Countdown to 2030
Strengthening Global, Regional and Country Analytic Capacity on Women’s, Children’s & Adolescents’ Health

Producing reliable national and subnational health statistics with a focus on maternal, newborn and child health and nutrition

Analysis workshop for sub-Saharan Africa
Countdown to 2030 / APHRC / GFF / UNICEF / WHO
June 13-17, 2022

Abdoulaye Maïga, PhD
SESSION #1 – PREPARING THE DATA SETS
Methodological Steps Data Preparation

1. Compile and standardize monthly district level data
2. Run data quality checks
3. Correct dataset
4. Analysis health system configuration, service provision (level, trends, distribution, coverage)

1. Reporting completeness
2. Consistency
3. Outliers and missings

1. Adjustment for completeness
2. Correction with a clear audit trial
3. Correction of outliers & missings

Clean dataset for analysis (${country}_master_adjusted_dataset.dta)

DQ report card
District & Primary Admin Unit Profiles for Annual Reviews
RHIS Data and Indicators

• Data consists of monthly service use numbers by district from January 2017 to December 2021
  – Extracted from DHIS-2 in each country
• Additional information on monthly completeness rate, population, region, area (urban/rural), health system strength data
Content Excel template

- **Name of Excel spreadsheet**
  - DHIS2_dataset_Countryname
    - e.g., “DHIS2_dataset_Tanzania”

- **Service_data_1** (ANC, Delivery, PNC, FP)
- **Service_data_2** (Vaccination, OPD, IPD)
- **Service_data_3** (Mortality)
- **Reporting_completeness** (All services by reporting forms)
- **Population_data** (Total and age-specific)
Indicators

- ANC 1st visit
- ANC 4th visit
- IPT 2\textsuperscript{nd} dose
- Institutional delivery
- Skilled birth attendant
- C-section
- PNC 48h
- Family planning: new visits / acceptors
- Family planning: revisits

- BCG given (infants)
- Pentavalent vaccination 1\textsuperscript{st} (infants)
- Pentavalent vaccination 3\textsuperscript{rd} (infants)
- Measles 1\textsuperscript{st} dose (infants)
- Number of OPD visits in under-5
- Total number of OPD visits
- Number of IPD admissions in under-5
- Total number of IPD admissions

Deaths in facilities
- Stillbirths (fresh)
- Stillbirths (macerated)
- Under-5 deaths
- Maternal deaths

Population
- Total population
- Total under-1
- Total under-5
- Total births
- Total live births
- Total women 15-49 years
- Annual population growth rate
Usual data issues

- Making changes to the template
  - Inserting new worksheets/removing existing worksheets
  - Inserting new columns/removing existing columns
  - Renaming worksheets
  - Reshaping template in wide format
  - Merging cells
- Different numbers of districts across worksheets
- Different spellings of district names across worksheets
  - e.g., “Nairobi”, “ Nairobi”, “Nairoby”, “NAIROBI”, “nairobi”, “Nai-robi”,
- (Invisible) extra spaces for string variables
- Different year and month format
- Duplicates
- Missing data
- Using 0 in lieu of missing data and vice versa
- Etc.

Follow general guidance in the « READ ME » as well as worksheet-specific instructions.

Fill out sheets completely and properly.
Data processing & quality checks scripts

- Stata do-files
  - 1_Code_RHIS_Data_Preparation
    - Import and transform Excel data into Stata format
    - Merge data and create new variables
  - 2_Code_RHIS_DQA_Completeness
    - Assess completeness issues
    - Adjust for incomplete reporting
  - 3_Code_RHIS_DQA_Internal_consistency
    - Assess internal consistency (outliers) and missingness
    - Correct internal consistency and missingness
    - Generate summary data quality scores for DQ report card
    - Generate clean and adjusted dataset for analysis
      ($\{country\}_master_adjusted_dataset.dta$)
Guidance before using the scripts

- Create a folder dedicated to the analysis which will include inputs and outputs data
- Rename your input Excel file: DHIS2_dataset_Countryname
  - e.g., DHIS2_dataset_Kenya
- Read and follow the instructions at the top of each do.file
- Carry out required changes according to the instructions
  - Change working folder directory according to location of input data on your computer
  - Set country name
  - Set first year (2017 default)
  - Set last year (2021 default)
  - Set threshold for low reporting (90 default)
  - Set adjustment factor for incomplete reporting (k-value) par service (0.25 default)
Expected outcomes from #1 for report/poster

**Administrative organization**
- Number of provinces / regions / counties
- Number of districts

**Health facilities**
- Number of health facilities in country
- Availability of data on core health professionals
- Availability of data on hospital beds

**Facility data analysis period**
- First month and year with health facility data
- Last month and year with health facility data

<table>
<thead>
<tr>
<th>Indicators available with facility data for the analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antenatal care first visit</td>
</tr>
<tr>
<td>Antenatal care 4th visit</td>
</tr>
<tr>
<td>IPT 2nd dose (malaria)</td>
</tr>
<tr>
<td>Institutional delivery or skilled birth attendant</td>
</tr>
<tr>
<td>Caesarean Section</td>
</tr>
<tr>
<td>Postnatal care</td>
</tr>
<tr>
<td>Family planning new and revisits</td>
</tr>
<tr>
<td>BCG vaccination</td>
</tr>
<tr>
<td>Pentavalent / DPT first dose</td>
</tr>
<tr>
<td>Pentavalent / DPT third dose</td>
</tr>
<tr>
<td>Measles vaccination</td>
</tr>
<tr>
<td>Stillbirths (fresh / macerated)</td>
</tr>
<tr>
<td>Maternal deaths</td>
</tr>
<tr>
<td>OPD visits children under 5 years</td>
</tr>
<tr>
<td>IPD admissions children under 5 years</td>
</tr>
<tr>
<td>Under 5 deaths in health facilities</td>
</tr>
</tbody>
</table>

**Population-based surveys (3 most recent health surveys)**

<table>
<thead>
<tr>
<th>Name of survey</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Population projection data in DHIS2**

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total population for every year</td>
<td>Yes / No</td>
</tr>
<tr>
<td>Live births for every year</td>
<td>Yes / No</td>
</tr>
<tr>
<td>Population under 1 year for every year</td>
<td>Yes / No</td>
</tr>
</tbody>
</table>
SESSION #2 – ASSESSMENT AND ADJUSTMENT FOR INCOMPLETE REPORTING
Assessment and adjustment for incomplete reporting by health facilities

<table>
<thead>
<tr>
<th>Service</th>
<th>2017</th>
<th>2018</th>
<th>2019</th>
<th>2020</th>
<th>2021</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family planning</td>
<td>12</td>
<td>11</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>ANC</td>
<td>19</td>
<td>12</td>
<td>8</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Delivery</td>
<td>19</td>
<td>12</td>
<td>8</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Vaccination</td>
<td>20</td>
<td>22</td>
<td>21</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>OPD visits</td>
<td>29</td>
<td>27</td>
<td>21</td>
<td>21</td>
<td>8</td>
</tr>
<tr>
<td>IPD admissions</td>
<td>20</td>
<td>22</td>
<td>21</td>
<td>21</td>
<td>22</td>
</tr>
</tbody>
</table>

Percentage of districts with low reporting rate (<90%) by service and by year, Tanzania

Low reporting rate (<90%)
Assessment and adjustment for incomplete reporting

- Incomplete reporting can have a major effect on levels and trends of coverage and other statistics derived from health facilities
  - If we do not consider completeness, we assume that no services are provided by the non-reporting facilities

- There are two levels of incompleteness to consider
  - Facility level: facilities that did not report data
  - Service level: expected level of services among facilities that did not report data

- Adjustment must account for both dimensions of incompleteness

- Adjustment depend on assumptions about the number of service outputs (pregnancy care, vaccinations, etc.) provided at non-reporting facilities compared to those that reported.
  - None? Some? About half as the reporting facilities? The same as the reporting facilities?
Assessment and adjustment for incomplete reporting

- Adjustment for incomplete reporting formula is expressed as follows:

\[
N_{(adjusted)} = N_{reported} + N_{reported} \times \frac{1}{(1/c)-1} \times k
\]

where \(N\) = number of service outputs, \(c\) = reporting completeness, \(k\) = adjustment factor.

- \(k=0\)  no services in non-reporting facilities
- \(k=0.25\) some services, but much lower than reporting facilities
- \(k=0.5\) half the rate compared to reporting facilities
- \(k=0.75\) nearly as much as reporting facilities
- \(k=1.0\) same rate of services as reporting facilities
Assessment and adjustment for incomplete reporting

- Choice of k is based on considerations by analysis experts in country:
  - Does reporting link to supplies? (e.g., vaccines)
  - Does not reporting mean stockouts?
  - How good is private facility reporting and what role do they play in service provision such as deliveries or vaccination?
  - Are large hospitals reporting and what role do they play in service provision?
Assessment and adjustment for incomplete reporting

- Double check the effects of the adjustment by comparing the reported number to the adjusted number for incomplete reporting.
SESSION #3 – ASSESSMENT AND CORRECTIONS OF EXTREME OUTLIERS AND INTERNAL CONSISTENCY; DATA QUALITY SCORECARD
Internal consistency checks over time & adjustments of extreme outliers

- In general, we expect limited year-to-year variation in the reported numbers of interventions, especially for interventions with high coverage (e.g., ANC1, DPT1)
- If there is a lot of variation and there is no good explanation, data are considered “noisy” and may have serious data quality issues
- For annual data, an extreme outlier is defined as any number in the dataset higher or lower than 5 STD from the median absolute deviation (MAD) calculated from the preceding 3 years. They were corrected by imputing a value based on the median value of the calendar year. Similar imputation for missing values.

\[
\text{Outlier if } \quad \text{Median} - 1.4826 \times 5 \times \text{MAD} < X_i > \text{Median} + 1.4826 \times 5 \times \text{MAD}
\]

\[\begin{align*}
\text{LB} &= \text{Median} - 1.4826 \times 3 \times \text{MAD} \\
\text{UB} &= \text{Median} + 1.4826 \times 3 \times \text{MAD}
\end{align*}\]

where \(X_i\) is the value of the observation for a particular time period (year) and the MAD is defined as the median absolute deviation (MAD = median(|\(X_i - X^\sim|)), where \(X^\sim\) is the median of the three preceding years).

\[\text{For a normal distribution, one standard deviation from the mean is about 1.4826 MADs (Hampel X84 rule)}\]
Internal consistency checks over time & adjustments of extreme outliers

Assessment of anc1 outliers over time by district, Zambia

anc1 after correcting outliers over time by district, Zambia
Internal consistency between services

- Internal consistency of services is typically assessed between the number of events for ANC1 to Penta1 and for Penta1 to Penta3.
- Consistency calculated as the ratio ANC1 / Penta1 numbers; and ratio Penta1/Penta3 numbers.
  - Questionable data when ratio out of 1.0 and 1.5.
  - % of districts with ANC1/Penta1 (Penta1/Penta3) ratio between 1.0 and 1.25 (85% or higher used threshold).
- This metric can also be calculated as the absolute difference between the expected and the reported ratios of the two indicators.
  - The larger the difference the lower the quality of the reported data (below 5 → good quality; 5-14.9 → moderate; 15+ → poor quality).
  - Required external data from household survey to calculate the expected ratio.
  - There are no recent household surveys in certain countries and level of disaggregation cannot be appropriate.
  - We expect some variation, as the survey has sampling error which affects the Penta1 – Penta3 values, so the expected ratio is only a general indication.
Internal consistency between services

- Scatter diagrams with one intervention on the X-axis and one on the Y-axis can also be helpful to examine district outliers
  - Provides a good visual of the correlation between service data
  - Pearson correlation coefficient can be added
  - Presence of potential outliers can be easily verified from the graph
Internal consistency between services

- Scatter points show districts for which data quality is questionable as a good alignment between data points is expected.
- In-depth data check is recommended to confirm any inconsistencies which should be corrected with a clear audit trail.

Comparison of adjusted numbers of ANC1 and Penta1 by year, Tanzania

Comparison of adjusted numbers of Penta1 and Penta3 by year, Tanzania

R-squared = 0.9027

R-squared = 0.8824
Internal consistency checks over time & adjustments of extreme outliers

- Fluctuations in service delivery may not necessarily be the result of data quality issues

- Important to consider:
  - Changes in programmatic activities such as intensified campaign, stock-outs (e.g., vaccination)
  - Seasonality of diseases of OPD visits or IPD admissions (e.g., malaria)
  - Impact of contextual factors such as population migration (e.g., recent displacements or refugees)

- Good knowledge of context and program data for interpreting trends data

- Account for expected annual increase of service due to population growth and improvements in service coverage
Expected outcomes from #2 & 3 for report/poster

Report on adjustments and corrections

- Report the threshold used to assess low reporting rates
  - If the default value is used, then report this and explain what this means for the reader.

- Report the adjustment factors (also known as factor k) that were used to adjust for incomplete reporting (by service).
  - If the default factor is used, then report this and explain what this means for the reader.

- Provide a brief report of the corrections that were made for extreme outliers or inconsistencies.
  - What values were given to correct the extreme outliers or inconsistencies?

- May include a graph to illustrate the impact of the correction in a district or region or national.
Expected outcomes from #2 & 3 for report/poster

Table: Data quality score card for national and subnational levels, 2017-2021

- Output from Stata code #3 can be used to develop a table with the annual score on indicators of data quality
- Possible modifications to the table to improve its presentation
- Selected data quality graphs can be added
- Add comments about interpretation of the data quality table, e.g.:
  - is there a data quality pattern by year for which there is an explanation? (include the explanation)
  - are there certain regions or other subnational units that are particularly problematic?
  - are there certain reporting forms or services (e.g., antenatal care, labor and delivery, immunization) that are problematic?
# Example of data quality scorecard

<table>
<thead>
<tr>
<th>Indicator</th>
<th>2017</th>
<th>2018</th>
<th>2019</th>
<th>2020</th>
<th>2021</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Completeness of monthly facility reporting (green &gt;90%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a % of expected monthly facility reports (mean, national)*</td>
<td>94</td>
<td>95</td>
<td>95</td>
<td>96</td>
<td>97</td>
</tr>
<tr>
<td>1b % of districts with completeness of facility reporting &gt;= 90%*</td>
<td>82</td>
<td>86</td>
<td>88</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>1c % of facilities with no missing monthly values in the year *</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td><strong>Extreme outliers (green &gt; 95%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2a % of monthly values that are not extreme outliers (mean, national)*</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99</td>
<td>97</td>
</tr>
<tr>
<td>2b % of districts with no extreme outliers in the year*</td>
<td>97</td>
<td>98</td>
<td>97</td>
<td>96</td>
<td>86</td>
</tr>
<tr>
<td><strong>Consistency of annual reporting (green&gt;85%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratio ANC1 – penta1 numbers (national)</td>
<td>1.06</td>
<td>1.10</td>
<td>1.09</td>
<td>1.20</td>
<td>1.18</td>
</tr>
<tr>
<td>3a % of districts with ANC1-penta1 ratio between 1.0 and 1.5</td>
<td>63</td>
<td>77</td>
<td>68</td>
<td>69</td>
<td>70</td>
</tr>
<tr>
<td>Ratio Penta1 – penta3 numbers (national)</td>
<td>1.11</td>
<td>1.10</td>
<td>1.09</td>
<td>1.09</td>
<td>1.10</td>
</tr>
<tr>
<td>3b % of districts with penta1-penta3 ratio between 1.0 and 1.5</td>
<td>91</td>
<td>93</td>
<td>85</td>
<td>88</td>
<td>88</td>
</tr>
<tr>
<td><strong>Annual data quality score (mean indicator 1a to 3b)</strong></td>
<td>90</td>
<td>93</td>
<td>91</td>
<td>91</td>
<td>90</td>
</tr>
</tbody>
</table>

*Mean for ANC, delivery, immunization and OPD services*