Tableau Prep Note

Download address: <https://www.tableau.com/academic/students>

A picture containing text, clock

Description automatically generated

Tableau Prep is used to preprocess and clean data, the example below is directly connected to csv files, and the output file containing the processed data needs to further connect to Tableau Desktop to do visualization.

Below is the demonstration for Kaplan Meier double curve with same cancer, different medication treatment example

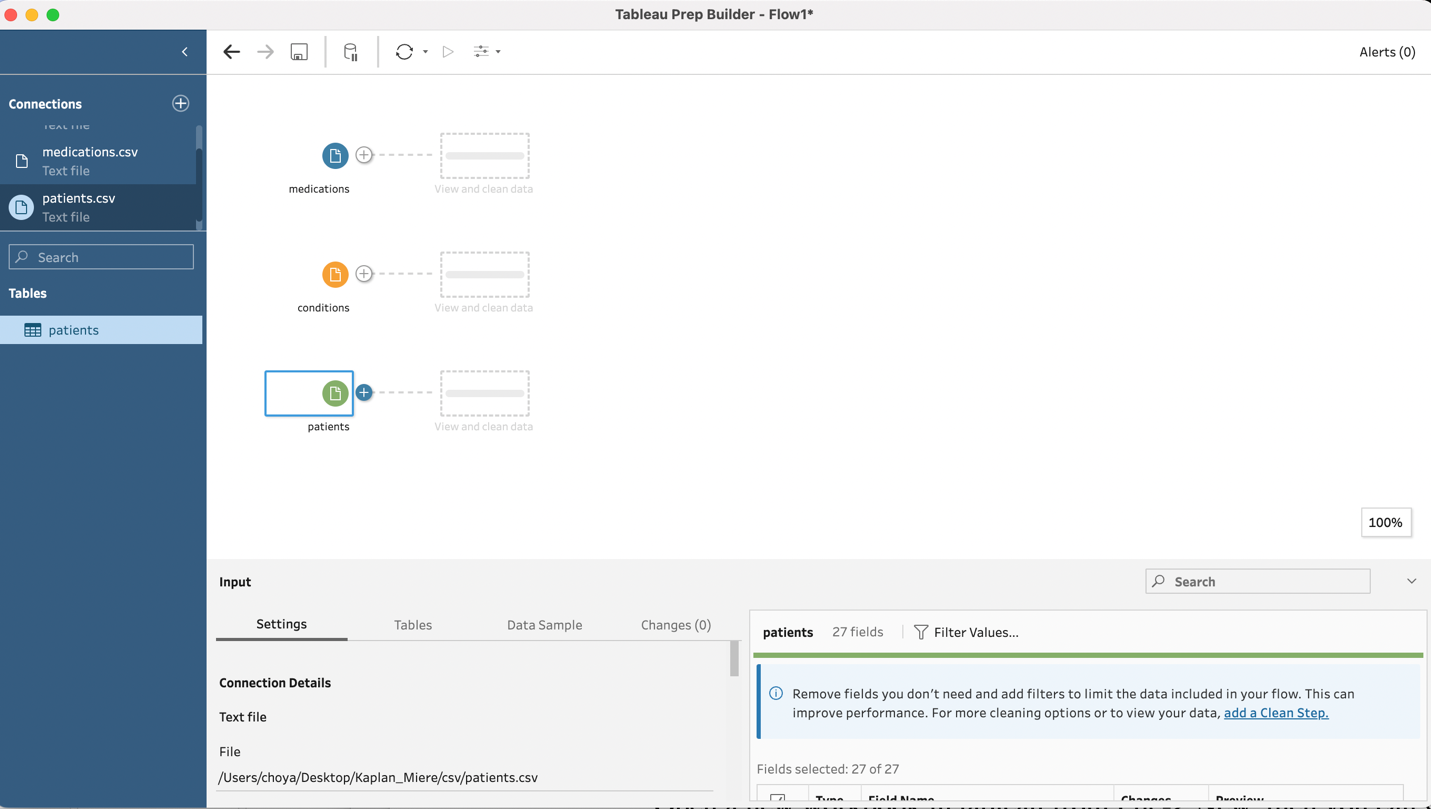
1. Create a Flow in Tableau Prep and Connect to the data

Open Tableau Prep and press the right button next to the Connections -> To a File -> Text file and select the needed csv files in the located path

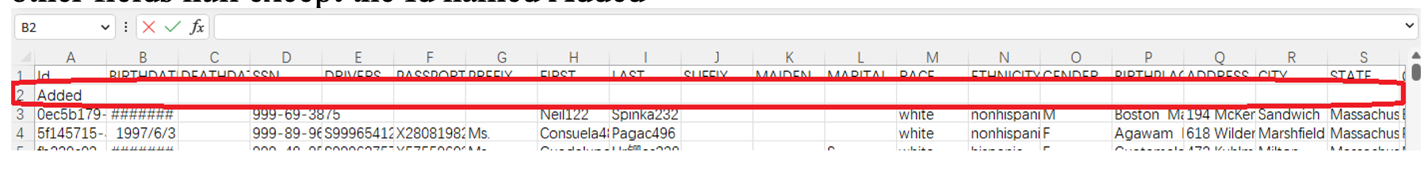
Graphical user interface

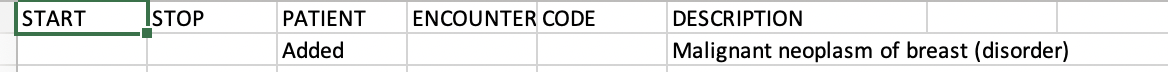
Description automatically generated

then get the window shown below.

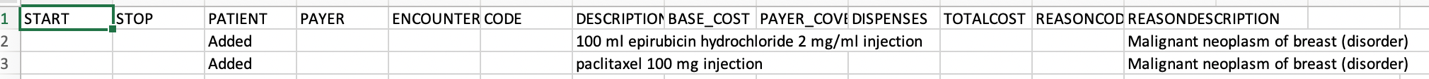


Note, before we connect to the CSV file, we need to do some extra modifications to the csv. file, we need to insert a few blank rows into each file.

* For patients.csv, we add an extra row between the column name and the first record(shown below marked with red). Set the ID as ‘Added’ and leave other columns blank. 
* For condition.csv, we add an extra row Set the ID as ‘Added’ , DESCRIPTION as ‘Malignant neoplasm of breast (disorder)’ and leave other columns blank.



* For medication.csv, we also add two extra rows. For first row set the ID as ‘Added’, DESCRIPTION as ‘Malignant neoplasm of breast (disorder)’ REASONDESCRIPTION as’100 ML Epirubicin Hydrochloride 2 MG/ML Injection’ and leave other columns blank, while second row set the ID as ‘Added’, DESCRIPTION as ‘Malignant neoplasm of breast (disorder)’ REASONDESCRIPTION Paclitaxel 100 MG Injection’ and leave other columns blank.

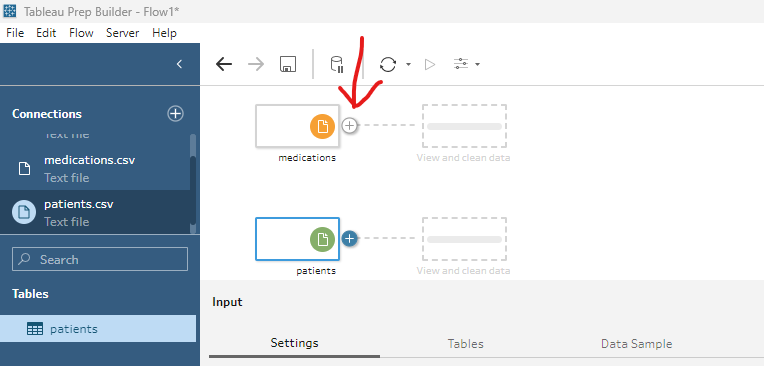


These steps are for the origin points in the Kaplan-Meier curve (Time at 0 the survival probability is 1)

Now we can go back to tableau Prep and choose the modified CSV files.

1. Add the process step into the Flow
   1. Remove duplicate rows by aggregate

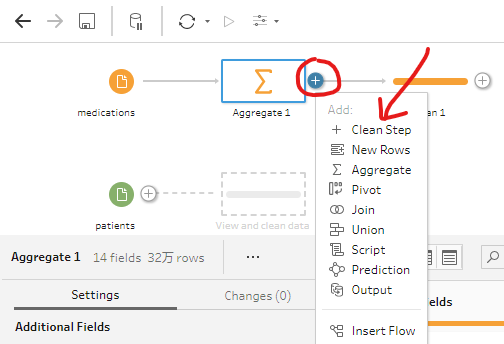
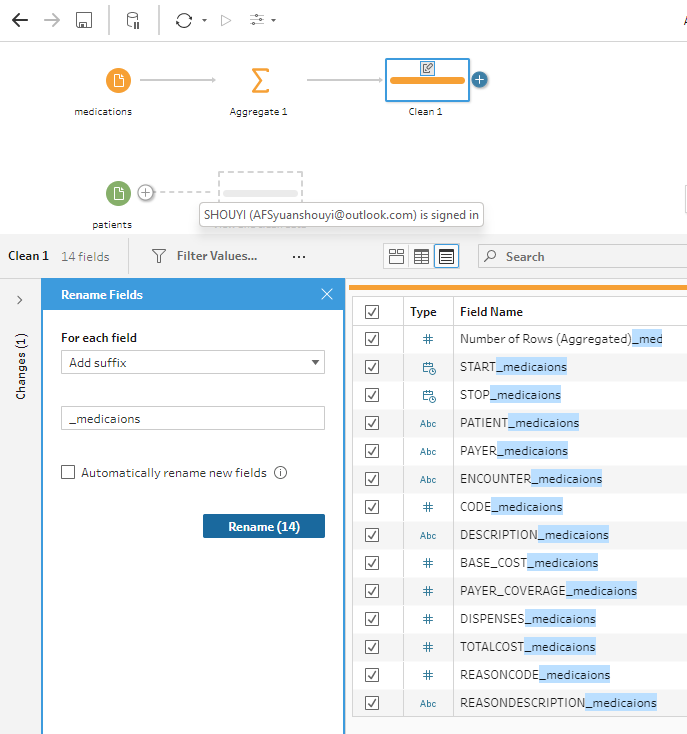
Click ‘+’ add button(shown as image below) near the data -> Aggregate -> choose Add All under Settings.



* 1. Rename column filed

As there are same column name across the table, we add suffix about the source table of the column.

Click ‘+’ add button -> Clean Steps :Rename Fields -> select Add suffix in the For each field in check box

We do the step both for condition and medication and add suffix ‘\_condition’ to column in the condition.csv and ‘\_medication’ to column in the medication’.csv.

* 1. Join files

Click ‘+’ add button -> Join ->

Diagram

Description automatically generated

select the columns that join condition on Condition.csv, patient.csv and medication.csv join with below condition in the applied join

* + - Condition.csv inner join with Patients.csv

on Patient(Condition.csv) == Id(from Patients.csv)

* + - Condition.csv inner join with Medication.csv on

on Patient(Condition.csv) = Patient (Medication.csv)

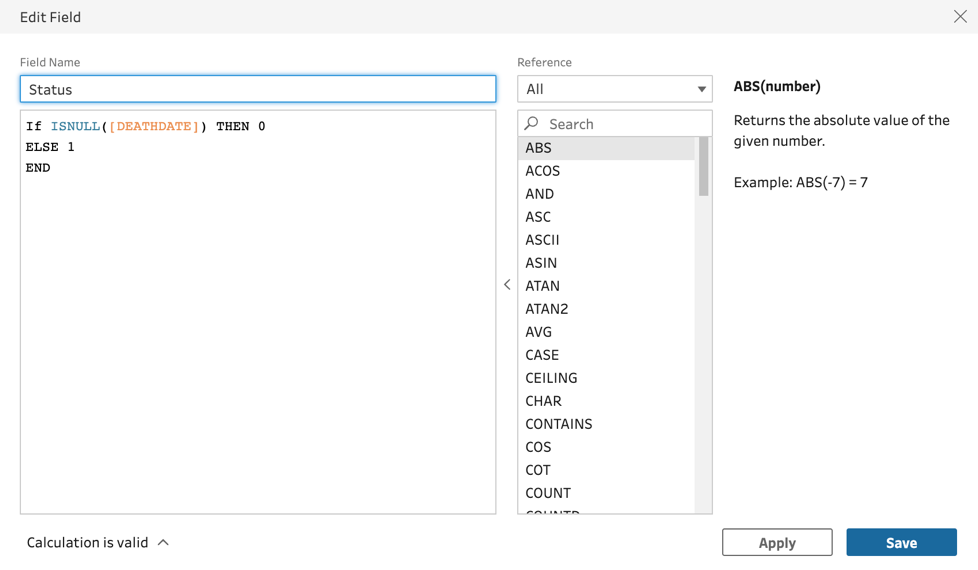
and Description(Condition.csv) == ReasonDescription(Medication.csv)

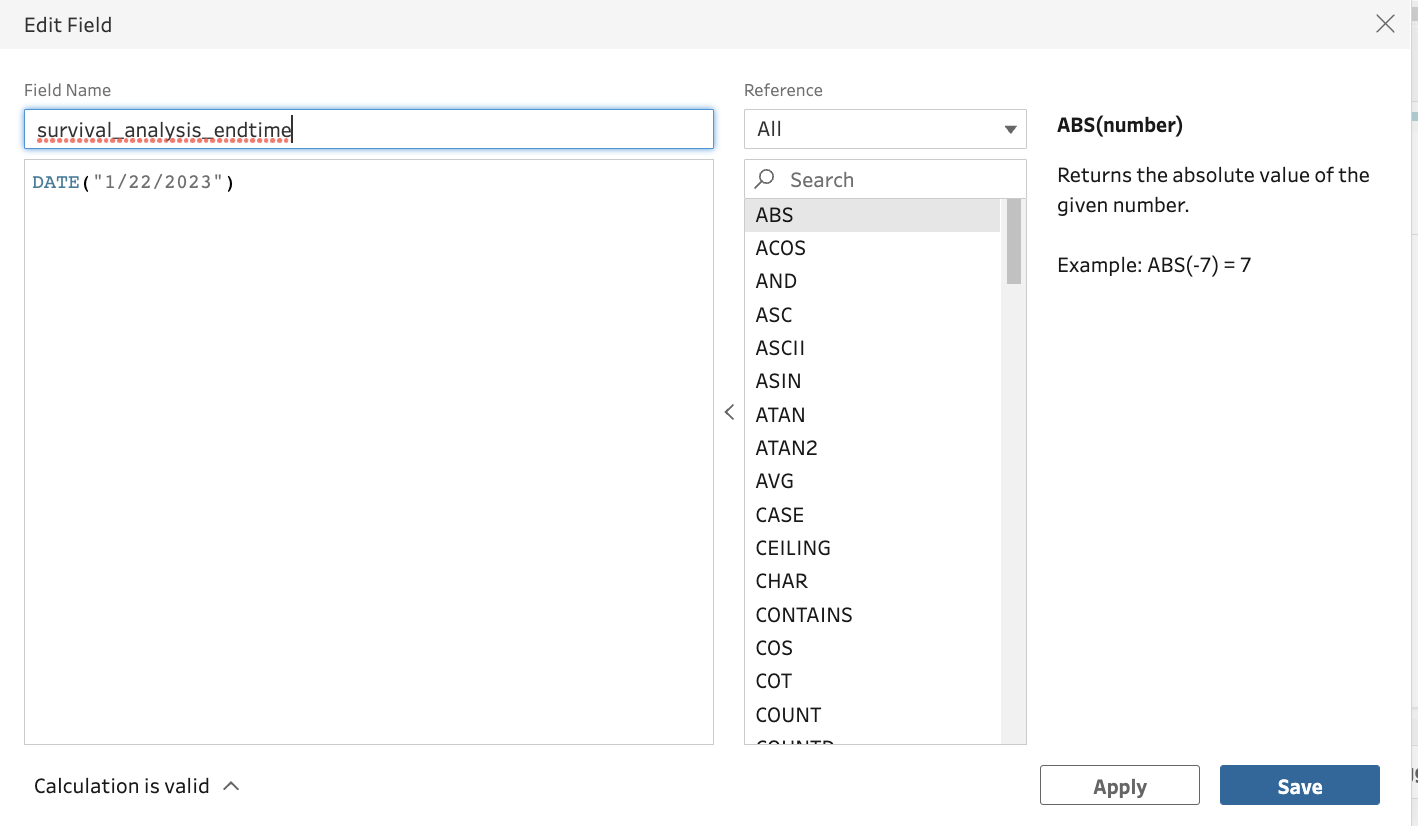
* 1. Add calculated fields

Click ‘+’ add button -> Clean Steps :Create Calculated Field

* + Status

The status of the patient, 1 represents alive while 0 represents death



* + survival\_analysis\_endtime (date when data was extracted) 
  + earliest\_medication\_starttime

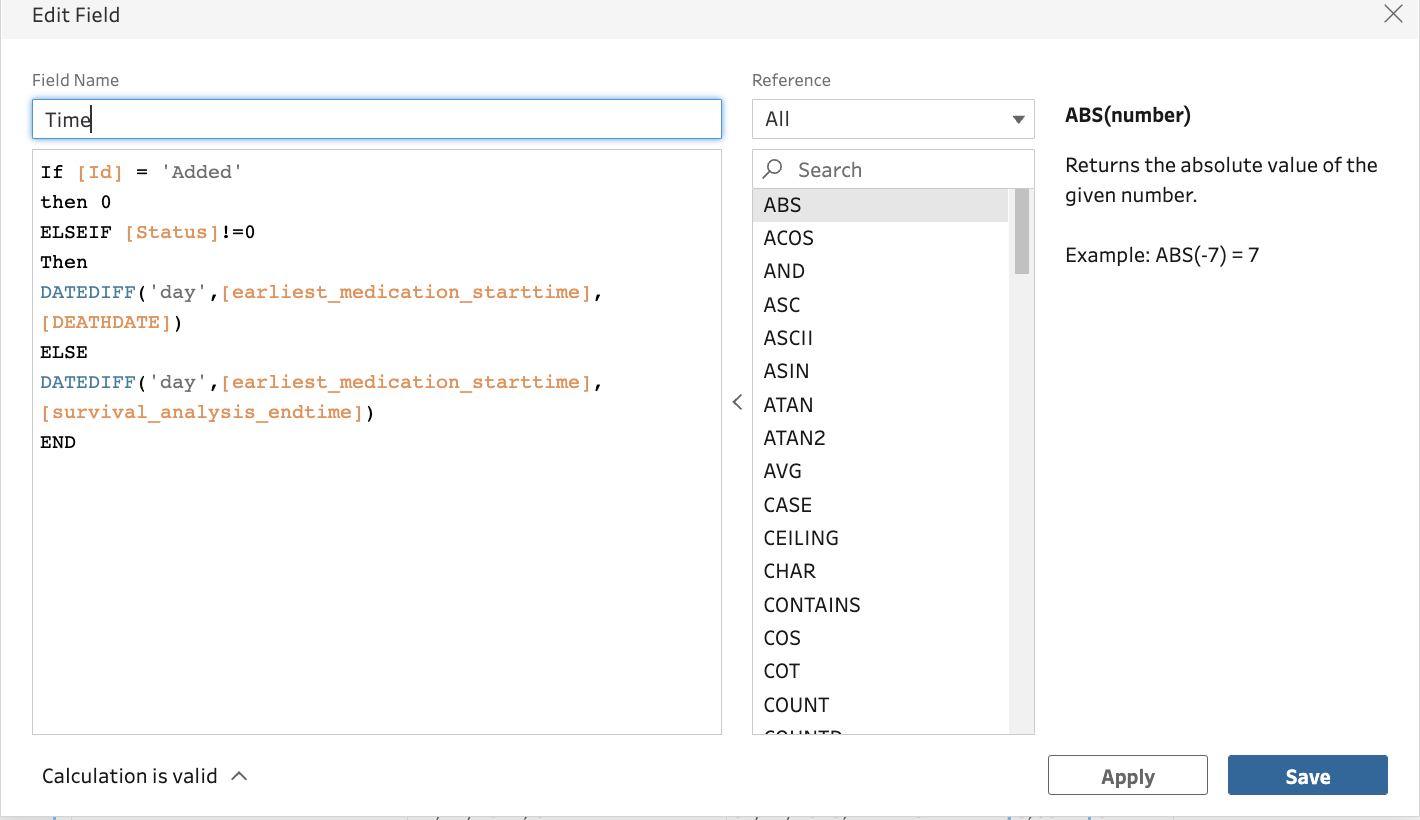
The earliest time for a patient to take the medication which considers as the start time in the given survival analysis

Graphical user interface, application, Word

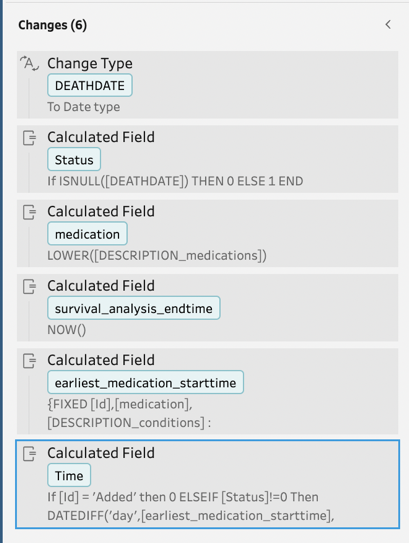
Description automatically generated

* + Time

The survival time for patient

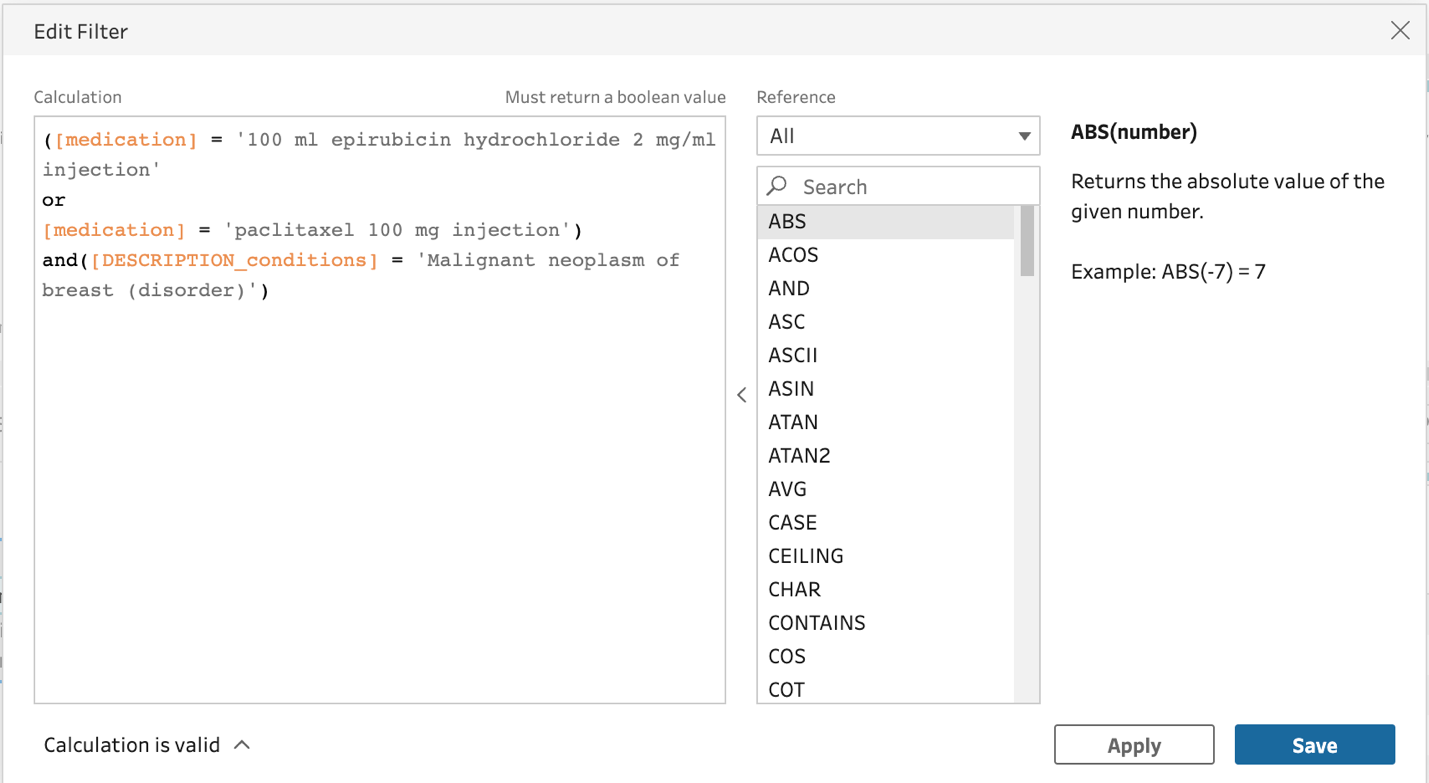


Note:

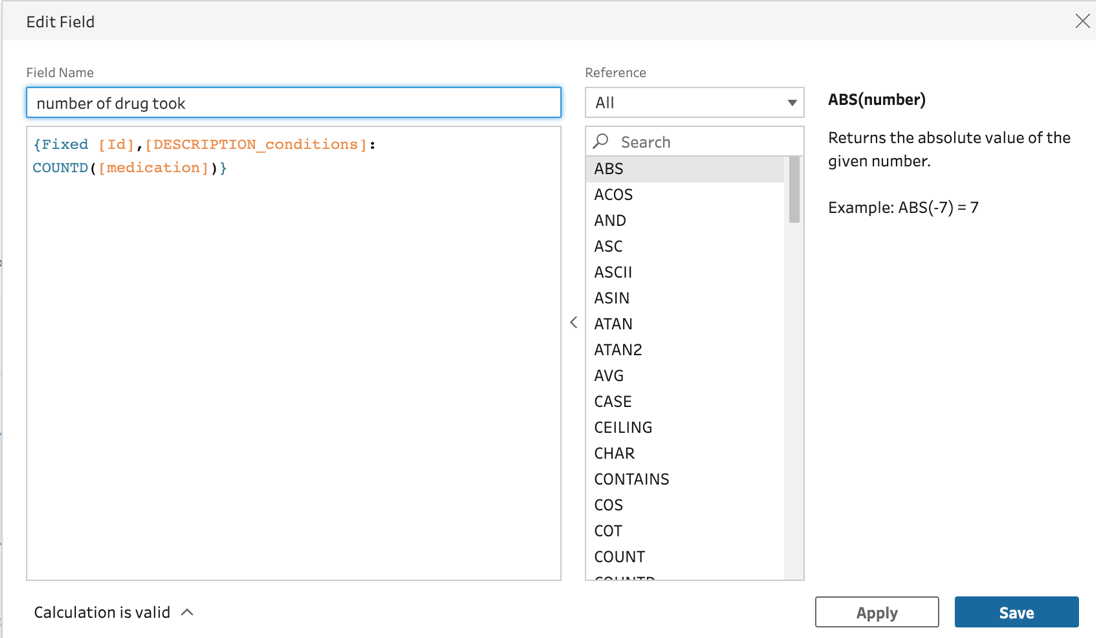


* + - After creating those calculated fields, all the steps will be shown under the Changes, the steps are run in order then need to make sure the calculated field does not use another calculated field created after it.
    - Also need to make sure that the column DEATHDATE is a type of Date, as in the extreme case, all the value in DEATHDATE is null. Move the Change Type step to the beginning of Changes so that the DEATHDATE can be recognized as a Date object
  1. Filter to choose data under specified conditions and medication

Click plus -> Clean Steps: Filter Values…



* 1. Filter to get cohort of patients only took exactly one of the drug
  + Number of drug took



Then filter on the value

Graphical user interface, application

Description automatically generated

* 1. Filter to data ensure each patient corresponds to one record about the survival time and its status to avoid duplicate counts the patient
  + Patient\_row\_number

Graphical user interface, application

Description automatically generated

Then filter on

Graphical user interface, application

Description automatically generated

* 1. Output the preprocessed data

Click plus -> Output and click the output(default name as Output.hyper), In mac, the default location that the output stored in /Users/<username>/Documents/My Tableau Prep Repository/Datasources folder

The final flow shown below (the flow file is named KM\_extracted\_Flow-medication)

A picture containing chart

Description automatically generated

* 1. In Tableau Desktop, it can connect the Output.hyper as data source

Through Data-> New Data Source, Under the To a File choose More… and find the located place to select the Output. hyper

Below is the demonstration for Kaplan Meier double curve with different cancers example

Follow similar steps described above, but for this plot only consider patients.csv and conditions.csv, the data join shown belowGraphical user interface

Description automatically generated with medium confidence

Also not forget to insert the row

* For patients.csv, we add an extra row set Id as ‘Added’ and left other columns empty
* For condition.csv, we add an extra row set Patient as ‘Added’, Description as ‘Malignant neoplasm of breast (disorder)’ which is the analyzed condition, and left column as empty

For Kaplan Meier curve compare the survival probability in different conditions (cancer type stored in description in condition table)

Example: Plotting the KM plot analyzing the survival probability in different cancer type with usage of the data in the two\_conditions.csv under Test Sets folder. And other test data set also preprocess through the flow

Modify the data

The first step is still to insert few rows into corresponding table so that we do not lose the starting point(survival probability is 1 when time at 0) in KM plot

* For patients.csv, we add an extra row set Id as ‘Added’ and left other columns empty



* For condition.csv, we add two extra rows, for first row set Patient as ‘Added’, Description as ‘Malignant neoplasm of breast (disorder)’, the second row set Patient as ‘Added’, Description as ‘Metastasis from malignant tumor of breast (disorder)’and all the left column as empty

