

How to use Predictor Rules to Identify Outliers

This document describes the detailed steps required to make an outlier for ANC1 following the instructions outlined in the exercise “Outlier Validation Rule Alert Notification Exercise.” Please follow the instructions in detail to create the predictor as part of this exercise.

Introduction

1. **Why use Predictor to identify extreme outliers?** “Extreme outliers” are values which are highly suspicious and which need to be double checked for accuracy. These are different from the values normally reported by a health facility. If extreme outliers are found to be erroneous, then they should be edited.
2. The WHO Data Quality Tool can rapidly identify extreme outliers; however, these values are relegated to the WHO Data Quality Tool itself and can not be used for any other purposes. By using predictors to identify extreme outliers, we can create notifications, view these values on the dashboard, or use it in combination with other DHIS2 functionality. Additionally, we can define the formula used to calculate or identify the outlier. This flexibility can accommodate a large variety of outlier detection methods.
3. **Overview of the steps involved:**
 - A. Configure a new data element and a new dataset.
 - Ensure you apply the proper sharing settings to the dataset!
 - B. Configure a new Predictor rule.
 - C. Configure the Scheduler (Whilst we will describe these steps, you will not perform this operation in this exercise; You manually run the predictor you create in this exercise instead).
 - D. Use the Ddata export or data entry to confirm that new data have been generated.
 - E. Update the Analytics tables (this will run every 15 minutes for you in this course; you will not be able to run it yourself).
 - F. Review the data outputs in one of the DHIS 2 visualization tools
4. **Using these instructions** - At key stages of the process (e.g. after configuration of the Predictor rule; after configuration of the Data Validation rule), it is advisable to pause to verify that the configuration is working correctly. Remember, this is not a graded exercise, so take your time, and have a go at completing the exercise.

In a live implementation, several hours may be required to run a Predictor rule. In addition, even after the Predictor rule has generated new data, the data may not be visible until the Analytics have been updated; this is another step for which more than an hour may be required (in this course it will run every 15 minutes). As a result, considerable patience and several prolonged pauses in the workflow are likely to be required the first time the guidance is followed. However, once you have successfully

configured one data element/indicator and confirmed that the resulting outputs are correct, the configuration process for each subsequent data elements or indicator can involve cloning and should take much less time.

Step 1: Configuring new data elements and a new dataset

Before a Predictor rule can be configured, a new data element must be configured which will be used to store the value of the outlier threshold. To get started, open the Maintenance app, and then select *Data Element* to define a new data element.










1. **Configure a data element for outliers from the last 12 months:** Give your new data element a name such as “ANC 1 outlier – your initials”;
 - 1.1. Set the Domain type to “Aggregate”
 - 1.2. Set the Value type to “Positive or Zero Integer”
 - 1.3. Set the Aggregation type to “Sum”
 - 1.4. Leave the “Store zero data values” unchecked
 - 1.5. Set the category combination to “None”
 - 1.6. Leave the Aggregation levels blank
2. **Configure additional data elements depending upon how you want to visualize the suspicious data:** Predictor rules can also be used to visualize outliers in other ways besides the one described in this document: validation rules, maps showing the locations of health facilities reporting extreme outliers, or others. Refer to separate documents for instructions on how to do this.
3. **Create a new data set called “Outliers – your initials” Add the new data element to the dataset you have just made:**
 - 3.1. Use the maintenance app to create a new data set.
 - 3.2 Leave the default expiry days, open future periods for data entry and days after period to qualify for timely submission
 - 3.3 Set the period to monthly (the same period as you are generating outliers for)
 - 3.4 Add in the data element(s) you have just made to store your predictor value(s). You should be able to find the data element quickly by using your initials. You may also want to add in the data element “ANC 1 visits” just so you can perform a comparison of the actual value against the generated outlier.
 - 3.2. Ensure you **assign this data set to the facility level.** (all facilities under the national level)
 - 3.4. After making the data set and adding in your data elements, share the data set with your user:
 - 3.3.1 - Metadata : Can edit and view
 - 3.3.2 - Data : Can capture and view

Sharing settings

SND_Outliers

Created by: Student Test

Who has access

	Public access No access		
	External access No access		
	Student Test		

Add users and user groups

Enter names

METADATA

☒ Can edit and view

Can view only

DATA

☒ Can capture and view

Can view only

No access

CLOSE

June 17, 2021

You can use this data set to review your predictors in the data entry and/or the data export app by exporting the values of the new data element to confirm that the outlier data have been generated. This is an interim measure until analytics has been successfully run.

Step 2: Configure the Predictor

1. Go to Maintenance, then Other, then Predictor.
2. Select the “+” icon to create a new predictor.
3. Give the predictor a name, for example “ANC 1 Outlier – Your initials”.
4. Click on “Output data element (*)” and a window will appear with the names of all existing data elements. On the “filter list”, find the new data element you have just created. You should be able to find them more easily by typing in your name. Find and click on the data element you created in the dropdown list that appears. The name of the new data element should now appear on the line beneath “Output data element (*)”.
5. Set the Period type to “Monthly” and Organisation unit level to “Facility”. This will

generate the predictor values monthly at the facility level.

6. Click on Generator (*). Here is where you enter the formula for generating the data for the new data element. Leave the Missing value strategy set to “Skip if all values are missing” (this is located above the description under the heading “Missing value strategy”).

6.1. We will define a threshold in this case as we did in the overview video; that is a value that is 3 standard deviations above the mean for the last 12 months, within the same facility. This will allow us to generate values that are 98% above the average value within a facility over a 12-month period.

6.2. To generate this value use the syntax:

6.2.1. $\text{avg}(\text{dataelement}) + 3 * \text{stddev}(\text{dataelement})$

6.2.2. It will look like this when using ANC 1 visits to generate the outliers

The screenshot shows the DHIS2 Generator interface. At the top, the 'Missing value strategy' is set to 'Skip if all values are missing'. Below this, the 'Generator (*)' section is active. On the left, the 'Description' field is labeled 'ANC1 Outlier'. The formula entered is `avg({RvArfQFKdXe}) + 3 * stddev({RvArfQFKdXe})`. Below the formula is a toolbar with symbols for parentheses, multiplication, division, addition, subtraction, and 'Days'. On the right, the 'DATA ELEMENTS' tab is selected, showing a list of data elements: 'ANC 1 visi', 'ANC 1 visits', 'ANC 1 visits expected', and 'ANC 1 visits present'. The 'ANC 1 visits' element is highlighted. At the bottom, the formula `avg(ANC 1 visits) + 3 * stddev(ANC 1 visits)` is displayed in a text box. Below this, the word 'Valid' is shown in green. At the bottom right, there are 'CANCEL' and 'SUBMIT' buttons.

6.2.3. **NOTE:** All statements (EX: “stddev” or “avg”) in DHIS 2.34 and above must be lower case. In versions lower than 2.34, it can be in either capital letters or lowercase. It is therefore recommended to use lower case as this will be compatible with all versions of DHIS 2.

6.2.4. In the next step when creating validation rules, we can use these values to compare to data as it is entered to see if an actual, entered value is beyond this extreme value that we are generating.

7. Once the expression is well-formed, click on Submit.
8. Ignore the button for “*Sample skip test*”.
9. Set the sequential sample count to 12. This is the number of months of data which are used for the average and the standard deviation of the Generator formula. If last month is, for example, July 2021, then to calculate the threshold, the Generator formula takes the average of the values for the same health facility for July 2020 to June 2021, then adds 3 times the standard deviation of the values for July 2020 to June 2021.
10. Leave the Annual sample count (*) set to 0.
11. Leave the Sequential skip count blank. It should look like this when all the details are filled in:

Sequential sample count (*)

12

Annual sample count (*)

0

Sequential skip count

SAVE

CANCEL

- 12 Save the new Predictor rule.

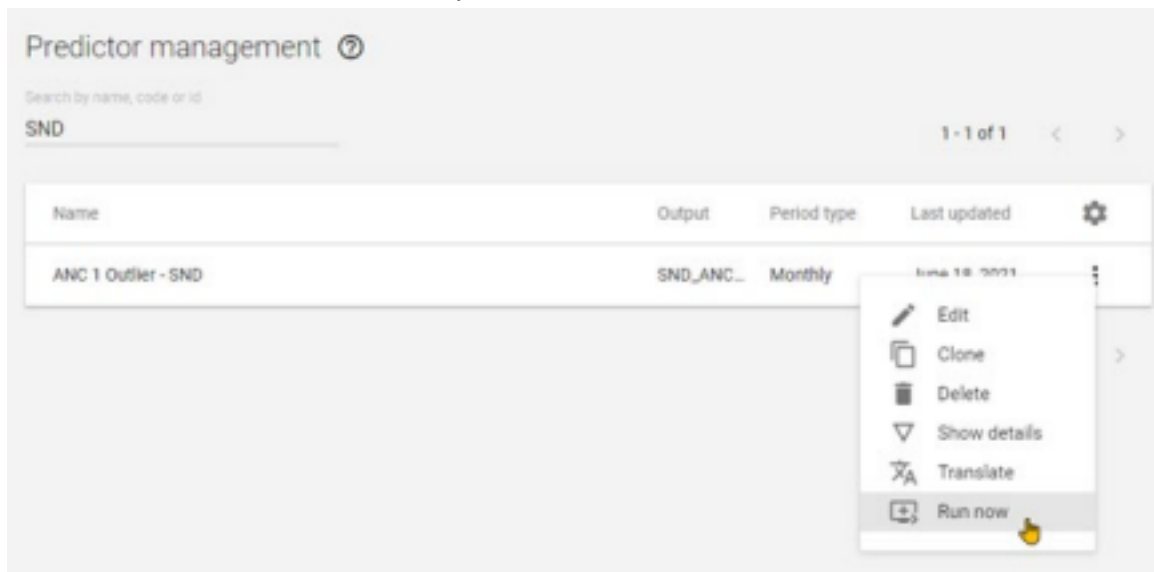
Step 3: Run the predictor

In this exercise, you will run the predictor to ensure it is generating values correctly. In a live setting, you can also use this process to test your predictor, but you will want to set up the predictor to run routinely via the scheduler app. Please refer to *Additional Information - Using the scheduler* at the bottom of this exercise for more information on how to configure predictors to run routinely.

1. To run/test the predictor you have made during this exercise:
 - 1.1. Filter your predictor from within the “Predictor management” screen. You can

use your name/initials to quickly filter it out from the list.

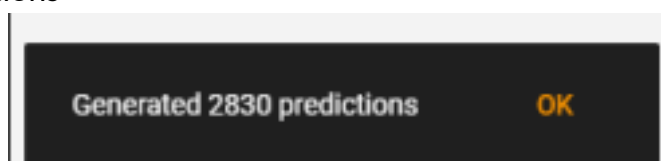
1.2. Select the 3 dots icon followed by “Run now”:



1.3. Run the predictor from January 1, 2021 to June 30, 2021

The screenshot shows the 'Run predictor' dialog box. It has two input fields: 'Start date (*)' with the value '2021-01-01' and 'End date (*)' with the value '2021-06-30'. At the bottom, there are two buttons: 'CANCEL' and 'RUN PREDICTOR'.

2. You should see a message that you have generated a certain number of predictions



Step 4: Check that your new predictor rule is running correctly.

Option 1 : Export the data

One way to determine whether the predictor has run correctly is to use the Data Export app to export data for the new outlier data element. This will allow you to quickly review all of the generated values at once. Otherwise, you may have to wait until the Analytics tables have been updated (this is done every 15 minutes in your demo instance). Once they have updated, the outlier threshold data can be visualized with the Pivot table or Data Visualizer apps. Meanwhile, use the following steps to check your predictor rule:

1. Launch the Import/Export app and select “Export data”:



2. Under Organisation unit, select the national level.
3. Then select *“Include the first level of organization units inside selections”*
4. Under Datasets, select the dataset that you have made. You can use your initials to filter out the list.
5. Set the *Start Date* to the January 1, 2021. Set *End Date* to June 30, 2021.
6. Set *Format* to CSV.
7. Set *Compression* to Uncompressed.
8. With all options selected, it will look like the following:

Overview

Import

Data import

Event import

GML import

Metadata import

TEI import

Export

Data export

Event export

Metadata dependency export

Metadata export

TEI export

Job overview

Data export

Export metadata, such as data elements and organisation units, in DXF 2 format.

Basic options

Organisation unit(s) to export data from

☒ National

☐ Region A
 ☐ Region B
 ☐ Region C
 ☐ Region D

Include first level units

☒ Include the first level of organisation units inside selections

Which data sets should be included in export?

Filter data sets

ANC

DHS coverage estimates

EPI

External data

HMIS

Immunisation

Malaria

➡

➡

⬅

⬅

Selected data sets

SND_Outliers

Date range to export data for

Start date

End date

12/31/2020

06/30/2021

What format should the data be exported as?

☐ JSON
 ☒ CSV
 ☐ XML

Compression mode

☐ Zip
 ☐ GZip
 ☒ Uncompressed

Advanced options

Export data

9. Click on *Export data* when you have made all of your selections and wait for the icon for the CSV file to appear in the lower left of the screen.
10. Open the CSV file and check to see whether you have generated your outliers (you should have 2830 values in the CSV file)

Option 2 : Check via data entry

1. Navigate to the data entry app
2. Select a facility
3. Select the dataset you have made
4. Select a period between January – June 2021

Data Entry

Facility 204 - February 2021 - No Data Element Selected

Organisation Unit: Facility 204

Data Set: SND_Outliers

Period: February 2021

Run validation, Print form, Print blank form

Filter in section	Value
ANC 1 visits	58
SND_ANC1 Outlier	78

Option 3: Data Visualizer

After analytics has run, you can create a pivot table at the facility level comparing ANC 1 visits with the ANC 1 threshold. You can do this around 15 minutes after you generated your outliers. If you do not immediately see results using this method, it means that analytics has not yet run. Please wait a few minutes and try again.

		January 2021	February 2021	March 2021	April 2021	May 2021	June 2021
A-1 District Hospital	ANC 1 visits	160	160	211	165	260	241
	SND_ANC1 Outlier	291	302	302	295	299	316
A-1 NGO Hospital	ANC 1 visits	114		101	119	149	85
	SND_ANC1 Outlier	231	233	225	235	236	232
Facility 1	ANC 1 visits	165	128	125	151	90	146
	SND_ANC1 Outlier	244	243	218	212	213	217
Facility 10	ANC 1 visits	72		56	67	60	67
	SND_ANC1 Outlier	105	105	105	105	105	102
Facility 100	ANC 1 visits	216	228	178	184	213	235
	SND_ANC1 Outlier	437	438	426	424	373	347
Facility 101	ANC 1 visits		68	81	69	66	69
	SND_ANC1 Outlier	114	116	112	112	112	112

Once you have completed this step and verified that your Predictor Rule is running correctly, you can return to Step 4 of Part 1 of this exercise.

Part 3: Additional Information – Using the scheduler

Note: This instruction is being provided to you as additional information. You will not be able to perform this operation within the instances accessible via the data quality course as this can result in these systems routinely crashing (in order to save everyone's work, we can not routinely reset these systems). You should try this however when creating your predictors in a development system as this process allows you to automate the generation of these predictors. Alternatively, if you do not have access to such a system, please use the [DHIS 2 play demo](#) to practice configuration of jobs within the scheduler app.

1. Launch the Scheduler app by selecting it from the apps menu.
2. Create a new scheduled job by using the “+” icon
3. Give the scheduled job a name such as “Generation of ANC 1 outliers”
4. Set the Job type to Predictor
4. Leave *Select frequency* set to *Custom*
5. The Cron expression determines when the job will start. Learn more about cron expressions here:
https://docs.oracle.com/cd/E12058_01/doc/doc.1014/e12030/cron_expressions.htm.
This is how all jobs in DHIS 2 are scheduled. We want the predictors to update once every day. Here is an example expression that will run at 15:31 every day:

0 31 15 ? * *

The expression reads as follows

- 0 - The start of the expression
- 31 - The minute which the job starts
- 15 - The hour in which the job starts
- ? * * - This tells the system to run the job once everyday

Based on this, you could change the expression to suit your needs, ensuring that it completes before your analytics job is performed. For example, if we wanted it to run at 16:40 everyday, we could change the expression to look like this :

0 40 16 ? * *

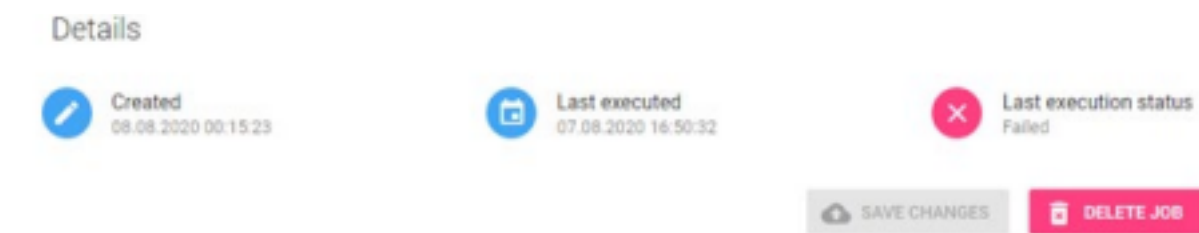
6. Set the Relative start to -395.
7. Set the relative end to 30; With this Relative start and Relative end, the Predictor job will identify any extreme outliers reported in the last 12 months (note this does not include the current month, it is the last 12 months starting from last month)
8. Search and add each of the predictors you want to run as part of the schedule from the *Predictors* line
9. Leave *Predictor groups* blank. Our example will look like this when completed

The screenshot shows a web interface for configuring a job titled "Generation of ANC 1 outliers". The form is divided into several sections:

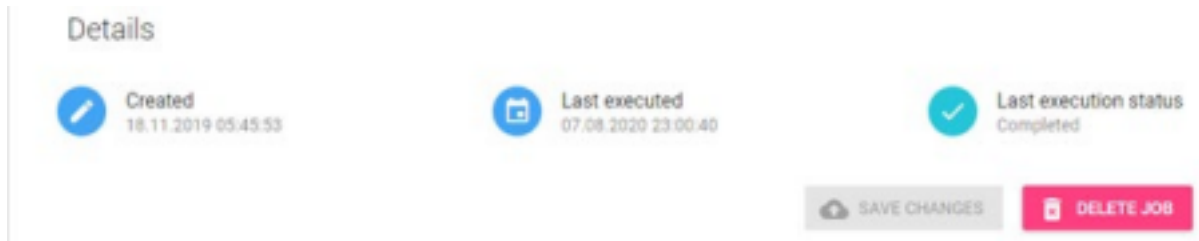
- Attributes**: Includes a "Name" field with the value "Generation of ANC 1 outliers", a "Job type" dropdown menu set to "Predictor", and a "Select frequency" dropdown menu set to "Custom". A "Cron expression" field contains the value "0 40 16 ? * *".
- Parameters**: Includes a "Relative start" field with the value "-395" and a "Relative end" field with the value "30".
- Predictors**: A list of selected predictors, currently showing "ANC 1 Outlier - SND" with a close button (X).
- Predictor groups**: A field for adding predictor groups, currently empty.

10. Click on *Add job* when you are finished
11.
 - The name of your new or modified Predictor job should now appear within the list of "Scheduled jobs".
 - The Type should be shown as Predictor
 - The Status should be Scheduled
 - Next execution should show the day and time it will be run next
12. You can run the predictor job to test it by selecting the "Run now" button between the status and next execution columns.
13. Once the Predictor has finished running (refresh your screen after about 2-3

minutes), select it and examine the “Last execution status” at the bottom of the screen. The status will either be “*Last execution status: Failed*”, as you can see on the right:



or it will be “*Last execution status: Completed*”:



If the job was completed successfully then the values should be generated correctly. Check this was done properly via exporting the data, viewing the data in data entry or viewing it in data visualizer after analytics has been run.