailed List of Frequency Check Procedures**

The following checks should be performed when checking frequencies both for the first and second review.

1. Checking the variables means checking that numbers connected to the value labels from raw to recode. The numbers should remain the same in both sets – meaning that care needs to be taken to select appropriate cases for variables when the raw data have more cases than will end up in the recode file. This is especially true for sections of questionnaires that need to be limited to births within certain periods.
2. Watch for missing labels on values that should have labels. English language and spelling errors should be checked in in variable and value labels. For recode variables that are constructed from 2-part questions (unit/number), ensure that the minimum value in the frequencies for each unit (100s, 200s, 300s, etc) has a label (ex: 105 Days:5) and that all ‘special type values’ (996, 997, 998, etc) have labels.
3. Sampling strata values must have labels
4. If a variable is listed as NA- in the doc, then NA- is at the front of the variable label in the dictionary and that all cases are not applicable. Also check that NA cases in variables that have values are correctly NA
5. Make sure that any labels in the men’s frequencies are sex-appropriate if dictionary entries have been copied from the woman’s dictionary .
6. Check for any raw/recode variable/value labels that are totally incorrect or incomplete based on the questionnaire.
7. Carry out the following checks for birth history variables:
8. check B0 (child is twin). The numbers for codes 1 and 2 should be equal.
9. B11 and B12 should not have any negative birth intervals this is a problem of imputation and must be corrected.
10. In the MIS, the birth history is truncated. If the questionnaire says births must have occurred after a particular date, check the cmcs of birth (B3).
11. Check the total numbers in the households of ‘eligible women’, ‘eligible men’ and ‘children < 5’ against the numbers in the frequencies by multiplying the number of households for each of the values and adding them up.
12. Make sure the total numbers for each population match from raw to recode – households, HH members, Anthropometry, any other country specific multiple record or group, women, birth histories, maternity, child health, anthropometry, maternal mortality, domestic violence, men.

# Preparation of Data Files and Documentation for Distribution and Archiving

Before copying the files to P:\DISTRIB, do the following checks in the file DistribCheck.docx, found in P:\DHS\Projects\Distrib\!Data Processing.

PRE-CHECK BEFORE SUBMITTING DATA FOR DISTRIBUTION

DHS/AIS/MIS/SPECIAL

|  |
| --- |
| ccIRvv.DCF - Set multiple **Household** records in the recode dictionary to their true maximum.  *Do not change the maximums of the Women records!*  ccIQvv.DCF – Set multiple records in the raw dictionary to their true maximum. |
| Add the Country Name, Year to all dictionary labels. |
| Internal names of IR and MR Dictionaries must read: RECODEn and MRECODEn respectively. |
| Are there 3 relations in the IR dictionary: PERSONS/CHILD5/ALLCHILD? |
| ccIRvv.DCF Delete unused non-standard Household Records **after RECH6.**  *If RECH5 & RECH6 are unused, do not delete - set MAX to 2*  ccIRvv.DCF Delete unused non-standard Women’s Records **after REC81.**  ccMRvv.DCF Delete unused non-standard Men’s Records **after MREC80.** |
| Run Fieldwkcheck.app on FieldWorker data.  Are the correct labels for the actual languages (FW113\*) present in the Field Worker dictionary?  Did you sort the fieldworker data by ID?  If you have done the above 3 steps for the fieldworker data, please send confirmation email to Bridgette. |
| Run country specificDataExplorer to check for stars in the data. |
| Are occupation codes in the raw and recode dictionaries (and not in the recode documentation)? |
| Run the utility in [VarValueOverlap](file://ICFI.icfconsulting.com/ORG/SAS/IHD/ISRE/DHS/Projects/Distrib/Data%20Processing/VarValueOverlap) to ensure that you have no variable valuesets overlapping. |
| Check for out of range values (search for the “@” symbol in IR and MR freqs) |
| Clean/update all \Data Processing directories in Project folder on SharePoint (or P:\). |
| Submit a signed PII Review Form by email to Bridgette (PII forms are on SharePoint). |
| Copy the documentation and the following folders to: [**P:\DHS\Projects\Distrib**](file://ICFI.icfconsulting.com/ORG/SAS/IHD/ISRE/DHS/Projects/Distrib)**.**  ..\FinData final raw/recode dat (HIVTest, OtherBiomarkers, FieldWorker - if applicable)  ..\Dicts final raw/recode dcf (HIVTest, OtherBiomarkers, FieldWorker - if applicable)  ..\Tables please ensure that all table apps compile successfully (include \*.rtf)  ..\Library  The above folders will be sent to the Implementing Agency, along with the distribution datasets. All apps must compile without errors. |

**Couples Note**: *Only women and men whose data for variables V034 and MV034(N) are consistent are matched in the couples data. If this is not the case here, please provide the logic for exporting the “couples”.*

**RECHML Note:** *Did you add Country Specific (country specific) RECHML2 in the DCF? If yes, please add to this record: HML? - “Line number of person who slept in this net”. This is neeeded during exports, for attaching to every person the characteristics of the net they slept in. Another option is to simply add these country specific variable(s) at the end of RECHML (and delete RECHML2).*

CHECK BEFORE ARCHIVING SURVEY DATA

Clean the SharePoint folder but leave the following information for archiving:

1. **ccIRvv.DIC** The dictionary for the recode should contain only the following sections:

Household: RECH0 – RECH2

RECH1 (RECH4) – please change the maximum number of occurrences from 90 to the maximum for this survey. Do the same for RECH4 (if applicable).

RECH3 – RECH6 (country specific sections)

Individual: REC11 – REC81

REC82 – REC85 (if applicable)

REC91 – REC99 (country specific sections)

**ccMRvv.DIC** The dictionary for the male recode should contain only the following sections:

Individual: MREC11 – MREC75

MREC82 – MREC85 (if applicable)

MREC91 – MREC99 (country specific sections)

2. **ccIRvv.DAT:** The final standard recode data file

1. **ccIRvv.DOC:** The country documentation in WORD format (contains male recode if applicable).
2. **IRFREQ, HRFREQ and MRFREQ** applications and attribute files with the correct country name and sections.
3. In a subdirectory RECODE all files related to the recoding process, such as:

**RECODEx.\* MCONSISx.\***

**MRECODEx.\* MICONSISx.\***

**CONSISx.\* HHSMALL.DIC**

**IRSMALL.DIC IDSMALL.\***

1. In separate subdirectories all other files used during data processing, e.g:

* COUNTRY: all files in the root directory or this country
* ALLWOMFACT: all files related to creation of all woman factors (if applicable)
* DATA: all versions of the raw and recode data files
* DOcountry specific: all documents produced for data processing
* EDITING: all files used during data editing
* ENTRY: all files used during data entry
* LIBRARY: all files in the library directory
* OTHERS: all files used during recoding of “OTHER” answers
* PRELIM: all files used during production of preliminary tables
* RECODE: all files as listed under 7
* SAMPLE: all files used for sample selection
* TABLES: all files used for the production of the final tabulations

1. If there are other data files (Service Availability, Health etc.) put them in different sub-directories.
2. **ccIQvv.DIC:** Last version of the raw dictionary. The version should correspond to the last version of the raw data file. Make sure that any multiple section in the household dictionary is reset to its true maximum to conserve space in the exported files. Do the same for the birth history and maternal mortality sections in the individual questionnaires (if exists). The final raw data **ccIQvv.DAT** should also be made available. Be sure to include any changes made during the recode process in the raw data file. Basic information in the raw and recode data files should be equivalent.
3. Guidelines for the production of the raw data files. The information provide should contain:

* The number of completed cases for the household, individual woman, male and service availability surveys if applicable.
* The selection criteria used for the export file. Select only completed interviews (skip cases with QHRESULT, QRESULT or QMRESELT <> 1). Indicate how to select the correct survey (e.g. skip cases when QQTYPE <> 2 for female survey etc.). Restrict case selection to de-facto respondents (e.g. skip cases when QH05(QLINE) <> 1).
* List section names used in the respective surveys.

# Archive All Available Country Documentation

After the country report is published all available written country documentation must be transferred to the Central Files. These files can be found in the hallway outside room 484. Confer with the country monitor to make sure that one copy of the following documentation is transferred to these files:

* All questionnaires used in the survey (if possible all language versions).
* All manuals and information provided by the sampling specialist.
* All manuals used during the training of the fieldwork supervisors, editors and enumerators (usually provided by country monitor).
* All forms used during fieldwork.
* All manuals and forms used during data processing (one copy of the Data Entry Guidelines, Office Editing Guidelines, Data Entry Supervision and Editing Guidelines).
* All country specific coding instructions (e.g. brand names of contraceptives, occupation categories etc.).

# Recode Specifications for Malaria Indicator Surveys (DHS7)

The Malaria Indicator Survey (MIS) recode file is a subset of the DHS recode file. A decision was made not to have a separate recode application mainly for maintenance purposes and because there may be instances when a DHS module that is not part of the MIS core questionnaire is applied in a MIS survey. The following considerations should be taken into account when producing a recode for a MIS survey:

* In general, the same principles as those applied to a DHS recode construction are applied to a MIS recode file.

* Recode file names (data and dictionary) will follow the same naming conventions as those of a DHS. The mask name is: CCIRXX, where CC is the country code, IR is a constant for standard recode, and XX represents the recode phase definition and the version of the data file (for instance, 7th phase version 1 or 7th phase version A).
* There are some mandatory records in a malaria recode dictionary and variables in that dictionary have exactly the same meaning as those in the DHS standard recode dictionary. The mandatory records are:

|  |
| --- |
| HOUSEHOLD |
|  |
| RECH0 Household's basic data |
| RECH1 Household Schedule |
| RECH2 Household Characteristics |
| RECH6 Children Height/Weight/Hemoglobin |
| RECHML Malaria: by Mosquito Bed Net |
| RECHMH Malaria: by Household Member |
|  |
| WOMAN |
|  |
| REC01 Respondent's basic data |
| REC11 Respondent's basic data (continued) |
| REC21 Reproduction and Birth History |
| REC22 Reproduction (continued) |
| REC41 Maternity |
| REC42 Health and Breastfeeding |
| REC4A Child's health |
| REC44 Child's Height and Weight |
| REC81 Characteristics of Interview |
| RECML Malaria |

* Household records RECH5 and/or RECHMA should only be included if either height and weight, hemoglobin or other biomarkers are collected for women and/or men. If that is the case only the relevant variables to the biomarkers collected by the survey should be included in the record
* If a variable is not a standard MIS recode variable (according to the variable list provided below on *List 2*), but it is a DHS recode variable, the DHS recode variable name should be kept. The variable should be placed in the same record where it is defined for DHS. If the variable is part of a mandatory MIS record, it should be placed at the end of the standard variable definitions. If a DHS module is collected as part of a MIS survey, the record(s) where the module is defined in the recode should be kept. For example, if a maternal mortality module is used in a MIS survey, records REC83 and REC84 should be present.
* Variables not used by a particular MIS survey but that are part of the standard MIS recode definition, should be deleted, except **(a) variables that are part of a multi-category variable**, i.e., VxxxA-Z and **(b)** **all variables in records RECH0, RECHML, RECHMH, REC01, and RECML must be kept**. For non-used variables described on exceptions (a) and (b), the variable label should be marked with “–NA”. For this purpose, use the DP utilities “NAvars” and “NAreplace”. NAvars identify not applicable variables based on the output of a recode frequency distribution whereas NAreplace assigns the characters “–NA” to the variable labels in a CSProdictionary.
* Any other variable not defined as part of the MIS recode definition and that is neither a DHS variable, should be kept as a survey specific variable. MIS survey specific variables follow the same naming conventions as those of a DHS recode. Essentially, names for survey specific variables from the household questionnaire should start with the letters SH and will be followed by the household question number. Names for survey specific variables from the biomarker questionnaire should also with the letters SB and will be followed by the biomarker question number. Names for survey specific variables from the individual questionnaire should start with the letter S and will be followed by the individual question number. Country specific variables for multiple records can be placed at the end of the standard variables in the multiple record. This can be done because the record length varies depending on the data collected by the survey. This is even desirable for survey specific variables in the net records (RECHML) where the net characteristics are associated to all individuals sleeping under that net. Survey specific single variables should be stored in survey specific records RECH3 and REC91 for household and individual levels respectively.
* The birth history in a Malaria Indicator Survey is normally constrained to births in the last five years. However, imputation is required and should include all DHS events that are present in the MIS survey. MIS standard events are: Respondent’s date of birth, children’s date of birth (constrained to births in the last five years), current pregnancy, and date of interview. If the date of marriage and the date of sterilization are included in the survey, they should also be imputed. The CSProbatch secondary editing application “ccEditx” should be used for this purpose. Again, cc represents the country code and x the recode phase.
* Make sure to properly link records REC41, REC42, REC4A, and REC44 to the birth history (REC21). Variables in the Maternity record (REC41) in MIS surveys are collected only for the last birth. Variables related to the Child’s health (REC4A) and the Child’s height and weight (REC44) records in MIS surveys are collected for all births in the last five years, similar to DHS.
* Non mandatory records in MIS where all variables are NA should be removed.
* An strategy to construct a MIS recode application will be as follows:
  + Run A-to-Q on the RECODEx application as input, where x is the recode phase
  + Remove MIS standard variables not used by the particular MIS survey, except if **(a)** **the variable is part of a multi-category variable**, i.e., VxxxA-Z, or **(b)** **if the variable belongs to any of these records: RECH0, RECHML, RECHMH, REC01, and RECML**. In the case of exceptions (a) and/or (b), assign “–NA” to the variable label of the MIS standard variable
  + Remove from the application those variables that are not part of the MIS, but leave those that despite the fact they are not part of the standard MIS, they are still DHS standard variables
  + Add survey specific variables and logic to the dictionary and application, respectively. Run and check the recode and its frequencies
  + Run utilities NAvars and NAreplace

**List 1. Standard records in a MIS recode file**

Following is a list of standard records that should be present in any MIS recode file. However, if there are no survey specific variables for a given type of record, the record should not be present (i.e. deleted). Likewise, RECHMA for Men height/weight/hemoglobin should be deleted if no biomarker were collected for men. It is important to note that the record length for the standard MIS variables portion is fixed. It could be longer if standard DHS variables were added after the standard portion. All other record characteristics should be kept the same as they are defined here (Type value, whether required or not, and number of occurrences). The number of occurrences for certain records (household schedule, etc.) should be adjusted at the end, the same way it is done for DHS surveys.

# Appendix A. Correspondence of Raw Data Dictionary Sections to Recode Data File Sections

|  |  |  |  |
| --- | --- | --- | --- |
| **Household questionnaire** | | | |
| **Record name in standard raw dictionary** | **Description of content** | **Record name in recode dictionary** | **Unit of analysis** |
| AHSECOVER | Household identification | RECH0, RECH3 | Household |
| AHSEC01 | Household schedule | RECH1 index = HVIDX, RECH4 index = HVIDX4 | Household member |
| AHSEC02 | Household characteristics | RECH2, RECH3 | Household |
| AHSEC03 | Mosquito nets | RECHML index = HMLIDX, RECHML2, index = HMLIDX2 1 | Mosquito net |
| AHSEC03 | Mosquito nets | RECHMH index = HMHIDX, RECHMH2, index = HMHIDX2 2 | Household member |
| AHSEC04 | Additional household characteristics | RECH2, RECH3 | Household |
| ABSECOVER | Biomarker identification | RECH0, RECH3 | Household |
| ABSEC01 | Biomarker children | RECH6 index = HC0, RECH6S index = HCA | Household member |
| ABSEC02 | Biomarker women | RECH5 index = HA0, RECH5S index = HAA | Household member |
| ABSEC03 | Biomarker men | RECHMA index = HBA, RECHMS index = HBA | Household member |
| AHSECcountry specific3 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Individual (woman) questionnaire** | | | |
| **Record name in standard raw dictionary** | **Description of content** | **Record name in recode dictionary** | **Unit of analysis** |
| ASECOVER | Woman identification | REC01, REC91 | Woman |
| AWSEC01 | Woman’s background | REC11, REC91 | Woman |
| AWSEC2A | Reproduction | REC22, REC91 | Woman |
| AWSEC2B | Birth history | REC21 index = BIDX, REC92 index = IXD92 | All children ever born |
| AWSEC2C | Calendar | REC82 1 | Months 5+ years prior to interview |
| AWSEC2D | Reproduction (continued) | REC22, REC91 | Woman |
| AWSEC3A | Contraceptive knowledge and use | REC31, REC91 | Woman |
| AWSEC3B | Contraceptive practice | REC32, REC91 | Woman |
| AWSEC04 | Pregnancy and postnatal care | REC41 index = MIDX, REC94 index = IDX94 | Children 0-59 months old |
| AWSEC05 | Child immunization | REC43 index = HIDX, REC95 index = IDX95 | Children 0-35 months old |
| AWSEC6A | Child health and nutrition | REC4A index = HIDXA, REC95 index = IDX95, RECML index = IDXML | Children 0-59 months old |
| AWSEC6B | Child health and nutrition (continued) | REC42 | Woman |
| ABSEC01 2 | Child height and weight | REC44 index = HWIDX, REC96 index = IDX96 | Children 0-59 months old |
| AWSEC07 | Marriage and sexual activity | REC51 | Woman |
| AWSEC08 | Fertility Preferences | REC61 | Woman |
| AWSEC09 | Husband’s background and woman’s work | REC71 | Woman |
| AWSEC10 | HIV/AIDS | REC75, REC80 | Woman |
| AWSEC11 | Other health issues | REC42 | Woman |
| AWSECMM | Maternal mortality | REC83 index = MMIDX, REC84 | Woman |
| AHSEC03 | Mosquito nets | RECML |  |
| AWSECDV | Domestic violence | RECDV | Woman |
| 1 Country specific variables in a standard calendar column or country specific columns are to be presented in columns 6 thru 9 in record 82.  2 The variables in REC44 are derived from the variables in record RECH6 of the household questionnaire. | | | |

# Appendix B. Variable Naming Conventions

|  |  |  |
| --- | --- | --- |
| **Recode Data File** | | |
| **Questionnaire or Section** | **Section** | **Variable Prefix** |
| **Household** | All (except those listed below) | HVxxx |
| Survey specific household | RECH3 | SHVxxx |
| Survey specific household roster | RECH4 | SHVxxx |
| Woman height/weight/hemoglobin | RECH5 | HAxxx |
| Child height/weight/hemoglobin | RECH6 | HCxxx |
| Men height/weight/hemoglobin | RECHMA | HBxxx |
| Malaria by mosquito bed net | RECHML | HMLxxx |
| Malaria by household member | RECHMH | HMLxxx |
| Disability module | RECHDIS | HDISxxx |
| Child labor | RECHCL2 | HCHLxxx |
| Child discipline | RECHCDIS | HCDIxxx |
| **Woman Questionnaire** | All except those listed below | Vxxx |
| Reproduction and birth history | REC21 | Bxxx |
| Maternity | REC41 | Mxxx |
| Child’s vaccinations | REC43 | Hxxx |
| Child’s health | REC4A | Hxxx |
| Child’s height and weight | REC44 | HWxxx |
| Maternal mortality | REC83 | MMxxx |
| Maternal mortality (suite) | REC84 | MMCxxx |
| Malaria | RECML | MLxxx |
| Domestic violence | RECDV | DVxxx |
| Female Genital Cutting | RECG1 | Gxxx |
| Female Genital Cutting – Roster | RECG2 | Gxxx |
| Early childhood development | RECECD | ECDxxx |
| Country specific sections | REC91-99 | SVxxx |
| **Male Questionnaire** | All except those listed below | MVxxx |
| Female genital cutting | MRECGC | MGxxx |
| Country specific sections | MREC91-99 | MVxxx |

|  |  |  |
| --- | --- | --- |
| **Raw Data File** | | |
| **Questionnaire or Section** | **Section** | **Variable Prefix** |
| Household | All | QHxxx |
| Woman | All | Qxxx |
| Men | All | QMxxx |