**T-MSIS DQ Measures Code**

This project aims to provide resources and tools that help the research community and other stakeholders (the end user community) obtain clear, concise information about the overall data quality, as well as granular detail that can inform specific research projects or information needs in the Transformed Medicaid Statistical Information System (T-MSIS).

**Getting the code**

The easiest way to obtain the code is to clone it with git. If you're not familiar with git, a tool like [Github Desktop](https://desktop.github.com/) or [SourceTree](https://www.sourcetreeapp.com/) can help make the experience easier. The HTTPS link is https://github.com/CMSgov/T-MSIS-Data-Quality-Measures-Generation-Code

If you're familiar with git and just want to work from the command line, you just need to run:

git clone https://github.com/CMSgov/T-MSIS-Data-Quality-Measures-Generation-Code.git

If you would prefer not to use git, you can also download the most recent code as a [ZIP file](https://github.com/CMSgov/T-MSIS-Data-Quality-Measures-Generation-Code/archive/master.zip).

**Running the code**

The DQ measures code is written in Databricks SQL executed through explicit SQL passthrough embedded within a SAS code wrapper on T-MSIS data in DataConnect (formerly known as AREMAC).

The code is executed on new state month submissions as soon as possible after data are available in DataConnect, or on historical data on an as-needed basis.

**More technical documentation**

Additional technical documentation can be found in this repository in the file [Technical\_Documentation.md](https://github.com/CMSgov/T-MSIS-Data-Quality-Measures-Generation-Code/blob/master/technical_documentation.md).

The specifications for a subset of the measures included in the code is available in the file [DQMeasure\_Specifications.md](https://github.com/CMSgov/T-MSIS-Data-Quality-Measures-Generation-Code/blob/master/dqmeasure_specifications.md).

**Contributing**

We would be happy to receive suggestions on how to fix bugs or make improvements, though we will not support changes made through this repository. Instead, please send your suggestions to [MACBISData@cms.hhs.gov](mailto:MACBISData@cms.hhs.gov).

**Public domain**

This project is in the worldwide [public domain](https://github.com/CMSgov/T-MSIS-Data-Quality-Measures-Generation-Code/blob/master/LICENSE.md).

This project is in the public domain within the United States, and copyright and related rights in the work worldwide are waived through the [CC0 1.0 Universal public domain dedication](https://creativecommons.org/publicdomain/zero/1.0/).

**Background on data quality measures**

A measure is a calculated statistic. There are several types of data quality measures:

* Percentage
* Count
* Sum
* Ratio
* Average number of occurrences
* Average
* Ratio of averages
* Index of dissimilarity
* Frequency

There are several types of data quality validations users can apply to the measures. Two types are most common:

1. General inferential validations check that a monthly statistic generated for a measure falls within a range of expected values for one point in time.
2. Longitudinal inferential validations compare the most recent month’s statistic for a measure against the average of the prior six months’ statistics for the same measure to see whether the most recent statistic deviates more than expected.

Both general inferential validations and longitudinal inferential validations rely on data quality standards, which are either tolerance thresholds or minimum-maximum ranges, depending upon the type of validation for the measure. The ranges and tolerances are based on historical and expected data patterns. Currently, CMS does not have state-specific thresholds, so the thresholds are the same for all states, regardless of the states’ programs or population size. In addition, some measures do not have any data quality standards because a threshold has not yet been established. Measure specific data quality standards are available in the [Thresholds\_File.md](https://github.com/CMSgov/T-MSIS-Data-Quality-Measures-Generation-Code/blob/master/thresholds_file.md).

Note that a data quality measure statistic falling outside of the data quality standard is not the same as an error. Errors indicate a business rule violation (i.e. something is wrong with the data) and are based on the validation rules in the T-MSIS Data Dictionary. The data quality standards are based on expected patterns in the data, and a statistic outside of the expected range may or may not reflect actual errors in the data.

Sometimes, the data are obviously wrong. But often, a measure statistic outside of the expected range indicates only a possible data problem. For example, if the data showed 100 percent of Medicaid-eligible beneficiaries died in a month, the data would be wrong. However, if the percentage of Medicaid eligible beneficiaries who died in a month was 5 percent, this would indicate a *possible* data problem because the measure statistic of 5 percent is outside the maximum threshold (i.e. the expected maximum of the percentage of Medicaid eligible beneficiaries who would die in one month) of 4 percent. It is not certain there is an issue with the data in that month; it would be important to consider both the value of the statistic and the expected range or longitudinal tolerance. In some cases, it may also be important to consider policy or operational factors in a state’s Medicaid or CHIP program.

There are two other less common types of validation users can apply to the measures:

1. Authoritative source validation compares reported values with the values that should have been reported based on state program characteristics—for example, the managed care operating authorities that a state has.
2. Index of dissimilarityvalidation checks that the frequency distribution of a data element does not change significantly compared with the prior month.