Why ”dummy data”?

Important in the euCanSHare Data Catalogue are:

- **The study descriptions**. These are entered directly into Mica and provide information on the study population, study design, study size etc. (See the tutorial ”How to upload data to Mica”)
- **The variable descriptions.** These describe the data variables available from the studies. The variables are described in data dictionary files which are uploaded to Opal, and linked to Mica, as described in the tutorial ”How to upload data to Opal”.
- **Annotation of the variables** according to Maelstrom ”Areas of Information” and additional euCanSHare taxonomies. These make it possible to search for variables in specified areas of interest, across the studies. The annotations of the variables are also entered into Opal as a part of the data dictionaries. The annotation process is described in the tutorial "Annotation of variables in Opal" (login required).

If individual level *real* data are available in Opal, these can be linked to Mica so that the Catalogue shows the distributions of the variables. However, linkage of the real data to a publicly accessible Mica can be interpreted as violating the General Data Protection Regulation (GDPR).

Although the distributions of the variables cannot be shown in the euCanSHare Catalogue, it is possible to show the amount and reasons of missing data of the variables. This can be done by replacing the real variables by dummy variables, which provide information on the availability and unavailability of the real data values, but not on the actual measured values. As these ”dummy data” do not convey information on the characteristics of the individuals persons, they are not ”personal data” in the sense of the GDPR.

Although the ”dummy data” are not as important in the euCanSHare Catalogue as the study descriptions, ”real” data dictionaries and the annotations of the variables, they are useful to those searching data for research.

This document describes how to create and upload the dummy data into Opal and Mica.

For the ”dummy data” (i.e. showing the missingness / availability information by variable), one has to enter

1. dummy data dictionary (Excel file) and
2. individual level dummy data (csv file), to Opal
3. update the dataset description in Mica.

How to code the dummy data?

Each study can define the dummy data categories according to what is relevant for the study and what metadata are available. There must be at least two dummy data categories (Valid value, Missing), but missingness can also be divided into multiple categories. For example, for the MORGAM studies, the following coding is used for all dummy variables:

1 = Valid value

2 = Missing, but replaceable to valid (Note: this typically relates to skip rules in data collection questionnaires. Be careful with this because it may reveal real values of the filter variable.)

3 = Missing (not by design) (Note: This typically refers to data which are not missing on purpose.)

4 = Missing by design
5 = Not applicable

Figure 1 shows the "Categories" sheet of the data dictionary Excel file based on real data and how the dummy data codes are derived from the metadata. However, please note that items EAGE and STOP are continuous and hence only the missing data categories are listed here.

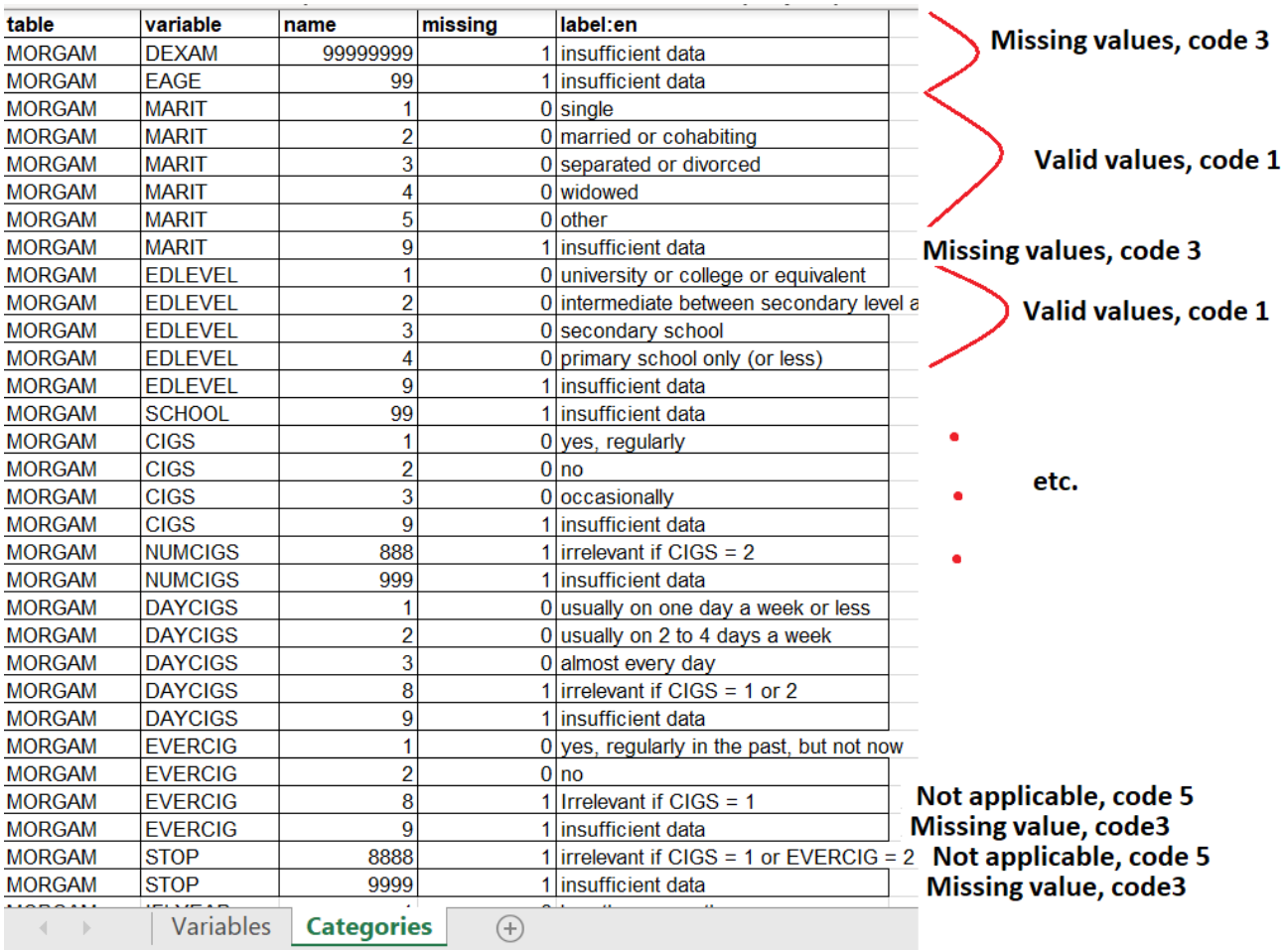

table	variable	name	missing	label:en
MORGAM	DEXAM	99999999	1	insufficient data
MORGAM	EAGE	99	1	insufficient data
MORGAM	MARIT	1	0	single
MORGAM	MARIT	2	0	married or cohabiting
MORGAM	MARIT	3	0	separated or divorced
MORGAM	MARIT	4	0	widowed
MORGAM	MARIT	5	0	other
MORGAM	MARIT	9	1	insufficient data
MORGAM	EDLEVEL	1	0	university or college or equivalent
MORGAM	EDLEVEL	2	0	intermediate between secondary level a
MORGAM	EDLEVEL	3	0	secondary school
MORGAM	EDLEVEL	4	0	primary school only (or less)
MORGAM	EDLEVEL	9	1	insufficient data
MORGAM	SCHOOL	99	1	insufficient data
MORGAM	CIGS	1	0	yes, regularly
MORGAM	CIGS	2	0	no
MORGAM	CIGS	3	0	occasionally
MORGAM	CIGS	9	1	insufficient data
MORGAM	NUMCIGS	888	1	irrelevant if CIGS = 2
MORGAM	NUMCIGS	999	1	insufficient data
MORGAM	DAYCIGS	1	0	usually on one day a week or less
MORGAM	DAYCIGS	2	0	usually on 2 to 4 days a week
MORGAM	DAYCIGS	3	0	almost every day
MORGAM	DAYCIGS	8	1	irrelevant if CIGS = 1 or 2
MORGAM	DAYCIGS	9	1	insufficient data
MORGAM	EVERCIG	1	0	yes, regularly in the past, but not now
MORGAM	EVERCIG	2	0	no
MORGAM	EVERCIG	8	1	Irrelevant if CIGS = 1
MORGAM	EVERCIG	9	1	insufficient data
MORGAM	STOP	8888	1	irrelevant if CIGS = 1 or EVERCIG = 2
MORGAM	STOP	9999	1	insufficient data

Figure 1: Data dictionary for the real data; "Categories" sheet in the data dictionary Excel file (i.e. the values and labels of the categories). Codes for dummy data, as defined for MORGAM studies, are marked to the right

Figure 2 shows the individual level dummy data after transformation according to the 5 categories of dummy codes (example taken from MORGAM studies). Please note that STOP (the year when smoking was stopped) takes value 5, i.e. not applicable, if the subject did not smoke. The data transformation here was done programmatically using R and saved as a csv file.

H	I	J	K	L	M	N	O	P	Q	R
SEX	DEXAM	EAGE	MARIT	EDLEVEL	SCHOOL	CIGS	NUMCIGS	DAYCIGS	EVERCIG	STOP
1	1	1	1	1	1	1	2	5	1	1
1	1	1	1	1	1	1	1	3	5	5
1	1	1	1	1	1	1	2	5	1	1
1	1	1	1	1	1	1	2	5	1	5
1	1	1	1	1	1	1	2	5	1	5
1	1	1	1	1	1	1	1	3	5	5
1	1	1	1	1	1	1	2	5	1	5
1	1	1	1	1	1	1	2	5	1	5
1	1	1	1	1	1	1	2	5	1	1
1	1	1	1	1	1	1	2	5	1	5
1	1	1	1	1	1	1	2	5	1	5
1	1	1	1	1	1	1	2	5	1	5
1	1	1	1	1	1	1	2	5	1	5
1	1	1	1	1	1	1	2	5	1	1
1	1	1	1	1	1	1	2	5	1	5
1	1	1	1	1	1	1	2	5	1	1

Figure 2: Individual level dummy data (to be saved as a csv file), example taken from MORGAM studies.

How to create a dummy data dictionary?

Dummy data dictionary Excel file is created similarly as described in the tutorial "How to upload data to Opal", with these exceptions:

1. Dummy data dictionary "table" column must be named differently from the real data dictionary (this name is given in both sheets; "Variables" and "Categories"),
2. "Variables" sheet should contain 6 columns (table, name, valueType, unit, label:en, description:en) - those can be copied from the real data dictionary Excel file (column "script" is not needed). For dummy data dictionary:
   - column "table" should be renamed, e.g. "xxx_dummy",
   - all valueTypes should be "integer" (if the 0,1,2,… are used for dummy codes), and
   - units are not needed (can be left empty).
3. "Categories" sheet must:
   - include the dummy codes for each variable (also for continuous / date / etc. variables) and the category values and labels are same for each variable,
   - column "table" should include the dummy table name, e.g. "xxx_dummy", and
   - column "missing" should be 0 for categories indicating valid and replaceable to valid values, and 1 for categories indicating missingness.

See figures 3 ("Variables" sheet) and 4 ("Categories" sheet) for details (examples taken from MORGAM studies).

table	name	valueType	unit	description:en	label:en
FINRISK2401_dummy	DEXAM	integer		"See the specific descriptio	Date of examination (day, month, year)
FINRISK2401_dummy	EAGE	integer		"See the specific descriptio	Age on the date of examination
FINRISK2401_dummy	MARIT	integer		"See the specific descriptio	Marital status
FINRISK2401_dummy	EDLEVEL	integer		"See the specific descriptio	"What is the highest level of education you have completed?"
FINRISK2401_dummy	SCHOOL	integer		"See the specific descriptio	"How many years have you spent at school or in full time study?"
FINRISK2401_dummy	CIGS	integer		"See the specific descriptio	"Do you smoke cigarettes now?"
FINRISK2401_dummy	NUMCIGS	integer		"See the specific descriptio	"On average how many cigarettes do you now smoke a day?" "
FINRISK2401_dummy	DAYCIGS	integer		"See the specific descriptio	"On how many days a week do you smoke cigarettes?"
FINRISK2401_dummy	EVERCIG	integer		"See the specific descriptio	"Did you ever smoke cigarettes regularly in the past?"
FINRISK2401_dummy	STOP	integer		"See the specific descriptio	"When did you stop smoking cigarettes regularly?" Record the year (fo

Variables | Categories

Figure 3: Variables for the dummy data dictionary (example taken from MORGAM).

table	variable	name	missing	label:en
FINRISK2401_dummy	DEXAM	1	0	Valid value
FINRISK2401_dummy	DEXAM	2	0	Missing, but replaceable to valid
FINRISK2401_dummy	DEXAM	3	1	Missing (not by design)
FINRISK2401_dummy	DEXAM	4	1	Missing by design
FINRISK2401_dummy	DEXAM	5	1	Not applicable
FINRISK2401_dummy	EAGE	1	0	Valid value
FINRISK2401_dummy	EAGE	2	0	Missing, but replaceable to valid
FINRISK2401_dummy	EAGE	3	1	Missing (not by design)
FINRISK2401_dummy	EAGE	4	1	Missing by design
FINRISK2401_dummy	EAGE	5	1	Not applicable
FINRISK2401_dummy	MARIT	1	0	Valid value
FINRISK2401_dummy	MARIT	2	0	Missing, but replaceable to valid
FINRISK2401_dummy	MARIT	3	1	Missing (not by design)
FINRISK2401_dummy	MARIT	4	1	Missing by design
FINRISK2401_dummy	MARIT	5	1	Not applicable
FINRISK2401_dummy	EDLEVEL	1	0	Valid value
FINRISK2401_dummy	EDLEVEL	2	0	Missing, but replaceable to valid
FINRISK2401_dummy	EDLEVEL	3	1	Missing (not by design)
FINRISK2401_dummy	EDLEVEL	4	1	Missing by design
FINRISK2401_dummy	EDLEVEL	5	1	Not applicable
FINRISK2401_dummy	SCHOOL	1	0	Valid value

Variables | Categories

Figure 4: Categories for the dummy data dictionary (example taken from MORGAM).

The dummy data dictionary Excel file is uploaded to Opal as given in the tutorial "How to upload data to Opal" (the same project folder for real data dictionary can be used, when the table names are not the same).

How to import the individual level dummy data to Opal?

The individual level dummy data (csv file) is imported into Opal similarly as given in the tutorial "How to upload data to Opal" in the section "Raw file". Note that in the tutorial the example data file contains the real data values for individual level data, but using the real data in linking to Mica can be interpreted as violating the GDPR.

"Destination table" for the individual level dummy data ("raw data") should be the dummy dictionary table.

Linking a dummy Opal data table to a dataset in Mica

To show the dummy data summary statistics in the variable description pages of the catalogue, the dummy data table should be added to the Mica dataset. The dummy data table uploaded to Opal is linked to the Mica dataset by ticking the box ”*This dataset is linked to a dummy table*”, and adding the name of the project and table exactly as given in Opal - see figure 5 below. After saving, remember also move the dataset description “to under review” and click “publish”.

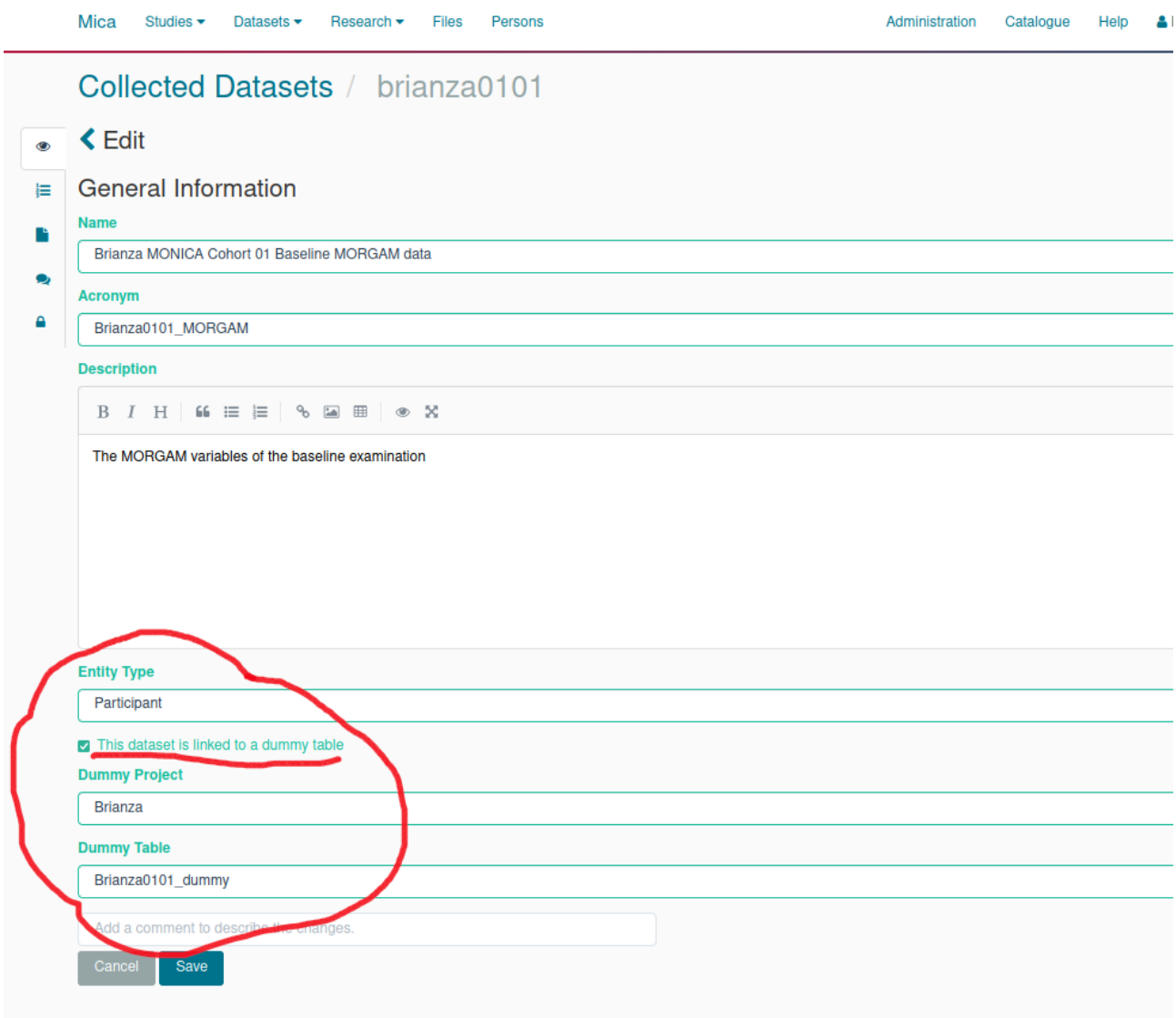

Figure 5: Link a dummy table to the Mica dataset.

Once a dummy table is linked to the Mica dataset and is published, the summary statistics for the variables in the dataset would be displayed as in figure 6 below. The top part, marked in red, displays the real data categories, while the section ’Summary statistics’ (marked in blue) displays the summary according to the dummy data categorisation.

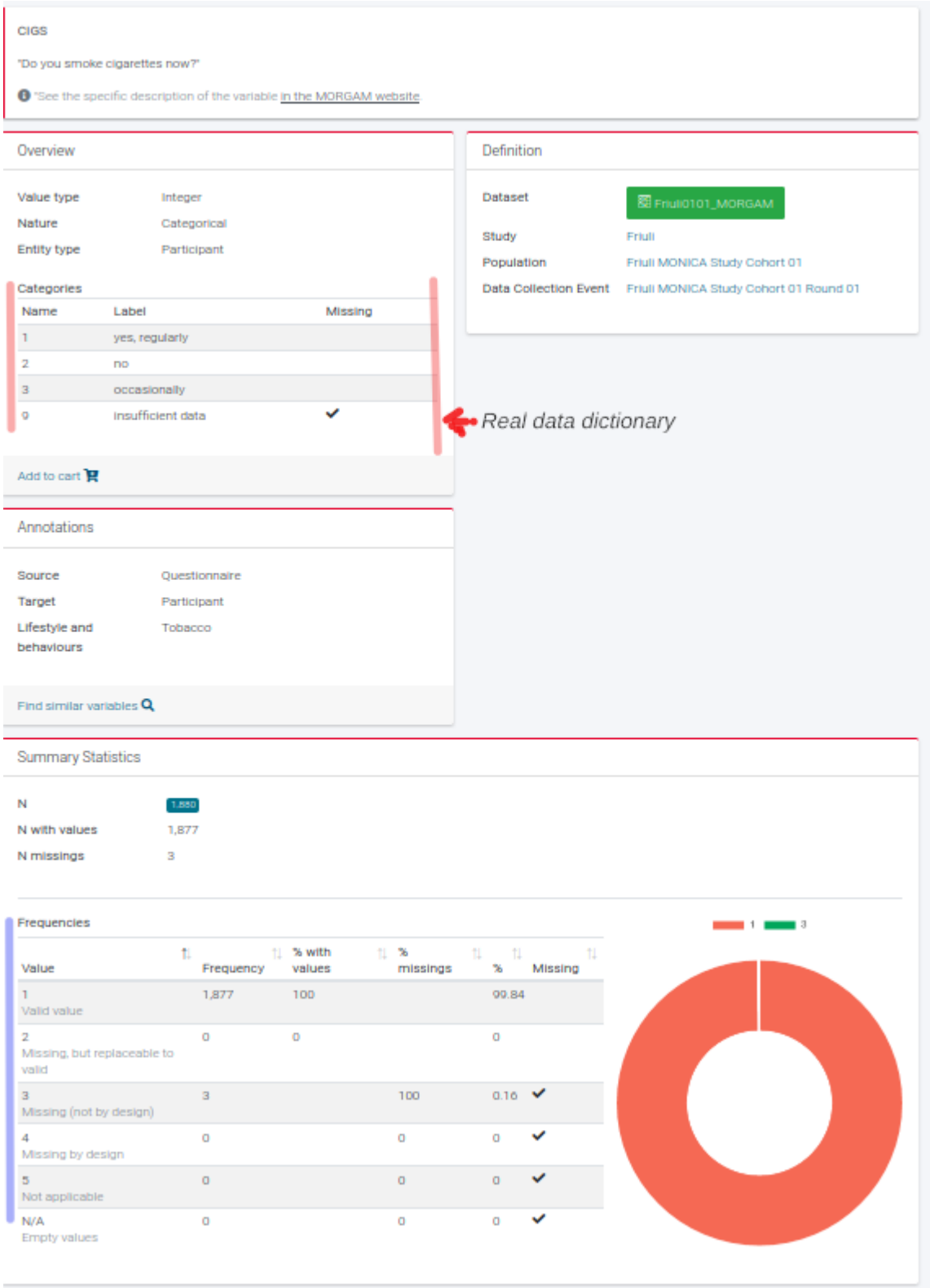

Figure 6: Summary statistics, with real categories marked in red, and dummy categories marked in blue.