

Explainable Predictive Analytics for Supporting Clinical Risk Assessment in Healthcare Information Systems

Dr. Martin Novak

Professor, Slovak Institute of Clinical Analytics, Slovakia

Abstract

Healthcare information systems contain increasingly large collections of longitudinal clinical and administrative data that can support predictive risk assessment. Conventional statistical models remain important because their structure can often be interpreted directly, while more flexible machine-learning models may capture complex nonlinear relationships but provide less transparent predictions. In clinical environments, predictive accuracy alone is insufficient because practitioners must understand the factors contributing to a risk estimate before deciding whether the prediction is clinically plausible and relevant. This hypothetical methodological study evaluates explainable predictive analytics for supporting clinical risk assessment using structured healthcare information-system data. A fictional dataset containing 2,400 patient encounters and twenty-six predictor variables is divided into training, validation, and independent testing subsets. Three conceptual models are compared: regularized logistic regression, an explainable generalized additive model, and a random forest accompanied by local model-agnostic explanations.

The target outcome is a fictional composite thirty-day adverse clinical outcome used only for methodological illustration. The hypothetical analysis evaluates discrimination and classification through area under the receiver operating characteristic curve, accuracy, sensitivity, specificity, precision, F1 score, and Brier score. The random forest produces the strongest assigned predictive performance, while the generalized additive model provides a favorable balance between accuracy and intrinsic interpretability. A second analysis examines explanation fidelity, stability, sparsity, and availability, illustrating that post-hoc explanations can increase transparency but do not make a complex model equivalent to an inherently interpretable model. The study concludes that explainable predictive analytics can support clinical risk assessment when predictive performance, calibration, explanation quality, data provenance, and clinician oversight are considered jointly. Explanations should assist verification and contextual interpretation rather than create unjustified confidence in model outputs. External validation, prospective assessment, subgroup evaluation, and rigorous governance would be required before any real healthcare deployment.

Keywords: Explainable Artificial Intelligence, Clinical Risk Assessment, Predictive Analytics, Healthcare Information Systems, Clinical Decision Support, Machine Learning, Electronic Health Records, Interpretability, Risk Prediction, Health Informatics.

1. Introduction

Healthcare information systems routinely collect structured and semi-structured information describing patient encounters, physiological observations, laboratory measurements, prior utilization, diagnoses, procedures, and other clinical events. These data can support prediction of future outcomes when appropriate analytical methods are applied. Electronic health records therefore provide an important foundation for clinical risk modeling, although their use introduces challenges involving missingness, measurement variability, coding practices, temporal dependence, and differences among healthcare settings.

Clinical prediction models traditionally rely heavily on regression-based approaches because coefficients and estimated effects can be inspected directly. Steyerberg emphasized that useful clinical prediction requires careful model development, validation, discrimination assessment, calibration, and consideration of intended use. Prediction research should therefore distinguish statistical performance from clinical usefulness.

Machine-learning methods offer greater flexibility for representing nonlinear effects and interactions among clinical variables. Random forests, boosting methods, and other flexible algorithms can potentially improve prediction when healthcare variables interact in complex ways. Churpek et al. compared machine-learning methods with conventional regression for clinical deterioration prediction, while Caruana et al. demonstrated that high-performing models in healthcare can reveal unexpected data relationships that require careful interpretation before clinical use.

This interpretability requirement has contributed to growing interest in explainable predictive analytics. Inherently interpretable models communicate their prediction structure directly through coefficients, rules, or additive contributions. Post-hoc explanation methods instead approximate or describe the behavior of a more complex predictive model after it has been trained. Ribeiro et al. introduced LIME as a model-agnostic technique that constructs an interpretable approximation locally around an individual prediction, allowing users to inspect which features influence a particular classification.

The present hypothetical study evaluates prediction and explanation simultaneously within a clinical risk-assessment framework. Logistic regression, a generalized additive model, and random forest with local explanation are compared using a fictional healthcare dataset. The objective is not to automate diagnosis or determine treatment. Instead, the study examines how predictive models could generate risk information accompanied by understandable evidence that allows clinicians to review, question, or reject model outputs before any decision-support use.

2. Objectives of the Study

The principal objective of this study is to examine how explainable predictive analytics can support transparent clinical risk assessment within healthcare information systems. The investigation focuses on predictive performance, intrinsic and post-hoc interpretability, explanation stability, clinical oversight, and methodological requirements for trustworthy risk-model evaluation.

2.1 Development of an Explainable Clinical Risk Framework

The first objective is to construct a hypothetical analytical framework in which risk prediction and explanation are generated together rather than treating interpretability as an optional stage. Logistic regression represents a conventional intrinsically interpretable model, while the generalized additive

model represents a more flexible yet transparent architecture. The random forest provides a higher-capacity nonlinear comparator accompanied by local surrogate explanations.

This comparison allows interpretability to be examined as a property of the overall analytical workflow rather than merely a visual presentation layer. Direct model explanations and post-hoc explanations have different methodological meanings. A coefficient or additive contribution belongs to the predictive model itself, whereas a local surrogate describes an approximation of a complex predictor in a limited neighborhood. Clinical users should therefore be informed about the type and limitations of each explanation.

2.2 Evaluation of Predictive Performance and Calibration

The second objective is to compare model discrimination, classification, and probabilistic accuracy using a common independent test set. Area under the receiver operating characteristic curve assesses discrimination across decision thresholds, while sensitivity and specificity describe performance at the selected classification threshold. Precision and F1 score provide additional information when the outcome is less common than the non-event class.

The Brier score is included because clinical risk assessment involves estimated probabilities rather than only binary decisions. A lower Brier score indicates greater overall probabilistic accuracy, although calibration should ideally also be examined graphically and through calibration slopes and intercepts. Predictive models intended for clinical use require both discrimination and calibration because an apparently well-ranked patient population can still receive systematically inaccurate absolute risk estimates.

2.3 Assessment of Explanation Quality and Clinical Usability

The third objective is to evaluate whether explanations remain sufficiently faithful, stable, concise, and consistently available to support clinical review. Fidelity concerns whether an explanation accurately represents the underlying model locally, while stability concerns whether small changes in a patient's input produce disproportionately different explanatory feature rankings. Sparsity concerns whether explanations can be presented using a manageable number of influential variables.

The study does not treat explanation quality as proof of clinical usefulness. A technically faithful explanation may highlight a variable whose association reflects data collection patterns rather than a clinically meaningful mechanism. Human review is therefore necessary to determine whether an explanation is plausible, actionable, or potentially misleading. The framework positions explanations as tools for model scrutiny rather than as automatic justifications for predictive outputs.

3. Review of Literature

3.1 Clinical Prediction Models and Healthcare Information Systems

Clinical prediction modeling has long emphasized transparent development and validation. Steyerberg and colleagues described discrimination, calibration, and prediction error as essential dimensions of model performance.

Steyerberg et al. further provided a framework for evaluating prediction models using traditional and newer measures. These principles remain relevant when machine learning is introduced because algorithmic complexity does not remove the need for statistically appropriate validation.

Electronic health records provide new opportunities for predictive modeling because they contain longitudinal and multidimensional clinical information. Jensen et al. described the potential of mining EHR data for biomedical research and clinical care.

Amarasingham et al. demonstrated an automated EHR-based model for identifying patients at risk of thirty-day readmission or death. Such studies illustrate the possibility of embedding predictive analytics within healthcare information infrastructures while also emphasizing the importance of data quality and appropriate outcome definition.

3.2 Interpretability and Explainable Predictive Models

Caruana et al. investigated intelligible high-accuracy models in healthcare through pneumonia-risk and hospital-readmission case studies. Their work showed that model transparency can reveal potentially misleading learned relationships that might otherwise remain hidden within a high-performing black-box model. This is especially important in healthcare because apparent predictive associations can arise from treatment processes, measurement practices, or selection effects rather than biological risk alone.

Ribeiro et al. introduced LIME for explaining predictions of arbitrary classifiers through local interpretable approximations. The approach provides an important conceptual distinction between global model transparency and local explanation of individual predictions. Although local explanations can help users investigate complex predictions, they should be evaluated for stability and fidelity because the explanation procedure introduces its own modeling assumptions.

3.3 Validation, Reporting, and Clinical Utility

The TRIPOD Statement established reporting standards for multivariable prediction models intended for diagnosis or prognosis. Collins et al. emphasized transparent documentation of data sources, participants, predictors, outcomes, model development, and validation. Moons et al. provided further explanation and elaboration of these principles. Such reporting is essential for reproducibility and for determining whether model performance is likely to transfer beyond the development dataset.

Prediction performance also needs to be connected to potential clinical consequences. Vickers and Elkin introduced decision curve analysis for evaluating net benefit across decision thresholds, while Steyerberg et al. emphasized that discrimination and calibration remain core evaluation components. Explainable AI does not replace these statistical requirements. A transparent but poorly validated model remains unsuitable for clinical use, just as a highly discriminative but unexplained black-box model may require substantial scrutiny before deployment.

4. Research Methodology

4.1 Research Design and Predictive Models

The present study adopts a hypothetical retrospective methodological design using structured healthcare information-system data. Three binary risk-prediction models are compared. Regularized logistic regression provides an intrinsically interpretable linear model. A generalized additive model represents nonlinear relationships through feature-specific smooth functions while maintaining transparent additive contributions.

A random forest represents a more flexible nonlinear ensemble and is accompanied by LIME-style local explanations. The fictional target represents occurrence of a composite adverse clinical outcome within thirty days of the index encounter. The target is used only for methodological illustration

and does not correspond to any real diagnostic category or treatment decision. Model hyperparameters are assumed to be selected using the validation dataset only. The independent test partition is reserved for final evaluation and explanation analysis.

4.2 Hypothetical Dataset and Clinical Variables

The fictional dataset contains 2,400 patient encounters and twenty-six structured predictors. One thousand six hundred observations are allocated to training, four hundred to validation, and four hundred to independent testing. The test partition contains 100 hypothetical adverse-outcome observations and 300 non-event observations. No record represents a real patient, hospital, clinician, or healthcare organization.

Predictors conceptually include age group, recent healthcare utilization, heart-rate summary, respiratory-rate summary, systolic blood-pressure trend, temperature, oxygen-saturation category, selected laboratory summaries, comorbidity count, prior admission frequency, recent medication-count category, and length-of-stay indicators. Direct identifiers and free-text clinical notes are excluded. All predictors are assumed to be available before the risk-estimation time to prevent temporal leakage.

4.3 Evaluation and Explainability Procedure

Predictive performance is evaluated using AUC, accuracy, sensitivity, specificity, precision, F1 score, and Brier score. A common classification threshold is conceptually selected from the validation set before final testing. Calibration would additionally be evaluated through calibration plots and regression-based calibration assessment in a full empirical study. No threshold is selected using the independent testing observations.

Explanation evaluation differs according to model type. Logistic regression and the generalized additive model generate direct feature contributions from their predictive structures. Random-forest predictions are accompanied by a five-feature local surrogate explanation. Explanation fidelity is assessed as agreement between the local surrogate and underlying model within sampled neighboring observations. Stability is assessed through consistency of the five highest-ranked features following small, clinically plausible perturbations to input variables.

5. Analysis

5.1 Comparative Clinical Risk Prediction

The first hypothetical analysis evaluates the three models on the 400-record independent test dataset. Logistic regression correctly identifies seventy-six of 100 event observations and 245 of 300 non-event observations. The generalized additive model identifies eighty-two events and 253 non-events correctly. The random forest records the strongest assigned classification performance, correctly identifying eighty-six events and 258 non-events.

The random forest achieves hypothetical accuracy of 86.00%, sensitivity of 86.00%, specificity of 86.00%, precision of 67.19%, and F1 score of 75.44%. Its assigned AUC is 0.88 and Brier score is 0.137. The generalized additive model performs slightly less strongly but provides intrinsic feature-level transparency. These values illustrate an accuracy–interpretability comparison and do not represent clinical evidence from actual patients.

Table 1: Hypothetical Predictive Performance for Clinical Risk Assessment

Model	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)	Brier Score
Regularized Logistic Regression	0.82	80.25	76.00	81.67	58.02	65.80	0.158
Explainable Generalized Additive Model	0.85	83.75	82.00	84.33	63.57	71.62	0.146
Random Forest + Local Explanation	0.88	86.00	86.00	86.00	67.19	75.44	0.137

5.2 Explanation Fidelity and Stability

The second hypothetical analysis examines the explanatory properties of the three predictive configurations. Logistic regression and the generalized additive model provide direct explanations because their displayed contributions are derived from the predictive function itself. Their conceptual fidelity is therefore represented as complete within the stated model parameterization. The generalized additive model demonstrates higher hypothetical feature-ranking stability than logistic regression because smooth nonlinear functions reduce some threshold-related explanatory variation.

Table 2: Hypothetical Explanation Characteristics of the Predictive Models

Model / Explanation Strategy	Explanation Type	Mean Local Fidelity	Top-5 Feature Stability	Mean Features Displayed	Explanation Availability (%)
Logistic Regression	Intrinsic coefficient contribution	1.00	0.82	5	100
Generalized Additive Model	Intrinsic additive contribution	1.00	0.88	5	100
Random Forest + Local Surrogate	Post-hoc local approximation	0.89	0.74	5	100

6. Discussion

6.1 Predictive Performance and Clinical Interpretability

The hypothetical results demonstrate that a more flexible model can provide stronger predictive performance than a simpler linear baseline while creating additional interpretability requirements. The random forest achieves the highest assigned AUC and lowest Brier score, but its internal decision structure is difficult to summarize directly for an individual prediction. The generalized additive model provides slightly weaker performance but maintains an interpretable additive structure.

This pattern is consistent with the healthcare modeling problem identified by Caruana et al., who showed that high predictive accuracy and intelligibility need not always be mutually exclusive. Model selection in clinical environments should therefore consider whether a modest change in discrimination justifies a substantial change in interpretability. The best model is not necessarily the one with the highest AUC if clinicians cannot adequately validate or contextualize its predictions.

6.2 Local Explanations as Clinical Decision-Support Information

Local explanations can help clinicians understand why a particular patient receives a high or low model-estimated risk. For example, an explanation may indicate that recent physiological trends and utilization history contributed more strongly to an estimated risk than other variables. This information can support critical review by allowing the user to compare the model's reasoning with the available clinical context.

However, local explanation should not be interpreted as causal evidence. If a model highlights one laboratory measurement, the explanation indicates that the variable influenced the model prediction, not that changing the measurement would necessarily change the patient's true clinical outcome. Ribeiro et al.'s local-surrogate framework is valuable for inspecting predictions, but explanatory faithfulness and stability remain important because the explanation is itself an approximation of the underlying classifier.

6.3 Calibration, Validation, and Clinical Governance

Clinical prediction requires more than discrimination. Steyerberg et al. emphasized calibration and decision-analytic evaluation in addition to AUC-based assessment. A model that systematically predicts a 40% probability when the true event frequency is closer to 20% may still rank patients effectively while providing misleading absolute risk estimates. Calibration should therefore be monitored whenever predicted probabilities are shown to clinicians.

External validation is similarly important because healthcare systems differ in patient populations, documentation, clinical workflows, and measurement practices. A model trained in one information system can exploit patterns that do not generalize elsewhere. Explainability may help identify implausible relationships, but it cannot substitute for independent validation. Deployment would additionally require model monitoring, access controls, documentation, change management, and clear responsibility for clinical interpretation.

7. Challenges

7.1 Data Quality, Missingness, and Temporal Leakage

Healthcare information systems contain data generated during care rather than through a perfectly controlled research protocol. Laboratory measurements may be ordered selectively, vital signs may be recorded at different frequencies, and missingness can itself reflect clinical decision-making.

Consequently, a predictive model may learn patterns associated with documentation processes instead of underlying health risk.

Temporal leakage creates an additional threat because variables recorded after the prediction point can inadvertently enter the feature set. This can make retrospective model performance appear much stronger than real-time performance would be. A rigorous pipeline must define an index time, include only information available before that time, and ensure that preprocessing decisions do not use future outcomes.

7.2 Explanation Reliability and Automation Bias

An explanation can increase user confidence even when the underlying model is inaccurate. This creates a risk of automation bias if clinicians interpret a plausible-looking feature list as evidence that the prediction must be correct. Explanation interfaces should therefore communicate uncertainty and should not visually imply that model-generated importance is equivalent to clinical causality.

Explanation instability can also reduce usability. If clinically insignificant changes in input measurements produce entirely different feature rankings, users may struggle to understand the model consistently. Explanation methods should therefore be evaluated empirically for fidelity, stability, sparsity, and user comprehension. Healthcare explainability research should assess whether explanations help clinicians identify erroneous predictions rather than merely whether clinicians report liking the interface.

7.3 Generalization, Fairness, and Clinical Accountability

Predictive performance may vary among hospitals, demographic groups, disease patterns, and levels of healthcare access. A model can achieve acceptable overall discrimination while performing poorly for a smaller subgroup. Subgroup evaluation and external validation are therefore necessary before clinical use. Differences should be investigated rather than assumed to reflect biological risk.

Clinical accountability must also remain explicit. A predictive system should not become the final authority for diagnosis, treatment, admission, discharge, or other consequential care decisions. Clinicians need access to the underlying clinical record, prediction uncertainty, relevant model information, and appropriate escalation pathways. Explainability is valuable only when embedded within a governance framework that preserves professional responsibility and patient-centered decision-making.

8. Conclusion

Explainable predictive analytics offers a methodological framework for integrating clinical risk prediction with transparent information about the factors influencing model outputs. The present hypothetical study compared regularized logistic regression, an explainable generalized additive model, and a random forest with local model-agnostic explanations using a fictional healthcare information-system dataset.

The illustrative analysis assigned the strongest predictive performance to the random forest, while the generalized additive model offered a favorable combination of accuracy and intrinsic interpretability. The explanation analysis demonstrated that post-hoc local explanations can provide useful approximations of complex predictions but may exhibit lower fidelity and stability than directly

interpretable model structures. Every dataset element and numerical result reported in this study is hypothetical.

The study concludes that explainability should not be treated as a cosmetic addition to a clinical prediction system. Reliable clinical decision support requires predictive discrimination, calibration, explanation quality, temporal integrity, external validation, subgroup assessment, and human oversight. Future empirical research should evaluate whether explanations improve clinicians' ability to detect incorrect predictions, understand model limitations, and use risk estimates appropriately without increasing automation bias.

References

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
2. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20, Article 310. <https://doi.org/10.1186/s12911-020-01332-6>
3. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
4. Barredo Arrieta, A., Gil-López, S., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence: Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
5. Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *JAMA*, 318(6), 517–518. <https://doi.org/10.1001/jama.2017.7797>
6. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721–1730). ACM. <https://doi.org/10.1145/2783258.2788613>
7. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor AI: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference* (pp. 301–318). PMLR.
8. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. <https://arxiv.org/abs/1702.08608>
9. Freyer, N., Groß, D., & Lipprandt, M. (2024). The ethical requirement of explainability for AI-DSS in healthcare: A systematic review of reasons. *BMC Medical Ethics*, 25, Article 104. <https://doi.org/10.1186/s12910-024-01103-2>
10. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745–e750. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9)
11. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4), Article e1312. <https://doi.org/10.1002/widm.1312>
12. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 30.
13. Lundberg, S. M., Erion, G. G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2, 56–67. <https://doi.org/10.1038/s42256-019-0138-9>

14. McCoy, L. G., Brenna, C. T. A., Chen, S. S., Vold, K., & Das, S. (2022). Believing in black boxes: Machine learning for healthcare does not need explainability to be evidence-based. *Journal of Clinical Epidemiology*, 142, 252–257.
15. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). ACM. <https://doi.org/10.1145/3287560.3287596>
16. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “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* (pp. 1135–1144). ACM. <https://doi.org/10.1145/2939672.2939778>
17. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
18. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Proceedings of the Machine Learning for Healthcare Conference* (pp. 359–380). PMLR.
19. Van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 79, Article 102470. <https://doi.org/10.1016/j.media.2022.102470>
20. Amann, J., Vayena, E., Blasimme, A., Frey, D., & Madai, V. I. (2022). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 22, Article 250.