Step- by- Step Guidance on IM Data Analysis

ADAPTED FOR ADDITION TO THE RHIZOME **POLIOK.IT** PLATFORM 16 JAN 2017

Rhizome INFORMATION FOR ERADICATION FROM POLIO GLOBAL ERADICATION INITIATIVE

Step-by-Step Guidance on IM Data Analysis

Table of Contents

Overall aim of analysis

As a general exercise, we usually first focus on the following topics:

- Campaign outcome over time (trends) with disaggregated reasons for missed children;
- For the most recent campaign, where were the absent children?
- For the most recent campaign, why did parents/caregivers refuse the OPV?
- For the most recent campaign, how were parents/caregivers informed?
- Immunity gap among npAFP cases (as supporting evidence of campaign coverage);

Given time and available data, further analysis could include:

- Correlation between campaign awareness and outcome (vaccination coverage);
- Correlation between IM data and Quick Survey data;
- Geographic analysis of campaign outcome and WPV case incidence (mapping);
- Geographic analysis of campaign outcome and immunity gap among npAFP cases (mapping);
- Association between specific social mobilization activities and campaign outcome;

Data organization

This is carried out using Excel program on available IM data sheets from multiple campaign rounds, resulting in a final data set in proper shape for further analysis.

- Combining single-campaign data sheets into one long-shape time-series sheet;
- Make sure variables are consistent across different campaigns (may have to combine variables to fit structure);
- Leave blank columns for additional variables just to align sheets from different campaigns;
- Add time series variable (Year & Month);
- Double check to ensure no mismatch when appending rows of data from different campaigns together;

Examples:
KEN data sets: 2013.05 – 2014.06
2013.05 & 2013.06 data structure very different from later rounds: not-combinable;
New variables of social reasons for missed children added at various time points (be careful when appending data);
Variable order slightly changed over time (again double check column headings when appending);

Data cleaning

- Once appended together, create "pivot" tables using all columns and select geographic area names as "row-labels" in the pivoting panel;
- Identify and correct name inconsistencies (slash/hyphen/space, upper/lower case, etc.);
- Shorten variable names for further analysis (create a "name-description" dictionary);
- Inspect data consistency by test-run summation of number of missed children and number of missed children by reason for each district (via pivot table);
- Note excessive "others" for reason of missed children;
- In statistical program check histogram of newly computed coverage and awareness indicators;
- In statistical program check scatter plot of newly computed coverage and number missed;

Example:
KEN data sets: in a number of districts, home monitoring data and public place data are exactly same;
In a number of districts, typo in data entry (other category >200 counts) → drop observation;

Data analysis

For efficiency in analysis and graphics, I use STATA program, but all of these can also be carried out using SPSS or Excel (using "pivot" tables). I have saved the coding program in STATA for future reference.

Cautionary question:
Does it make sense to combine "inside home" monitoring data and "public place" monitoring data? If the sampling procedures (and resulting sampling weights) are different, then the resulting "average" could be biased and misleading. For the description and results presented below, analysis was restricted to "inside home" numbers. The same procedure can be repeated using "public place" numbers.

Defining variables:

From IM data set

Awareness: Proportion of parents/caregivers aware of the most recent campaign
Calculated as a ratio expressed in percentage points

- **Numerator**: Number of parents/caregivers aware of the most recent campaign
- **Denominator**: Total number parents/caregivers interviewed during IM

Awareness can be calculated at the lowest geographic level where data are representative and also can be aggregated to higher geographic level.

Source information: Means by which parent/caregiver was informed of the previous campaign
This is calculated semi-quantitatively as a lead frequency ranking of all types of source information type listed in independent monitoring data set
Source information can be calculated at the lowest geographic level where data are representative and also can be aggregated (summed up) to higher levels.

Coverage: Proportion of under-five children vaccinated during the most recent campaign
Calculated as a ratio expressed in percentage points

- **Numerator**: Number of children vaccinated (either confirmed by finger marking or verbal history)
- **Denominator**: Total number of eligible children who were present during IM

Coverage can be calculated at the lowest geographic level where data are representative and also can be aggregated to higher geographic level.

Missed children by reason: Proportion of children who missed the most recent campaign due to each of the following reason: Absence, Refusal (by caregiver), Household not visited, Child asleep and Other reasons
Calculated as a ratio expressed in percentage points

- **Numerator**: Number of children missed the campaign due to reason X
- **Denominator**: Total number of eligible children who were present during IM

Technical note

1. *Missed children by reason can be calculated at the lowest geographic level where data are representative. However, reporting trends for each district can be cumbersome. In addition, often times sample size is very small at district level (below 5 children missed). Therefore it may be preferable to aggregate Missed children by reason at higher level(s).*
2. *During data quality review, it was noted that often times the variable "total number of unvaccinated children" doesn't match with the aggregated number of children summed*

from the five categories of reasons. The mismatch is very small in magnitude. Most likely the latter is smaller than the former. This could be the result of a number of issues, such as:

- *During IM, a small fraction of unvaccinated children were not further queried for reason of missing vaccination.*
- *Human error during data compilation from individual line item to "district-level" aggregate, under-counting or over-counting some numbers.*

To adjust for the discrepancy, an additional step was taken in calculating the proportion of children missing campaign for each reason, as follows:

(1 – Coverage) x [(# missed children with reason X)/(sum of # missed children due to each reason)]

Absence pattern: Distribution of absent child locations
Calculated as a ratio expressed in percentage points

- **Numerator**: Number of children not available for vaccination as they were at location X
- **Denominator**: Total number of children not available for vaccination

The percentage of location categories should add up to 100%.

Refusal pattern: Distribution of refusal reasons
Calculated as a ratio expressed in percentage points

- **Numerator**: Number of children whose caregiver refused OPV citing reason X
- **Denominator**: Total number of children whose caregiver refused OPV

The percentage of reason categories should add up to 100%.

From WHO AFP data set

At the current set up, the AFP data set managed by the WHO is updated on a monthly basis and shared with UNICEF (and other GPEI partner agencies) via secure file download (login required). This may change later as UNICEF and WHO collaborate on improved online data platforms.

The AFP data set is well organized as one long sheet where each row is an AFP case and columns recording geographic, demographic and clinical characteristics of the case, such as country, province, district, age, sex, time of onset, time of diagnosis, number of OPV doses and diagnosis.

0 dose npAFP: Proportion of non-polio AFP cases who had 0 dose of OPV
Calculated as a ratio expressed in percentage points

- **Numerator**: Number of npAFP cases with 0 dose of OPV
- **Denominator**: Total number of npAFP cases

4+ doses npAFP: Proportion of non-polio AFP cases who had 4 or more doses of OPV
Calculated as a ratio expressed in percentage points

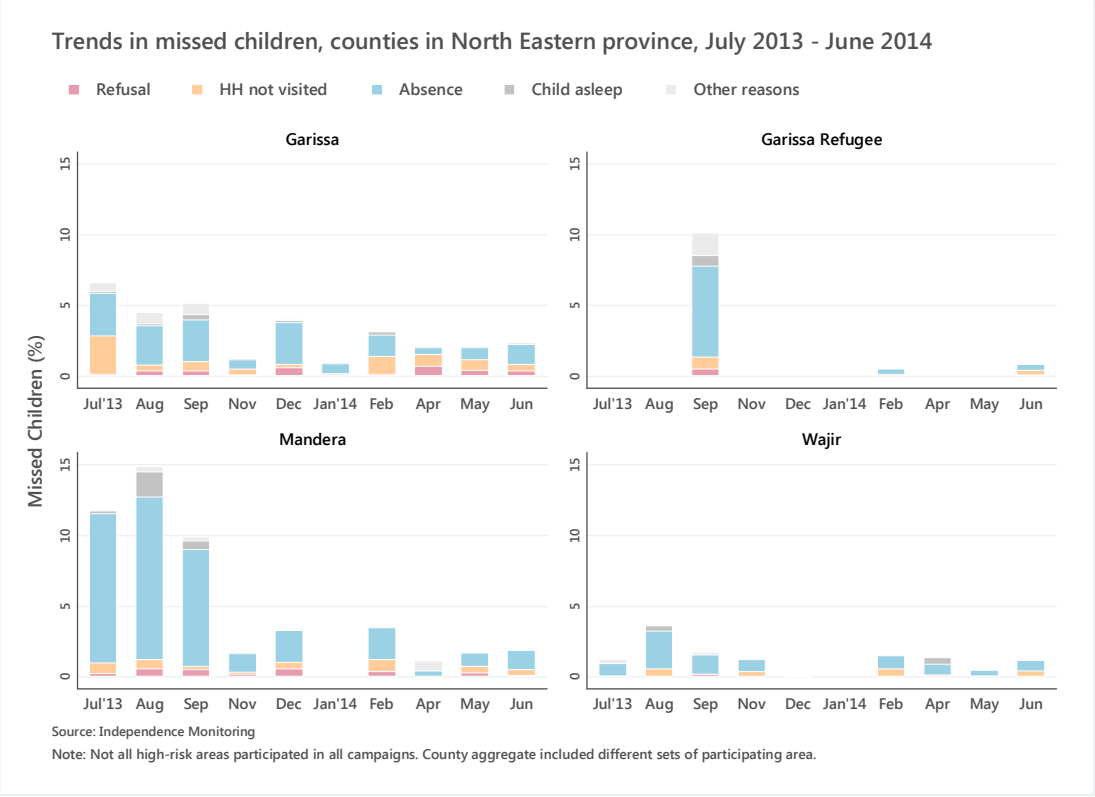
- **Numerator**: Number of npAFP cases with 4 or more doses of OPV
- **Denominator**: Total number of npAFP cases

Technical note

1. *It is important to set the age group of npAFP cases. Usually it is 6 – 59 months of age.*
2. *The two indicators from AFP data set are often aggregated at province level due to small sample size at sub-province levels.*

Charting trends

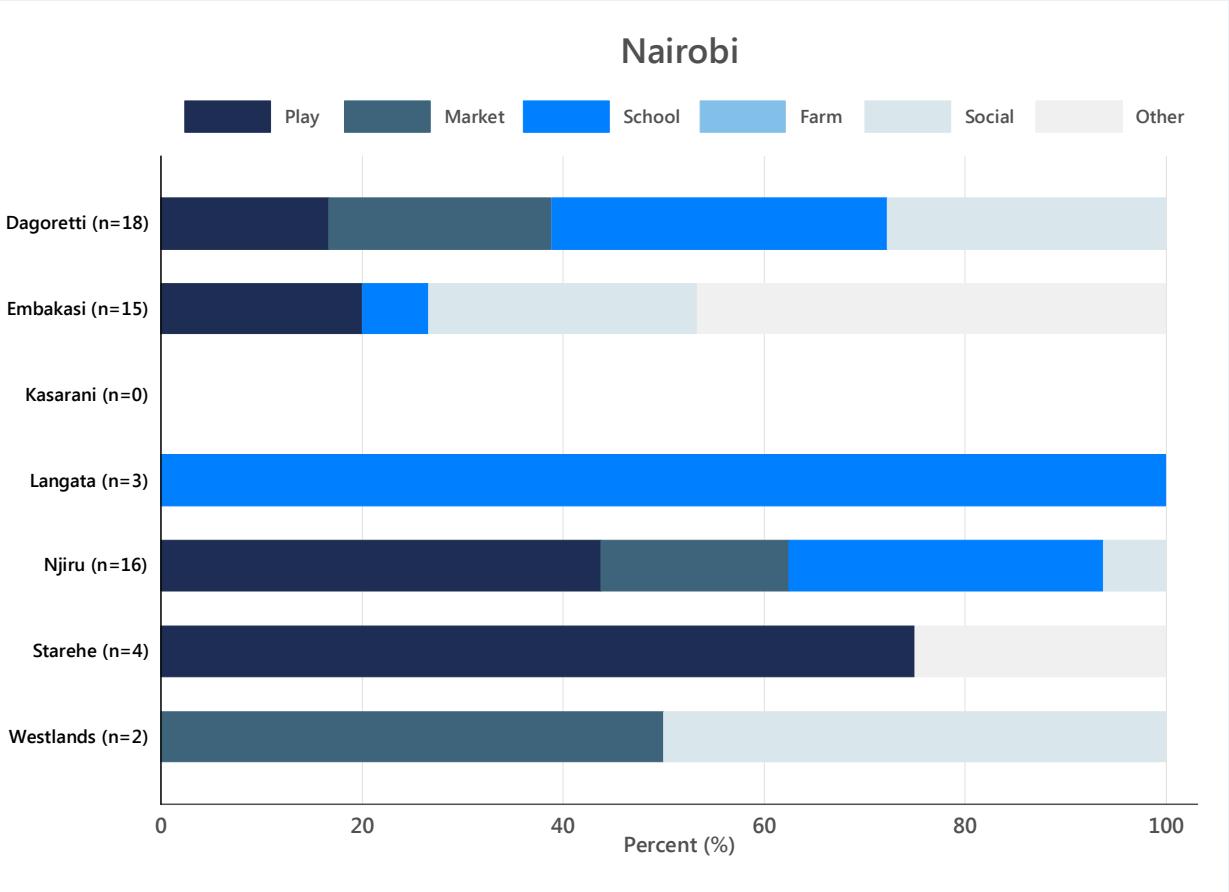
We prefer to use stacked bars showing proportion of missed children over time to show trends, where the overall height of the bar represents total proportion of missed children in a given campaign and the disaggregated bars represent reasons of missed children, color-coded conventionally as such: pink – refusal; light orange – no visit; light blue – absence; dark gray – asleep; light gray – others;
For reliability and simplicity, this is often presented at the 2nd lowest level.
Example



This is carried out using STATA for efficiency but can also be done using SPSS or EXCEL.

Charting absence pattern

For a given geographic area, we can show a snapshot of absence pattern during the most recent campaign with disaggregation. In the following example, Nairobi County is shown as disaggregated by districts. On a national level, however, we can show "province" snapshot with disaggregation at "county" level. Here the horizontal stacked bars are color-coded using a blue gradient.

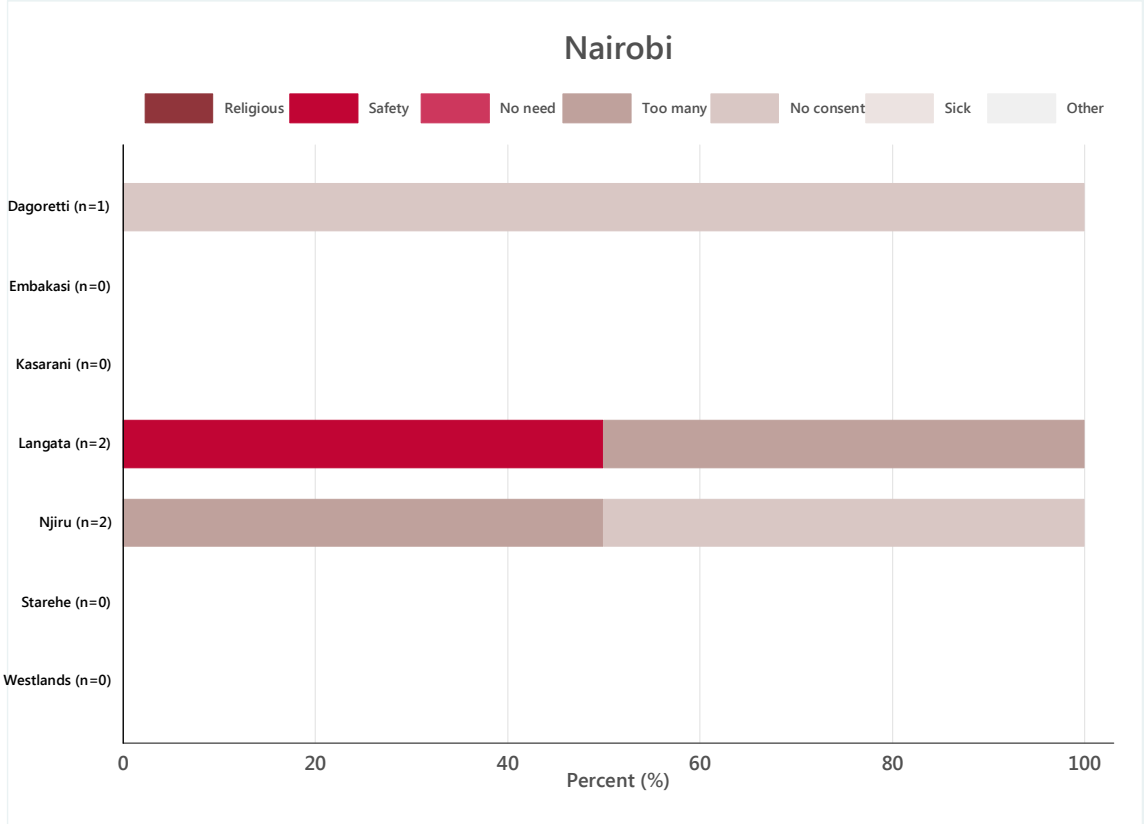


It should be noted that such proportion composition chart needs to be viewed/interpreted together with the overall sample size in each instance, as relative proportions can be "artificially" inflated merely due to small sample size. In the example given above, in Langata district in Nairobi county, one could say that "100% of absent children during last campaign were at school". Does this immediately suggest that we should change our strategy next time accordingly? Keep in mind that the "100%" was based on only 3 children. This could very well be random "noise".

This is carried out using STATA for efficiency but can also be done using SPSS or EXCEL.

Charting refusal pattern

For a given geographic area, we can show a snapshot of refusal pattern during the most recent campaign with disaggregation. In the following example, Nairobi County is shown as disaggregated by districts. On a national level, however, we can show "province" snapshot with disaggregation at "county" level. Here the horizontal stacked bars are color-coded using a red gradient.



Note that due to very few refusals, at district level this is often empty.

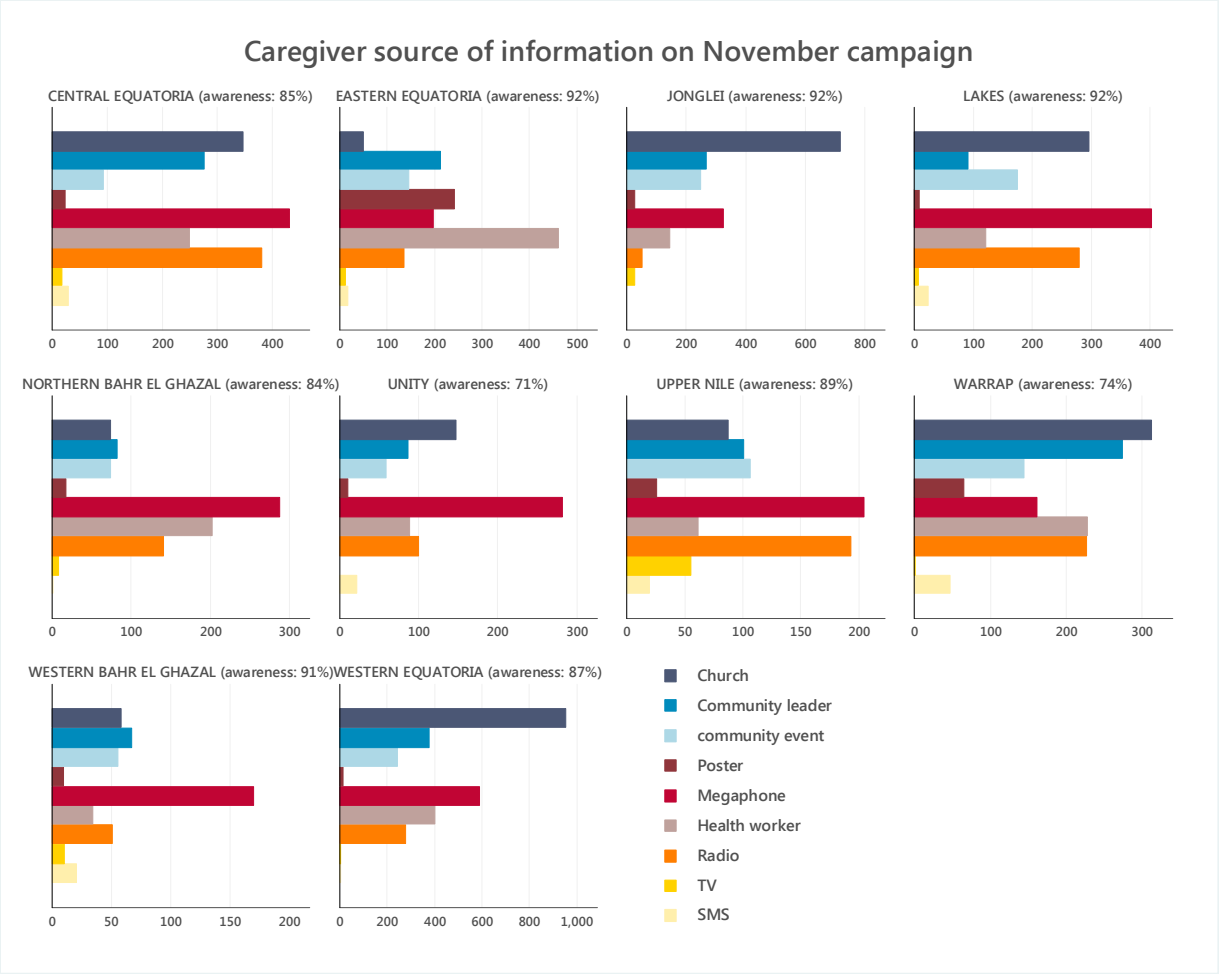
This is carried out using STATA for efficiency but can also be done using SPSS or EXCEL.

Charting source information

For a given geographic area, we can show a snapshot of refusal pattern during the most recent campaign with disaggregation. In the following example, data from South Sudan is shown as disaggregated by provinces.

Source information type can be categorized slightly differently in different countries. When there are too many types, maybe some infrequent types can be combined under the condition that no important information is lost.

Example of source of information chart



This is carried out using STATA for efficiency but can also be done using SPSS or EXCEL.

Analyzing AFP data

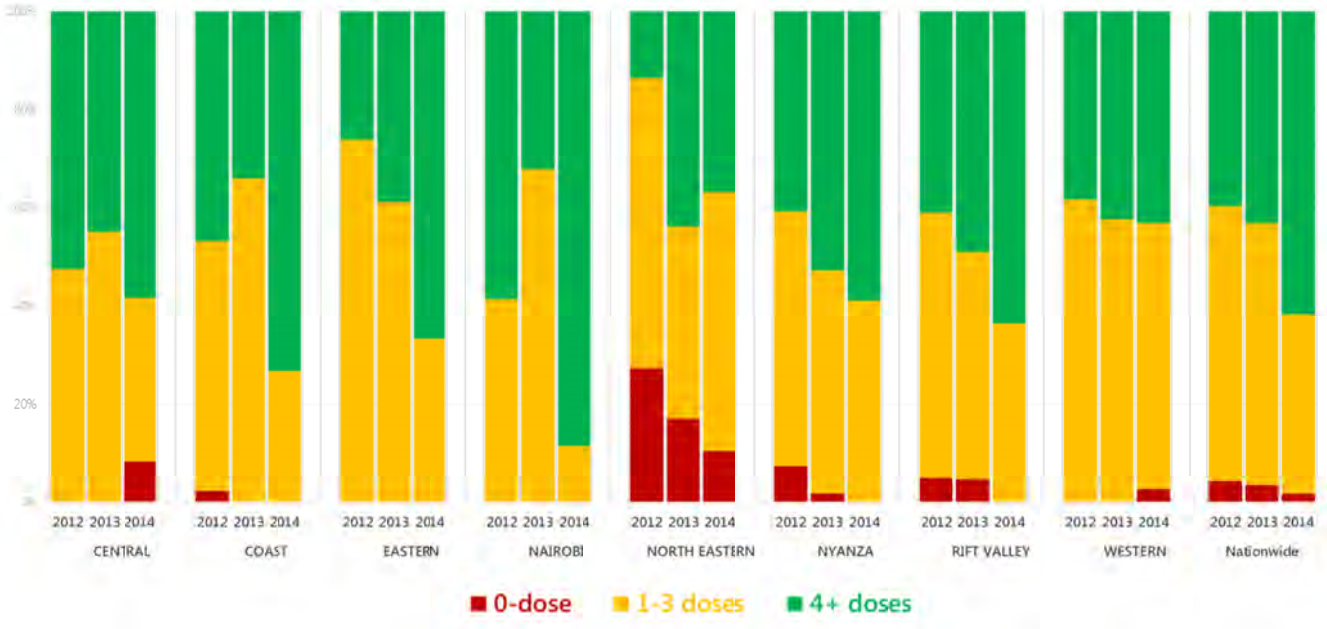
There is usually not enough npAFP case data at district level for trend analysis. Depending on country population size, we may even need to aggregate at 2nd highest administrative level (province).

Using "pivot" table in EXCEL, we can calculate and present the indicators of 0-dose npAFP and 4+ dose npAFP in a simple table, with geographic disaggregation and time series, as follows:

Table. Immunity gap among npAFP cases, 2012 - 2014

	2012				2013				2014		
Province	**# cases**	**0-dose**	**4+ dose**		**# cases**	**0-dose**	**4+ dose**		**# cases**	**0-dose**	**4+ dose**
CENTRAL	21	0%	52%		20	0%	45%		12	8%	58%
COAST	45	2%	47%		47	0%	34%		30	0%	73%
EASTERN	46	0%	26%		18	0%	39%		27	0%	67%
NAIROBI	29	0%	59%		28	0%	32%		26	0%	88%
NORTH EASTERN	22	27%	14%		41	17%	44%		19	11%	37%
NYANZA	54	7%	41%		57	2%	53%		34	0%	59%
RIFT VALLEY	61	5%	41%		43	5%	49%		55	0%	64%
WESTERN	47	0%	38%		33	0%	42%		37	3%	43%
Nationwide	325	4%	40%		287	3%	43%		240	2%	62%

Alternatively, this can also be shown by a color-coded stacked bar chart



Interpreting results

This is the most intriguing part of the exercise. It requires comprehensive understanding of the program and deep insight of the local situation. One may frequently go back to the raw data to look for more evidence or a missing piece of information.
Without knowing the background of each country program, here are some initial general questions (primers) to ask when trying to interpret results for each research aim (going back to the first section of this document).

Data quality

Were there too many missing values?
Is the definition of variables constant over time?
Were sample sizes large enough for meaningful pattern identification? At district level? At higher levels?
Were there particular geographic areas where data look suspicious?
How can data collection and compilation be improved for next time?

Trends in coverage (outcome)

Campaign outcome over time (trends) with disaggregated reasons for missed children;
Looking over the trend charts, is there any particular pattern of coverage overall?
Is there any reason for missed children standing out among others?
What does this mean for program performance?
What could be potential causes? What additional information is needed?
Is there any geographic region to pay closer attention to than others?

Absence pattern

Where were the children who were not available during last campaign?
Any dominant locations?
What are the implications for the next round?

Refusal pattern

Were there any dominant reasons for refusal?
Were there areas that needed special attention?
What are the implications for the next round?

Source of information

How were parents/caregivers informed about the last campaign?
Considering mobilization activities carried out before the campaign, how effective they were in achieving the results?
Is the pattern shown in IM data consistent with activities on the ground?
Is there a need to change information promotion strategy for next round?

Immunity gap

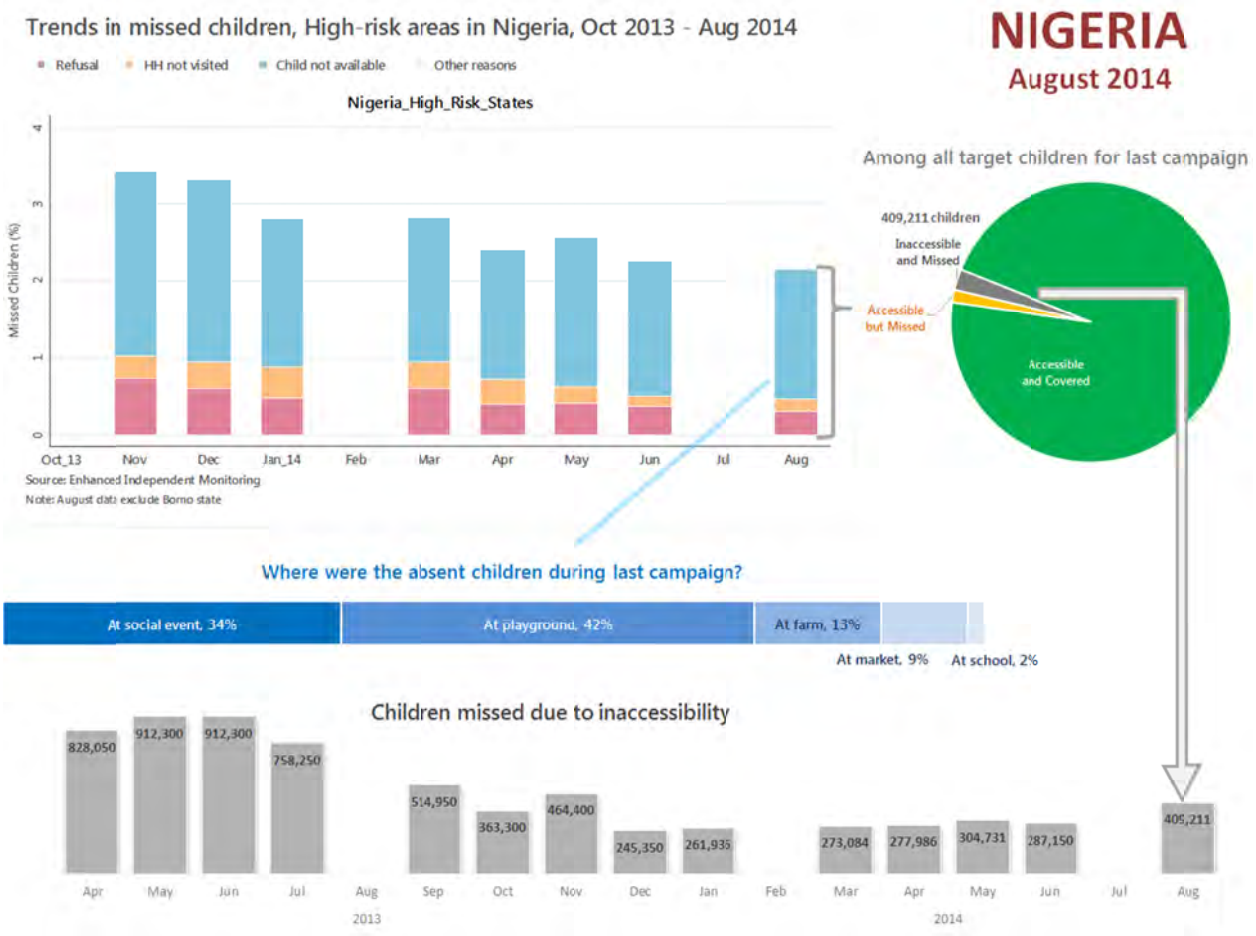
What are the overall patterns over time?
Do these patterns support the findings from previous steps?
If there is discrepancy between IM data and AFP data in terms of OPV campaign coverage and immunity level, what could be potential causes?
What additional information is needed?

Next steps

Given time and available data, further analysis could include:

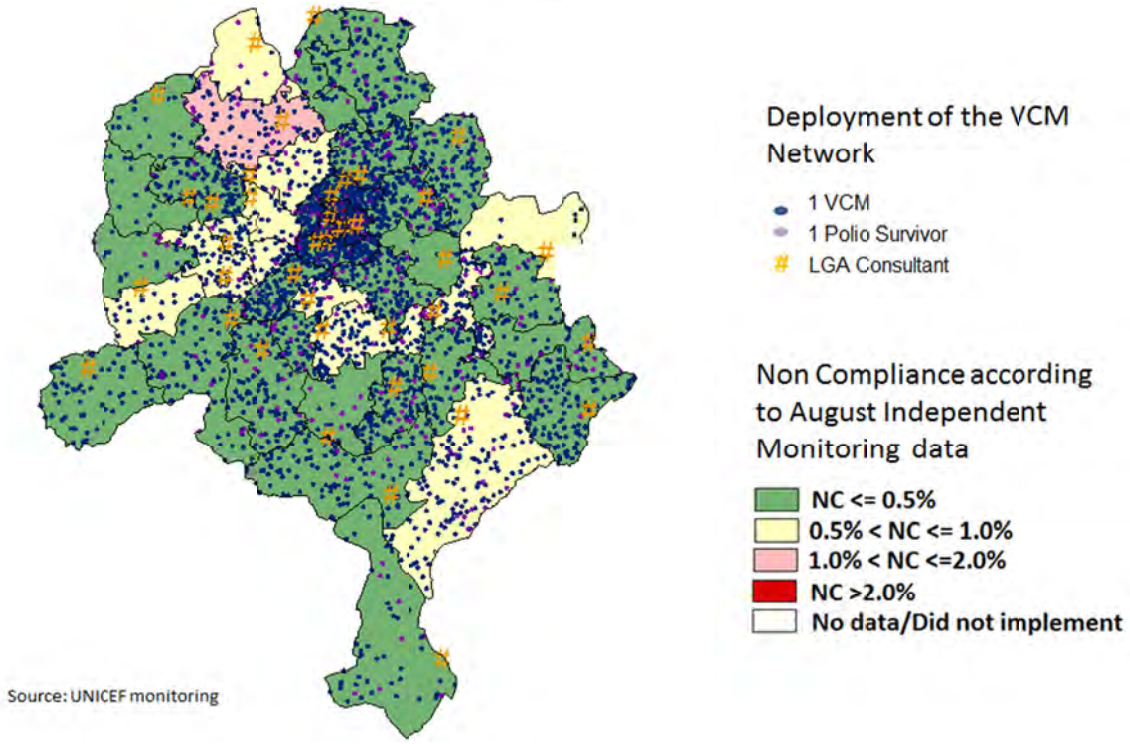
- Correlation between campaign awareness and outcome (vaccination coverage);
- Correlation between IM data and Quick Survey data;
- Geographic analysis of campaign outcome and WPV case incidence (mapping);
- Geographic analysis of campaign outcome and immunity gap among npAFP cases (mapping);
- Association between specific social mobilization activities and campaign outcome;

Some examples of combining data from multiple sources

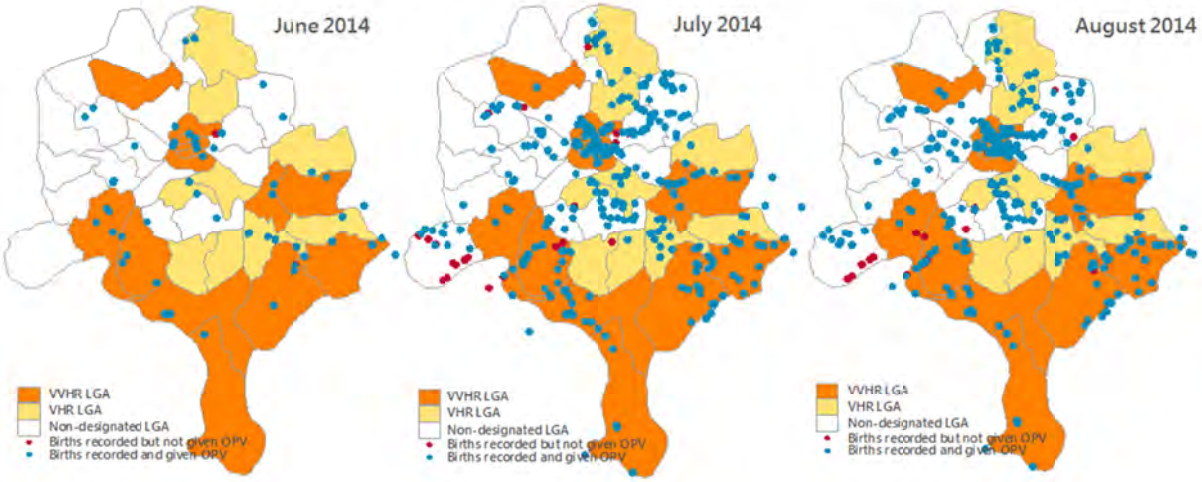


3,500 VCMs in Kano contribute to reducing non-compliance

August 2014



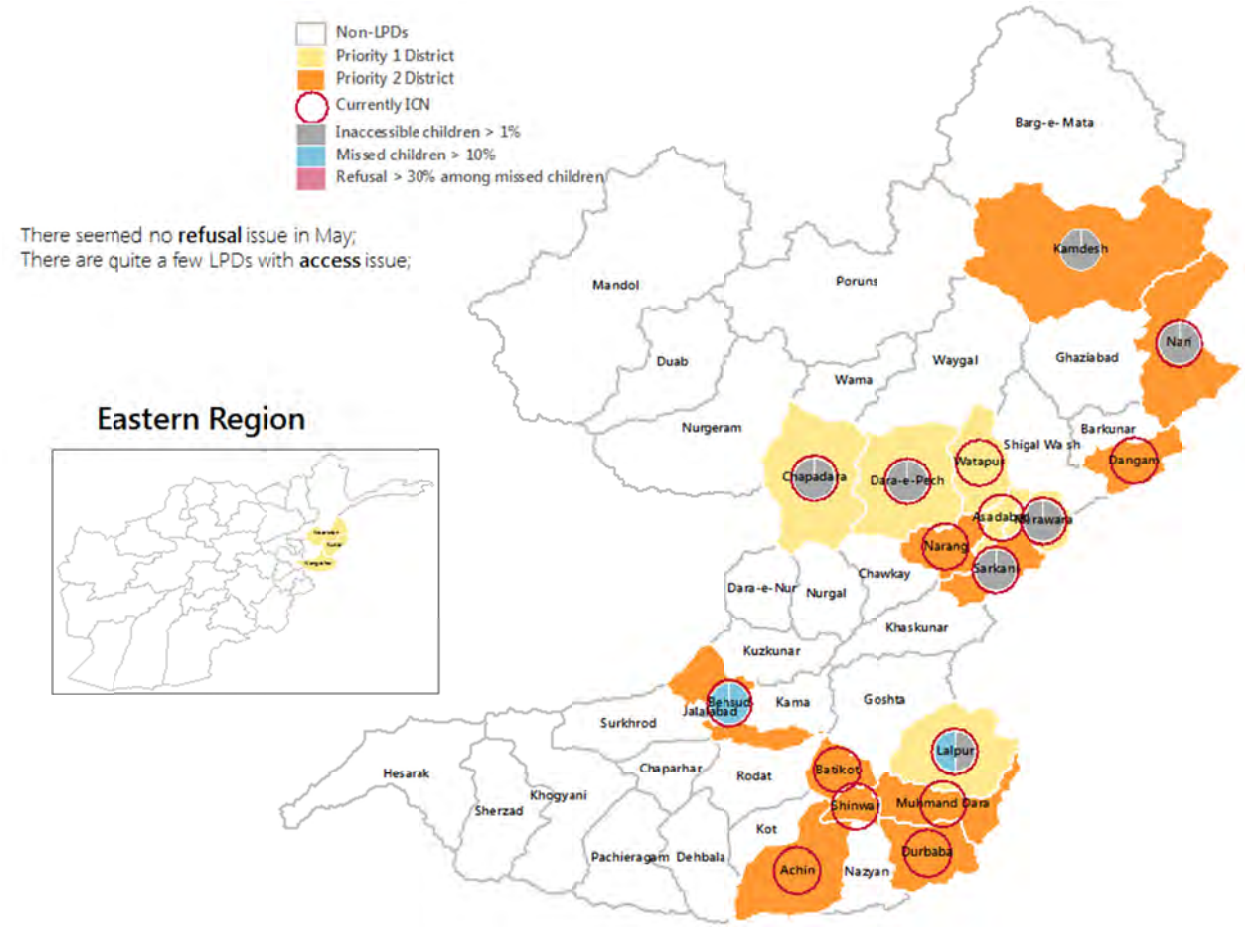
Recording Newborns in Kano



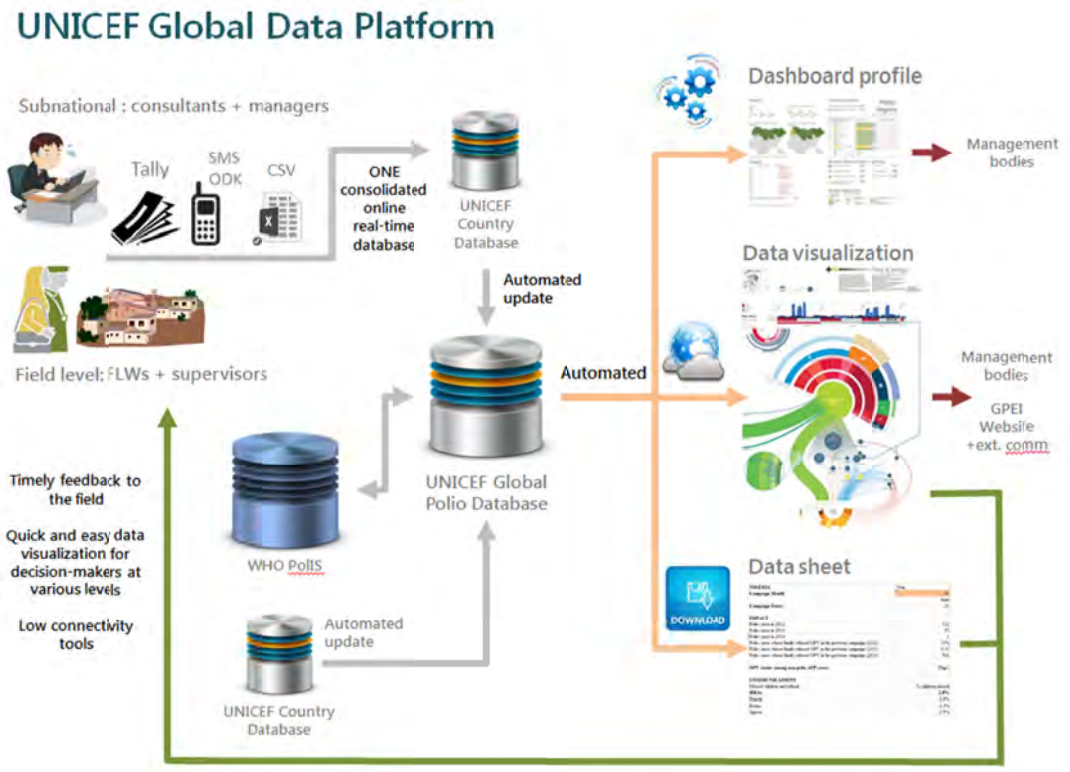
One of the activities of the VCMs during IPD interval is recording newborns in the settlement. They also attend the naming ceremony of the baby (usually on the 7th day after birth) and give the 1st dose of OPV. They also sensitize the caregiver to initiate routine immunization.

In Kano state, there are almost **7,000** newborns recorded by VCMs in 2014 (as of August). More than **90%** of these newborns have been given the 1st dose of OPV by VCMs.

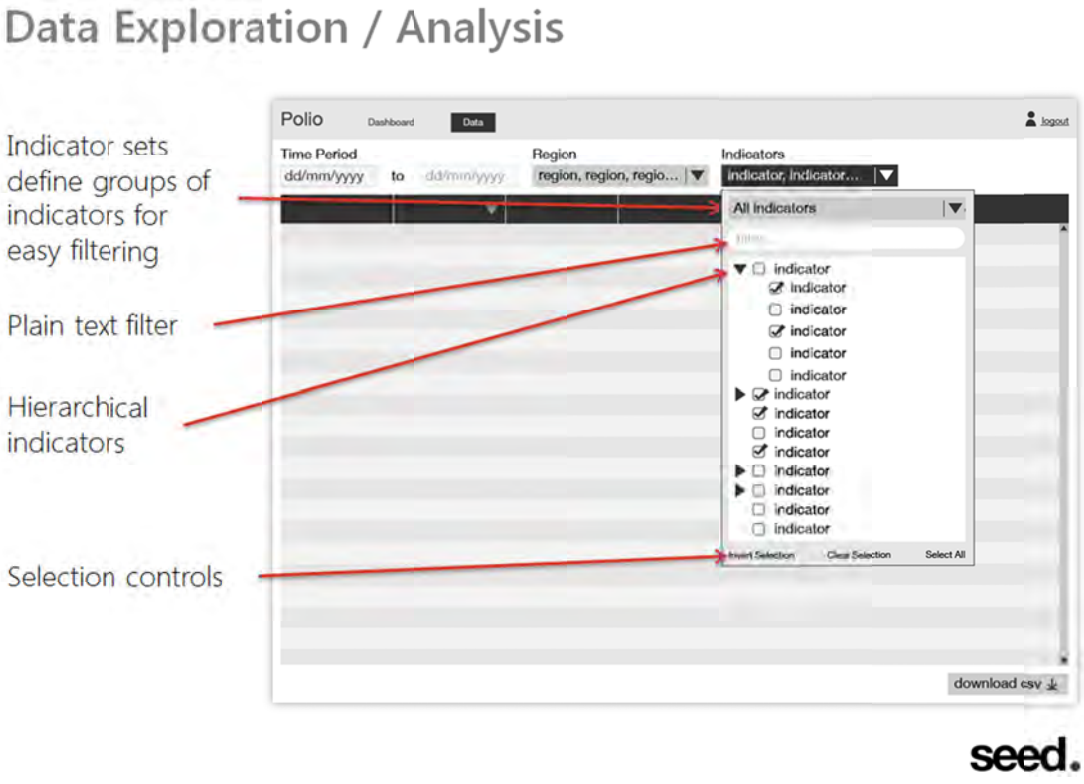
Assessing situation in Afghanistan for deployment of social mobilization network



The future of UNICEF Global Polio Data Platform



Data organization tool



Country dashboard

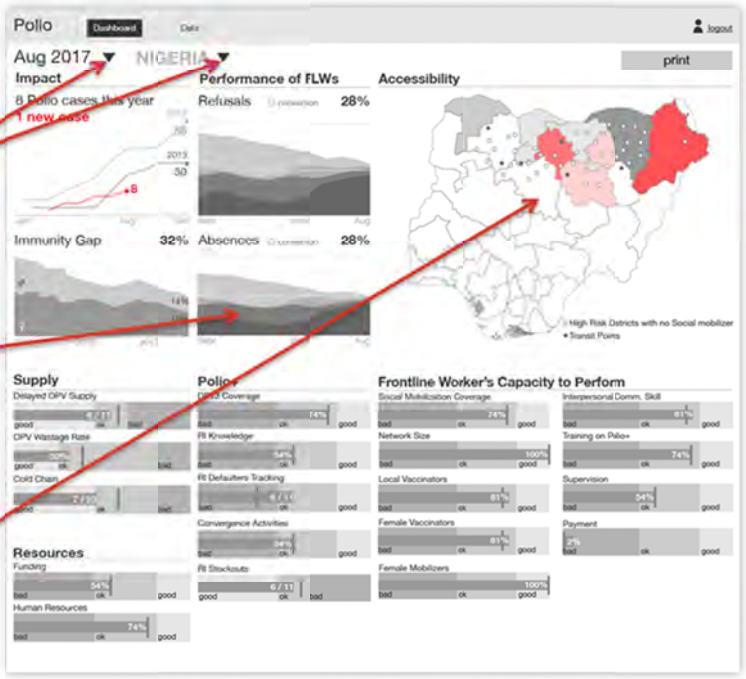
Management Dashboard

Fits 1024 px wide screen

Title doubles as dropdowns to select campaigns and countries

Interactive charts have more space and can show more information

Hover over a region to view its data in the dashboard, click to zoom into that region



seed.