SoftwareX 16 (2021) 100828

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Original software publication

mics-library: A Python package for reproducible studies on the Multiple Indicator Cluster Survey

Andrea Bizzego [a], Mengyu Lim [b], Gianluca Esposito [a,b,c,*]

[a] Department of Psychology and Cognitive Science, University of Trento, Rovereto, Italy
[b] Psychology Program, School of Social Sciences, Nanyang Technological University, Singapore, Singapore
[c] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore

ARTICLE INFO

Article history:
Received 12 April 2021
Received in revised form 21 September 2021
Accepted 21 September 2021

Keywords:
Low- and middle- income countries
Child development
Publich health
Multiple Indicator Cluster Survey

ABSTRACT

Psycho-sociological research has historically shown a lack of representation towards Low- and Middle Income Countries (LMIC), yet the issues faced by these countries, especially in the domains of child development and public health, are much more severe and prevalent. To close this research gap, the Multiple Indicator Cluster Survey (MICS) is an appropriate and comprehensive large dataset that captures information on LMIC health and human development. We therefore introduce `mics_library`, a tool designed to help researchers using the MICS dataset by allowing data preview, organizing files and extracting relevant data.

Code metadata

Current code version	v1.0
Permanent link to code/repository used for this code version	https://github.com/ElsevierSoftwareX/SOFTX-D-21-00065
Code Ocean compute capsule	
Legal Code License	GPL v3
Code versioning system used	git
Software code languages, tools, and services used	Python v3.9+
Compilation requirements, operating environments & dependencies	The following Python packages are required: numpy, pandas, pyreadstats
If available Link to developer documentation/manual	
Support email for questions	andrea.bizzego@unitn.it

1. Motivation and significance

Low- and Middle-Income Countries (LMIC) is a term coined by the World Bank and World Health Organization (WHO) to group countries with lower gross national income per capita. Unfortunately, economical disparities between countries reflects into different incidences on public health issues and human development.

It is well-established that there are unequal rates of prevalence of child disability distributed globally, with the largest proportions of child disability occurring in LMIC [1]. Furthermore, child development paths in LMIC are different from high-income countries due to parenting and caregiving differences [2–4]. For example, parents belonging to lower socioeconomic statuses often experience greater parenting stress, which leads to poor child outcomes [5]. Additionally, research has suggested that there is substantial psychological and behavioural variation among human populations, due to external factors in child development such as different environments, epigenetics, cultural evolution and internal factors such as differential distribution of genes. Therefore, findings from one population cannot be automatically applied to another [6]. These differences restrict the generalizability of existing child development research, and a means of obtaining findings that are more ecologically valid and generalizable across LMIC is needed. However, much of the research surrounding child development and disability is conducted within cultural contexts that are more aligned with high-income countries [7,8], resulting in disproportionate representation from LMIC in child research. An analysis by Arnett [9] of several top journals

* Corresponding author.
E-mail addresses: gesposito79@gmail.com, gianluca.esposito@unitn.it (Gianluca Esposito).

https://doi.org/10.1016/j.softx.2021.100828

representing diverse areas of psychology revealed that 68% of subjects came from the United States, and a full 96% of subjects were from Western industrialized countries, specifically those in North America and Europe, as well as Australia and Israel, countries with only 12% of the world's population [6].

At the same time, exacerbating the impact of a lack of research, resources dedicated to child development are severely lacking in LMIC. For example, in 2012, only 0.01% of gross national product in 27 sub-Saharan African countries was spent on preschool education [10]. Therefore, LMICs are not represented in child and human development research, and at the same time face a lack of resources in this area. The scientific awareness on this issue is growing, with some initiatives focusing on conducting objective cross-cultural research in child development [11], so as to bridge the research gap and empower children in LMIC [12].

1.1. Multiple indicator cluster survey

A fundamental resource for research on LMIC is represented by the Multiple Indicator Cluster Survey (MICS), a nationally representative household survey developed by UNICEF and administered internationally. A majority of household surveys conducted in LMIC are conducted under MICS [13]. Collected data are open access, so they can be compared across countries, and they can be used across research fields [14]. The MICS program was started by UNICEF in 1995, as a response to the need to develop a household survey in order to improve the evidence base on children's, men's and women's development. The MICS provides a full suite of tools and technical assistance for implementation at national, regional and international levels. After 25 years since its beginning, the MICS includes data on 118 LMIC, and is one of the most reliable sources of quantitative data to enable comparable measures across many nations in the world [14].

The MICS is periodically administered, each administration wave is called round. In general, each MICS round consists of a set of items derived from questions about physical and social conditions related to human and child development, and household management. The items are grouped by questionnaires, each one focusing on a topic, such as: "Questionnaire for Children Under Five", "Household Questionnaire". In turn, questionnaires are divided into modules, each one targeting specific aspects of the main topic, such as the "Discipline module" in "Household Questionnaire", or the "Caregiving module" in the "Questionnaire for Children Under Five"). Each module focuses on specific members of the household within a specific age range. The range of MICS coverage is wide, including indicators on health, development and standard of living. Participants in the MICS are recruited by household units, randomly selected within a hierarchical process that starts from census enumeration areas that are then broken down to segments within each area.

A range of scientific publications are based on the MICS, ranging from topics such as child development, disability and education [15–18], maternal health [19,20] and public health issues such as HIV/AIDS, water and sanitation. For example, specific topics include the monitoring of health in relation to different issues (such as unimproved water supplies [21] or tobacco and alcohol abuse among youth [22]), as well as the investigation of children's education [23] or protection [24] around the world. Other studies focus on the association between disability and parental practices [25–28], child mortality [29] and parenting practices [30].

The access to MICS data files is open, prior registration at mics.unicef.org to obtain the permissions to download and use data for research purposes. License files should be consulted for information about data usage permissions and restrictions for each country.

1.2. Issues with MICS

Notwithstanding the importance of the information provided, the use of the MICS in research is still limited. In fact, the MICS present some issues that complicate its usage, which we summarize into:

1. Format of the data files (.sav), which requires specific (typically, commercial) software to be opened and processed;
2. Complexity of the dataset, which includes hundreds of different files and folders. In addition, the names of files and folders may change between MICS rounds and countries;
3. Information about each participant is spread across different questionnaires and modules;
4. Inconsistencies in the way questions and answers are administered and recoded. For instance, numerical values indicating nominal answers to a multiple choice question might differ across countries, or reflect specific characteristic of one country (e.g.: the educational levels).

1.3. Aims of `mics_library`

`mics_library` aims at facilitating the use of MICS data by addressing these issues. Specifically, it offers functionalities to accomplish the following steps: (a) preview of available data; (b) solve inconsistencies; (c) extract and export datasets.

`mics_library` is meant to be used at the beginning of the MICS data analysis pipeline to extract the information of interest and obtain a coherent and consistent dataset. It can be used to create complex pre-processing pipelines, or by running a sequence of template scripts that only require the user to define the indicators of interest. `mics_library` has been developed while conducting studies on MICS data to investigate the effects of child disabilities on parenting [25–27] and the causes of child mortality [29].

2. Software description

`mics_library` is a Python (v. 3.9+) package that offers several functions to work with MICS data, in particular to select, preprocess and extract the items of interest for a specific study. It also includes utilities and metadata for the advanced customization of the data processing pipeline.

2.1. Software architecture

Functions of mics_library target the MICS dataset and are organized according to the three steps of the data processing pipeline:

- `mics_library.preview` to screen the items available in a specific round;
- `mics_library.recode` to fix inconsistencies in the data;
- `mics_library.loaders` to extract and export data.

Other functions are provided in `mics_library.utils`.

`mics_library` is based on the Python packages `pandas` and `numpy` to manage and process data using DataFrames, and on `pyreadstat` to read the original MICS data files.

2.2. Software functionalities

To comply with `mics_library` functions, MICS data files should be grouped following a tree structure. The root folder should contain the folders of each MICS round (e.g., MICS4, MICS5). The round folders should contain the folders of each country (e.g., Mozambique, Argentina), which, in turn, contain the data files.

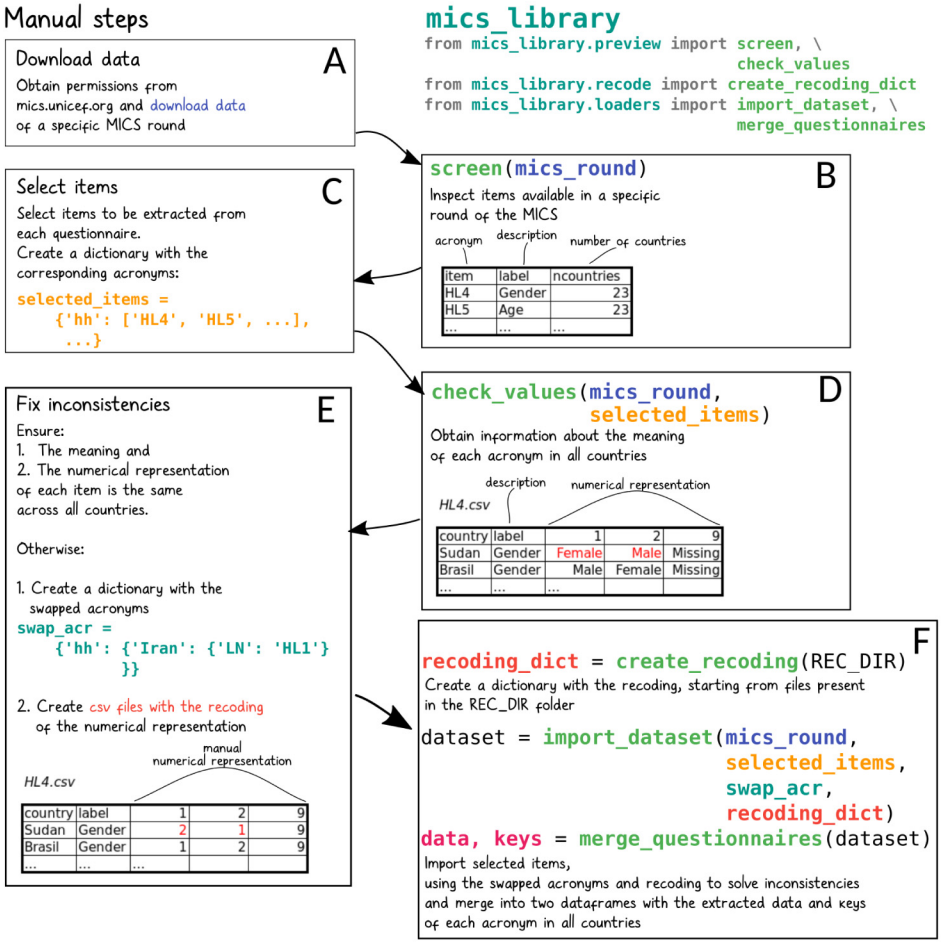

Fig. 1. Pipeline to extract data from the MICS dataset, using `mics_library`. Left: steps that are manually performed (A–C–E); Right: steps that are performed using `mics_library` functions (B–D–F).

`mics_library` can be used to support the three steps required to extract a coherent dataset from the MICS (Fig. 1): (a) preview of available data; (b) solve inconsistencies; (c) extract and export datasets.

Preview

The first step of the pipeline aims at obtaining an overall description of the information available in a round of the MICS. The `mics_library.preview.screen` (Fig. 1B) function scans all datafiles provided in a round of the MICS and generates a DataFrame for each questionnaire, listing the acronyms of available items, with their description and the number of countries that provide the item. The DataFrames are then inspected by the user to shortlist the MICS items that are relevant to the research topic of interest (Fig. 1C).

Then, the `mics_library.preview.check_values` function (Fig. 1D) allows the user to identify inconsistencies in the data. For each selected item, this function provides, for each country, a description of the item and the numerical representation of the answers.

Solve inconsistencies

When the same acronym indicates different questions in different countries, the user should inspect the original data files and identify the correct acronyms that should be used for each country. These can be defined as a dictionary, which will be used by `mics_library` to load the correct acronyms (Fig. 1E).

The numerical representation used by the different country needs to be the same. In case of inconsistencies, the user can use formatted *.csv* files that indicate how the numerical values should be recoded to obtain consistent information. The `mics_library.recode.create_recoding_dict` function automatically reads the *.csv* files and creates a dictionary that will be used by `mics_library` to correct the values (Fig. 1E).

Extract and export

The final step extracts the selected items from the MICS (Fig. 1F). The function `mics_library.loaders.import_dataset` uses the dictionaries with corrected acronyms and numerical representations to consistently extract the selected items; then the function `mics_library.loaders.merge_questionnaires` is used to merge the data from different questionnaires.

Two DataFrames are created: the first with the extracted items; the second with keys that allow linking information between the questionnaires. Keys can be used to both track the same participant across questionnaires and to link information from related participants; for instance, to link mothers and children.

3. Illustrative example

To better describe how `mics_library` can be adopted to leverage on MICS data, we present the code (Fig. 2) to extract a dataset to investigate whether being a child or grandchild of a household's head affects the access to educational resources. We focus on the fifth round of the MICS (years 2013–2017) which involved 47 LMIC countries.

```python
import mics_library

mics_library.set_rootdir(ROOTDIR)

#%% PREVIEW ------------------------------------------------------------
from mics_library.preview import screen, check_values

# screen available items
mics_questionnaires = screen(micsround=5)

# save csv with items of the hh questionnaire
# (to be repeated for all questionnaires)
mics_questionnaires['hh'].to_csv('hh.csv')

# select target indicators
selected_items = {'hh': ['HELEVEL'],
                  'hl': ['HL3'],
                  'ch': ['EC1', 'EC5', 'AG2']}

# check inconsistencies
mics_items = check_values(micsround=5, selected_items)

# save csv with numerical representation of item HELEVEL
# (to be repeated for all items)
mics_items['hh']['HELEVEL'].to_csv('HELEVEL.csv')

#%% SOLVE INCONSISTENCIES ----------------------------------------------

# MANUALLY create csv files with the correct numerical representation

from mics_library.recode import create_recoding_dict

# automatically create the recoding dictionary
# using the csv files with the correct numerical representation
recoding_dict = create_recoding_dict(RECODING_DIR)

# manually create dictionary with the correct acronyms
# {Country : {TARGET_INDICATOR : CORRECT_INDICATOR}}
swap_acr = {'ch' : {'Mexico' : {'LN': 'UF4'}},
            'hh' : {'Panama' : {'SL9B': 'CD9'}}}

#%% EXTRACT AND EXPORT -------------------------------------------------
from mics_library.loaders import import_dataset, merge_questionnaires

dataset = import_dataset(micsround=5, selected_items,
                         swap_acr, recoding_dict)

data, keys = merge_questionnaires(dataset)

data.to_csv('data.csv')
keys.to_csv('keys.csv')
```

Fig. 2. Example of a Python script using `mics_library` to extract a dataset from the MICS.

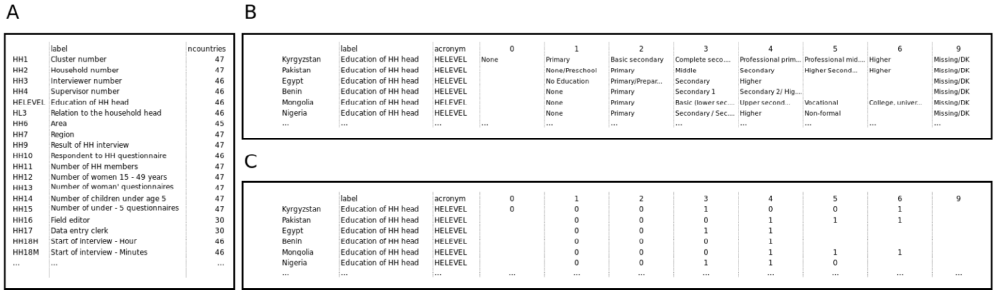

Fig. 3. Examples of DataFrames created in a typical pipeline using `mics_library` to extract data from the MICS. A: Result of `mics_library.preview.screen` with available items in a MICS questionnaire; B: Result of `mics_library.preview.check_values` with the description and numerical representation of an item in each country; C: DataFrame used by the `mics_library.recode.create_recoding_dict` to fix inconsistencies in the numerical representations of the item. Empty cells indicate that a numerical value is not used for the specific country and do not constitute an issue.

After inspecting the DataFrames (Fig. 3A) obtained from `mics_library.preview.screen` (Fig. 2, Line 6), we define a dictionary (Fig. 2, Line 11) to select the following items: "Education level of the household head" (HELEVEL) from the Household questionnaire (hh); "Relation to the household head" (HL3) from the Household Listing (hl); "Number of books available for the child" (EC1), "Child attends early education programme" (EC5) and "Age of Child" (AG2) from the Children Under 5 questionnaire (ch). EC1 and EC5 can be used to indicate the access to resources, HL3 to categorize the children as "child" or "grandchild", while HELEVEL and AG2 will be considered as possible confounders.

Looking at the DataFrames (Fig. 3B) obtained from `mics_library.preview.check_values` (Fig. 2, Line 13) we

note some inconsistencies. For instance, HELEVEL associates different education levels to the same numerical values. We correct these inconsistencies by editing the created DataFrames, replacing the meaning of the answer with the correct numerical representation (Fig. 3C). At this stage we can already implement some data analytics decision. For instance, to simplify the analysis, we avoid considering all education levels and only focus on whether or not the household head has completed the secondary level of education. For this reason we need recode as "1" the numerical representations in the cells that indicate a secondary level of instruction or higher, and as "0" all other levels (Fig. 3C).

`mics_library.recode.create_recoding_dict` (Fig. 2, Line 22) is used to create a dictionary with the recoded numerical representations, based on the formatted *.csv* files present in a specified folder (`RECODING_DIR`). Another source of inconsistency is the use of the same acronym to indicate different questions. This can be solved by creating a dictionary with the correct acronyms (Fig. 2, Line 25).

In the last step, we extract the selected items, with the correct acronyms and numerical representation (`mics_library.loaders.import_dataset`, Fig. 2, Line 28) and merge all DataFrames (`mics_library.loaders.merge_questionnaires`, Fig. 2, Line 29). After this, the resulting DataFrames can be used within a standard statistical pipeline, or exported for usage with other tools (Fig. 2, Lines 30–31).

4. Impact

A search of the keywords "multiple indicator cluster survey" on PubMED, returns 189 studies published between the years 2002–2020. This indicates that the MICS is a well-known and important dataset for the scientific community. Given the high amount of resources and skills invested in its management and administration, the MICS still represents one of the key resources to monitor and investigate human development in LMIC; in particular, of children and women.

However, the MICS is a complex dataset: for instance the MICS5 round is composed of more than 400 files and folders, differences exist in the administration of the MICS in different countries and in the way data are reported.

`mics_library` aims at simplifying the extraction of coherent dataset from the MICS and can be a key tool to expand the research based on this dataset.

Using `mics_library` allows a standardized and reproducible data extraction process, which can be easily accomplished by users with low computation or programming skills. By significantly reducing the time required to extract consistent datasets from the MICS, `mics_library` allows researchers to focus on the extraction of new knowledge from the MICS. For instance focusing on the design of the study, on new computational approaches, or interpretation of the results.

`mics_library`, being an open-source and non-commercial package, can be a valuable resource in terms of reproducibility of the results. First, users can share the code of the data extraction and preprocessing steps, with intermediate results: typically these steps are manually performed, based on hardly reproducible interactions with a software user interface, and, sometimes, annotated on separate files. Second, by allowing the export of data to non-proprietary file formats, the use of `mics_library` is expected to favour the adoption of open-source pipelines for the data analysis, for instance based on Python or R. In turn, these pipelines can be shared; thus boosting the development of new research and validation of the results. Finally, being a reference tool for studies on the MICS, `mics_library` is open to contributions from research teams, and future releases may improve the functionalities and add specialized processing functions.

The impact of `mics_library` is not limited to practical aspects. In fact, by facilitating the study of parenting and child development in LMIC, `mics_library` contributes to shed new light on how similar aspects come into play in High Income Countries as well. `mics_libary` can finally be used by policy makers and NGOs to inform and drive operational decisions and on-field projects in LMIC.

5. Conclusions

`mics_library` is a Python package, developed to standardize and facilitate the use of data from the MICS. In particular, the package aims at providing functions to extract consistent datasets: `mics_library` efficiently simplifies and regulates the steps that are required to solve data inconsistencies, allowing the user to focus on the identification of the information of interest and on the downstream data analysis. We expect the use of `mics_library` will significantly impact the research on human development in LMIC, based on the MICS, by favouring the reproducibility of the results and the development of new studies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

G.E. was supported by grants from the NAP SUG, Singapore (M4081597, 2015–2021). A.B. was supported by a Post-doctoral Fellowship within the programme framework "Dipartimenti di Eccellenza", Ministry of University, Italy.

References

[1] Department of Health UK Government. Valuing people: A new strategy for learning disability for the 21st century. UK: HM Stationery Office London; 2001.
[2] Fatimilehin Iyabo A, Nadirshaw Zenobia. A cross-cultural study of parental attitudes and beliefs about learning disability (mental handicap). Mental Handicap Res 1994;7(3):202–27.
[3] Gaad Eman. Cross-cultural perspectives on the effect of cultural attitudes towards inclusion for children with intellectual disabilities. Int J Incl Educ 2004;8(3):311–28.
[4] Wong Siu Yung, Wong Thomas KS, Martinson Ida, Lai Alice C, Chen WJ, He YS. Needs of Chinese parents of children with developmental disability. J Learn Disabil 2004;8(2):141–58.
[5] Noel Melanie, Peterson Carole, Jesso Beulah. The relationship of parenting stress and child temperament to language development among economically disadvantaged preschoolers. J. Child Lang 2008;35(4):823–43.
[6] Henrich Joseph, Heine Steven J, Norenzayan Ara. The weirdest people in the world? Behav Brain Sci 2010;33(2–3):61–83.
[7] Azhari Atiqah, Truzzi Anna, Neoh Michelle Jin-Yee, Balagtas Jan Paolo M, Tan HuiAnn Hannah, Goh PeiLin Pamela, et al. A decade of infant neuroimaging research: What have we learned and where are we going? Infant Behav Dev 2020;58:101389.
[8] Henrich Joseph, Heine Steven J, Norenzayan Ara. Most people are not WEIRD. Nature 2010;466(7302):29.
[9] Arnett Jeffrey Jensen. The neglected 95%, a challenge to psychology's philosophy of science. Am Psychol 2009;64(6):571–4.
[10] UNICEF. Early childhood development. 2021, https://www.unicef.org/early-childhood-development.
[11] Bornstein Marc H. Handbook of cultural developmental science. Psychology Press; 2014.
[12] Ki-Moon Ban. Children and the millennium development goals : progress towards a world fit for children. UNICEF; 2007.
[13] Hancioglu Attila, Arnold Fred. Measuring coverage in MNCH: tracking progress in health for women and children using DHS and MICS household surveys. PLoS Med 2013;10(5):e1001391.

[14] Khan Shane, Hancioglu Attila. Multiple indicator cluster surveys: delivering robust data on children and women across the globe. Stud Fam Plan 2019;50(3):279–86.
[15] Bornstein Marc H, Putnick Diane L, Bradley Robert H, Deater-Deckard Kirby, Lansford Jennifer E. Gender in low- and middle-income countries. Monogr Soc Res Child Dev 2016;81(1).
[16] Bornstein Marc H, Putnick Diane L, Bradley Robert H, Lansford Jennifer E, Deater-Deckard Kirby. Pathways among caregiver education, household resources, and infant growth in 39 low-and middle-income countries. Infancy 2015;20(4):353–76.
[17] Bornstein Marc H, Hendricks Charlene. Screening for developmental disabilities in developing countries. Soc Sci Med 2013;97:307–15.
[18] Gottlieb Carissa A, Maenner Matthew J, Cappa Claudia, Durkin Maureen S. Child disability screening, nutrition, and early learning in 18 countries with low and middle incomes: data from the third round of UNICEF's Multiple Indicator Cluster Survey (2005–06). Lancet 2009;374(9704):1831–9.
[19] Cuartas Jorge, Jeong Joshua, Rey-Guerra Catalina, McCoy Dana Charles, Yoshikawa Hirokazu. Maternal, paternal, and other caregivers' stimulation in low-and-middle-income countries. PLoS One 2020;15(7):e0236107.
[20] Lansford Jennifer E, Zietz Susannah, Putnick Diane L, Deater-Deckard Kirby, Bradley Robert H, Costa Megan, et al. Men's and women's views on acceptability of husband-to-wife violence and use of corporal punishment with children in 21 low-and middle-income countries. Child Abuse Negl 2020;108:104692.
[21] Geere Jo-Anne L, Hunter Paul R. The association of water carriage, water supply and sanitation usage with maternal and child health. A combined analysis of 49 multiple indicator cluster surveys from 41 countries. Int J Hyg Environ Health 2020;223(1):238–47.
[22] Wasil Akash R, Venturo-Conerly Katherine E, Shinde Sachin, Patel Vikram, Jones Payton J. Applying network analysis to understand depression and substance use in Indian adolescents. J Affect Disord 2020;265:278–86.
[23] Graetz Nicholas, Friedman Joseph, Osgood-Zimmerman Aaron, Burstein Roy, Biehl Molly H, Shields Chloe, et al. Mapping local variation in educational attainment across Africa. Nature 2018;555(7694):48–53.
[24] Cuartas Jorge, McCoy Dana Charles, Rey-Guerra Catalina, Britto Pia Rebello, Beatriz Elizabeth, Salhi Carmel. Early childhood exposure to non-violent discipline and physical and psychological aggression in low-and middle-income countries: national, regional, and global prevalence estimates. Child Abuse Negl 2019;92:93–105.
[25] Bizzego Andrea, Lim Mengyu, Schiavon Greta, Esposito Gianluca. Children with developmental disabilities in low-and middle-income countries: More neglected and physically punished. Int J Environ Res Public Health 2020;17(19):7009.
[26] Bizzego Andrea, Lim Mengyu, Schiavon Greta, Setoh Peipei, Gabrieli Giulio, Dimitriou Dagmara, et al. Child disability and caregiving in low and middle income countries: Big data approach on open data. Res Dev Disabil 2020;107:103795.
[27] Bizzego Andrea, Lim Mengyu, Dimitriou Dagmara, Esposito Gianluca. The role of the family network when raising a child with a disability in low-and middle-income countries. Disabilities 2021;1(1):58–68.
[28] Hendricks Charlene, Lansford Jennifer E, Deater-Deckard Kirby, Bornstein Marc H. Associations between child disabilities and caregiver discipline and violence in low-and middle-income countries. Child Dev 2014;85(2):513–31.
[29] Bizzego Andrea, Gabrieli Giulio, Bornstein Marc H, Deater-Deckard Kirby, Lansford Jennifer E, Bradley Robert H, et al. Predictors of contemporary under-5 child mortality in low-and middle-income countries: a machine learning approach. Int J Environ Res Public Health 2021;18(3):1315.
[30] Jeong Joshua, McCoy Dana Charles, Fink Günther. Pathways between paternal and maternal education, caregivers' support for learning, and early child development in 44 low-and middle-income countries. Early Child Res Q 2017;41:136–48.