IQ International DQ Metrics Working Group- Healthcare Playbook

IQ International

Version 1

Contents

Section 1: Introduction and How to Use this Document

This "playbook" is a compilation of thousands of hours of real-world practitioner experience in the field of data quality specific to each domain. Volunteers that specialized in Data Quality facilitated the creation of the core of this playbook, and domain specific practitioners of DQ contributed their industry specific insights (e.g. Healthcare, Manufacturing, Telecommunications…etc.). All of the metrics and use cases developed by each IQ International industry working groups are targeted at challenges that these practitioners are facing today and solutions they've employed and recommend.

This playbook is helpful in the following ways: first, it is a vendor-neutral platform for practitioners to share methods of identifying poor DQ, then improving it, and even hopefully prevention. Secondly, as a collaborative work- it includes only the best ideas and enables DQ practitioners to educate themselves about how DQ is handled in industries they don't usually work in. This highlights truly multi-domain best practices used around the globe. Lastly, the format of this playbook is not overly technical, so all levels of the organization should feel comfortable reading it and gleaning actionable ideas to improve their own data landscapes. As such, it is the goal of IQ International to make this a very cost-effective educational resource for both individuals and organizations.

Data Quality Measurement

Determining whether data is of poor, acceptable or high quality has been problematic for many organizations. Often, data consumers will conclude that a dataset contains bad data. However, they are unable to articulate exactly what it is that makes the data "bad". Although many data quality practitioners suggest leveraging Data Quality Dimensions as a means of determining whether data is of acceptable quality, few offer a process or method for doing so.

The first section of the Conformed Dimensions of Data Quality Playbook addresses basic concepts and the context of measurement - why we measure, what we measure, when and how we measure. It is important to understand the concepts of measurement and the context in which they are applied.

Dimensions of Data Quality

As the practice of data quality management evolved, so did the concept of using "data quality dimensions" to describe and categorize specific characteristics of quality data. Perhaps the earliest well-known research in this area was done by Wang and Strong in 1996[1]. The rationale for adopting this concept was simple. If "quality" is defined as *a "*distinctive attribute or characteristic possessed by someone or something", then distinctive characteristics possessed by data could be grouped into categories and measured using related metrics.

The categories, or Data Quality Dimensions, frequently cited by data quality practitioners describe data that is *valid, timely, complete, accurate…etc.* In order to ensure that each of the dimensions of data quality, used to document the recommended DQ metrics provided in this playbook, are objective and comprehensively documented we've used the Conformed Dimensions of Data Quality (http://dimensionsOfDataQuality.com) which is an open source (Creative Commons) licensed framework based on the work of DQ authors such as, Redman, English, Loshin, McGilvray, Sebastian-Coleman…etc. The current version is provided in the appendix, but please refer to the website for the most current version.

[1] Richard Wang, Diane Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers" Internet: http://courses.washington.edu/geog482/resource/14_Beyond_Accuracy.pdf

Context of Measurement

Measurement helps with the understanding of abstract concepts. It condenses concepts into things we can understand such as numbers or representation of numbers such as graphs, charts or maps. Further, the process of measurement allows us to understand a characteristic of one thing in terms of a characteristic of another thing. The ability to compare or contrast measurements helps to simplify concepts and provides the ability to compare objects which were measured in the same way.

Next, it is important to understand the context within which data quality is being measured in order to apply the most appropriate metrics. Context relates to the environment where data currently exists such as a database or a specific database, as well as how data was created which could be as a result of data processing, data entry, machine generation or an extraction from disparate systems.

Context also relates to data structure. For example, data quality can be measured at various hierarchical levels such as the database level and at the record or attribute levels. Even data elements which are part of a specific attribute can be measured. In addition, data quality can be measured based on a statistically significant sample which is becoming more prevalent particularly for those organizations who capture massive amounts of data such as click stream data or output from medical monitoring devices.

Last, context relates to the frequency of measurement. Data quality can be measured during routine, scheduled processing such as updating or extraction procedures which is referred to as "in-line" processing. Data quality measurement can also occur periodically such as quarterly, bi-monthly or yearly. Alternatively, measuring the quality of data may occur only once to set a baseline or to determine if candidate datasets are viable.

DQ Metrics versus DQ Rules

For the purpose of alignment during the working groups we make a distinction between DQ metrics and DQ rules in that rules are the atomic Boolean identification about whether the data is as expected (documented ahead of time via business rules). DQ Metrics, aggregate up the occurrences of these DQ rules into meaningful measures (typically numeric) so that managers can evaluate trends, prioritize resources, and predict the impacts of poor DQ on business outcomes.

Typically, DQ Rules are highly specific to an organization, and therefore are harder to prescribe at an industry level. Therefore, the DQ metrics (as discussed in the next section) are a level higher than DQ rules and are usually broader and more agnostic.

Data Quality Metrics

The characteristics of a data quality dimension form the basis of a ***metric.*** Completeness, for instance, reflects both attribute (column) population and record (row) population. A metric defines ***which*** data, within a specific dimension will be measured, ***what*** will be measured about the data and ***how*** the quantitative result will be expressed such as a "count" or "percentage". Using data quality metrics is a pragmatic method for answering a very basic question – "is this data good enough to use"?

Intended use of a target dataset could be a specific initiative such as a data integration or data migration project, development of an operational data hub or data warehouse, creation of required reports or the need to demonstrate the validity of existing business metrics. Because each stakeholder group (e.g. department, partner, regulator) will use the data in different ways, we need to be able to measure a multitude of data quality aspects. We accomplish this through use of the dimensions of data quality to categorize them.

Regardless of intended purpose, the first step in selecting specific metrics is to have a clear understanding of the expectations of end-users relative to the quality of that data. End-users should be encouraged to explicitly, as possible, articulate their assumptions. These assumptions can be formulated as data quality rules such as a rule

that a field should always be populated or a more complex rule that a field should be populated only under specific conditions.

The next step is to determine how to measure the ***degree*** to which data meets business expectations. However, determining whether the current state meets expectations is the most challenging. End-users tend to find it much easier to articulate what's wrong with data rather than describe what's right about it or, specifically, the characteristics of "good" data.

Metric Thresholds

A ***metric threshold*** is a numeric representation of the acceptable limit of a metric. Without a threshold, it would be impossible to judge whether data is of acceptable quality for its intended use. Business stakeholders, who are familiar with the data and who ultimately are the end-users, should be the key decision-makers in defining metric thresholds. They inherently understand how data will be used and what is most critical in terms of data quality requirements. A conclusion can be drawn based on how well the data meets each defined threshold.
In addition, the specificity of a threshold can be increased by simply assigning a weight to each dimension as an expression of its relative value. It is imperative however, that thresholds are set *Before* metrics are calculated to control the risk of introducing any bias.

Data quality, like project management, lean/continuous improvement and similar problem-resolution jobs includes involves a cycle of identifying issues and them fixing them. DQ professionals often refer to this lifecycle (more in appendix I) and the uniquely DQ measurement types that are conducted within this cycle are as follows:

- **In-Line Processing-** Data quality is measured during one or more production processes to detect any significant changes in data.
- **Periodic Measurement-** Data quality is measured at periodic intervals such as quarterly, twice a year or annually.
- **One-Time DQ Measurement-** Data quality is measured for a single, specific point in time.

Note: For the sake of simplicity, the use case provided in this paper is a One-Time DQ Assessment.

Root Cause Analysis

Root Cause analysis (RCA) is a process designed for use in investigating and categorizing the root causes of events with safety, health, environmental, quality, reliability and production impacts. Simply stated, RCA is a tool designed to help identify not only what and how an event occurred, but also why it happened. Only when investigators are able to determine why an event or failure occurred will they be able to specify workable corrective measures that prevent future events of the type observed.

An important part of taking action on any data quality metric is to make sure time and effort is taken to find the root cause of the data quality error. This could have come from a code defect, a missed business requirement, lack of governance or data definition, or it could be just human error. Without putting plans in place to attack the root cause of the error, metrics over time will continue to erode and an organization will be put in a constant state of data remediation.

There are several helpful types of root cause analysis and specific processes to follow such as, 6Sigma & DMAIC methods are very popular and accessible. IQ International's IQCP Test preparation study guide includes works by Deming, Juran and others with helpful discussion relating to Root Cause.

Section 2: DQ Metrics in the Healthcare Field

Healthcare analytics is based on data, and data sets in particular. The purpose of health care data sets is to identify the data elements to be collected for each patient and to provide uniform definitions for common terms. When data is accurate, physicians at any practice are thoroughly informed of patient history, tendencies, previous complications, current conditions and likely responses to treatment. Data accuracy allows relevant healthcare staff to treat patients promptly and in the most effective and appropriate way possible. If generated across a variety of sources, data collection in healthcare can also encourage efficient communication between doctors and patients, and increases the overall quality of patient care.

Generally, data quality refers to the extent that the fitness for data fulfill users' expectations and suit its intended purposes. Data quality is an important issue in healthcare arena. The unique characteristics of health services, such as non-reversible services, life critical scenario, repetitive assignment of care from one provider to another along with high turnover at rapid pace increases the possibility of making errors in the related settings.

For health care organizations, data is central to both effective health care and to financial survival. Data about the effectiveness of treatments, the accuracy of diagnoses, and the practices of health care providers is crucial to organizations that strive to maintain and improve health care delivery. A single, accurate view of patient records is critical across time and physical space. Often faulty identity matchings produce inaccurate patient records. Accurate data leads to improved patient communications and operations efficiencies. Hence, assessment of data quality in regular intervals is necessary to avoid the consequences of low quality data. In fact, the quality assurance of the data in healthcare systems is an emphasis on the continuity of the quality of care, and the technologies that support clinical care need accurate and complete data. A solid Data Quality/Governance plan should be followed to implement and preserve Data Quality.

Working Group Recommended Metrics

Metrics are meaningful measurements and calculations that are used to direct, track or control processes or performance goals of an organization. There are several basic types of metrics. But, within the realm of data quality measurement, ***Data Quality Metrics*** define the specific data that is being measured and what is being

measured about it. Within the realm of data quality, metrics are used to measure and track specific aspects or features within each dimension and are generally expressed numerically. When specific features or aspects are observed to be present in data, the data is declared to be of acceptable quality.

Thresholds specific to Healthcare

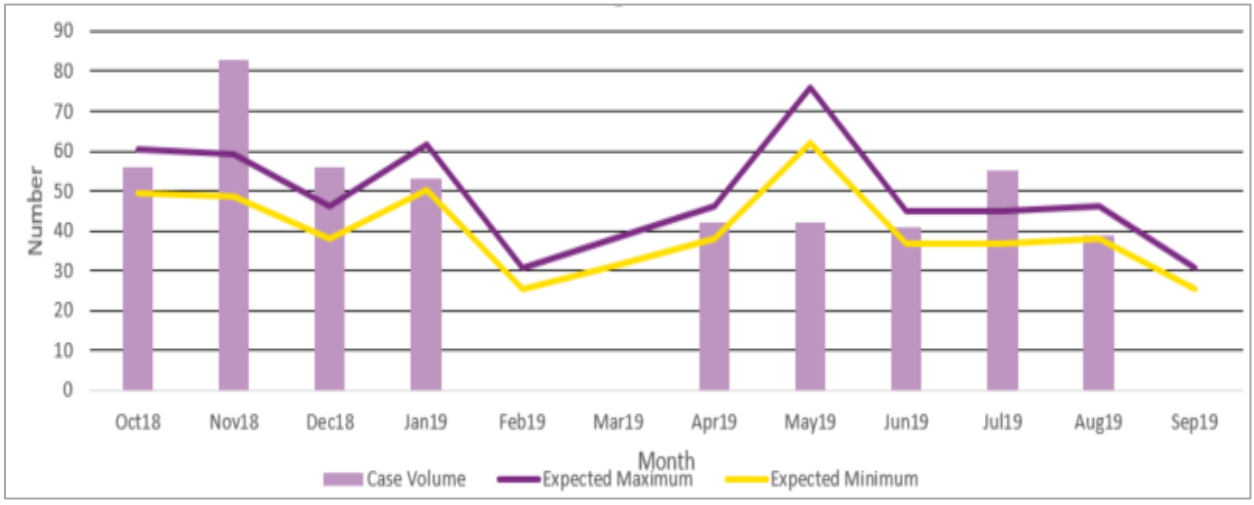
Healthcare specific thresholds aren't particularly unique, but the context may be specific to the data in which they pertain. The following are components that must be defined:

- Defining what is the desired state of specific attributes (i.e., "targets")
- Setting what is the minimum or maximum levels of quality acceptable (i.e., "thresholds", "upper" or "lower" limits)

For example, healthcare thresholds may be used to monitor volume of records specific to a patient population, or provider's capacity. Setting thresholds identifies if the data are 'good'. In the example below, when the percentage of targeted volume is greater than 75%, or when the percentage of target is less than 75%, the result is unacceptable. Highlighted values trigger action to improve compliance with set targets.

Indicator	Actual Value	Target	% of Target
Volume of Surgical Procedures 1	**265**	275	**96.4%**
Volume of Surgical Procedures 2	**0**	2	**0%**
Volume of Surgical Procedures 3	**190**	260	**73.1%**

An example of a Metric Threshold is depicted in the table below.



DQ Metrics for Healthcare

At time of publication, there are approximately **23 example metrics** created by the IQ International Healthcare DQ Metrics working group. Application of a few of these metrics are demonstrated below in the following use case.

Example Metric Using a Real-World Use case

A provider practice currently has n=439 active patients and is using an Electronic Health Record (EHR) system to capture patient administrative and clinical data. The clinical data which is relevant to a patient's care includes the following: progress notes, problem list, medications, vital signs, past medical history, immunizations, allergies, lab results, and radiology reports.

For example, prior to a patient exam, physicians review the patient's medical history and the doctor expects all relevant attributes to be populated and current. But, during review of patient data within the EHR before scheduled visits, multiple fields were observed to be missing across a substantial number of patients.

Measurements	Object of Measurement	Metric	Threshold
Measurement 1	Record population - % of populated records	# of observed populated records / Total # of expected populated records * 100	Threshold parameter set by business team
Measurement 2	Attribute population - % of populated attributes	# of observed populated attributes (Not Null Fields) / Total # of expected populated attributes * 100	Threshold parameter set by business team
Measurement 3	Attribute Existence - % of attributes present	# of observed required attributes / total # of required attributes * 100	Threshold parameter set by business team

The total number of patients is approximately 439. But, to simplify the initial analysis, only **data for 421 patients**, that have been active in the last ninety-days, was examined. The extracted data was loaded into a relational table for initial analysis. Metrics related to the ***Dimension of Completeness*** were calculated to determine the prevalence of missing data.

Table 2. Partial View of Extracted Patient Data - Recent 90 -Day Time Period

Attributes

Records

OBS #	Patient Name	Patient ID	Med Rec #	RX Name	Fill Date	Lab Test	Test Date	Test Result(s)	Partner Privacy Status
01	Fred Stokes	3442S4485	265	Vicodin	02/21/19	A1c			Full
02	Mary Jones	1184J5777	399	Lipitor		HPV	4/30/19	Negative	Partial
03	Sue Green	9991G1234	038	Amoxil	03/15/19		05/09/19	Normal	UNKNOWN
04	Chris Allison	4524A4929		Amoxil					UNKNOWN
05	Sangita Lorenz					Pap			UNKNOWN

Missing row

OBS #	Patient Name	Patient ID	Med Rec #	RX Name	Fill Date	Lab Test	Test Date	Test Result(s)	Marital Status
06	Tom Jenkins		8273						UNKNOWN

Data Quality Types Observe Above

Red: Record Population issue

Example(s): Observation row 06 (shown below table)

Green: Attribute Population

Example(s): all rows have at least one attribute missing (Observations 01…06)

Blue: Existence

Example(s): Observations 01…06, where there isn't a column to hold the Marital Status.

Calculations

Measurements	Healthcare Metric Name	Metric	Threshold
Measurement 1. Record Population Concept	Percent of Active Patients	# of patient records present/ Total # of active patient records * 100	Threshold parameter set by business team
Calculation	Record population - % of populated records	(20-1) / 20 * 100 = 95%	Threshold = 100% 5% of patients missing (error)

Record population is the measure of whether all the expected rows are in the dataset. For instance, if a physician pulls one day's worth of cases for his ambulatory care from the list of 421 active patients, he'd expect to see all of them? He met 20 patients, and expects to find all 20 rows listing the patients. From our prior example it could be said that he is missing data for one patient which is presented as ROW 6 (from Table 2 above). The goal is to have them all displayed this is why the threshold is 100%. If one is missing the healthcare metric, ***percent of active patients*** is 99.8% as further described above.

Measurements	Healthcare Metric Name	Metric	Threshold
Measurement 2. Attribute Population Concept	Percentage of Patients with Patient ID	Number of patients with Patient ID is complete = Count of patients with Patient ID NOT NULL / Total Number of Patients in dataset * 100	Threshold parameter set by business team
Calculation	# of Expected Observations N=421	(421-3)/421*100=99.5%	Threshold = 100% 0.5% Threshold not met

When doctors request additional demographic or even clinical data **all** of the patient identifiers (Patient ID) are required otherwise the files can't be linked to the original patient. In this case, all but three of the patients had incomplete Patient ID values. Because the doctor needs this information, all records are required, meaning the threshold is again 100%. In our case, three patient IDs are missing, therefore the ***percentage of patients with patient ID*** is 99.5%.

Measurements	Healthcare Metric Name	Metric	Threshold
Measurement 3. Existence Concept	Percent of New Attributes Required by Doctor	Percent of New Attributes Required by Doctor = (Number of attributes currently available to the Doctor – (Number of patient attributes available today + those desired in the future)) / (Number of patient attributes available today + those desired in the future)	Threshold parameter set by business team
Calculation	# of Expected Observations N=421	(8-(8+1))/(8+1)*100=11.1%	Threshold = 10% 10% Threshold exceeded so flagged for executive review

Although many organizations don't consider the lack of data to be a data quality issue, Information Quality seeks to ensure that all business decisions can be made using data. This includes usage of new data that may need to be collected going forward. From a task perspective, a physician's work may be incomplete if additional attributes of data aren't provided. For example, if a test for a disease has been requested by a patient, outside the knowledge of his/her partner, the revelation of the result for that test should not to be revealed when the other partner is present. In accordance with privacy laws- the physician may need to know this status during a consultation.

As shown in measurement 3, example above, the *Partner Privacy Status* attribute has not been collected (see table 2 data), so the doctor must ask the patient prior to discussion in front of a partner. Ultimately, organizations will want to limit the number of open (new attributes requested by stakeholders) to a percentage of existing attributes available. So, if either a high importance attribute (e.g. for compliance) or a large number of attributes are requested but not collected, (e.g. 10% of or more missing) then an executive report, triggered by a 10% threshold or sensitivity weighting, may be warranted.

Data Remediation

Data remediation is an activity focused on cleansing data which has been identified as corrupt or inaccurate and encompasses replacing, modifying or deleting the compromised data. There are various reasons why it is important to improve the quality of data. Typically, data remediation is driven by a specific initiative such as a data integration or migration project; development of an operational data hub or enterprise data warehouse. Regardless, the overarching driver is to ensure that data is "fit for purpose".

The data remediation process includes a "detection/validation" stage focused on identifying problematic data or validation of reported data issues. Remediation can be performed manually, with cleansing tools, or included as part of a batch production process.

Preventing Poor Data Quality

The cheapest way to improve data quality is to prevent it from occurring to begin with. It is said that the cost of remediation increases by a factor of ten each time you move to the next significant phase of the data lifecycle (e.g. conceptualization to requirements development, and then from requirements to design…etc.). The most advanced data-driven organizations define ways to ensure high quality data from point of collection.

One of the important data collection steps in health care data collection is to create data validation rules that provide logical checks on data entered into the database against predefined rules for either value ranges (e.g., systolic blood pressure less than 300 mmHg) or logical consistency with respect to other data fields for the same patient. The rules could be programmed into the data collection tool and can help data providers to test data before recording it in the database. Validation rules could be embedded into the data collection tools/interfaces so that the tool will "reject" invalid records or will create "warnings" for data providers. The benefit of performing data validation ensures that cost savings and satisfaction by end-users.

Preventive, up front logical validation checks, have been established for data collection used in Clinical Trials and Clinical registries. Some governmental institutions also use these datasets for policy development, planning, funding, wait times monitoring and quality of care evaluation, so logical checks have been built into the data collection tools. Sometimes literature review is used to define common data elements and validated patient-reported outcomes are used to ensure that the data is sufficiently generic in nature and comparable with clinical trial data. From both a clinical and practical standpoint, a multidisciplinary advisory board could help to ensure collection of key data elements in an appropriate manner.

Creating data dictionaries with data definitions and parameters (used to create data validation rules) describe both the data elements and how those data elements are interpreted. The data dictionary contains a detailed description of each variable used by the registry, including the source of the variable, coding information if used, and normal ranges if relevant.

Perhaps one of the most helpful tools to preventing poor data quality is to ensure that staff who input data are properly trained. If the reason for validity checks and seemingly useless front-end checks aren't explained to the end-users of an application, they will ignore instructions and even find ways to evade edit checks to save time and simplify their lives. One of the best ways to improve data quality is to observe current the data collection methods (without the knowledge of the staff), in a comprehensive way. Measure quality levels prior to changing business processes and training, and then measure quality levels after changes in order to document improvements based purely on improvements to training and communication.

Proactive Monitoring of Data Quality

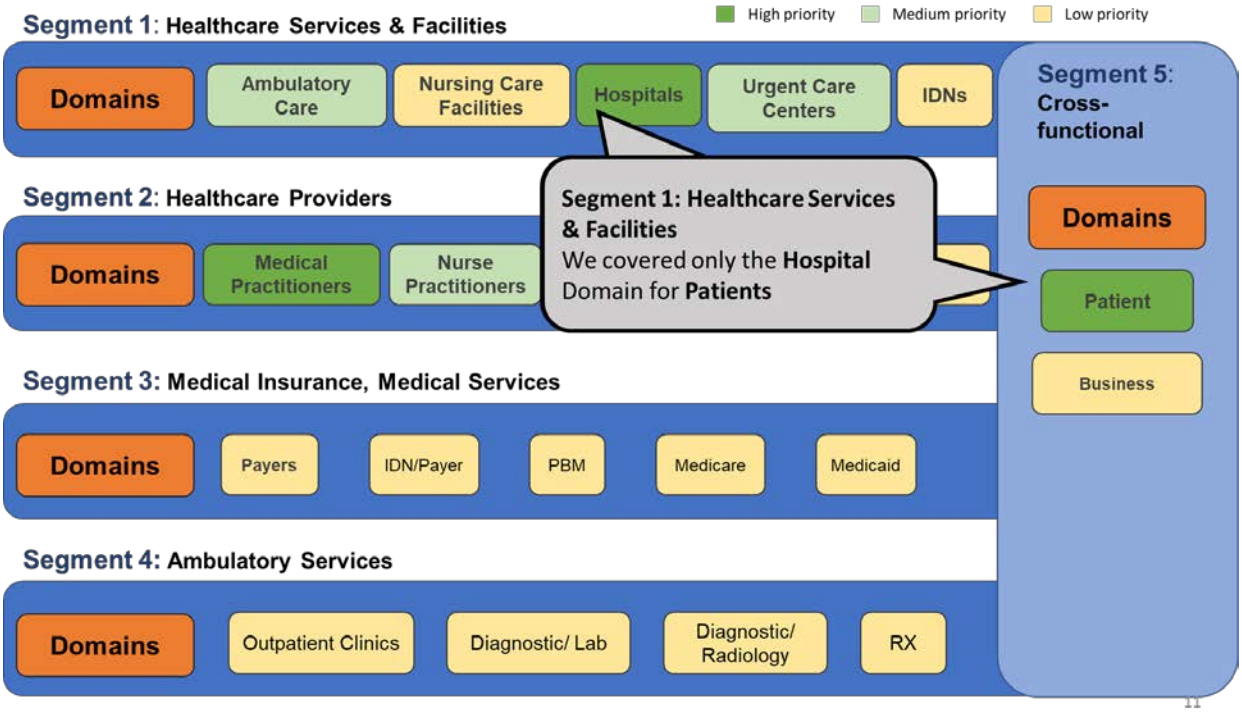
Various vendors offer software that proactively monitors data quality and delivers alerts when data quality issues are found. Most products provide a set of preconfigured, prepackaged data quality monitoring rules, as well as the functional capability of rule customization. A variety of alerting metrics can also be created which compare historical values to incoming data. Alerts can be sent proactively via email, text or to a dashboard.

The following are some typical methods of proactively identifying data quality issues but this isn't an exhaustive list. Rather we encourage you to use the Underlying Concepts of the Conformed Dimensions to help you identify the most valuable indicators of quality.

- **Validity:** We use may different types of validity to flag poor data quality. This includes use of reference data (lists of predefined values) both in forms (when data is collected) and later downstream when processed. Additionally, predefined ranges (with upper or lower limits) are useful for lab test results, vital signs, and date/time related data. During drug development, historical implications become very important.
- **Completeness:** When possible, provide sites with immediate feedback on issues such as missing or out-of-range values and logical inconsistencies. Collection of data at a later time is challenging and often impossible, so completeness controls are best imbedded into data collection processes.
- **Consistency**
    - **Logical Consistency:** Compare care metrics between different departments, doctors, remedies, geographies to identify anomalies that are likely driven by poor data quality. Note that not all anomalies are data quality issues but caused by natural changes in business/patient behavior or intentional policy decisions of leaders.
    - **Temporal Consistency:** potential sites at higher suspicion of accidental errors or even intentional errors, such as discrepancies between enrollment and screening logs, narrow data ranges, and overly high or low enrollment.

Section 3: Conclusion and Next Steps

The Healthcare DQ Metrics Working Group has only just begun it's work to define relevant and valuable metrics specific to this field. As described in the release presentation on November 29th, 2019, we have only defined metrics for the Hospital domain within the Healthcare Services and Facilities Segment as well as the Patient domain within the cross functional segment (see below). We will continue to develop metrics and look forward to additional volunteers joining the team soon. If you're interested, register here.



In conclusion, the value of a practitioner defined set of Healthcare specific DQ metrics is of immense value. We have use these (or similar) metrics and plan to add valuable metrics going forward. Please consider submitting your own metrics to the team for review and addition to the list.

The value of having this list also provides some of the following:

- Faster implementation times for Healthcare organizations worldwide
- Consistency of usage between similar organizations
- Consistency between heterogeneous industries (format and documentation of metrics will be the same for all IQ International DQ Metrics working groups).
- Enables future benchmarking between organizational units, medical facilities, companies…etc.

IQ International is proud to publish this paper, and humbly requests your input and contribution moving forward. Please direct recommendations, questions and concerns to the president of IQ International, Dan Myers (dan.myers@iqint.org).

Section 4: Acknowledgement and Appendix

Acknowledgement of DQ Working Group Members

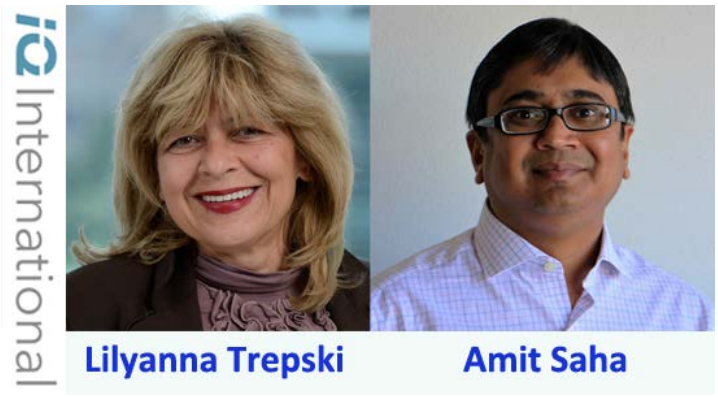
IQ International would like to acknowledge the hard work and valuable examples that other healthcare professionals contributed to the development of this playbook. The following people participated in many of the conference calls, metric formation and provision of real-world scenarios.

Melanie A. Allison (LinkedIn Profile)
Melanie was one of the strongest contributors to the DQ Metrics Working Group effort and provided a number of initial examples specific to the healthcare industry. Melanie also was helpful in defining the layout (list of descriptive information for each metric), but her signature contribution was the recommendation that we include a "Playbook" which describes the industry context with examples of metrics in order to bring the metrics alive. This contribution can't be understated given that all of the other industry working groups have adopted this deliverable format. We believe this IQ International proprietary material will be valued by members for years to come.

Sheryll Parsons (LinkedIn Profile)
Sheryll attended many of the working group conference calls, sacrificing precious personal time. She offered a great sounding board and veteran healthcare perspective. This was especially helpful during the development of the scenarios and conceptualization of the playbook. She also contributed some DQ metrics based her work in the healthcare.

Appendix 1. About the Authors and Editor of this Playbook

The principal authors who developed the first version of the IQ International DQ Metrics for Healthcare and authored the content of this report are named below. IQ International owes a debt of gratitude to them for selflessly giving time to share their experiences, thoughts for on standardization and useful content to IQ Healthcare professionals around the world. Here is a little bit about them.



Lilyanna Trpeski,

Data Quality lead at the Canadian Cancer Care Ontario (CCO) in Toronto

Lilyanna Trpeski is Data Quality lead at Cancer Care Ontario in Toronto, a provincial Agency of the Ministry of Health that Advise Ministry on various cancer and non cancer programs using breath of the health care data collected through the Ontario from Health Providers. Lilyanna works on data quality management and reporting at CCO. Lilyanna came to Canada as trained physician and epidemiologist, and worked in various environments, such as Hospitals, Canadian Institute for Health Information and Cancer Care in Health Information Field as a Lead Reporting person, on Health Research Projects and lately in Data Quality Role.

Dr. Amit K Saha,

Assistant Professor, Clinical Data Scientist in the Department of Anesthesiology at Wake Forest School of Medicine, Winston-Salem, NC

Dr. Amit K Saha working as working as Assistant Professor, Clinical Data Scientist in the Department of Anesthesiology at Wake Forest School of Medicine, Winston-Salem, NC. He is responsible to drive an optimal data architecture that efficiently and effectively presents an integrated view of clinical data for operational, financial and research processes. Also serving as a leader for data governance, best practices and optimal use of information to support the goals of the department. Supporting the research interests of faculty in the department. Previously he worked as Worked as the member of Center for Distance Health evaluation and analysis group at University of Arkansas Medical Sciences with the focus on achievement of key quality metrics to provide better patients care and outcome. His research interest includes Healthcare Informatics, Patient Care Data Management, Quality Improvement, natural language processing (NLP) and Social Computing. His long-term goal is to build a patient care experience research model using informatics approaches of data mining and predictive analytics methodology which will be of help for patient care decision support and help researchers answer key research questions. He along with his advisor Dr. Agarwal has examined various cyber campaigns such as Autism support groups' efforts to debunk misinformation campaigns. The research has resulted in the best paper award at Fifth International Conference on Social Media Technologies, Communication, and Informatics (SOTICS 2015), Barcelona, Spain and publication in other journal articles, and conference proceeding papers. Dr. Saha obtained Ph.D. from the University of Arkansas at Little Rock with outstanding dissertation recognition.

The authors can be contacted via LinkedIn.

LinkedIn for Ms. Trpeski: https://linkedin.com/in/lilyanna-trpeski-822a5313

LinkedIn Mr. Saha: https://linkedin.com/in/aksahaphd

About the Editor:
Dan Myers is the president of IQ International and acts as the coordinator for a number of the IQ International DQ Metric working groups. As editor, Mr. Myers, combined content, prepared illustrations, presentations and merged contributions from authors and other working group members into the final playbook.

Dan Myers (IQCP, MBA),
Principal Information Quality Educator at DQMatters.com and Founder of the Conformed Dimensions of Data Quality
Dan Myers is Principal Info Quality Educator at DQMatters- an eLearning organization focused on Information Quality training and consulting. Dan speaks internationally on topics of Information Quality, Blockchain, Data Governance, and Metadata Management. In previous roles, Dan has managed business intelligence teams, and lead architecture reviews of metadata management repositories, data management tools and implemented data governance programs. In his role at Farmers Insurance, he authored the Finance-led data governance policies, established the enterprise glossary and toolsets. Dan's fluency in Japanese enabled him to work in both the public and private sectors in Japan, and he speaks there annually. Dan received his MBA from the U.S.C. Marshall School of Business in 2009.

Mr. Myers can be contacted via:
Email: dan@DQMatters.com
Twitter: @kiwidankun (personal) or @dqmatters (work)
LinkedIn: https://linkedin.com/in/dan-myers-mba-iqcp-33110

Appendix 2. Tables, Records, and Fields

Tables

A database table is composed of records and fields that hold data. Each table in a database holds data about a different, but related, subject.

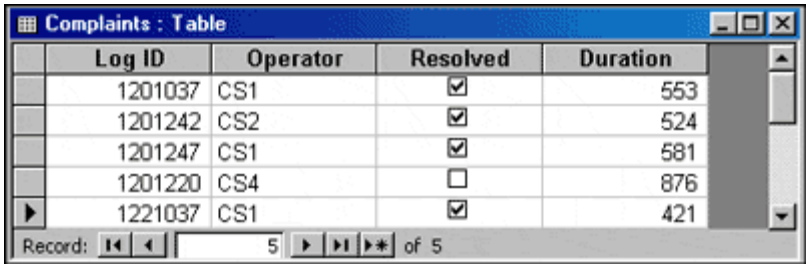


Figure 2 Database Table

Records

Data is stored in records. A record is composed of fields and contains all the data about one specific person, company, or item in a database. In this database, a record contains the data for one customer support incident report. Records appear as rows in the database table. A record for Log ID 1201242 is highlighted in Figure 3.

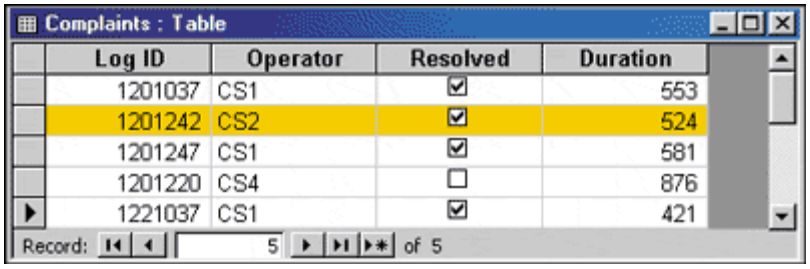


Figure 3 Records appear as rows in a database table.

Fields

The word 'field' is normally used interchangeably with 'column'. A field is part of a record and contains a single piece of data for the subject of the record. In the database table illustrated in Figure 4, each record contains four fields:

Log ID	A number assigned to this customer support incident for identification purposes
Operator	The code for the customer support operator who handled this incident
Resolved	A check box to indicate whether the incident was resolved
Duration	The time in seconds the operator spent on this incident

Fields appear as columns in a database table. Data from the Log ID field for five records is highlighted in the Figure 4.

Complaints : Table

Log ID	Operator	Resolved	Duration
1201037	CS1	☑	553
1201242	CS2	☑	524
1201247	CS1	☑	581
1201220	CS4	☐	876
1221037	CS1	☑	421

Record: 5 of 5

Figure 4 Fields appear as columns in a database table.

Column

In a relational database, a **column** is a set of data values of a particular simple type, one value for each row of the database. A column may contain text values, numbers, or even pointers to files in the operating system. Some relational database systems allow columns to contain more complex data types; whole documents, images or even video clips are examples.[3] A column can also be called an **attribute**.

Each row would provide a data value for each column and would then be understood as a single structured data value. For example, a database that represents company contact information might have the following columns: ID, Company Name, Address Line 1, Address Line 2, City, and Postal Code.

Appendix 3. Conformed Dimensions of Data Quality

Provided below is a complete list of the conformed dimensions with a definition for each. We also include the names of the Underling Concepts that list features, or aspects, related to each category. It is these features, or aspects, that can be measured and through which data quality can be quantified. For definitions of each of the Underlying Concepts themselves, please refer to the associated website, DimensionsOfDataQuality.com.

Conformed Dimensions of Data Quality (Release 4.3)

Conformed Dimension	Conformed Dimension Definition	Underlying Concepts	Non Standard Terminology for Dimension
Completeness	Completeness measures the degree of population of data values in a data set.	Record Population, Attribute Population, Truncation, Existence	Fill Rate, Coverage, Usability, Scope
Accuracy	Accuracy measures the degree to which data factually represents its associated real-world object, event, concept or alternatively matches the agreed upon source(s).	Agree with Real-world, Match to Agreed Source	Consistency
Consistency	Consistency measures whether or not data is equivalent across systems or location of storage.	Equivalence of Redundant or Distributed Data, Format Consistency, Logical Consistency, Temporal Consistency	Integrity, Concurrence, Coherence
Validity	Validity measures whether a value conforms to a preset standard.	Values in Specified Range, Values Conform to Business Rule, Domain of Predefined Values, Values Conform to Data Type, Values Conform to Format	Conformity, Accuracy, Integrity, Reasonableness, Compliance
Timeliness	Timeliness is a measure of time between when data is expected versus made available.	Time Expectation for Availability, Manual Float, Electronic Float	Currency, Lag Time, Latency, Information Float
Currency	Currency measures how quickly data reflects the real-world concept that it represents.	Current with World it Models	Timeliness
Integrity	Integrity measures the structural or relational quality of data sets.	Referential Integrity, Uniqueness, Cardinality	Validity, Duplication
Accessibility	Accessibility measures how easy it is to acquire data when needed, how long it is retained, and how access is controlled.	Ease of Obtaining Data, Access Control, Retention	Availability, Security
Precision	Precision is the measurement or classification detail used in specifying an attribute's domain.	Precision of Data Value, Granularity, Domain Precision	Coverage, Detail
Lineage	Lineage measures whether factual documentation exists about where data came from, how it was transformed, where it went and end-to-end graphical illustration.	Source Documentation, Segment Documentation, Target Documentation, End-to-End Graphical Documentation	
Representation	Representation measures ease of understanding data, consistency of presentation, appropriate media choice, and availability of documentation (metadata).	Easy to Read and Interpret, Presentation Language, Media Appropriate, Metadata Availability, Includes Measurement Units	Presentation