The following is an excerpt from the new publication:
Pipeline Risk Assessment:
The Definitive Approach and Its Role in Risk Management

This 550+ page book presents the definitive approach to assessing risks from pipelines—an approach that overcomes the limitations of previous methodologies. The methodology detailed here is a practical, easy-to-apply technique for measuring risks associated with operating any type of pipeline in any environment. Management of those risks is then optimized by proper utilization of the assessment results, as is fully discussed herein.

For more information, visit www.pipelinerisk.net.

3.7 VERIFICATION, CALIBRATION, AND VALIDATION

Given enough time, a risk assessment can be proven by comparing predicted pipeline failures against actual. This is the basis of the testing of the risk assessment as a diagnostic tool, as discussed elsewhere. Pipeline failures on any specific system are usually not frequent enough to provide sample sizes sufficient to test the assessment performance. In most cases, initial examination of the assessment is best done by ensuring that risk estimates are consistent with all available information (including actual pipeline failures and near-failures) and consistent with the experiences and judgments of the most knowledgeable experts. The latter can be at least partially tested via structured testing sessions and/or model sensitivity analyses (discussed in Chapter 3.7.4 SME Validation and Chapter 3.7.12 Diagnosing Disconnects Between Results and 'Reali-

ty'). Additionally, the output of a risk model can be carefully examined for the behavior of the risk values compared with our knowledge of behavior of numbers in general.

More formal examinations of the risk assessment are also possible. The processes of verification, calibration, and validation are likely not familiar to most readers and, based on a brief literature search, are not even standardized among those who more routinely deal with them. Some background discussion to these processes, especially as they relate to pipeline risk management, are warranted.

In this text, a distinction is made between verification, validation, and calibration. Verification is the process of 'de-bugging' a model—ensuring that functions operate as intended. Calibration is tuning model output so that it mirrors actual event frequencies. This is a practical necessity when knowledge of underlying factors is incomplete (as it almost always is in natural systems). Validation is ensuring consistent and believable output from the model by comparing model prediction with actual observation. Defining these terms in the context of this discussion is important since they seem to have no universally accepted definitions.

An important aspect of proving a risk assessment is agreement with SME beliefs. Users should be vigilant against becoming too confident in using any risk assessment output without initial and periodic 'reality checks'. But users should also recognize that SME beliefs can be wrong. Disconnects between risk assessment results and SME beliefs are opportunities for both to improve, as is discussed in Chapter 3.7.4 SME Validation.

Note also that the conclusions of any risk assessment can be no stronger than the inputs used. Especially when confidence in inputs is low, calibration to a judged performance is warranted.

3.7.1 Verification

Especially where software is used, verification ensures that the model has been programmed correctly and is, to the extent tested, error-free (no bugs). In a pre-acceptance review of a risk assessment, confirmation of calculations should be performed. Therefore, verification—checks to ensure that intended results are produced by the risk algorithms—confirms that the intended routines are functioning properly.

To ensure that all equations and point assignments are working as intended, some tools can be developed to produce test results using random or extreme value inputs.

3.7.2 Calibration

Risk assessment should be performed on individual pipe segments due to the changes along a pipeline route. These individual risk estimates can be combined (into 'populations') and compared to the known behavior of similar populations. For a variety of reasons, discrepancies in predicted population behavior will usually exist. Calibration serves to rectify the inappropriate discrepancies by adjusting the individual estimates en masse so that credible population characteristics emerge.

The process of calibrating risk assessment results begins with establishing plausible future leak rates of populations based on relevant historical experience, adjusted for relevance and other considerations. These rates become 'targets' for risk assessment outputs, with the belief that large populations of pipeline segments, over long periods of time, would have their overall failure estimates approach these targets. The risk assessment model is then adjusted so that its outputs do indeed approximate the target values for behavior of populations.

The choice of representative population is challenging. It is difficult to find a collection of components similar enough to the system being assessed and with a long enough history to make comparisons relevant. A selection of a population that is not sufficiently representative will weaken the calibration process.

Calibration is done using both a representative population and a target level of conservatism. Both are required as illustrated by this thought exercise. Imagine you could run experiments on real pipelines over long periods of time. Say you chose a 70 mile pipeline operating for 50 years. You would run multiple, maybe hundreds or thousands, of trials to see how the 70 mile pipeline performs over many different 50 year lifetimes. Each trial—that is, each 50 year lifetime—is influenced by random influences of exposures, mitigations, resistance, and consequence scenarios over its 50 years in service. In some of those lifetimes, there would be no incidents, so no actual consequences at all. Choosing these trial results as representative of future behavior of the next trial might reflect a P10 level of conservatism. Some of your multiple trials would result in dozens of leaks and ruptures, some producing very consequential results. Using this set of trials to represent future behavior would be choosing a P90 or so level of conservatism. The results of the majority of your trials would form the P50 portion of the distribution of all results, perhaps the center point of a normal or bell-shaped distribution.

With an appropriate comparison population, the chief goal of a calibration will often be the removal of unwanted conservatism. As discussed, conservatism plays an important and useful role in risk assessment for individual components. P90+ inputs are recommended for many initial risk assessments. However, the need for estimates as close as possible to actual risk levels is also important, especially for populations—collections of individuals. A decision-maker gains more insight from a P50 type risk assessment of a pipeline system than a system summary incorporating multiple P90+ inputs. The P50 estimate can become a part of company-wide strategic planning while the P90+ estimates ensure proper attention to risk management for each component.

In a simple calibration exercise, we seek a single factor representing the amount of conservatism included in a risk assessment's P90+ estimates. This factor can be used to reduce the conservative estimates of each component's risk to best-estimates of risk. The resulting collection of 'best estimates' should be close to the representative population's historical risk levels.

One can track differences between P50 and P99 to see, at least partially, reduction in uncertainty. P50 and P90+ have both natural variability (apparent randomness) and uncertainty. Each PXX produces a distribution. The difference between, for instance,

the midpoint of the P50 and P90+ distributions can be called the conservatism bias multiple.

Both P50 and P90+ risk assessments will often be needed—the former to represent likely system wide behavior and the latter to use in risk management. Some practitioners choose to run parallel P50 and P90+ assessments. Others perform the P90+ assessment, estimate the conservatism bias, and then use it to 'back calculate' P50 results.

Once calibrated, estimates could represent a wide range of possibilities. For example, a US natural gas transmission pipeline may have components with P50 PoF estimates from perhaps 0.00001 to 0.1 reportable events per mile-year, assuming that segments' actual PoF's could range from about 100 times higher or lower than the US average for reportable incidents on natural gas pipelines.

A similar process can be performed on overall risk values or any intermediate calculations. More calibration—calibrating to lower level algorithms—should produce more confidence in the overall correlation. This essentially provides more intermediate correlating points from which a correlation curve can be better developed.

3.7.3 Validation

Validation of a model is achieved by ensuring that appropriate relationships exist among input data and that produced outputs are representative of real-world experience. Validation seeks to authenticate or verify that the model produces risk estimates that are accurate.

While pipeline industry documents do not generally detail these processes, examples of how the pipeline industry typically uses the term 'validation' are noted in PHMSA and PRCI documents:

US Gas IMP Protocol C.04

> Verify that the validation process includes a check that the risk results are logical and consistent with the operator's and other industry experience. [§192.917(c) and ASME B31.8S-2004, Section 5.12] (**http://primis.phmsa.dot.gov/gasimp/QstHome.gim?qst=145**)

From PRCI, discussing validation of a risk-based model for pipelines:

> The fault tree model and basic event probabilities were validated by analyzing a representative cross-country gas transmission pipeline and confirming that the results are in general agreement with relevant historical information.

Validation of risk assessment is also noted in US IMP documents.

ASME B31.8s

"*…experience-based reviews should validate risk assessment output with other relevant factors not included in the process, the impact of assumptions, or the potential risk variability caused by missing or estimated data.*"

As a part of the validation effort, the general relationship between model output and reality should be examined. When new or altered theories are proposed as part of a model, examination of those must be included in the validation process.

Theories applicable to pipeline risk assessment typically include:

- Metallic corrosion
- Mitigation of metallic corrosion—coatings and cathodic protection
- Stresses in a shell structure (pipe)
- Effect of wall loss on pressure-containing capability
- Component rupture potential
- Probability theory
- Probability distributions as applied to observed phenomena
- Structural theory
- Materials science
- Plastics and coatings performance.

The risk assessment methodology described in this book does not propose new theories of failure mechanisms. It relies upon thoroughly documented models of the above theories including widely accepted beliefs about impacts of certain factors on certain aspects of risk; for example, 'increases in Factor X lead to increased risk'.

3.7.4 SME Validation

Similar to the use of a benchmark for model calibration, a carefully structured interview with SME's can also identify model weaknesses (and also often be a learning experience for SME's). If an SME reaches a risk conclusion that is different from the risk assessment results, a drill down (for example, a deeper examination) into both the model and the SME's basis of belief should be done. Any disconnect between the two represents either a model error or an inappropriate conclusion by the SME. Either can be readily corrected. The key is to identify exactly where the model and the SME first diverge in their assumptions and/or conclusions.

An important step in validation is therefore to identify and correct 'disconnects' between subject matter experts' beliefs and model outputs. This is similar to calibration discussed previously but differs in that validation should occur after calibration has been done. In the absence of calibration of risk results, validation can still be performed on intermediate calculations but the role of conservatism must be factored in. For relative, scoring models, validation can only be done in general terms, where SME's can agree in relative changes to risk accompanying certain changes in inputs.

SME concurrence with assessment outputs should be a part of model validation. Risk assessment-identified higher—and lower—risk segments should comport with SME-identified higher— and lower—risk segments.

SME review should include concurrence with aspects such as:

- Direction and magnitude of risk changes accompanying changes in factors and groups of factors
- identified locations of higher- and lower-threats, considering each threat independently
- identified locations of higher- and lower-consequences.

A good objective of risk assessment should be to have the risk assessment model capture the collective knowledge of the organization—anything that anyone knows about a pipeline's condition or environment, or any new knowledge of how risk variables actually behave and interact, can and should be included in the analysis protocol.

3.7.5 Predictive Capability

Implicit in the notions of validation and verification is the idea of predictive capability. A good risk assessment always produces some estimate of failure probability. Theoretically, this can forecast, to some degree of accuracy, future failures on specific pipeline segments. Except in extreme cases, this is not a realistic expectation. A more realistic expectation is for the assessment to forecast behavior of populations of segments rather than individuals. A good risk assessment will, however, highlight areas where probability and consequence combinations warrant special attention.

Leak/break rate is related to estimated failure probability. In most transmission pipelines, insufficient system-specific information exists to build a meaningful prediction model solely from leak/break rate—events are so rare that any such prediction will have very large uncertainty bounds. Distribution systems, where leaks are precursors to "failures," are often more viable candidates for producing predictions directly from leak/break rates.

A leak/break rate assessment may show both time-dependent failure mechanisms such as corrosion and fatigue and more random failure mechanisms such as third-party damages and seismic events. The random events will normally occur at a relatively constant rate over time for a constant set of conditions.

A leak/break rate is called a "deterioration" rate by some, but that phrase seems to be best applied specifically to time-dependent failure mechanisms only (corrosion and fatigue).

Even though they are commonly expressed as a single value, each failure probability estimate really represents an underlying distribution—all possible failure rates with associated probability of occurrence—with an average, median, and standard deviation. This distribution describes the range of failure rates that would accompany any pipeline section with a particular predicted failure rate.

Nonetheless, to test the predictive power of the risk assessment model, the incident and inspection history in recent years could be examined. Knowing what the risk assessment 'thought' about the risk on the day before the incident (or the day before an inspection) would provide insight into the predictive power of the assessment. Given the role of probability, spot samples from individual segments may appear to show inaccurate predictions, but actual accuracy can only be verified after sufficient data has been accumulated to compare the predicted versus actual long term behavior of a large population.

3.7.6 Evaluating a risk assessment technique

Note: Locating this discussion in this book was challenging. On one hand, a reader is often not terribly interested in this aspect until he is an active practitioner. On the other hand, a reader who has an existing risk assessment approach may need early incentivization to investigate alternative approaches. This latter rationale has obviously determined the issue for purposes of organizing this book. The early discussion has a further advantage of setting the stage—arming the reader with criteria that will later determine the quality of his assessments, even as he works his way through this text to learn about pipeline risk assessment.

In general, proving or confirming a risk assessment methodology addresses the extent to which the underlying model represents and correctly reproduces the actual system being modeled. Another view is that validation involves two main aspects:

1) ensuring that the model correctly uses its inputs and
2) model produces outputs that are useful representations of the underlying real-world processes being modeled.

Ref [1046] focuses on the need for transparency in any risk assessment:

> Transparency provides explicitness in the risk assessment process. It ensures that any reader understands all the steps, logic, key assumptions, limitations, and decisions in the risk assessment, and comprehends the supporting rationale that lead to the outcome. Transparency achieves full disclosure in terms of:
>
> a. the assessment approach employed
> b. the use of assumptions and their impact on the assessment
> c. the use of extrapolations and their impact on the assessment
> d. the use of models vs. measurements and their impact on the assessment
> e. plausible alternatives and the choices made among those alternatives
> f. the impacts of one choice vs. another on the assessment
> g. significant data gaps and their implications for the assessment
> h. the scientific conclusions identified separately from default assumptions and policy calls
> i. the major risk conclusions and the assessor's confidence and uncertainties in them;

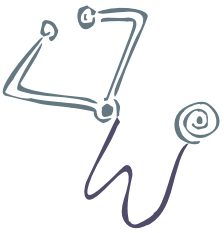
j. the relative strength of each risk assessment component and its impact on the overall assessment (e.g., the case for the agent posing a hazard is strong, but the overall assessment of risk is weak because the case for exposure is weak)

Process transparent and the risk characterization products clear, consistent and reasonable" (TCCR) became the underlying principle for a good risk characterization. [1046]

To properly support risk management, the superior risk assessment process will have additional characteristics, including:

- QA/QC and error checking capabilities, perhaps automated
- Ability to rapidly integrate new information and refresh risk estimates
- Be able to rapidly incorporate new information on emerging threats, new mitigation opportunities, or any other changing aspect of risk.
- Seamless integration with other databases and legacy data systems
- Accessible, understandable to all decision-makers.

3.7.7 Diagnostic tool—Operator Characteristic Curve



For those seeking a more structured approach to proving a risk assessment, techniques are available. A pipeline risk assessment is really a diagnostic tool. Similar to a diagnostic test used by a doctor, the idea is to determine, with the least amount of cost and patient discomfort, whether the patient has the disease or doesn't. He knows that in any population, a certain fraction of individuals will have the disease and most won't. For a diagnosis to be successful, he must correctly determine into which group to place his patient. In making this determination, the doctor can choose a whole battery of expensive and intrusive tests and procedures in order to have the highest confidence of his diagnosis. On the other hand, he can choose minimal tests and accept a higher error rate in diagnoses. The most accurate test or set of tests will minimize the rate of false positives and false negatives. But there is a cost associated with such testing.

In the case of pipeline risk management, the manager is trying to determine which pipeline segments and components have the 'disease' of higher risk among the hopefully many which do not. His choice of tests to help in the diagnosis goes beyond the risk assessment itself. He can request surveys and inspections to improve the diagnosis, but with an accompanying expense and the potential for inefficient use of resources. The latter occurs when expensive 'tests' do not add much certainty to the assessment.

Both the doctor and the risk manager will be balancing the costs of the diagnostics and the costs of being wrong—false positives and false negatives.

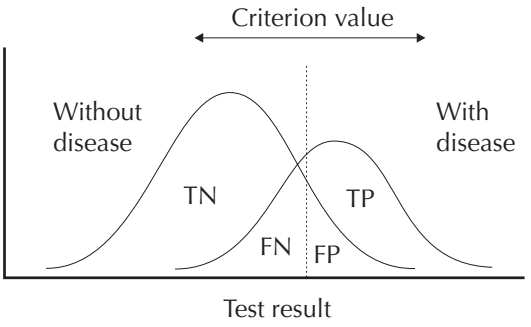


Figure 3.4 Risk assessment as a diagnostic tool; trade-offs among true positives, true negative, false positives, false negatives (TP, TN, FP, FN)

3.7.8 Possible Outcomes from a Diagnosis

The tuner of a leak detection system is well aware of the false alarm phenomena. In order to find smaller leaks, it is necessary to alarm and investigate apparent smaller leaks that later prove to be only transient conditions. After too many false alarms, the investigators grow weary of responding and are less attentive, thereby increasing their error rate when an actual leak does appear. It is standard to sacrifice some leak detectability in order to avoid too many nuisance alarms. A modern leak detection system will state a probability associated with the indication, to assist the investigator in setting his response urgency.

Advanced applications of these ideas is found in signal detection theory, receiver operating characteristic curves, artificial intelligence (machine learning), and others. For our purposes here, it is useful to simply bear in mind the diagnostic intent behind a risk assessment and the corresponding ability to test its diagnostic power over time.

3.7.9 Risk model performance

Some sophisticated routines can be used to evaluate risk assessment outputs. For instance, a Monte Carlo simulation uses random numbers as part of the assessment inputs in order to produce distributions of all possible outputs from a set of risk algorithms. The resulting distribution of risk estimates might help evaluate the "fairness" of the assessment. In many cases a normal, or bell-shaped, distribution would be expected since this is a very common distribution of properties of materials and engineered structures as well as many naturally occurring characteristics. Alternative distributions are also common, such as those often used to represent rare events. All distributions that emerge should be explainable. If some implausible distribution appears, further examination may be warranted. For instance, excessive tails or gaps in the distributions might indicate discontinuities or biases in the results being generated.

3.7.10 Sensitivity analysis

The algorithms that underlie a risk assessment model must react appropriately—neither too much nor too little—to changes in any and all variables. In the absence of reliable data, this appropriate reaction is gauged to a large extent by expert judgment as to how the real-world risk is really impacted by a variable change.

A single variable can play a role as both risk increaser and risk reducer. A casing protects a pipe segment from external force damage but complicates corrosion control; in the offshore environment, water depth is a risk reducer when it makes anchoring damage less likely but it is a risk increaser when it heightens the chance for buckling. So the same variable, water depth, is a "good" thing in one part of the model and a "bad" thing somewhere else.

See discussion of data collection in Chapter 4 Data Management and Analyses for a deeper examination into types and roles of information

Some variables such as pressure and population density impact both the probability (often linked to lower resistance and higher activity levels) and consequence (larger hazard zone and more receptor damage) sides of the risk algorithm. In these cases, the impact on overall risk is not always obvious. When a variable is used in a more complex mathematical relationship, such as those sometimes used in resistance estimates, then influences of changes on final risk estimates will also not be apparent.

Sensitivity quantifications can be utilized for evaluating effects of changing factorsbut require fairly sophisticated analyses procedures. It is important to recognize that many variables will *usually* play lessor roles in overall risk but may occasionally be the single greatest determinant of risk.

3.7.11 Weightings

FOCUS POINT
The use of weightings in a risk assessment will almost certainly result in serious analyses errors.

The use of 'weightings' should be a target of critical review of any risk assessment practice. Weightings have been used in some older risk assessments to give more importance to certain factors. They were usually based on a factor's perceived importance in the majority of historical pipeline failure scenarios. For instance, the potential for AC induced corrosion is usually very low for many kilometers of pipeline, so assigning a low numerical weighting appeared appropriate for that phenomenon. This was intended to show that AC induced corrosion is a rare threat.

Used in this way, weightings steer risk assessment results towards pre-determined outcomes. Implicit in this use is the assumption of a predictable distribution of future incidents and, most often, an accompanying assumption that the future distribution

will closely track the past distribution. This practice introduces a bias that will almost always lead to very wrong conclusions for some pipeline segments.

The first problem with the use of weightings is finding a representative basis for the weightings. Weightings were usually based on historical incident statistics—"20% of pipeline failures from external corrosion"; "30% from third party damage"; etc. These statistics were usually derived from experience with many kilometers of pipelines over many years of operation. However, different sets of pipeline kilometer-years shows different experience. Which past experience best represents the pipeline being assessed? What about changes in maintenance, inspection, and operation over time? Shouldn't those influence which data sets are most representative to future expectations?

It is difficult if not impossible to know what set of historical population behavior best represents the future behavior of the segments undergoing the current risk assessment. If weightings are based on, say, average country-wide history, the non-average behavior of many miles of pipeline is discounted. Using national statistics means including many pipelines with vastly different characteristics from the system you are assessing.

If the weightings are based on a specific operator's experience, then (hopefully) only a very limited amount of failure data is available. Statistics using small data sets is always problematic. Furthermore, a specific pipeline's accident experience will probably change with the operator's changing risk management focus. When an operator experiences many corrosion failures, he will presumably take actions to specifically reduce corrosion potential. Over time, a different mechanism should then become the chief failure cause. So, the weightings would need to change periodically and would always lag behind actual experience, therefore having no predictive contribution to risk management.

The bigger issue with the use of weightings is the underlying assumption that the past behavior of a large population will reliably predict the future of an individual. Even if an assumed distribution is valid for the long term population behavior, there will be many locations along a pipeline where the pre-set distribution is not representative of the particular mechanisms at work there. In fact, the weightings can fully obscure the true threat. The weighted modeling of risk may fail to highlight the most important threats when certain numerical values are kept artificially low, making them virtually unnoticeable.

The use of weightings as a significant source of inappropriate bias in risk assessment is readily demonstrated. One can easily envision numerous scenarios where, in some segments, a single failure mode should dominate the risk assessment and result in a very high probability of failure rather than only some percentage of the total.

Consider threats such as landslides, erosion, or subsidence as a class of failure mechanisms called geohazards. An assumed distribution of all failure mechanisms will almost certainly assign a very low weighting to this class since most pipelines are not significantly threatened by the phenomena and, hence, incidents are rare. For example, to match a historical record that shows 30% of pipeline incidents are caused by corro-

sion and 2% by geohazards, weightings might have been used to make corrosion point totals 15 times higher than geohazard point totals (assuming more points means higher risk) in an older scoring methodology.

But a geohazard phenomenon is a very localized and very significant threat for some pipelines. It will dominate all other threats in some segments. Assigning a 2% weighting masks the reality that, perhaps 90% of the failure probability on this segment is due to geohazards. So, while the assumed distribution may be valid on average, there will be locations along some pipelines where the pre-set distribution is very wrong. It would not at all be representative of the dominant failure mechanism at work there. The weightings will often completely mask the real threat at such locations.

This is a classic difficulty in moving between behaviors of statistical populations and individual behaviors. The former is often a reliable predictor—hence the success of the insurance actuarial analyses—but the latter is not.

In addition to masking location-specific failure potential, use of weightings can force only the higher weighted threats to be perceived 'drivers' of risk, at all points along all pipelines. This is rarely realistic. Risk management can become driven solely by the pre-set weightings rather than actual data and conditions along the pipelines. Forcing risk assessment results to resemble a pre-determined incident history will almost certainly create errors.

Since weightings can obscure the real risks and interfere with risk management, their use should be discontinued. Using actual measurements of risk factors avoids the incentive to apply artificial weightings (see previous column on the need for measurements). Therefore, migration away from older scoring or indexing approaches to a modern risk assessment methodology will automatically avoid the misstep of weightings.

3.7.12 Diagnosing Disconnects Between Results and 'Reality'

FOCUS POINT
A 'gut check' is a reasonable and prudent aspect of validation

PRMM provides a useful discussion of types of disconnects between reality and assessed results that may arise in a risk evaluation. Disconnects discussed there include those that may emerge from:

- New inspection results, including visual inspections
- Incident investigations, including root cause analyses
- Leak history analyses
- Populations vs individuals disconnects.

An important step in validation is to identify and correct 'disconnects' between sources such as subject matter experts' beliefs and risk assessment outputs. Two types

of potential disconnects should be explored. The first is comparisons of populations—the behavior of an assessed collection of components (for example, a pipeline system) with a representative population of similar components (other pipeline systems). The representative population will be called a benchmark for purposes here. Common benchmarks include average incident rates for many km of pipelines over several years, often country-wide (for example, US, Canadian, European, etc).

The second comparison disconnect type involves a risk assessment of a component or several components whose risk estimates do not comport with SME beliefs or other evidence. Other evidence includes results of inspections not available prior to the risk assessment.

If assessment results are not consistent with a benchmark believed to closely represent future performance of the system or when a discrepancy arises in a comparison of a component- or location-specific assessment with an SME belief or other evidence, any of several things might be happening:

- Benchmark is not representative of the assessed segments
- Effects of conservatism are not being fully considered
- Both are correct (ie, within the range of expectations), but probability effects make them appear contradictory
- Exposure estimates were too high or too low,
- Mitigation effectiveness was judged too high or too low,
- Resistance to failure was judged too high or too low.
- Consequences estimates were too high or low
- SME belief or contrarian evidence is flawed.

The distinction between PoF and probability of damage (damage without failure) can be useful in diagnosing where the assessment is not reflecting perceived reality. If damages are predicted but not occurring, then the exposure is overestimated and/or the mitigation is underestimated. Alternatively, consider a situation where damage potential is modeled as being very low but an inspection (perhaps ILI) discovers certain damages. It is often difficult to determine which estimate—exposure or mitigation—was most contributory to the damage underestimate, but insight has been gained nonetheless.

Mitigation measures have several aspects that can be tuned. The orders of magnitude range established for measuring mitigation is critical to the result, as is the maximum benefit from each mitigation, and the currently judged effectiveness of each. More research is becoming available and can often be used directly in judging the effectiveness of a mitigation measure.

Note that calibration might also be contributing to such disconnects. Calibrating to a target population of pipeline segments includes ‘outliers’ in the target distribution. So, disconnects involving very few segments may be only due to the outlier effect. More widespread disconnects may indicate that the target population used in calibration is not representative of the pipeline segments being assessed.

A trial and error procedure might be required to balance all these aspects so the assessment produces credible results for all inputs.

3.7.13 Incident Investigation

Incident investigation is both a useful input into a risk assessment and a consumer of risk assessment results. In the former, learnings from the incident are almost always relevant to other portions of other pipelines. In the latter, especially when responsibility (blame) is to be assigned, what should have been known, via risk assessment, prior to the incident is almost always relevant. From this, the risk management decision-making will normally be challenged by parties having suffered damage from the incident.

Retro-fitting a risk assessment for this type of application uses the same steps as any other risk assessment. Care must be exercised to not introduce hindsight, if the assessment is to truly reflect what was/should-have-been known immediately prior to the incident.

When evaluating what should have or 'could have' been known and what should have (or 'could have') been done prior to an accident, the investigation often seeks to determine if decision-makers acted in a reasonable and prudent manner. For more extreme behavior, the legal concept of negligence may also be applicable and some investigations will seek to demonstrate that.

The risk aspect of the investigation can focus on these issues by including the following:

1. List of evidence available prior to incident. This includes information that was readily available to decision-makers prior to the incident. Less available information—determining to what extents research, data collection, investigation, etc, should have been done—is a later consideration.
2. Risk implications of this evidence. This can be demonstrated via a translation, showing how each piece of evidence is translated into a measurement of exposure, mitigation, resistance, or consequence.
3. P50 and P90+ risk assessments prior to incident, using all available information, again, prior to incident. The assessment should model uncertainty as increased risk, reflecting a prudent decision-making practice of erring on the side of over-protection.
4. Decision-making context. Here, the risk report puts the assessment results into context for the reader. This can include at least two types of context:
   Relative: how did the risk of the subject segment—the failed component—compare to other risks under the control of the risk manager, immediately prior to the incident? Should this have been a priority segment for the decision-makers? Did the failure mechanism that actually precipitated the event appear as a dominant threat? Should it have, given the information available at the time?
   Acceptability Criteria: immediately prior to the incident, would the risk from this segment have been deemed 'acceptable' by any common measure of

risk acceptability? Even when numerical criteria for 'risk acceptability' or 'tolerable risk' are unavailable for a specific pipeline, inferred and comparative criteria are always available. Examples are numerous and include:

- Risk criteria used in similar applications; for example, siting of pipelines near public schools [1048].
- General industrial risk criteria used in other countries; for example, ALARP
- Land use and setback criteria suggested in some guidelines [1047] and applied in some municipalities
- Risk criteria employed in other industries
- Suggested target reliability levels. [95, 333]

Risk criteria often use fatalities as the consequence of interest. So, even if not directly applicable to the subject pipeline, the fact that a fatality-based risk level is tolerable (or not) in a similar area or for a similar application, may be relevant to the subject incident.

Care should be exercised to emphasize the probabilistic nature of a risk assessment. A risk assessment can easily fail to highlight a threat that later turns out to cause the next failure. But that does not mean that the assessment is incorrect. A 1% probability event can occur before a 90% probability event, but they may still be accurately depicted as 1% and 90% probability events, respectively. Of course, if several events assessed at 1% each happen before the 90% event, the assessment results should become increasingly suspect.

5. Mitigation options prior to the incident. A listing of all risk reduction opportunities available to decision-makers prior to the incident will be useful to the analyses. The reasonableness of each should not be a consideration at this stage—rather the focus should be on a comprehensive list.
6. Cost/benefit analyses of available mitigation prior to the incident. This addresses reasonableness and is also captured in ALARP. See Chapter 13 Risk Management. While spending to prevent consequences that are difficult to monetize (for example, fatality, threatened and endangered species harm, etc) evokes emotionalism in decision-making, there is nonetheless a concept of reasonableness in spending to prevent any type of potential loss. Monetization of all types of consequence is becoming more common. But even expressed in qualitative (non-monetized) ways, the costs of opportunities for consequence avoidance prior to the incident, will still be of use in the investigation.

3.7.14 Use of Inspection and Integrity Assessment Data

The first and primary use of inspection and integrity assessment data, including investigations from failures and damage incidents, is in determining resistance. This is detailed in Chapter 10 Resistance Modeling. A secondary, but also very important use

of this information is in revisiting previous assumptions used in the risk assessment. Since this latter use permeates so many inputs into a risk assessment, this topic is explored here in an early chapter.

When inspection does not find damages where they had been predicted by the risk assessment, a common cause is conservatism in the risk estimates. However, one should not discount the possibility of damages present but undetected by the inspection. In the case of ILI, such disconnects may warrant a re-examination of factors such as:

- Assumed detection capabilities to various ILI types regarding various anomaly types and configurations.
- Assumed reductions in detection capabilities to various types of ILI excursions.

When an inspection detects corrosion or cracking damage, it is logical to conclude that damage potential existed at one time and may still exist. When there is actual damage, but risk assessment results do not indicate a significant potential for such damage, then a conflict seemingly exists between the direct and the indirect evidence. Such conflicts are discussed in Chapter 3.7 Verification, Calibration, and Validation, especially Chapter 3.7.12 Diagnosing Disconnects Between Results and 'Reality'.

Identifying the location of the inconsistency is necessary. The conflict could reflect an overly optimistic assessment of effectiveness of mitigation measures (coatings, CP, etc.) or it could reflect an underestimate of the harshness of the environment. Another possibility is that detected damages do not reflect active mechanisms but only old and now-inactive mechanisms. For instance, replacing anode beds, increasing current output from rectifiers, eliminating interferences, and re-coating are all actions that could halt previously active external corrosion. Finally, the apparent disconnect might not be a disconnect at all. It could simply be an actually very rare occurrence whose time had come. Even very low probability events will occur eventually.

The degradation estimates in a risk assessment should always include the best available inspection information. The risk assessment should preferentially use recent direct evidence over previous assumptions, until the conflicts between the two are investigated.

For example, suppose that, using information available prior to an ILI, the assessment concluded a low probability of subsurface corrosion because both coating and CP were estimated to be fully effective. If the ILI recent inspection, indicates that some external metal loss has occurred, then the subsurface corrosion assessment would be suspect, pending an investigation. The previous assessment based on indirect evidence should probably be initially overridden by the results of the ILI pending an investigation to determine the cause of the damage—how the mitigation measures may have failed and how the risk assessment failed to reflect that.

If the risk assessment is modified based upon un-verified ILI results, it can later be improved with results from more detailed examinations, that is, excavation, inspection, and verifications that anomalies are present and represent loss of resistance. If a root cause analysis of the detected damages concludes that active corrosion is not present,

the original risk assessment may have been correct. The root cause analysis might demonstrate that corrosion damage is old and corrosion has been mitigated and values may have to again be revised.

A similar approach is used for integrity assessments such as pressure tests. If test results were not predicted by the risk assessment, investigation is warranted.

Techniques to assimilate ILI and other direct inspection information into risk estimates are discussed in Chapter 10 Resistance Modeling.