**How to use create predictor rules to identify outliers**

## Description

This document described 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 very different from the values normally reported by a health facility. If extreme outliers are found to be erroneous, then they should be edited.

The WHO Data Quality Tool (DQ Tool) does an excellent job of rapidly identifying extreme outliers; however these values are relegated to the DQ 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 any other DHIS2 functionality. Additionally, we can define the formula used to calculate or identify our outlier. This ends up being rather flexible in practice to accommodate a large variety of outlier detection methods.

* 1. **The steps involved**
     1. Configure a new data element and a new dataset
     2. Configure a new Predictor rule
     3. Configure Scheduler [we will describe these steps; however you will not perform this operation in this exercise]
        + We will have you manually run the predictor you create in this exercise instead
     4. Use Data Export or data entry to confirm that new data have been generated
     5. Update the Analytics tables (this will run every 15 minutes for you in this course; you will not be able to run it yourself)
     6. Configure the pivot table to visualize the outlier values
  2. **The key to successful use of Predictor** -- For success, the following instructions must be followed with great attention to the details. 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. Instructions are provided on how to do this.

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, these data may not be visible until the Analytics have been updated – 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 for one data element/indicator and confirmed that the resulting outputs are correct, the configuration process for each subsequent data elements/indicator can involve cloning and should take much less time.

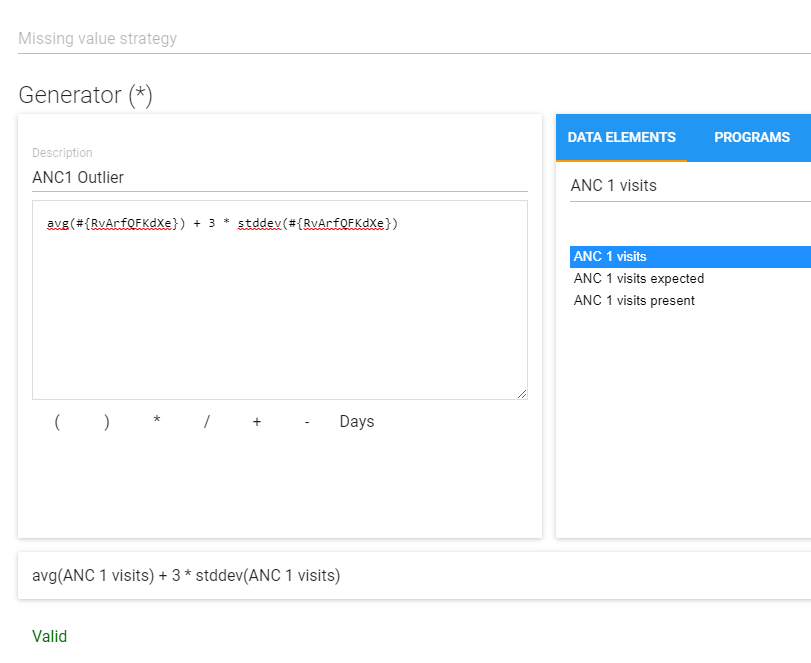
### Step 1: Configuring new data elements and a new dataset

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. Use Maintenance – 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 name”
     1. Set Domain type to “Aggregate”
     2. Set Value type to “Positive or Zero Integer”
     3. Set Aggregation type to “Sum”
     4. Leave “Store zero data values” unchecked
     5. Set category combination to “None”
     6. Leave 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, etc … Refer to separate documents for instructions on how to do this.
  3. **Create a new data set called “Outliers – your name” Add the new data element to the dataset you have just made** –
     1. Use the maintenance app to create a new data set.
     2. Ensure you *assign this data set to the facility level*.
     3. Once you have made the data set, add in the data element(s) you have just made to store your predictor value(s). You can use this data set to review your predictors in data entry or the Data Export app can 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

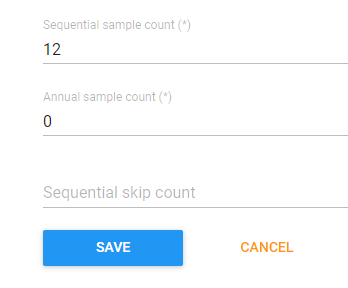
* 1. Go to Maintenance – Other – Predictor
  2. Select the “+” icon to create a new predictor
  3. Give the predictor a name, for example “ANC 1 Outlier – Your name”
  4. Click on “Output data element (\*)” and a window will appear with the names of all data existing data elements. On the “filter list” 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.” By doing this you 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.”
* 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.
* 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.
* To generate this value use the syntax:
  + avg(dataelement) + 3 \* stddev(dataelement)
  + It will look like this when using ANC 1 as to generate the outliers

****

**NOTE: All statements (EX: “stddev” or “avg”) in 2.34 and above must be lower case. In versions lower then 2.34 it can be in eather capital letters or lowercase.**

* 1. Once the expression is well-formed, click on Submit.
  2. Ignore the button for “Sample skip test”
  3. 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.
  4. Leave Annual sample count (\*) set to 0
  5. Leave Sequential skip count blank.

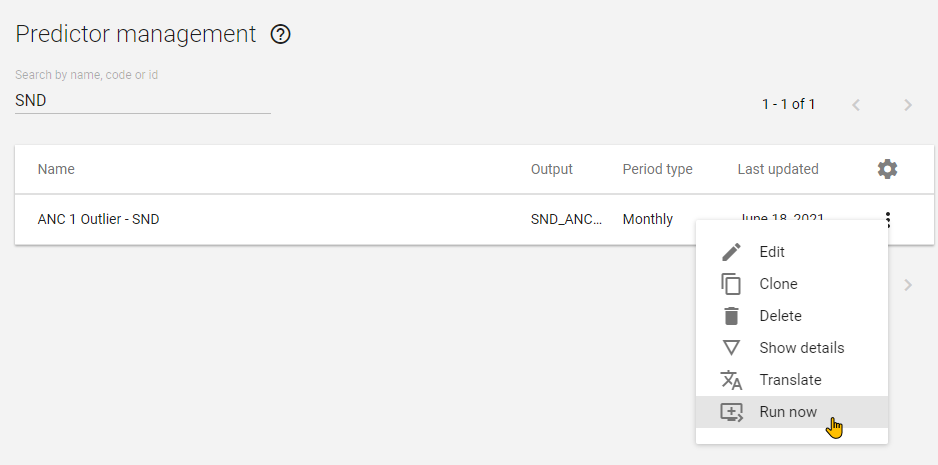
It should look like this when all the details are filled in



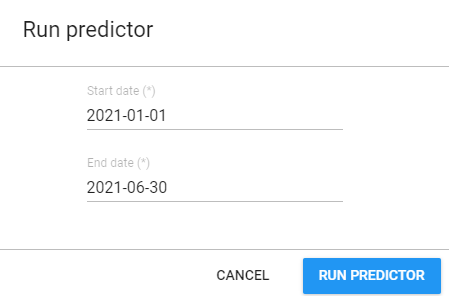
* 1. Save the new Predictor rule.

### Step 3: Run the predictor

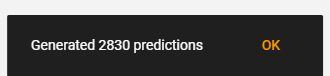
* 1. In this exercise, you will run the predictor to ensure it is generating values correctly. In a live setting, you can also use this process to test your predictor, but you will want to set up the predictor to run routinely via the scheduler app. Please refer to Annex 1 – Using the scheduler for more information on how to configure predictors to run routinely.
  2. To run/test the predictor you have made during this exercise
     1. Filter your predictor from within the “Predictor management” screen. You can use your name to quickly filter it out from the list.
     2. Select the 3 dots icon followed by “Run now”



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



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

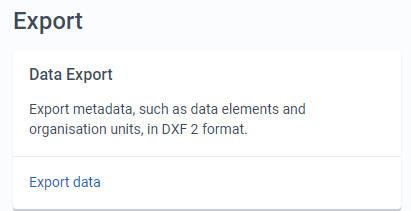


### Step 4: Check that your new predictor rule is running correctly.

#### Option 1 : Export the data

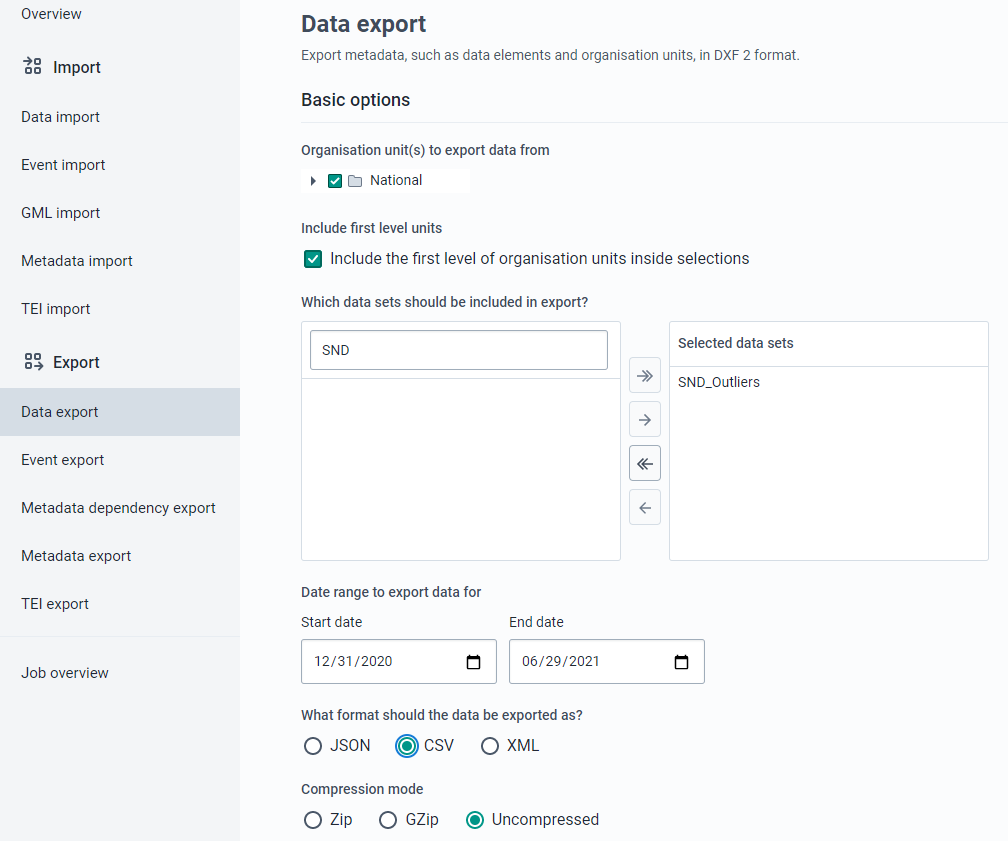
The best way to determine whether the predictor has run correctly is to use the Data Export app to export data for the new outlier data element. Otherwise, you may have to wait until the Analytics tables have been updated (this is done every 15 minutes in your demo instance).. Only then can the outlier threshold data 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”



* 1. Under Organisation unit, select the national level.
  2. Also select “Include the first level of organization units inside selections”
  3. Under Datasets, select the dataset that you have made. You can use your initials to filter out the list.
  4. Set START DATE to the January 1, 2021. Set END DATE to June 30, 2021.
  5. Set FORMAT to CSV.
  6. Set COMPRESSION to Uncompressed.

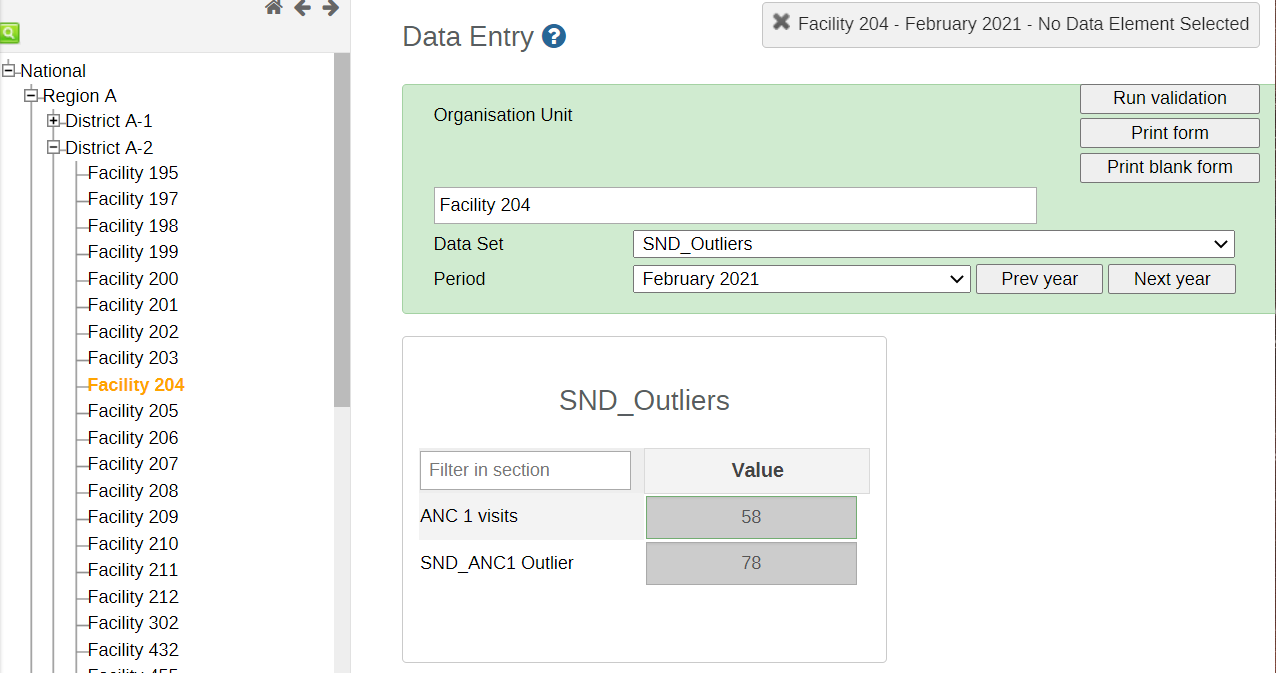
With all options selected, it will look like the following:



* 1. 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.
  2. Open the CSV file and check to see whether you have generated your outliers

#### 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



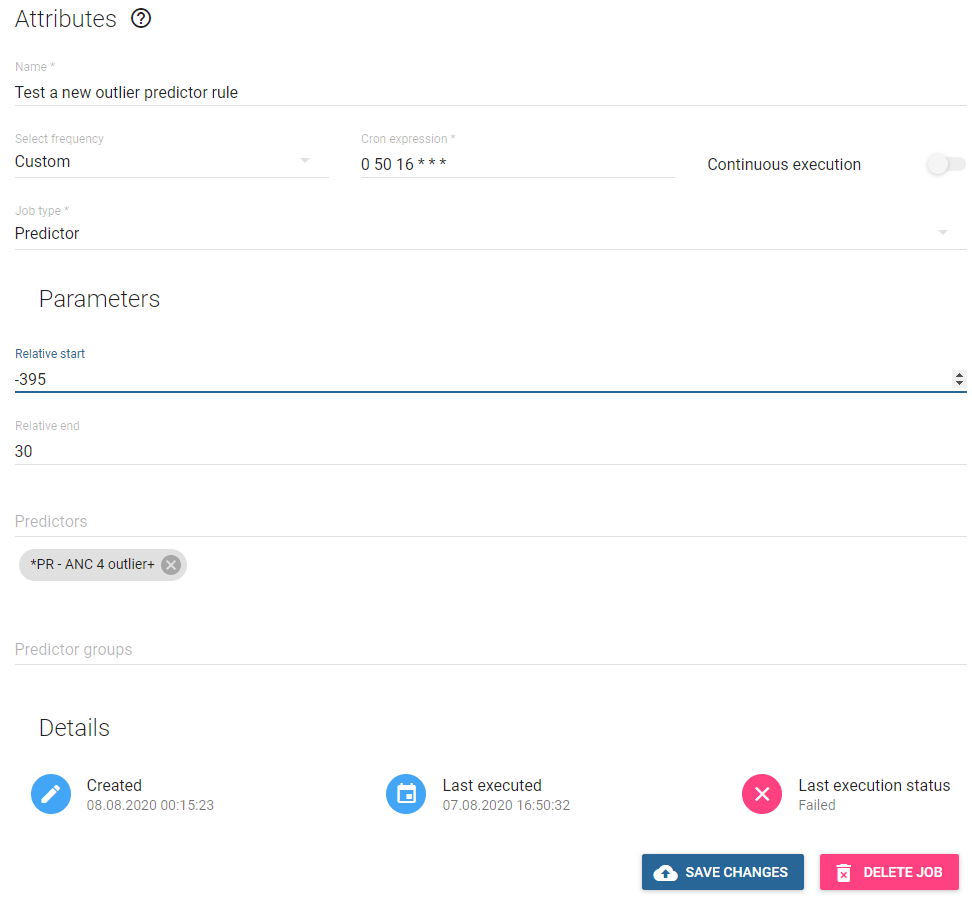
#### Final Check : Data Visualizer

***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.***

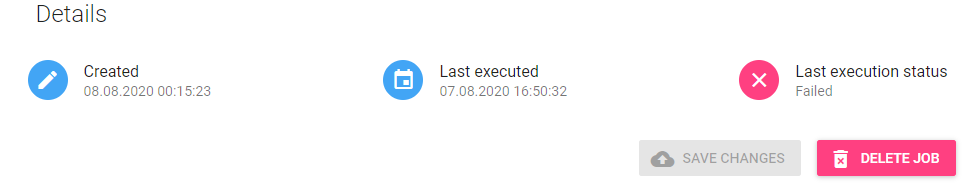


### Annex 1 – Using the scheduler

* 1. Launch the Scheduler app by selecting it from the apps menu
  2. To test a new Predictor rule, edit the existing job named “Test a new outlier predictor rule”. Click on the name of this job and a window will open.



* 1. Leave Name as it is.
  2. Leave Select frequency set to Custom
  3. The Cron expression determines when the job will start. Set this to 00 XX YY \* \* \* where XX is 5 minutes after the current minutes and YY is the current hour. Make sure that there is a single space between each part of this cron expression: 00 and XX and YY and \* and \* and \*. With this expression, the job will run 5 minutes after you set the cron expression.
  4. LEAVE “Continuous execution” SET TO OFF
  5. Leave Job type\* set to Predictor
  6. Leave Relative start set to -395.
  7. Leave Relative end set 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 whatever the last month happens to be)
  8. Leave Predictor groups blank
  9. Click on SAVE CHANGES.
  10. The name of your new or modified Predictor job should now appear under “Scheduled jobs”. The Type should be shown as Predictor. The Status should be Scheduled. Next execution should say something like “07:08:2020 18:10” (if the job were configured in August of 2020). Enabled should be set to on.
  11. Wait 15 minutes, then refresh the screen. If the Predictor is still running the Scheduler will show the Status of the job as Running. If the Predictor has finished running, the status will shows as Scheduled, but the date that it is scheduled will be tomorrow.
  12. Once the Predictor has finished running, click on it and examine the “Last execution status” at the bottom of the screen. The status will either be Failed



or it will be Completed.

