Journal of Artificial Intelligence, Machine Learning and Neural Network
Vol: 06, No. 02, July-Dec 2026
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.62.98.105

Research Paper

Stability and robustness of explainable artificial intelligence methods in clinical prediction models: Protocol for a systematic review

Surjeet Dalal[1*], Danish Ather[2]
[1*]Professor & Head, Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Gurugram, Haryana, India; Visiting Professor, City Institute of Higher Learning (C-IHL), Canada.
[2]Professor, Department of Artificial Intelligence, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan.

Article Info

Article History:
Received: 25 June 2026
Revised: 24 August 2026
Accepted: 15 September 2026
Published: 22 October 2026

Keywords:
Systematic Review Protocol
Prisma-P
Explainable Artificial Intelligence
Explanation Stability
Clinical Prediction Models

ABSTRACT

Explainable artificial intelligence (XAI) methods such as SHAP, LIME and permutation feature importance are widely used to justify machine learning models for clinical prediction, yet an explanation is useful only if it is reproducible. Explanations may change when a model is refitted on resampled data, when inputs are slightly perturbed, when a different but equally accurate model is chosen, or when a different explanation method is applied. It is not known how often published clinical prediction studies assess these forms of stability, how they measure them, or how stable the reported explanations are. This protocol describes a systematic review that will address these questions. Scopus, Web of Science Core Collection, PubMed/MEDLINE and IEEE Xplore will be searched for primary studies published from January 2016 onwards that develop or validate a clinical prediction model using machine learning and report a quantitative assessment of the stability, robustness, consistency or agreement of feature-attribution explanations. Two reviewers will independently screen records and full texts, extract data with a form based on the CHARMS checklist, and assess risk of bias with PROBAST. Outcomes will include the type of stability assessed, the metrics used, the magnitude of reported stability, and factors associated with instability, such as sample size, feature correlation and model family. Results will be synthesised narratively and, where at least five studies report comparable metrics, summarised quantitatively. The review will be reported according to PRISMA 2020 and PRISMA-S and is expected to inform reporting standards for explanations in clinical machine learning.

Corresponding Author:
Surjeet Dalal

Journal of Artificial Intelligence, Machine Learning and Neural

Professor & Head, Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Gurugram, Haryana, India; Visiting Professor, City Institute of Higher Learning (C-IHL), Canada.
Email: sdalal@ggn.amity.edu

Copyright © 2026 The Author(s). This is an open access article distributed under the Creative Commons Attribution License, (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

1.1 Rationale

Machine learning (ML) models for diagnosis, prognosis and risk stratification are increasingly accompanied by explanations of which input variables drive their predictions [1], [2]. Post hoc feature-attribution methods, in particular SHAP [3], LIME [4] and permutation feature importance [5], [6], have become a routine part of clinical ML papers and are often interpreted as evidence about clinical risk factors or as a basis for clinician trust [7]. An explanation, however, is only informative if it is reproducible. Four sources of instability have been described: sampling variability, when an explanation changes after the model is refitted on resampled data [8], [9]; input sensitivity, when small perturbations of an instance change its local explanation [10]; model multiplicity, when equally accurate models explain the data differently [6], [11]; and method disagreement, when different explanation methods applied to the same model disagree [12]. Feature dependence can further distort attribution methods based on permutation [13], [14].

Critics have argued that current explanation practice in health care may give a false sense of assurance [15], [16], and updated reporting guidance for clinical prediction models encourages transparency about model interpretation [17]. Yet there is no systematic overview of how often published clinical prediction studies assess the stability of their explanations, which types of stability they examine, which metrics they use, and how stable the reported explanations actually are.

1.2 Objectives

The review will answer four questions: (RQ1) What proportion of eligible clinical ML studies that report feature-attribution explanations also assess the stability of those explanations, and which of the four types of stability do they assess? (RQ2) Which metrics and experimental protocols are used to quantify stability? (RQ3) What magnitudes of stability are reported, and how do they differ between explanation methods and model families? (RQ4) Which study characteristics, such as sample size, number and correlation of predictors, or model family, are associated with lower stability?

2. METHODOLOGY

This protocol follows the PRISMA-P 2015 statement [18]. The completed review will be reported according to PRISMA 2020 [19] and its search extension PRISMA-S [20]. The protocol will be registered in PROSPERO before screening begins, and any amendments will be dated and justified in the final report.

2.1 Eligibility Criteria

Table 1. Eligibility Criteria (PICOS-Style Framework)

Element	Inclusion	Exclusion
Population / data	Human clinical or health data (electronic health records, registries, trials, imaging-derived or tabular clinical features, physiological signals)	Non-human data; purely simulated data without any clinical dataset
Intervention	Development or validation of a clinical diagnostic, prognostic or risk prediction model using machine learning, accompanied by a feature-attribution explanation (e.g., SHAP, LIME, permutation importance, integrated gradients, saliency)	Explanations limited to model coefficients of a pre-specified regression model without an ML component; qualitative explanations only
Comparator	Not required	–
Outcome	Any quantitative assessment of explanation stability, robustness, consistency or agreement: across resamples, perturbations, models, random seeds or explanation methods	Explanations reported from a single fit without any stability assessment (these are counted for RQ1 in a sampling frame, see Section 2.7, but not included for RQ2–RQ4)
Study design	Primary research articles and full conference papers	Reviews, editorials, commentaries, protocols, abstracts only, theses
Other	Published January 2016 or later; English language	Retracted articles

2.2 Information Sources

Scopus, Web of Science Core Collection, PubMed/MEDLINE and IEEE Xplore will be searched from 1 January 2016 to the date of the search, which will be reported exactly. 2016 was chosen because LIME [4] and SHAP [3], the most widely used attribution methods, were published in 2016 and 2017. Reference lists of included studies and of relevant reviews will be screened (backward citation searching), and studies citing the included studies will be identified in Scopus (forward citation searching).

2.3 Search Strategy

The search combines three concept blocks: explanation methods, stability-related terms, and clinical prediction. The full strings for each database are given in Table *2*; they will be piloted, and their sensitivity checked against a set of known relevant studies identified before the search. Search results will be exported with abstracts in RIS or CSV format, and the date, database, interface and number of records will be logged for each search [20].

Table 2. Database-Specific Search Strings

Database	Search String
Scopus	TITLE-ABS-KEY (("explainab" OR "interpretab" OR "XAI" OR "SHAP" OR "Shapley" OR "LIME" OR "feature importance" OR "feature attribution" OR "variable importance" OR "saliency" OR "integrated gradients") AND ("stabilit" OR "robust" OR "consisten" OR "reproducib" OR "agreement" OR "disagreement" OR "variabilit" OR "sensitivit") AND ("clinical" OR "patient" OR "diagnos" OR "prognos" OR "mortality" OR "risk prediction" OR "health" OR "medical") AND ("machine learning" OR "deep learning" OR "neural network*" OR "random forest" OR "gradient boosting" OR "XGBoost")) AND PUBYEAR > 2015 AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp")) AND (LIMIT-TO (LANGUAGE , "English"))

Database	Search String
Web of Science Core Collection	TS=((explainab OR interpretab OR XAI OR SHAP OR Shapley OR LIME OR "feature importance" OR "feature attribution" OR "variable importance" OR saliency OR "integrated gradients") AND (stabilit OR robust OR consisten OR reproducib OR agreement OR disagreement OR variabilit OR sensitivit) AND (clinical OR patient OR diagnos OR prognos OR mortality OR "risk prediction" OR health OR medical) AND ("machine learning" OR "deep learning" OR "neural network*" OR "random forest" OR "gradient boosting" OR XGBoost)); Publication years 2016–present; Document types: Article, Proceedings Paper; Language: English
PubMed/MEDLINE	(explainab[tiab] OR interpretab[tiab] OR XAI[tiab] OR SHAP[tiab] OR Shapley[tiab] OR LIME[tiab] OR "feature importance"[tiab] OR "feature attribution"[tiab] OR "variable importance"[tiab] OR saliency[tiab]) AND (stabilit[tiab] OR robust[tiab] OR consisten[tiab] OR reproducib[tiab] OR agreement[tiab] OR disagreement[tiab] OR variabilit[tiab]) AND ("Machine Learning"[Mesh] OR "machine learning"[tiab] OR "deep learning"[tiab] OR "neural network"[tiab] OR "random forest"[tiab] OR "gradient boosting"[tiab] OR XGBoost[tiab]) AND 2016:3000[dp] AND english[la]
IEEE Xplore	("All Metadata": explainab OR interpretab OR SHAP OR LIME OR "feature importance" OR "feature attribution") AND ("All Metadata": stabilit OR robust OR consisten OR reproducib OR agreement) AND ("All Metadata": clinical OR patient OR medical OR diagnos OR prognos*); Filters: 2016–present; Journals and Conferences

2.4 Study Selection

Records will be de-duplicated using DOIs and normalised titles, and duplicates will be checked manually. Two reviewers will independently screen titles and abstracts in Rayyan [21] against the eligibility criteria after a calibration round of 50 records; agreement will be measured with Cohen's κ [22], and screening will proceed when $\kappa \geq 0.70$. Full texts of potentially eligible records will be assessed independently by the same reviewers, with reasons for exclusion recorded for every excluded report. Disagreements will be resolved by discussion or, if necessary, by a third reviewer. Reports that cannot be retrieved after contacting the authors will be listed as not retrieved. The selection process will be summarised in a PRISMA 2020 flow diagram Figure *1*.

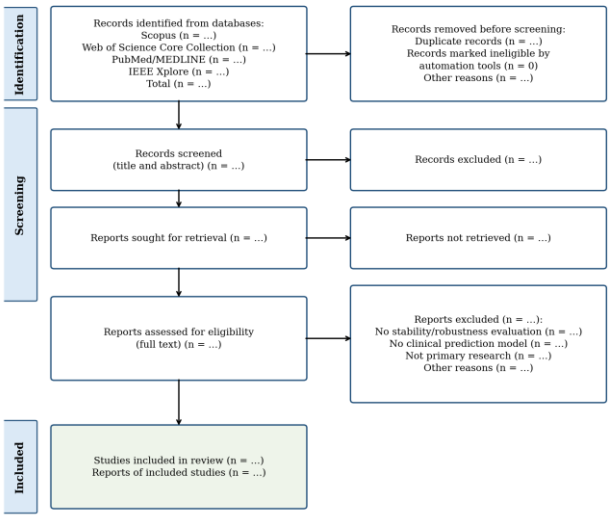

Figure 1. PRISMA 2020 Flow Diagram Template for the Review. All Counts (n = ...) Will be Completed from the Logged Search and Screening Records

2.5 Data Extraction

Data will be extracted with a piloted form Table *3* based on the CHARMS checklist [23] and extended with items specific to explanations. Two reviewers will extract data independently from a random 20% of included studies to check consistency; the remainder will be extracted by one reviewer and verified by a second.

Table 3. Data Extraction Items

Domain	Items
Study	Authors, year, country, venue, funding, code and data availability
Data and population	Clinical domain, prediction target, data source, sample size, number of events, number of candidate predictors, reported predictor correlation
Model	Model family (linear, tree ensemble, kernel, neural network, other), tuning procedure, validation design, discrimination and calibration results
Explanation	Attribution method(s), global or local explanation, data used to compute explanations (training, validation, test), software
Stability assessment	Type (sampling, input perturbation, model multiplicity, method disagreement, random seed), protocol (e.g., bootstrap, repeated cross-validation, noise level), number of repetitions
Stability results	Metric (e.g., rank correlation, top-k overlap, Jaccard index, cosine similarity, coefficient of variation), reported values, comparison between methods or models, authors' interpretation
Reporting	Whether unstable explanations were acknowledged in the conclusions; whether clinical claims were based on explanations

2.6 Risk of Bias and Quality of Reporting

Risk of bias and applicability of the prediction models will be assessed with PROBAST [24] across the participants, predictors, outcome and analysis domains. Because no validated tool exists for the quality of stability assessments, we will additionally record five descriptive indicators: whether stability was assessed on held-out data, whether the number of repetitions was stated, whether a quantitative metric was reported, whether uncertainty in the metric was reported, and whether correlated predictors were handled. These indicators will be reported descriptively and will not be combined into a score.

2.7 Data Synthesis

For RQ1, a random sample of 200 clinical ML studies that report feature-attribution explanations, drawn from the records retrieved by the first and third concept blocks without the stability block, will be classified by whether any stability assessment was reported; this sampling frame avoids the bias of estimating prevalence only from studies selected for mentioning stability. For RQ2 and RQ3, stability types, protocols, metrics and reported values will be tabulated and synthesised narratively, grouped by explanation method and model family. Where at least five studies report the same metric for the same type of stability (for example, mean pairwise Spearman correlation of global importance across resamples), values will be summarised with medians and interquartile ranges and, if appropriate, pooled with a random-effects model on the Fisher z scale. For RQ4, associations between study characteristics and reported stability will be explored descriptively and, if data permit, with meta-regression. Heterogeneity is

expected to be substantial, and no pooling will be performed where metrics or protocols are not comparable.

2.8 Meta-Bias and Certainty of Evidence

Selective reporting will be considered by comparing methods and results sections within included studies and, where available, published protocols or code. Because studies that find unstable explanations may be less likely to report them, the RQ1 sampling frame will also be used to estimate how often stability is assessed but not quantified. Certainty will be discussed qualitatively in view of the risk-of-bias assessments and the consistency of findings.

2.9 Discussion

This review will provide the first systematic evidence on how the stability of feature-attribution explanations is assessed in clinical ML, and on how stable those explanations are when it is assessed. Its findings are expected to inform reporting recommendations, for example that explanations used to support clinical claims should be accompanied by resampling-based stability estimates, and to identify the settings, such as small samples or highly correlated predictors, in which explanations are least reliable. Anticipated limitations include heterogeneous terminology, which may reduce search sensitivity and is mitigated by broad search terms and citation searching; heterogeneous stability metrics, which may preclude quantitative synthesis; and the restriction to English-language publications.

2.9.1 Timeline

Table 4. Planned Timeline

Stage	Planned Period
Protocol registration (PROSPERO)	Month 1
Search, de-duplication and calibration	Month 1
Title/abstract and full-text screening	Months 2–3
Data extraction and risk of bias	Months 3–4
Synthesis and manuscript preparation	Months 5–6

Acknowledgments

The authors thank the researchers whose work informed this protocol.

Funding Information

This research has received no external funding.

Author Contribution Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Surjeet Dalal	✓	✓	✓		✓				✓		✓			
Danish Ather		✓		✓		✓		✓		✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

Conflict of Interest Statement

No conflict of interest.

Informed Consent

Not Applicable.

Ethical Approval

Not Applicable. This review will use only published data.

Data Availability

No data were generated for this protocol. Search logs, screening decisions and extracted data from the completed review will be made publicly available.

REFERENCES

[1] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018. doi.org/10.1109/ACCESS.2018.2870052

[2] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793–4813, 2021. doi.org/10.1109/TNNLS.2020.3027314

[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. doi.org/10.48550/arXiv.1705.07874

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144. doi.org/10.1145/2939672.2939778

[5] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi.org/10.1023/A:1010933404324

[6] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," Journal of Machine Learning Research, vol. 20, no. 177, pp. 1–81, 2019. jmlr.org/papers/v20/18-760.html

[7] 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, p. 310, 2020. doi.org/10.1186/s12911-020-01332-6

[8] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: A study on high-dimensional spaces," Knowledge and Information Systems, vol. 12, no. 1, pp. 95–116, 2007. doi.org/10.1007/s10115-006-0040-8

[9] S. Nogueira, K. Sechidis, and G. Brown, "On the stability of feature selection algorithms," Journal of Machine Learning Research, vol. 18, no. 174, pp. 1–54, 2018. jmlr.org/papers/v18/17-514.html

[10] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," arXiv preprint arXiv:1806.08049, 2018. doi.org/10.48550/arXiv.1806.08049

[11] A. D'Amour et al., "Underspecification presents challenges for credibility in modern machine learning," Journal of Machine Learning Research, vol. 23, no. 226, pp. 1–61, 2022. jmlr.org/papers/v23/20-1335.html

[12] S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju, "The disagreement problem in explainable machine learning: A practitioner's perspective," arXiv preprint arXiv:2202.01602, 2022. doi.org/10.48550/arXiv.2202.01602

[13] G. Hooker, L. Mentch, and S. Zhou, "Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance," Statistics and Computing, vol. 31, no. 6, p. 82, 2021. doi.org/10.1007/s11222-021-10057-z

[14] C. Molnar et al., "General pitfalls of model-agnostic interpretation methods for machine learning models," in xxAI – Beyond Explainable AI, Lecture Notes in Computer Science, vol. 13200. Cham: Springer, 2022, pp. 39–68. doi.org/10.1007/978-3-031-04083-2_4

[15] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," The Lancet Digital Health, vol. 3, no. 11, pp. e745–e750, 2021. doi.org/10.1016/S2589-7500(21)00208-9

[16] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019. doi.org/10.1038/s42256-019-0048-x

[17] G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, p. e078378, 2024. doi.org/10.1136/bmj-2023-078378

[18] D. Moher et al., "Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement," Systematic Reviews, vol. 4, p. 1, 2015. doi.org/10.1186/2046-4053-4-1

[19] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, p. n71, 2021. doi.org/10.1136/bmj.n71

[20] M. L. Rethlefsen et al., "PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews," Systematic Reviews, vol. 10, p. 39, 2021. doi.org/10.1186/s13643-020-01542-z

[21] M. Ouzzani, H. Hammady, Z. Fedorowicz, and A. Elmagarmid, "Rayyan — a web and mobile app for systematic reviews," Systematic Reviews, vol. 5, p. 210, 2016. doi.org/10.1186/s13643-016-0384-4

[22] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, 1960. doi.org/10.1177/001316446002000104

[23] K. G. M. Moons et al., "Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS checklist," PLoS Medicine, vol. 11, no. 10, p. e1001744, 2014. doi.org/10.1371/journal.pmed.1001744

[24] R. F. Wolff et al., "PROBAST: A tool to assess the risk of bias and applicability of prediction model studies," Annals of Internal Medicine, vol. 170, no. 1, pp. 51–58, 2019. doi.org/10.7326/M18-1376

How to Cite: Surjeet Dalal (2026). Stability and robustness of explainable artificial intelligence methods in clinical prediction models: Protocol for a systematic review. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), 6(2), 98-105. https://doi.org/10.55529/jaimlnn.62.98.105

BIOGRAPHIES OF AUTHORS

Prof. (Dr.) Surjeet Dalal is a Professor and Head of the Department of Computer Science and Engineering at Amity School of Engineering and Technology, Amity University Gurugram, Haryana, India, and a Visiting Professor at the City Institute of Higher Learning, Canada. His research concerns artificial intelligence and machine learning for clinical and health-care applications, including predictive modelling, decision support and the interpretability of models used in medical settings.

Prof. (Dr.) Danish Ather is a Professor in the Department of Artificial Intelligence at Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan. His research covers artificial intelligence methods and their application across engineering and health domains, including predictive modelling, uncertainty estimation and the evaluation of model explanations. He has a continuing interest in the methodology of AI research and in the standards by which such work is reviewed.