Paper AS10

Statistical Endpoints and Analysis in Diagnostic Imaging: The Matrix Has You... and Your Data

James Sykes, Blue Earth Diagnostics, Oxford, United Kingdom
Srinivas Lanka, Blue Earth Diagnostics, Oxford, United Kingdom

ABSTRACT

Diagnostic imaging studies are essential for assessing the accuracy, reliability, and diagnostic performance of new imaging technologies, particularly in oncology. As the demand for improved imaging grows, the ability to accurately detect lesions and support early clinical decisions becomes increasingly important. Effective imaging can lead to faster, more appropriate treatment pathways. These studies often rely on endpoints derived from a standard 2 x 2 matrix—most notably, sensitivity and specificity, which are key co-primary endpoints in Phase 3 trials. A well-defined standard of truth is also critical for evaluating diagnostic accuracy. But how do we efficiently collect and summarize these endpoints? This paper explores how to map diagnostic imaging data to the SDTM structure, integrate it into ADaM efficacy datasets, and derive relevant performance statistics. Given SAS's ongoing role in clinical research, we will also provide practical SAS code examples to support analysis, ensuring transparency and reproducibility in diagnostic performance evaluation.
Keywords: Diagnostic imaging, detection rates, statistical endpoints, SDTM, ADaM, SAS Programming

INTRODUCTION

In oncology, precise lesion detection plays a critical role in shaping participant outcomes (Sharkey et al., 2020). Key impacts include:

1. Early Detection - Improved detection facilitates timely treatment initiation.
2. Treatment Planning - Accurate identification of lesions guides surgical and therapeutic strategies.
3. Disease Monitoring - Consistent detection enables reliable assessment of treatment response.
4. Participant Management - Detection rates influence clinical decision-making and resource allocation.

Therefore, modern imaging technologies must demonstrate clinical value through rigorous evaluation of their diagnostic performance.

In current clinical trials, diagnostic performance generally describes how well a novel imaging agent (e.g., prostate-specific membrane antigen [PSMA]--targeted positron emission tomography [PET] in prostate cancer) can accurately identify, characterize, and exclude disease. It is typically assessed using endpoints such as sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) reflecting the ability of an imaging test to correctly identify true findings while avoiding false findings compared to a truth standard e.g. standard of truth (SoT) or standard of reference (SoR).

Diagnostic performance metrics often act as co-primary endpoints in Phase 2 and 3 trials, for example sensitivity and specificity. Regulatory authorities often recommend co-primary endpoints in diagnostic imaging trials because together they capture a test's ability to both detect disease when it is present (sensitivity) and rule it out when it is absent (specificity). Considering them jointly ensures a balanced evaluation of diagnostic performance, preventing a test from appearing effective by excelling in only one aspect while failing in the other.

At the core of evaluating diagnostic imaging is a basic 2×2 contingency table, which compares diagnostic results to an established truth standard (FDA Guidance for Industry, 2022). This paper aims to tackle the practical question: "How can we efficiently gather and summarize diagnostic imaging data, to efficiently create the 2x2 table?" The answer to this question is not as simple as you might think. We will examine the entire process from mapping diagnostic imaging data to Study Data Tabulation Model (SDTM) structure, integrating it into Analysis Data Model (ADaM) efficacy datasets, and deriving the relevant counts so that the diagnostic performance statistics can be calculated (CDISC SDTM Implementation Guide, 2023; CDISC ADaM Implementation Guide, 2021).

DIAGNOSTIC PERFORMANCE FRAMEWORK

Building The 2×2 Matrix

When we evaluate how well a diagnostic imaging agent performs, we need measurable endpoints that capture the essence of what makes an imaging agent valuable. These endpoints help us understand whether an imaging agent truly delivers meaningful clinical benefit. These endpoints utilise the following categories when comparing an imaging agent to a truth standard:

- True Positive (TP): The imaging agent correctly detects the presence of disease.
- True Negative (TN): The imaging agent correctly detects the absence of disease.
- False Positive (FP): The imaging agent incorrectly indicates the presence of disease.
- False Negative (FN): The imaging agent incorrectly indicates the absence of disease.

These form the fundamental 2×2 contingency table:

	Disease Present (+)	**Disease Absent (-)**	**Total**
Test Positive (+)	TP (a)	FP (b)	m1 = a+b = TP+FP
Test Negative (-)	FN (c)	TN (d)	m2 = c+d = FN+TN
Total	n1 = a+c = TP+FN	n2 = b+d = FP+TN	N = a+b+c+d = TP+FP+FN+TN

The variables from this table can be used to estimate our diagnostic performance parameters defined below:

- Sensitivity (True Positive Rate): TP / (TP + FN) - The probability that a test result is positive when the participant has the disease.
- Specificity (True Negative Rate): TN / (TN + FP) - The probability that a test result is negative when the participant does not have the disease.
- Positive Predictive Value (PPV): TP / (TP + FP) - The probability that a participant has disease when the test result is positive.
- Negative Predictive Value (NPV): d/m2 = TN / (TN + FN) - The probability that a participant does not have the disease when the test result is negative.

Establishing Truth Standards

One of the most complex aspects of diagnostic imaging research involves defining what we accept as the truth standard i.e. the SoT or SoR (Bossuyt et al., 2015). A truth standard serves as an independent method for evaluating the same variable assessed by an investigational medical imaging agent. It is defined as a measure that reflects, or is strongly believed to reflect, the actual state of the participant or the true value of a measurement. Truth standards are essential for demonstrating that results from the imaging agent are valid and reliable, and for calculating key performance metrics such as sensitivity, specificity, and predictive values.

Defining SoT vs. SoR:

- Standard of Truth (SoT): Represents the most definitive determination of disease status, typically histopathology or other gold-standard diagnostic procedures. The SoT is considered the closest approximation to actual disease state.
- Standard of Reference (SoR): May include broader evidence beyond histopathology, such as clinical follow-up or composite endpoints when histopathology is impractical. The SoR acknowledges that obtaining histopathological confirmation for every lesion is often not possible.

From a practical standpoint, diagnostic standards are based on procedures considered more definitive than the investigational test in approximating the true disease state. For example, histopathology or long-term clinical outcomes may serve as acceptable standards for determining whether a mass is malignant. While diagnostic standards are not entirely error-free, they are generally regarded as definitive within the context of a clinical trial. However, any misclassification by the truth standard can introduce positive or negative biases into diagnostic performance measures (i.e., misclassification bias).

For this paper, we will focus on two standards:

Tissue-Based (Histopathology): Traditional pathological examination provides the most definitive diagnosis, but obtaining tissue samples can be invasive and sometimes impossible depending on lesion location or participant condition.

Long-term Clinical Follow-up (Follow-Up Imaging): This refers to repeat imaging performed after an initial scan to verify or clarify suspicious findings. For example, a screening MRI may be followed by a PET/CT with an approved PSMA tracer to confirm whether a lesion represents metastatic disease.

Each approach brings its own strengths and limitations that researchers must carefully consider during study design. The chosen reference standard fundamentally affects study validity and must align with the diagnostic tool's intended clinical application.
In clinical studies, histopathology is often seen as the greater body of evidence, but either or both truth sources may be used to classify lesions, regions or participants disease presence.

Assessing Diagnostic Performance at the Participant, Region and Sub-region Levels

Diagnostic performance can be evaluated at multiple hierarchical levels, each serving different clinical purposes. At the participant level, performance measures such as sensitivity and specificity indicate how accurately a diagnostic method can identify whether an individual has disease which can be a critical outcome for guiding participant management and treatment decisions. At the region level, we are assessing how well the imaging agent can localize disease within specific regions, which can be equally important in conditions where treatment planning depends on accurate mapping of disease extent or distribution e.g. identifying whether disease has re-appeared or spread.
Participant-Level Analysis: Correctly classifies individual participants as having or not having disease overall.
Region-Level Analysis: Correctly classifies disease presence within defined anatomical regions (e.g., pelvic lymph nodes [PLNs], prostate bed, distant metastases).
Subregion-Level Analysis: Provides granular assessment at specific anatomical locations within regions (e.g., lung vs liver vs bones in distant metastatic disease).
Often when sub-regions and regions are analyzed, only the lowest level of data is evaluated and collected, and these results are “rolled up” to the higher levels for the analysis. For example, if we have a region for metastatic disease, this can include many different sub-regions (or anatomical locations) such as skull, rib, shoulder and many others. The external imaging database would be setup to collect whether disease was present in the skull, rib, shoulder etc. and if any of those sub-regions were evaluated as positive, then the region for metastatic disease would be derived as positive. The full set of rules for analysis would be presented in the protocol and/or the Statistical Analysis Plan (SAP).

Central Evaluation and Reader Agreement

Central evaluation involves systematic review by independent and qualified readers external to the study, rather than relying on local site interpretations. This approach is used in diagnostic imaging trials to minimize variability and bias that can occur when multiple sites interpret the images with varying reader expertise. By having all imaging studies evaluated centrally by trained, independent readers following standardized criteria, the study ensures consistent assessment across all participants. This reduces inter-site variability, ensures consistent application of evaluation criteria, and provides standardized assessments critical for regulatory submissions by minimizing reader bias. Central evaluation is particularly important for pivotal trials where the imaging results serve as primary or co-primary endpoints, as regulatory authorities require evidence that the diagnostic performance is reliable and reproducible.
In later phase studies, three independent readers are often used. The majority read result may be used for primary and secondary endpoints, where the result where two out of the three readers agree is summarized. Or in other cases, the 3 readers results are reported separately, and the study is a win only if you have 2 out of 3 readers meet the endpoint. Another method is the 2+1 central read methodology which uses two independent readers who evaluate all participants images. When they agree, their consensus becomes final or the majority result. When they disagree, a third independent reader (adjudicator) evaluates to establish the majority result.

Example 1: Central Evaluation by Region (With Adjudication Required)

Region	Reader 1	Reader 2	Reader 3 (Adjudicator)	Majority Result
Prostatic	Positive	Negative	Positive	Positive
PLNs	Negative	Negative	---	Negative
Extra-pelvic sites (Metastases)	Positive	Positive	---	Positive

Example 2: Central Evaluation by Region (No Adjudication Required)

Region	Reader 1	Reader 2	Reader 3 (Adjudicator)	Majority Result
Prostatic	Positive	Positive	---	Positive
PLNs	Negative	Negative	---	Negative
Extra-pelvic sites (Metastases)	Positive	Positive	---	Positive

Measuring Agreement Between Readers

Agreement measures tell us how consistently different readers read the same imaging studies, providing us with information on reliability and reproducibility of detection rate assessment (Landis & Koch, 1977).

Reader concordance measures agreement between multiple readers. This is especially useful in larger clinical trials involving multiple readers.

Kappa Statistic (κ) provides an agreement score with chance agreement:

$\kappa = (\text{Observed Agreement} - \text{Expected Agreement}) / (1 - \text{Expected Agreement})$,

where Observed Agreement is the proportion of cases where readers reached the same conclusion and Expected Agreement is the proportion of agreement occurring by chance alone (e.g. by rolling up from different regions/sub-regions to obtain the same overall or region-level result).

We typically interpret Kappa values as follows (Landis & Koch, 1977):

- $\kappa < 0.20$: Poor agreement
- $\kappa = 0.21$-0.40: Fair agreement
- $\kappa = 0.41$-0.60: Moderate agreement
- $\kappa = 0.61$-0.80: Good agreement
- $\kappa = 0.81$-1.00: Excellent agreement

High inter-reader agreement ($\kappa > 0.60$) often demonstrates that imaging assessment criteria are well-defined and readers can apply them consistently. However, it's important to note that this isn't always true. For instance, when the marginal totals for either the positive or negative cases are very low, it's possible to observe high agreement but a low kappa value.

CASE STUDY: BED-PRO-301

BED-PRO-301 is a simulated Phase 3 open-label diagnostic imaging study enrolling 50 males with newly diagnosed prostate cancer. The study evaluates whether BED-PRO-001 (an investigational PSMA imaging agent) followed by a Positron Emission Tomography–Computed Tomography (PET-CT) scan achieves ≥50% sensitivity and ≥50% specificity for detection of prostate cancer in the PLNs, compared to a composite standard of truth.

Study Design Overview:

- Imaging agent: BED-PRO-001 (investigational PSMA imaging agent)
- Population: 50 participants with newly diagnosed prostate cancer
- Co-Primary Endpoints: Sensitivity and Specificity in PLNs
- Hypothesis: Sensitivity ≥50% and Specificity ≥50%
- Analysis Levels: Region-level: PLNs
- Truth Standard: Composite of histopathology and conventional imaging
- Central Evaluation: 3 independent readers

Defined region for primary endpoint: PLN - left and right PLNs are assessed for regional spread.

Simulated Efficacy Data

For this illustrative case study, central imaging results for the left and right PLNs were simulated for three independent readers. The simulation assumed the following for each reader:

- Reader 1
  - Left PLN Sensitivity 80% and Specificity 76%
  - Right PLN Sensitivity 72% and Specificity 68%
- Reader 2
  - Left PLN Sensitivity 68% and Specificity 88%
  - Right PLN Sensitivity 84% and Specificity 76%
- Reader 3
  - Left PLN Sensitivity 60% and Specificity 88%
  - Right PLN Sensitivity 48% and Specificity 88%

Data was simulated using the %simdata macro shown in Appendix 1 – Code for Simulation. The %simdata macro generated simulated diagnostic performance data for the PLNs for a given imaging reader and PLN laterality. The macro uses user-defined input parameters corresponding to the number of TPs, FNs, FPs, and TNs, along with the reader identifier (reader) and the laterality (lat).

For example, to produce the Reader 1 left and right PLN results the number of TPs, FNs, FPs and TNs were first visualized using the following 2x2 tables:

Example 3: Left and Right PLN 2x2 Tables

Left PLN	SoT (+)	SoT (-)	Total
BED-PRO-001 (+)	20 (TP)	6 (FP)	26
BED-PRO-001 (-)	5 (FN)	19 (TN)	24
Total	25	25	50

Sens = 20/25 (80%); Spec = 19/25 (76%)

Right PLN	SoT (+)	SoT (-)	Total
BED-PRO-001 (+)	18 (TP)	8 (FP)	26
BED-PRO-001 (-)	7 (FN)	17 (TN)	24
Total	25	25	50

Sens = 18/25 (72%); Spec = 17/25 (68%)

And then the following macro calls were used:

```
*** Example for Reader 1, Right PLN;
%simdata(tp=20,fn=5,fp=6,tn=19,reader=1,lat=Left)
*** Example for Reader 1, Left PLN;
%simdata(tp=18,fn=7,fp=8,tn=17,reader=1,lat=Right)
```

The data was then set together and separated into PET/CT reader imaging data and the SoT.

Example 4: Simulated Efficacy Data

USUBJID	TEST	TESTCD	READER	REGION	LOCATION	LATERALITY	RESULT
BED-PRO-301-001	PET/CT Qualitative Results	PETCT	Reader 1	PLN	Iliac Nodes	Right	Positive

USUBJID	TEST	TESTCD	READER	REGION	LOCATION	LATERALITY	RESULT
BEDPRO-301-001	SoT Qualitative Results	SoT		PLN	Iliac Nodes	Right	Positive

SDTM DATA ORGANIZATION

SDTM structuring is fundamental to capturing the complexity of diagnostic imaging trials while maintaining regulatory compliance. Understanding how central imaging review is performed, including which regions and subregions are assessed, what variables define anatomical locations, and how reader assessments are structured is critical for appropriate SDTM implementation. The key to successful SDTM mapping is recognizing that the data structure must reflect the hierarchical nature of imaging assessments (participant, region, subregion/laterality) and accommodate multiple independent reader evaluations. Variables such as TRLOC (location), TRLAT (laterality), and TREVAL (evaluator) become essential for capturing the granularity required for both regulatory review and subsequent statistical analysis. This careful structuring in SDTM provides the foundation for efficient ADaM dataset construction and ultimately enables accurate calculation of diagnostic performance metrics.

Core SDTM Domains

Procedures (PR): Documents treatment and imaging procedures (e.g., radical prostatectomy, conventional imaging such as Computed Tomography [CT] or Magnetic Resonance Imaging [MRI]). The PR domain provides the foundation for linking imaging findings with corresponding procedures.

Example 5: PR Domain

DOMAIN	USUBJID	PRSEQ	PRTRT	PRCAT	PRSCAT	PRSTDTC	PRLNKID
PR	BEDPRO-301-018	1	BED-PRO-001 PET/CT	IMAGING	BASELINE	2023-06-28	IMG-018
PR	BEDPRO-301-018	2	PLN Biopsy	PROCEDURE	BASELINE	2023-06-33	SOT-018

Tumor Results (TR): Documents imaging assessment results at region and subregion levels. Each row represents one anatomical assessment by a named reader, enabling tracking of individual reader determinations and subsequent agreement analysis.

Example 6: TR Domain

DOMAIN	USUBJID	TRSEQ	TRLNKID	TRTESTCD	TRSTRESC	TREVAL	TRLOC	TRLAT
TR	BEDPRO-301-018	1	IMG-018-PLN-Right	DETECT	Positive	READER 1	PLN	RIGHT
TR	BEDPRO-301-018	2	IMG-018-PLN-Left	DETECT	Positive	READER 1	PLN	LEFT
TR	BEDPRO-301-018	3	IMG-018-PLN-Right	DETECT	Positive	READER 2	PLN	RIGHT
TR	BEDPRO-301-018	4	IMG-018-PLN-Left	DETECT	Negative	READER 2	PLN	LEFT
TR	BEDPRO-301-018	5	IMG-018-PLN-Right	DETECT	Negative	READER 3	PLN	RIGHT
TR	BEDPRO-301-018	6	IMG-018-PLN-Left	DETECT	Negative	READER 3	PLN	LEFT

Microscopic Findings (MI): Records histopathological results when tissue confirmation is available. The domain structure mirrors the regional organization used in TR to support direct comparison between imaging and histological findings.

Note on Domain Usage: Per CDISC standards, the biopsy procedure itself is recorded in PR (the act of performing the procedure), while the histopathology results are recorded in MI (the findings from microscopic examination). This separation maintains appropriate domain usage and traceability.

Example 7: MI Domain

DOMAIN	USUBJID	MISEQ	MILNKID	MITESTCD	MISTRESC	MINAM	MILOC	MILAT
MI	BEDPRO-301-018	1	BIOPSY-018-PLN-Left	GLEASON	8	Grade Group 4	PLN	LEFT
MI	BEDPRO-301-018	2	BIOPSY-018-PLN-Left	MALIGN	POSITIVE	Adenocarcinoma	PLN	LEFT
MI	BEDPRO-301-018	3	BIOPSY-018-PLN-Right	GLEASON	7	Grade Group 2	PLN	RIGHT
MI	BEDPRO-301-018	4	BIOPSY-018-PLN-Right	MALIGN	POSITIVE	Adenocarcinoma	PLN	RIGHT

Adjudicated Truth Results (ZT): Is a custom domain and combines various sources of evidence including histopathology, follow-up imaging, and expert panel agreement to resolve final reference standard determinations per region or subregion being evaluated.

Example 8: ZT Domain

DOMAIN	USUBJID	ZTSEQ	ZTLNKID	ZTTESTCD	ZTSTRESC	ZTLOC	ZTLAT	ZTMETHOD
ZT	BEDPRO-301-018	1	SOT-018-PLN-Right	TRUTHDET	POSITIVE	PLN	RIGHT	HISTOPATHOLOGY
ZT	BEDPRO-301-018	2	SOT-018-PLN-Left	TRUTHDET	POSITIVE	PLN	LEFT	HISTOPATHOLOGY

The ZTMETHOD variable documents whether truth standard was established via histopathology or clinical follow-up, maintaining transparency and traceability essential for regulatory review.

ADAM DATASET CONSTRUCTION

ADEFF Structure

ADEFF (Analysis Dataset for Efficacy) combines imaging assessments with truth standard determinations to enable diagnostic performance computation (CDISC ADaM IG, 2021). The dataset maintains the hierarchical structure from SDTM while adding derived variables necessary for analysis.

Key ADEFF Variables:

- USUBJID: Unique subject identifier
- PARAMCD/PARAM: Analysis parameter identification. DPERPAR = Diagnostic Performance (Participant-Level), DPERREG = Diagnostic Performance (Region-Level), DPERSREG = Diagnostic Performance (Subregion-Level)
- AVALC: Character result from imaging agent test (e.g., Positive, Negative, Not Evaluable)
- BASEC: Character result from truth standard (e.g., Positive, Negative, Not Evaluable)
- AVALCAT1: Outcome classification (e.g., TRUE POSITIVE, TRUE NEGATIVE, FALSE POSITIVE, FALSE NEGATIVE)
- AREGION: Anatomical defined region (e.g., PLN, Prostatic, Extra-pelvic)
- ALOC: Specific anatomical location within region (e.g., Iliac Nodes, Obturator, Seminal Vesicle Bed) - populated at subregion level (DPERSREG)
- ALAT: Laterality within location (e.g., Left, Right, Bilateral) - populated at subregion level when applicable
- EVAL/EVALID: Reader type and identification. EVAL="Independent Assessor" for all central readers. EVALID distinguishes individual readers: "Reader 1", "Reader 2", "Adjudicator" (when disagreement occurs between Reader 1 and Reader 2)

Example 9: ADEFF Structure:

USUBJID	PARAMCD	AREGION	ALAT	AVALC	BASEC	AVALCAT1	EVALID
BEDPRO-301-018	DPERPAR			Positive	Positive	True Positive	Reader 1
BEDPRO-301-018	DPERPAR			Positive	Positive	True Positive	Reader 2
BEDPRO-301-018	DPERPAR			Negative	Positive	False Negative	Reader 3
BEDPRO-301-018	DPERPAR			Positive	Positive	True Positive	Majority
BEDPRO-301-018	DPERREG	PLN		Positive	Positive	True Positive	Reader 1
BEDPRO-301-018	DPERREG	PLN		Positive	Positive	True Positive	Reader 2
BEDPRO-301-018	DPERREG	PLN		Negative	Positive	False Negative	Reader 3
BEDPRO-301-018	DPERREG	PLN		Positive	Positive	True Positive	Majority
BEDPRO-301-018	DPERREG	PLN	LEFT	Positive	Positive	True Positive	Reader 1
BEDPRO-301-018	DPERREG	PLN	LEFT	Negative	Positive	False Negative	Reader 2
BEDPRO-301-018	DPERREG	PLN	LEFT	Negative	Positive	False Negative	Reader 3

USUBJID	PARAMCD	AREGION	ALAT	AVALC	BASEC	AVALCAT1	EVALID
BEDPRO-301-018	DPERREG	PLN	LEFT	Negative	Positive	False Negative	Majority
BEDPRO-301-018	DPERREG	PLN	RIGHT	Positive	Positive	True Positive	Reader 1
BEDPRO-301-018	DPERREG	PLN	RIGHT	Positive	Positive	True Positive	Reader 2
BEDPRO-301-018	DPERREG	PLN	RIGHT	Negative	Positive	False Negative	Reader 3
BEDPRO-301-018	DPERREG	PLN	RIGHT	Positive	Positive	True Positive	Majority

Note: PARAMCD: DPERPAR=Diagnostic Performance (Participant-Level); DPERREG=Diagnostic Performance (Region-Level). ALOC specifies the anatomical location within the region (e.g., Iliac Nodes). ALAT indicates left/right/bilateral and can be populated at region (e.g. for PLNs) or subregion level.

Data Flow:

- TR Domain → multiple records per reader (TREVAL identifies each reader);
- MI Domain → histopathology contributes to truth standard;
- ZT Domain → final truth standard populates BASEC; PARAMCD
- Hierarchy → organizes data at three levels (DPERPAR→DPERREG→ALAT within DPERREG);
- Reader Structure → EVAL="Independent Assessor" for all central readers, with EVALID identifying Reader 1, Reader 2, and Reader 3.
    - For the primary analysis, the majority read is presented and this is derived as a new EVALID (in our case Reader 4).
- AREGION, ALOC, and ALAT variables capture the anatomical hierarchy;
- Outcome Classification → AVALCAT1 derived by comparing AVALC to BASEC.

"Rolling Up" Results for Hierarchical Analysis

The assessments for PLNs are aggregated to regional levels using rules defined in the SAP and or/protocol. For example, in this case we use the following rule for left and right PLNs to derive the regional results:

- If either the left PLN or the right PLN is TP, then the PLN region is TP
- Otherwise, if both the left PLN and right PLN are negative, then the PLN region is TN
- Otherwise, if there is at least one FN, then the PLN region is FN
- Otherwise, the PLN region is a FP

Subject-level disease status is determined by aggregating all regional assessments. A similar rule to the PLN region is followed:

- If any REGION is TP, then the subject-level is TP
- Otherwise, if all REGIONs are negative, then the subject-level is TN
- Otherwise, if there is at least one TN REGION, then the subject-level is FN
- Otherwise, the REGION region is a FP

Note that the ordering in both cases may be dependent on the disease under study. In other cases, FP may be more clinically important than FN.

The rules for rolling up results from what is collected during central imaging review must be clearly defined in the protocol and/or SAP before data collection begins. This hierarchical aggregation maintains traceability from detailed subregion findings through regional summaries to overall subject-level disease status. The aggregation logic should reflect clinical practice and be justified based on how treatment decisions are made.

STATISTICAL ANALYSIS FOR DIAGNOSTIC PERFORMANCE

The statistical approach enables standardized analysis supporting regulatory submission requirements. Key considerations include appropriate confidence interval methodology and hypothesis testing frameworks for co-primary endpoints.

Primary Analysis Strategy:

1. Construct 2×2 contingency table comparing majority read to truth standard
2. Calculate sensitivity, specificity, PPV, NPV using standard formulae
3. Compute confidence intervals e.g. using Exact Clopper-Pearson or Wilson Score methods
4. Perform statistical tests against pre-specified thresholds (in our example, 50% and 50% for sensitivity and specificity respectively)

Primary Analysis Population

Populations for studies can differ in name and in definition. We give a couple of examples below.
The **Diagnostic Evaluable Set** (DES) population includes participants with both evaluable imaging assessments and evaluable truth standard determination. Participants with non-evaluable or missing results are excluded from primary analysis but may be included in sensitivity analyses where missing data is imputed.
The **Intent-to-Image** (ITI) population includes all enrolled and dosed participants who are expected to have a diagnostic scan. This analysis set may include participants with non-evaluable or missing results, and if this analysis set was used as the primary analysis the missing data would need to be imputed.
Either population can be included as the primary, but this can vary depending on the expected amount of missing data and discussion with regulatory agencies.

Handling Indeterminate Results

Participants or regions classified as "Not Evaluable" by either the imaging test or truth standard may be excluded from the primary efficacy analysis. Non-evaluable results can occur when imaging quality is inadequate for interpretation or when anatomical regions cannot be adequately assessed due to artifacts or prior surgery, or when truth standard determination cannot be established within the protocol-specified timeframe.
When imputation is performed on missing data, a worst-case analysis could be performed to assess robustness of primary conclusions and ensure that excluding indeterminate results does not artificially inflate performance metrics. For example, when assessing sensitivity if a participant had the disease by SoT but had a missing or non-evaluable scan, the worst-case imputation here would be imputing a FN which would be assuming the assessment using the imaging agent reported a negative result and thus reducing the sensitivity. A tipping point analysis may also be used to systematically explore how the results of a study might change under different assumptions for missing data. The analysis would look at varying the probability that missing data are imputed as a positive outcome in different increments from 0 to 1.

Hypothesis Testing

For co-primary endpoints, both hypotheses must be rejected to declare the study success:
Sensitivity Hypothesis:

- H_0: Sensitivity $\leq \sigma_{Sen}$
- H_a: Sensitivity $> \sigma_{Sen}$

Specificity Hypothesis:

- H_0: Specificity $\leq \sigma_{Spec}$
- H_a: Specificity $> \sigma_{Spec}$

Where σ_{Sen} and σ_{Spec} are performance thresholds for sensitivity and specificity respectively. For example, in our simulated study σ_{Sen} and σ_{Spec} are equal to 50% and 50% respectively. A statistical test is used to assess each endpoint for success, and the lower bound of the confidence interval must be greater than the respective threshold to class the endpoint as a success.

Statistical Tests and Confidence Intervals

There are various methods that can be used to calculate p-values and confidence intervals, and both should be pre-specified in the protocol and/or the SAP. Different methods may be used dependent on the study design, and we give a couple of examples below:

P-Value Test	Confidence Interval	Used When…
Exact binomial test	Exact (Clopper–Pearson)	Sample size is small and/or proportions are extreme (near 0 or 1).
Z-test	Wilson Score or Wald	Performs adequately when n is large and p is not close to 0 or 1.

SAS Example 1 – Calculating Exact Confidence Intervals and P-Values for Sensitivity:

```
/* Subset on SoT Positive participants for Sensitivity */
proc sort data=adeff out=adeff_sens;
    by evalid;
    where paramcd = "DPERREG"
          and aregion = "PLN"
          and alat = ""
          and basec = "POSITIVE";
run;

/* Frequency counts of response categories by evaluation ID */
proc freq data=adeff_sens noprint;
    tables avalc / out=counts01(drop=percent);
    by evalid;
run;

/* Calculate exact binomial confidence intervals and test vs. H0: Sens <= 50% */
proc freq data=counts01 noprint;
    tables avalc / binomial(level="Positive") alpha=0.05 binomial(p=0.50);
    by evalid;
    weight count / zeros;
    exact binomial;
    output out=sum01(keep=evalid n xl_bin xu_bin pr_bin) binomial;
run;

/* Format CI and p-value for presentation */
data sum02;
    set sum01;
    length sum $50;

    if xu_bin ne 1 then
        sum = compress("[" || put(round(100*xl_bin,0.1),9.1) || "% - " ||
                        put(round(100*xu_bin,0.1),9.1) || "%]");
    else
        sum = compress("[" || put(round(100*xl_bin,0.1),9.1) || "% - 100%]");

    pval = put(round(pr_bin,0.001),6.3);
    if pval = " 0.000" then pval = "<0.001";
run;
```

BED-PRO-301 RESULTS

Primary Analysis Results

The following is the 2×2 Table from the primary analysis: Region-Level PLN (N=50) [Majority Read]:

PLNs	SoT (+)	SoT (-)	Total
BED-PRO-001 (+)	18 (TP)	6 (FP)	24
BED-PRO-001 (-)	7 (FN)	19 (TN)	26
Total	25	25	50

The below is the summary table that includes the results from all readers, and the majority read.

Example 10: Table Summary of Primary Analysis Results

	Reader 1 (N=50)	Reader 2 (N=50)	Reader 3 (N=50)	Majority (N=50)
Number of Participants in the Analysis	50	50	50	50
True Positive (TP)	20 (40.0%)	21 (42.0%)	15 (30.0%)	18 (36.0%)
False Positive (FP)	8 (16.0%)	6 (12.0%)	3 (6.0%)	6 (12.0%)
True Negative (TN)	17 (34.0%)	19 (38.0%)	22 (44.0%)	19 (38.0%)
False Negative (FN)	5 (10.0%)	4 (8.0%)	10 (20.0%)	7 (14.0%)
Sensitivity (%)	20/25 (80.0%)	21/25 (84.0%)	15/25 (60.0%)	18/25 (72.0%)
(95% CI)	[59.3, 93.2]	[63.9, 95.5]	[38.7, 78.9]	[50.6, 87.9]
p-value (H0: Sens <= 50%)	0.0020	<0.001	0.2122	0.0216
Specificity (%)	17/25 (68.0%)	19/25 (76.0%)	22/25 (88.0%)	19/25 (76.0%)
(95% CI)	[46.5, 85.1]	[54.9, 90.6]	[68.8, 97.5]	[54.9, 90.6]
p-value (H0: Spec <= 50%)	0.054	0.007	<0.001	0.007

Primary Efficacy Analysis

The primary analysis evaluated the sensitivity and specificity of BED-PRO-001 for PLN disease based on majority read at the region level. Among the 50 participants in the analysis population, 25 had histologically confirmed PLN disease (SoT positive). BED-PRO-001 correctly identified 18 of these 25 participants with disease, yielding a sensitivity of 72.0% (95% exact CI: 50.6%, 87.9%). Statistical testing against the pre-specified threshold of 50% resulted in a one-sided p-value of 0.0216, demonstrating that the sensitivity significantly exceeded the threshold of 50%. Among the 25 participants with negative SoT, BED-PRO-001 correctly identified 19 participants as disease-free, yielding a specificity of 76.0% (95% exact CI: 54.9%, 90.6%). Statistical testing against the pre-specified threshold of 50% resulted in a one-sided p-value of 0.0073, demonstrating that the specificity significantly exceeded the threshold of 50%. The study successfully rejected the null hypothesis for both endpoints. It is worth noting that for sensitivity, Reader 1 and 2 met the endpoint, whereas for specificity, Readers 2 and 3 met the endpoint. This could be flagged by regulators, but the criteria for confirming attainment of the co-primary endpoints should be prospectively defined in the protocol and/or the SAP.

Inter-Reader Reliability

Inter-reader reliability was assessed to evaluate the reproducibility of BED-PRO-001 image interpretation across the three independent readers. Pairwise Cohen's kappa statistics ranged from 0.61 to 0.88, with an average of 0.71, indicating agreement according to Landis & Koch. Pairwise agreement percentages were high: Reader 1 vs Reader 2 (94.0%), Reader 1 vs Reader 3 (80.0%), and Reader 2 vs Reader 3 (82.0%), yielding an overall agreement of 85.3%. Concordance among all three readers was observed in 39 of 50 cases (78.0%), demonstrating that most participants received consistent interpretations across readers. These findings demonstrate reliable and reproducible assessments, supporting the validity of the majority read approach used for the primary efficacy analysis. Code for reporting Cohen’s Kappa can be found in Appendix 2 - Calculating Cohen's Kappa.

DISCUSSION

This paper describes how diagnostic imaging data can be collected and summarized to produce standard diagnostic performance endpoints and how those results are summarized. The simulated study BED-PRO-301 presents how this framework can be used to resent robust study results.

The key to successful implementation is establishing a consistent data flow from the very beginning of study design. The framework presented here is one approach to organizing imaging data in SDTM/ADaM, but it requires that the Case Report Form (CRF) and external imaging data are structured consistently with clear rules for how participant, region, and location assessments should be conducted and aggregated. Statisticians, with support from programmers, should be involved from the beginning of study design to ensure the data flow is clear and feasible.

Key considerations include:

1. Define assessment structure (participant-level vs. region-level vs. both) based on clinical objectives before finalizing CRF and external imaging database
2. Establish clear anatomical definitions for all regions and subregions and define these in the protocol and/or SAP
3. Pre-specify aggregation rules (i.e. rolling up rules) and define these in the protocol and/or SAP
4. Coordinate with imaging vendors early to ensure data transfer specifications align with planned EDC and SDTM structure
5. Create comprehensive mock datasets for external imaging data (e.g. using %simdata). Often it can take external imaging vendors a long time to produce usable test data, so creating your own early means programming can start earlier.

While this paper focuses on prostate cancer with region-level analysis of PLN, the principles adapt across indications and type of disease (e.g. initial diagnosis vs. metastatic disease), though the hierarchical structure remains consistent. Similarly, some studies may require participant-level analysis only, simplifying the PARAMCD structure while maintaining the same SDTM foundation.

REFERENCES

1. Bossuyt PM, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. Radiology. 2015;277(3):826-832.
2. Clinical Data Interchange Standards Consortium (CDISC). Study Data Tabulation Model Implementation Guide (SDTM-IG), Version 3.4. 2023.
3. Clinical Data Interchange Standards Consortium (CDISC). Analysis Data Model Implementation Guide (ADaM-IG), Version 1.3. 2021.
4. Faroz L, Kola S. Demystifying SDTM OE, MI, and PR Domains. PharmaSUG 2020 - Paper DS-110.
5. FDA Guidance for Industry: Developing Medical Imaging Drug and Biological Products Part 3. June 2022.
6. Inácio V, et al. Statistical Evaluation of Medical Tests. Annual Review of Statistics and Its Application. 2020;7:1-30.
7. Koduru B., Pitchuka B. Know your PET: From the Scans to SDTM Datasets. PharmaSUG 2018 - Paper MD-03.
8. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977;33(1):159-174.
9. Sharkey AR, et al. Novel Molecular Imaging Approaches in Oncology. Current Opinion in Oncology. 2020;32(1):45-53.
10. Zhou XH, et al. Statistical Methods in Diagnostic Medicine. 2nd ed. Wiley; 2011.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

James Sykes
Blue Earth Diagnostics Ltd
Oxford, United Kingdom
Email: James.Sykes@blueearthdx.com

Srinivas Lanka
Blue Earth Diagnostics Ltd
Oxford, United Kingdom
Email: Shri.Lanka@blueearthdx.com

Appendix 1 – Code for Simulation

```
/*************************************************************************/
/* Macro: simdata                                                        */
/* Purpose:                                                              */
/*   Simulate diagnostic performance data for a single anatomical region */
/*   (Pelvic Lymph Nodes, PLN) for a given laterality and image reader.  */
/*                                                                       */
/*   The macro generates individual subject-level records representing   */
/*   true positives, false negatives, true negatives, and false positives */
/*   based on specified counts, and then derives performance metrics     */
/*   including sensitivity, specificity, PPV, and NPV.                   */
/*                                                                       */
/* Parameters:                                                           */
/*   tp      = Number of True Positives                                  */
/*   fn      = Number of False Negatives                                 */
/*   tn      = Number of True Negatives                                  */
/*   fp      = Number of False Positives                                 */
/*   reader  = Reader number (numeric identifier)                        */
/*   lat     = Laterality (e.g., 'Left' or 'Right')                      */
/*************************************************************************/
%macro simdata(tp=, fn=, fp=, tn=, reader=, lat=);

  /*********************************************************************/
  /* STEP 1: Generate subject-level simulated data                     */
  /* Each record represents a unique subject-reader-region combination.  */
  /*********************************************************************/
  data simdata_pln_&lat._r&reader.;
    length result_ovr $20 usubjid $14 reader $10;
    testcd = 'OVERALL';
    test   = 'Overall Qualitative Results';
    region = 'PLN';
    loc    = 'Iliac Nodes';
    lat    = "&lat.";
    readern = &reader.;
    reader  = cats('Reader ', put(readern, 1.));
    i = 0;
    /* --- Generate True Positive cases --- */
    do j = 1 to &tp.;
        i + 1;
        subj = j;
        usubjid = cats('BEDPRO-301-', put(subj, z3.));
        result_petct = 'Positive';
        result_sot   = 'Positive';
        result_ovr   = 'True Positive';
        output;
    end;
    /* --- Generate False Negative cases --- */
    do j = 1 to &fn.;
        i + 1;
        subj = &tp. + j;
        usubjid = cats('BEDPRO-301-', put(subj, z3.));
        result_petct = 'Negative';
        result_sot   = 'Positive';
        result_ovr   = 'False Negative';
        output;
    end;
```

```
  /* --- Generate True Negative cases --- */
  do j = 1 to &tn.;
      i + 1;
      subj = &tp. + &fn. + j;
      usubjid = cats('BEDPRO-301-', put(subj, z3.));
      result_petct = 'Negative';
      result_sot   = 'Negative';
      result_ovr   = 'True Negative';
      output;
  end;
  /* --- Generate False Positive cases --- */
  do j = 1 to &fp.;
      i + 1;
      subj = &tp. + &fn. + &tn. + j;
      usubjid = cats('BEDPRO-301-', put(subj, z3.));
      result_petct = 'Positive';
      result_sot   = 'Negative';
      result_ovr   = 'False Positive';
      output;
  end;

  /* Clean up temporary variables */
  drop i j subj;
run;

/*********************************************************************/
/* STEP 2: Summarise counts by reader, region, laterality, and result  */
/*********************************************************************/
proc freq data=simdata_pln_&lat._r&reader. noprint;
  tables reader*region*lat*result_ovr
         / out=results_pln_&lat._r&reader.01(drop=percent);
run;

/*********************************************************************/
/* STEP 3: Transpose counts to a wide format for metric calculation    */
/* Each diagnostic category (TP, FN, TN, FP) becomes a separate column.*/
/*********************************************************************/
proc transpose data=results_pln_&lat._r&reader.01
               out=results_pln_&lat._r&reader.02(drop=_:);
  id result_ovr;   /* Column headers become result categories */
  var count;       /* The variable to transpose */
  by reader region lat;
run;

/*********************************************************************/
/* STEP 4: Calculate diagnostic performance metrics                    */
/* Metrics are calculated as percentages and rounded to one decimal.   */
/*********************************************************************/
data results_pln_&lat._r&reader.;
  set results_pln_&lat._r&reader.02;

  /* Sensitivity: TP / (TP + FN) */
  sens = round(100 * true_positive / (true_positive + false_negative), 0.1);

  /* Specificity: TN / (TN + FP) */
  spec = round(100 * true_negative / (true_negative + false_positive), 0.1);

  /* Positive Predictive Value (PPV): TP / (TP + FP) */
  ppv  = round(100 * true_positive / (true_positive + false_positive), 0.1);
```

```
    /* Negative Predictive Value (NPV): TN / (TN + FN) */
    npv  = round(100 * true_negative / (true_negative + false_negative), 0.1);
  run;


  /*************************************************************************/
  /* STEP 5: Clean up intermediate datasets to maintain a tidy workspace */
  /*************************************************************************/
  proc datasets library=work nolist;
    delete results_pln_&lat._r&reader.01 results_pln_&lat._r&reader.02;
  quit;

%mend simdata;

*** Example for Reader 1, Right PLN;
%simdata(tp=20,fn=5,fp=6,tn=19,reader=1,lat=Left)

*** Example for Reader 1, Left PLN;
%simdata(tp=18,fn=7,fp=8,tn=17,reader=1,lat=Right)

*** Example for Reader 2, Right PLN;
%simdata(tp=17,fn=8,fp=3,tn=22,reader=2,lat=Left)

*** Example for Reader 2, Left PLN;
%simdata(tp=21,fn=4,fp=6,tn=19,reader=2,lat=Right)

*** Example for Reader 3, Right PLN;
%simdata(tp=15,fn=10,fp=3,tn=22,reader=3,lat=Left)

*** Example for Reader 3, Left PLN;
%simdata(tp=12,fn=13,fp=3,tn=22,reader=3,lat=Right)
```

Appendix 2 - Calculating Cohen's Kappa

```
/*********************************************************************************
* Inter-Reader Reliability - Cohen's Kappa
*********************************************************************************/

/* Get data in wide format */
proc sql;
    create table work.rw as
    select a.USUBJID, a.AVALC as R1, b.AVALC as R2, c.AVALC as R3
    from (select * from adam.adeff where PARAMCD='DPERREG' and ALAT='' and
AREGION='PLN' and EVALID='Reader 1') a
    inner join (select * from adam.adeff where PARAMCD='DPERREG' and ALAT='' and
AREGION='PLN' and EVALID='Reader 2') b on a.USUBJID=b.USUBJID
    inner join (select * from adam.adeff where PARAMCD='DPERREG' and ALAT='' and
AREGION='PLN' and EVALID='Reader 3') c on a.USUBJID=c.USUBJID;
quit;

/* Calculate kappa  */
proc freq data=work.rw;
    tables R1*R2 / agree;
    ods output KappaStatistics=k12;
run;
proc freq data=work.rw;
    tables R1*R3 / agree;
    ods output KappaStatistics=k13;
run;
proc freq data=work.rw;
    tables R2*R3 / agree;
    ods output KappaStatistics=k23;
run;

/* Extract kappa values - where Statistic='Simple Kappa' */
proc sql noprint;
    select Value into :k12
    from k12 where Statistic = 'Simple Kappa';
    select Value into :k13 from k13
    where Statistic = 'Simple Kappa';
    select Value into :k23 from k23
    where Statistic = 'Simple Kappa';
quit;

/* Calculate agreement */
data work.agr;
    set work.rw;
    a12 = (R1=R2);
    a13 = (R1=R3);
    a23 = (R2=R3);
    all3 = (R1=R2 and R2=R3);
run;
proc sql noprint;
    select count(*),sum(a12),sum(a13),sum(a23),sum(all3)
    into :n,:n12,:n13,:n23,:nall
    from work.agr;
quit;
/* Display results */
data _null_;
    file print;
    avgk = (&k12 + &k13 + &k23) / 3;
```

```
    overall = (&n12 + &n13 + &n23) / (3 * &n) * 100;
    perfect = &nall / &n * 100;
    length interp $30;
    if avgk >= 0.81 then interp = "Almost perfect agreement";
    else if avgk >= 0.61 then interp = "Substantial agreement";
    else if avgk >= 0.41 then interp = "Moderate agreement";
    else if avgk >= 0.21 then interp = "Fair agreement";
    else interp = "Slight agreement";
    put "Average Kappa:                  " avgk 5.3 " (" interp +(-1) ")";
    put "Overall Agreement:              " overall 4.1 "%";
    put "Perfect Concordance (all 3):    " perfect 4.1 "% (" "&nall" "/" "&n" "
subjects)";
run;
```