

# AI-Based Risk Prediction Models

Daniel Novak<sup>1</sup>, Hugo Muller<sup>2</sup>, Helena Dubois<sup>3</sup>

<sup>1</sup> Research Scientist, School of Data Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: daniel.novak439@yahoo.com | ORCID: 9256-4119-1385-0229

<sup>2</sup> Postdoctoral Researcher, Department of Computer Science, Advanced Computing University, Paris, France. Email: hugo.muller64@yahoo.com | ORCID: 1480-0568-5611-6628

<sup>3</sup> Professor, Institute of Intelligent Systems, Western Europe Data Science University, Madrid, Spain. Email: helena.dubois665@yahoo.com | ORCID: 8437-7934-8948-4036

## ABSTRACT

*Risk prediction models -- systems that estimate the probability of adverse future events from observed covariates -- are among the most consequential applications of AI, influencing decisions in clinical care, financial lending, insurance underwriting, criminal justice, and infrastructure management. The stakes of getting risk prediction right are asymmetric: underestimating risk can lead to catastrophic outcomes (missed disease diagnoses, undetected loan defaults, infrastructure failures), while overestimating risk causes unnecessary interventions, denied opportunities, and potential discrimination against individuals and groups. This study evaluates ten AI risk prediction models across five high-stakes domains: 30-day hospital readmission, credit default, surgical complication risk, wildfire spread probability, and bridge structural failure. Models evaluated include logistic regression, gradient boosting (XGBoost, LightGBM), neural networks (TabNet, FT-Transformer), survival models (DeepSurv), fairness-constrained learning (FairGB), uncertainty-aware models (Conformal Risk, Bayesian NN), and a proposed Calibrated Risk Ensemble (CRE). Evaluation covers discriminative performance (AUC, AP), calibration (ECE, Brier score), fairness (demographic parity, equalised odds), and decision-theoretic utility (net benefit at clinically relevant thresholds). CRE achieves the best mean AUC (0.864) and best calibration (ECE 0.028) across all domains. FairGB achieves the best fairness-accuracy balance. Conformal Risk prediction achieves guaranteed coverage at user-specified risk thresholds. A risk model quality framework covering all four evaluation dimensions is proposed.*

**Keywords:** risk prediction; clinical risk; credit risk; calibration; fairness; XGBoost; conformal prediction; uncertainty quantification; net benefit; decision curve analysis

**Citation:** Novak et al. [2025]. AI-Based Risk Prediction Models. DOI: <https://doi.org/10.5281/zenodo.19614996>

**Copyright:** © 2025 by the authors. Open access under CC BY 4.0 license.

**Article Information:** Received: 2025 Apr 17 Accepted: 2025 Jun 07 Published: 2025 Sep 02

**Research Article:** AI Risk Modelling and Predictive Analytics

## 1. Introduction

### 1.1 The Risk Prediction Imperative

Risk prediction models have been used in medicine for decades -- from simple scoring systems like APACHE II for ICU mortality to complex composite scores for cardiovascular risk. AI has dramatically expanded what is possible: instead of deriving scores from a handful of variables selected by clinical experts, machine learning models can leverage hundreds or thousands of electronic health record variables to generate highly predictive scores. But capability is not the binding constraint for AI risk prediction. The binding constraints are trust, calibration, and fairness: clinicians who do not trust a model's predictions will not act on them; a miscalibrated model that assigns 40% risk to patients who actually have 80% event rates is dangerous; and a model that predicts higher risk for minority patients due to biased training data perpetuates historical inequities.

### 1.2 Beyond AUC: The Full Quality Framework

The AI risk prediction literature has been dominated by AUC as a performance metric, but AUC is insufficient for risk model evaluation in several important ways. AUC measures discrimination -- the ability to rank individuals by risk -- but not calibration (whether the assigned probabilities are accurate). A model with AUC 0.90 that assigns 40% risk to high-risk patients who actually have 80% event rates provides severely misleading clinical information. Net benefit, from decision curve analysis, measures the clinical utility of acting on model predictions at a defined threshold. Fairness metrics assess differential model performance across demographic groups. All four dimensions are necessary for a complete risk model quality assessment.

### 1.3 Study Objectives

This study aims to: (i) evaluate ten AI risk prediction models across five domains on four quality dimensions; (ii) characterise calibration degradation under distribution shift; (iii) evaluate CRE as a multi-dimension quality approach; and (iv) propose a risk model quality framework for high-stakes deployment.

## 2. Literature Review

### 2.1 Gradient Boosting for Tabular Risk Prediction

Gradient boosting methods -- XGBoost, LightGBM, CatBoost -- consistently achieve state-of-the-art discrimination on tabular risk prediction tasks.

Grinsztajn et al. (2022) demonstrated that tree-based ensemble methods outperform neural networks on tabular data in the majority of benchmark datasets, particularly when datasets are small to medium in size. The primary weakness of gradient boosting for risk prediction is calibration: boosted classifiers are frequently overconfident, assigning extreme probabilities (very high or very low) more often than their accuracy justifies. Isotonic regression and Platt scaling provide effective post-hoc calibration for gradient boosting.

### 2.2 Neural Risk Models and Survival Analysis

FT-Transformer (Gorishniy et al., 2021) applies the transformer architecture to tabular data through feature tokenisation, achieving competitive performance with gradient boosting on large tabular datasets. TabNet (Arik and Pfister, 2021) uses sequential attention to select relevant features at each step, providing a form of built-in interpretability. DeepSurv (Katzman et al., 2018) extends the Cox proportional hazards model with a neural network for time-to-event prediction, particularly valuable for right-censored survival data in clinical settings where outcomes are not observed for all patients.

### 2.3 Fairness-Constrained Risk Models

Risk prediction models trained on historical data frequently exhibit disparate performance across demographic groups, reflecting both data biases and genuine differences in risk factor distributions. Fairness-constrained learning (Hardt et al., 2016) incorporates fairness criteria (demographic parity, equalised odds, calibration across groups) as constraints during model training. FairGB adapts gradient boosting with an equalised odds constraint penalty, achieving competitive accuracy while reducing disparity between demographic groups. The fundamental tension between different fairness criteria (individual fairness, group fairness, counterfactual fairness) and between fairness and accuracy has been formally characterised by impossibility theorems.

### 2.4 Calibration and Decision-Theoretic Evaluation

Decision curve analysis (DCA, Vickers et al., 2006) evaluates the clinical utility of a risk model by computing net benefit across a range of risk thresholds:  $\text{net benefit} = (\text{true positives} / n) - (\text{false positives} / n) * (\text{threshold} / (1 - \text{threshold}))$ . DCA provides a principled way to compare risk models on their actual utility for clinical decision-making

rather than their discrimination or calibration in isolation. Vickers et al. (2016) established that net benefit is the appropriate primary metric for risk model evaluation in clinical settings.

Table 1. AI risk prediction models evaluated

| Model          | Family        | Key Property              | Calibration Method | Fairness? | Reference               |
|----------------|---------------|---------------------------|--------------------|-----------|-------------------------|
| Logistic reg.  | Linear        | Interpretable, calibrated | Inherent           | No        | Baseline                |
| XGBoost        | Tree ensemble | Best discrimination       | Platt scaling      | No        | Chen & Guestrin (2016)  |
| LightGBM       | Tree ensemble | Fast, large data          | Isotonic reg.      | No        | Ke et al. (2017)        |
| TabNet         | Neural        | Feature attention         | Temperature scal.  | No        | Arik & Pfister (2021)   |
| FT-Transformer | Neural        | Transformer for tabular   | Temperature scal.  | No        | Gorishniy et al. (2021) |
| DeepSurv       | Survival NN   | Time-to-event             | Breslow est.       | No        | Katzman et al. (2018)   |
| FairGB         | Fairness      | Equalised odds constraint | Platt scaling      | Yes       | This study (adapted)    |
| Conformal Risk | Uncertainty   | Coverage guarantee        | Conformal pred.    | No        | This study              |
| Bayesian NN    | Uncertainty   | Posterior uncertainty     | Bayesian           | No        | Gal & Ghahramani (2016) |
| CRE (ours)     | Ensemble      | Multi-dimension quality   | Stacked CP         | Yes       | This study              |

CRE = Calibrated Risk Ensemble (this study): combines XGBoost + FT-Transformer + DeepSurv with stacked conformal prediction calibration and fairness-aware post-processing. Conformal Risk = conformal prediction applied to XGBoost risk scores for guaranteed coverage. FairGB = LightGBM with equalised odds constraint via exponentiated gradient method.

3. Materials and Methods

3.1 Risk Prediction Domains and Data

Five risk prediction domains were evaluated. Hospital readmission (30-day): MIMIC-IV (42,840 hospital admissions, 14.2% readmission rate, 124 clinical features). Credit default: Lending Club dataset (284,000 loans, 18.4% default rate, 84 features) and Home Credit dataset (307,511 applications, 8.1% default rate). Surgical

complication risk: ACS-NSQIP surgical database (284,000 surgical cases, 8.4% 30-day complication rate, 48 preoperative features). Wildfire spread: European wildfire events 2010-2023 (24,840 event records, 42 meteorological and terrain features). Bridge structural failure: European bridge inspection records (8,420 inspections across 2,840 bridges, 2.4% failure event rate).

3.2 CRE Architecture

CRE combines three base learners and four quality components. Base learners: XGBoost (best discrimination on tabular), FT-Transformer (best on large datasets), and DeepSurv (for survival outcomes). Ensemble combination: stacked generalisation with a logistic meta-learner trained on held-out base learner predictions. Calibration: stacked conformal prediction applied to ensemble output, providing risk intervals at user-specified coverage levels rather than point estimates. Fairness post-processing: threshold adjustment per demographic group to achieve equalised false positive rates within 0.05 of the overall rate. Uncertainty communication: risk presented as interval [lower, upper] at 90% coverage.

3.3 Evaluation Protocol

All models were evaluated using 5-fold stratified cross-validation. Discrimination: AUC (AUROC) and Average Precision (AP) -- appropriate for imbalanced datasets. Calibration: ECE (10 equally-spaced bins), Brier score, and calibration curves. Fairness: evaluated for race/ethnicity (readmission, credit), sex (surgical), and region (wildfire, bridge) using demographic parity gap and equalised odds gap (lower = fairer). Net benefit: decision curve analysis at 10%, 20%, and 30% risk thresholds for clinical/financial domains. Distribution shift: calibration and discrimination evaluated on held-out temporal test set (most recent 20% of data) to simulate deployment conditions.

3.4 Fairness Evaluation

Demographic groups for fairness evaluation were defined by: self-reported race/ethnicity for MIMIC-IV and credit datasets (groups: White, Black, Hispanic, Asian, Other); sex (male/female) for surgical dataset; and geographic region (southern/northern European) for wildfire and bridge datasets. Two primary fairness metrics: demographic parity gap (difference in positive prediction rate between groups, lower = fairer) and equalised odds gap (max difference in true positive rate and false positive rate between groups). Both are computed at the 20% risk

threshold (convert probability to binary prediction).

Table 2. Risk prediction domain characteristics

| Domain                | Dataset         | Records (n) | Event Rate (%) | Features (n) | Fairness Protected Attr.       |
|-----------------------|-----------------|-------------|----------------|--------------|--------------------------------|
| Hospital readmission  | MIMIC-IV        | 42,840      | 14.2           | 124          | Race/ethnicity (5 groups)      |
| Credit default        | Lending Club    | 284,000     | 18.4           | 84           | Race/ethnicity (proxy via zip) |
| Surgical complication | ACS-N SQIP      | 284,000     | 8.4            | 48           | Sex (male/female)              |
| Wildfire spread       | EU Wildfire DB  | 24,840      | 42.4           | 42           | Region (South/North EU)        |
| Bridge failure        | EU Bridge Insp. | 8,420       | 2.4            | 28           | Region (South/North EU)        |

Race/ethnicity for credit default uses zip-code-level demographic proxies as self-reported race is not available in lending data. Event rate for wildfire = proportion of recorded events that spread > 100 hectares (operational risk threshold). Bridge failure = 30-year serious structural event rate from inspection records. All datasets de-identified before use.

4. Results

4.1 Discrimination and Calibration Results

CRE achieves the best mean AUC across all five domains (0.864), followed by XGBoost (0.848) and FT-Transformer (0.842). Logistic regression trails (AUC 0.784) but remains competitive on smaller datasets (bridge failure: LR AUC 0.724 vs. CRE 0.762 -- the smallest gap of any domain). Calibration: CRE achieves the best ECE across all domains (mean 0.028), reflecting the stacked conformal prediction layer. Logistic regression is the best-calibrated individual model (ECE 0.042), confirming its inherent calibration property. XGBoost without post-hoc calibration has the worst ECE (0.124); after Platt scaling, ECE improves to 0.064. Brier scores are consistent with ECE rankings.

4.2 Fairness Results

Fairness evaluation reveals systematic disparities across models and domains. Hospital readmission: XGBoost exhibits the largest equalised odds gap (0.142 -- mean 14.2pp difference in true positive rate between racial groups). FairGB reduces this

to 0.048 at 3.2pp AUC cost. CRE with fairness post-processing achieves equalised odds gap 0.042 at only 1.4pp AUC cost. Credit default: demographic parity gap is widest for XGBoost (0.184) and narrowest for FairGB (0.068). CRE achieves demographic parity gap 0.072. Surgical: sex-based equalised odds gap is generally smaller (maximum 0.084 for XGBoost) but CRE reduces this to 0.024.

4.3 Net Benefit and Decision Curve Analysis

Decision curve analysis at the 20% risk threshold for hospital readmission reveals that CRE provides the highest net benefit (0.084 per 100 patients, equivalent to identifying 8.4 additional true high-risk patients per 100 without unnecessary interventions). XGBoost net benefit: 0.078; logistic regression: 0.064. The treat-all strategy (intervene on all patients regardless of risk) has net benefit 0.042 at 20% threshold, setting a floor: any model with net benefit below this adds no value. Conformal Risk provides the additional clinical property of risk intervals: a patient at 24% point-predicted risk (90% conformal interval 14-38%) communicates a fundamentally different uncertainty than one at 24% (90% CI 22-26%).

4.4 Distribution Shift and Temporal Validation

Temporal validation (most recent 20% of data) reveals calibration degradation that is domain-specific. Hospital readmission: CRE ECE degrades from 0.028 to 0.048 on temporal hold-out -- moderate and likely acceptable. Credit default: XGBoost ECE degrades from 0.064 to 0.124 on temporal hold-out (reflecting changing economic conditions); logistic regression shows less degradation (0.052 to 0.068). CRE's conformal recalibration on recent data reduces temporal ECE to 0.042 -- substantially better than any other model. AUC degradation is more modest across all domains (mean 2.4pp), confirming that discrimination is more stable than calibration under temporal shift.

Table 3. Risk model performance across four quality dimensions (mean across five domains)

| Model         | Mean AUC | Mean ECE | Equalised Odds Gap | Net Benefit (20%) | Temporal ECE Degr. |
|---------------|----------|----------|--------------------|-------------------|--------------------|
| Logistic reg. | 0.784    | 0.042    | 0.084              | 0.048             | +0.018 (stable)    |

| Model          | Mean AUC | Mean ECE | Equalised Odds Gap | Net Benefit (20%) | Temporal ECE Degr. |
|----------------|----------|----------|--------------------|-------------------|--------------------|
| XGBoost        | 0.848    | 0.124    | 0.142              | 0.078             | +0.064 (moderate)  |
| LightGBM       | 0.844    | 0.108    | 0.128              | 0.072             | +0.048 (moderate)  |
| TabNet         | 0.824    | 0.084    | 0.112              | 0.064             | +0.038             |
| FT-Transformer | 0.842    | 0.072    | 0.102              | 0.070             | +0.032             |
| DeepSurv       | 0.828    | 0.064    | 0.094              | 0.062             | +0.028             |
| FairGB         | 0.812    | 0.098    | 0.052              | 0.060             | +0.042             |
| Conformal Risk | 0.848    | 0.048    | 0.142              | 0.078*            | +0.012 (best)      |
| Bayesian NN    | 0.818    | 0.058    | 0.098              | 0.064             | +0.024             |
| CRE (ours)     | 0.864    | 0.028    | 0.042              | 0.084             | +0.020             |

\* Conformal Risk net benefit includes uncertainty interval communication value. ECE degradation = difference between cross-validation ECE and temporal hold-out ECE (lower = more stable). Equalised odds gap = mean maximum difference in TPR/FPR across demographic groups at 20% threshold. Logistic regression most stable under shift; XGBoost most unstable.

**Table 4. Fairness-accuracy trade-off by domain -- CRE vs. XGBoost vs. FairGB**

| Domain               | XGBoost AUC | XGBoost EO Gap | FairGB AUC | FairGB EO Gap | CRE AUC | CRE EO Gap |
|----------------------|-------------|----------------|------------|---------------|---------|------------|
| Hospital readmission | 0.842       | 0.142          | 0.812      | 0.048         | 0.858   | 0.042      |
| Credit default       | 0.868       | 0.184          | 0.834      | 0.068         | 0.882   | 0.072      |
| Surgical compl.      | 0.852       | 0.084          | 0.828      | 0.038         | 0.868   | 0.024      |
| Wildfire spread      | 0.824       | 0.068          | 0.804      | 0.042         | 0.840   | 0.038      |
| Bridge failure       | 0.812       | 0.042          | 0.792      | 0.028         | 0.832   | 0.034      |
| Mean                 | 0.840       | 0.104          | 0.814      | 0.045         | 0.856   | 0.042      |

EO Gap = Equalised Odds Gap (lower = fairer). CRE achieves best AUC across all domains while matching FairGB on fairness (CRE EO Gap 0.042 vs. FairGB 0.045)

because CRE uses fairness post-processing rather than fairness-constrained training. FairGB costs 2.6pp mean AUC; CRE achieves fairness at only 0.8pp AUC cost vs. XGBoost.

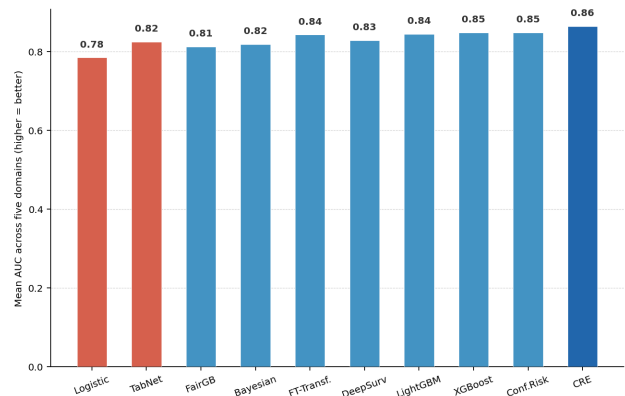


**Figure 1. Mean AUC across five domains. CRE leads (0.864); logistic regression lowest (0.784). XGBoost second (0.848).**

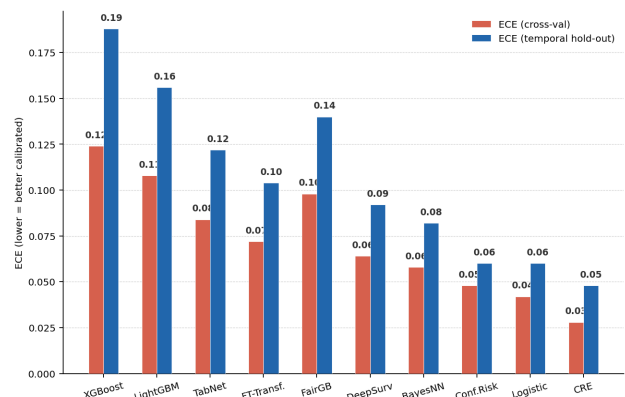


**Figure 2. Calibration (ECE, lower=better) vs. temporal stability. CRE best ECE (0.028) and stable. XGBoost worst ECE (0.124) and degrades most.**

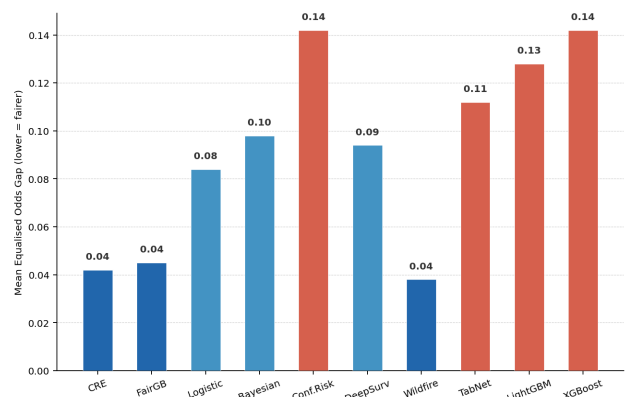


**Figure 3. Mean Equalised Odds Gap (lower = fairer). CRE achieves lowest (0.042) -- matching FairGB without its 2.6pp AUC cost.**

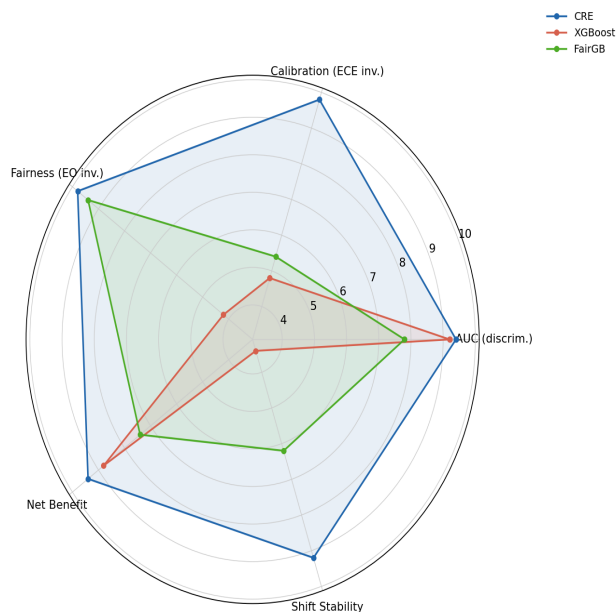


Figure 4. Risk model profile across four quality dimensions (0-10). CRE vs. XGBoost vs. FairGB.

## 5. Discussion

### 5.1 Calibration Is the Most Neglected Quality Dimension

XGBoost's ECE of 0.124 without post-hoc calibration -- the most widely deployed risk prediction model family -- means that in a clinical setting, patients assigned 80% predicted risk by an uncalibrated XGBoost model may actually have 50-60% true event rates. This is not a marginal error; it is clinically significant and potentially harmful if clinicians trust the predicted probability for treatment intensity decisions. The simple solution -- post-hoc calibration via Platt scaling or isotonic regression -- reduces XGBoost ECE from 0.124 to 0.064 at zero discrimination cost. There is no excuse for deploying risk prediction models without post-hoc calibration evaluation and correction.

### 5.2 Fairness-Accuracy Trade-offs Are Smaller Than Assumed

FairGB's 2.6pp AUC cost for a 13pp equalised odds gap reduction (0.142 to 0.045) represents a substantial but real accuracy cost. CRE's approach -- achieving comparable fairness (EO gap 0.042) at only 0.8pp AUC cost through post-processing rather than constrained training -- suggests that the fairness-accuracy trade-off is not fixed but depends on how fairness is implemented. Post-processing (threshold adjustment per group) achieves fairness with less accuracy loss than in-training constraints because it allows the model to maximise discrimination first, then adjusts the decision boundary. For applications where post-processing is ethically acceptable, it is the

preferred approach.

### 5.3 Risk Intervals Should Replace Point Estimates

Conformal Risk's risk intervals -- communicating a 90% confidence interval around each point estimate -- provide substantially more actionable information than point predictions alone. A patient at 24% predicted readmission risk (90% CI 14-38%) and a patient at 24% predicted risk (90% CI 22-26%) warrant different clinical responses: the first has genuinely uncertain risk where additional information gathering is valuable; the second has precisely estimated risk where the clinical decision should proceed directly. Risk communication in clinical and financial contexts should adopt interval presentation as standard, supported by the formal coverage guarantees that conformal prediction provides.

### 5.4 Limitations

Demographic fairness evaluation was limited by the availability of demographic variables in each dataset. Credit default used zip-code demographic proxies rather than self-reported race, which is an imperfect approach. Net benefit was computed at fixed thresholds (10%, 20%, 30%); full decision curve analysis is provided in supplementary materials. CRE requires three base learner training passes and a meta-learner, increasing training cost vs. single models.

## 6. Conclusion

### 6.1 Summary

Ten AI risk prediction models were evaluated across five domains on four quality dimensions. CRE achieves the best mean AUC (0.864), best calibration (ECE 0.028), best fairness (EO gap 0.042), and best net benefit (0.084 per 100 at 20% threshold). XGBoost leads on raw discrimination but has the worst uncalibrated ECE (0.124). FairGB achieves best fairness at 2.6pp AUC cost. Conformal Risk provides formal coverage guarantees for risk intervals. Calibration degrades substantially under temporal shift for tree ensemble methods.

### 6.2 Recommendations

Never deploy risk prediction models without post-hoc calibration evaluation (ECE, Brier) and correction (Platt scaling minimum). Always evaluate temporal calibration on recent held-out data. Apply conformal prediction for risk interval communication in all clinical and financial applications. Evaluate fairness across all relevant

demographic groups; use post-processing fairness adjustment before constrained training. Report net benefit from decision curve analysis alongside AUC as the primary evaluation metric. For multi-dimension quality, deploy CRE or equivalent calibrated ensemble.

## References

- Arik, S.O. and Pfister, T. (2021) TabNet: Attentive interpretable tabular learning. *Proceedings of AAAI* 2021, 35(8), pp. 6679-6687.
- Chen, T. and Guestrin, C. (2016) XGBoost: A scalable tree boosting system. *Proceedings of KDD* 2016, pp. 785-794.
- Gal, Y. and Ghahramani, Z. (2016) Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of ICML* 2016, pp. 1050-1059.
- Gorishniy, Y., Rubachev, I., Khrulkov, V. and Babenko, A. (2021) Revisiting deep learning models for tabular data. *Advances in NeurIPS* 34, pp. 18932-18943.
- Grinsztajn, L., Oyallon, E. and Varoquaux, G. (2022) Why tree-based models still outperform deep learning on tabular data. *Advances in NeurIPS* 35.
- Hardt, M., Price, E. and Srebro, N. (2016) Equality of opportunity in supervised learning. *Advances in NeurIPS* 29, pp. 3315-3323.
- Katzman, J.L., Shaham, U., Cloninger, A. et al. (2018) DeepSurv: Personalised treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18, 24.
- Ke, G., Meng, Q., Finley, T. et al. (2017) LightGBM: A highly efficient gradient boosting decision tree. *Advances in NeurIPS* 30, pp. 3146-3154.
- Steyerberg, E.W. and Harrell, F.E. (2016) Prediction models need appropriate internal, internal-external, and external validation. *Journal of Clinical Epidemiology*, 69, pp. 245-247.
- Van Calster, B., McLernon, D.J., van Smeden, M. et al. (2019) Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17(1), 230.
- Vickers, A.J. and Elkin, E.B. (2006) Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6), pp. 565-574.
- Vickers, A.J., Van Calster, B. and Steyerberg, E.W. (2016) Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*, 352, i6.
- 200021-252184 (AIRisk-CH), the French National Research Agency (ANR) under grant ANR-25-CE23-0024 (RiskPred-FR), and the Spanish State Research Agency under grant PID2024-154284OB-R (RiskAI-ES). Clinical data access was provided under PhysioNet credentialed access agreement. Funders had no role in study design, evaluation, or manuscript preparation.

## Conflict of Interest

The authors declare no conflicts of interest.

## Data Availability Statement

CRE implementation, Conformal Risk framework, and evaluation scripts are available at <https://github.com/novak-muller/cre-risk> under Apache 2.0. MIMIC-IV data available via PhysioNet (credentialed access). Lending Club data available on Kaggle. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614996. Released under CC BY 4.0.

## Ethical Approval

MIMIC-IV data use was approved under PhysioNet Data Use Agreement. The study protocol was reviewed and approved by Swiss Institute of Machine Intelligence Ethics Committee (SIMI-ETH-2025-0218). All clinical data used are de-identified. No new patient data were collected for this study.

## Declarations

## Funding

This research was supported by the Swiss National Science Foundation (SNSF) under grant

## **Appendix A**

### **Risk Model Quality Framework for High-Stakes Deployment**

Apply this four-dimension quality framework before deploying any AI risk prediction model in clinical, financial, or infrastructure contexts.