

## QAMyData: Table of QA tests included (V1.0)

Type of check available	Specific test	User note
Basic file checks	File opens	Checks whether acceptable format
	Bad filename check, regular expression via RegEx pattern	Regex requires quotes "[a-z]". To use a special characters, e.g. a backslash (\) a backslash before is required e.g. \\
Metadata checks	Report on number of cases and variables	Always run
	Count of grouping variables	
	Missing variable labels	Must be set to true, or the test will not run
	No label for user defined missing values e.g. -9 not labelled	SPSS only
	'Odd' characters in variable names and labels	User specifies the characters
	'Odd' characters in value labels	User specifies the characters
	Maximum length of variable labels, e.g. >79 characters	User specifies the length
	Maximum length of value labels, e.g. >39 characters	User specifies the length
	Spelling mistakes (non-dictionary words) in variable labels using a dictionary file	User specifies a dictionary file

## TABLE OF TESTS

<b>Data integrity checks</b>	Report number of numeric and string variables	
	Check for duplicate IDs	User specifies the variables. Multiple variables can be added on new lines e.g. Caseno or AnotherVariable
	'Odd' characters in string data	User specifies the characters
	Spelling mistakes (non-dictionary words) in string data using a dictionary file	User specifies a dictionary file
	Percentage of values missing ('Sys miss' and undefined missing)	User sets the threshold, e.g. more than 25%
<b>Disclosure risk checks</b>	Identifying disclosure risk from unique values or low thresholds (frequencies of categorical variables or minimum values)	User sets the threshold value, e.g. 5
	Direct identifiers using a RegEx pattern search	User runs separately for postcodes, telephone numbers etc. Advise tests run separately as resource intensive
<b>Useful checks to add</b>	<b>Specific test</b>	<b>User note</b>
<b>Metadata checks</b>	Export a Code book DDI	
<b>Data integrity checks</b>	Expected format for a variable for coded data	User specifies field type/format e.g. ICD code
	Values outlying the listed code values	
<b>Disclosure risk checks</b>	Direct identifiers/named entities in string data using a dictionary file	Specify a dictionary file containing lists of stop words or named entities e.g. for places, names etc. Advise tests are separately as may be resource intensive