unicefData: Trilingual Library for UNICEF SDMX Indicators

João Pedro Azevedo
UNICEF
Division of Data, Analytics, Planning and Monitoring
3 United Nations Plaza
New York, NY 10017
jpazevedo@unicef.org

Abstract. This article introduces `unicefdata`, a Stata package for accessing over 733 child welfare indicators from the UNICEF Data Warehouse via the SDMX API. The package provides discovery commands (search, list, info), data retrieval with flexible filtering (countries, years, disaggregations), and multiple output formats (long, wide, wide_indicators). Automatic dataflow detection from indicator codes eliminates the need for manual API navigation. Part of a trilingual ecosystem with R and Python implementations, `unicefdata` enables reproducible research on child health, nutrition, education, protection, and WASH indicators. The package requires `yaml.ado` for metadata parsing and is available via SSC.

Keywords: st0001, unicefdata, SDMX, UNICEF, child monitoring, custodianship, open data

1 Introduction

The United Nations Children's Fund (UNICEF) maintains one of the world's most comprehensive databases on child welfare, covering health, nutrition, education, protection, HIV/AIDS, and water/sanitation (WASH) (UNICEF 2024). The UNICEF Data Warehouse uses the Statistical Data and Metadata eXchange (SDMX) standard (SDMX 2021), an ISO-certified framework for exchanging statistical information. While powerful, direct SDMX API interaction requires knowledge of dataflow structures, dimension codes, and RESTful query syntax—barriers for researchers focused on substantive analysis rather than data engineering.

The `unicefdata` package removes these barriers by providing a Stata interface modeled after `wbopendata` (Azevedo 2011). Users specify indicators, countries, and time periods in familiar syntax; the package handles API queries, dataflow detection, and data transformation. Discovery commands enable exploration without prior knowledge of SDMX structures (SDMX 2021). The package is part of a trilingual ecosystem with R (`get_unicef()`) and Python (`unicef_api`) implementations sharing identical function names and parameter structures (UNICEF Division of Data, Analytics, Planning and Monitoring 2025a,b).

The package includes comprehensive help files accessible via `help unicefdata` and `help unicefdata_sync`. This article provides conceptual overview and extended examples; the help files document all options and stored results.

1.1 Relationship to other tools

Direct SDMX API access offers full control but requires knowledge of dataflow IDs, dimension codes, and REST query syntax. `unicefdata` abstracts this complexity behind Stata-friendly options.

UNICEF Data Portal (web) is ideal for quick downloads and visual exploration, but it does not provide reproducible, scriptable workflows.

R (`get_unicef`) and Python (`unicef_api`) clients expose the same interface as `unicefdata`. Choose the language that fits your analytical stack; all three are designed for parity to ease cross-team collaboration.

1.2 Requirements

The package requires Stata 14.0 or higher and depends on `yaml.ado` for parsing metadata files. The `yaml` package provides robust YAML reading/writing capabilities and can be installed from SSC:

```
. ssc install yaml, replace
```

No additional dependencies or external software are required. The package works entirely within Stata's native environment.

2 Data model: Indicators and dimensions

2.1 Indicator structure and definitions

The UNICEF Data Warehouse follows the Statistical Data and Metadata eXchange (SDMX) standard, an ISO 20922-certified framework for exchanging statistical information. Within SDMX, an *indicator* is a time-series observation of a specific phenomenon (e.g., under-5 mortality rate, prevalence of stunting, immunization coverage). Each indicator is uniquely identified by a code (e.g., `CME_MRY0T4` for under-5 mortality) and belongs to a *dataflow*, which is a logical grouping of related indicators.

The UNICEF data warehouse currently maintains 733+ indicators organized across 11 thematic dataflows:

- **CME** (Child Mortality Estimates) - 39 mortality indicators
- **NUTRITION** - 112 nutrition indicators (stunting, wasting, underweight)
- **WASH** - 57 water/sanitation indicators

- **EDUCATION** - 38 education indicators
- **HIV/AIDS** - 38 HIV-related indicators
- **IMMUNISATION** - 18 immunization coverage indicators
- And 6 others (MNCH, CAUSE_OF_DEATH, CHLD_PVTY, PT, ECON, GENDER)

Each indicator is categorized by thematic domain, enabling discovery and cross-cutting analysis of child welfare across health, nutrition, education, and protection domains.

Governance note: UNICEF's custodial and co-custodial arrangements for major child-related series shape available disaggregations and provenance in SDMX responses; see Section 3 for a brief overview.

2.2 Dimensions and attributes

Conceptual framework

At the heart of the UNICEF data warehouse is a conceptual distinction between how data are *structured for analysis* and how data are *contextualized and interpreted.* An indicator (such as "under-5 mortality rate") is not a single number but rather a collection of observations, each representing a specific combination of context and reference.

Table 1: Sample SDMX Observations: Raw API Response Structure

Country	**Year**	**Sex**	**Wealth**	**Value**	**Lower**	**Upper**	**Src**	**Stat**
(REF_AREA)	**(TIME)**	**(SEX)**	**(QUINTILE)**	**(OBS)**	**(CI)**	**(CI)**	**(ATT)**	**(ATT)**
BGD	2019	F	_T	30.2	28.5	31.8	UN_IGME	A
BGD	2020	F	_T	29.4	27.8	30.8	UN_IGME	A
BGD	2019	M	_T	34.5	32.6	36.2	UN_IGME	A
BGD	2020	M	_T	33.6	31.9	35.1	UN_IGME	A
BGD	2019	_T	Q1	42.2	38.2	46.3	UN_IGME	A
BGD	2020	_T	Q1	41.0	37.0	45.1	UN_IGME	A
BGD	2019	_T	Q2	37.1	33.7	40.7	UN_IGME	A
BGD	2020	_T	Q2	36.1	32.7	39.5	UN_IGME	A

Note: Each row represents one SDMX observation. **Dimensions** (Country, Year, Sex, Wealth Quintile) define the observation's coordinates and enable filtering. **Attributes** (confidence intervals, source, status) provide context but are not used for query filtering. Value is under-5 mortality rate per 1,000 live births.
Source: UNICEF SDMX API, indicator CME_MRY0T4, retrieved January 6, 2026.

For example, the under-5 mortality rate for Bangladesh in 2020 disaggregated by sex yields three observations: the total rate (`_T` = 31.5 deaths per 1,000 live births), the rate for males (`M` = 33.6), and the rate for females (`F` = 29.4). The rate varies not just by country and year, but also by demographic group (sex, age), socioeconomic status (wealth quintile), and geography (urban vs. rural). Each dimension includes a total category (`_T`) alongside disaggregated values, enabling both aggregate and stratified analysis. These *structural dimensions* are essential for analysis because they define the axes along which data can be filtered, aggregated, and compared.

From an SDMX data modeling perspective, the columns in Table 1 map directly to the data structure definition (DSD): `REF_AREA`, `TIME`, `SEX`, and `QUINTILE` are dimensions that jointly identify the observation key; `OBS` is the primary measure; `Src` (often `SOURCE`) and `Stat` (often `OBS_STATUS`) are observation-level attributes that travel with each value. The DSD binds these concept roles to code lists so that all SDMX responses are schema-valid and interoperable across agencies (SDMX 2021). This distinction motivates the interface design in `unicefdata`: dimensions are exposed as filters because they define the retrieval key, while attributes remain metadata for interpretation and quality assessment.

Table 2 (Table 2) shows how these concepts play out for a single indicator (CME_MRY0T4) in Bangladesh. Wealth quintiles are available only for the total sex category, and sex-specific estimates are available only at the total wealth level. The empty cells in the sex rows make it visually clear that the SDMX source does not publish intersections of sex by wealth for this series—a limitation users must respect when requesting disaggregations.

Table 2: Under-5 Mortality Rate for Bangladesh in 2020 by Sex and Wealth Quintile

Sex	Total	Q1	Q2	Q3	Q4	Q5
	(`_T`)	(`Q1`)	(`Q2`)	(`Q3`)	(`Q4`)	(`Q5`)
	(All)	(Poorest)				(Richest)
Total (`_T`)	31.5	41.0	36.1	32.0	27.1	21.5
Male (`M`)	33.6	–	–	–	–	–
Female (`F`)	29.4	–	–	–	–	–

Note: Values are deaths per 1,000 live births.
Wealth quintile disaggregation available only for total (both sexes combined).
Sex-specific rates (Male, Female) available only as totals across all wealth quintiles.
Source: UNICEF SDMX API, retrieved January 6, 2026.

However, an observation also carries metadata that clarifies its meaning and reliability: the source survey (e.g., DHS, census), the confidence interval around the estimate, whether the value is observed or modeled, and the date of collection. These *attributes* provide context and quality assessment but are not used for filtering or disaggregation. A user querying for all observations of stunting prevalence in Ethiopia does not specify "include DHS data but exclude MICS data" at the query level; instead, the data returned carry attributes indicating the source, allowing the user to filter or weight results post-hoc.

The `unicefdata` package exposes dimensions as command-line filtering options (because users commonly need to select specific subsets of data), while attributes are stored in metadata files and YAML configuration for reference and interpretation.

Dimensions

Dimensions define how observations are disaggregated. The `unicefdata` package exposes the most commonly used dimensions as filtering options:

Geographic dimension - Specifies the `country` or regional aggregate. The package automatically classifies entities as countries (e.g., `BRA`) or aggregates (e.g., `SSA` for Sub-Saharan Africa), enabling selective filtering.

Time dimension - Specifies the `period` (year or year-month combination). The package automatically converts sub-annual periods to decimal years (e.g., June 2020 becomes 2020.5) for consistent time-series analysis.

Sex dimension (`sex`) - Values: `_T` (total/both sexes), `M` (male), `F` (female). Some indicators lack sex disaggregation.

Age dimension (`age`) - Age group codes vary by indicator. Common values: `0T4` (under-5), `0T17` (under-18), `15T49` (reproductive age females).

Wealth dimension (`wealth`) - Wealth quintiles: `Q1`–`Q5`. Indicates income/asset-based household wealth stratification. Available only for survey-based indicators.

Residence dimension (`residence`) - Values: `URBAN`, `RURAL`, or `_T` (total). Distinguishes urban-rural disparities for select indicators.

Maternal education dimension (`maternal_edu`) - Education level of mother (typically for child-related indicators). Values depend on the survey protocol.

Attributes

Attributes are non-structural metadata that contextualize observations without serving as filtering dimensions. The package stores attributes in YAML metadata files but does not expose them as command-line options:

Data source - The survey, administrative system, or estimation model producing the observation (e.g., "Demographic and Health Survey", "UN IGME").

Confidence intervals - Upper and lower bounds (where available) for uncertainty quantification, critical for estimates from mortality models.

Data type - "Observed" (from surveys/administrative systems) or "Estimated" (from models, e.g., UN mortality models).

Geographic classification - Whether the entity is a country or regional aggregate (automatically applied by the package).

Temporal metadata - Survey year, collection period, or model reference year.

2.3 Indicator-to-dataflow mapping and automatic detection

Unlike direct SDMX API access, which requires users to know both the indicator code *and* the dataflow it belongs to, the `unicefdata` package maintains a YAML-based index mapping all 733+ indicators to their parent dataflows. This enables automatic dataflow detection.

When a user specifies an indicator code (e.g., CME_MRY0T4), the package:

1. Queries the local indicator-to-dataflow index (_unicefdata_indicators.yaml)
2. Retrieves the associated dataflow ID (CME)
3. Constructs the SDMX query URL with appropriate filter parameters
4. Submits the query to the UNICEF SDMX API
5. Automatically falls back to alternative dataflows if the primary returns HTTP 404

This design prioritizes user convenience: researchers can request indicators by code alone, without understanding SDMX dataflow semantics. The indicator index is synchronized every 30 days, or less as per manual request, via the unicefdata_sync command, ensuring new indicators and definitions remain current.

2.4 Output data structure

By default, the package returns data in *long format*, one row per country-year-indicator combination with the following columns:

countrycode ISO3 country code (e.g., BRA)

country Country name (e.g., Brazil)

indicator Indicator code (e.g., CME_MRY0T4)

year Time period (decimal year after conversion)

value Numeric observation

sex Sex disaggregation (if applicable)

age Age disaggregation (if applicable)

wealth Wealth quintile (if applicable)

residence Urban/rural (if applicable)

maternal_edu Maternal education level (if applicable)

Dimension columns are omitted for indicators lacking that dimension, and missing values (typically represented as . in Stata) indicate unavailable observations. Users can reshape data to wide format (years as columns) or wide-indicators format (indicators as columns) using the format options.

2.5 Key features

- Access to 733+ UNICEF indicators across 11 thematic dataflows
- Discovery commands: search indicators, list dataflows, display metadata
- Automatic dataflow detection from indicator codes using YAML metadata
- Flexible filtering: countries, years, sex, age, wealth quintiles, residence
- Multiple output formats: long (default), wide (years as columns), wide_indicators (series as columns)
- Geographic classification: automatic country vs aggregate region detection
- Metadata synchronization via `unicefdata_sync` command
- XML-to-YAML conversion for offline schema analysis (`unicefdata_xmltoyaml`)

2.6 Disaggregation availability by dataflow

Compatibility Matrix (v1.0, updated 2025-12-20)

The following table shows which disaggregation dimensions are available for each dataflow (as of 2025-12-20):

Table 3: Disaggregation availability by dataflow (v1.0, updated 2025-12-20)

Dataflow	SEX	AGE	WEALTH	RESIDENCE	MATERNAL_EDU
CAUSE_OF_DEATH	✓	✓	–	–	–
CCRI	–	–	–	–	–
CHILD_RELATED_SDG	✓	✓	✓	✓	–
CHLD_PVTY	✓	–	–	✓	–
CME	✓	–	✓	–	–
CME_CAUSE_OF_DEATH	✓	–	–	–	–
CME_COUNTRY_PROFILES_DATA	–	–	–	–	–
CME_DF_2021_WQ	✓	–	✓	–	–
CME_SUBNATIONAL	✓	–	✓	–	–
COVID	✓	✓	✓	✓	–
COVID_CASES	✓	✓	–	–	–
DM	✓	✓	–	✓	–
DM_PROJECTIONS	✓	✓	–	✓	–
ECD	✓	✓	✓	✓	✓
ECONOMIC	–	–	–	–	–
EDUCATION	✓	–	✓	✓	–
EDUCATION_FLS	✓	–	–	–	–
EDUCATION_UIS_SDG	✓	–	✓	✓	–
FUNCTIONAL_DIFF	✓	✓	✓	✓	–
GENDER	✓	✓	–	✓	–
GLOBAL_DATAFLOW	✓	✓	–	–	–
HIV_AIDS	✓	✓	✓	✓	–
IMMUNISATION	–	✓	–	–	–
MG (Migration)	–	✓	–	–	–
MNCH	✓	✓	✓	✓	✓
NUTRITION	✓	✓	✓	✓	✓
PT (Child Protection)	✓	✓	✓	✓	–
PT_CM (Child Marriage)	✓	✓	✓	✓	–
PT_CONFLICT	✓	✓	–	–	–
PT_FGM	–	✓	✓	✓	–
SDG_PROG_ASSESSMENT	–	–	–	–	–
SOC_PROTECTION	✓	–	✓	✓	–
WASH_HEALTHCARE_FACILITY	–	–	–	✓	–
WASH_HOUSEHOLDS	–	–	✓	✓	–
WASH_HOUSEHOLD_MH	✓	✓	–	✓	–
WASH_HOUSEHOLD_SUBNAT	–	–	✓	✓	–
WASH_SCHOOLS	–	–	–	✓	–

Notes: ✓ = Dimension available for disaggregation; – = Dimension not available.
CME_SUBNAT_* dataflows (country-specific subnational) all support SEX and WEALTH_QUINTILE.
MATERNAL_EDU includes both `MATERNAL_EDU_LVL` and `MOTHER_EDUCATION` dimension names.
Availability does not guarantee data exists for all values; use API to check actual data coverage.

3 Data governance and custodianship

The UNICEF Data Warehouse is governed by the UNICEF Office of the Chief Statistician, which serves as custodian or co-custodian for most child-related SDG indicators and other global child monitoring series. Because UNICEF produces or co-produces the vast majority of indicators in the warehouse, definitions, code lists, and quality standards are harmonized at the point of production. When indicators originate from partner agencies, the warehouse ingests them via SDMX while preserving agency-specific source attributes so that provenance is transparent to users. This custodianship model explains the consistency users see in dimension code lists (for example, sex, wealth quintile, residence) and why `unicefdata` can rely on stable SDMX structures across dataflows.

Custodianship arrangements: key examples

UNICEF's custodianship and co-custodianship span several flagship initiatives. A brief overview helps users understand why specific dimensions, release cadences, and attribution conventions differ across series:

UN Inter-agency Group for Child Mortality Estimation (IGME) — Co-led by UNICEF, IGME produces child mortality indicators (for example, under-five, infant). IGME sets methods, validates inputs, and releases modelled estimates with uncertainty bounds. In SDMX responses, observation attributes typically include source (IGME) and status (modelled), with dimensions such as sex available but intersections with wealth generally not published.

Joint Malnutrition Estimates (JME) — A partnership including UNICEF produces global and country estimates for stunting, wasting, and underweight. JME governs age-group definitions and survey inclusion criteria, leading to harmonized age dimensions and standard uncertainty reporting across nutrition series.

WHO/UNICEF Joint Monitoring Programme (JMP) — For water, sanitation, and hygiene (WASH), JMP defines service ladders and urban–rural residence classifications. This governance is reflected in the residence dimension and the categorical code lists found in WASH dataflows.

WHO/UNICEF Estimates of National Immunization Coverage (WUENIC) — Immunization coverage indicators (for example, DTP3) follow WUENIC's methods and validation rules. Sub-annual timing and administrative versus survey sources are captured via observation attributes, while the primary disaggregation is typically age-appropriate cohort totals rather than wealth or sex intersections.

Across these initiatives, UNICEF's custodial or co-custodial role ensures coherence at the concept and code-list level, while SDMX attributes preserve provenance (agency, status, method) for transparency. Users should interpret disaggregation availability in light of each initiative's methodological stance—for example, whether sex-by-wealth intersections are conceptually supported or withheld to protect statistical reliability.

Additional initiatives

Beyond IGME, JME, JMP and WUENIC, several inter-agency programmes and custodianship contexts shape indicator definitions and disaggregation availability:

UNAIDS Global AIDS Monitoring (GAM) — Annual monitoring of commitments in the Political Declaration on HIV and AIDS provides standardized indicators and reporting tools. UNICEF, as a co-sponsor of UNAIDS, engages on child and adolescent HIV outcomes. The `HIV_AIDS` dataflows reflect alignment with GAM concepts (UNAIDS 2025).

UIS SDG 4 monitoring — The UNESCO Institute for Statistics (UIS) is the custodian for SDG 4 indicators (education), including out-of-school rates and learning metrics. Related dataflows (for example, `EDUCATION_UIS_SDG`) adopt UIS definitions and code lists (UNESCO Institute for Statistics 2025).

UNFPA–UNICEF programmes on Child Marriage and FGM — Joint programmes focus on ending harmful practices and complement statistical monitoring. The `PT_CM` and `PT_FGM` dataflows align to UNICEF Data statistics (UNICEF 2023, 2026c).

Child protection domains (Child labour, Violence against children) —

UNICEF’s child protection portfolio includes indicators and evidence on child labour and violence against children, informing programme design and monitoring in related `PT` dataflows (UNICEF 2026a,b).

Checking Indicator Disaggregations

```
* Stata: Check what disaggregations are supported
unicefdata, info(CME_MRY0T4)

* Output shows:
*  Supported Disaggregations:
*    sex:          Yes (SEX)
*    age:          No
*    wealth:       Yes (WEALTH_QUINTILE)
*    residence:    No
*    maternal_edu: No
```

4 The unicefdata command

4.1 Syntax

Data retrieval:

`unicefdata , indicator(`*string*`)` [[`options`]]

`unicefdata , dataflow(`*string*`)` [[`options`]]

Discovery commands:

`unicefdata , categories` [`verbose`]

`unicefdata , flows` [`detail verbose`]

`unicefdata , dataflow(`*string*`)`

`unicefdata , search(`*string*`)` [`limit(`#`) dataflow(`*string*`)`]

`unicefdata , indicators(`*string*`)`

`unicefdata , info(`*string*`)`

4.2 Main options

`indicator(`*string*`)` specifies indicator code(s) to download (for example, `CME_MRY0T4` for under-5 mortality rate). Multiple indicators can be specified space-separated.

`dataflow(`*string*`)` specifies the SDMX dataflow ID (for example, `CME` or `NUTRITION`). Required only if not using `indicator()`. The package can retrieve all indicators from a dataflow.

`countries(`*string*`)` specifies ISO3 country codes, space or comma separated (for example, `BRA USA IND`). Omit to retrieve all countries and regional aggregates.

`year(`*string*`)` specifies the year range or list. Supports single year (`2020`), range (`2015:2023`), or comma-separated list (`2015,2018,2020`).

`circa` requests the closest available year for each country when the exact year is unavailable. Useful for harmonizing panel data with missing years.

4.3 Disaggregation filters

`sex(`*string*`)` specifies the sex dimension: `_T` (total, default), `F` (female), `M` (male), or `ALL` (retrieve all sex disaggregations).

`age(`*string*`)` specifies the age group filter. Default is `_T` (total). Use `ALL` to retrieve all age disaggregations. Age group codes vary by indicator.

`wealth(`*string*`)` specifies the wealth quintile filter. Default is `_T` (total). Use `ALL` to retrieve all quintiles (Q1–Q5).

`residence(`*string*`)` specifies the residence dimension: `URBAN`, `RURAL`, or `_T` (default, total).

`maternal_edu(`*string*`)` specifies the maternal education level filter. Default is `_T` (total).

4.4 Output options

`long` requests long format (default): one row per country-year-indicator combination. This is the standard format for panel data analysis.

`wide` requests wide format with years as columns. Variables will be named `yr2020`, `yr2021`, etc.

`wide_indicators` requests wide format with indicators as columns. Useful for multi-indicator comparative analysis.

`latest` requests only the most recent value per country-indicator combination. Returns a cross-section.

`mrv(`#`)` requests the N most recent values per country. For example, `mrv(5)` returns the five most recent observations.

dropna drops observations with missing values. By default, missing values are retained.

simplify keeps only essential columns: country code, year, and value. Removes dimension columns.

addmeta(*string*) adds geographic metadata columns. Options include region, income_group, and continent.

clear replaces data in memory. Required unless memory is empty.

4.5 Discovery options

categories List all thematic categories with indicator counts.

flows List available SDMX dataflows. Use with detail for full descriptions.

dataflow(name) Display dataflow schema (dimensions and attributes).

search(keyword) Search indicators by keyword. Use limit(#) to control results.

indicators(dataflow) List all indicators in a specific dataflow.

info(indicator) Display detailed metadata for an indicator.

4.6 Stored results

unicefdata stores the following in r():

r(N) Number of observations downloaded

r(countries) Number of unique countries

r(years) Number of unique years

r(indicators) Number of unique indicators

r(dataflow) Dataflow ID used for query

Discovery commands (info, dataflow) return additional metadata in scalars and macros.

4.7 Common errors and solutions

extbfIndicator not found

If you see "indicator XXX not found", verify the code is spelled correctly. Use exttunicefdata, search(keyword) to locate the indicator, then retry with the exact code.

extbfDisaggregation not available

Some indicators lack certain dimensions (for example, CME has sex but not age). Run `unicefdata, info(INDICATOR)` to check available disaggregations and adjust filters accordingly. Passing `ALL` to a missing dimension returns an empty result set.

extbfNetwork timeout or rate limit

If the SDMX API is slow or returns HTTP 429 (rate limited), increase retries with extttmax_retries(5) or rerun after a short pause. Large queries can be batched by reducing the country list or years, then concatenating results in Stata.

5 Examples

5.1 Discovery: Finding indicators

Before downloading data, explore available indicators and metadata.

extbfCategories (with counts)

```
. unicefdata, categories
  category                indicators
------------------------------------
  Child Mortality (CME)            39
  Nutrition                       112
  Water and Sanitation             57
  Education                        38
  HIV/AIDS                         38
  ... (11 categories listed)
```

extbfKeyword search

```
. unicefdata, search(mortality) limit(5)
indicator      name
---------------------------------------------
CME_MRY0T4     Under-five mortality rate
CME_TMY0T4     Infant mortality rate
CME_ANN_DTHS   Annual deaths (under-5)
... (5 results displayed)
```

extbfList indicators in a dataflow

```
. unicefdata, indicators(CME)
indicator      name
---------------------------------------------
CME_MRY0T4     Under-five mortality rate
CME_MRY0T17    Under-17 mortality rate
... (39 CME indicators listed)
```

extbfIndicator metadata

```
. unicefdata, info(CME_MRY0T4)
Indicator: CME_MRY0T4 (Under-five mortality rate)
```

```
Dataflow: CME (Child Mortality Estimates)
Dimensions: sex (YES), age (NO), wealth (YES), residence (NO)
Source: UNICEF/IGME modelled estimates
```

5.2 Basic data retrieval

Download under-5 mortality for selected countries:

```
. unicefdata, indicator(CME_MRY0T4) countries(BRA IND CHN) year(2015:2023) clear
. describe
. summarize
```

The `clear` option replaces data in memory. Omitting `countries()` retrieves all countries. The default output is long format with variables: `countrycode`, `year`, `value`, `indicator`, `sex`, `age`, etc.

5.3 Geographic filtering

Filter by a targeted set of countries or regions:

```
. unicefdata, indicator(CME_MRY0T4) countries(NGA ETH KEN UGA TZA) year(2020) clear
. tabulate countrycode
```

Use `countries()` to focus on priority geographies without downloading the full global dataset.

5.4 Temporal aggregation (latest / MRV)

Retrieve the most recent observation per country or a rolling window of recent years:

```
. * Latest value only
. unicefdata, indicator(CME_MRY0T4) countries(IND BGD PAK) latest clear

. * Three most recent values per country
. unicefdata, indicator(CME_MRY0T4) countries(IND BGD PAK) mrv(3) clear
```

5.5 Disaggregated analysis

Retrieve wasting prevalence by wealth quintile and sex:

```
. unicefdata, indicator(NT_ANT_WHZ_NE2) countries(IND BGD PAK) ///
  year(2020) wealth(ALL) sex(ALL) clear wide_indicators
```

The `wealth(ALL)` option retrieves all quintiles, `sex(ALL)` gets both male and female disaggregations, and `wide_indicators` reshapes data with one column per indicator-dimension combination.

5.6 Wide format (time series)

Create year columns for panel analysis:

```
. unicefdata, indicator(CME_MRY0T4) countries(USA GBR FRA DEU) ///
  year(2000:2023) clear wide
. list countrycode yr2000 yr2010 yr2020 yr2023
```

5.7 Multiple indicators

Download several related indicators:

```
. unicefdata, indicator(CME_MRY0T4 CME_MRY0T17 CME_TMY0T4) ///
  countries(ETH KEN UGA) year(2020) clear
. table indicator, statistic(mean value) statistic(sd value)
```

5.8 Data validation and quality checks

After downloading data, verify structure and missingness:

```
. unicefdata, indicator(CME_MRY0T4) countries(BRA USA) year(2015:2023) clear
. describe
. codebook countrycode year value
. misstable summarize
. duplicates report countrycode year indicator sex age wealth
```

Key checks:

- confirm variable types and labels (ISO3 codes as strings, years numeric);
- inspect missing patterns to ensure gaps align with known data coverage;
- detect duplicates before reshaping or merging with other datasets;
- review value ranges for plausibility given the indicator definition.

5.9 Metadata enrichment

Add regional and income group classifications:

```
. unicefdata, indicator(ED_CR_L2OR3_PPP) year(2020) ///
  addmeta(region income_group) dropna clear
. tabulate region
. bysort income_group: summarize value
```

6 Implementation notes

6.1 Automatic dataflow detection

Unlike direct SDMX API usage, users need not know which dataflow contains an indicator. The package maintains a YAML index mapping 733+ indicators to their parent dataflows. When `indicator()` is specified, the package:

1. Looks up the indicator in `_unicefdata_indicators.yaml`
2. Retrieves the associated dataflow ID
3. Constructs the appropriate SDMX query
4. Falls back to alternative dataflows if the primary fails (HTTP 404)

This design mirrors `wbopendata`'s approach, prioritizing user convenience over strict SDMX protocol adherence.

6.2 Performance and API limits

The UNICEF SDMX API enforces rate limits to protect service availability. `unicefdata` reduces the risk of throttling by:

- batching indicators into single API requests when possible;
- retrying failed calls with backoff when the server returns transient errors;
- caching metadata locally to avoid redundant metadata fetches.

Large queries (for example, all countries × many years × all disaggregations) can take several minutes and produce sizable datasets. To improve responsiveness:

- start with a small country list or recent years, then expand once verified;
- use `mrv()` or `latest` when only the most recent values are needed;
- split very large pulls into batches (by region or indicator group) and append in Stata.

6.3 YAML metadata architecture

The package uses five YAML files stored in `src/_/`:

- `_unicefdata_indicators.yaml`: Indicator-to-dataflow mapping (733+ indicators)

- `_unicefdata_dataflows.yaml`: Dataflow definitions and schemas (11 dataflows)
- `_unicefdata_countries.yaml`: Country codes and names (ISO 3166-1 alpha-3)
- `_unicefdata_regions.yaml`: Regional aggregates (geographic/economic groups)
- `_unicefdata_codelists.yaml`: Dimension value codes (sex, age, wealth, etc.)

These files are generated via `unicefdata_sync` and can be inspected/edited manually. YAML was chosen for its human-readable format and widespread use in API documentation. The `yaml.ado` package (Azevedo 2025) provides robust parsing, handling nested structures, lists, and unicode characters. This dependency is automatically installed when `unicefdata` is installed via SSC.

Inspecting and editing metadata files

YAML files live in the Stata PLUS ado directory. To locate and inspect them:

```
. sysdir PLUS
. ! notepad "_/_unicefdata_indicators.yaml"
```

The metadata files start with a small header documenting version and update date, followed by key-value pairs for indicators, dataflows, and codelists. Users may add notes or annotations, but should avoid changing field names or indentation. If a file becomes corrupted, rerun `unicefdata_sync` to regenerate clean copies from the API.

6.4 Geographic classification

The package automatically classifies geographic entities as:

- **Country**: Sovereign states and territories (e.g., BRA, IND)
- **Aggregate**: Regional/economic groupings (e.g., SSA for Sub-Saharan Africa)

This classification enables filtering (e.g., `if geo_type == "country"`) and prevents mixing countries with aggregates in regression analysis.

6.5 Metadata synchronization

The package maintains synchronized YAML metadata files that cache information from the UNICEF SDMX API. Users can update local metadata using the `unicefdata_sync` command:

```
. unicefdata_sync, dataflows
. unicefdata_sync, indicators
```

Synchronization downloads the latest indicator and dataflow definitions from the UNICEF SDMX API and updates the local YAML cache files (located in the installed ado plus directory). This is useful when new indicators are added or metadata changes. By default, the package refreshes cached metadata every 30 days automatically; manual synchronization ensures immediate updates. The `unicefdata_xmltoyaml` utility command converts SDMX XML schemas to YAML format for offline inspection, leveraging the same `yaml.ado` infrastructure used for metadata reading. For details on metadata structure and generation, refer to the METADATA_GENERATION_GUIDE.md in the package repository.

7 Conclusion

The `unicefdata` package provides Stata users with streamlined access to UNICEF's comprehensive child welfare database. By abstracting SDMX API complexity into familiar Stata syntax, the package enables researchers to focus on substantive analysis rather than data engineering. Discovery commands facilitate exploration without prior knowledge of indicator codes or dataflow structures. Automatic dataflow detection, flexible filtering, and multiple output formats accommodate diverse research workflows.

The package's YAML-based metadata architecture supports offline operation and manual inspection/editing of indicator definitions. Integration with the broader trilingual ecosystem (R, Python) enables cross-platform reproducibility—analyses written in one language can be translated to others with minimal syntax changes.

Future development will focus on: (1) expanded disaggregation options as UNICEF adds new dimensions; (2) improved handling of temporal misalignment across indicators; (3) integration with spatial data packages for geographic visualization; and (4) performance optimization for large multi-indicator queries.

The `unicefdata` package demonstrates that API-based data access need not sacrifice ease of use. By following `wbopendata`'s design philosophy while adapting to SDMX-specific requirements, the package lowers barriers to evidence-based research on child welfare worldwide.

Acknowledgments

The author thanks the UNICEF Data Warehouse team for maintaining the SDMX API infrastructure, and the many users who have contributed feedback and feature suggestions.

About the author

João Pedro Azevedo is Chief Statistician at the United Nations Children's Fund (UNICEF), Division of Data, Analytics, Planning and Monitoring, New York. His research interests include poverty measurement, child welfare indicators, and open data infrastructure for development research. He is the author of `wbopendata` and principal architect of the trilingual unicefData ecosystem.

8 References

Azevedo, J. P. 2011. WBOPENDATA: Stata Module to Access World Bank Databases. Statistical Software Components S457234, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s457234.html.

———. 2025. yaml: Stata Module for Reading and Writing YAML Files. Statistical Software Components, Boston College Department of Economics. Available from SSC.

SDMX. 2021. Statistical Data and Metadata eXchange (SDMX) Technical Specifications. Technical standard, Statistical Data and Metadata eXchange Initiative. https://sdmx.org/.

UNAIDS. 2025. Global AIDS Monitoring 2025: Indicators and guidance. Website. Accessed 2026-01-06. https://www.unaids.org/en/global-aids-monitoring.

UNESCO Institute for Statistics. 2025. UNESCO Institute for Statistics — Data for the Sustainable Development Goals. Website. Accessed 2026-01-06. https://www.uis.unesco.org/en.

UNICEF. 2023. Child marriage. Website. Last updated July 2023; Accessed 2026-01-06. https://www.unicef.org/protection/child-marriage.

———. 2024. *UNICEF Data Warehouse.* United Nations Children's Fund. https://data.unicef.org/.

———. 2026a. Child labour. Website. Accessed 2026-01-06. https://www.unicef.org/protection/child-labour.

———. 2026b. Violence against children. Website. Accessed 2026-01-06. https://www.unicef.org/protection/violence-against-children.

———. 2026c. What is female genital mutilation? Website. Accessed 2026-01-06. https://www.unicef.org/protection/female-genital-mutilation.

UNICEF Division of Data, Analytics, Planning and Monitoring. 2025a. *get_unicef: R Client for the UNICEF SDMX API.* https://github.com/unicef-drp/unicefData.

———. 2025b. *unicef_api: Python Client for the UNICEF SDMX API.* https://github.com/unicef-drp/unicefData.

Supplementary materials

The software repository (https://github.com/unicef-drp/unicefData) includes example scripts in R, Python, and Stata, test suites, and complete documentation. The library is available on CRAN (R), PyPI (Python), and SSC (Stata).

Suggested citation: Azevedo, João Pedro. 2025. "unicefData: Trilingual library for accessing UNICEF SDMX indicators." UNICEF Division of Data, Analytics, Planning and Monitoring.