Project Number: *101057390*

Project Acronym: HappyMums

Project title:

Understanding, predicting, and treating depression in pregnancy to improve mothers and offspring mental health outcomes.

D2.3 Data Harmonization Procedures and Protocols

Research and Innovation Action

HORIZON-HLTH-2021-STAYHLTH-01-02

Work Package:	2
Due date of deliverable:	31/08/2023
Actual submission date:	06/09/2023
Lead beneficiary:	OSR
Contributors:	Benedetta Vai (OSR), Federico Calesella (OSR), Federica Colombo (OSR), Francesco Benedetti (OSR) Noussair Lazrak (UB) Libera Cavaliere (UMIL).
Reviewers:	Annamaria Cattaneo (UMIL)

Disclaimer

Table of Contents

Executive Summary

The deliverable reports the results of the work carried out in *Task 2.3. Data Harmonisation.* This task aims to develop data harmonization procedures and protocols to allow interoperability across the different cohorts that will be used in *HappyMums,* specifically in WP3 *Multimodal biomarkers for depressive symptoms in pregnancy* and WP4 *Biological mediators & environmental moderators of offspring risk and resilience*.

These guidelines have been developed to unify and bridge the different coding standards of target variables between sites to cover:

a) Presence of depressive symptoms, their severity, and diagnoses;
b) Biological data;
c) Lifestyle, psychosocial and environmental factors.

To integrate biological and clinical data, a specific harmonization pipeline has been defined integrating sematic, phenotype and batch effect harmonization, to reduce the impact of between and within centers.

The implementation of the harmonization of variables will be achieved through collaboration between

- Task 2.3 Leader, Ospedale San Raffaele (OSR);
- Leader of Work Packages (WPL) 2,3,4. (respectively, University of Barcelona - UB, Charite - Universitaetsmedizin Berlin - CHARITE, and Erasmus Medical Center- EMC).
- Task leaders (TL) as data harmonization techniques and the evaluation of their applicability and quality are dependent on the specific research questions formulated in the different WP3/WP4 tasks and what studies will be targeted.
- Cohort Owner (CO), who are the partners with direct access to the data.

Three Actions have been defined to obtain harmonization:

- Action 1 – Developing harmonization guideline – (Involved partners: OSR and WPL).
- Action 2 – Developing specific harmonization protocols for each demographic, clinical and biological data in *HappyMums* – (Involved partners: OSR and TL)
- Action 3 – Application of harmonization protocols on cohorts – (Involved partners: TL and CO).

A dedicated folder called *Variable repository* has been created in the private area of the *HappyMmums* website (https://www.happymums.eu/) specifically for this task, where all partners can have access once they register. The area is a workspace where partners can upload and download files and work directly on them with track changes on.

For the moment, the *variable repository* is shared among the WP2, 3 and 4 and related tasks and accessible to all partners of *HappyMums,* since there are no references to actual data and thus no ethical implication. However, if needed, dedicated new folders will be created and specific access restrictions implemented.

The folder contains:

- *Variables overview* files*:* an excel file describing each variable of interest, it will be updated from WPL and TL;
- *BA harmonization* folder: a folder where stored scripts and guidelines for the batch effects harmonization of biological variables;
- *Phenotype harmonization* folder: a folder where there will be stored scripts and guidelines for the harmonization of clinical measures.

Acronyms

Abbreviation	Full term
AU	Aarhus University
CA	Consortium Agreement
CO	Cohort Owner
cRCT	cluster Randomized Clinical Trial
DPO	Data Protection Officer
EMC	Erasmus Medical Center
EU	European Union
FI	Finland
GA	Grant Agreement
GWAS	Genome-Wide Association Study
IRB	Internal Review Board
IT	Italy
KCL	King's College London
OSR	Ospedale San Raffaele
PR	Project Reporting
RIA	Research and Innovation Action
TL	Task Leader
UB	University of Barcelona
UH	University of Helsinki
UMIL	University of Milan
WP	Work Package
WPL	Work package Leader

1. Introduction

HappyMums aims to improve the understanding of the biological mechanisms that underlie the development of depression during pregnancy and its treatment and to investigate the mechanisms that may affect the foetal environmental biology, shaping the risk of the offspring to develop negative outcomes (cognitive, neurodevelopmental and psychiatric symptoms) later in life.

One of the main objectives of *HappyMums* is to capitalize on longitudinal data and biological samples from large population-based birth cohorts and case-control cohort studies with in-depth measures of genetic, biological, environmental, lifestyle and demographic factors (e.g. ethnic background, socio-economic status) in mothers and offspring to:

- Identify risk and protective factors for the development of depressive symptoms in pregnancy and associated biological blood signatures (WP2, WP3).
- Characterize the impact of prenatal maternal depressive symptoms on offspring outcomes across different developmental domains and identify underlying biological mechanisms as well as pre- and postnatal moderators (WP2, WP4).

This means that data integration, standardization and harmonization procedures across cohorts are key, since pulling more data together increases statistical power, allows more advanced subgroup analyses, enhance generalizability of findings and supports cross validation or replication of findings across datasets.

In particular, for these analyses, *HappyMums* will have access to

1. "HappyMums completed studies": Already existing cohorts, where the recruitment has been already completed, biological samples already collected and some of the biological data needed in *HappyMums* already obtained (PREDO, ITU, PRAMD, Berlin Birth Cohort Study, Generation-R BHRCS).
2. "HappyMums ongoing studies" Already existing cohorts, where the recruitment is still ongoing, and biological samples are being collected (PRESeNt and IMPRINT).

Table 1 provides a summary of the different cohorts, with their acronym (if available), and with details on the recruitment status, collection of biological samples and data.

***Table 1**. Cohorts available in HappyMums.*

"HappyMums completed studies"			
Name of the Study (and acronym)	**Country of the Study**	**Cohort Owner**	**Recruitment Status and data availability**
Prediction and prevention of preeclampsia and intrauterine growth restriction (PREDO) Study	Finland	UH	Recruitment already completed; Biological Samples collected. Some of the biological data are generated.
The Intrauterine sampling in early pregnancy study (ITU).	Finland	UH	Recruitment already completed; Biological Samples collected. Some of the biological data are generated
Psychiatry Research and Motherhood - Depression (PRAM-D)	UK	KCL	Recruitment already completed; Biological Samples collected. Some of the biological data are generated
Berlin Birth Cohort Study	Germany	CHARITE	Recruitment already completed; Biological Samples collected. Some of the biological data are generated
Generation R and GenerationR *Next*	Netherlands	EMC	Recruitment already completed; Biological Samples collected. Some of the biological data are generated
Brazilian High Risk Cohort Study (BHRCS)	Brazil	AU	Recruitment already completed; Biological Samples collected. Some of the biological data are generated
"HappyMums ongoing studies"			
Name of the Study (and acronym)	**Country of the Study**	**Cohort Owner**	**Recruitment Status**
Imprint	Finland	UH	Recruitment, data and biological samples collection is ongoing.
PRESeNT Study	Italy	UMIL	Recruitment, data and biological samples collection is ongoing.

Table 2 describes the data already available or to be generated in the *HappyMums* cohorts.

***Table 2**. Data and info in the different cohorts available in HappyMums*

Name	**Partner**	**Infl**	**GWAS**	**DNAm**	**Horm**	**Size**	**Study**	**Medical, Clinical, Sociodemographic, Environmental lifestyle Info**
PREDO & ITU	UH	✓	✓	✓	✓	6000	LC	✓
PRAM	KCL	✓	TBG	✓	✓	400	LC	✓
Berlin Birth Cohort Study	Charite	✓	✓	TBG	TBG	200	LC	✓
GenR	EMC	✓	✓	✓	✓	9,778	LC	✓
GenR Next	EMC	✓	✓	✓	✓	>4000	LC	✓
BHRCS	AU	TBG	✓	✓	TBG	2500	LC	✓
PRESeNT	UMIL	✓	TBG	✓	TBG	200	LC	✓
IMPRINT	UH	✓	TBG	✓		1000	LC	✓

✓ = Already Available or that analyses will be done through other available funding; LC= Longitudinal Cohort; DNAm=DNA methylome; Hormones=Glucocorticoids and Reproductive Hormones; Infl = pro and anti-inflammatory cytokines (by Luminexor Mesoscale). TBG: Data to be generated in HappyMums.

1. Action 1 - Harmonization protocols

Harmonization protocols allow to harmonize demographic, clinical and biological data. They will include:

1. **Semantic harmonization**

Semantic harmonization is the process of combining multiple sources and representations of data into a form where items of data share meaning. Most of the variables of interest can be harmonized only using semantic harmonization, allowing to generate comparable items among cohorts. Semantic harmonization will rely on a standardization of i) variable names (e.g., original label: Education Degree; harmonized label consistent among cohorts: h_education), ii) concept and meaning and iii) unit of measures (Almeida et al., 2021).

Each TL is required:

1. to firstly define research hypotheses.
2. to list the variables that will be used in the analyses.
3. to check if these variables of interest have been already harmonized in the *Variable Overview.*
4. if not already harmonized, to indicate for each variable the following details needed for harmonization:
    a. Concept ID: indicate a progressive number for the new added variable.
    b. Concept name: indicate the name of the variable (eg. height).
    c. Harmonized name: indicate the name of the variable with the prefix h_ (eg. h_height). These names will be used by CO to generate harmonized variables to perform the analyses.
    d. Domain id: indicate the domain of the variables among Demographic, Biological, Clinical.
    e. Description: if necessary, briefly describe the variable.
    f. Unit of Measure: indicate the unit of measure to uniform the datasets (eg. height in cm).
    g. Type of variable: indicate if continuous, categorical or ordinal.
    h. Range or possible value: indicate possible min or max values, or coding for categorical ones (eg. for sex, 0 = female and 1 =male).
    i. Harmonization type: indicate if this variable will rely only on *semantic* or needs further *BA* or *phenotype* harmonizations as described in the following sections.
    j. Support from OSR: indicate 1 if you require support from OSR for harmonizing that variable.

See below an example of details for different variables.

concept_id	concept_name	harmonised_name	domain_id	description	unit_of_measure	type_of_variable	range or possibile value
1	age	h_age	demographic		years	countinuos	>0
2	height	h_height	demographic		cm	countinuos	>0
3	sex	h_sex	demographic			categorical	1=male; 0=female

In case of any issues in harmonizing a variable, TL can complete the *Variables Overview* with details available and flag the last column indicating *Support from OSR* and adding an email address for the person that OSR should contact.

The OSR team will check the table weekly to give support. If no feedback is provided in the following two weeks, concerned partners are requested to contact OSR via email (vai.benedetta@hsr.it and calesella.federico@hsr.it).

By using this pipeline for semantic harmonization, each consortium member will then be able to generate harmonized variables in their dataset by consulting the *Variable Overview*.

2. Batch effects harmonization for biological data

Biological data may be affected by within and between centers batch effects (i.e., systematic technical artifacts), which are known to be sources of bias and variance (Leek et al.,2010; Johnson et al., 2007). Such non-biological effects can negatively affect the consistency and reproducibility of the downstream analyses and findings. The ComBat method was introduced by Johnson et al. in 2007 and has since been shown to be effective in batch effects removal on various types of data, including genomics, transcriptomics, and neuroimaging (Leek et al., 2010; Fortin et al., 2017; Fortin et al., 2017; Yi et al., 2018).

In case of biological data, the TL interested in exploring batch effects harmonization for their biological variables should:

1) Firstly, define research hypotheses.
2) List the biological markers that will be used in the analyses.
3) Collect from partners meta-data on technology and array used to extract biological data, sample size, and their availability to share data. Different scenarios have been envisaged:
    a) Available to send raw pseudo-anonymized data;
    b) Available to upload raw data in a protected server: authorized partners' researchers can see the raw data;
    c) Available to upload raw data in a protected server: but no partners' researchers can see the raw data;
    d) Data analyzable only locally with delivered scripts.

For scenarios b and c, the Project Coordinator (UMIL) and the WP2 Leader (UB) are consulting with their ICT Departments in order to create a shared storing space while ensuring institutional and personal privacy by limiting data integration within the scope of the applied models and preserving further characteristics of the data for its respective owner.

4) Contact OSR with previous collected information and fill the *variable overview* with available information by flagging *Support from HSR* and *Harmonization Type* (Batch effect).

OSR will respond to the TL with a specific harmonization strategy in approximately 2-3 weeks, providing the scripts that CO will run on each cohort. This will require sharing pooled estimates, avoiding the need to transfer data between partners. Various phases and runs will be needed, so OSR will coordinate the run and upload phases. Scripts will allow TL to evaluate if their data can be effectively affected by within and/or between centers batch effects. This can be assessed by running a principal component analysis and evaluating whether the first few principal components differ between cohorts and batches. If data transfer will not be allowed, also in this case, OSR will provide scripts and coordinate CO to run Principal Component Analysis (PCA) locally on each cohort.

In the meantime, TL can proceed in generating datasets that will be used for harmonization by:

5) semantically harmonizing the variables that will be of interest in the analyses (eg. age, sex) (see section 1. Semantic harmonization). These variables will be indeed included in the harmonization model, but only to avoid the erroneous removal of their information.

6) coordinating the creation of overlapping datasets among COs. Datasets must have the same features in the same order, so it will be crucial to map each cohort's dataset with the others (see also Action 3 – Generate harmonized dataset).

7) adding a variable named "batch", indicating the cohort and if multiple batches are present within a cohort (e.g., neuroimaging acquisitions with different scanners, or blood markers estimated with different plates) -> each batch will have the name of the cohort followed by a numbered identifier (eg. GENR_01). In case of no expected within batch effects do not report any number.

Cohort	**Batch**
GENR	GENR
BBCS	BBCS
PREDO	PRED
ITU	ITUU
PRAM	PRAM

PRES	PRES
IMPRINT	IMPR

If successful, harmonization protocols will generate harmonization scripts runnable by CO that will be stored in the *BA Harmonization folder.*

3. Phenotype harmonization

Phenotype harmonization of clinical scales will be performed through a test linking approach, which allows mapping items from different questionnaires to the same underlying construct (Kolen & Brennan, 2014). This approach has been previously applied for the harmonization of neuroticism, extroversion (Jovic et al., 2022; Van Den Berg et al., 2014) and depression clinical scales (Choi et al., 2014; McCabe-Beane et al., 2016), providing a common measure of the same underlying construct across different scales and cohorts.

Harmonization on clinical measures will require the following steps:

1. First, define the construct that you are interested in.
2. Collect meta-data among cohorts in terms of specific instruments used to rate the phenotype (e.g. clinical diagnosis, interviews- SCID I, questionnaires - BDI-13, BSI, etc.);
3. If more than an instrument was used in a cohort, explore if there is a subsample of individuals that was assessed with more than one instrument.
4. Ask CO to report, if they know, validated clinical cut-off for their continuous scales.
5. Contact OSR with previous collected information and fill the *variable overview* with available info flagging *Support from OSR* and *Harmonization Type* (Phenotype).

Considering the provided information, different strategies will be then laid out by OSR:

1. Looking for crosswalk tables for published test linking harmonization (Wahl et al., 2014; McCabe-Beane et al., 2016; Furukawa et al., 2019).
2. If no previous validated harmonization protocols have been developed in literature, or if some scales have not been already harmonized, generating crosswalk tables from ex novo test linking harmonization (possible only if two continuous scales have been administered in at least one subsample of individuals in at least one cohort).
3. Performing categorical harmonization with standard cut-offs.
4. Performing meta-analyses to assess pooled estimates, Q and I2 statistics to check for heterogeneity among cohorts.

To generate crosswalk tables from ex novo test linking harmonization, TLs should check that (Fig. 1):

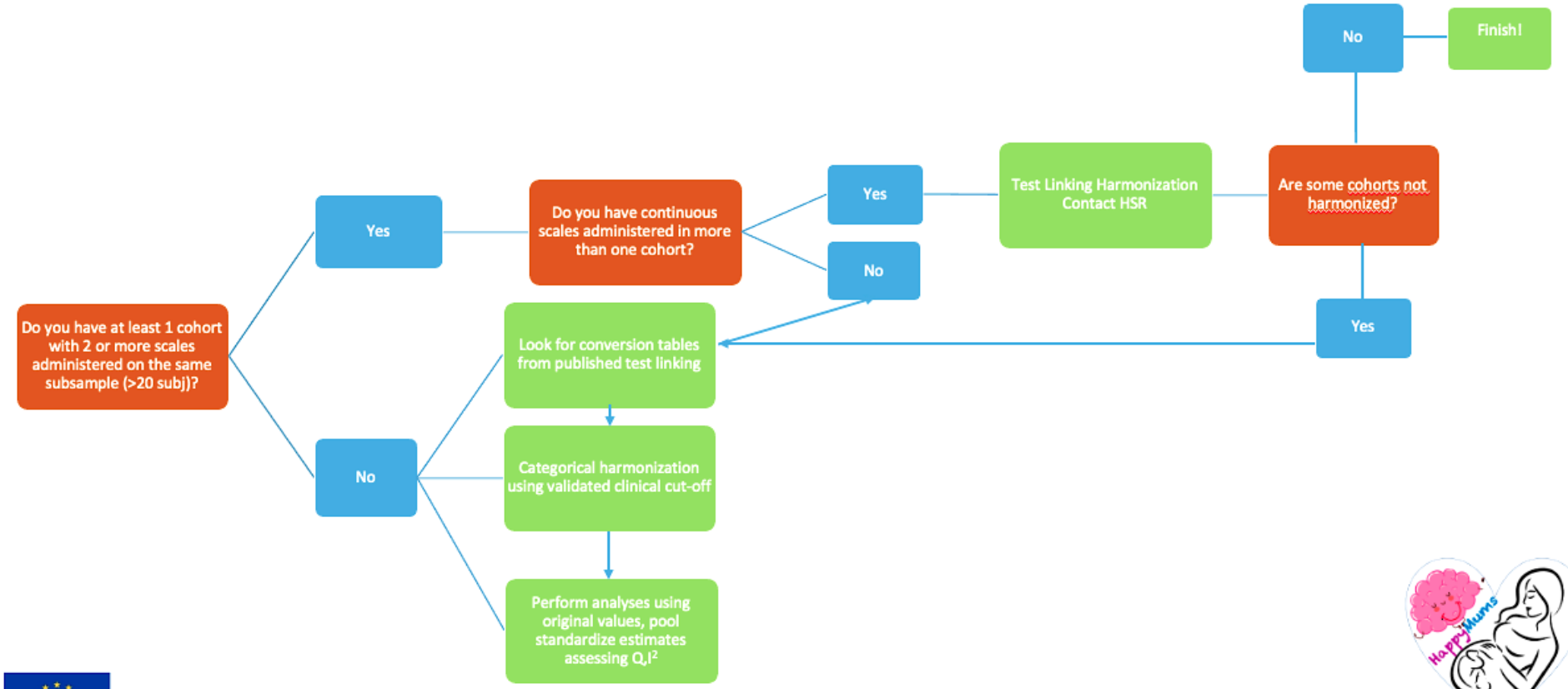
1. At least 1 cohort has at least 2 different scales administered on the same subjects, that will define the linking sample.
2. The linking sample should have a minimum sample size of 50 and it should be representative for the phenotype (scores should cover the min and max obtainable scores to the test, e.g. no depression vs severe depression).
3. At least 1 of the administered scale should be present in at least another cohort.

Then, TLs should contact COs to:

4. Create a database in .csv format in which each row corresponds to a single subject and item scores are reported in columns (e.g. BDI_01; BDI_02, etc.). The database should not contain any missing value.

OSR will deliver runnable scripts to generate crosswalk tables, which will be shared among COs in the *Phenotype harmonization* folder. Only the finally harmonized total scores will be shared among cohorts, thus preventing movement of sensible data.

Figure 1. *Flowchart of the phenotype harmonization pipeline*



Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101057390.

2. Action 2- Harmonization protocols

Following the previous guidelines, each variable of interest can be harmonized through ad hoc generated protocols as specified in the *Variable Repository* accessible to the partners including semantic harmonization, batch effect harmonization with runnable scripts, and crosswalk tables for Phenotype harmonization.

When all the necessary variables will be generated, the TL will contact the CO to move to the next action, generating harmonized datasets.

3. Action 3 – Generate harmonized dataset

When the COs prepare their datasets for data sharing or running the analyses, they are advised to follow these recommendations:

1) Prepare the dataset in comma separated .csv format, with quotes as string delimiter.
2) Files should be organized as subjects-by-features (e.g. subjects in rows, variables in columns).
3) For each variable, check the *Variables Overview* and *Repository* and follow instructions to generate harmonized variables.
4) Name the variables using the labels specified in *variable overview file* (eg. age will be transformed in h_age).
5) Variables should be ordered following the consecutive order in the variable *overview file.*
6) Different times can be entered in subsequent columns.
7) Decimals: use . to indicate decimals units.
8) Participants ID: generate ID for participants 0000 including the following suffix for each cohort (e.g., 0001_GENR)

Cohort	**Suffix**
GENR	_GENR
BBCS	_ BBCS
PREDO	_PRED
ITU	_ITUU
PRAM	_PRAM
PRES	_PRES
IMPRINT	_IMPR

4. Conclusion

HappyMums will bring together longitudinal data from large population-based birth cohorts and case-control cohort studies of perinatal samples, with in-depth measures of genetic, biological, environmental, lifestyle and demographic factors (e.g. ethnic background, socio-economic status) in mothers and offspring to generate key new insights into the impact of prenatal maternal depressive symptoms on offspring health.

The overall objective of the deliverable is to provide a solid foundation for in-depth data harmonization and development and implementation of guidelines for standardized data management and processing protocols across sites.

This is a live document that will be updated according to necessities under the supervision of OSR but in continual collaboration with all consortium partners working with data. These updates will ensure the smooth functioning of data analysis across the different work packages and tasks according to the different research questions.

5. References

Almeida, João Rafael, et al. "A methodology for cohort harmonisation in multicentre clinical research." Informatics in Medicine Unlocked 27 (2021): 100760.

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., & McGrath, P. J. (2018). Harmonization of cortical thickness measurements across scanners and sites. Neuroimage, 167, 104–120.

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., & Gur, R. E. (2017). Harmonization of multi-site diffusion tensor imaging data. Neuroimage, 161, 149–170.

Furukawa TA, Reijnders M, Kishimoto S, Sakata M, DeRubeis RJ, Dimidjian S, Dozois DJA, Hegerl U, Hollon SD, Jarrett RB, Lespérance F, Segal ZV, Mohr DC, Simons AD, Quilty LC, Reynolds CF, Gentili C, Leucht S, Engel RR, Cuijpers P. Translating the BDI and BDI-II into the HAMD and vice versa with equipercentile linking. Epidemiol Psychiatr Sci. 2019 Mar 14;29:e24. doi: 10.1017/S2045796019000088. PMID: 30867082; PMCID: PMC8061209

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics, 11(10), 733–739.

McCabe-Beane, J. E., Segre, L. S., Perkhounkova, Y., Stuart, S., & O'Hara, M. W. (2016). The identification of severity ranges for the Edinburgh Postnatal Depression Scale. Journal of Reproductive and Infant Psychology, 34(3), 293–303. https://doi.org/10.1080/02646838.2016.1141346

Yi, H., Raman, A. T., Zhang, H., Allen, G. I., & Liu, Z. (2018). Detecting hidden batch factors through data-adaptive adjustment for biological effects. Bioinformatics, 34(7), 1141–1147.

Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 Jan;8(1):118-27. doi: 10.1093/biostatistics/kxj037. Epub 2006 Apr 21. PMID: 16632515.

Jović, Miljan, et al. "Harmonized Phenotypes for Anxiety, Depression, and Attention-Deficit Hyperactivity Disorder (ADHD)." Journal of psychopathology and behavioral assessment 44.3 (2022): 663-678.

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking: Methods and practices (3rd ed.). Springer Science + Business Media. https://doi.org/10.1007/978-1-4939-0317-7

Van Den Berg, Stéphanie M., et al. "Harmonization of Neuroticism and Extraversion phenotypes across inventories and cohorts in the Genetics of Personality Consortium: an application of Item Response Theory." Behavior genetics 44 (2014): 295-313.

Wahl I, Löwe B, Bjorner JB, Fischer F, Langs G, Voderholzer U, Aita SA, Bergemann N, Brähler E, Rose M. Standardization of depression measurement: a common metric was developed for 11 self-report depression measures. J Clin Epidemiol. 2014 Jan;67(1):73-86. doi: 10.1016/j.jclinepi.2013.04.019. PMID: 24262771