Deployer-Side XAI Instrumentation for Regulated AI: A Clinical Case Study in ICL Sizing

Dorleta Urrutia-Onate[1], Enrique Onieva[2], Asier Perallos[3]
University of Deusto, Avda. Universidades 24, 48007 Bilbao, Spain[1,2,3]
Faculty of Business, Mondragon University, Ibarra Zelaia 2, 20560 Oñati, Spain[1]

***Abstract*—Regulated AI creates a monitoring problem for deployers who must organise human oversight, log-retention and post-market surveillance while often having access only to the prediction interface. This study specifies a deployer-side XAI instrumentation protocol for the output→action boundary, where a model output becomes a reason for action. The protocol reorganises KernelSHAP, nearest-neighbour envelope checks, bounded perturbation, and rank-order stability into a per-decision evidence record computed from `predict()` calls. We instantiate the protocol in a clinical case of phakic Implantable Collamer Lens sizing, using a 55-eye held-out cohort and an Extra Trees regressor for post-operative vault prediction. The record contains five signals: `score_margin`, `constraint_enforcement`, `envelope_validity`, `decision_robustness`, and `record_integrity`, plus two cohort-level oversight aggregates. The case study shows how the same record can support decision-time human oversight, later audit and post-market surveillance under the EU AI Act and the Medical Devices Regulation.**

***Keywords*—Explainable AI; XAI instrumentation; human oversight; Medical Device Software; EU AI Act; ICL sizing**

I. Introduction

Regulated AI creates a practical problem for deployers. They must monitor systems they often cannot inspect. Source code, gradients, training data, weights, tree structures and provider-side documentation may remain unavailable because of proprietary, contractual or organisational restrictions. In many deployments, the deployer can only call the prediction interface. This limited-access setting has to be reconciled with legal duties of human oversight, monitoring, log retention and post-market surveillance under the EU AI Act [1] and the Medical Devices Regulation (MDR) [2]. Explainable AI (XAI) methods can operate from this interface: KernelSHAP [3] against a fixed background sample, distance-based envelope checks and bounded perturbation tests. What is still missing is a specification of which method should produce which evidence, for which deployer-side function, and in what computable form when a prediction is used as a reason for action.

This study specifies that mapping. Its objective is to make a single prediction usable as defensible evidence for action under regulation, using only the prediction interface. The study defines the output→action boundary—the point at which a prediction becomes a reason for action—as its unit of analysis, maps four documented failure modes at that boundary to measurement requirements, and specifies a decision-boundary instrumentation protocol that computes the required evidence from `predict()` calls alone.

This study tests that protocol in a clinical case. Phakic Implantable Collamer Lens (ICL) sizing is a pre-operative decision in which the surgeon chooses one of four available lens sizes. The main safety variable is the post-operative vault, which should remain within a clinically acceptable band to avoid contact- and pressure-related complications [4]. The band used in this study is defined in Section IV-A. A model that predicts post-operative vault from pre-operative biometry and anterior-segment OCT can therefore influence a high-stakes clinical decision. In the regulatory setting used here, the tool is treated as Medical Device Software under the MDR and as high-risk Medical Device Artificial Intelligence under Annex III of the AI Act. The clinical deployer is then responsible for organising oversight in use.

The protocol emits a per-decision telemetry record whose five core signals are `score_margin`, `constraint_enforcement`, `envelope_validity`, `decision_robustness` and `record_integrity`, complemented by two cohort-level oversight fields, `sustainability_impact` and `human_ai_teaming`. It is instantiated end-to-end on a 55-eye held-out cohort of an ICL vault prediction regressor (Extra Trees, held-out mean absolute error 160.37 µm), producing per-decision JSON records and a 4×55 counterfactual matrix over the lens-size decision space, and the resulting record is mapped to EU AI Act Article 14 and MDR Articles 83 and 88. Model performance and lens-size superiority are outside the study's claim.

The study is organised as follows: Section II reviews XAI methods, the deployer's position under regulation, the standards frame and the clinical context of ICL sizing. Section III defines the methodological framework and the output→action boundary as the unit of analysis. Section IV introduces the case study, dataset, deployment context and deployed regressor. Section V develops the signal architecture under limited-inspection deployment. Section VI applies the protocol to the held-out cohort, reports the per-decision records, estimates cohort-level review workload and analyse the counterfactual sweep over the four lens sizes. Section VII and Section VIII discuss the findings and conclude.

II. Background and Related Work

A. XAI Methods and their Classical Destination

Post-hoc XAI is usually grouped into four method families. Feature-attribution methods, including KernelSHAP [3] and LIME [5], assign a per-feature contribution to a single prediction. Counterfactual-explanation methods [6], including DiCE-style procedures, search the input neighbourhood for changes

that would alter the model output. Prototype- and influence-based methods locate a prediction in relation to training examples. Surrogate-tree methods approximate an opaque model with a globally interpretable structure. Arrieta et al. [7] provide the taxonomy used here as background.

Recent work has made the model-centred use of post-hoc XAI harder to defend without qualification. Adebayo et al. [8] showed that several saliency methods can produce visually similar attributions even after model randomisation. Slack et al. [9] showed that LIME and SHAP can be manipulated by an adversary controlling the prediction interface, leaving the explanation blind to fairness-relevant features. Rudin [10] argued that high-stakes decisions should use interpretable models where possible instead of post-hoc explanations for opaque models.

A second critique matters more for deployment: who is meant to use the explanation, and for what. Bhatt et al. [11] document that XAI outputs in many organisations are still mainly consumed by model developers, rather than by the operators who must act on model outputs. Liao and Varshney [12] argue that XAI remains incomplete when the user question is not specified. Doshi-Velez and Kim [13] anticipated this separation by distinguishing functionally grounded proxies from human-grounded and application-grounded evaluations.

The literature, therefore, separates two uses of post-hoc XAI. One use serves development: model construction, debugging and selection. It works close to the model and characterises model-level behaviour. The other use serves deployment: a specific output is about to be used as a reason for action. It works from the prediction interface and must characterise one decision in context.

Table I summarises this distinction. Fig. 1 locates the point where the two uses meet: the output→action boundary. Most XAI methods, including the methods used in this study, were designed for model-centred work. The question here is how to reorganise them into deployer-side evidence records.

For the deployer, the useful destination of XAI is the decision record. A SHAP attribution, a counterfactual explanation or a surrogate model is useful in regulated deployment only if it helps document why a specific output could be used, disregarded, overridden or reviewed at that point in the workflow. This study, therefore, treats XAI as instrumentation for functional characterisation, observable behaviour analysis, validation and technical trust evidence at decision time.

B. The Deployer's Predicament Under Regulation

The EU AI Act distinguishes the provider that develops or places an AI system on the market from the deployer that uses the system under its authority. Article 26 assigns deployers of high-risk AI systems duties around use according to instructions, assignment of human oversight, monitoring of operation, reporting of risks or serious incidents, and retention of automatically generated logs when those logs are under their control [1]. Article 14 defines the human-oversight capabilities that high-risk systems must enable: understanding capacities and limitations, remaining aware of automation bias, interpreting outputs and deciding when to disregard, override or reverse an output [1].

The MDR adds clinical evaluation, post-market surveillance and vigilance [2]. For AI-enabled medical devices, the MDCG–AIB guidance [14] clarifies that the AI Act and the MDR/IVDR apply together, while separating the AI Act deployer from the MDR manufacturer. In many clinical uses, the surgeon, hospital or clinic operating the tool is the deployer.

This creates a gap between responsibility and access. The deployer must monitor a system in use, while model internals, training data, source code and provider-side validation material may remain unavailable for technical, legal, organisational or commercial reasons [15], [16], [17]. This study addresses that access regime: the deployer has the case input, a fixed background sample and the ability to call `predict()`.

C. Regulatory and Standards Frame

The operational frame combines the binding regulation of Section II-B with voluntary management standards.

Voluntary frameworks give organisations procedures for organising AI risk and evidence. The NIST AI Risk Management Framework 1.0 [18] organises that work through four functions: *Govern*, *Map*, *Measure* and *Manage*. ISO/IEC 42001:2023 [19] specifies an AI management system based on a plan-do-check-act cycle and auditable operational records.

These frameworks define the organisational need: risks must be governed, mapped, measured, managed and recorded. The protocol developed here supplies an operational mechanism for producing such records at the point where an AI output may become a reason for action.

D. Clinical Context: ICL Sizing as a Regulated AI Decision

The case study uses phakic ICL sizing for refractive correction. The ICL outcomes literature commonly treats the band $[250, 750]$ µm as the range of clinical sufficiency for post-operative vault [4]. This study adopts that band as the operational decision boundary for the case study (Section IV-A).

An ICL sizing tool can influence a high-stakes clinical decision. The model output can shape the lens size selected before surgery. If the predicted vault lies outside the clinically acceptable band, the output becomes part of the reasoning behind a surgical decision with safety implications. Under the medical-device regime, that use needs evidence for oversight, validation and post-market monitoring.

Clinical XAI literature raises the same concern from the healthcare side. Babic et al. [20] warn that explanations in healthcare can overstate what they show, especially when post-hoc rationalisations do not reflect the basis of the prediction. Tonekaboni et al. [21] show that clinicians ask for explanations tied to the clinical situation, timing and decision context, while many XAI methods provide outputs that do not match those needs. Amann et al. [22] frame explainability in healthcare as a multidisciplinary requirement linked to ethical values, safe use and accountability. These contributions support the move from model-centred explanation to decision-time evidence.

The dataset used here comes from clinical operations at three centres of a multi-site ophthalmology network. It was first organised for predictive modelling in earlier clinically

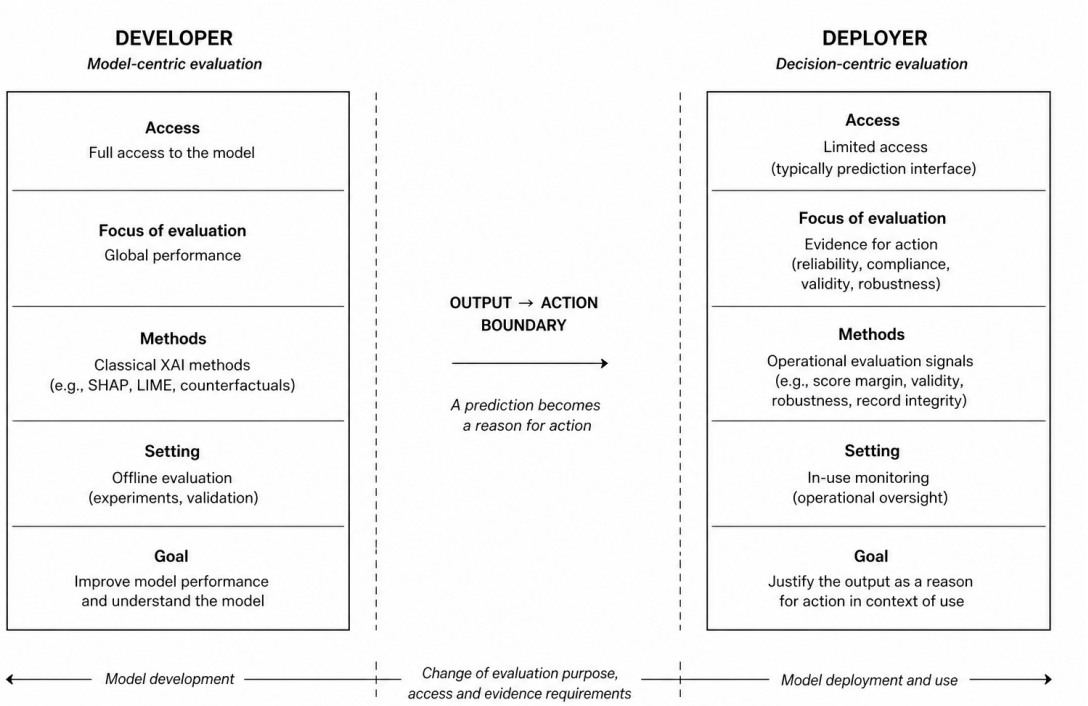

Fig. 1. Developer-side and deployer-side destinations for post-hoc XAI under regulated deployment. The developer-side destination is model-centred: it assumes broad access to the model and supports development, debugging and model selection. The deployer-side destination is decision-centred: it usually works from the prediction interface and supports evidence for action, in-use monitoring and audit. The two destinations meet at the output→action boundary, where a prediction becomes a reason for action.

TABLE I. DEVELOPER-SIDE XAI VS. DEPLOYER-SIDE INSTRUMENTATION: TWO DESTINATIONS FOR POST-HOC EXPLANATION UNDER REGULATED DEPLOYMENT.

Aspect	Developer-side XAI	Deployer-side instrumentation
Access to the model	weights, gradients, training data, architecture, internals	prediction interface only (`predict()` calls)
Unit of analysis	the model	the individual decision
Phase of the lifecycle	development, debugging, model selection	deployment, decision time, post-market surveillance
Primary consumer	model developer, ML researcher	deployer: operator, clinician, audit team
Output form	global explanations, fidelity diagnostics, debugging outputs	per-decision evidence record + cohort aggregates
Regulatory anchor	usually indirect	EU AI Act Art. 14 and Art. 26; MDR Art. 83 and Art. 88
Question answered	does this model behave as expected in general?	is this specific output defensible as a reason for action now?

supervised work. This study inherits the clinical motivation, feature inventory and filtering criteria from that work.

The dataset and the regression model are used to expose the AI-decision structure on which the instrumentation protocol is tested.

E. Accountability Artefacts and Decision Logging

A parallel line of work proposes documentation and logging artefacts for accountable AI. Model Cards [23], FactSheets [24] and Datasheets for Datasets [25] standardise what a provider should disclose about a model, a service or a dataset, while internal audit frameworks such as the one of Raji et al. [26] organise accountability as an organisational process. Decision provenance [27] extends this from static documentation towards logging the data flow behind individual decisions. These artefacts are mostly produced at model or dataset release, on the provider or developer side, and they characterise the system rather than a single output. The protocol developed here is complementary and sits at a different point: it is a per-decision, deployer-side evidence record computed at prediction time from `predict()` calls, so that the actor who acts on an output can attach defensible, auditable evidence to that specific decision under limited inspection.

F. Position of this Work

The architecture maps the deployer's monitoring problem to four failure modes at the output→action boundary, each turned into a measurement computable from `predict()` calls alone (Section III). This study adapts it to a clinical regression case and applies it end to end to a real ophthalmology cohort.

The ICL case has three useful properties for the protocol.

First, the predictive task is regression on a continuous safety variable, so the relevant boundary is a clinical band rather than a probability threshold. Second, the dataset comes from a real multi-centre clinical cohort operating under the medical-device regime. Third, the decision space is small and discrete: the surgeon chooses one of four lens sizes. That discreteness allows the counterfactual component to enumerate all decision alternatives. For each held-out eye, the protocol queries the model under each candidate lens size while holding anatomy and biometry fixed, and then recomputes the boundary-relevant signals for each alternative. The premise is narrow. XAI outputs used in regulated deployment must be structured well enough to count as evidence. The question is whether each decision record supports functional characterisation, observable behaviour analysis, validation and technical trust evidence at the moment when a prediction is used as a reason for action.

III. Methods

This section defines the framework applied in the clinical case. The framework was introduced in the companion manuscript [28]. Here it is adapted to a regression task on a continuous safety variable and later implemented for the ICL vault prediction model.

A. Unit of Analysis: The Output→Action Boundary

The framework takes the output→action boundary as its unit of analysis. This is the point at which a single prediction becomes a reason for action in a clinical, financial or industrial workflow.

At that point, the question is evidential. The deployer responsible for monitoring the system in use must be able to attach evidence to the prediction before the decision is taken. Under the access conditions described in Section II-B, that evidence must be computable from the prediction interface. The framework therefore specifies what evidence is needed, how it is structured and which part of the output→action relation it characterises.

B. From Failure Modes to Measurement Requirements

Four safety and human-factors perspectives identify failure modes at the output→action boundary.

Leveson [29] identifies constraint-enforcement failure: the action proceeds on grounds that the institution would not recognise as legitimate. Rasmussen [30] identifies envelope migration: the case has moved beyond the population on which the model was validated. Parasuraman [31] identifies automation bias: a fragile output is treated as robust because the operator has no evidence of fragility. A fourth failure is procedural: accountability failure, where no reproducible record links the output, the decision context and the evidence used at the time of action.

Each failure mode defines a measurement requirement. Constraint-enforcement failure requires evidence about whether the local prediction is supported by legitimate variables or by decision-linked variables. Envelope migration requires evidence about the distance between the case and the validated population. Automation bias requires evidence about whether decision-band membership is stable under bounded perturbation. Accountability failure requires evidence about whether the explanation is stable enough to be used as a record.

The framework also includes a domain-derived signal. In the ICL case, the predicted vault must be located relative to the operational clinical band. This signal indicates whether the predicted vault is safely inside the band, close to a threshold or outside the clinically acceptable range.

C. From Measurement Requirements to Prediction-Interface Methods

The deployer's access condition restricts the available methods. The framework therefore selects XAI and validation methods that can operate from the prediction interface.

KernelSHAP [3] computes feature attributions from `predict()` calls against a fixed background sample. It can therefore measure attribution mass on legitimate and decision-linked variables. *k*-nearest-neighbour distance operates on the input feature space and measures envelope validity. Bounded perturbation repeatedly calls `predict()` on perturbed versions of the same input and estimates the flip rate used for `decision_robustness`. Rank-order stability under micro-noise recomputes the attribution procedure on small perturbations of the input and produces `record_integrity`.

The methodological step is the binding between failure mode, measurement requirement, computable signal and prediction-interface method. Each XAI or validation method receives a specific evidential function at the output→action boundary.

D. Properties of the Resulting Protocol

The protocol has three methodological properties.

First, the protocol is compatible with prediction-interface-only access. It requires the case input, a fixed background sample and the ability to call `predict()`. It does not require model weights, gradients, trees, source code, training data or provider-side documentation.

Second, the protocol is non-compensatory. Each signal addresses a different failure mode. A high value on one signal does not offset a failure on another. A stable attribution does not compensate for a prediction outside the clinical band. Robust band membership does not compensate for reliance on decision-linked variables.

Third, the protocol is decision-boundary local. Each signal is computed for the specific case on which the deployer is acting. The protocol produces a per-decision evidence record attached to one prediction at the moment that prediction may support action. Model-level validation remains necessary, but it answers a different question.

IV. Case Study

A. Clinical Target and Decision Space

The clinical target of an ICL sizing model is the post-operative vault: the distance between the implanted lens and the natural crystalline lens. A vault below the clinically acceptable range may increase the risk of contact-related

complications. A vault above that range may increase intraocular pressure and the risk of secondary glaucoma. The ICL outcomes literature commonly uses $[250, 750]$ µm as the range of clinical sufficiency [4].

The decision supported by the tool is the pre-operative choice of lens size. In this case, the available sizes are the four standard options used at the contributing sites: {12.1, 12.6, 13.2, 13.7} mm. The decision uses anterior-segment optical coherence tomography (OCT), biometric measurements and the clinician's selected refractive correction.

The regression model analysed here evaluates candidate decisions. For each candidate lens size and selected refractive correction, the model uses the available anatomical and biometric measurements to predict the post-operative vault expected from that combination. The model supports sizing by estimating the safety variable associated with each candidate decision.

B. Dataset

The dataset comes from clinical ophthalmology operations at three centres of a multi-site ophthalmology network. The data were first organised for predictive modelling in earlier clinically supervised work. This study inherits three elements from that work: clinical motivation, feature inventory and cohort filtering criteria.

The clinical motivation is vault prediction, as the principal safety variable in ICL sizing. The feature inventory contains 25 pre-operative variables combining anterior-segment OCT indices, biometric and refractive measurements, and categorical descriptors of the implant decision. The filtering criteria retain eyes with complete pre-operative OCT, biometry and post-operative vault measurement. After filtering, the cohort contains 545 eyes.

The dataset is reused to evaluate whether the instrumentation protocol can produce decision-level evidence from a deployed regressor.

The feature set contains 25 pre-operative variables grouped into three sets.

The OCT set includes anterior-segment indices: lens rise, anterior depth, limbus diameter and centre, pupil diameter and centre, and the temporal/nasal AOD/TISA pairs at the 750 µm reference.

The biometric and refractive set includes k1 keratometry, patient age at operation, pre-operative spherical equivalent and selected refractive correction.

The decision-descriptor set includes categorical variables encoded as one-hot indicators: lens family (`ICL`, `TICL`), refractive type (`Myopic`, `Toric Myopic`) and lens-size dummies (`lens_12.1`, `_12.6`, `_13.2`, `_13.7`).

The lens-size variables are encoded as four independent indicators, not as one numerical variable. This avoids treating lens size as a continuous ordered scale. It also aligns the feature representation with the discrete decision space defined in Section IV-A. The same encoding supports the counterfactual sweep in Section VI-C, where each eye is evaluated under the four candidate lens sizes.

All 25 features are available at the moment of the sizing decision. This matters for the deployment-time evaluation in Section VI: the model is evaluated using only information available before the clinical decision. Continuous features are imputed with the median of the training partition and the same imputers are applied to the held-out partition, avoiding held-out information in the imputation step.

Case identifiers used in the body of the study are anonymised as `case1`, `case2`, ..., `case5`. These labels do not correspond to clinical-record database entries.

C. Deployment Context

ICL sizing models are no longer only isolated research prototypes. Since 2021, several machine-learning pipelines have been published for vault prediction and lens-size selection on cohorts ranging from hundreds to thousands of eyes. These include Random Forest and gradient-boosted regressors trained on large EVO-ICL records [32], regression-tree ensembles trained on multicentre OCT–MS39 data [33], and deep-learning models trained on raw SS-OCT images and made publicly available through the `safevaulticl.com` interface [34].

The model studied here belongs to that emerging class of clinical decision-support software. Its relevance for this study is the operational structure of the decision: a model output can become part of the reasoning behind a surgical choice.

Section II set out the regulatory frame. In this case, the relevant actor is the clinical user of the system. In a clinical setting, the operator of the sizing AI is typically the refractive surgeon or the clinic's data team, not the model developer. This actor is the deployer under Article 26 of the EU AI Act and must organise system use, monitoring and human oversight in practice [1].

Article 14 defines the human-oversight capacities that the system must enable. Article 26 places the deployer in the operational position where those capacities must be used. MDR Articles 83 and 88 add the post-market surveillance and trend-reporting frame for the medical-device setting [2].

The problem is operational. The deployer is expected to monitor the system in use, while access to the model's internal design, training process and provider-side documentation may be limited [15], [16], [17]. The case study asks what evidence the deployer can document when an ICL vault prediction is used as a reason for a sizing decision.

D. The Deployed Regressor

From the 545-eye dataset described in Section IV-B, 490 eyes are used for training and 55 for held-out evaluation. The split is reproducible, with `test_size = 0.10` and `seed = 42`.

Five standard scikit-learn regressors are evaluated on the training partition using 10-fold cross-validation and mean absolute error (MAE) as the scoring metric: Extra Trees [35], Random Forest [36], Ridge regression, histogram-based gradient boosting and standard gradient boosting [37]. Table II reports the cross-validated MAE for each candidate.

TABLE II. CROSS-VALIDATED COMPARISON OF CANDIDATE REGRESSORS ON THE TRAINING PARTITION (490 EYES, 10-FOLD CV).

Rank	Model	MAE_cv (μm)
1	**Extra Trees** *(selected)*	**167.24**
2	Ridge	175.73
3	Random Forest	176.86
4	HistGradientBoosting	180.50
5	Gradient Boosting	182.99

The Extra Trees regressor is selected for the lowest cross-validated MAE; on the held-out partition its MAE is 160.37 μm, close to the cross-validated training value, which we read as evidence of no obvious train/held-out drift rather than a precision claim.

The model is serialised with its column order, per-feature imputation medians and lens-size column name; this object is the only model input consumed by the protocol in Section V. The regressor is not meant to outperform the clinical state-of-the-art: it is the test bed for the instrumentation protocol.

V. SIGNAL ARCHITECTURE UNDER LIMITED-INSPECTION DEPLOYMENT

A. Failure Modes at the Output→Action Boundary and the Measurements they Require

Sections II-B and Section II-C define the deployer-side access condition. Section III defines the failure-mode/measurement-requirement chain. This section turns that chain into operational signals and prediction-interface methods. The clinical case adds one domain signal: the distance between the predicted vault and the centre of the operational clinical band. That domain signal orients the four failure-mode signals. The methodological contribution is the binding itself: existing XAI and validation methods are assigned, one by one, to specific failure modes at the output→action boundary under prediction-interface-only access. Table III summarises the chain.

Each signal follows from a failure mode and from the access conditions of deployment. The chain is: failure mode, required measurement, computable signal, prediction-interface method.

B. Cohort-Level Aggregates: From Per-Decision Telemetry to Oversight Workload

The five signals in Section V-A, together with case identifier, timestamp, model version and lens-size selection, form the telemetry record emitted at the output→action boundary. The operator can use this record at decision time and persist it afterwards as the audit unit.

The record also contains two oversight fields: `sustainability_impact` and `human_ai_teaming`. Their cohort-level aggregates estimate how many outputs remain inside the operational band and how many can proceed under the AI-supported workflow before human review. This matters because Article 14 requires oversight measures that allow assigned persons to understand system capacities and limits, remain aware of automation bias, interpret outputs and decide when to disregard, override or reverse them [1].

`human_ai_teaming_cohort` measures the proportion of decisions that satisfy the joint sufficiency criterion envelope_validity $\geq \tau_{\text{env}}$ and score_margin $\geq \tau_{\text{margin}}$. Its complement is the expected human-review workload under Article 14. Both aggregates are computed from the same per-decision telemetry and require no additional measurements. The thresholds $(\tau_{\text{env}}, \tau_{\text{margin}})$ are set by the deployment context, balancing unnecessary routing against missed review.

C. Pipeline and Per-Decision Evidence Record

The three methodological properties in Section III-D instantiate, as shown in Fig. 2. The Listing given below gives one telemetry record from the held-out cohort. The same record is the decision-time view for the clinician and the later audit unit used by post-market surveillance.

Section VI reports what the protocol produces on every held-out eye.

VI. APPLICATION TO THE ICL VAULT PREDICTION MODEL

This section applies the protocol of [28] to the deployed ICL vault regressor, showing the evidence it produces from `predict()` calls only, for one sizing decision and for cohort-level oversight. Box 1 presents the glossary signals and parameters used in Section VI.

Section VI-A defines the calibration choices; Section VI-B reports the output on the 55-eye held-out cohort; Section VI-C evaluates the four lens sizes and the counterfactual matrix for oversight and post-market surveillance; Section VI-D states reproducibility.

A. Operational Calibration

Applying the architecture (Section V) to the deployed regressor (Section IV-D) requires four calibration choices: the clinical band (`score_margin`), the feature taxonomy (`constraint_enforcement`), the perturbation parameters (`decision_robustness`, `record_integrity`), and the operating point for cohort-level oversight. The clinical band $[250, 750]$ μm is taken from Section IV-A unchanged; Table IV summarises the calibration.

Clinical band. The `score_margin` signal reports the distance of the predicted vault from the centre of the clinical band, normalised by the band half-width. With centre $c = 500$ μm and half-width $h = 250$ μm, the signal is:

$$\text{score\_margin}(\hat{y}) = 1 - \frac{|\hat{y} - c|}{h} \tag{1}$$

A prediction at the band centre yields 1.0, at either edge 0.0, and outside the band a negative value whose magnitude indicates how far outside the case lies.

Feature taxonomy. The legitimate/decision-linked split defines the clinical meaning of `constraint_enforcement`. In this case, the split separates variables that describe the eye from variables that encode, or partly encode, the clinical decision. The deployer needs to know whether the local prediction is mainly supported by anatomical and biometric

TABLE III. FAILURE MODES AT THE OUTPUT→ACTION BOUNDARY, THE MEASUREMENT THEY REQUIRE, THE SIGNAL THAT DELIVERS IT, AND THE PREDICTION-INTERFACE-COMPUTABLE METHOD.

Perspective	Failure	Measure	Signal	Method
Domain (clinical band)	Case far from operational threshold	Distance from band centre	`score_margin`	scalar of the prediction
Leveson (STAMP) [29]	Action proceeds on non-legitimate or decision-linked features	Share of attribution mass on legitimate features	`constraint_enforcement`	KernelSHAP [3] vs fixed background
Rasmussen [30]	Case outside validated envelope	Distance to validated population	`envelope_validity`	k-NN distance in feature space
Parasuraman [31]	Brittle output read as robust	Flip-rate under bounded perturbation	`decision_robustness`	repeated `predict()` over perturbations
Cross-cutting accountability	No reproducible record of evidential basis	Rank-order stability of the attribution	`record_integrity`	KernelSHAP recomputed under micro-noise

Listing. Example structure of one per-decision telemetry record.

```
{
  "case_id": "case1",
  "timestamp": "2026-04-30T12:34:56Z",
  "model_version": "extra_trees_icl_v1",
  "predicted_vault_um": 498.5,
  "lens_option_mm": 13.2,
  "core_signals": {
    "score_margin": 0.99,
    "constraint_enforcement": 0.73,
    "envelope_validity": 1.00,
    "decision_robustness": 1.00,
    "record_integrity": 0.86
  },
  "oversight_fields": {
    "sustainability_impact": "inside_clinical_band",
    "human_ai_teaming": "ai_supported_workflow_sufficient"
  },
  "counterfactual_by_lens_mm": {
    "12.1": {"vault_um": 316, "score_margin": 0.27, "in_band": true},
    "12.6": {"vault_um": 351, "score_margin": 0.40, "in_band": true},
    "13.2": {"vault_um": 499, "score_margin": 0.99, "in_band": true},
    "13.7": {"vault_um": 684, "score_margin": 0.26, "in_band": true}
  }
}
```

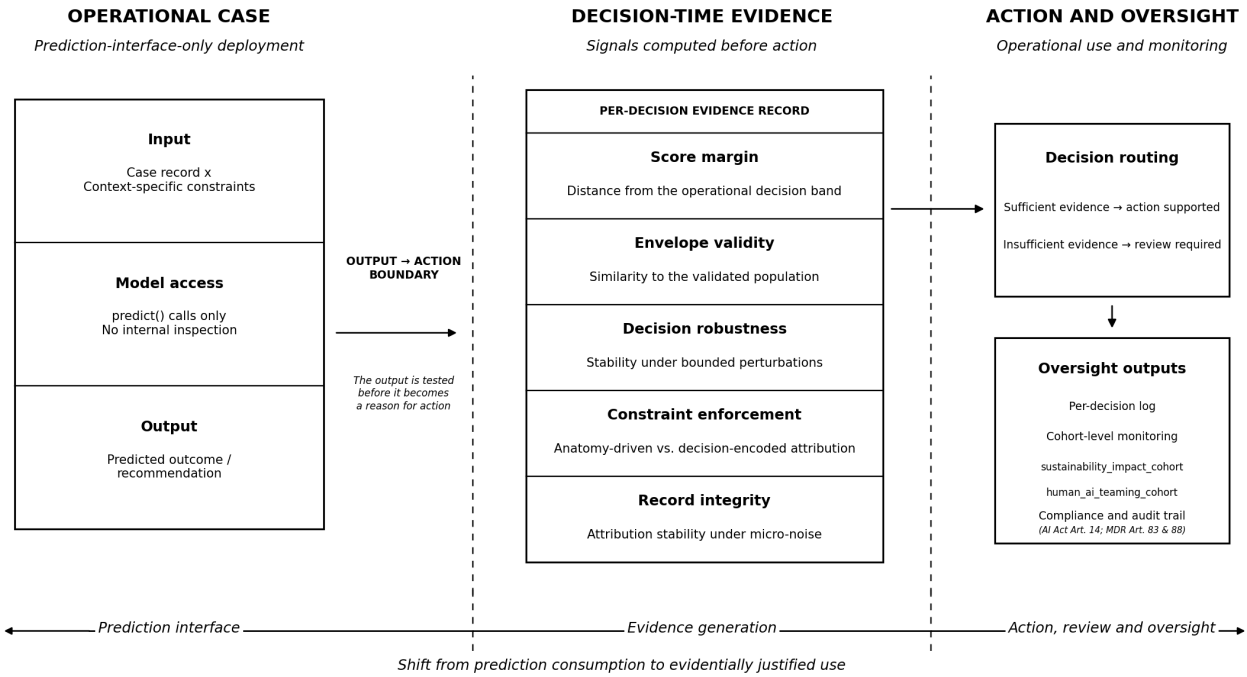

Fig. 2. Pipeline of the protocol on a single decision. The deployer queries the model through `predict()` only. The five core signals and two per-decision oversight fields are emitted as a JSON evidence record. Cohort-level aggregates are computed later by aggregating persisted records across cases.

Box 1 — Glossary of signals and parameters used in Section VI. The five core signals, the two cohort-level oversight aggregates and the calibration parameters used throughout this section are summarised below. Operational definitions and formulas are given in Section VI-A. Cohort-level results are reported in Section VI-B to Section VI-C. Dataset feature names with their legitimate/decision-linked classification appear in Table V.

Symbol / field	Description	Range / value
`score_margin`	distance to clinical-band centre, normalised by half-width	$(-\infty, 1]$
`constraint_enforcement`	KernelSHAP attribution mass on legitimate features	$[0, 1]$
`envelope_validity`	k-NN distance beyond validated population, clipped	$[0, 1]$
`decision_robustness`	fraction of perturbations preserving band membership	$[0, 1]$
`record_integrity`	mean Spearman rank-order stability under micro-noise	$[-1, 1]$
`sustainability_impact_cohort`	cohort fraction with prediction inside [250, 750] µm	$[0, 1]$
`human_ai_teaming_cohort`	cohort fraction satisfying joint sufficiency	$[0, 1]$
τ_{env}	sufficiency threshold for `envelope_validity`	0.5
τ_{margin}	sufficiency threshold for `score_margin`	0.3
σ	bounded-perturbation magnitude for `decision_robustness`	0.10 × std
ε	micro-noise magnitude for `record_integrity`	0.01 × std
N	bounded-perturbation samples per case	100
M	micro-noise replicates per case	50
k	k-NN neighbourhood size	5
$d_{\max}$	envelope distance threshold	90th percentile of train 5-NN mean
X_{bg}	KernelSHAP background sample size	50
$\mathcal{L}, \mathcal{D}$	legitimate / decision-linked feature partition (Table V)	14 / 11

Box 2 — Justification of calibration parameters. Each non-domain parameter used in Section VI-A is summarised here with its operational rationale and references.

Parameter	Value	Rationale	Reference / source
Clinical band $[v_{\text{lo}}, v_{\text{hi}}]$	$[250, 750]$ µm	domain-derived; clinical literature on ICL vault outcomes	[4]
X_{bg}	50	background sample size for KernelSHAP; in the recommended 50–1000 range, with empirical stability of SHAP explanations documented for sample sizes of this order	SHAP documentation; [38]
k	5	small neighbourhood size for k-NN-based out-of-distribution detection on a 490-eye reference set; consistent with deep k-NN OOD-detection literature	[39]
$d_{\max}$ rule	90th percentile of train 5-NN mean	percentile-based threshold for out-of-distribution detection; less sensitive to training-set outliers than a maximum-distance rule	[39]
σ	$0.10\times$ std	bounded-perturbation magnitude; of the order of inter-operator anterior-segment OCT measurement variability reported in the imaging literature	domain-derived; calibrated against OCT inter-operator variability
ε	$0.01\times$ std	micro-noise magnitude one order of magnitude below σ; tests rank-order stability of the explanation under noise that should leave clinical interpretation unchanged	derived; clinical-interpretation invariance
N	100	bounded-perturbation samples per case; standard Monte Carlo budget for flip-rate estimates at the precision used by `decision_robustness`	empirical; pilot stability test
M	50	micro-noise replicates per case; sufficient sample for the Spearman rank-order correlation estimate used by `record_integrity`	empirical; pilot stability test
$(\tau_{\text{env}}, \tau_{\text{margin}})$	$(0.5, 0.3)$	deployer-side operating point; conservative defaults used throughout this study. Sensitivity at three alternative values is reported in Table VI	deployer-side domain choice

TABLE IV. OPERATIONAL MAPPING OF THE 5 + 2 SIGNALS ON THE ICL VAULT PREDICTION MODEL

Signal	Domain choice	Parameters
`score_margin`	clinical band [250, 750] µm	half-width $h = 250$, centre $c = 500$
`constraint_enforcement`	L = 14, D = 11 (Table V)	$X_{\text{bg}} = 50$, `seed = 42`
`envelope_validity`	25-D feature space, post-imputation	$k = 5$, $d_{\max}$ = max train 5-NN mean
`decision_robustness`	bounded Gaussian perturbation on continuous features	$N = 100$, $\sigma = 0.10$ std
`record_integrity`	rank-order stability under micro-noise on continuous features	$M = 50$, $\varepsilon = 0.01$ std
`sustainability_impact_cohort`	inside clinical band	population fraction
`human_ai_teaming_cohort`	sufficiency thresholds	$\tau_{\text{env}} = 0.5$, $\tau_{\text{margin}} = 0.3$

measurements, or whether it is reproducing variables derived from the clinician's prospective choice.

The 25 pre-operative features are divided into a legitimate set $\mathcal{L}$ of fourteen anatomical, biometric, age and laterality variables, and a decision-linked/report-dependent set $\mathcal{D}$ of eleven variables. The full taxonomy is shown in Table V.

The legitimate set contains pre-operative anatomical and biometric measurements that exist independently of the sizing decision. The decision-linked/report-dependent set contains features whose values encode the clinician's prospective lens-size selection (`lens_*`, `selected_lens_size`, `selected_sph`), lens-family and refractive-type dummies linked to the clinical decision (`ICL`, `TICL`, `Myopic`, `Toric Myopic`), and patient-reported refraction, which is subject to instrument-of-reporting variability (`preop_sph`).

The *legitimate/decision-linked* partition is informational, not a judgement of clinical validity: lens size physically influences vault, so decision-linked variables are clinically relevant inputs. `constraint_enforcement` asks whether the prediction differentiates cases mainly through anatomical evidence (a high value: anatomy and biometry carry most of the attribution mass) or by tracking the lens choice already encoded in the input (a low value: dominated by decision-linked or report-dependent variables). Both modes can be clinically plausible and provide different evidence at the output→action boundary.

In the counterfactual sweep of Section VI-C, the lens-size variables are queried under each of the four candidate values as intervention variables. By contrast, `constraint_enforcement` is computed at the originally observed prediction, with all input features held at their recorded values.

Two caveats about the KernelSHAP background sample X_{bg}: it is drawn from the training partition here for reproducibility, but in deployment without training-data access the provider should supply an anonymised representative sample, since KernelSHAP needs a reference distribution for per-feature contributions [3]; and because KernelSHAP interpretations depend on that sample [38], the protocol relies on a provider-supplied or training-derived sample the deployer cannot fully verify from the prediction interface alone.

The `constraint_enforcement` signal is computed as:

$$\text{constraint\_enforcement} = \frac{\sum_{i\in\mathcal{L}} |\phi_i|}{\sum_{i\in\mathcal{L}} |\phi_i| + \sum_{i\in\mathcal{D}} |\phi_i|}. \quad (2)$$

Here, ϕ_i is the per-feature KernelSHAP attribution computed by the protocol's regression-aware KernelSHAP variant. The cohort-level aggregates reported in Sections VI-B–VI-C are computed from these per-decision records.

1) Validation envelope: The `envelope_validity` signal reports whether the case remains inside the feature-space envelope of the training partition. The signal uses a *k*-nearest-neighbour distance estimator on the 25-D post-imputation feature space, with $k = 5$. Features are standardised with the training-partition mean and standard deviation before distance computation, so each variable contributes comparably to the Euclidean distance.

The threshold $d_{\max}$ is the 90th percentile of the per-training-point 5-NN mean distance over the 490-eye training partition, a choice consistent with out-of-distribution detection practice [39]. The signal is:

$$\text{envelope\_validity}(x) = 1 - \max\left(0,\ \frac{d_{k\text{NN}}(x)}{d_{\max}} - 1\right), \quad (3)$$

clipped to $[0, 1]$: a value of 1.0 means the case remains within the calibrated training envelope, and lower values indicate greater distance beyond the threshold.

2) Perturbation parameters: The bounded-perturbation magnitude used by `decision_robustness` is set to $\sigma = 0.10$ of each continuous feature's training-partition standard deviation. Perturbations are applied only to continuous anatomical, biometric and refractive variables. One-hot categorical indicators and the candidate lens-size encoding are held fixed during bounded-perturbation and micro-noise tests.

This magnitude is of the same order as inter-operator OCT measurement variability, so a band-membership flip under it means the decision can change under deployment-level measurement noise.

Let s_j denote the training-partition standard deviation of continuous feature j. With $N = 100$ perturbations $\epsilon_n \sim \mathcal{N}(0, \text{diag}((\sigma s_j)^2))$ and band-membership indicator $b(\hat{y}) = \mathbf{1}[250 \le \hat{y} \le 750]$, the signal is:

$$\text{decision\_robustness}(x) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\Big[b(\hat{f}(x+\epsilon_n)) = b(\hat{f}(x))\Big]. \quad (4)$$

A value of 1.0 indicates that no perturbation flips the case's band membership; lower values report the share that do.

The micro-noise magnitude used by `record_integrity` is set to $\varepsilon = 0.01$ of each continuous feature's training-partition standard deviation. This tighter regime tests whether the attribution ranking stays stable under noise that should leave the clinical interpretation unchanged.

With $M = 50$ micro-perturbed copies $\delta_m \sim \mathcal{N}(0, \text{diag}((\varepsilon s_j)^2))$ and Spearman rank-order correlation ρ_S, the signal is the mean correlation between the original attribution vector and each perturbed replicate:

$$\text{record\_integrity}(x) = \frac{1}{M}\sum_{m=1}^{M} \rho_S(\phi(x),\ \phi(x+\delta_m)) \quad (5)$$

A value of 1.0 indicates a perfectly stable explanation under micro-noise; lower values flag an attribution whose ranking should be cited only with its noise envelope.

TABLE V. FEATURE TAXONOMY USED BY CONSTRAINT_ENFORCEMENT (25 FEATURES PARTITIONED INTO 14 LEGITIMATE + 11 DECISION-LINKED/REPORT-DEPENDENT VARIABLES).

#	Feature	Class	Domain group
1	lens_rise	L	anatomical OCT
2	anterior_depth	L	anatomical OCT
3	aod_nasal	L	anatomical OCT
4	aod_temporal	L	anatomical OCT
5	tisa_nasal	L	anatomical OCT
6	corneal_volume	L	anatomical OCT
7	pupil_x	L	anatomical OCT
8	pupil_y	L	anatomical OCT
9	pupil_diameter	L	anatomical OCT
10	limbus_y	L	anatomical OCT
11	limbus_diameter	L	anatomical OCT
12	age	L	biometric / age
13	k1	L	biometric
14	laterality	L	laterality
15	preop_sph	D	reported refraction
16	selected_sph	D	clinician's selection
17	selected_lens_size	D	clinician's selection (numeric)
18–21	lens_12.1, lens_12.6, lens_13.2, lens_13.7	D	one-hot encoding of the four candidate lens sizes
22–23	ICL, TICL	D	lens-family one-hot
24–25	Myopic, Toric Myopic	D	refractive-type one-hot

3) Operating point: The two sufficiency thresholds are set to $(\tau_{\text{env}}, \tau_{\text{margin}}) = (0.5, 0.3)$. This is the conservative operating point used throughout the study.

The routed-workload aggregate in Section VI-B is also reported at two alternative values of τ_{margin} (Table VI). Routed workload depends on the selected operating point. The architecture defines the aggregate. The deployer chooses the operating point. Among the three operating points reported in Table VI, $\tau_{\text{margin}} = 0.30$ was selected because it routes the fragile mid-band cases that $\tau_{\text{margin}} = 0.20$ leaves unrouted, while holding the human-review workload at 20% (11 of 55), below the 23.6% implied by $\tau_{\text{margin}} = 0.40$; this balances oversight coverage against the review capacity a single clinic can absorb.

B. Output of the Protocol on the Held-Out Cohort

The protocol is run on the 55-eye held-out cohort. Each prediction produces one JSON record with the five core signals, two per-decision oversight fields and case metadata.

The results are read at three levels: the individual decision record, the cohort aggregate used to estimate human-review workload, and the protocol-calibration level where signal distributions are checked on real data. The protocol output is produced at prediction time; if the record is persisted, the same record can later be audited.

Fig. 4 and Fig. 3 reproduce two records as they would appear at decision time, with retrospective outcome information shown separately. Case case4 shows counterfactual evidence supporting deviation from the evaluated lens option. Case case1 shows operational sufficiency despite a later large prediction error, separating decision-time evidence from model-quality assessment.

1) Per-decision evidence: The individual JSON record is the unit at which the deployer documents the output→action relation. Table VII reports five records from the held-out cohort. They were selected because each one shows a different behaviour that the protocol is designed to separate: clean, borderline, flip-prone, out-of-band and attribution-unstable.

Each column of Table VII is one JSON record produced at decision time and persisted for later audit. Two records are reproduced in full in Fig. 3 and Fig. 4. Retrospective measured vault is included only for the discussion of predictive accuracy in Section VII-A.

The five columns show different operational behaviours. The Clean column has every signal above threshold, although its later absolute prediction error is 308 µm. The Borderline and Flip-prone columns show fragile band membership under OCT-grade perturbation, with robustness values of 0.79 and 0.38. The Out-of-band column has stable attribution and band membership, but a negative score margin (-0.38). The Unstable column has the lowest record integrity in the cohort, so its SHAP ranking should be cited only with its noise envelope. A high value on one signal is not used to compensate for failure on another; the axes remain separate.

2) Per-cohort workload: The two cohort-level oversight aggregates in Table VIII convert the 55 decision records into an estimate of human-review workload. This is the cohort-level view relevant to Article 14 oversight. The aggregates are defined as

Definitions. For $N = 55$, the sustainability aggregate is the share of cases with $250 \leq \hat{y}_i \leq 750$. The teaming aggregate is the share of cases satisfying $\text{env}_i \geq \tau_{\text{env}}$ and $\text{margin}_i \geq \tau_{\text{margin}}$. Review required is the complement of the teaming aggregate.

The sustainability aggregate shows that 52 of 55 eyes receive a prediction inside the clinically plausible band. The human-AI-teaming aggregate shows that 44 eyes meet the joint sufficiency criterion and 11 do not. Under the operating point $(\tau_{\text{env}}, \tau_{\text{margin}}) = (0.5, 0.3)$, those 11 eyes represent 20% of the cohort flow. This is the workload routed to human review.

TABLE VI. COHORT ROUTING AGGREGATE UNDER THREE OPERATING POINTS OF THE SCORE-MARGIN SUFFICIENCY THRESHOLD ($\tau_{\text{ENV}} = 0.5$ FIXED).

τ_{margin}	Sufficient (AI-supported workflow)	Routed to review	Rationale
0.20	46 / 55 (0.836)	9 / 55 (0.164)	permissive: only clearly out-of-band cases routed
0.30	**44 / 55 (0.800)**	**11 / 55 (0.200)**	**operating point used in this study**
0.40	42 / 55 (0.764)	13 / 55 (0.236)	conservative: routes mid-band cases for review

Field	Value
CASE	`case1`
Lens option evaluated at decision time	13.2 mm
Predicted vault	498.5 µm
Retrospective measured vault (post-op; not decision-time evidence)	807.0 µm ($\Delta = 308$ µm)
PER-DECISION SIGNALS	
`score_margin`	0.99 (threshold 0.30) ✓ OK
`constraint_enforcement`	0.73 (legitimate share)
`envelope_validity`	1.00 (threshold 0.50) ✓ OK
`decision_robustness`	1.00 (flip rate 0.000)
`record_integrity`	0.86 (rank-order stability)
CONSERVATIVE OPERATION POINT	
`sustainability_impact`	`inside_clinical_band` ✓
`human_ai_teaming`	`ai_supported_workflow_sufficient` ✓
COUNTERFACTUAL BY LENS	
12.1 mm	316 µm margin 0.27 ✓ in band
12.6 mm	351 µm margin 0.40 ✓ in band
13.2 mm (evaluated option)	**499 µm margin 0.99 ✓ in band**
13.7 mm	684 µm margin 0.26 ✓ in band
HUMAN-OVERSIGHT EVIDENCE FUNCTIONS	
14(4)(b) capacities / limits	`score_margin`, `envelope_validity`
14(4)(c) automation bias	`decision_robustness`
14(4)(d) disregard / override / reverse	`constraint_enforcement`, counterfactual
14(4)(e) meaningful intervention	`counterfactual_by_lens`

`predict()` calls only – model internals not accessed

Fig. 3. Decision-time evidence record for case `case1`, with retrospective outcome shown separately. The protocol reports operational sufficiency on all five signals. The measured post-operative vault and absolute error are included only as retrospective outcome information.

Field	Value
CASE	`case4`
Lens option evaluated at decision time	13.7 mm
Predicted vault	845.4 µm
Retrospective measured vault (post-op; not decision-time evidence)	534.0 µm
PER-DECISION SIGNALS	
`score_margin`	-0.38 (threshold 0.30) × FAIL
`constraint_enforcement`	0.46 (legitimate share)
`envelope_validity`	1.00 (threshold 0.50) ✓ OK
`decision_robustness`	1.00 (flip rate 0.000)
`record_integrity`	0.94 (rank-order stability)
CONSERVATIVE OPERATION POINT	
`sustainability_impact`	`outside_clinical_band` ×
`human_ai_teaming`	`review_required` ×
COUNTERFACTUAL BY LENS	
12.1 mm	584 µm margin 0.66 ✓ in band
12.6 mm	531 µm margin 0.88 ✓ in band
13.2 mm	649 µm margin 0.40 ✓ in band
13.7 mm (evaluated option)	**845 µm margin -0.38 × out of band**
HUMAN-OVERSIGHT EVIDENCE FUNCTIONS	
14(4)(b) capacities / limits	`score_margin`, `envelope_validity`
14(4)(c) automation bias	`decision_robustness`
14(4)(d) disregard / override / reverse	`constraint_enforcement`, counterfactual
14(4)(e) meaningful intervention	`counterfactual_by_lens`

`predict()` calls only – model internals not accessed

Fig. 4. Decision-time evidence record for case `case4`, with retrospective outcome shown separately. Status indicators flag the signals failing the operating point of Section VI-A. The measured post-operative vault is included only as retrospective outcome information and is not part of the decision-time record.

TABLE VII. FIVE ILLUSTRATIVE OUTPUT RECORDS FROM THE HELD-OUT COHORT, ONE PER DECISION REGIME. TABLE IS TRANSPOSED FOR READABILITY: EACH COLUMN IS A SINGLE JSON RECORD PRODUCED BY THE PROTOCOL, AND EACH ROW REPORTS ONE FIELD OF THE RECORD. THE COLUMN IS THE AUDIT UNIT: IT IS AVAILABLE TO THE CLINICIAN AT DECISION TIME AND TO THE AUDITOR AFTERWARDS. RETROSPECTIVE MEASURED VAULT IS SHOWN ONLY AS OUTCOME INFORMATION. CASE IDENTIFIERS (`case1`–`case5`) ARE ANONYMISED AND DO NOT CORRESPOND TO CLINICAL-RECORDS DATABASE ENTRIES.

Attribute	Clean	Borderline	Flip-prone	Out-of-band	Unstable
case_id	case1	case2	case3	case4	case5
retrospective vault_real (µm)	807	310	690	534	321
vault_pred (µm)	498	251	774	845	446
score_margin	**0.99**	**0.00**	**-0.09**	**-0.38**	0.78
constraint_enforcement	0.73	0.47	0.79	0.46	0.78
envelope_validity	1.00	1.00	1.00	1.00	1.00
decision_robustness	1.00	0.79	**0.38**	1.00	1.00
record_integrity	0.86	0.93	0.87	0.94	**0.77**
lens option evaluated (mm)	13.2	12.6	13.2	13.7	13.2

TABLE VIII. COHORT-LEVEL OVERSIGHT AGGREGATES AND IMPLIED HUMAN-REVIEW WORKLOAD

Metric	Value	Reading
`sustainability_impact_cohort`	0.945	52 / 55 eyes inside [250, 750] µm
`human_ai_teaming_cohort`	0.800	44 / 55 eyes sufficient under AI-supported workflow
`review_required` (= 1 - teaming)	0.200	11 / 55 eyes routed to human review

Table IX maps each part of the protocol output to a human-oversight or post-market-surveillance function. The table is a traceability map between protocol outputs and the evidence that the deployer must organise and retain.

The clinician uses the record at decision time; persisted records can later be audited and aggregated for post-market surveillance. Both uses rely on `predict()` calls only.

The 11/55 routing result in Table VIII depends on the selected operating point. The deployer calibrates τ_{margin} against two costs: unnecessary routing and unrouted error.

Table VI reports the routed workload at three values of τ_{margin} while keeping τ_{env} fixed at 0.5. The routed workload increases monotonically from 9/55 at the permissive setting to 13/55 at the conservative setting.

3) Signal calibration over the cohort: Table X reports the empirical distribution of the core signals across the held-out cohort. This table gives protocol-level evidence. The clinician and the later auditor use the individual record, as shown in Table VII and Fig. 4 and Fig. 3.

The distribution addresses a methodological objection: the signals could collapse to constants on real data. Here they do not. `score_margin` ranges from -0.38 to 0.99, `constraint_enforcement` from 0.45 to 0.91, `record_integrity` from 0.77 to 0.95, and `decision_robustness` identifies six eyes with flip rates below unity.

4) Envelope-validity calibration: Under the percentile-based threshold, $d_{\max}$ is the 90th percentile of the per-training-point 5-NN mean distance ($d_{\max} = 5.06$). The held-out 5-NN mean distance ranges from 2.20 to 5.97. Fifty eyes remain saturated at `envelope_validity` $= 1.00$; five return values between 0.60 and 1.00. None falls below $\tau_{\text{env}} = 0.5$, so the 11 routed cases are still driven by the score-margin threshold in this cohort. The signal would become more decisive under genuine deployment shift, for example after changes in OCT device, surgeon protocol or patient population.

The cohort mean of `constraint_enforcement` is 0.676. The aggregate attribution share of the legitimate set is 0.681, agreeing to two decimals. Six of the ten highest-ranked features are legitimate, and the largest contributor is the anatomical feature `lens_rise` (Fig. 5). This independent agreement supports the claim that `constraint_enforcement` measures the attribution split it was defined to measure.

C. Counterfactual Over the Discrete Decision Space

Bounded perturbation tests local fragility. It does not compare the four surgical alternatives. The sizing decision is discrete: the surgeon chooses one lens size from four available options. The protocol can therefore query the model under each candidate lens size while holding all other features fixed.

For each held-out eye, the model is queried four times, once per available size. For each query, the protocol recomputes the three sweep-relevant signals: `score_margin`, `envelope_validity` and `decision_robustness`. The result is a 4×55 matrix. Each cell records the predicted vault for one eye under one candidate lens size, together with the three boundary-relevant signals recomputed for that hypothetical decision while all anatomical and biometric inputs remain fixed.

The body of the study reports two slices of the complete 220-row long-format matrix: one decision-level slice and one cohort-level slice.

The decision-level slice gives the operator evidence for disregarding, overriding or intervening under Article 14(4)(d–e). Fig. 4 shows this for case `case4`. The cohort-level slice aggregates across all 55 eyes by lens size and supports post-market surveillance under MDR Article 88. Table XI reports this aggregate.

1) Per-decision counterfactual evidence: Consider eye `case4` at the moment when the surgeon must commit to a lens

TABLE IX. MAPPING OF PROTOCOL OUTPUTS TO HUMAN-OVERSIGHT AND POST-MARKET-SURVEILLANCE FUNCTIONS

Protocol output	Evidence function supported
Per-decision JSON record (Figs. 3, 4)	`Art 14(4)(b)` capacities and limits of the system
`decision_robustness` flag	`Art 14(4)(c)` awareness of automation bias
`constraint_enforcement` taxonomy (legitimate / decision-linked)	`Art 14(4)(d)` basis for disregard, override or reversal
Counterfactual-by-lens matrix (Section VI-C)	`Art 14(4)(e)` basis for meaningful intervention
Cohort routing aggregate (Table VIII)	`Art 14(3)` oversight measures proportionate to risk
Cohort signal distribution (Table X)	`MDR Art 83 & 88` post-market surveillance and trend reporting

TABLE X. DISTRIBUTION OF THE FOUR INFORMATIVE CORE SIGNALS ACROSS THE 55 RECORDS (SIGNAL-CALIBRATION EVIDENCE; NOT PART OF THE PER-DECISION AUDIT UNIT). ENVELOPE_VALIDITY IS CONSTANT AT 1.00. SEE CALIBRATION PARAGRAPH ABOVE.

Statistic	score_margin	constraint_enf	decision_rob	record_int
min	-0.38	0.45	0.38	0.77
Q1	0.42	0.62	1.00	0.81
median	0.65	0.68	1.00	0.85
Q3	0.84	0.74	1.00	0.87
max	0.99	0.91	1.00	0.95
mean $\pm$ std	0.59 $\pm$ 0.33	0.68 $\pm$ 0.11	0.97 $\pm$ 0.10	0.85 $\pm$ 0.05

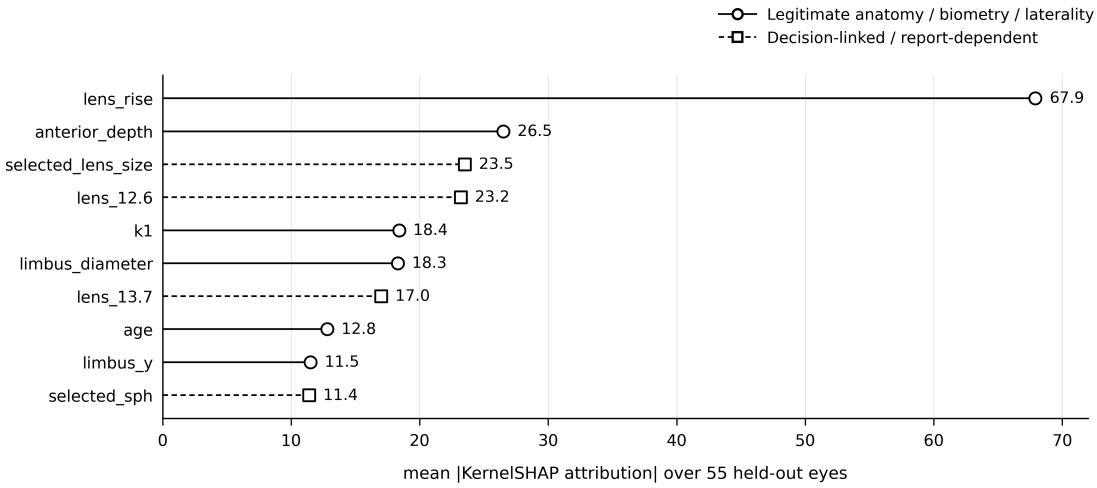

Fig. 5. Top-10 features by mean |KernelSHAP attribution| over the 55-eye held-out cohort. The legitimate/decision-linked split used by `constraint_enforcement` is visible here at the level of individual features.

TABLE XI. COUNTERFACTUAL-BY-LENS-SIZE COHORT AGGREGATES

Lens (mm)	Mean vault (µm)	% in band	Mean margin	Mean robust.
12.1	464.7	98.2	0.66	0.99
12.6	461.6	96.4	0.67	0.98
13.2	542.3	98.2	0.63	0.98
13.7	**684.8**	**63.6**	**0.20**	**0.94**

size. The protocol returns four predictions, one per available size, with their score margin, envelope validity and robustness.

The clinician sees that the 13.7 mm option gives a predicted vault of 845 µm and a score margin of -0.38. The same record also shows that the 12.1 mm and 12.6 mm options produce predictions inside the $[250, 750]$ µm band: 584 µm and 531 µm, with score margins of 0.66 and 0.88.

The decision record and the later audit record are identical. Fig. 4 reproduces the full record. The clinician decides with that evidence available. The auditor later reviews the same evidence.

The counterfactual matrix gives the deployer structured evidence about the available alternatives. It shows the model's prediction under each candidate lens size. This makes intervention under Article 14(4)(e) evidence-based rather than dependent on a single recommendation.

2) Per-cohort post-market surveillance evidence: Aggregating the 4×55 matrix over the held-out cohort reveals a systematic pattern in the deployed regressor (Table XI).

Three of the four lens sizes return predictions inside the clinical band for 96–98% of eyes. The largest size, 13.7 mm, returns predictions inside the band for only 63.6% of eyes. It also has the highest mean predicted vault, 684.8 µm, and the lowest mean robustness in the four-way comparison.

This result should be read as operational monitoring evi-

dence. The deployer does not redesign or retrain the regressor. The value of the matrix is that it identifies a pattern that can be reported through post-market surveillance if it appears in deployment.

The matrix produces this result from `predict()` calls only. It is generated alongside the per-decision telemetry reported in Section VI-B.

Vault standard deviations across the cohort are $\sigma(12.1) = 97.1$ µm, $\sigma(12.6) = 98.1$ µm, $\sigma(13.2) = 109.2$ µm and $\sigma(13.7) = 126.8$ µm.

The counterfactual matrix supports two evidence functions. At decision time, it supports the operator's intervention under Article 14(4)(e). At cohort level, it supports post-market surveillance under MDR Article 88. Both readings are produced from `predict()` calls only, during the time required to predict the four candidate lens sizes, and are persisted with the per-decision telemetry of Section VI-B.

Taken together, the five-signal cohort profile of Section VI-B and the by-lens-size counterfactual matrix of Section VI-C specify the evidence needed by the deployer. The protocol produces per-decision evidence for human oversight under EU AI Act Article 14(3)–(4), and cohort-level signal distributions for post-market surveillance under MDR Articles 83 and 88.

The clinician uses the evidence record at decision time. The auditor and the post-market surveillance process use the same persisted record afterwards. The deployer can produce both readings without access to model internals and without cooperation from the model provider.

D. Reproducibility

The clinical dataset described in Section IV-B cannot be published because it contains personal health information governed by the Medical Devices Regulation and by contractual restrictions across the three contributing centres.

The instrumentation protocol is reproducible at method level. The five core signals are defined operationally in Section VI-A through expressions over `predict()` calls and a fixed background sample. The calibration parameters – clinical band $[250, 750]$ µm, $\sigma = 0.10$, $\varepsilon = 0.01$, $k = 5$, $N = 100$, $M = 50$ and $X_{\text{bg}} = 50$ – are listed in Table IV. The train/held-out split is reproducible with `test_size = 0.10` and `seed = 42`, as stated in Section IV-D.

The deployed model is an Extra Trees regressor over 25 input features partitioned into 14 legitimate and 11 proxy features (Table V); the two sufficiency thresholds are $(\tau_{\text{env}}, \tau_{\text{margin}}) = (0.5, 0.3)$. The protocol is computationally lightweight: feature attribution uses regression-aware KernelSHAP with a 50-sample background, and the perturbation and counterfactual tests require only repeated `predict()` calls, so the full per-decision evidence record is produced without specialised hardware. Experiments were run on Google Colab (CPU runtime) using scikit-learn, NumPy, pandas and SciPy.

A deployer with access to the prediction interface of any regression model on a comparable feature space can rerun the protocol and reproduce the structure of the per-decision evidence reported here. The cohort-level numerical results in Section VI-B and Section VI-C are specific to the contributing cohort; the protocol structure is reusable.

VII. Discussion

A. Operational Sufficiency and Predictive Accuracy are Separate Readings

The main result is that predictive accuracy and operational sufficiency do not collapse into the same judgement.

Case `case3` (Table VII, Flip-prone column) shows the direction in which this separation is most useful. The predicted vault is 774 µm and the absolute prediction error against the measured post-operative vault is 84 µm, which is small by the standards of this cohort. The prediction lies just above the upper edge of the clinical band (`score_margin` = -0.09), but the boundary reading is sharper: `decision_robustness` is 0.38, which means that 62% of bounded perturbations flip the band membership of the prediction. The protocol therefore routes this case to human review even though the regression prediction is accurate by the model's own standards. Good predictive accuracy does not constitute operational sufficiency at the decision boundary.

The reverse direction appears in case `case1` (Fig. 3, Clean column). Every per-decision signal is above its operating threshold, and both per-decision oversight fields report sufficiency. The later measured vault is 807 µm, producing a 308 µm error. The protocol did not fail to predict the outcome; it was not designed to estimate post-operative error after the fact. It was designed to document whether the prediction, at decision time, had the structural properties required for AI-supported workflow.

This separation is methodologically important. Model validation asks whether the regressor predicts vault well enough at population level. Decision-boundary instrumentation asks whether a specific output is supported by enough evidence to be used as a reason for action. Both questions are necessary. They are not interchangeable.

B. Calibration of `Envelope_validity` Under the Percentile Threshold

Under the percentile-based threshold ($d_{\max}$ = 90th percentile of the per-training-point 5-NN mean distance), `envelope_validity` is informative on the present cohort. The threshold value is $d_{\max} = 5.06$ in the standardised feature space. The held-out 5-NN mean distance ranges from 2.20 to 5.97. Fifty of the 55 held-out eyes remain saturated at `envelope_validity` = 1.00; five eyes return values between 0.60 and 1.00. None falls below the sufficiency threshold $\tau_{\text{env}} = 0.5$. The joint criterion of `human_ai_teaming` therefore routes the same 11 cases as the score-margin criterion alone, and the cohort-level workload is unchanged by the percentile calibration.

The informative regime for `envelope_validity` is genuine deployment shift. Changes in OCT device, surgeon protocol or patient population would move more cases beyond the calibrated envelope and yield values below 0.50. Those values would route additional cases through `human_ai_teaming`. A controlled drift experiment on the

input feature space would be the next step for testing that behaviour, but it is outside the scope of this study.

C. One Evidence Record Supports Three Regulatory Evidence Functions

The protocol produces one persisted evidence record that can serve three regulatory evidence functions: decision-time human oversight, later audit and post-market surveillance.

At decision time, the per-decision JSON record (Fig. 4 and Fig. 3) makes the relevant signal values available to the operator. `score_margin` and `envelope_validity` provide evidence for the Article 14(4)(b) function of understanding system capacities and limits in use. `decision_robustness` provides evidence for Article 14(4)(c) by making output fragility visible and reducing blind reliance on the prediction.

The legitimate/decision-linked taxonomy used by `constraint_enforcement` provides evidence for Article 14(4)(d). It shows whether the local prediction is mainly supported by anatomical and biometric variables, or whether it is mainly using variables linked to the clinician's prospective decision. This gives the operator evidence for disregarding, overriding or reversing the output.

The counterfactual-by-lens matrix (Fig. 6, Table XI) provides evidence for Article 14(4)(e). It shows what the model would predict under each available lens size while anatomy and biometry are held fixed. That comparison makes intervention meaningful because the operator can inspect the available alternatives, not only the evaluated option.

After the decision, the same persisted record becomes an audit unit. Across cases, the same records become post-market surveillance evidence under MDR Articles 83 and 88. The 13.7 mm pattern in Table XI is a concrete example: it is not a single adverse event, but a cohort-level signal that could justify closer monitoring if reproduced in deployment.

D. Boundary Conditions for Deployer-Side Use

The protocol should be read under four boundary conditions. First, decision-boundary instrumentation presupposes a model that has already passed model-level validation. The signals reported here evaluate whether a specific output has the observable properties required for use in a workflow. Population-level validation, clinical evaluation and model selection remain prerequisites. In this case, that prerequisite is represented by the regression comparison and cross-validated MAE reported in Table II.

Second, several elements of the protocol require explicit configuration before deployment. The legitimate/decision-linked feature taxonomy defines how `constraint_enforcement` is interpreted. The KernelSHAP background sample defines the reference distribution for local attribution. The operating thresholds define the review workload accepted by the organisation. They are explicit elements of the deployer-side evidence package and should be documented before operational use.

Third, the empirical results are case-study results from a 55-eye held-out cohort within one multi-site ophthalmology network. The workload estimate, the 13.7 mm counterfactual pattern and the observed signal distributions should therefore be read as evidence of how the protocol behaves in this deployment setting. Transfer to other ICL datasets, devices, surgeons or patient populations would require recalibration and prospective monitoring.

Fourth, the protocol specifies five core signals derived from four documented failure modes; it does not claim that these five are necessary and sufficient. Failure modes that the AI-safety literature has identified but that the present protocol does not instrument – for instance, calibration drift, label distribution shift or systematic disagreement with an expert reference – would require additional signals not covered here. The heuristic parameter values reported in Box 2 reflect operational choices grounded in the cited literature; a systematic sensitivity study across all heuristic parameters, beyond the τ_{margin} axis reported in Table VI, is left for future work.

VIII. Conclusion

This study treats post-hoc XAI as deployer-side evidence rather than as explanation alone. Much of the XAI literature characterises the model, which is mainly the developer's question. Regulated deployment creates a different question for the deployer: when a prediction supports an action, what evidence is attached to that prediction, and can another party audit it later?

The study specifies a decision-boundary instrumentation architecture that answers that question by reorganising existing XAI and validation methods – KernelSHAP, k-nearest-neighbour distance, bounded perturbation and rank-order stability – into a per-decision telemetry record. The record is computed from `predict()` calls only and is designed to support decision-time human oversight, later audit and post-market surveillance.

The protocol was applied to a 55-eye held-out cohort of an ICL vault prediction regressor. At the conservative operating point, 11 cases (20%) were routed to human review and the rest were classified as operationally sufficient. Each case produced a JSON record that maps to deployer obligations under EU AI Act Article 14 and to MDR post-market surveillance reporting under Articles 83 and 88. The counterfactual-by-lens matrix showed how the same prediction interface can support meaningful intervention at decision time and cohort-level monitoring afterwards.

These results hold under clear limitations. The empirical findings come from a single 55-eye held-out cohort in one ophthalmology network, so the workload estimate and counterfactual pattern are cohort-specific and require multi-site, prospective validation before generalisation. Some signals saturate in this cohort: `envelope_validity` equals 1.00 for 50 of the 55 cases (Section VII-B), so its discriminative contribution should be re-assessed under more varied deployment conditions. The protocol is also not yet compared against alternative evidence-generation approaches such as confidence-threshold routing or clinical audit checklists. Consolidating the future work noted above, the main directions are: multi-site and prospective validation, a systematic sensitivity study beyond the τ_{margin} axis, comparison against baseline evidence-generation approaches, and additional signals for failure modes

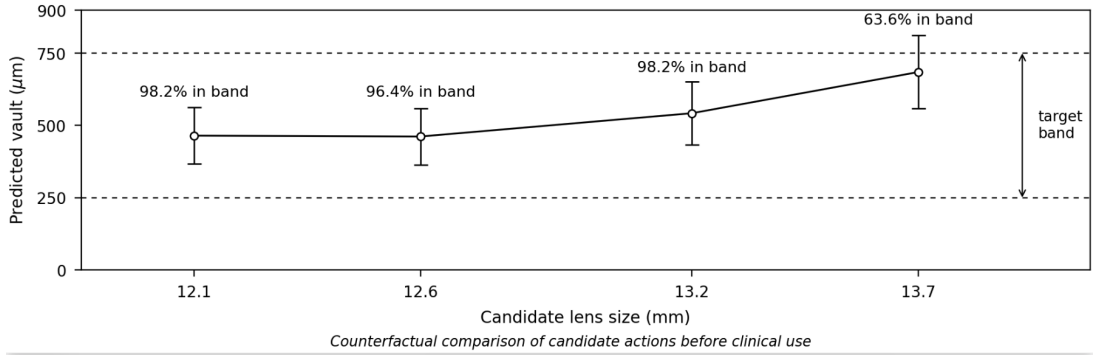

Fig. 6. Cohort-level visualisation of the by-lens-size counterfactual matrix of Table XI. Each bar is the mean predicted vault under one candidate lens size, computed across all 55 held-out eyes; the vertical lines mark ± one standard deviation. The horizontal shaded band marks the clinical sufficiency interval [250, 750] µm. The three smaller lens sizes place their mean predictions near the centre of the band; the 13.7 mm option pushes the mean towards the upper edge, with the largest standard deviation in the four-way comparison and the lowest in-band fraction (63.6%). The pattern is generated from `predict()` calls only and is the visual companion to the post-market surveillance reading of Section VI-C.

not instrumented here (for example calibration drift, label-distribution shift, or systematic disagreement with an expert reference).

The main methodological claim is limited. Under limited-inspection conditions, XAI can function as instrumentation without requiring access to model internals: it characterises observable behaviour, tests boundary stability and creates technical trust evidence at the moment when an output may become a reason for action.

ACKNOWLEDGMENT

This research has been supported by the Ministerio de Ciencia, Innovación y Universidades, Gobierno de España, under research grant PID2024-158490OB-C33 (ECEAAS).

DECLARATION ON GENERATIVE AI

The authors used Claude (Anthropic, `claude-opus-4`) to assist in drafting prose from author-specified outlines and in generating Python code and replication scripts. The development of the scientific ideas, theoretical framework, research design, results and conclusions was carried out by the authors. The authors critically reviewed and edited all AI-assisted text and code, and assume full responsibility for the accuracy, originality and integrity of the manuscript. No generative AI tool is listed as an author.

REFERENCES

[1] European Parliament and Council of the European Union, "Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence," Official Journal of the European Union, 2024. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[2] Co of the European Union and European Parli, "Regulation (eu) 2017/745 of the european parliament and of the council of 5 april 2017 on medical devices," Official Journal of the European Union, 2017. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2017/745/oj

[3] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765–4774.

[4] M. Ouchi *et al.*, "Vault changes in eyes with a vertically implanted implantable collamer lens," *Scientific Reports*, 2024, clinical reference using the commonly cited ICL vault safety interval of approximately 250–750 micrometres.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[6] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.

[7] A. B. Arrieta, N. Diaz-Rodriguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[8] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems 31*, 2018, pp. 9505–9515.

[9] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 180–186.

[10] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019.

[11] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 648–657.

[12] Q. V. Liao and K. R. Varshney, "Human-centered explainable AI (XAI): From algorithms to user experiences," arXiv:2110.10790, 2022.

[13] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.

[14] Artificial Intelligence Board and Medical Device Coordination Group, "Aib 2025-1 / mdcg 2025-6: Interplay between the medical devices regulation (mdr) & in vitro diagnostic medical devices regulation (ivdr) and the artificial intelligence act (aia)," European Commission guidance document, Jun. 2025. [Online]. Available: https://health.ec.europa.eu/latest-updates/mdcg-2025-6-faq-interplay-between-medical-devices-regulation-vitro-diagnostic-medical-devices-2025-06-19_en

[15] J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big Data & Society*, vol. 3, no. 1, 2016.

[16] F. Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, 2015.

[17] J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu, "Accountable algorithms," *University of Pennsylvania Law Review*, vol. 165, no. 3, pp. 633–705, 2017.

[18] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," NIST AI 100-1, 2023.

[19] International Organization for Standardization, "ISO/IEC 42001:2023: Information technology — artificial intelligence — management system," International standard, 2023. [Online]. Available: https://www.iso.org/standard/81230.html

[20] B. Babic, S. Gerke, T. Evgeniou, and I. G. Cohen, "Beware explanations from AI in health care," *Science*, vol. 373, no. 6552, pp. 284–286, 2021.

[21] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," in *Proceedings of the 4th Machine Learning for Healthcare Conference*, ser. Proceedings of Machine Learning Research, vol. 106, 2019, pp. 359–380.

[22] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 310, 2020.

[23] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 220–229.

[24] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. Natesan Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney, "FactSheets: Increasing trust in AI services through supplier's declarations of conformity," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 6:1–6:13, 2019.

[25] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[26] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 33–44.

[27] J. Singh, J. Cobbe, and C. Norval, "Decision provenance: Harnessing data flow for accountable systems," *IEEE Access*, vol. 7, pp. 6562–6574, 2019.

[28] D. Urrutia-Onate, A. Perallos, and E. Onieva, "XAI as decision-boundary instrumentation for AI governance," Zenodo preprint, 2026.

[29] N. G. Leveson, *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press, 2011.

[30] J. Rasmussen, "Risk management in a dynamic society: A modelling problem," *Safety Science*, vol. 27, no. 2–3, pp. 183–213, 1997.

[31] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors*, vol. 52, no. 3, pp. 381–410, 2010.

[32] Y. Shen, L. Wang, W. Jian *et al.*, "Big-data and artificial-intelligence-assisted vault prediction and EVO-ICL size selection for myopia correction," *British Journal of Ophthalmology*, vol. 107, no. 2, pp. 201–206, 2023.

[33] A. Russo, O. Filini, G. Savini, G. Festa, F. Morescalchi, A. Boldini, and F. Semeraro, "Predictability of the vault after implantable collamer lens implantation using OCT and artificial intelligence in white patient eyes," *Journal of Cataract & Refractive Surgery*, vol. 49, no. 7, pp. 724–731, 2023.

[34] P. Zeboulon *et al.*, "Validation of a new implantable collamer lens sizing algorithm based on SS-OCT images," *Journal of Cataract & Refractive Surgery*, vol. 52, no. 1, pp. 61–66, 2026.

[35] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[36] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[38] H. Yuan, M. Liu, L. Kang, C. Miao, and Y. Wu, "An empirical study of the effect of background data size on the stability of SHapley Additive exPlanations (SHAP) for deep learning models," 2022.

[39] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, ser. PMLR, vol. 162, 2022, pp. 20 827–20 840. [Online]. Available: https://proceedings.mlr.press/v162/sun22d.html