Assumptions for exposure variables

In this document we describe the range of plausible values considered for each of the variables used in the imputation model. In addition, we provide the rules we use for combination variables, which were created in order to include as much information in the model as possible without the need to add many variables with a high proportion of missing values.

Plausibility bounds

The following table describes the plausibility ranges for the continuous and time-related variables used in the imputation process. It shows the minimum and maximum value for each variable. When an observation has values greater than or equal to the maximum value or less than or equal to the minimum value, we set the value of the variable out of range as NA.

Table 1: Plausibility bounds

Classification	Variable name	Variable type	Units	Min	Max
Pwoman	age	Continuous	(years)	10	55
Pwoman	weight	Continuous	(Kg)	30	200
Pwoman	pre_pregweight	Continuous	(Kg)	30	200
Pwoman	height	Continuous	(cm)	100	210
Exposure	zikv_pcr_ga_1	Time	(weeks)	0	46
Exposure	zikv_pcr_vl_1	Continuous	(copies/microL)	0	Inf
Exposure	zikv_elisa_ga_1	Time	(weeks)	0	46
Exposure	zikv_ga	Time	(weeks)	0	46
Exposure	symp_ga	Time	(weeks)	0	46
Exposure	arb_clindiag_ga	Time	(weeks)	0	46
Outcome	miscarriage_ga	Time	(weeks)	0	20
Outcome	loss_ga	Time	(weeks)	20	46
Outcome	endga	Time	(weeks)	0	46
Outcome	birth_ga	Time	(weeks)	0	46
Outcome	inf_weight	Continuous	(grams)	100	6000
Outcome	inf_length	Continuous	(cm)	18	70
Outcome	inf_head_circ_birth	Continuous	(cm)	0	Inf
Outcome	inf_head_circ_fu1	Continuous	(cm)	0	Inf
Outcome	inf_head_circ_age_fu1	Time	(months)	0	Inf
Outcome	inf_head_circ_fu2	Continuous	(cm)	0	Inf
Outcome	inf_head_circ_age_fu2	Time	(months)	0	Inf
Outcome	inf_head_circ_fu3	Continuous	(cm)	0	Inf
Outcome	inf_head_circ_age_fu3	Time	(months)	0	Inf

When running this code, the file "obs_to_check.RData" with the observations outside the plausibility range is automatically generated. This file contains all the information of the conflicting observations and additionally has the column "Var_to_check" with the name of the variable outside the limits and the column "Var_bound" with the possible outlier value. This file can be found in the folder "3_Output_data" in the folder "imputation process".

Logical rules for exposures

Set to NA values

SAS and R handle missing values differently. To facilitate the imputation process, we set to NA all values designated in the harmonisation process as 666,777,888,999. However, this assignment is not generalised as sometimes these NA values contain useful information in the definition of other variables.

Zika test measurement time (zikv_ga)

In case the date of the Zika test "zikv_ga" is not specified but information about the date of the PCR test "zikv_pcr_ga_1" or the elisa test "zikv_elisa_ga_1" is available. We set the value of zika_ga as the earliest date of the PCR and ELISA tests. if is.na (zikv_ga) then zikv_ga = min(zikv_pcr_ga_1,zikv_elisa_ga_1)

After the last meeting with the exposures group, we will include the test value in the assignment, i.e. for tests with positive values, the date of the first positive test, and in case of tests with negative values, the date of the last test performed.

Creation of combination variables

For the following variables, we combine different variables into a combination variable x_c in order to include as much information as possible without affecting the imputation process. These variables try to describe globally the presence of a certain characteristic or disease in the subject.

Table 2: Example of combination variable

Patient	x1	x2	x3	xn	xc
1	1	0	NA	NA	1
2	NA	0	NA	0	0
3	NA	NA	NA	NA	NA

For example, given the combination variable x_c created from the variables x_1,x_2,...,x_n. We assign a presence value $x_m = 1$, if any of the variables x_i has a presence value $x_i = 1$. (i.e. Patient 1, Table 2)

$$if(any(x_1 == 1, x_2 == 1, ..x_n = 1)), then, x_c = 1.$$

We assign a non-presence value $x_m = 0$ in case all variables has a non-presence value $x_i = 0$ (i.e. Patient 2, Table 2).

$$if(all(is.na(x_1), is.na(x_2), ..is.na(x_n = 1))), then, x_c = NA.$$

We assume that in case of missing values in some variable $x_i = NA$ its value is equal to non-presence $x_i = 0$. When there are not observable values for all the x_i we assign NA to the combination variable $x_c = NA$ (i.e. Patient 3, Table 2).

1. storch_patho

Denotes intrauterine exposure to storch pathogens. It was created by combining the following variables with the rule explained above.

- ifelse(storch==0,0,ifelse(!is.na(storch),1,0))]
- storch_bin

- toxo
- toxo_treat
- syphilis
- syphilis_treat
- varicella
- parvo
- rubella
- cmv
- herpes
- listeria
- chlamydia
- gonorrhea
- genitalwarts

2. arb_ever

Refers to the presence of a previous arb virus infection, and the following variables were used to create it

- ifelse(zikv_pcr_everpos==1,1,ifelse(zikv_pcr_everpos==0,0,NA))]
- zikv_elisa_everpos
- denv_ever
- chikv_ever

3. flavi_alpha_virus

This variable denotes the presence of a concurrent or previous flavi or alpha virus infection, and we combine the following variables

- ifelse(arb_clindiag_plus==0,0,ifelse(!is.na(arb_clindiag_plus),1,0))
- ifelse(arb_clindiag!=0&arb_clindiag!=1,0,ifelse(!is.na(arb_clindiag),1,0))
- denv_ever
- chikv_ever

4. arb_preg_nz

This variable is related to the presence of any arbovirus in the current pregnancy without considering zika virus, as the information on the presence of zika is found in the variable zika_preg.

- ifelse(arb_clindiag==0|arb_clindiag==1,0,ifelse(!is.na(arb_clindiag),1,NA))
- denv_preg
- chikv_preg