Data Quality Composite Score Tool

Zola Allen, Jenny Mwanza, and Fredrick Onyango, Data.FI

PEPFAR

USAID
FROM THE AMERICAN PEOPLE

Data.FI
Data for Implementation

Overview

- Development of the Data Quality Composite Score (DQS)
- Customized Data Quality Score Tool

Introduction

- There is a range of strategies designed to improve quality of routinely reported aggregate data*
- We recommend use of less resource-intensive approaches before deploying more resource-intensive approaches
- We acknowledge that a desk review using data quality scoring has limitations

	M&E Personnel	External Auditors	Field Visits	Desk Review	Resources
Data quality scoring	x			x	Low
Routine data quality audits	x		x		Medium
Data quality audits		x	x		High

* Individual longitudinal records—such as electronic medical records (EMRs)—require different approaches for desk review and data quality scoring (not covered here).

Components

COMPLETENESS

measures the number of submitted records against the number of expected records

TX_CURR was used as proxy measure for whether a facility has submitted any of the data elements in a given week.

TX_CURR was chosen because it is assumed that once a facility has reported TX_CURR it will continue to report on TX_CURR, while in a given week it may or may not report TX_NEW, for example.

COHERENCE

measures the degree to which data elements fit together

Data coherence score is used to measure possible data quality issues arising from an indicator numerator being greater than its denominator. This score is assigned to the following indicators: HTS_TST & HTS_TST_POS, TX_NEW & TLD_NEW, TX_CURR & TLD_CURR and TX_PVLS_D & TX_PVLS_N.

CONSISTENCY

measures the extent to which data elements are consistent over time

This measures trend performance on high-frequency reporting (HFR) indicators reporting over a period and identifies outliers on data reporting.

Indicators of interest

In customizing the DQS approach, we work with stakeholders to identify the most meaningful indicators to review given the country context.

In the following examples, we have reviewed HFR data and combinations of these indicators. The tool can be modified to include other indicators as appropriate.

HTS_TST	HTS_TST_POS
TX_NEW	TLD_NEW
TX_CURR	TLD_CURR
TX_PVLS_D	TX_PVLS_N

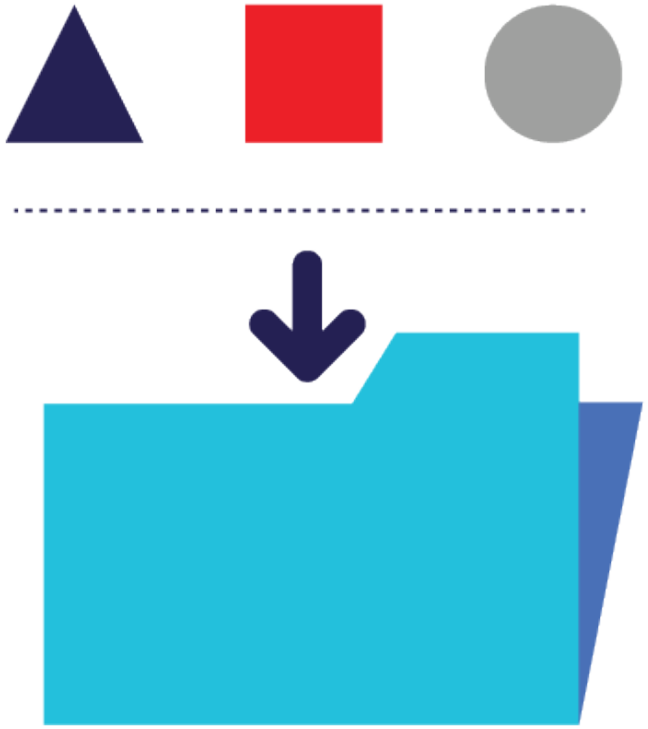
Data quality scores

	Completeness	Coherence	Consistency	Data Quality Score
Implementing Partner A	97%	0%	50%	**49%**
Implementing Partner B	93%	0%	50%	**48%**
Implementing Partner C	100%	100%	50%	**83%**
USAID				**60%**

The data quality score allows us to quantify performance across implementing partners, subnational organizational units, or countries within a region.

The guiding principle is that when scores are transparent and calculated consistently, there will be increased accountability and responsible parties will work to improve scores.

Completeness



Methods

- HTS_TST is used as proxy measure for whether a facility has submitted any of the data elements in a given week.
- Expected report = Total number of facilities * number of reports for the period.
- Report received is the sum of all report for the period (weeks).
- Percent score is determined by dividing reports received by expected reports x 100.

$$\frac{\text{Number reports received}}{\text{Number reports expected x 100}}$$

Completeness scores

Implementing partner	Health facilities	Number of weeks	Number of reports expected	Number of reports received	Score
Partner A	25	16	400	388	97%
Partner B	17	16	272	202	74%

Coherence

Methods

COHERENCE measures the degree to which data elements fit together as measured by validation rules.

- HTS_TST_POS =< HTST_TST
- TLD_NEW =< TX_NEW
- TLD_CURR =< TX_CURR
- TX_PVLS_N =< TX_PVLS_D

- Each pair of validation rules is assigned a maximum percentage coherence score which is calculated as 100% divided by the number of rules.
  - If there are four rules, each validation rule is assigned a 25% coherence score.
- This maximum score is kept if a validation rule is met by all facilities. A zero score is assigned if at least one facility did not meet a validation rule.
- The total coherence score is calculated as a sum of all validation rule scores.
- Each pair of validation rules that have been validated must necessarily be corrected.

Coherence scores

Coherence	HTS_TST ≥ HTS_TST_POS	TX_NEW ≥ TLD_NEW	TX_CURR ≥ TLD_CURR	TX_PVLS_D ≥ TX_PVLS_N	Score
Implementing Partner A	3	52	5	17	**0%**
Implementing Partner B	4	13	13	1	**0%**
Implementing Partner C	0	0	0	0	**100%**

- The data coherence score is used to measure possible data quality issues arising from an indicator numerator being greater than its denominator.
- Each validation rule is assigned a 25% coherence score

Examples of coherence

TESTING

TX_PVLS_D	TX_PVLS_N	TX_PVLS_D ≥ TX_PVLS_N
135	**149**	**VIOLATION** The number of clients with viral suppression testing should be greater than or equal to the number of clients virally suppressed.

TREATMENT

TX_NEW	TLD_NEW	TX_NEW ≥ TLD_NEW
2	**6**	**VIOLATION** The number of clients initiated on ART treatment should be greater than or equal to the number of clients initiated on the TLD regimen.

Tables above with highlighted cells show the indicators with data quality issues on coherence.

Coherence scores | baseline

BASELINE

Country	HTS_TST_POS > HTS_TST					Index HTS_TST_POS > Index HTS_TST					Score
	F <15	M <15	F 15+	M 15+	ALL	F <15	M <15	F 15+	M 15+	ALL	
TOGO	0	0	0	0	0	0	1	1	2	0	70%
BURKINA	0	0	0	0	0	0	1	2	1	1	60%

Each validation rule is assigned a 10% coherence score (there are 10 validation rules in this example).

Facilities with coherence issues, Togo | baseline

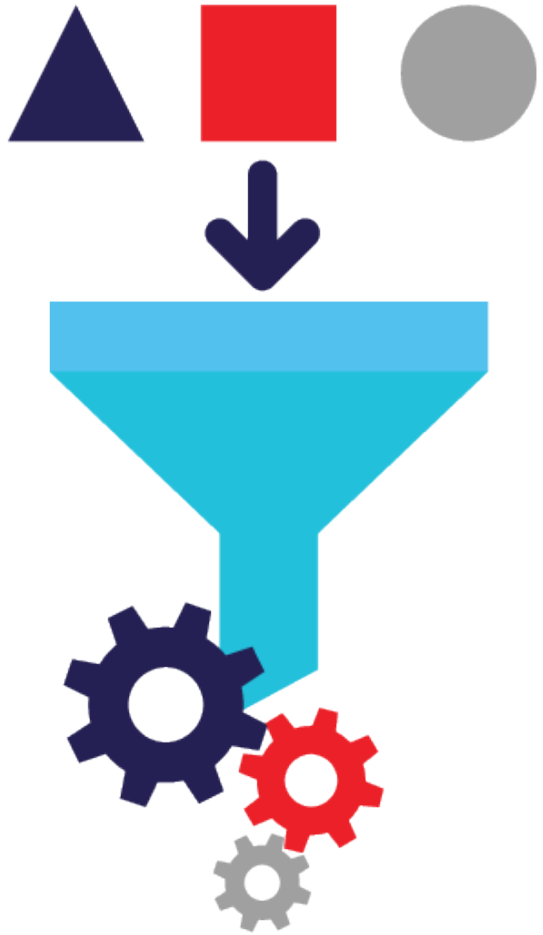
CMS Lucia (EVT) | BASELINE

Date	HTS_INDEX_TST					HTS_INDEX_POS				
	F < 15	M < 15	F 15+	M 15+	ALL	F < 15	M < 15	F 15+	M 15+	ALL
13/04/2020	2	0	2	0	4	0	1	0	1	2

Hôpital de Bè | BASELINE

Date	HTS_INDEX_TST					HTS_INDEX_POS				
	F < 15	M < 15	F 15+	M 15+	ALL	F < 15	M < 15	F 15+	M 15+	ALL
29/06/2020	4	0	0	3	7	0	0	3	3	6
18/05/2020	0	4	3	0	7	0	0	1	1	2

Consistency



Methods

This is the measure of **CONSISTENCY** in records over time or simply the absence of outliers:

1. TX_NEW
2. TX_CURR

- The data values identified are not consistent with the trend reported by that facility.

Methods, cont.

- Identify the presence of outliers:
  - For each indicator, calculate the minimum and maximum values over the time period.
  - Use conditional formatting to highlight a facility as having an outlier if the maximum value is at least five times greater than the minimum value.
  - Have a human review all highlighted values and make the final determination on a facility having an outlier, based on the provided guidance.
- Each indicator measure of consistency is assigned a maximum score, which is calculated as 100% divided by the number of indicators under review.
  - If there are two indicators, each measure is assigned a 50% consistency score.
- This maximum score is kept if there are no outliers in any of the facilities. A zero score is assigned if at least one facility has outlier/s.
- The total consistency score is calculated as a sum of all indicators' consistency scores.

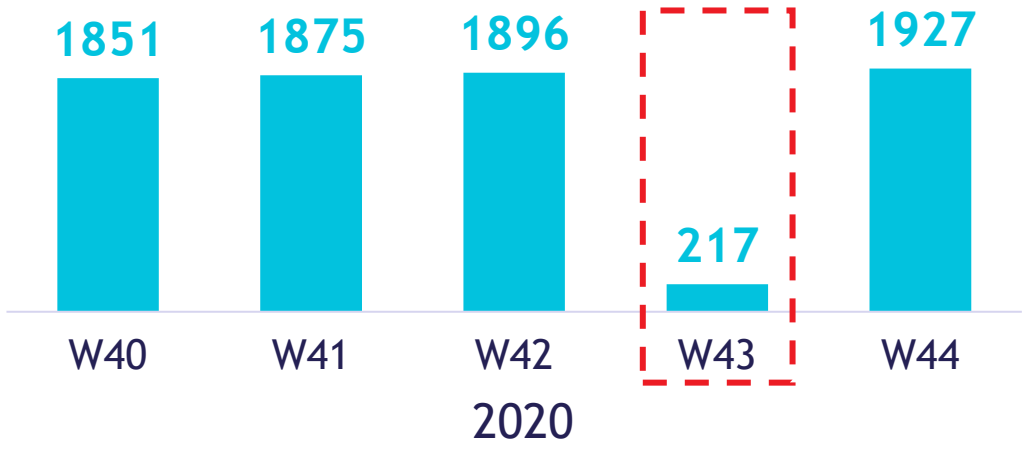
Consistency scores

Consistency	TX_NEW	TX_CURR	Score
Implementing Partner A	2	0	**50%**
Implementing Partner B	0	5	**50%**
Implementing Partner C	0	2	**50%**

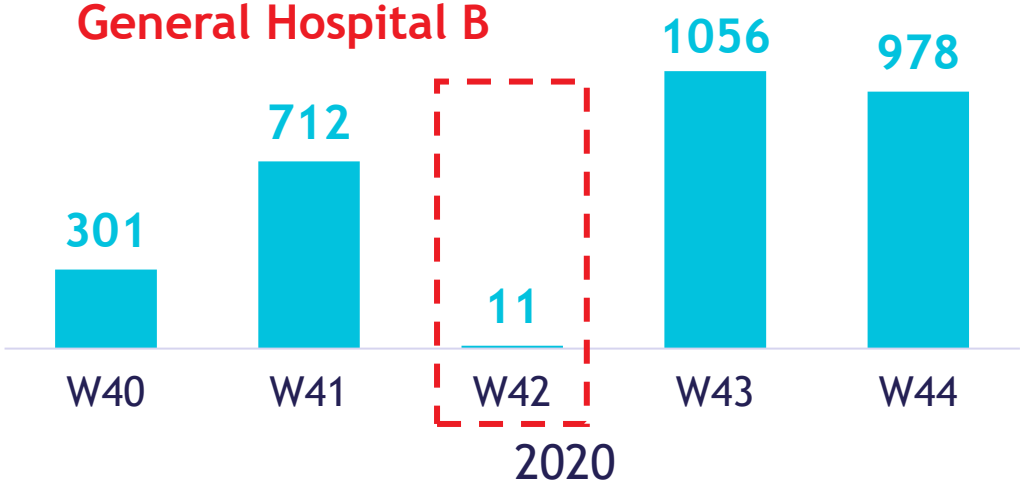
Each measure of consistency on a specific indicator is attributed 50%.

Examples of consistency

General Hospital A

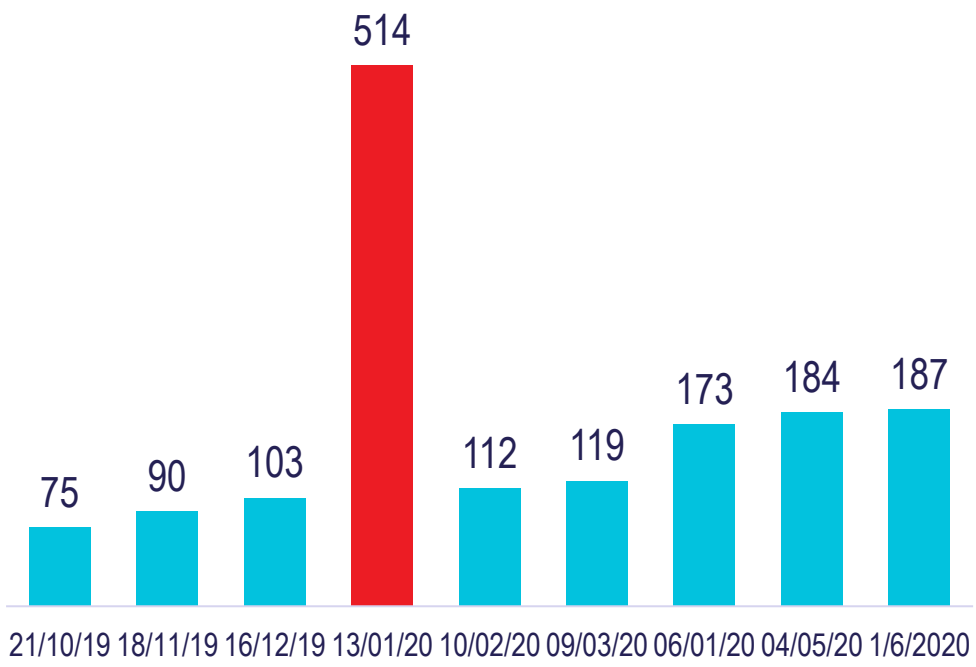


General Hospital B

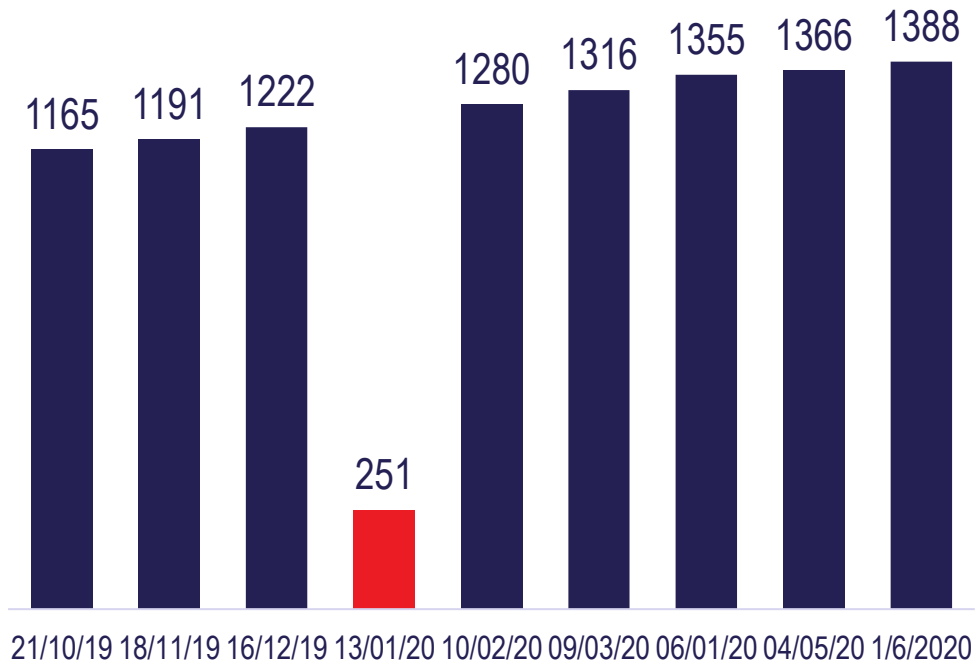


Facilities with consistency issues, TX_CURR |baseline

Hospital A

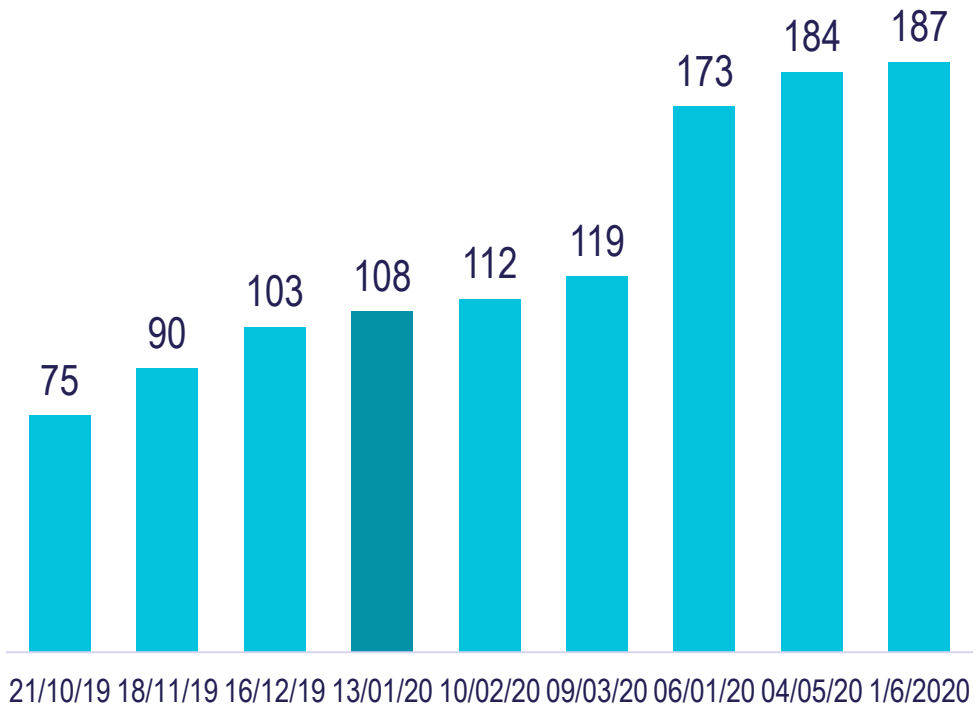


Hospital B

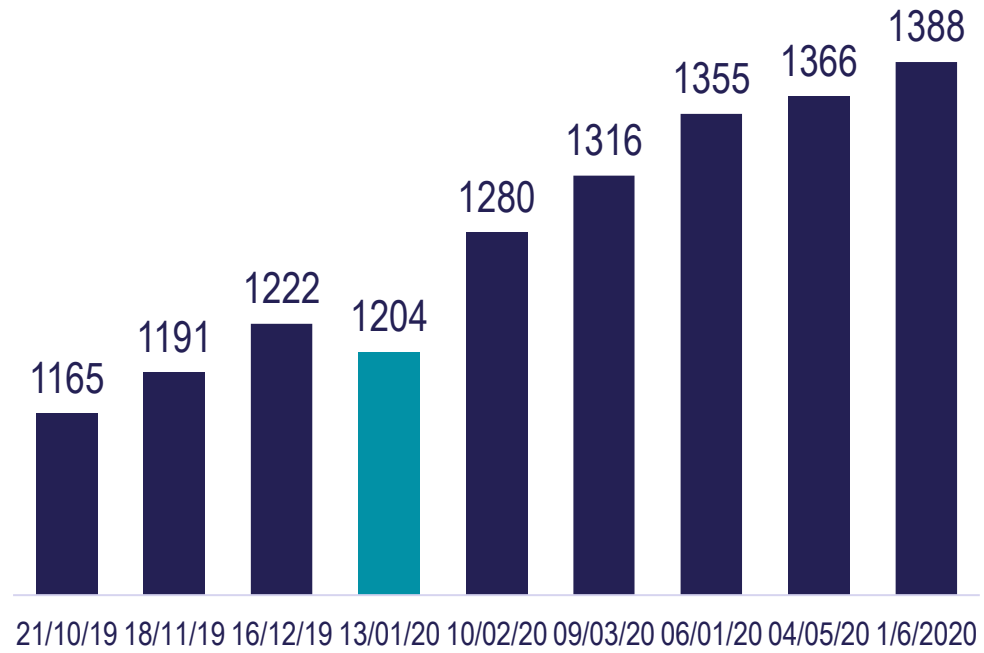


Facilities with consistency issues, TX_CURR |endline

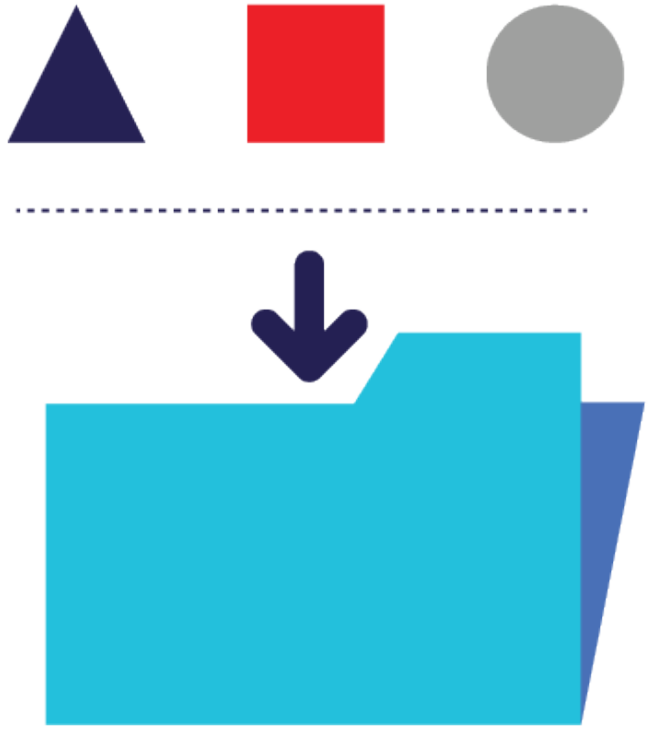
Hospital A



Hospital B



Customized Data Quality Scoring Tool



Selection of platform

Initially, we conducted data quality scoring manually—a tedious and labor-intensive process. When selecting a tool to automate this process, our team reviewed the following existing tools:

1. Data Cleaner - http://datacleaner.org/
2. DataPreparator - http://www.datapreparator.com/
3. Data Quality Analyzer - https://www.ataccama.com/product/data-discovery-and-profiling/dqa
4. WHO DHIS2 Data Quality Tool - https://who.dhis2.org/dq/dhis-web-commons/security/login.action

Selection criteria

Requirements for the tool
Extracts and shares summary data quality scores
Extracts and shares raw data assessed for data quality
Has the ability to automatically import new (Excel) data provided and to auto-compute scores without further manual input
Measures completeness—the degree to which all required data are available in the dataset; a measure of the percentage of data entries from units
Measures coherence—the extent to which data adhere to business rules
Measures consistency—measures outliers
Minimizes risk of duplication, errors, and maintains integrity
Proposed tool should be robust, with customizable scoring criteria to cater for different/additional measurements in the long run
Has the ability to store data at least until utility is complete
Can import and export data, at least in Excel—has the ability to share reports, as required
Is user-friendly
Is easy to install
Excludes null values

Customized tool developed

When none of the existing data quality platforms met our selection criteria, Microsoft Excel with Visual Basic Application (VBA) was chosen!

To use the resulting Data Quality Score Tool follow these steps

1. Prepare raw data files
2. Import raw data files
3. Execute script
4. Review results

The tool produces both detailed reports and summary analyses that can easily be shared with stakeholders

The tool can accommodate 10,000 records

- Up to 12 reporting periods
- Up to 10 indicators
- Up to 80 reporting units

The tool easily be adapted for data sets other than HFR PEPFAR data! The possible applications are infinite!

Current DQS applications

- Calculated DQS using HIV high-frequency reporting data
  - Nigeria, Mali, Ghana, Togo, Burkina Faso, Liberia, Senegal

- Indicators included
  - HTS_TST, HTS_TST_POS
  - HTS_Index_TST, HTS_Index_TST_POS
  - KP_PREV
  - TX_NEW
  - TX_CURR

- Average baseline score across countries – 74%
- Average end line score – 100%

Current DQS Applications, cont.

Country	Time period under review	Number of weeks under review	Number of health facilities included	Areas needed data validation or cleaning		
				Completeness	Coherence	Consistency
Ghana	Sept 30, 2019–June 06, 2020	37	25		X	X
Burkina Faso	March 16–June 29, 2020	16	17		X	X
Liberia	January 20–June 29, 2020	24	18		X	
Mali	December 16, 2019–March 02, 2020	12	13		X	
Senegal	January 20–June 01, 2020	12	8		X	
Togo	March 16–June 29, 2020	16	25		X	X
Nigeria	December 30, 2020–March 23, 2020	13	479	X	X	X

Results of selection process (Annex)

Requirements	Data Cleaner	DataPreparator	Data Quality Analyzer	WHO DHIS Quality Tool
Extract and share summary data quality scores	**No**	Install failed	Yes	Yes
Extract and share raw data assessed for data quality	Yes		Yes	Yes
Ability to automatically import new (excel) data provided and auto compute scores without further manual input.	Yes		**No**	Yes - Must be prepared into DHIS2 import format
Measure Completeness : degree to all required data are available in the dataset. A measure of the percentage of expected data entries from units	No		No	No
Measure Coherence : Extent to which data adheres to the business rules	Yes		Yes	**No**
Measure Consistency : Measure outliers. Data should be constant in time	**No**		No	Yes
Minimize risk of duplication, errors and maintain integrity	Yes		Yes	Yes
Proposed tool should be robust with customizable scoring criteria to cater for different/additional measurements in the long run	Yes - Not fully customizable		Yes - To some extent	No
Ability to store data at least until utility is complete.	Yes		Yes	Yes
Data import and export at least in excel. Ability to share report as required	Yes		**No**	Yes
User Friendly	To some extent		No	Yes
Easy to Install	Yes	No	Yes	No
Exclude Null Values	Yes		Yes	Yes

Data.FI
Data for Implementation

FOR MORE INFORMATION

Emily Harris, Data.FI AOR, USAID Office of HIV/AIDS
emharris@usaid.gov

Jenifer Chapman, Data.FI Project Director
datafiproject@thepalladiumgroup.com

Data for Implementation (Data.FI) is a five-year cooperative agreement funded by the U.S. President's Emergency Plan for AIDS Relief through the U.S. Agency for International Development under Agreement No. 7200AA19CA0004, beginning April 15, 2019. It is implemented by Palladium, in partnership with JSI Research & Training Institute (JSI), Johns Hopkins University (JHU) Department of Epidemiology, Right to Care (RTC), Cooper/Smith, IMC Worldwide, Jembi Health Systems and Macro-Eyes, and supported by expert local resource partners.

Data for Implementation (Data.FI) is a five-year cooperative agreement funded by the U.S. President's Emergency Plan for AIDS Relief through the U.S. Agency for International Development under Agreement No. 7200AA19CA0004, beginning April 15, 2019. It is implemented by Palladium, in partnership with JSI Research & Training Institute (JSI), Johns Hopkins University (JHU) Department of Epidemiology, Right to Care (RTC), Cooper/Smith, IMC Worldwide, Jembi Health Systems and Macro-Eyes, and supported by expert local resource partners.