Session 3B (3) – Inspection Qualification

Chairman – ***M Mienczakowski***

11.15 Derivation And Use Of Probability Of Detection Curves In The Nuclear Industry

Authors - ***Luca Gandoss & Kaisa Simola***

The use of probability of detection curves to quantify NDT reliability is common in the aeronautical industry, but relatively less so in the nuclear industry, at least in European countries. The main reason for this lies in the very nature of the components being inspected. Sample sizes of inspected cracks tend to be much lower, and it is often very difficult to procure or manufacture representative flaws in test pieces in an high enough number to allow drawing statistical conclusions on the capability of the NDT system being investigated. Similar considerations led to the development of the ENIQ inspection qualification methodology, based on the idea of the Technical Justification, i.e. a document assembling evidence and reasoning providing assurance that the NDT system is indeed capable of finding the flaws which is designed to detect. The ENIQ methodology has become widely used in many European countries, and is gaining appreciation outside Europe as well, but the assurance it provides is usually of qualitative nature. The need to quantify the output of inspection qualification has become more and more important, especially as risk-informed in-service inspection methodologies become more widely used. To quantify risk reduction after an inspection, a measure of the NDT reliability is necessary. A probability of detection (POD) curve provides such metric.

The purpose of this paper is to briefly review the statistical models proposed to quantify NDT reliability, to highlight the potential problems that can arise if the main underlying assumptions and requirements are not verified, and to clarify the confusion that can arise over the true nature of the POD curve and associated confidence bounds.

DERIVATION AND USE OF PROBABILITY OF DETECTION CURVES IN THE NUCLEAR INDUSTRY

Luca Gandossi
European Commission, Joint Research Centre, Institute for Energy
Westerduinweg 3, 1755 LE, Petten, the Netherlands
Tel. 0031-224-565250
Luca.gandossi@jrc.nl

Kaisa Simola
VTT Technical Research Centre of Finland
P.O. Box 1000, FI-02044 VTT, Espoo, Finland

Abstract

The use of probability of detection curves to quantify NDT reliability is common in the aeronautical industry, but relatively less so in the nuclear industry, at least in European countries. The main reason for this lies in the very nature of the components being inspected. Sample sizes of inspected flaws tend to be much lower, and it is often very difficult to procure or manufacture representative flaws in test pieces in a high enough number to draw statistical conclusions on the reliability of the NDT system being investigated. Similar considerations led to the development of the ENIQ inspection qualification methodology, based on the idea of the Technical Justification, i.e. a document assembling evidence and reasoning providing assurance that the NDT system is indeed capable of finding the flaws which it is designed to detect. The ENIQ methodology has become widely used in many European countries, and is gaining appreciation outside Europe as well, but the assurance it provides is usually of qualitative nature. The need to quantify the output of inspection qualification has become more and more important, especially as structural reliability modelling and quantitative risk-informed in-service inspection methodologies become more widely used. To credit the inspections in structural reliability evaluations, a measure of the NDT reliability is necessary. A probability of detection (POD) curve provides such metric.
The purpose of this paper is to briefly review the statistical models proposed to quantify NDT reliability, to highlight the potential problems that can arise if the main underlying assumptions and requirements are not verified, and to clarify the confusion that can arise over the nature of the POD curve and associated confidence bounds.

1. Introduction

Many factors influence whether or not the application of a non-destructive evaluation (NDE) system will result in the correct flaw detection. In general, NDE involves the application of a stimulus to a structure and the subsequent interpretation of the response to the stimulus, but great variability is inherent in the process. Repeated inspections of a single flaw can produce different responses to the stimulus response because of very

small variations in setup and calibration. Inspection of different flaws characterised by the same size can produce different responses because of differences in flaw geometry, flaw location, material properties, etc. The interpretation of the response can be influenced by the capability of the interpreter (manual or automatic). Further, human factors, such as fatigue or stress, and a challenging inspection environment, affect the NDE result [(1)].

Much of the modern literature on inspection reliability has been developed within the aeronautical industry and nearly all refer to a small set of seminal papers by Berens and Hovey [(2) (3)], which were produced in the early 1980s. These works recognised the fundamentally stochastic nature of flaw detection, and still today provide an excellent account of the analytical framework devised to treat NDE data in order to obtain POD curves. A very good review of more recent NDE reliability assessment practices is given in [(4)], a US Department of Defense handbook that provides guidance for establishing NDE procedures for inspecting flight propulsion systems, airframe components, etc.

Most of the inspection reliability issues arising in the aeronautical industry are clearly very similar to those in the nuclear industry. For instance, the statistical treatment of experimental data from test pieces is the same. One notable exception is possibly the fact that in the nuclear industry, given the very nature of the components being inspected, the sample sizes of inspected flaws tend to be much lower. In Europe, the ENIQ methodology for inspection qualification [(5)] was specifically developed because of the difficulty (or impossibility) of procuring or manufacturing representative flaws in test pieces in an high enough number to draw quantitative (statistical) conclusions on the reliability of the NDE system being investigated. The fundament of the ENIQ methodology is the Technical Justification, a document assembling evidence and reasoning providing assurance that the NDE system at hand is capable of finding - with suitably high reliability - the flaws it is meant to detect. This assurance is qualitative, and comes usually in the form of statements such as: "*The experimental evidence and supporting theoretical modelling results presented show that application of the pulse-echo method in accordance with the proposed procedure provides a reliable means of detecting all plausible defects of concern*".

The importance of obtaining a quantitative measure of inspection reliability is justified by the fact that structural reliability modelling and quantitative risk-informed in-service inspection methodologies are becoming more widely used within the nuclear industry in Europe [(6) (7)]. A measure of inspection reliability is essential to quantify the reduction of failure probability, and hence risk reduction, after inspection.

This paper is mostly aimed at engineers and experts working in the nuclear industry, especially in the inspection qualification field and in the structural integrity/structural reliability field. Our aim is to review the statistical models that have been proposed to quantify inspection reliability from experimental data. As stated above, the majority of the published literature on inspection reliability refers to the seminal papers by Berens at Hovey [(2) (3)] and uses the models presented therein, especially for what it concerns the choice of a parametric model to fit the data. The purpose of this paper is to argue that care must be used when using such parametric models and when interpreting the results,

as the main assumptions and requirements for the use of the models can be easily ignored. Also, confusion can arise over the true nature of the POD curve and associated confidence bounds obtained.

2. Statistical approaches to evaluate NDE reliability

As discussed above, the probability of detecting a flaw depends on many factors: not only factors intrinsic to the flaw itself (its shape, location, roughness, material, etc.) but also factors related to the inspection system (procedure, hardware, software, operator capability, etc.). For structural integrity reasons, probability of detection is nearly always derived (and plotted) against flaw size, in particular the through-wall extent of the flaw. It is thus not surprising that flaws having the same through-wall extent may have different detection probabilities. In this paper, we use "flaw size" as equivalent to "flaw through-wall extent". We assume, without further specification, that we are dealing with an NDE system which has been designed to detect flaws having specified characteristics. In the following we will describe different statistical approaches to deal with experimental data. With experimental data we mean a set of inspection results, where the NDE system at hand has been applied to either real or artificial flaws whose size is exactly known. In case of artificial flaws, these are assumed to be representative of the real flaws that the NDE system will have to find.

In all NDE systems a decision on whether a flaw has been detected is made by interpreting the response to an inspection stimulus. In ultrasonic testing (UT) the stimulus is an ultrasonic wave which is transmitted into the component under inspection, and the response is the reflected echo from the flaw. The amplitude of the response is normally indicated by $\hat{a}$. When such response is obtained, the operator (or the automated system) compares it with a threshold value. If $\hat{a}$ exceeds the threshold value, a positive indication is recorded. If $\hat{a}$ does not exceed the threshold value, a negative indication is recorded. In some inspections (or experiments), the response is in a non-recordable form (for instance, in a visual inspection): $\hat{a}$ has no quantitative value and the flaw is either detected or not detected. In these cases, the inspection reliability data are in a binary form: hit or miss, 0 or 1. Inspection data of this kind are usually called hit/miss data, and the analysis of such data hit/miss analysis. In other inspections, the response is in a recordable and quantitative form. For instance, is the peak eddy current voltage, or the amplitude (in dB) of the ultrasonic signal returning to the UT probe. Detection versus no detection decisions are made by comparing the magnitude of $\hat{a}$ to a decision threshold value, $\hat{a}_{dec}$. The $\hat{a}$ versus flaw size analysis is a method of estimating the *POD(a)* function based on the correlation between $\hat{a}$ and flaws of known size, a, and it is called $\hat{a}$ versus a analysis. Obviously, more information is carried by $\hat{a}$ data than hit/miss data, and therefore a POD curve can be obtained with a smaller number of experiments if $\hat{a}$ data is available.

In one statistical framework, the probability of detection for flaws of a given size can be thought as the population proportion of flaws of that size that will be found when many are inspected. This assumption leads to a statistical model based on the binomial distribution, which will be discussed in section 2.1. In this model, a great deal of experimental data is required to "prove" that the NDE system under consideration has a high enough capability of detecting flaws of a particular type. For example, if a 95%

POD must be shown at a 95% confidence level, this would require procuring or manufacturing test specimens containing 59 examples of this flaw type and the detection of all 59 in the test piece trial. These very considerations led the ENIQ network to state that *"[…] ENIQ argues that it is normally not possible or practicable to construct a convincing argument for inspection capability using results from test piece trials alone"* [8] and led to development of the concept of Technical Justification. If the above approach is repeated at different flaw sizes, a POD curve can be obtained in a point-wise fashion. This is discussed in section 2.2.

If multiple inspections were carried out on individual flaws, i.e. if the NDE data is in the form of detection frequencies for individual flaws, a liner regression model can be used to derive the POD curve. This is discussed in section 2.3.

Berens and Hovey [2] [3] realised that individual flaws of the same size had different probabilities of being detected, and hence proposed a model based on the idea that, at each flaw size, a distribution of flaw probabilities exist. The POD curve is then seen as the curve through the mean of detection probabilities. The additional idea they proposed was to assume that the (whole) POD curve has a functional form which is dependent on a small number of parameters (usually 2). Using this approach, discussed in section 2.4, the number of data points required to obtain a full POD curve becomes much smaller. We will also discus the fact that this is an important assumption to make, and that it is very important that the user is well aware of the implications.

2.1 Binomial model: NDE reliability at one flaw size

A single POD for all flaws of a given size can be modelled in terms of the probability of detecting a randomly selected flaw from the population of all flaws of that given size. In this framework, the proportion of flaws detected in a random sample is an estimate of the POD for that size. Each experiment is seen as a Bernoulli trial, and binomial distribution theory can be used to calculate a lower confidence bound on the estimate. This is the model we used in [9] and [10]. In particular, Appendix 1 of [10] describes in more detail the estimation of the relevant statistical quantities (such as confidence bounds), both in the classical and Bayesian statistical frameworks. It is interesting to note that in the aeronautical industry this model was extensively considered in the 1970s but is virtually not applied anymore for the quantification of NDE reliability [11], for reasons we are going to review in the following.

Let us suppose that N flaws of size a are inspected, and N_s are successfully detected. If p is the true (but unknown) probability of detection for the population of flaws, the number of detections is modelled by the binomial distribution. The probability of N_s detections (successes) in N independent inspections is:

$$p(successes = N_s) = \binom{N_s}{N} p^{N_s} (1-p)^{N-N_s} \tag{1}$$

The unknown quantity p can be approximate by the ratio N_s/N, which is not only a very natural choice, but also the unbiased, maximum likelihood estimate of the true value of p.

$$\hat{p} = \frac{N_s}{N} \tag{2}$$

$\hat{p}$ is the best estimate of p that can be obtained after carrying out the set of N trials, but alone does not tell much about the true value of p. A confidence interval must therefore be built around $\hat{p}$ to have a quantified likelihood of capturing the true value of p.
The α percent lower confidence bound, p_{CL}, on the estimate of POD can be obtained as the solution to the following equation [12]:

$$p_{CL} = \sup\left\{ p : \sum_{i=0}^{N_s-1} \binom{N}{i} p^i (1-p)^{N-i} \geq 1-\alpha \right\} \tag{3}$$

The interpretation of p_{CL} as a lower confidence bound is as follows [11]. If the experiment (comprising the inspection of N flaws of size a) was completely and independently repeated a large number of times, α percent of the calculated lower bounds would be less than the true value of p. In other words, there is α percent confidence that p_{CL} from a single experiment will be less than the true value.

Solutions to (Eq. 3) are plotted in Figure 1 for α = 90, 95, and 99 percent confidence limits, sample sizes ranging between N=20 and N=80, and assuming that all flaws were detected (N_s=N). To obtain the required sample size necessary to guarantee a specific lower confidence bound p_{CL}, one needs to enter the plot drawing a horizontal line at that value and find the intercept with the curve with the required confidence limit. For instance, to guarantee that the probability of detection is higher than 0.9 with 95% confidence, one needs to inspect and find 29 out of 29 flaws. Similarly, one finds that to guarantee that the probability of detection is higher than 0.95 with 95% confidence, one needs to inspect and find 59 out of 59 flaws.

Conversely, drawing a vertical line at a given sample size allows the determination of what p_{CL} can be guaranteed at different confidence levels. If, for instance, 50 flaws have been inspected and all found, we can state that p_{CL}=0.955 at 90% confidence level, p_{CL}=0.942 at 95% confidence level, and p_{CL}=0.912 at 99% confidence level. The obligatory trade-off is evident: increasing the confidence level in the estimate results in a decrease of the absolute value of p_{CL}.

Several objections to the use of this approach to quantifying inspection reliability have been raised [2] [3] [11].

1. The choice of a particular POD and confidence limits are normally made on a rather arbitrary basis. Often a 95% confidence limit is assumed to provide the required degree of conservatism, but there is no sound justification for such a choice.

2. A p_{CL} estimate is not a single, uniquely defined number but rather a statistical quantity. Any particular estimate is only one realisation from a conceptually large number of repeats of the demonstration program. Berens and Hovey [2] showed there can be a large degree of scatter in such estimates, depending on the POD function, analysis method, POD value, confidence level and number of flaws.

3. The p_{CL} characterization is not related to the size of flaws that may be present in the structure after an inspection. To calculate the probability of missing a large flaw requires knowledge of both the POD curve, POD(a), for all flaw sizes and the flaw size distribution of the flaws in the population of structural components being inspected.

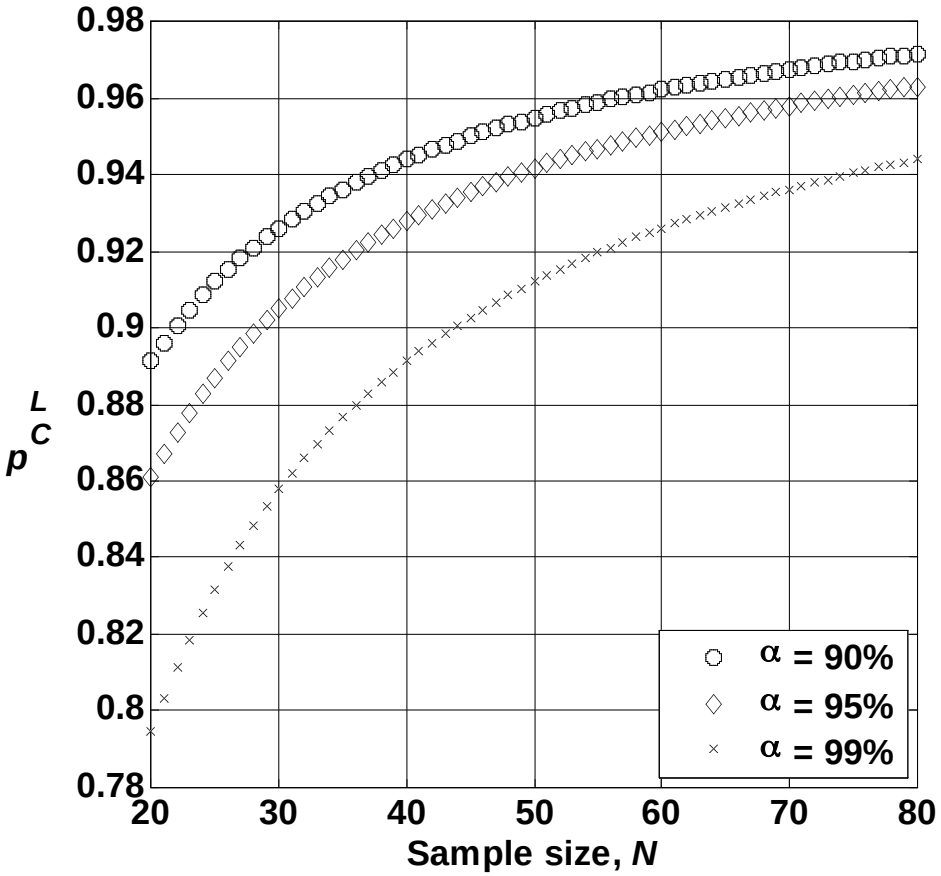

Figure 1. Lower bound (90, 95 and 99%) probabilities of detection for case $N=N_s$ (all flaws detected)

For the reasons outlined above, quantifying inspection reliability in terms of the entire POD(a) function (section 2.4) has evolved as the preferred method.

2.2 Binomial model: NDE reliability at multiple flaw sizes

In this category of experiments, many components covering a range of flaw sizes are inspected once and the results are used to estimate the POD as a function of flaw size with confidence limits. Statistically, this approach can be seen as an extension of the binomial model described in the previous section, repeated at several different flaw sizes.

In general, it will not be possible to have flaws that can be neatly partitioned into sets, each set being characterised by a different flaw size and with all flaws of the set having exactly the same size. More likely, the flaws will span a continuous interval of flaw sizes. The main idea of this approach is still to group the flaws in intervals of flaw size,

and to assume that all flaws within a specified interval have approximately the same POD. The number of detections for the interval is then modelled with the binomial distribution, exactly as described in the previous section, and the lower confidence bound is usually assigned to the flaw size at the upper end of the interval.

Various methods have been developed on how to form these flaw size intervals, ranging from very simple (partitioning the range of data into equal intervals) to more sophisticated ones. These are reviewed in [2]. In simplest method, called the Range Interval Method, flaw size intervals are defined with equal lengths across the range of data. Due to the somewhat random nature of the flaw sizes in the components being inspected, the interval constructed according to this method will very likely contain different numbers of flaws. For this reason, the estimate of the POD curve and its lower confidence bound can exhibit an erratic behaviour.

An example is shown in Figure 2. In this example, the inspection results for eddy current inspections of 361 etched fatigue flaws in aluminium flat plates [13] were grouped in 29 flaw intervals. The solid dots represent, for each flaw size interval, the value of the point estimate, $\hat{p}$. The empty dots represent the value of the 95% lower confidence bound, p_{CL}. The latter show an extremely erratic behaviour, due to small sample sizes in certain intervals, even though all flaws greater than 3.6mm were detected. For example, the very low confidence bound at 5.89mm resulted from the fact that particular interval contained only one flaw. Even though that flaw was detected (and therefore $\hat{p}=1$ for the interval), the lower 95% confidence bound on p is 5%. The same occurred at flaw sizes of 10-11mm.

To obtain narrower intervals with a relatively large number of flaws in each interval, and therefore smoother confidence limits, the intervals can be defined to overlap. For instance, a method can be devised so that each interval contains the same number of flaws (clearly, many flaws will end up belonging to more than one interval [13]. Other more sophisticated methods for assigning flaws to intervals have been proposed, but the conclusion is that they all suffer important deficiencies. The POD curves obtained can show a very erratic behaviour. Since they are based on the binomial distribution, the confidence bounds are greatly influenced by the method used for assigning flaws to the intervals. In other words, the confidence bounds are as much influenced by the analysis method as they are by the data. Finally, a very high number of flaws in test pieces are required to obtain a whole POD curve, because the required sample sizes discussed in section 2.1 (e.g. inspecting and finding 29 out of 29 flaws to guarantee a POD higher than 0.9 with 95% confidence) must be here multiplied by the number of points needed to define a reasonable POD curve covering all flaw sizes of interest.

2.3 Regression analysis: estimation of the POD curve with multiple observations per flaw

This type of analysis was originally developed to model the results of a large NDE reliability program performed by the US Air Force [14]. In this program, known as "Have-Cracks-Will-Travel" program, a substantial amount of data became available. In particular, sections of retired aircraft were transported to Air Force depots and inspected by representative personnel, using a variety of inspection methods. At the end of this

phase, the components and specimen were examined destructively to determine the real flaw sizes. For this reason, single flaws were inspected many times and it was possible to determine probabilities of detection for individual flaws.

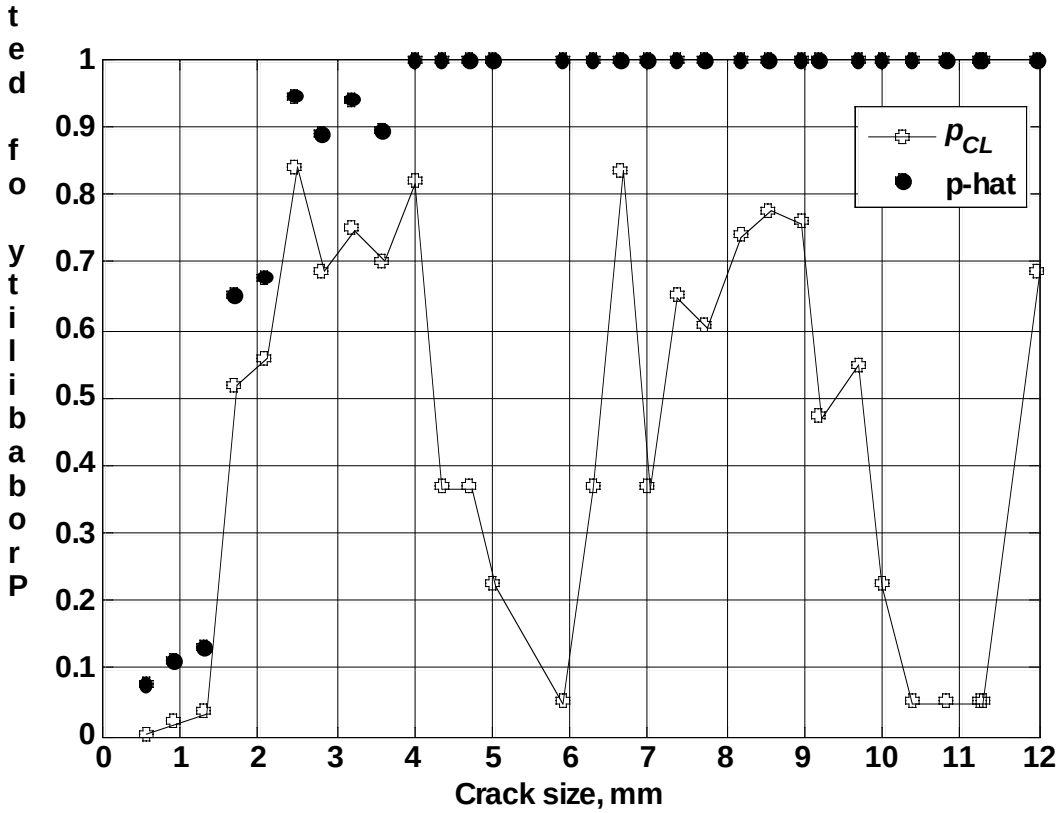

Figure 2. Probability of detection of etched fatigue cracks in aluminium flat plates (eddy current inspections), data from [13], Range Interval Method.

References [1] [2] discuss an example of such data, presented in [14] and reproduced here in Figure 3. The points in the plot represent the detection probabilities of a set of 41 flaws inspected by 60 different inspectors with the eddy current method. The flaws were located around fastener holes in a segment of a C-130 center wing box. Each data point in the figure represents the proportion of times that a flaw was found when subjected to these 60 independent inspections.

To analyse the data collected from this category of experiments, a regression analysis can be performed in which a model curve is fit to the data points. A lower confidence limit is then placed on the regression equation. For example, in reference [14] the following model ("Lockheed" model) was selected as providing the best fit:

$$\mathrm{POD}(a) = \exp(-\alpha \cdot a^{(1-\beta)}) \tag{4}$$

Where the parameters α and β are estimated by a linear regression of transformations of the flaw sizes and observed probabilities of detection. In other words, the flaw sizes and observed probabilities of detection are transformed according to the following equations:

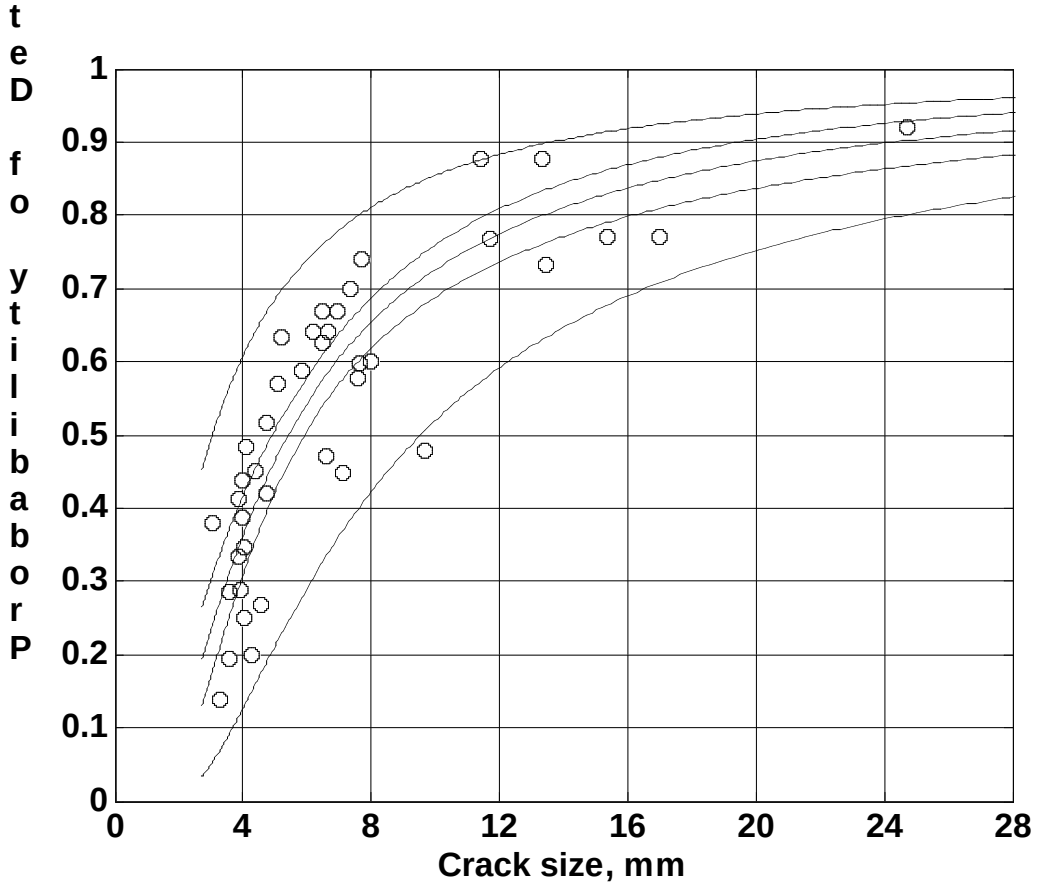

Figure 3. Flaw detection probabilities from 60 eddy current inspections of each flaw (reproduced from [14]). The solid line represents the mean of the regression. The dotted lines represent the 95% upper and lower confidence bounds on the data. The dashed lines represent the 95% upper and lower confidence bounds on the mean trend.

$$X = -\log(a)$$
$$Y = \log(\frac{-\log(p)}{a}) \quad (5)$$

Many other POD models (i.e. transformation equations) may be used. In a regression analysis, a linear relationship is then postulated to exist between the predictor variable X and the measured variable Y.

$$Y = A + B \cdot X + e \quad (6)$$

where e is the term (a random variable) that accounts for the deviations from the regression equation.

The curve obtained by fitting the model of (Eq. 4) to the data is plotted in Figure 3 as a solid line. This is the mean of the regression. In [14] the view was taken that the lower confidence limit on the POD curve at a particular flaw size is some (for instance 95%) low percentile of the distribution of detection probabilities at that flaw size. To calculate such lower confidence limit, a confidence bound on the population of detection probabilities is used. The details of this calculation are not discussed here, for lack of

space. Both lower and upper 95% confidence limits are represented in Figure 3 (dotted lines), and it is important to note that these two curves bound the near totality of the data points. Indeed, this confidence bound is a bound on the whole set of flaw detection data, and at a 95% confidence level one would expect only 5% of data points to fall outside it.

The view taken by Berens and Hovey in [2] is that the POD should actually be seen as the mean of the detection probability distribution. They argued that the regression approach described here (and used in [14]) has no rationale other than goodness of fit. We will describe this different approach in the next section.

2.4 Parametric models

The realisation that different flaws of virtually the same size can have significantly different detection probabilities led Berens and Hovey [2] [3] to use a different statistical framework to model the POD curve.

The main idea is to assume that there is a distribution of detection probabilities at each flaw size. The scatter of this distribution is caused by the non reproducibility of all factors other than flaw size (flaw orientation, geometry, location, operator, environment of inspection, etc.). Figure 4 shows a schematic representation of this distribution of detection probabilities. Let $f_a(p)$ be the density function of the probability of detection at flaw size a. The subscript a is meant to remind that the density function is defined at each flaw size a, i.e. this is not necessarily the same function as we move from a to a different size a_1.

Let us focus exclusively on all the flaws of exactly size a. The probability of randomly selecting a flaw in such set whose detection probability is between p and $p+dp$ is $f_a(p)\cdot dp$. The detection probability of such a flaw is, of course, p.

The conditional probability that the detection probability is p and that the flaw will be detected is thus $p\cdot f_a(p)\cdot dp$. The unconditional probability that a randomly selected flaw of size a will be detected, $POD(a)$, is obtained by summing over the conditional probabilities over the whole range of detection probabilities. Thus:

$$POD(a) = \int_0^1 p \cdot f_a(p) dp \qquad (7)$$

The meaning of (Eq. 7) is that the true POD function is the curve through the means of the individual density functions of detection probabilities. Such a curve is also the traditional regression equation of POD as a function of flaw size. Therefore, regression analysis techniques can be used to estimate the POD function when individual estimates of the detection probabilities are available. However, confidence limits on the true POD would be calculated from the confidence limits for the average of the prediction and not the individual detection probability estimates.

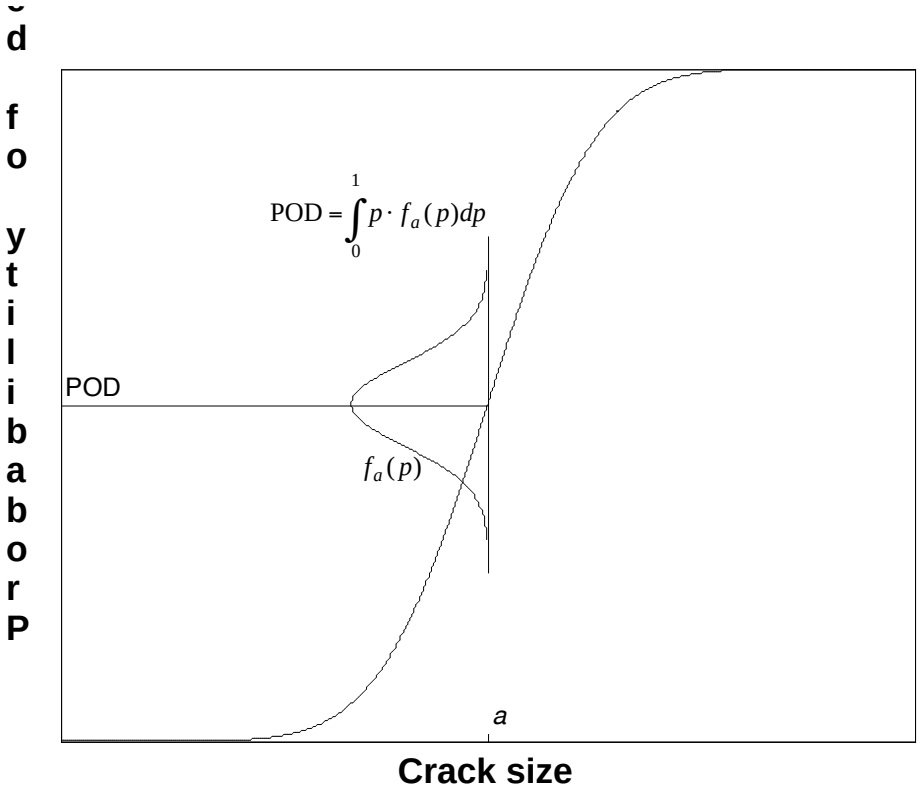

Figure 4. Schematic representation of distribution of detection probabilities for flaws of fixed size.

In general, NDE systems are such that the response signal, $\hat{a}$, increases with flaw size. This is a very reasonable assumption to make, well grounded in the underlying physical processes of the majority of NDE systems, but it is not always true. For instance, in a time-of-flight diffraction inspection, a longer embedded flaw may have its tip closer to the surface as compared to a shorter flaw, and therefore it may be more difficult to detect.

The statistical model proposed by Berens and Hovey [3] is based on the idea of grouping the sources of uncertainties affecting the response $\hat{a}$ into two distinct groups:

1. the variability in the mean $\hat{a}$ from flaw to flaw;
2. the variability in $\hat{a}$ from inspection to inspection of the same flaw.

The material properties, the flaw location, geometry, orientation, etc. are strictly associated with individual flaws, and do not change from inspection to inspection. All these causes of uncertainties affect the first type of source of variation. Human and equipment factors usually vary from inspection to inspection even of the same flaw, and therefore affect the second type of source of variation.

Because of this dual nature of the sources of variation, the response signal has a compound distribution. This can be seen in the steps of the inspection process. First, a flaw is randomly chosen along with its individual mean $\hat{a}$. Secondly, the human and equipment factors come into play resulting in a random deviation from the flaw mean

for an individual inspection. These are two distinct random processes with distinct random variables.

Berens and Hovey proposed a statistical model called a variance component model [3], according to which the response signal is broken down into components that can be attributed to specific sources of variation.

$$\hat{a} = h(a) + \delta + \varepsilon \tag{8}$$

Where *h(a)* represents the overall mean trend in $\hat{a}$ as a function of *a*, δ represents the flaw to flaw variation, and ε represent the variation from inspection to inspection of the same flaw. *h(a)* is fixed, whereas δ and ε are random variables with zero mean.

The model presented in (Eq. 8) is completely generic, and in this form alone does not allow drawing any quantitative conclusion. The important part comes when one has to start making assumption about the functional form of *h(a)* and the type of distributions of δ and ε . Note that (Eq. 8) describes the behaviour of $\hat{a}$, but this is not necessarily the POD function yet. An equivalent form of (Eq. 8) can be written to model the uncertainties directly on the POD function itself.

$$POD(a) = g(a) + \delta + \varepsilon \tag{9}$$

Clearly δ and ε in (Eq. 8) have a slightly different meaning than the same terms in (Eq. 9). For instance ε in (Eq. 8) represents the variation in $\hat{a}$ from inspection to inspection of the same flaw, whereas in (Eq. 9) it represents the variation in POD. For simplicity, we have chosen to avoid introducing new nomenclature. Notably, (Eq. 8) or (Eq. 9) point out that when more than one inspection is made of an individual flaw, the results obtained are not totally independent. Correlations between inspections of the same flaw occur because δ is the same. Specialised methods are required to analyse NDE reliability data that include multiple inspections of individual flaws.

We calculated confidence bounds for the same data of Figure 3 (from reference [14], flaws in fastener holes of a C-130 center wing box.), using the same "Lockheed" model presented in (Eq. 4) to model the shape of the POD curve, *g(a)*. The mean POD curve obtained is exactly the same as the one using the regression analysis (solid line of Figure 3). The new confidence bounds are plotted in Figure 3 with two dashed lines, and are clearly much narrower than in the previous case (dotted lines). Their meaning is the following: if we were to repeat the whole experiment many times over and we were to plot the mean trend (solid line) each time, we would expect it to fall inside the band 95% of the time. In this approach, the POD confidence limit is placed on the mean regression line and not on the total population of flaws.

3. Selection of a POD(a) model

The model presented by Berens and Hovey, expressed for instance in (Eq. 9), is completely generic. At this stage, no assumption has been made on the shape of the function *g(a)* or on the statistical properties of the random quantities δ and ε.

Choosing a model for the POD function means choosing a functional form for *g(a)*, which is usually dependent on a small number of parameters. Further, assumption must be made on the nature of δ and ε so that confidence bounds on the estimates can be derived. The parameters that determine the exact shape of *g(a)*, (and hence of the POD curve) are determined so that they provide the best fit (in some statistical sense, such as the maximum likelihood principle) to the available experimental data.

Berens and Hovey [2] discussed three criteria for the definition of an acceptable POD model: (1) goodness of fit, (2) normality of deviation from fit, and (3) equality of variance of deviations from fit at all flaw sizes. The latter two criteria are necessary statistical assumptions for the validity of conventional confidence limits derived from regression analyses. It is very important to point out that these criteria are assumptions, and there is no guarantee that they hold true in reality. They are indeed convenient assumptions, because the ability to derive confidence bounds is extremely desirable. As discussed above, goodness of fit is an intuitively desirable condition, but there is no guarantee that a model that fits the data better than another is necessarily better at capturing and describing the true underlying physical phenomenon.

Normality of deviation from fit and equality of variance of deviations from fit at all flaw sizes are also assumptions that should be made in full awareness. There is no guarantee that in reality the spread in flaw detection probabilities at two different flaw sizes (say one very small and one very large) are the same, yet this assumption is always made (strictly speaking, in most logistic regression models it is the transformed variable Y which is assumed to have constant variance [15]).

Berens and Hovey [2] investigated seven potential functional forms for the POD curve *g(a)*. Regression analysis was used to fit all seven models to the "Have Cracks Will Fly" data, which comprised 13 data sets (each data set defined in terms of an NDE method and a type of structure). The detection probabilities, p_i, and the flaw sizes, a_i, for each flaw i were used to obtained transformed variables X_i and Y_i. These transformed variables were then used in a linear regression analysis of the form already presented in (Eq. 6). The so-called "log-logistic" (or "log-odds") model is obtained by choosing the following transformation:

$$\begin{aligned} X &= \log(a) \\ Y &= \log\left(\frac{p}{1-p}\right) \end{aligned} \tag{10}$$

Which gives rise to the following *POD(a)* function, dependent on two parameters, α and β.

$$POD(a) = \frac{e^{\alpha+\beta\cdot\log(a)}}{1+e^{\alpha+\beta\cdot\log(a)}} \tag{11}$$

Ref. 2 concluded that the log-logistic model was the best (among the seven investigated) to fit the available data, mainly because of goodness of fit of the mean trend and because of the structure of the deviations from the mean. Furthermore, also based on the analysis of an additional eight data sets from Yee et al. [13], reference [2] concluded that

the log-logistic model had the additional characteristic of providing the lowest POD estimate at the longer flaw sizes. This was considered a positive feature, because it entails providing a conservative estimate of the POD at longer flaw sizes which can have the greatest impact from a structural integrity point of view. Finally, a very desirable feature was identified in its analytical simplicity.

This conclusion has been relied upon by many authors in later works. Most of the published NDE reliability literature makes use of the model presented in (Eq. 9), and of the log-logistic transformation expressed by (Eq. 10) and (Eq. 11). In virtually all cases, reference [3] is quoted as having established that the log-logistic model is the best to fit NDE reliability data. This is not exactly what Berens and Hovey stated, who very explicitly made the point that "*however, the evidence is still limited to this study*" [3]. Very interestingly, Berens and Hovey also stated that "*one of the most controversial aspects on NDI reliability estimation is the selection of a model for the POD function*" [3]. We do not intend to criticise the use of the log-logistic transformation (or, more generally, of the choice of (Eq. 9) as POD model) in other works, but we argue that, when adopting it, the NDE reliability practitioner should be very well aware of the limitations implicit in the choice.

4. Discussion

We start the discussion by arguing that a single correct statistical model to treat the data simply does not exist. A model can be more sophisticated than another, but this does not necessarily imply that its results will be better or more credible. Choosing a model over another often implies a philosophical choice that reflects the way the user is interpreting the underlying physical nature of the phenomenon being analysed. For instance, we have seen in section 2.1 how the probability of detection at one flaw size can be seen as the frequency of detection of many randomly selected flaws from the population of all flaws of that given size. We have also seen that this binomial model has more or less been completely abandoned, in favour of the parametric model described in section 2.4 (notable exceptions are ASME V Article 14 and ASME XI Appendix VIII). This must not be taken to imply that the former is not adequate. Very often, a statistical model is chosen because it provides results with much less data, or because it is particularly easy to treat analytically or to implement numerically, or because it provides confidence bounds in a straightforward way. The parametric model has very desirable qualities. Most notably it provides a smooth POD curve over the entire range of flaw sizes whilst requiring less experimental data. This does not automatically guarantee that it is the "best" model to use.

The use of an assumed functional form for the POD mean trend *g(a)* greatly simplifies the task of estimating NDE reliability, but comes with a hidden price, as discussed below. The approaches described in sections 2.1, 2.2 and 2.3 did not assume any underlying model for the POD function. The binomial methods described in sections 2.1 and 2.2 grouped the flaws into size intervals, and confidence bounds were derived for each interval. The approach described in section 2.3 used a regression analysis simply to fit the data, without assuming an underlying model such as the one described by (Eq. 8).

The great advantages of using the model of (Eq. 8) are twofold. First of all, substantially fewer data points are required to obtain a POD curve. For instance, [2] states that 60 points (covering the whole range of flaw sizes) should be enough to obtain a reliable POD curve. This is much less than the 29 flaws needed to obtain just one meaningful point (a p_{CL} greater than 0.9 at a 95% confidence level) on the POD curve in the binomial analysis. Secondly, the smoothness of the POD curve and its confidence bound is supplied by the model rather than by artificial interpolation techniques.

We argue that such reduction in the amount of data needed does not come for free. It is not achieved by the greater sophistication of the model, but simply by the extra "information" which is injected by choosing a POD function over another one. The model selected is often a purely arbitrary choice, a justification for which cannot be usually found in the available data. For this reason, it is essential that the choice of a POD function is sound and well grounded. The model chosen should be physically meaningful. Further, results at one flaw size will influence the POD curve at other flaw sizes. If flaws in the data set are over-represented for a particular sub-interval of the flaw size range, these will unduly affect the probabilities of detection at flaw sizes outside this interval. For instance, if many data points are available at very small flaw sizes and practically no data exist for larger flaws, the behaviour of the POD curve at larger flaw sizes will not be reliable.

The confidence bounds obtained may not be realistic if the assumptions made on δ and ε are not verified. Normality of a random variable is an assumption that is frequently made. It just seems right that the variation around the mean value is well described by a bell-shaped curve, but this is not necessarily so. It is quite easy for instance to devise physically meaningful situations in which the crack-to-crack variability is not normally distributed. For cracks whose mean probability of detection is very close to 0 or to 1, the crack-to-crack variability cannot be normally distributed, as the POD has a lower (0) or upper (1) limit. Indeed, so-called ‘logistic’ regression models usually assume that the POD follows a binomial distribution [15] and/or that a transformed variable (such as Y in Eq. 10) follows a standard unbounded distribution (typically logistic, normal or Gompertz) [16] [17].

We also want to stress the importance of correctly understanding the nature of confidence bounds. In Figure 3, both confidence bounds are entirely correct, but it is essential that the interested user understand the difference. The exact choice of the confidence bounds to be used depends precisely on the intended use. The correct curve to use when estimating risk reduction following an inspection is the mean trend. The spread around the mean value will not affect the risk reduction. Therefore, the correct confidence bounds to derive on the estimated POD curve would be those on the mean trend, and a conservative choice for the POD curve would be the lower dashed line of Figure 3. If assurance is required that individual flaws will be detected with at least a given detection probability, the correct conservative choice for the POD curve would be the lower dotted line of Figure 3. These two curves are considerably different.

5. Conclusions

The use of probability of detection curves to quantify NDT reliability is common in the aeronautical industry, but relatively less so in the nuclear industry, at least in European countries. The ENIQ inspection qualification methodology, which is widely used in European countries, is based on the idea of the Technical Justification and thus nearly always provides qualitative statements on the capability of the NDE system to find flaws. The need to quantify the output of inspection qualification is becoming increasingly important, as utilities move to risk-informed in-service inspection programmes, to quantify (and thus obtain real benefit from) the risk reduction achieved after carrying out an inspection.

Probability of detection (POD) curves provide the required metric. The statistical models used to derive POD curves from experimental data are straightforward, but require that the user is very well aware of the assumptions made. In this paper we have reviewed such models and have attempted to highlight some of the potential problems that can arise if the main underlying assumptions are not verified. Further, we have attempted to clarify the confusion that often arise over the nature of the POD curve and associated confidence bounds.

References

1. A.P. Berens, NDE Reliability Data Analysis, in Non-destructive Evaluation and Quality Control: Qualitative Non-destructive Evaluation, ASM Metals Data Book, Volume 17, ASM International (1989) 689.
2. Berens A P and Hovey P W: ‘Evaluation of NDE reliability characterisation’, AFWAL-TR-81-4160, Vol. 1, Air Force Wright-Aeronautical Laboratories, Wright-Patterson Air Force Base, December 1981.
3. Berens, A. P. and Hovey, P. W., Flaw Detection Reliability Criteria. Volume 1. Methods and Results, AD-A142 001, DAYTON UNIV OH RESEARCH INST, Final technical report., April 1984.
4. Department of Defense Handbook: Nondestructive Evaluation System Reliability Assessment, MIL-HDBK-1823A, 7 April 2009. Available for download at http://statistical-engineering.com/mh1823/MIL-HDBK-1823A%20(2009).pdf.
5. European methodology for qualification of non-destructive testing: third issue. EUR 22906 EN, 2007.
6. The nuclear regulators working group—task force on risk-informed in-service inspection, report on the regulatory experience of risk-informed in-service inspection of nuclear power plant components and common views, EUR21320 EN; August 2004.
7. European framework document for risk-informed in-service inspection. ENIQ Report nr. 23, EUR 21581 EN; 2005.
8. ENIQ recommended practice 3: strategy document for technical justification. ENIQ Report No. 5, JRC-Petten, EUR 18100/EN; 1998.
9. L Gandossi & K Simola, Framework for the quantitative modelling of the European methodology for qualification of non-destructive testing, International Journal of Pressure Vessels and Piping, Volume 82, Issue 11 , November 2005, Pages 814-824.

10. L Gandossi & K Simola, A Bayesian Framework for the Quantitative Modelling of the ENIQ Methodology for Qualification of Non-Destructive Testing, JRC Technical report, EUR 22675 EN, 2007.
11. Damage Tolerance Design Handbook, available online at http://www.afgrow.net/applications/DTDHandbook/
12. G. Casella, R. Berger, Statistical Inference, Duxbury Press, 1990.
13. Yee, B. G. W.; Chang, F. H.; Covchman, J. C.; Lemon, G. H.; Packman, P. F., Assessment of NDE reliability data, NASA report NASA-CR-134991, Oct 1, 1976.
14. W.H. Lewis, B.D. Dodd, W.H. Sproat, and J.M. Hamilton (1978). Reliability of Nondestructive Inspections – Final Report. Report No. SA-ALC/MEE 76-6-38-1. United States Air Force, San Antonio Air Logistics Center, Kelly Air Force Base, Texas.
15. D W Hosmer and S Lemeshow, 'Applied logistic regression', John Wiley & Sons, New York, 1989.
16. P McCullagh and J A Nelder (1992), 'Generalised linear models'. Chapman & Hall.
17. Minitab Inc. (2003). MINITAB Statistical Software, Release 14 for Windows, State College, Pennsylvania.