Reliability Analysis Combining Multiple Inspection Techniques

Dag HORN, AECL Chalk River Laboratories, Chalk River, Ontario, Canada

Abstract. The benefits of combining several independent inspection techniques must be quantified before clients are willing to incur the additional expense and the perceived potential for conflicting results inherent in the application of multiple techniques. Examples from the Canadian nuclear industry illustrate the quantitative improvements to reliability offered by combining the results of eddy current and ultrasonic testing through a number of statistical techniques. When inspection frequency and sample size are calculated on the basis of probabilistic assessments, the improvements to detection probability attainable with multiple techniques can result in reduced inspection requirements.

1 Introduction

1.1 Multiple techniques

Inspecting by multiple independent techniques permits decisions to be made on the basis of more information. While that additional information can be helpful in resolving critical questions, practical considerations often dissuade the client from pursuing inspection by multiple techniques. The concerns include:

- additional expense and time,
- the potential for conflicting results, and
- difficulty in quantitatively interpreting multiple results.

Mathematical methods for combining data can consolidate multiple-technique results for comparison to single-technique results. Assessment of the alternatives by means of reliability analysis may then justify the expense and time required for additional inspections. Even the observation of conflicting findings from different techniques can, rather than discrediting one or another of the techniques, provide quantitative information on the confidence that should be assigned to the inspection results. This is valuable because knowing the sensitivity or accuracy of results is often more important than improving their sensitivity or accuracy: in a probabilistic safety assessment, the uncertainty 'tail' from a low confidence limit can significantly influence event frequency.

1.2 Topics addressed

Section 2 of this paper describes a number of techniques for combining inspection results from multiple independent measurements. These range from a simple logical OR of hit – miss detection results to sophisticated data fusion algorithms. Reliability analysis tools available for assessment of results are discussed in Section 3. We present examples from the Canadian nuclear industry to illustrate the potential gains attainable (Sections 4 and 5) and conclude in Section 6 with a forward look at the potential of reliability analysis for assessing combined results from multiple inspection techniques.

2 NDE Data Fusion

2.1 Independent measurements

An additional independent measurement increases the available information in a different way than a repeat of the same measurement does. A repeat merely improves the statistical precision of the aggregate result, reducing random effects. By contrast, an additional independent measurement can reveal phenomena to which the first technique might be insensitive. For example, eddy current testing (ET) and ultrasonic testing (UT) may preferentially respond to defects of different orientation or geometry. Examining both types of data then increases the overall probability of detecting a defect. Even within the parameters of one non-destructive evaluation (NDE) “method”, one specific technique, *e.g.,* ultrasonic pulse-echo detection, may respond to different defects than another (such as UT pitch-catch measurements).

2.2 Combining data

Statistical tools for combining data from separate measurements are standard in science and engineering [1] and techniques for making decisions based on multiple inputs have evolved under the label “data fusion” [2]. Data fusion concepts are eminently applicable to NDE; see, for example, Gros, Strachan, and Lowden [3]. Techniques for combining ET and UT results were explored in [4], beginning with standard statistical techniques:

- In its simplest form, combination can consist of viewing two yes-no (zeroth order) results separately and letting a response from either (*i.e.* a logical OR of the methods) define an indication.
- First-order techniques involve summing or averaging to combine measured values, each assumed related to a defect characteristic like depth or area. The result ignores the width of the distribution in defect characteristic for a measured value and deals only with the first moment of the distribution.
- The second moment, namely the width of the distribution in defect characteristics for a given measurement value, can be used in producing an average that is weighted by the relative accuracy of each method. This also provides an uncertainty estimate for the combined value.

Knowledge of higher moments, such as skewness, can further refine the combined value. However, when this level of detail is known, the probability distributions can be used directly through more sophisticated data fusion techniques, listed in the next sub-section.

2.3 Data fusion

Data fusion [2] provides mathematical techniques for combining incommensurate, uncertain, or conflicting information into joint probability distributions. When constructed from NDE data, these distributions may provide sizing information for accept/reject decisions.

- In *classical inference*, the probability distribution for signal amplitudes given a defect size is obtained from tests of each method. Their product is a joint probability distribution, which must then be inverted to obtain the probability of a defect size, given a set of responses from the different inspection methods.
- *Bayesian inference* produces a similar joint probability distribution, but one that is multiplied by the *a priori* distribution of flaws. In other words, when measurement uncertainties allow a range of outcomes, the calculated likelihood of rare events is reduced relative to that of common events.
- The *Dempster-Shafer* method offers a means of dealing with conflict between measurements. The approach most commonly used in NDE work assumes that probabilities outside a measurement bin represent no particular belief about the primary interval.

3 Assessing the Combined Results

3.1 Reliability analysis

Reliability analysis is most familiar to the NDE industry from probability of detection (POD) studies. See, for example, Berens [5]. In fact, reliability analysis is a field of study in its own right [6], with applications in process control, accelerated life tests, and life cycle management. In the present work, reliability analysis is used to quantify:

- the gain in POD for a combination of techniques, compared with the detection probability for each technique individually,
- the increase in false positives associated with the detection probability gain, and
- the optimum trade-off between improvement in POD and reduction of false positives.

3.2 Probability of detection

Detection probability is traditionally calculated from the ratio of detected indications to the total number of indications present in a particular size interval and the binomial distribution is invoked to obtain the lower-bound POD at a specific confidence level. Only with a large number of test samples in each defect size interval does this method permit a narrow band of detection probabilities to be established with high confidence. To avoid the need for numerous samples in size ranges where detection is not in doubt, the assumption is often made that, all other things equal, detectability increases with defect size. One manifestation of this assumption is the “method of optimised probabilities” [7], in which assigning credit for detection of smaller defects reduces the need for larger flaws in the test sample set. Another approach is to postulate the general form of the POD curve and fit the parameters of the curve to test outcomes, either as “hit-miss” results or as amplitude data, recorded as a function of defect size [5]. An advantage of these methods over the method of optimised probabilities is that they use the continuum of defect sizes rather than binning them.

4 Example from Manufacturing Inspection

4.1 Data set

The manufacturing inspection of CANDU® pressure tubes is performed by both eddy current and ultrasonic testing. An extensive study of dual-technique manufacturing inspection [8] was analysed in terms of standard and data fusion combination techniques in reference [4]. The richness of that data set makes it ideal for assessing the effectiveness of combination of results from different NDE methods, and we return to it in the present work. In the data set there are 69 flaws of depth 0.08 mm or greater, which we shall arbitrarily define as rejectable, and 39 indications less than 0.08 mm, which we shall here consider acceptable.

4.2 Inspection errors

The two errors of concern are therefore:

- missing a rejectable flaw (type-1 error) or
- falsely rejecting an acceptable indication (type-2 error).

With the rejection thresholds set to make acceptance of a rejectable flaw and rejection of an acceptable flaw equally likely, the tally of inspection errors is 9 missed defects and 9 false calls for ET and 10 missed defects and 11 false calls for UT. The error rates for fixed threshold are listed in Table 1. Lower-bound POD curves for the 95% confidence level, determined by a maximum likelihood analysis of hit-miss data [5], are shown in Figure 1 for ET and UT.

Table 1. Error rates for fixed threshold

	ET	UT	ET.OR.UT	ET.AND.UT	AVG (ET,UT)
missed callls	9	10	3	16	3
false calls	9	11	14	6	9

4.3 Data combination

A logical OR of the methods performs the union of the two sets of calls, rejecting flaws called by one or both of the techniques, and reducing the number missed to 3. This results in a more sharply rising POD curve, as shown by the heavy line in Figure 1, and indicates a significant detectability increase in the critical threshold region. In this case, the combination technique increases the number of false calls, recording all those seen by either technique for a total of 14. A logical AND performs the intersection of the two sets of calls. Since a reject decision requires rejection by both methods, more flaws (16) are missed but fewer false calls (6) are made. These results are summarized in Table 1.

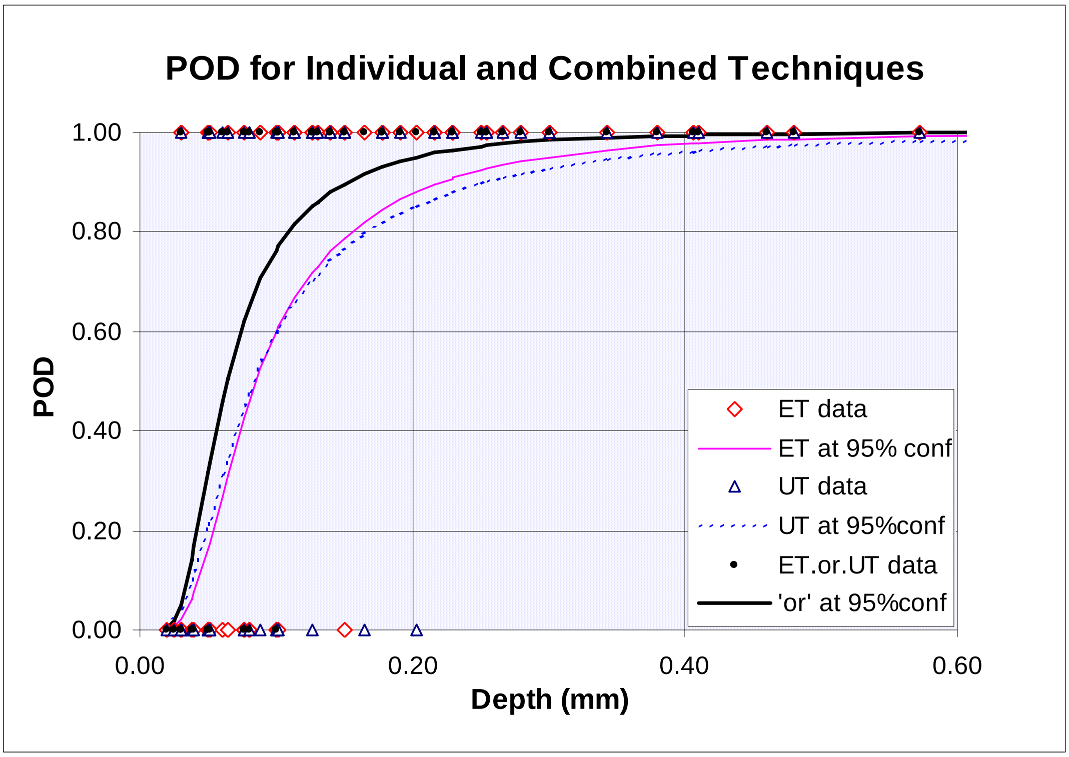

Figure 1. Comparison of combined lower-bound (95% confidence) POD (heavy curve) to detection probability for two methods individually (thin curves).

Finally, an average of the ET and UT signal amplitudes and threshold values should also lead to nearly equal rates of type-1 and type-2 errors. As technique-dependent fluctuations in signal distribution for a specific flaw are averaged, a narrower signal distribution is expected, giving fewer of each error type. Results listed in the rightmost column of the table agree with this prediction.

4.4 Optimised threshold

Simply lowering the rejection threshold can often increase POD, but improving detection rates at the expense of increased false calls may be counterproductive, and a quantitative assessment of the trade-off is needed. Earlier work [4] addressed this issue by means of a relative operating characteristic, which plots detection rate as a function of false call rate for varying rejection threshold [9]. Table 2 shows the sum of false calls and missed defects. The minimum total error rates obtained by optimising the thresholds are lower for combined data than for the techniques viewed individually, with the AND being the most effective of the "logical" combinations. An equally effective quantity is the averaged amplitude of ET and UT signals, which reduces the signal fluctuations, as discussed above.

Table 2. Minimum error rates for varied threshold

	ET	UT	ET.OR.UT	ET.AND.UT	AVG (ET,UT)
minimum of (type-1 + type-2 errors)	18	16	15	12	12

4.5 Observations

For this data set we observe that:

- combining results from different techniques increases the POD,
- for fixed threshold, POD improvement may come at the cost of additional false calls,
- with optimised thresholds, the total number of inspection errors is reduced for combined results,
- a logical OR is best for minimizing the rate of missed defects,
- a logical AND is best for reducing the rate of false calls, and
- averaging permits an amplitude-based POD analysis, which offers improved precision over a hit-miss analysis by incorporating additional information.

5 Example from Field Inspection

5.1 Data set

Cracking data from a difficult-to-access carbon steel component is used to illustrate some of the challenges encountered in inspection for rare events. Less than a dozen field data meet the combined requirements that:

- crack geometries are known from destructive examination,
- field inspections are done with techniques that produce commensurate quantities on comparable amplitude scales (same method, probe response, reference notch, etc.), and
- field detection amplitudes are known.

Furthermore, laboratory simulations of such cracks have more favourable ultrasonic detection properties than the field data; therefore, as a minimum, enough field data must be available for scaling of the laboratory data. To maximize the number of real flaws in the sample set, a scaling factor was applied to adjust for the difference between amplitudes from inside and outside surface cracks, permitting aggregation of the two phenomena. The available data, collected over several years of inspection, are plotted in Figure 2, with missed calls assigned an amplitude value of 1.0, representing the reporting level for the inspection.

5.2 False calls

With few real detected flaws, POD can be poorly defined, so to ensure a conservative assessment in the critical size region of 20 mm^2, a low rejection threshold was specified. In a recent inspection, post-remediation analysis showed 6 false calls and only one real defect in a set of 7 calls. The amplitude distribution for these false calls, plotted at a minimum nominal value of 1mm^2 for crack size, is superimposed on the flaw data of Figure 2; a normal distribution fitted to the false call amplitude distribution has a standard deviation of 3 times the reporting threshold. Evidently, to halve the false call rate, the decision threshold would have to be raised from 1.0 to 3.0. The effect such a high threshold would have on POD is shown in Figure 3: detection capability for flaws in the critical size range drops by a factor of two, making this an unsatisfactory means of reducing false calls.

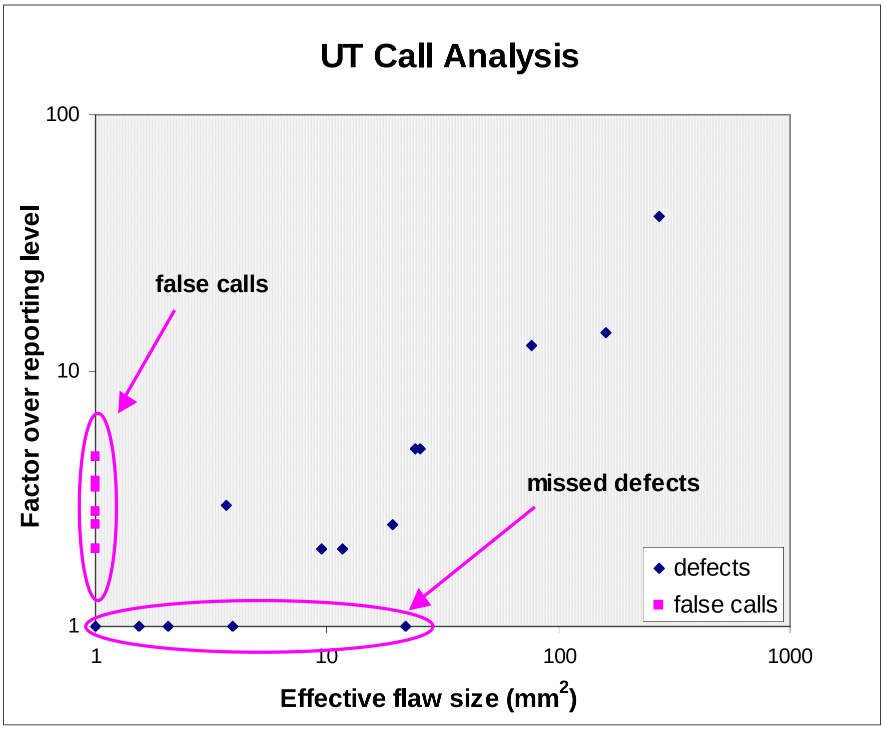

Figure 2. UT signal amplitude as function of defect size. Missed calls are assigned an upper limit amplitude value of 1, representing the reporting level; false calls are plotted at a negligible flaw size value of 1 mm^2.

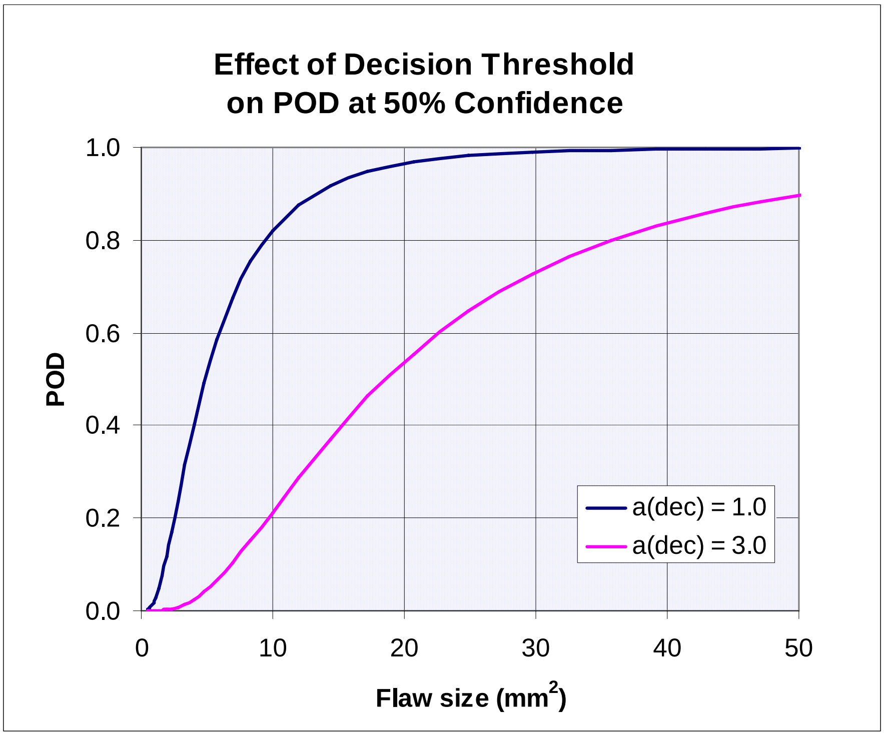

Figure 3. POD at 50% confidence level showing effect of raising decision threshold.

5.3 Performance of secondary technique

As seen in Section 4, the use of multiple techniques offers the potential for false call mitigation. On a trial basis, 78 of the above locations were inspected by ET, representing approximately 10% of the sample size for the standard UT inspection and including the three locations at which the UT inspection detected potential outside-surface indications. Considering only the subset of the data for which both techniques were applied, and only the performance for outside-surface cracking, to which ET would be sensitive, three indications were observed out of 78 possible locations. The secondary technique identified two of the three indications that were also called by the primary technique; one was a true defect and one a false call. The third indication was a false call and was not confirmed by ET. Threshold optimisation of the type performed in Section 4 requires a more extensive data set than available here, but results for the 78 locations inspected by both techniques can be tabulated for fixed threshold. The missed call rate is not known, since only locations with an NDE response were destructively analysed.

Table 3. Error rates for fixed threshold

	ET	UT	ET.OR.UT	ET.AND.UT
true call	1	1	1	1
false call	1	2	2	1

5.4 Observations on sparse data from field inspection

For this sparse data set:

- the number of events is statistically adequate only to estimate inspection error rates within a factor of two,
- the false call rate observed for ET is one in 78, which is comparable to the observed UT rate of six for a sample size one order of magnitude larger,
- the rate of missed calls is unknown, since most components with no NDE indications are left in service,
- an estimate of the size distribution for undetected flaws may be based on analysis of removed components,
- a logical AND reduces the rate of false calls, and
- a logical AND does not reduce the rate of true calls.

6 Conclusions

6.1 Motivation for improved NDE

A quantitative demonstration of increased probability of detection, reduced false call rates, and narrower confidence bands about the POD can help clients optimise inspection intervals and reduce inspection extent, thus providing the schedule, quality, and economic drivers for investment in improved inspections. In sampling inspections, a reduction in the number of missed defects justifies a smaller inspection scope; in 100% inspections, reduction in the size of the largest missed defect extends the operating period before the next inspection is required. Reduction of the false call rate saves remediation time and expense. In this paper we have illustrated the potential of multiple inspection techniques, data combination methods, and reliability analysis to attain these ends.

6.2 Findings

The balance between the rates of two types of inspection errors, missed defects and false calls, can be shifted by raising or lowering the detection threshold. However, when the signal amplitudes associated with a particular flaw size have a broad distribution, it may be that satisfactory performance with respect to one type of error can only be achieved with unacceptable rates of the other. Inclusion of a second, independent inspection method can reduce the total number of errors. Two data sets, one nearly ideal and one realistically incomplete, have been used to illustrate that various means of data combination have different overall effectiveness and can be chosen to preferentially reduce one type of inspection error or the other.

6.3 Future work

- A major challenge in the implementation of multiple-technique inspection is the extensive set of data required for qualification and the detailed analysis of flaws needed to validate the NDE results. Since field data are scarce, it may be necessary to establish a scaling relationship between signals from artificial flaws and those from real defects to permit use of the former in qualification work.
- Logical combinations (AND, OR) and averaging are appropriate to limited data sets where detailed probability distributions are not known. For large, well-understood data sets, more sophisticated data fusion techniques may offer an incremental advantage.

7 References

[1] See, for example, Meyer, S.L., Data Analysis for Scientists and Engineers, (Wiley, Toronto, 1975).

[2] Hall, D.L., Mathematical Techniques in Multisensor Data Fusion, (Artech House, Boston, 1992) and references therein.

[3] Gros, X.E., P. Strachan, and D.W. Lowden, *Theory and Implementation of NDT Data Fusion*, Res. Nondestr. Eval. **6** (1995) 227.

[4] Horn, D., and Mayo, W.R., *NDE Reliability Gains from Combining Eddy Current and Ultrasonic Testing*, NDT&E International **33** (2001)351.

[5] A.P. Berens, *NDE Reliability Data Analysis*, Metals Handbook, Vol. **17**, ASM International (1989) 689.

[6] W. Q. Meeker and L. A. Escobar, *Statistical Methods for Reliability Data*, John Wiley & Sons, New York, 1998.

[7] Packman, P.F., S.J. Klima, R.L. Davies, J. Malpani, J. Moyzis, W. Walker, B.G.W. Yee, and D.P. Johnson, *Reliability of Flaw Detection by Nondestructive Inspection*, ASM Metals Handbook, 8th Ed., **11** (ASM, 1976).

[8] Mayo, W.R., Reliability *of Nondestructive Testing*, CSNDT Journal 12, (1991)14.

[9] Heasler, P.G., S.R. Doctor, and T.T. Taylor, *Quantifying Inspection Performance Using Relative Operating Characteristic Curves*, 10th Int'l. Conf, on NDE in the Nuclear and Pressure Vessel Industries, M. J. Whittle et al., eds. (ASM International, 1990) 481.