Accounting for Data Gaps and Uncertainty in Pipeline Risk Assessment and Integrity Assurance

Richie Joseph , Saul Chirinos, Eduardo Munoz, Pushpendra Tomar
Dynamic Risk

Organized by

https://doi.org/10.52202/078572-0091

https://doi.org/10.52202/078572-0091

Abstract

The viability of a risk model is dependent on the availability and quality of the input parameters. The data gathering effort can either push the risk model (*i. e.*, the model is built around the available data) or be pulled by the risk model (the model dictates the data needs). The former option is not available to US pipeline gas operators that must now adhere to a list of data items prescribed in 45 CFR § 192.917(c). This work produced data quality diagnostic tools, Key Performance Indicators (KPI), and an approach to incorporate data uncertainty into the different types of risk models. A Data Quality Score was developed to allow the internal stakeholders to assess the suitability of the input database before running risk, which is also useful to demonstrate the progress with the data acquisition effort. Data quality KPIs can be evaluated in many dimensions; their development was based on a review of data quality systems for scientific and engineering processes, which had many coincidences with the parameters in the guideline in API Bulletin 1178. The multidimensionality of the data uncertainty makes the definition of the associated meta data a complex task and a possible issue for the database definition. The sensitivity analysis can be leveraged to assess the data importance and minimize the amount of meta data stored. Finally, a guideline for modifying the risk model to compensate for data with high uncertainty. For probabilistic models, the distribution of the input parameter with high uncertainty needs to be modified depending on the nature of source of uncertainty.

https://doi.org/10.52202/078572-0091

Introduction

The current state of the art in pipeline risk modelling in North America is characterized by (1) being data-driven, (2) transitioning to fully quantitative models, (3) introducing the "As Low As Reasonably Practicable (ALARP) concept, (4) demonstrating risk reduction through Preventive and Mitigative Measure (PMMs) selection and (5) incorporating model and data uncertainty to the risk assessment. The last listed item, *i.e.*, incorporation of model and data uncertainty, is now a requirement for natural gas operators in the United States as per CFR 45 § 192.917(c)3.
Data Quality (DQ), defined as a measure of the current state of data, is particularly difficult to evaluate and report because of the multiple dimensions of data uncertainty. DQ limits the performance of any model or assessment; a model, being a simplified representation of reality, can produce realistic results as good as its algorithms, its assumptions and its inputs. The present work proposes transitioning from a data gathering plan feeding a risk model to a data quality management framework.

The development process of a Data Quality Management System is shown in Figure 1. The determination of the data quality Key Performance Indicators (KPIs) was based on a literature review of existing data quality systems and the dimensions of data uncertainty considered. Based on the selected KPIs, two tools were developed for data assessment: a Data Quality Scorecard and a Data Quality Dashboard. Scoring the different dimensions of data uncertainty is not enough for model correction; hence the scores were normalized by introducing the concept of Data Quality Target (*i.e.*, the realistic data quality that can be achieved) and then combined with the results of the risk model sensitivity analysis to come up with a Data Quality Risk Vector (*i. e.*, a measure of the risk introduced by the data uncertainty). Data Quality Risk was used to optimize the Probabilistic Risk Assessment (PRA) by modifying the input distributions. Data Quality should be leveraged to improve the data-gathering process. Hence, it is incorporated into a DQ management framework that updates the mapping of data inside the pipeline operator's organization and the data gathering plan for the risk model.

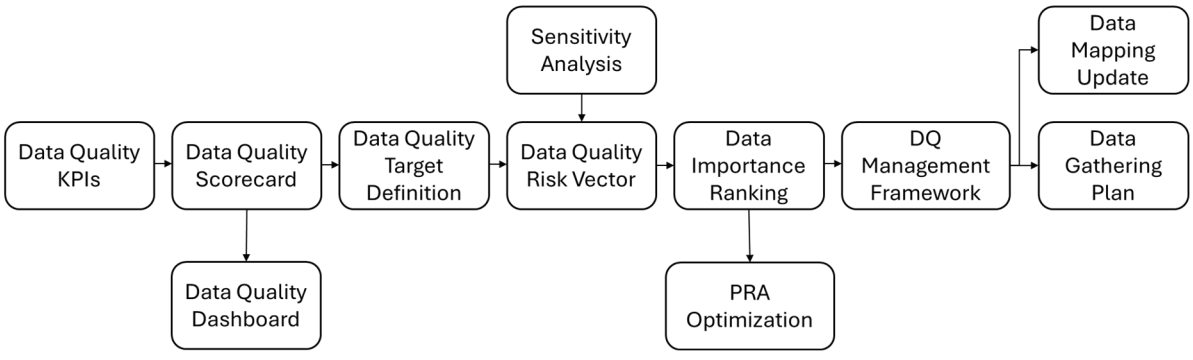


Figure 1. Data Quality Management System development process

https://doi.org/10.52202/078572-0091

Data Quality KPIs

API Bulletin 1178 Second Edition (2024)

API Bulletin 1178 Second Ed. "Integrity Data Management and Integration" §5.2.3 proposes *considering* the following dimensions of data quality:

1. Accuracy – the data represents reality.
2. Precision – the data are as exact as needed.
3. Completeness – all needed data is available.
4. Consistency – the data is free of internal conflicts
5. Timeliness – the data are as current as needed and are retained until no longer needed.
6. Granularity – the data are kept and represented at the right level of detail to meet the ends.
7. Integrity – the data are structurally sound.

The proposed DQ dimensions are based on a Standard Data Quality Classification developed by the Data Warehouse Institute (TDWI), a private initiative providing resources for data management and AI adoption [1]. The definitions implicitly introduce the notion of defining a target to satisfy; it is impractical and unrealistic to try to achieve perfect DQ in all dimensions, such as expecting perfect measurements or maintaining all information up to date. The concept of DQ target was integrated into the developed methodology. DQ management is context dependent. Hence a standard DQ system such as the one proposed by TDWI needs to be adapted to the information systems and applications of interest.

Existing DQ Systems

A review of existing data quality systems was performed for the present work and identified mature DQ methods applied to the fields of information technology, [2, 3, 4, 5, 6], business intelligence [7, 8, 9, 10], national security [11, 12], supply chain management [13], engineering asset management [14] and geosciences models [15, 16]. Those DQ methods were tailored to their specific information systems. For example, data warehouses and peer-to-peer information systems are common in IT applications. On the other hand, monolithic information systems and raw data systems (*e.g.*, printed records) are rare DQ applications. Pipeline risk models are served by distributed information systems, which are a collection of initiatives and stakeholders coordinated by a workflow (*aka*. the data gathering plan) with data originating from different repositories and with a degree of interoperability between applications and databases.

Each reviewed DQ method provides a classification of data dimensions, qualities, or attributes. A large variety of DQ dimensions have been defined for the existing DQ methods, proving the complexity of assessing data uncertainty. The developed data dimension classification is context-dependent (*e.g.,* Duplicity is extremely detrimental in supply chain systems, while analytic laboratories can limit their DQ system to only consider Accuracy and Precision). Table 1 presents the most

https://doi.org/10.52202/078572-0091

common data attributes considered in existing DQ methods, along with a general definition and alternate names. Two dimensions appear in all reviewed DQ methods: Accuracy and Precision. Timeliness is the third most frequent dimension used. Completeness is the attribute with most ambiguous definition. Some reviews list up more than sixteen possible data dimensions. Batini *et al.* [17] propose a basic set of data dimensions composed of Accuracy, Precision, Completeness, Consistency and Timeliness.

Table 1. List of Data Attributes from Literature Review

Attribute	Definition	Alternative Names
Accuracy	Extent to which the data represents reality	Exactness
Precision	Exactness of the measurement	Exactness, Exactitude
Timeliness	Certainty time-dependent data is still valid	Currency
Coverage	Percentage of non-null values for a given parameter.	Completeness
Completeness	Degree of all required values for an assessment being available.	
Conformity	Degree data is structurally sound	Integrity
Consistency	Degree data is free of internal conflicts	
Uniqueness	Measure of the duplicity of records	Duplicity
Lineage	Uncertainty introduced by derivation methods	
Subjectivity	Human interpretation	
Usability	Degree data is accessible and navigable	
Granularity	Degree data is kept at the right level of detail	

https://doi.org/10.52202/078572-0091

Accuracy and Precision

Accuracy has a syntactic component (the degree data is free of format errors) and a semantic component (the degree a measurement represents real world data). This work only considers the semantic component, though the syntactic accuracy of the risk model database is part of the preliminary checks performed by the database administrator. Precision is the exactness of a measurement, and many statistics handbooks resource to a four-pane illustration with archery targets to illustrate the difference between accuracy and precision. Sizing measurement methods used in pipeline integrity are validated based on Accuracy and Precision (*e. g.*, the unity plots used for ILI validation). Quantitative Risk Assessment (QRA) users, feeding the model with fixed value inputs, are faced with the challenge of combining these two attributes in one parameter. Should they feed the model with the most common value or the value corresponding to the worst-case scenario? In the authors' experience, most quantitative models' inputs are not consistent and are a mix of the two cases. PRA have the advantage of combining accuracy and precision in the input distributions: for example, in a normal distribution accuracy will be associated to the mean value and precision to the standard deviation.

Timeliness

Data quality deteriorates with the passage of time; for some applications it is associated with the constant improvement of data acquisition methods and the challenges associated with data and record retention. In pipeline integrity, Timeliness is associated to the time dependent threats, such as corrosion or pressure fatigue. Integrity assessments, such as ILI runs, hydrostatic testing and direct assessments, provide integrity assurance for a given moment in time. The information provided by this snapshot of the pipeline integrity, mainly flaw dimensions, is projected into the future performing an engineering assessment that is often performed offline (*e. g.*, growing crack like features by pressure fatigue). The engineering assessments used for the projections introduce a degree of error and/or uncertainty that often goes unaccounted in the risk model. Timeliness can also be used to control values that are no longer valid and should be deprecated or archived.

Completeness and Coverage

Completeness is a recurrent data attribute in DQ, yet its definition is dependent of the context. In the context of databases and data warehouses, it can be understood as the degree of values being included in the data collection or a measure of non-null values in the data collection (*e. g.*, a database where 50% of the pipe is missing manufacturing year). It can also be understood as a measure of the data sufficiency of a given parameter (*e. g.*, the pipe seam type parameter using the *ERW – Unknown* label would provide incomplete information) or the degree the information has all required parts (*e. g.*, pipe susceptibility assessment requires values for manufacturing year, pipe manufacturer, and pipe seam type, among other parameters). In this work, the two concepts were differentiated with the introduction of the Completeness and Coverage attributes, which correspond to semantic and database completeness respectively.

https://doi.org/10.52202/078572-0091

Conformity, Consistency, Uniqueness and Lineage

As mentioned before, monolithic data sources are rare, hence the input database used for the PRA should not be considered the Source of Truth, *i.e.*, the reference used to verify the format and value of a given input. In pipeline integrity, the risk model can be several steps apart from the original repositories and the Source of Truth (SoT) for a given parameter can be an intermediary repository (*e. g.*, the digitalized original construction records or an export of the model used for geotechnical studies). In many instances, a Source of Authentication (SoA) is also required to certify the data has been verified or generated according to a standard or requirement. For example, the MAOP could be looked up from an export of the operational parameters, but part of its uncertainty is associated to its method of determination or re-confirmation. This work adopted Conformity to control the degree a parameter complies with the designated format and value range, and Consistency as a control of the parameter value has been authenticated or verified (by comparing the SoT with the SoA).

Vintage pipeline information systems tend to be a mix of legacy records and data generated with modern acquisition and analysis methods, which leads to data gaps and duplicate records. Data duplicity impact is attenuated by mapping data repositories and designated the SoT. Still this work considered Uniqueness to control the existence of multiple records for a given parameter across the data repositories.

Data uncertainty introduced by derivation methods is generally oversighted, since engineering methods for pipeline integrity are generally validated for Fitness for Service (FFS) and their degree of conservatism is demonstrated. Lineage, the measure of uncertainty introduced by the derivation methods was dropped for this work but is particularly important for fracture toughness. Crack Tip Opening Displacement (CTOD) toughness is considered more representative of crack growth conditions in pipelines than Charpy V-Notch (CVN) toughness, in addition all available equations for transforming CVN to CTOD toughness introduce a different of degree of error.

Miscellaneous Data Dimensions

Not all data attributes developed for DQ systems are relevant to PRAs:

- Subjectivity is critical for SME based assessments but can be omitted for fully probabilistic models.
- Usability is related to discovery and is relevant for systems with multiple applications or users with different qualifications or expertise.
- Granularity measures how finely data is divided within a data structure and is relevant for resource assessment.
- Other dimensions such as conciseness, clarity, interactivity and security do not apply for PRAs.

https://doi.org/10.52202/078572-0091

Table 2 illustrates some of the most common DQ issues in pipeline PRAs along with the associated KPI used for control.

Table 2. Examples of Data Quality Challenges and Derived Requirements for Pipeline PRAs

Data Domain	Data Challenges	Derived Data Quality Requirements
Pipe Dimensions	Uncertain nominal dimensions	Accuracy
	Reported dimensions but no TVC source	Consistency
Pipe Mechanical Properties	YS rather than SMYS for PRA models	Accuracy and Precision
	CVN rather than CTOD toughness	Lineage
Pipe Manufacturing	Unknown Manufacturer	Completeness
	Unknown Seam Type	Completeness
Line Construction	Unknown line construction year	Completeness
	Missing weld inspection information	Completeness
Line Operation	Pressure Spectrum only available for recent years	Completeness
	MAOP reconfirmation pending	Consistency
Flaw Dimensions	Correct flaw dimensioning	Accuracy and Precision
	Representative feature growth	Timeliness
Corrosion Monitoring	Missing monitoring data for extended period	Coverage
	Surveys missing lines or segments	Coverage
Fluid properties	Average fluid temperature used for all line	Completeness
	Infrequent fluid chemistry sampling	Coverage and Timeliness

Table 3 presents the eight data dimensions adopted for the DQ method developed for this work, along with the matching data dimensions proposed in API Bulletin 1178.

https://doi.org/10.52202/078572-0091

Table 3. DQ Method Data Dimensions and Corresponding API Bulletin 1178 Dimensions

DQ Method	API Bulletin 1178
Accuracy	Accuracy
Precision	Precision
Completeness	Completeness
Coverage	
Timeliness	Timeliness
Conformity	Integrity
Consistency	Consistency
Uniqueness	
n/a	Granularity

The DQ method developed for this work uses similar KPIs for Accuracy, Precision, Consistency and Timeliness as those defined in API Bulletin1178. The DQ system uses KPIs for semantic and syntactic completeness, hence the introduction of Coverage. Consistency was supplemented with Uniqueness. Integrity was substituted by Conformity (The degree of which data align to the pre-established format and value range rather than the degree data is structurally sound). Granularity, measures how finely data is divided within a data structure, was considered relevant for resource assessment and was dropped for the developed DQ method.

Data Quality Scoring

An essential part of the data quality framework developed, a data quality control document was created, this document formalizes data flow within the data systems, as well as controls in place on all variables related to the pipeline risk models. Each variable is critically detailed, outlining its Source of Truth (SoT), data type, and priority level, ranked from one through three, *i.e.*, signifying its influence on the accuracy and reliability of the risk algorithm results. The document also identifies the data owner and specifies where the variable is stored within the database. To ensure data integrity, the document provides guidelines for assessing data quality across key dimensions, including completeness, consistency, conformity, coverage, timeliness, and uniqueness. Additionally, it links each variable and the specific risk algorithm it is used in.

A dashboard was designed to visually present the KPIs of each variable for better DQ monitoring and management. The dashboard provides an easy-to-use interface for tracking key dimensions of data quality in real time or on a regular interval. The dashboard should be allowed by this view to provide

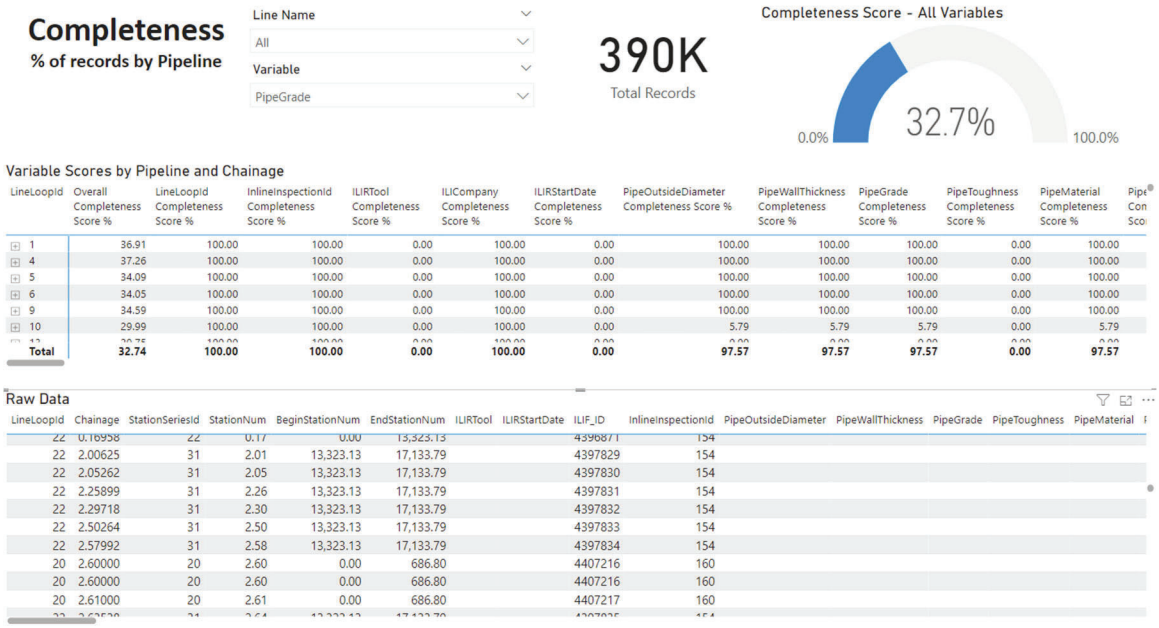
https://doi.org/10.52202/078572-0091

clear and swift identification and prioritization of data quality improvements through variable importance. The dashboard alongside the control document ultimately reinforces the importance of maintaining high data quality standards for informed decision-making, transparency, and robust risk assessments. Figure 2 and Figure 3 illustrate an example dashboard that can be developed for monitoring data quality.

Overall Scores by Pipeline

LineLoopId	Overall Completeness Score %	Overall Conformity Score %	Overall Coverage Score %	Overall Timeliness Score %	Overall Uniqueness Score %	Overall Score %
1	36.91	94.70	35.33	100.00	99.98	53.01
4	37.26	94.70	35.46	99.99	99.99	53.18
5	34.09	94.70	32.30	66.67	99.72	50.72
6	34.05	94.70	32.22	93.66	100.00	50.84
9	34.59	94.70	32.78	66.67	100.00	51.08
10	29.99	94.70	28.19	3.86	99.99	47.60
11	28.65	94.70	26.67	0.00	100.00	46.56
13	29.75	94.70	27.79	0.00	99.99	47.36
14	30.27	94.70	28.33	0.00	72.73	47.15
16	35.17	94.70	33.36	66.67	100.00	51.49
17	34.83	94.70	33.47	66.67	100.00	51.40
Total	32.74	94.70	30.89	92.00	99.94	49.88

Figure 2. Data quality dashboard summary



Variable Scores by Pipeline and Chainage

LineLoopId	Overall Completeness Score %	LineLoopId Completeness Score %	InlineInspectionId Completeness Score %	ILIRTool Completeness Score %	ILICompany Completeness Score %	ILIRStartDate Completeness Score %	PipeOutsideDiameter Completeness Score %	PipeWallThickness Completeness Score %	PipeGrade Completeness Score %	PipeToughness Completeness Score %	PipeMaterial Completeness Score %	Pipe Con Sco
1	36.91	100.00	100.00	0.00	100.00	0.00	100.00	100.00	100.00	0.00	100.00	
4	37.26	100.00	100.00	0.00	100.00	0.00	100.00	100.00	100.00	0.00	100.00	
5	34.09	100.00	100.00	0.00	100.00	0.00	100.00	100.00	100.00	0.00	100.00	
6	34.05	100.00	100.00	0.00	100.00	0.00	100.00	100.00	100.00	0.00	100.00	
9	34.59	100.00	100.00	0.00	100.00	0.00	100.00	100.00	100.00	0.00	100.00	
10	29.99	100.00	100.00	0.00	100.00	0.00	5.79	5.79	5.79	0.00	5.79	
Total	32.74	100.00	100.00	0.00	100.00	0.00	97.57	97.57	97.57	0.00	97.57	

Raw Data

LineLoopId	Chainage	StationSeriesId	StationNum	BeginStationNum	EndStationNum	ILIRTool	ILIRStartDate	ILIF_ID	InlineInspectionId	PipeOutsideDiameter	PipeWallThickness	PipeGrade	PipeToughness	PipeMaterial
22	0.16958	22	0.17	0.00	13,323.13			4396871	154					
22	2.00625	31	2.01	13,323.13	17,133.79			4397829	154					
22	2.05262	31	2.05	13,323.13	17,133.79			4397830	154					
22	2.25899	31	2.26	13,323.13	17,133.79			4397831	154					
22	2.29718	31	2.30	13,323.13	17,133.79			4397832	154					
22	2.50264	31	2.50	13,323.13	17,133.79			4397833	154					
22	2.57992	31	2.58	13,323.13	17,133.79			4397834	154					
20	2.60000	20	2.60	0.00	686.80			4407216	160					
20	2.60000	20	2.60	0.00	686.80			4407216	160					
20	2.61000	20	2.61	0.00	686.80			4407217	160					

Figure 3. Data quality dashboard extract for completeness

https://doi.org/10.52202/078572-0091

Data Quality Risk

The eight KPIs selected for the DQ system (Accuracy, Precision, Completeness, Coverage, Timeliness, Conformity, Consistency and Uniqueness) cannot be used directly to modify the probabilistic risk model and compensate for data uncertainty. The developed DQ method picked up and adapted the concepts of Data Quality Target and Data Risk Vector from a risk-based method to quantify the impact of data uncertainty in corporate governance systems [18].

Data Quality Target

The definitions of data dimensions in API Bulletin 1178 refer to the needed data quality rather than the best possible quality. It would be unrealistic and impractical to expect the highest scores in all eight KPIs for all parameters. For example, it would be unrealistic to expect full coverage of flaw detection and sizing using ILI tools in a pipeline system with sections that cannot be made ILI piggable; a realistic expectation would be to maximize the fraction of ILI piggable sections and have flaws reported for all piggable sections. Many operators settle for CVN toughness while others are generating CTOD toughness values; the target data quality for that parameter would be different for the two cases.

The Data Quality Target consists in setting the optimal KPI scores for each parameter: Figure 4 illustrates this concept with a radial plot; in the illustrative case the target scores for Accuracy, Precision, Completeness and Coverage are not maximized yet they are satisfactory for the risk model.

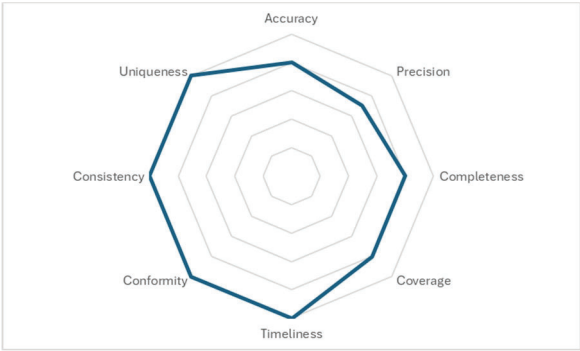


Figure 4. Data quality target for a parameter (illustrative)

Figure 5 illustrates sample KPI scores compared to the DQ target in a radial plot; the actual KPI scores would always form a polygon contained within the polygon defined by the DQ target.

https://doi.org/10.52202/078572-0091

Figure 5. Data quality target and KPI score for a parameter (illustrative)

The definition of the DQ Target allows the normalization of the KPI scores per the following equation:

$$QS_i = \frac{S_i}{T_i}$$

Where,

QS_i is the Quality Score for KPI(i) with a value between (0;1);

S_i is the Score for KPI(i);

T_i is the target for KPI(i).

The normalized Quality Score can be combined with the results of the sensitivity analysis to rank the impact of data uncertainty. Sensitivity analysis determines a Sensitivity Factor for each parameter that is a measure of the impact of the input parameter upon the model output, as illustrated in Figure 6.

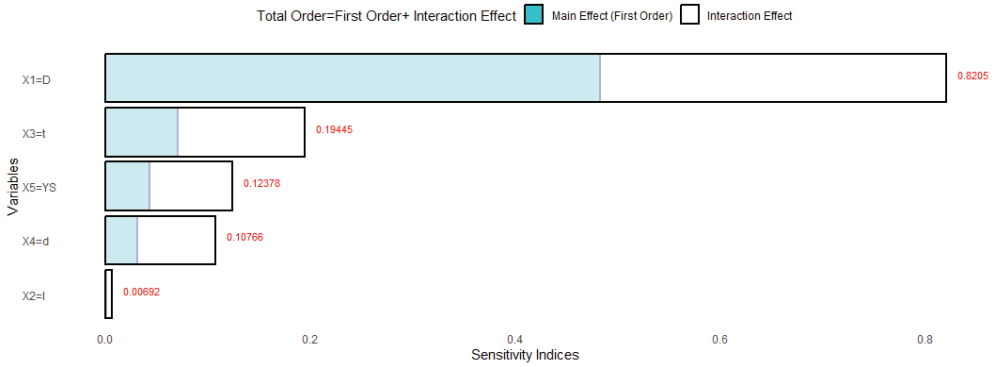


Figure 6. Sample sensitivity factors for Mod B31G [19]

https://doi.org/10.52202/078572-0091

The developed DQ method proposes determining Data Quality Risk, the impact of data uncertainty upon the model output, as follows:

$$DQR_i = (1 - QS_i) \cdot SF$$

Where,

DQR_i is Data Quality Risk for KPI(i), with a value between (0;1);

QS_i is the Quality Score for KP(i);

SF is the Sensitivity Factor for the input parameter.

DQR_i is a measure of the importance of data uncertainty and will tend to the upper limit of 1 when (1) the data quality is low and (2) the sensitivity factor is high.

The above relationship defines a Data Risk Vector as follows:

$$DQR = \begin{bmatrix} (1 - QS_{Accuracy}) \\ (1 - QS_{Precision}) \\ (1 - QS_{Completeness}) \\ (1 - QS_{Coverage}) \\ (1 - QS_{Timeliness}) \\ (1 - QS_{Conformity}) \\ (1 - QS_{Consistency}) \\ (1 - QS_{Uniqueness}) \end{bmatrix} \cdot SF = \begin{bmatrix} DQR_{Accuracy} \\ DQR_{Precision} \\ DQR_{Completeness} \\ DQR_{Coverage} \\ DQR_{Timeliness} \\ DQR_{Conformity} \\ DQR_{Consistency} \\ DQR_{Uniqueness} \end{bmatrix}$$

Where DQR is the Data Risk Vector for a given input parameter.

Integrating Data Uncertainty to Probabilistic Risk Models

The Data Risk Vector is defined for each input parameter and ideally should be used to compensate for data uncertainty by adjusting the distribution of the input parameter. The present section assumes a PRA input with a normal distribution, with mean value α and standard deviation σ.

Accuracy and Precision uncertainty is already incorporated for flaw sizing, since ILI validation is based in bounding both dimensions. Figure 7 shows four unity plots for flaw sizing using an ILI tool that illustrate how Accuracy and Precision should be considered to correct the flaw size input distribution used as an input. Case (a) corresponds to an ideal ILI tool inspection with all field verified flaw measurements within the specified toll error tolerance (10% WT) with no obvious under or over calling; under the unity plot a corresponding normal distribution for a flaw is illustrated. Cases (b) and (c) would correspond to a valid ILI inspection with the tool under calling and overcalling the flaw size respectively. Finally, case (d) would correspond to a ILI inspection with excessive scatter (less than 80% of the readings outside the 10% band).

https://doi.org/10.52202/078572-0091

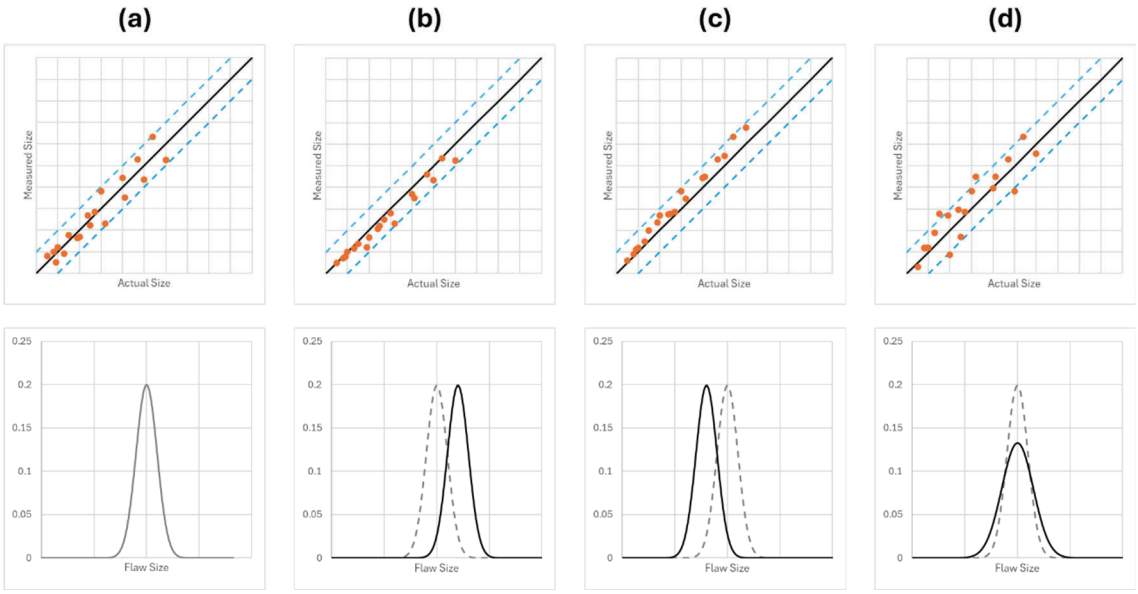


Figure 7. Illustration of precision and accuracy corrections for flaw sizing with ILI tool. (a) ideal validated run, (b) under calling tool, (c) over calling tool, (d) excessive scatter.

Case (a) in Figure 7 corresponds to the most flaw sizes used in Probability of Exceedance analysis for PRAs: a normal distribution is considered with a men value, α, equal to the measured value and a standard deviation, σ, equal to in-third the specified tool tolerance. No correction for the mean value (i. e., the tool's Accuracy) is considered in case (a). A correction to the measurement's Accuracy for when the tool under or overcalls, can be introduced by decreasing or increasing the men value, as shown in cases (b) and(c). Finally, a correction to the measurement's Precision is to be introduced by modifying the standard deviation, as illustrated in case (d) in Figure 7. The general equation for the correction due to Accuracy and Precision uncertainty would be as follows:

$$\alpha_{Corrected} = (1 + \varepsilon_{Accuracy}) \cdot \alpha$$

and

$$\sigma_{Corrected} = (1 + \varepsilon_{Precision}) \cdot \sigma$$

Where:

$$\varepsilon_{Accuracy} \propto DQR_{Accuracy}$$

and

$$\varepsilon_{Precision} \propto DQR_{Precision}$$

Though Accuracy and Precision are generally compensated for input distributions. The effect of Timeliness is often disregarded: flaw sizing projection into the future, (e. g., crack growth due to pressure cycling and wall loss due to corrosion) are often performed offline (i. e. not within the risk model) and using deterministic methods with conservative assumptions developed for fitness-for-

https://doi.org/10.52202/078572-0091

service. To be consistent with the PRA, flaw growth/projection should also be a probabilistic analysis. Let's considered a common scenario; the pressure spectrum for a limited period for a location upstream from the flaw (the compression station outlet) is to be used for pressure fatigue analysis. The analysis for fitness-for-service would assume the pressure spectrum is representative for all the pipeline operation since the flaw was characterized and that the pressure cycling at the compressor station outlet is more severe than what is experienced at the flaw location. The equivalent pressure fatigue analysis for a PRA should move away from conservative assumptions and perform a Monte Carlo analysis with representative distributions for all inputs; this is not currently a common practice. The question remains, as what type of correction would need to be considered to compensate for the projection (Timeliness) of a flaw size using deterministic analysis. Figure 8 illustrates the two options available for a flaw distribution determined by inspection and then growth by deterministic methods (i. e., the mean value is increased but the distribution shape remains unchanged).

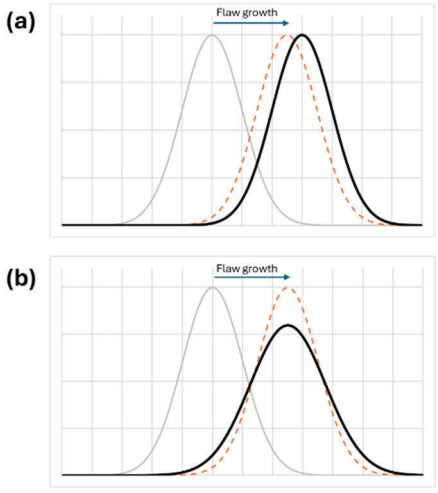


Figure 8. Illustration of timeliness correction options: (a) mean correction and (b) standard deviation correction

Case (a) in Figure 8 would correspond to a mean vale correction; the distribution can be shifted to higher or lower values. Case (b) corresponds to a standard deviation correction by increasing the scatter of the resulting flaw sizes. As previously mentioned, the deterministic analysis is required to be conservative, and a shift to higher values would exacerbate the level of conservativeness of the results. On the other hand, it would be difficult to determine the decrease in mean value required to compensate for the level of conservativeness introduced by the deterministic analysis. Hence, the method illustrated in case (a) is inappropriate to compensate for Timeliness. The correction illustrated in case (b) represents better how data uncertainty propagates in linear methods. Hence, the proposed correction for Timeliness uncertainty for PRAs is the following:

$$\sigma_{Corrected} = (1 + \varepsilon_{Timeliness}) \cdot \sigma$$

Where:

$$\varepsilon_{Timeliness} \propto DQR_{Timeliness}$$

https://doi.org/10.52202/078572-0091

Default Distributions for PRAs

The adequate application of the Completeness and Coverage KPIs require differentiating real values (generated through a data acquisition method) and default values. Default values in deterministic models tend to represent the worst-case scenario[1]; for example, in FFS analysis when the nature of a pipe seam is uncertain, the minimum CVN toughness values specified by PHMSA are adopted. The default values determined for deterministic methods and quantitative models are often adopted for PRAs; this is not adequate and leads to hybrid risk models running on a mix of fixed value and distribution inputs. Compensating for Completeness and Coverage requires developing default distributions representative of the worst-case scenario. In the case of material toughness, the adoption of the default CTOD toughness distribution in API 1178 Annex E would be more appropriate than the adoption of fixed toughness values when running PRAs. It's important to note, some authors have proposed a correction to the distribution corresponding to the most probable case [REF]; this approach seemed complicated for the nature of pipeline systems. Hence the preference to determine default distributions corresponding to the worst case based on the vast guidance available for FFS analysis.

Synthetization

The DQ methodology presented here in would seem overwhelming with eight KPIs and the requirement to define target for each dimension and for each parameter. The amount of meta-data required to be stored seems excessive. In practice, many simplifications to the general DQR equations occurred during the application of the proposed methodology. For example, it has been mentioned that Lineage was originally considered as a KPI, but during the application of the methodology, it was found to be relevant only to material toughness; it was then decided to drop Lineage and document the inherent limitations of using CVN toughness distributions. DQ method users should also be reminded that the determination of the Sensitivity Factor and the DQR vector for each parameter should be leveraged to concentrate in the relevant parameters: In general, pipe dimensions and material properties have been found to be the most impactful parameters for Likelihood of Failure (LoF) assessment.

Consistency is relevant to US gas operators due to the requirements for TVC files but might not be relevant for operators in other parts of the world. In addition, other simplifications can be done for mature information system; Uniqueness was considered not to be an issue in some applications.

Depending on the reviewer and application, the recommended minimum KPIs to consider are Accuracy, Precision, Completeness and Timeliness, with Coverage often added to this shortlist. The authors agree these five KPIs are the minimum set of dimensions to consider for pipeline PRAs.

[1] Some QRA users use two sets of default values corresponding to the worst case and the most probable case and generate two separate risk runs.

https://doi.org/10.52202/078572-0091

A Framework for Data Quality Management.

In the US, the implementation of DQ methods for PRAs is currently driven by requirements from the Federal Agencies; if the impact of the DQ methods is limited to the risk assessment, then it is equivalent to an audit [17]. Incorporating the DQ method into the continuous improvement cycle creates a framework capable of addressing the operational issues impacting DQ. Data Risk determination contributes to identifying the parameters with low DQ and their KPI or data dimensions to improve.

A past work on sensitivity analysis highlighted how its results can help the data gathering efforts by determining the inputs with highest impact on the model outcome [19]. The developed DQ management methods supplement those results by providing actual metrics of the data adequacy and determining the actual information issues that need to be addressed. Figure 9 presents a schematic representation of the continuous improvement cycle developed around DQ management.

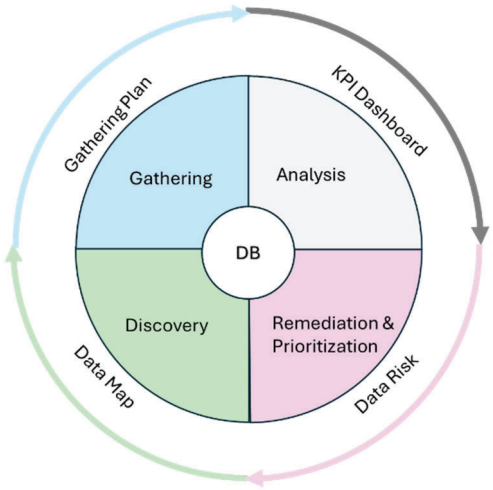


Figure 9. Data Quality Management Cycle for PRA

The four stages defined for DQ management consist in calculating and presenting the eight DQ KPIs with the help of the tools developed. The KPIs are further processed to determine Data Quality Risk and determine the issues with the current information (the database feeding the risk model). Data Risk can be used to propose remediating actions for the next data gathering effort in two categories: data mapping (determine or update the SoT and SoA) and the updated data gathering plan with the prioritized items to collect. *Shein et al.* have developed a detailed data quality framework for asset management that can be adopted for future iterations of the continuous improvement cycle [20].

https://doi.org/10.52202/078572-0091

Conclusions

A Data Quality Management framework was developed for pipeline risk assessments, in alignment with the proposed methods in API Bulletin 1178. The method scores eight KPIs (Accuracy, Precision, Completeness, Coverage, Timeliness, Conformity, Consistency and Uniqueness) with the help of a scorecard and a Dashboard.

A method based on Data Quality Risk was developed to compensate for data uncertainty in Probabilistic Risk Assessments. The method combines the score relative to a preset target for each data dimension with the sensitivity factor determined for a given input parameter. The resulting Data Quality Risk for each dimension can be used for the correction of the input distribution required to compensate for data uncertainty.

A continuous improvement cycle around data quality management is proposed to steer all data related initiatives related to pipeline integrity. The developed system is adequate for detecting data issues and prioritize the remediating data gathering actions.

https://doi.org/10.52202/078572-0091

References

[1] The Data Warehouse Institute, "Data and Analytics Education and Research," 01 01 2025. [Online]. Available: https://tdwi.org/Home.aspx.

[2] J. You, S. Lou, R. Mao and T. Xu, "An improved FMEA quality risk assessment framework for Enterprise Data Assets," *Journal of Digital Economy,* no. 1, pp. 141-152, 2022.

[3] E. A. M. Borglund, "RQAM: A recordkeeping quality assessment model," *International Journal of Information Quality,* vol. 1, no. 3, pp. 326-344, 2007.

[4] Z. Yanjun, "Component comparison based information quality," *International Journal of Information Quality,* vol. 1, no. 3, 2007.

[5] L. Berti-Équille, I. Comyn-Wattiau, M. Cosquer, Z. Kead, S. Nugire, V. Peralta and S. Si-Saïd Cherfi, "Assessment and analysis of information quality: a multidimensional model and case studies," *International Journal of Information Quality,* vol. 2, no. 4, 2011.

[6] N. Zellal and A. Zaouia, "A Measurement Model for Factors Influencing Data Quality in Data Warehhouse," *IEEE Transactions,* no. 78-1-5090-0751-6/16, 2016.

[7] J. Corcoran and M. Scott, "Measuring information quality and success in business intelligence and analytics: key dimensions and impacts," *International Journal of Information Quality,* vol. 4, no. 2, 2017.

[8] P. Woodall, M. Oberhofer and A. Borek, "A classification of data quality assessment and improvement methods," *International Journal of Information Quality,* vol. 3, no. 4, 2014.

[9] P. Woodall, B. Alexander and A. K. Parlikad, "Evaluation criteria for information quality research," *International Journal of Information Quality,* vol. 4, no. 2, 2016.

[10] J. Wang, Y. Liu, P. Li, Z. Lin and S. Sindakis, "Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality," *Journal of the Knowledge Economy,* no. https://doi.org/10.1007/s13132-022-01096-6, 2022.

[11] J. Thomson, E. Hetzler, A. MacEachren, M. Gahegan and M. Pavel, "A Typology for Visualizing Uncertainty," *Visualization and Data Analysis,* no. Proc. of SPIE-IS&T, 2005.

[12] R. Mead and S. Kenett, "Development of assurance techniques for information quality on technical advice," *International Journal of Information Quality,* vol. 3, no. 3, 2014.

[13] R. Silvola, J. Harkonen, O. Vilppola, H. Kropsu-Vehkapera and H. Haapasalo, "Data quality assessment and improvement," *International Journal of Business Information Systems,* vol. 22, no. 1, 2016.

[14] V. Masayna, A. Koronios, J. Gao and M. Gendron, "Data Quality and KPIs: A Link to be Established," in *The 2nd World Congress on Engineering Asset Management (EAM) and The 4th International Conference on Condition Monitoring*, Harrogate, United Kingdom, 2007.

[15] V. Simard, M. Rönnqvist, L. Lebel and N. Lehoux, "A Method to Classify Data Quality for Decision Making Under Uncertainty," *Journal of Data and Information Quality,* no. https://doi.org/10.1145/3592534, 2023.

[16] B. S. Daya Sagar, F. Agterberg and Q. Cheng, Handbook of Mathematical Geosciences, Springer Open, 2018.

[17] C. Batini, C. Capiello, C. Francalanci and A. Maurino, "Methodologies for Data Quality Assessment and Improvement," *CM Computing Surveys, Vol. 41, No. 3, Article 16,* July 2009.

[18] A. Borek, A. K. Parlikad, P. Woodall and M. Tomasella, "A Risk Based Model for Quantifying the Impact of Information Quality," *Computers in Industry (DOI:10.1016/j.compind.2013.12.004),* 2013.

[19] H. Fateminia and E. Munoz, "Sensitivity Analysis in the Oil and Gas Pipeline Risk Assessment:," in *Pipeline Pigging and Integrity Management Conferences*, Houston, 2024.

[20] S. Lin, J. Gao and A. Koronios, "Validating a Data Quality Framework in Engineering Asset Management," *ACIS 2006 Proceedings,* vol. 75, 2006.

https://doi.org/10.52202/078572-0091