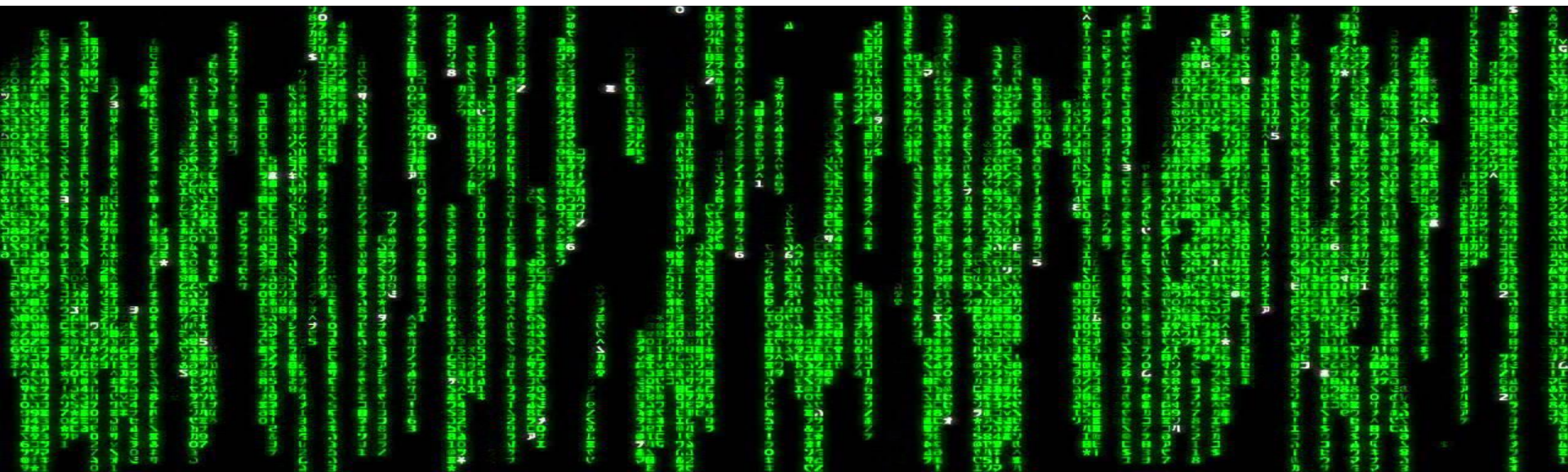


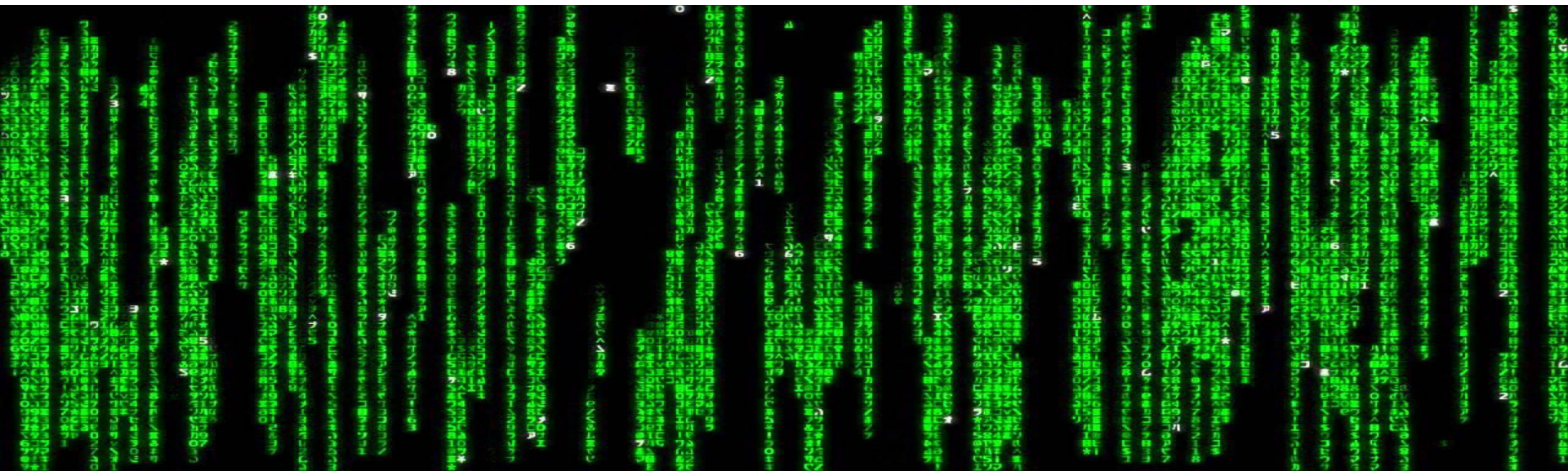
Statistical Endpoints and Analysis in Diagnostic Imaging: The Matrix Has You... and Your Data

James Sykes and Srinivas Lanka



Today's Journey

- Introduction to **Diagnostic Imaging**
- **Diagnostic Performance and Endpoints**
- Introducing the **BED-PRO-301** study
- **SDTM and ADaM Standards**
- **Statistical Analysis & Results**
- **Key Takeaways**





BLUE EARTH
DIAGNOSTICS
A Bracco Company

Authors

James Sykes

Director of Biometrics

Blue Earth Diagnostics Ltd

Srinivas Lanka

Principal Statistical Programmer

Blue Earth Diagnostics Ltd



Introduction to Diagnostic Imaging

In oncology, **precise lesion detection** via **diagnostic imaging** plays a **critical role** in **shaping** participant **outcomes**.

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 of lesions **influences clinical decision-making** and resource allocation.

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

Introduction to Diagnostic Imaging

- **PET/CT** imaging **combines** positron emission tomography (**PET**) with computed tomography (**CT**) to provide both **metabolic** and **anatomical information** in a **single scan**.
- **PET** uses a small amount of **radioactive tracer** to **highlight areas of increased cellular activity**, which can **indicate early disease processes** such as cancer.
- The **CT** portion supplies **detailed anatomical images** that help **precisely localize** these active regions, **improving diagnostic accuracy** and allowing clinicians to **monitor treatment response** and disease **progression**.



Diagnostic Performance and Endpoints

- **How do we assess** if a new diagnostic imaging agent is better than what's already on the market?
- **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**.



Diagnostic Performance and Endpoints

- Diagnostic performance metrics often act as **co-primary endpoints** in **Phase 2** and **3** trials.
- **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** (e.g. sensitivity) and **rule it out when it is absent** (e.g. 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.

Diagnostic Performance and Endpoints

At the core of evaluating diagnostic imaging is a basic 2×2 contingency table...

	Disease Present (+)	Disease Absent (-)	Total
Test Positive (+)	TP	FP	TP+FP
Test Negative (-)	FN	TN	FN+TN
Total	TP+FN	FP+TN	N

- 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.

Diagnostic Performance and Endpoints

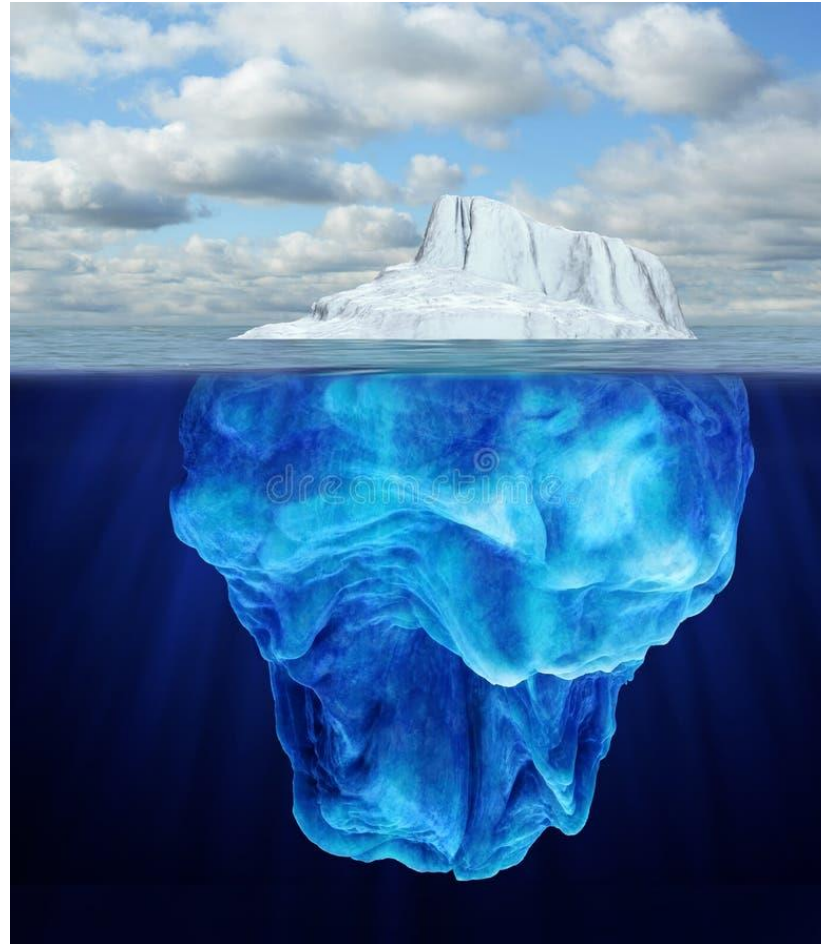
At the core of evaluating diagnostic imaging is a basic 2×2 contingency table...

	Disease Present (+)	Disease Absent (-)	Total
Test Positive (+)	TP	FP	TP+FP
Test Negative (-)	FN	TN	FN+TN
Total	TP+FN	FP+TN	N

- **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**.

Diagnostic Performance and Endpoints

At the core of evaluating diagnostic imaging is a basic 2×2 contingency table...



Diagnostic Performance and Endpoints



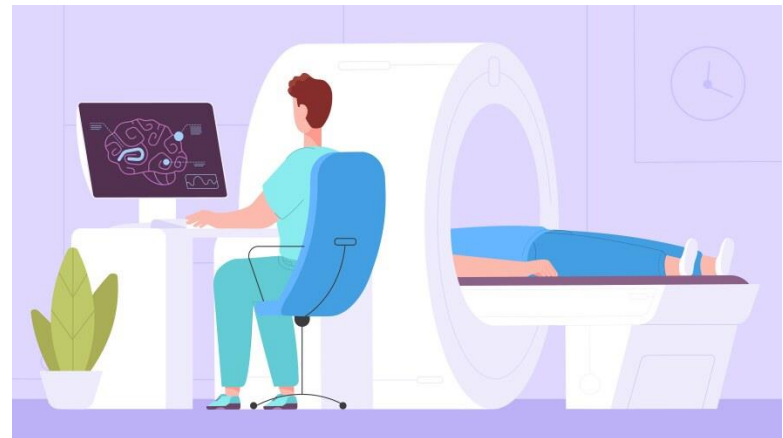
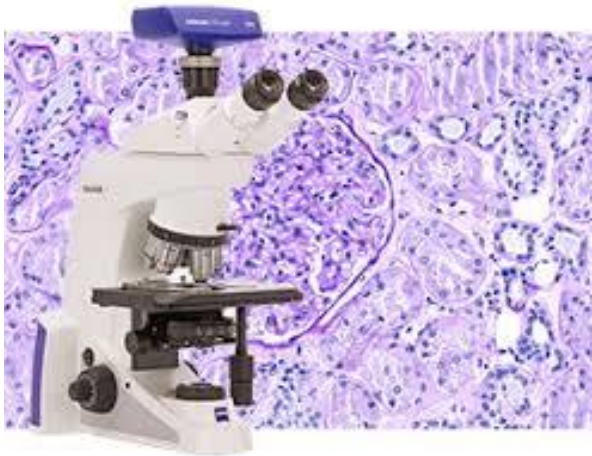
Establishing Truth Standards

- One of the **most complex aspects** of diagnostic imaging research involves defining what we accept as the **truth standard**
- 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.

Diagnostic Performance and Endpoints

Establishing Truth Standards

- Two examples in Oncology:
 - 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**.



Diagnostic Performance and Endpoints

Central Evaluation and Reader Agreement

- Involves **systematic review by independent and qualified readers** external to the study, rather than relying on local site interpretations.
- **Minimize variability and bias** that can occur when multiple sites interpret the images with varying reader expertise.
- **Ensures consistent assessment across all participants.**
- **Reduces inter-site variability, ensures consistent application of evaluation criteria**, and provides **standardized assessments** critical for regulatory submissions by **minimizing reader bias**.
- In later phase studies, **three independent readers** are often used with the **majority read result** used for **primary** and **secondary endpoints**, where the result where two out of the three readers agree is summarized.

Introducing the BED-PRO-301 study

- **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 PET-CT scan** achieves **≥50% sensitivity** and **≥50% specificity** for **detection of prostate cancer** in the Pelvic Lymph Nodes (**PLNs**), compared to standard of truth.



Introducing the BED-PRO-301 study

- **Central imaging results** for the left and right PLNs were **simulated** for three independent readers. The **simulations used** the number of **T**Ps, **F**Ns, **F**Ps, and **T**Ns from case examples for each Reader.
- E.g. Reader 1

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%)

- Data was simulated using the **%simdata macro** which is presented in the paper. The %simdata macro **generates simulated diagnostic performance data** for the PLNs for a given imaging reader and PLN laterality.

SDTM Standards

- SDTM structuring is fundamental to capturing the complexity of diagnostic imaging trials
- 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
- This careful structuring in SDTM provides the **foundation for efficient ADaM dataset construction** and ultimately enables accurate **calculation of diagnostic performance** metrics

SDTM Standards

Domain	Purpose	Structure	Key Variables
TR (Tumor Results)	Individual reader assessments	One record per reader per region	TRSTRESC, TRLOC, TRLAT, TREVAL
MI (Microscopic)	Histopathology results	When tissue available	MISTRESC, MILOC, MIMETHOD
ATR (Truth Results)	Final truth standard	One record per region	ATRSTRESC, ATRLOC, ATRMETHOD
PR (Procedures)	Imaging procedures performed	Conventional imaging follow-up	PRSTRESC, PRLOC, PRDAT

SDTM Standards

TR Domain: Multiple Reader Assessments

Example: Subject 001, Left Pelvic Lymph Node

USUBJID	TRLOC	TRLAT	TREVAL	TRSTRESC
BEDPRO-301-001	PELVIC LN	LEFT	READER 1	Positive
BEDPRO-301-001	PELVIC LN	LEFT	READER 2	Positive
BEDPRO-301-001	PELVIC LN	LEFT	READER 3	Negative

Majority Read Result: Positive (2/3 agreement)

SDTM Standards

ATR Domain: Truth Standard Determination

Data for standard of truth can come directly from the CRF or it may come externally after review of all data by an external independent reader.

Example ATR Records

USUBJID	ATRLOC	ATRLAT	ATRSTRES C	ATRMETH OD
001	PELVIC LN	LEFT	Present	HISTOPAT HOLOGY
002	PELVIC LN	RIGHT	Absent	IMAGING FOLLOWU P

BED-PRO-301: 100% had definitive truth standard

Example Truth Standard Algorithm

1. If histopathology available & conclusive
→ **Use histopathology**
2. If histopathology not available
→ **Use conventional imaging (e.g. CT/MRI)**
3. If no definitive determination
→ **Mark NOT EVALUABLE (exclude)**

Rules should be defined in the protocol and/or SAP

ADaM Standards

ADEFF: Analysis Dataset Structure

Variable	Purpose	Example Values
PARAMCD	Analysis level	DPERREG (region), DPERSREG (subregion)
AREGION	Anatomical region	PLN, PLN-L, PLN-R
EVALID	Reader identification	"Majority Read" (primary analysis)
AVALC / AVAL	Imaging result	"Positive"/"Negative"
BASEC / BASE	Truth standard	"Positive"/"Negative"
AVALCAT1	Outcome classification	"True Positive" "False Negative" etc.

AVALCAT1 is derived based on AVALC and BASEC values

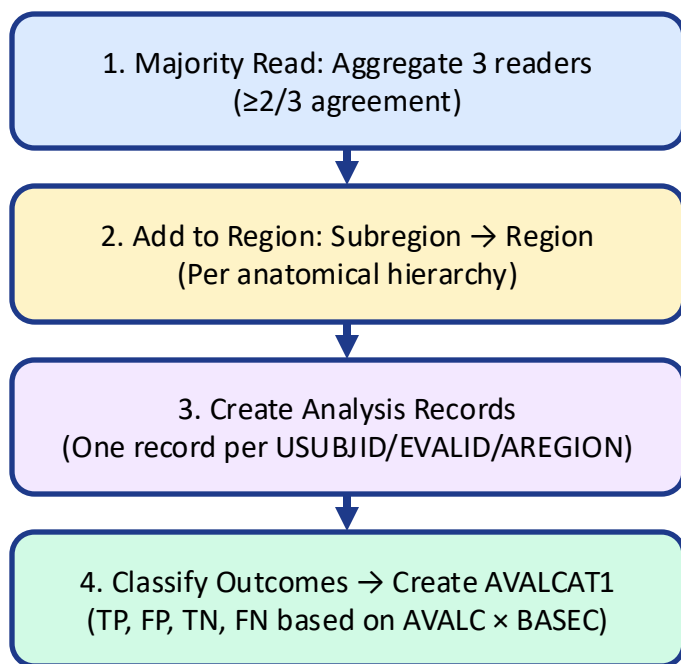
ADaM Standards

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
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

ADEFF: Analysis Dataset Structure

ADaM Standards

Data Flow: SDTM → ADaM Derivation Process



**Result: ADEFF dataset ready for analysis with
AVALCAT1 classification**

Study Objectives & Statistical Framework

Objective Type	Description
Primary	Co-primary endpoints: • Sensitivity >50% (H_0: Sens $\leq$ 50%) • Specificity >50% (H_0: Spec $\leq$ 50%)

Statistical Methods:

- One-sided exact binomial tests at $\alpha=0.025$
- Exact 95% CIs using Clopper-Pearson method

Statistical Analysis Approach

Analysis Population:

- All participants with EVALUFL='Y' in analysis dataset
- Majority read (primary) and individual readers (secondary)

Endpoints:

- For Sensitivity: Test proportion >70% (one-sided $\alpha=0.025$)
- For Specificity: Test proportion >50% (one-sided $\alpha=0.025$)

Example SAS Calculation:

```
SENS_PVAL = 1 - probbnml(0.70, N_DIS, TP - 1);  
SPEC_PVAL = 1 - probbnml(0.50, N_NOTDIS, TN - 1);
```

Hypothesis Testing

Sensitivity Hypothesis:

- • H_0 : Sensitivity $\leq 70\%$
- • H_1 : Sensitivity $> 70\%$
- • One-sided test at $\alpha = 0.025$

Specificity Hypothesis:

- • H_0 : Specificity $\leq 50\%$
- • H_1 : Specificity $> 50\%$
- • One-sided test at $\alpha = 0.025$

Confidence Intervals:

- • Exact 95% CIs using Clopper-Pearson method
- • Conservative coverage appropriate for regulatory submissions
- • Valid for small sample sizes

Primary Endpoint: Majority Read

2×2 Contingency Table (N=50)

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

Performance Metrics: Detailed Results

Metric	Estimate (%)	95% CI (Exact)	p-value	Endpoint Met?
Sensitivity*	72.0	[50.6, 87.9]	0.512	Yes ✓
Specificity*	76.0	[54.9, 90.6]	0.007	Yes ✓

** Co-primary endpoints. Both must meet thresholds for complete study success.*

One-sided tests: Sensitivity >70% (failed), Specificity >50% (passed)

Sensitivity and Specificity by Reader and Majority Assessment

Metric	Reader 1 (N = 50)	Reader 2 (N = 50)	Reader 3 (N = 50)	Majority (N = 50)
Number of Patients in 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	0.1935	0.0905	0.9022	0.5118
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	0.054	0.007	<0.001	0.007

*Note: All estimates derived using exact binomial (Clopper–Pearson) confidence intervals.
 Thresholds: Sensitivity > 70%, Specificity > 50%.*

Reader Agreement

- Multiple readers ensure unbiased assessments
- 2+1 methodology: Adjudicator resolves disagreements
- Kappa Statistic: Good agreement ($\kappa = 0.76$)
- Nearly 9/10 cases consistent across readers

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

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

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

```
proc freq data=reader_comparison;  
  tables reader1*reader2 / agree;  
  output out=kappa_results agree;  
run;
```

Reader Agreement Analysis

- Multiple readers ensure unbiased assessments
- 3+1 methodology: Adjudicator (if no majority)
- Kappa Statistic: Good agreement ($\kappa = 0.74$)
- Nearly 80% of cases: All readers consistent

Metric	Value	Interpretation
Fleiss' Kappa	0.78	Good (>0.60)
Overall agreement	87.5%	High consistency
Perfect 3-way	78%	39 of 50 cases
Pair: R1 vs R2	94.0%	Excellent
Pair: R1 vs R3	80.0%	Good
Pair: R2 vs R3	82.0%	Good

Discussion

- Our 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 present 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 defined in the protocol and/or SAP
- Statisticians, with support from programmers, should be involved from the beginning of study design to ensure the data flow is clear and feasible.

Discussion

Key considerations include:

1. Define assessment structure (participant vs. region level) 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.

Thank You

For detailed methodology, SAS code examples, and comprehensive case study results, please contact the authors.

PRE_AS10 PHUSE EU Connect 2025