AGE and BIRTH YEAR - General Target Variable Report (GVR)

1. General Information

The target variable T_AGE measures respondent's age in years (cf. Table 1). To construct T_AGE we use two types of source variables: respondents' declared age and their year of birth. Our first choice is to harmonize information about declared age whenever available. If source datasets do not provide age variables but respondent's year of birth is available, we use the latter to derive target values of age. We flag procedures of deriving T_AGE values through the harmonization control variable C_AGE_DERIVED (see Sections 3.2 & 3.4).

The target variable T_BIRTHYR records respondents' year of birth (cf. Table 1). To construct it, we follow the same criteria and apply the same procedures as for T_AGE, this time on age derived from reported birth year and interview year. Specifically we use source data on birth year if available, otherwise we derive respondents' year of birth from their reported age and flag this procedure via C_BIRTHYR_DERIVED (see Sections 3.2 & 3.4).

T_AGE and T_BIRTHYR are accompanied by a set of methodological indicators (cf. Table 1 and Section 3.4.). First, the harmonization controls C_AGE_DERIVED and C_BIRTHYR_DERIVED identify age and birth year target values that have been derived by us. Second, we provide two control variables for quality of data records in the source files: QR_AGE_SUSPECT and QR_BITRTHYR_SUSPECT identify values of the target measures that we define as suspicious (in general, instances when age - provided or derived - is outside the 14-99 years range).[1] Third, the quality control variables for processing error types, QP_AGE and QP_BIRTHYR, measure the occurrence of processing errors on source variables used to construct T_AGE and T_BIRTHYR. Finally, the quality flag variable QF_AGE_BIRTHYR_DIFF measures the difference between source age values (i.e. respondents' declared age) and age values derived from source birth year, when both source variables for age and year of birth are available. We discuss all methodological variables in Section 3.4.

The target variable report for T_AGE and T_BIRTHYR is accompanied by the following Excel documents:

- The Detailed Variable Report (DVR): T_AGE_BIRTHYR_DVR_SDR2.xlsx. DVR Excel files in SDR2 systemize all information about source variables that were used for harmonization into a given target variable of the SDR2 database;
- Three Crosswalk Tables (CWT): T_AGE_CWT_SDR2.xlsx, T_BIRTHYR_CWT_SDR2.xlsx and T_INT_YEAR_CWT_SDR2.xlsx. CWT Excel files in SDR2 contain details about mapping of source values to target values.

[1]The decision to control for, rather than remove, suspicious values on T_AGE and T_BIRTHYR is informed by the SDR2 rule to preserve as much of the source data as possible in the process of creating harmonized measures. Thus, researchers can define age according to their substantive interests and use control variables to examine the extent to which removing suspicious values matters for empirical analyses.

Table 1.1 Description of the Target Variables on AGE and BIRTH YEAR

	Variable description	Variable name	Variable values [a]
Target variable	Respondent's age in years	T_AGE	2 = minimum age 130 = maximum age
	Respondent's year of birth	T_BIRTHYR	1859 = minimum birth year 2008 = maximum birth year
Source variables			See: T_AGE_BIRTHYR_DVR_SDR2.xlsx T_AGE_CWT_SDR2.xlsx T_BIRTHYR_CWT_SDR2.xlsx T_INT_YEAR_CWT_SDR2.xlsx
Harmonization control variables	Values of age are derived	C_AGE_DERIVED	0 = Not derived 1 = Derived from grouped values (e.g. 18-28) 2 = Derived from birth year
	Values of birth year are derived	C_BIRTHYR_DERIVED	0 = Not derived 1 = Derived from grouped values (e.g. 9980) 2 = Derived from age
Quality control variables for data records	Values of T_AGE are suspicious	QR_AGE_SUSPECT	0 = Not suspect 1= Suspect
	Values of T_BIRTHYR are suspicious	QR_BIRTHYR_SUSPECT	0 = Not suspect 1= Suspect
Quality control variables for processing error types	Processing error types for values of age [nominal variable, see Section 3.4]	QP_AGE	0 = No processing errors … 13 = Illegitimate & discrepant & lack of variable value labels
	Processing error types for values of birth year [nominal variable, see Section 3.4]	QP_BIRTHYR	0 = No processing errors … 13 = Illegitimate & discrepant & lack of variable value labels
Quality flag variable for data records	Differences between source age and age derived from source birth year	QF_AGE_BIRTHYR_DIFF	-80 = Minimum difference 81 = Maximum difference NULL – (i.e. empty value) when difference cannot be calculated due to missing source age or birth year

[a] Missing values are assigned according to the SDR2 missing codes schema, provided in the Appendix.

2. Survey Projects

Source variables that we used for T_AGE appear in 23 international survey projects: ABS, AFB, AMB, ARB, ASES, CB, CDCEE, CNEP, EB, EQLS, ESS, EVS, ISJP, ISSP, LB, LITS, NBB, NEB, PA1, PA2, PPE7N, VPCPCE, WVS, 174 waves and 3322 national surveys. The data cover 156 countries and years from 1966 to 2017.

Source variables that we used for T_BIRTHYR appear in 15 international survey projects: ABS, AMB, CB, CDCEE, CNEP, EB, ESS, EVS, ISJP, ISSP, LB, PA1, PA2, PPE7N, WVS, 47 waves and 947 national surveys. The data cover 131 countries and years from 1971 to 2017.

3. General Rules and Procedures

3.1. Source data description

In source data files, respondents' age is provided (1) as exact values (usually), (2) as values grouped in brackets (rarely), and (3) occasionally in both ways within a single variable. The following is the complete list where scenarios (2) and (3) occur:

- (2) Grouped age values (i.e. in brackets): `V5` in EB/0.1; in PPE7N – `V260` for India, `V90` for Japan, and `V192` for Yugoslavia e.g. 18-28 years, 29-38 years etc.
- (3) Age in exact values and in brackets for some surveys within a single variable:
  - `X003` in the integrated WVS/1-6 file for WVS/2 Brazil and South Africa and WVS/3 Puerto Rico and South Africa;
  - `V117` in ISSP/1985, `V80` in ISSP/1986, and `V82` in ISSP/1987 provides age in brackets only for Italy.

Some multi-country source data files provide both the variable with exact age values and an auxiliary measure with grouped values. Yet, in a few instances that we list below, for specific countries only the auxiliary variable is available:

- Auxiliary (i.e. age in brackets) variables: `SE003` instead of `SE003A` for ABS/1 Mongolia; `Q701AGECATEGORIES` instead of `Q701` for ARB/1 Morocco; `I100` instead of `V100` for ISSP/1992 Italy; `DK_AGE` instead of `AGE` for ISSP/2015 Denmark; `S3ROM` instead of `S3` for NEB/3 Romania; `S2COL` instead of `S2` for NEB/4 Croatia; `S2COL` instead of `S2` for NEB/5 Croatia and Poland.

We encounter a similar situation for birth year source variables. Among data files that provide this type of information, the vast majority does so using exact values. There are some exceptions that we list below:

- Both birth year in exact values (usually) and in approximate values (pointing to an undefined range of years) appear within a single variable: `X002` for WVS/3 Puerto Rico and South Africa, e.g. `1927`, `1938`, `1955`, and `1965` are the only source values for WVS/3 Puerto Rico.
- As above, but grouped values for birth year are only for some respondents in a given national survey: `RESPDOB` in CB/2009, 2010, and 2015; for example a code `9980` is used for respondents born in the 1980s (for whom we assign a middle value, 1985),

while for other respondents we have exact birth year values, e.g. 1981. 1982, and the like.

In seven instances, information about respondents' age is not present in the source data: AMB/2008 United States; CDCEE/1 Lithuania and Romania; PPE7N Netherlands; WVS/1 Finland, Hungary, and South Korea. Of these, only PPE7N Netherlands provides an alternative measure of respondents' year of birth, for other cases we have a missing code, QNA.

Compared to availability of respondent's age measures, source information on year of birth is much sparser. 71 percent (2371 of the 3329 national surveys) do not provide data about respondent's birth year.

It is important to note that information about age and year of birth in the source data files does not always match information provided in the source documentation.

3.2. Rules of transformation of source variables into target variable

1. In constructing T_AGE and T_BIRTHYR, we give priority to source variables that record exact values of age, and of birth year, respectively. The use of this type of information constitutes the most common procedure of building both target variables.
2. If exact age or year of birth values are not available (either for all, or a subgroup of respondents, cf. Section 3.1), but grouped values exist, we use the latter to derive T_AGE and T_BIRTHYR values, respectively. Our decisions are as follows:
   - For T_AGE we assign as target values the values of groups' midpoints (e.g. a respondent aged '25 – 34' in the source data is assigned the target value 29), their edge points (e.g. we assign the target value 19 for source group values '19 and less', target value 75 for source group values '75 and more'), and their first value if group values consist of two consecutive values (e.g. target value 18 for source group values '18-19'). In one case (ISSP_2015 `DK_AGE`), where the source value is 70 while its label says '65 and older,' we assign the target value 70.
   - For T_BIRTHYR, we assign as target values the midpoints of given decades (e.g. 1985 for the source `9980` meaning 1980s in Caucasus Barometer), or treat source values of which we know they are approximate as the target (e.g. `1927`, `1938`, `1955`, and `1965` which are the only source values for WVS/3 Puerto Rico).

     Whenever target values on T_AGE and T_BIRTHYR are derived using this procedure, the harmonization control variables C_AGE_DERIVED and C_ BIRTHYR _DERIVED, respectively, take the value 1 (see Section 3.4).
3. If no source variables for age are available, but there is information regarding respondent's year of birth, we derive values of T_AGE by subtracting reported birth year from interview year (T_INT_YEAR).[2] This situation is captured with the control value C_AGE_DERIVED = 2.

[2] For calculating age, we gave preference to the variable "year of interview" over the variable recording year of survey, since fieldwork sometimes lasts for more than one year. We took only a year of interview into consideration, and omitted additional variables concerning day/month of interview. In the case of missing

We follow the same procedure for T_BIRTHYR. If source variable on birth year is not available, we derive it by subtracting age from the interview year (T_INT_YEAR) and flag such cases with the control value C_ BIRTHYR_DERIVED = 2.

3.3. Rules for identifying suspicious values on T_AGE and T_BIRTHYR

1. First, we analyze the distribution of T_AGE. For T_AGE values lower than 14, we code QR_AGE_SUSPECT = 1, since the lowest documented age in the source data is 14 years. For T_AGE larger than 99, we also code QR_AGE_SUSPECT = 1. In doing so, we take into consideration that the relative number of centenarians in 100 000 population is very low (UN World Population Prospects, https://population.un.org/wpp/) and serious illnesses in old age (e.g. Corrada et al 2010, Yang et al. 2013) decrease the probability that centenarians fully complete a survey interview.

   We flag suspicious target values of T_BIRTHYR using the control variable QR_BIRTHYR_SUSPECT. We apply the same rules to identify suspicious year of birth target values, by examining the distribution of age derived from source birth year.

2. Next, we analyze two separate age distributions, of T_AGE values and of age values derived from T_BIRTHYR, for all national surveys that provide source birth year data. The purpose is to screen out values that we define as suspicious.

   In the case of T_AGE:

   - For the left-hand tail of the distribution (i.e. for the youngest respondents), we observe if there is a gap in the sequence of age *values*. If we find a gap, we treat the values to its left as suspicious, and code QR_AGE_SUSPECT as 1.

     Example: Assume that, for a given national survey, the left-hand tail of the distribution of T_AGE is given by the sequence 14, 15, 17, 18, 19, 20… There is a gap between values 15 and 17. We code QR_AGE_SUSPECT = 1 for values to the left of the gap, which are, in this case, 14 and 15; we code QR_AGE_SUSPECT = 0 for target values 17 and 18 (even if the documented minimum age is higher, e.g. 19). We assume that the observed continuity in the sequence of values to the right of the gap increases the likelihood that age was misreported in documentation rather than it was misattributed to respondents in a dataset.

   - For the right-hand tail of the age distribution (i.e. for the oldest respondents) we observe if there is a bump in age *frequencies* in the values 97, 98, and 99. We treat these values with suspicion, because they often are missing codes in source datasets and, if not documented, we could accidentally take them as valid age values. So, if we find a bump, and the "bumped" value is over three times more frequent than the previous one, we treat this and the next values from the 97 – 99 range as suspicious and code QR_AGE_SUSPECT = 1.

information about interview dates, we used fieldwork dates from the survey documentation. For more details on interview year, see SDR2 Master File Overview document.

In the case of T_BIRTHYR we apply similar solutions to the distribution of age derived from birth year, and record them in the control variable QR_BIRTHYR_SUSPECT.

Information about documented minimum and maximum age as reported in the source documentation and investigations of age distributions on the national level are stored in a separate sheet "Age-Birthyr(Age2)-distributions" of the Detailed Variable Report (DVR) Excel file T_AGE_BIRTHYR_DVR_SDR2.xlsx, which is part of the joint target documentation on age and birth year.

3.4. Methodological variables that accompany T_AGE and T_BIRTHYR

Two harmonization control variables: C_AGE_DERIVED and C_BIRTHYR_DERIVED identify cases when a target value for T_AGE and T_BIRTHYR, respectively, is derived through harmonization procedures.

C_AGE_DERIVED is coded as:

- 0 when T_AGE values are based on source variables providing exact values of age. These are the majority of our cases.
- 1 if we derive age from source values of grouped age values (i.e. age 'in brackets');
- 2 if we derive age by subtracting reported birth year from the interview year. This procedure is applied when source data provide respondent's year of birth only.

C_BIRTHYR_DERIVED is coded as:

- 0 when T_BIRTHYR values are based on source variables providing exact values of respondents' birth year. Same as in age, these are the majority of our cases;
- 1 if we derive birth year from source values of grouped birth year values;
- 2 if we derive birth year by subtracting reported age in years from the interview year. This procedure is applied when source data provide respondent's age in years only.

Two quality controls for source data records: The variables QR_AGE_SUSPECT and QR_BITRTHYR_SUSPECT are dichotomous measures. They take the value 1 for any value of T_AGE and T_BIRTHYR, respectively, that meets the SDR2 rule of being a suspicious (cf. Section 3.3). The quality control variables are coded 0 when values of T_AGE and T_BIRTHYR, respectively, are not considered suspicious.

Two quality control variables for processing error types: QP_AGE and QP_BIRTHYR are nominal variables that can take values from 0 (no processing errors) to 13. Values larger than 0 identify what type, or combination of types, of processing errors occur in the source variables associated with T_AGE and T_BIRTHYR, respectively.

QP_AGE and QP_BIRTHYR, stored in the SDR2 PLUG_SURVEY file, measure processing error types that we identified when comparing information about the source variables available in the survey documentation (codebook, questionnaire) with information about the same source variables in data records in the source data files (data dictionaries). Put differently, QP_AGE and QP_BIRTHYR account for inconsistencies between different metadata elements, such as data records on the one hand (i.e. variable values, variable labels, value labels – in the source data files), and information in the codebook or questionnaires, on the other hand.

Generally, in the SDR2 project we check for five types of processing errors that we define as follows:

- **Illegitimate Variable Values** captures values of the source variable that are outside of the range that SDR proposes as acceptable, e.g. respondents' age below 14 and above 99 years.
- **Misleading Variable Values** shows that variable values, as coded in the questionnaires/codebooks (e.g., age cohorts), are not congruent with the data records (actual age in years mixed with age cohorts) in the source data files.
- **Contradictory Variable Values** captures inconsistencies in how the same variable value is labeled in a codebook vs. a questionnaire or a data dictionary.
- **Variable Values Discrepancy** concerns situations when several or all values in the same variable are inconsistently labeled in codebook, vs. questionnaire or a data dictionary (i.e. there are reversed scales or contradictory value labels for more than one variable value).
- **Lack of Variable Value Labels** concerns undefined "nulls" and codes (variable values) in the source data file that are not explained in any source of documentation defining variables and their values (i.e. "wild codes" for a given variable, such as value 7 in data records while the documentation describes the variable as a 1 to 5 scale).

Correspondingly, the full list of SDR2 codes for Processing Error Types is::

0 = No processing errors
1 = Illegitimate variable values
2 = Misleading variable values
3 = Contradictory variable values
4 = Variable values discrepancies
5 = Lack of variable value labels
6 = Illegitimate & misleading variable values
7 = Illegitimate & contradictory variable values
8 = Illegitimate & discrepancy variable values
9 = Illegitimate & lack of variable value labels
10 = Contradictory variable values & lack of variable value labels
11 = Discrepant & lack of variable value labels
12 = Illegitimate & contradictory & lack of variable value labels
13 = Illegitimate & discrepant & lack of variable value labels

Values 6 to 13 indicate that a combination of different processing error types occurs within one source variable.

Quality flag for T_AGE and T_BIRTHYR: QF_AGE_BIRTHYR_DIFF indicates the difference between source age, and age calculated using source birth year, for each respondent for whom both age and birth year are available. The variable takes the minimum value -80 and the maximum value 81, while 0 means there is no difference. If comparison of source age and age derived from birth year is not possible – due to missing data on at least one of the sources – the control takes a NULL value (which translates to `sysmiss` in statistical packages, e.g. SPSS and STATA). Figure 1 presents the distribution of these differences. For the purpose of clarity, the graph excludes extreme values (beyond +/-50); it also excludes values between -2

and 2 because of their high frequency. An interesting fact is that higher differences are generally less frequent than smaller ones, except for the values of +/-5, +/-10, +/-20 etc.

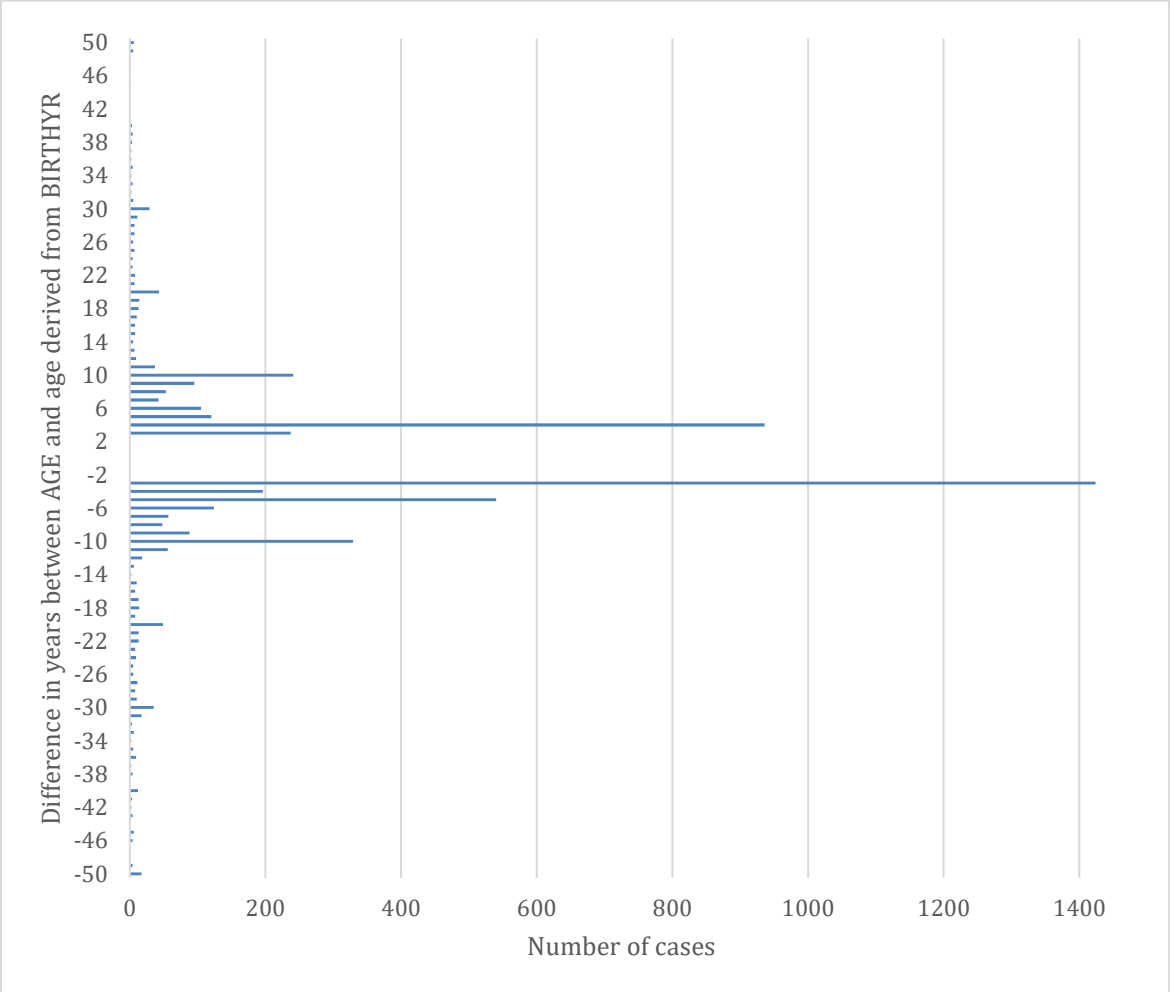

Figure 1. Distribution of differences in years between T_AGE and age derived from T_BIRTHYR for surveys where both variables are available.

4. Special cases

- CNEP/3 Hungary, the questionnaire contains a question about respondent's birth year, but the data file records store only values on age in years, derived from the birth year.
- In LB/1995-2004, variable values from the master questionnaire suggest that age is grouped in cohorts, but it appears in exact years in the data.
- In many surveys in ABS/1; ABS/4; AFB/1-6; AMB/2008,2014,2016; CB/2010,2011; CNEP/3; ESS/1-8; EVS/1,3,4; ISSP/2001,2011-2015; PPE7N Netherlands (after derivation from birth year) data contain improbable age values: well over 100 years old (up to 130). In the case of CDCEE/1 Ukraine; EB/0.2 Denmark, France, United Kingdom; EB/13 Denmark; EB/42 Luxembourg; ESS/2 Greece; LITS/2 Hungary, Lithuania, Russia; PPE7N Nigeria; WVS/1 Mexico, there are values in the data that are too low for respondents' age (below 14) and that are unaccounted for in documentation. For other surveys, we coded some values that broke our rules described in Section 3.2 (gaps and bumps) as ERR (see Appendix on missing codes schema); for details see

"Age-Birthyr(Age2)-distributions" Excel sheet in the file: T_AGE_BIRTHYR_DVR_SDR2.xlsx.

- In similar way, we identified birth years which lead to improbable (derived) age values: over 100 years (ABS/4; AMB/2012,2014,2016; CB/2010-2012,2015; ESS/1-8; EVS/1,3,4; ISSP/2011-2015; PPE7N Netherlands; WVS/5,6) and below 14 years (AMB/2012 Mexico; CDCEE/1 Ukraine; CNEP/3 Mexico; ESS/2 Greece; ISSP/2010 Chile; WVS/1 Mexico; WVS/2 Chile; WVS/6 Rwanda). For some other surveys additional criteria described in Section 3.2 were violated; for details see "Age-Birthyr(Age2)-distributions" Excel sheet in the file: T_AGE_BIRTHYR_DVR_SDR2.xlsx .
- In ISSP/2010 the codebook indicates DK and NA are 2 digits, in data – in 3 digits.
- In AMB/2016 the values of 'Don't know' and 'No answer' ('No response') in national level questionnaires differ from the values in the data dictionary.
- LB/2016 contains a birth year format that most probably corresponds to day-month-year (six-digit numbers), but also includes values like 300119998, which are unexplained elsewhere in the survey documentation. We recoded day-month-year format into a year format and assigned a code ERR to unexplained source values in LB/2016.
- In EB/0.2 source birth year values are last digits of the years so we added 18XX or 19XX depending on a value.
- In EB/0.2 one case of birth year in Great Britain is more recent (1975) than the interview year (1973). We flagged this case as QR_BIRTHYR_SUSPECT=1.
- In all waves of Caucasus Barometer data contain years in four-digit numbers, whereas documentation indicates the years in cohorts.

References

Corrada Maria M., Ron Brookmeyer, Annlia Paganini-Hill, Daniel Berlau. 2010. Dementia incidence continues to increase with age in the oldest old: The 90+ study. Annals of Neurology 67 (1): 114-121.

Yang, Zixuan., Melissa J Slavin, Perminder S. Sachdev. 2013. Dementia in the oldest old. Nature Reviews Neurology 9 (6): 382–393.

UN World Population Prospects, https://population.un.org/wpp/.

Appendix: Codes for missing values in SDR2

In the SDR database v.2 we identify different situations that warrant to be treated as missing data. Table A.1 lists all SDR2 missing value codes:

Table A.1. Codes for missing values in SDR2

SDR tag [a]	**SPSS (STATA) codes**	**Label**
		Standardized source codes for missing values
DK	-1 (.a)	Don't know
NA	-2 (.b)	No answer
REF	-3 (.c)	Refusal
DU	-4 (.d)	Don't understand the question
DNR	-5 (.e)	Any combination of DK, NA, REF, DU
INAP	-6 (.f)	Inapplicable
NEC	-7 (.g)	Not elsewhere classified
		SDR created codes for missing values
UNFIT	-8 (.h)	Source value does not fit to target
ERR	-9 (.i)	Errors in source data and undocumented source values
COMBI	-10 (.j)	Different missing codes on multiple sources taken for a target
CINAP	-11 (.k)	For control variables only: inapplicable
INSUF	-12 (.l)	For survey: Insufficiently defined response categories
QNA	-20 (.t)	For survey: Question not available

[a] Abbreviations for the labels corresponding to the SDR2 codes for missing values. These tags are used in the Crosswalk Table (CWT) files (Excel) that accompany documentation of SDR2 target variables.

In exceptional situations when codes for missing data listed in Table A.1 cannot be used, we apply a system missing <null> value.