sanaitio

Causally Informed *Dynamic Borrowing* for Multi-Site AI Diagnostic Deployment

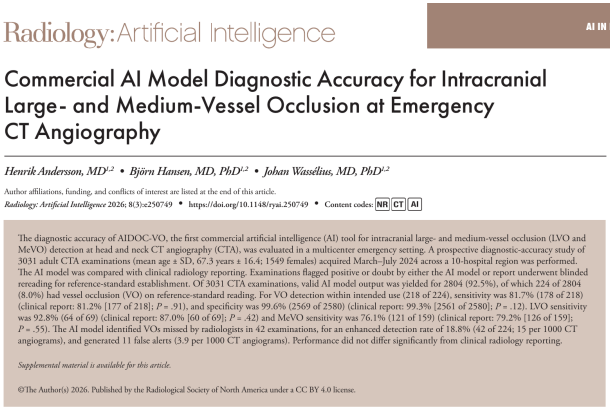
Testing exchangeability **mechanism by mechanism** with causal DAGs and Bayesian model comparison

Bruno Boulanger Nils Boulanger
SANAITIO, Belgium

BAYES 2026 | Leiden | 21 to 23 October 2026

CiValAI Project: Causal Integrated Validation for AI in Medicine
Trust by Design for AI-based medical devices

AIDOC-VO: a commercial stroke-detection device



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 Wasselius, 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

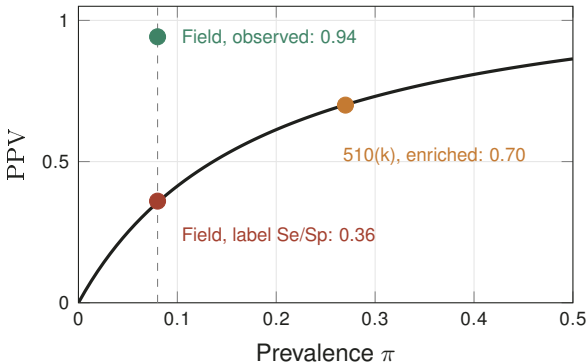
The diagnostic accuracy of AIDOC-VO, the first commercial artificial intelligence (AI) tool for intracranial large- and medium-vessel occlusion (LVO and MeVO) detection at head and neck CT angiography (CTA), was evaluated in a multicenter emergency setting. A prospective diagnostic-accuracy study of 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.

©The Author(s) 2026. Published by the Radiological Society of North America under a CC BY 4.0 license.

Andersson et al. (2026), CC BY 4.0. Prospective study, 10 hospitals, 3031 emergency CTAs, field prevalence $\pi = 8.0\%$.

A label PPV of 0.36 turned out to be 0.94 in the field



Curve: label $\mathrm{Se} = 0.913$, $\mathrm{Sp} = 0.856$. AIDOC-VO stroke-detection device (Andersson et al., 2026): 3031 CTAs, 10 hospitals.

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

- Prevalence fell (0.27 to 0.08) and Se, Sp moved too: the label misled in both directions.
- Borrowing across sites assumes exchangeability, **rarely tested**.

No prediction model is ever truly validated



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*]

Abstract

Background Clinical prediction models should be validated before implementation in clinical practice. But is favorable performance at internal validation or one external validation sufficient to claim that a prediction model works well in the intended clinical context?

Main body We argue to the contrary because (1) patient populations vary, (2) measurement procedures vary, and (3) populations and measurements change over time. Hence, we have to expect heterogeneity in model performance between locations and settings, and across time. It follows that prediction models are never truly validated. This does not imply that validation is not important. Rather, the current focus on developing new models should shift to a focus on more extensive, well-conducted, and well-reported validation studies of promising models.

Conclusion Principled validation strategies are needed to understand and quantify heterogeneity, monitor performance over time, and update prediction models when appropriate. Such strategies will help to ensure that prediction models stay up-to-date and safe to support clinical decision-making.

Keywords Risk prediction models, Predictive analytics, Internal validation, External validation, Heterogeneity, Model performance, Calibration, Discrimination

Van Calster et al. (2023), *BMC Medicine* 21:70, CC BY 4.0. Populations, measurements and time all vary, so performance differs between settings.

Borrowing methods say that sites differ, not why

What exists

Borrowing: robust MAP (Schmidli et al., 2014), EXNEX and commensurate priors (Neuenschwander et al., 2016; Hobbs et al., 2011).

Causal and transport: transportability (Pearl and Bareinboim, 2014), label shift (Lipton et al., 2018).

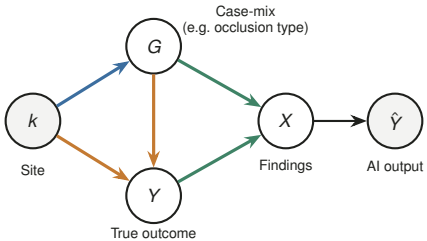
What is missing for AI diagnostics

Conflict is scalar. The answer to "why" changes the action: **pool, reweight, or re-estimate locally**.

Idea: decompose between-site variation into causal mechanisms, learn *which* ones shifted, and let that set the prior.

Three causal mechanisms can shift, each with its own indicator

$$P(G, Y, X \mid k) \;=\; P(G \mid k) \;\cdot\; P(Y \mid G, k) \;\cdot\; P(X \mid Y, G)$$



- **Case-mix**: who is referred.
- **Prevalence**: how often disease occurs, given case-mix.
- **Presentation**: how disease looks in the data; changes Se_g, Sp_g.

Each mechanism is a conditional probability distribution; a shift means its parameters differ between sites.

One indicator $\delta_m \in \{0, 1\}$ per mechanism (0: shared across sites, 1: site-specific), so $2^3 = 8$ models.
Sparse prior $\delta_m \sim \mathrm{Bernoulli}(0.3)$: a new site tends to change few mechanisms.

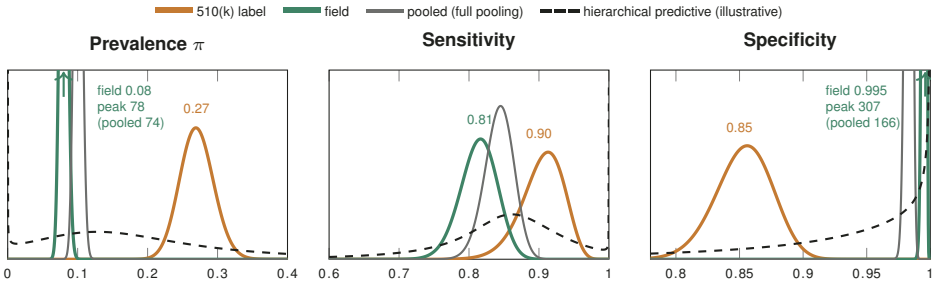
The posterior shift probabilities set how much to borrow

$$p(\theta_{k^*}) = \prod_{m \in \{G,Y,X\}} \Big[\mathrm{PIP}_m \; p(\theta_m \mid \delta_m = 1) + (1 - \mathrm{PIP}_m) \; p(\theta_m \mid \delta_m = 0) \Big]$$

- k^* is the new site and $\theta_{k^*} = (\theta_G, \theta_Y, \theta_X)$ its parameters, in one block per mechanism: $\theta_G = \rho$, $\theta_Y = (\pi_g)$, $\theta_X = (\mathrm{Se}_g, \mathrm{Sp}_g)$. $p(\theta_m \mid \delta_m)$ is the prior of block m: shared if $\delta_m = 0$, shifted if $\delta_m = 1$.
- **PIP** (posterior inclusion probability) $\mathrm{PIP}_m = P(\delta_m = 1 \mid \mathcal{D})$ is the probability that mechanism m is site-specific at k^*; $\mathrm{logit}\,\mathrm{PIP}_m = \mathrm{logit}(0.3) + \ln \mathrm{BF}_m$. Example: $\mathrm{PIP}_Y = 0.2$ gives weight 0.2 to the hierarchical predictive (site-specific) and 0.8 to the pooled prior (shared).
- **Hierarchical predictive:** the prior for a new site when sites vary around a common value, each parameter of θ_{k^*} follows $\mathrm{Beta}(\mu\kappa, (1-\mu)\kappa)$ with (μ, κ) learned from the other sites. It is centred on the pooled value, and its width reflects how much sites differ.

Mechanism	Shared ($\delta = 0$)	Shifted ($\delta = 1$)
Prevalence π_g	pool across sites	hierarchical Beta predictive
Case-mix ρ	pool, standardise to k^* mix	propensity reweighting $w(g)$, overlap check (no labels needed)
Presentation $\mathrm{Se}_g, \mathrm{Sp}_g$	pool across sites	wide hierarchical predictive

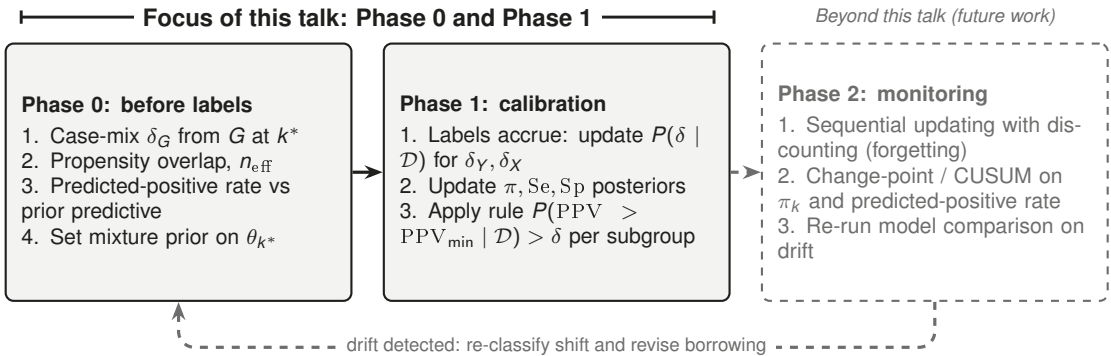
The label and field Beta posteriors barely overlap



True Beta densities (area 1 each); vertical axis clipped, arrows show true peaks. Label: 92/342, 84/92, 214/250 (from 510(k) rates); field: 224/2804, 178/218, 2569/2580.

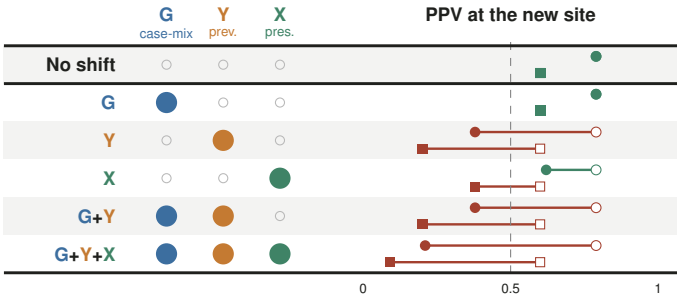
- π **and Sp: no overlap** between label and field (0.27 against 0.08, 0.85 against 0.995). The field curves are far narrower (2804 against 342 exams). Pooling lands on the field curve, and for π (0.10) matches neither source.
- **Se: partial overlap** (0.90 against 0.81); the label rests on 92 positives, so evidence of a shift is weaker.
- **Dashed:** illustrative hierarchical predictive for a new site from these two sites only. Its spread comes from the hyperprior, not from the data.

Adaptive site calibration protocol



This talk covers Phases 0 and 1. The protocol maps to the regulatory lifecycle for AI as a medical device: FDA Predetermined Change Control Plans; EU AI Act risk management (Art. 9) and post-market monitoring (Art. 72).

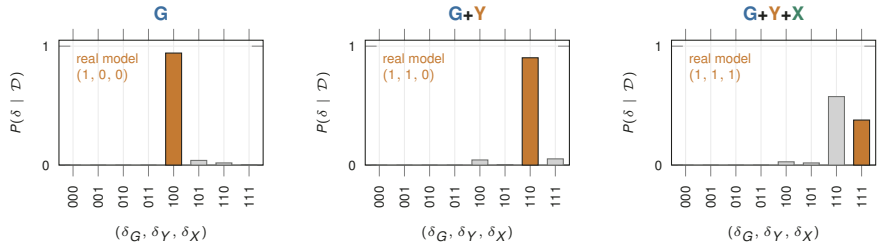
Simulation: six scenarios, three sample sizes, five methods



Dots: mechanism that really shifted. Hollow: PPV at the reference sites; filled: real PPV at the new site (green: deploy, red: do not deploy). Circle $g = 1$, square $g = 2$.

- $K = 5$ reference sites of 300 patients, 2 subgroups; new site $n_{k^*} \in \{50, 200, 1000\}$; 500 replicates.
- Logit shifts: $P(G{=}2)$ 0.30 to 0.70; π −1.8; Se −1.2, Sp −0.7.
- Deploy g if $P(\mathrm{PPV}_{k^*,g} > 0.5 \mid \mathcal{D}) > 0.8$.
- Methods: no pooling, full pooling, fixed hierarchical, robust MAP, proposed ($p = 0.3$).
- Metrics: inclusion probabilities, RMSE, coverage, wrong decisions.

One replicate: what the method sees and decides



PIP (G / Y / X)	1.00 / 0.02 / 0.04	1.00 / 0.95 / 0.05	1.00 / 0.95 / 0.40
Full pooling	1.00 / 1.00	**1.00** / **0.97**	**1.00** / 0.74
Fixed hierarchical	1.00 / 0.98	**0.96** / 0.11	**0.83** / 0.01
No pooling	1.00 / 0.94	0.52 / 0.02	0.11 / 0.00
Robust MAP	1.00 / 0.97	**0.86** / 0.09	0.60 / 0.00
Proposed	1.00 / 1.00	**0.98** / 0.12	**0.91** / 0.04
Real answer	deploy / deploy	no / no	no / no

Replicate 119, $n_{k*} = 200$. Method rows: $P(\mathrm{PPV} > 0.5 \mid \mathcal{D})$ for $g = 1$ / $g = 2$; deploy if > 0.8; red: wrong decision. One replicate only: the 500-replicate rates are on the accuracy slide. Case-mix weights are identical in the three scenarios: $w(1) = 0.49$, $w(2) = 2.07$ (exact 0.43, 2.33).

Simulation results: recovery of the shift set

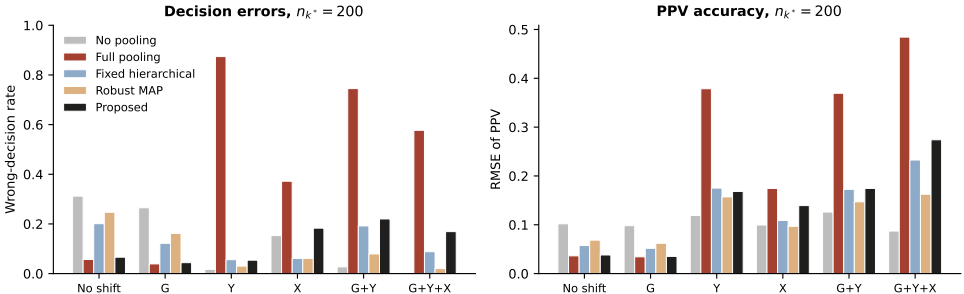
	δ_G (case-mix)			δ_Y (prevalence)			δ_X (presentation)		
n_{k*}	50	200	1000	50	200	1000	50	200	1000
No shift	0.09	0.10	0.07	0.07	0.06	0.04	0.06	0.06	0.04
G	**0.95**	**1.00**	**1.00**	0.07	0.05	0.04	0.06	0.06	0.03
Y	0.09	0.08	0.07	**0.19**	**1.00**	**1.00**	0.06	0.05	0.03
X	0.10	0.08	0.07	0.07	0.06	0.04	**0.10**	**0.47**	**0.99**
G+Y	**0.94**	**1.00**	**1.00**	**0.14**	**0.96**	**1.00**	0.06	0.05	0.04
G+Y+X	**0.94**	**1.00**	**1.00**	**0.15**	**0.95**	**1.00**	**0.10**	**0.26**	**0.89**

Mean posterior inclusion probability $\mathrm{PIP}_m = P(\delta_m = 1 \mid \mathcal{D})$ over 500 replicates. Rows: scenario that generated the data.

Boxed cells = real shift: read high is good. Other cells = false inclusion: read low is good.

- False inclusion is low: mean PIP ≤ 0.10 for every mechanism that did not shift.
- Detection depends on the mechanism and on n_{k*}: case-mix at $n = 50$ (PIP ≥ 0.94), prevalence from $n = 200$ (0.19 at $n = 50$, 1.00 at $n = 200$), presentation only near $n = 1000$ (0.47 at $n = 200$, 0.99 at $n = 1000$).

Simulation results: accuracy and decisions ($n_{k^*} = 200$)



- **Exchangeable or case-mix only:** the proposed method stays close to full pooling (wrong decisions 0.07 and 0.04, against 0.25 and 0.16 for robust MAP). Under case-mix shift it also has the lowest error on π_{k*} (RMSE 0.012 vs 0.019 for robust MAP; full pooling 0.037).
- **Prevalence or presentation shift:** full pooling fails (0.37 to 0.87). The proposed method is comparable to fixed hierarchical, but robust MAP and no pooling protect better when several mechanisms move or presentation shifts (e.g. G+Y: 0.22 vs 0.08 vs 0.03).

Limitations and open questions

- **DAG misspecification:** model comparison ranks the shifts *we specified*. AIDOC-VO specificity (0.86 to 0.996) cannot be split into case-mix and presentation without subtype-level negatives.
- **Power and priors:** with small n_{k^*} the inclusion probabilities are diffuse and the prior dominates (prevalence 0.19 at $n = 50$). A sensitivity analysis on p and on shift magnitudes is still needed.
- **Data quality:** label delay and verification bias; when overlap fails, stop borrowing and collect local data.
- **Simulation scope:** identical reference sites and a plain hierarchical shifted component; a vague (robust MAP type) component is the next step.

Summary

- Exchangeability becomes checkable, mechanism by mechanism, under explicit causal assumptions.
- Borrowing is dynamic: derived from posterior shift probabilities, per subgroup, not tuned.
- Protocol follows the regulatory lifecycle; simulation: pooling-level efficiency when borrowing is valid, protection under shift that grows with n_{k^*}.

References

- Andersson, H., Hansen, B., Wassélius, J. (2026). Commercial AI model diagnostic accuracy for intracranial large- and medium-vessel occlusion at emergency CT angiography. *Radiology: Artificial Intelligence*, 8(3), e250749.
- Chouldechova, A. (2017). Fair prediction with disparate impact. *Big Data*, 5(2), 153-163.
- Dahabreh, I. J. et al. (2020). Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*, 39(14), 1999-2014.
- Hobbs, B. P. et al. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3), 1047-1056.
- Lipton, Z., Wang, Y.-X., Smola, A. (2018). Detecting and correcting for label shift with black box predictors. *ICML*.
- Neuenschwander, B. et al. (2016). Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharmaceutical Statistics*, 15(2), 123-134.
- Pearl, J., Bareinboim, E. (2014). External validity: from do-calculus to transportability across populations. *Statistical Science*, 29(4), 579-595.
- Schmidli, H. et al. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4), 1023-1032.
- Schölkopf, B. et al. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5), 612-634.
- Van Calster, B., Steyerberg, E. W., Wynants, L., van Smeden, M. (2023). There is no such thing as a validated prediction model. *BMC Medicine*, 21, 70.

Thank you

Exchangeability becomes checkable, mechanism by mechanism



Bruno Boulanger and Nils Boulanger
SANAITIO, Belgium
bruno@sanaitio.com

BAYES 2026 | Leiden | 21 to 23 October 2026