ARKA
ARKA - Drift, calibration, and certainty monitoring
==========
v1.0.0 22 August 2026 as of 2026-08-22
Method and ARKA's own published figures for certainty burn-down, calibration status, drift under three monitoring strategies, the four appropriateness dimensions, and the corrections feed. Customer-specific rates stay behind an authenticated packet.
This public report carries no customer names and no site-specific customer rates.
Web: /monitoring PDF: /api/monitoring/report

1. CERTAINTY BURN-DOWN

The knowledge base is a strong prototype and an inadequate production state - of 1536 matrix ratings, 40 are high, 119 moderate, 22 low, and 1355 very low GRADE certainty, with 21 of 1536 abstaining - and Volume III said advisors forgive a small knowledge base and do not forgive a founder who claims completeness.
The volume-weighted very_low share is 88.1%, below the raw share of 88.2%, because the knowledge base was built from the most common scenarios outward - higher-certainty ratings concentrate where order volume is highest.
high: raw 40 volume-weighted 1694
moderate: raw 119 volume-weighted 5130
low: raw 22 volume-weighted 1764
very_low: raw 1355 volume-weighted 63636
Scheduled (90d): 50 reviewed: 6 (raised 0, lowered 6, unchanged 0)

2. CALIBRATION

To evaluate calibration in this population, provide at least 100 paired rows with predictedProbability (0-1), observedBinaryOutcome (0 or 1), and careSetting (ED | inpatient | ambulatory | virtual | unknown). Synthetic benchmarks are for reproducibility only and never fill this panel.
Required fields: predictedProbability, observedBinaryOutcome, careSetting minimum paired outcomes: 100

2a. OUTCOME-SIGNAL LEDGER

Calibration and drift figures on this page are weighted by arm-exposure covariates (justification, default-order, and peer-packet exposure) so that a change in who is scored after ARKA changes behaviour is not read as calibration drift. Unweighted, adherence-weighted, and sampling-weighted strategies are shown together; the unweighted figure is never published alone once an intervention is live (prompt 34.4).
Downstream utilization within 90 days is a utilization fact read from claims. It is not harm, not an adverse event, and not a quality adjudication.
n=115 windows 90d / 90d provenance=measured
study_performed: n=115 observed rate 1.000
downstream_utilization_90d: n=115 observed rate 0.478

3. DRIFT - THREE STRATEGIES

Illustrative method panel site arka-method 2026-06 provenance=illustrative
Intervention live: yes arm exposure 0.400
[unweighted] Na?ve average over observed rows. Biased once an intervention changes who is scored or labelled; published with an explicit caution when interventionLive is true.
PSI 0.1408 label delta 0.1450 ECE delta 0.0875
CAUTION: Intervention is live at this site - the unweighted figure is contaminated by the feedback loop and must be read beside the adherence-weighted and sampling-weighted figures, not alone.
[adherence_weighted] Reweights by the inverse of arm-adherence probability so evaluated rows resemble the pre-intervention population under selective uptake.
PSI 0.1408 label delta 0.1450 ECE delta 0.0875
[sampling_weighted] Reweights by the inverse of outcome-sampling probability so label drift is not confused with selective outcome capture.
PSI 0.1408 label delta 0.1450 ECE delta 0.0875

4. APPROPRIATENESS - FOUR DIMENSIONS

Not a single scalar. Public surface shows ARKA evidence-adequacy from the certainty census; indication, shared decision, and patient understanding remain not_measured here.
Indication supported: not_measured illustrative
Evidence adequate: not met 181/1536 (11.8%) measured
Shared decision occurred: not_measured illustrative

- imaging-peer-comparison-existence: We said no imaging study of peer comparison existed. Two had been published. We had not looked hard enough.
Now: Imaging-specific peer-comparison evidence exists and is uncontrolled. The open questions are how much of the published reduction was the intervention, whether imaging behaves like the antibiotic trials, and whether any effect persists - not whether a study had been published.
Commit: e400d3ad504037d63ab4f0aa8b16dcbf7d20f949
- benchmark-published-without-doi: We announced the ARKA-IP inpatient imaging benchmark as published on 18 August 2026 with a citation that pointed at a private repository and no DOI.
Now: A dataset without a resolvable identifier is not published. The DOI was minted on 21 August 2026 (10.5281/zenodo.22051812). CHECK-BENCH-1 now fails any asset marked published without one.
Commit: f809f97dd4531f61a86a8e3fe752b999b7f5b3b5
- abstention-by-provenance-not-consequence: We treated evidence-quality abstention as a provenance filter: AIIE abstained only when GRADE certainty was very_low and the sole anchor was single-specialty AUC or payer-authored. Everything else, including strong recommendations on very-low certainty, could still emit a confident card.
Now: evidenceQualityAbstain is the disjunction of weakEvidenceMustAbstain (unchanged) and strongRecOnVeryLowMustAbstain, which now fires only for a strong recommendation to image (rating ? 7) on very_low certainty. A strong recommendation not to image on very_low certainty still emits, and must carry formatClinicianCertaintyPhrase and STRONG_REC_LOW_CERTAINTY_CAUTION. Certainty values are not raised to shrink abstainCount - that is step 33.2 only.
Commit: 7427e2fb88c8f4c3e5c17ee21b87e22c3679ba16
- per-clinician-ledger-ownership: We claimed that nobody owns the per-clinician risk-adjusted ledger - that the measurement substrate for low-value imaging in risk-bearing primary care was unowned.
Now: We no longer claim ledger exclusivity. We claim the intervention stack - locked pre-period, monthly peer packet outside the EHR, record-write verification - is what none of the named incumbents ship as a published product. The withdrawal is recorded beside the nudge write-up with sources that resolve.
Commit: d3405a09a2e7543766df9dbfb4b473b2911293ad