Questions on ZIKA dataset

Exposures

* According to the metadata file, the Brazil\_RiodeJaneiro\_Cunha study includes women of reproductive age. This means it can also include nonpregnant women. Is this indeed the case? If yes, is there a variable indicating whether a woman was pregnant or not? All outcomes are only applicable to pregnant women so nonpregnant women should be excluded.
* Variables zikv\_pcr\_ga\_1, zikv\_elisa\_ga\_1, zikv\_ga seem to have outliers -> can these be cleaned?
* For maternal zika status we are using zikv\_preg variable (=as defined by the study). There are many more variables on exposure. Can we create a new (dichotomous) version of this variable that is more satisfying to everyone? How can we construct this?
* For the following variables we were told there is no match in the pilot dataset. Is this indeed still the case, or are they now added? Workplace or environmental exposures to teratogenic substances (e.g. maternal exposure to lead, mercury), viral genotype, Maternal history of Japanese encephalitis vaccination.
* We impute only the exposure variables that are described in the sheet "Imputed exposure variables" in the attached “imputed\_variables” excel file. Over there, there are definitions on the variables that we created to facilitate the imputation process. Could you please review them and tell us if you agree with the variables included?

Outcomes

* Variable “loss\_etiology” – what do values of 4 mean?
* Outcome: ‘postnatal microcephaly’. We can use head-circumference variables to check for any changes. How should we do this? See if there is microcephaly at any time point and compare that to presence of microcephaly at birth?
* Variables “inf\_weight” (min 3.4 gram, max 23500 gram), “inf\_length” (values of 0, 4, 19 cm), inf\_head\_circ\_birth, inf\_head\_circ\_age\_fu1, age, all contain outliers -> can these be cleaned?
* Which are maximum and minimum realistic values for born babies? We set the following boundaries could you please check them inf\_weigh (100g,7000 g), inf\_length (18,60) inf\_head\_circ\_birth (15,50)
* Many outcome variables needed to be created from other variables in the dataset. We have done that, but are not sure whether we did it correctly. Below it is presented how we constructed these. Can someone check our work and indicate if things need to be changed?

|  |  |  |
| --- | --- | --- |
| **Variable** | **Variable is 1 (present) if:** | **Variable is 0 (absent) if:** |
| Neuroimaging abnormalities (intracranial calcification, lissencephaly, hydranencephaly, porencephaly, ventriculomegaly, posterior fossa abnormalities, cerebellar hypoplasia, corpus callosal and vermian dysgenesis; focal cortical dysplasia) | if one of the following is 0:  - fet\_us\_abn\_spec\_tri1  - fet\_us\_abn\_spec\_tri2  - fet\_us\_abn\_spec\_tri3,  or if one of the following is 1:  - hydrocephaly  - calcifications  - ventriculomegaly  - fet\_us\_cns\_tri2  - fet\_us\_cns\_tri3. | If any of the variables in the middle column is not missing and the variable is not yet coded as 1. |
| Cardiovascular abnormalities | if one of the following is 2:  - fet\_us\_abn\_spec\_tri1  - fet\_us\_abn\_spec\_tri2  - fet\_us\_abn\_spec\_tri3,  or if one of the following is 1:  - fet\_us\_cardio\_tri2  - fet\_us\_cardio\_tri3. | If any of the variables in the middle column is not missing and the variable is not yet coded as 1. |
| Gastrointestinal abnormalities | if one of the following is 3:  - fet\_us\_abn\_spec\_tri1  - fet\_us\_abn\_spec\_tri2  - fet\_us\_abn\_spec\_tri3,  or if one of the following is 1:  - fet\_us\_gastro\_tri2  - fet\_us\_gastro\_tri3 | If any of the variables in the middle column is not missing and the variable is not yet coded as 1. |
| Orofacial abnormalities | if one of the following is 4:  - fet\_us\_abn\_spec\_tri1  - fet\_us\_abn\_spec\_tri2  - fet\_us\_abn\_spec\_tri3,  or if one of the following is 1:  - fet\_us\_orofac\_tri2  - fet\_us\_orofac\_tri3 | If any of the variables in the middle column is not missing and the variable is not yet coded as 1. |
| Ocular abnormalities (blindness, other)  Or  Congenital deafness or hearing loss  (these are two separate variables in the protocol, however are combined in the dataset) | if one of the following is 5:  - fet\_us\_abn\_spec\_tri1  - fet\_us\_abn\_spec\_tri2  - fet\_us\_abn\_spec\_tri3,  or if one of the following is 1:  - fet\_us\_eyeear\_tri2  - fet\_us\_eyeear\_tri3 | If any of the variables in the middle column is not missing and the variable is not yet coded as 1. |
| Congenital contractures (arthrogryposis, uni or bilateral clubfoot) | if one of the following is 1:  - fet\_us\_abn\_spec\_tri1  - fet\_us\_abn\_spec\_tri2  - fet\_us\_abn\_spec\_tri3,  - fet\_us\_msk\_tri2  - fet\_us\_msk\_tri3. | If any of the variables in the middle column is not missing and the variable is not yet coded as 1. |
| Genitourinary abnormalities | if one of the following is 6:  - fet\_us\_abn\_spec\_tri1  - fet\_us\_abn\_spec\_tri2  - fet\_us\_abn\_spec\_tri3,  or if one of the following is 1:  - fet\_us\_genur\_tri2  - fet\_us\_genur\_tri3 | If any of the variables in the middle column is not missing and the variable is not yet coded as 1. |

Table 1. Created abnormalities variables

* For the following variables we were told there is no match in the pilot dataset. Is this indeed still the case, or are they now added? Placental insufficienty, Postnatal intraventricular hemorrhage, Motor abnormalities (hypotonia, hypertonia, hyperreflexia, spasticity, clonus, extrapyramidal symptoms), Seizures / epilepsy, Cortical auditory processing, Neurodevelopment (expressive and receptive language, fine and gross motor skills, attention and executive function, memory and learning, socioemotional development, overall neurodevelopmental score), Vision (Cardiff test)
* We created a new variable CZSn that replaces the incomplete value of the CSZ with a value given by a WHO definition (excel file), then the variable CZSn was used as input to the imputation model. This week, Mabel was kind enough to share with us detailed information about the CZS variable, we found that we used most of the information she proposed to use except the variables: othabnorm\_spec, storch test (later included as a predictor) and arthrogryposis. We are not sure how to include these variables, especially othabnorm\_spec, whose specification is impossible to understand. Do you have any suggestions in this regard? In cases where the CZS value is given should we prioritize the CZS diagnosis given by the study or the CZSn calculated according to the WHO definition?
* We impute only the outcome variables that are described in the sheet "Imputed outcome variables" in the attached “imputed\_variables” excel file. Over there, there are definitions on the variables that we created to facilitate the imputation process. Could you please review them and tell us if you agree with the variables included?

Effect modifiers / confounders / covariates

* Variable “tobacco” – what do values of 3 mean?
* Effect modifier: “Genetic anomalies, metabolic disorders, perinatal brain injury” -> how can we construct this?
* Effect modifier: “Clinical/subclinical illness” -> how should we construct this?
* Effect modifier: “Presence and severity of maternal and infant clinical symptoms” -> how should we construct this?
* Variables symp \_ga and arb\_clindiag\_ga seem to have outliers -> can these be cleaned?
* For the following variables we were told there is no match in the pilot dataset. Is this indeed still the case, or are they now added? Maternal experience of violence during pregnancy; infant or child exposure to intimate partner violence
* Which are the maximum and minimum plausible values on mother age?
* We impute only the modifiers/confounder variables that are described in the sheet "modifiers counfounder variables" in the attached “imputed\_variables” excel file. Over there, there are definitions on the variables that we created to facilitate the imputation process. Could you please review them and tell us if you agree with the variables included?