



# Datafit Toolkit

## User Guide

A Practical Approach for  
Population Data  
Quality Assessment

January 2020

### Authors

Mahmoud Azimaee

Gangamma Kalappa

Sean Ji

Milton Hu

Cheng Qian

## Publication Information

© 2020 ICES. All rights reserved.

This publication may be reproduced in whole or in part for non-commercial purposes only and on the condition that the original content of the publication or portion of the publication not be altered in any way without the express written permission of ICES. To seek this permission, please contact [communications@ices.on.ca](mailto:communications@ices.on.ca).

### ICES

G1 06, 2075 Bayview Avenue  
Toronto, ON M4N 3M5  
Telephone: 416-480-4055  
Email: [DQIMDept@ices.on.ca](mailto:DQIMDept@ices.on.ca)

### How to cite this publication

Azimaee M. Kalappa G, Ji S, Hu M. Qian C. *Datafit Toolkit User Guide: A Practical Approach for Population Data Quality Assessment*. Toronto, ON: ICES; 2019.

ISBN 978-1-926850-89-4 (online)

ICES (formerly the Institute for Clinical Evaluative Sciences) is an independent, not-for-profit organization that produces knowledge to enhance the effectiveness of health care for Ontarians. Internationally recognized for its innovative use of population-based health data and information, ICES evidence supports health policy development and guides changes to the organization and delivery of health care services in Ontario.

ICES is funded in part by an annual grant from the Ontario Ministry of Health. The opinions, results and conclusions included in this report are those of the authors. No endorsement by the Ministry of Health is intended or should be inferred.

# Contents

<b>Background</b>	3
<b>1.0 ICES Data Quality Framework</b>	4
1.1 Accuracy	5
1.2 Internal validity	6
1.3 External validity	8
1.4 Timeliness	10
1.5 Interpretability	11
1.6 Relevance	11
<b>2.0 DataFit Toolkit Package</b>	13
2.1 Technical requirements	13
2.2 Parameters and options in the main macros	15
2.2.1. Meta macro	15
2.2.2. Dictionary macro	16
2.2.3. VIMO macro	17
2.2.4. TIM macro	20
2.2.5. Trend macro	22
2.2.6. Linkability macro	24
2.2.7. Agreement macro	26
<b>3.0 Tutorial</b>	30
<b>References</b>	33

## Background

In 2010, a data quality framework was initiated and developed by the Manitoba Centre for Health Policy (MCHP), a research unit in the Faculty of Medicine at the University of Manitoba.<sup>1,2</sup> The framework measured and evaluated five dimensions of database-specific quality of administrative data, including:

- Accuracy
- Internal validity
- External validity
- Timeliness
- Interpretability

Each of these dimensions has its own components, such as completeness, correctness, internal consistency, linkability and stability across time. Based on the framework, a package of 18 SAS® macros was designed to generate standard data quality reports for MCHP's various data holdings.

Beginning in 2012, the MCHP data quality framework and tools were further developed and enhanced at ICES (in collaboration with MCHP). ICES' data quality reports were formatted in HTML instead of Excel, and the tools were modified to handle Ontario's ten-fold larger population. Also, ICES added a sixth dimension, relevance, to MCHP's original framework.

In 2013, MCHP licensed the tools under a GNU General Public License. Since then, the two versions of the tools have evolved in slightly different directions to fit the unique needs of MCHP and ICES. In January 2020, ICES licensed its version of the tools under a GNU General Public License, naming it the DataFit Toolkit.

## 1.0 ICES Data Quality Framework

Health administrative data are routinely collected for various administrative and billing purposes, such as health system management and provider remuneration. Data collected for these purposes may not always be of the best quality for research. Because the use of inaccurate or incomplete data can impede the research process and lead to false conclusions, evaluation of data quality is a crucial and integral step before conducting research.

The ICES data quality framework differentiates between *database-specific* quality and *research-specific* quality (Figure 1). This is to acknowledge that compliance with specific privacy and legal regulations is required to link two or more data sources for research purposes. In addition, the engagement of a subject matter expert (SME) is critical to bring knowledge of the data contents to the data quality exercise.

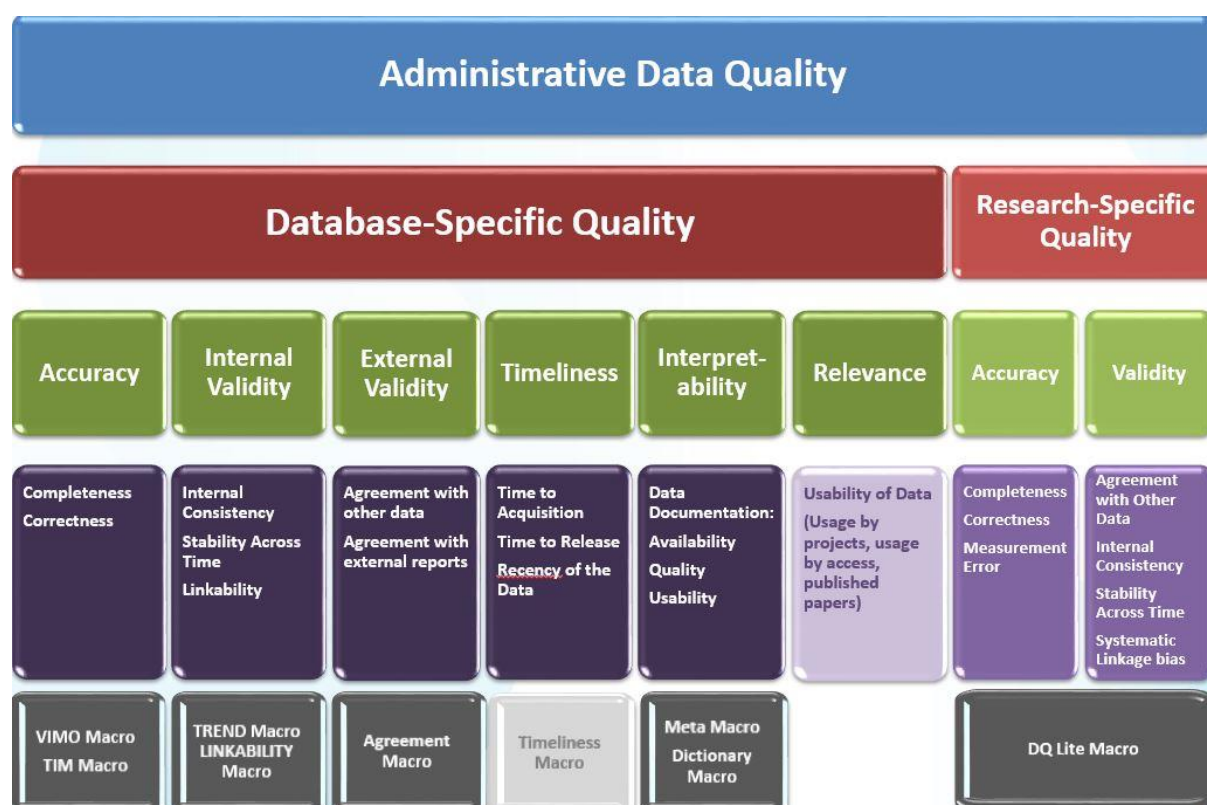


Figure 1. ICES data quality framework

Each new data holding or each update or refresh of an existing data holding is assessed for all the dimensions of the database-specific quality, and all the data quality reports are generated for this data. Depending on the sources of the data quality issues identified during this exercise, the following actions will be taken by ICES' data quality team:

- If the issue appears to have been caused by the data provider during the data extraction, a new data extract will be requested from the data provider.

- If the issue appears to have been caused during ICES' data anonymization, record linkage or standardization processes, the internal process will be repeated to address the issue.
- If the issue is related to the nature of the data, it will be highlighted in ICES' internal data documentation.

The final data quality reports will be made available internally to all ICES data users, including scientists, methodologists, research analysts and epidemiologists.

## 1.1 Accuracy

Accuracy refers to the degree to which data correctly describe the phenomenon they are designed to measure.<sup>3</sup> The accuracy dimension includes completeness and correctness, defined as follows<sup>5</sup>:

- Completeness is measured by the rate of missing values.
- Correctness is measured by the percentage of valid values (values within the domain of possible or plausible values) and the percentage of outliers (values that are potentially invalid because they violate physical, logical or metadata-based constraints). An assessment of the validity of data values requires documentation about plausible values as well as knowledge gained through exploratory analyses of the data.<sup>4</sup>

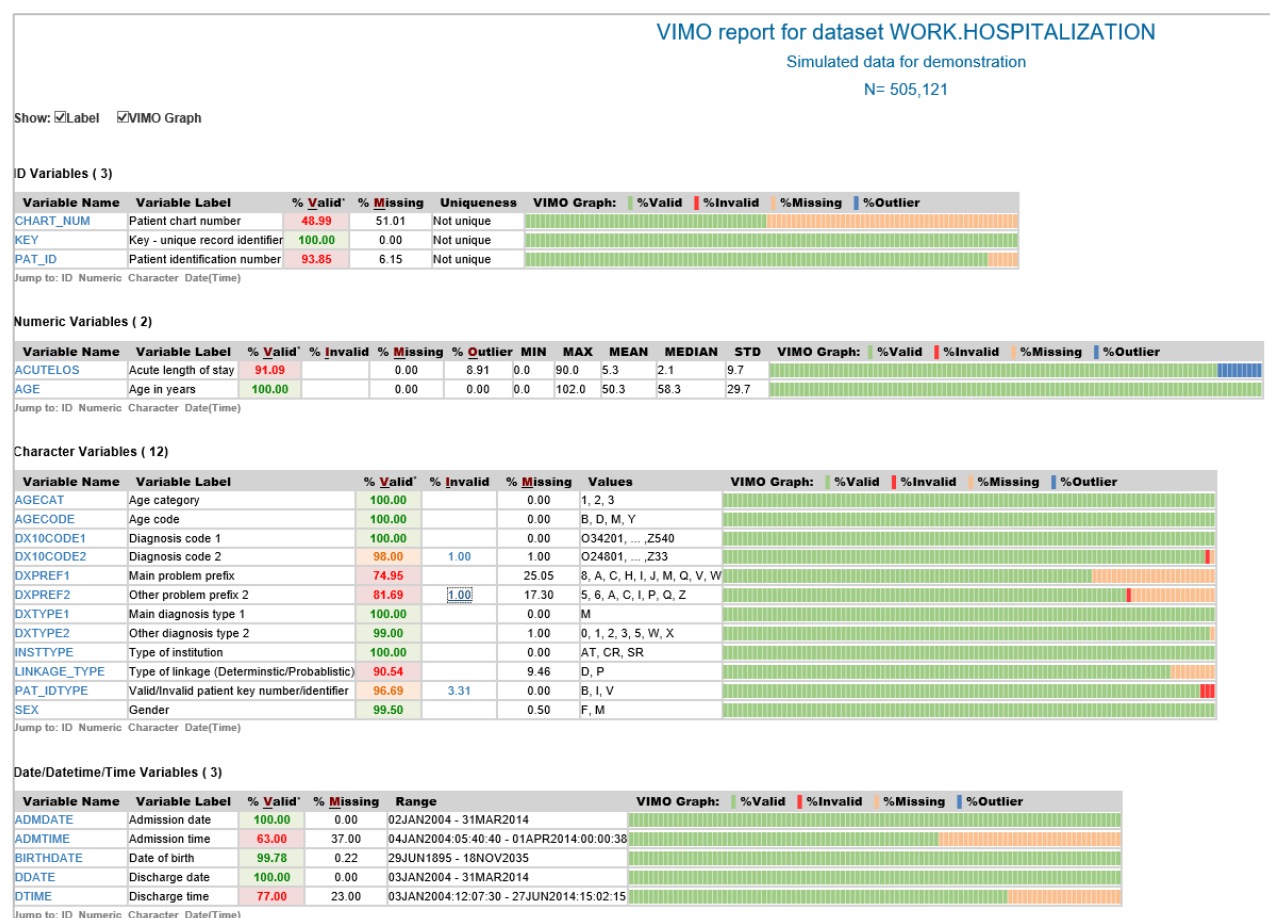


Figure 2. A sample VIMO report (see also [in HTML](#))

The ICES data quality framework uses two standard reports to summarize the accuracy of data: VIMO and TIM. (This documentation has been borrowed from a similar technique called VODIM, introduced by the United Kingdom's National Health Service.<sup>6</sup> VODIM stands for Valid, Other, Default, Invalid and Missing.) VIMO (Valid, Invalid, Missing and Outlier) reports a percentage for each of these four components as well as descriptive statistics such as the mean, median, standard deviation and frequency of values for each data element in a given dataset (Figure 2).

While VIMO provides an overall missing rate for each data element, it does not tell us how the missing rate changes over time. TIM (Trends in Missingness) provides this information over time (Figure 3).

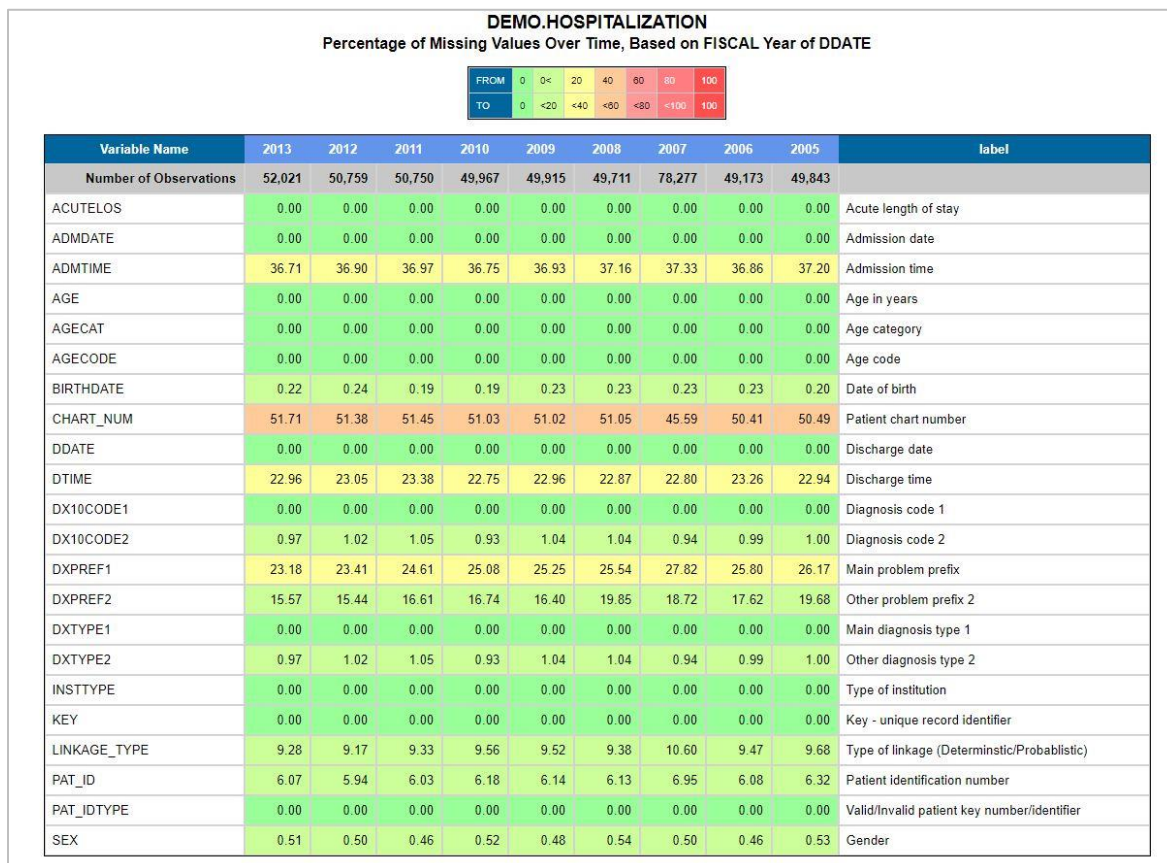


Figure 3. A sample TIM report (see also [in HTML](#))

## 1.2 Internal validity

Temporal consistency is measured by the degree to which a set of time-related observations conforms to a smooth line or curve over time and the percentage of observations that are classified as outliers from that line or curve.<sup>4</sup> Stability over time is assessed using trend analysis, which involves fitting different types of lines or curves to a set of data and applying graphic or inferential techniques to compare observed values with expected values. The ICES data quality framework uses trend reporting to assess the temporal consistency of data over time (Figure 4).

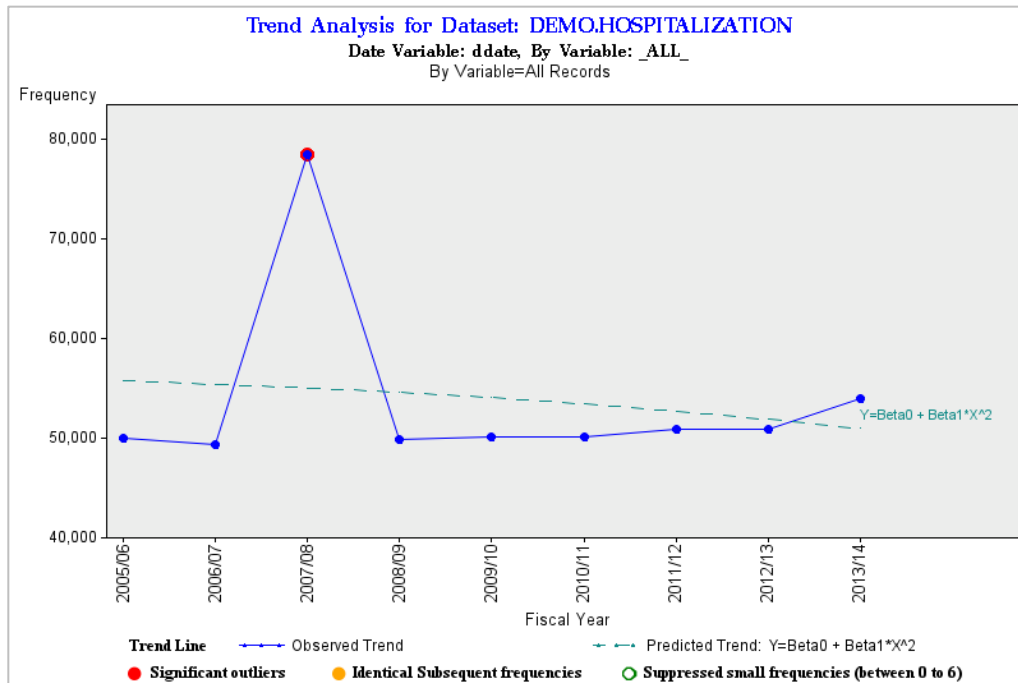


Figure 4. A sample trend report by fiscal year

Linkability is another report provided by this framework to assess a different aspect of Internal validity in a given dataset. Linkability measures the ability to connect one data file to another data file using a unique subject-specific identifier.<sup>5</sup> Linkability is an important data quality indicator because it determines the extent to which different databases can be used in research-specific analyse<sup>4</sup> (Figure 5).



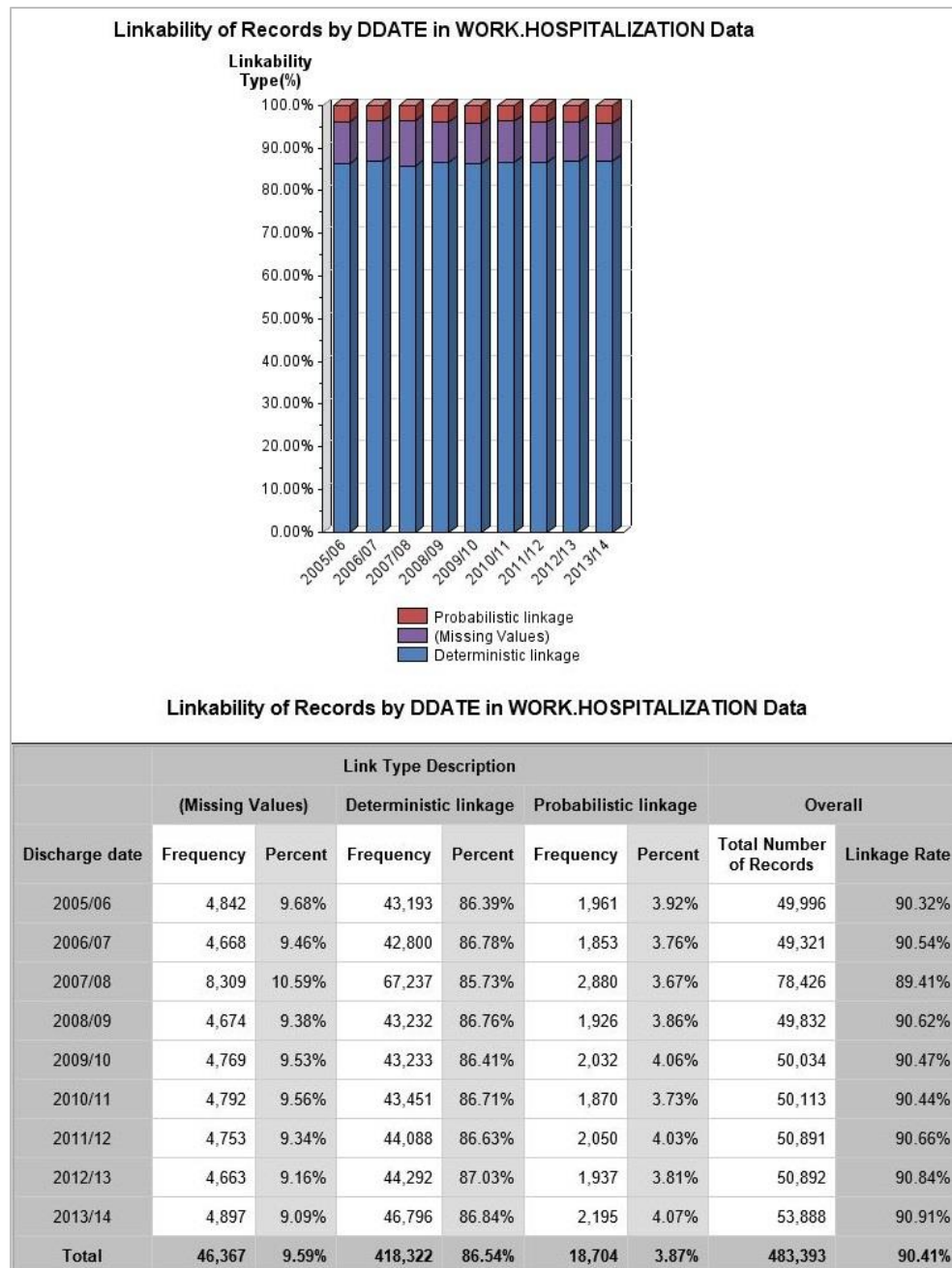


Figure 5. A sample linkability report by fiscal year (see also [in HTML](#))

### 1.3 External validity

“External validity of data can sometimes be quantified by comparison with a “gold standard,” that is, an external data source that contains error-free information about the measure or construct under investigation. Literature, reports and general knowledge of the data can also be used to assess external validity.”<sup>1</sup>

Often it is difficult or impossible to provide a standard method for assessing external validity, as doing so directly depends on the contents of the data. While the ICES data quality framework leaves this to the data quality team to come up with the best approach for external validity in each case, it does provide one standard approach to compare the demography of individuals with a gold standard. Assuming a registry database is available to the data quality team, the agreement report is a standard report for comparing sex and date of birth in a given input dataset with reference data (Figure 6).

**Agreement report**  
Comparison of DEMO.HOSPITALIZATION and DEMO.REFERENCEDATA

Agreement level - summary:

Linkage Type	Agreement Variable	Number of Subjects in the Stratum	Percent Agreement (%)					Kappa Statistics $\kappa$ (95% CI)
			Missing	No Match	Poor Match*	Good Match**	Perfect Match	
Deterministic	HOSPITALIZATION_BIRTHDATE	473,922	0.22	0.43	0.12	0.10	99.12	
	HOSPITALIZATION_SEX	473,922	0.50	1.12	0	0	98.38	0.9771 (0.9765, 0.9777)

Poor Match\* : Agreement on only one date component(m,d,y)  
Good Match\*\* : Agreement on two out of three date components(m,d,y)  
As ds\_link\_type variable was null, all calculations were done assuming linkage type was Deterministic

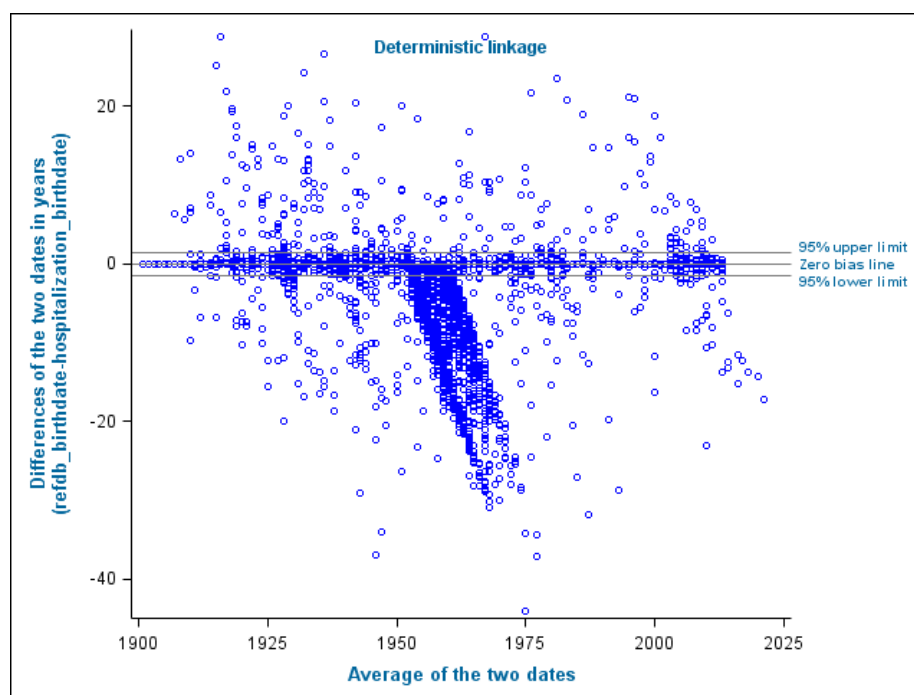


Figure 6. A sample agreement report (see also [in HTML](#))

In the agreement report, three metrics for two variables (sex and date of birth) in the input dataset are compared with corresponding variables in the registry database (i.e., the gold standard). The three metrics are:

- **Percent agreement**
  - It is easily calculated and directly interpretable; however, it does not take into account the role of chance in the agreement of values.
  - A modified percent agreement has been implemented for date variables.

- The modified implementation uses the three components of the date (year/month/day) and determines the agreement for each component, calculates a weight based on the number of component agreement and presents the percent agreement according to the calculated weight.
  - This modified method for calculating? Identifying? characterizing? date variables can differentiate between good agreement (two of three components agree), poor agreement (only one component agrees) and no agreement.
- **Cohen's kappa coefficient**
    - The kappa coefficient ranges from -1 to +1, with 0 representing the agreement that can be expected from random chance and 1 representing perfect agreement.
    - The strength of the kappa coefficient is that it takes into account the agreement that happens by pure chance.
    - It applies only to nominal variables (sex), and categories are limited to male and female.
  - **Bland-Altman plot<sup>9</sup>**
    - It is a method of quantification of the agreement between two quantitative measurements by studying the mean difference and constructing limits of agreement.
    - It is calculated by plotting the difference of the two measurements (birthdate A – birthdate B) [the Y-axis of the plot] to the mean of the two measurements  $((\text{birthdate A} + \text{birthdate B})/2)$  [the X-axis of the plot].
    - To apply this method, the agreement report considers date variables as continuous and displays a scatterplot with 95% confidence interval.
    - The Bland-Altman plot provides the following information:
      - Whether the date variable in the dataset of interest is skewed in one direction.
      - Whether the confidence interval is wide (e.g., too many date variables do not match, making it relatively unreliable).

## 1.4 Timeliness

While the ICES data quality framework does not provide a standard report or tool to measure timeliness of the data, it emphasizes the importance of this dimension. Timeliness reporting should provide information on how current the data are in a dataset.<sup>4</sup> This is indicated by:

- Time until a dataset is acquired.
- Time until the data are released to the organization.
- Time until updates to the data are in place.

Timeliness will be included in ICES' to-be-developed report and tools for data quality assessment.

## 1.5 Interpretability

Interpretability focuses on the data documentation, including historical and current documentation.<sup>3</sup> The concept of interpretability focuses on including documents that are maintained over time and developing documents during the process of data quality assessment.<sup>3</sup>

Changes in program, inclusion criteria, data collection methods and reporting criteria are all critical information for data analysts and researchers to identify potential data quality issues.<sup>4</sup>

Rich metadata are essential for creating good and interpretable data documentation. Metadata is defined as data about data. The ICES data quality framework provides an approach to generating a standard metadata repository that will become essential for other data quality dimensions and a very rich source of information with which to generate a data dictionary (Figure 7).

Data Dictionary			
Dataset Name: hospitalization Library Name: DEMO			
Variable Name	Label	Type/Length	Format
<a href="#">ACUTELOS</a>	Acute length of stay	Num8	
<a href="#">ADMDATE</a>	Admission date	Num8	DATE9.
<a href="#">ADMTIME</a>	Admission time	Num8	DATETIME15.
<a href="#">AGE</a>	Age in years	Num8	
<a href="#">AGECAT</a>	Age category	Char1	\$HOSP_AGE CAT.
<a href="#">AGECODE</a>	Age code	Char1	\$HOSP_AGE CODE.
<a href="#">BIRTHDATE</a>	Date of birth	Num8	DATE9.
<a href="#">CHART_NUM</a>	Patient chart number	Char20	
<a href="#">DDATE</a>	Discharge date	Num8	DATE9.
<a href="#">DTIME</a>	Discharge time	Num8	DATETIME15.
<a href="#">DX10CODE1</a>	Diagnosis code 1	Char7	\$HOSP_DX10 CODE.
<a href="#">DX10CODE2</a>	Diagnosis code 2	Char7	\$HOSP_DX10 CODE.
<a href="#">DXPREF1</a>	Main problem prefix	Char1	\$HOSP_DXPREF.
<a href="#">DXPREF2</a>	Other problem prefix 2	Char1	\$HOSP_DXPREF.
<a href="#">DXTYPE1</a>	Main diagnosis type 1	Char1	\$HOSP_DXTYPE.
<a href="#">DXTYPE2</a>	Other diagnosis type 2	Char1	\$HOSP_DXTYPE.
<a href="#">INSTTYPE</a>	Type of institution	Char2	\$HOSP_INSTTYPE.
<a href="#">KEY</a>	Key - unique record identifier	Char20	
<a href="#">LINKAGE_TYPE</a>	Type of linkage (Deterministic/Probabilistic)	Char1	\$HOSP_LINK_TYPE.
<a href="#">PAT_ID</a>	Patient identification number	Char12	
<a href="#">PAT_IDTYPE</a>	Valid/Invalid patient key number/identifier	Char1	\$HOSP_VALPATID.
<a href="#">SEX</a>	Gender	Char1	\$HOSP_SEX.

Variable DXPREF1	
Label:	Main problem prefix
Type/Length:	Char1
Available in:	HOSPITALIZATION
Format:	\$HOSP_DXPREF.
Values:	<a href="#">Click here for values and descriptions.</a>

Figure 7. A sample data dictionary (see also [in HTML](#))

## 1.6 Relevance

Relevance, a recent addition to the ICES data quality framework, helps data users and managers understand usability and the degree to which the data meet the current needs of the organization.

At ICES, the relevance report is generated by merging various information such as the date the data were approved for use by the privacy and legal department, analytics on user access to and usage of each data source, and the date the research findings were published (Figure 8).

This report, which is updated on an annual basis, generates granular information on how the data are touched, approved and used.



Figure 8. A sample relevance report

## 2.0 DataFit Toolkit Package

The DataFit Toolkit package currently includes 33 SAS® macros implemented for the database-specific quality assessment; seven of them are main macros that users run to generate standard data quality reports. The other 26 macros, termed intermediate macros, are invoked by the three main macros (Figure 9). (The main macros are displayed at the bottom of the data quality framework presented in Figure 1; note that the timeliness macro is not currently available in the package.)

### 2.1 Technical requirements

- Macros in the DataFit Toolkit have been tested on SAS® version 9.4 and will work in both Windows and Linux environments. The macros can assess the quality of a data repository (or a dataset) in SAS BASE format (sas7bdat file extension).
- The DataFit Toolkit can generate quality assessment reports for both cumulative data (historical observations stored in a single dataset) and annual data (a dataset for each fiscal or calendar year of observation). However, certain naming conventions need to be followed for annual datasets, especially for creating TIM and TREND reports. Annual datasets must always end with a four-digit year representing the year of the dataset (e.g. <datasetname><yyyy>).
- A standard SAS format catalog corresponding to variables in the base SAS datasets is necessary for complete and comprehensive data quality reports.
- A SAS metadata library is essential for creating any data quality assessment report. The %meta macro in this package is provided for this purpose.
- The DataFit Toolkit assumes that the naming conventions used in each SAS library are consistently applied and are specific to that library. If a SAS library includes multiple datasets, variable X represents the same data element in every dataset in the library. That is, variable X should have the same properties (length, type, description, format, etc.) across different datasets in the library. If this assumption is not met, inaccurate results, or in some cases run-time errors, can occur in some of the macros.
- Users working in a moated environment with no internet access must download Bootstrap and JQuery and make them available to the DataFit Toolkit. The location of these scripts must be defined properly in %VIMO's parameters. The required scripts and their URLs are as follows:
  - Bootstrap
    - URL: <https://getbootstrap.com/>
    - Required script: "bootstrap.min"
  - JQuery
    - URL: <https://code.jquery.com/>
    - Required scripts: "jquery.popupoverlay" and "jquery-1.8.2.min"



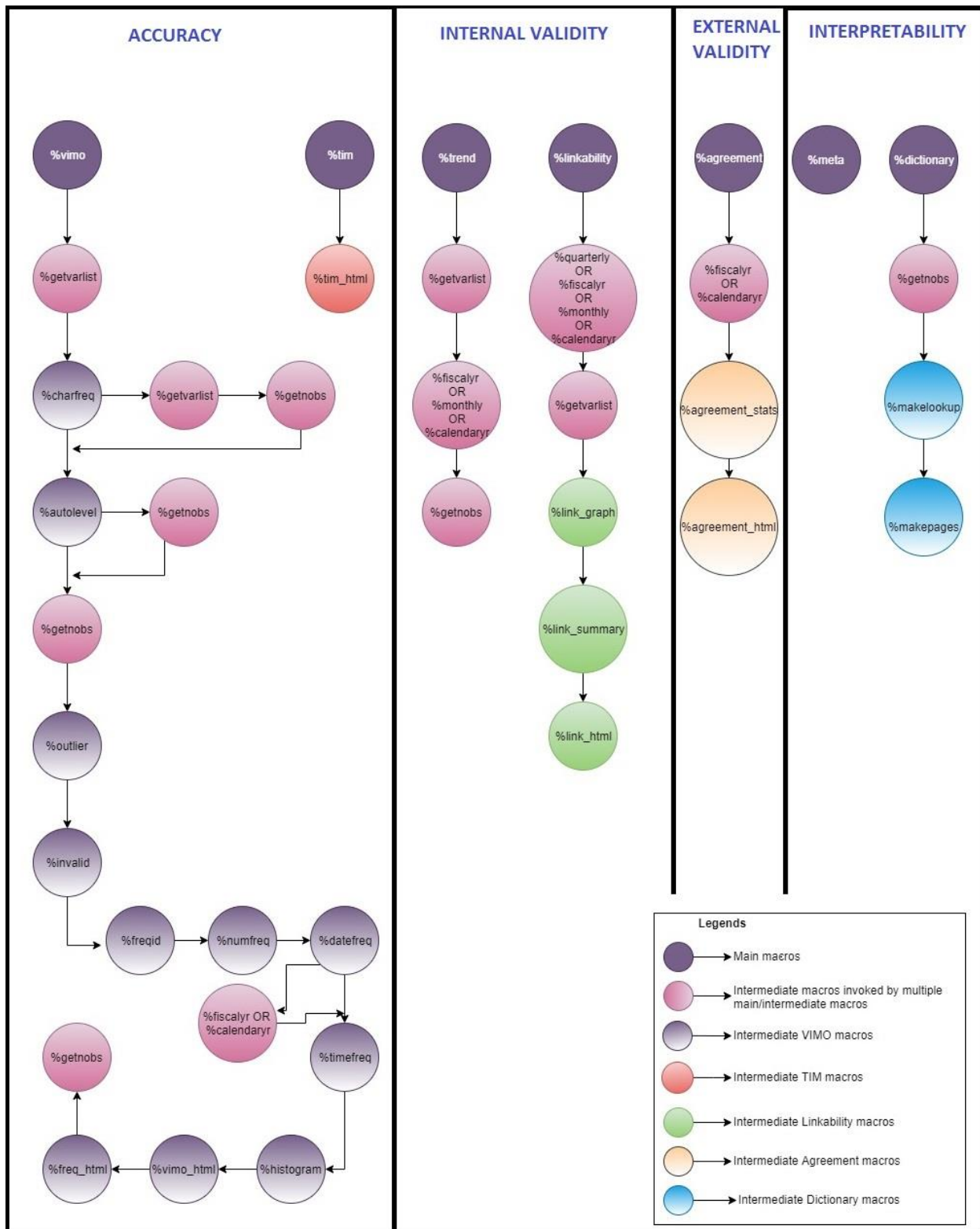


Figure 9. Interdependencies of macros (main and intermediate) included in the toolkit

## 2.2 Parameters and options in the main macros

### 2.2.1. Meta macro

The Meta macro generates a metadata dataset for a single dataset or a series of datasets in a specified library. It generates a reference table that includes names, labels, formats, length and other attributes of the variables in each dataset. Additional information such as the physical location of data on the disk, the access control group associated with each data, the data owner and permission settings are also captured in the table generated by the Meta macro.

According to best practices in SAS programming, embedding SAS formats with variables in a dataset is not recommended. However, to generate comprehensive metadata for the purposes of this data quality framework, it is essential to associate variables with their corresponding SAS formats. The framework recommends maintaining a crosswalk between the variable list and its associated SAS formats. The Meta macro expects a simple space/tab-separated text file called Varlist.

Parameter	Mandatory	Default value	Parameter Description
<b>lib</b>	Yes		SAS library name that you want to create metadata for.
<b>Ds</b>			Dataset prefix or complete name of dataset in SAS library for which the metadata are created.  Note: If it is left blank, then metadata will be generated for the entire SAS library specified in the LIB parameter.
<b>Excl</b>			A SAS dataset name to be excluded from the metadata
<b>fmlib</b>		formats	SAS library name(s) containing the format catalogs
<b>path</b>	Yes		Location for a (VARLIST) text file containing variable names and their associated formats.
<b>Outlib</b>		meta	SAS library name for output dataset.  Note: A SAS dataset called meta_<data> is created.
<b>Log</b>		OFF	Turn the listing of the entire log off or on during execution of the macro.  Valid values: ON/OFF.

**Example 1.** Creates a metadata called “meta\_hospitalization” in the WORK library for the simulated administrative dataset called “hospitalization” that is present in the DEMO library.

```
%meta (lib=demo,  
       ds=hospitalization,  
       fmlib=formats,  
       outlib=work,  
       path='/metadata/varlists/hospitalization_varlist.txt');
```

**Example 2.** Creates a metadata for all datasets in the “demo” library except for the “ref” dataset.

```
%meta (lib=demo,  
       excl=ref,  
       fmlib=formats,  
       path='/metadata/varlists/hospitalization_varlist.txt');
```



## 2.2.2 Dictionary macro

The Dictionary macro generates HTML pages to display the data dictionary for a single dataset or for all the datasets in a given SAS library. This macro uses the following three sources of information to create a complete data dictionary.

- METADATA is basically appended to all of the metadata datasets that were created by multiple runs of %META (as described earlier) through a simple data step:

```
data meta.metadata;  
  set meta_ ;  
run;
```

- A SAS dataset version of the entire format catalog that stores information about informats or formats:

```
proc format library=formats cntlout=meta.formats;  
run;
```

- The NOTES dataset includes all additional notes, comments, warnings and hyperlinks for data elements (variables). This component is optional and could be created manually by entry into an Excel spreadsheet or through a data entry interface and imported to SAS. The NOTES SAS dataset must include the following variables:
  - Libname – a key variable containing the library name; it can be linked to the metadata dataset
  - Name – a key variable containing the variable name; it can be linked to the metadata dataset
  - Notes – contains a free-text note for the specific variable (Name) in the library (Libname)
  - Urltitle – contains the text that needs to be hyperlinked
  - Url – contains the destination link

Parameter	Mandatory	Default value	Parameter Description
libname	Yes		The SAS Library that a data dictionary is being created for.
dataset			A specific dataset in a specified library (&libname). If left blank, then a data dictionary will be generated for all the datasets in a specified library (&libname).
metadata		meta.metadata	Name of the metadata dataset along with the SAS library in which it is present.
fmtlib		formats	Name of the SAS library containing the format catalog.
path	Yes		The physical location of the data dictionary.
lookupsubdir		Lookup	A subdirectory within the PATH to store lookup table pages.  Note: This subdirectory should be created in advance.

Parameter	Mandatory	Default value	Parameter Description
<b>varsubdir</b>		Variables	A subdirectory within the PATH to store variable pages.  Note: This subdirectory should be created in advance.
<b>shownvalue</b>		20	The maximum number of rows for “Values”. If the number of values exceeds this number, then instead of the list of values, a link will be shown pointing the user to a separate page containing the values or lookup tables.
<b>title</b>	Yes		Title for the main HTML page.

**Example 1.** Creates a complete data dictionary HTML page called “DEMO\_hospitalization\_dictionary.html” in \users\temp for “hospitalization” dataset present in the “demo” library.

```
%dictionary(libname=demo,
              dataset=hospitalization,
              metadata=meta_hospitalization,
              fmtlib=work,
              path=\users\temp,
              lookupsubdir=Lookup,
              varsubdir=Variables,
              shownvalue=20,
              title=The Data Dictionary for demo.hospitalization);
```

**Example 2.** Creates complete data dictionary HTML pages in \users\temp for all the datasets present in the “demo” library.

```
%dictionary(libname=demo,
              path=\users\temp,
              title=The Data Dictionary for all datasets in demo library);
```

### 2.2.3. VIMO macro

The VIMO macro measures the correctness of values. It captures the percentage of Valid, Invalid, Missing and Outlier values for each data element. It also generates basic descriptive statistics and the frequency of coded values or histograms for each data elements. This macro’s final report will be a clickable HTML report.

When there is a format associated with a variable, any value in the data that is not defined in its format is considered invalid. Invalid values could represent the gaps in data documentation and metadata and not necessarily invalid values. When the data element is numerical, the proportion of outliers (extreme values) is calculated to reflect a potential data quality issue.

The VIMO macro groups variables in a given dataset into one of four categories, including:

- **ID variables**
  - The variables, which are identified as ID, must be explicitly provided to the macro.
  - The ID parameter in the macro accepts a space separated list of variable names.
  - For each of the variables passed to the ID parameter, the macro determines the uniqueness as follows:
    - Unique – the variable had unique non-missing values.
    - Unique when not missing – the variable could have a missing value, but it had unique non-missing values.
    - Not unique – the variable had some non-unique non-missing values.
- **Numeric variables**
  - The VIMO macro is coded to identify numeric variables automatically.
  - It performs a specific process and analysis for each numeric variable to report descriptive statistics (min, max, mean, median and std).
- **Character variables**
  - The VIMO macro is coded to identify character variables automatically.
  - It reports all (or the first and last level of) values.
  - Character variables in the VIMO table are hyperlinked to an HTML page, which is also generated by the VIMO macro. When a user clicks on the variable' hyperlink, a frequency table, which contains the actual values and their formatted values, will pop up.
- **Date/time variables**
  - The VIMO macro is coded to identify date/time variables automatically.
  - The earliest and latest values formatted in date/time will be reported.

The VIMO macro is able to handle large datasets. This has been achieved by:

- Using a proc sql procedure instead of proc freq.
- To identify outliers, using a piecewise parabolic ( $P^2$ ) algorithm to calculate quartiles.<sup>3</sup>

Parameter	Mandatory	Default value	Parameter Description
<b>ds</b>	Yes		Name of input SAS dataset.
<b>invalids</b>		ON	Turn on or off invalid checks. Valid values: ON or OFF.
<b>path</b>	Yes		Output location/physical path for the HTML VIMO report.
<b>fmtlib</b>		formats	SAS library name(s) containing format catalog(s).
<b>metalib</b>		meta	SAS library name containing metadata created by using meta macro.
<b>freq</b>		ON	Turn on/off the frequency tables for character variables in the HTML output. These tables can be displayed by clicking on character variable names in the VIMO report. Valid values: ON or OFF.  Note: A sub-folder called 'Freq' must exist under the given path specified as value to path parameter.
<b>excludefreq</b>		pstlcode	Space delimited list of variables to be excluded from the frequency tables.
<b>id</b>			Space-delimited list of ID variables.
<b>time</b>		fiscal	Get Date or Datetime variable frequencies by fiscal year or by calendar year or by quarterly or by monthly. Valid values: FISCAL, CALENDAR, QUARTERLY, MONTHLY.
<b>log</b>		OFF	Turn on or off the SAS log. Valid values: ON or OFF.
<b>scriptsRAEpath</b>			LINUX location of the JQuery related scripts mentioned in table.
<b>scriptsWINpath</b>			Windows location of the JQuery related scripts mentioned in table.

**Example 1.** Creates hospitalization\_vimo.html page in S:\DQIM\DataFit Toolkit\DQ Reports using for "hospitalization" dataset in "demo" library with variables key, pat\_id and chart\_num as ID variables. Invalid value checks are done and frequency tables are also generated.

```
%vimo(ds=demo.hospitalization,
      path= S:\DQIM\DataFit Toolkit\DQ Reports,
      metalib=demo,
      fmtlib=work,
      id= key pat_id chart_num);
```

**Example 2.** Creates hospitalization\_vimo.html page in S:\DQIM\DataFit Toolkit\DQ Reports using for "hospitalization" dataset in "demo" library with variables key, pat\_id and chart\_num as ID variables. Invalid value checks are not done and frequency tables are not generated.

```
%vimo(ds=demo.hospitalization,
      path=S:\DQIM\DataFit Toolkit\DQ Reports,
      metalib=demo,
      fmtlib=work,
      invalids=OFF,
      freq=OFF
      id= key pat_id chart_num);
```

## 2.2.4. TIM macro

The TIM (Trends In Missingness) macro reports the percentage of missing values over time for each data element. TIM is a complimentary report for VIMO to dive into more details for completeness.

The TIM macro can be run in two different ways depending on the type of dataset in the SAS library (yearly or cumulative):

- TIM macro for a yearly dataset
  - Datasets in the SAS library must follow the naming convention <datasetname><yyyy> where <yyyy> represents the year of data.
  - The rate of missing values is determined based on the year of the dataset. In a given SAS library, the missing percentage is calculated for all variables in all datasets in a specified time period.
- TIM macro for a cumulative dataset
  - The rate of missing values is determined based on the year in the reference date variable (refdate), which is provided as input to the macro.

Parameter	Mandatory	Default value	Parameter Description
<b>library</b>	Yes		SAS library where the input dataset can be found.
<b>data</b>	Yes		<p>Actual dataset Name or prefix of name for input SAS dataset(s).</p> <p>This parameter is specified in two different ways, depending on if the macro is to be run on yearly datasets or a single cumulative dataset.</p> <p>Yearly datasets: Prefix name of the input SAS datasets. The expected naming convention to be followed for naming yearly dataset is &lt;datasetname&gt;&lt;yyyy&gt;; thus &lt;datasetname&gt; would be the value for the data parameter.</p> <p>Cumulative datasets: Name of input SAS dataset.</p>

Parameter	Mandatory	Default value	Parameter Description
<b>start</b>	Yes (if macro is run on yearly datasets)		<p>First year of data to report on.</p> <p>This parameter is specified in two different ways, depending on whether the macro is to be run on yearly datasets or a single cumulative dataset.</p> <p>Yearly datasets: Mandatory; the first year of the data to report on.</p> <p>Cumulative datasets: Need not be specified; however if specified, the TIM report generated will be starting from the year passed to this parameter.</p>
<b>end</b>	Yes (if macro is run on yearly datasets)		<p>Last year of data to report on.</p> <p>This parameter is specified in two different ways, depending on whether the macro is to be run on yearly datasets or a single cumulative dataset.</p> <p>Yearly datasets: Mandatory; the last year of the data to report on.</p> <p>Cumulative datasets: Need not be specified; however if specified, the TIM report generated will be ending at the year passed to this parameter.</p>
<b>refdate</b>	Yes (if macro is run on cumulative datasets)		<p>Name of index date variable.</p> <p>This parameter is specified in two different ways, depending on whether the macro is to be run on yearly datasets or a single cumulative dataset.</p> <p>Yearly datasets: Not applicable.</p> <p>Cumulative datasets: The date variable in the dataset will be used as the index date for the report. While choosing the index date; ensure that it has good coverage and is good representative of dataset.</p>
<b>altdir</b>			Alternate directory to read data from.
<b>path</b>			Output location/physical directory path for the HTML TIM report.
<b>testn</b>			<p>Number of observations to limit input SAS dataset(s) to. This parameter is to be used for testing purposes only.</p> <p>Example: testn=100.</p>
<b>time</b>		FISCAL	<p>Generate the TIM report based on calendar year or fiscal year.</p> <p>Valid values: FISCAL, CALENDAR.</p>

**Example 1.** Creates hospitalization\_tim.html page in S:\DQIM\DataFit Toolkit\DQ Reports using "hospitalization" cumulative dataset in "demo" library based on ddate as index date.

```
%tim(library=demo,
      data=hospitalization,
```

```
refdate=ddate,
path= S:\DQIM\DataFit Toolkit\DQ Reports);
```

**Example 2.** Creates hospitalization\_tim.html page in S:\DQIM\DataFit Toolkit\DQ Reports using yearly hospitalization datasets from 2005 to 2014 in demo.

```
%tim(library=demo,
      data=hospitalization,
      start=2005,
      end=2014,
      path= S:\DQIM\DataFit Toolkit\DQ Reports);
```

### 2.2.5. Trend macro

The Trend macro fits the following series of seven different smooth lines or curves to the aggregated number of observations in a given dataset by a date variable:

- Simple linear:  $Y = \beta_0 + \beta_1 X$
- Quadratic:  $Y = \beta_0 + \beta_1 X^2$
- Exponential:  $Y = \beta_0 + \beta_1 \exp(X)$
- Logarithmic:  $Y = \beta_0 + \beta_1 \log(X)$
- SQRT:  $Y = \beta_0 + \beta_1 \sqrt{X}$
- Inverse:  $Y = \beta_0 + \beta_1$
- Negative exponential:  $Y = \beta_0 + \beta_1 \exp(-X)$

To choose a date variable that describes changes in the number of records over time in the best way, the date variable must be identified by a user who knows the data well. For each model, the root mean square error (RMSE) between the model's output and observations is calculated, and the model with the minimum RMSE is selected as the optimum model to represent the temporal changes in the number of observations by the chosen date variable. Then, the chosen model is refitted to the data to perform outlier analysis. This is done by calculating studentized residuals for each observation when that observation is removed from the model. Studentized residuals that are statistically significant (i.e., larger or smaller than expected) are identified as potential outliers or unusual data points. The Trend macro also identifies subsequent identical observations (indicating no change over time) and flags these as potential errors in data extraction.<sup>8</sup>

The Trend macro checks for small absolute numbers of records (between 1 and 5 inclusive) in each time period. If the macro finds any small frequencies, it replaces them with 3 (the average of all possible small numbers as an estimated value). The modeling and outlier analysis are done based on the actual numbers, but for demonstration of the trend graphs, small numbers are changed to 3 to comply with privacy policies.

Parameter	Mandatory	Default value	Parameter Description
<b>ds</b>	Yes		<p>Actual dataset Name or prefix of name for input SAS dataset(s) along with the SAS library name.</p> <p>This parameter is specified in two different ways, depending on whether the macro is to be run on yearly datasets or a single cumulative dataset.</p> <p>Yearly datasets: Prefix name of the input SAS datasets. The expected naming convention to be followed for naming yearly dataset is &lt;datasetname&gt;&lt;yyyy&gt;. Thus, &lt;datasetname&gt; would be the value for the parameter.</p> <p>Cumulative datasets: Name of input SAS dataset.</p>
<b>Dsprefix</b>	Yes	OFF	<p>Flag to differentiate between yearly and cumulative datasets.</p> <p>Yearly datasets: Value ON is specified for DSPREFIX if macro is run for yearly dataset.</p> <p>Cumulative datasets: Value OFF is specified for DSPREFIX if macro is run for yearly dataset.</p> <p>Valid values: ON or OFF.</p>
<b>Startyr</b>	Yes		The first year (4-digit year) of the data being reported on.
<b>Endyr</b>	Yes		The last year (4-digit year) of the data being reported on.
<b>Bydate</b>	Yes		<p>Name of index date variable, based on which the trend report will be generated.</p> <p>While choosing the index date, ensure that it has good coverage and is good representative of dataset.</p>
<b>Byvar</b>		_ALL_	When it's provided, multiple Trend graphs will be generated for each value of this categorical variable.
<b>Byfmt</b>			Format for the categorical variable specified in BYVAR parameter.
<b>Time</b>			<p>Parameter to specify how the trend report must be generated.</p> <p>Valid values: FISCAL, CALENDAR, MONTHLY, QUARTERLY.</p>
<b>Path</b>	Yes		Output location of PNG format of graph.

**Example 1.** Creates a fiscal year-based trend report called hospitalization\_trend.png image in P:\Work\DataFit Toolkit\DQReports using “hospitalization” cumulative dataset in a “demo” library with ddate as the index date.

```
%trend(ds=demo.hospitalization,
      startyr=2005,
```



```

endyr=2014,
bydate=ddate,
time=fiscal,
path=P:\Work\DataFit Toolkit\DQReports);

```

**Example 2.** Creates a fiscal year–based trend report called work\_all\_trend.png image in P:\Work\DataFit Toolkit\DQReports using yearly hospitalization datasets from 2005 to 2014 in a “work” library with ddate as the index date.

```

%trend(ds=work.hospitalization,
      dsprefix=ON,
      startyr=2005,
      endyr=2014,
      bydate=ddate,
      time=fiscal,
      path=P:\Work\DataFit Toolkit\DQReports);

```

### 2.2.6. Linkability macro

The Linkability macro can be run on one or a series of datasets with the same prefix name within a given SAS library.

The Linkability macro requires an index date and a linkage type variable to be present in the input dataset.

- The linkage type variable could contain the validation information of a health card number (i.e., if it follows a valid health card pattern):
  - B = blank or other health number
  - N = invalid health number
  - V = encoded health card number based on valid health number
- Linkage type variables contain coded value to represent the method used to add encoded health card numbers in the dataset:
  - D = deterministic linkage
  - H = valid health card number supplied
  - P = probabilistic linkage
- A format corresponding to coded value are created in format catalog

The macro generates an HTML report and it can be based on calendar year, fiscal year, month or quarter.

Parameter	Mandatory	Default value	Parameter Description
<b>ds</b>	Yes		<p>Actual dataset name or prefix of name for input SAS dataset(s) including the SAS library.</p> <p>This parameter is specified in two different ways, depending on if macro is to be run on yearly datasets or a single cumulative dataset.</p> <p>Yearly datasets: Prefix name of the input SAS datasets. The expected naming convention to be followed for naming yearly dataset is &lt;datasetname&gt;&lt;yyyy&gt;. Thus, &lt;datasetname&gt; would be the value for the parameter.</p> <p>Cumulative datasets: Name of the input SAS dataset.</p>
<b>dsprefix</b>	Yes	OFF	<p>Flag for the macro to differentiate between yearly and cumulative datasets.</p> <p>Yearly datasets: Value ON is specified for DSPREFIX if macro is run for yearly dataset.</p> <p>Cumulative datasets: Value OFF is specified for DSPREFIX if macro is run for yearly dataset.</p> <p>Valid values: ON or OFF.</p>
<b>bydate</b>	Yes		Name of index date variable; based on which the linkability report will be generated.
<b>fmtlib</b>		FORMATS	Name of the format catalog where format specified in linktypeformat parameter can be found.
<b>linktype</b>	Yes		Linkage type variable. The variable which contains the type of linkage (or primary ID type).
<b>linktypefmt</b>	Yes		Format of linktype variable.
<b>linkvalue</b>		D, P, H, V	Valid values used for data linkage.
<b>startyr</b>	Yes		The first year (4-digit year) of the data being reported on.
<b>endyr</b>	Yes		The last year (4-digit year) of the data being reported on.
<b>time</b>		FISCAL	<p>Parameter to specify how the linkability report must be generated. Should it be calendar year, fiscal year, monthly or quarterly based.</p> <p>Valid values: FISCAL/CALENDAR/MONTHLY/QUARTERLY.</p>
<b>path</b>	Yes		Output location for HTML linkability report.

**Example 1.** Creates fiscal year–based linkability report called hospitalization\_linkability.html in P:\Work\DataFit Toolkit\DQReports using “hospitalization” cumulative dataset in “demo” library with ddate as the index date.

```
%linkability(ds=demo.hospitalization,
            bydate=ddate,
```

```

linktype=linkage_type,
linktypefmt=$Hosp_link_type.,
startyr=2005,
endyr=2014,
time=fiscal,
fmtlib=work,
path=P:\Work\DataFit Toolkit\DQReports);

```

**Example 2.** Creates fiscal year–based linkability report called hospitalization\_linkability.html in P:\Work\DataFit Toolkit\DQReports using yearly hospitalization datasets from 2005 to 2014 in "work" library with ddate as index date.

```

%linkability(ds=work.hospitalization,
            dsprefix=ON,
            bydate=ddate,
            linktype=linkage_type,
            linktypefmt=$Hosp_link_type.,
            startyr=2005,
            endyr=2014,
            time=fiscal,
            fmtlib=work,
            path=P:\Work\DataFit Toolkit\DQReports);

```

### 2.2.7. Agreement macro

The agreement macro generates a standard HTML report for comparing sex and date of birth in a given input data with a registry data. The macro accepts a single cumulated dataset or a series of annual datasets.

Parameter	Mandatory	Default value	Parameter Description
<b>lib</b>	Yes	work	SAS library containing input dataset.
<b>templib</b>		work	SAS library for creating intermediate datasets during macro execution.
<b>ds</b>	Yes		Name or prefix of name for input SAS dataset(s) including the SAS library.
<b>ds_prefix</b>	Yes	OFF	Flag for the macro to differentiate between yearly and cumulative datasets.  Yearly datasets: Value ON is specified for DSPREFIX if the macro is run for yearly dataset.  Cumulative datasets: Value OFF is specified for DSPREFIX if macro is run for yearly dataset.  Valid values: ON/OFF.
<b>ds_crosswalk</b>			If sex and date of birth variables are not present in the dataset specified in the DS parameter but are in another dataset in the SAS library specified in the LIB parameter, then the name of the dataset containing those variables must be specified here. The macro has been

Parameter	Mandatory	Default value	Parameter Description
			implemented to merge the two datasets to retrieve relevant variables.
<b>ds_mergeby</b>	Yes (if ds_crosswalk parameter has been passed non-null value)		If sex and date of birth variables are not present in the dataset specified in DS parameter; but is in another dataset within the SAS library specified in LIB parameter; the two datasets are merged to retrieve relevant variables. The ID variable that links the two datasets must be specified in this parameter.
<b>ds_linktype</b>			Linkage type variable. The variable which contains the type of linkage (or primary ID type).  Note: If the ds_linktype variable is not specified, the macro creates a link type variable with value as "D" (Deterministic).
<b>ds_startyr</b>	Yes		The first year (4-digit year) of the data being reported on.  Note: For yearly datasets, this is a mandatory parameter. For cumulative datasets, it is optional; however, if it is not specified, the agreement HTML report generated will contain a summary table but not a year-based-report table.
<b>ds_endyr</b>	Yes		The last year (4-digit year) of the data being reported on.  Note: For yearly datasets, this is a mandatory parameter. For cumulative datasets, it is optional; however, if it is not specified, the agreement HTML report generated will contain a summary table but not a year-based-report table.
<b>ds_bydate</b>	Yes		Name of index date variable, based on which the agreement report will be generated.  Note: For cumulative datasets, if ds_startyr and ds_endyr parameters are not specified but ds_bydate is specified, the macro will determine the appropriate year range and generate a year-based- report table.
<b>ds_byvar</b>	Yes		ID variable in input dataset that can be used to link input dataset with the "gold standard" reference dataset.
<b>ds_datevar</b>			Date variable in input dataset for which the agreement report needs to be generated.  Note: Date of birth variable is specified here.

Parameter	Mandatory	Default value	Parameter Description
<b>ds_categvar</b>			Categorical variable in the input dataset for which the agreement report needs to be generated.  Note: Sex variable is specified here.
<b>ref_data</b>	Yes		Name of SAS dataset considered as the "gold standard" along with the SAS library.
<b>ref_byvar</b>	Yes		Variable in the "gold standard" reference dataset that can be used to link with the input dataset.
<b>ref_datevar</b>		bdate	Date variable in the "gold standard" reference dataset for which the agreement report needs to be generated.  Note: Date of birth variable is specified here
<b>ref_categvar</b>		sex	Categorical variable in the "gold standard" reference dataset for which the agreement report needs to be generated.  Note: Sex variable is specified here.
<b>time</b>	Yes	fiscal	Parameter to specify how the agreement report must be generated.  Valid values: FISCAL/CALENDAR.
<b>path</b>	Yes		Output location for HTML agreement report.

**Example 1.** Creates a fiscal year–based agreement report called hospitalization\_agreement.html in P:\Work\DataFit Toolkit\DQReports using the “hospitalization” cumulative dataset in the "demo" library, with ddate as the index date.

```
%agreement(lib=demo,
            templib=work,
            ds=hospitalization,
            ds_prefix=off,
            ds_startyr=2005,
            ds_endyr=2014,
            ds_byvar=pat_id,
            ds_bydate=ddate,
            ds_datevar=birthdate,
            ds_categvar=sex,
            ref_data=demo.referencedata,
            ref_byvar=pat_id,
            ref_datevar=birthdate,
            ref_categvar=sex,
            time=fiscal,
            path=P:\Work\DataFit Toolkit\DQReports);
```

**Example 2.** Creates a fiscal year–based agreement report called hospitalization\_agreement.html in P:\Work\DataFit Toolkit\DQReports using yearly hospitalization datasets from 2005 to 2014 in a "demo" library with ddate as the index date.

```
%agreement(lib=work,  
            templib=work,  
            ds=hospitalization,  
            ds_prefix=off,  
            ds_startyr=2005,  
            ds_endyr=2014,  
            ds_byvar=pat_id,  
            ds_bydate=ddate,  
            ds_datevar=birthdate,  
            ds_categvar=sex,  
            ref_data=demo.referencedata,  
            ref_byvar=pat_id,  
            ref_datevar=birthdate,  
            ref_categvar=sex,  
            time=fiscal,  
            path=P:\Work\DataFit Toolkit\DQReports);
```

## 3.0 Tutorial

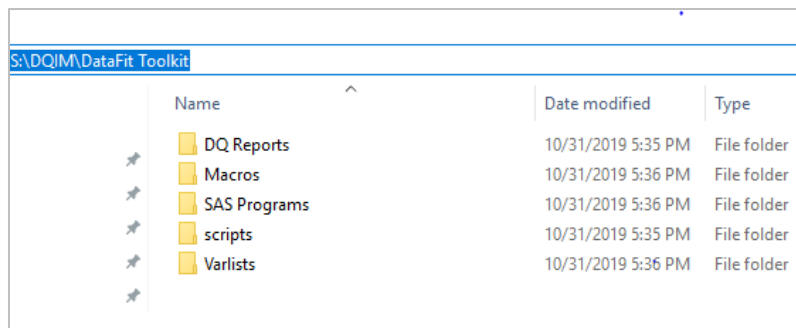
In this chapter, you will be walked through the entire package and guided step by step to create simulated data and metadata and generate data quality reports for your simulated data.

### Step 1. Download the DataFit Toolkit

Download the latest version of the DataFit Toolkit from <https://github.com/icescentral/DataFit-Toolkit>

### Step 2. Configure the folder structure

Create the folder structure shown in Figure 10 in a folder called “DataFit Toolkit” in a drive (here S). Copy the SAS macros to the Macros folder and the four SAS program files to the SAS Programs folder.



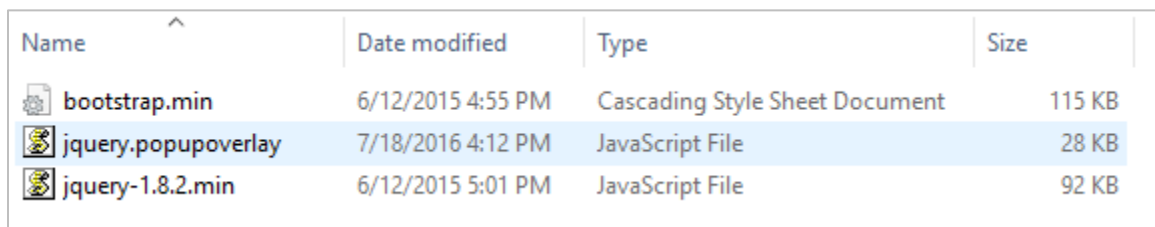
Name	Date modified	Type
DQ Reports	10/31/2019 5:35 PM	File folder
Macros	10/31/2019 5:36 PM	File folder
SAS Programs	10/31/2019 5:36 PM	File folder
scripts	10/31/2019 5:35 PM	File folder
Varlists	10/31/2019 5:36 PM	File folder

Figure 10. Folder structure

### Step 3. Download the scripts

If the environment in which you will be viewing the final data quality reports is moated and therefore has no access to the Internet, you will need to make Bootstrap and JQuery available in offline mode. These reports provide interactive functionalities in the HTML reports, such as pop-up windows and hide/show features for the VIMOs. To make these open-source scripts available offline, download them from the following URLs and copy them to your “scripts” subfolder (Figure 11).

- Bootstrap
  - URL: <https://getbootstrap.com/>
  - Required script: “bootstrap.min”
- JQuery
  - URL: <https://code.jquery.com/>
  - Required scripts: “jquery.popupoverlay” and “jquery-1.8.2.min”



Name	Date modified	Type	Size
bootstrap.min	6/12/2015 4:55 PM	Cascading Style Sheet Document	115 KB
jquery.popupoverlay	7/18/2016 4:12 PM	JavaScript File	28 KB
jquery-1.8.2.min	6/12/2015 5:01 PM	JavaScript File	92 KB

Figure 11. Required contents of the scripts sub-folder

#### Step 4. Simulate hospitalization data and its format catalog

Run the first three SAS programs in the “SAS Programs” folder in order. Here is a brief description of each program:

- **01\_createSimulatedData.sas**

This file contains the createSimulatedData macro code. This macro can generate simulated hospitalization data for a given number of observations. When you run this program, SAS compiles it and makes it available in the memory. You don’t need to modify anything in this program.

- **02\_initiate\_simulateddata\_creation.sas**

This program invokes createSimulatedData macro to create a specific simulated datasets representing hospitalization. For the purpose of this tutorial, the following parameters are chosen and you don’t need to modify anything:

- nobs = 505,062 (number of records)
- endyr = 2014 (end year)
- noyrs = 10 (number of years)

Notes:

- The minimum value that can be passed as a value for parameter NOBS is 500,000.
- Rerunning the program for the same input will not result in the same dataset because the date and time variables are generated randomly.
- Errors are also randomly incorporated in the simulated dataset by the macro for demonstrating various data quality assessment tools.

- **03\_create\_formatcatalog\_simulateddata.sas**

This program generates a format catalog for the variables in the simulated dataset “hospitalization.” You don’t need to modify anything in this program.

After running these three programs in order, you should have a SAS dataset called Hospitalization in your SAS “Work” library and also all the associated formats in the SAS Format Catalog under your SAS “Work” library.

#### Step 5. Review the VARLIST file

In step 4, you created a SAS Format Catalog for your Hospitalization dataset; however, the SAS system does not know which variable is associated with which format. To fill that gap, a “VARLIST” file under “Varlist” subfolder is provided to you. Open the hospitalization\_varlist.txt to see its contents (Figure 12).

Note that for your actual data quality projects, you will need to create similar “VARLIST” files for each of your SAS libraries or datasets.



```

File Edit Format View Help
INSTTYPE      $HOSP_INSTTYPE.
LINKAGE_TYPE  $HOSP_LINK_TYPE.
DXTYPE2       $HOSP_DXTYPE.
AGECODE       $HOSP_AGECODE.
SEX           $HOSP_SEX.
AGECAT        $HOSP_AGECAT.
DX10CODE1     $HOSP_DX10CODE.
DXPREF1       $HOSP_DXPREF.
DXTYPE1       $HOSP_DXTYPE.
DX10CODE2     $HOSP_DX10CODE.
DXPREF2       $HOSP_DXPREF.
PAT_IDTYPE    $HOSP_VALPATID.

```

Figure 12. Contents of hospitalization\_varlist.txt

### Step 6. Run the DataFit Toolkit and generate data quality reports

The last SAS program, which is **04\_generate\_dataquality\_reports.sas**, contains eight blocks of codes. In the codes where a path is mentioned, you will need to modify it to match the path in your computer or analytical environment. For a first-time user, it is strongly recommended that this program be run in blocks. The first block will set up the DataFit Toolkit in your environment. The second block will create metadata for your project. The remaining blocks will create different data quality reports, such as VIMO, Linkability, Trend, TIM and Agreement, and a data dictionary for your project.

### Step 7. Review the data quality reports

If you were able to complete the previous steps without any errors or issues, you should find your data quality reports in your DQ reports subfolder (Figure 13). With the exception of the Trend report, which is a PNG image, all reports are in HTML format.

Name	Date modified	Type	Size
Freq	10/31/2019 5:35 PM	File folder	
Lookup	10/31/2019 5:35 PM	File folder	
Variables	10/31/2019 5:35 PM	File folder	
hospitalization_agreement	5/17/2018 8:11 AM	HTML Document	84 KB
hospitalization_bland_altman_D1	5/17/2018 8:11 AM	PNG File	25 KB
hospitalization_linkability	5/17/2018 8:09 AM	HTML Document	48 KB
hospitalization_linkability	5/17/2018 8:09 AM	PNG File	28 KB
HOSPITALIZATION_tim	5/17/2018 8:10 AM	HTML Document	74 KB
hospitalization_trend	5/17/2018 8:10 AM	PNG File	28 KB
hospitalization_VIMO	9/23/2019 3:03 PM	HTML Document	43 KB
WORK_hospitalization_dictionary	5/17/2018 8:12 AM	HTML Document	48 KB

Figure 13. Browsing generated data quality reports

## References

1. Azimaee M, Smith M, Lix L, Ostapyk T, Burchill C, Orr J. *MCHP Data Quality Framework*. Winnipeg, MB: Manitoba Centre for Health Policy; 2018. Accessed November 18, 2019 at [http://umanitoba.ca/faculties/medicine/units/chs/departamental\\_units/mchp/protocol/media/Data\\_Quality\\_Framework.pdf](http://umanitoba.ca/faculties/medicine/units/chs/departamental_units/mchp/protocol/media/Data_Quality_Framework.pdf).
2. Smith M, Lix LM, Azimaee M, Enns JE, Orr J, Hong S, Roos LL. Assessing the quality of administrative data for research: a framework from the Manitoba Centre for Health Policy. *J Am Med Inform Assoc*. 2018; 25(3):224–9.
3. Azimaee M. Data fitness: a SAS® macro-based application for data quality of large health administrative data. Proceedings of the SAS Global Forum 2013, San Francisco, CA, April 28–May 1, 2013. Accessed November 18, 2019 at <http://support.sas.com/resources/papers/proceedings13/075-2013.pdf>.
4. Lix LM, Smith M, Azimaee M, Dahl M, Nicol P, Burchill C, Burland E, Goh CY, Schultz J, Bailly A. *A Systematic Investigation of Manitoba's Provincial Laboratory Data*. Winnipeg, MB: Manitoba Centre for Health Policy; 2012. Accessed November 18, 2019 at [http://mchp-appserv.cpe.umanitoba.ca/reference/cadham\\_report\\_WEB.pdf](http://mchp-appserv.cpe.umanitoba.ca/reference/cadham_report_WEB.pdf).
5. Iron K, Manuel DG. *Quality Assessment of Administrative Data (QuAAD): An Opportunity for Enhancing Ontario's Health Data*. Toronto, ON: Institute for Clinical Evaluative Sciences; 2007. Accessed November 18, 2019 at <https://www.ices.on.ca/~media/Files/Atlases-Reports/2007/Quality-assessment-of-administrative-data/Full-report.ashx>.
6. NHS Digital. Mental health services data quality at <https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/mental-health-data-hub/mental-health-services-data-quality>
7. Canadian Institute for Health Information. The CIHI Data Quality Framework, 2009. Ottawa, ON: CIHI; 2009. Accessed November 18, 2019 at [https://www.cihi.ca/sites/default/files/data\\_quality\\_frsmework\\_2009\\_en\\_0.pdf](https://www.cihi.ca/sites/default/files/data_quality_frsmework_2009_en_0.pdf).
8. Azimaee M. *Trend analysis: an automated data quality approach for large health administrative databases*. Proceedings of the SAS Global Forum 2012, Orlando, FL, April 22-25, 2012. Accessed November 18, 2019 at <https://support.sas.com/resources/papers/proceedings12/123-2012.pdf>.
9. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician*. 1983; 32(3):307–17.