INPLASY

Protocol for a Scoped Systematic Review and Meta-Regression of Resampling Methods in Imbalanced Medical Datasets: Data-Level and Algorithm-Level Strategies in Clinical Prediction Models

INPLASY202550026
doi: 10.37766/inplasy2025.5.0026
Received: 12 May 2025
Published: 12 May 2025

Abdelhay, O; Shatnawi, A; Najadat, H; Altamimi, T.

Corresponding author:
Osama Abdelhay

osamaabdelhay@gmail.com

Author Affiliation:
Jordan University for Science and Technology.

ADMINISTRATIVE INFORMATION

Support - No financial support.

Review Stage at time of this submission - Completed but not published.

Conflicts of interest - None declared.

INPLASY registration number: INPLASY202550026

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 12 May 2025 and was last updated on 12 May 2025.

INTRODUCTION

Review question / Objective Review Question (PICOTS) Population (P): Clinical datasets with binary outcomes where the minority class constitutes less than 30% of the total observations.
Intervention (I):
Resampling strategies including:
Oversampling (e.g., SMOTE, ROS)
Undersampling (e.g., RUS, NearMiss)
Hybrid approaches (e.g., SMOTE+ENN)
Algorithm-level strategies (e.g., cost-sensitive learning, focal loss)
Comparator (C):
Models trained on original (imbalanced) data without any correction
Logistic regression models as reference classifiers
Alternative balancing methods
Outcomes (O):
Discriminative performance (AUC, sensitivity, specificity, F1-score, accuracy)
Calibration (Brier score, calibration slope)
Reported misclassification costs (if applicable)
Timing (T):
Studies published from January 2009 to December 31, 2024.
Setting (S):
Clinical or healthcare prediction settings; primary research using retrospective or prospective study designs.

Rationale Clinical prediction models are increasingly applied across various healthcare domains to support diagnosis, prognosis, and treatment decisions. However, a persistent methodological challenge in this field is class imbalance, where the number of observations in one outcome class (e.g., disease-positive cases) is much smaller than the other. This imbalance often leads to biased model performance, typically favouring the majority class and reducing the model's ability to detect minority events, frequently the most clinically important ones (e.g., rare adverse outcomes, disease onset).
To address this, a wide array of resampling techniques has been developed. Data-level approaches such as oversampling (e.g., SMOTE),

undersampling, and hybrid methods are widely adopted. Additionally, algorithm-level strategies like cost-sensitive learning have emerged as principled alternatives, especially for small or highly imbalanced datasets. Despite their popularity, the empirical evidence guiding when and how these techniques improve performance in real-world medical applications remains fragmented, often derived from small-scale studies with inconsistent reporting standards.

Most prior reviews have either lacked quantitative synthesis or focused narrowly on specific diseases or models. There is a pressing need for a comprehensive and methodologically rigorous synthesis that compares these strategies across clinical domains and quantifies their impact on discriminative and calibration performance. Furthermore, the increasing use of machine learning in medicine demands guidance on the interplay between model choice, imbalance correction, and dataset characteristics, such as sample size and feature dimensionality.

This review addresses that gap by systematically synthesising 15 years of research on resampling methods applied to imbalanced clinical datasets. A scoped systematic review and meta-regression aim to provide evidence-based guidance on the conditions under which resampling methods improve model validity and when alternative strategies should be considered. This review will help stakeholders—researchers, clinicians, and policymakers—develop more reliable, fair, and clinically useful predictive models by evaluating methodological moderators and adjusting for publication bias. A scoping review was selected for this study because of the heterogeneity and breadth of existing research on resampling methods in imbalanced medical datasets. The literature spans multiple clinical domains, modelling frameworks (e.g., logistic regression, random forests, neural networks), and balancing strategies (e.g., SMOTE, undersampling, cost-sensitive learning). Many studies vary in design, reporting standards, and metrics (e.g., AUC, F1-score, calibration), making a narrow systematic review unsuitable for capturing the whole methodological landscape.

Scoping reviews are particularly appropriate when:

The evidence base is complex and diverse.

The goal includes mapping concepts, practices, and methodological patterns;

A comprehensive understanding is needed to inform future systematic reviews or empirical work.

In this context, a scoping review enables us to:

Catalogue and classify the range of resampling strategies used.

Describe trends across time, medical domains, and modelling approaches;

Identify methodological gaps and reporting inconsistencies.

To complement the mapping function of the scoping review, we employ meta-regression to quantitatively assess how study-level moderators —such as imbalance ratio, sample size, resampling method, and model type—affect reported performance metrics (mainly AUC). This mixed-methods approach allows us to extract descriptive and inferential insights, offering a more nuanced understanding than traditional reviews limited to vote-counting or narrative synthesis.

This hybrid design is aligned with emerging best practices for evidence synthesis in computational health research and addresses the lack of quantitative generalisability in prior reviews of resampling methods.

Condition being studied This review investigates imbalanced binary classification problems in clinical prediction models, where one outcome class (typically the diseased or adverse outcome group) is significantly underrepresented relative to the other. Class imbalance is pervasive in medical datasets, particularly in predicting rare conditions, complications, adverse events, or disease onset.

The clinical conditions represented in the included studies span a broad range of medical domains, including but not limited to:

Cardiovascular diseases (e.g., heart failure, stroke, coronary artery disease)

Cancer (e.g., breast, cervical, lung)

Diabetes and metabolic disorders

Neuropsychiatric conditions (e.g., Parkinson's, Alzheimer's, mental health diagnoses)

Infectious diseases (e.g., hepatitis, COVID-19, malaria)

By focusing on the methodological challenge of class imbalance rather than a specific disease, this review seeks to provide generalizable evidence across medical prediction contexts, addressing a statistical condition that threatens the validity, sensitivity, and fairness of clinical AI and machine learning models.

METHODS

Search strategy Databases Searched

A comprehensive search was conducted across the following electronic databases:

MEDLINE (via PubMed)

EMBASE

Scopus

Web of Science

IEEE Xplore

These sources were selected to ensure broad coverage of clinical and computational research on predictive modeling in healthcare.

Time Frame
Eligible studies were those published between January 2009 and December 31, 2024. This time frame captures the rapid growth in machine learning and the increasing adoption of resampling strategies in medical AI research.
Language
No restrictions were placed on language. However, the feasibility of translation was considered at the full-text screening stage.
Search Terms and Query Structure
Search terms were developed through expert consultation, pilot searches, and iterative refinement. The final query combined controlled vocabulary (e.g., MeSH) and keywords related to three key concepts:
Imbalance or class imbalance
Resampling or balancing techniques
Medical or clinical prediction studies
Example PubMed (MEDLINE) query:
(("class imbalance" OR imbalanced OR "minority class" OR "data imbalance") AND
(resampling OR oversampling OR undersampling OR SMOTE OR "synthetic minority" OR "sampling strategy" OR "balancing technique") AND
("medical dataset" OR "clinical data" OR "healthcare prediction" OR diagnosis OR prognosis OR "clinical decision support"))
Queries were adapted to the syntax of each database (e.g., MeSH terms in PubMed, Emtree in EMBASE, subject headings in IEEE).
Supplementary Searches
To ensure comprehensive inclusion:

Backwards and forward citation tracking was conducted for all included full-text studies.
Grey literature sources such as medRxiv, arXiv, and bioRxiv were screened for preprints using the inclusion criteria.
Code repositories (e.g., GitHub) were searched when studies referred to external implementations of resampling methods, provided sufficient documentation, or an associated paper was available.
Search Management
Deduplication was performed using Zotero v7.0.15.
Two independent reviewers conducted the Title and abstract screening, supported by ASReview software for prioritisation.
Full-text screening followed predefined eligibility criteria, with discrepancies resolved by consensus.
All steps were documented in a PRISMA flow diagram.

Participant or population This review targets clinical prediction models applied to medical datasets with binary outcomes, where one class (typically representing the event of interest, such as disease presence or complication) constitutes a minority class (<30%) of the sample. The populations are not restricted to a specific disease or demographic but encompass a broad range of clinical contexts in which class imbalance is explicitly reported.
The included studies span diverse patient populations, such as:
Individuals with or at risk of cardiovascular diseases (e.g., stroke, myocardial infarction)
Cancer patients, across various tumour types
Persons with diabetes mellitus and metabolic disorders
Patients diagnosed with neuropsychiatric disorders (e.g., Parkinson's, Alzheimer's)
Populations affected by infectious diseases (e.g., hepatitis, COVID-19, malaria)
These datasets may include hospital records, registry data, cohort studies, or EHR-based prediction models. The common denominator across all included studies is applying a binary classification task in a clinically relevant prediction setting where class imbalance poses methodological concerns.

Intervention The interventions under review are strategies to correct class imbalance in medical datasets used for binary clinical prediction. These include data-level (resampling) methods and algorithm-level (cost-sensitive) methods to improve model sensitivity, discrimination, and fairness for the minority class.
1. Data-Level (Resampling) Methods
These modify the dataset before model training:
Oversampling Techniques
Random Oversampling (ROS): Duplicates existing minority class samples.
SMOTE (Synthetic Minority Oversampling Technique) and variants (e.g., Borderline-SMOTE, ADASYN): Generate synthetic samples based on minority instances.
Undersampling Techniques
Random Undersampling (RUS): Removes samples from the majority class.
NearMiss, K-Medoids: Select specific majority instances to retain.
Hybrid Methods
SMOTE+ENN, SMOTE+Tomek Links: Combine oversampling with instance cleaning or undersampling for noise reduction and improved balance.
2. Algorithm-Level (Cost-Sensitive) Methods
These incorporate imbalance correction directly into the model training process:
Weighted Loss Functions: Assign higher misclassification penalties to the minority class.

Focal Loss: Adjusts learning to focus on hard-to-classify cases.
Custom Penalty Matrices or Threshold Adjustments
These methods improve sensitivity, AUC, and calibration by addressing the data imbalance without altering the original dataset distribution.

Comparator 1. No Resampling (Original Dataset Models) Predictive models were trained directly on the original, imbalanced dataset without balancing intervention.
These serve as baseline models, enabling assessment of the marginal benefit or harm introduced by resampling or cost-sensitive methods.
2. Traditional Statistical Models
Most frequently, logistic regression models are trained without class-balancing.
Chosen due to their widespread use in clinical research and interpretability, logistic regression is a reference standard for comparing the added value of more complex or machine learning models when imbalance correction is applied.
3. Alternative Resampling or Balancing Strategies
When studies compare multiple resampling methods (e.g., SMOTE vs. RUS or SMOTE+ENN vs. ROS), they are evaluated to explore relative effectiveness.
In algorithm-level comparisons, models using cost-sensitive learning may be compared to those using data-level balancing.
Comparative analyses in the meta-regression will assess how model performance varies with and without balancing strategies, across techniques, model types, and dataset characteristics.

Study designs to be included This review includes primary research studies that apply binary classification models to imbalanced clinical datasets and report on the use and impact of resampling or cost-sensitive strategies.Eligible study designs are:Retrospective or prospective observational studies, including cohort or registry-based studies, that develop or validate predictive models.Clinical prediction modeling studies using machine learning or traditional statistical methods (e.g., logistic regression) with a reported imbalance correction strategy.

Eligibility criteria Inclusion Criteria
Studies will be included if they meet all the following conditions:
Population:
Clinical or biomedical datasets with binary outcomes (e.g., disease vs. no disease, event vs. no event) should be included.
Class imbalance is explicitly reported, typically with a minority class representing <30% of the total sample.
Intervention:
Apply at least one resampling or cost-sensitive learning method to address class imbalance.
Includes data-level strategies (e.g., SMOTE, ROS, RUS, hybrid methods) and/or algorithm-level approaches (e.g., weighted losses, focal loss).
Outcomes:
Report at least one model performance metric (e.g., AUC, sensitivity, specificity, F1-score, calibration).
Study Type:
Empirical primary research (retrospective or prospective), or
Systematic reviews/meta-analyses that include re-analysis or meta-estimates on resampling strategies.
Time Frame:
Published between January 2009 and April 30, 2024.
Language:
There is no language restriction at the search level; however, only studies with extractable English data or translatable full texts will be analysed.
Exclusion Criteria
Studies do not involve binary classification tasks (e.g., regression or multi-class classification only).
Studies not reporting a resampling or balancing intervention.
Studies without quantitative model performance metrics.
Simulation-only studies with no real clinical data.
Conference abstracts that lack full methodological or outcome details.
Only the most comprehensive version will be retained for duplicate studies using the same dataset and model.

Information sources The following information sources were used to ensure a comprehensive and reproducible evidence base for this review:
Electronic Databases
The search was conducted in five major bibliographic databases covering both clinical and computational literature:
MEDLINE (via PubMed)
EMBASE
Scopus
Web of Science (Core Collection)
IEEE Xplore
These databases capture peer-reviewed studies across medicine, health informatics, machine learning, and engineering.
Grey Literature
To reduce publication bias and identify recent or ongoing studies:

medRxiv
arXiv
bioRxiv
were screened for eligible preprints and unpublished studies that meet the inclusion criteria.
Supplementary Sources
Backwards and forward citation tracking was performed for all included full-text studies.
Code repositories such as GitHub were examined when referenced in included papers, particularly for resampling method implementations.
Bibliographies of relevant systematic reviews and meta-analyses were also reviewed.
Search Timeline
The final database searches were conducted in February 2025 and included all studies published between January 2009 and December 31, 2024.

Main outcome(s) Area Under the Receiver Operating Characteristic Curve (AUC or AUROC):
Chosen as the main outcome due to its widespread use in binary classification for assessing a model's ability to distinguish between positive and negative cases, independent of class distribution.
Sensitivity (Recall):
This is especially critical in clinical applications where identifying the minority class (e.g., disease or adverse event) is of high importance.
F1-Score (if available):
It balances sensitivity and precision, which is relevant in highly skewed datasets.
These outcomes were included in the meta-regression model, using logit-transformed AUC as the effect size, to quantify the impact of resampling strategies and dataset characteristics on model performance.

Additional outcome(s) In addition to the primary outcomes (AUC, sensitivity, F1-score), the following secondary outcomes will be collected and synthesised where reported:
1. Calibration Measures
Brier Score: A proper scoring rule reflecting both discrimination and calibration.
Calibration Slope or Calibration-in-the-Large: To assess how well predicted probabilities agree with observed outcomes.
These outcomes are critical for evaluating clinical trustworthiness, especially in risk prediction models.
2. Misclassification Costs or Cost-Sensitive Metrics
When studies report cost-based performance measures (e.g., cost matrices, weighted error rates), these will be extracted to understand trade-offs in classifying minority events.
Includes studies applying custom loss functions, focal loss, or weighted decision frameworks.
3. Specificity and Balanced Accuracy
Specificity: Complements sensitivity for evaluating error balance.
Balanced Accuracy: Accounts for class imbalance directly by averaging sensitivity and specificity.
4. External Validation Metrics (if available)
Performance metrics reported on external datasets will be separately noted to assess generalisability.

Data management All references retrieved from the database and supplementary searches were imported into Zotero (version 7.0.15) for citation management and deduplication. Titles and abstracts were initially screened using Zotero and the open-source AI-assisted tool ASReview, facilitating transparent prioritisation and tagging based on inclusion criteria.
Screening and Selection Process
Two reviewers independently screened all titles and abstracts.
Full texts of potentially eligible studies were retrieved and assessed in parallel.
Any disagreements were resolved through discussion or third-party adjudication.
A PRISMA flow diagram documented all screening decisions and reasons for exclusions.
Data Extraction
A standardised data extraction form was developed and piloted on a subset of studies to ensure clarity and consistency.
The extracted fields included study metadata, clinical condition, sample size, imbalance ratio, model type, resampling method, performance metrics, calibration statistics, and cost-sensitive measures.
Data were entered into structured spreadsheet formats (.xlsx/.csv) and analysed using R (particularly the metafor, dplyr, and ggplot2 packages).
Data Storage and Security
All data were stored in institutionally approved secure cloud storage (e.g., OneDrive, institutional Google Drive).
Data integrity was maintained via version control, regular backups, and controlled access rights.

Quality assessment / Risk of bias analysis Given the scoping nature of this review and its focus on methodological strategies (resampling and cost-sensitive learning) rather than clinical interventions or treatment effects, a formal risk of bias assessment was considered optional, in line with the PRISMA-ScR and Joanna Briggs Institute (JBI) guidance.
However, to enhance transparency and rigour, the following steps were taken:

1. Design-Level Screening for Bias Sensitivity
Only primary research studies that reported sufficient methodological details (e.g., dataset characteristics, modeling pipeline, performance metrics) were included.
Studies lacking reproducible methods (e.g., no mention of resampling ratio, model hyperparameters, or outcome definitions) were excluded.
2. Meta-Regression Adjustment
In the meta-analytical component, performance variability was explained using moderators such as sample size, imbalance ratio, and model type.
Influence diagnostics (e.g., Cook's distance, studentised residuals) were applied to detect overly influential or outlier studies.
3. Small-Study Effects and Publication Bias
A funnel plot, Egger's regression test, and trim-and-fill analysis were conducted to detect and adjust for small-study effects or reporting bias.
The Vevea and Hedges weight-function model was also applied to adjust for potential publication bias in effect estimates.
4. Sensitivity Analyses
Studies contributing extreme values or methodological outliers were identified, and their impact on overall findings was assessed through a leave-one-out sensitivity analysis.
5. Transparency in Reporting
For future reproducibility, a structured data extraction table documented each study's inclusion, dataset overlap, and deviations from protocol.

Strategy of data synthesis The data synthesis will be conducted in two phases, combining descriptive mapping with quantitative meta-regression to assess the effectiveness of resampling strategies across diverse clinical prediction studies.
Phase 1: Qualitative and Descriptive Synthesis (Scoping Review)
All included studies will be mapped according to key variables, including:
Clinical domain
Sample size
Imbalance ratio
Resampling method (e.g., SMOTE, ROS, RUS, hybrids)
Model type (e.g., logistic regression, random forest, SVM)
Performance metrics (AUC, sensitivity, F1-score, calibration)
Visual summaries will include:
Temporal trends in resampling usage
Distribution of models across imbalance severities
Comparison of reported performance across sampling groups
Interaction plots (e.g., performance vs. sample size and IR)
These findings will be summarised in figures and tables to highlight empirical patterns, research gaps, and inconsistencies.
Phase 2: Quantitative Synthesis (Meta-Regression)
A random-effects meta-regression model will be performed on a subset of studies that:
Report a standard performance metric (primarily AUC)
Provide sufficient statistical data (e.g., standard errors or 95% CIs)
Details:
Effect size: Logit-transformed AUC
Model: Mixed-effects meta-regression using the metafor package in R
Moderators: Sample size (log scale), imbalance ratio, number of features, resampling method, model type, and clinical condition
Exploratory analyses will include:
Funnel plots, Egger's test, trim-and-fill, and weight-function models to assess small-study and publication bias
Influence diagnostics to detect outliers
Subgroup analyses by resampling group and imbalance severity
This dual synthesis strategy ensures comprehensive coverage of the literature and quantifies how data-level and algorithm-level strategies perform under varying modeling conditions.

Subgroup analysis Subgroup analyses will be conducted to explore sources of heterogeneity in model performance and to assess whether specific resampling strategies or dataset characteristics systematically influence the effectiveness of predictive models trained on imbalanced medical data.
Planned subgroups include:
1. Type of Resampling Strategy
Oversampling (e.g., SMOTE, ROS)
Undersampling (e.g., RUS, NearMiss)
Hybrid approaches (e.g., SMOTE+ENN)
No resampling (original data)
2. Model Type
Traditional statistical models (e.g., logistic regression)
Tree-based models (e.g., random forest, XGBoost)
Neural networks (ANNs)
Support vector machines (SVMs)
Ensemble learners
3. Clinical Domain
Cardiovascular diseases
Cancer
Diabetes
Neuropsychiatric conditions
Infectious diseases

Other or mixed domains
4. Imbalance Severity
Low imbalance (IR < 5)
Moderate imbalance (IR 5–20)
Severe imbalance (IR > 20)
5. Sample Size Categories
Small (N < 500)
Moderate (500 ≤ N < 10,000)
Large (N ≥ 10,000)
Each subgroup will be analysed using stratified summary statistics and meta-regression interaction terms where sufficient data exist. The purpose is to assess whether performance differences are attributable to dataset size, modelling approach, imbalance level, or clinical context.

Sensitivity analysis Sensitivity analyses will be performed to assess the pooled performance estimates' robustness and stability and determine the influence of specific studies or methodological choices on the meta-regression outcomes.
Planned sensitivity analyses include:
1. Influence Diagnostics
Apply Cook's distance, studentised residuals, and hat values to identify influential or outlier studies.
Re-run the meta-regression with and without influential studies to evaluate changes in effect sizes and model heterogeneity (I^2, τ^2).
2. Publication Bias Adjustments
Use trim-and-fill methods to impute potentially missing studies and adjust the pooled AUC.
Apply the Vevea and Hedges weight-function model to test the impact of small-study effects and selective reporting on performance metrics.
3. Subset Analyses
Restrict meta-regression to studies with:
External validation
Reported calibration metrics
Clearly defined resampling ratios or hyperparameters
Compare results to the full dataset to assess bias due to poor reporting or internal validation.
4. Exclusion of Non-Standard Metrics
Exclude studies that report non-AUC outcomes only, or derive performance from non-test datasets, to confirm the consistency of AUC-based inferences.
These analyses will support transparent reporting and enhance the credibility of the quantitative findings.

Language restriction No language restrictions were applied during the initial database searches to ensure broad and inclusive coverage of the literature. Titles and abstracts in all languages were screened.

Country(ies) involved Jordan.

Other relevant information Registration and Reporting Standards
This review protocol follows the PRISMA-P 2015 (Preferred Reporting Items for Systematic review and Meta-Analysis Protocols) and PRISMA-ScR (for scoping reviews) guidelines.
When applicable, the final review will be reported according to the PRISMA 2020 statement and the Meta-Analysis of Observational Studies in Epidemiology (MOOSE) guideline.
Ethics and Dissemination
Ethical approval is not required as the study uses publicly available, non-identifiable secondary data.
Results will be submitted for peer-reviewed publication and presented at relevant health data science and medical informatics conferences.
Software and Tools
Reference management: Zotero v7.0.15
Screening: ASReview (open-source machine learning-aided screening)
Statistical analysis and meta-regression: R (metafor, dplyr, ggplot2)
Visualisation: ggplot2, patchwork, and gridExtra
Team and Roles
Review conception and protocol drafting: Osama Abdelhay
Screening and data extraction: Osama Abdelhay, Adam Shatnawi
Meta-regression and statistical analysis: Osama Abdelhay
Interpretaion and methodological validation: Hassan Najadat
All reviewers contributed to protocol development and approved the final version.
Funding and Conflicts of Interest
This review is unfunded.
The authors declare no competing interests.

Keywords Class imbalance; Resampling methods; SMOTE; Undersampling; Cost-sensitive learning; Medical prediction models; Binary classification; Machine learning; Meta-regression; Systematic review; Scoping review.

Dissemination plans Peer-Reviewed Publication:
The completed review and meta-regression will be submitted to a high-impact journal in the fields of medical informatics, machine learning in healthcare, or evidence-based medicine (e.g., Journal of Biomedical Informatics, BMC Medical Informatics and Decision Making, BMJ Health & Care Informatics).
Open Access Repository:
Preprint versions will be shared on medRxiv or arXiv to allow early access and transparency.
Code and Data Sharing:
Analytical scripts (e.g., R code for meta-regression and visualisation) and the curated dataset of

included studies will be made available via GitHub and linked to the final publication, subject to license compatibility and journal policies.
Policy and Practice Outreach:
A plain-language summary will be developed and shared with clinical AI researchers and practitioners, highlighting actionable recommendations for handling class imbalance in real-world prediction modeling.

Contributions of each author

Author 1 - Osama Abdelhay - Conceptualisation, Methodology, Writing – Original Draft, Supervision.
Email: o.abdelhay@psut.edu.jo
Author 2 - Adam Shatnawi - Conceptualisation, Methodology, Investigation, Writing – Original Draft, Validation.
Email: azshatnawi20@cit.just.edu.jo
Author 3 - Hasan Najadat - Writing – Review & Editing, Supervision, Project Administration.
Email: najadat@just.edu.jo
Author 4 - Taghreed Altamimi - Writing – Review & Editing, Validation, Resources.
Email: taaltamimi@alfaisal.edu