**Describing the harmonisation process**

Most analyses in ICAD will rely, to a greater or lesser degree, on the use of harmonised variables. That is, variables that have been recoded or otherwise transformed from their original format to enable their inclusion in pooled analyses. This document is intended to provide ICAD users with guidance on how to document the process they have followed in creating new harmonised variables. The aim is to provide transparency of methodology for projects that are subsequently published and allow other users to review how existing variables were derived and evaluate whether they meet their needs or whether new variables will be required to address their research questions. This is not intended as a ‘How to’ guide for data harmonisation. Information on issues to consider when conducting retrospective data harmonisation can be found elsewhere.(1)

We recommend that ICAD user’s follow the template outlined below for describing the harmonisation process, amending or adding to the steps described as appropriate. The template comprises two sections: Notes and Tables. Completion of either one or both sections may be appropriate depending upon the variable under consideration. For example, the ‘Tables’ section is well suited to describing the creation of categorical variables. For the creation of harmonised variables that are continuous, it may be possible to convey sufficient information on the process followed within the ‘Notes’ section. Users are advised to review the documents provided for existing harmonised variables. A blank template, including only the main subheadings for each section, is provided on the ICAD website.

**Notes**

The notes section is intended to provide a general overview of what data was available for use in harmonisation, what variables were created and what decisions were made in the process of creating them. The following sub-headings are used:

* Studies (wave) with relevant data (n=X)
* Assessment characteristics: Respondent, Constructs, Timing.
* Variable(s) created
* Studies included in each harmonised variable
* Excluded studies / waves
* Item selection / prioritisation
* Study specific notes
* Missing data

Not all headings will be relevant to all harmonised variables and others can / should be added where they aid clarity. Details on how to complete each section are provided below. Exemplar text (in yellow) is included to facilitate completion of the relevant sections. Guidance on what information should be provided under each heading is provided in *italics*.

**Tables**

The tables provide a more detailed description of the process followed in creating a harmonised variable for each available study / wave. This includes descriptive information on both the source and harmonised variables and any transformations / recoding undertaken. For any study / wave where the harmonisation process is more complex than can be clearly conveyed within the table, a more detailed description should be provided under the ‘Study specific notes’ section of ‘Notes’. The following column headings are used:

* Study / wave
* Source data: Variable name, description, respondent
* Harmonisation: category, processing, summary

The general form of the harmonisation table is provided below, with column headings in bold. A short description (in italics) of what is required under each heading is provided. The notes provided refer to creation of categorical variables, but (where appropriate) the table can be adapted to accommodate continuous variables.

**Harmonisation Notes**

Studies (wave) with relevant data (n=X)

ALSPAC (1,2), CLAN (1,2), HEAPS (1,2), PEACH (1,2,3)

*List under this heading those studies that provided data relevant to the harmonised variable being created. Where a variable will be derived for multiple time-points, include the relevant waves in brackets after the study name.*

Assessment characteristics

Respondent: Parent, Child, Researcher assessed.

*For all studies / waves, list under ‘respondent’ the person(s) who completed the questionnaire or conducted the assessment. This serves to convey the diversity of sources of data.*

Constructs: Years of education completed, Schools attended, Qualifications obtained.

*For all studies / waves, list under ‘construct’ the specific variable or concept assessed. In the example above, the various constructs listed may be used to indicate level of education.*

Timing: No. of waves of assessment, proximity to accelerometry.

*For all studies / waves, note here any relevant considerations related to the timing of assessment. This should include whether assessments were undertaken at multiple time-points and the proximity of assessment to other variables in the analysis. For example, if considering the potential influence of car ownership on children’s physical activity, it is necessary (or at least highly preferable) that both of these constructs were assessed at the same point in calendar time (or that assessment of car ownership precedes activity assessment).*

Variable(s) created

|  |  |
| --- | --- |
| **Name** | **Description / Coding** |
| ICAD\_MotherEducation1 | Up to and including completion of compulsory education (coded 0)  Any post-compulsory education including vocational training (1) Missing (999) |

*Use one row in the table for each new harmonised variable created. Report the variable name, a short label or description where necessary, and the value labels for categorical variables.*

*Studies / waves included in each harmonised variable*

|  |  |
| --- | --- |
| **Name** | **Study** |
| ICAD\_Ethnicity1 | ALSPAC, CHAMPS UK, CHAMPS US, EYHS Denmark, EYHS Portugal, EYHS Norway, EYHS Estonia, IBDS, NHANES 2003-04, NHANES 2005-06, PEACH, Pelotas, SPEEDY, TAAG. |

*Use one row in the table for each new harmonised variable created. In the ‘study’ column, list each study (along with the corresponding wave(s) of assessment where applicable) that provided the necessary data to create the harmonised variable.*

Excluded studies / waves

|  |  |
| --- | --- |
| **Study / wave**  **Variable** | **Rationale** |
| SPEEDY / wave 2  ICAD\_SchoolTravel3 | No information collected / shared on duration of journey to school. |

*Use one row in the table per study / wave. Use this table describe why a particular study (or wave of assessment within a study) was not included in any newly derived variables. Report the study / wave and name of the harmonised variable in the first column and describe the rationale for exclusion in the second column.*

Item selection / prioritisation

* Assuming the same construct was assessed, respondent was prioritised as follows: parent, child.

*In some cases, multiple variables within a study may be relevant to the variable being created. Use this section to describe which variables were used preferentially. For example, it may be that information was provided on a construct by both the child and his / her parent. In such cases, it is necessary to decide which variable to use in creating the harmonised variable.*

Study specific notes

*See notes for previously created variables. Use this section to describe any study-specific issues relevant to creation of the harmonised variable(s). This might include complexities related to the use of a particular questionnaire / assessment method, differences in methodology between waves of the same study or specific decisions made regarding variable selection / prioritisation.*

Missing data

*Provide details here of any general issues or actions taken related to missing data. For example, some researchers may wish to impute missing values or impose a general rule on a maximum permissible amount of missing data for inclusion in any newly derived variables. This section may also be used to notify other ICAD users of particular studies / waves where the scale or nature of missing data may require special consideration.*

**Harmonisation Table**

**Construct**: *Short description of the construct of interest*

**Variable**: *Name of variable created (as listed under ‘Variable(s) created’ in ‘Notes’)*

**Coding**: *Value labels for categorical variables (as listed under ‘Variable(s) created’ in ‘Notes’)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Study / Wave** | **Source data**  *These columns refer to the data submitted from each study in its original format. These data (variables) are subsequently transformed / recoded to derive the harmonised variable* | |  | **Harmonisation**  *These columns provide information on how the harmonised variable was derived and summary information* | | |
| **Variable(s): name(s), respondent, description** | **Summary** |  | **Category** | **Processing** | **Summary** |
| *State the study name and wave (as appropriate) here. Use one row per study/wave*. | *Report the name(s) of the variable(s) in the source dataset (Var’ name:), the respondent and a short description of the construct assessed*. | *Provide descriptive information (summary statistics) for the source variable named in the preceding column.* |  | *Split this cell to provide one row for each category of the harmonised variable. Label each row with the category name and code.* | *Split this cell to provide one row for each category of the harmonised variable. Describe how responses from the source variable were mapped to those of the harmonised variable.* | *Split this cell to provide one row for each category of the harmonised variable. Report summary statistics for the derived variable (e.g. n in each category).* |

**References**

1. Fortier I, Raina P, Van den Heuvel ER, Griffith LE, Craig C, Saliba M, et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. Int J Epidemiol. Oxford University Press; 2016 Jun 6;dyw075.