Tutorial: Data Validation and Cleaning Tool

This tutorial overviews how to use the data validation and cleaning tool to interactively clean up excel data files.

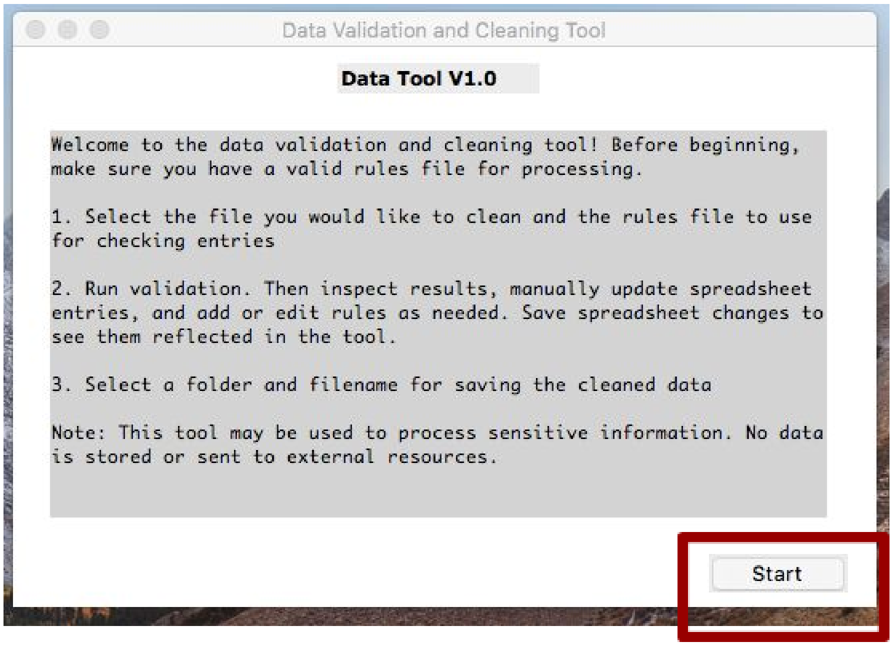
The data validation tool is a GUI that guides you through an automated data cleaning process. The rules file determines what data columns we check for valid entries, and how we flag errors. Errors will either be autocorrected or flagged for manual review.

Topics in this tutorial:

- How to clean datafiles if you already have a rules file set up
- How to update a rules file
- How to set up a new rules file for a new data source or data format
- How to address common errors

Cleaning datafiles using an existing rules file

Launch the tool and click the start button on the welcome screen.



Welcome to the data validation and cleaning tool! Before beginning, make sure you have a valid rules file for processing.

1. Select the file you would like to clean and the rules file to use for checking entries

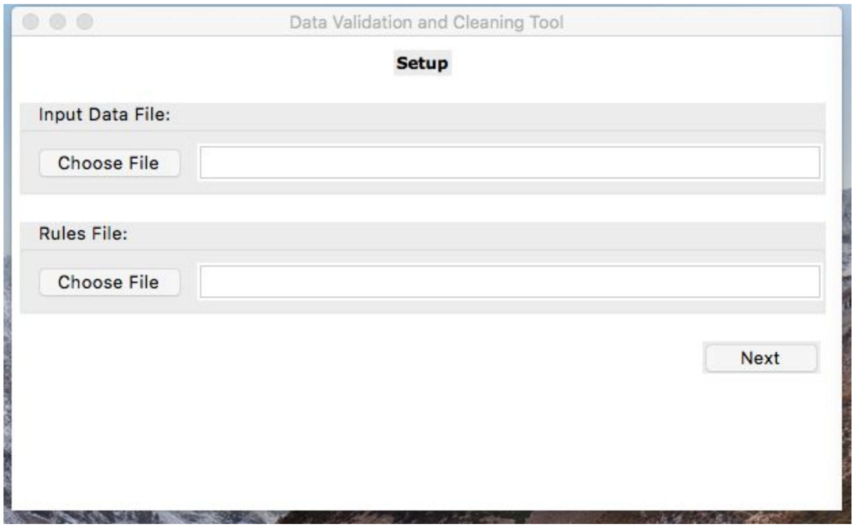
2. Run validation. Then inspect results, manually update spreadsheet entries, and add or edit rules as needed. Save spreadsheet changes to see them reflected in the tool.

3. Select a folder and filename for saving the cleaned data

Note: This tool may be used to process sensitive information. No data is stored or sent to external resources.

Load your data and rules template

The ‘Setup’ screen is where you select the data file you want to clean, and the data cleaning rules file you wish to apply. You can select the data file and rules files using the Choose File buttons. The tool will save the last rules file you used, so if you don’t want to change it, you can leave it as is.



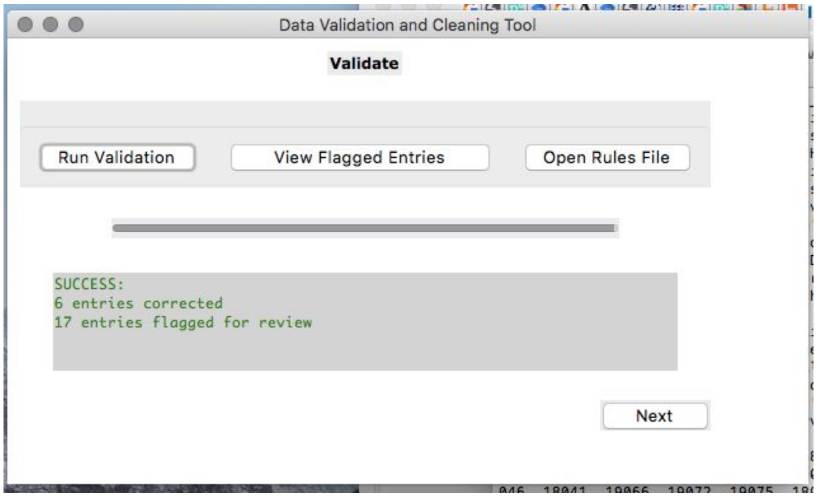
Once you are happy with your file selections, click next.

Run the validation and iteratively clean up the data

The validation screen lets you run the validation, review and update your rules file, and view your flagged entries. The view flagged entries button will not work until you have run the data validation at least once.

To clean the data:

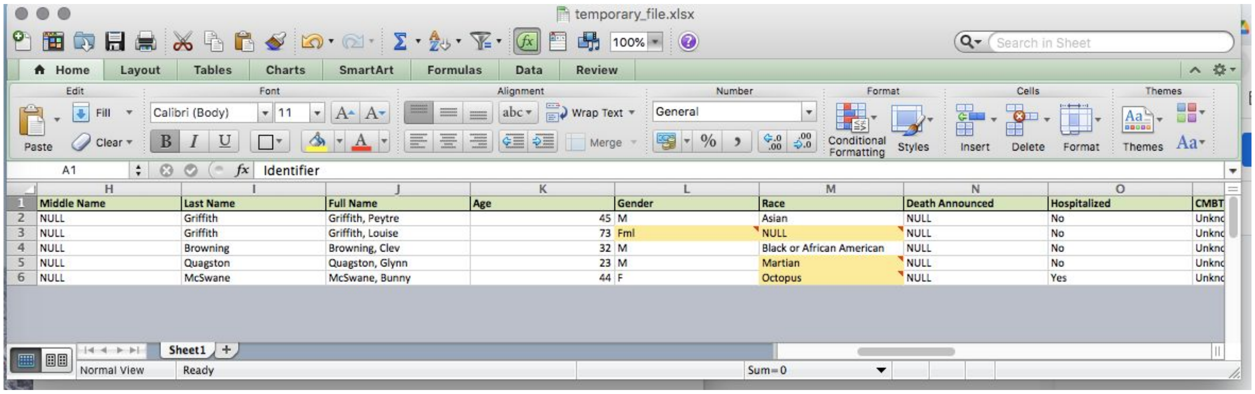
Click the ‘Run Validation’ button.



```
SUCCESS:
6 entries corrected
17 entries flagged for review
```

Once the validation is complete, you should see a “SUCCESS” message in the Validation screen, and the ‘View Flagged Entries’ button should become available if the data needs additional review.

In the example above, 6 entries were automatically corrected by the rules file and 17 entries were flagged for manual review. To manually review the 17 flagged entries, click the ‘View Flagged Entries’ button. This will open your review excel file.



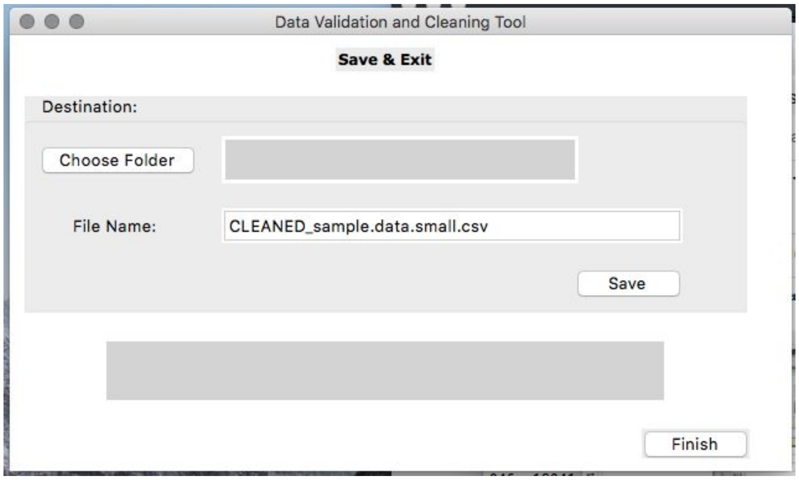
A1 Identifier

	H	I	J	K	L	M	N	O
1	Middle Name	Last Name	Full Name	Age	Gender	Race	Death Announced	Hospitalized	CMBT
2	NULL	Griffith	Griffith, Peytre	45	M	Asian	NULL	No	Unknc
3	NULL	Griffith	Griffith, Louise	73	Fml	NULL	NULL	No	Unknc
4	NULL	Browning	Browning, Clev	32	M	Black or African American	NULL	No	Unknc
5	NULL	Quagston	Quagston, Glynn	23	M	Martian	NULL	No	Unknc
6	NULL	McSwane	McSwane, Bunny	44	F	Octopus	NULL	Yes	Unknc

Manually correct the entries as needed. Save and close the temporary file (just save, don't save as) and re-run the "Run Validation" button to check for any errors you have missed. If there are still "entries flagged for review", repeat the clean up and save. You can repeat this process until you are satisfied with the results. If there are flagged entries you don't wish to correct, that is ok. It will not cause any issues with the final cleaned data.

Note: If an error is recurring, you have the option to automate the correction by adding it to the rules file. If a flagged record is not really an error, you can address this in the rules file as well. (See Update a rules file)

Save the cleaned data and exit

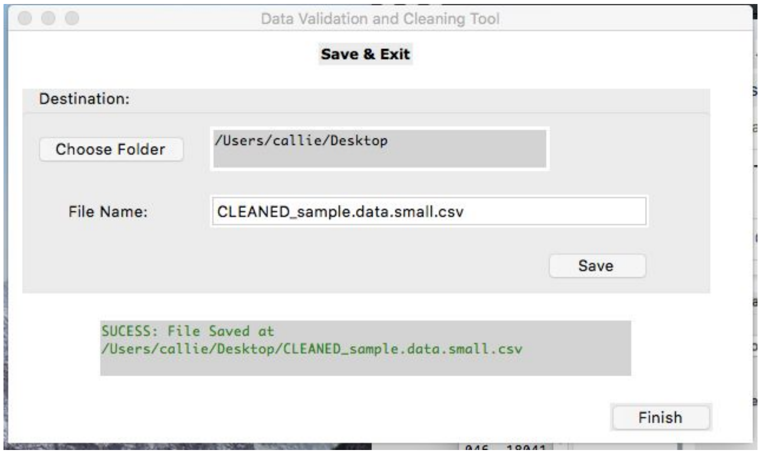


Destination:

File Name: CLEANED_sample.data.small.csv

Choose which folder you wish to save the clean data in. A file name will auto-populate, but you can manually edit it. Follow the naming convention determined by your team. Click 'Save' to save the file.

If you file is successfully saved you should see a success message with the file path. Click the 'Finish' button to exit the program.



Destination:

/Users/callie/Desktop

File Name: CLEANED_sample.data.small.csv

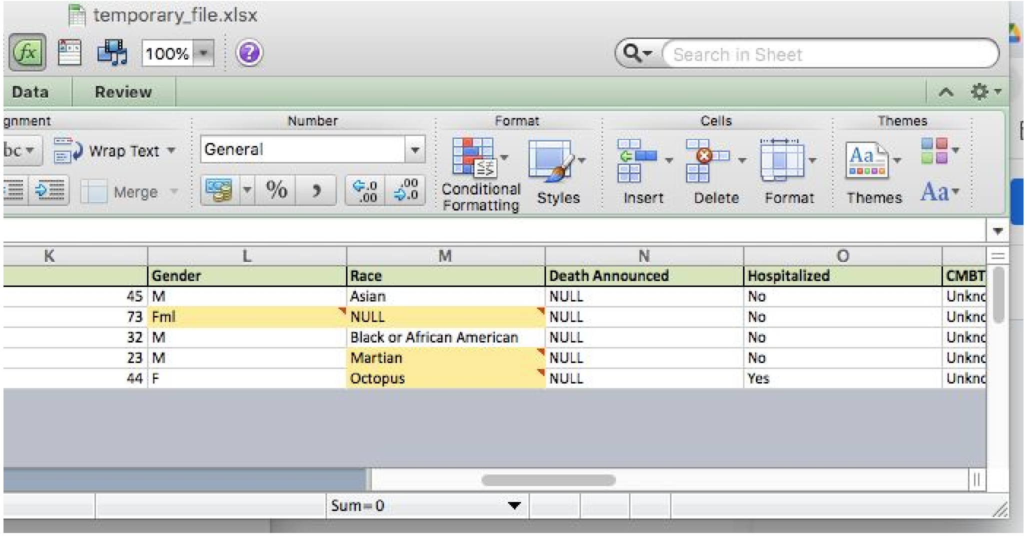
```
SUCESS: File Saved at
/Users/callie/Desktop/CLEANED_sample.data.small.csv
```

Update a rules file

If you want to add a new automatic data field correction, you can do so by editing the rules file. You can do this manually at any time by opening and editing the rules file, or you can do so during the iterative cleaning process by using the tool. In this example, we'll run through how to edit the rules file during the data cleaning process.

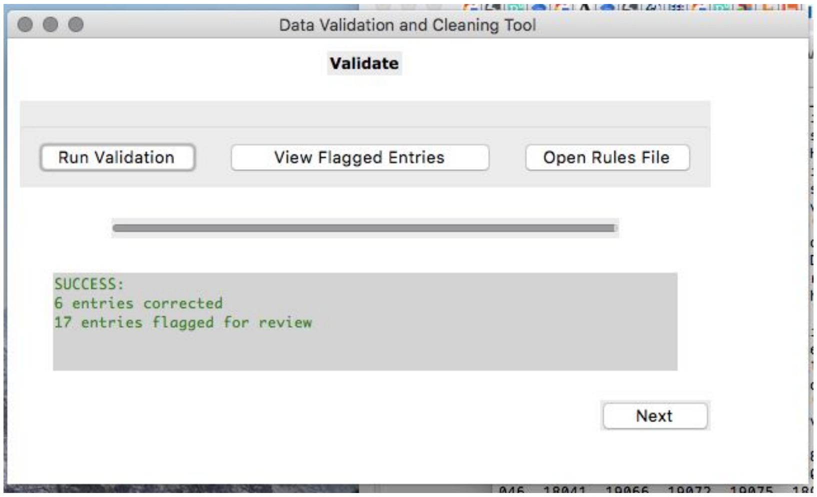
Note: we recommend keeping a copy of the original/standard rules file in a safe place in case you need to revert to a prior version.

In this example, we are going to edit the rules file so that “Fml” is autocorrected to “F” in the Gender column.



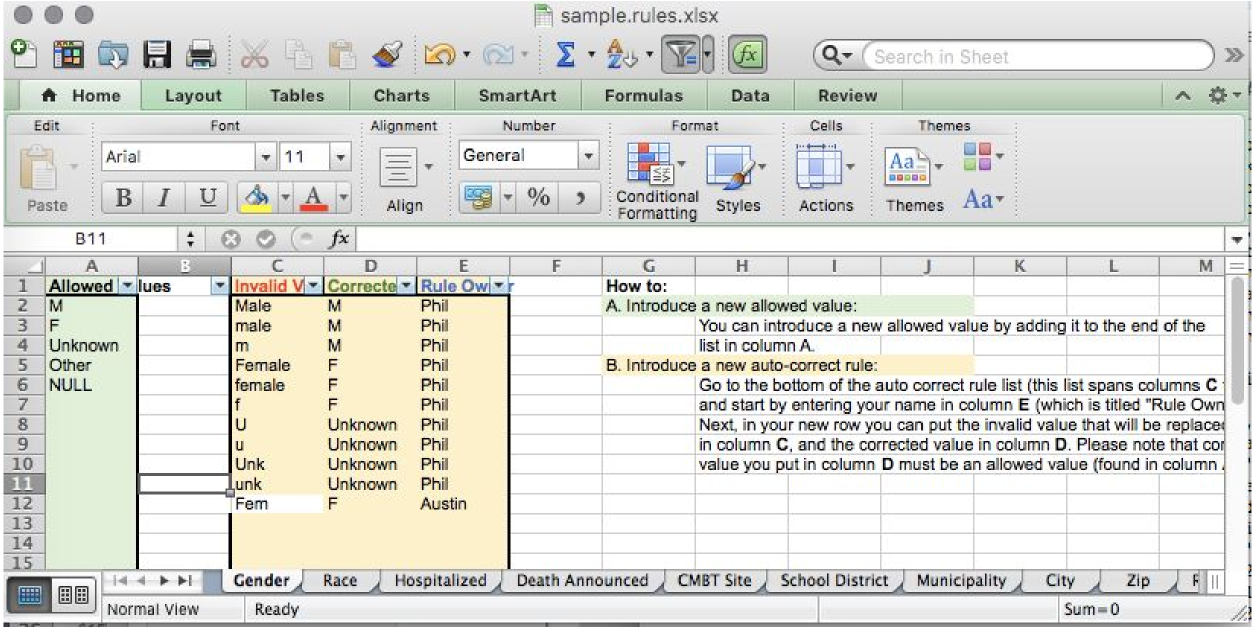
K	L	M	N	O	
	Gender	Race	Death Announced	Hospitalized	CMBT
45	M	Asian	NULL	No	Unknc
73	Fml	NULL	NULL	No	Unknc
32	M	Black or African American	NULL	No	Unknc
23	M	Martian	NULL	No	Unknc
44	F	Octopus	NULL	Yes	Unknc

From the validation screen, choose “Open Rules File”



```
SUCCESS:
6 entries corrected
17 entries flagged for review
```

This will open the rules file in excel. Navigate to the “Gender” column.



	A	B	C	D	E
1	Allowed	lues	Invalid V	Correcte	Rule Ow r
2	M		Male	M	Phil
3	F		male	M	Phil
4	Unknown		m	M	Phil
5	Other		Female	F	Phil
6	NULL		female	F	Phil
7			f	F	Phil
8			U	Unknown	Phil
9			u	Unknown	Phil
10			Unk	Unknown	Phil
11			unk	Unknown	Phil
12			Fem	F	Austin
13					
14					
15					

How to:

A. Introduce a new allowed value:

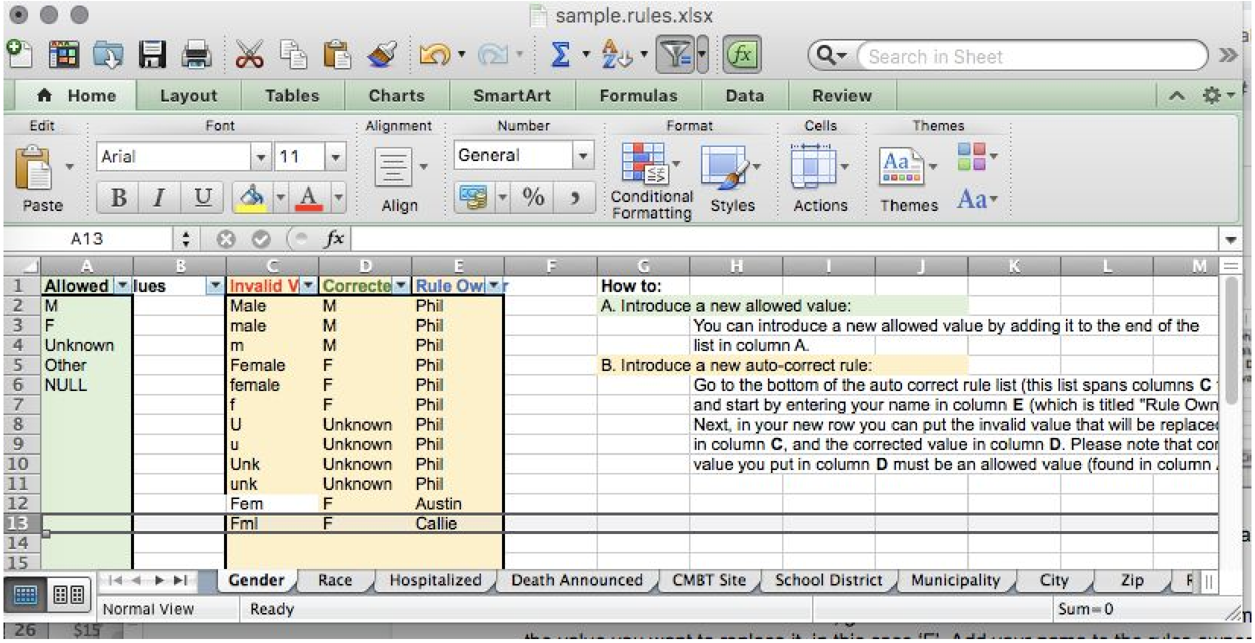
You can introduce a new allowed value by adding it to the end of the list in column A.

B. Introduce a new auto-correct rule:

Go to the bottom of the auto correct rule list (this list spans columns **C** and start by entering your name in column **E** (which is titled "Rule Own Next, in your new row you can put the invalid value that will be replace in column **C**, and the corrected value in column **D**. Please note that co value you put in column **D** must be an allowed value (found in column

Note: In this example, we won’t change the allowed values, but you can also expand Allowed values to include new options that you would like to stop flagging as incorrect.

To add an auto-correction for “Fml”, go to the “Invalid Values” column and add ‘Fml’. Next add the value you want to replace it, in this case ‘F’. Add your name to the rules owner column.



	A	B	C	D	E
1	Allowed	lues	Invalid V	Correcte	Rule Ow r
2	M		Male	M	Phil
3	F		male	M	Phil
4	Unknown		m	M	Phil
5	Other		Female	F	Phil
6	NULL		female	F	Phil
7			f	F	Phil
8			U	Unknown	Phil
9			u	Unknown	Phil
10			Unk	Unknown	Phil
11			unk	Unknown	Phil
12			Fem	F	Austin
13			Fml	F	Callie
14					
15					

How to:

A. Introduce a new allowed value:

You can introduce a new allowed value by adding it to the end of the list in column A.

B. Introduce a new auto-correct rule:

Go to the bottom of the auto correct rule list (this list spans columns **C** and start by entering your name in column **E** (which is titled "Rule Own Next, in your new row you can put the invalid value that will be replace in column **C**, and the corrected value in column **D**. Please note that co value you put in column **D** must be an allowed value (found in column

Save the rules file. Now you can re-run the validation (‘Run Validation’) using the updated rules file, and the new auto-correction will be applied.

Create a new rules file

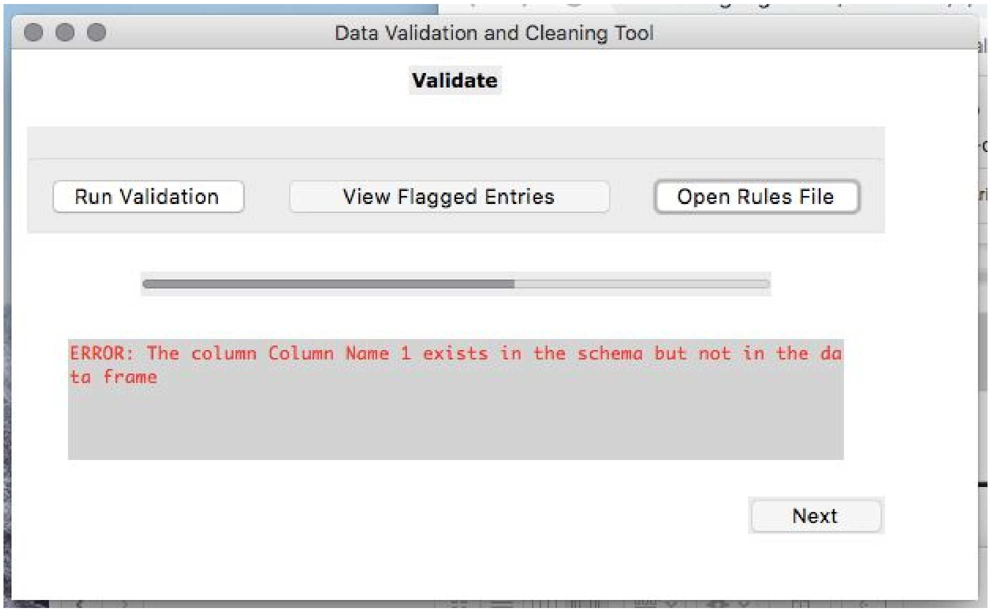
In the top menu bar of the tool, select File > Create new rules file from template. You should see a warning letting you know that you need to save the template spreadsheet in a new location and under a new name for your changes to be stored.

An excel spreadsheet will open with a sample worksheet for data column for ‘Gender’, as well as a blank worksheet. Rename worksheets to match column names of spreadsheets you would like to clean with the tool. Add allowed values and autocorrect mapping rules as needed.

Common errors

Helpful hints for error messages

Error example 1: There is a mismatch between the columns in your rules and th e columns in your data. You need to update the rules file so that any included tab is also in the data you want to clean.



```
ERROR: The column Column Name 1 exists in the schema but not in the da
ta frame
```