Designation: E3327/E3327M – 21

Standard Guide for the Qualification and Control of the Assisted Defect Recognition of Digital Radiographic Test Data[1]

This standard is issued under the fixed designation E3327/E3327M; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope

1.1 Assisted defect recognition (AssistDR) describes a class of computer algorithms that assist a human operator in making a determination about nondestructive test data. This guide uses the term AssistDR to describe those computer assisted evaluation algorithms and associated software. For the purposes of this guide, the usage of the words "defect," "evaluate," "evaluation," etc., in no way implies that the algorithms are dispositioning or otherwise making an unaided final disposition. Depending on the application, AssistDR computer algorithms detect and optionally classify indications of defects, flaws, discontinuities, or other anomalous signals in the acquired images. Software that does make an unaided final disposition is classified as automated defect recognition (AutoDR). While the concepts discussed in this guide are pertinent to AutoDR applications, additional validation tests or controls may be necessary when implementing AutoDR.

1.2 This guide establishes the minimum considerations for the radiographical examination of components using AssistDR for non-film radiographic test data. Most of the examples and discussion in this guide are built around two-dimensional test data for simplicity. The principles can be applied to three (volumetric computed tomography, for example) or higher dimensional test data.

1.3 The methods and practices described in this guide are intended for the application of AssistDR where image analysis will aid a human operator in the detection and evaluation of indications. The degree to which AssistDR is integrated into the testing and evaluation process will help the user determine the appropriate levels of process qualification and control required. This guide is not intended for applications wishing to employ AutoDR in which there is no human review of the results.

1.4 This guide applies to radiographic examination using an X-ray source. Some of the concepts presented may be appropriate for other nondestructive test methods when approved by the AssistDR system purchaser.

1.5 *Units*—The values stated in either SI units or inch-pound units are to be regarded separately as standard. The values stated in each AssistDR system may not be exact equivalents; therefore, each AssistDR system should be used independently of the other.

1.6 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate safety, health, and environmental practices and determine the applicability of regulatory limitations prior to use.*

1.7 *This international standard was developed in accordance with internationally recognized principles on standardization established in the Decision on Principles for the Development of International Standards, Guides and Recommendations issued by the World Trade Organization Technical Barriers to Trade (TBT) Committee.*

2. Referenced Documents

2.1 *ASTM Standards:*[2]

E1316 Terminology for Nondestructive Examinations

E1441 Guide for Computed Tomography (CT)

E1695 Test Method for Measurement of Computed Tomography (CT) System Performance

E2033 Practice for Radiographic Examination Using Computed Radiography (Photostimulable Luminescence Method)

E2339 Practice for Digital Imaging and Communication in Nondestructive Evaluation (DICONDE)

E2422 Digital Reference Images for Inspection of Aluminum Castings

E2445/E2445M Practice for Performance Evaluation and Long-Term Stability of Computed Radiography Systems

E2586 Practice for Calculating and Using Basic Statistics

E2597/E2597M Practice for Manufacturing Characterization of Digital Detector Arrays

[1] This guide is under the jurisdiction of ASTM Committee E07 on Nondestructive Testing and is the direct responsibility of Subcommittee E07.01 on Radiology (X and Gamma) Method.

Current edition approved Dec. 1, 2021. Published February 2022. DOI: 10.1520/E3327_E3327M-21.

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

E2698 Practice for Radiographic Examination Using Digital Detector Arrays

E2862 Practice for Probability of Detection Analysis for Hit/Miss Data

E2737 Practice for Digital Detector Array Performance Evaluation and Long-Term Stability

E3023 Practice for Probability of Detection Analysis for $\hat{a}$ Versus a Data

E3169 Guide for Digital Imaging and Communication in Nondestructive Evaluation (DICONDE)

2.2 *ISO Standards:*[3]

ISO 9000 Family Quality Management

2.3 *Other Documents:*

NEMA PS3 / ISO 12052Digital Imaging and Communications in Medicine (DICOM) Standard, National Electrical Manufacturers Association, Rosslyn, VA, USA (available free at http://www.dicomstandard.org/)

MIL-HDBK-1823A Nondestructive Evaluation System Reliability Assessment[4]

3. Terminology

3.1 See Terminology E1316 as well as the ASTM CT, CR, and DDA standards listed in Section 2 for a complete set of standard non-film radiographic test method definitions.

3.2 *Definitions of Terms Specific to This Standard:*

3.2.1 *assisted defect recognition (AssistDR), n*—the software or computer algorithms, typically involving image segmentation, feature identification, classification, and measurement, that aid operators in detecting and optionally the evaluation of indications in digital nondestructive testing data. Also referred to as software assisted evaluation, computer assisted evaluation, computer assisted detection, semi-automated defect recognition, supervised automated defect recognition, or computer aided detection.

3.2.2 *automated defect recognition (AutoDR), n*—the software or computer algorithms, typically involving image segmentation, feature identification, classification, and measurement, that classify the part being tested as acceptable or rejectable without the involvement of an operator.

3.2.3 *confidence interval, n*—a range of values such that there is a specified probability that the value of an unknown constant parameter of interest is contained in the range.

3.2.4 *confidence level, n*—the probability that a specified range of values covers an unknown constant parameter of interest.

3.2.5 *data curation, v*—the organization and integration of data collected from various sources, involving annotation, publication, and presentation of the data such that the value of the data is maintained over time, and the data remains available for reuse and preservation.

3.2.6 *failure mode and effects analysis (FMEA), n*—the systematic process of reviewing as many components and subsystems as possible to identify potential risks in a system and their root causes and impact.

3.2.7 *false negative (FN), n*—an examination result that reports that no indication is present when there is an indication in the ground truth, that is, a missed identification of an indication that should have been detected, sometimes called a "miss."

3.2.8 *false positive (FP), n*—an examination result that reports that an indication is present when there is no corresponding indication in the ground truth, that is, identification of an indication that should not have been identified, sometimes called a "false call."

3.2.9 *false positive rate (FPR), n*—number of false positive examination results divided by the total opportunities.

3.2.10 *ground truth, n*—list of indications and associated metadata present in the test data as determined by the process expert.

3.2.11 *negative, n*—an examination result that does not report the presence of an indication.

3.2.12 *negative predictive value (NPV), n*—the probability that a negative is a true negative.

3.2.13 *opportunity, n*—a single occurrence of the unit of measure for the examination, for example, a part, an image, or a pixel.

3.2.14 *positive, n*—an examination result that reports the presence of an indication.

3.2.15 *positive predictive value (PPV), n*—the probability that a positive is a true positive.

3.2.16 *probability of detection (POD), n*—a method to quantitatively assess the performance of a test method described in Practice E2862, Practice E3023, and MIL-HDBK-1823A.

3.2.17 *process expert, n*—the individual or group of individuals responsible for establishing ground truth for the data used in the examination, for example, subject matter experts, AssistDR system experts, certified Level 3s, statisticians, etc.

3.2.18 *receiver operator characteristic curve (ROC), n*—a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

3.2.19 *sensitivity, n*—measure of the portion of true positives that are correctly identified in an examination. Sensitivity is synonymous with true positive rate (TPR).

3.2.20 *specificity, n*—measure of the portion of true negatives which are correctly identified in an examination. Specificity is synonymous with true negative rate (TNR).

3.2.21 *total indications, n*—the number of indications present in the ground truth.

3.2.22 *total opportunities, n*—the number of possibilities for indications to be identified in the ground truth.

[3] Available from International Organization for Standardization (ISO), ISO Central Secretariat, Chemin de Blandonnet 8, CP 401, 1214 Vernier, Geneva, Switzerland, https://www.iso.org.

[4] Available from http://everyspec.com/MIL-HDBK/MIL-HDBK-1800-1999/MIL-HDBK-1823A_33187/.

3.2.23 *true negative (TN), n*—an examination result that reports that no indication is present when there is no corresponding indication in the ground truth.

3.2.24 *true negative rate (TNR), n*—the number of true negative examination results divided by the total opportunities.

3.2.25 *true positive (TP), n*—an indication in the examination results that corresponds to an indication in the ground truth, that is, identification of an indication that should have been identified, sometimes called a "hit."

3.2.26 *true positive rate (TPR), n*—the number of true positive examination results divided by the total indications.

3.2.27 *type I error, n*—the incorrect rejection of a true null hypothesis (a "false positive").

3.2.28 *type II error, n*—the incorrect acceptance of a false null hypothesis (a "false negative").

3.2.29 *yield, n*—the percentage of manufactured components that are evaluated as conforming.

4. Summary of Guide

4.1 This document is written in sections that correspond to the stages in the life cycle of an AssistDR implementation from initial concept through production operation as shown in Fig. 1.

4.1.1 Following the life cycle, first the performance measurements for the system are defined and agreed upon. Next, data collection, image quality, data availability, and data curation are initiated. Once a curated data set is available, process training and qualification for use can occur. Production use of the AssistDR system occurs after qualification is complete. The production use of the AssistDR system needs to be controlled in a manner similar to other nondestructive testing processes. At some point during the life cycle, the AssistDR system will need to be maintained or the manufacturing process will change for the part being inspected. When AssistDR system maintenance or upgrade occurs, or changes to the part manufacturing process occur, the AssistDR system performance needs to be verified. If the changes to the software or process are significant enough, the system may need to be requalified. A summary of each of the steps in the life cycle follows below.

4.2 *Performance Measurement*—In order to determine if an inspection process utilizing AssistDR is equivalent to or better than the existing inspection process, the performance of the existing inspection process needs to be understood. Common definitions for inspection system performance metrics and methods for measuring those metrics for both operator and AssistDR are described in this guide.

4.3 *Initial Considerations*—Several often-overlooked items should be considered before undertaking a project to implement an AssistDR process. Are the image chain and X-ray technique optimized for software evaluation? Is a statistically significant amount of the data available from the inspection process? If so, is curated ground truth available for that data, and is that data representative of all indication types? Are the target part manufacturing and inspection processes mature enough to execute AssistDR development without repeated requalification?

4.4 *Data Collection*—Significant amounts of data both with and without indications is needed to have a successful implementation of AssistDR. Hundreds, if not thousands, of images will be needed for both the development/training of the software algorithms as well as the initial qualification and future requalifications. This guide describes the different data types needed for AssistDR and best practices for assembling that data.

4.5 *Process Qualification*—Once the AssistDR system is trained using a set of data for which ground truth has been provided, it should then be qualified without knowledge of the ground truth to understand its real performance. After training, an expected TPR and FPR are known from a Receiver Operator Characteristic (ROC) chart, but the confidence in those results may be low because the training is inherently biased by knowledge of the ground truth for the data. Therefore, qualification should be conducted with a statistically significant data set. The size of the qualification data set is determined by the sample size calculated based on expected TPR and FPR, the required confidence level, and the required confidence interval as described in 6.3. It should be noted that the qualification database should incorporate the same types of indications from the FMEA as the training database, supplementing with synthetic data where necessary. Also, any test data that results in the AssistDR system being changed should be added to the training database and removed from the qualification database. The process of qualification is shown in Fig. 2.

4.5.1 The operating point on the ROC chart generated at the qualification phase therefore should be the reportable TPR and FPR for the AssistDR process, and these values should have confidence intervals associated with them that meet the requirements of the application. For instance, some applications that utilize AssistDR only as a tool to call attention to an abnormal condition would require a significantly different operating point than an application where AssistDR is used as required or critical input to the operator's decision. Once qualified, all test data that has influenced AssistDR system

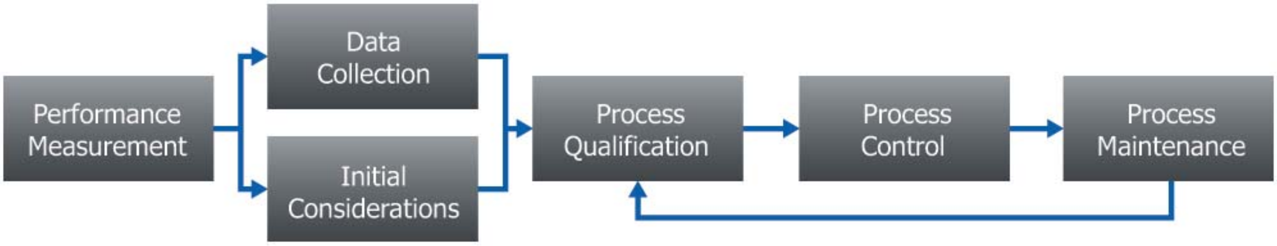


FIG. 1 Overview of an AssistDR System Life Cycle

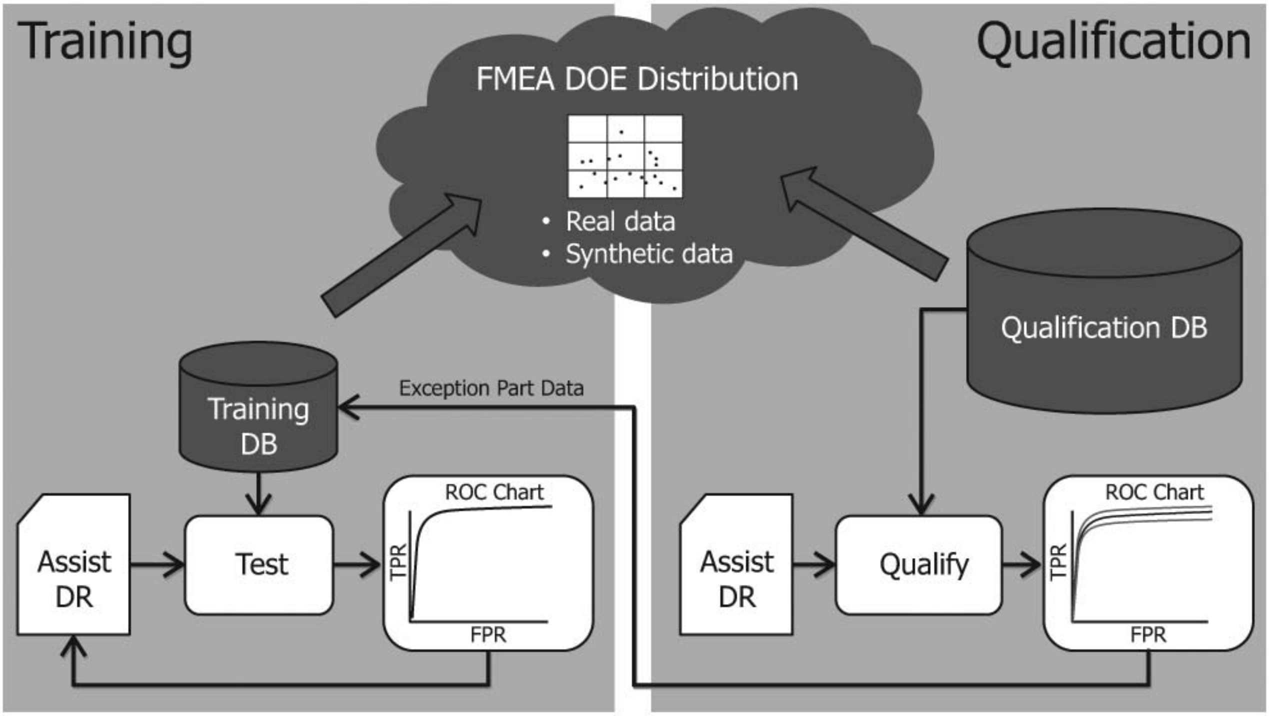


FIG. 2 Qualification Process for AssistDR

performance has been incorporated into the training database and removed from the qualification database. The qualification database can serve as a basis for the regression and requalification test databases for future requalification when AssistDR system changes or improvements occur.

4.6 *Process Control*—Once an AssistDR process has been qualified for use in production, a method for monitoring and controlling the performance of the process will be needed. Strategies for both the monitoring and control of an inspection process using AssistDR are presented in this guide.

4.7 *Process Maintenance*—Similar to other inspection processes, both routine and special cause maintenance occur for AssistDR processes. Equipment or software upgrades or replacements may also occur. When these events occur, the potential impact of those events on the performance of the system needs to be assessed. Recommendations for assessing and measuring the performance impact of maintenance events are detailed in this guide.

5. Significance and Use

5.1 This guide describes the recommended procedure for using software to assist with the identification of indications in digital radiographic images. Some of the concepts presented may be appropriate for other nondestructive test methods.

5.2 When properly applied, the methods and techniques outlined in this guide offer radiographic testing practitioners the potential to improve inspection reliability, reduce inspection cycle time, and harness inspection statistics for improving manufacturing processes.

5.3 The typical goal of a nondestructive test is to identify flaws that exceed the acceptance criteria. Due to the variability and uncertainty present in any inspection process, acceptance thresholds are established so that some acceptable components are discarded in an effort to prevent parts with discontinuities that exceed the acceptance criteria from entering service. This type of error, called a false positive, is considered less critical than a false negative error which would allow a nonconforming part into service. A successful application of AssistDR minimizes the false positive rate while reducing the false negative rate to levels appropriate for the intended application. The methods and techniques described in this guide facilitate achieving this desired outcome.

5.4 With the advent of deep learning, convolutional neural networks, and other forms of artificial intelligence, scenarios become possible where an AssistDR system continues to evolve or learn after qualification for production use. This guide does not address learning-based AssistDR systems. This guide addresses only deterministic systems that have software code and parameters that are fixed after qualification. Note that this limitation does not prohibit the use of this guide for developing a qualification and usage strategy for software using deep learning technology. The training or learning process for the deep learning system would need to be completed before qualification and all parameters of the deep learning system held fixed (as with deterministic software approaches based on traditional image processing) after qualification and during use.

6. Performance Measurement

6.1 The ability of an AssistDR system to find relevant indications in the nondestructive test data and to ignore nonrelevant ones is the cornerstone of a successful implementation. This section defines common performance measures for

AssistDR processes. Since those measures will be calculated from a sample of the processed data, the relevant statistical measures on the confidence of those metrics are described next. An AssistDR project will need to set targets for both the performance measures and statistical confidence during initial phases. Finally, some guidance on determining the sample size needed to meet those targets is presented. The AssistDR process should be evaluated under the same conditions that it is intended to be used. For example, when measuring the performance of AssistDR, it should include operator input if that is the intended production use.

6.2 *Quantitative System Performance Evaluation*

6.2.1 The first step in analyzing AssistDR system performance is to organize its results into the format shown in Fig. 3. In this table, the number of positive examination results that had a corresponding indication in the ground truth is put in the first row and first column and the number of positive examination results that did not have a corresponding indication in the ground truth goes in the second row of column 1. The same process is followed for the second column of the table for the negative examination results. A table in this format is often referred to as a truth table or confusion matrix.

6.2.2 The number of total indications for the examination result in Fig. 3 is *a+b*.

6.2.3 The number of true positive results for the examination in Fig. 3 is *a*. The true positive rate can be calculated as *a / (a+b)*.

6.2.4 The number of false negative results for the examination in Fig. 3 is *b*. These are Type II errors. The false negative rate can be calculated as *b / (a+b)*.

6.2.5 The number of opportunities for a false positive for the examination result in Fig. 3 is *c+d*.

6.2.6 The number of false positive results for the examination in Fig. 3 is *c*. These are Type I errors. The false positive rate can be calculated as *c / (c+d)*.

6.2.7 The number of true negative results for the examination in Fig. 3 is *d*. The true negative rate can be calculated as *d / (c+d)*.

6.2.8 The fraction of examinations that correctly tested positive for an indication is the positive predictive value (PPV). The PPV can be calculated as *a / (a+c)*.

6.2.9 The fraction of examinations that correctly tested negative for an indication is the negative predictive value (NPV). The NPV can be calculated as *d / (b+d)*.

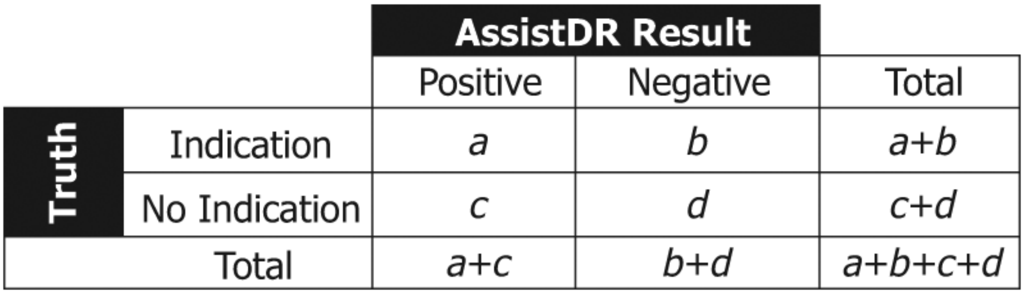
6.3 *Use of Confidence Intervals in AssistDR Validation*

6.3.1 The true value of the performance metrics described in 6.2 is never completely known. The performance of operators or AssistDR systems can never be measured on every part produced. Instead, the performance metrics are estimated using a sample of parts. This single measurement does not necessarily reflect the true performance, but its statistical confidence can also be measured and reported. Due to this measurement uncertainty, the statistical significance of a sample is expressed using a confidence interval and confidence level. This allows for the precision of the performance measurement to be reported.

6.3.2 The interpretation of confidence interval and confidence level on TPR is shown in Fig. 4. The endpoints of the confidence interval are referred to as the upper and lower confidence bounds. For simplicity, Fig. 4 shows a special case of upper and lower confidence bounds that are symmetrically distributed about the sample estimate. Because the distribution is symmetric, the confidence level is divided into two equal halves to create both the upper and lower confidence bounds. For a TPR estimate t with confidence bounds cb_u and cb_l, and a confidence level c%, the probability that the sample's upper and lower bounds, cb_u and cb_l, contain the population TPR is c%.

6.3.3 For the measurement of TPR, the symmetric probability distribution above is appropriate only for TPRs measured at or near 50 %, or for very large sample sizes. As the TPR approaches 100 %, the distribution skews left and sample TPR does not coincide with the mode or peak of the distribution. Additionally, neither the confidence level nor the confidence intervals are symmetric. The width of the confidence interval depends on the sample size, N. Fig. 5 shows this relationship as TPR deviates from the simple case, for a relatively low sample size of 30. For more information regarding skewness of probability density functions and appropriate probability distribution models for TPR and FPR estimates, refer to Practice E2586.

6.3.4 It is the lower confidence bound that is of interest for the measurement of the TPR of an AssistDR application. The lower confidence bound on the TPR estimate represents the lowest value of TPR that could be expected on future measurements of that AssistDR process at a given confidence level. Note that for skewed sensitivity distributions as shown in Fig. 6, the confidence interval is asymmetric including more TPR estimates below the measured sensitivity. Hence, the lower bound is further below the measured TPR than the upper bound



		AssistDR Result		
		Positive	Negative	Total
Truth	Indication	a	b	a+b
	No Indication	c	d	c+d
	Total	a+c	b+d	a+b+c+d

FIG. 3 AssistDR Test Results Compared to Ground Truth

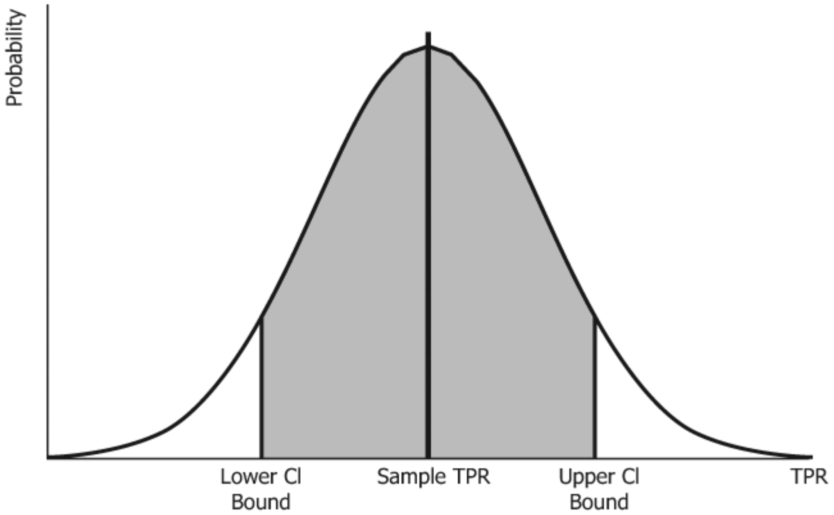


FIG. 4 Illustration of Confidence Bounds and Confidence Level

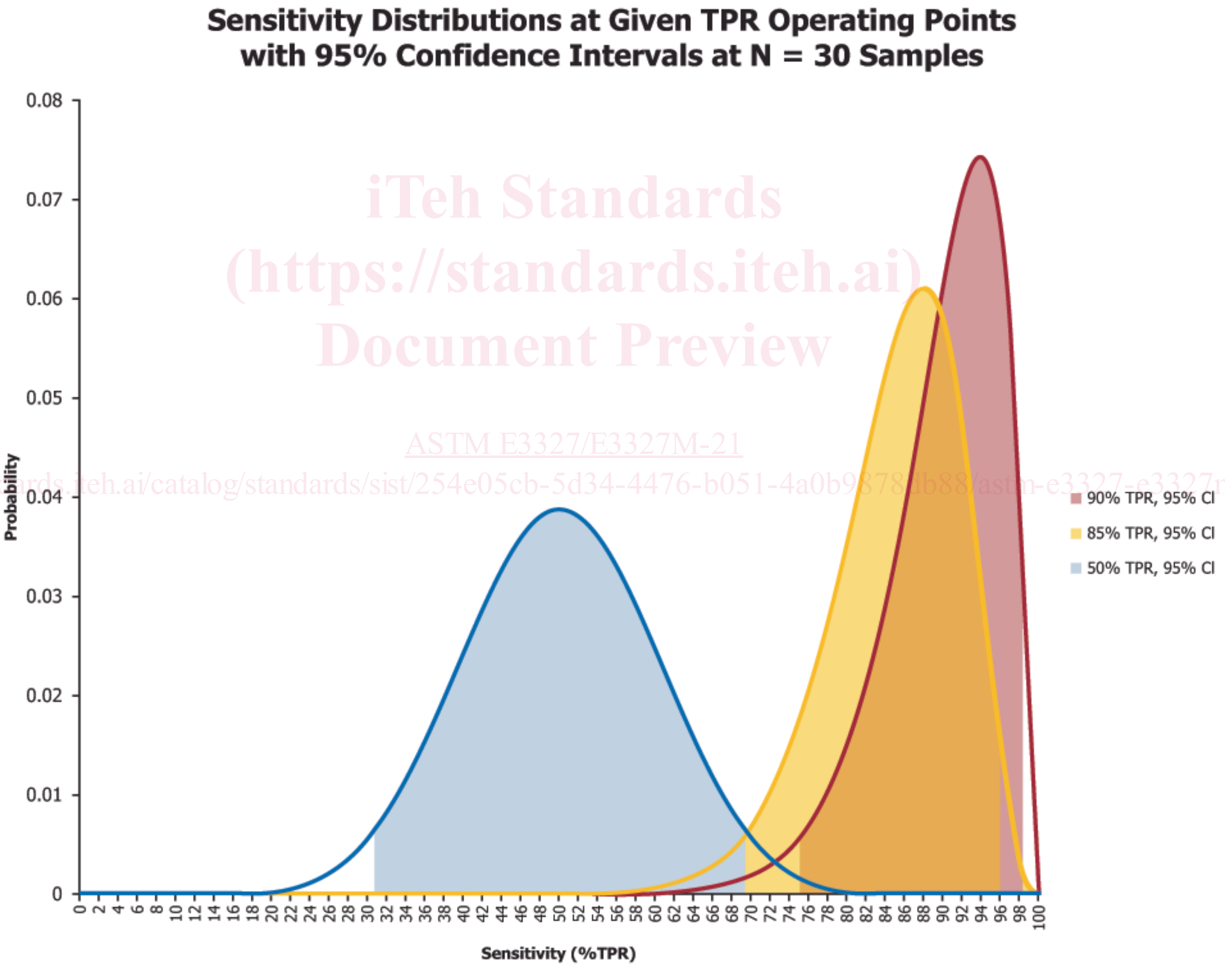


FIG. 5 Illustration of Confidence Level and Confidence Intervals for Increasing TPR Measurements

is above. Increasing the sample size reduces this effect on the sensitivity measurement's lower bound. Similarly, it is the upper confidence bound that is of most interest for FPR estimates.

6.4 *Sample Size Requirements for Validation of Initial Performance Assessments*

6.4.1 In order to provide a clear, quantitative validation of an AssistDR system's performance (in terms of TPR, for

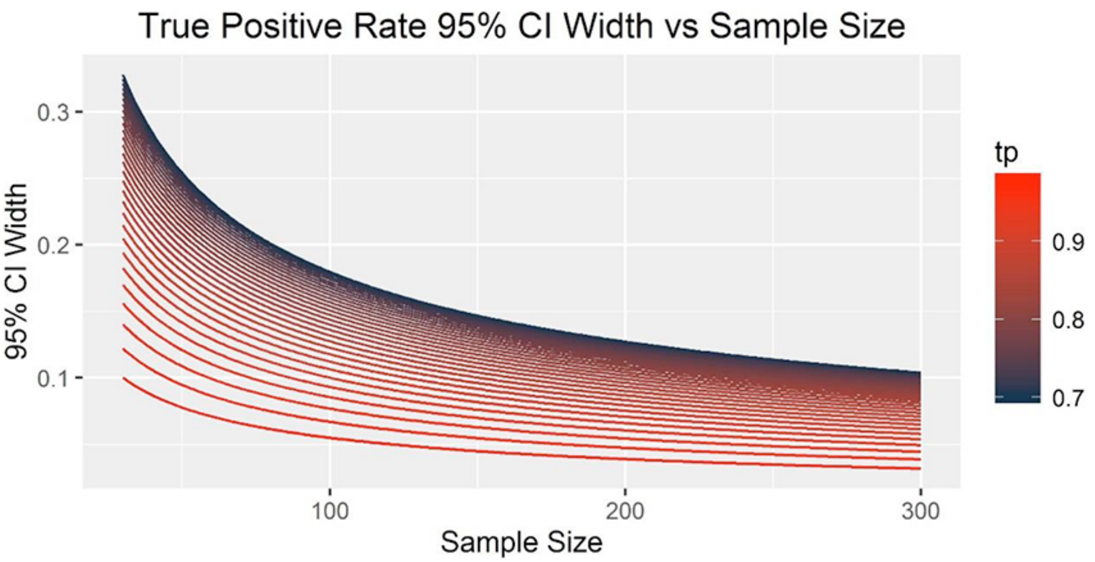


FIG. 6 95 % Confidence Interval Width for TPR Estimates as A Function of Sample Size and TPR; Lines Are Colored by Estimated TPR, Increasing From 70 % (in Black) to 98 % (in Red); Note That Confidence Interval Width Decreases as A Function of Sample Size And That Confidence Interval Widths Are Smaller for Larger TPR Estimates Than for Lower TPR Estimates For A Given Sample Size

example) on a given part, two quantities are required beforehand. The first of these quantities is an estimate of the anticipated performance (say TPR) and the second is the desired width of the resulting confidence interval. These two quantities are needed because the relationship between sample size and confidence interval width is a function of anticipated performance. That is, the number of samples required to meet a given confidence interval width for one value of TPR will differ from the number required for a higher value of TPR. Fig. 6 illustrates this relationship for TPR.

6.4.2 The lower confidence bound for a TPR estimate can be estimated given the sample size used to calculate the estimate and the desired confidence level. The approximate lower confidence bound for a 95 % confidence level for a given TPR and sample size in the associated study can be determined from Table 1. To use Table 1, first find the row that is closest to, but not greater than, the TPR of the study. Next, find the column with the number of samples that is closest to, but not greater than, the number of samples in the study. The heading of the column is the lower confidence interval. To determine the lower confidence bound, simply subtract the lower confidence interval from the TPR of the study. For example, consider a study with a 90.5 % TPR measured from a qualification data set with 250 indications. The row of Table 1 that is closest to, but not greater than, the TPR is the 90 % TPR row. The column with the number of indications that is closest to, but not greater than, the 250 indications used in the study is the 5 % Confidence Interval column. The 95 % lower confidence bound on TPR for this study is 85.5 % (90.5 % - 5 %).

6.4.3 Similarly, an upper confidence bound for an FPR estimate can be estimated given the sample size used to calculate the estimate and the desired confidence interval. If the number of opportunities for a false positive is defined, FPR can be bounded. By noting that FPR = 1 – TNR, Table 1 can also be used to compute a lower bound on TNR. Analogously, this lower bound on TNR equates to an upper bound on FPR.

6.4.4 If the number of opportunities for a false positive is not defined, FPR may be unbounded. For example, if it is desired to measure the number of false positives per opportunity, the upper confidence bound should be calculated from a counting process distribution, as described in Practice E2586. The approximate upper confidence bound for a 95 % confidence level for a given number of false positives in the associated study can be determined from Table 2. To use Table 2, first find the row that is closest to, but not less than, the number of false positives per opportunity of the study. Next,

TABLE 1 Number of Indications Required in the Qualification Data Set for a 95 % Confidence Level (CL) on True Positive Rate

True Positive Rate	Confidence Interval (CI)						
	1 %	2 %	3 %	4 %	5 %	10 %	20 %
97.0 %	1484	461	244	159	115	45	19
96.0 %	1837	548	282	180	129	49	20
95.0 %	2181	632	319	201	142	52	20
94.0 %	2517	715	356	221	155	55	21
93.0 %	2846	796	391	240	167	57	21
92.0 %	3167	875	425	260	179	60	22
91.0 %	3480	952	459	278	191	63	23
90.0 %	3786	1027	492	296	202	65	23
87.5 %	4516	1206	570	339	229	71	24
85.0 %	5198	1373	643	379	254	77	25
82.5 %	5832	1528	710	416	277	82	26
80.0 %	6418	1671	772	450	298	87	27

TABLE 2 Number of Opportunities Required in the Qualification Data Set for a 95 % Confidence Level (CL) on False Positives per Opportunity

False Positives Per Opportunity	Confidence Interval (CI)						
	0.05	0.10	0.20	0.30	0.40	0.50	1.00
0.10	231	77	29	17	12	-	-
0.20	385	116	39	22	15	7	-
0.30	539	154	48	26	17	8	-
0.40	693	193	58	30	20	8	6
0.50	846	231	68	35	22	9	6
0.60	1000	270	77	39	24	10	7
0.70	1154	308	87	43	27	10	7
0.80	1307	347	97	47	29	11	7
0.90	1461	385	106	52	32	12	8
1.00	1615	423	116	56	34	12	8
1.25	1999	520	140	67	40	14	9
1.50	2383	616	164	77	46	16	10
1.75	2767	712	188	88	52	18	11
2.00	3151	808	212	99	48	19	12
4.00	6225	1576	404	184	106	33	20

find the column with the number of samples that is closest to, but not less than, the number of samples in the study. The heading of the column is the upper confidence interval. To determine the upper confidence bound, add the upper confidence interval from the number of false positives per opportunity of the study. For example, consider a study with 0.83 false positives per clean component measured from a qualification data set with 100 clean components. The row of Table 2 that is closest to, but not less than, the false positives per clean component is the 0.9 false positives row. The column with the number of opportunities that is closest to, but not less than, the 100 opportunities used in the study is the 0.2 Confidence Interval column. The 95 % upper confidence bound on false positives per component for this study is 1.1 (0.9 + 0.2).

6.5 An alternative approach to the quantitative assessment of the performance of a nondestructive testing method is probability of detection (POD). Like TPR, the POD approach provides estimates of the probability that an inspection method correctly detects defects. The POD approach is based upon a statistical model with several important assumptions. Perhaps the most important of these assumptions is that there is a single aspect of true defects (often defect size) that strongly influences probability of detection. TPR, on the other hand, rests on very few assumptions. The assumptions behind both methods have both benefits and risks. The strongly parametric nature of the POD approach means that smaller data sets are required to get meaningful estimates, while the simpler TPR approach requires larger data sets. However, TPR is ultimately more robust in the presence of multiple defect types or when image properties other than indication size have strong influence on performance (for example, strong gradients or variation in contrast). Detailed information on POD can be found in Practice E2862, Practice E3023, and MIL-HDBK-1823A.

7. Data Collection

7.1 Several different data sets are necessary to qualify an AssistDR system for use in a production environment. These data sets correspond to the overall process described by Fig. 1 and are broken down into more detail in Fig. 7 below. In Fig. 7 , the first step is to carefully define the space of NDT data that is representative of the manufacturing and inspection process. This is described in detail in 7.2. This definition should be used as guidance to collect data sets for assessing operator performance, software training, and software performance. For this data to be useful in conducting performance assessments or training, it will need to have the ground truth determined. Ground truth determination is discussed in 7.3.

7.1.1 Before describing the elements of a comparison for evaluating an AssistDR system, it should be noted that every inspection system and manufacturing context is different, and those differences can have significant implications for conducting a successful and meaningful comparison study. To handle these nuances and to ensure the accuracy of results, it is recommended that a trained statistician be consulted whenever possible. If the AssistDR system purchaser or provider lacks a staff statistician, there are a variety of fully qualified statistical consulting firms that can be contacted to assist in the design of these studies.

7.2 *Representative Process Data*

7.2.1 Often, faults in AssistDR systems are not detected until the later stages of development. The late detection is due to the random nature of the variation of manufacturing processes. The image or type of indication that caused the fault was simply not produced by the process during the previous testing periods. However, finding a problem at this point in the development process can add significant cost and delays to schedules. The challenge is to create a training data set that is as robust and comprehensive as possible to uncover these faults during the training phase. One way to create a robust training set is to use Failure Mode and Effects Analysis (FMEA), a tool for identifying potential problems and their impact. FMEA is a qualitative and systematic tool, usually created within a spreadsheet, to help practitioners anticipate what might go wrong with a product or process. Since the effects of failure for AssistDR systems are limited, the focus is on identifying modes (indication types, indication locations, background level, background gradient, etc.) that may cause the software to provide an incorrect result.

7.2.2 A strong cross-functional team is necessary to perform an effective FMEA. This team should include manufacturing process experts to identify type, size, and location of indications; inspection process experts to identify variation in image

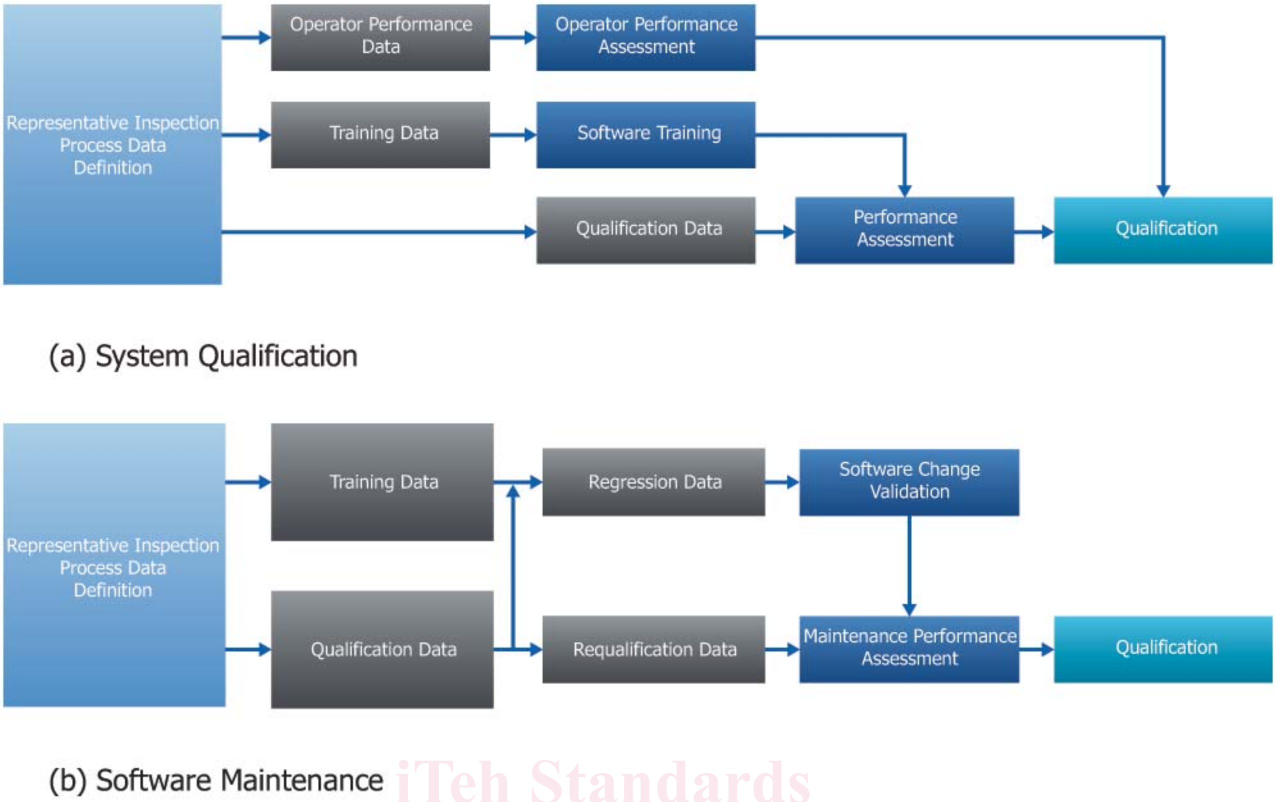

Representative Inspection Process Data Definition
Operator Performance Data
Operator Performance Assessment
Training Data
Software Training
Qualification Data
Performance Assessment
Qualification
(a) System Qualification
Representative Inspection Process Data Definition
Training Data
Regression Data
Software Change Validation
Qualification Data
Requalification Data
Maintenance Performance Assessment
Qualification
(b) Software Maintenance

FIG. 7 Data Sets Used in AssistDR System Qualification and Software Maintenance

view angle, quality, and noise; and AssistDR system experts to identify weak points of algorithms and filters. Such a group is needed to identify the broadest set of failure modes possible to include in the training data set.

7.2.3 The objective of a FMEA for the qualification and control of AssistDR software is to identify the range of input data to the software that could impact performance. A more extensive FMEA considering a broader range factors impacting implementation, such as computation time, network speeds, operator acceptance, etc., is recommended. The analysis includes expected variations in the manufacturing process, inspection process, and software algorithms. The output of the FMEA should be a table of factors that could cause a failure of the AssistDR software. An example table of the output of a FMEA is found in Table 3.

7.2.3.1 This example is based on the positive image in Fig. 8 where air is shown as white and the part being inspected as shades of gray. The darker shades of gray indicate thicker areas of the part. There are two indication types (foreign material and porosity) that the AssistDR system is required to detect. In this example, each of these types has a 0.020 in. minimum interpretable size, but the manufacturing process produces indications over a wide range of sizes. Due to the nature of the foreign material and porosity, the digital signals produced by the indications vary in both magnitude and signal-to-background ratio. There are two boundaries of interest in the image, the part/air boundary (1) and the part/end of image boundary (3, 4). There are areas of thinner material (2), thicker material (3), and a transition from thin to thick (5).

7.2.4 After the failure modes are identified, they should be incorporated into the training and qualification databases for the AssistDR system. For elements of the FMEA having a range of values, an efficient way to put together a data set that spans the entire space is by applying the Design of Experiments (DOE) methodology. Design of Experiments is a systematic method to determine the relationship between factors affecting a process and the output of that process. The most common DOE method uses a full factorial design where the boundary points on the range of the input parameters are combined in a systematic manner to form a set of experiments. For the example in Table 3, this would result in a 48 (2x2x2x2x3) run DOE for each indication type. Each of the runs for this design is shown in Table 4.

7.2.5 After creating a DOE design, images are assembled representing each of the runs in the DOE. It is preferable for these to be indications that are produced by the manufacturing process and created on the production inspection system. Due to the nature of manufacturing processes, not all indications that may occur in the process occur on a regular basis. These infrequently occurring indication types may make completing the DOE image matrix in a timely manner impossible. In these cases, it is necessary to use carefully created simulated images to fill out the DOE matrix.

TABLE 3 Example FMEA Output For A Radiographic Inspection Process Based on Notional Image in Fig. 8

Failure Modes	Impact of Failure		Range of Causing Condition	Failure Mode Mitigation
	Miss	False Positive		
Foreign Material Too Small to Detect	×		0.020 in. Min Interpretable	10 indications between 0.025 in. and 0.015 in. major dimension in qualification data set
Foreign Material Too Large to Detect	×		Indications greater than 0.150 in. may cause algorithm error	10 indications greater than 0.150 in. in qualification data set
Low Contrast Foreign Material	×		Foreign material less than CNR threshold possible	10 indications within 5 % of CNRthreshold in qualification data set
High Contrast Foreign Material	×		Large areas of foreign material rare but possible	10 indications greater than 200 % of CNRthreshold in qualification data set
Porosity Too Small to Detect	×		0.020 in. Min Interpretable	10 indications between 0.025 in. and 0.015 in. major dimension in qualification data set
Porosity Too Large to Detect	×		Indications greater than 0.150 in. may cause algorithm error	10 indications greater than 0.150 in. in qualification data set
Low Contrast Porosity Detection	×		Porosity less than CNR threshold rare but possible	10 indications between within 5 % of CNRthreshold in qualification data set
High Contrast Porosity Detection	×		Large pores rare but possible	10 indications greater than 200 % of CNRthreshold in qualification data set
Algorithm Performance at Part/Air Interfaces	×	×	2 regions of part/air interface	20 indications located within 0.010 in. of a part/air interface in qualification data set
Algorithm Performance at Part/Image Edge Interfaces	×	×	2 regions of part/image edge interface	20 indications located within 0.010 in. of a part/image edge interface in qualification data set
Algorithm Performance in Areas of Thickness Transition	×	×	1 region of thickness transition between platform and airfoil	20 indications located within 0.020 in. of a thickness transition in qualification data set (10 in thicker region, 10 in thinner in region)
Complete Detector Failure	×		All data in image occupies less than 5 % of dynamic range	Error check for detector failure incorporated into software before detection algorithm
Detector Degradation	×	×	Areas of low contrast or higher than acceptable noise	Performance check of system using a reference part at the start and end of each shift
Operator Does Not Evaluate Identified Indication	×	×	All software identified indications must be evaluated	Software implements check box for each indication and all check boxes must be set before software will register the inspection complete

7.2.6 The results of the AssistDR system on the DOE should be expressed as a TPR for the indications. While other metrics such as FPR may be of interest and provide insight into the system's performance, the DOE design method described in this section is intended to investigate the impact of process variation on TPR.

7.2.7 Requirements on the performance of the AssistDR system on the DOE matrix will vary from project to project. If a large, statistically significant DOE matrix like the one specified in Table 4 is used, a TPR requirement based on the production inspection processes can be applied to evaluate the DOE results. For some projects, a statistically significant DOE may not be possible. In this case, the experiment can use a reduced size DOE focused on specific indication conditions (for example, location, contrast, size) that do not occur in typical production but are identified by the FMEA as conditions the algorithm needs to address. If a reduced size DOE is used, a higher TPR should be required of the software on the DOE data to ensure that the AssistDR system has some degree of sensitivity to those conditions.

7.2.8 The nature of the assessment data set used in a qualification study is vital to determining the efficacy of an AssistDR system. This data set should consist of a sufficiently large (see discussion in 6.4) collection of indications and inspection opportunities to provide estimates of statistical sensitivity and specificity for a purely manual inspection system and the associated AssistDR-enabled inspection system. In practice, determining how many indications are required is often a tradeoff between statistical requirements and practical matters such as availability of indications in production, operator availability, cost, etc. If a reasonable initial estimate of operator sensitivity and specificity is available, the method described in 6.4.2 can be employed to identify the number of indications and indication-free inspection system outputs required to provide real sensitivity and specificity estimates with a specified degree of uncertainty. Generally speaking, if a statistician is unavailable, it is prudent to err on the side of larger data sets. Beyond the quantity of indications and indication-free inspection opportunities, it is extremely important to ensure that the assessment data set

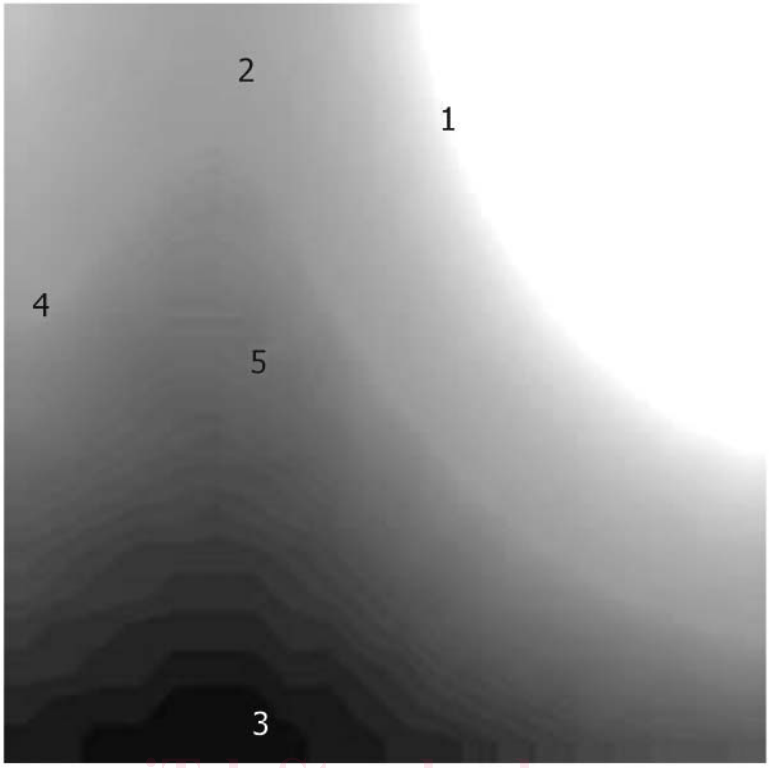


FIG. 8 Image for FMEA Example

TABLE 4 Example DOE based on FMEA in Table 3

Run Order	Size	SBR	Image Location	Part Location	Gradient	Run Order	Size	SBR	Image Location	Part Location	Gradient
1	Small	Low	Edge	Edge	B->G	25	Large	Low	Edge	Edge	B->G
2	Small	Low	Edge	Edge	W->G	26	Large	Low	Edge	Edge	W->G
3	Small	Low	Edge	Edge	Flat	27	Large	Low	Edge	Edge	Flat
4	Small	Low	Edge	Center	B->G	28	Large	Low	Edge	Center	B->G
5	Small	Low	Edge	Center	W->G	29	Large	Low	Edge	Center	W->G
6	Small	Low	Edge	Center	Flat	30	Large	Low	Edge	Center	Flat
7	Small	Low	Center	Edge	B->G	31	Large	Low	Center	Edge	B->G
8	Small	Low	Center	Edge	W->G	32	Large	Low	Center	Edge	W->G
9	Small	Low	Center	Edge	Flat	33	Large	Low	Center	Edge	Flat
10	Small	Low	Center	Center	B->G	34	Large	Low	Center	Center	B->G
11	Small	Low	Center	Center	W->G	35	Large	Low	Center	Center	W->G
12	Small	Low	Center	Center	Flat	36	Large	Low	Center	Center	Flat
13	Small	High	Edge	Edge	B->G	37	Large	High	Edge	Edge	B->G
14	Small	High	Edge	Edge	W->G	38	Large	High	Edge	Edge	W->G
15	Small	High	Edge	Edge	Flat	39	Large	High	Edge	Edge	Flat
16	Small	High	Edge	Center	B->G	40	Large	High	Edge	Center	B->G
17	Small	High	Edge	Center	W->G	41	Large	High	Edge	Center	W->G
18	Small	High	Edge	Center	Flat	42	Large	High	Edge	Center	Flat
19	Small	High	Center	Edge	B->G	43	Large	High	Center	Edge	B->G
20	Small	High	Center	Edge	W->G	44	Large	High	Center	Edge	W->G
21	Small	High	Center	Edge	Flat	45	Large	High	Center	Edge	Flat
22	Small	High	Center	Center	B->G	46	Large	High	Center	Center	B->G
23	Small	High	Center	Center	W->G	47	Large	High	Center	Center	W->G
24	Small	High	Center	Center	Flat	48	Large	High	Center	Center	Flat

covers the full space of possible inspection contexts. Thus, when developing an assessment data set, it is recommended that the AssistDR system purchaser or provider conduct an FMEA Design of Experiments as described above.

7.3 *Ground Truth*—Assembly of the associated ground truth data set is one of the most important, most time consuming, and most challenging elements of creating a training or qualification data set. Assembling this ground truth is challenging because it requires that all indications in the image data set larger than some minimum interpretable size be located, measured, and assigned a unique identification. Doing this task accurately usually requires significant manual intervention. The accuracy of the performance measurements for the AssistDR algorithm is directly dependent on the correctness of the