

A Data Quality Assessment

*Evaluating the major safety data programs
for pipeline and hazardous materials safety*

November 10, 2009

Prepared by Rick Kowalewski
Senior Policy Advisor
Pipeline and Hazardous Materials Safety Administration
U.S. Department of Transportation

A Data Quality Assessment

Evaluating the Major Safety Data Programs for Pipeline and Hazardous Materials Safety

November 10, 2009

Objectives and scope of the evaluation: The purpose of this data quality assessment is to ensure our safety data provide a sound basis for risk-based decision making. The assessment focused on the major data collection programs we use to assess and manage risk in the pipeline and hazardous materials safety programs. These data collection programs—together with shared, professional experience—comprise the core of our knowledge base about systems and program performance. The data are used by PHMSA, states, communities, other agencies, researchers, the private sector (companies and trade associations), and the general public.

Background: In the 2008 safety culture survey, only 46% of PHMSA employees agreed that *“our available safety data is useful for decision making.”* More recent concerns about the quality of our safety data have amplified the need for assessing data quality, but this assessment was really begun as an outgrowth of PHMSA’s 2007-2011 strategic plan (the first of four general strategies was focused on data-driven risk management), and DOT’s Information Quality Guidelines—which recommend periodic reviews of mission-critical data systems.

All data systems have error, and errors tend to accumulate through the development and operating cycle of a data program. We examined the life cycle of incident, activity, and exposure data from the definition of requirements for information through system design and data collection/processing to the interpretation and use of analytical results. The aim was to identify the major sources of error in the data or its use—as a basis for continuous improvement in our programs.

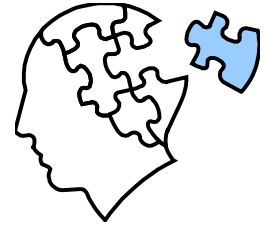
We recognize that there are some significant efforts underway to address many of the data quality issues outlined in this report. We can’t judge the likely outcomes of all these ongoing efforts (our evaluation didn’t probe them deeply). But we believe that many of the findings from this assessment could be used to inform, reinforce, redirect, or extend the scope of these efforts; and that this assessment can provide a useful baseline for evaluating the results of these efforts.

Overview of the results:

Data gaps limit our ability to analyze risks.

Missing data often compromises our ability to draw useful conclusions. A 2007 review estimated we are missing 60-90% of the hazmat incidents that occur. Incident reports in both programs are often missing some important data, including a higher number of “other/unknown” causes for the most serious incidents—exactly where we need good data most. Reporting lags (months, or years in some cases) compromise our performance reporting and time series analyses. But more serious are gaps in the scope of the data we collect. Our failure data focuses at the top of a much larger pyramid, and

small numbers make it difficult to detect meaningful trends. We are missing information about many of the precursors to larger failures—leading indicators that could help reduce risk.



Our “peripheral vision” is limited too: for several “invisible risks” (e.g., gas pipeline master meter operators or failures of DOT packages “outside transportation”), we have little/no risk data. Lacking data, we can’t quantify or address these risks effectively. We have limited data on risk exposure to help us understand failure rates or trends. At the same time, we collect very little useful data from our own inspections and investigations—which could provide our primary *sensory information* on what’s going on and our primary *feedback loop* on where things have not worked. At the root of all these issues: *We haven’t thought through what we need to know in a systematic way.* As a result, we often don’t have the data we need to understand risk, while we don’t use much of the data we collect.

We don’t have a good conceptual model for understanding failures.

We don’t capture the chain of failures—especially the root causes—that typically are associated with significant incidents. We don’t capture all the latent conditions, circumstances and interactions that might reveal hidden explanations. We can’t use incident data effectively to focus our inspections because we don’t capture failure data in a form that is very useful for inspections. And we don’t capture inspection deficiencies in a way that allows us to connect latent conditions to later failures. The relationship between conditions (or processes) and outcomes goes to the heart of regulation. With weak data on both sides of the equation, we might be hampered in our ability to impact outcomes by regulating processes.

Errors and biases in the data contribute to a misleading picture.

In practice, the quality of our data reviews and edit checking is mixed. Underreporting and blank data fields are more serious than just *reducing the numbers*. It appears likely the reports and data we get are *not representative* of all the incidents that actually occur. When data gaps are not random, it can be misleading to draw general conclusions from the data we have.

We rely heavily on the regulated industry to help us acquire information. This is convenient, and goes directly to the source. It also introduces a natural, inherent bias in the data we collect. Our accident investigations have shown some significant differences between what the company reports and an objective view of these events. Our processes do not effectively reconcile these discrepancies. There is a historical understanding that the data we get from industry is “their” data. Even when we know (or believe) their data to be wrong, we don’t modify our data until we get revised reports from the companies—which can result in releasing and using bad data for months or years after an incident.

We don’t regularly monitor data quality indicators. We invest substantial resources in improving the accuracy of our data. But we do not track error rates or how much our efforts *change* the quality of our data. We could be over-investing in report-level accuracy compared to the more general problems with data gaps, concepts, and analysis of the data.

Analytical gaps limit our ability to direct safety programs effectively.

Today, we have limited organizational capacity and focus for risk analysis; and almost no capacity for program evaluation. These should provide the *primary intelligence function* for interpreting the data we get on systems, program effectiveness, and failures; and turning it into useful program information. These gaps are increasingly important as we achieve diminishing returns in our safety programs. More importantly, these evaluation processes are usually the drivers for a wide range of other processes aimed at ensuring high quality data.

Over the years, we have assumed our programs are effective (or not effective) with no clear analytical basis for that assumption. We modify or develop new programs without a systematic evaluation of what we have now, or a plan to collect data to evaluate the impact of changes. We monitor key safety outcomes, but we have limited understanding of the safety trends we are seeing, and little data or analyses that might be used to identify emerging risks or leading indicators of safety. As a result, we don't use our performance measures to drive our programs. Our regulatory evaluations generally begin too late in the process to affect decision making. Our inspection targeting models combine judgment and data in ways that can degrade the quality of the original data. Good data, with clear predictive value, can be overwhelmed by less-important data variables. Our grant allocation models use risk data or performance data, but not both in a direct way. To target resources effectively, we need information about where the risk is *and* what works in reducing risk.

Our approach to analysis is uneven and is not guided by a strategic view. We sometimes frame our questions about risk in very general ways, leading to analytical results or tabulations that are disappointing and not very useful for decision making. We do not have any standard practices in presenting the results of our analyses. Some of our analyses highlight the limitations of the data and methods; many don't. We rarely quantify the uncertainty in our analyses, and the uncertainty is often large—which could undermine the basis for important program decisions. Because of these analytical gaps, we often make program decisions without good analytical input.

Much of the data we release are difficult to use.

Data will never be perfect, but those using it have to know its imperfections; otherwise, misleading conclusions can lead to poor decisions. We release data without documentation to help analyze it, and our own tabulations and analyses are often misleading. Good metadata would include how the data are collected, what the data elements mean, how the reporting has changed over time, and where there are known data quality issues. We have no data documentation for our major safety data systems beyond simple record layouts. For pipeline incidents, we don't release narrative descriptions of the incident, which are often the most useful information for analysis. And often the data we have are difficult to integrate across data systems when they lack common identifiers or common data architecture.

Our tabulations and graphics reflect a wide range of practices. The pipeline safety website demonstrates good practices in every aspect addressed in the Information Quality Guidelines. Hazmat safety tabulations, by comparison, are missing many key elements of good presentation. Some of our

statistical tabulations normalize the data, and some don't. The data summaries we publish do not differentiate public vs. occupational (or private sector) risk—providing a misleading view of public risk. Generally we have no standard, pre-dissemination review process for the statistics we release.

We're missing some key skills.

We lack some of the analytical skills/expertise needed to focus our data collection. While our compliance program is aimed at influencing company behavior, we have no social/behavioral scientist positions to help guide our efforts. Our regulatory evaluations estimate costs and benefits of alternative approaches, but we have no economist positions to guide the work. Many of our hazmat accident investigators have no training in root cause analysis. Our use of established statistical methods is generally weak—reflecting a broader lack of analytical capacity—and we have little expertise in presenting quantitative information effectively. We have almost no professional experience in the specialized discipline of program evaluation, and no positions requiring these skills.

While many efforts are underway to address data quality, these skill gaps could seriously limit our ability to think through what we really need to know, fill all the important data gaps, develop good conceptual models of failures, build a strong analytical capability, and provide data that are well-organized and easy to use for analysis.

General methodology:

The evaluation concentrated on periodic reports (pipeline annual reports, hazardous materials registrations) and event-driven reports (particularly incidents, inspections and investigations)—and the basic questions that need to be answered from these data. The general approach we followed was to evaluate the extent to which our data programs follow DOT's Information Quality Guidelines, identify the sources of error and other potential quality problems in the data, and assess the utility of the data for decision making.

Our overall strategy was to be *comprehensive*, not *exhaustive*. We aimed to keep a high level view of the problem, to maintain a perspective on the relative importance of the data quality issues we face. In general, we looked at the larger context of the *decisions* that need to be made using the data, and how errors in data collection and analysis could ultimately affect these decisions.

The evidence we considered:

- We reviewed previous studies to identify data issues; existing documentation on our data systems; and some of the literature on data quality, risk management, and safety regulatory programs.
- We interviewed analysts and managers—including the executive leadership for each operating program—to get their perspectives on information needs and known data issues.
- We reviewed several recent program analyses to get a sense of the important data limitations and their causes, and we examined a sample of internal reports from inspections and accident investigations.
- We evaluated reporting forms and instructions, and tested incident reporting.
- We reviewed publicly-available micro-data releases, and evaluated published data and statistics on the PHMSA website.
- We also analyzed the data directly, including more in-depth analysis of selected data quality issues.

The attached *Findings* provide the core of this report—a summary of the evidence considered, with a discussion of some of the likely causes and implications. Draft findings were reviewed by the program managers for factual accuracy; their comments and suggestions have been considered and largely addressed in the final report. The Associate Administrators for both programs and the Chief Information Officer acknowledged the general validity of the findings. In fact, many of the issues were identified by program directors and staff independently—before this assessment, and as input into it.

Some options for improving our data and analyses:

In our judgment, we should not try to fix every problem identified here in a near-term action plan. Instead, we recommend a step-wise approach to improving data quality—focused on structural changes, not specific solutions that are limited to the data problems we can identify today. The aim would be to build an engine that can drive continuous improvement in our data, re-setting priorities along the way, in a way that is sustainable. To do this, here are several elements/options we might consider:

- *Build our analytical capacity* (especially hazmat risk evaluation, broad program evaluation, and data quality analysis)—with multi-disciplinary expertise. This should include economists, social and behavioral scientists, statisticians, engineers, and professional evaluators. Clearly there are alternatives in the level and timing of resource investment, organizational placement of functions and positions, and the degree of centralization that should be considered.
- *Develop an analytical agenda* (draft attached) to guide our work/priorities in answering the most pressing questions. Set aside some resources to simply explore the data for trends. We could certainly do more to analyze the data we've already got, as we learn how to make the data better.
- *Expand our accident/failure investigation programs* to develop better data on the causes and circumstances of safety failures. Develop criteria, guidance, and training for conducting failure investigations; and review processes for making use of the information. Give the program an organizational home—to focus the application of knowledge and data broadly from accidents to decision making. Consider redirecting resources from inspections as needed to do this.
- *Explore options for re-casting our inspections* to put more emphasis on learning about the organizations and systems we regulate, identifying safety issues and deficiencies—apart from whether they are compliance issues—identifying good practices, and capturing more data on these systems and processes. This could range from putting more of what we find in the form of data to a broader change in how we conduct inspections. As a first step, program logic modeling could help clarify the assumptions we're making and the linkages between our activities and outcomes.
- *Make the data our own*. Reduce/limit the reporting burden for incidents to a few key pieces of information we need, and follow up with our own investigations of more significant failures to develop good data on causes and circumstances. We might, instead, retain the current reporting systems with greater levels of review of the data, and/or supplement the data with our own. However, in any event we need to better address the inherent biases in industry reporting and disseminate data that are clear and easy to use.

- *Expand our collection of failure data* beyond reportable incidents and beyond the currently-regulated community. Identify all the gaps we can, and try to fill them in to get at least some risk data on the “invisible risks” we know about. Each risk area probably needs to be considered on its own, given that the costs and benefits of acquiring data on them will be context-dependent.
- *Require analytical input* for all major policy, program, and rulemaking initiatives—*before* we decide on our approach. How we go about this, of course, must take into account the analytical capacity we have at the time. But we also need to keep in mind the general principle: *using data is the best way to improve it*.
- *Develop priorities for targeted data quality analyses*, including clarifying key concepts (failures, etc.) and adding the chain of events in our data, developing data profiles and metadata to release with our safety data, developing a common model for Federal/State inspection data, and initiating a targeted evaluation study of hazmat incident under-reporting. A good conceptual model for failures might be the highest priority here. We might use the ongoing program logic modeling (in pipeline safety) as one input, but developing a strong data model probably requires some intensive research to make it really useful.
- *Develop/establish a governance structure* to clarify authorities and responsibilities for data quality, including the key issues of data ownership and maintenance.

A concluding observation: We recognize that data quality has a cost, and there is no such thing as perfect data. We also recognize that data and analysis are part of a larger picture in managing and carrying out an effective safety program. Addressing the data quality issues identified in this assessment might require further research and evaluation in some areas; it certainly will require careful evaluation of the options and tradeoffs—in resources, organization, processes, and technology.

Respectfully submitted November 10, 2009

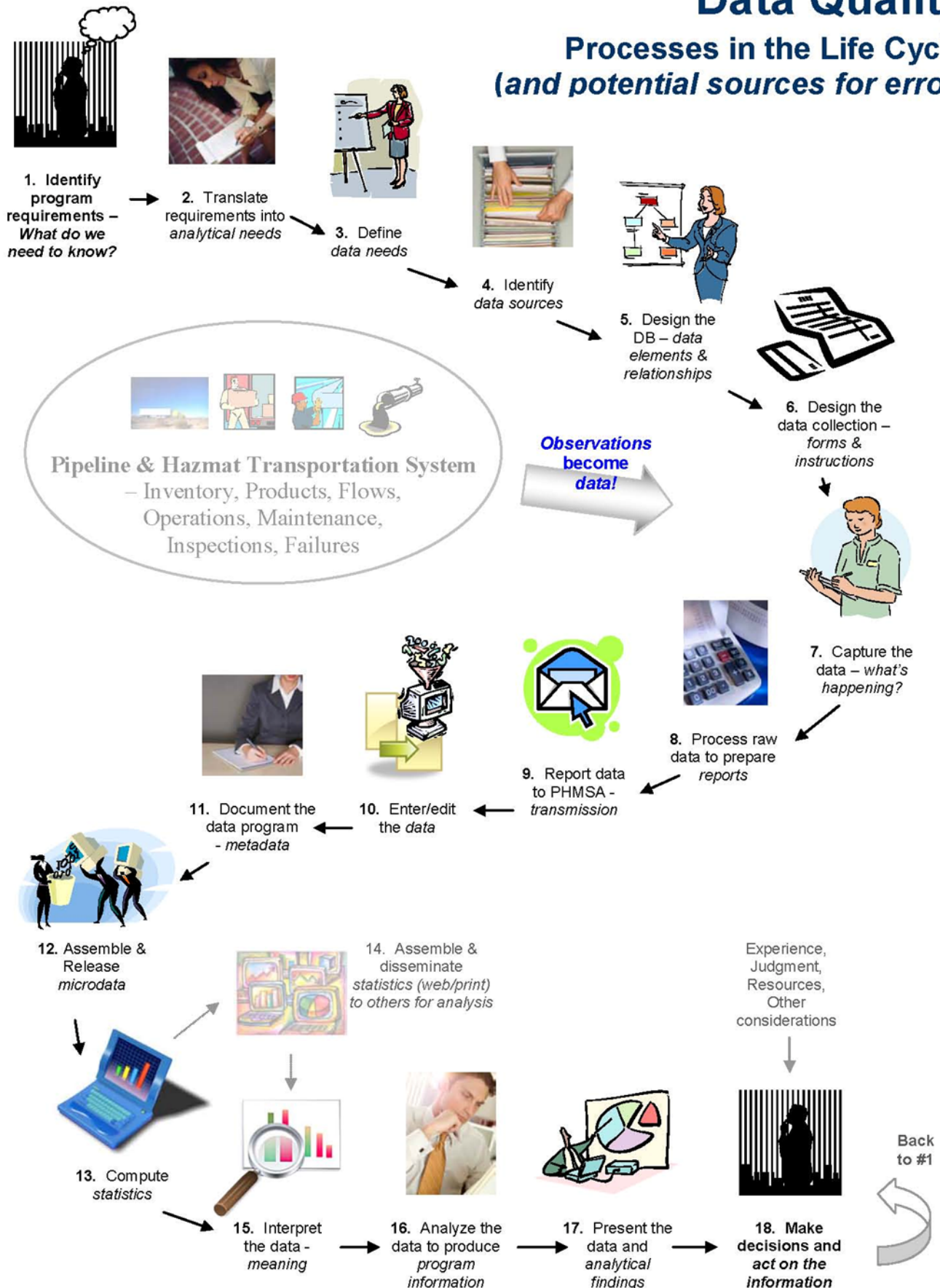
Rick Kowalewski
Senior Policy Advisor
Pipeline and Hazardous Materials Safety Administration

Attachments:

- A. Data quality – processes in the life cycle
- B. A summary of data quality risks
- C. Findings
- D. Information sources
- E. A summary of data quality issues from recent analyses
- F. A summary of past reviews of our data programs
- G. A draft analytical agenda

Data Quality

Processes in the Life Cycle (and potential sources for error)



Data Quality Assessment - A Summary of Risks

November 2009

"Risk" is defined here as the risk of error in the data or its use. The "degree of risk" is a judgment based on the cumulative evidence from the data quality assessment. Processes with "higher risk" reflect specific, significant weaknesses that suggest relatively greater impact on the use of the data in decision making. Each of these is addressed in the Findings section of the assessment.

Relative Degree of Risk

Processes in the life cycle of data quality	DOT's Information Quality Guidelines	Who is responsible for the process?	Common sources of error	Primary Effects	Possible results of errors	Hazmat Safety Data	Pipeline Safety Data
Getting the concepts right ...							
1	Identifying program requirements: <i>What do we need to know?</i>	2.1 Data system objectives	Decision makers	Requirements not defined, or too vague to be useful.	Relevance, Completeness	Invisible risks; no basis for program assumptions.	Higher Risk Higher Risk
2	Translating requirements into <i>analytical needs</i>	2.2 Data requirements (empirical indicators)	Pgm analysts	Analytical questions not formulated, overly focused on current programs.	Relevance	No analytical framework for data reqmts; important program questions can't be answered.	Higher Risk Higher Risk
Design ...							
3	Defining <i>data needs</i>	2.2 Data requirements (data needs)	Pgm analysts	Missing key data (e.g., contributing causes)	Completeness	Data gaps - Important questions can't be answered, or are answered incorrectly.	Medium Risk Medium Risk
4	Identify <i>data sources</i>	2.3 Methods to acquire data; 2.4 Sources of data	Statisticians, DB developers	Alternative data sources not adequately considered to meet quality needs	Accuracy	Biases in data reporting, inconsistency in data concepts, inefficient collection	Higher Risk Higher Risk
5	Designing the data base - <i>data elements and relationships</i>	2.5 Data collection design	DB developers	Lack of standards; DB does not reflect or accommodate important relationships.	Comparability, Utility	Data not consistent, and can't be integrated; analytical results may be misleading.	Higher Risk Higher Risk
6	Designing the data collection - <i>forms and instructions</i>	3.1 Data collection operations	Statisticians, DB developers	Forms/instructions are unclear; scope of data collection is incomplete.	Accuracy, Completeness	Miscommunication - Reported data might not reflect what is intended.	Lower Risk Lower Risk
Measurement ...							
7	Capturing the data - <i>What's happening?</i>	3.1 Data collection operations	Operators, inspectors, investigators	Errors in judgment or measurement, gaps in knowledge, biases.	Accuracy, Completeness	Incorrect data are reported and used in subsequent analyses; information gaps.	Higher Risk Higher Risk
8	Processing raw data to prepare <i>reports</i>	3.1 Data collection operations	Operators, inspectors, investigators	Errors in transcribing or processing the data; under-reporting.	Completeness, Accuracy	Analyses are based on biased, incomplete, or inaccurate data.	Higher Risk Lower Risk
9	Reporting data to PHMSA - <i>transmission</i>	3.1 Data collection operations	Operators	Low response rates, duplicate reporting.	Accuracy, Completeness	Statistical analyses are biased, may be invalid.	Lower Risk Lower Risk
Processing ...							
10	Entering/editing the data - <i>quality control</i>	3.2 Missing data avoidance; 4.1 Data editing and coding; 4.2 Handling missing data	DB mgrs	Invalid entries or incomplete data not caught.	Accuracy, Completeness	Incorrect/incomplete data are not fixed at last best opportunity.	Lower Risk Lower Risk
11	Documenting the data program - <i>metadata</i>	5.3 Source and accuracy statements	Data system owners	Missing or poor documentation.	Utility	Analysts don't understand the limitations of the data; results go beyond the data.	Medium Risk Medium Risk
12	Assembling & releasing <i>microdata</i>	5.2 Micro data releases; 5.4 Pre-dissemination reviews	DB mgrs	Processing errors, problems with comparability of data over time.	Accuracy, Comparability	Data are misinterpreted.	Medium Risk Medium Risk
13	Computing <i>statistics</i>	4.3 Production of estimates and projections	DB mgrs, Analytical groups	Not addressing uncertainty or known data limitations, or normalizing data for comparison.	Comparability, Utility	Statistics are misleading and misused.	Medium Risk Lower Risk
14	Assemble & disseminate <i>statistics</i>	5.1 Publications and disseminated summaries; 5.4 Pre-dissemination reviews	DB mgrs	Presentation is confusing, incomplete, or misleading.	Comparability, Utility	Statistics are misleading and misused.	Lower Risk Lower Risk
Interpretation ...							
15	Interpreting the data - <i>deriving meaning</i>	4.4 Data analysis and interpretation	Analysts (govt and public)	Wrong data are used, data are misinterpreted, or stakeholders are not consulted.	Utility	Decisions are not grounded in the data, or don't consider multiple perspectives.	Higher Risk Medium Risk
16	Analyzing the data to produce <i>program information</i>	4.4 Data analysis and interpretation	Analysts (govt and public)	Analysis focuses on the wrong issues, or uses inappropriate methods.	Relevance, Utility	Analytical results are wrong, misleading, or irrelevant to program decisions.	Higher Risk Medium Risk
17	Presenting the data and <i>analytical findings</i>	5.1 Publications and disseminated summaries	Analysts (govt and public)	Presentation is confusing or misleading.	Utility	Decision makers cannot use the results of analysis for decisions.	Higher Risk Medium Risk
Use ...							
18	Using data - Making decisions and <i>acting on the information</i>	[Not addressed specifically in the guidelines.]	Decision makers	Data are not valued or demanded; misunderstanding limits use of the information.	Relevance, Utility	Decisions are not grounded in the data, or conflict with the data.	Higher Risk Higher Risk

Findings from the Data Quality Assessment

November 10, 2009

Evaluating Data Quality:

This review is intended to be a broad data quality assessment to ensure our safety data provide a sound basis for risk-based decision making. The assessment focused on the major data collection programs we use to assess and manage risk in the pipeline and hazardous materials safety programs. These data collection programs—together with shared, professional experience—comprise the core of our knowledge base about systems and program performance. The data are used by PHMSA, states, communities, other agencies, researchers, the private sector (companies and trade associations), and the general public.

The DOT Information Quality Guidelines (*Guidelines*) suggest periodic assessments of data quality to assess sources of possible error and other potential quality problems in the data—ultimately to help data system owners improve data quality. The *Guidelines* also suggest more targeted evaluation studies to evaluate particular aspects of data quality—periodically, when analysis of the data reveals a significant problem, and especially after a major system redesign.

The two characteristics most likely to distinguish safe organizations from less safe ones are, firstly, top-level commitment and, secondly, the possession of an adequate safety information system.

- Managing the Risks of Organizational Accidents (James Reason, 1997)

For this evaluation, we identified 18 processes affecting data quality (attachment A). Each of these presents an opportunity for error in the data or its use, and *errors tend to accumulate* through the life cycle. The findings from the assessment trace this sequence of 18 processes to help identify risks.

The risks for introducing error appear to be concentrated in the early and later stages in this life cycle. From our review, the higher risks are most evident in identifying program requirements; translating requirements into analytical needs; identifying data sources; designing the data base; capturing the data; interpreting, analyzing, and presenting the data; and using the data.

Considerable work to improve data quality is ongoing. We recognized this at the outset, but we did not evaluate the design or likely effectiveness of these efforts. We examined existing processes, the resulting data, and challenges in using the data today. This establishes a baseline to help evaluate the results from any of these efforts in the future.

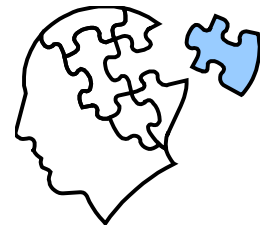
- *The pipeline safety program has invested considerable resources and effort* to document and address known data quality issues. A Data Team was chartered in 2007 to improve the quality of pipeline safety data. By December 2007, it had documented 13 pages of known problems and was soliciting input on other areas. The program has developed regional review teams for incident data, created and staffed a Performance Evaluation Group with six professional analysts, staffed

each region with accident investigators, developed a rulemaking to modify data reporting and new forms and instructions for data collection, and completed a year-long effort to validate seven years of enforcement data. The pipeline safety program has advanced its risk modeling, developed new procedures for investigations, and started development of program logic models for all of its programs to help identify information requirements. Still, some significant challenges remain.

- *The hazmat safety program has made more limited progress* in evaluating or addressing data quality issues. The program added three new analyst positions over the past two years, but has not otherwise allocated substantial resources for reviewing or modifying its data collections. One project under the Hazardous Materials Cooperative Research Program – *Incident Data for Root Cause Analysis* – produced disappointing results in the first draft. However, the hazmat safety program has developed systems for identifying missing reports and data errors. And it has invested considerable resources in development of the Hazmat Intelligence Portal (HIP) as a data warehouse for hazmat safety data, with a suite of tools for presenting data summaries—that has been well-received by data users in both pipeline and hazmat safety programs.
- *The IT program review addressed data quality in a general way*, and recommended establishing a data governance structure as a high priority for the agency. The CIO has begun the process of developing a data governance structure, and developed a broader Data Management and IT Modernization roadmap to support data quality improvements.

Identifying program requirements—*what do we need to know?*

Safety data need to be, above all, *relevant* to the needs of decision makers in managing and executing their programs. This means getting the concepts right from the outset. Information needs should drive our analyses, which in turn should drive our data collection. When requirements are not well-defined or too vague to be useful, important questions ultimately can't be answered with the data we have. In fact, we found that many of the data quality problems we see today might be traced back ultimately to shortcomings in defining what we need to know.



The DOT Information Quality Guidelines (*Guidelines*) suggest that data system objectives should be written in terms of the questions that need to be answered by the data, traceable to user needs, updated in partnership with key users and stakeholders, documented, and made available to the public. We found some strengths in identifying program requirements, and several shortcomings.

- *We have no comprehensive requirements documents for our safety data collections.* We have very detailed requirements for our *systems*, but not in terms of the *information needs* of the programs. We have a series of Federal Register Notices explaining *changes* in our data collections and providing very general objectives for the information collected on each form, but no basic outline of our overall requirements and how they fit together. The lack of documentation means that much of the history—why we're doing things the way we're doing them—is in a few people's

heads, scattered among many different papers, or lost. That creates significant challenges in interpreting the data in our analyses. It also reflects a deeper problem ...



- *We haven't thought through what we need to know in a systematic way.* Some past efforts attempted to apply some structure to the definition of information needs; some of these were not completed, and others were overtaken by growth (requirements “creep”) in what people wanted from the data systems. Managers, major users, and system owners have substantial difficulty re-constructing or describing data requirements in common terms, and in terms of questions that need to be answered—often by several different users. Our data programs *generally* address safety, but decision makers express frustration with their ability to get useful information from the data. Many who are deeply involved in our data collections and analysis believe this gap is one of the core challenges to improving data quality.
- *At a strategic level, we can probably distill three basic things we need to know* to manage our safety programs effectively:
 1. *Where is the safety risk (probability, consequences, and exposure)?*
 2. *Where are the critical points where we might address risk effectively?*
 3. *What really works in reducing risk?*

At a more operational level, we need to determine—among other things—how and why things fail, why some failures result in more serious consequences than others, what to regulate, how to regulate (or otherwise intervene), how to target our resources, who to inspect and how frequently, where to focus our efforts, how much to penalize non-compliance, whether our programs are having the intended effect, and which programs work best in what situations.



- *We have tended to concentrate more of our efforts on risk evaluation than program evaluation.* This is probably a common emphasis for safety programs, but there is a hidden hazard: “The most common failures in problem solving stem from the tendency to leap straight to action” (Sparrow, 2000). Understanding risks, even in great detail, might be futile if the program interventions we apply are ineffective. More regulations, more inspections, more training, more procedures, etc. could even be counterproductive in some cases. The discipline of program evaluation can serve as a feedback loop to help discover flaws in design or implementation of our programs, identify external factors and unintended effects, and help assess the value and impacts of a program. Today, we do not have any significant capability for program evaluation in the agency. This means we are operating with limited information to answer the third strategic question—*what really works in reducing risk?*

The most common failures in problem solving stem from the tendency to leap straight to action.

- *The Regulatory Craft* (Sparrow, 2000)

- *We have a growing understanding of safety culture* and the upstream organizational processes and circumstances that lead to failures. Over the past two years, the agency has taken a lead role in working with the pipeline industry and other agencies to explore safety culture and its relationship

with process safety. We *know* we need to know more about this. But our questions still reflect an early stage in the learning process, with considerable work to do before we can connect safety culture with our data systems.

- *We have relatively strong history of tracking safety outcome indicators* that are tied directly to DOT's strategic goals. At a very high level, we have invested considerable effort over the years to developing and refining the concepts driving our performance measures—which we use in guiding priorities for the agency, justifying budget requests, and reporting to Congress. However, we have already recognized some significant shortcomings in these indicators (discussed later in the findings), and we have made limited progress beyond monitoring these indicators.

Building a High-Performing Government:

A reformed performance improvement and analysis framework will switch the focus from grading programs as successful or unsuccessful to requiring agency leaders to set priority goals, demonstrate progress in achieving goals, and explain performance trends.

- *Analytical Perspectives*
Budget of the U.S. Government, FY 2010

- *We have limited understanding of the safety trends we are seeing.* When our safety indicators reflect unexpected trends or emerging problems, we know we need to understand *why*, but we do not have a well-developed analytical program to help us answer the question. This is a significant gap in view of the overall scheme of performance management, which is focused on setting goals, demonstrating progress, *and explaining performance trends* (FY2010 Budget: *Analytical Perspectives*; and PHMSA Strategic Plan: *How We'll Manage Our Work*).

- *There are several "invisible risks"* (within our statutory authority but not necessarily regulated) where we have little/no risk data—for example:
 - LNG facility incidents (exempt from incident reporting, subject to change in a proposed rule),
 - hazmat incidents in the maritime mode (particularly in intermodal containers),
 - certain low stress pipelines in rural areas,
 - bulk loading and unloading of rail tank cars,
 - non-jurisdictional failures that are tied to jurisdictional pipeline systems,
 - exclusion of state/local governments (e.g., highway maintenance) from one-call reporting,
 - failures of DOT packages, cylinders, or containers "outside transportation,"
 - tank truck wetlines (not coded),
 - gas pipeline master meter operators,
 - LP gas systems,
 - hazmat response preparedness and effectiveness,
 - greenhouse gas emissions (all releases) from pipelines,
 - environmental effects from hazmat releases, and
 - hazardous "materials of trade."

Invisible risks ... are non-self-revealing problems—issues that either by conscious design or by a quirk of their nature are not adequately represented in the organization's process workloads. These problems do not present themselves; if an agency wants to control them, they must first deliberately uncover them.

The heart of the analytic challenge for invisible risks is to help agencies avoid the circularity trap, in which they fish in the same parts of the river day after day because that is where they caught fish before.

- *The Regulatory Craft* (Sparrow, 2000)

In some cases, we have explicitly exempted certain operations from reporting; in other cases, we might not have fully considered the potential risks or the benefits of casting more widely for failure data so we can understand the risks before a big accident occurs. Lacking good data on these risks, we can't quantify the risks or address them effectively.



- *Our regulatory evaluations generally begin too late in the process to affect decision making.* We appear to have reasonably good estimates of the costs and benefits of our preferred approach in rulemaking. In both of our operating programs, we have engaged strong contract support and we have an iterative process to develop and review regulatory evaluations as outlined in OMB requirements. This tells us generally whether our intended approach is economically-viable—as one element of the decision process. However, generally we have developed rulemaking proposals *before* we had a good estimate of costs and benefits, or a real understanding of the *alternatives* that might address the risks. The proposals have become the preferred alternative by default. Across several rulemakings in both programs, our regulatory evaluations have served to justify the decisions we have made rather than as an input into the decision making process. Both operating programs have recognized this gap, and have acknowledged the need to demand earlier input into the process.
- *We don't know what our rules cost or what benefit the public gets* because we don't assess the *actual* costs and benefits after implementation. Section 610 of the Regulatory Flexibility Act requires agencies to review certain regulations within ten years of their adoption. But the scope of this requirement is limited, and in practice our evaluations under Section 610 are limited. At the same time, this might be one of the easiest kinds of program evaluation we could do; we have a clear “before” picture and estimates of what we expected to happen. Retrospective analysis could help us redirect programs where we found unexpected consequences, refine our estimates for future benefit-cost analysis, and generally better understand what's happening in the regulated industry.
- *We sometimes frame our questions about risk in overly general ways* (e.g., “we need to find out everything we can about the risks ...”). This approach can lead to “slicing and dicing” the data, producing analytical results or simple tabulations that are disappointing and not very useful for decision making. Some extensive compilations of hazmat statistics and some targeted analyses of particular hazmat risks reflect this problem. The common theme seems to be this: decision makers often aren't sure exactly what they need to know about a problem until they begin to explore it. This can be perceived as a “bring me a rock” approach, which seems inefficient. It actually reflects *a shortcoming in our analytical capacity.*

The objective is not to produce research; it is to produce insight with a view to action.

*- The Regulatory Craft
(Sparrow, 2000)*

An analytical-deliberative approach to risk evaluation requires a partnership between decision makers and analysts to help formulate the questions in a way that they can be analyzed effectively (*Understanding Risk*, NRC Committee, 2000).



- *Over the years, we have assumed most of our programs are effective (or not effective) with no clear analytical basis for that assumption*, and—from recent evaluations—substantial evidence that some of our assumptions were wrong. Of course, program managers have a strong advocacy role for their programs, and conventional wisdom exerts a strong pull. At the same time, there are strong incentives in government to change our processes and create new initiatives.

Well-run programs reflect a curiosity—even a skepticism—about program effectiveness that can drive continuous improvement while tempering the impulse to abandon existing programs without understanding their value. There is certainly a growing awareness of this need within the agency. But our resources and capability for credible program evaluation still lag. This also presents a risk at the front end as we undertake new initiatives ...

- *When we design/implement new programs, we build-in limited capabilities for evaluation*. We often pilot test new approaches to assess workability, and build in data collection to assess implementation. These can provide useful checkpoints. But generally we do not build in the performance measures and data collection we would need to evaluate effectiveness and impacts.

The hazmat safety program uses re-inspections to re-assess compliance. The pipeline integrity management program uses inspection and other data to track implementation. But neither of our inspection programs includes, for example, a random component (like comprehensive IRS audits). The principal value of random audits is that they provide information about types of non-compliance that existing targeting strategies miss (Sparrow, 2000). Generally, we are not anticipating this sort of program evaluation.

- *We lack some of the analytical skills/expertise needed to focus our data collection*. For example, our compliance program fundamentally is aimed at influencing *company behavior*, but we have no social/behavioral scientist positions in the agency to help guide our efforts (and particularly our data collection). Our regulatory evaluations are aimed at evaluating the costs and benefits of alternative approaches, while we have no economist positions to guide/evaluate the work of contractors or provide feedback into our data collection. We have almost no professional experience in the specialized discipline of program evaluation, and no positions requiring these program evaluation skills.

Translating requirements into *analytical needs*:

No data system can answer directly the program questions we might ask—that is a function best done by program analysts who understand how to gather relevant data, interpret the data to provide meaning, analyze the data to produce program information, and present the data and analytical findings for use in decision making. This suggests an important translation function that needs to be in place in order to clarify the concepts and get the right data in the first place. Without it, we have no analytical framework for developing data requirements, and the resulting data are unlikely to be useful in answering important program questions.

The *Guidelines* suggest translating basic information requirements into indicators that need to be measured. But the more general need here is to bridge the gap between general information requirements and detailed data requirements. This is a critical process in the life cycle of data quality. At the same time, we have been lacking a core analytical capability for many years, so there is effectively little communication between decision makers and data collectors.

- *The requirements documents that we have for our information systems are far too complex* for managers to understand their significance, and they don't really bridge the information needs to data needs; they generally address only system performance requirements. In system development, we tend to go from broad outlines of need straight to data base design.



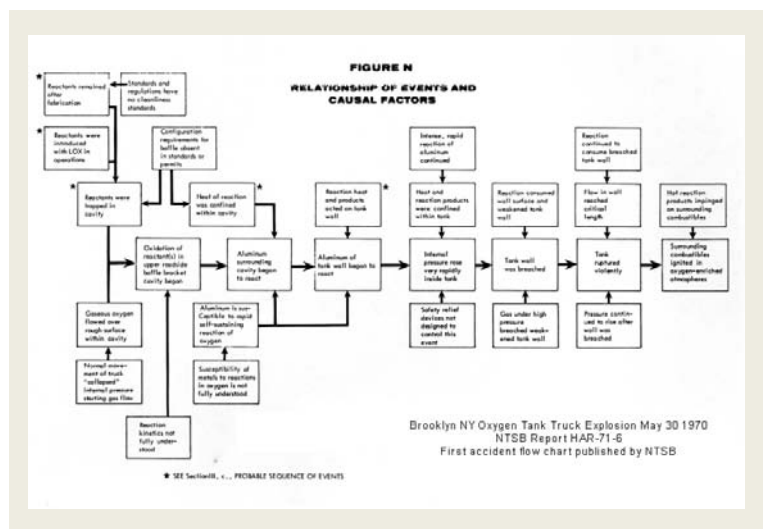
- *Our data systems don't provide a place to capture all the latent conditions that lead to accidents.* Social science research over the last 30 years has highlighted the depth of interactions between human and organizational factors that contribute to major accidents. In *Managing the Risks of Organizational Accidents (1997)*, James Reason explains:

"Latent conditions are to technological organizations what resident pathogens are to the human body. Like pathogens, latent conditions—such as poor design, gaps in supervision, undetected manufacturing defects or maintenance failures, unworkable procedures, clumsy automation, shortfalls in training, less than adequate tools and equipment—may be present for years before they combine with local circumstances and active failures to penetrate the system's many layers of defenses. Latent conditions are always present in complex systems."

In fact, we often look for and *find* these latent conditions in the course of our business. But we don't convert them into *data* at the point of an incident investigation or inspection.



- *We don't have a good conceptual model for understanding failures.* We don't capture *the chain* of failures, including especially the root causes, that typically are associated with any accident or incident; we don't capture all the relevant circumstances and interactions that might (through statistical analysis) reveal hidden explanations; we don't capture inspection deficiencies in a way that would allow us to tie together our inspections and accident investigations—by getting at common explanations for failures. Most analyses struggle with the data to find patterns and meaning, but they are severely limited by the basic conceptual models.



The risk literature provides several possible models to consider. In *Barriers and Accident Prevention*, for example, Erik Hollnagel describes three of these models—sequential, epidemiological, and systemic accident models—and discusses their relative value in helping to understand failures with a view to action. Without a complete picture of causes and circumstances, we don't understand the full extent of the problems (e.g., human error, corrosion, etc.) or the interrelationships of causes, and we have greater difficulty identifying critical control points and targeting the risks effectively. The solution to this is not at all trivial, and might include both data structure and narrative. But this is probably a major risk in our data quality, with potentially broad implications.



- *Our primary outcome measures do not really reflect changes in risk over time.* The numbers are now too small to find meaningful patterns in the data. We must also deal with the natural tension between the need to monitor outcomes and the desire to attribute outcomes to what we do. We have begun work to make better use of our data—to identify risk factors that might be used to better monitor risk, but performance measurement “presents formidable intellectual challenges that have never been solved in a way that provides clear guidance to practitioners” (Sparrow, 2000).

Outcome data provide an unreliable indication of a system's intrinsic safety.

This is especially the case when the number of adverse events has fallen to some low asymptotic value around which small fluctuations ... are more noise than signal.

- Managing the Risks of Organizational Accidents, Reason (1997)



- *We have little data that might be used to identify emerging risks or leading indicators* of performance. Incidents are often considered to be lagging indicators, in the sense that they tell us what has already happened, while people want to understand what is likely to happen in the future so we can target these risks effectively. This might include a turning point in company performance or investment, a new risk from new materials or processes, or a change in external factors affecting the systems.

It's a significant challenge to anticipate these risks. One of the tools we have is situational awareness—watching carefully for trends and using professional judgment to focus attention. For this to work well, it's important that we translate *individual* learning into *group* learning (by sharing lessons learned, which we do), and ultimately into *organizational* learning (by turning judgment into data where possible, which we don't really do in any systematic way now). With both failure data and risk exposure data, a set of statistical indicators might provide early warning of risks that could help in managing program priorities.



- *We have limited data on risk exposure*, making it very difficult to identify and evaluate relative risks. NTSB, in its report on *Transportation Safety Databases* (September 2002), highlighted the need for exposure data in understanding risk. Exposure data can be useful also in helping to forecast changes in risk, and comparing safety across modes. In pipeline safety, we have some exposure data with the identification of high consequence areas, but some challenges in matching product

throughput with location. For hazmat safety, better commodity flow data—including route information—could be helpful especially for response planning.

- *We don't have a good way (yet) to characterize low-probability, high-consequence risks.* Many kinds of risks we can see in our incident data over a baseline of about 20 years. They occur with enough frequency to allow us to estimate future frequency with some confidence. We can't do this with Low Probability High Consequence (LPHC) events; by definition, they are infrequent. More importantly, we simply don't accept any realistic probability of a *very* high consequence failure. This is a key area where the standard risk model does not work for us, and where historical data cannot provide an adequate baseline.

While we recognize this problem generally, our analyses often mischaracterize LPHC risks by projecting directly from incident histories. Some recent regulatory evaluations provide some examples of how we might do this better. We need to identify outliers in the data (e.g., the consequences of Hurricane Katrina in 2005) and spread out the effects of these over a longer period of time, and we need to estimate the probability and consequences of events we can envision but that *don't* appear in the data, and add these into our analyses.

Defining *data needs*:

Data collection is expensive. We cannot collect all of the data we might ever want or need, so we need a systematic process for prioritizing what we will collect—tied to our analytical needs. We have a lot of experience in developing and refining our data needs (despite the lack of clear analytical needs). In practice, this has become largely an incremental approach to modifying the data we already collect.

Data needs should include requirements for accuracy, timeliness, comparability, etc. Errors at this stage of the process can result in significant data gaps or, conversely, significant costs that are not justified.



- *Our failure data focuses on the top layer of a much larger pyramid*—we record 60 excavation damage incidents/year while operators record over 100,000 excavation-caused leaks on pipelines; we have data on fewer than 100 lithium battery failures aboard aircraft over 17 years, while these batteries spend a miniscule fraction of their lives aboard aircraft (i.e., there must be thousands of battery failures); we record 17,000 hazmat incidents/year while most hazmat spends a small fraction of its life in the transportation system; we have no data on close calls (or near-misses), which are believed to represent from 10-to-1,000 failures for every one injury incident.



Figure 1 - From ConocoPhillips Marine (2003)

Small numbers make it difficult to detect meaningful trends or prove the effectiveness of our programs. At the same time, we are missing information about many of the precursors to larger failures. Other safety agencies (FAA, Coast Guard) have recognized the value of near-miss reporting systems and expanded collection of failure data, including the possibility of finding out *what prevented* many failures from becoming more serious incidents. There can be a significant

costs and challenges in expanding our data collections, and there are important differences in the operating environment in different modes that need to be taken into account. But there are also potentially valuable opportunities in expanding the scope of the data we consider in risk evaluation.

- *Our incident reporting criteria are not based on an analytical accuracy requirement.* We use different incident reporting criteria across our programs without a clear *requirements*-based rationale for the difference. Both programs average about 10-20 deaths per year, but we collect 17,000 incident reports for hazmat and fewer than 1,000 reports for pipelines, based on independently-developed reporting criteria. For hazmat incidents, any unintentional release must be reported; for pipeline incidents, reporting requirements are based on the amount released (5 gallons or more) or severity of consequences (death, injury, damage >\$50,000). From an analytical view, we need enough data to determine patterns in risk with reasonable confidence—an accuracy standard that should drive our choices for data collection.

This issue is complicated considerably by other data accuracy issues (discussed in other findings), including significant underreporting of some incidents, bias in reporting from the regulated industry, changes in the dollar value of lost product from pipelines, and weak coding schemes that don't help us answer the most critical questions. Both programs are in the process of reviewing the reporting criteria in ways that might bring them more in line with each other, but these other issues need to be addressed as well.

- *We don't capture some of the most important incident consequences in a useful way.* For example, we distinguish the severity of injuries based on in-patient hospitalization, but this is not sufficient to estimate the potential benefits of new programs in our regulatory evaluations. We collect data on estimated costs associated with property damage and emergency response, but (for liquid pipelines) this has been limited to costs reimbursed by the operator; this limitation, in fact, is inconsistent with the basic reporting criteria, and results in an underestimate of total damages. The problem would be corrected with new reporting forms that have been proposed, but even so we will have a data comparability issue with the data we have collected through 2009.



- *We don't capture consistent data on some key risk factors like fire/explosion.* For hazmat incidents, we require reporting (with some key gaps, described previously) for every unintentional release. For each of these releases, we ask for information on the presence of fire or explosion. And from an analysis of these data using conditional probabilities, we have found that in the highway mode, the presence of fire alone increases the probability of death or major injury by a factor of 340-to-1 (34,000 percent). This is an astonishing number, and potentially very useful in targeting intervention strategies. We can't do this same analysis effectively for pipeline safety, because we

Conditional Probabilities of death or major injury, given two risk factors

Highway Mode	Ignition	No Ignition
Gas Dispersion	16.67 %	1.70 %
No Gas Dispersion	22.60 %	0.07 %

don't capture the first condition—a release—or the second condition—a fire—consistently. The reporting criteria for pipeline incidents depend on the consequences, which of course are affected by the presence of fire/explosion in the first place.



- *Our treatment of the human element is particularly ambiguous as a causal factor in incidents.* The human error literature provides ample evidence that human error itself is not a useful way to characterize cause. Nearly every failure can be traced to human error at some level. At the same time, there are always other factors and circumstances like time and operational pressures, organizational culture, system design, clarity of procedures, etc.

The adaptability and flexibility of human work is the reason for its efficiency. At the same time it is also the reason for the failures that occur, although it is never the cause of the failures.

*-Barriers and Accident Prevention
(Erik Hollnagel, 2004)*

In the pipeline safety program, incorrect operation is one of eight cause codes, and (according to the instructions for reporting) it can include human error or faulty procedures; at least two other cause codes (excavation damage and other outside force damage) might also involve human error, although the data won't show this further level of detail. In the hazmat safety program, human error is one of 37 cause codes, but the concept is embedded in several other codes (e.g., improper preparation, inadequate preparation, inadequate training, overfilled) that suggest individual fault. Over-attribution to the human element can lead to program interventions (like more training) that do not address the root cause of failure.

- *Reporting lags inhibit the effective use of some of our data.* The key safety indicators we track in the annual Performance and Accountability Report to Congress require estimates in October for the previous fiscal year. But we routinely revise and add to our hazmat incident data for months, and on a trickle basis up to several years, after any reporting period. Recent research conducted by program staff has shown that about 10% of the serious hazmat incidents are reported more than four months after the incident. In pipeline safety, annual reports from pipeline operators do not coincide with our planning annual pipeline inspections. Incident reports can be updated many months after the incident occurred (and many months after better information is available). Our own enforcement data can lag for months as a case is processed. In both programs, we have not developed a timeliness requirement and applied it to our data collection and forecasting.

Identifying data sources:

If existing data can be found that address (or, with some modification, could address) our data needs, this is usually the most efficient approach to getting data. But we also need to consider the quality, timeliness, comparability, relevance, and utility of the data we might get. And if we use existing sources of data, we need to make sure we have common identifiers (like company ID, for example) to integrate the data with our own data systems. We need to understand the potential biases in data reporting, and the consistency of the concepts (e.g., *what is a shipper?*, or *a failure?*).

The *Guidelines* suggest casting the net widely to consider possible sources of data and processes for collecting it after we have identified data needs. In general, we consider other potential sources of data, but in practice we generally default to the way we have always done it.



- *Most of our data collection relies on third-party reporting from regulated companies.* This is convenient, and it goes directly to the source. It also introduces *serious biases and gaps* in the data we collect. Despite the best intentions and professionalism, the regulated industry has an institutional bias (and probably a liability aversion) in determining the causes, circumstances, and consequences of failures. Accident investigations—the limited number that we do—have shown some significant differences between what a company reports and an objective view of these events. Reports from companies also reflect large numbers of blanks and “unknown” data, particularly in the most serious cases—exactly where it is most critical that we have good data. Our collection of system data (as in the annual reports from pipeline operators) is further constrained by the need to minimize reporting burden on the industry, so much of the data are aggregated to a level that cannot be used in risk evaluation. An alternative approach—collecting much of the data ourselves in the course of our inspections or investigations—has been discussed but never evaluated fully.

We have ample authority to collect data directly as part of our inspections or accident investigations, but many in the organization see data collection as a distraction from more important safety oversight activities. There is also an ownership issue with the data ...



- *There is a historical understanding that the data we get from industry is “their” data.* Even when we believe (or know) data to be wrong, we don’t modify our data until we get revised reports from the company. Even now, as we recognize the need for more accurate information, we generally *augment* the data with our own information rather than modifying the basic data in our system. This practice, however, creates ambiguity in the data that analysts might use, expand the opportunity for misinterpretation, and doesn’t really solve the problem.

Autonomy and Independence as Constraints on the Regulatory Process

Regulators, for their part, attempt to penetrate the boundaries of the regulated organizations by requesting certain kinds of information and by making periodic site visits. But these strategies can only provide isolated glimpses of the organization’s activities. Size, complexity, the peculiarities of organizational jargon, the rapid development of technology and, on occasions, deliberate obfuscation all combine to make it difficult for the regulator to gain a comprehensive and in-depth view of the way in which an organization really conducts its business ...

In an effort to work around these obstacles, regulators tend to become dependent upon the regulated organizations to help them acquire and interpret information. Such interdependence can undermine the regulatory process in various ways ...

- *Managing the Risks of Organizational Accidents (Reason, 1997)*

The data we disseminate for analysis should reflect our best understanding of reality. We can certainly keep a separate file of reports that have been submitted by companies if anyone wants to see those (under FOIA or otherwise), but we should differentiate *reported* data from *agency* data and encourage analysts to use agency data that we can vouch for. We don’t need to disseminate a data base of reported data; those data do not meet DOT’s information quality guidelines.



- *Our own independent accident investigations are very limited in number and scope.* We have completed 19 investigations (about 3%) of the 664 reported pipeline incidents in 2008, and about 40 (or 0.5%) of the 8,000 reported hazmat incidents over a six month period in 2009. More generally, the information we get from our investigations is not converted into data that could be used for statistical analysis or engineering reference. We often collect more data during the course of an investigation than we require in the incident report from an operator, but this information does not get entered into any data base. It appears to be collected primarily for enforcement purposes related to individual companies, not to build our knowledge base. Hazmat investigation reports, in particular, are almost indistinguishable from inspection reports—both use the same form, and both are focused on the identification of violations. The incident is simply a *trigger* for an inspection.
- *Our processes do not effectively reconcile discrepancies between our investigation reports and the accident reports submitted by operators.* The discrepancies can be significant. In one case, a pipeline operator reported \$0 damage (and that is what we showed in our data); the investigator reported lost product, a fire, and an estimated \$588,000 in property damage, but the data base was not updated or corrected until 9 months after the incident. In other cases, the data base was not updated to reflect design pressure, operating pressure at the time of the accident, or year of installation—from the more detailed investigation reports. These discrepancies are just from a cursory review of the 10 most recently closed cases.




- *We have difficulty integrating data* from police and fire department reports, reports to the National Response Center, data from CDC, and many other systems. We can't easily use the data from most other data systems because we lack common identifiers (in many cases), and many of these external systems are not sufficiently transparent to allow us to understand their limitations. These are common problems in safety programs government-wide, and data integration is often encouraged as one of the best ways to leverage resources. But we still don't know about all the relevant systems that might be out there; we regularly find other systems through serendipitous interactions with other agencies and organizations.

Designing the DB—*data elements and relationships*:


A strong data architecture—defining standards for data elements; what values are allowable; how data elements, records, and files relate to one another—is the first technical step in creating a *useful* data base for analysis. One of the biggest challenges in data base design is ensuring consistency or comparability of data, especially with changes over time or with data that are collected through different systems. Comparisons are the essence of measurement and analysis. Good standards can help provide the basis for comparability.

The IT program review completed in 2008 identified a series of projects to strengthen our information architecture, including the need for a data architecture and management plan and a strong technical architecture. This was aimed at helping overcome many shortcomings in our data.



- 
 • *We often have built new data systems to meet pressing requirements without a view toward integration* of these systems with our existing data. The pipeline safety state grants program has a self-contained application and data base for state pipeline safety activities; it was not built with the same data architecture as the system we use for Federal pipeline safety activities. Data bases for pipeline safety integrity management inspections and the hazmat information center illustrate the development of prototypes that have become production systems, without a common data taxonomy and with non-standard technologies and tools. As a result, these data bases have sometimes proven difficult to reconcile/use with other safety data. There are legitimate program needs for rapid solutions that cannot always be met with existing data systems. But there is a missing data architecture that could provide the basis for standardization and smooth integration.

The most valuable database will be the database of databases, which will grow as analysts accumulate knowledge from a variety of sources ...

- The Regulatory Craft (Sparrow, 2000)

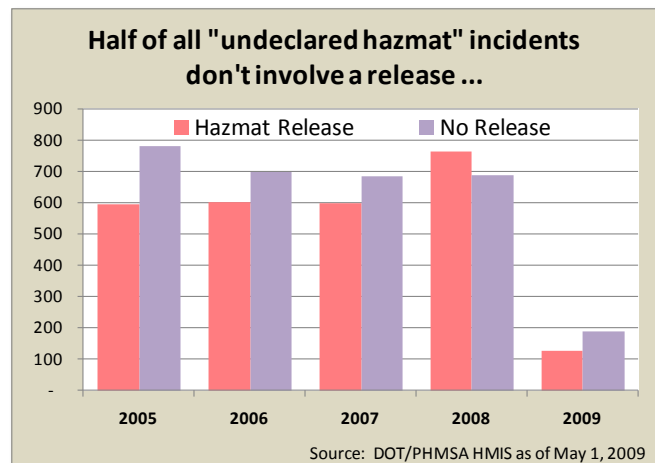
- 
 • *We cannot accurately and consistently identify companies*—companies often use multiple IDs for different purposes; entities change through mergers, acquisitions, and partnerships; and companies might be related through several corporate layers. We have limited mechanisms in place to track company relationships or changes over time. As a result, we can't measure operator performance consistently or target our resources most effectively based on comparative risks; performance histories are sometimes fragmented.

The proposed "One-Rule" in pipeline safety includes an attempt to resolve this issue for pipeline operators, but there is not yet a corresponding effort to address it for hazmat shippers and carriers. The Hazmat Intelligence Portal, for example, returns 165 companies from a search on "UPS." As a result, performance histories can be fragmented and misleading. We have incomplete data on shippers, because the hazmat registration program was not designed to capture all shippers (many/most are exempt). For motor carriers, in particular, there is an even broader problem of multiple DOT numbers issued to the same company—limiting our ability to assess the fitness of operators from FMCSA safety data.

- 
 • *Inspection deficiencies are not captured in a way that can be tied to the causes of incidents/failures.* Deficiencies are coded to capture the regulation that was violated; they are reflections of *non-compliance only*, and we don't collect data on other safety weaknesses (e.g., all the kinds of latent defects and safety culture issues identified previously). Incidents are coded to capture the component that failed and the general cause of the failure; there is no direct data connection with the regulations or non-compliance. "Without an ability to connect process indicators (safety deficiencies) with outcome indicators (failures), we have limited ability to impact outcomes by regulating processes" (Von Hermann, 2007).
- 
 • *Incident cause codes cannot deal effectively with multiple failures or sequences of failures.* Hazmat incident reports permit entry of multiple cause codes, but these are not tied to what failed, how it

failed, or in what sequence; they are simply listed, and often in duplicate within a single record. Pipeline incident reports permit only a single primary cause with one additional level of explanatory codes (e.g., corrosion – internal or external).

- *Hazmat incident data can include multiple records for each incident.* In cases where an incident involved different commodities or multiple tank cars, the incident data base includes a separate record for each release. Given the design of HMIS and technology limitations when it was developed, this was probably a reasonable compromise to make sure we got data for each failure. But it can result in misleading tabulations or analyses if an analyst is not familiar with the way records relate to incidents. There are alternative data base designs that might be explored, and/or better metadata might be provided for analysis.
- *Hazmat incident data are compromised by the inclusion of some violations that are not “incidents” in the common understanding of the term.* In 2005, incident reporting criteria were modified to include the discovery of undeclared hazmat. These now comprise about 8% of the total reported incidents in our data base, although about half of these do not indicate a release of hazmat or any other criteria for incident reporting—no deaths, injuries, property damage, fire, explosion, evacuation, closure of a transportation artery, or any other consequences. The inclusion of these data here increase the potential for error and misleading conclusions in conducting safety analysis.



- *We don't capture data from states in a form that is comparable to the federal program*—limiting our ability to evaluate the effectiveness of state pipeline safety programs or operators and systems that are inspected by states. States use our data systems when they are acting as interstate agents for pipeline safety, but this represents a small fraction of their work. They do not report their inspections of intrastate gas pipeline operators with any detail (they report the number of inspections conducted). As a result, we have very limited information about the condition of 80% of the national pipeline system—where about 80% of the incidents involving death/injury occur.

Designing the data collection—*forms and instructions*:

The forms and instructions for capturing data provide the outline of what we expect, a visual reflection of the logic, and interpretations to help clarify what we want. Good data collection instruments help minimize errors and missing data by making the process easy. They are communication tools that can have a high, direct impact on the quality of data.



Data collection methods need to be appropriate to the complexity and size of the data collection, data requirements, and the amount of time available. The *Guidelines* suggest electronic reporting where possible, logical sequencing of the data, minimal need for calculations or conversions by the reporting source, clearly posted procedures, and tracking processes. This stage of data quality is one where we have invested considerable time/attention, and have fairly good systems set up.

- *Reporting forms and instructions are generally clear, logically organized, and posted online.* The PHMSA website clearly points visitors to the reporting forms. The forms appear to walk through the data we have asked for in a logical sequence.
- *Electronic reporting has increased significantly, and helps automate edit checking.* About 50% of hazmat registrations and incidents are reported online; another 40% of hazmat incident reports are provided in machine-readable form. Most pipeline incident reports are filed online (about 95% for hazardous liquid, 65% for gas transmission, and 40% for gas distribution pipelines), and we have proposed mandatory electronic reporting. In both programs, we also use these same systems for entering data from paper forms, to take advantage of the built-in edit checking.
- *Summary data on pipeline systems (in the annual reports from operators) requires aggregation* of detailed data—generally asking for the total number of pipeline miles with certain characteristics (e.g., by year of installation, coating, cathodic protection, diameter, onshore/offshore, miles inspected, etc.). Aggregation reduces the volume of reported data dramatically, but introduces opportunity for error and—more importantly—sacrifices the ability to see combinations (e.g., coated, cathodically-protected, installed in the 1970s) that might be useful in evaluating relative risk. As a result, we do not have exposure data at the same level of resolution that we capture in incident reports. Getting these data through GIS systems, which many companies maintain, might provide the cross-sectional data that would be most useful.
- *We have limited processes for tracking the status of reports.* Pipeline incident reports can be filed as original, supplemental, or final, but there is no requirement that reports ever be finalized (and some companies have refused to do so) and no tickler to track this. For more significant incidents, regional review teams follow up with the operator to address potential errors in incident reports that are flagged in our edit checks; a system tracks the review process, but there are some inconsistencies across regions in how this is used. Hazmat incident reports can also be filed as an original or update, but there is no followup tracking after the first report is filed. Hazmat registration includes followup in subsequent years, but this is used only for the purpose of generating reminder letters, not tracking submission.




Capturing the data—*what’s happening?*

At some point, observations become data. Ideally, the data reflect accurately what is really happening. In practice, there are all sorts of things that can go wrong at this stage—



errors (or omissions) in measurement, misunderstanding and biases affecting what we observe, memory failures, missing evidence, errors in communication, etc. We are relying on people here, and people are notoriously poor witnesses. Company procedures can add another layer.

The *Guidelines* suggest formal training for observers. Where our own inspectors or investigators are collecting information, we provide some fairly extensive training in how to look for things. But most of our data come from third-party observers (or collectors)—mostly from the regulated industry—and this presents some special challenges.

-  • *There is a natural, inherent bias in reporting from the regulated industry.* When most failures result from multiple causes and operators are asked to report one, or when the primary cause is ambiguous, it would be natural to report the failures of others or natural forces first. When legal liability is at issue, it would be natural for operators to be more cautious in reporting “facts” when they can choose “unknown” or leave a data field blank. It would also be natural for operators to choose the low end of an estimate of a release and the high end of an estimate of recovery. These would not be not false reports, but they might not be objective representations either. We have very little research to help estimate the extent or magnitude of reporting bias, but some clear indicators that it exists.
-  • *Even when we do accident investigations, we generally do not develop the root causes of these failures*—despite extensive training (in the pipeline safety program, particularly) in root cause investigations. In reviewing the ten most recent investigation reports for each program, only one (an investigation of multiple hazmat incidents) provided any insight or conclusions on the root causes of failures.
-  • *We capture only limited data from our inspections.* During the course of an inspection, we might observe many things worth commenting on but not rising to the level of a violation—maintenance issues, clarity of records, concerns about training, etc. This presents a useful opportunity to build a broader performance profile, to capture the kinds of little failures that sometimes lead to bigger failures, even to help create leading indicators or a body of good practices. But—except for Integrity Management inspections—we don’t capture the broad range of inspection findings as data. We haven’t set up our data systems to create a home for the data. As a result, we have limited ability to rank operator or system performance from our inspection data.
- *Inspections also present an opportunity to capture more data on the systems we regulate.* The annual reports from pipeline operators provide only very aggregate information on their systems, with no information on any cross-section of pipe. During our inspections, we often have this information available to us, and we could use it to build a more detailed system profile for risk evaluation.

Processing raw data to prepare reports:

Processing raw data and preparing reports might appear to be simple, straightforward steps with limited opportunity for problems. But when we rely on the regulated industry for so much of our data, there is

a decision point here with some substantial consequences—by choice (or omission), much of the data we should get is not compiled into reports and sent to us. This creates significant data gaps. And when the missing data is not random, these gaps can create a substantial bias in the data affecting the basic reliability of our estimates and the soundness of our safety decisions.

- ➔ • *Incidents might be substantially under-reported*—preliminary analysis of hazmat incidents suggests we might be missing 60-90% of all reportable incidents. Some of these missing reports might be the result of a lack of knowledge about the reporting requirements (these tend to be smaller companies); some might be the result of a decision to avoid reporting. The effect is the same. The most troubling aspect of this is that the missing incidents appear to be *different in kind* (different patterns of causes, circumstances, consequences, etc.) from the reported incidents, so projecting from what we know might be giving us a *distorted* picture, as well as an incomplete one. We believe we have much higher reliability of reporting for pipeline incidents—since pipeline operators comprise a much smaller, identified universe—but we don’t have an analytical basis to demonstrate this.

- ➔ • *Incident reports are often missing important data.* A significant number of reports indicate *Other/Unknown* causes—and the prevalence of these causes is higher for the most serious incidents—exactly the cases where we need complete and accurate data most. Program staff suggest that concerns about legal liability could be motivating operators to limit their reporting in these cases. But here the missing data are not random. So the resulting data and the patterns we might find in our analyses are likely to be biased and misleading.

An analysis of blanks in our pipeline safety accident data shows that over half of the fields where we are expecting data included at least some blanks. We are missing 2.3% of all generally-expected data, and 6.8% of all conditionally-required data (e.g., if “Other”, explain). Many of the blanks might simply reflect a lack of available information; they might reflect difficulty understanding what we’re asking for; and in some cases, they might reflect zero values (e.g., no property damages). But our data don’t differentiate between “unknown” values and “zero” values. Part of the problem here is that we don’t have enough information to help analysts interpret the data we get.

- *Late incident reports compromise our performance reporting and time series analyses* for hazmat safety. In past years, our annual performance report to Congress has included multiple revisions to prior year hazmat data, and our monthly reports to agency leadership typically include revised data for up to (and sometimes more than) a year. The data are never considered final. This results in chronic underestimates of recent risk data.
- *We’re missing an unknown number of pipeline safety-related condition reports* because we have exempted the reporting of any safety-related condition that is corrected by repair or replacement in accordance with the applicable safety standards before the deadline for filing the report. That makes sense if the purpose of the report is to identify problems where we might want to intervene to correct the situation. It seriously limits the value of the report as a window into safety weaknesses we might need to know about. We don’t know what we’re missing.

Reporting data to PHMSA—*transmission*:



Getting data to us is a small step in the overall process, and one that generally doesn't affect the accuracy of the information. We continue to receive some paper reports, by mail and fax. But electronic filing is increasing substantially, so it should be getting easier to transmit reports. Still, there are at least two minor data quality issues associated with transmission of data.

- *Electronic filing requires manual entry of data.* Many larger companies have their own systems for capturing failure data, but we do not yet have XML processes for accepting pipeline safety data directly (we do for hazardous materials safety). As a result, companies must manually transcribe and submit the same data to PHMSA in another form. This increases the opportunity for error, and imposes a reporting burden that might be reduced with a technical fix.
- *Paper reports are commonly mailed to the agency*—probably with substantial delay (and an unknown risk of loss) associated with the X-ray process for our incoming mail.

Entering/editing the data—*quality control*:



Data editing is probably the most familiar step in managing data quality—the one that many people think of first in any discussion of quality. Edit checks are certainly important, and they are relatively easy to develop and implement. This can help control the entry of invalid data (e.g., text in numeric fields, numbers outside the possible range, mismatches between county and state, formatting errors). It can also help reduce missing data by flagging blanks and/or missing reports.

The *Guidelines* suggest a checking process to reduce missing reports, identification of critical/required data fields and follow up on these missing data, development of standard coding schemes for data, separate values to distinguish zeroes from unknown information (not blanks for both), an automated editing process to reduce errors, and publication of editing statistics. To deal with missing data, the *Guidelines* suggest adjusting the data with weights or imputation to reduce bias in the data, and analysis of the effects of missing data.

- *We make limited use of other data sources to help reduce under-reporting.* The hazmat safety program regularly screens media reports, complaints, information from emergency responders, and reports to the National Response Center to identify potentially unreported incidents. In 2005, for example, this process found an additional nine fatal incidents—75% more than were reported to PHMS directly. However, this created an anomaly in the data for 2005, and followup on these incidents was suspended for several years as a result. There is no corresponding program in pipeline safety. Reporting from pipeline operators is believed to be more reliable, since it is a much more concentrated industry, but a review like this could help quantify the reliability of reporting (and is suggested in the *Guidelines*).
- *Both operating programs have developed edit checking procedures to reduce blanks and errors.* Hazmat incident reports are run through an automated edit process to detect some kinds of basic errors, and flagged items for the more serious cases (involving death, injury, or evacuation) are

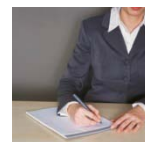
followed up by phone to verify the data. Hazmat registrations are processed directly by Wachovia Bank (which also processes registration fees); we have no information on their quality controls. Pipeline incident reports are run through a separate automated edit process to detect errors and flag them for review by region review teams and follow up with operators.

- *In practice, the quality of edit checking is mixed.* To test online reporting of a hazmat incident, we input several clear errors or inconsistencies that were caught and some that were not; found some areas where the valid entries were not accepted; and encountered some error messages that were difficult (or impossible) to correct. The interface was reasonably clear, although many terms and acronyms might require reference to the instructions, and error messages were not shown alongside the data in error. Individually, these issues appear to be relatively minor; collectively, they seem to add to the burden of reporting and tend to compromise the quality of data we get.
- *Our edit checks don't address all of the missing data.* An analysis of blanks in pipeline safety accident reports from 2002-2009 found that many fields were blank in at least some reports. Blanks present an interpretation problem for the analyst. Not counting these reports assumes that the remaining reports are a representative sample, and that might not be the case.
- *Our data bases generally do not distinguish zero values from unknown values,* because we don't usually get that information from the reporting source.
- *Our data include a large fraction of "other/unknown" values for incident cause,* particularly in more serious incidents—exactly the cases where we most need good data. In many/most cases, we get amplifying information about "other" values, which could be used (and have been used) to distill new codes. The prevalence of unknown values might be reduced with increased investigations.
- *We do not make any adjustments to the data to account for missing data.* For these kinds of data, where we expect 100% reporting, it probably does not make sense to impute missing values. But it might be very useful to evaluate the extent and effects of missing data. The results could be used to help target data collection improvements, and to help analysts understand the limitations of the data.
- *We don't consistently and promptly reconcile our data* with the [presumably better] information we obtain through investigations—including NTSB's or our own. One of the major hazmat incidents over the past 10 years—a rail accident at McDona, TX in 2004—resulted in \$5.7 million in property damage, according to NTSB's report on the incident; our data base shows \$0 in property damage. In one of three pipeline safety investigation reports reviewed as part of this assessment, the investigating officer found \$588,000 in property damage; the operator reported none, and our data base showed none for nine months.
- *We don't regularly monitor data quality indicators.* It is arguable whether the number of blanks we see in the data is a "good" sign (so few) or a "bad" sign (so many). The more important point is that we are not regularly measuring this and evaluating the overall accuracy of the data we have. We invest substantial resources in improving the accuracy of our data (particularly with region

reviews in pipeline safety, and generally with all of our edit checking). But we do not track or publish statistics from our edit checking, and we don't retain this information for analysis. From a management perspective, it would be helpful to know how much these efforts *change* the quality of our data. We could be over-investing in report-level accuracy compared to the more general problems with data gaps.

Documenting the data program—*metadata*:

Data documentation provides the “instruction manual” for analysts to help interpret and use data correctly. Analysts need information about how the data are collected, what the data elements mean, how the reporting has changed over time, how the data have been edited, and where there are known data quality issues. Without this documentation, there is a high risk that analysts will develop misleading results and that different analysts will get different answers to the same question. Comparisons over time might be invalid. The wrong data might be used. Conclusions might extend substantially beyond the data—misleading decision makers about what the data really tell us.



The *Guidelines* suggest publishing source and accuracy statements for major data systems, and providing detailed file information to accompany data releases.



- *We have no metadata for our major safety data systems beyond simple record layouts* and a description of data details for our safety performance measures. Internally, we rely on institutional knowledge about our data systems, but there are many instances where we have discovered—through peer review, often much later—errors in our analysis because we did not account for the peculiarities in our data collections. The risk for analysts outside the agency (public users of our data) is probably much greater. In peer reviewing a proposed journal article recently, I found extensive errors in the assumptions about our hazmat incident data. Metadata might not prevent this, but it would reduce the risk.

Assembling and releasing *microdata*:

Public dissemination of our data is important in many ways—it provides an authoritative source for public research on safety issues, it makes our program operations more transparent for public oversight and participation, it provides information others can use to make safety-related decisions, and it can lead to corrections which improve the quality of our data. All of these help us leverage our own resources and can make government programs work better.



The *Guidelines* suggest releasing microdata (essentially record-level data) in many accessible formats, accompanied by data documentation (metadata), a clear description of revision information, and a contact point to facilitate feedback.

- *We update the data we release at regular intervals*—daily for hazmat incidents, monthly for pipeline incidents, and annually for pipeline annual reports. This frequency appears to be adequate for most analytical purposes, given the reporting frequencies and lags, and we produce special data

files where we need more timely data for internal analysis. The process for controlling data releases is incorporated into the normal data entry/editing processes—i.e., when the data are considered “good” they are entered into the data base and that triggers release to the public with minimal additional checking.

- ➔ • *We do not provide data documentation* beyond the data collection forms, instructions, reporting criteria in regulation, and record layout. We have not developed detailed metadata (described previously). Where we have information—for example, changes in reporting over time, estimates of underreporting, supplementary guidance we have given to operators—we have no mechanism for including this with our microdata releases.
- *We provide limited information on data revisions*—for hazmat incidents, we provide the date of the report or update, but we do not flag changes since a previous report. For pipeline incidents we do not provide any information about revisions to the data. This lack of information presents a common problem of competing findings. Analysts can refer to the data-date as a way of clarifying the scope and limitations of their analysis, but without any indication of what has changed when, other analysts would find it difficult to replicate an analysis without a same-day copy of the data.
- ➔ • *We handle narrative text inconsistently*. Analysts often find the narrative text to be the single most useful source of information about an incident. It was critical in our own recent risk evaluations of wetlines and lithium batteries, and it would be equally useful to others to replicate our analysis or to evaluate other risks. For hazmat incident reports, we remove personally-identifiable information before releasing the data, but otherwise include the narrative description of events in the data we release. In contrast, we remove all narrative text from pipeline incident reports before releasing the data; the data program manager refers to previous advice from counsel for this decision to withhold the information.
- *We could make our data easier to use and help avoid analytical errors* by adding more computed variables in the data set. For example, we might add the total number of fatalities, injuries, and property damage for hazmat incidents (we’ve seen analyses that misinterpret these data). We could tag TIH materials or any other useful grouping of commodities. For data comparability, we might differentiate which pipeline incidents meet the quantitative reporting criteria vs. those that are submitted because they are significant “in the judgment of the operator.” There are undoubtedly many other areas where we might improve the data for analysis before we release it. To some extent this is tied to the direction we take on the *data ownership* issue described earlier.

Computing statistics:

Agencies usually are the first to tabulate statistics from their own data, and these tabulations are widely viewed as authoritative sources of data in their own right.

Tabulations provide an opportunity to highlight the things we believe are important. This can include, for example, information about the scope and extent of the regulated industry, the causes and



consequences of accidents, and the results of enforcement actions. However, there is at least one special pitfall in tabulating these seemingly straightforward statistics.

We have to compare *apples to apples*. Anytime we are looking for trends over time, we need to adjust the data for inflation, changes in reporting criteria, seasonal patterns, or other similar changes that would prevent an accurate comparison. This is often called “normalizing” the data—to make the data comparable. We should use normalized data, generally, in tabulations, graphs, analyses, or presentations of data when we are trying to show trends over time.

- *Some of our statistical tabulations normalize the data, and some don't.* Tabulations of pipeline safety incidents (posted on our website) provide a particularly good example of how to normalize the data for comparability. By contrast, tabulations of hazmat safety incidents (posted on our website and in the Hazmat Intelligence Portal) do not adjust for changes in reporting criteria (which occurred in 2005), inflationary effects on dollar damages over time, seasonal patterns by month, or incomplete data for a year. As a result, the hazmat tabulations can be misleading.

Assembling and disseminating *statistics*:

We disseminate statistics through the PHMSA website, in publications, and in various forums like public meetings and conferences or workshops. Ideally, the process for assembling and disseminating these statistics would provide enough controls to give us (and others) some confidence that people will interpret them correctly.

The *Guidelines* provide some detailed suggestions for organizing, formatting, labeling, citing, footnoting, and documenting the statistics we release. They also suggest a process for pre-dissemination reviews to help ensure that publications and summaries meet minimal levels of quality.

- *The Pipeline Safety website demonstrates good practices in every aspect addressed in the Guidelines.* On the pipeline safety website, the presentation of incident data is clear, well-organized, and internally consistent. It includes complete source references for others to duplicate the results, good labeling of graphics and tables, appropriate footnotes to clarify the data, and explanations for outlier data. This appears to be a good example of how to do it right.
- *Hazmat safety tabulations, by comparison, are missing many key elements of good presentation.* At the Hazmat Safety website and in the Hazmat Intelligence Portal, the choice of graphics is often poorly suited to the data, and much of the labeling and other documentation are missing. As a result, the data are generally confusing, and the tables and graphics cannot be replicated easily from the data.
- *We have no standard, pre-dissemination review process for the statistics we release.* Each release is managed independently. There is a process for pre-dissemination review of the incident tabulations on the Pipeline Safety website, but this is not extended to other disseminations.

- *Most of the statistics we publish—in tables and graphics—are 508-compliant*; there are limited exceptions in some graphics and pdf files that are not. These have been identified previously by the CIO, who is working with the program to resolve the issues.
- *The data summaries we publish do not differentiate public vs. occupational (or private sector) risk.* We aggregate deaths/injuries affecting the general public *together* with those affecting workers and emergency responders, and we aggregate spills and dollar damages affecting company property with those affecting the general public or rights of way. But the risks are different, and risk exposure is different. Aggregating the data can be useful in some analyses, but it can also present a misleading picture of public risk.

A Public View: *Public vs. Industrial Risk*

People usually want to know what *their* risk is. The general public faces certain kinds of risks from pipelines and hazardous materials; industrial workers face other kinds of risks. Adding these together can be misleading. Risk exposures are different, people's knowledge of the risks is different, and the interventions are sometimes different.

- *Evaluating Risk (Feb 2008)*

Interpreting the data—*deriving meaning*:

It's often said that statistics don't lie ... but they can be interpreted many ways, and not all of them are useful or objective reflections of reality. It is the analyst's job to derive meaning from the data.

The *Guidelines* suggest involving other concerned parties in complex analyses, starting with the questions that need to be answered (rather than showing all data results from a collection), taking into account how the data were collected and the stability of the underlying processes, and using established statistical methods to distinguish information from uncertainty. In many ways, this echoes the recommendations in *Understanding Risk (2002)*, which suggested an analytic-deliberative process for identifying the questions and helping to guide the analytical effort in the first place.

- ➡ • *We don't typically involve others (i.e., affected stakeholders) in our hazmat risk evaluations.* We commonly invite others to review/comment on our analyses in draft form, but that isn't the same as involving others at the outset—helping to define the questions, shape the evaluation design and even the choice of measures. The risk here is that our approach can complicate the deliberation process by failing to incorporate others' perspectives from the beginning. This might be more an issue for the hazmat safety program, since the issues can involve so many stakeholders (shippers, carriers, many regulators, public interests) with different perspectives. The National Research Council Committee on Risk Characterization was clear: "*Science alone can never be an adequate basis for a risk decision ... risk decisions are, ultimately, public policy choices.*" There usually isn't a *right* answer about risk.
- ➡ • *We often use reported data as though it accurately reflected what actually occurred* – even in the face of contrary evidence. Our regulatory evaluations, for example, summarize our incident data (number of incidents, causes, and consequences) without accounting for missing data or potential biases in the data. This likely underestimates the potential benefits of our rulemaking, and might

result in poor assumptions about the causes and circumstances of failures. By this point in the process, however, it can be a major challenge to go beyond the data we have.




- *Incidents present a tempting—but misleading—measure of failures.* In the past, we have projected some conclusions about risk based on *all*-incidents when we were in fact much more concerned about the safety consequences affecting people (the Secretary’s safety goal is aimed explicitly at reducing deaths and injuries). Several analyses have shown that the patterns of causes, circumstances, and consequences for all-incidents are different from those for more serious incidents. More recent studies and performance measures generally make this distinction, and the pipeline safety website highlights the differences. The hazmat safety program has focused more narrowly on serious incidents, but even that definition includes a mixed collection of outcomes that is not tied analytically to the risk of death or injury.
- ➔ • *Interpreting the data is substantially complicated by the lack of metadata.* Many of the analyses using our data (some we’ve done, some others have done) have used the wrong data, or have not accounted for changes or known limitations in the data. Fortunately, all of these errors were discovered before the analyses were completed. What we don’t know is what we didn’t find. Without good metadata, the probability is high that the data have been misinterpreted frequently.
- *Pipelines reflect considerably more stable processes than hazmat transportation.* Pipelines carry essentially the same commodities they carried 20 years ago, and much of the infrastructure in place today was in place then. Changes in risk management approaches might be evolving more rapidly, but the same is probably true for hazmat transportation. At the same time, hazmat transportation involves changes in the technology of what is shipped (e.g., lithium batteries), and in how it is shipped (packaging, four different modes of transportation, etc.). This means that comparisons over time are probably much more challenging with hazmat.
- ➔ • *Our use of established statistical methods is generally weak—reflecting a broader lack of analytical capacity.* Few of our analyses address uncertainty in any systematic way, even though it is a very large factor in many cases. Few analyses address potential bias or missing data, and very few (if any) have used imputation to compensate for missing data. Analyses much more frequently use averages than distributions, numbers rather than rates, and rolling averages instead of time series models. This is largely just a reflection of the expertise we have. But it can result in misleading conclusions, and our conclusions may lack credibility.
- *Compilations of data that are not driven by a decision need are of limited value.* For example, an analysis of the HMIS data base in June 2005 provided hundreds of tables and graphs, providing breakouts of incident data by mode, year, month, state, hazard class, amount released, causes, amount of damage, and various combinations of these factors. However, most of the data presented all-incidents, when the program focus is on more serious consequences; there was no accounting for underreporting or potential bias in the data; the problem of multiple causes was not addressed; and the graphics (pie charts and bar charts, with no particular sequence in the data) generally obscured seasonal and other potentially important patterns. “Slicing and dicing” the data does not make for good, useful analysis.

Analyzing the data to produce *program information*:

Good analysis can compensate for many of the shortcomings in data quality at earlier stages in the process. At the same time, poorly-planned and poorly-conducted analysis can introduce new errors that degrade the quality of the data, distort our picture of reality, mislead decision makers, and result in unjustified confidence in what we “know.”

This stage of the life cycle in data quality brings us nearly full-circle back to the first step—identifying program requirements (*what do we need to know?*). In a well-defined program, we might even already have a picture in mind of what the program information might look like, or at least hypotheses that we could test, as we undertake the analysis. When we lack good program requirements, it is often the analyst that determines (or guesses) what someone else needs to know.

The *Guidelines* suggest a project plan before beginning all but the most simple analyses, with review of the plan by subject matter experts and data analysis experts to help ensure an appropriate focus and methods. Our *Strategic Plan* aims to make good use of information to help reduce risk, and specifically directs that we build a standing analytical capability to strengthen our understanding of risk based on sound data. We are moving in this direction. But we cannot yet claim to have sound data or a strong analytical capability to help guide decision making.

-  • *Recent analyses have identified many shortcomings in our data.* In a review of about a dozen recent analyses and the data quality issues discovered in the process, we noted lots of missing data, insufficient data models, difficulty in integrating data from multiple sources, data that are difficult to use, and gaps in how we manage data collection (See attached *Data Quality Issues from Recent Analyses*.) More fundamentally, these analyses reflected some big gaps in our analytical capacity.
-  • *Our approach to analysis is uneven and is not guided by a strategic view.* The pipeline safety program has substantially advanced its analytical capacity over the past two years through the Performance Evaluation Group (PEG). This group was created as a multi-disciplinary team of statisticians, engineers, program analysts, and other research specialties to help bring new insights to the program. By comparison, the hazmat safety program is just beginning to develop a core analytical capacity. Both analytical programs, however, appear to be somewhat disconnected from the needs of the program leadership. Their work is largely self-directed. There is no agreed-upon analytical agenda to guide their work or their priorities.
- *Each of our analyses generally is developed without a project plan.* In one case (an analysis of the risks of carrying flammable liquids by air), the analysis was preceded by development of hypotheses and an approach to testing them. Few other analyses have followed this model.
-  • *Our risk models use data; they are not data-driven.* In fact, they combine judgment and data in ways that can degrade the quality of the original data.

For example, one part of the Pipeline Inspection and Prioritization Programs (PIPP) was reviewed in 2008 as part of the Inspection Integration initiative—to explore whether we could develop a better data-driven model. The results demonstrated that the core risk model in PIPP tended to underestimate some risks and overestimate others. The review also showed that *a single variable* (recent incidents) within the model was better at predicting risk than the sum of all 10-12 weighted variables combined in the model. The *best* data were overwhelmed by about ten less-important data variables.

The hazmat strategy for prioritizing inspections (and other field activities) is broader in its scope (it prioritizes companies as well as kinds of activities), but similarly lacks an analytical basis for weighting different risk factors. Risk factors are identified through judgment, and weighted by judgment. The pipeline risk model is the subject of continuing research to improve it, and a new model might be released soon; the hazmat model is not (yet) subject to a similar, rigorous review.

<p>The Pipeline Inspection and Prioritization Program (PIPP-1) is used with other information to help set scheduling priorities for standard inspections.</p>	<p>The National Business Strategy for Hazmat is used to prioritize field activities based on risk.</p>
<p>How it works: PIPP is a data-based model using 10-12 data variables (like past accidents) that are transformed into nine indexes, which are added together for an overall risk score. The variables were selected using expert judgment, and the transformations that determine the weight for each variable also used expert judgment.</p>	<p>How it works: The NBS rank-orders 15 different activities—including several kinds of inspections, investigations, and outreach—and groups these into 5 priority categories based on judgment. Accident/incident investigations, failure analysis, and complaints are judged to be the maximum priority.</p>
<p>Some limitations: <i>Data</i>-weighting has been demonstrated to be superior to <i>judgment</i>-weighting in predicting future risk, and many of the PIPP-1 variables cannot be correlated with operator risk. PIPP-1 is only 1/5 factors in scheduling inspections; all are judgment-based.</p>	<p>Some limitations: None of the criteria for prioritizing activities have been tied analytically (and quantitatively) to differences in risk. High accident frequency, for example, is rated a medium priority, although this has been shown to provide the strongest indicator for pipeline safety.</p>

Presenting the data and *analytical findings*:

In a compelling (and chilling) review of the space shuttle Challenger accident, Edward Tufte has shown that engineers had all the information they needed to demonstrate the risk of launch the night before Challenger exploded, but that their *presentation* of the data was fatally flawed. Presentation is often critical in how others understand what the analyst found. It can also expose weaknesses in the analysis—if it can't be clearly communicated, we have to ask if the analysis is adequate.

The *Guidelines* provide some general direction for organizing information, presenting graphics and tabulations, and describing limitations of the data and the analysis. They also suggest development and use of a style manual.

- *We do not have (or follow) any standard practices in presenting the results of our analyses.* Reports, including graphics and tabulations, are developed by many different groups within the agency, with varying degrees of skill and experience in presenting data.

- ➡ • *Some of our analyses highlight the limitations of the data and methods; many don't.* The Data Details for the agency's principal performance measures provides probably the most extensive example of a discussion of the limitations of each measure, although even these do not address every limitation. Some of our analyses do not address limitations at all.
 - ➡ • *We rarely quantify the uncertainty in our analyses,* and the uncertainty is often large—which could seriously undermine the basis for important program decisions. This is basic in statistical methods. In some draft regulatory evaluations, we have estimated costs and benefits through many successive computations (each with assumptions) to find a single point estimate without addressing the cumulative error that this might introduce. In these cases, we are presenting conclusions for decision makers without addressing the degree of confidence we have in those conclusions.
- Dealing with Uncertainty: *Point Estimates vs. Ranges***

Decision makers need *estimates* of risk. There is no value in waiting for certainty because we'll never achieve it. At the same time, it is often very misleading to provide a point estimate and nothing more. At worst, it could lead to the wrong decisions in cases where it is important to avoid certain scenarios.

- *Evaluating Risk* (Feb. 2008)
- ➡ • *We have little expertise in presenting quantitative information effectively.* Even professional statisticians—whose livelihood is all about data—often lack the more specialized skill in presenting data for decision making. But it can make all the difference between effective and ineffective use of data. We also generally lack the technology and tools to present data effectively, and we make limited use of peer review to help compensate for these gaps.

Using Data – Making decisions and *acting on the information*:

The decision making process brings us to the end-point of data quality, although it is clearly beyond the direct scope of the Data Quality Act and the *Guidelines* published under the Act. This last step is included here because—to some extent—every other step presumes it, and because a gap or error here can make all of the other quality processes moot. The results of our analyses need to be *received and understood* in order to be useful. And we have set the standard ourselves for making decisions based on good data.

The agency's strategic plan envisions a risk-based, data-driven organization where we *use data to help drive program priorities, improve our ability to detect emerging risks and target/focus our prevention activities, and evaluate the effectiveness of our programs to help improve them as a means of reducing risk*. We recognized in the plan that this requires sound data and a strong analytical capability. This does not mean using data *alone* to make decisions; there are many other legitimate factors that go into public policy. It means inquiring, exploring the data, piecing together the best picture we can with the data that are available, demanding analytical input, and being able to explain how the data were used in the decision making process. So we strive to be a data-driven organization, even as we fall short in some ways.



- *We often make program decisions and use data to support them*, rather than demanding data as input for our decision making.
 - We don't use regulatory evaluations to drive our regulatory approach. We routinely have developed an analysis of alternatives *after* we decided on the preferred approach.
 - We don't use safety data to prioritize the regulatory agenda, because we don't have any estimates of the likely costs or benefits at this stage in the process.
 - We modify or develop new programs generally without a systematic evaluation of what we have now, and we don't create baselines or build program evaluation or measures of success into the new program design.
 - We can't use incident data effectively to focus our inspections—to zero in on what to inspect when we get there—because we don't capture failure data from incidents in a form that is very useful for inspections.

If those responsible for controlling risks lack the analytic fabric to disaggregate the overall problem into actionable projects, then they cannot work on them intelligently; nobody will know what to do tomorrow—except to do the same things they did yesterday.

- *The Regulatory Craft* (Sparrow, 2000)



- *We don't use our performance measures to drive our programs*. This might seem like the most basic, high-level use of data in a risk-based, data-driven organization—to start with the key measures of success, set goals, and drive our operations toward achieving our goals. But we don't really do that.

For many years, our pipeline safety program used the number of incidents as its principal safety performance measure; at the same time, it redirected its efforts in a major way toward integrity management—which was based on the premise that certain kinds of incidents were more important. The IM program *could not have been derived* from our goals at the time; it was, in fact, inconsistent with our goals, but we pursued it anyway because we knew intuitively that it was focused on the right thing.

The hazmat safety program for many years has used the number of serious incidents as its principal safety performance measure, even though it too reflects a mix of conditions. If we were to actually use this measure to drive our programs, it would suggest a focus on all bulk releases, which constitute about 2/3 of all serious incidents. But there are other risk factors like fire, explosion, and gas dispersion that we can show are much more important factors leading to death or major injury.



- *We don't combine/use both risk and performance data to allocate grants to states*. Both programs administer grant programs to states to help reduce safety risk. The hazmat allocation formula includes several risk-related variables to help target the greatest risks; the pipeline allocation formula does not address comparative risk. On the other hand, the pipeline formula adjusts grant

allocations based on a review of several performance factors; the hazmat formula does not address differences in performance or capability (i.e., need). To target resources effectively, we need information about where the risk is *and* what works in reducing risk.

- *We don't use much of the data we collect.* This is a judgment—a hypothesis that is fairly widely shared and that should probably be explored more fully. Some data might be critical in a future analysis even if we have never used it before. But some of the contrasts between what we collect and what we need are stark – we collect (just in case?) information on manufacturer, model and size of a pump that failed, but we don't have information on the mode or causes of the failure.
- *This might be viewed as a chicken/egg problem.* As a practical matter, we often make decisions without data input because we have substantial shortcomings in our data systems and analytical capacity. These shortcomings, in turn, reflect a lack of analysis and use. However, it's widely acknowledged in the statistical community that the best way to improve data is to use it. Statistical agencies develop strong analytical programs to help understand the data, demonstrate appropriate uses of the data, and provide feedback for improving it. The value of that model probably applies equally to any data program.

Some limitations of the evaluation:

This evaluation is intended to be reasonably comprehensive, as outlined in the *Guidelines* for conducting data quality assessments. This required ranging widely into the processes that might affect the quality and use of safety data. At the same time, this wide range limited the depth we could address in the evaluation. Some of the more significant limitations of the evaluation:

- The assessment does not cover all of our safety data. It generally addresses our incident/failure data more extensively than our system/exposure data, and it does not address in any significant way our mapping data, or the data systems we use to manage headquarters activities (like special permit processing).
- The assessment does not address all the known uses of our safety data. We concentrated more on programs with significant investment of resources—including rulemaking, inspection, investigations, enforcement, and risk evaluations—and did not address use of the data for managing training and outreach programs, or research and development, for example.
- In many areas, the findings are based on a sample of the evidence. For example, we conducted an analysis of blanks in the hazardous liquid pipeline incident data, but not in other sectors; we filed a hazmat incident report to test the online system of reporting and editing, but did not file a pipeline incident report; we sampled the most recent ten inspection and incident reports from each program. We tried to be careful not to overreach the findings beyond the evidence we considered unless we also had evidence that the sample was reasonably representative.
- We did not get input from external users of our data, beyond what we generally know from recent experience. The *Guidelines* suggest getting input from external users, but we believed we needed

to limit the assessment at this point because of time/resource limitations. This might be an important part of the next data quality assessment, and/or part of any followup analyses of specific data quality issues.

- To some extent, the data and processes we evaluated in this assessment have been a moving target, as many initiatives have been underway to improve data quality. This includes (in the pipeline safety program) building logic models to help understand information needs, extensive changes proposed in incident reporting criteria and forms, development of new risk models for targeting inspections, development of new processes and data collection for incident investigations, and examination of the organizational roles and responsibilities for data. We did not fully evaluate these ongoing initiatives to assess the likely outcomes.
- There is always some subjectivity in evaluating programs. We used the *Guidelines* and several other widely-used references to help establish criteria to evaluate against. But judging the seriousness of each weakness inevitably required some judgment as well. To help minimize the risks associated with these judgments, we included many experts in data quality on the review team, we considered the judgments of others in the interview process, we considered the general values suggested in the literature, and we provided opportunity for program review of all draft findings to help ensure factual accuracy and reasonable characterization of the issues. We also included some peer review of the draft report by data and evaluation experts in FRA and FMCSA.

Information Sources used in the Data Quality Assessment

Legislation

1. *Paperwork Reduction Act of 1995* (44 U.S.C. 3504 (e)(3)).
2. *Data Quality Act* – Section 515 of the Consolidated Appropriations Act of 2001.

Guidelines

3. *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies* – OMB (2001).
4. *Secretary's Policy Statement on Information Quality* – DOT (August 2002).
5. *The Department of Transportation's Information Dissemination Quality Guidelines* (2002) - <http://docketsinfo.dot.gov/ombfinal092502.pdf>.
6. *Updated Principles for Risk Analysis* - OMB Memorandum M-07-24, September 19, 2007.
7. *PHMSA Strategic Plan 2007-2011* (Aug. 2007).
8. *Evaluating Risk: A Working Paper for Pipeline and Hazardous Materials Safety Programs* - Pipeline and Hazardous Materials Safety Administration (February 8, 2008).
9. *Principles and Practices for a Federal Statistical Agency* - National Research Council, Committee on National Statistics (Fourth Edition).
10. *Analytical Perspectives, Budget of the U.S.* – Office of Management and Budget (FY 2010).
11. *Increased Emphasis on Program Evaluations* – OMB Memorandum M-10-01 (October 7, 2009).

Risk/Evaluation Literature

12. *Managing the Risks of Organizational Accidents* – James Reason (1997).
13. *Understanding Risk: Informing Decisions in a Democratic Society* – National Research Council, Committee on Risk Characterization (1996).
14. *The Regulatory Craft* – Malcolm K. Sparrow (2000).
15. *Barriers and Accident Prevention* – Erik Hollnagel (2004).
16. *Managing the Unexpected* – Karl E. Weick and Kathleen M. Sutcliffe (2007).

17. *An Evaluation Roadmap for a More Effective Government* – American Evaluation Association, Policy Task Force (February 2009).
18. *Data-Driven Risk Models Could Help Target Pipeline Safety Inspections* – BTS Special Report (July 2008).
19. *Visual Explanations: Images and Quantities, Evidence and Narrative* – Edward Tufte (Feb. 1997).
20. *Handbook of Practical Program Evaluation* – Wholey, Hatry, and Newcomer (Ed., 2004).
21. *Increasing Evaluation Use Among Policymakers Through Performance Measurement* – Mohan, Tikoo, Capela, and Bernstein in *New Directions for Evaluation* (Winter 2006).
22. *Missing Data* – McKnight, McKnight, Sidani and Figueredo (2007).

Data/Statistical compilations

23. *Hazmat incident statistics*: <http://www.phmsa.dot.gov/hazmat/library/data-stats/incidents>
24. *Pipeline data and statistics*: <http://www.phmsa.dot.gov/pipeline/library/data-stats>
25. *Analysis of Hazardous Materials Information System Database: An Interim Analysis* – PHMSA (June 2005).

Past Reviews of our Data

26. *Pipeline Safety and Security: Federal Programs* – Congressional Research Service Report for Congress (Jan. 5, 2007)
27. *Using or Creating Incident Databases for Natural Gas Transmission Pipelines (A Guideline)* – Report of Study Group 3.4 from the 23rd World Gas Conference (June 2006).
28. *Integrity Threats to Hazardous Liquid Pipelines* – Inspector General Report AV-2006-071 (Sep. 18, 2006).
29. *New Risk Assessment Program Could Help Evaluate Inspection Cycle* – GAO Report RCED-89-107 (March 1989).
30. *The Office of Pipeline Safety Is Changing How It Oversees the Pipeline Industry* – GAO Report RCD-00-128 (May 2000).
31. *Information Strategy Needed for Hazardous Materials* – GAO Report IMTEC-91-50 (Sep. 1991).
32. *Estimating the Extent of Under-reporting of Hazmat Incidents* – Preliminary Findings (May 11, 2007).
33. PHMSA IT Program Review (2008).

34. *Safety Incidents on Natural Gas Distribution Systems: Understanding the Hazards* – Allegro Energy Consulting Report for Office of Pipeline Safety (April 2005).
35. *The U.S. Oil Pipeline Industry's Safety Performance* - Allegro Energy Consulting Report for Association of Oil Pipelines and the API Pipeline Committee (Feb. 2003).
36. *Data-Driven Risk Models Could Help Target Pipeline Safety Inspections* – BTS Special Report (Kowalewski and Young, July 2008).

Program and Data Systems Documentation

37. *The Pipeline Inspection Priority Program used by the Office of Pipeline Safety* (1994).
38. *National Business Strategy* – PHMSA Office of Hazardous Materials Enforcement (Feb. 2008).
39. *Intermodal Hazardous Materials Intelligence Portal, Functional Requirements Document* – DOT (March 23, 2007).
40. *Hazardous Materials Information System (HMIS) Requirements Definition* – PHMSA Office of Hazardous Materials Safety (rev. May 19, 2000).
41. *Safety Monitoring and Reporting Tool (SMART), System Requirements Document* – Office of Pipeline Safety (April 15, 2004).
42. *Safety Monitoring & Reporting Tool (SMART) Implementation Strategy* – Office of Pipeline Safety (rev. May 9, 2006).

Data Quality Issues from Recent Analyses

Hazmat Bulk loading/Unloading Analysis (March 23, 2007)

- **We lack data on one of the central risks**—unloading of rail tank cars at fixed facilities—because our reporting requirements for hazmat incidents are limited to events that occur “in transportation.”
- **The causes and circumstances of serious hazmat incidents are different from all hazmat incidents.** Breakouts by mode, transportation phase, and cause all showed different patterns, which means that an analysis of all-incidents is not representative of the smaller subset with more serious consequences. This presents a small-numbers issue too—an analysis of serious incidents is subject to greater uncertainty.
- **Our failure codes do not point to the transportation phase at which a failure occurred.** This analysis was focused on loading and unloading, but many of the in-transit incidents also could be traced to failures during loading (e.g., over-pressurization, failure to tighten couplings, etc.). However, many of the codes for *what* failed, *how* it failed, and *why* it failed were inconclusive with respect to the key variable of phase.
- **We have large uncertainty in our conclusions** because of substantial underreporting of hazmat incidents, and failure codes which cannot be used to nail down the transport phase during which a failure occurred. Some large effects were due to a small number of companies reporting, suggesting missing data.
- **The number of incidents is a little ambiguous** since the data base contains separate records for each shipment/commodity/tank involved. For some analytical purposes, this is helpful; for others, it might appear to inflate the number of incidents by a small percentage.
- **Long reporting lags mean that recent-year data might be incomplete.** For analyses aimed at breaking out percentages, this is less of a problem than when we compare numbers over time.

Pipeline Safety Risk Modeling for Inspection Integration (2007)

- **Risk scores from PIPP don't correlate with actual risk of incidents** because the variables and weighting were determined by judgment, instead of by the data. PIPP predicts future risk better than random sampling, but some of the individual variables used in PIPP (like past accidents) are a better predictor than the more complex PIPP algorithm. That means that PIPP is degrading the quality of the input data by combining good data with much weaker variables.
- **We have limited ability to rank operator performance from our inspection data,** because—except for Integrity Management inspections—we don't capture inspection findings

(deficiencies) as data. We record only those few deficiencies that rise to the level of a notice of probable violation.

- **We can't combine many risk variables** (like age and corrosion protection) to assess inherent pipe risk because exposure data (from annual reports) doesn't include combinations of characteristics like we have for incidents.
- **Many of the risks we assign to operators are outdated** by as much as a year because we don't have a mechanism for tracking changes in operator, when pipelines change hands, etc. between annual reports.
- **Some of the annual report and incident data are unusable** in our risk model because operators file incident reports under OPIDs with no annual report.
- **The data we collect on systems, inspections, and incidents appear to provide little value** in identifying risk factors for targeting inspections.

Hazmat Inspection Prioritization: National Business Strategy (Feb. 2008)

- **We can highlight high risk but not high risk rates** because we don't have exposure data to normalize incidents and deficiencies.
- **We can't identify risk factors except by expert judgment** because we lack exposure data, and sufficient incident data, to draw correlations.

Performance and Accountability Report (FY 2008)

- **We were unable to say definitively how risk has changed** because the number of incidents fluctuates annually apart from real, underlying changes in risk, and we have no indicator of risk.
- **We were stuck using old measures** that we replaced internally two years ago because the performance system discourages experimentation and rapid improvement.
- **Our performance projections were very preliminary** and we revised data for several past years (again) because we use CY measures in a FY report, we have a significant reporting lag, and we never call a report "final."
- **Our performance trends might be substantially biased** (for hazmat) because we have significant under-reporting (estimated up to 90%).
- **Our reported performance is binary (met/not met)** without accounting for uncertainty because the performance system emphasizes accountability over learning.

Low Stress Pipelines and Gathering Lines Rule

- **We can't estimate the potential benefits** for gathering lines because we have no data on exposure (mileage) or safety (incidents).
- **We have very uncertain estimates of the benefits** for low stress pipelines because we have no data of our own on exposure or safety, and we have to combine eight separate estimates to compute benefits.
- **Our estimates are misleading** with respect to the policy choices because they do not show the range of uncertainty; we have limited capacity to do this.

Lithium Battery Risk Analysis

- **We can't (yet) quantify most of the risk factors** we have identified because we lack exposure data and incident data on most variables.
- **The analysis relies on very limited failure data.** Any given lithium battery spends a vanishingly small fraction of its life aboard an aircraft; but the failure data we have is constrained to just aircraft incidents.

Pipeline Safety State Grant Allocation

- **We can't use incidents** directly as an indicator of relative risk because there are so few incidents in individual states each year.
- **We have difficulty using mileage** to allocate grants by risk because we don't capture gas transmission mileage by state.
- **We have limited information on other risk factors** (e.g., geography) because we don't capture data on these factors.

Organizational Assessment Monitoring (Monthly Updates)

- **We can't say (with any confidence) where we stand** on our performance re: hazmat risk because the number of hazmat incidents keeps changing over time - the data are not final for years.
- **Monthly updates are labor-intensive** and prone to error because monitoring requires extensive, manual pre-processing to trim records, add computed variables, and aggregate across sectors.
- **Adverse trends (Spills in HCAs, Undeclared hazmat incidents) can be misleading** because our measures mix together data without a clean, unifying concept of risk.

Tank Truck Wetline Incidents

- **We can't easily distinguish wetline incidents** because our coding scheme for hazmat incidents doesn't address them specifically. Determining which incidents were caused by wetline failure required extensive, manual review of incident reports, including narrative descriptions.
- **We aren't capturing all wetline incidents** because carriers tend to underreport hazmat incidents generally.

Analysis of Undeclared Hazmat Incidents (May 4, 2009)

- **Reported incidents are misleading, since half of the reports are not "incidents"** in the traditional sense. They are simply discoveries of undeclared hazmat somewhere in the transportation system, with no release of product, no deaths or injuries, no evacuations, no closure of roads, and no other consequences which would trigger reporting of an incident. These were brought under the incident reporting system in 2005 when we simply changed the reporting criteria—effectively redefining an "incident" to include these particular findings of non-compliance.

Analysis of Serious Hazmat Incidents (May 28, 2009)

- **The criteria for "serious" hazmat incidents reflects a mixed collection of outcomes.** Most (over 60%) of the serious incidents are "serious" only because they involved a bulk release—with no other serious consequences; this dominates the measure.
- **The criteria do not reflect some of the most important risk factors** that make death/major injury more likely. Fire, explosion, and gas dispersion, for example, are all more significant risk factors than bulk release.
- **The causes and circumstances of serious incidents are different from death/injury incidents**, so focusing on serious incidents could divert our attention from the most important safety interventions.
- **Three of the criteria for serious incidents reflect *positive* actions people take** (evacuations, road closures, alterations of a flight plan) to limit safety consequences.
- **All of the criteria for serious incidents are weighted equally**, even though it's fairly obvious that we don't value the consequences or risks equally in how we manage the program.

Distribution Integrity Management Program – Data Group Final Report (Undated)

- **The causes of incidents and leaks were both limited to a small set of cause categories.** The categories were changed in 2004 (new causes added to incident reports), but this made leak causes different from incident causes.
- **Inconsistent reporting**—over-reporting of incidents involving property damage and inconsistent classification of leak severity—limited the analysis.
- **Missing data**—no annual report data on master meter or LPG operations, no data on leaks removed (by material), and insufficient incident detail to determine whether excess flow valves would have mitigated an incident—limited the analysis.
- **We have insufficient incident/failure data by state** to provide meaningful baseline performance measures for operators or for individual states. Most operators and many states experience zero distribution incidents in a typical year.

Safety Incidents on Natural Gas Distribution Systems: Understanding the Hazards (April 2005)

- **Narrative description of the incident** is needed to reclassify older incidents into the new cause categories. Some (especially older) cases fall into cause categories which are too general to allow for effective analysis.
- **Greater detail on incidents** (through the PPTS data) was needed to get a finer understanding of the role of different hazards and issues impacting public safety. The conventional wisdom that most gas distribution incidents are caused by outside force damage was correct, but it was based on cause categories that are too broad to allow development of effective strategies for performance improvement.
- **Ambiguous reporting instructions** for data collected from 1999-2003 limited reporting of certain incidents (fire first) based on a narrow definition of “facilities.” This presents a comparability issue today.
- **There is inconsistent reporting** of incidents that involve facilities outside PHMSA jurisdiction, so data cannot be compared state-to-state or utility-to-utility. The inconsistency also obscures the real picture of failures.

Past Reviews of our Data Programs:

Several evaluations over the past ten years have examined quality issues with PHMSA's safety data. Each of these was examined to get perspective on the issues that have been identified previously.

- **A cross-modal team evaluated DOT's Hazardous Materials program** in 1999, and in the process offered several observations and recommendations for data analysis and improvement. The team found that DOT's hazmat safety programs were hampered by a lack of sufficient, reliable, and timely information; that our data systems did not contain adequate information on hazmat shippers and carriers; and that DOT did not have reliable hazmat flow estimates. The report recommended that BTS evaluate existing hazmat data bases and identify additional data needs.
- **The Bureau of Transportation Statistics (BTS) assessed the quality of data** in PHMSA's Hazardous Materials Information System (HMIS) in 2002. BTS' assessment profiled the 6 data bases in HMIS from written documentation and interviews concerning data processing, analyzed the data against published tabulations, and assessed the system against the six major attributes of data quality published by BTS: relevance, completeness, quality, timeliness, comparability, and utility. The system was judged strong in utility and adequate in all other areas. BTS offered eight recommendations for improving quality.
- **The National Transportation Safety Board (NTSB) reviewed transportation safety data bases** in 2002 to evaluate data quality issues. This study generally advocated a data-driven approach (like the one used by CDC) that is grounded in the science of epidemiology—consisting of surveillance, identification of risk factors, and development and evaluation of prevention strategies. NTSB surveyed its past data recommendations, concluded that the Board's ability to study important safety issues is often affected by poor data quality, and recommended development of better safety exposure data to help evaluate risk and safety interventions.
- **PHMSA conducted an IT program review** in 2008 to identify business and technology performance gaps that might inhibit the achievement of our safety mission objectives. The review proposed a series of projects to strengthen PHMSA's information architecture in four general areas: *data governance* (to develop common standards, processes, and procedures); *data architecture and management*; *organization* (roles and responsibilities, rules of engagement); and *technical architecture*.

A Draft Analytical Agenda for PHMSA

November 2009

Program Evaluation

Program evaluation is a systematic study of the logic, design, implementation, and effectiveness of a program—what works (in addressing safety risks) and why. It provides evidence that can be used to compare alternative approaches, guide program development and decision making, identify unintended effects, redirect programs that are not working as expected, and identify effective practices. There are several important questions we need to answer with program evaluation:

1. Do our risk models target the highest risks?

We target inspections to focus resources on the highest risks. Our risk models use data from incidents, previous inspections, and other sources but the weighting of risk factors is based largely on judgment. There is some evidence—from recent reviews of risk models—that data-weighting could be superior to judgment-weighting in identifying companies or systems with the highest risk. We invest considerable resources in inspections of the regulated community, so the value of a good risk model could be high. We might explore the potential for transferring the knowledge from pipeline safety modeling to hazmat safety, and also consider extending the work in pipeline safety to better focus on specific risks. We should include a broader review of the risk models used by others to help identify good practices.

2. Do our enforcement actions have the intended effect?

The program is designed to achieve correction and deterrence, but we believe that these results might be limited by the lack of timeliness in our enforcement action, the offsetting benefits of non-compliance, the limited use of enforcement as a tool to fix the problems we find during our inspections, and limited awareness by others of their exposure. In the hazmat safety program, the sheer numbers of companies to inspect presents a significant obstacle to effective deterrence. If correction is enhanced by our enforcement actions, we should see improved safety performance (fewer accidents, lesser consequences) in those companies. What would we expect to see if deterrence was working? We might estimate the value and extent of non-compliance, and compare that to the penalty for non-compliance, as a first cut threshold for effective deterrence.

3. Could our inspections be re-oriented to collect better data on latent safety conditions?

Social science research over the last 30 years has highlighted the depth of interactions between human and organizational factors that contribute to major accidents. Our data systems reflect [almost] none of these advances. Latent conditions—such as poor design, gaps in supervision, undetected manufacturing defects or maintenance failures, unworkable procedures, clumsy automation, shortfalls in training, less than adequate tools and equipment—are *always* present in complex systems. We don't have data on latent conditions at the point of an incident, nor at the point of inspection. What is the opportunity to re-orient our inspections to collect such data, and what might be some of the unintended consequences of doing so?

4. Could we get all the incident data we need from our own investigations?

We rely on industry reporting for our incident data. But the regulated industry has an institutional bias in determining the causes, circumstances, and consequences of failures. Accident investigations—the limited number that we do—have shown some significant differences between what a company reports and an objective view of these events. Reports from companies also reflect large numbers of blanks and “unknown” data, particularly in the most serious cases—exactly where it is most critical that we have good data. There are about 700 pipeline incidents and 17,000 hazmat incidents reported each year. If we investigated all incidents above some threshold, could we reduce industry reporting to a simple notification or postcard, and develop the data we needed ourselves?

5. How could we measure strong safety culture in companies?

We have a growing understanding of safety culture and the upstream organizational processes and circumstances that lead to failures—especially the kinds of process failures that can have catastrophic consequences. Over the past two years, the agency has taken a lead role in working with the pipeline industry and other agencies to explore safety culture and its relationship with process safety. We *know* we need to know more about this. But our questions still reflect an early stage in the learning process, and so far they remain disconnected from our data systems.

6. How can we better measure the overall changes in safety risk over time?

At a very high level, we have invested considerable effort over the years to developing and refining the concepts driving our performance measures—which we use in guiding priorities for the agency, justifying budget requests, and reporting to Congress. However, we have already recognized some significant shortcomings—our primary outcome measures do not really reflect changes in risk over time. The numbers are now too small to draw meaningful conclusions about the trends, or to disaggregate the data to find meaningful patterns. We must also deal with the natural tension between the need to monitor outcomes and the desire to attribute outcomes to what we do.

7. Monthly monitoring of agency performance—how can we better explain the trends?

We need to track progress in the key metrics in our Organizational Assessment, and investigate unusual patterns to understand what’s happening and why, and to redirect our programs if needed. We have limited understanding of the safety trends we are seeing. Understanding the trends requires disaggregation of the data, but the data are often very “thin.” We need to explore other approaches.

8. What are the actual costs and benefits of the rules we have published?

We don’t know what our rules cost or what benefit the public gets because we don’t assess the *actual* costs and benefits after implementation. This might be one of the easiest kinds of program evaluation we could do; we have a clear “before” picture and estimates of what we expected to happen. Retrospective analysis could help us redirect programs where we found unexpected consequences, refine our estimates for future benefit-cost analysis, and generally better understand what’s happening in the regulated industry.

9. Allocation model for State grants—how should we combine risk and performance data?

We don't combine/use both risk and performance data to allocate grants to states. Both programs administer grant programs to states to help reduce safety risk. The hazmat allocation formula includes several risk-related variables to help target the greatest risks; the pipeline allocation formula does not address comparative risk. On the other hand, the pipeline formula adjusts grant allocations based on a review of several performance factors; the hazmat formula does not address differences in performance or capability (i.e., need). To target resources effectively, we need information about where the risk is *and* what works in reducing risk.

10. What is the value of a “statistical injury?”

Our regulatory evaluations frequently require that we monetize deaths and injuries; we have an accepted range for the value of a statistical life (VSL) from OST, but we do not have a consistent and supportable value for injuries. There are other models (e.g., NHTSA's analysis for crash-related injuries) we might use as a starting point. Since the nature of the injuries might be different for pipeline vs. hazmat, we might also end up with different values across our programs, but we should consider a consistent approach in our analysis.

Risk Evaluation

Risk evaluation is a systematic assessment of safety risks to help us understand the scope, magnitude, and nature of the problem we are trying to address with federal interventions. It requires analysis of technologies, operating practices, and failures to help focus our efforts on the most significant risks and on the root causes of failures. There are several important questions we need to answer with risk evaluation:

1. What is the underlying risk of low probability high consequence (LPHC) accidents?

We need to better estimate the risk of low probability high consequence (LPHC) events—the kind of risk that is potentially hidden when we simply focus on the recent historical record. There is evidence that the public cares disproportionately about this kind of risk, and it probably presents a greater strategic risk for the agency as well. We need to identify outliers in the data (e.g., the consequences of Hurricane Katrina in 2005) and spread out the effects of these over a longer period of time, and we need to estimate the probability and consequences of risks that *don't* appear in the data and add these into our analyses. We might approach this through a broader failure mode and effects analysis, or through review of incidents to find areas where simple luck prevented more serious consequences.

2. How is risk exposure changing over time?

Our performance measures track the negative consequences of pipeline and hazmat transportation, but we lack good data on some of the changes in risk exposure that would help provide context for deaths, injuries, etc. Exposure measures are often used to *normalize* data. One measure that would be useful to have is ton-miles—to help draw comparisons across freight modes of transportation. Ton-mile estimates for hazmat are limited to the five-year Commodity Flow Survey with a 2-year time lag; it

would be more useful to have annual data. Estimates for pipelines were developed by BTS several years ago but rest on some key assumptions that should be validated or revised.

3. Leading indicators—how can we detect increasing risk or emerging risks in advance of failure?

Failures are lagging indicators—they tell us about problems after the fact. In some ways, past accidents can be used to help predict future accidents, but we also need better indicators of emerging risks and other conditions that might provide early warning signals of increasing risk. Safety investment, safety culture, financial health, etc. might provide useful early warning signals.

4. Risk vs. outcomes—what are the key risk factors?

The number of incidents involving death or major injury is small, and annual variations often do not reflect real changes in risk. To take better advantage of the data we have, we need to identify the most important risk factors—those conditions/circumstances that increase the likelihood that an incident will result in death or major injury. This should be a data-driven analysis, demonstrating the significance of a risk factor with conditional probabilities, and using these results to weight all incidents based on their *potential* for serious consequences (regardless of actual consequences). Weighted risk measures like this could help smooth out annual variations in the data, help explain the trends in the most serious incidents, and provide clues to the risk factors we might target for program intervention.

5. Undeclared hazmat—how much risk is this really?

In 2005, incident reporting criteria were modified to include the discovery of undeclared hazmat. These now comprise about 8% of the total reported incidents in our data base, although about half of these do not indicate a release of hazmat or any other criteria for incident reporting—no deaths, injuries, property damage, fire, explosion, evacuation, closure of a transportation artery, or any other consequences. The inclusion of these data here increase the potential for error and misleading conclusions in conducting safety analysis. We need to explore undeclared hazmat more deeply—especially looking at the problem of underreporting, and the nature of the risk. One potential new program we might look at is for voluntary reporting of leaks involving undeclared hazmat at the receiving end of a shipment.

6. What is the public risk from pipelines and hazmat vs. releases industrial/occupational risk?

The data summaries we publish do not differentiate public vs. occupational (or private sector) risk. We aggregate deaths/injuries affecting the general public together with those affecting workers, and we aggregate spills and dollar damages affecting company property with those affecting the general public or rights of way. The risks are different, and risk exposure is different. Aggregating data can be useful in some analyses, but it can also present a misleading picture of public risk.

7. What is the extent/consequence of greenhouse gas emissions from pipelines?

About 2% of the natural gas carried through pipelines is reported as “unaccounted for” due to metering error, leaks, accidents, and intentional releases associated with maintenance. It’s not clear

how much of this is actually released into the atmosphere, but the question is an important one to answer in the context of global warming. Methane is a greenhouse gas, with much greater warming potential than carbon dioxide—it is more 25 times more potent than CO₂ over the long term (100 years) and 72 times more potent over the short term (20 years). EPA's estimates of methane emissions from pipelines are somewhat less than the "losses" reported to PHMSA, but neither agency has reliable estimates.

8. What is the basic information we need to assess risks outside our current regulations?

There are several "invisible risks" (within our statutory authority but not necessarily regulated) where we have little/no risk data—for example: non-jurisdictional failures that are tied to jurisdictional pipeline systems; failures of DOT packages, cylinders, or containers "outside transportation" generally; gas pipeline master meter operators, or hazardous "materials of trade." In some cases, we have explicitly exempted certain operations from reporting; in other cases, we might not have fully considered the potential risks or the benefits of casting more widely for failure data so we can understand the risks before a big accident occurs. Lacking good data on these "invisible risks", we can't quantify the risks or address them effectively. What is the scope/magnitude of these kinds of risks, and what would we need to know to assess the relative risks?

Data Evaluation

Data evaluation is an analysis of data quality issues to help ensure the relevance, accuracy, completeness, comparability, timeliness, and utility of data to support effective decision making. It provides a basis for developing the data profiles or metadata that analysts need to understand the limitations of the data, and a basis for continuous improvement in data quality. There are several important questions we need to answer with data evaluation:

1. What do we need to know about safety failures?

We don't have a good conceptual model for understanding failures. We don't capture *the chain* of failures, including especially the root causes, that typically are associated with any accident or incident; we don't capture all the relevant circumstances that might (through statistical analysis) reveal hidden causes; we don't capture inspection deficiencies in a way that would allow us to tie together our inspections and accident investigations. Most analyses struggle with the data to find patterns and meaning, but they are severely limited by the basic conceptual models. Without a complete picture of causes and circumstances, we don't understand the full extent of the problems (e.g., human error, corrosion, etc.) or the interrelationships of causes, and we have greater difficulty identifying critical control points and targeting the risks effectively.

2. What is the likely extent of underreporting hazmat incidents?

Preliminary analysis of hazmat incidents suggests we might be missing 60-90% of all reportable incidents. Some of these missing reports might be the result of a lack of knowledge about the reporting requirements (these tend to be smaller companies); some might be the result of a decision

to avoid reporting. The effect is the same. The most troubling aspect of this is that the missing incidents appear to be *different in kind* (different patterns of causes, circumstances, consequences, etc.) from the reported incidents, so projecting from what we know might be giving us a *distorted* picture, as well as an incomplete one.

3. What is the scope/magnitude of the problem in identifying companies?

We cannot accurately and consistently identify companies—companies often use multiple IDs for different purposes; entities change through mergers, acquisitions, and partnerships; and companies might be related through several corporate layers. We have limited mechanisms in place to track company relationships or changes over time. As a result, we can't measure operator performance consistently or target our resources most effectively based on comparative risks; performance histories are sometimes fragmented. The proposed "One-Rule" in pipeline safety includes an attempt to resolve this issue for pipeline operators, but there is not yet a corresponding effort to address it for hazmat shippers and carriers. The Hazmat Intelligence Portal, for example, returns 165 companies from a search on "UPS." As a result, performance histories can be fragmented and misleading. We have incomplete data on shippers, because the hazmat registration program was not designed to capture all shippers (many/most are exempt).

4. Are there significant opportunities to leverage data from other sources?

We have difficulty integrating data from police and fire department reports, reports to the National Response Center, data from CDC, and many other systems. We can't easily use the data from most other data systems because we lack common identifiers (in many cases), and many of these external systems are not sufficiently transparent to allow us to understand their limitations. These are common problems in safety programs government-wide, and data integration is often encouraged as one of the best ways to leverage resources.

5. How might we collect comparable data for both federal and state programs?

We don't capture data from states in a form that is comparable to the federal program—limiting our ability to evaluate the effectiveness of state pipeline safety programs or operators and systems that are inspected by states. States use our data systems when they are acting as interstate agents for pipeline safety, but this represents a small fraction of their work. They do not report their inspections of intrastate gas pipeline operators with any detail (they report the number of inspections conducted). As a result, we have very limited information about the condition of 80% of the national pipeline system—where about 80% of the incidents involving death/injury occur.

6. What can our edit checking processes tell us about the quality of the resulting data?

We do not publish editing statistics, and we don't retain editing information for analysis. For hazmat incidents we use editing information to grade contractor performance, but there are no flags in any of our data bases to indicate where we commonly see errors. But tracking the areas where people have problems reporting accurate data could help identify areas of potential problems in the coding schemes and the resulting data we get.

7. What do analysts need to know about our data to avoid misinterpretation or misuse?

Without good metadata, the probability is high that the data have been misinterpreted frequently. We have no metadata for our major safety data systems beyond simple record layouts and a description of data details for our safety performance measures. Internally, we rely on institutional knowledge about our data systems, but there are many instances where we have discovered—through peer review, often much later—errors in our analysis because we did not account for the peculiarities in our data collections. The risk for analysts outside the agency (public users of our data) is probably much greater. Metadata might not prevent this, but it would reduce the risk.

8. How can we make narrative data more useful for analysis?

Analysts often find the narrative text to be the single most useful source of information about an incident. It was critical in our own recent risk evaluations of wetlines and lithium batteries, and it would be equally useful to others to replicate our analysis or to evaluate other risks. For hazmat incident reports, we remove personally-identifiable information before releasing the data, but otherwise include the narrative description of events in the data we release. In contrast, we remove all narrative text from pipeline incident reports before releasing the data.