A Unified Bayesian Approach to Uncertainty, Performance and Fairness

Trustworthy AI in Medicine

Bruno Boulanger, Nils Boulanger

SANAITIO – Belgium

PSI Conference 2026 | Belfast | June 16

Sanaitio

The Question: Will the Device Work on New Sites and Regions?

An AI diagnostic device returns, for a new patient x_{new}, a probability/decision. Three questions actually matter:

- **Inverse question (routine use):** given the device output, what is the true status?
  $\rightarrow\ p(y_{\text{new}} \mid x_{\text{new}}, D)$
- **Transportability:** what will performance be at a site/region *not* seen in training?
- **Exchangeability:** can sites be pooled, or has the clinical relationship itself changed?

"There Is No Such Thing as a Validated Prediction Model"

Van Calster *et al. BMC Medicine* (2023) 21:70
https://doi.org/10.1186/s12916-023-02779-w

BMC Medicine

OPINION Open Access

There is no such thing as a validated prediction model

Ben Van Calster[1,2,3], Ewout W. Steyerberg[1], Laure Wynants[1,2,4] and Maarten van Smeden[5*]

Performance is inherently heterogeneous (Van Calster et al., 2023):

1. **Populations vary**
2. **Measurements vary**: scanners, assays, operators
3. **Change over time**

A single external validation is only a *snapshot*.

$\Rightarrow$ Assessment must be **distributional** and **local**: quantify heterogeneity, borrow across sites, monitor and update, on $p(y_{\text{new}} \mid x_{\text{new}}, \text{site}, D, \text{data}_{\text{site}})$.

The Problem: A Babel of Disconnected Measures

Discrimination	Calibration	Classification	Clinical utility
AUROC	O:E ratio	Accuracy	Net benefit
AUPRC	Cal. intercept	Balanced acc.	Standardized NB
pAUROC	Cal. slope	Youden J	Expected cost
	ECE	Cohen κ	
	ICI	Diagnostic OR	
	ECI	Sensitivity	
	Hosmer–Lemeshow	Specificity	
		PPV (precision)	
		NPV	
		F1 ($F\beta$)	
		MCC	

Van Calster et al. (2025): 32 measures across 5 domains, with conflicting recommendations across guidelines.
One distributional object unifies them all.

The Foundational Object: The Bayesian Predictive Distribution

$$p(y_{\text{new}} \mid x_{\text{new}}, D) = \int p(y_{\text{new}} \mid x_{\text{new}}, \theta)\; p(\theta \mid D)\, d\theta$$

- Integrates over the posterior probability $p(\theta \mid D)$
- The *complete* statement of what the AI "knows" about a patient's outcome
- **Every concept in this talk is a functional of this single object**

AI = Measurement Instrument | Training domain = Validated operating range

Bridge from Analytical Validation: TAE & MU

Measurement science, ISO/GUM metrology and USP general chapters, already formalised the same inverse problem:

Validation → "forward" direction

$$P(X \mid \mu)$$

$$\text{Se} = P(\text{Result}+ \mid \text{Sample}+)$$
$$\text{Sp} = P(\text{Result}- \mid \text{Sample}-)$$

condition on the (known) truth → **total error**

Routine use → "inverse" direction

$$P(\mu \mid X)$$

$$\text{PPV} = P(\text{Sample}+ \mid \text{Result}+)$$
$$\text{NPV} = P(\text{Sample}- \mid \text{Result}-)$$

condition on the result → **measurement uncertainty**

- β-expectation tolerance interval = prediction interval = credible interval of the posterior predictive = **uncertainty of measurement** = total error
- It integrates *precision* + uncertainty of the bias estimate + uncertainty of the precision estimate, so uncertainty $\neq$ precision

1. Uncertainty and Personalised Uncertainty

Epistemic (reducible)

$$\mathrm{Var}_{\text{epist}} = \mathrm{Var}_{\theta|D}\Big[\,\mathbb{E}[y \mid x, \theta]\,\Big]$$

Limited data / model knowledge.
Shrinks with more data.

Aleatoric (irreducible)

$$\mathrm{Var}_{\text{aleat}} = \mathbb{E}_{\theta|D}\Big[\,\mathrm{Var}[y \mid x, \theta]\,\Big]$$

Inherent variability at known input.

$$\mathrm{Var}[y_{\text{new}} \mid x_{\text{new}}, D] = \mathrm{Var}_{\text{epist}} + \mathrm{Var}_{\text{aleat}}$$

PUQ = the predictive distribution for a specific patient

$$\mathrm{PUQ}(x_{\text{new}}) = p(y_{\text{new}} \mid x_{\text{new}}, D)$$

Within training domain: tight PUQ → low epistemic uncertainty
At boundary / outside: wide PUQ → model flags its own ignorance

2. PPV / NPV and Prevalence: The Mechanism

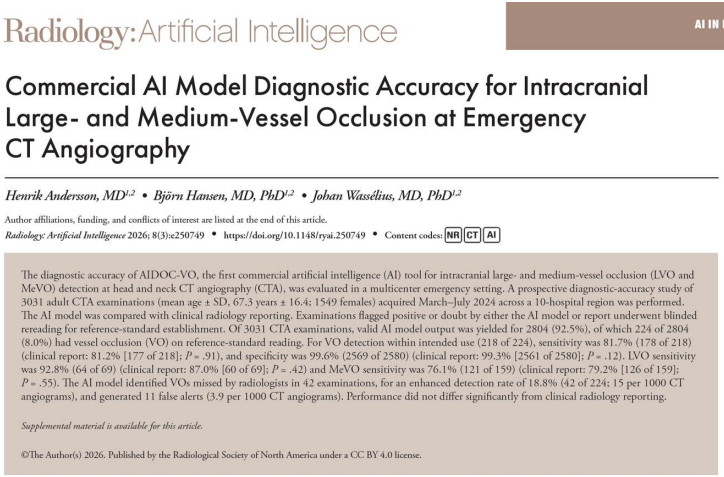
$$\mathrm{PPV}(\pi) = \frac{\mathrm{Se} \cdot \pi}{\mathrm{Se} \cdot \pi + (1 - \mathrm{Sp}) \cdot (1 - \pi)} \qquad \mathrm{NPV}(\pi) = \frac{\mathrm{Sp} \cdot (1 - \pi)}{\mathrm{Sp} \cdot (1 - \pi) + (1 - \mathrm{Se}) \cdot \pi}$$

- PPV and NPV are explicit functions of prevalence π
- Even with fixed Se, Sp: PPV collapses as $\pi \to 0$
- AUROC is prevalence-independent, PPV/NPV are not

PPV/NPV explain the *mechanism* of prevalence dependence, but are not the right *evaluation criterion*.

Assume exchangeability and homogeneity between groups or sites.

AIDOC-VO: A Commercial Stroke-Detection Device (Andersson et al., 2026)



Radiology: Artificial Intelligence

AI IN BRIEF

Commercial AI Model Diagnostic Accuracy for Intracranial Large- and Medium-Vessel Occlusion at Emergency CT Angiography

Henrik Andersson, MD[1,2] • *Björn Hansen, MD, PhD*[1,2] • *Johan Wassélius, MD, PhD*[1,2]

Author affiliations, funding, and conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2026; 8(3):e250749 • https://doi.org/10.1148/ryai.250749 • Content codes: NR CT AI

The diagnostic accuracy of AIDOC-VO, the first commercial artificial intelligence (AI) tool for intracranial large- and medium-vessel occlusion (LVO and MeVO) detection at head and neck CT angiography (CTA), was evaluated in a multicenter emergency setting. A prospective diagnostic-accuracy study of 3031 adult CTA examinations (mean age ± SD, 67.3 years ± 16.4; 1549 females) acquired March–July 2024 across a 10-hospital region was performed. The AI model was compared with clinical radiology reporting. Examinations flagged positive or doubt by either the AI model or report underwent blinded rereading for reference-standard establishment. Of 3031 CTA examinations, valid AI model output was yielded for 2804 (92.5%), of which 224 of 2804 (8.0%) had vessel occlusion (VO) on reference-standard reading. For VO detection within intended use (218 of 224), sensitivity was 81.7% (178 of 218) (clinical report: 81.2% [177 of 218]; $P = .91$), and specificity was 99.6% (2569 of 2580) (clinical report: 99.3% [2561 of 2580]; $P = .12$). LVO sensitivity was 92.8% (64 of 69) (clinical report: 87.0% [60 of 69]; $P = .42$) and MeVO sensitivity was 76.1% (121 of 159) (clinical report: 79.2% [126 of 159]; $P = .55$). The AI model identified VOs missed by radiologists in 42 examinations, for an enhanced detection rate of 18.8% (42 of 224; 15 per 1000 CT angiograms), and generated 11 false alerts (3.9 per 1000 CT angiograms). Performance did not differ significantly from clinical radiology reporting.

Supplemental material is available for this article.

Prospective study, 10 hospitals, 3031 emergency CTAs, field prevalence $\pi = 8.0\%$. CC BY 4.0.

Case in Point: One Validation Number Does Not Transport

A CE-marked, FDA-cleared stroke-detection device (AIDOC-VO), evaluated prospectively on 3031 CT-angiograms across 10 hospitals ($\pi = 8.0\%$).

	510(k) (enriched)	**Field** (π=8%)
Sensitivity	91.3%	81.7%
Specificity	85.6%	99.6%
Prevalence π	0.27	0.08
At each column's own π:		
PPV	70%	94.2%
NPV	96%	98.5%
Both Se/Sp at field π=8%:		
PPV	$\approx$ 36%	94.2%
NPV	99%	98.5%

Not device constants, and they moved in **opposite** directions:

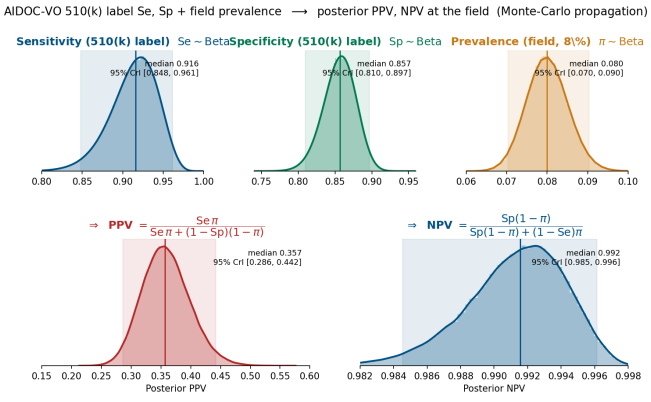
- **Se**$\downarrow$ (91.3$\rightarrow$81.7), a harder field case-mix it misses more often: MeVO, ICA-I, posterior circulation
- **Sp**$\uparrow$ (85.6$\rightarrow$99.6), but on a routine (easier) negative stream, not the enriched clearance mix; not a like-for-like gain
- **Prevalence** then drives PPV (mechanism slide)

The point

The label misled in *both* directions: optimistic on sensitivity, pessimistic on PPV ($\approx$36% vs 94%). A single validation is a *snapshot*; honest deployment needs a **distributional, local** estimate.

Andersson et al., *Radiol Artif Intell* 2026; 8(3):e250749 (CC BY 4.0). FDA 510(k) K220709.

PPV / NPV as Posterior Distributions: the Benefit of Being Bayesian



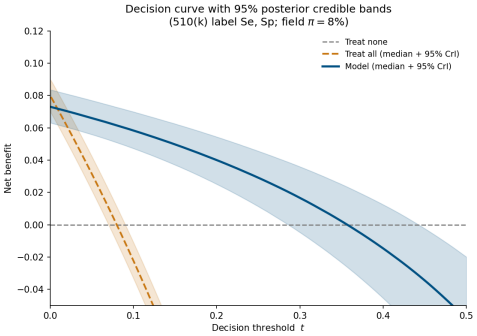
Se, Sp and prevalence each carry a **Beta posterior**:

$$\text{Se} \sim \text{Beta},\ \text{Sp} \sim \text{Beta},\ \pi \sim \text{Beta}$$

Pushed through the Bayes inversion, PPV and NPV become **full posterior distributions**, not point estimates.

- Width here is driven by the 510(k) label Se/Sp uncertainty (small enriched sample), carried to the field π
- Enables direct statements: $P(\text{PPV} > \text{PPV}_{\text{min}} \mid \text{data})$
- No risk of being *confidently wrong*

Decision Curve Analysis: NB as a Function of t



Three strategies compared

- Model: $\text{Se} \cdot \pi - (1 - \text{Sp})(1 - \pi) \cdot \frac{t}{1-t}$
- Treat all: $\pi - (1 - \pi) \cdot \frac{t}{1-t}$
- Treat none: $\text{NB} = 0$

How to read the curve

- Model is useful at t if its NB exceeds both baselines
- As t grows, harm of FP grows, treat-all crashes fast
- Flatter model curve = robust across reasonable t

Benefit of being Bayesian

With Se, Sp, π as posteriors, $\text{NB}(t)$ inherits a **full posterior** at every t. We can therefore *probabilise* the decision:

$$P(\text{NB}_{\text{model}}(t) > \text{NB}_{\text{treat all}}(t) \mid \text{data}) > \delta$$

→ a clinical-utility *decision*, not a point estimate.

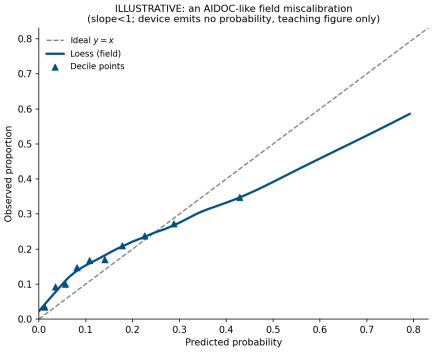
3. Calibration in the Field (illustrative)

Elements on one canvas

- - - - Ideal diagonal $y = x$
- — Loess flexible calibration curve
- ■ 95% pointwise CI band
- ▲ Decile points + Wilson CIs
- Rug: events (top) / non-events (bottom)

Three summary statistics

- **O:E ratio** $= \sum Y / \sum \hat{p}$ (mean cal., $\to 1$)
- **Cal. intercept** $a\,(\to 0)$ & **slope** $b\,(\to 1)$ in $\text{logit}(Y) = a + b\,\text{logit}(\hat{p})$



The x-axis $\hat{p}$ is the **posterior predictive** probability $P(Y_{\text{new}}{=}1 \mid x, D) = \int P(Y{=}1 \mid x, \theta)\, p(\theta \mid D)\, d\theta$.

4. Prevalence Mismatch Across Sites

$$\mathrm{PPV}(\pi) = \frac{\mathrm{Se} \cdot \pi}{\mathrm{Se} \cdot \pi + (1 - \mathrm{Sp})(1 - \pi)} \qquad \Longrightarrow \qquad \pi_1 \neq \pi_2 \;\Rightarrow\; \mathrm{PPV}_1 \neq \mathrm{PPV}_2$$

Development site

$\pi_{\mathrm{dev}} = 0.269 \;\Rightarrow\;$ **PPV** $= 0.70$, **NPV** $= 0.96$

Enriched 510(k) test set (92/342)

Deployment site s^*

$\pi_{s^*} = 0.08 \;\Rightarrow\;$ **PPV** $= 0.355$, **NPV** $= 0.99$

Field deployment ($\pi = 8\%$)

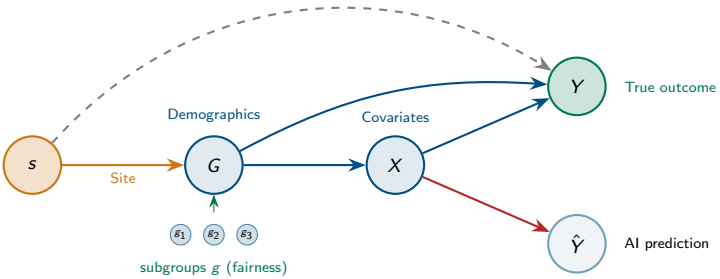
- Same model, Se $= 0.913$, Sp $= 0.856$ (510(k) label): PPV drops $0.70 \to 0.36$ as π falls $0.27 \to 0.08$; NPV barely moves ($0.96 \to 0.99$)
- Natural response: model $\pi_s \sim \mathrm{Beta}(a, b)$ hierarchically, borrow across sites

But: the exchangeability assumption is rarely tested. Standard conflict measures detect *that* sites differ, but not *why*.

Source: FDA 510(k) K220709 pivotal set, 92/342 VO ($\pi = 27\%$); its reported PPV 41.3% was standardised to an assumed $\pi \approx 10\%$, neither the 27% test set nor the 8% field.

5. The Causal DAG for Multi-Site Deployment

Decomposing site variation into interpretable mechanisms



ICM factorisation: $P(G \mid s) \cdot P(X \mid G, s) \cdot P(Y \mid X, G)$

ICM = Independent Causal Mechanisms: each factor can change on its own (Schölkopf et al.).

M1 prevalence π_s · **M2** case-mix $P(G \mid s), P(X \mid G, s)$ · **M3** concept $P(Y \mid X, G)$.

Fairness = stability across subgroups g (same M2/M3 split as across sites; 510(k)→field = this with s for g). Notation: s site, g subgroup, k subtype.

Transportability: Which Mechanism Moves Which Metric

From the DAG, site variation enters through three knobs that can move independently.

Mechanism	$P(X \mid Y)$	**Se, Sp**	**PPV, NPV**
M1 prevalence π_s	unchanged	invariant	shift
M2 case-mix $P(X \mid G, s)$	shifts (within class)	shift	shift
M3 concept $P(Y \mid X, G)$	shifts (relationship)	shift	shift

Read-off, and the AIDOC-VO case

Only **M1** spares Se/Sp, so *any* observed Se/Sp shift implicates **M2 or M3**, never prevalence alone. For AIDOC-VO, Se↓ with Sp↑ is most parsimoniously **M2** (harder occlusion mix; easier routine negatives) on top of an **M1** prevalence gap, with *no* firm evidence of M3.

Fairness is the same problem, across subgroups g

Replace site s by subgroup g: subgroup case-mix gaps are **M2**; residual gaps at a fixed case-mix are **M3**, the genuine algorithmic-bias term.

6. Hierarchical Prevalence + Dynamic Borrowing

Site-level prevalence prior with hierarchical structure

$$\pi_s \mid a, b \sim \text{Beta}(a, b) \qquad \pi_s \mid \text{data} \sim \text{Beta}(a + y_s,\ b + n_s - y_s)$$

Dynamic borrowing of the prior across sites:

$$p(\pi_{s^*}) = w \cdot p_{\text{pooled}}(\pi) + (1 - w) \cdot p_{\text{vague}}(\pi)$$

p_{pooled}
Hierarchical posterior from existing sites. Encodes "typical" prevalence.

p_{vague}
Weakly informative. Protects if new site diverges.

w (weight)
Data-driven; $w \to 0$ as local n_s grows.

Analogy: Bayesian adaptive clinical trials, borrowing strength across arms / sites.

7. Net Benefit per Subgroup: The Evaluation Criterion

Prevalence-uncertainty-adjusted PPV (the mechanism, posterior)

$$\mathrm{PPV}_s = \int \mathrm{PPV}(\pi) \cdot \mathrm{Beta}(\pi \mid a + y_s,\ b + n_s - y_s)\, d\pi$$

→ a **distribution** over PPV, not a point estimate. Computed per subgroup: $\mathrm{PPV}_{s,g}$.

Net benefit per subgroup: the evaluation criterion

$$\mathrm{NB}(t, g) = \frac{\mathrm{TP}_g}{N_g} - \frac{\mathrm{FP}_g}{N_g} \cdot \frac{t}{1-t} = \mathrm{Se}_g\, \pi_g - (1 - \mathrm{Sp}_g)(1 - \pi_g) \cdot \frac{t}{1-t}$$

- **Left**: confusion-matrix form (directly observable)
- **Right**: prevalence-explicit form (shows *why* NB differs across subgroups)
- $\mathrm{Se}_g, \mathrm{Sp}_g$: a *fixed* subgroup-specific operating point (binary flag); the threshold t enters *only* via the false-positive weight $t/(1-t)$
- NB weights false positives by the threshold odds $t/(1-t)$; PPV ignores utilities
- Recalibration can *worsen* F1/accuracy even as calibration improves, a regulatory pitfall

8. From Model Selection to the Borrowing Decision

M1: Prevalence shift

Exchangeability supported

π_s differs but $P(Y \mid X, G)$ stable.

$\Rightarrow$ **Full borrowing**

$p(\pi_{s^*}) = w\, p_{\text{pooled}} + (1-w)\, p_{\text{vague}}$

Hierarchical Beta model handles this naturally.

M2: Covariate shift

Partial exchangeability

Case-mix differs, concept preserved.

$\Rightarrow$ **Borrow with propensity-score reweighting**

Correct the population mismatch before setting the prior.

M3: Concept shift

Exchangeability violated

Clinical relationship changed.

$\Rightarrow$ **Refuse to borrow**

$w \to 0$. Trigger local re-estimation or model recalibration.

Bayesian model comparison turns exchangeability from an untestable assumption into a **mechanism-specific, formally testable diagnostic**.

How to borrow under M2: fit a propensity score $e(x) = P(\text{target} \mid x)$; weight source cases by $w_i \propto e(x_i)/(1 - e(x_i))$ to match the target case-mix, then form the prior from the *reweighted* sample.

Fairness as a Borrowing Decision Across Subgroups g

The same M1/M2/M3 logic, now *within* a site, across demographic subgroups g:

M1: Prevalence differs by g

Rates exchangeable

π_g differs; $\mathrm{Se}_g, \mathrm{Sp}_g$ stable.

$\Rightarrow$ **Borrow Se/Sp; report PPV/NPV per π_g**

Different PPV across g is prevalence, not bias.

M2: Case-mix differs by g

Partial exchangeability

$P(X \mid Y, g)$ composition differs.

$\Rightarrow$ **Borrow with standardisation**

Compare Se/Sp at a common case-mix before judging fairness.

M3: Concept differs by g

Exchangeability violated

$P(Y \mid X, g)$ genuinely differs.

$\Rightarrow$ **Do not borrow**

$w \to 0$: estimate g separately, flag the disparity, this is the bias.

Fairness is not a separate metric: it is the **borrowing decision across subgroups**. Borrow (with adjustment) under M1/M2; refuse under M3, where intervention is required.

How to borrow under M2 (standardisation): estimate stratum rates Se_k pooled across g, recombine with each subgroup's weights $\mathrm{Se}_g = \sum_k w_{g,k}\,\mathrm{Se}_k$, and compare these *standardised* rates; a residual gap is M3.

The Unification

$$p(y_{\text{new}} \mid x_{\text{new}}, \text{site } s, \text{group } g, D, \text{data}_{s,g})$$

Uncertainty	→ spread of the predictive distribution
PUQ	→ patient-conditional predictive distribution
Analytical validation (TAE/MU/ATP/QTPP)	→ forward $P(X \mid \mu)$ vs inverse $P(\mu \mid X)$
Calibration (4 levels)	→ Van Calster hierarchy as Bayesian calibration checks
PPV / NPV	→ prevalence-dependent Bayesian inversion (posterior)
Net Benefit	→ **decision-analytic criterion; posterior ⇒ probabilised**
Conformal Prediction	→ frequentist safety net when strong calibration fails
AI Fairness	→ subgroup-conditional functionals
Site & subgroup calibration	→ hierarchical + dynamic borrowing (by mechanism)

One distributional object. Consistency all the way through.

Thank you

Questions & discussion welcome

Bruno Boulanger | SANAITIO – Belgium

bruno.boulanger@sanaitio.com | sanaitio.com

Trust by Design for AI-based medical devices

Backup: Two Levels of Evidence for Non-Transportability

	AIDOC-VO (Andersson 2026)	**Scoping review** (Dorochowicz 2026)
	1 device, 10 sites, 3031 CTAs	4 devices, 29 studies, 2019–25
Design	prospective, controlled *within-device* cross-site contrast	between-study map; formal meta-analysis *unfeasible*
Se	91.3 → 81.7 (label → field)	78–97% across studies
Sp	85.6 → 99.6 (enriched → routine)	74–97%
PPV / π	π=8%: ≈36% → 94% (clean)	π heterogeneous/enriched ⇒ *not poolable*
Shift axes	M1/M2/M3, population only (f fixed)	+ vendor (f differs) + version/time (f drifts; PCCP)

Same thesis, opposite ends of the evidence telescope

The controlled contrast lets you *attribute* the shift to a mechanism (M2); the literature band shows what happens when you *don't* condition, marginal Se/Sp that cannot be pooled. Both point to one need: a **local, distributional** estimate.

Andersson et al., *Radiol Artif Intell* 2026; 8(3):e250749. Dorochowicz et al., *Medicina* 2026; 62(3):582 (both CC BY 4.0).