Paper 075-2013

Data Fitness: A SAS® Macro-based Application for Data Quality of Large Health Administrative Data

Mahmoud Azimaee, Institute for Clinical Evaluative Sciences (ISEC)

ABSTRACT

This paper introduces a SAS® macro-based application package as a solution for creating automated data quality assurance reports for large health administrative data. It includes methods and tools for developing metadata for a SAS data holding, for measuring different data quality indicators using a Data Quality Framework, and for generating automated visual data quality reports. Because quality of data documentation should be considered as a usability and interpretability factor for good quality data, this application uses the same metadata developed for data quality purposes to generate an automated web-based data dictionary as well.

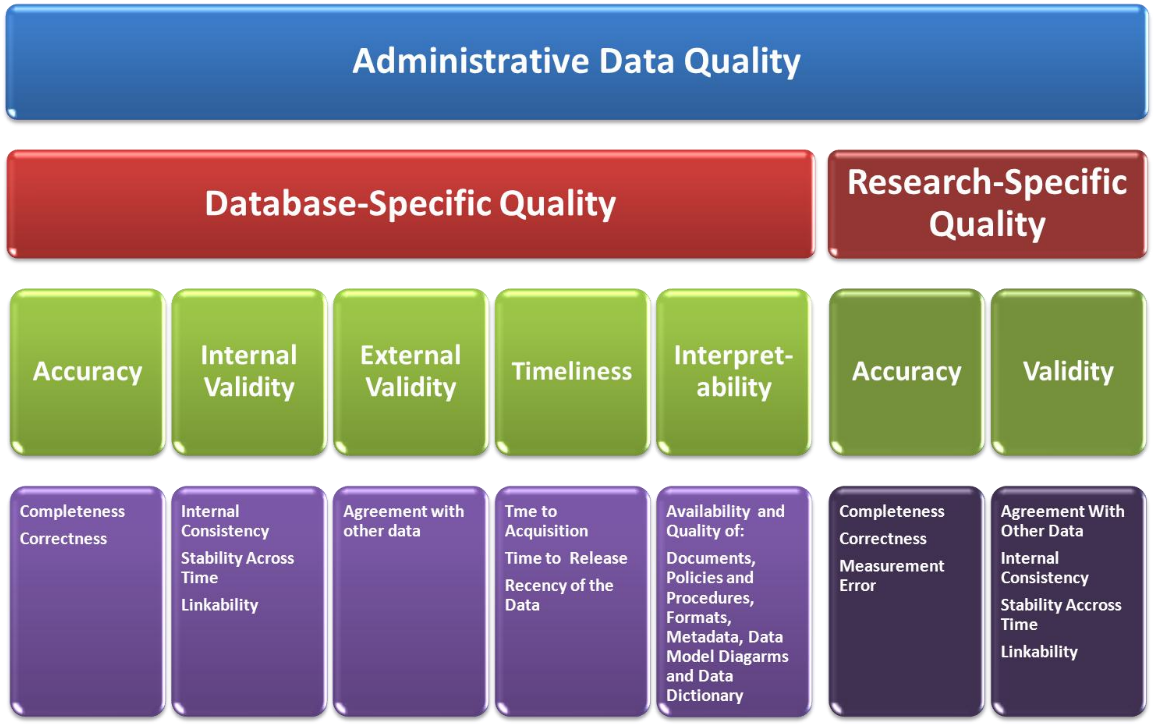
INTRODUCTION

Quality evaluation of research-ready data is an important factor in conducting medical research project using large health administrative data. During a research project started in the summer of 2010, Manitoba Centre for Health Policy (MCHP), a research unit within the faculty of Medicine, University of Manitoba, initiated creating a Data Quality Framework, which suited their need for data quality assurance of their data repository. After studying different data quality frameworks from similar data holding organizations in Canada [4, 9], the UK [10] and Australia [1, 2], a specific data quality framework for MCHP was designed [7]. This framework measures and evaluates five dimensions of database-specific quality of administrative data:

1. Accuracy
2. Internal Validity
3. External Validity
4. Timeliness
5. Interpretability

Each of these dimensions has its own components such as Completeness, Correctness, Internal Consistency, Stability Across time and Likability. Figure 1 shows a diagram of MCHP Data Quality Framework.

Figure 1: MCHP Data Quality Framework

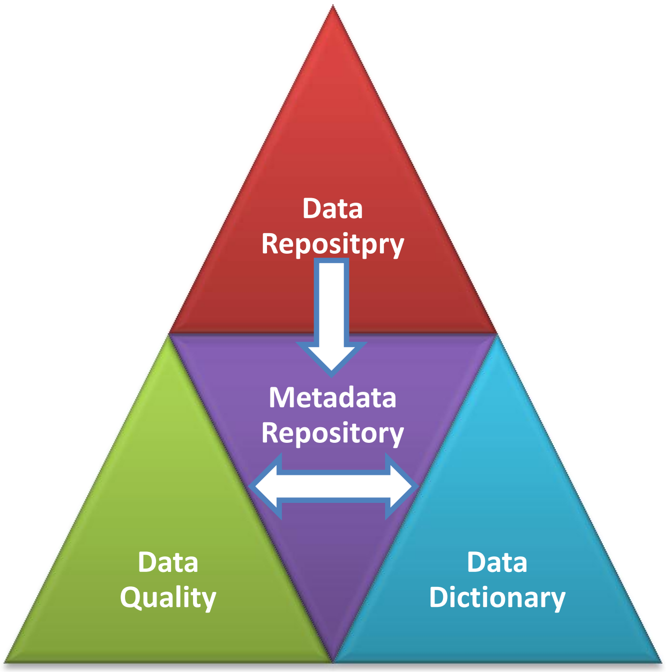


Considering the growing number of databases including the new data and annual updates to the existing ones, it was clear that the process had to work with the minimum amount of manual interference. Therefore a package of 18 SAS macros was designed to apply this framework to the data in the MCHP repository. Two years later, the author had the opportunity to adopt and generalize this package for the data repository of the Institute of Clinical Evaluative Sciences (ICES). Although the nature of the data in ICES data repository is very similar to MCHP, but considering the population of Ontario which is almost 10 times of Manitoba's population, some development in the macro were necessary in order to handle larger datasets. Also, many new features were added to the application. Data Fitness application includes 18 SAS macros that altogether generate a metadata repository, different data quality reports and also a web-ready Data Dictionary in HTML format.

This paper discusses the methodology and techniques that were used in developing this Data Quality Application. Data Fitness package has been written to work under UNIX, however it can be easily modified to work under Windows operating system too.

FIRST STEP: METADATA REPOSITORY

Figure 2: Data Fitness workflow



Metadata is defined as "data about data". The main key in implementing a successful Data Quality system is to know everything about the data in the data repository. The very first macro creates a metadata repository.

METADATA MACRO

Meta macro is a PROC CONTENTS with OUT option to generate a reference table including names, labels, formats, length and other attributes of variables within each dataset. It also gathers some additional information, such as physical location of data on the disk, UNIX group associated with each data, the owner of data and permission settings which are not being generated using PROC CONTENTS.

For technical reasons, most organizations which own a data holding prefer not to apply user defined formats to the data. This means that, SAS system does not know which variable is associated with which formats. META macro uses a reference table to get this information and incorporate it into the final metadata. These tables, that we call them VARLIST, must be created and maintained manually before running META macro. VARLISTs are simple plain text files which include a variable name and its format separated with space or tab in each row. META macro reads them using a %INC statement to apply formats to the variables on the fly when generating CONTENTS output.

%META PARAMETERS AND DESCRIPTIONS

LIB:	SAS library name that you want to create metadata for
DS:	Dataset prefix or complete name of a dataset in &LIB. (If left blank, then metadata will be generated for the entire directory)
EXCL:	Dataset name to exclude from metadata
FMTLIB:	SAS library name(s) containing the format catalog (default is Formats)
OUTLIB:	SAS library name for output dataset (default is Meta)
METALIB:	SAS library name containing METADATA (default is Meta)
PATH:	Location of VarList file (a text, tab/space delimited file, containing variable names and their associated formats

EXAMPLES:

```
%META (LIB=ohip,
        DS =ohip ,
        FMTLIB=meta,
        PATH=/metadata/varlists/ohip_varlist.txt
        )
```

```
%META (LIB=ohip,
        EXCL =ref ,
        FMTLIB=meta,
        PATH=/metadata/varlists/ohip_varlist.txt
        )
```

EVALUATING ACCURACY OF DATA

Accuracy refers to the degree to which the data correctly describe the phenomenon they were designed to measure. This is an important component of quality as it relates to how well the data portray reality, which has clear implications for how useful and meaningful the data will be for interpretation or further analysis. [3]

Accuracy includes Completeness and Correctness. Completeness can be measured by the rate of missing values. However, completeness or comprehensiveness can also be measured by investigating database exclusions. If selected sub–groups are missing from a database because of exclusions based on age, stage/type of disease, or geography, then the databases will result in incomplete estimates of the target outcome (e.g., incidence or prevalence).[7] These kinds of exclusion in database-specific data quality are not taken into account for evaluating completeness; instead, they were left for research-specific data quality.

Correctness is measured by the percentage of valid values, that is, values within the domain of possible or plausible values. Values may be invalid because they violate physical, logical, or metadata–based constraints. An assessment of validity of data values requires documentation about plausible values as well as knowledge gained through exploratory analyses of the data.[7] In the Data Fitness application, rate of invalid values are mainly calculated by comparing the values of each data element (variable) with its described values in the format catalog. When a variable is associated with a specific format but it has a value without a description in the format catalog, that value is considered as Invalid. For the numeric data elements, rate of outliers (extreme values) is also calculated as potential invalid values.

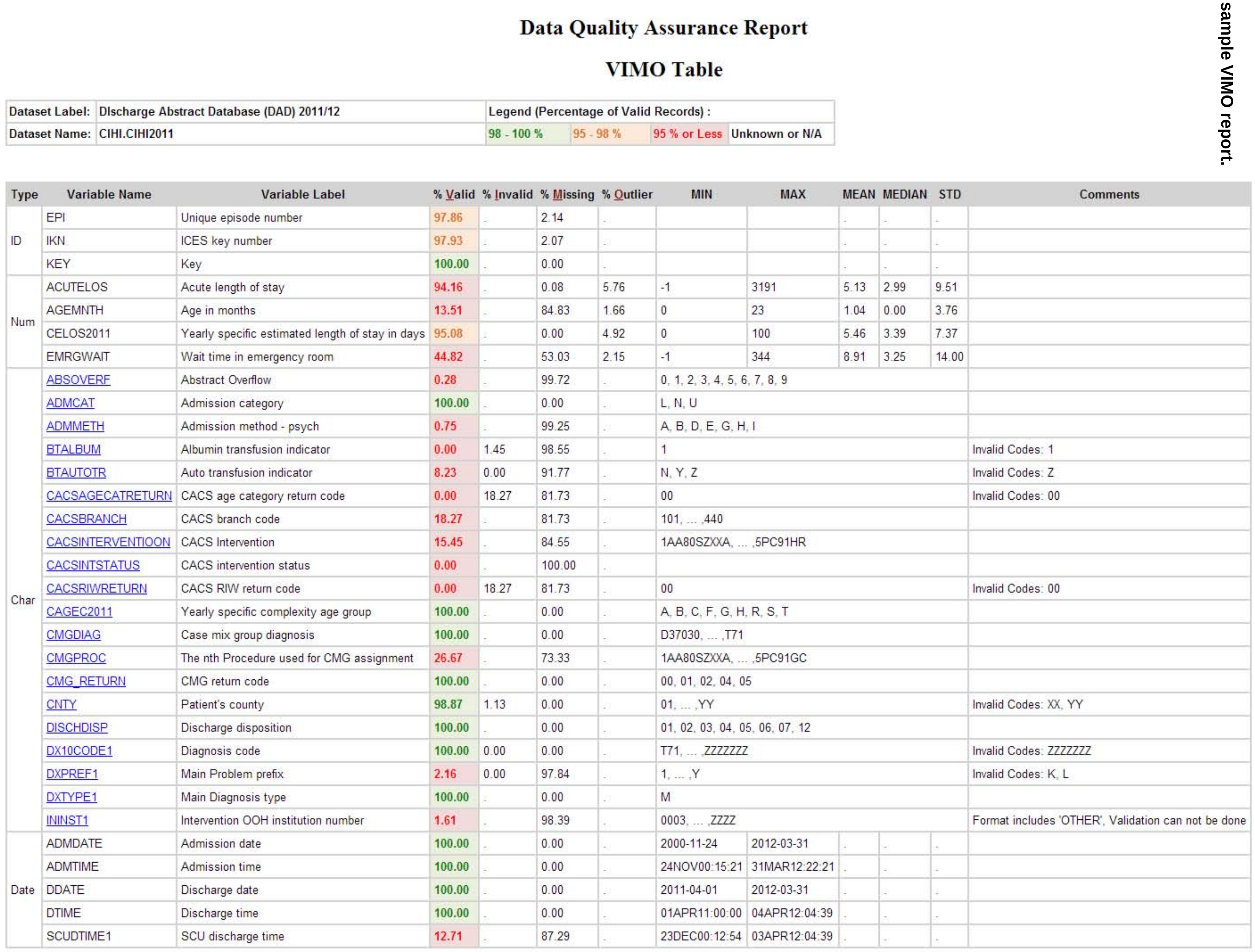
VIMO TABLE

After studying many different styles for Data Quality Reports in different organizations, we found VODIM Test Analysis Methodology as the most informative, visual and compact way of presenting “Accuracy Evaluation” of data quality.[10] VOMID which stands for Valid, Other, Default, Invalid, and Missing; is a data quality assessment conducted by the United Kingdom’s National Health Service (NHS). We modified the idea of VODIM by incorporating Canadian Institute for Health Information (CIHI) guidelines [4] for quantifying data quality indicators and proposed VIMO table which stands for Valid, Invalid, Missing and Outlier.

VIMO MACRO

VIMO macro in Data Fitness package generates a clickable HTML report, ready to publish in an internal (or external) website. Figure 3 shows an example of VIMO report. It groups all the data elements in four categories: ID variables (e.g. record id, record number, etc.), numeric variables, character variables, and date/time variables. Except the ID variables, that user should identify them while invoking the macro; VIMO is able to recognize other three types automatically. In addition to, percentage of valid, invalid, missing, and outlier values, depends on the type of variable; VIMO performs specific process and analysis for each variable to report some descriptive statistics for each variables. For character variables, instead, it reports all (or first and last level of) values. For the date/time variables, it reports the earliest and the latest value formatted in date/time. Character variables in VIMO table are hyperlinked. By clicking on their names, user will be taken to another page which presents a frequency table of the actual values and their formatted values.

Figure 3: A sample VIMO report.



Data Quality Assurance Report

VIMO Table

Dataset Label:	Discharge Abstract Database (DAD) 2011/12	Legend (Percentage of Valid Records) :			
Dataset Name:	CIHI.CIHI2011	98 - 100 %	95 - 98 %	95 % or Less	Unknown or N/A

Type	Variable Name	Variable Label	% Valid	% Invalid	% Missing	% Outlier	MIN	MAX	MEAN	MEDIAN	STD	Comments
ID	EPI	Unique episode number	97.86	.	2.14	.			.	.	.	
	IKN	ICES key number	97.93	.	2.07	.			.	.	.	
	KEY	Key	100.00	.	0.00	.			.	.	.	
Num	ACUTELOS	Acute length of stay	94.16	.	0.08	5.76	-1	3191	5.13	2.99	9.51	
	AGEMNTH	Age in months	13.51	.	84.83	1.66	0	23	1.04	0.00	3.76	
	CELOS2011	Yearly specific estimated length of stay in days	95.08	.	0.00	4.92	0	100	5.46	3.39	7.37	
	EMRGWAIT	Wait time in emergency room	44.82	.	53.03	2.15	-1	344	8.91	3.25	14.00	
Char	ABSOVERF	Abstract Overflow	0.28	.	99.72	.	0, 1, 2, 3, 4, 5, 6, 7, 8, 9					
	ADMCAT	Admission category	100.00	.	0.00	.	L, N, U					
	ADMMETH	Admission method - psych	0.75	.	99.25	.	A, B, D, E, G, H, I					
	BTALBUM	Albumin transfusion indicator	0.00	1.45	98.55	.	1					Invalid Codes: 1
	BTAUTOTR	Auto transfusion indicator	8.23	0.00	91.77	.	N, Y, Z					Invalid Codes: Z
	CACSAGECATRETURN	CACS age category return code	0.00	18.27	81.73	.	00					Invalid Codes: 00
	CACSBRANCH	CACS branch code	18.27	.	81.73	.	101, ... ,440					
	CACSINTERVENTIOON	CACS Intervention	15.45	.	84.55	.	1AA80SZXXA, ... ,5PC91HR					
	CACSINTSTATUS	CACS intervention status	0.00	.	100.00	.						
	CACSRIWRETURN	CACS RIW return code	0.00	18.27	81.73	.	00					Invalid Codes: 00
	CAGEC2011	Yearly specific complexity age group	100.00	.	0.00	.	A, B, C, F, G, H, R, S, T					
	CMGDIAG	Case mix group diagnosis	100.00	.	0.00	.	D37030, ... ,T71					
	CMGPROC	The nth Procedure used for CMG assignment	26.67	.	73.33	.	1AA80SZXXA, ... ,5PC91GC					
	CMG_RETURN	CMG return code	100.00	.	0.00	.	00, 01, 02, 04, 05					
	CNTY	Patient's county	98.87	1.13	0.00	.	01, ... ,YY					Invalid Codes: XX, YY
	DISCHDISP	Discharge disposition	100.00	.	0.00	.	01, 02, 03, 04, 05, 06, 07, 12					
	DX10CODE1	Diagnosis code	100.00	0.00	0.00	.	T71, ... ,ZZZZZZZ					Invalid Codes: ZZZZZZZ
	DXPREF1	Main Problem prefix	2.16	0.00	97.84	.	1, ... ,Y					Invalid Codes: K, L
	DXTYPE1	Main Diagnosis type	100.00	.	0.00	.	M					
	ININST1	Intervention OOH institution number	1.61	.	98.39	.	0003, ... ,ZZZZ					Format includes 'OTHER', Validation can not be done
Date	ADMDATE	Admission date	100.00	.	0.00	.	2000-11-24	2012-03-31	.	.	.	
	ADMTIME	Admission time	100.00	.	0.00	.	24NOV00:15:21	31MAR12:22:21	.	.	.	
	DDATE	Discharge date	100.00	.	0.00	.	2011-04-01	2012-03-31	.	.	.	
	DTIME	Discharge time	100.00	.	0.00	.	01APR11:00:00	04APR12:04:39	.	.	.	
	SCUDTIME1	SCU discharge time	12.71	.	87.29	.	23DEC00:12:54	03APR12:04:39	.	.	.	

%VIMO PARAMETERS AND DESCRIPTIONS

DS:	Name of Dataset
INVALIDS:	Option to turn Invalid checks ON/OFF (default value is ON)
PATH:	Location for saving the HTML VIMO report
POSTALS:	List of variables containing postal codes for validation separated by blank
FMTLIB:	SAS library name(s) containing the format catalogs (default is FORMATS)
METALIB:	SAS library name containing METADATA (default is Meta)
FREQ:	An option for turning ON/OFF the FREQ feature. This feature creates frequency tables for character variables in HTML format. These tables can be displayed by clicking on character variable names in VIMO report. Note that a subfolder called "Freq" must exist under the given &PATH (default value is ON).
EXCLUDEFREQ:	List of variables for exclusion from frequency tables, separated by blank (default value is pstlcode)
ID:	List ID variables, separated by blank. (default value is IKN)

EXAMPLES:

```
%VIMO (DS=nrs.epidemo,
            PATH=~/bkup/DQ/NRS/epidemo,
            POSTAL=postcode,
            FMTLIB=Meta,
            METALIB=Meta,
            ID=IKN epi_id
            )
```

```
%VIMO (DS=nrs.epidemo,
            INVALIDS=off,
            PATH=~/bkup/DQ/NRS/epidemo,
            FMTLIB=Meta,
            METALIB=Meta,
            ID=IKN epi_id,
            FREQ=off
            )
```

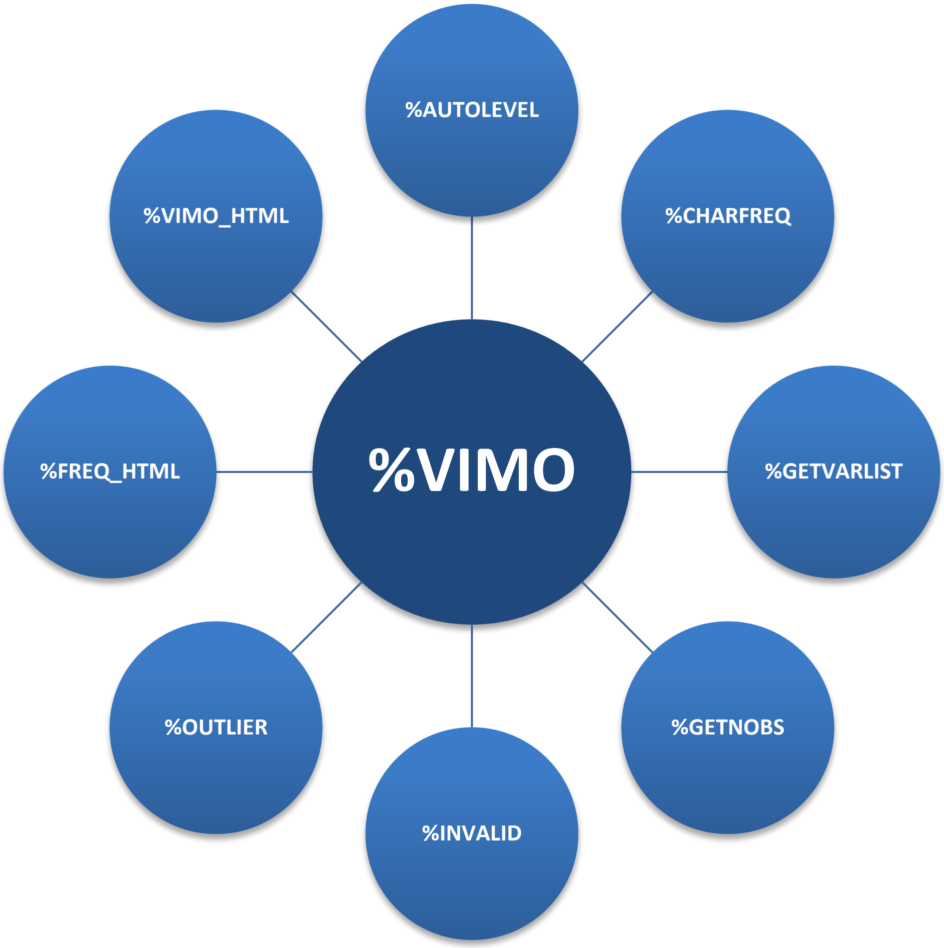
In %VIMO Different tasks such as identifying outliers, identifying invalid values, check for the postal codes, generating frequency tables, generating final html reports, and etc. have been assigned to smaller macros which are invoked by %VIMO. The diagram in figure 4 shows the relation between intermediate macros and %VIMO.

While working with extremely large datasets, it is common to deal with some sort of limitations; from a memory error while running a simple PROC FREQ to calculating median and quarters to find an interquartile range. In VIMO macro, these limitations have been taken into account. All the PROC FREQs were replaced with equivalent PROC SQL. In calculating quartiles with PROC MEANS, piecewise-parabolic (P^2) algorithm was selected:

```
PROC MEANS DATA=&LIB..&DSN NOPRINT QMETHOD=P2;
```

In order to create HTML report, VIMO macro writes CSS and HTML codes directly to a text file and incorporates the results of VIMO into that file. CSS code was used to colorize the level of validity for each variable.

Figure 4: VIMO macro’s connection with other intermediate macros



EVALUATING INTERNAL VALIDITY OF DATA

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. [7]

Stability over time can be 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 Canadian Institute for Health Information (CIHI)’s Data Quality Framework indicates [4]:

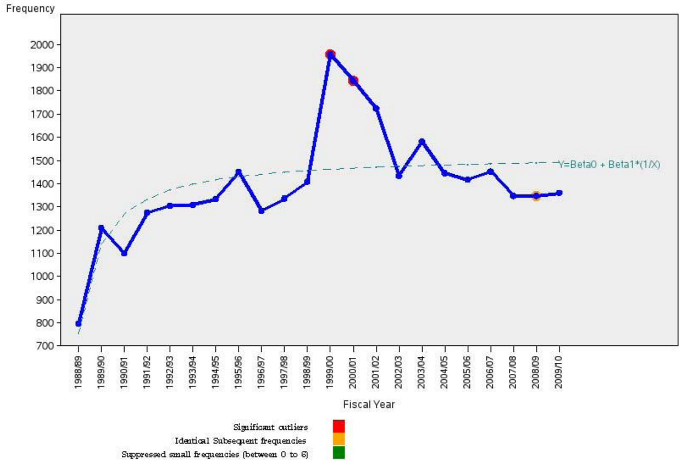
- Trend analysis is used to examine changes in core data elements over time
- Changes in methodology or inclusion/exclusion criteria should be taken into account to determine whether the observed changes were real or not
- Trend analysis includes comparisons of counts or proportions over time, as well as more sophisticated time series analysis, smoothing or curve fitting.
- Graphing data is usually particularly helpful for investigating temporal changes.
- When data is expected to naturally trend upward or downward due to policies implemented or social or economic changes, “no change” across years may also be an indication of a problem

TREND MACRO

TREND macro evaluates temporal consistency. This macro fits the following series of seven smooth lines or curves to a set of observations and can compute the mean square error (MSE) for the statistical model associated with a line or curve; this information can be used to identify the best–fit line for a set of data:

1. Simple Linear: $Y=\beta_0 + \beta_1 X$
2. Quadratic: $Y= \beta_0 + \beta_1 X^2$
3. Exponential: $Y= \beta_0 + \beta_1 \exp(X)$
4. Logarithmic: $Y= \beta_0 + \beta_1 \log(X)$
5. SQRT: $Y= \beta_0 + \beta_1 \sqrt{x}$
6. Inverse: $Y= \beta_0 + \beta_1 \frac{1}{x}$
7. Negative Exponential: $Y= \beta_0 + \beta_1 \text{Exp}(-X)$

Figure 5: A Sample TREND graph



The macro estimates studentized residuals, which are the standardized differences between observed and predicted values. Studentized residuals that are statistically significant (i.e., larger or smaller than expected) are identified. The macro also identifies repeated observations with the exact same value (indicating no change over time) and will flag these as potential coding errors.

Trend macro uses annotation method in PROC GPLOT to flag unusual points with different colors on the graph. For this purpose four different annotation datasets are created and appended to each other;PROC GPLOT then uses them to overlay the flagged point on the graph. A sample of TREND graph has been shown in figure 5.

%TREND PARAMETERS AND DESCRIPTIONS

DS:	Name of Dataset
STARTYR:	Beginning year (calendar/fiscal year, 4 digits)
ENDYR:	Ending year (calendar/fiscal year, 4 digits)
BYDATE:	Desired Date variable (Must be SAS Date variable)
BYVAR:	An optional categorical variable. If omitted only one trend analysis will be done for all the records in the dataset.
BYFMT:	An optional Format for BYVAR, if there exists any.
TIME:	Must be one of these values: FISCAL, MONTHLY, CALENDAR (default value is FISCAL)
PATH:	Physical location for storing PNG format of the graph.

EXAMPLES:

```
%TREND (DS=health.ccic_med,
            STARTYR=2003,
            ENDYR=2010,
            BYDATE=admitdt,
            BYVAR=HOSP,
            BYFMT=$CCIC_HOSP.
            PATH=~/DQ/
            )
```

```
%TREND (DS=health.ccic_med,
            STARTYR=2003,
            ENDYR=2010,
            BYDATE=admitdt,
            TIME=calendar,
            PATH=~/DQ/
            )
```

LIKABILITY MACRO

Linkability measures the ability to connect one data file to another data file using a unique subject–specific identifier.[6] In the MCHP Data Quality Framework, linkability is defined as the percentage of records that have common identifiers in two or more administrative databases. Linkability is an important data quality indicator because it determines the extent to which different databases can be used in research–specific analyses. [7]

% LINKABILITY PARAMETERS AND DESCRIPTIONS

DS:	Complete name of a dataset or the first few common characters of a series of datasets as a prefix
DSPREFIX:	If DS is a prefix then DSPRIFIX should be ON (Default value is OFF)
BYDATE:	Desired Date variable (Must be a SAS Date variable)
LINKTYPE:	The variable which contains type of linkage (or primary ID type)
STARTYR:	Beginning year (calendar/fiscal year, 4 digits)
ENDYR:	Ending year (calendar/fiscal year, 4 digits)
TIME:	Must be either FISCAL,CALENDAR (default value is FISCAL)
PATH:	Specify a location for storing HTML Linkability report.

EXAMPLES:

```
%LINKABILITY (DS=CIHI.CIHI,
                DSPREFIX=ON,
                BYDATE=DDATE,
                LINKTYPE=VALIKN,
               STARTYR=1988,
               ENDYR=2011,
               PATH=~/temp
              )
```

```
%LINKABILITY (DS= cic.cic2010,
                BYDATE= landing_date,
                LINKTYPE= link_type,
                STARTYR=1980,
                ENDYR=2011,
                TIME=Calendar,
                PATH=~/temp
               )
```

Table 1: A sample Linkability report

Death Year	IKN based on valid Health Number		Blank or other Health Number		Invalid Health Number		Total Number of Records	Linkage Rate (%)
	#	%	#	%	#	%		
2000/01	1,140,372	97.72	26,596	2.28	0	0	1,166,968	97.72
2001/02	1,152,604	97.71	27,059	2.29	0	0	1,179,663	97.71
2002/03	1,128,322	97.71	26,402	2.29	0	0	1,154,724	97.71
2003/04	1,102,213	97.7	25,987	2.3	0	0	1,128,200	97.7
2004/05	1,124,922	97.77	25,633	2.23	0	0	1,150,555	97.77
2005/06	1,125,024	97.79	25,386	2.21	0	0	1,150,410	97.79
2006/07	1,067,631	97.96	22,281	2.04	0	0	1,089,912	97.96
2007/08	1,067,828	98.02	21,616	1.98	0	0	1,089,444	98.02
2008/09	1,064,668	98	21,701	2	0	0	1,086,369	98
2009/10	1,069,416	97.99	21,934	2.01	0	0	1,091,350	97.99
2010/11	1,074,605	97.99	22,072	2.01	0	0	1,096,677	97.99
2011/12	1,097,698	97.93	23,204	2.07	0	0	1,120,902	97.93
Total	**13,215,303**	**97.85**	**289,871**	**2.15**	**0**	**0**	**13,505,174**	**97.85**

INTERPRETABILITY OF DATA

The concept of interpretability focuses on the documentation for a data file, including historical and concurrent documentation. The former refers to documentation that is maintained over time, while the latter is developed as the database is examined for inclusion in the Repository. Changes in program inclusion criteria, data collection methods, or reporting criteria may confound an analyst's or researcher's ability to identify data quality problems. [7] Most of this information can be included in a well-organized Data Dictionary. Metadata and the data quality reports are the other components of interpretability dimension.

Data Fitness package uses the same metadata which was created for evaluating accuracy of the data in order to generate a data dictionary in HTML format.

DATADIC MACRO

%DATADIC is a SAS macro under Windows which uses three sources of data to generate a complete data dictionary for a single dataset or for the entire datasets within a SAS Library. These three sources are:

1. Metadata, which is basically all the metadata created by %META (as described earlier) appended together using a data step:

```
DATA meta.metadata;
        SET meta_: ;
RUN;
```

2. A dataset version of the entire format catalog using a PROC FORMAT:

```
ORIC FORMAT LIBRARY=formats CNTLOUT=meta.formats;
RUN;
```

3. NOTES dataset: this is a dataset which includes any additional notes, comments, warnings and hyperlinks to other pages/websites for data elements (variables).Obviously, this component should be created manually. It can be either entered into an Excel spreadsheet or through a data entry interface written in Access format or any other way. The final data should be imported to SAS with the following components:

   LIBANME, NAME, NOTES, URLTITLE, URL

   All the variables are TEXT; LIBNAME and NAME are the key variables that make this dataset connected to

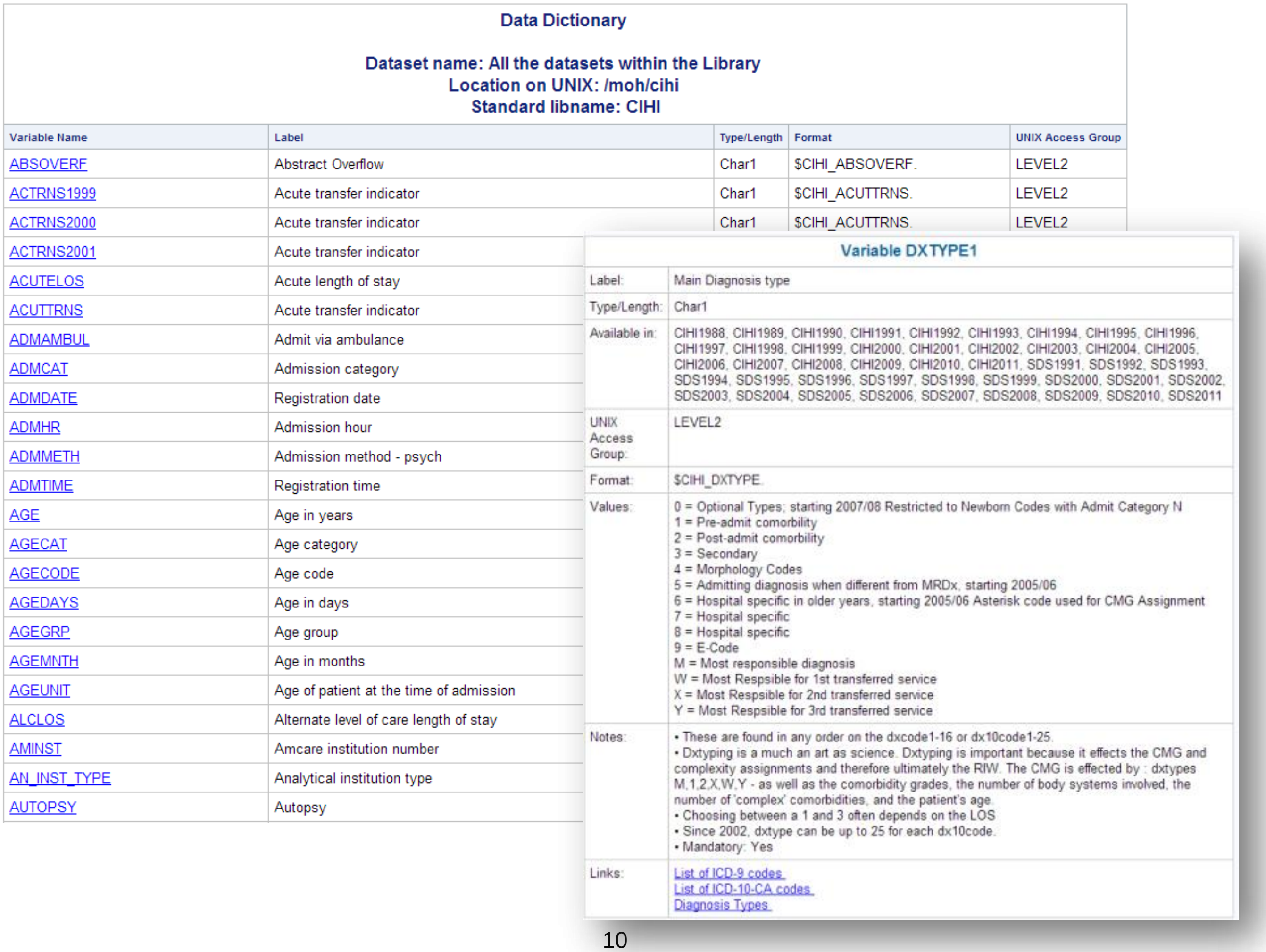
METADATA. NOTES contains a free-text note for the specific variable (NAME) within the library (LIBNAME). URLTITLE contains the text you want to be hyperlinked and URL will be the destination link.

Having these three components, %DATADIC creates an HTML website to display a data dictionary for the data.

% DATADIC PARAMETERS AND DESCRIPTIONS

LIB:	The SAS Library that a data dictionary is being created for.
DS:	A specific dataset within &LIB. If left blank, then a data dictionary will be generated for all the datasets within the &LIB
META:	Name of the METADATA. (default value is meta.metadata)
FORMATS:	Name of the FORMATS. (default value is meta.formats)
PATH:	The physical location to save the data dictionary
LOOKUPSUBDIR:	A sub-directory within the PATH to store lookup table pages. This sub-directory should be created in advance. (default value is Lookup Tables)
VARSUBDIR:	A sub-directory within the PATH to store variable pages. This sub-directory should be created in advance. (default value is Variables)
SHOWNVALUE:	The maximum number of the rows for “Values”. If number of values exceeds this number, then instead of the list of values, a link will be shown to take the user to a separate page containing the values or lookup tables (default is 20)
TITLE:	Title for the main html page.

Figure 6: A Sample Data Dictionary created by DATADIC macro



Data Dictionary

Dataset name: All the datasets within the Library
Location on UNIX: /moh/cihi
Standard libname: CIHI

Variable Name	Label	Type/Length	Format	UNIX Access Group
ABSOVERF	Abstract Overflow	Char1	$CIHI_ABSOVERF.	LEVEL2
ACTRNS1999	Acute transfer indicator	Char1	$CIHI_ACUTTRNS.	LEVEL2
ACTRNS2000	Acute transfer indicator	Char1	$CIHI_ACUTTRNS.	LEVEL2
ACTRNS2001	Acute transfer indicator			
ACUTELOS	Acute length of stay			
ACUTTRNS	Acute transfer indicator			
ADMAMBUL	Admit via ambulance			
ADMCAT	Admission category			
ADMDATE	Registration date			
ADMHR	Admission hour			
ADMMETH	Admission method - psych			
ADMTIME	Registration time			
AGE	Age in years			
AGECAT	Age category			
AGECODE	Age code			
AGEDAYS	Age in days			
AGEGRP	Age group			
AGEMNTH	Age in months			
AGEUNIT	Age of patient at the time of admission			
ALCLOS	Alternate level of care length of stay			
AMINST	Amcare institution number			
AN_INST_TYPE	Analytical institution type			
AUTOPSY	Autopsy			

Variable DXTYPE1

Label:	Main Diagnosis type
Type/Length:	Char1
Available in:	CIHI1988, CIHI1989, CIHI1990, CIHI1991, CIHI1992, CIHI1993, CIHI1994, CIHI1995, CIHI1996, CIHI1997, CIHI1998, CIHI1999, CIHI2000, CIHI2001, CIHI2002, CIHI2003, CIHI2004, CIHI2005, CIHI2006, CIHI2007, CIHI2008, CIHI2009, CIHI2010, CIHI2011, SDS1991, SDS1992, SDS1993, SDS1994, SDS1995, SDS1996, SDS1997, SDS1998, SDS1999, SDS2000, SDS2001, SDS2002, SDS2003, SDS2004, SDS2005, SDS2006, SDS2007, SDS2008, SDS2009, SDS2010, SDS2011
UNIX Access Group:	LEVEL2
Format:	$CIHI_DXTYPE.
Values:	0 = Optional Types; starting 2007/08 Restricted to Newborn Codes with Admit Category N 1 = Pre-admit comorbility 2 = Post-admit comorbility 3 = Secondary 4 = Morphology Codes 5 = Admitting diagnosis when different from MRDx, starting 2005/06 6 = Hospital specific in older years, starting 2005/06 Asterisk code used for CMG Assignment 7 = Hospital specific 8 = Hospital specific 9 = E-Code M = Most responsible diagnosis W = Most Respsible for 1st transferred service X = Most Respsible for 2nd transferred service Y = Most Respsible for 3rd transferred service
Notes:	• These are found in any order on the dxcode1-16 or dx10code1-25. • Dxtyping is a much an art as science. Dxtyping is important because it effects the CMG and complexity assignments and therefore ultimately the RIW. The CMG is effected by : dxtypes M,1,2,X,W,Y - as well as the comorbidity grades, the number of body systems involved, the number of 'complex' comorbidities, and the patient's age. • Choosing between a 1 and 3 often depends on the LOS • Since 2002, dxtype can be up to 25 for each dx10code. • Mandatory: Yes
Links:	List of ICD-9 codes List of ICD-10-CA codes Diagnosis Types

EXAMPLES:

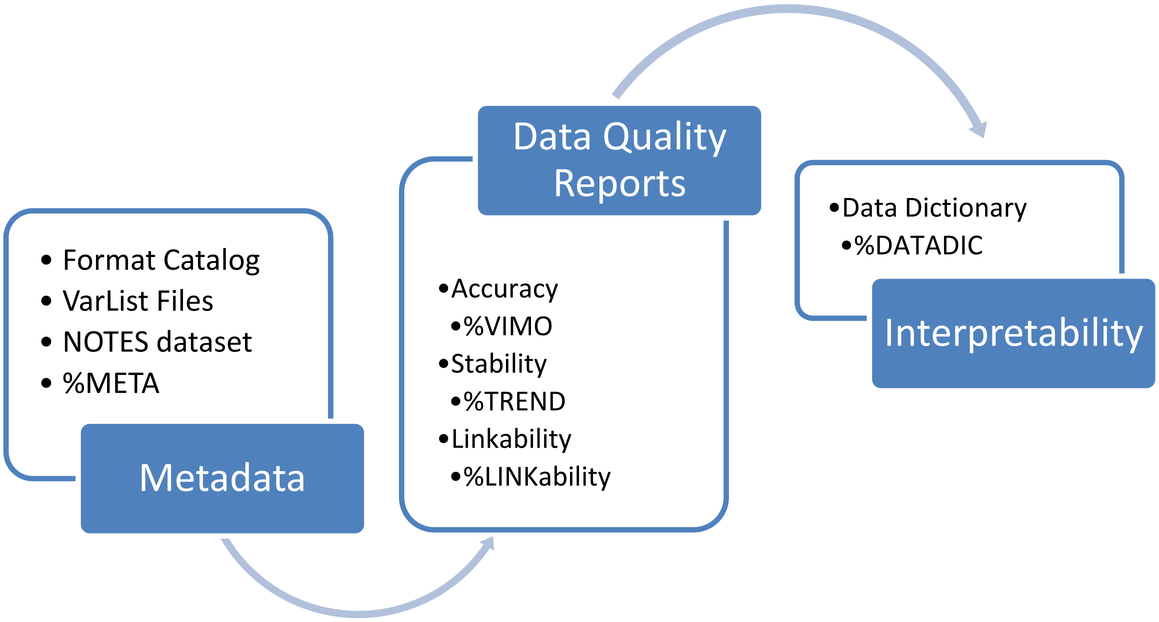
```
%DATADIC (LIB = CIHI,
          PATH =P:\Projects\Metadata\html\CIHI,
          TITLE =Discharge Abstract Data (DAD)
         )
```

```
%DATADIC (LIB = CIC,
          DS = cic2010,
          PATH =P:\Projects\Metadata\html\CIC,
          TITLE = Citizenship and Immigration Data (CIC)
         )
```

%DATADIC invokes three other macros which are %GETNOBS, %MAKELOOKUP and %MAKEPAGES.

This macro uses PROC TEMPLATE to define the style of HTML pages and then using ODS HTML statement to print the report into an HTML file. Figure 6 shows a sample Data Dictionary created by DATADIC macro.

Figure 7: The overall workflow in Data Fitness Packages



CONCLUSION

Data Fitness package provides a comprehensive set of database-specific data quality reports. It works like a system which requires all the components which are developed and maintained in a specific order. For example, without a metadata you cannot generate a complete VIMO table.

It also needs a set of standardization rules to be in place. For example, all the variables with the same library must be identical. Also, any two variables with the same name within the same library should be assigned to one single format.

This package is still under development. Some new features and functionalities will be added to the macros including “Sort by” and “Index by” information to be added to the metadata, enabling Trend macro to accept aggregated datasets and also yearly datasets, improving the output file for Linkability macro and flagging unique ID variables in VIMO.

REFERENCES

1. Australian Bureau of Statistics, “ABS Data Quality Framework”, 2009 (online document available at: Link)
2. Australian Bureau of Statistics, “Data Fitness: A guide to keeping your data in good shape”, 2009
3. Australian National Statistics Service (NSS), “Handbook”, (online document available at: Link) http://www.nss.gov.au/nss/home.nsf/NSS/7A3193EA236B927ACA25763F00096581?opendocument
4. Canadian Institute for Health Information, “The CIHI Data Quality Framework”, 2009 Edition.
5. Don Edwards, “DATA QUALITY CONTROL / QUALITY ASSURANCE” (online paper available at: Link)
6. Iron K, Manuel DG. “Quality assessment of administrative data (QuAAD): an opportunity for enhancing Ontario's health data.” Institute for Clinical Evaluative Sciences. 2007.
7. Lisa Lix, Mark Smith, Mahmoud Azimaee, et al. “A Systematic Investigation of Manitoba’s Provincial Laboratory Data”, Manitoba Centre for Health policy, December 2012.
8. Mahmoud Azimaee, “Trend Analysis: An Automated Data Quality Approach for Large Health Administrative Databases”, SAS Global Forum 2012.
9. Public Health Agency of Canada, “PHAC Data Quality Framework”, March 2009.
10. UK’s National Health Services, Data Quality Report for Independent Sector NHS funded treatment Q1 – Q2 2007/08 (online document available at: Link)
11. Ron Cody, Cody’s Data Cleaning Techniques Using SAS, SAS Inc., 2008.

ACKNOWLEDGMENTS

Author would like to thank Lisa Lix and Mark Smith at the University of Manitoba for their supervision on this project and Kathy Sykora, Nicholas Gnidziejko, and Kinwah Fung at the Institute for Clinical Evaluative Sciences for their great support and invaluable advice on development of this project.

CONTACT INFORMATION

Mahmoud Azimaee
Institute for Clinical Evaluative Sciences (ICES)
2075 Bayview Ave,
Toronto, ON, M4N 3M5
Work Phone: (416) 480 - 4055 (Ex 85266)
Fax: (416) 480 - 6821
E-mail: mahmoud.azimaee@ices.on.ca
Web: dastneveshteha.com