SESUG 2023
SHARE • LEARN • CONNECT
- OCTOBER, 22-24 -
CHARLOTTE, NC

2023 Southeast SAS User Group (SESUG) Conference
October 22nd - October 24th, 2023, Charlotte, NC

Ensuring Accurate Data Linkages with Metadata Tables

Yeats Ye, Jessie Parker, Cindy Zhang, Cordell Golden

The findings and conclusions are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Background:

National Center for Health Statistics (NCHS) Data Linkage Program

- Links NCHS survey data with administrative records
- Expands the scientific value of the NCHS population-based surveys
- Uses personally identifiable information (PII) for survey participants to facilitate linkages
- Follows a secure and complex data storage and management process for maintaining PII
- Utilizes a series of relational SAS tables, known as the Record Linkage Repository (RLR) for data storage and management
- Uses metadata tables for data quality assurance

NCHS surveys used in linkages

National Health Interview Survey (NHIS):

- Nationally representative, cross-sectional household survey
- Serves as an important source of information on the health of the civilian, noninstitutionalized population in the US.

Since 1957

National Health and Nutrition Examination Survey (NHANES):

- Nationally representative, cross-sectional household survey
- Includes a household interview and examination component
- Data are released in 2-year cycles
- Serves as an important source of information on the health and nutritional status of civilian, noninstitutionalized adults and children in the US.

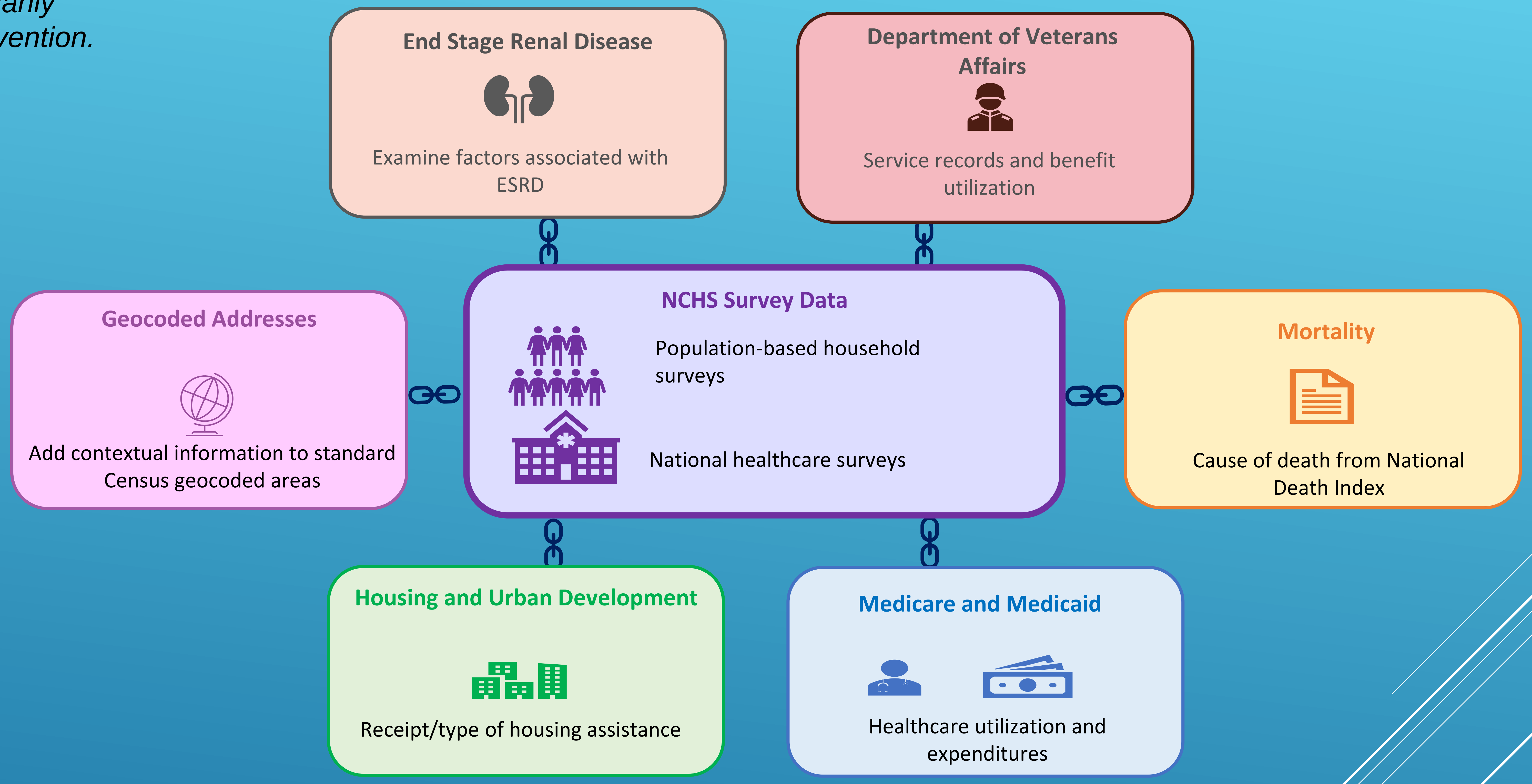
nhanes
The Nation's Mobile Health Survey

National Hospital Care Survey (NHCS):

- Collects data on patient care in hospital-based settings (inpatient, emergency, and outpatient departments) to describe patterns of health care delivery and utilization in the US.

NHCS
National Hospital Care Survey

Current NCHS Data Linkages



End Stage Renal Disease
Examine factors associated with ESRD

Department of Veterans Affairs
Service records and benefit utilization

Geocoded Addresses
Add contextual information to standard Census geocoded areas

NCHS Survey Data
Population-based household surveys
National healthcare surveys

Mortality
Cause of death from National Death Index

Housing and Urban Development
Receipt/type of housing assistance

Medicare and Medicaid
Healthcare utilization and expenditures

Record Linkage Repository (RLR)

- A series of relational SAS tables containing PII and other data collected for NCHS survey participants. Contains over 12 SAS tables for 60+ survey years/cycles, totaling 500+ datasets.
- Each SAS table contains at least one record for each survey participant
- IDS and SUBJECT tables contain a single record for each survey participant
- Other tables may contain multiple records per survey participant if alternate information (such as name or address) is provided during the initial data collection or obtained during a follow up data collection activity (survey follow up or linkage)
- Each survey participant is assigned a unique identification code that can be used to merge formation from multiple tables
- Survey participant data can be easily extracted from the RLR to facilitate entity-level record linkages

CHARLOTTE, NC

SESUG 2023

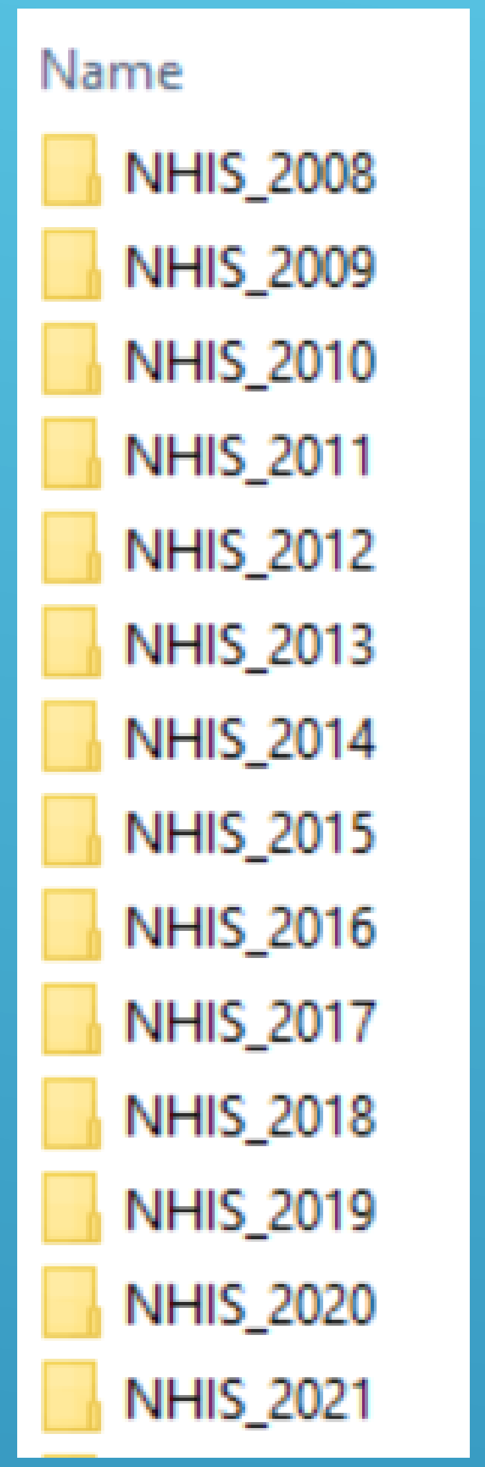
SHARE • LEARN • CONNECT

- OCTOBER, 22-24 -

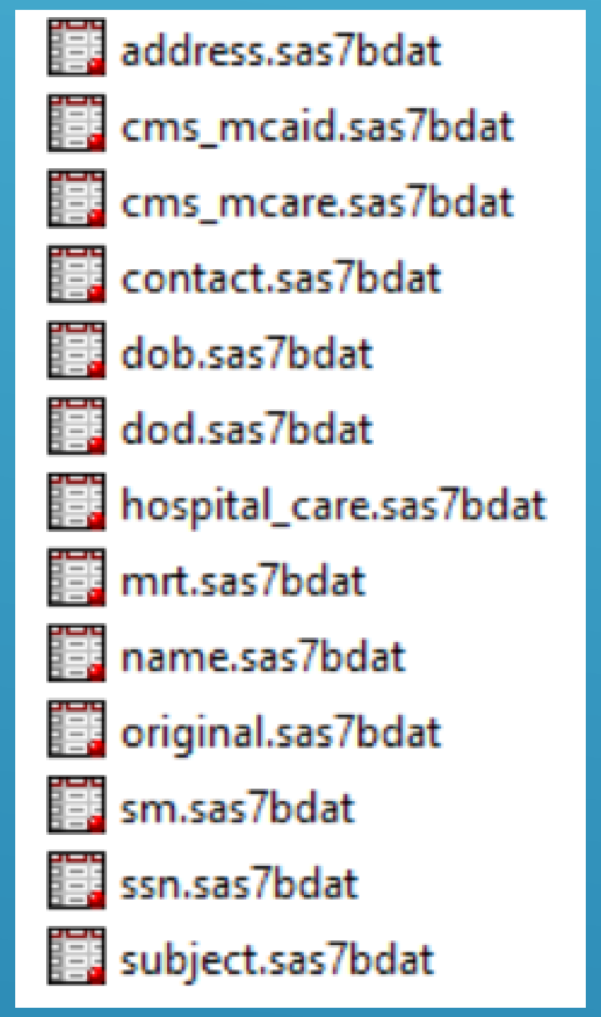
Monthly metadata tables are created with Proc Contents from RLR data.

We use the monthly metadata tables and comparison reports based on those tables to check the number of variables, number of observations, and the date on which a file was last modified to ensure the accuracy of all RLR data.

RLR folder structure



Name
NHIS_2008
NHIS_2009
NHIS_2010
NHIS_2011
NHIS_2012
NHIS_2013
NHIS_2014
NHIS_2015
NHIS_2016
NHIS_2017
NHIS_2018
NHIS_2019
NHIS_2020
NHIS_2021



address.sas7bdat
cms_mcaid.sas7bdat
cms_mcare.sas7bdat
contact.sas7bdat
dob.sas7bdat
dod.sas7bdat
hospital_care.sas7bdat
mrt.sas7bdat
name.sas7bdat
original.sas7bdat
sm.sas7bdat
ssn.sas7bdat
subject.sas7bdat

We use the metadata tables to ensure the accuracy of the number of observations for linked data products, such as our match rate tables. The publicly-available match rate tables report counts of successfully linked beneficiaries by survey year and age group to aid researchers in using our data.

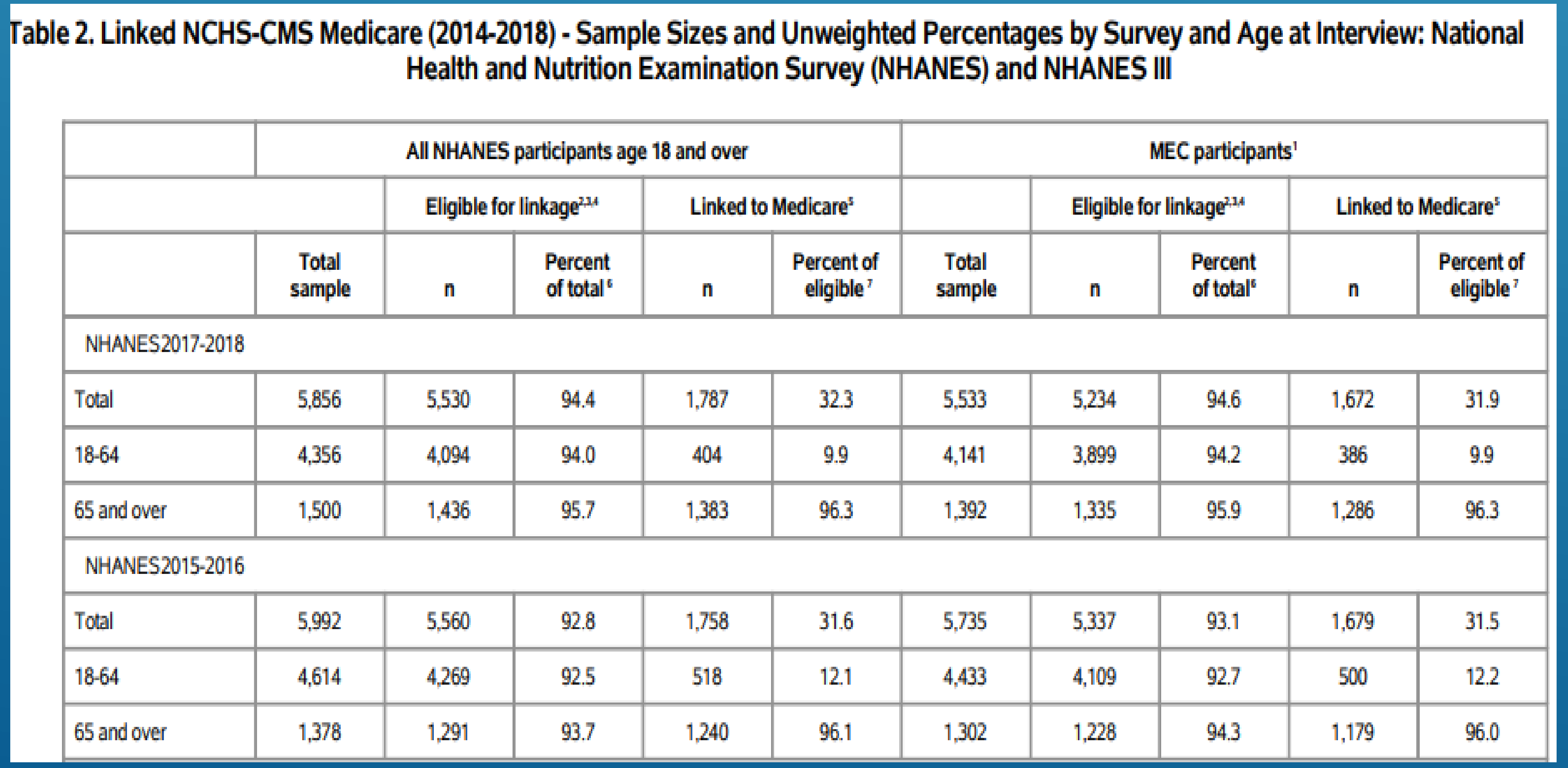


Table 2. Linked NCHS-CMS Medicare (2014-2018) - Sample Sizes and Unweighted Percentages by Survey and Age at Interview: National Health and Nutrition Examination Survey (NHANES) and NHANES III

	All NHANES participants age 18 and over					MEC participants[1]				
		Eligible for linkage[2,3,4]		Linked to Medicare[5]			Eligible for linkage[2,3,4]		Linked to Medicare[5]	
	Total sample	n	Percent of total[6]	n	Percent of eligible[7]	Total sample	n	Percent of total[6]	n	Percent of eligible[7]
NHANES2017-2018										
Total	5,856	5,530	94.4	1,787	32.3	5,533	5,234	94.6	1,672	31.9
18-64	4,356	4,094	94.0	404	9.9	4,141	3,899	94.2	386	9.9
65 and over	1,500	1,436	95.7	1,383	96.3	1,392	1,335	95.9	1,286	96.3
NHANES2015-2016										
Total	5,992	5,560	92.8	1,758	31.6	5,735	5,337	93.1	1,679	31.5
18-64	4,614	4,269	92.5	518	12.1	4,433	4,109	92.7	500	12.2
65 and over	1,378	1,291	93.7	1,240	96.1	1,302	1,228	94.3	1,179	96.0

Metadata tables and comparison reports

In example 1: The comparison report for February 2023 shows some changes from the previous month (January 2023)

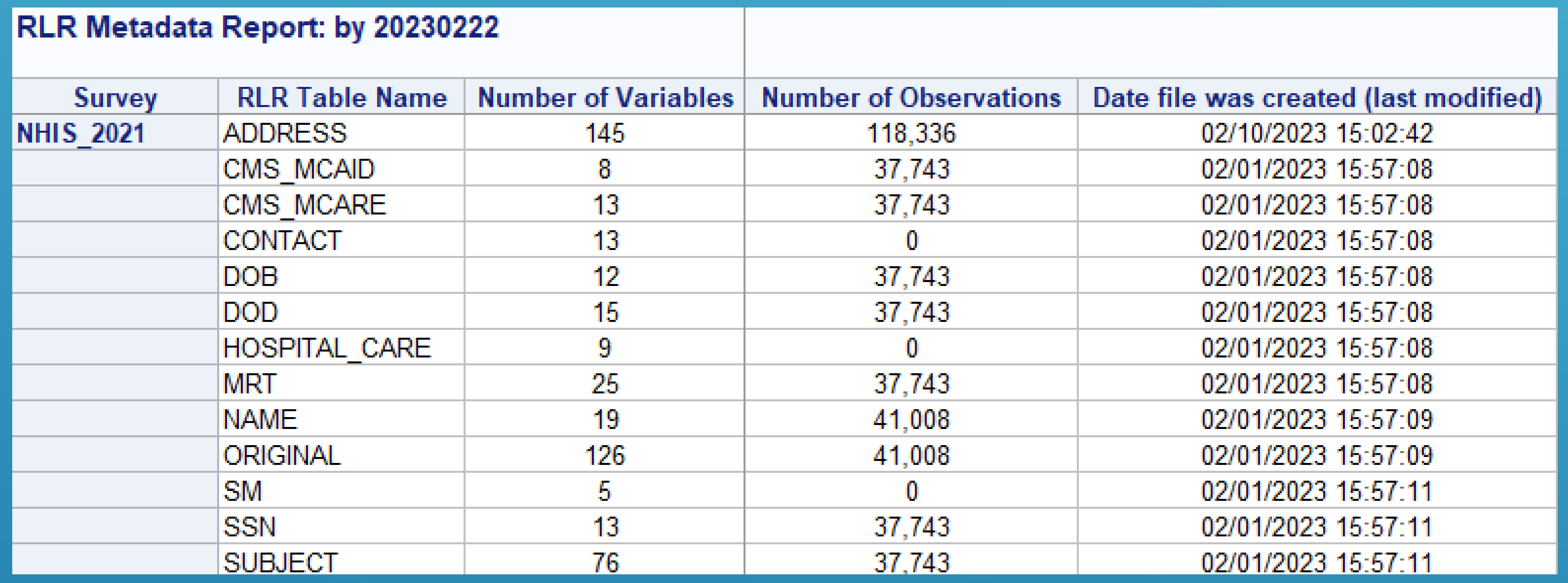
- Survey data from NHIS 2021 was loaded into the RLR this month (February 2023)

In example 2: The comparison report for May 2023 shows no changes from the previous month (April 2023)

In example 3: The comparison report for July 2023 shows some changes were made from the previous month (June 2023)

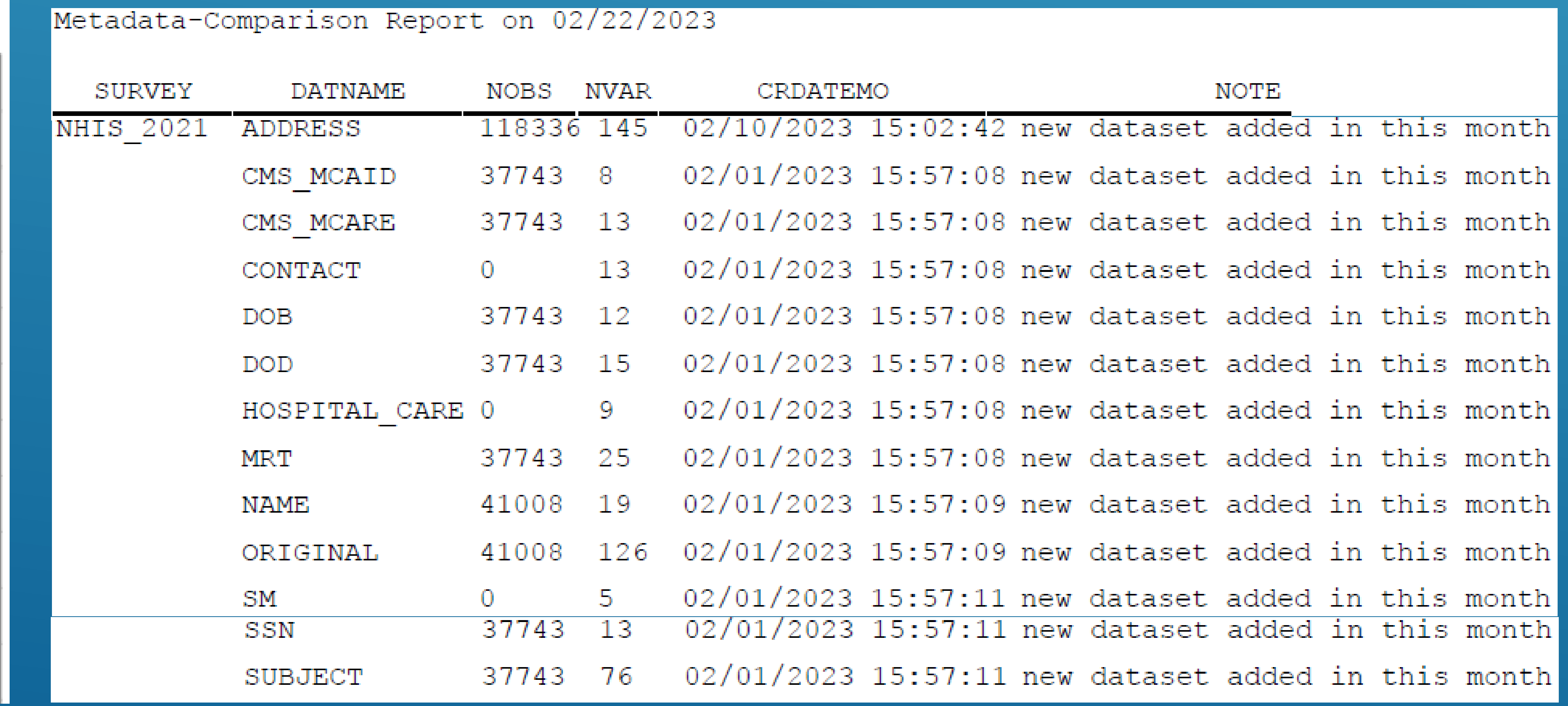
- NHIS_2019 SUBJECT table was updated on July 14, 2023
- NHIS_2020 SUBJECT table was updated on July 14, 2023

Example 1: Feb. 2023 metadata table and comparison report



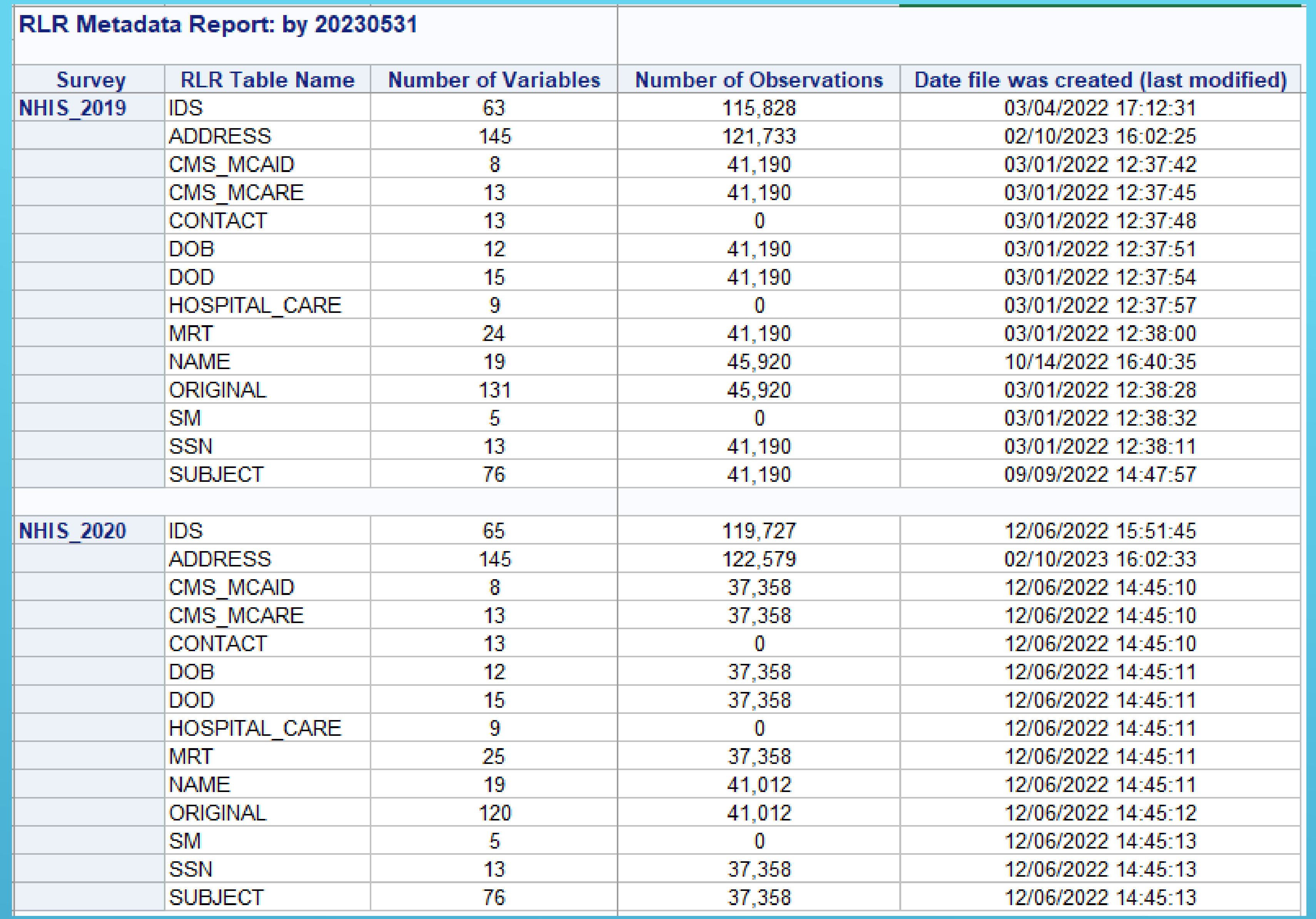
RLR Metadata Report: by 20230222

Survey	RLR Table Name	Number of Variables	Number of Observations	Date file was created (last modified)
NHIS_2021	ADDRESS	145	118,336	02/10/2023 15:02:42
	CMS_MCAID	8	37,743	02/01/2023 15:57:08
	CMS_MCARE	13	37,743	02/01/2023 15:57:08
	CONTACT	13	0	02/01/2023 15:57:08
	DOB	12	37,743	02/01/2023 15:57:08
	DOD	15	37,743	02/01/2023 15:57:08
	HOSPITAL_CARE	9	0	02/01/2023 15:57:08
	MRT	25	37,743	02/01/2023 15:57:08
	NAME	19	41,008	02/01/2023 15:57:09
	ORIGINAL	126	41,008	02/01/2023 15:57:09
	SM	5	0	02/01/2023 15:57:11
	SSN	13	37,743	02/01/2023 15:57:11
	SUBJECT	76	37,743	02/01/2023 15:57:11



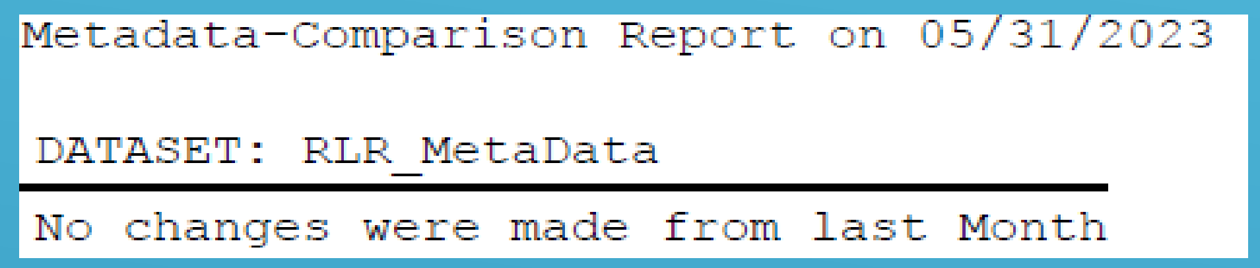
```
Metadata-Comparison Report on 02/22/2023

  SURVEY      DATNAME       NOBS  NVAR      CRDATEMO                 NOTE
NHIS_2021  ADDRESS        118336 145  02/10/2023 15:02:42 new dataset added in this month
           CMS_MCAID      37743  8    02/01/2023 15:57:08 new dataset added in this month
           CMS_MCARE      37743  13   02/01/2023 15:57:08 new dataset added in this month
           CONTACT        0      13   02/01/2023 15:57:08 new dataset added in this month
           DOB            37743  12   02/01/2023 15:57:08 new dataset added in this month
           DOD            37743  15   02/01/2023 15:57:08 new dataset added in this month
           HOSPITAL_CARE  0      9    02/01/2023 15:57:08 new dataset added in this month
           MRT            37743  25   02/01/2023 15:57:08 new dataset added in this month
           NAME           41008  19   02/01/2023 15:57:09 new dataset added in this month
           ORIGINAL       41008  126  02/01/2023 15:57:09 new dataset added in this month
           SM             0      5    02/01/2023 15:57:11 new dataset added in this month
           SSN            37743  13   02/01/2023 15:57:11 new dataset added in this month
           SUBJECT        37743  76   02/01/2023 15:57:11 new dataset added in this month
```

Example 2: May 2023 metadata table and comparison report



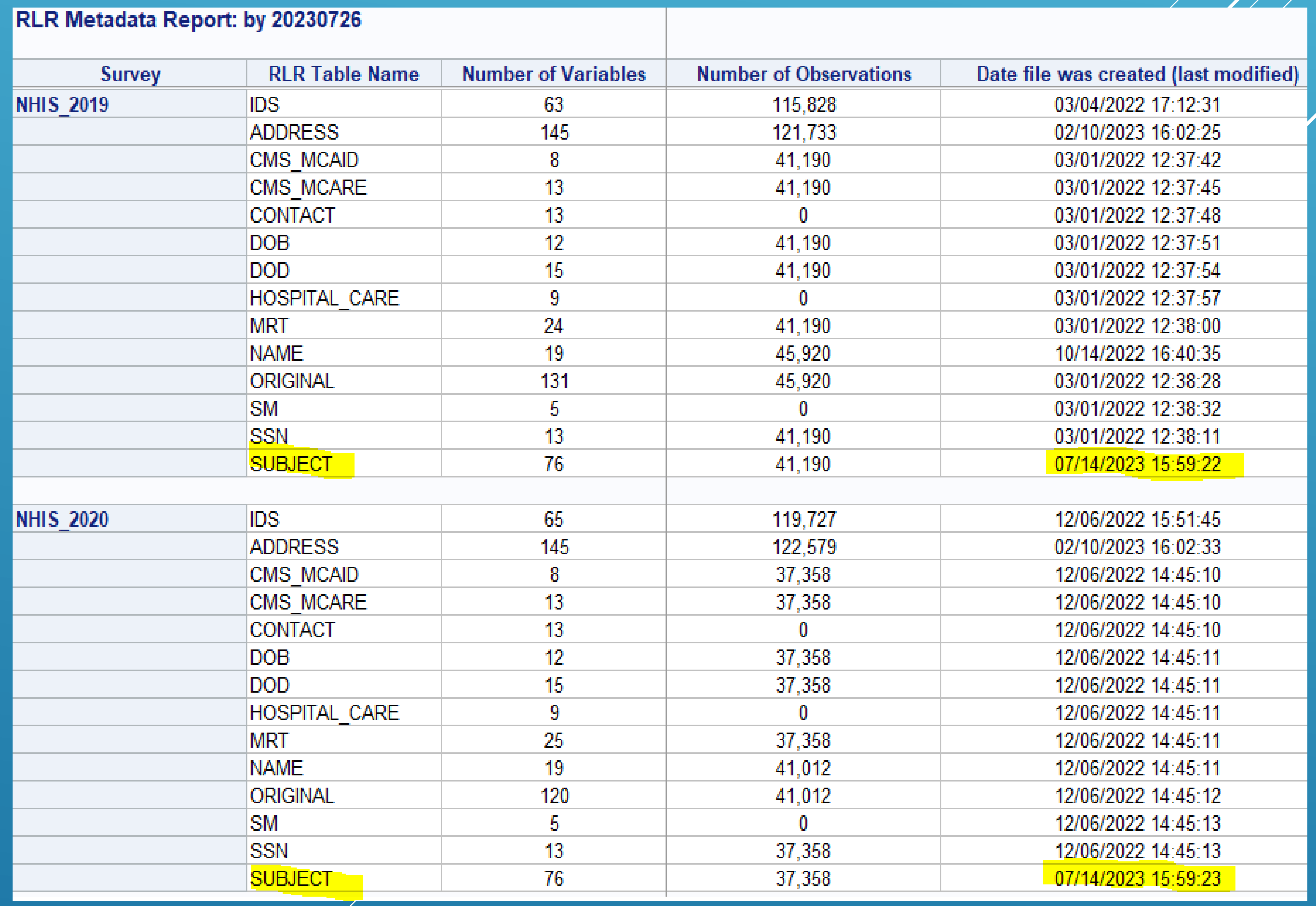
RLR Metadata Report: by 20230531

Survey	RLR Table Name	Number of Variables	Number of Observations	Date file was created (last modified)
NHIS_2019	IDS	63	115,828	03/04/2022 17:12:31
	ADDRESS	145	121,733	02/10/2023 16:02:25
	CMS_MCAID	8	41,190	03/01/2022 12:37:42
	CMS_MCARE	13	41,190	03/01/2022 12:37:45
	CONTACT	13	0	03/01/2022 12:37:48
	DOB	12	41,190	03/01/2022 12:37:51
	DOD	15	41,190	03/01/2022 12:37:54
	HOSPITAL_CARE	9	0	03/01/2022 12:37:57
	MRT	24	41,190	03/01/2022 12:38:00
	NAME	19	45,920	10/14/2022 16:40:35
	ORIGINAL	131	45,920	03/01/2022 12:38:28
	SM	5	0	03/01/2022 12:38:32
	SSN	13	41,190	03/01/2022 12:38:11
	SUBJECT	76	41,190	09/09/2022 14:47:57
NHIS_2020	IDS	65	119,727	12/06/2022 15:51:45
	ADDRESS	145	122,579	02/10/2023 16:02:33
	CMS_MCAID	8	37,358	12/06/2022 14:45:10
	CMS_MCARE	13	37,358	12/06/2022 14:45:10
	CONTACT	13	0	12/06/2022 14:45:10
	DOB	12	37,358	12/06/2022 14:45:11
	DOD	15	37,358	12/06/2022 14:45:11
	HOSPITAL_CARE	9	0	12/06/2022 14:45:11
	MRT	25	37,358	12/06/2022 14:45:11
	NAME	19	41,012	12/06/2022 14:45:11
	ORIGINAL	120	41,012	12/06/2022 14:45:12
	SM	5	0	12/06/2022 14:45:13
	SSN	13	37,358	12/06/2022 14:45:13
	SUBJECT	76	37,358	12/06/2022 14:45:13



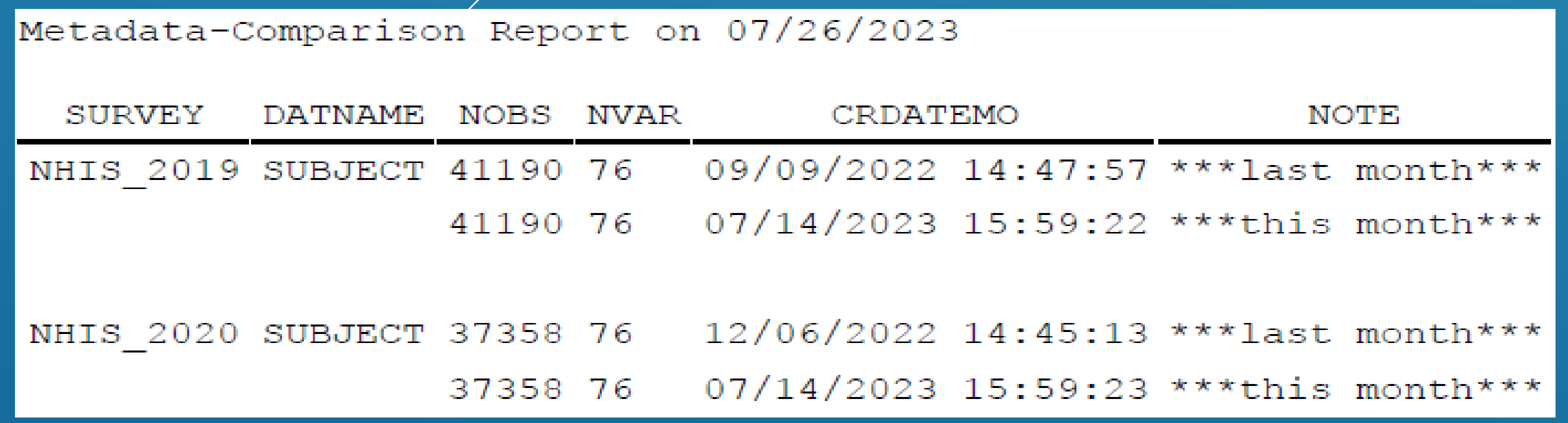
```
Metadata-Comparison Report on 05/31/2023

 DATASET: RLR_MetaData
 No changes were made from last Month
```

Example 3: July 2023 metadata table and comparison report



RLR Metadata Report: by 20230726

Survey	RLR Table Name	Number of Variables	Number of Observations	Date file was created (last modified)
NHIS_2019	IDS	63	115,828	03/04/2022 17:12:31
	ADDRESS	145	121,733	02/10/2023 16:02:25
	CMS_MCAID	8	41,190	03/01/2022 12:37:42
	CMS_MCARE	13	41,190	03/01/2022 12:37:45
	CONTACT	13	0	03/01/2022 12:37:48
	DOB	12	41,190	03/01/2022 12:37:51
	DOD	15	41,190	03/01/2022 12:37:54
	HOSPITAL_CARE	9	0	03/01/2022 12:37:57
	MRT	24	41,190	03/01/2022 12:38:00
	NAME	19	45,920	10/14/2022 16:40:35
	ORIGINAL	131	45,920	03/01/2022 12:38:28
	SM	5	0	03/01/2022 12:38:32
	SSN	13	41,190	03/01/2022 12:38:11
	SUBJECT	76	41,190	07/14/2023 15:59:22
NHIS_2020	IDS	65	119,727	12/06/2022 15:51:45
	ADDRESS	145	122,579	02/10/2023 16:02:33
	CMS_MCAID	8	37,358	12/06/2022 14:45:10
	CMS_MCARE	13	37,358	12/06/2022 14:45:10
	CONTACT	13	0	12/06/2022 14:45:10
	DOB	12	37,358	12/06/2022 14:45:11
	DOD	15	37,358	12/06/2022 14:45:11
	HOSPITAL_CARE	9	0	12/06/2022 14:45:11
	MRT	25	37,358	12/06/2022 14:45:11
	NAME	19	41,012	12/06/2022 14:45:11
	ORIGINAL	120	41,012	12/06/2022 14:45:12
	SM	5	0	12/06/2022 14:45:13
	SSN	13	37,358	12/06/2022 14:45:13
	SUBJECT	76	37,358	07/14/2023 15:59:23



```
Metadata-Comparison Report on 07/26/2023

  SURVEY   DATNAME NOBS  NVAR      CRDATEMO              NOTE
 NHIS_2019 SUBJECT 41190 76   09/09/2022 14:47:57 ***last month***
                   41190 76   07/14/2023 15:59:22 ***this month***

 NHIS_2020 SUBJECT 37358 76   12/06/2022 14:45:13 ***last month***
                   37358 76   07/14/2023 15:59:23 ***this month***
```

Example code

```
*Back up this month's metadata as last month's metadata in our folder.  Afterwards, get the PROC
CONTENTS output from about 100 survey years into an ODS output ATTRIBUTE dataset for current
month. Save the resulting metadata as this month's metadata;
%LET TODAY=%SYSFUNC(TODAY(),YYMMDDN8.) ;
%LET ROOT_OUT = folder name here ;
LIBNAME METADATA "folder name here" ;
DATA METADATA.LASTMON ; SET METADATA.THISMON ; RUN ;
%MACRO CREATE_ATR(LIB,DAT) ;
    ODS OUTPUT ATTRIBUTES=ATR ;
        PROC CONTENTS DATA=&LIB.._ALL_ MEMTYPE=DATA ; RUN ;
        PROC SQL ;
        CREATE TABLE MEMBER AS SELECT "&SURVEY" AS Survey LENGTH=20,
        SCAN(MEMBER,2) AS DATNAME LENGTH=20, CVALUE2 AS NOBS FROM ATR
        WHERE LABEL2='Observations' ;
        CREATE TABLE VAR AS SELECT CVALUE2 AS NVAR FROM ATR WHERE
        LABEL2='Variables' ;
        CREATE TABLE MODATE AS SELECT CVALUE1 AS CRDATEMO FROM ATR WHERE
        LABEL1='Last Modified'; QUIT ;
        DATA &SURVEY._&DAT. ; MERGE MEMBER VAR MODATE ; RUN ;
    ODS _ALL_ CLOSE ;
%MEND CREATE_ATR ;
%MACRO GETMETA(SURVEYLIST) ;
    %LOCAL I ;
    %LET I=1 ;
    %LET SURVEY = %SCAN(&SURVEYLIST, &I) ;
    %DO %WHILE("&SURVEY" NE "") ;
    LIBNAME IDS "\\cdc\csp_project\CIPSEA_PII_OAE_LINK_MASTER\MASTER_IDS\&survey."
        ACCESS=READONLY ;
    LIBNAME RLR "\\cdc\csp_project\CIPSEA_PII_OAE_LINK_MASTER\MASTER_DATA\&survey."
        ACCESS=READONLY ;
                %CREATE_ATR(IDS,1)
                %CREATE_ATR(RLR,2)
                *append data;
                DATA &SURVEY. ;
                LENGTH NOBS $ 10 ;
                SET &SURVEY._1 &SURVEY._2 ;
                *del unwanted datasets;
                PROC DATASETS LIB=WORK ;
                DELETE ATR MEMBER VAR MODATE &SURVEY._1 &SURVEY._2 ;
                QUIT ;
                %LET I = %EVAL(&I + 1) ;
                %LET SURVEY = %SCAN(&SURVEYLIST, &I) ;
    %END ;
%MEND GETMETA ;
%LET SURVEYLIST=NHIS_2019 NHIS_2020 …… ;
```

```
%GETMETA(&SURVEYLIST)
DATA METADATA.THISMON ;
     LENGTH NOBS $ 12 NVAR $ 20 MODATE $ 50 ;
     SET &SURVEYLIST ;
    NOBS2=INPUT(NOBS,8.) ; RUN ;

*Use ODS to export current month's metadata to two Excel files. One of the files is a report for
the end user while the other is for tracking purposes;
%MACRO REPORT(file) ;
     ODS EXCEL FILE="&ROOT_OUT.\&file..xlsx"
         OPTIONS(START_AT="1,1"
                 FROZEN_HEADERS="3"
                 FROZEN_ROWHEADERS="3"
                 SHEET_NAME="RLR Metadata Report"
                 ROW_REPEAT="2"
                 EMBEDDED_TITLES="YES") ;
         TITLE "RLR Metadata Report: by &TODAY." ;
         PROC REPORT DATA=METADATA.THISMON ;
            COLUMN SURVEY DATNAME NVAR NOBS2 CRDATEMO ;
            DEFINE SURVEY / ORDER STYLE(COLUMN)=HEADER "Survey" ;
            DEFINE NOBS2 / CENTER  "Number of Observations" FORMAT=COMMA10. ;
            DEFINE NVAR / DISPLAY CENTER "Number of Variables";
            DEFINE CRDATEMO / DISPLAY CENTER "Date file was created (last modified)" ;
            DEFINE DATNAME / "RLR Table Name" ;
                 COMPUTE AFTER SURVEY ;
                          LINE ' ' ;
                 ENDCOMP ;
         RUN;
    ODS EXCEL CLOSE ;
%Mend REPORT ;
%REPORT(RLR_MetaData)
%REPORT(RLR_MetaData_&TODAY.)

*Using Proc Compare by SURVEY and DATNAME, compare this month's metadata against
last month's metadata;
PROC SORT DATA=METADATA.LASTMON ;
         BY SURVEY DATNAME NOBS NVAR CRDATEMO ;
PROC SORT DATA=METADATA.THISMON; ;
         BY SURVEY DATNAME NOBS NVAR CRDATEMO ;
RUN ;
PROC COMPARE BASE=METADATA.LASTMON (KEEP=SURVEY DATNAME NOBS NVAR
CRDATEMO)  COMPARE=METADATA.THISMON(KEEP=SURVEY DATNAME NOBS NVAR
CRDATEMO)  OUTNOEQUAL OUT=TOPRINT OUTCOMP OUTBASE ; ID SURVEY
DATNAME;
RUN ;
```

CHARLOTTE, NC
SESUG 2023
SHARE • LEARN • CONNECT
- OCTOBER, 22-24 -

Example code

```
DATA TOPRINT ;
        LENGTH NOTE $ 40 ;
        RETAIN SURVEY DATNAME NOBS NVAR ;
        SET TOPRINT ;
        IF _TYPE_= "COMPARE" THEN NOTE="     ***this month***     " ;
                ELSE NOTE="     ***last month***     " ;
        BY SURVEY DATNAME ;
        IF FIRST.DATNAME AND LAST.DATNAME THEN DO;
                IF _TYPE_="COMPARE" THEN NOTE="new dataset added in this month" ;
                ELSE IF _TYPE_="BASE" THEN NOTE="dataset only exist in last month" ;
         END ;
        DROP _TYPE_ _OBS_ ;
RUN ;

*Use ODS to export the comparison report as an Excel file;
%MACRO DELDAT ;
         PROC SQL NOPRINT ;
                SELECT COUNT(*) INTO :N FROM TOPRINT ;
         QUIT ;
         %PUT &N ;
         %IF &N = 0 %THEN %DO ;
                PROC DATASETS LIB = WORK KILL ;
                QUIT ;
         %END ;
%MEND DELDAT ;
%DELDAT
%MACRO CHANGED ;
        TITLE "Metadata-Comparison Report on %SYSFUNC(DATE(),MMDDYY10.)" ;
        ODS PDF FILE="&ROOT_OUT.\RLR_ComparisonReport_&TODAY..pdf"
        STYLE=Monospace ;
        ODS ESCAPECHAR='^' ;
        ODS NOPROCTITLE ;
        PROC REPORT DATA=&DAT NOWD ;
        COLUMN SURVEY DATNAME NOBS NVAR CRDATEMO NOTE ;
        DEFINE SURVEY/ORDER STYLE(COLUMN)=HEADER "SURVEY" ;
        DEFINE DATNAME/ORDER;
                DEFINE NOBS/DISPLAY WIDTH=20 FORMAT=$20. ;
                DEFINE NVAR/DISPLAY WIDTH=20 FORMAT=$20.;
                COMPUTE AFTER SURVEY ;
                  LINE ' ' ;
                ENDCOMP ;
        RUN;
        ODS PDF CLOSE ;
%MEND CHANGED ;
```

```
%MACRO NOCHANGE ;
        DATA NULL ;
          REPORT="No changes were made from last Month" ;
        RUN ;
        TITLE "Metadata-Comparison Report on %SYSFUNC(DATE(),MMDDYY10.)" ;
        ODS PDF FILE="&ROOT_OUT.\RLR_ComparisonReport_&TODAY..pdf"
        STYLE=Monospace ;
        ODS ESCAPECHAR='^' ;
        ODS NOPROCTITLE ;
        OPTIONS NONUMBER NODATE ;
        PROC PRINT NOOBS LABEL ;
                LABEL REPORT="DATASET: RLR_Meta_Data" ;
        RUN ;
        ODS PDF CLOSE ;
%MEND NOCHANGE ;
%MACRO CHECKDS(DAT) ;
        %IF %SYSFUNC(EXIST(&DAT)) %THEN %DO ;
                %CHANGED
        %END ;
                %ELSE %DO ;
                        %NOCHANGE ;
                %END;
%MEND CHECKDS ;
%CHECKDS(TOPRINT)
```

Conclusions

- The NCHS Data Linkage Program utilizes metadata tables to verify updates to the RLR and ensure the accuracy of the linked data files developed.
- Metadata tables serve as a critical tool for data quality assurance when maintaining complex databases.

Contact Information

Contact me at: Yeats.Ye@cdc.hhs.gov
Contact the Data Linkage Program: datalinkage@cdc.gov
Visit our website: www.cdc.gov/nchs/data-linkage

Subscribe to the NCHS Data Linkage Program LISTSERV to receive updates!
Email a message to list@cdc.gov. Leave the subject line blank. In the body of the message, type:
SUBSCRIBE NCHS-DATA-LINKAGE-PROGRAM last name, first name