Risk Analysis and Ordinal Risk Rating Scales—A Closer Look

By

Kelly Waldron

Dec 21, 2015 8:00 am EST

Peer Reviewed: Risk

ABSTRACT

This paper offers practical advice on how to improve the risk analysis portion of the QRM lifecycle, with particular attention paid to the typical ordinal risk rating scales that are used during risk analysis. Rating scales are criticality important in the risk process. A wealth of academic and industry literature exists that imply that rating scales are often flawed and sometimes inherently invalid. Guidelines to improve the quantitative risk analysis process are proposed as follows:

1. Spend time developing risk rating scales—construct detailed and meaningful criteria for each numeric score
2. Train risk team members on heuristics at a practical level -- not a theoretical level
3. Require a documented rationale for each selected score, with references to scientific studies, data, and trending
4. Require that sources of uncertainty, assumptions, or gaps in knowledge be discussed during the scoring process and explicitly disclosed in the risk management documentation
5. Avoid using *only* risk priority numbers (rpns) in risk reduction and acceptance decisions.

These concepts can be applied to qualitative approaches as well to improve the conduct of risk analyses. A journey towards true QRM maturity will require a revisit of the most commonly applied portions of the QRM lifecycle.

INTRODUCTION

At the ten year anniversary at ICH Q9, any risk practitioner worth his or her salt can recite the components of a risk assessment by rote: Risk identification, risk analysis, and risk evaluation (1). These steps are perhaps the most commonly invoked throughout the Quality Risk Management (QRM) lifecycle. Without a thorough understanding of the risks that may exist (risk identification), the nature and gravity of those risks (risk analysis), and whether they require additional attention or reduction (risk evaluation), steps towards improving process and product quality cannot be taken and QRM activities will not succeed. Nevertheless it would be folly for the pharmaceutical and biopharmaceutical industries to continue to mature in QRM without seeking opportunities to further optimize the risk assessment process (see Figure 1) underpinning science- and risk-based thinking.

This paper offers practical advice on how to improve the risk analysis portion of the QRM lifecycle, with particular attention paid to the typical ordinal risk rating scales that are used during risk analysis. While the terms risk analysis and risk assessment are sometimes used interchangeably by some companies, this paper treats risk analysis in the same way that ICH Q9 does – namely, that it is a sub-part of the risk assessment process that involves estimating the risk associated with identified hazards. Risk analysis uses either a qualitative or quantitative approach that links the likelihood of occurrence and
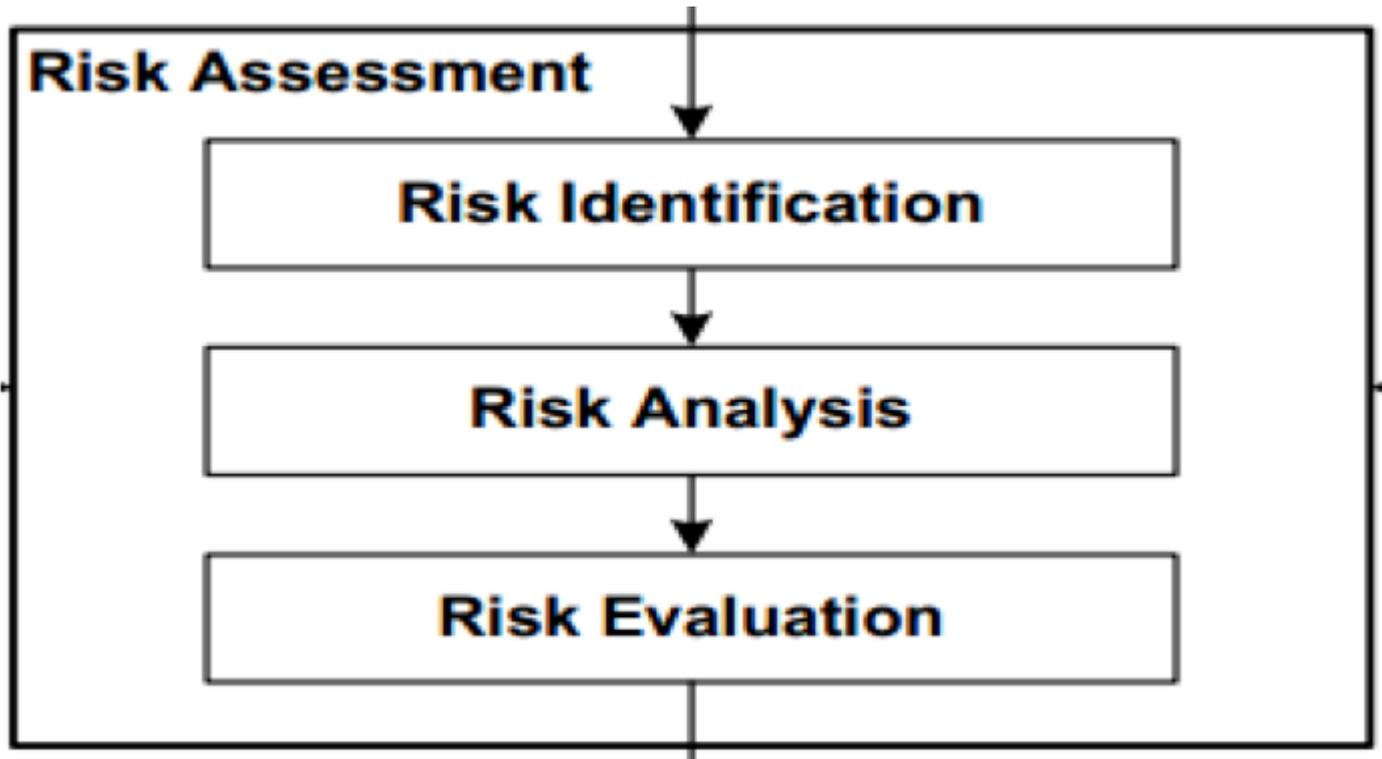
severity of harm.

FIGURE 1. THE RISK ASSESSMENT PROCESS PER ICH Q9 (1)

RISK RATING SCALES AS THE FOUNDATION OF RISK ANALYSIS

As noted above, ICH Q9 defines risk analysis as "the estimation of the risk associated with… identified hazards. It is the qualitative or quantitative process of linking the likelihood of occurrence and severity of harms" (1). Most risk management tools enable the estimation of likelihood and severity through a risk rating or scoring process where each individual risk is assigned a "score" based on pre-defined criteria. As a result, the success or failure (or more pertinently, the relative robustness) of the risk analysis process hinges upon the underlying rating scales that are employed. Based on their relative importance to the overarching risk assessment, one might expect these rating scales and their utility in a risk analysis to have been carefully vetted, refined, and tested such that they can withstand a rigorous level of scrutiny. After all, these scales are the foundation for the assignment of the risk levels that will determine the acceptability of the risks to the process, product, and ultimately the patient. Yet, a wealth of academic and industry literature exists that imply just the opposite -- The risk rating scales employed in a risk analysis are often flawed and sometimes inherently invalid (2-5).

A personal anecdote may prove helpful to illustrate the point. I live in northern New Jersey, a swath of the country that in recent years has been the victim of several dramatically and unusually cold and snowy winters. In one such year, the snowpack averaged around 3 feet of accumulation, and I decided to wade across my yard to refill my bird feeders. This effort ended poorly, as an unfortunate combination of snow depth and foot placement led to a slight but extremely painful dislocation of my knee. During the triage process in the hospital emergency room, I was shown the following diagram and asked to choose the appropriate number to describe my pain.

0 NO HURT | 2 HURTS LITTLE BIT | 4 HURTS LITTLE MORE | 6 HURTS EVEN MORE | 8 HURTS WHOLE LOT | 10 HURTS WORST

FIGURE 2. EXAMPLE PAIN RATING SCALE (6)

I considered the task at hand—certainly I was in quite a bit of pain, so I quickly ruled out the pain scores for 0, 2, and 4. The ride to the hospital had a soundtrack of moaning and crying, likely accompanied by a face similar to that for the pain score of 10. But was this really the worst pain I could be in? I imagined gory stories of tragedy and adjusted my gaze further down the scale. I wondered, is this *close* to the worst pain that would be possible? Likely not! So after a few thoughtful minutes, I reported to the nurse that I had a pain score of 6. She raised an eyebrow, made a note in my file, and sent me back outside to the waiting room.

Later on, I took a moment to reflect on this experience relative to my work in QRM. I quickly noticed that the numerical score I had relayed to the nurse was essentially meaningless. A pain score of 8 was not 20% less painful that a score of 10, just as a score of 8 was not precisely twice as painful as a 4. Nor would it be sufficient to note that a score of 2 required a 200mg dose of acetaminophen and four 10-minute sessions with an ice pack in order to move the pain to a score of 0. This scale, I realized, was what is referred to as an "ordinal" scale, with the numerical scores serving as stand-ins for a qualitative concept. The approach was not quantitative at all, but rather *semi-quantitative* "constituting or involving less than quantitative precision" (7). Further, this semi-quantitative scale was riddled with subjectivity; the process necessary to employ the scale required that I invoke several common cognitive shortcuts, or heuristics, in order to choose the best fit representation of my experience (8). In order to understand what a pain score of 10 might be, I was required to imagine scenarios that I had not personally experienced, that might feel as though they were the worst pain possible (*availability* heuristic). I had to also my own experiences with pain, through past illness or injury, that might be similar in some form of this particular instance (*representativeness* heuristic). I was initially fixed on the pain score of 10, since the illustration included a very unhappy person with a face full of tears, just as I was at the time (*anchoring* bias). I had to rely upon subjectivity in order to use the rating scale that was provided to me. And in the end, it failed to serve its intended purpose—to communicate my experience of pain to a third party.

This experience, and my subsequent reflection on its relationship with QRM, led me to several insights as follows:

- Ordinal scale risk tools often considered "quantitative" by virtue of the fact that numbers are involved are actually "semi-quantitative," and therefore functionally equivalent to qualitative rating scales, including the accompanying problems of subjectivity.
- Any rating scale that requires heuristics to be invoked in order for the scale to be used is necessarily flawed (refer to Ramnarine for additional information on heuristics (9).
- Provided we understand these limitations, manage them relative to our expectations and objectives within a risk analysis, and use additional techniques to preserve the integrity of the risk assessment, each of these flaws can be overcome.

GUIDELINES FOR THOROUGH RISK ANALYSIS

In the author's experience, the application of the following guidelines can help minimize the potential impact of subjectivity associated with the use of ordinal risk rating scales and ensure the risk analysis process is sufficiently rigorous.

1. Spend Time Developing Risk Rating Scales—Construct Detailed and Meaningful Criteria for Each Numeric Score

Some of the subjectivity associated with the use of semi-quantitative tools can be minimized if the associated ordinal risk

ranking scales are thoughtfully established. To do so, each numeric score should have a detailed list of criteria that can level-set team members as they seek to apply the scoring model in an actual assessment exercise. An ordinal rating scale should include:

- A qualitative term corresponding to the general principles of the scoring model (e.g. likelihood, probability, frequency, detection, severity, uncertainty, etc.).
- Quantitative reference points for each numeric scale, where at all possible. In the event quantitation cannot be determined (e.g. early in development, introduction of new technology), detailed criteria using actual, tangible examples of instances where a given score would be the appropriate choice should be included.

Typical ordinal rating scales for occurrence often look something like that in Table 1 – these kinds of scales have few if any design features to minimize subjectivity and heuristics-related problems.

RATING	CRITERIA
1	Remote likelihood of occurrence
2	Failure unlikely to occur
3	Average likelihood of occurrence
4	Failure likely to occur
5	Failure extremely likely to occur

TABLE 1: EXAMPLE OF A TYPICAL RISK RATING SCALE FOR OCCURRENCE THAT IS VULNERABLE TO THE PROBLEMS OF SUBJECTIVITY AND HEURISTICS

However, construction of more detailed criteria using the points noted above might look something like Table 2.

RATING	LIKELIHOOD	CRITERIA
1	Remote	<0.01% probability of failure. Preventive controls are designed into the process, and sign supporting data exists to demonstrate effectiveness.
2	Unlikely	0.01% probability of failure, or 1 in 10,000 batches/lots. Preventive controls are in place and effective, however pote exists for control failure under certain circumstances that ha not been fully defined.
3	Average	0.1% probability of failure, or 1 in 1000 batches/lots. Preventive controls are in place but minimal data exists to support effectiveness claims.
4	Likely	1% probability of failure, or 1 in 100 batches/lots. Minimal preventive controls defined, or preventive controls minimally effective.
5	Extremely likely	10% probability of failure, or 1 in 10 batches/lots. No defined preventive controls. No available data regarding potential rate of failure.

TABLE 2. EXAMPLE OF A RISK RATING SCALE FOR OCCURRENCE THAT ATTEMPTS TO MINIMIZE SUBJECTIVITY AND HEURISTICS

The addition of the quantitative reference points in the above example, along with a discussion on the types and effectiveness of preventive actions in reducing failure rates, can help ground the discussion in data and lead to more evidence-based risk ranking decisions.

2. Train Risk Team Members on Heuristics at a Practical Level -- Not a Theoretical Level

Risk team members should be familiar with the general concepts of heuristics and be able to identify when these might be in play during a given risk assessment. An action plan should also be in place to help the team minimize the impact of heuristics, especially during brainstorming and risk ranking. This might entail the use of a simple job aid that each team member can refer to during the assessment. Table 3 provides an example [see O'Donnell (10)].

HEURISTIC	DESCRIPTION	SOUNDS LIKE...	ACTION PLAN
Anchoring	Tendency for people to rely too much on the first piece of information offered	"I think the detectability is a 4!" followed by the team discussing differences between a 3, 4, and 5.	Ask for a package of supporting evidence regarding the risk under discussion. Ensure sufficient time for each team member to review the data and develop their own opinion. Allow for discussion on the rationale for each opinion. Consider "silent balloting" for risk ratings so each team member can write down his or her preferred score without being influenced by others.
Availability	Tendency to over/underestimate the probability of occurrence based on the ease with which it can be remembered or imagined	"This issue happened just last quarter!" followed by a team discussion regarding whether the likelihood is a 4 or a 5.	Challenge the team member who recalled experiencing a similar issue. Ask when it might have last happened, to help determine frequency and understand the circumstances. Ask someone who directly experienced the issue to share their knowledge, including historical failures and recent CAPAs/improvements.
Representativeness	Tendency to over/underestimate the probability of occurrence on a given failure based on generalizations or imprecise deduction reasoning, or where certain important pieces of information are overlooked as one seeks to emphasize the general over the specific.	"That department *always* messes up!" or similar sweeping generalization, followed by a team discussion regarding whether the likelihood is a 4 or a 5.	Beware of words like "always" or "never." Watch for scope creep, where team members might rely on anecdotes or information from unrelated risks and project those into a separate issue. Refocus the team on the specific risk under discussion, rather than impressions of the rate of failure in general. Ask a neutral party without knowledge of the specific background context to review the data and draw conclusions regarding the risk score.

TABLE 3. EXAMPLE OF JOB AID TO ASSIST WITH HEURISTIC IDENTIFICATION AND MINIMIZATION OF IMPACT ON RISK ANALYSIS OUTCOMES

Each of the above action plans for dealing with heuristics has a common approach: "understand how the heuristic may manifest, and then seek to counteract it." A related and important means of counteracting heuristics in any quality risk management activity is to compile and thoroughly review the available scientific evidence and supporting data for the issue at hand. Of course, a well-trained risk facilitator can serve as the watchdog to protect against the intrusion of heuristics, and can refresh the risk team on these concepts just prior to commencing the risk analysis portion of the risk assessment.

3. Require a documented rationale for each selected score, with references to scientific studies, data, and trending

A third party should be able to arrive at the same conclusions as the risk team, when given the same information. A rationale for the risk rating score, including identifying information for relevant studies, trending data and time periods, summary reports, and scientific literature should therefore be included in the risk analysis and assessment documentation. Risk analysis worksheets should include a space to document this rationale, and any strategies employed for risk ranking throughout the assessment should be described in the resultant report. A brief statement such as "team scored likelihood a 2 based on environmental control trending data covering calendar year 2014; see Quality Management Review slide deck and meeting minutes dated January 18, 2015" is often sufficient. The practice of documenting conclusions and risk judgments is not only good QRM practice, but also good knowledge management.

4. Require That Sources of Uncertainty, Assumptions, or Gaps in Knowledge be Discussed During the Scoring Process and Explicitly Disclosed in the Risk Management Documentation

Some level of uncertainty is inherent in any risk analysis. Indeed, ISO 31000 defines risk as "the effect of uncertainty on objectives," (11) implying that were certainty present, no risk assessment would be needed! The risk team should have a candid discussion about any gaps in knowledge, sources of uncertainty, and underlying assumptions that may bias the risk analysis, both prior to commencing the analysis and at any point during the activity when the team becomes aware of such a gap. Sources of uncertainty may include unknown root causes, lack of clarity regarding the effectiveness of an individual preventive or detection control, pending scientific inquiries (e.g. experiments or studies), or sources of variability within a process that have not been fully characterized. These knowledge gaps should be documented and any potential influence on the outcomes of the risk analysis should be disclosed. In the event the level of risk may be underestimated due to uncertainty, it is best to err on the side of conservative and assign higher risk ratings, particularly when the risk could potentially affect the patient.

5. Avoid Using *Only* Risk Priority Numbers (RPNs) in Risk Reduction and Acceptance Decisions

Though RPNs offer a strong indication of the level of risk involved with a particular scenario, they should not be used in isolation to determine the need for risk reduction or the acceptability of the risk. The rationale behind this is due to fact that determining an RPN through the multiplication of individual likelihood, severity, and detectability risk ratings, is a mathematically invalid operation given the ordinal nature of the individual ratings (3). The relative risk ranking associated with each discrete element is blurred once they are manipulated in the RPN calculation; such is the problem with multiplying ordinal scales (3). Multiplying ordinal numbers is invalid because their magnitudes are not meaningful in a mathematical sense – for example, a hazard that is assigned a 4 on an ordinal scale for its probability of occurrence is not necessarily twice as likely to occur as a hazard with a score of 2 on the same scale. Multiplication (or addition, or any other mathematical operation) dissolves the rank order associated with the different scores.

It is also important to realize that RPNs that are the result of multiplied ordinal numbers can often represent very different risk situations, but this is often disregarded (or not noticed) by virtue of a specific number being assigned to the risk. For example, an RPN of 20 in an FMEA that employed 1-5 rating scales for frequency, detectability, and severity can be the result of 8 possible combinations, as follows:

SCENARIO	FREQUENCY	DETECTABILITY	SEVERITY	RPN
A	1	4	5	20
B	1	5	4	20
C	2	2	5	20
D	2	5	2	20
E	4	1	5	20
F	4	5	1	20
G	5	2	2	20
H	5	4	1	20

While the RPN is the same in each case, the 8 situations above represent quite different risk scenarios that should be

considered individually with regards to acceptability. For example, scenarios A, C, and E all represent instances in which the consequences are incredibly dire (severity = 5). However, the narratives associated with these scenarios differ significantly. In scenario A, for example, we have what might be considered a "black swan" event—one that is rare yet catastrophic. The scenario is very unlikely to occur (frequency = 1), however if it did, it's improbable we would know about it (detectability = 4) before the catastrophic outcome (severity = 5) is realized. Conversely, in scenario E, we have a risk or failure that occurs quire often (frequency = 4), but we can detect it readily (detectability = 1) before the catastrophic consequence (severity = 5) transpires. Scenario C represents a middle ground, where a failure that could pose a severe consequence (severity = 5) is fairly infrequent (frequency = 2) and readily detectable (detectability = 2).

These risk scenarios are of course all different, but this is not evident when one only considers the RPNs, without taking into account the individual scores that gave rise to those RPNs. The RPNs alone do not indicate where the differences lie. Thus, when RPN thresholds are used to determine whether risk reduction or acceptance is warranted, flawed decision-making is often the result.

Would we advise the risk team to treat all scenarios A, C, and E the same with regard to recommending risk mitigation actions for them, or with regard to accepting the risks presented by each? What if we compared them with scenario H, where a failure is extremely likely to occur (frequency = 5) and is very difficult to detect (detectability = 4), but which would result in negligible consequences (severity = 1)? This scenario has the same RPN of 20, but is very different and has a potentially different level of acceptability. Is each of these scenarios the same? Would we advise the risk team treat them the same with regarding to recommending risk mitigation, or accepting the risk? What if we compared these with scenario H, were a failure is extremely likely to occur (frequency = 5) and very difficult to detect (detectability = 4), but would result in negligible consequences (severity = 1)? Each of these has the same RPN but very different derivations, and therefore potentially different levels of acceptability.

In order to ensure the appropriate risk acceptance decisions are made, some companies apply a second set of considerations to be used in concert with the RPN. For example, a company may require any risk with a severity score of 5 to be reduced as low as possible, irrespective of the RPN, since such scenarios could result in a significant impact to process, product, or patient and should be handled with those consequences in mind. A company may also require risks with a frequency or detectability of 5 to be reduced as well, as in many cases these are indicative of lack of effective process controls. These considerations can be added to an analysis of the RPN to assist with decisions regarding the need for risk control and overall risk acceptance; in this way, the RPN remains useful but the rank-order masking effect caused by the multiplication of ordinal scales can be overcome.

Another option to ensure the outcomes of semi-quantitative risk assessments facilitate more robust decision-making is to abandon the RPN approach altogether and replace it with what Wheeler calls an "SOD code." (4) Rather than multiplying the frequency, detection, and severity ratings together to yield an RPN, Wheeler's method involves creating a three-digit code for each risk, listing the severity (S) rating first, followed by the occurrence (O; aka likelihood / frequency / probability) rating, and finally the detection (D) rating. When the SOD codes are then listed in numerical order, each risk will be prioritized first according to severity, then occurrence, with detection as the final consideration. This method allows for prioritization of risk control to be weighted towards those risks that could impact the patient, thereby eliminating any mathematical fallacy while preserving the focus of risk control.

Whatever the methods chosen to address the limitations of RPNs, the risk team should acknowledge that Risk Priority Numbers, as the name suggests, are intended to be used to prioritize risks for reduction; they should not be used in isolation to accept risk (5). The decision to accept risk should be made only after a thorough evaluation of the risk scenario (i.e. combination of failure mode, cause, and effect), risk level (i.e. RPN), and potential harm involved (e.g. patient exposure), and where applicable, a balance against the benefits of proceeding with the process or product under the risk conditions.

CONCLUSION

With the ten year anniversary of the publication of ICH Q9 and a decade of Quality Risk Management learning and expeeience behind us, our industry has an opportunity to revisit its application of QRM principles and methods to optimize outcomes. This paper explored the challenges associated with traditional risk rating scales and methods, and provided suggestions to optimize risk estimation criteria and the risk analysis process. While it focused on the use of semi-quantitative risk tools and ordinal risk rating scales, the concepts can be applied to qualitative approaches as well to improve the conduct of risk

analyses. A journey towards true QRM maturity will require a revisit of the most commonly applied portions of the QRM lifecycle.

REFERENCES

1. International Conference on Harmonization (ICH) Guideline Q9, "Quality Risk Management." 2005.
2. Hubbard, D. *The Failure of Risk Management: Why Its Broken and How to Fix It*. Wiley. 2009.
3. Hubbard, D. and D. Evans. "Problems with scoring methods and ordinal scales in risk assessment." *IBM Journal of Research and Development*. Volume 54, Number 3. May/June 2010.
4. Wheeler, D. "Problems with Risk Priority Numbers." *Quality Digest*. 27Jun2011. http://www.qualitydigest.com/inside/quality-insider-column/problems-risk-priority-numbers.html# Retrieved on 14Nov 2015.
5. Kaplan, S. and B.J. Garrick. "On the Quantitative Definition of Risk." *Risk Analysis*. Volume 1, Number 1. 1981.
6. Wong-Baker FACES® Pain Rating Scale. http://wongbakerfaces.org/ Retrieved 14Nov2015
7. Definition of "semi-quantitative" taken from Merriam-Webster, http://www.merriam-webster.com/dictionary/semiquantitative, on 14Nov2015.
8. Kahneman, D. *Thinking, Fast and Slow*. Farrar, Straus and Giroux. 2011.
9. Ramnarine, E. "Understanding Problems of Subjectivity and Uncertainty in QRM." *Journal of Validation Technology*. Volume 21, Number 4. Dec 2015.
10. O'Donnell, K. "Strategies for Addressing the Problems of Subjectivity and Uncertainty in Quality Risk Management Exercises. Part I – The Role of Human Heuristics." *Journal of GxP Compliance*. Volume 14, Number 4. Autumn 2010.
11. International Organization for Standards (ISO) 31000:2009. "Risk management – principles and guidelines."

ACKNOWLEDGEMENT

Source URL: http://www.ivtnetwork.com/article/risk-analysis-and-ordinal-risk-rating-scales%E2%80%94-closer-look