DOCUMENT 1 OF 3 · 14 PAGES

Clinical AI Validation Report & Model Card

Clinical validation evidence, model card and human-sign-off contract for healthcare AI — subgroup performance, calibration, dataset-shift sensitivity, FDA SaMD / MHRA alignment and named clinical sign-off.

Audience · Chief Medical Officer, model owner, clinical safety officer, quality / regulatory, IRB, regulator
Companion documents · HIPAA AI DPIA & Sub-processor Inventory · Clinical Agent Incident-Response Runbook
Use as · Audit-grade artefact requiring named human sign-off
Version · v1.0 · 2026-05

§ 1

Intended use & statement of contra-indications

A clinical AI model is only safe if the boundary of its intended use is written down before clinicians use it. This section establishes the boundary.

Intended use statement

Element	Specification
Clinical decision supported	[e.g. triage of incoming ED presentations into 5 acuity tiers]
Patient population	Age range, sex, conditions in / out of scope
Care setting	Inpatient / outpatient / ED / telemedicine
Clinical autonomy	Decision-support; requires clinician confirmation. Never disposition-final.
Inputs	Structured EHR fields; chief complaint free text; vitals
Output	Acuity tier + supporting features + confidence
Time horizon	Acute (within current encounter)

Contra-indications & out-of-scope

- Paediatric < 2 yo: insufficient training data
- Trauma activation cases: bypass AI, use trauma protocol
- Mass-casualty incident: bypass AI, use START / SALT triage
- Patients with prior protocol-defined exception flags

Hard rule. The model never closes a clinical decision unilaterally. Every recommendation requires an attending or trained clinician to confirm; disagreement is logged and used for monitoring.

§ 2

Training & validation cohort description

Cohort table

Cohort	Source	Period	N	Inclusion / exclusion
Training	EHR · Sites A, B	2023-01 – 2024-12		Adults ≥ 18, completed encounters
Internal validation	EHR · Sites A, B	2025-01 – 2025-06		Same; held out by time
External validation	EHR · Site C	2025-01 – 2025-06		Geographically distinct
Prospective validation	Site A · shadow mode	2025-09 – 2026-02		Live data; no clinical use

Demographic composition

Attribute	Training	Internal	External
Age (mean / IQR)			
Sex (% female)			
Race / ethnicity (top 3)			
Comorbidity (CCI median)			
Insurance class			
Language preference			

Distributional differences between cohorts must be reported and reflected in the limitations register (§ 9).

§ 3

Performance · headline & by subgroup

Headline performance (validation)

AUROC	AUPRC	SENSITIVITY	SPECIFICITY
—	—	—	—

All metrics reported with 95% confidence intervals. Operating threshold documented and traceable to the clinical decision it supports.

Subgroup performance

Subgroup	AUROC (95% CI)	Sensitivity	Specificity	Notes
Overall				
Age < 40 / 40–65 / ≥ 65				
Sex (F / M)				
Race / ethnicity				
High comorbidity (CCI ≥ 4)				
Limited English proficiency				
Medicaid / underinsured				

Equity guardrail. Material under-performance in any subgroup requires investigation before clinical rollout. Acceptable mitigations: re-training with cohort upweighting, decision threshold per subgroup with documented rationale, scope reduction excluding the affected subgroup.

§ 4

Calibration & reliability

Discrimination measures (AUROC) tell you the model orders cases correctly; calibration tells you whether the predicted probabilities mean what they say. Both are required for clinical safety.

Calibration evidence required

- Reliability diagram (predicted vs observed) across deciles, with 95% CI
- Calibration-in-the-large (mean predicted vs mean observed)
- Brier score
- Subgroup reliability for the same subgroups as § 3
- Pre / post any threshold change

Decision threshold derivation

Operating point	Sensitivity	Specificity	PPV	NPV	Cost rationale
Default					[Why this point?]
High-sensitivity					e.g. screening
High-specificity					e.g. confirmatory

The operating point sits at the intersection of false-negative cost and false-positive cost in the clinical pathway. This rationale is documented in the model card.

§ 5

Dataset-shift sensitivity

Healthcare data shifts: case-mix moves, documentation conventions change, EHR upgrades alter data flows, public-health events reshape presentations. Sensitivity to such shifts must be tested explicitly.

Shift type	How tested	Result & mitigation
Covariate shift	External validation cohort (different site)	
Prior shift	Period with elevated base-rate disease	
Documentation shift	Pre / post EHR upgrade	
Demographic shift	Subgroup composition change	
Concept shift	Definition of acuity / outcome changes	
Adversarial	Perturbed inputs in plausible ranges	

Post-deployment. Same shift tests run continuously in production. Triggered alerts fire into the monitoring contract (§ 8) and may prompt re-validation per the incident runbook (companion doc).

§ 6

Human-in-the-loop contract

The HITL contract states exactly what the AI does, what the clinician does, and how disagreement is captured. Without this, the model and the clinician share responsibility ambiguously — the worst possible outcome.

Boundaries of automation

Action	AI alone	AI + clinician	Clinician alone
Compute risk score	Yes	—	—
Surface differential considerations	Yes (with confidence)	—	—
Recommend acuity tier	—	Yes (clinician confirms)	—
Finalise disposition	Never	—	Yes
Order tests / treatment	Never	—	Yes
Inform patient of diagnosis	Never	—	Yes

Disagreement capture

1. When clinician chooses an acuity tier different from the AI recommendation, the system prompts for a brief reason (free text or coded).
2. Disagreement records are aggregated weekly and reviewed by the model owner + clinical lead.
3. Persistent disagreement patterns trigger root-cause review — model defect, training data, UI affordance, clinician training, or genuine clinical edge case.

§ 7

Outcome tracking & longitudinal accuracy

A clinical AI is only as accurate as the truth label it is measured against. Outcome tracking links AI recommendation to final diagnosis (and downstream outcome) so accuracy can be evaluated longitudinally.

Link	Source of truth	Lag	Owner
Recommendation → clinician decision	HITL log	Same encounter	Model owner
Decision → discharge diagnosis	Coded diagnosis	Days	Quality / coding
Discharge → 30-day outcome	Readmission, mortality, return to ED	30 days	Quality
Discharge → long-term outcome	Disease-specific registry / claims	Months	Quality

All linkages performed within HIPAA bounds (see companion DPIA doc); identifiers held in a controlled linkage table; analytic outputs use limited data sets.

§ 8

Post-deployment monitoring contract

Monitor	SLI	Threshold	Action on breach
Score distribution	PSI vs baseline	0.25	Investigate
Per-subgroup performance	AUROC vs baseline by subgroup	≥ 10% relative decline	Re-validate
Calibration drift	Reliability deviation	Material	Re-calibrate
Disagreement rate	Clinician override rate	Sustained increase	Root cause
Time-to-decision	Workflow KPI	Material change vs baseline	UX / model review
30-day adverse outcome	Rate vs pre-AI baseline	Material increase	Patient-safety event (incident runbook)
Fairness drift	Subgroup performance gap	Increase beyond tolerance	Equity review
System SLA	Latency / availability	SLA breach	SRE on-call

§ 9

Limitations register

#	Limitation	Impact	Mitigation	Owner
1	External validation limited to 1 site	Unknown generalisation	Multi-site rollout phased; site-1 monitoring intensified	Model owner
2	Limited representation of LEP patients	Subgroup under-performance risk	LEP cohort upweighting next cycle; clinical override emphasised	Model owner + Equity
3	EHR upgrade scheduled mid-deployment	Documentation shift	Shadow re-validation pre / post upgrade	Engineering + MRM
4	Outcome label noisy	Calibration uncertainty	Adjudicated subsample; CI widened	Quality
5	Long-term outcome registry incomplete	Long-term accuracy unknown	Registry linkage roadmap stated	Quality
6	Cold-start on new diagnostic codes	Hidden generalisation gap	Code-set version pinned; alert on new codes	Engineering

§ 10

Regulatory alignment (FDA SaMD / MHRA / EU AI Act)

Framework	Classification	Implication
FDA SaMD (IMDRF)	State of healthcare situation × significance of information → Class I–IV	Drives clearance / approval path (510(k), De Novo, PMA) where applicable
FDA Predetermined Change Control Plan	Submitted ahead of model update if applicable	Defines what changes can occur without resubmission
MHRA SaMD guidance (UK)	Similar risk-based scheme	Mirrors many FDA expectations; check post-Brexit specifics
EU AI Act	High-risk (most clinical AI)	Conformity assessment, post-market monitoring, transparency, human oversight
EU MDR (CE marking)	Where qualifying as medical device	Notified-body conformity
HIPAA	Privacy & Security Rules	Covered separately in companion DPIA

A formal regulatory determination is required for every clinical AI before live use. Determination is signed by the regulatory officer and recorded in this section.

§ 11

Change-control & re-validation triggers

- Material change to model architecture, target definition, or training population
- Material change to inputs (new feature, deprecated feature, EHR mapping)
- Sustained breach of any monitor in § 8
- Subgroup performance regression on equity guardrails
- Patient-safety event linked to AI recommendation
- External regulatory change (FDA, MHRA, EU AI Act, HIPAA)
- Annual review cadence

Change-control gate

Change type	Model owner	Clinical lead	Quality	Regulatory
Re-calibration	Plan	Notified	Sign	Notified
Re-train (same architecture)	Develop	Review	Sign	Sign
New version (structural)	Develop	Approve	Sign	Sign + re-submit if applicable
Retirement	Propose	Approve	Sign	Notified

§ 12

Model card · summary block

Identity

Model ID · version · owner · date · regulatory class (if any)

Intended use

Decision · population · setting · out-of-scope

Cohorts

Training · internal · external · prospective

Performance

Headline + subgroup with CI

Calibration

Reliability evidence + operating point rationale

Shift sensitivity

Test outcomes and mitigations

HITL contract

Automation boundary table

Monitoring

SLIs + thresholds + escalation

Limitations

Register with mitigations + owner

Approvals

Sign-off block (next page)

§ 13

Evidence index

Evidence #	Description	Section	Location	Owner
CV-01	Intended use & contra-indications	§ 1	Doc	Clinical lead
CV-02	Cohort & demographics	§ 2	Repo	Data eng
CV-03	Performance & subgroup tables	§ 3	Validation report	Model owner
CV-04	Calibration evidence	§ 4	Validation report	Model owner
CV-05	Dataset-shift tests	§ 5	Validation report	Model owner
CV-06	HITL contract	§ 6	Doc	Clinical lead
CV-07	Outcome linkage spec	§ 7	Quality repo	Quality
CV-08	Monitoring contract	§ 8	MRM repo	Model owner + MRM
CV-09	Limitations register	§ 9	MRM repo	Model owner
CV-10	Regulatory determination	§ 10	Regulatory repo	Regulatory
CV-11	Change-control records	§ 11	Registry	Engineering
CV-12	IRB / ethics review (if applicable)	—	IRB	Clinical

§ 14

Sign-off & attestation

Sign-off block

Role	Name	Date	Signature
Chief Medical Officer			
Model owner			
Clinical safety officer			
Quality / patient safety			
Regulatory officer			
Privacy / HIPAA officer			

Production go-live requires every row above to be completed. The signed validation report is filed in the regulatory repository and re-affirmed at every change-control gate (§ 11).