1 November 2021: Update on Exposure Workgroup response to Statistics Workgroup questions

Responses to questions on ZIKA dataset indicated in blue font below:

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.

EWG Proposal: Contact PI directly to determine if all were pregnant women. Janet can facilitate these communications.

* Variables zikv\_pcr\_ga\_1, zikv\_elisa\_ga\_1, zikv\_ga (gestational age when the ZIKV diagnosis was made) seem to have outliers -> can these be cleaned?

EWG: It will be helpful to know how many studies contain outliers in these variables. In the 20-20-20 key variables spreadsheet (<https://docs.google.com/spreadsheets/d/1rVmO-TLsAm-cno2crYFmAK-zFS5LUFFDdytT1LwpOOg/edit#gid=898781629>) there are valid reference ranges included. For the EWG to get a better idea of which variables have illogical outliers it would help if the SWG could provide the ranges of the values. Based on this we could differentiate between plausible outliers and invalid values. Given iterative additions of study results to the overall dataset, it is not feasible to send email updates on the ranges for each iteration. However, they could set up an output update file with variables and range that can be kept current.

The variable zikv\_pcr\_ga\_1 can probably be used to populate/impute zikv\_ga but there needs to be a review of other related variables.

* 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? Analysis thus far only relied on zikv\_preg as the main exposure variable looking at absolute vs relative risk of infection in pregnancy and outcome. However, there are other variables that may be worth incorporating. In the 20-20-20 key variables spreadsheet, it stipulates that woman had diagnosis of ZIKV infection as identified by the study. The metadata indicates which laboratory criteria were used (<https://docs.google.com/spreadsheets/d/1okXSnixnUutrzS2OQVdrR_cz-7a9Z4ODyIwZVQI23lY/edit#gid=1885512341>). Heterogeneity of this variable was captured in metadata survey. We could use our own definition e.g., WHO definition, by extracting laboratory data and creating the classification based on that. Different diagnostic tests have different levels of accuracy/specificity for diagnosis thus we may need to reclassify cases into different categories of infection certainty. The paper by Ricardo Ximenes et al proposes an algorithm for classification and will be circulated. Ximenes RAA, Miranda-Filho DB, Brickley EB, Montarroyos UR, Martelli CMT, Araújo TVB, Rodrigues LC, de Albuquerque MFPM, de Souza WV, Castanha PMDS, França RFO, Dhália R, Marques ETA; Microcephaly Epidemic Research Group (MERG). Zika virus infection in pregnancy: Establishing a case definition for clinical research on pregnant women with rash in an active transmission setting. *PLoS Negl Trop Dis*. 2019 Oct 7;13(10):e0007763. doi: 10.1371/journal.pntd.0007763. PMID: 31589611; PMCID: PMC6797234.

In addition to the primary analysis of ZIKV infection status and risk of adverse infant outcomes, there also needs to be consideration of timing of maternal infection and risk, as the risk likely differs depending on gestational age at the time of infection. Other variables with temporal components include:

1. miscarriage or abortions could not occur after 20 weeks of GA.

2. any loss after 20 weeks is called fetal loss (fetal demise, etc).

3. we do have a variable for induced abortion, which is another question, so we should not combine the two questions in the same variable.

4. stillbirth/Intrapartum death are only known at delivery.

* 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.

These variables may be less pertinent to the current analysis given the level of evidence linking ZIKV to developmental birth defects.

* 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? Need to identify and variables that would be important to impute.

EWG members should please review the spreadsheet attached to the meeting invitation to suggest any other exposure variables that should be imputed.

Effect modifiers / confounders / covariates

* Variable “tobacco” – what do values of 3 mean?

The 2020 key variable list only contains values up to “2”; if “3” is listed, this is an error.

* Effect modifier: “Genetic anomalies, metabolic disorders, perinatal brain injury” -> how can we construct this?

Variables pertinent to this question are in the “Pregnant Women” tab, other in “Outcomes” tab

* Effect modifier: “Clinical/subclinical illness” -> how should we construct this?

This will also benefit from a review of the Ximenes et al paper to determine spectrum of severity. The question was posed if this is too generic of a variable. This point can be discussed on the next EWG call.

In general, we should make use of the data we have and conduct sensitivity analyses for inclusion of the data.

* Effect modifier: “Presence and severity of maternal and infant clinical symptoms” -> how should we construct this?

For maternal symptoms, see discussion above.

* Variables symp \_ga and arb\_clindiag\_ga seem to have outliers -> can these be cleaned?

As for the other outlier question above, this should be checked against the 20-20-20 key variables dataset and if plausible, ranges shared with the EWG on the output update.

* 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?