Pacific Northwest
NATIONAL LABORATORY

PNNL-32908

Guidance for Performing Probability of Detection Analysis for Nuclear Power Component Inspections

May 2022

Ryan M. Meyer
Aimee E. Holmes

Prepared for the U.S. Nuclear Regulatory Commission
Office of Nuclear Regulatory Research
Under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

PNNL-32908

Guidance for Performing Probability of Detection Analysis for Nuclear Power Component Inspections

May 2022

Ryan M. Meyer
Aimee E. Holmes

Prepared for the U.S. Nuclear Regulatory Commission
Office of Nuclear Regulatory Research
Under Contract DE-AC05-76RL01830
Interagency Agreement: NRC-HQ-60-17-D-0010

Carol A. Nove, NRC Contracting Officer Representative

Pacific Northwest National Laboratory
Richland, Washington 99354

Summary

This document provides guidance for probability of detection (POD) analysis specific to nuclear power plant (NPP) component applications. The document was prepared as part of an ongoing international research collaboration for topics related to nondestructive examination (NDE) in commercial NPPs called the Program for Investigation of NDE by International Collaboration (PIONIC). PIONIC was established by the U.S. Nuclear Regulatory Commission (NRC) to facilitate information sharing and leveraging of resources toward research topics of mutual interest. PIONIC follows preceding NRC sponsored international collaborations, including the Program to Assess the Reliability of Emerging Nondestructive Technologies (PARENT) (Meyer and Heasler 2017) and the Program for Inspection of Nickel Alloy Components (PINC) (Cumblidge et al. 2010). PINC and PARENT both focused on the generation of empirical POD data for dissimilar weld components in light water reactors. The experiences from PINC and PARENT motivated the development of the guidance provided in this document.

Conducting empirical POD studies requires the development of test blocks with material, dimensional, and geometrical relevancy to the intended components of study. For NPP applications, acquiring such test blocks is cost intensive relative to other industries. As a result, the studies under PINC and PARENT were constrained by the availability of such test blocks. This led to limitations on the sets of data that could be created for POD estimations. To overcome dataset challenges, non-standard approaches or techniques were implemented in the analyses of POD data.

This report does not attempt to be a sole source of guidance for POD analysis of NPP components but is meant to provide guidance for approaching POD analysis under PINC and PARENT that are not addressed or endorsed by existing guidance. Therefore, the guidance in this report is meant to supplement existing guidance, specifically, MIL-HDBK-1823A – *Nondestructive Evaluation System Reliability Assessment* (MIL-HDBK-1823A 2009) and ASTM E2862-18 – *Standard Practice for Probability of Detection Analysis for Hit/Miss Data* (ASTM E2862-818 2018).

Guidance is provided related to the following approaches to POD analysis that are either not covered or endorsed by standards MIL-HDBK-1823A and ASTM E2862-18:

- Using false call data in the logistic regression fitting of the continuous POD curve
- Using artificial data points (termed “pseudopoints”) in the logistic regression fitting of the continuous POD curve
- Using virtual flaw data for POD estimation
  - In this context, virtual flaw data refers to simulated flaw responses that are created from digital alterations made to responses acquired from physical flaws. This technique was not employed in PINC or PARENT, but an exercise was conducted under PIONIC to demonstrate the concept.
- Representing POD for aggregated data
  - The most appropriate representation of POD depends on the intended end-use. It has been the convention to represent aggregated POD as the average of the population. However, aggregate POD could be represented by other quantiles of the population to emphasize “worst-case” or “best-case” performance. It has also been the convention to calculate 95% confidence bounds for the mean POD. However, the confidence bounds

only convey information about uncertainty in the estimated mean POD and do not convey information about distribution of performance.

A general conclusion of this effort is that estimating POD for NPP components is complex and that there is not a “one size fits all” approach. The best approach depends on the characteristics of the available dataset and intended end-use of the POD data. If multiple end-uses are anticipated, it may be desirable to tailor the approach and representation of POD for each end-use.

Acronyms and Abbreviations

DMW	Dissimilar metal weld
DOE	Department of Energy
ECT	Eddy current testing
ENIQ	European Network for Inspection and Qualification
EPRI	Electric Power Research Institute
FCP	False call probability
ID	Inner diameter
IGSCC	Intergranular stress corrosion cracking
ISI	Inservice inspection
LBDMW	Large bore dissimilar metal welds
MLE	Maximum likelihood estimation
NDE	Nondestructive examination
NPP	Nuclear power plant
OD	Outer diameter
PARENT	Program to Assess the Reliability of Emerging Nondestructive Techniques
PAUT	Phased-array ultrasonic testing
PDI	Performance Demonstration Initiative
PINC	Program for Inspection of Nickel Alloy Components
PIONIC	Program for Investigation of NDE by International Collaboration
PNNL	Pacific Northwest National Laboratory
POD	Probability of detection
PWSCC	Primary water stress corrosion cracking
SBDMW	Small bore dissimilar metal weld
SF	Scaling factor
TOFD	Time-of-flight diffraction
TWD	Through-wall depth
VRR	Virtual round robin

Contents

Figures

Tables

1.0 Introduction

The purpose of this report is to provide guidance specific to probability of detection (POD) analysis for quantifying performance estimates of nondestructive examination (NDE) for nuclear power plant (NPP) components. This report is motivated by results and experiences from the Program for Inspection of Nickel Alloy Components (PINC) (Cumblidge et al. 2010), the Program to Assess the Reliability of Emerging Nondestructive Techniques (PARENT) (Meyer and Heasler 2017), and the documentation of POD models for use in the probabilistic fracture mechanics code, eXtremely Low Probability of Rupture (xLPR) (Meyer and Holmes 2019). It has multiple objectives, including:

- Highlighting limitations of POD models provided in Meyer and Holmes (2019) for use in probabilistic fracture mechanics codes (i.e., xLPR)
- Elucidating limitations of past POD studies to inform future POD estimation efforts
- Facilitating comparisons and benchmarking with results from PARENT (Meyer and Heasler 2017) by clarifying the methods of analysis
- Documenting the outcome of a virtual round robin (VRR) activity that implemented a novel virtual flaw methodology in the attempt to overcome challenges with empirical POD estimation for NPP components
- Summarizing guidance and recommendations for performing POD analysis for NPP components based on the review of PARENT data analysis, existing standards, and experience from the VRR activity.

The guidance provided in this report was developed under the auspices of the Program for Investigation of NDE by International Collaboration (PIONIC). PIONIC was established to facilitate collaborative information sharing and the leveraging of resources by organizations in government, industry, and academia with technical expertise in the NDE of commercial nuclear power facilities. PIONIC follows the examples of its predecessor programs PINC (Cumblidge et al. 2010) and PARENT (Meyer and Heasler 2017), which focused on the collection of empirical NDE reliability data. In these programs, participants provide resources in the form of in-kind contributions. The participants in PIONIC are currently associated with the following countries: the United States of America, Finland, Japan, Sweden, South Korea, and Switzerland.

1.1 POD Representation

POD is the standard metric for quantifying the performance of NDE. Although POD can be represented as a function of multiple independent variables, it is customary to represent POD as a function of a flaw depth for most NPP applications. The context of a flaw's significance can be more easily understood by normalizing the flaw depth, d, with the component wall thickness, T. Thus, flaw depth is represented as "through-wall-depth" (TWD). Scaled in this way, TWD = d/T and can range from TWD = 0 to 1.

POD is represented as a monotonic curve versus TWD that can range from 0 for flaws below the detection threshold and up to 1 (i.e., 100% detection) for larger flaws. An ideal POD relationship is illustrated in Figure 1-1 as a step function with perfect (POD = 1) detection of all flaws with TWD greater than a threshold flaw depth and no detection (POD = 0) for flaws with TWD less than a threshold flaw depth. In practice, the transition from low to high POD occurs over a finite interval of TWD and has the appearance of an S-like shape, as illustrated in Figure

1-1. Less than ideal NDE performance can be indicated by deviations from the ideal POD relationship. Such deviations may include 1) a transition that is more gradual, 2) a maximum POD < 1 for large TWD flaws, or 3) a minimum POD > 0 for small TWD flaws.

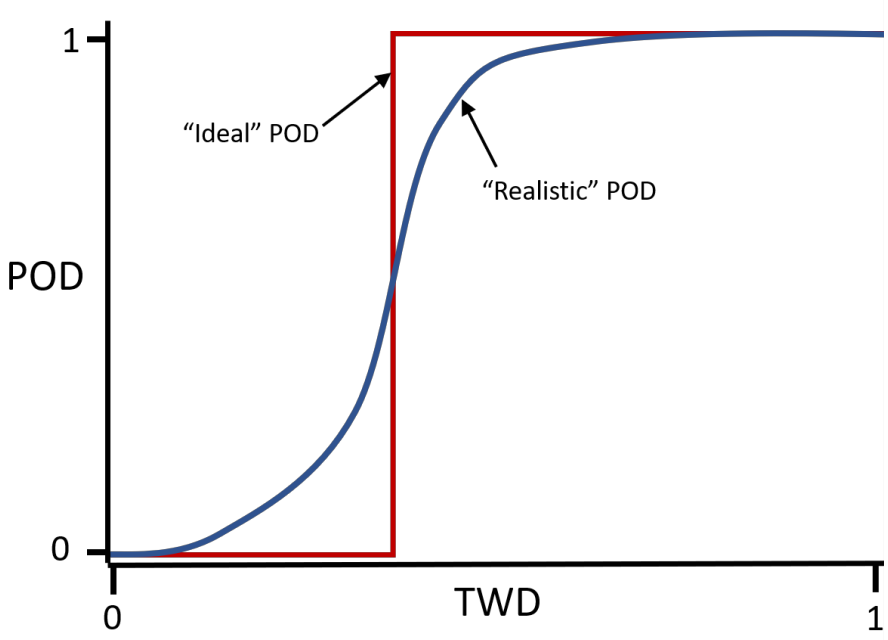

Figure 1-1 Illustration of a POD curve

Mathematical models representing the general relationship of the S-like curve in Figure 1-1 are fit to empirically derived data to define an expression of POD that is continuous with respect to TWD. For empirical studies that generate binary NDE responses (i.e., hit/miss data; response data represented as hit = 1, miss = 0), the data can be fit with the logistic function (Berens 1989),

$$\mathrm{POD}(\mathrm{TWD}) = \frac{1}{1 + exp(-\beta_1 - \beta_2\ \mathrm{TWD})} = \frac{exp(\beta_1 + \beta_2\ \mathrm{TWD})}{1 + exp(\beta_1 + \beta_2\ \mathrm{TWD})} \qquad \text{Equation 1.1}$$

This model includes two parameters, β_1, and β_2, to be determined from curve fitting with empirical data using maximum likelihood estimation (MLE) (Forsyth and Fahr 1998). The β_1 and β_2 parameters in Equation 3.5 are dimensionless because TWD is dimensionless. This model was used to create POD curves from data collected in PINC (Cumblidge et al. 2010) and PARENT (Meyer et al. 2017; Meyer and Heasler 2017) and was also the primary model used to fit data from the industry's Performance Demonstration Initiative (PDI) as described in MRP-262 Rev. 3 (EPRI 2017). In PARENT, the MLE procedure for logistic regression was implemented in the statistical software package known as "R" (R Core Team 2021).

The typical approach to obtaining a POD curve for an application is to create test blocks to simulate relevant components and their potential flaws and then perform examinations of the test blocks in a laboratory setting. The collected data is then analyzed by fitting data using logistic regression as described above to obtain an estimate of POD as a function of TWD.

Standards have been developed to cover various aspects of POD studies, including MIL-HDBK-1823A – *Nondestructive Evaluation System Reliability Assessment* (MIL-HDBK-1823A 2009) and ASTM E2862-18 – *Standard Practice for Probability of Detection Analysis for Hit/Miss Data* (ASTM E2862-818 2018). These standards provide guidance for performing a POD study and performing POD analysis, including recommendations on the number and size distribution of flaws in test blocks.

1.2 Sources of Empirical POD Data for NPP Components

The data collected by PDI is the most extensive source of empirical data for estimating POD for NPP components. A description of the analysis performed to develop POD estimates from the PDI database is provided in report MRP-262 Rev. 3 (EPRI 2017), which provides these estimates for several component types. The component types are categorized and include pressurizer surge and hot-leg surge components (referred to as Category A) and reactor pressure vessel inlet and outlet nozzles (referred to as Category B1).

Empirical POD data has also been generated for dissimilar metal weld (DMW) components as part of the U.S. Nuclear Regulatory Commission (NRC)-supported round robin studies—PINC (Cumblidge et al. 2010) and PARENT (Meyer and Heasler 2017). In PINC and PARENT, studies were performed on test blocks categorized as small bore and large bore DMWs (SBDMW and LBDMW, respectively). The diameter and wall thickness of the test blocks used for MRP-262 Rev. 3, PINC, and PARENT are summarized in Table 1-1. This table also indicates if outer diameter (OD) or inner diameter (ID) examinations were performed. Throughout this report, results obtained from Category A components in MRP-262 Rev. 3 are compared to results from SBDMWs in PINC and PARENT, and results from Category B1 components in MRP-262 Rev. 3 are compared to results from LBDMWs in PINC and PARENT. The MRP-262 Rev. 3 results were generated from only ultrasonic testing data. While PINC and PARENT results for OD examinations were also derived from datasets that only included results from ultrasonic testing, many of the ID examination results were based on datasets that included results of eddy current testing.

Table 1-1 Summary of general test block dimensions for data collected as part of MRP-262 Rev. 3, PINC, and PARENT

	MRP-262 Rev. 3		PINC	PARENT	
	Pressurizer Surge (Category A)	Reactor Pressure Vessel (Category B1)	SBDMW	SBDMW	LBDMW
Outer Diameter (mm)	305–356	686–787	386–390	289 and 815	852–895
Wall Thickness (mm)	30–58	64–76	42–46	35 and 39.5	68–78
Exam Surface	OD	ID	OD and ID	OD	OD and ID

Only the OD surface can be accessed for most piping weld joints in the primary coolant circuits of NPPs. However, the welds joining primary coolant piping to the reactor vessel safe ends may be accessible from the ID surface by entering the piping from inside the reactor vessel. Plant and site-specific factors are considered when determining if these welds will be examined from

the OD or ID surface. Access to the ID or OD surface may be obstructed by other components or structures. For instance, access to the ID surface may be obstructed by the core barrel. Even if the ID surface is accessible, special tooling is required to facilitate performing the examination remotely. Although examinations from the OD surface may not require special tooling, potential radiation dosage to personnel is a factor that can complicate OD examinations.

The number of data points (i.e., recorded hits or misses) fit to the logistic function model of POD for MRP-262 Rev. 3, PINC, and PARENT are summarized in Table 1-2 for circumferential and axial flaws. The number of data points is referred to as the "NOBS," which stands for number of observations in Table 1-2. In MRP-262 Rev. 3, this is referred to as the number of "Attempts." Regardless of the nomenclature, this number includes multiple examination attempts of individual flaws. For instance, in PARENT, if ten teams each examined the same 10 flaws, NOBS would equal 100. The source data used for POD analysis in MRP-262 Rev. 3 were approximately one order of magnitude greater than the source data in PINC or PARENT, based on NOBS.

Table 1-2 Summary of flaw populations and depths in MRP-262 Rev. 3, PINC, and PARENT.

	MRP-262 Rev. 3		PINC	PARENT	
	Pressurizer Surge (Category A)	Reactor Pressure Vessel (Category B1)	SBDMW	SBDMW	LBDMW
NOBS – circumferential flaws	1675[1]	553[1]	150	183	50
NOBS – axial flaws	611[1]	288[1]	100	45	45
Circumferential flaw depth range (TWD)	10%–100%[2]	10%–100%[2]	10%–83%	3%–72%	1%–36%
Axial flaw depth range (TWD)	10%–100%[2]	10%–100%[2]	11%–71%	11%–74%	1%–36%

POD studies for NPP components face limitations on the number of test blocks that can be supplied because acquiring test blocks that simulate relevant features of an NPP component can be expensive relative to other industries. Additionally, the potential degradation mechanisms of concern, such as primary water stress corrosion cracking (PWSCC) or intergranular stress corrosion cracking (IGSCC), result in cracks with complex morphologies expressing NDE-relevant features that aren't simulated with saw cuts, electro-discharge machine notches, or mechanical fatigue cracks. As a result, more sophisticated manufacturing processes are needed to produce simulated flaws. Consequently, POD studies performed under

[1] Table 6-1 in MRP 262 Rev. 3 (EPRI 2017).

[2] Flaw size distribution requirements specified in ASME Boiler and Pressure Vessel Code Section XI, Appendix VIII, Supplement 10.

PINC and PARENT do not meet the recommendations in MIL-HDBK-1823A for the number of unique simulated flaws (i.e., minimum of 60 unique flaws). Approximately 10–25 unique circumferential flaws and ≤ 10 unique axial flaws were incorporated in PINC and PARENT.

Flaw depth ranges are also listed in Table 1-2 for MRP-262 Rev. 3, PINC and PARENT. Although PINC and PARENT incorporated smaller flaw depths in comparison to MRP-262 Rev. 3, the number of flaws with TWD < 10% was limited. The limitations of empirical data sources are evident when comparing the results of POD analysis from MRP-262 Rev. 3, PINC, and PARENT (Meyer and Holmes 2019). Meyer and Holmes (2019) compiled POD models for examinations performed on LBDMW and SBDMW components for OD and ID examinations provided as input to the inservice inspection module of xLPR. Generally, the models for MRP-262 Rev. 3, PINC, and PARENT do not converge to singular POD representations for components of similar size and type, for similar flaw orientations, and for ID or OD examinations.

1.3 Pseudopoints in PARENT Data Analysis

Pseudopoints refer to artificial data points added to POD datasets to assist with some aspect of analysis. In binary hit/miss datasets, detections are assigned a value of 1 at associated flaw depths and misses are assigned a value of 0 at associated flaw depths. Pseudopoints may be assigned any value from 0 to 1 and assigned to any flaw depth over $0 \leq \text{TWD} \geq 1$. Pseudopoints were used in the Quickblind study (Braatz et al. 2014) performed under PARENT (Meyer and Heasler 2017) to facilitate convergence of the logistic function fit because data was sparse. The Quickblind study was performed under PARENT on a set of test blocks with laboratory grown stress corrosion cracks contributed by Japan. Four test blocks were examined in the Quickblind study, including one test block that was blank and three test blocks that each contained a single flaw. The test blocks were scheduled to be destructively analyzed before the rest of the test blocks in PARENT could be examined, which motivated the separate Quickblind activity.

The pseudopoints used in the analysis of data from the Quickblind study were carried over to the analysis of the larger PARENT dataset (Meyer and Heasler 2017) and the generation of POD curves reported in Meyer and Holmes (2019). This usage of pseudopoints is examined in Section 5.0 to understand their impact on the POD curves reported for PARENT in Meyer and Holmes (2019).

1.4 Absolute POD Estimation Versus Comparative Analysis

A distinction between an “absolute” POD” estimate and “comparative” POD analysis should be made when considering the adequacy of requirements for a POD study and for POD analysis. An absolute POD estimation refers to an absolute measure of POD, which is an attempt to determine the intrinsic quantitative POD value for an examination scenario. Absolute POD estimations are relevant for input in xLPR (Meyer and Holmes 2019) in which the quantitative values of POD for examination scenarios have an influence on calculating piping failure probabilities. In comparative POD analysis, relative comparisons of quantitative POD estimations for different examination scenarios are made to emphasize the influence that examination variables have on POD values. For instance, a comparative analysis can be performed to determine if an ID examination or OD examination is likely to result in better performance (i.e., higher POD value) and how much better performance can be expected. Comparative analysis may also be desirable to determine how the adoption of new technology

or procedures affects performance or how performance is affected by limited coverage scenarios (i.e., scenarios in which access is partially limited by geometry or material factors).

Requirements are more stringent for absolute POD estimations than for comparative POD analyses because the actual POD value has greater importance. The actual POD value is less important in a comparative POD analysis where error in absolute POD estimations may be tolerated so long as the relationship between two or more POD estimations is accurate.

1.5 Virtual POD Estimation

The creation of digital (or virtual) test blocks is a potential approach to overcoming challenges encountered in acquiring enough test blocks with manufactured flaws for NPP component POD studies. A so-called virtual flaw method was one approach investigated in PIONIC. This approach involves the creation of multiple virtual flaw responses from a small number of physical flaw responses using digital manipulation techniques.

A VRR activity was organized in PIONIC to demonstrate the capabilities of a virtual flaw approach. The VRR was designed based on an SBDMW test block that was used in the open testing of PARENT (Meyer et al. 2017). Circumferential flaw responses were obtained with a phased-array ultrasonic testing (PAUT) technique also employed in PARENT open testing. Several virtual flaw responses exceeding the minimum recommendation in MIL-HDBK-1823A were created from the small sample of physical flaw responses. Multiple teams then performed detection analysis on virtual test blocks created with the virtual flaw responses, and these results were used to estimate POD using logistic regression.

1.6 Contents of Report

Section 2.0 of this report provides a summary of the results of the POD analyses conducted in PINC, PARENT, and MRP 262 Rev. 3. The purpose of this summary is to provide readers with the state of POD estimations for SBDMW and LBDMW NPP components. In Section 3.0, a detailed description of the POD analysis procedure for PARENT is provided to compare results from other POD studies, such as the efforts reported in MRP 262 Rev. 3 and the PIONIC VRR, to the results of PARENT. Sections 4.0 and 5.0 include analyses of the influence of false call data and pseudopoints on the results of PARENT POD analysis, respectively. False call data was utilized in the analysis of PARENT data, in part, to compensate for the limited number of unique flaws. Section 6.0 includes an overview of the MIL-HDBK-1823A and ASTM E2862-18 standards, highlighting recommendations that were not fulfilled in the PARENT data analysis and providing additional analysis to fill gaps. A description and summary of the results of the VRR activity conducted under PIONIC is provided in Section 7.0, and guidance derived from the review of standards, review of the PARENT analysis, and the VRR experience is provided in Section 8.0. Finally, a summary and conclusion are included in Section 9.0.

2.0 Summary of POD Analysis Results

This section is a summary of the results of the POD analyses conducted in PINC, PARENT, and MRP-262 Rev. 3 and provides an overview of the comparison of results found in report PNNL-28090 (Meyer and Holmes 2019). This summary provides a status of empirical POD estimations for SBDMW and LBDMW components and highlights the lack of convergence of POD estimations for similar components, flaw orientations, and examination type (i.e., OD or ID).

The objective of PNNL-28090 (Meyer and Holmes 2019) was to document POD models to use with the xLPR probabilistic fracture mechanics code. In PNNL-28090, models are summarized for OD and ID examinations on SBDMW and LBDMW components. Furthermore, results are presented separately for axial and circumferential flaws.

Results are presented for multiple scenarios in MRP-262 Rev. 3 (EPRI 2017). Generally, results are presented based on analysis performed on: 1) passing examinations, and 2) all examinations, including both passing and failing examinations. Furthermore, outlier data are considered in MRP-262 Rev. 3 (EPRI 2017), and the results of analyses are provided for both cases of inclusion and exclusion of outlier data. In this report, and in PNNL-28090 (Meyer and Holmes 2019), the results from MRP-262 Rev. 3 that are presented were based on passing examination data and for which outlier data was included in the analysis (denoted as “All Data”).

2.1 Discrepancies Between PINC, PARENT, and MRP-262 Rev. 3

The plots in the following sections show that results documented in MRP-262 Rev. 3 deviate significantly from PINC and PARENT results, except for the LBDMW ID results for circumferential flaws shown in Figure 2-5. There are many possible explanations for the differences in these results, and some of them are outlined here:

- PDI is designed for qualification and not for fully characterizing NDE performance. It is a process for evaluating if procedures meet performance requirements defined in the American Society of Mechanical Engineers Boiler and Pressure Vessel Code, Section XI, Appendix VIII One consequence is that the databases analyzed in MRP-262 Rev. 3 are “unbalanced” and have many more detections than misses.
- MRP-262 Rev. 3 results are generated solely from ultrasonic testing data. In PINC and PARENT, many of the ID examination results were based on datasets that included results from eddy current testing.
- Flaw depth size distributions between MRP-262 Rev. 3, PINC, and PARENT were different, with the distributions of flaw depths in PINC and PARENT skewing to lower flaw depths relative to MRP-262 Rev. 3.
- Analysis methods deviated for MRP-262 Rev. 3, PINC, and PARENT. Specifically, false calls were not used in fitting the logistic regression model in MRP-262 Rev. 3 but were used in fitting the logistic regression model in PINC and PARENT.
- The MRP-262 Rev. 3 datasets were generally much larger (approximately by an order of magnitude) than the datasets for PINC and PARENT. This results in narrower confidence bounds for the MRP-262 Rev. 3 curves when compared to curves from PINC or PARENT.
- Between PINC and PARENT, there were some differences as well. Notably, a different grading unit size was used in the analysis for both datasets.

There were other factors that could have contributed to the differences in results, such as differences in the way flaws were manufactured for each of the studies and differences in which the tests were administered.

2.2 POD Results for SBDMWs

The POD results for SBDMWs are provided in this section. Results are displayed separately for circumferential and axial flaws for OD examinations in Figure 2-1 and Figure 2-2, respectively. In Figure 2-1, the result from MRP-262, Rev. 3 appears much flatter, resulting in POD estimations that are higher at smaller flaw depths in comparison to PINC and PARENT results but lower at larger flaw depths. In Figure 2-2, the curve for MRP-262, Rev. 3 results exhibits a greater dependence on flaw size but still estimates higher POD for small flaw sizes and lower POD at larger flaw sizes in comparison to PINC and PARENT results. As highlighted above (Section 2.1), there are several possible reasons for the results to deviate. Specifically, the distribution of flaw depths and unbalanced nature of the source data can be expected to contribute to POD curves exhibiting less variation with flaw depth. The results for PINC and PARENT also do not match; however, both studies produced curves displaying variation with flaw depth and a transition from low to high POD values. Some of the deviation between PINC and PARENT results can be attributed to the selection of grading unit sizes in the analyses procedures. The influence of grading unit sizes is explored further in Section 4.0.

Inspections of Figure 2-1 and Figure 2-2 reveal an estimation of higher POD for circumferential flaws relative to axial flaws for each respective study (i.e., MRP 262 Rev. 3, PINC, PARENT). This outcome is expected because most or all the sound path for an ultrasonic examination of an axial flaw must be within the weld material. For examination of circumferential flaws, only a portion of the sound path will be in the weld material and that portion will depend on the location of the flaw relative to the weld and the probe. For both the OD and ID examinations, the confidence bounds are wider for axial flaws because the NOBS of axial flaws was smaller than the NOBS of circumferential flaws for each study.

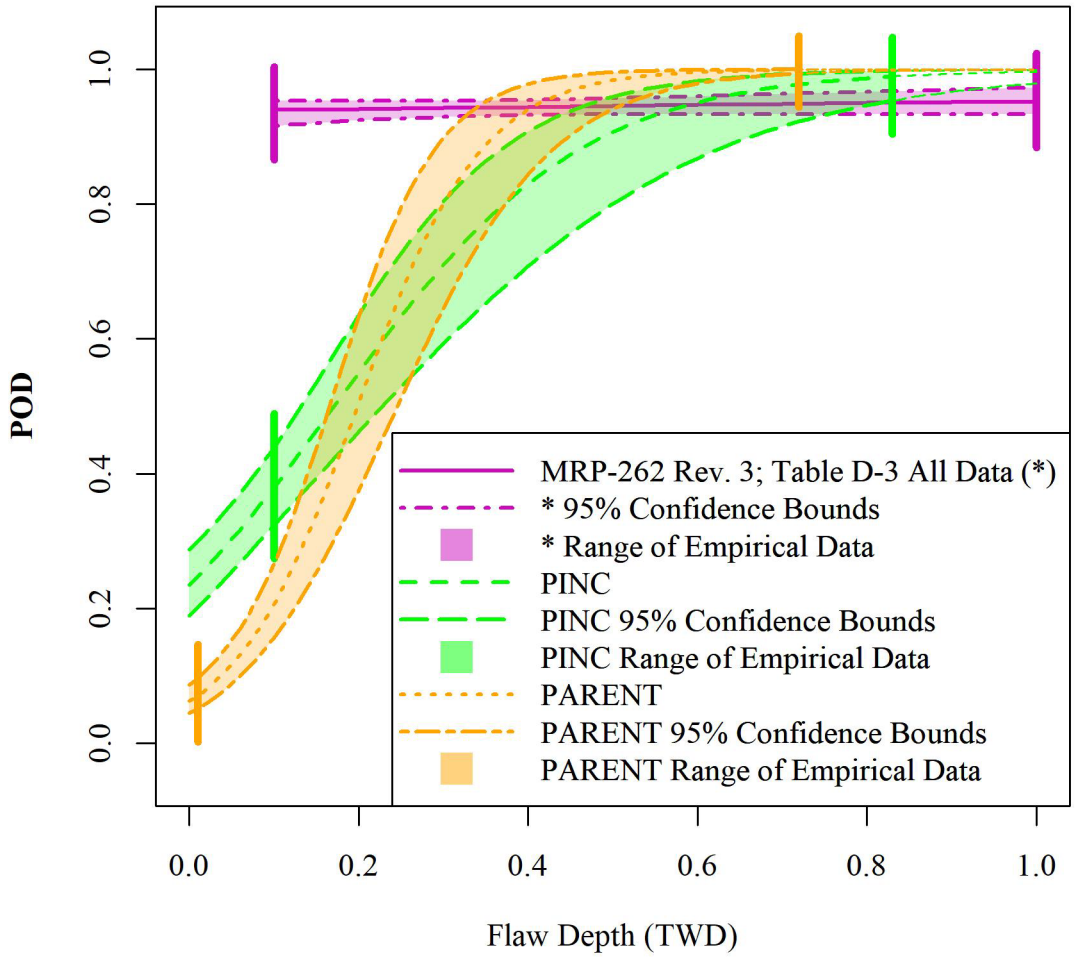

Figure 2-1 POD as a function of flaw depth (TWD) for circumferentially oriented flaws in SBDMW test blocks in PINC, PARENT, and MRP-262 Rev. 3 efforts (OD exams). The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

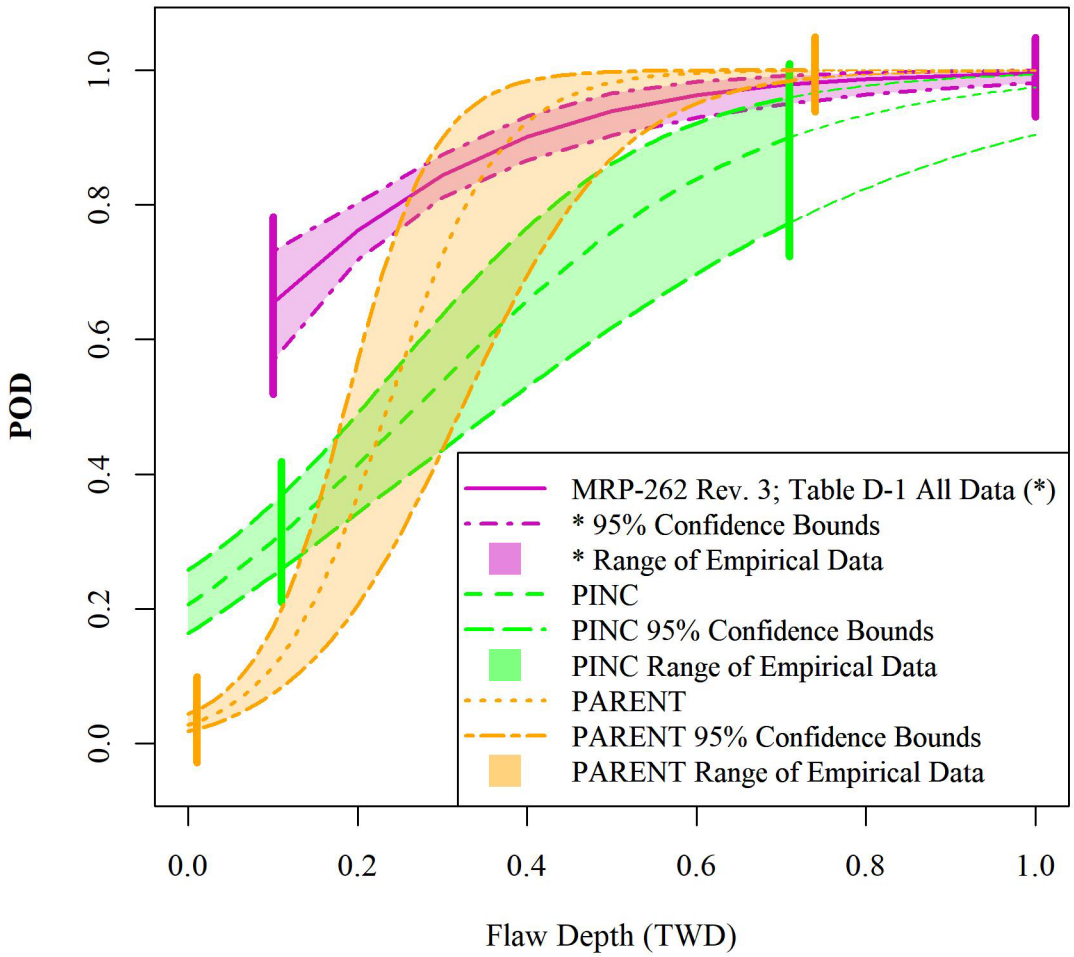

Figure 2-2 POD as a function of flaw depth (TWD) for axially oriented flaws in SBDMW test blocks in PINC, PARENT, and MRP-262 Rev. 3 efforts (OD exams). The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

2.3 POD Results for LBDMWs

The POD results for LBDMWs for OD and ID examinations are provided in this section. For each type of exam (OD or ID), results are displayed separately for circumferential and axial flaws. The results for OD examinations for circumferential and axial flaws are displayed in Figure 2-3 and Figure 2-4, respectively, and the results of ID examinations for circumferential and axial flaws are displayed in Figure 2-5 and Figure 2-6, respectively. Figure 2-3 and Figure 2-4 only contain plots from PARENT because MRP 262, Rev. 3 and PINC did not include ID examinations for LBDMWs. As expected, it is evident that POD for ID examinations is

significantly higher than POD for OD examinations when comparing the PARENT OD and ID results for LBDMWs.

Figure 2-5 and Figure 2-6 display results from MRP 262, Rev. 3 and PARENT. No plot is included for PINC because PINC did not include examinations of LBDMWs. From Figure 2-5, it is apparent that the curves from MRP 262, Rev. 3 and PARENT both predict higher POD for ID examinations of LBDMWs. The data from MRP 262, Rev. 3 was collected from a population of larger flaw sizes (> 10% TWD), whereas the PARENT data was limited to flaws 1%–36% TWD. The curves from MRP 262, Rev. 3 appear relatively flat for both circumferential flaws, whereas the empirical flaw depth range for PARENT appears to coincide with a transition from low to high POD.

A comparison of Figure 2-6 to Figure 2-5 shows that the confidence bounds appear larger for the axial flaws in both PARENT and MRP 262, Rev. 3 data. For both studies, the NOBS of axial flaws was less than the NOBS of circumferential flaws. The PARENT curve in Figure 2-6 also indicates lower POD for the axial flaws in comparison to the circumferential flaws. The MRP 262, Rev. 3 results also appear to show lower POD values for axial flaws relative to circumferential flaws, although the difference is modest. For the axial flaws, the portions of MRP 262, Rev. 3 and PARENT curves exhibit substantial disagreement for flaw sizes approximately < 20% TWD.

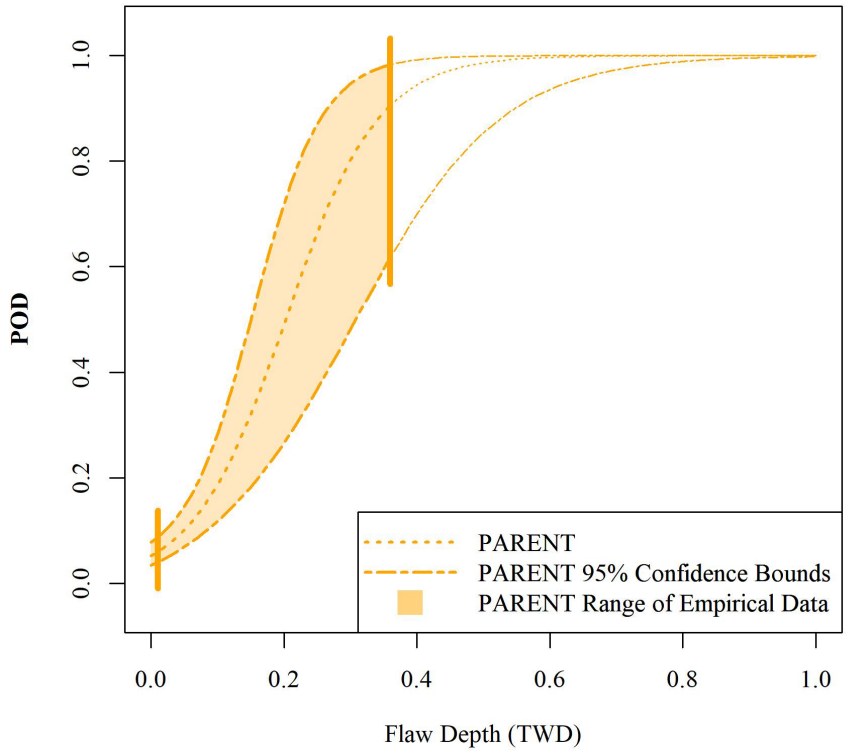

Figure 2-3 POD as a function of flaw depth (TWD) for circumferentially oriented flaws in LBDMW test blocks in PARENT (OD exams). The vertical lines indicate the range from minimum to maximum flaw depth. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

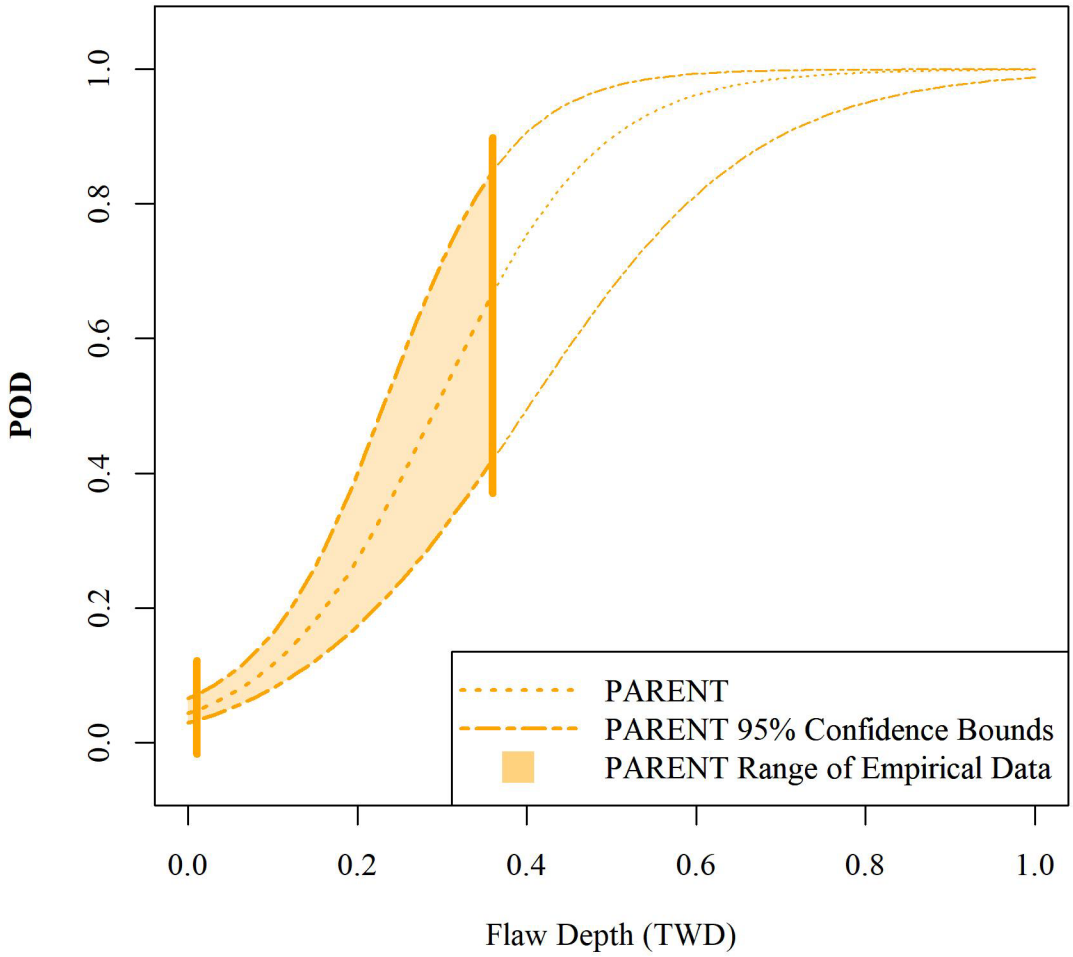

Figure 2-4 POD as a function of flaw depth (fraction of TW, *x*) for axially oriented flaws in LBDMW test blocks in PARENT (OD access). The vertical lines indicate the range from minimum to maximum flaw depth. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

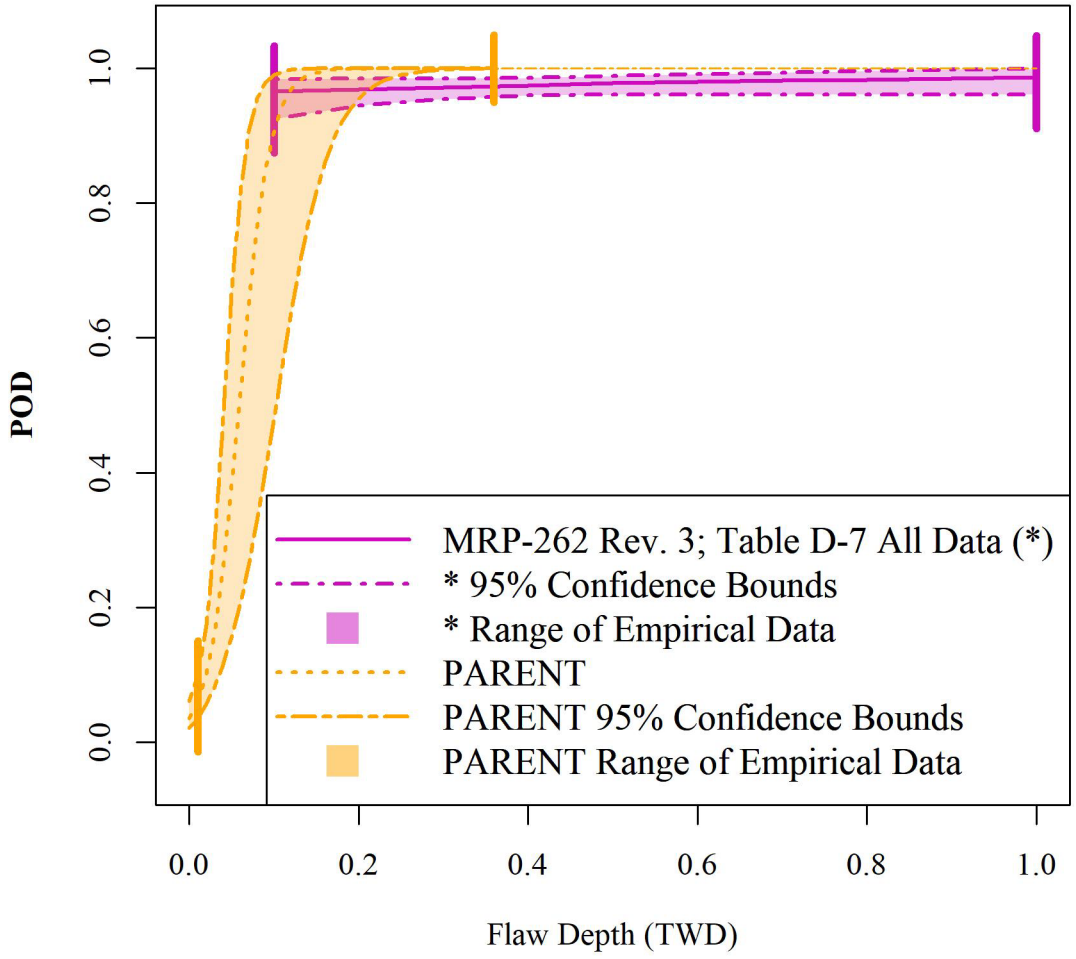

Figure 2-5 POD as a function of flaw depth (TWD) for circumferentially oriented flaws in LBDMW test blocks in PARENT and MRP 262, Rev. 3 (ID exams). The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

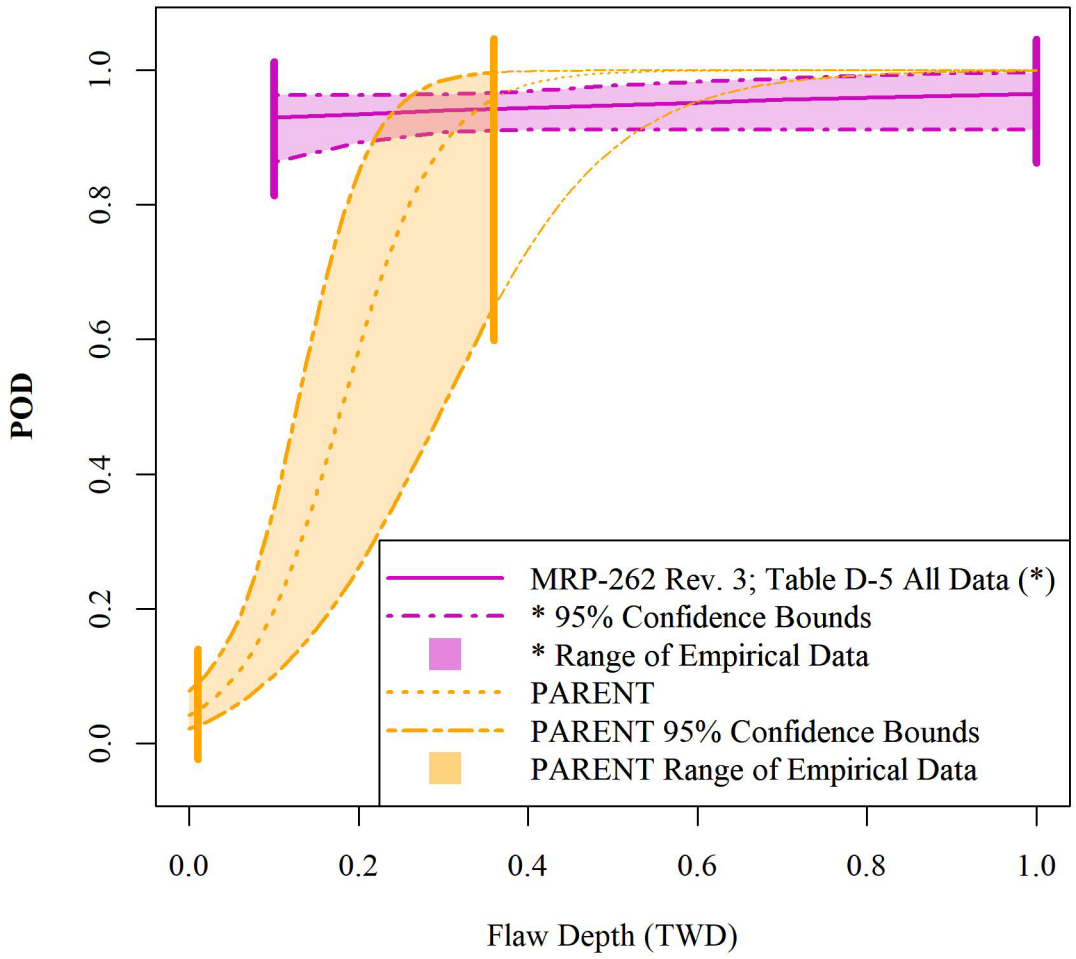

Figure 2-6 POD as a function of flaw depth (TWD) for axially oriented flaws in LBDMW test blocks in PARENT and MRP 262, Rev. 3 (ID exams). The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

3.0 Description of PARENT Analysis

In this section, a detailed description of the POD analysis for PARENT data is provided to compare results from other studies, such as MRP 262 Rev. 3 and the PIONIC VRR, to PARENT. In addition to differences in the underlying data, as summarized in Section 2.1, differences in the way the data were analyzed for MRP 262 Rev. 3, PINC, and PARENT could contribute to inconsistencies between POD estimations for each study. The inclusion of false call data in fitting the logistic function for POD is one aspect of POD analysis that was different for MRP 262 Rev. 3 in comparison to PINC and PARENT. The false call data was included for fitting the logistic function in PINC and PARENT and was not included for MRP 262 Rev. 3. Thus, the influence of false call data on the results from PARENT is evaluated in Section 4.0.

3.1 Data Scoring

Flaws and indications were represented as rectangles with dimensions bounding their sizes on the test block surface, for the purpose of scoring test results in PARENT. The flaw and indication bounding rectangles were defined by dimensions in the circumferential direction (X) and axial direction (Y) of the simulated components, as depicted in Figure 3-1. “Flaws” refers to the defects intentionally manufactured into the test blocks for simulating PWSCC or IGSCC degradation. “Indications” refers to NDE responses reported by the examination teams as being caused by flaws. NDE responses caused by known geometric features were not reported as “indications.”

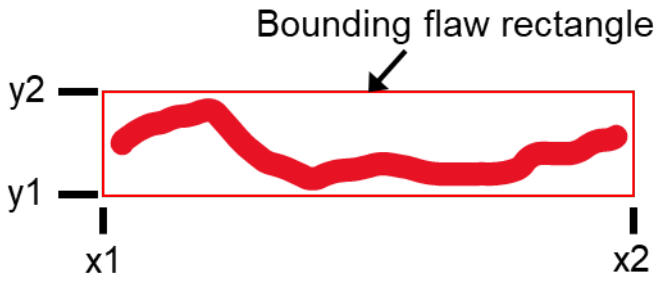

Figure 3-1 A 2-D bounding rectangle representing a circumferential flaw for data scoring in PARENT.

Tolerance was included in the scoring procedure to allow some margin for error. This tolerance was incorporated by adding to the overall dimensions of the bounding rectangles for flaws and indications in the X and Y dimensions. In the analysis of PARENT blind testing data, the tolerance added to the X dimension, δx, and the Y dimension, δy, was the same. An illustration of the tolerance applied to a flaw and an indication is shown in Figure 3-2. In PARENT, δx = δy = 5 mm (0.2 inch).

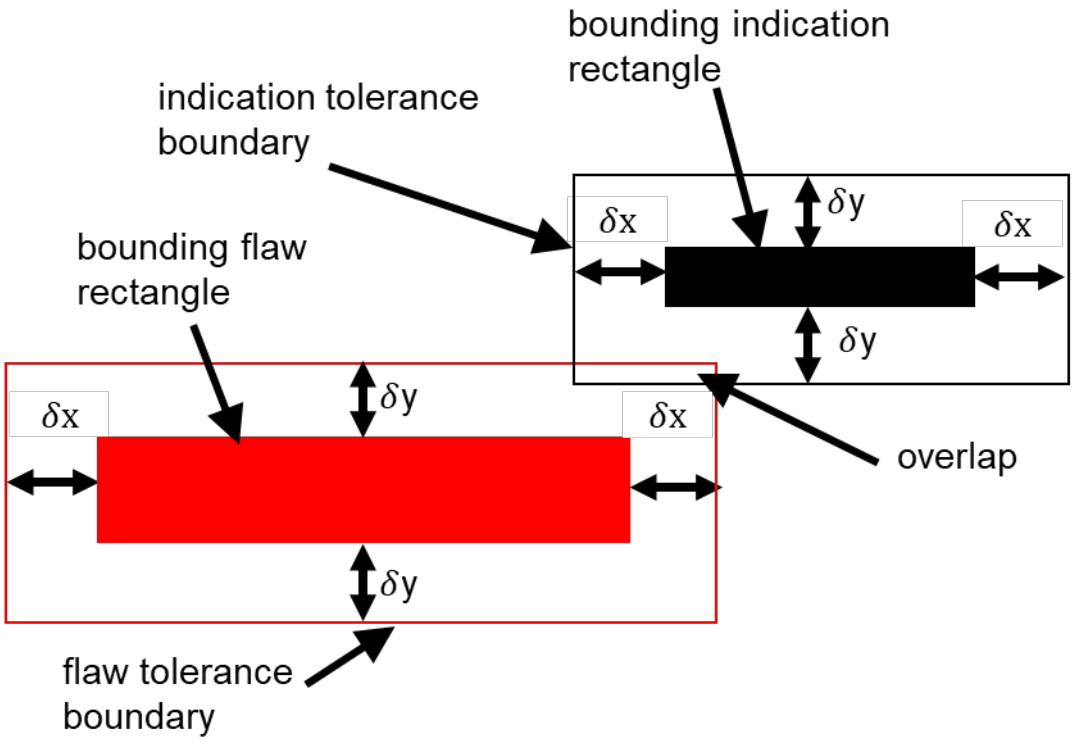

Figure 3-2 Application of tolerance to bounding flaw and bounding indication rectangles. Overlap of the tolerance boundary for an indication with the tolerance boundary of a flaw classifies the indication as a "hit" (i.e., true detection).

The outcomes of a test block examination could include "hits" (true detections), "misses" (missed detections), and "false calls" (false detections). An indication was scored as a "hit" when the tolerance boundary of an indication overlapped with the tolerance boundary of a flaw, an indication was scored as a "false call" when its tolerance boundary did not overlap with the tolerance boundary of any flaws, and a "miss" was scored when a flaw was not associated with any overlapping indications. These outcomes are each illustrated in Figure 3-3. For the hit example, two indications are shown overlapping with Flaw A. Although not typical, the scenario was encountered in the PARENT data. To score this scenario, the indications were considered as a single indication and scored as a hit.

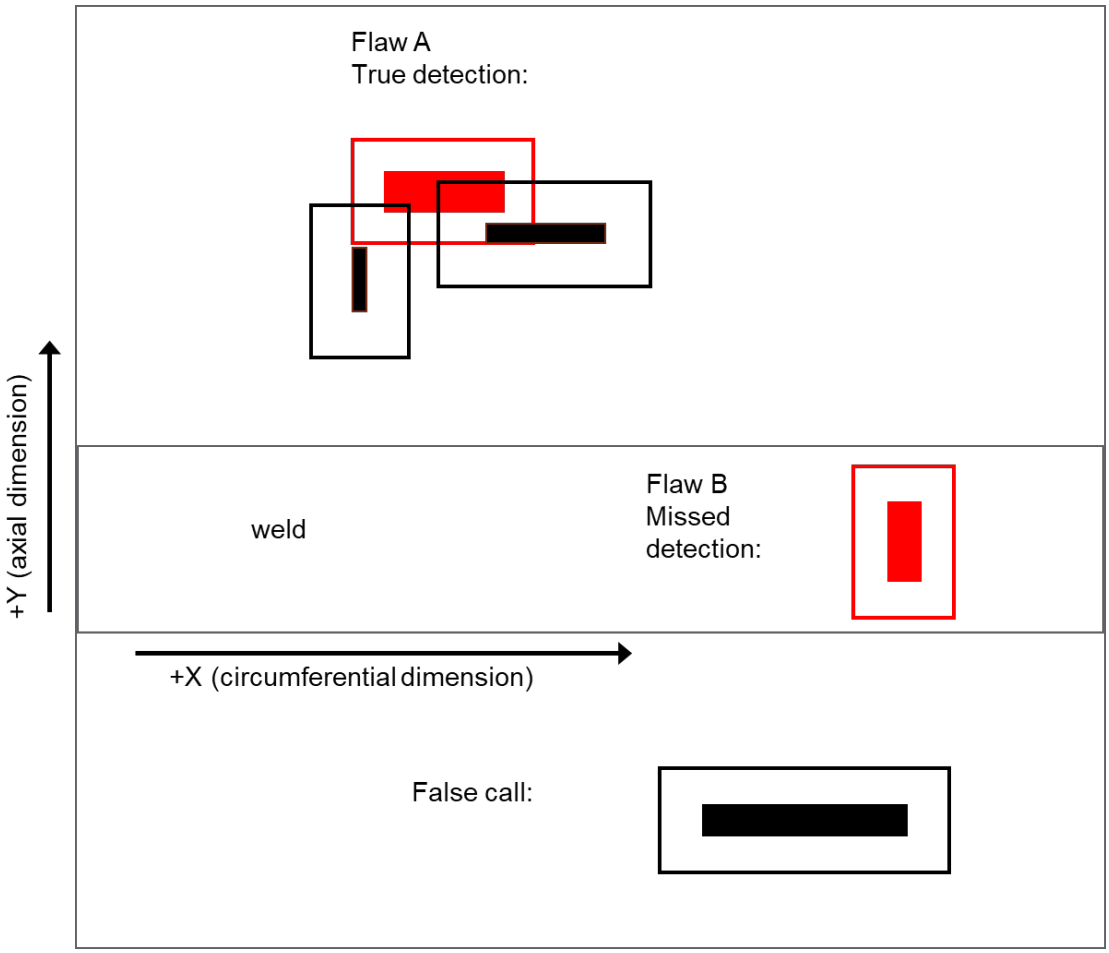

Figure 3-3 Depiction of a hit (true detection), miss (missed detection), and a false call (false detection). Flaws are shown with red color and indications are shown in black.

In the analysis of the PARENT blind test data, the scoring process was automated. However, indication plots were generated to provide a convenient visual record of examination results to allow manual review as a check. The indication plots are useful for identifying uncommon scoring scenarios, as highlighted in Sections 4.4, 4.5, and 6.4 of Meyer and Heasler (2017). These uncommon scenarios can be described as having multiple indications associated with one flaw, multiple flaws associated with one indication, and systematic positioning errors, respectively. Methods were defined for scoring each of those scenarios, and the method for scoring multiple indications associated with a single flaw is described in the previous paragraph. The method implemented for scoring multiple flaws associated with a single indication was to count both flaws as detections. No method was implemented for the systemic positioning error, but an analysis of the effect of increasing the tolerance boundary size on the estimated performance for procedure UT.ECT.144 is included in Section 6.4 of Meyer and Heasler (2017). Indication plots for all examinations performed in PARENT blind testing are compiled in Appendix E of Meyer and Heasler (2017).

3.2 Treatment of False Calls

In PARENT (Meyer and Heasler 2017), false call data was used in regression fitting to generate logistic function POD curves. This required converting the false call data to a probability that was consistent with the concept of probability of flaw detection. False call data was converted into false call probability (FCP) through Equation 3.1 where λ_{fc} is the false call rate and L_{gu} is the average length of an unflawed (or blank) grading unit. The false call rate, λ_{fc}, can be expressed as the number of false calls per length of unflawed material that was examined (Equation 3.2).

$$\text{FCP} = 1 - exp(-\lambda_{fc} * L_{gu}), \quad \text{Equation 3.1}$$

$$\lambda_{fc} = \frac{\#\text{False Calls}}{\text{Length of Unflawed Material Examined}} \quad \text{Equation 3.2}$$

Equation 3.1 is derived by representing FCP as a Poisson distribution, which is a discrete probability distribution expressing the probability that a random event will occur in fixed spatial intervals. In this application, false calls are the random events of interest, and unflawed grading units are fixed spatial intervals. To use Equation 3.1, the rate of false calls must be expressible as a constant mean rate (i.e., λ_{fc}), and each false call must be independent of other false calls. The number of false calls is simply the total number of false calls observed for an examination. In PARENT, the grading unit concept was introduced for the purpose of analysis, and the test blocks were not actually divided into grading units for the examinations. Thus, Equation 3.1 can be interpreted as the probability of a false call occurring in any section of unflawed material with length, L_{gu}.

The length of unflawed inspected material is calculated by subtracting the lengths of each flaw plus the tolerance added to each flaw from the total length of inspected material. In this case, “length” refers to the dimensions of the flaws in the circumferential direction, and the tolerance included in the calculation is the tolerance added in both the positive and negative circumferential directions for the flaw.

The length of a blank grading unit, L_{gu}, should be the same as the average size of flawed grading units so that the flawed and unflawed regions are weighted equally. In PARENT blind testing, the average size of the flawed grading units was calculated by summing the length of each flaw in the circumferential direction plus the tolerance added to each flaw in both the positive and negative circumferential directions and dividing this sum by the total number of flaws in the inspection region. In PARENT, this value was approximately 25 mm (1.0 inch).

FCP can also be estimated by dividing the number of false calls by the total number of blank grading units. The total number of blank grading units can be calculated by dividing the length of unflawed material that was examined by L_{gu}. The number of blank grading units and FCP can be expressed in the following equations:

$$\#\text{ Blank Grading Units} = \frac{Length\ of\ Unflawed\ Inspected\ Material}{L_{gu}} \quad \text{Equation 3.3}$$

$$FCP = \frac{\#False\ Calls}{\#Blank\ Grading\ Units} \qquad \text{Equation 3.4}$$

Figure 3-4 depicts a section of unflawed material that is divided into blank grading units consistent with Equation 3.3. In this example, two false calls are observed over seven blank grading units, so the FCP = 2/7 = 29% according to Equation 3.4. Figure 3-4 depicts a false call overlapping the boundary of two blank grading units. If each overlapped blank grading unit were credited with a false call, then FCP = 3/7 = 43%. Each false call was counted only one time in PARENT, regardless of if a false call overlapped more than one grading unit.

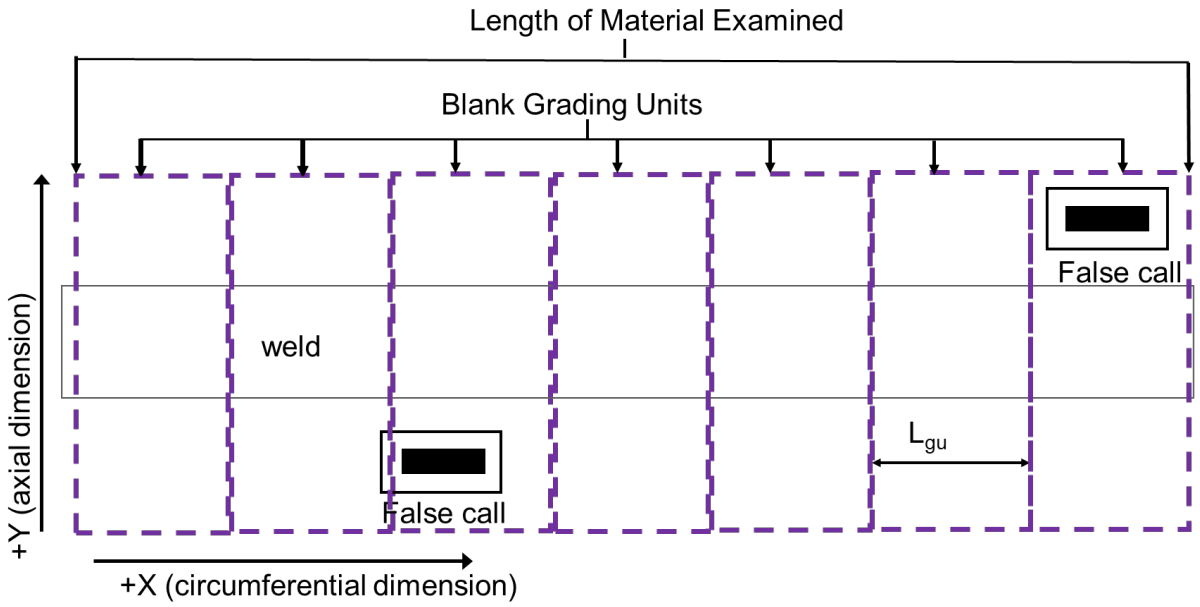

Figure 3-4 Section of examined unflawed material divided into blank grading units for calculation of FCP in PARENT.

3.3 Data Scoring Examples

Three case studies are provided in this section to give examples of the scoring procedure applied to PARENT data. These three case studies were derived from examinations performed in PARENT and documented in Meyer and Heasler (2017). The case studies are outlined below by defining the test block, procedure, and notable features of the examination. The category of examination is also noted according to test block type (SBDMW or LBDMW) and if the examination was OD or ID. A procedure could include one or more types of NDE techniques. The types of techniques were "coded" into the procedure identifier with acronyms. Phased-array ultrasonic techniques were coded as PAUT, eddy current testing techniques as ECT, conventional UT techniques as UT, and time-of-flight diffraction ultrasonic testing techniques as TOFD. The number at the end of the procedure identifier was associated with the team that performed the examination in PARENT.

- Case 1: SBDMW OD
  - Test Block: P35

  - Procedure: PAUT.108
  - Note: One indication that overlaps two flaws
- Case 2: LBDMW ID
  - Test.Block: P33
  - Procedure: ECT.135
  - Note: Two indications that overlap one flaw
- Case 3: LBDMW ID
  - Test Block: P33
  - Procedure: UT.TOFD.ECT.101
  - Note: Three false calls

3.3.1 Case 1: SBDMW OD (Test Block P35, Procedure PAUT.108)

Case 1 highlights an SBDMW OD examination performed on test block P35 by procedure PAUT.108. This example includes the unusual scenario of one indication overlapping two flaw boundaries. The indication plot for this examination is divided into three plots over the length of the circumferential weld and is presented in Figure 3-5 through Figure 3-7. In these plots, the flaws are depicted by solid red bounding rectangles surrounded by a tolerance boundary, which is shown by a red line. The indications are depicted as solid black bounding rectangles with a black line tolerance boundary. Flaws are numbered from 23 to 34 in the positive X direction, and indications are numbered from 664 to 670 in the positive X direction. The three plots (Figure 3-5 through Figure 3-7) depict the full circumference of the weld in separate plots ranging from X = 0 to 300 mm, X = 300 to 600 mm, and X = 600 mm to 900 mm, respectively. There is no break in the weld, so the position at X = 900 mm in Figure 3-7 corresponds to the same position as X = 0 in Figure 3-5.

From Figure 3-5 through Figure 3-7, it is evident that there were four missed detections in this examination (flaws 23, 29, 32, and 33), and there were no false calls. There were a total of nine hits in this examination and two hits could be associated with indication 665 because it overlapped flaws 25 and 26. Indication 665 appears to be circumferential and consistent with the orientation of flaw 26, while flaw 25 is axial. While the apparent orientation of the indication could have also been factored into the scoring procedure, it was not in PARENT. A summary of the scoring results that were the basis for the regression analysis is provided in Table 3-1. The table includes the number of detections (NDET) and the number of observations (NOBS) as used in Section 1.2.

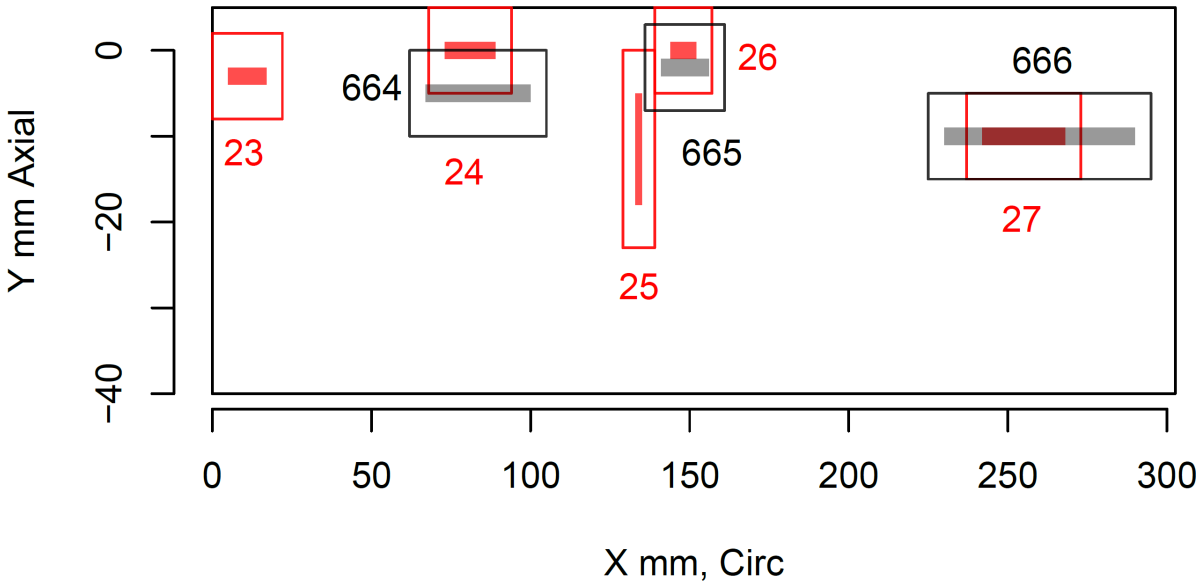

Figure 3-5 Indication plot for SBDMW OD examination on P35 by procedure PAUT.108 from X = 0 to 300 mm.

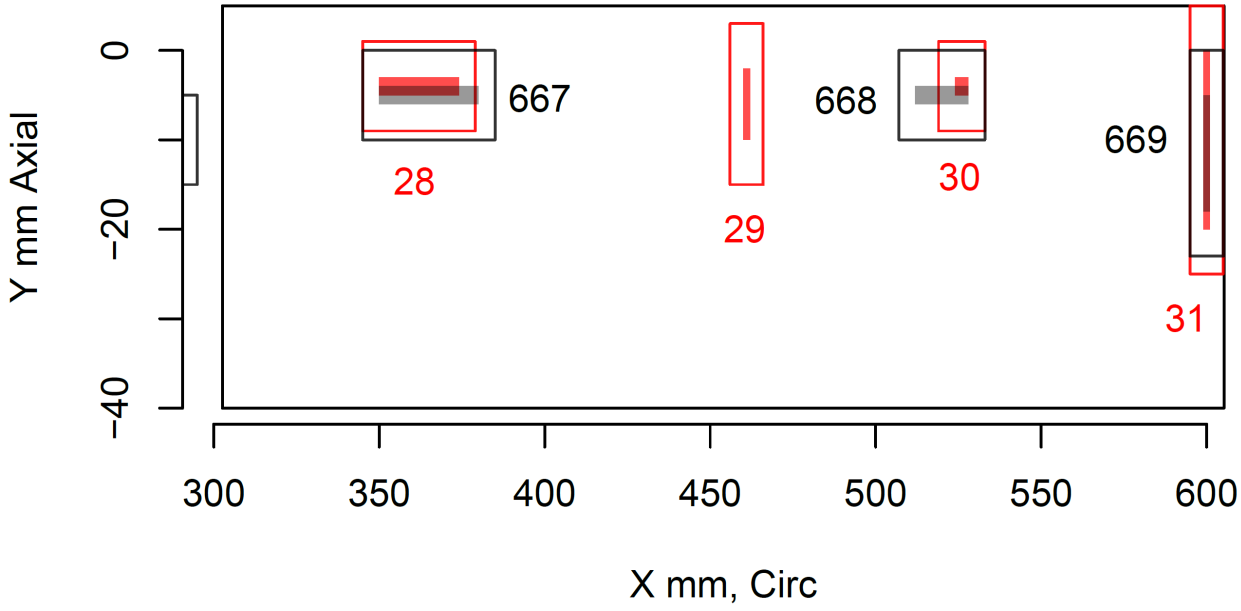

Figure 3-6 Indication plot for SBDMW OD examination on P35 by procedure PAUT.108 from X = 300 to 600 mm.

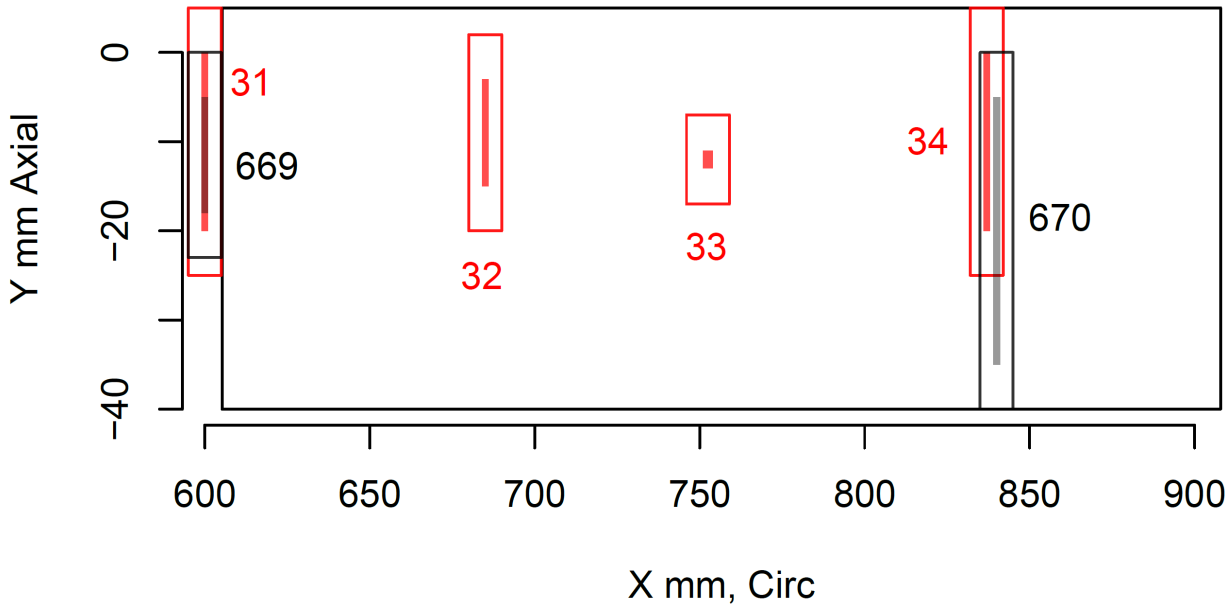

Figure 3-7 Indication plot for SBDMW OD examination on P35 by procedure PAUT.108 from X = 600 to 900 mm.

Table 3-1 Scoring results for examination of test block P35 by procedure PAUT.108.

Flaw ID	Indication ID	NDET	NOBS	Grading Unit
23	NA	0	1	Flawed
24	664	1	1	Flawed
25	665	1	1	Flawed
26	665	1	1	Flawed
27	666	1	1	Flawed
28	667	1	1	Flawed
29	NA	0	1	Flawed
30	668	1	1	Flawed
31	669	1	1	Flawed
32	NA	0	1	Flawed
33	NA	0	1	Flawed
34	670	1	1	Flawed
NA	NA	0	28	Unflawed

3.3.2 Case 2: LBDMW ID (Test Block ID P33, Procedure ID ECT.135)

Case 2 highlights an LBDMW ID examination performed on test block P33 by procedure ECT.135. The result for this examination was divided into three plots over the length of the circumferential weld and is presented in Figure 3-8 through Figure 3-10. Flaws are numbered

from 5 to 22 in the positive X direction, and indications are numbered from 463 to 476 in the positive X direction.

As illustrated Figure 3-8 through Figure 3-10, there were four missed detections in this examination (flaws 6, 8, 14, and 22), and there was one false call (indication #465). It appeared that there were two more missed detections (flaws 13 and 18); however, because these flaws were not surface breaking, they were excluded from the analysis of ECT detection performance. Finally, there were a total of 14 hits in this examination.

Figure 3-10 displays two indications (#472 and #473) that overlap flaw 17. In this case, because indication #472 had a larger area overlapping with flaw 17, it was associated with the detection of flaw 17 and indication #473 was discarded. A summary of the scoring results for this examination is provided in Table 3-2.

As noted, this examination resulted in one false call. The FCP was calculated using Equation 3.1 and estimated by Equation 3.4 for comparison. The first step in calculating FCP by Equation 3.1 is to calculate the false call rate, λ_{fc}, by Equation 3.2. This was determined by dividing the total number of false calls observed by the length of unflawed material that was examined, which was 2,471.5 mm (97.3 inches).

The false call rate for this scenario was calculated from Equation 3.2 as,

$$\lambda_{fc} = \frac{1 \text{ false call}}{2471.5 \text{ mm}} = 0.000405 \; false \; calls/mm,$$

and this result can be plugged into Equation 3.1 to obtain the FCP,

$$\text{FCP} = \; 1 - exp(-0.000405 \; false \; calls/mm * 25 \; mm) = 0.0101 = \underline{1.0\%}.$$

Equation 3.4 can also be used to estimate FCP by dividing the number of false calls by the total number of grading units. The total number of grading units was calculated by dividing the length of unflawed inspected material by the blank grading unit size,

$$Number \; of \; Blank \; Grading \; Units = \frac{2471.5 \; mm}{25 \; mm} \approx 99.$$

Thus,

$$FCP = \frac{1 \; false \; call}{99 \; blank \; grading \; units} = 0.0101 \approx \underline{1.0\%}.$$

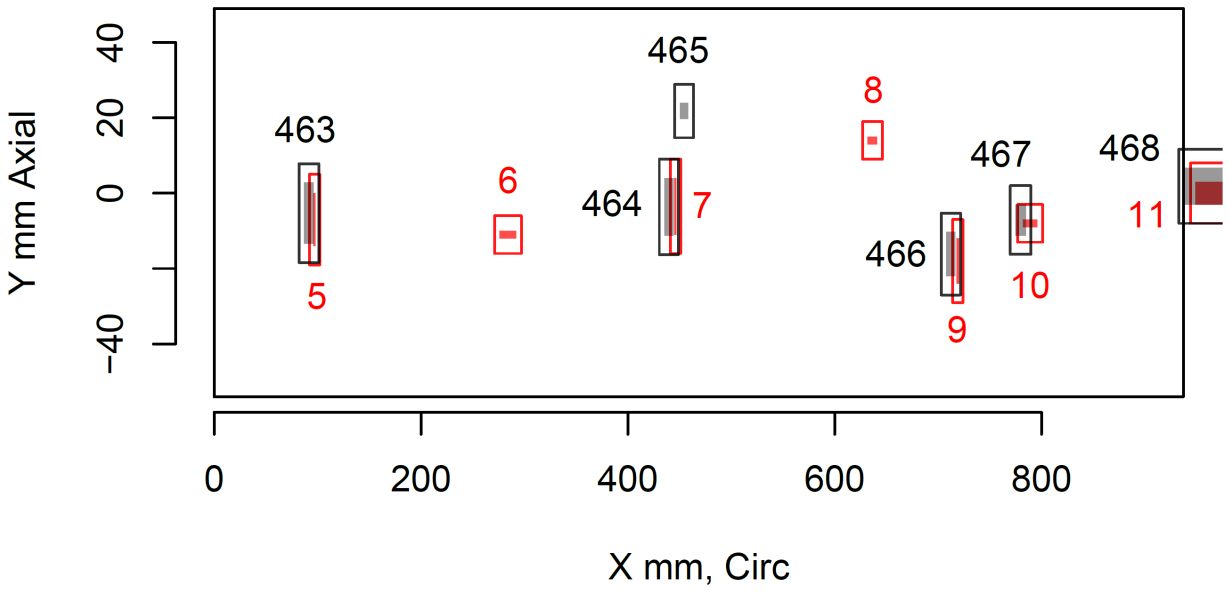

Figure 3-8 Indication plot for LBDMW ID examination on P33 by procedure ECT.135 from X = 0 to 1,000 mm.

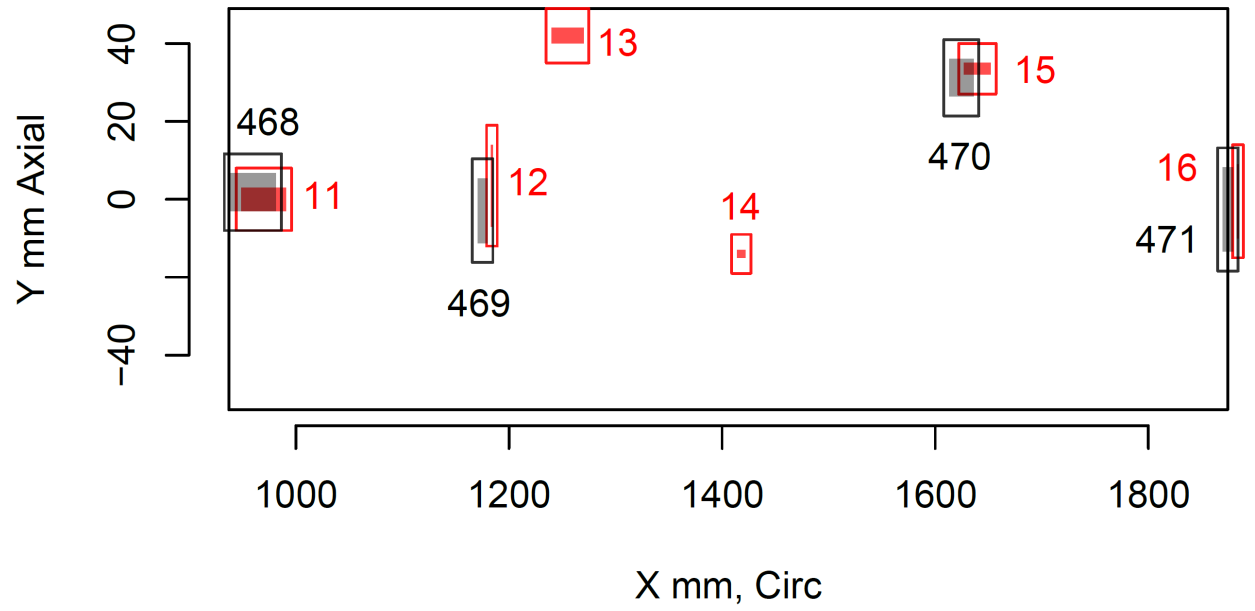

Figure 3-9 Indication plot for LBDMW ID examination on P33 by procedure ECT.135 from X = 1,000 to 2,000 mm.

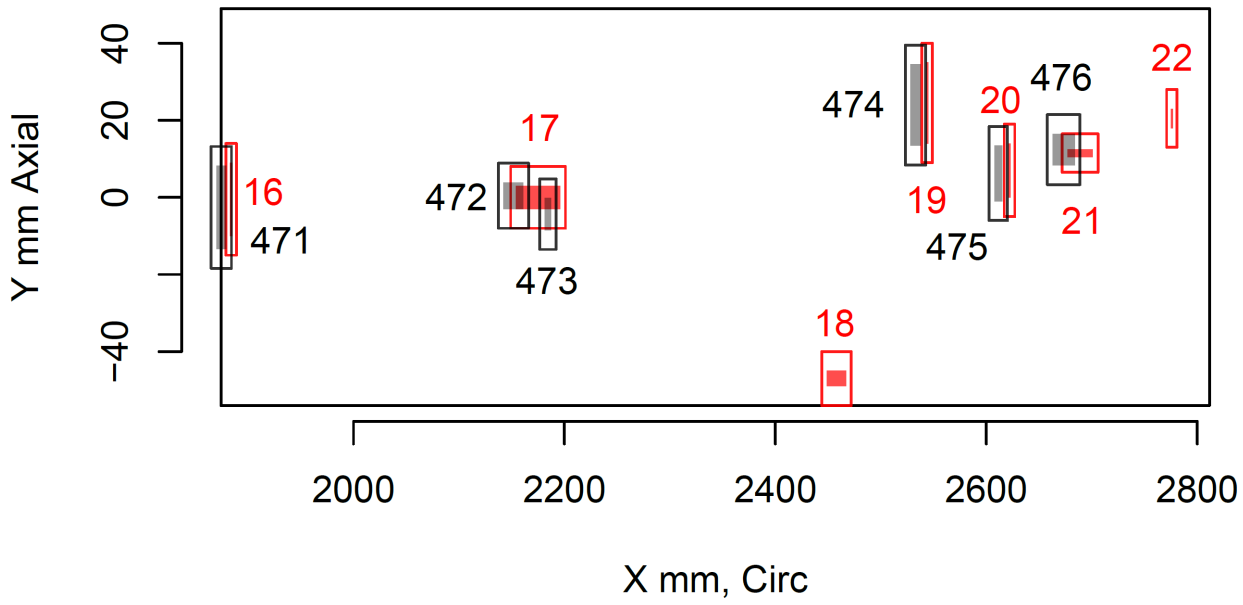

Figure 3-10 Indication plot for LBDMW ID examination on P33 by procedure ECT.135 from X = 2,000 to 2,800 mm.

Table 3-2 Scoring results for examination of test block P33 by procedure ECT.135.

Flaw ID	Indication ID	# of Detections	# of Observations	Grading Unit
5	463	1	1	Flawed
6	NA	0	1	Flawed
7	464	1	1	Flawed
8	NA	0	1	Flawed
9	466	1	1	Flawed
10	467	1	1	Flawed
11	468	1	1	Flawed
12	469	1	1	Flawed
14	NA	0	1	Flawed
15	470	1	1	Flawed
16	471	1	1	Flawed
17	472	1	1	Flawed
19	474	1	1	Flawed
20	475	1	1	Flawed
21	476	1	1	Flawed
22	NA	0	1	Flawed
NA	NA	1	99	Unflawed

3.3.3 Case 3: LBDMW ID (Test Block ID P33, Procedure ID UT.TOFD.ECT.101)

Case 3 highlights an ID examination performed on LBDMW test block P33 by procedure UT.TOFD.ECT.101. The indication plot for this examination is divided into three plots over the length of the circumferential weld and is presented in Figure 3-11 through Figure 3-13. Flaws are numbered from 5 to 22 in the positive X direction and indications are numbered from 57 to 73 in the positive X direction.

Figure 3-11 through Figure 3-13 show there were four missed detections in this examination (flaws 8, 14, 16, and 20) and there were three false calls (indication 60, 67, and 71). Finally, there were a total of 14 hits in this examination. A summary of the scoring results for this examination is provided in Table 3-3.

In this example, the length of unflawed material that was examined was 2,403.5 mm (94.6 inches). It may be noticed that this value is 68 mm (2.7 inches) smaller than the value calculated in Case 2, even though they consider the same test block. The discrepancy was due to the exclusion of flaws 13 and 18 in the analysis for Case 2 because they were not surface connecting flaws. In this case, these flaws were considered in the analysis because the inspection procedure, UT.TOFD.ECT.101, incorporates volumetric techniques.

The false call rate for this scenario was calculated from Equation 3.2 as,

$$\lambda_{fc} = \frac{3}{2403.5\ \text{mm}} = 0.00125\ false\ calls/mm,$$

and this result was plugged into Equation 3.1 to obtain the FCP,

$$\text{FCP} = 1 - exp(-0.00125\ false\ calls/mm \times 25\ mm) = 0.03079 = \underline{3.1\%}.$$

Equation 3.4 can also be used to estimate FCP by dividing the number of false calls by the total number of grading units. The total number of grading units was calculated by dividing the length of unflawed inspected material by the blank grading unit size,

$$Number\ of\ Blank\ Grading\ Units = \frac{2403.5\ mm}{25\ mm} \approx 96.$$

Thus,

$$FCP = \frac{3\ false\ calls}{96\ blank\ grading\ units} = 0.03125 \approx \underline{3.1\%}.$$

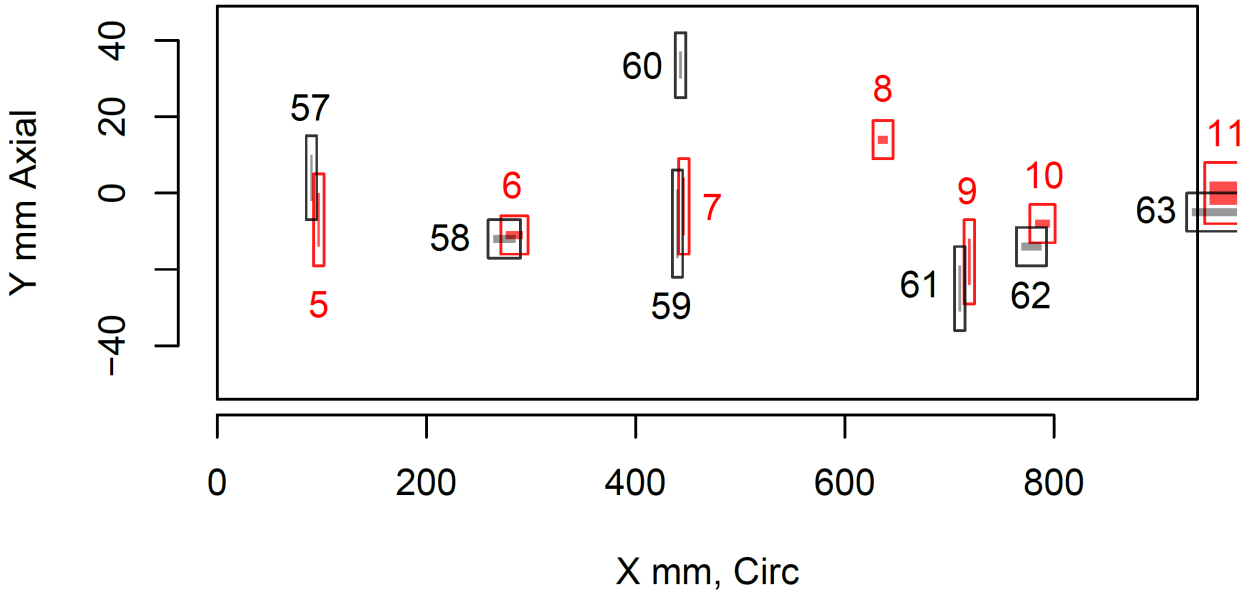

Figure 3-11 Indication plot for LBDMW ID examination on P33 by procedure UT.TOFD.ECT.101 from X = 0 to 1000 mm.

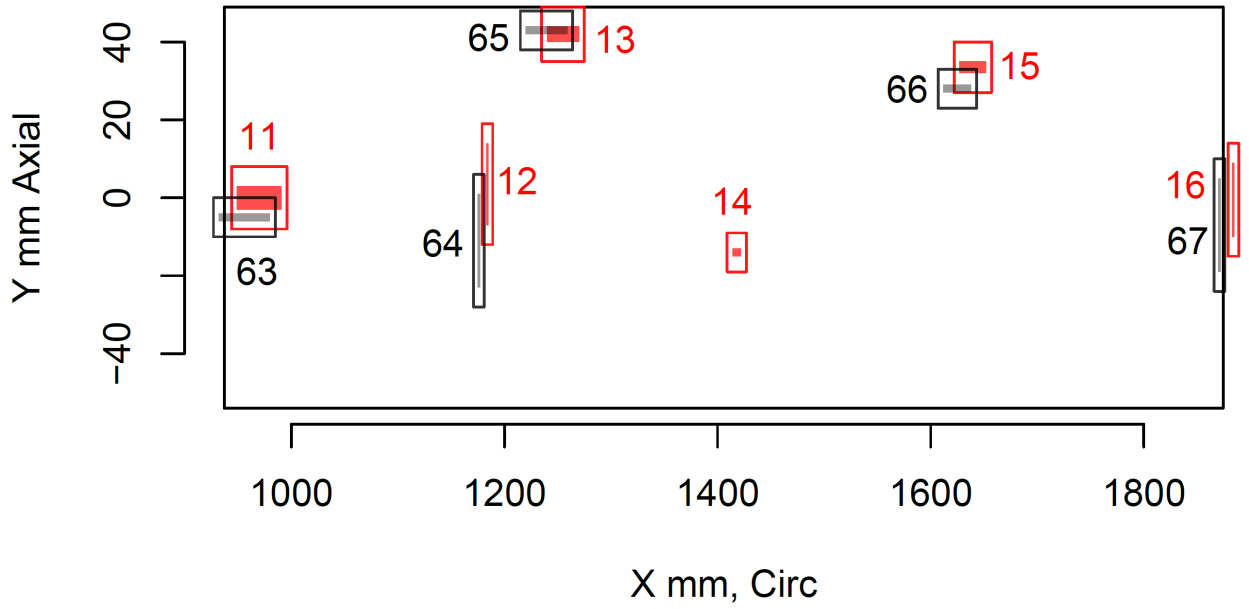

Figure 3-12 Indication plot for LBDMW ID examination on P33 by procedure UT.TOFD.ECT.101 from X = 1,000 to 2,000 mm.

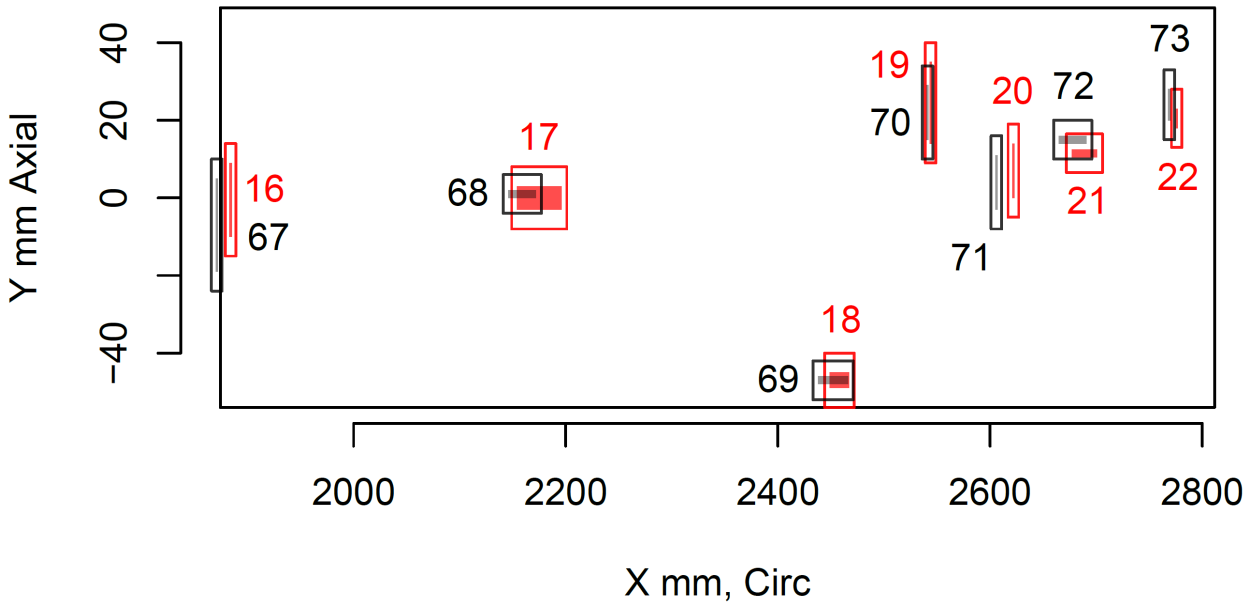

Figure 3-13 Indication plot for LBDMW ID examination on P33 by procedure UT.TOFD.ECT.101 from X = 2,000 to 2,800 mm.

Table 3-3 Scoring results for examination of test block P33 by procedure UT.TOFD.ECT.101.

Flaw ID	Indication ID	NDET	NOBS	Grading Unit
5	57	1	1	Flawed
6	58	1	1	Flawed
7	59	1	1	Flawed
8	NA	0	1	Flawed
9	61	1	1	Flawed
10	62	1	1	Flawed
11	63	1	1	Flawed
12	64	1	1	Flawed
13	65	1	1	Flawed
14	NA	0	1	Flawed
15	66	1	1	Flawed
16	NA	0	1	Flawed
17	68	1	1	Flawed
18	69	1	1	Flawed
19	70	1	1	Flawed
20	NA	0	1	Flawed
21	72	1	1	Flawed
22	73	1	1	Flawed
NA	NA	3	96	Unflawed

3.4 Maximum Likelihood Estimation

Maximum likelihood estimation (Rossi 2018) was used to fit the logistic function to the binary (hit/miss) POD data from PARENT. MLE for logistic regression is like least squares optimization for performing a linear regression. With MLE, initial guesses were made for the β_1 and β_2 parameters in Equation 1.1 to fit to the data, and the likelihood of the curve estimate was calculated for each flaw depth using Equation 3.5.

$$L(p; NDET) = \frac{NOBS!}{NDET!(NOBS-NDET)!} p^{NDET}(1-p)^{NOBS-NDET}. \quad \text{Equation 3.5}$$

In this equation, L is the likelihood, NDET and NOBS are the number of detections and the number of observations for a flaw at a given depth, and p is the POD estimated by the logistic curve fit at that flaw depth. The exclamation mark (!) represents the factorial operator. The likelihoods for each flaw depth were multiplied together to obtain an overall likelihood for the logistic curve. The β_1 and β_2 parameters were then updated, and this process was iterated until a logistic curve fit that provides the largest likelihood value for the data was achieved. In some

cases, this process can fail to converge if the data is not well represented by the shape of the logistic regression curve. ASTM E2862-818 (2018) recommends that convergence should be achieved within 20 iterations, otherwise, the logistic curve representation could be unreliable.

3.5 Pseudopoints

The analysis of PARENT and PINC data included the use of a pair of "pseudopoints" in the generation of POD curves. Pseudopoints are artificial data points that were inserted at the 0% TWD and 100% TWD locations, and each point had a value of 50%. Essentially, the inclusion of pseudopoints could be conceived of as modifying the initial assumption for all flaw sizes to 50% POD. The purpose of using pseudopoints in the analysis was to aid scenarios of small sample sizes, allowing realistic POD curves to converge with less data. As the sample size of actual data increased, the relative influence of the pseudopoints decreased.

As discussed in Section 1.3, the pseudopoints were carried over to the PARENT analysis (Meyer and Heasler 2017) and the generation of POD curves reported in Meyer and Holmes (2019) after they were used in the analysis of the much smaller dataset from the Quickblind study (Braatz et al. 2014). Section 5.0 analyzes whether the pseudopoints had a significant impact on the POD curves from PARENT reported in Meyer and Holmes (2019).

3.6 Limitations of the Logistic Regression Model

The use of the logistic regression model for POD representation does have limitations related to assumptions regarding the data to which it is fit. Fundamentally, the logistic regression model assumes the POD curve is monotonic and is a non-decreasing function of some variable (in this case, TWD). These are assumptions that may not be true in practice. For instance, the tip response of deep flaws can sometimes be detected without an associated corner response. In these cases, the tip responses can be misinterpreted as responses from manufacturing defects, resulting in missed detection of flaws of significant depth. It has been postulated that this type of scenario could lead to a POD that is not monotonic with flaw depth. The shape of the logistic expression is not compatible with a trend in the data of this nature and would not be able to represent it well.

3.7 Confidence Bounds to PARENT POD Curves

The logistic model of POD is typically displayed with upper and lower confidence bounds. These bounds were shown for plots of POD in PINC, PARENT, and MRP-262 Rev. 3 and are also included in Figure 2-1 through Figure 2-6. The POD curve for a specific scenario in PARENT (e.g., OD examinations of circumferential flaws in SBDMW test blocks) represents the mean POD for one sample. The confidence bounds define the range within which the mean POD for additional samples would be likely to occur. Ninety-five percent (95%) confidence bounds represent the range within which the average POD for 95 out of 100 samples should reside.

Confidence bounds on the POD expression in Equation 1.1 are calculated by first producing confidence bounds on the linear expression, $\hat{\beta}_0 + \hat{\beta}_1 \, TWD,$ and then evaluating the POD expression in Equation 1.1 at those bounds. For each flaw depth represented in the data, the lower and upper confidence bounds on the log-odds scale are computed from,

$$(l, u) = \hat{\beta}_0 + \hat{\beta}_1 \, TWD \pm SF \times z_{1-\alpha/2} \times se.fit(TWD), \quad \text{Equation 3.6}$$

where *l* and *u* are the lower and upper confidence bounds of the linear expression $\hat{\beta}_0 + \hat{\beta}_1 \, TWD$, *se.fit*(*TWD*) is the standard error of the linear expression, and $z_{1-\alpha/2}$ is the *z* score for the 95% confidence interval, which is 1.96. The variable "*SF*" is a scaling factor that is applied to account for deviation in the variance of the linear expression from the expected variance.

The confidence bounds for the POD curve are then obtained through transformation of the lower and upper confidence bounds for $\hat{\beta}_0 + \hat{\beta}_1 \, TWD$ by Equation 3.7,

$$\text{POD}(TWD)_{l,u} = \frac{1}{1+exp(-(l,u))}. \quad \text{Equation 3.7}$$

4.0 Influence of False Call Data

In Section 2.0, results of the POD analyses from PINC, PARENT, and MRP-262 Rev. 3 (EPRI 2017) are presented together for convenient comparison. There were significant differences, in general, in POD results from the three studies and potential reasons for the differences are outlined in 2.1. One significant difference in the way the studies were analyzed relates to the handling of false call data. In the analysis of PDI data documented in MRP-262 Rev. 3, false call rate data were excluded. In PINC and PARENT, false call rate data were handled as described in Section 3.2. However, in PINC, a blank grading unit of length 100 mm (L_{gu} = 100 mm) was used for analysis, while L_{gu} = 25 mm was used in the PARENT analysis. It can be anticipated that the smaller L_{gu} in PARENT would lead to smaller values of FCP in PARENT relative to PINC. However, it is difficult to anticipate how these analysis decisions affect the overall POD curves. Therefore, the influences of false call data and grading unit length assumptions on POD estimations are evaluated in this section by re-analyzing PARENT data using assumptions that are consistent with analysis performed for MRP-262 Rev. 3 and PINC.

4.1 PARENT Data Analysis Excluding False Call Data

The analysis of PARENT data is presented in this section. The PARENT data was analyzed by excluding false call data and comparing it to the MRP-262 Rev. 3 analyses to observe remaining differences in POD estimates and to judge the significance of the assumption on results. POD estimates are displayed in the form of plots for both circumferential and axial flaws for two cases: 1) OD examinations of SBDMWs and 2) ID examinations of LBDMWs.

4.1.1 OD Examinations of SBDMWs

Plots of POD estimations without false call data from PARENT and MRP-262 Rev. 3 are provided from OD examinations of SBDMWs for circumferential flaws and axial flaws in Figure 4-1 and Figure 4-2, respectively. PARENT POD plots created with false call data are included for comparison. It is evident that some significant difference remains between PARENT and MRP-262 Rev. 3 POD estimations even when false call data is excluded for the analysis of both datasets. It can also be concluded that the decision to exclude false call data (or not) has a significant influence on the resulting POD estimations. Not only did excluding false call data from POD analysis result in a higher estimate of FCP for PARENT, but the exclusion of false call data also widened the 95% confidence bounds for all portions of the curves, increasing the uncertainty.

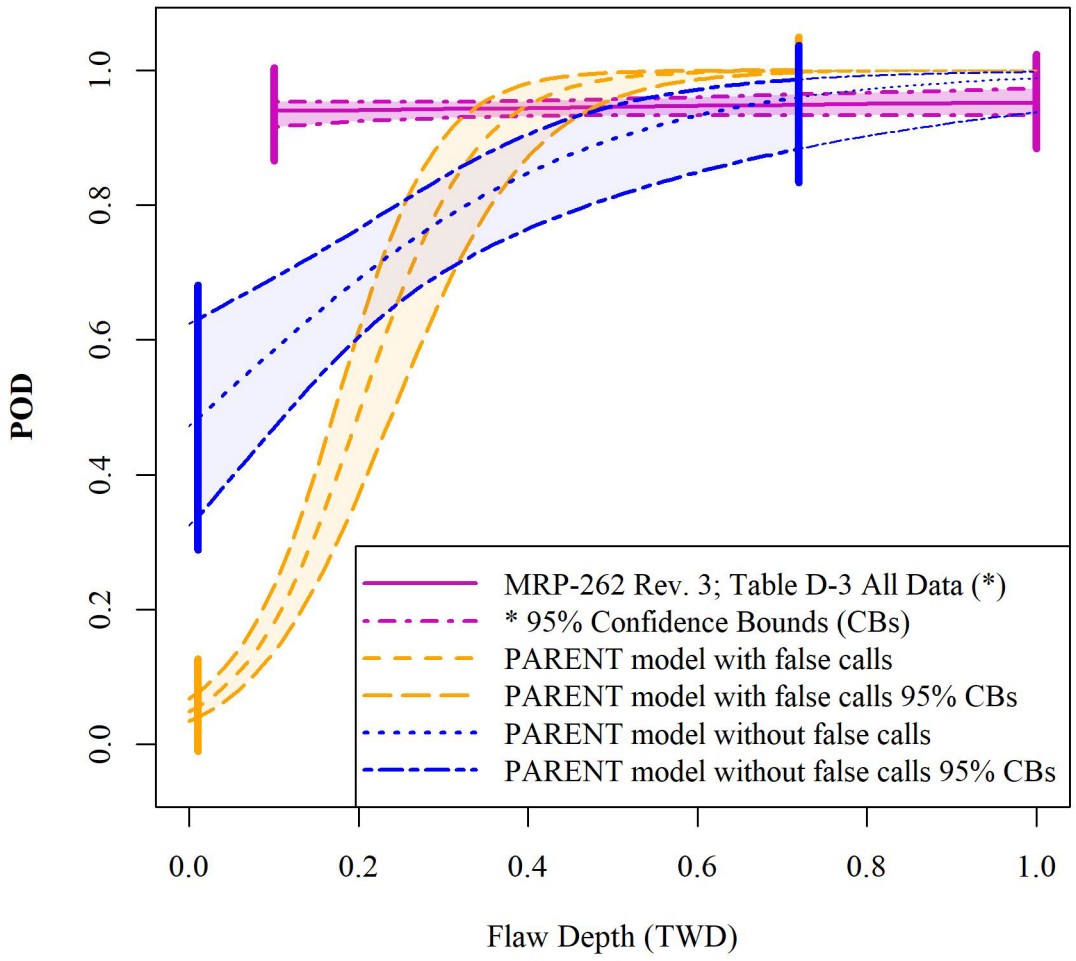

Figure 4-1 POD as a function of flaw depth (TWD) for circumferentially oriented flaws in SBDMW test blocks in MRP-262 Rev. 3 and for PARENT with and without false call data included (OD exams). The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

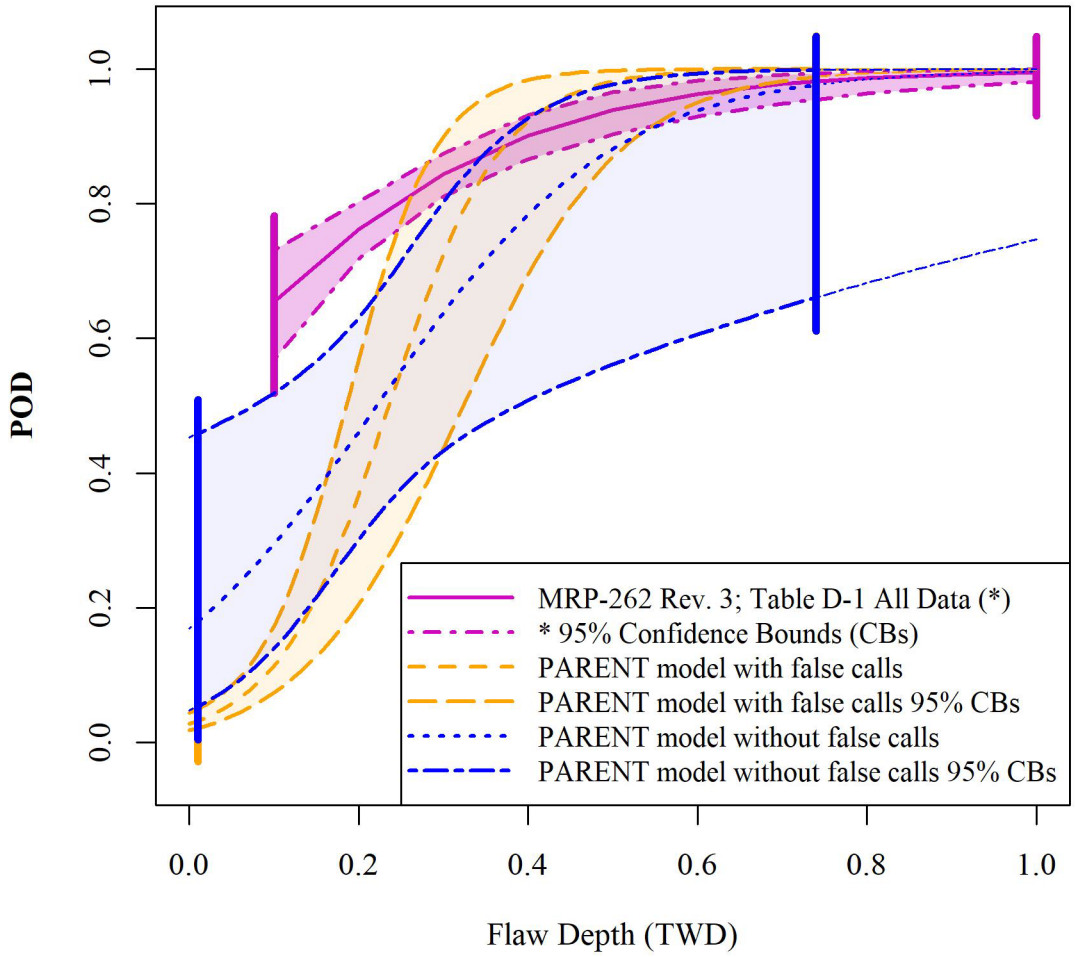

Figure 4-2 POD as a function of flaw depth (TWD) for axially oriented flaws in SBDMW test blocks in MRP-262 Rev. 3 and for PARENT with and without false call data included (OD exams). The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

4.1.2 ID examinations of LBDMWs

Plots of POD estimations without false call data from PARENT and MRP-262 Rev. 3 are provided from LBDMW exams performed with ID access for circumferential flaws and axial flaws in Figure 4-3 and Figure 4-4, respectively. PARENT POD plots created with false call data are, again, included for comparison. In this case, the POD estimation from MRP-262 Rev. 3 was within the confidence bounds of the POD estimations for PARENT (for > 0.1 TW) for circumferential flaws. However, estimations of POD for axial flaws exhibit greater differences. Like the case for SBDMW examinations by OD access, the exclusion of false call data from PARENT raises the small flaw depth portion of the curve and increases the uncertainty in the POD estimations.

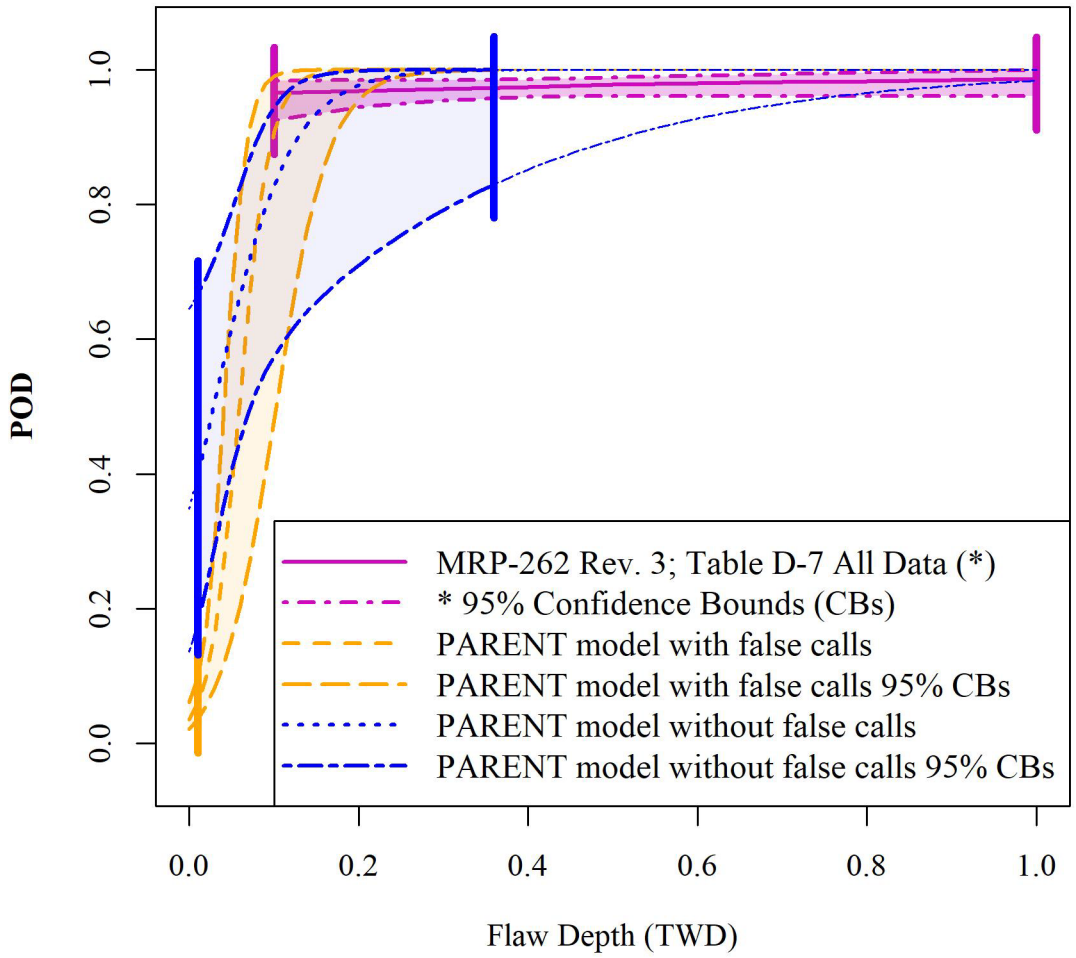

Figure 4-3 POD as a function of flaw depth (TWD) for circumferentially oriented flaws in LBDMW test blocks in MRP-262 Rev. 3 and for PARENT with and without false call data included (ID exams). The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

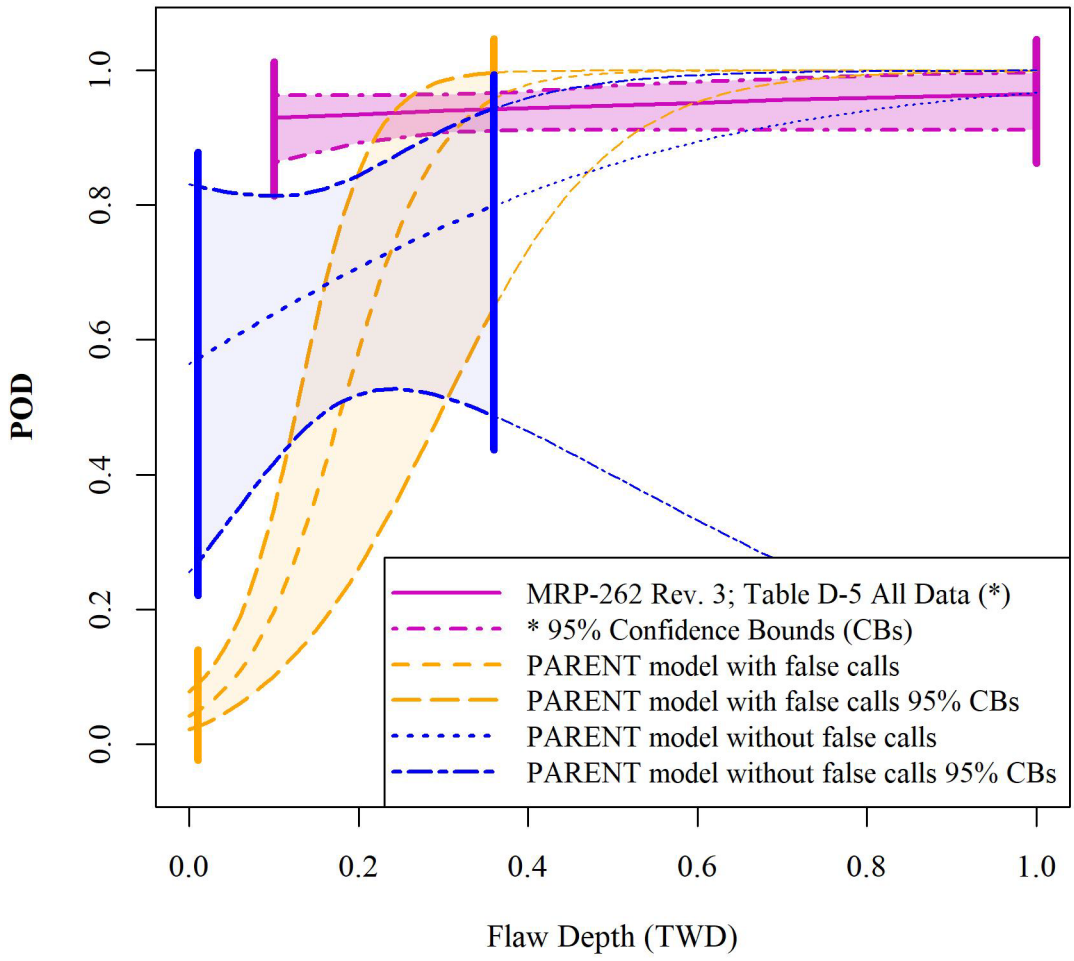

Figure 4-4 POD as a function of flaw depth (TWD) for axially oriented flaws in LBDMW test blocks in MRP-262 Rev. 3 and for PARENT with and without false call data included (ID exams). The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

4.2 PARENT Data Analysis Assuming 100 mm Blank Grading Unit Length

In this section, PARENT data was analyzed by setting L_{gu} = 100 mm to be consistent with the analysis of data from PINC to judge the significance of the assumption on results. POD estimates are displayed for both circumferential and axial flaws for SBDMW exams performed with OD access.

Plots of POD estimations from PARENT and PINC are provided from SBDMW exams performed with OD access for circumferential flaws and axial flaws in Figure 4-5 and Figure 4-6, respectively. It appears that PINC and PARENT estimations for POD agreed very well for

circumferential flaws when both datasets were analyzed with L_{gu} = 100 mm. For axial flaws, the adjustment of L_{gu} for PARENT appears to reduce the difference in POD estimations but not completely. It can be concluded that assumptions regarding L_{gu} had a significant influence on the resulting POD estimations.

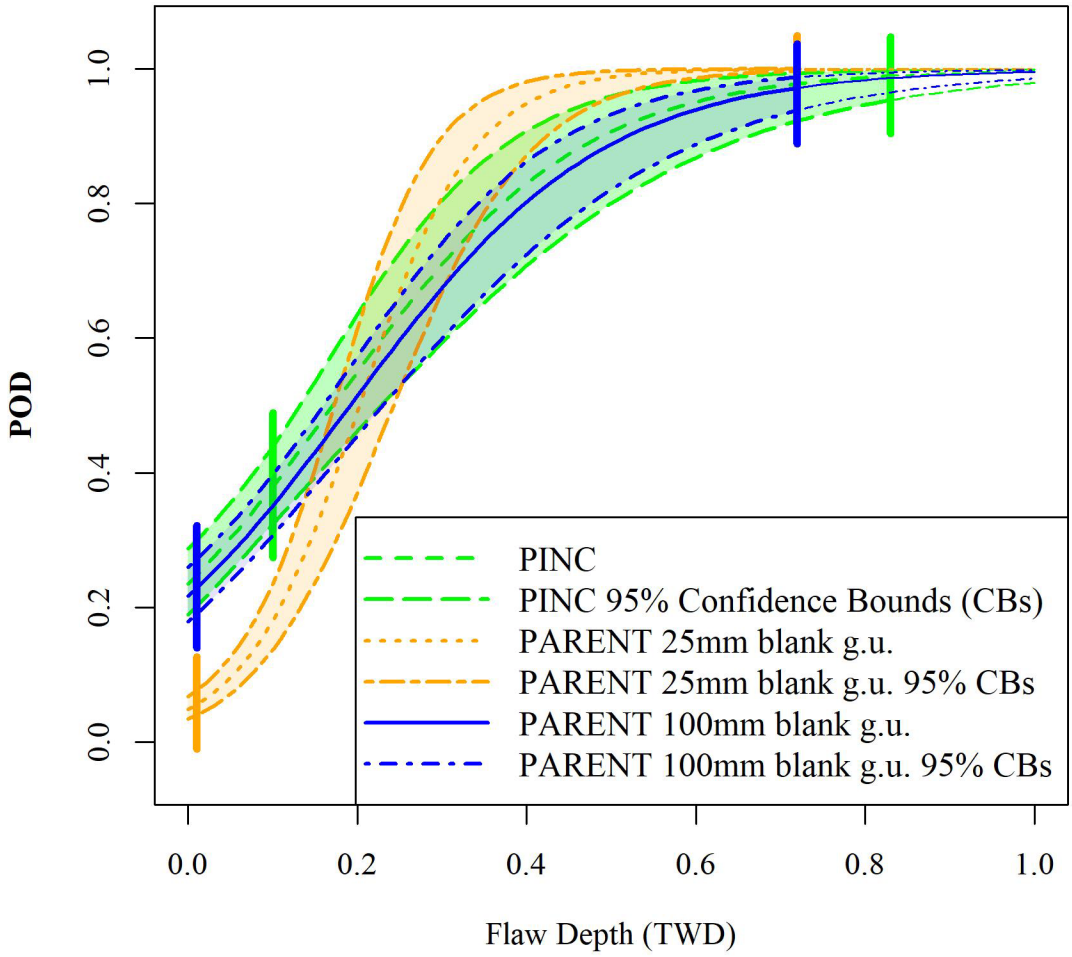

Figure 4-5 POD as a function of flaw depth (TWD) for circumferentially oriented flaws in SBDMW test blocks in PINC and for PARENT with L_{gu} = 100 mm (OD exams). The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

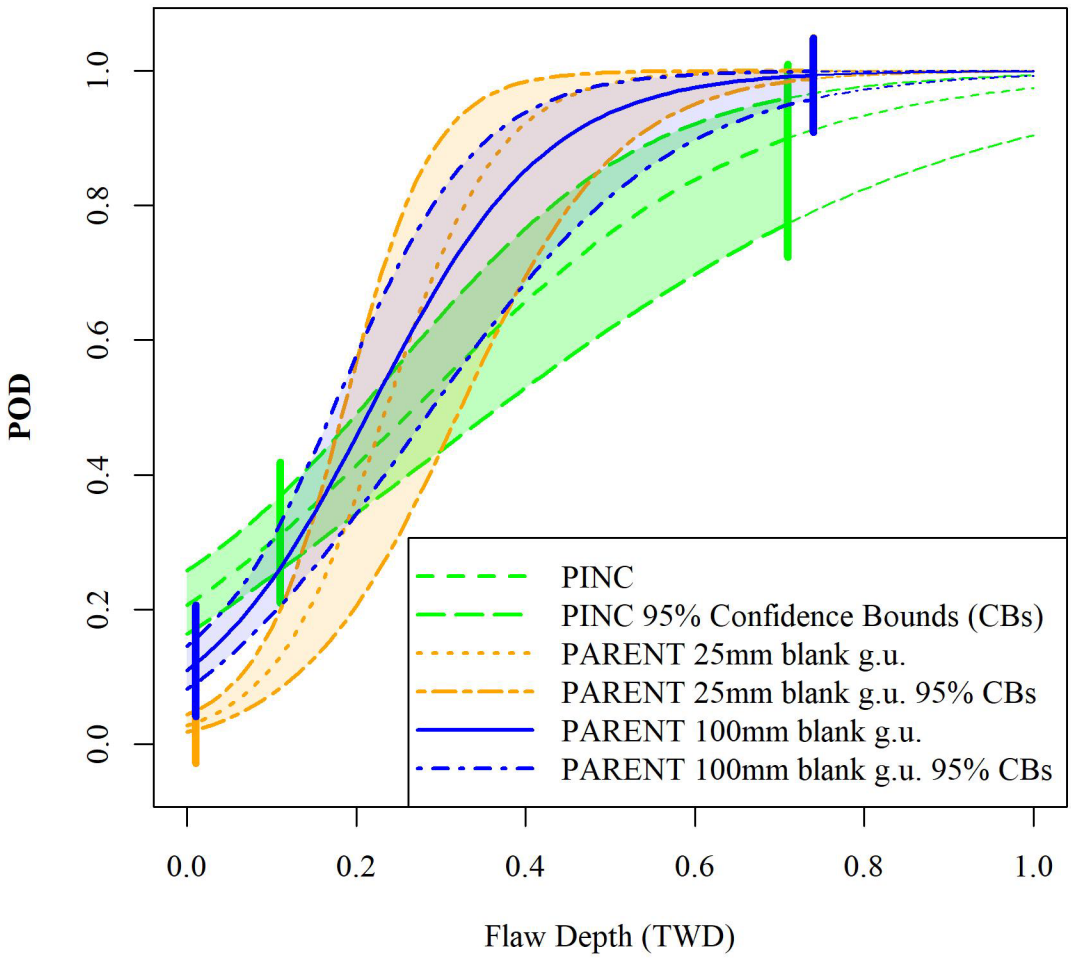

Figure 4-6 POD as a function of flaw depth (TWD) for axially oriented flaws in SBDMW test blocks in PINC and for PARENT with L_{gu} = 100 mm (OD exams). The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

5.0 Influence of Pseudopoints

As noted in Section 3.5, pseudopoints were used in the analysis of PARENT blind test data. Pseudopoints refer to artificial data points added to POD datasets to assist with some aspect of analysis. Pseudopoints were included to assist in analysis of quick blind testing (Braatz et al. 2014) and carried over to the analysis of the full blind test data in PARENT (Meyer and Heasler 2017). Pseudopoints with POD values of 50% were included at 0% TWD and the maximum flaw depth. Adding these pseudopoints biased the initial assumption of POD to random chance. As actual data points were added, the curve was updated from that initial starting point. With increasing amounts of real data, the influence of the pseudopoints on the resulting POD curves decreased. To understand how pseudopoints may have affected the results of the analysis of PARENT blind test data, the data was reanalyzed without the use of pseudopoints and compared to the results with pseudopoints.

The comparisons are presented as plots of the POD curves generated with and without pseudopoints in Figure 5-1 through Figure 5-6. The plots in Figure 5-1 and Figure 5-2 show the results for OD examinations of SBDMW test blocks for circumferential and axial flaws, respectively; the plots in Figure 5-3 and Figure 5-4 show results for OD examinations of LBDMW test blocks for circumferential and axial flaws, respectively; and the plots in Figure 5-5 and Figure 5-6 show results for ID examinations of LBDMW test blocks for circumferential and axial flaws, respectively.

For most of these figures, it appears that the pseudopoints have a minor influence on the resulting POD curves. A substantial influence is shown in Figure 5-5 for ID examinations of LBDMW test blocks for circumferential flaws. It is notable that the pseudopoints did not have a large influence for examinations for axial flaws (OD or ID) in LBDMW test blocks (Figure 5-4 and Figure 5-6) because the NOBS for axial flaws was significantly lower compared to circumferential flaws (refer to Table 1-2). Further, the impact of pseudopoints on the curve is also minor for OD examinations for the circumferential flaw population in LBDMWs (Figure 5-3). Therefore, it seems that the high detection rate for ID examinations of LBDMW test blocks for circumferential flaws contributes to the pseudopoints having greater effect.

The influence of pseudopoints in Figure 5-2 is also noticeable but not as substantial as in Figure 5-5. It does not appear that NOBS or a high rate of detection can explain a greater influence of pseudopoints in Figure 5-2.

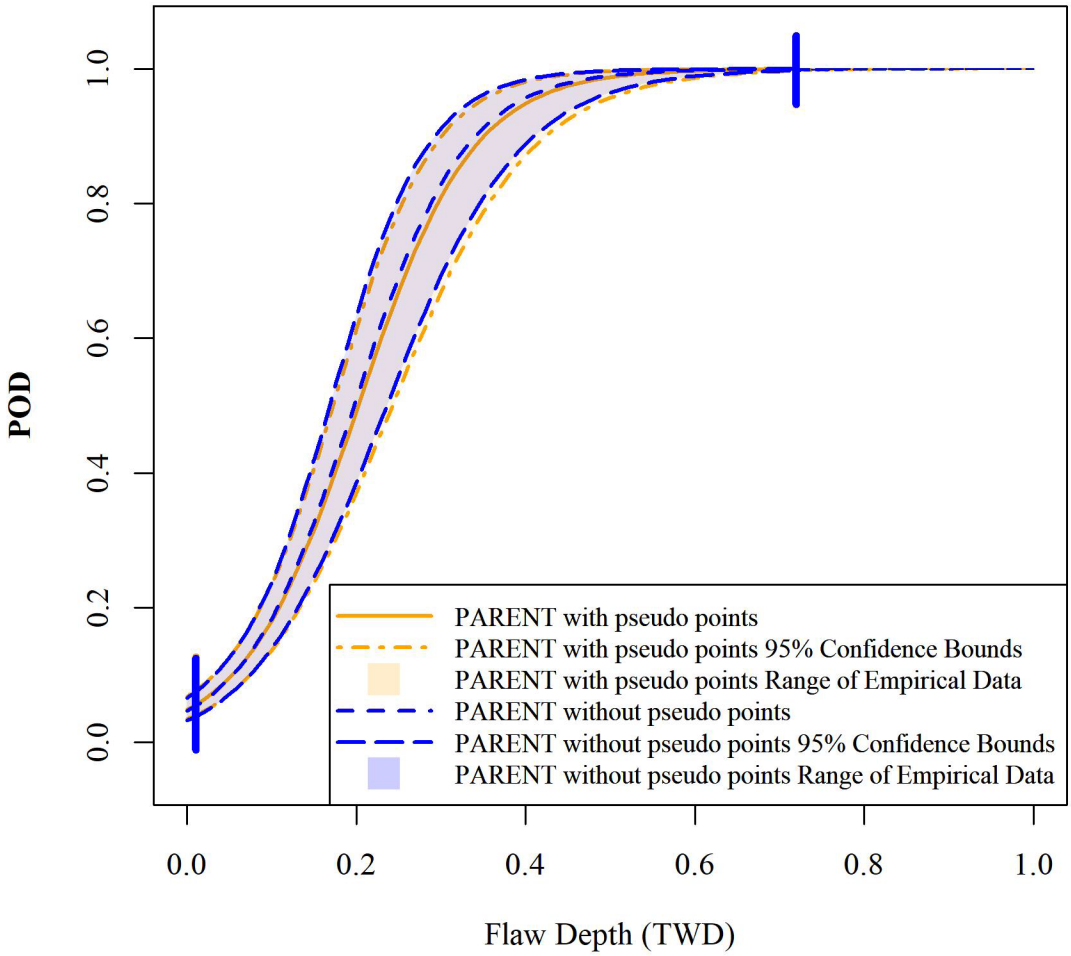

Figure 5-1 POD curves for SBDMW test blocks for circumferential flaws from PARENT (OD exams) for curves generated with and without pseudopoints. The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

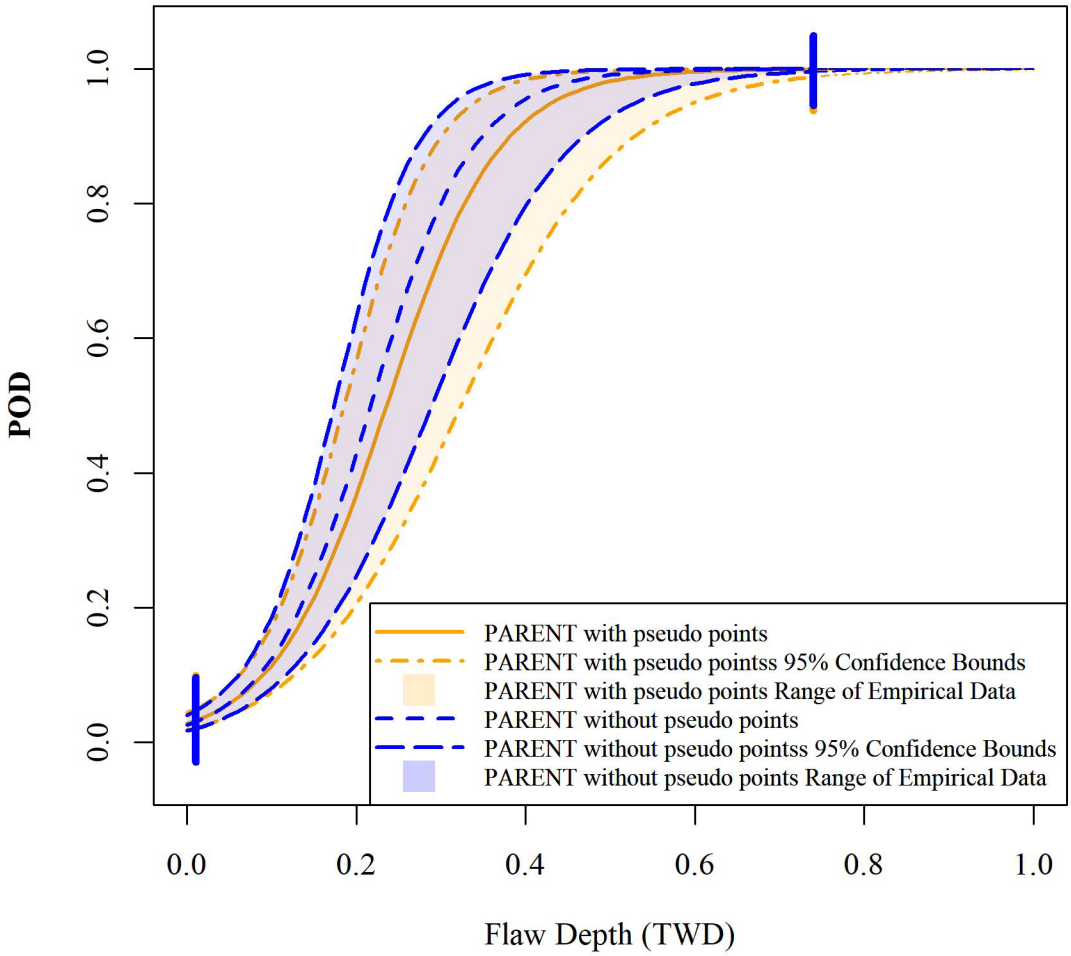

Figure 5-2 POD curves for SBDMW test blocks for axial flaws from PARENT (OD exams) blind testing for curves generated with and without pseudopoints. The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

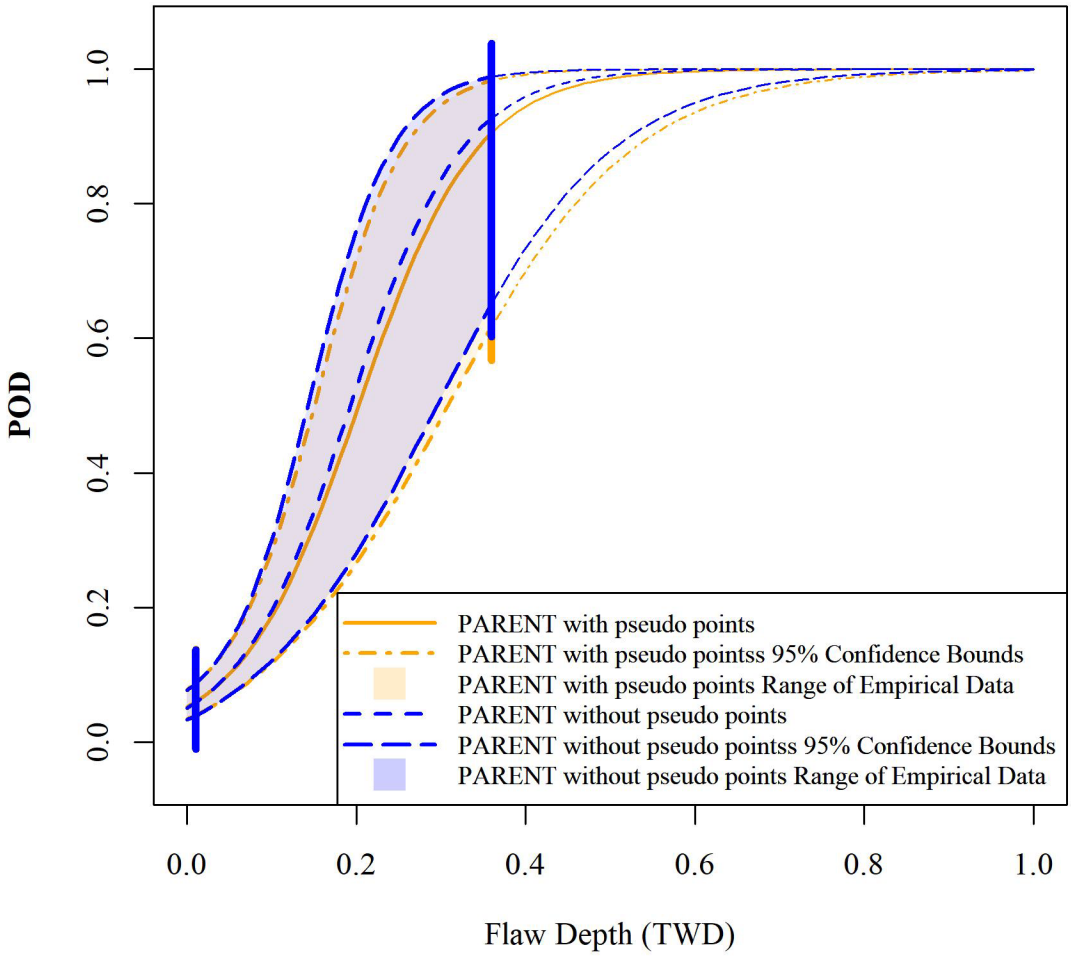

Figure 5-3 POD curves for LBDMW test blocks for circumferential flaws from PARENT (OD exams) for curves generated with and without pseudopoints. The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

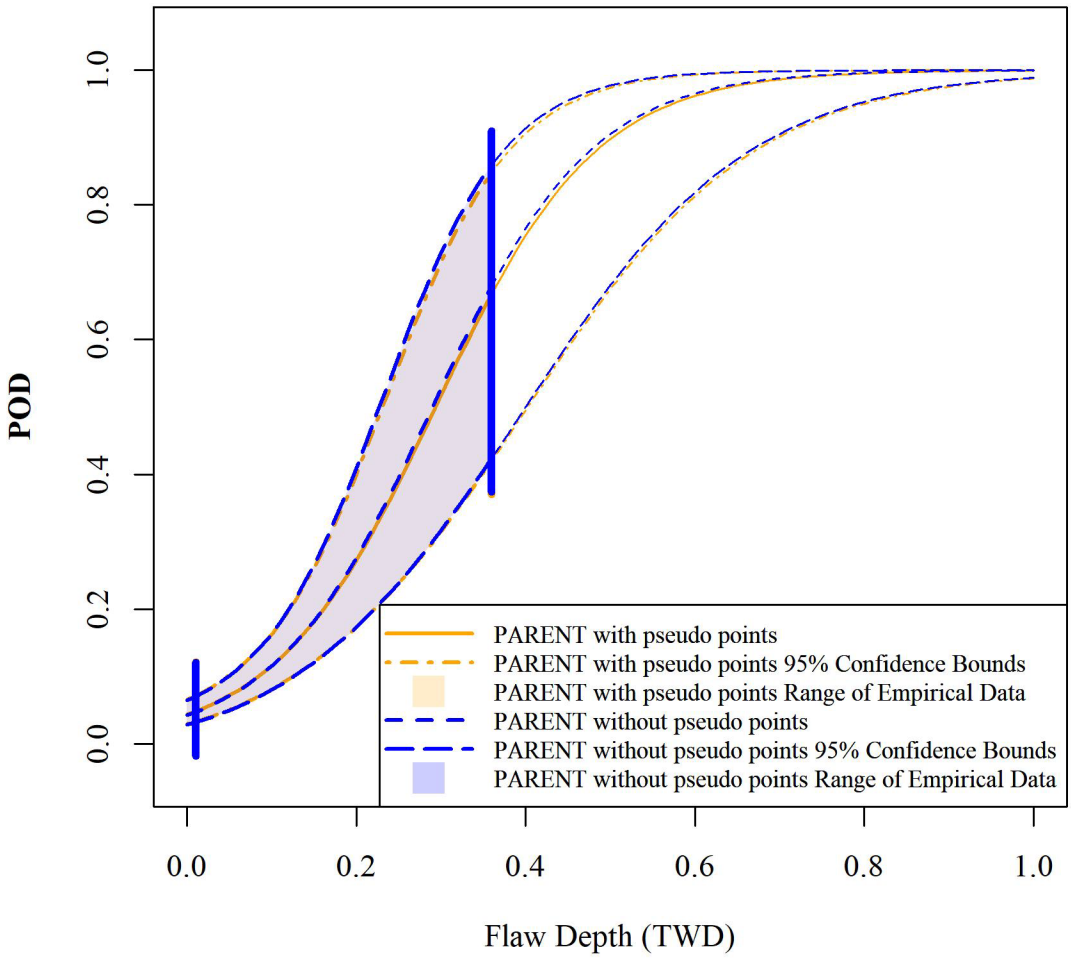

Figure 5-4 POD curves for LBDMW test blocks for axial flaws from PARENT (OD exams) for curves generated with and without pseudopoints. The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

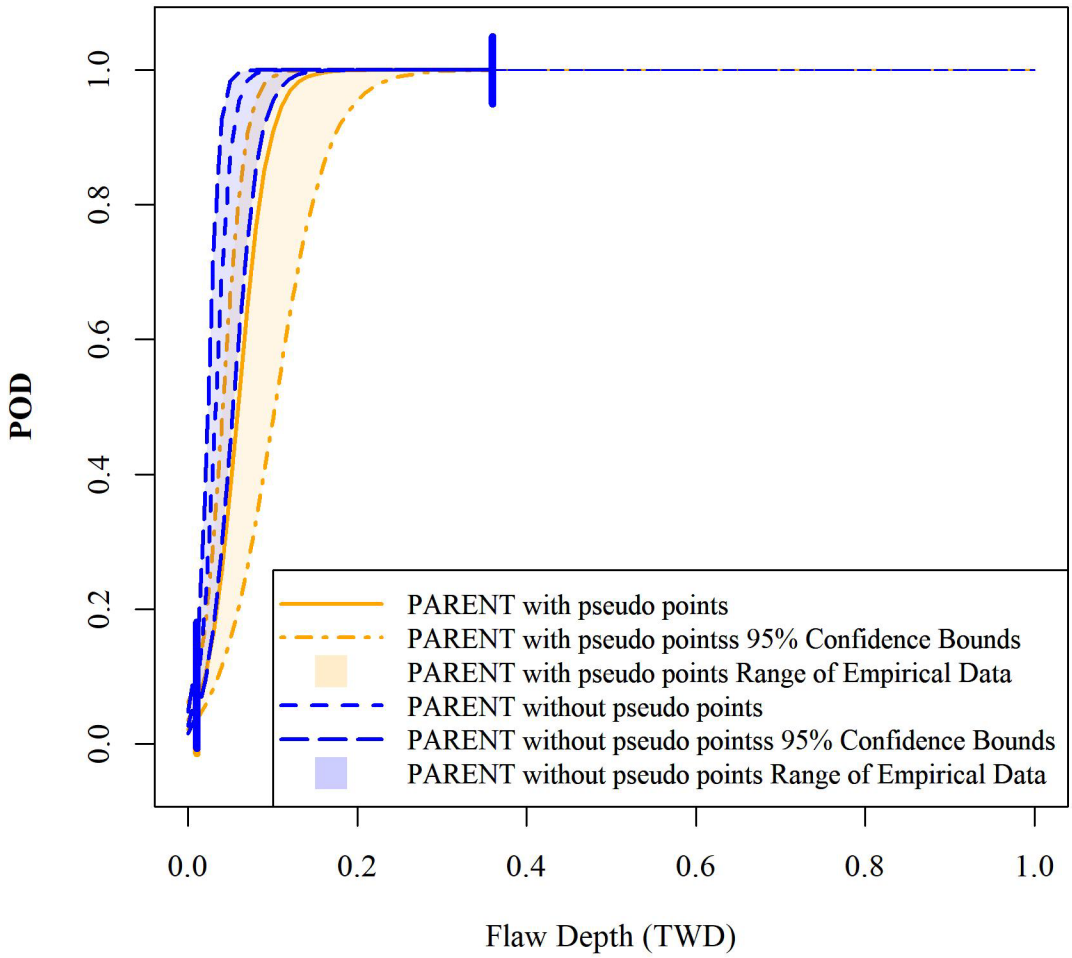

Figure 5-5 POD curves for LBDMW test blocks for circumferential flaws from PARENT (ID exams) for curves generated with and without pseudopoints. The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

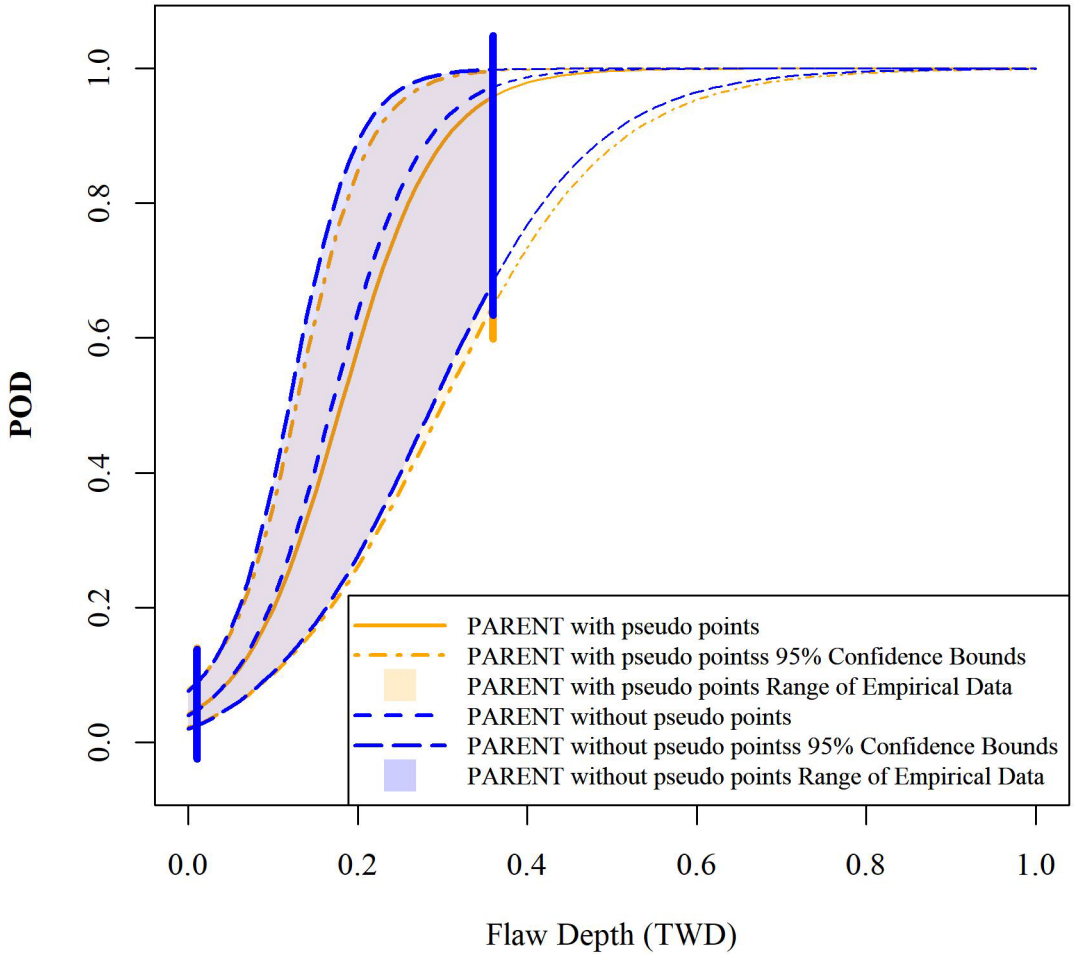

Figure 5-6 POD curves for LBDMW test blocks for axial flaws from PARENT (ID exams) for curves generated with and without pseudopoints. The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

6.0 Standards for POD Analysis

Generic standards for performing POD studies and analyzing data from POD studies have been developed to support the quality of results. Satisfying all the guidelines of these standards can be challenging for NPP applications because acquiring or fabricating suitable test blocks can be cost prohibitive and the limiting factor in designing studies. In this section, the analysis of PARENT data is reviewed against the guidelines in two existing standards to highlight gaps.

The following two standards are reviewed in this section:

- ASTM E2862-18: *Standard Practice for Probability of Detection Analysis for Hit/Miss Data* (ASTM E2862-818 2018)
- MIL-HDBK-1823A: *Nondestructive Evaluation System Reliability Assessment* (MIL-HDBK-1823A 2009)

Of these two standards, ASTM E2862-18 is more narrowly focused on the analysis of hit/miss binary data. MIL-HDBK-1823A applies more broadly to POD study "lifecycle" and includes guidelines on study design through POD data analysis. In this section, a selection of guidelines from ASTM E2862-18 and MIL-HDBK-1823A are highlighted and reviewed with respect to data analysis performed for PARENT. The highlighted guidelines are summarized in Table 6-1.

Table 6-1 Highlighted guidelines from ASTM E2862-18 and MIL-HDBK-1823A.

MIL-HDBK-1823A: 4.5.2.2 Target sizes and number of "flawed" and "unflawed" inspection sites (MIL-HDBK-1823A 2009)
4.5.2.2 a.(1) "Given that $a_{90/95}$ has become a de facto design criterion it may be more important to estimate the 90th percentile more precisely than lower parts of the curve."
4.5.2.2 a.(3) "If it is possible to estimate in advance the specific region where the POD(a) function rises rapidly, then it is advantageous to concentrate more targets in that size range. It should be noted that there is a tendency to include too many 'large' targets in NDE reliability demonstrations, as a result of the difficulties in producing small targets in specimens."
4.5.2.2 b. "To provide reasonable precision in the estimates of the POD(a) function, experience suggests that the specimen test set contain at least 60 targeted sites if the system provides only a binary, hit/miss response and at least 40 targeted sites if the system provides a quantitative target response, â. These numbers are minimums. For binary responses, 120 inspection opportunities will result in a significantly more precise estimate of a_{50}, and thus a smaller value for $a_{90/95}$."
ASTM E2862-18: Standard Practice for Probability of Detection Analysis for Hit/Miss Data (ASTM E2862-818 2018)
(6.4.3) Outlier data should be identified and investigated.
(6.4.4) POD cannot be modeled as a continuous function of discontinuity size if there is a complete separation of misses and hits as crack size increases.
(6.4.5) POD cannot be modeled as a continuous function of discontinuity size if all the discontinuities are found or if all the discontinuities are missed.
(6.7) False call data shall not be included in the development of the generalized linear model.
(6.9) After running the analysis, the analyst shall verify that convergence has been achieved.
(6.11.1) If more than 20 iterations are required for convergence, the model may not be reliable.
(6.10) The analyst shall assess the significance of the predictor variable in the regression model. Only significant variables shall be included in the regression model.
(6.11.2) The analyst shall visually assess the shape of the POD curve.
(6.11.3) The analyst shall visually assess how well the POD curve fits the data by comparing the range over which the POD curve is rising with the range over which misses begin to overlap with and transition to hits as the size increases.
(6.11.4) The analyst shall compare the empirical POD curve to the generalized linear model.

6.1 MIL-HDBK-1823A Guideline Highlights

Guideline 4.5.2.2 a.(1) from MIL-HDBK-1823A is stated in the first row of Table 6-1. In this discussion, the $a_{90/95}$ metric relates to the flaw depth (TWD) and represents TWD at which the lower 95% confidence boundary of a POD curve crosses the POD = 90% level. Thus, it is the TWD at which the POD is equal to or greater than 90% with 95% confidence. Although the $a_{90/95}$ metric has become the de facto design criterion for other industries, it has not been emphasized in POD studies for NPP components.

Guideline 4.5.2.2 a.(1) states that the high POD region (larger flaw depths) of a POD curve may be more important than the low POD region (smaller flaw depths) of a POD curve given emphasis placed on the $a_{90/95}$ metric. For some of scenarios analyzed in the studies in PINC, PARENT, and MRP-262 Rev. 3, the $a_{90/95}$ level of performance is not reached within the range

of empirical data. Further, the $a_{90/95}$ is arbitrary (ENIQ 2010) and this metric alone may not provide a meaningful characterization of NDE performance for many NPP component applications. Nevertheless, the small TWD regions of POD curves are most difficult to estimate and analysis to determine the relative importance of the lower POD region of a POD curve with respect to the higher POD curve, in the broader context of NPP safety, could help to clarify the need for better estimations of the lower POD regions of curves.

The second row of Table 6-1 states guideline 4.5.2.2 a.(3) from MIL-HDBK-1823A, which relates to the depth distribution of flaws that should be implanted in test blocks. This guideline is met with PARENT testing; Figure 5-1 through Figure 5-6 show that flaw depths in PARENT typically spanned the range where the POD curve rose rapidly, and the flaw depth distribution was not dominated by flaws in the saturated portion of the POD curve. Although PARENT included flaws as small as 1% TWD, only a handful of flaws were below 10% TWD.

Finally, guideline 4.5.2.2 b, stated in the third row of Table 6-1, recommends a minimum of 60 unique flaws for binary hit/miss response data. The number of unique flaws for PARENT testing cases in Figure 5-1 through Figure 5-6 range from approximately 5 for axial flaws to 20 for circumferential flaws.

The limited number of flaws in the < 10% TWD flaw depth range and the limited number of unique flaws, overall, motivated the inclusion of false call data in PARENT analysis and contributed to difficulties in characterizing POD at the lower flaw depth range.

6.2 ASTM E2862-18 Guideline Highlights

Several ASTM E2862-18 guidelines for the analysis of binary hit/miss POD data are also included in Table 6-1. The first three guidelines stated in the table (6.4.3, 6.4.4, and 6.4.5) were met in PARENT, while the fourth guideline (6.7) was not met, as false call data were used in the development of the logistic curve fits in PARENT.

The PARENT analysis was evaluated to determine if the next three ASTM E2862-18 guidelines in Table 6-1 were met (i.e., 6.9, 6.11.1, and 6.10). A summary of this analysis is provided in Table 6-2 for two analysis scenarios: 1) PARENT data was analyzed using false call data with the 25 mm blank grading unit size, and 2) PARENT data was analyzed while excluding the false call data. Both scenarios were considered because guideline 6.7 of ASTM E2862-18 states that false call data should not be used in the development of the generalized linear model and because false call data can significantly affect the resulting POD curves, as shown in Section 4.0. From Table 6-2, it is apparent that models converge within the required number of iterations for all cases, satisfying 6.9 and 6.11.1. In addition, the predictor variable (i.e., flaw depth) is significant for all the models except for ID examinations for axial flaws in LBDMW test blocks when false calls are excluded from the model development.

The next two ASTM E2862-18 guidelines in Table 6-1 recommend a visual assessment of the POD curves and the POD data, (i.e., 6.11.2 and 6.11.3), while the last guideline recommends a comparison of the logistic curve fit with an empirically derived POD (6.11.4). No further guidance is provided in 6.11.2 regarding visual assessments of the shapes of POD curves, but it can be noted that the POD curves in Figure 5-1 through Figure 5-6 do exhibit the expected S-shape and the exhibited expected relationships between flaw depth and POD.

Table 6-2 Evaluation of PARENT data analysis with respect to meeting ASTM E2862-18 guidelines 6.9, 6.11.1, and 6.10.

	Convergence?	Number of iterations	Significance of predictor variable
SBDMW OD Circ. (25 mm gu)	Yes	6	p-value = 0
SBDMW OD Ax. (25 mm gu)	Yes	5	p-value = 5.40e-11
LBDMW OD Circ. (25 mm gu)	Yes	5	p-value = 9.63e-08
LBDMW OD Ax. (25 mm gu)	Yes	5	p-value = 9.24e-11
LBDMW ID Circ. (25 mm gu)	Yes	7	p-value = 1.09e-05
LBDMW ID Ax. (25 mm gu)	Yes	5	p-value = 6.09e-06
SBDMW OD Circ. (no false calls)	Yes	5	p-value = 4.81e-05
SBDMW OD Ax. (no false calls)	Yes	5	p-value = 1.33e-02
LBDMW OD Circ. (no false calls)	Yes	4	p-value = 1.05e-02
LBDMW OD Ax. (no false calls)	Yes	4	p-value = 2.50e-03
LBDMW ID Circ. (no false calls)	Yes	7	p-value = 1.72e-02
LBDMW ID Ax. (no false calls)	Yes	4	p-value = 3.33e-01 Not significant

6.3 Comparisons of Empirical PODs

Empircal PODs are compared to POD curves from PARENT in this section in accordance with ASTM E2862-18 guideline 6.11.4. Empirical PODs are computed as the average POD over 10% TWD flaw depth intervals from 0% TWD to the maximum flaw depth. One exception is the empirical POD computed for OD examinations for axial flaws in SBDMW test blocks. In this case, the 10% TWD – 20% TWD interval is split into two 5% TWD intervals, and there is no data in the 20% TWD to 30% TWD range and the 40% TWD to 70% TWD range. The tabular summaries (Table 6-3 through Table 6-8) of empirical POD data indicate the NOBS and NDET for each flaw size interval. Interval notation is used in these tables to denote that the endpoint is excluded “( )” or included “[ ]” within the range. The POD estimate is obtained by dividing NDET by NOBS, and the lower and upper confidence bounds for the estimate are also included as columns in the tables.

Guideline 6.7 of ASTM E2862-18 recommends that false call data should not be used in the fitting of the continuous logistic curve to POD data. In both PINC and PARENT, the false call data was used for fitting these curves, and Section 4.0 shows how the decision to employ false call data affected the results. Comparisons of empirical PODs with POD curves for PARENT can be made by comparing the plots in Figure 6-1 through Figure 6-6. These plots display empirical PODs with POD curves from PARENT for two scenarios: 1) in which the false call data was included in the generation of the POD curve, and 2) the false call data was excluded from generation of the POD curve.

Agreement between the empirical POD and the POD curves for PARENT was assessed visually. Agreement is assessed primarily by observing how close individual empirical POD data points (black diamonds) are to the middle lines (average value) of the POD curves in Figure 6-1 through Figure 6-6. Based on this approach, the agreement between the empirical PODs and POD curves with and without false call data are similar for most of the scenarios. Exceptions include scenarios shown in Figure 6-1 (OD examinations for circumferential flaws in SBDMW

test blocks) and Figure 6-6 (ID examinations for axial flaws in LBDMW test blocks) in which the POD curves generated without false call data appear to be more consistent with the empirical POD. However, the widths of the confidence intervals on the empirical POD data points and on the POD curves generated without false calls are relatively large. Unfortunately, the data limitations in PARENT that motivated the use of false call data also cause large confidence bounds in calculation of empirical POD. This inhibits the usefulness of empirical POD for determining if POD curves generated with or without false call data provide a better representation.

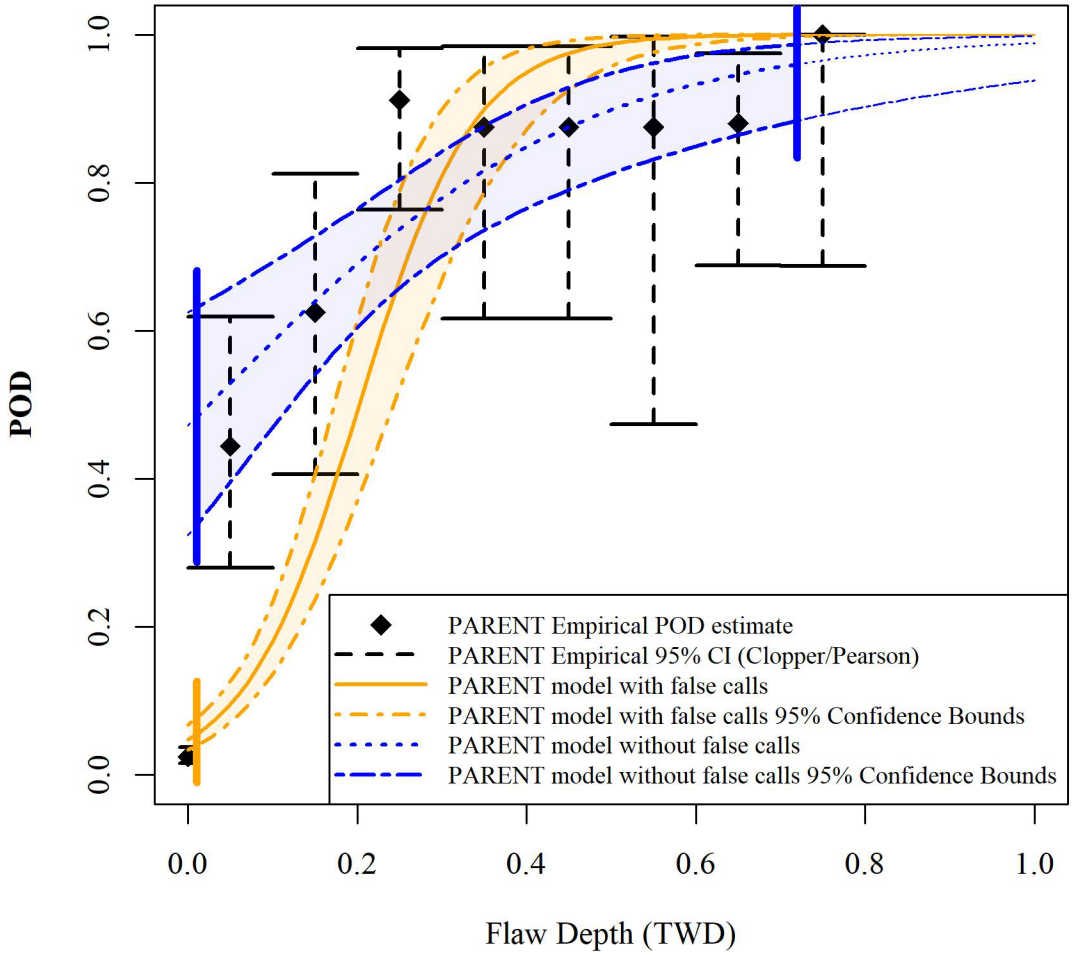

Figure 6-1 Empirical POD with POD curves for SBDMW test blocks for circumferential flaws from PARENT with and without false call data (OD exams). The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

Table 6-3 Empirical POD for SBDMW test blocks or circumferential flaws from PARENT (OD exams).

Interval	NDET	NOBS	POD	POD low	POD high
0	21	844	0.025	0.015	0.038
(0, 0.1]	16	36	0.44	0.28	0.62
(0.1, 0.2]	15	24	0.625	0.41	0.81
(0.2, 0.3]	31	34	0.91	0.76	0.98
(0.3, 0.4]	14	16	0.875	0.62	0.98
(0.4, 0.5]	14	16	0.875	0.62	0.98
(0.5, 0.6]	7	8	0.875	0.47	1.0
(0.6, 0.7]	22	25	0.88	0.69	0.97
(0.7, 0.8]	8	8	1.0	0.69	1.0

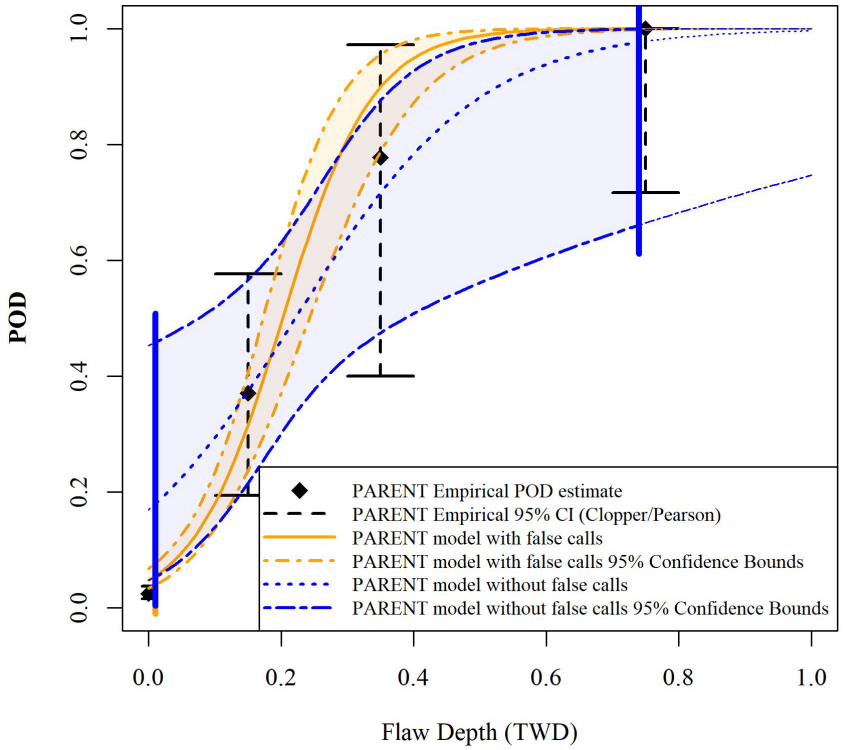

Figure 6-2 Empirical POD with POD curves for SBDMW test blocks for axial flaws from PARENT with and without false call data (OD exams). The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

Table 6-4 Empirical POD for SBDMW test blocks for axial flaws from PARENT (OD exams).

Interval	NDET	NOBS	POD	POD low	POD high
0	21	844	0.025	0.015	0.038
(0.1, 0.15]	2	9	0.22	0.028	0.60
(0.15, 0.2]	8	18	0.43	0.22	0.69
(0.3, 0.4]	7	9	0.78	0.40	0.97
(0.7, 0.8]	9	9	1.0	0.72	1.0

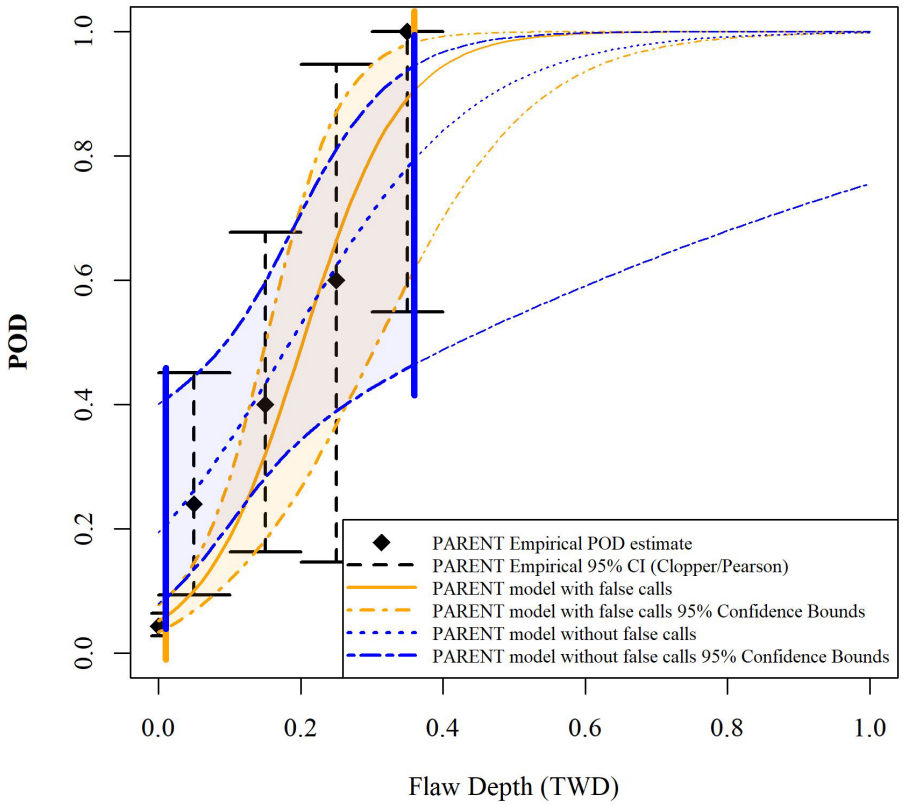

Figure 6-3 Empirical POD with POD curves for LBDMW test blocks for circumferential flaws from PARENT with and without false call data (OD exams). The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

Table 6-5 Empirical POD for LBDMW test blocks for circumferential flaws from PARENT (OD exams).

Interval	NDET	NOBS	POD	POD low	POD high
0	23	530	0.043	0.028	0.064
(0, 0.1]	6	25	0.24	0.094	0.45
(0.1, 0.2]	6	15	0.40	0.16	0.68
(0.2, 0.3]	3	5	0.60	0.15	0.95
(0.3, 0.4]	5	5	1.0	0.55	1.0

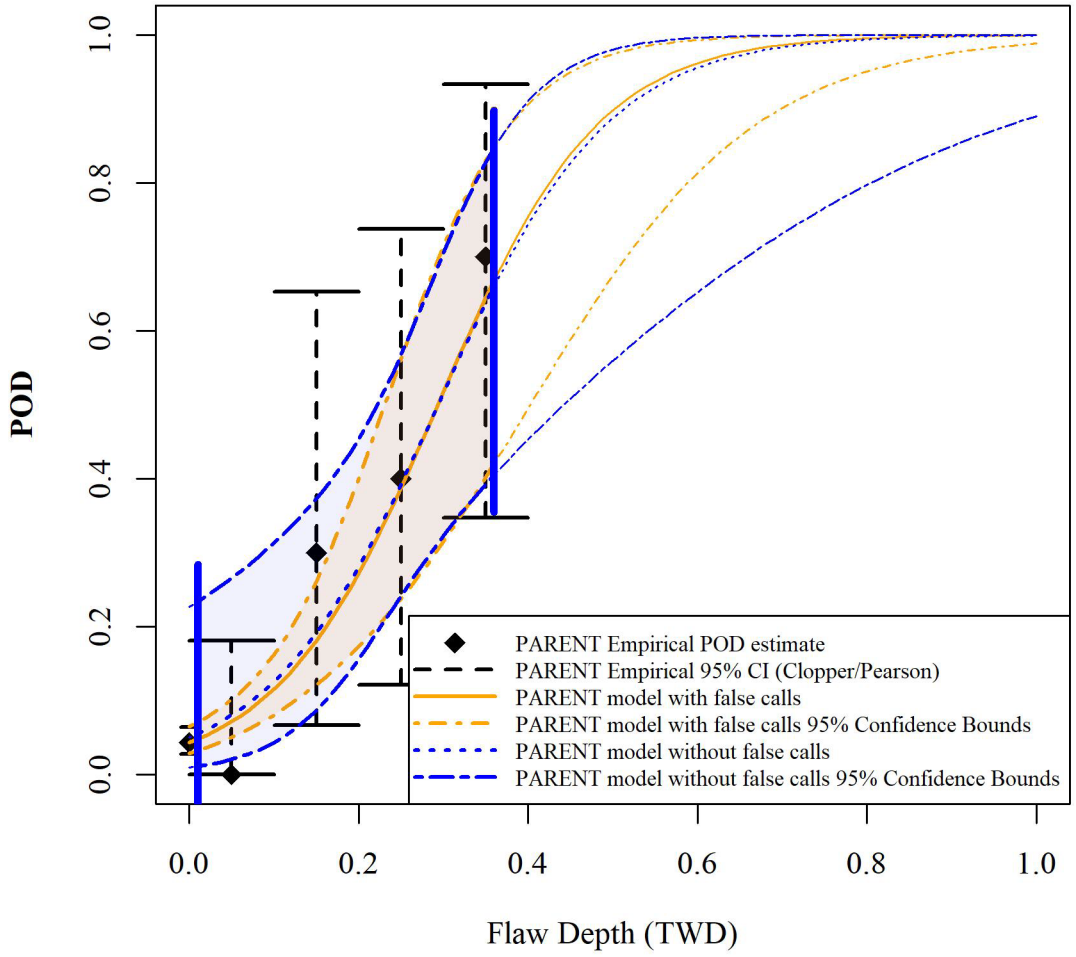

Figure 6-4 Empirical POD with POD curves for LBDMW test blocks for axial flaws from PARENT with and without false call data (OD exams). The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

Table 6-6 Empirical POD for LBDMW test blocks for axial flaws from PARENT (OD exams).

Interval	NDET	NOBS	POD	POD low	POD high
0	23	530	0.043	0.028	0.064
(0, 0.1]	0	15	0.0	0.0	0.18
(0.1, 0.2]	3	10	0.3	0.067	0.65
(0.2, 0.3]	4	10	0.4	0.12	0.74
(0.3, 0.4]	7	10	0.7	0.35	0.93

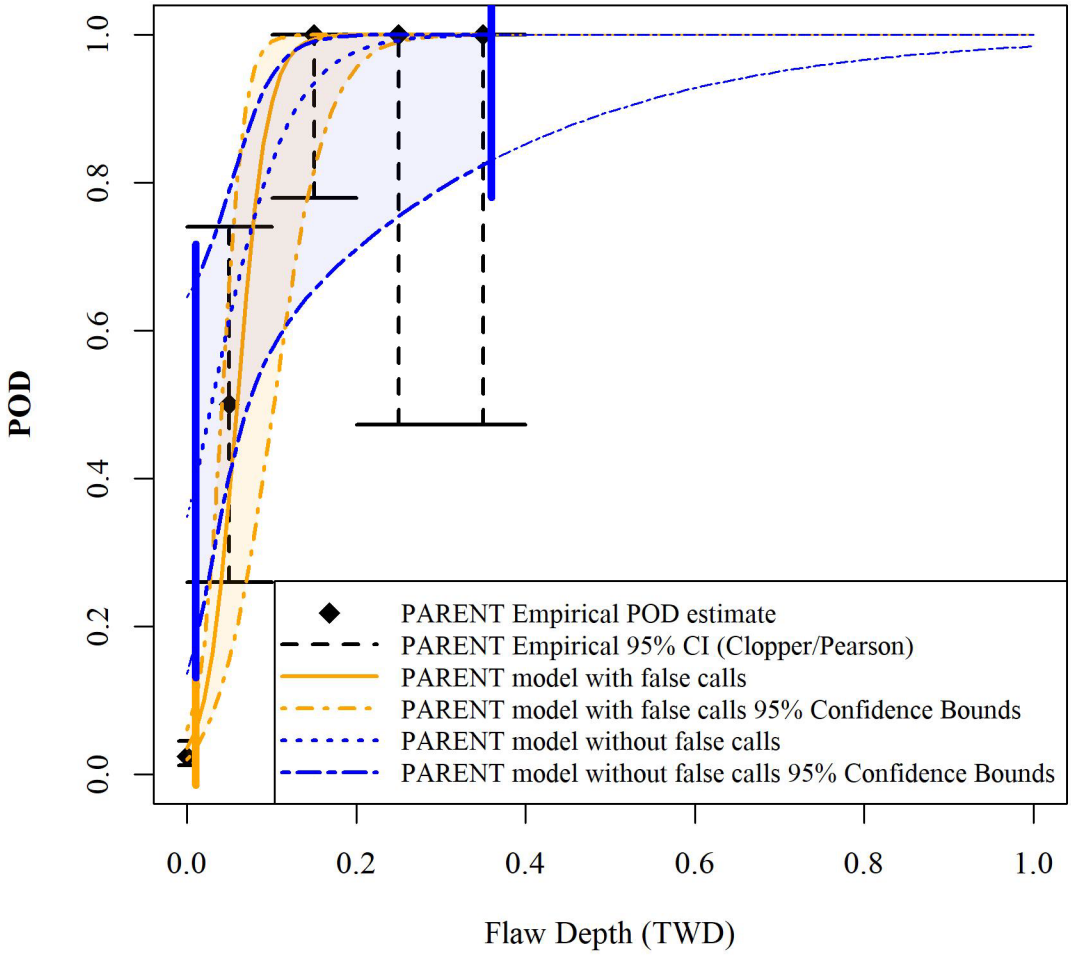

Figure 6-5 Empirical POD with POD curves for LBDMW test blocks for circumferential flaws from PARENT blind testing with and without false call data (ID exams). The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

Table 6-7 Empirical POD for LBDMW test blocks for circumferential flaws from PARENT (ID exams).

Interval	NDET	NOBS	POD	POD low	POD high
0	10	404	0.025	0.012	0.045
(0, 0.1]	9	18	0.5	0.26	0.74
(0.1, 0.2]	12	12	1.0	0.78	1.0
(0.2, 0.3]	4	4	1.0	0.478	1.0
(0.3, 0.4]	4	4	1.0	0.478	1.0

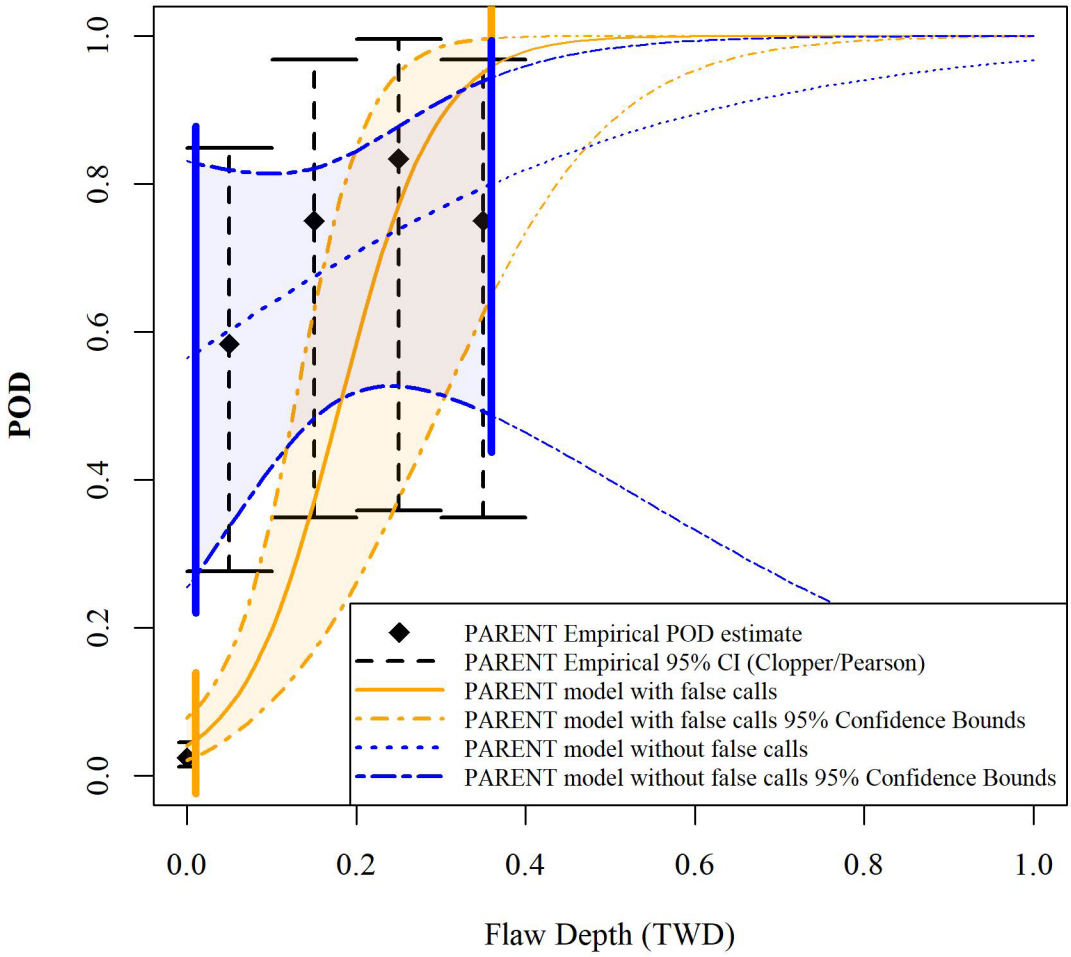

Figure 6-6 Empirical POD with POD curves for LBDMW test blocks for axial flaws from PARENT with and without false call data (ID exams). The vertical lines indicate the range from minimum to maximum flaw depth for the respective datasets. Shading between the lower and upper confidence boundaries is used to show the range of empirical data fit to the logistic function curve. Unshaded regions indicate extrapolation of the curve beyond the empirical data. Shading that extends below the minimum flaw depth indicates that false call data is used to fit the logistic function curve.

Table 6-8 Empirical POD for LBDMW test blocks for axial flaws from PARENT (ID exams).

Interval	NDET	NOBS	POD	POD low	POD high
0	10	404	0.025	0.012	0.045
(0, 0.1]	7	12	0.58	0.28	0.85
(0.1, 0.2]	6	8	0.75	0.35	0.97
(0.2, 0.3]	5	6	0.83	0.36	1.0
(0.3, 0.4]	6	8	0.75	0.35	0.97

7.0 Virtual Round Robin

A VRR activity was conducted in late 2019 and early 2020 using a virtual flaw tool (Virkkunen et al. 2021). The concept of the tool is the digital manipulation of physical flaw responses to generate artificial flaw responses to facilitate the creation of essentially arbitrarily large and varied flaw response datasets. It is envisioned that NDE performance could be derived from a database comprised solely of artificial flaw responses, or that the artificial flaw responses could be used to supplement a database of physical flaw responses. A canvas file is initially created that consists of the background response of a test block. This can be created by collecting response data from unflawed material or by digitally removing the flaw responses so that only the background response remains. Artificial flaw responses can be created from the physical flaw responses and manipulated by digitally scaling the size or manipulating other features of the flaw response. Copies of both the canvas and flaw responses can be created, allowing for the generation of essentially unlimited unique flaw response patterns on a canvas.

In this activity, the virtual flaw tool was implemented for the OD PAUT examination of circumferential flaws in an SBDMW test block (P41 from PARENT). The remainder of this section provides some background details of the activity, including a description of the test block and PAUT procedure that was implemented to create the physical flaw response data, the number of participants, and the creation of data files. Results from the aggregate analysis of the VRR data are also provided and discussed with respect to empirical results obtained in PARENT. Further descriptions of the individual data files and the performance analysis for individual participants is provided in (Virkkunen et al. 2021).

7.1 Background

The physical scan data used to create the virtual flaw files was developed by scanning SBDMW test block P41 from PARENT open testing (Meyer et al. 2017). The mockup contained six circumferential flaws. The procedure implemented was based on team 122-PA1's open technique, which utilized a 32 element transmit-receive longitudinal linear array probe operating at 1.5 MHz. Because all of the circumferential flaws in P41 were readily detectable, small thermal fatigue cracks were manufactured into separate plate specimens to generate responses from challenging, or barely detectable, flaws.

Participants for the VRR activity were recruited throughout 2019 through announcements at professional conferences, standards committees, and private communications. In all, 12 teams participated in the activity and included a mixture of participants from industry, research organizations, and universities.

A total of 66 virtual flaws were generated from the limited set of real flaw responses to create 12 simulated test block files each with 0–9 flaw responses in each file. Thus, a population of unique flaws exceeded the minimum target population of 60 specified in MIL-HDBK-1823A (2009) was able to be generated. The NOBS for this activity can be estimated by multiplying the number of unique virtual flaws by the number of participating teams, which gives NOBS = 792. Further details regarding the 12 data files can be found in Virkkunen et al. (2021).

7.2 Aggregate Analysis Methods

An aggregate analysis of the VRR data was performed by analysts at PNNL, Electric Power Research Institute (EPRI), and Aalto University. The multiple analyses were performed as a way

to verify the results and to see how sensitive the results would be with respect to the method of analysis. The influence of false call data and pseudopoints on the resulting POD representations was one area of focus. PNNL incorporated false call data by converting the false call rates to FCP as described in Section 3.2. The analysis was performed this way to replicate the method of analysis for PARENT blind data (Meyer and Heasler 2017). Similarly, pseudopoints were used in the PNNL analysis as described in Section 3.5. These pseudopoints were inserted at the 0% TWD and 100% TWD locations, and each point had a value of 50%.

EPRI performed the logistic regression analysis and computed confidence bounds using a standard R package (mh1823 R) that is referenced in MIL-HDBK-1823A (2009). “R” is a widely used programming language developed for statistical computing (R Core Team 2021). The EPRI team did not incorporate false call data into the analyses.

PNNL used in-house developed R scripts to perform the analysis. Significant differences between the in-house scripts and the standard mh1823 R package are that the mh1823 R package does not allow the input of false call data for the generation of the logistic regression curves and does not accommodate the use of pseudopoints. The in-house R script employed by PNNL allows using both false call data and pseudopoints. Otherwise, standard logistic regression functions in R were used for generating the curves.

Aalto University applied synthetic misses (i.e., pseudopoints with 0% POD values at TWD = 0) in individual analyses to compensate for the lack of miss data in some datasets (Virkkunen et al. 2021). These were carried over to the aggregate analysis presented here. Aalto University used in-house code consistent with the mh1823 R package. The Aalto University team did not incorporate false call data into the POD curve generations.

7.3 Aggregate Analysis Results

Aggregate analysis results are presented in Figure 7-1 and Figure 7-2 below for the PNNL, EPRI, and Aalto University teams. Figure 7-1 displays the portions of the POD curves where most of the transition from low to high POD occurs, while Figure 7-2 displays the POD curves over the full range of flaw depths. The 95% confidence bounds are also displayed in each plot for each team’s analysis. In Figure 7-1, differences in the aggregate analysis are apparent. Discrepancies in the resulting curves were expected since the PNNL analysis incorporated false call data and the analyses performed by EPRI and Aalto University teams did not. In comparing PNNL and EPRI generated curves, it appears that the inclusion of false call data results in lowering the value of the y-intercept at flaw depth = 0 and the curve rising with a steep slope over the transition region. Thus, below a flaw depth of ~ 2 mm, the EPRI curve predicts a higher POD while the PNNL curve predicts higher POD above that flaw depth. This is consistent with Figure 2-1 with respect to the POD curve generated in PARENT (Meyer and Heasler 2017) and the POD curve generated from MRP-262 (EPRI 2017).

Differences between Aalto University, EPRI, and PNNL POD analyses can be attributed to Aalto University’s use of synthetic misses. Figure 7-1 compares the Aalto University results to the PNNL results, showing that the Aalto University curve predicts a consistently higher POD over all flaw depths until both curves saturate.

Overall, this comparison of analyses shows that variability in POD estimations is introduced by the method of analysis that is performed. However, when comparing the differences between curves in Figure 7-2 to the results in Figure 2-1 and with results presented in the figures in Section 1.0, the variation is much smaller. It is postulated that the variance is much less for the

VRR data because of efforts to generate at least the minimum number of unique flaw responses recommended by guideline 6.7 in ASTM E2862-18 and efforts to include challenging flaw responses (i.e., virtual flaw responses representing shallow flaws with low POD). Thus, it's also hypothesized that the influence of the analysis method can be minimized through appropriate study design.

Overall, the POD estimations from the VRR show much better performance than what was observed empirically in Figure 2-1 and the results are not considered realistic. A possible explanation for this is that the physical flaw response data used to create the virtual flaw responses were created through a fingerprinting-type process. Fingerprinting refers to the application of an optimized NDE procedure to characterize a known flaw. Optimization of the NDE procedure is facilitated by knowledge of flaw location, manufacturing method, and expected characteristics. As a consequence, flaw responses generated by a fingerprinting process are expected to be of higher quality in comparison to flaw responses obtained through a blind examination.

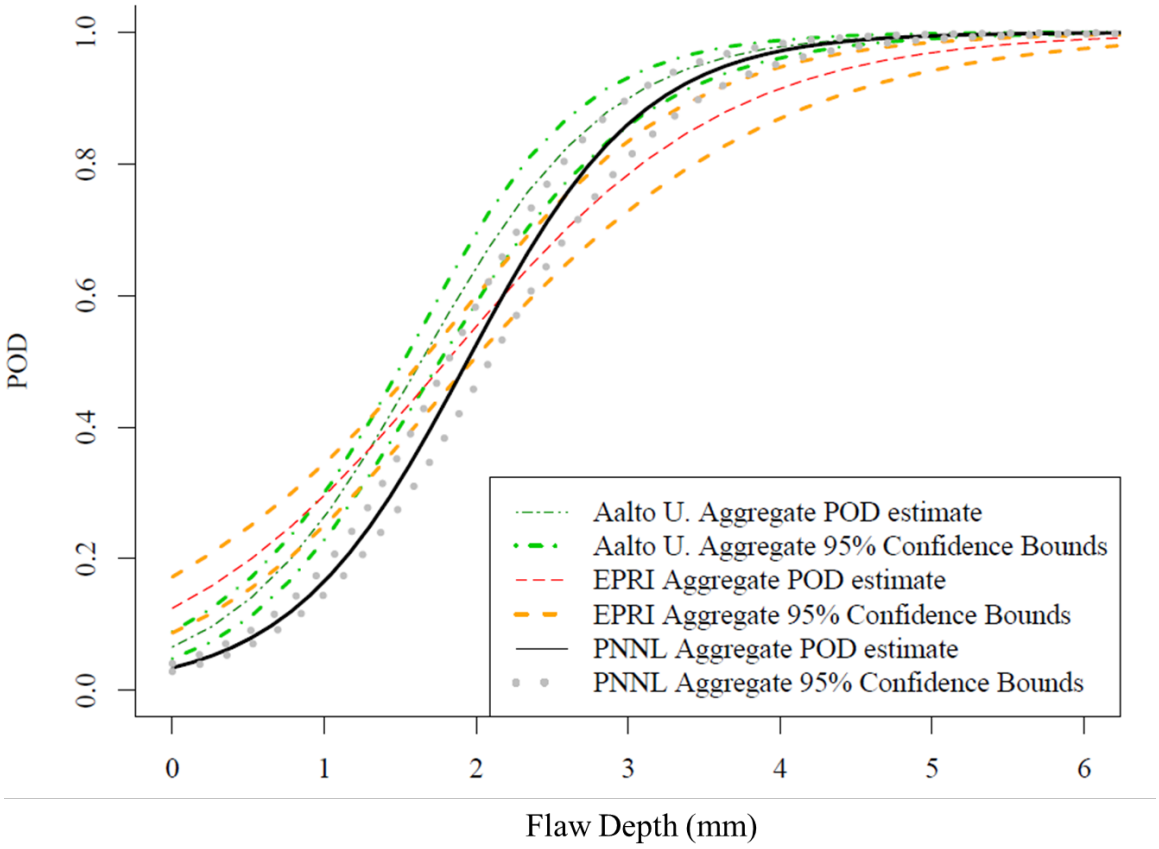

Figure 7-1 Aggregate POD curves and 95% confidence bounds for analyses by Aalto University, EPRI, and PNNL over a limited flaw depth range to emphasize the transition from low POD to high POD.

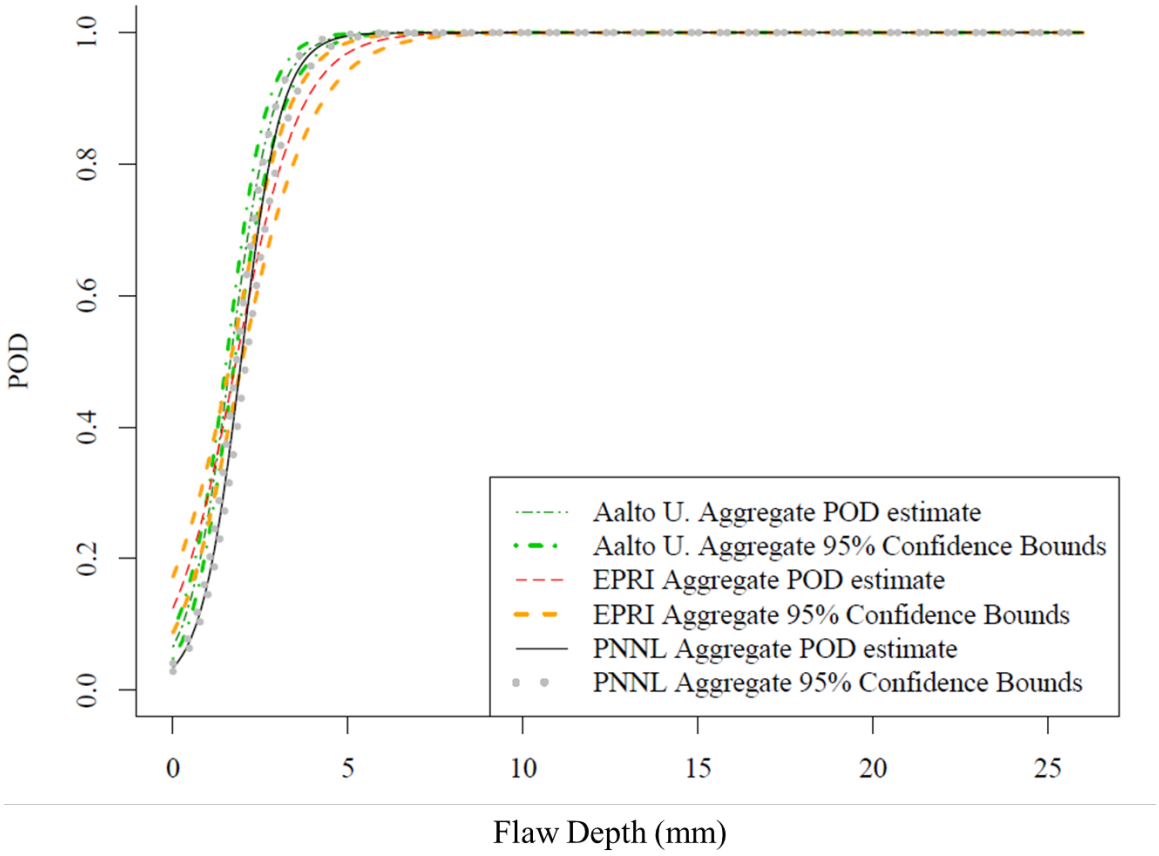

Figure 7-2 Aggregate POD curves and 95% confidence bounds for analyses by Aalto University, EPRI, and PNNL over the full test block thickness.

8.0 Guidance and Recommendations

In this section, some guidance and recommendations for performing POD analysis for NPP components are provided based on the review of PARENT data analysis, review of MIL-HDBK-1823A and ASTM E2862-18 standards, and the experience of the VRR activity. In general, NPP applications present challenges with respect to the acquisition of test blocks. Thus, POD studies of NPP components have faced data limitations and recommendations in MIL-HDBK-1823A and ASTM E2862-18 standards cannot always be met. To handle this, non-standard techniques have been adopted in the POD analysis of NPP components. The guidance and recommendations, in this section, relate to these non-standard techniques that are either not addressed or not recommended in MIL-HDBK-1823A and ASTM E2862-18.

8.1 Use of Pseudopoints

Pseuodopoints are artificial data points added to the real detection data, often for the purpose of assisting with convergence of the logistic regression curve fit. Aalto University and PNNL teams used pseuodopoints in the analysis of the VRR data in the following ways:

- Aalto University used pseodopoints in the analysis of individual participant data to help with numerical convergence of the logistic curve fits (Virkkunen et al. 2021).
- Aalto University applied “synthetic misses,” which are artificial POD data points with a value of zero at flaw depth of zero.
- PNNL employed two artificial data points each with 50% detection (one located at 0 flaw depth, another at the maximum flaw depth represented in the data) to bias the initial assumption of POD as 50% for all flaw depths. These pseudopoints were used to help logistic curve fits to converge faster (i.e., with fewer data points).

The influence that artificial data points can have on the resulting POD introduces the possibility that pseuodpoints could be applied to manipulate a POD result in arbitrary ways. Justifications were provided by Aalto University and PNNL for each of the use cases outlined in the bullets above and will be briefly described.

During Aalto University’s attempt to analyze the performances of individual participants in the VRR, it was discovered that some of the participants’ data exhibited a sharp separation between hits and misses, likely caused by a rigid thresholding procedure implemented by the participants. This sharp transition in the data prevented convergence to a logistic mode by MLE. To overcome this, pseudopoints were added to the data with an artificial hit added at 0.1 mm less than the $a_{50/50}$ value and an artificial miss added at 0.1 mm greater than the $a_{50/50}$ value. The addition of these pseudopoints increased the scatter in the results sufficiently to allow convergence (Virkkunen et al. 2021). These pseudopoints were not carried over to the aggregate VRR data analysis as the aggregated data exhibited sufficient scatter that MLE convergence issues were not encountered.

Aalto University also discovered that some teams generated data with poor separation between hits and misses and what is considered an unrealistically high POD at zero flaw depth. Pseudopoints were added in the form of 20 synthetic misses at zero flaw depth to the real detection data. The synthetic misses were added based on the assumption that POD at zero flaw depth should be zero. These synthetic misses were carried over to the aggregate analysis of the VRR data. There is conceptual similarity between the use of synthetic misses and PNNL’s

use of false call data as both are attempts to apply a constraint to what POD(TWD = 0) should be.

PNNL used pseudopoints in the analysis of VRR data, as described in the third bullet above, to maintain consistency with the analysis performed in PARENT. The use of these pseudopoints was initially motivated by the Quickblind study performed under PARENT, which incorporated a small sample size and resulted in an unbalanced dataset (no misses). The pseudopoints were carried over to the full test under PARENT and in the aggregate analysis. The influence of the pseudopoints on the aggregate analysis of PARENT data is analyzed in Section 5.0 where it is shown that they had minimal influence due to the larger sample of aggregated data.

Pseudopoint Guidance

Based on the experience with pseudopoint usage, the following guidance is recommended:

- If pseudopoints are used, a description of their implementation should be stated along with the justification for their use.
- Analysis should be performed to document the influence of pseudopoints on the results.
- The selections of pseudopoint values (POD = 0 through 1), pseudopoint locations (TWD = 0 through 1), and the number of pseudopoints can appear arbitrary although these choices are often arrived at through informal trial-and-error. Analysis should be performed and documented to justify specific selections of pseudopoint values, locations, and number of pseudopoints.

8.2 Use of False Call Data

The use of false call data in fitting the logistic model has been emphasized in descriptions of PARENT data analysis and in PNNL's analysis of the VRR data. This method has also been used in POD analysis performed by PNNL on data collected in a cast austenitic stainless steel round robin study (Jacob et al. 2021). An analysis of how the false call data influence the PARENT results is included in Section 4.0. Fitting the logistic model with false call data generally lowers the left (smaller flaw depth) portion of the curve and raises the right (larger flaw depth) portion of the curve, resulting in a steeper transition region and narrowing of the confidence bounds.

Using false call data to fit the logistic model is not recommended by standard ASTM E2862-18, *Standard Practice for Probability of Detection Analysis for Hit/Miss Data*, as discussed in Section 6.0. Table 6-1 lists several guidelines, including 6.7, which states, "False call data shall not be included in the development of the generalized linear model." False call data was used to fit the logistic model to data in PARENT despite the recommendation provided in ASTM E2862-18.

The position opposing using false call data to fit the logistic model was a topic of discussion with PIONIC participants because of the conflict with ASTM E2862-18 and because some participants of PIONIC also expressed that fitting the logistic model to false call data was inappropriate. The rationale for opposing the use of false call data was explored through discussions with PIONIC participants and communications with one of the committee members that developed ASTM E2862-18. The opposing rationale expressed through these communications was that the detection of targeted flaws is a different physical phenomenon than false indications. In other words, the interpretation of FCP as the probability of detecting

targeted flaws with zero depth is not correct because the false call indications are not actually responses to targeted flaws with zero depth but are actually the responses to inhomogeneities or other anomalies within the material.

However, an alternative viewpoint can still be expressed in support of using false call data if POD is instead interpreted as probability-of-indication. Using this point-of-view, the false call rate represents the rate at which indications are observed in unflawed material, and FCP is the probability of indications occurring in blank grading units. This view also implies that it is possible for indications to occur in flawed grading units by chance and not because the target flaw was detected. Depending on the application, the chance of an indication in a flawed grading unit being caused by something other than the target flaw may be negligible. In the context of analysis performed as part of a round robin effort to evaluate techniques for inspecting cast austenitic stainless steel specimens, the scenario is considered relevant (Jacob et al. 2021).

In addition to the conceptual challenge, false call data introduces a source of ambiguity into the analysis through the selection of L_{gu}. The analysis presented in Section 4.2 shows that the resulting POD curve is sensitive to the selection of L_{gu}. Although justifications can be provided for using specific values of L_{gu}, the selections ultimately involve some subjectivity. This can be avoided if the POD curve is only created from target flaws.

The synthetic misses incorporated by Aalto University in the analysis of the VRR data represent another way of approaching the handling of the POD curve at shallow flaw depths. As already discussed, the synthetic misses are pseudopoints with POD = 0 added at TWD = 0. This is justified by viewing POD as the probability-of-true-detection. Thus, since a flaw of zero depth is a flaw that does not exist, it cannot be detected and its POD = 0. This technique has a similar influence on POD curves as using false call data. It generally results in a steeper transition region and narrower confidence bounds. It should be noted that this method also introduces a source of ambiguity into the analysis, like false call data, through selection of the number of synthetic misses.

False Call Data Guidance

The following guidance is recommended if the false call data is used for generating a POD curve:

- An analysis of blank grading unit size, L_{gu}, should be performed and documented to show how choice of L_{gu} affects the overall POD curve and to justify selection of a specific value of L_{gu}.
- Analysis should be performed to document the influence that including the false call data has on the resulting POD curve, in comparison to the POD curve generated without false call data.

8.3 Using Virtual Flaw Data for POD Estimations

Virtual flaw methods may provide pathways for overcoming the challenges of empirical POD estimation for nuclear power applications. A VRR was conducted under PIONIC, using a virtual flaw tool as described in Section 7.0 and in (Virkkunen et al. 2021). The concept of utilizing virtual flaw data for generation of performance data was demonstrated in the VRR. Although the activity is not considered to have produced accurate performance data, virtual flaw technologies

are in early stages of development and further improvements can be expected. Some possible reasons the VRR was not able to provide realistic responses include:

- Flaw responses were obtained for the VRR through a fingerprinting-like process. Therefore, the quality of flaw responses from which the virtual flaw population was created were not representative of what would be obtained in a typical blind exam.
- Obtaining the flaw responses to populate the small flaw depths was a challenge (Virkkunen et al. 2021).

In the current form, virtual flaw methods can still support POD evaluations as a tool to aid in study design and data analysis. More specifically, virtual flaw methods can be used in support of the following:

- Evaluating minimum flaw population size requirements for POD studies
- Evaluating the effects of flaw depth distribution on POD estimates
- Evaluating the implementation of pseudopoints in POD analysis
- Evaluating the use of false call data in POD analysis.

Virtual Flaw Data Guidance

The VRR conducted under PIONIC represented a proof of concept for estimating POD with a virtual flaw method, and many aspects were not optimized. Based on the VRR experience, the following guidance is provided regarding the use of virtual flaw data for POD estimates:

- The quality of empirical flaw responses used to seed a virtual flaw population should be representative of the quality of flaw responses that would be anticipated for the application considered
- Empirical flaw responses obtained from challenging flaws (i.e., flaws that are difficult to detect) should be used for creating some of the virtual flaw population.

8.4 Aggregate POD Representation

The convention employed throughout the PINC and PARENT studies and the analysis of PDI data documented in MRP-262 Rev. 3 is to represent aggregate POD performances as an average. Average performance can be a useful representation if the goal is to compare performances among various techniques. For instance, comparing manual versus automated scanning performance or comparing the performance of examinations conducted from the component ID versus OD. The average performance can also be a useful metric for tracking how performance evolves over time with adoption of new technologies and improved procedures.

On the other hand, average performance representation does not convey information about the variation in performance among individual inspectors. POD curves derived by averaging the performance results of many inspection teams may not be appropriate because such averaging is non-conservative as several teams perform worse than the average. Thus, Li et al. (2012) propose that it is more appropriate to compute POD based on the 0.05 quantile instead of average or median (the 0.5 quantile). A curve based on the 0.05 quantile indicates the performance that would be achieved by 95% of the teams, and thus, is more conservative than POD curves based on the median or average. If the POD estimations will be used as input to

safety analysis, then derivation of POD curves based on a lower quantile, as suggested by Li et al. (2012), may be more appropriate than derivation based on the mean.

Aggregate POD curves derived based on the average performance can also have limited utility for predicting performance if the variance in the data is large. In PINC, PARENT, and MRP-262 Rev. 3, confidence intervals were used to estimate a parameter of interest, such as a mean or regression coefficient. In this case, the confidence interval is about the mean POD as a function of TWD. Confidence intervals only account for the uncertainty in the estimate of the population mean. As the number of samples increases, the width of the confidence interval decreases. Therefore, it is possible for many or most individual samples to lie outside the confidence intervals. Even though the average value is known well, information about individual performance is not conveyed.

Calculating prediction intervals may be desirable if the intent is to make best estimate predictions of future performance based on POD results. If the 95% confidence intervals for independent samples are calculated repeatedly, then 95% of those intervals should contain the true mean. On the other hand, a 95% prediction interval indicates the range of values one would expect a new sample data point to fall within. If a new independent sample data point was collected, 95% of such new data points would be expected to be contained in the 95% prediction interval. Prediction intervals account for both the uncertainty in the estimate of the population mean and the random variation of individual values. Prediction intervals incorporate information about the distribution of individual values and therefore are wider than confidence intervals. If the sample size increases sufficiently, confidence intervals are expected to get narrower and narrower, eventually becoming a line, whereas a prediction interval will converge to a constant width/thickness.

Aggregate POD Representation Guidance

There are multiple ways in which aggregate POD can be represented, and the best aggregate POD representation is dependent on the intended end-use of the POD data. The following guidance is provided regarding aggregate POD representation:

- Identify the intended end-use of aggregate POD and provide an appropriate representation of aggregate POD for that end-use. Determine if the mean or if a different quantile of POD will provide the most appropriate representation. In addition, determine if confidence intervals or prediction intervals are more appropriate for the end-use.
- If the potential end-uses of POD analysis cannot be fully anticipated, consider providing multiple representations of aggregate POD. Describe the end-use scenarios to which each POD representation applies.

9.0 Summary and Conclusion

This report provides guidance specific to POD analysis for quantifying performance estimates of NDE for NPP components and was developed as part of the PIONIC collaboration. It was motivated by experiences in estimating POD for NPP components, including PINC (Cumblidge et al. 2010), PARENT (Meyer and Heasler 2017), and the effort documented in report MRP-262 Rev. 3 (EPRI 2017). Estimating POD for NPP components is challenging because acquiring and/or manufacturing test blocks is cost prohibitive. This has led to the following primary challenges in developing test blocks and sets of flaws for POD estimations:

- The overall number of unique flaws is limited (PINC and PARENT)
- The distribution of flaw depths does not sufficiently cover the range of flaw depths for which detection is challenging (PINC, PARENT, and MRP-262 Rev. 3).

In the absence of these challenges, recommendations in MIL-HDBK-1823A (2009) and ASTM E2862-818 (2018) provide guidance for POD estimation. However, these challenges have motivated alternative methods for POD estimation for which no guidance exists. A primary example of an alternative method is the inclusion of false call data in the POD analysis in PINC and PARENT. Other examples include the use of pseudopoints or the use of virtual flaw data.

Guidance specific to POD analysis for quantifying performance of NDE for NPP components was developed by: 1) reviewing the analysis performed in PARENT and assessing the impact of the inclusion of false call data and pseudopoints on results, 2) providing an overview of the MIL-HDBK-1823A and ASTM E2862-18 standards to highlight areas where additional guidance is needed for POD analysis performed for NPP components, and 3) summarizing the outcome of a VRR activity performed with a virtual flaw tool. Guidance was then provided for several topics, including: 1) the use of pseudopoints, 2) the use of false call data, 3) the use of virtual flaw data, and 4) representation of aggregate POD.

The primary objectives achieved by this report can be summarized as follows:

- Highlighting limitations of POD models provided in Meyer and Holmes (2019) for use in probabilistic fracture mechanics codes;
- Elucidating limitations of past POD studies to inform future POD estimation efforts;
- Facilitating comparisons and benchmarking with results from PARENT (Meyer and Heasler 2017) by describing the methods of analysis;
- Documenting the outcome of a VRR activity that implemented a novel virtual flaw methodology in the attempt to overcome challenges with empirical POD estimation for NPP components;
- Summarizing guidance and recommendations for performing POD analysis for NPP components based on the review of PARENT data analysis, existing standards, and experience from the VRR activity.

Several conclusions can be made from these efforts, including:

- Each of the significant sources of data (PINC, PARENT, and MRP-262 Rev. 3) that have been utilized to estimate POD for DMWs for NPP components have some limitations that inhibit satisfying all the guidelines in the MIL-HDBK-1823A and ASTM E2862-18 standards.

- PARENT results are sensitive to the use of false call data in fitting of the logistic POD model. Use of false call data has a significant impact on the resulting POD curve in comparison to the POD curve generated with the false call data excluded. When false call data is excluded, the curve does not exhibit a steep transition from low to high POD and indicates a high false call probability. The false call data has the effect of lowering the POD at the small flaw size portion of the curve and creating a steeper transition from low to high POD.
- When comparing the logistic POD curve fits for PARENT data to binned empirical POD calculations over flaw size, the POD curve fits developed without false call data appear to fit the binned empirical POD values for flaw size > 0 but are inconsistent with the observed false call data.
- Aalto University introduced the concept of synthetic misses in the analysis of VRR data. Synthetic misses are assumed misses at flaw size = 0 and were added to mitigate the impact of having insufficient small challenging flaws in the target population. This is a potential alternative to using false call data.
- Three independent POD analyses were performed on the VRR data by PNNL, EPRI, and Aalto University. PNNL used false call data in the analysis, EPRI did not use false call data in the analysis, and Aalto University used synthetic misses in the analysis. Although the independently generated POD curves were not identical, their differences were modest in comparison to the effect that false call data had on PARENT POD analysis results. This suggests that the quality of the dataset affects the influence of POD analysis methodologies and that the influence may be minimized with datasets of sufficient quality.
- The POD estimations from the VRR show much better performance than what was observed in empirical round-robin studies and the VRR results are not considered realistic.
- Justification should be documented for adopting POD analysis methods that are not covered by existing guidance (e.g., use of false call data, synthetic misses). Further, the impact of adopting those methods should be documented by comparing with results of POD analysis performed only in accordance with existing guidance.

Finally, a general conclusion of this effort is that estimating POD for NPP components is complex and that there is not a “one size fits all” approach. The best approach depends on the characteristics of the dataset and intended end-use of the POD data. If multiple end-uses are anticipated, it may be desirable to tailor the approach and representation of POD for each end-use.

10.0 References

ASTM E2862-818. 2018. *Standard Practice for Probability of Detection Analysis for Hit/Miss Data.* West Conshohocken, PA: ASTM International.

Berens, A.P. 1989. "NDE Reliability Data Analysis." In *ASM Handbook, Volume 17: Nondestructive and Quality Control*, 689-701. Materials Park, Ohio: ASM International.

Braatz, B.G., P.G. Heasler, and R.M. Meyer. 2014. PNNL-22677. *PARENT Quick Blind Round-Robin Test Report.* Richland, WA: Pacific Northwest National Laboratory. ADAMS Accession No. ML14276A052.

Cumblidge, S.E., S.R. Doctor, P.G. Heasler, and T.T. Taylor. 2010. *Results of the Program for the Inspection of Nickel Alloy Components.* NUREG/CR-7019; PNNL-18713, Rev. 1. Washington, DC: U.S. Nuclear Regulatory Commission.

ENIQ. 2010. *Probability of Detection Curves: Statistical Best-Practices.* ENIQ Report No. 41, EUR 24429 EN. Luxembourg, Germany: European Network for Inspection and Qualification (ENIQ).

EPRI. 2017. *Development of Probability of Detection Curves for Ultrasonic Inspection of Dissimilar Metal Welds: Typical PWR Leak-Before-Break Line Locations.* EPRI Report 3002010988 (MRP-262, Revision 3). Palo Alto, California: Electric Power Research Institute.

Forsyth, D.S., and A. Fahr. 1998. "An Evaluation of Probability of Detection Statistics." Research and Technology Organization Applied Vehicle Technology Workshop on Airframe Inspection Reliability under Field/Depot Conditions (RTO MP-10)Brussels, Belgium, May 13-14, 1998.

Jacob, R.E., A.E. Holmes, M.S. Prowant, T.L. Moran, W.E Norris, A.A Diaz, and C.A. Nove. 2021. *Final Analysis of the EPRI CASS Round-Robin Study.* PNNL-32218. Pacific Northwest National Laboratory. Richland, WA.

Li, M., F.W. Spencer, and W.Q. Meeker. 2012. "Distinguishing Between Uncertainty and Variability in Nondestructive Evaluation." Proceedings of 38th Annual Review of Progress in Quantitative Nondestructive Evaluation, Burlington, Vermont, July 17-22, 2011. American Institute of Physics. AIP Vol. 1430 1725-1732. 10.1063/1.4716420.

Meyer, R.M., and P.G. Heasler. 2017. *Results of Blind Testing for the Program to Assess the Reliability of Emerging Nondestructive Techniques.* NUREG/CR-7235, PNNL-24196. Washington, D.C.: U.S. Nuclear Regulatory Commission. ADAMS Accession No. ML17159A466.

Meyer, RM, and AE Holmes. 2019. *Analysis of Empirical Probability of Detection Data for Dissimilar Metal Welds.* PNNL-28090. Richland, WA: Pacific Northwest National Laboratory.

Meyer, RM, AE Holmes, and PG Heasler. 2017. *Results of Open Testing for the Program to Assess the Reliability of Emerging Nondestructive Techniques.* NUREG/CR-7236, PNNL-24708. Washington, DC: U.S. Nuclear Regulatory Commission. ADAMS Accession No. ML17223A700 and ML17223A704.

MIL-HDBK-1823A. 2009. *Department of Defense Handbook: Nondestructive Evaluation System Reliability Assessment.* The United States Department of Defense.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rossi, R.J. 2018. *Mathematical Statistics : An Introduction to Likelihood Based Inference*. New York: John Wiley & Sons. 227.

Virkkunen, I., T. Koskinen, and O. Jessen-Juhler. 2021. "Virtual round robin – A new opportunity to study NDT reliability." *Nuclear Engineering and Design* 380. https://doi.org/10.1016/j.nucengdes.2021.111297.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov* | *www.nrc.gov