QUBAHAN JOURNAL OF SOCIAL SCIENCE
VOL. 3, NO. 4, November 2026
https://doi.org/10.48161/qjss.v3n4a16

A Robust Probabilistic Framework for Handling Missing and Noisy Data in Empirical Research

Koulla Parpa[1*]

[1] Department of Sport Science, Faculty of the School of Sciences, University of Central Lancashire, School of Sciences, Pyla, Cyprus;

* **Corresponding author:** kparpa@uclan.ac.uk.

ABSTRACT: Missing and noisy data remain major threats to the validity, reliability, and interpretability of empirical research. Incomplete observations, measurement error, unstable records, and contaminated predictors can distort parameter estimates, weaken predictive performance, and produce misleading conclusions. This study develops and tests a robust probabilistic framework for handling missing and noisy data in empirical research. The framework integrates data-quality diagnosis, probabilistic recovery, likelihood-based estimation, robust probability weighting, and uncertainty reporting. The empirical design uses two public real-world datasets as separate validation cases: 4,240 healthcare records for ten-year coronary heart disease prediction and 7,043 business records for customer churn prediction. We first profiled native missingness, then introduced controlled missingness and noise scenarios into the training data while keeping the test data fixed. We compared complete-case analysis, median/mode probability modeling, iterative likelihood imputation, and robust probability modeling using the area under the receiver operating characteristic curve (ROC-AUC), accuracy, F1 score, Brier score, log loss, predictive entropy, and retained training sample size. The results show that complete-case analysis becomes unstable when data degradation increases, retaining only 11 healthcare training records under the strongest degradation setting and failing to provide a stable complete-case model in the corresponding business scenario. Probability-based approaches preserved discrimination as missingness and noise increased, with ROC-AUC values remaining around 0.70 in the healthcare case and around 0.83 in the business case. The robust probability model produced the strongest calibration evidence in the business robustness case, while iterative likelihood modeling improved probability quality in the healthcare case under moderate and severe degradation. The study contributes an empirical framework for uncertainty-aware analysis of imperfect datasets.

Keywords: missing data; noisy data; probability models; probabilistic imputation; healthcare.

I. INTRODUCTION

Empirical research assumes that observed data provide a sufficiently accurate representation of the phenomenon under investigation. In practice, this assumption often weakens due to missing values, measurement error, unstable responses, outliers, inconsistent coding, and noise introduced during data collection or processing. These problems are not limited to one field. They appear in healthcare records, educational assessments, social surveys, business transactions, public administration databases, and many other sources used for applied statistical analysis. Ignoring missingness and noise can bias statistical results, make them unstable, and make them difficult to interpret.

A common response to incomplete data is to remove records with missing values or to replace missing cells using simple deterministic rules. These procedures are convenient but can be analytically fragile. Complete-case analysis reduces the sample size and may change the study population's composition, especially when missingness is concentrated in clinically, socially, or economically meaningful variables.

Simple replacement rules preserve the number of rows but may underestimate uncertainty and distort relationships among variables. Noisy observations create a related problem. They may remain hidden inside apparently complete datasets and can influence model estimates, classification boundaries, and probability calibration.

Probability-based modeling provides a stronger basis for handling imperfect data because it treats missingness and noise as sources of uncertainty rather than as isolated preprocessing defects. Instead of assuming that a single repaired dataset is fully known, probability models estimate plausible values, represent uncertainty around incomplete observations, and compare prediction quality under controlled degradation. This orientation matters for empirical research because the goal is not only to obtain an accurate prediction, but also to generate evidence that remains reliable when data are incomplete or unstable.

This study proposes and evaluates a robust probabilistic framework for handling missing and noisy data in empirical research. The framework combines data-quality diagnosis, probabilistic imputation, likelihood-based estimation, robust probability weighting, and uncertainty reporting. We test it on two real-world empirical datasets rather than relying only on simulated data. The primary case uses a healthcare prediction problem, while the secondary robustness case uses a business classification problem. This two-case design allows us to evaluate whether the proposed framework remains useful across different variable structures, outcome distributions, and missingness patterns.

The study is guided by three research questions. First, how do missing and noisy observations affect the stability of conventional empirical models? Second, do probability-based methods preserve predictive discrimination and inference quality better than complete-case analysis under increasing data degradation? Third, does robust probability weighting improve model reliability when missingness and noise occur together? The study tests three empirical hypotheses: H1, complete-case analysis becomes less stable as missingness and noise increase; H2, probabilistic imputation and likelihood-based estimation preserve prediction quality better than record deletion; and H3, robust probability modeling improves calibration and robustness when noisy observations are present.

II. LITERATURE REVIEW

The literature first shows that imperfect datasets may include both incompleteness and contamination. Robust estimation research shows that cellwise and casewise contamination can bias multivariate estimates when researchers treat unreliable observations as ordinary records [1]. Multilevel imputation research indicates that missing values in continuous and binary variables require methods aligned with data structure, variable type, and dependence among records [2]. Substantive-model-compatible imputation further shows that the imputation procedure should preserve the relationships required by the final analytic model [3].

Indicator-augmented imputation and applied statistical frameworks extend this view by showing that missingness may contain useful information but can also represent a data-generation process rather than a substantive effect [4]. Practical statistical guidance emphasizes connecting assumptions about missingness, model fitting, and interpretation within a transparent analytical workflow [5]. Prediction-model reporting standards similarly require clear documentation of data sources, missing-data handling, candidate predictors, validation, and interpretation [6].

Computational studies show that missing-data treatment in predictive modeling is sensitive to the structure of the dataset and the objective of the model [7]. Healthcare-focused reviews confirm that imputation methods matter especially when incomplete clinical values affect risk estimation and decision-making [8]. Research on systematically missing predictors also shows that empirical studies combining information from heterogeneous sources need generalized recovery strategies [9].

Foundational missing-data guidance treats incomplete observations as a substantive threat to empirical validity rather than a minor technical inconvenience [10]. Evaluations of predictive mean matching under violated distributional assumptions show that imputation can remain useful when the method respects the empirical shape of the data [11]. Contemporary reporting frameworks further emphasize that observational studies should disclose missingness assumptions, treatment decisions, diagnostics, and sensitivity procedures [12].

Bayesian and likelihood-oriented approaches provide a coherent probability basis for incomplete empirical data. Missing-not-at-random modeling highlights the importance of identifying restrictions when missingness may depend on unobserved values [13]. Clinical-data evaluations show that no universally best imputation method exists and that benchmarking should be tied to realistic applied conditions [14]. Broader Bayesian reviews show that one uncertainty-aware inferential system can estimate missing values, parameters, and predictions [15].

Small-sample research adds an important caution because missing-data methods may behave differently when the available evidence is limited or when missingness occurs arbitrarily across variables [16]. Transparent prediction-model guidance reinforces the need to report missing-data procedures, development steps, validation decisions, and interpretation rules [17]. Work on multiple imputation tuning shows that predictive mean matching and local residual draws can improve plausibility when distributional assumptions are imperfect [18].

Practical missing-data guidelines emphasize that arbitrary deletion may be defensible only under restrictive assumptions and that the amount, pattern, and treatment of missingness should be reported [19]. Clinical epidemiological guidance supports multiple imputation as more informative than deletion when assumptions and diagnostics are made explicit [20]. Multilevel chained-equation research further shows that sporadic and systematic missingness may require models that reflect hierarchical or structured data conditions [21].

Noisy data create a related but distinct threat to empirical inference. Research on robust statistics shows that anomalous observations can influence estimates and lead to misleading conclusions when treated as reliable records [22]. Work on deviating data cells shows that contamination can occur within individual cells as well as across complete cases, supporting methods that reduce the influence of unstable observations without discarding all available evidence [23].

Systematic reviews of computational imputation methods show that performance depends on missingness level, variable structure, modeling objective, and the evaluation metric used [24]. Calibration research demonstrates that a model can appear useful by discrimination or accuracy while still producing probabilities that are too high or too low for empirical decision-making [25]. This supports evaluating the Brier score, log loss, and uncertainty measures rather than relying only on classification accuracy.

Risk-model calibration frameworks show that predicted probabilities should be assessed by their agreement with observed outcomes, not only by their ability to rank cases [26]. Taken together, the literature supports the present study's integrated framework, which treats missingness and noise as connected sources of uncertainty and evaluates competing strategies through discrimination, probability error, calibration, retained evidence, and sensitivity to controlled degradation.

Table 1. Literature-review evidence matrix.

Ref.	Study	Domain focus	Methodological focus	Relevance to current study
[1]	Agostinelli, Leung, Yohai, and Zamar (2015)	Robustness and model reporting	Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination	Supports robust estimation under combined cellwise and casewise contamination.
[2]	Audigier etal. (2018)	Imputation and likelihood modeling	Multiple imputation for multilevel data with continuous and binary variables	Supports structured imputation across variable types and dependent records.
[3]	Bartlett et.al (2015)	Missing-data foundations	Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model	Supports alignment between imputation and the final empirical model.
[4]	Beesley etal. (2021)	Applied/computational imputation	Multiple imputation with missing data indicators	Supports careful interpretation of missingness indicators in prediction settings.

Ref.	Study	Domain focus	Methodological focus	Relevance to current study
[5]	Carpenter and Smuk (2021)	Imputation and likelihood modeling	Missing data: A statistical framework for practice	Supports transparent missing-data assumptions, modeling, and interpretation.
[6]	Collins, et.al (2015)	Robustness and model reporting	Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis: The TRIPOD statement	Supports transparent reporting of predictors, missing-data handling, and validation.
[7]	Emmanuel et al. (2021)	Applied/computational imputation	A survey on missing data in machine learning	Supports benchmarking under different data structures and modeling objectives.
[8]	Ismail, Abidin, and Maen (2022)	Applied/computational imputation	Systematic review on missing data imputation techniques with machine learning algorithms for healthcare	Supports healthcare-focused comparison of imputation strategies.
[9]	Jolani et al. al (2015)	Missing-data foundations	Imputation of systematically missing predictors in an individual participant data meta-analysis: A generalized approach using MICE	Supports recovery strategies for systematically missing predictors.
[10]	Kang (2013)	Missing-data foundations	The prevention and handling of missing data	Supports explicit recognition of missingness as a validity threat.
[11]	Kleinke (2017)	Imputation and likelihood modeling	Multiple imputation under violated distributional assumptions: A systematic evaluation of the assumed robustness of predictive mean matching	Supports robust imputation when distributional assumptions are imperfect.
[12]	Lee etal. (2021)	Imputation and likelihood modeling	Framework for the treatment and reporting of missing data in observational studies: The TARMOS framework	Supports reporting of assumptions, diagnostics, and sensitivity decisions.
[13]	Linero and Daniels (2018)	Imputation and likelihood modeling	Bayesian approaches for missing not at random outcome data: The role of identifying restrictions	Supports probabilistic handling of non-ignorable missingness.
[14]	Luo (2022)	Applied/computational imputation	Evaluating the state of the art in missing data imputation for clinical data	Supports realistic clinical-data benchmarking of imputation methods.
[15]	Ma and Chen (2018)	Imputation and likelihood modeling	Bayesian methods for dealing with missing data problems	Supports joint uncertainty modeling of missing values and parameters.
[16]	McNeish (2017)	Missing-data foundations	Missing data methods for arbitrary missingness with small samples	Supports caution when missingness occurs in limited samples.
[17]	Moons et al. al (2015)	Robustness and model reporting	Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis: Explanation and elaboration	Supports detailed explanation of prediction-model reporting decisions.
[18]	Morris, et.al (2014)	Missing-data foundations	Tuning multiple imputation by predictive mean matching and local residual draws	Supports improved imputation plausibility under imperfect distributions.
[19]	Newman (2014)	Missing-data foundations	Missing data: Five practical guidelines	Supports reporting of amount, pattern, and treatment of missing data.
[20]	Pedersen et al. (2017)	Missing-data foundations	Missing data and multiple imputation in clinical epidemiological research	Supports explicit assumptions and diagnostics in clinical missing-data analysis.

Ref.	Study	Domain focus	Methodological focus	Relevance to current study
[21]	Resche-Rigon and White (2018)	Imputation and likelihood modeling	Multiple imputation by chained equations for systematically and sporadically missing multilevel data	Supports imputation for structured sporadic and systematic missingness.
[22]	Rousseeuw and Hubert (2018)	Robustness and model reporting	Anomaly detection by robust statistics	Supports robust treatment of anomalous and unstable observations.
[23]	Rousseeuw and Van den Bossche (2018)	Robustness and model reporting	Detecting deviating data cells	Supports cell-level detection of contamination in empirical datasets.
[24]	Thomas and Rajabi (2021)	Applied/computational imputation	A systematic review of machine learning-based missing value imputation techniques	Supports comparison of computational imputation methods across conditions.
[25]	Van Calster, et.al (2019)	Imputation and likelihood modeling	Calibration: The Achilles heel of predictive analytics	Supports probability-quality evaluation beyond classification accuracy.
[26]	Van Calster, et al. (2016)	Robustness and model reporting	A calibration hierarchy for risk models was defined: From utopia to empirical data	Supports calibration-based assessment of empirical risk probabilities.

III. RESEARCH GAP

The reviewed literature provides strong guidance on missing-data mechanisms, imputation, Bayesian inference, robustness, calibration, and reporting. However, three gaps remain important for empirical researchers. First, missing data and noisy data are often handled as separate problems, although real datasets frequently contain both. Second, many applied studies still evaluate missing-data methods using one final repaired dataset, without examining how performance changes under increasing degradation. Third, researchers often judge probability-based models mainly by discrimination or accuracy, even though uncertainty, calibration, probability error, and retained evidence are central to reliable interpretation.

This study addresses these gaps by developing an integrated framework that treats missingness and noise as sources of uncertainty. We evaluate the framework on two real empirical datasets using native missingness and controlled degradation. The aim is not to claim that one method is universally superior, but to show how different probability-based strategies behave when the same data are exposed to increasing missingness and measurement noise. This design connects methodological principles to observable empirical consequences.

IV.STUDY CONTRIBUTION

The contribution of this study is methodological and empirical. Methodologically, it presents a compact framework that researchers working with imperfect observational datasets can implement. Empirically, it demonstrates the framework on real healthcare and business datasets with native missingness and controlled degradation scenarios. The paper therefore moves beyond general recommendations about data cleaning by showing how model performance, uncertainty, and retained evidence change across competing missing-data and noise-handling strategies.

V. PROPOSED METHOD

The proposed method is a robust probabilistic framework for empirical datasets affected by incomplete and noisy observations. It consists of four linked stages: data-quality diagnosis, probabilistic recovery, noise-resistant estimation, and uncertainty reporting. The framework begins by diagnosing native missingness, variable type, class balance, and duplicate records. It then estimates incomplete values using probability-based recovery strategies. After recovery, models are fitted under conventional and robust probability specifications. Finally, it reports results using discrimination, probability error, uncertainty, and sensitivity measures.

The framework deliberately avoids treating missing-value imputation as a single repair operation. Instead, it evaluates how predictions change when missingness and noise increase. This makes the framework useful for empirical research because it shows whether conclusions are stable under plausible data-quality deterioration. The approach is compatible with healthcare, education, social science, and business datasets because it requires only tabular records with an empirical outcome variable.

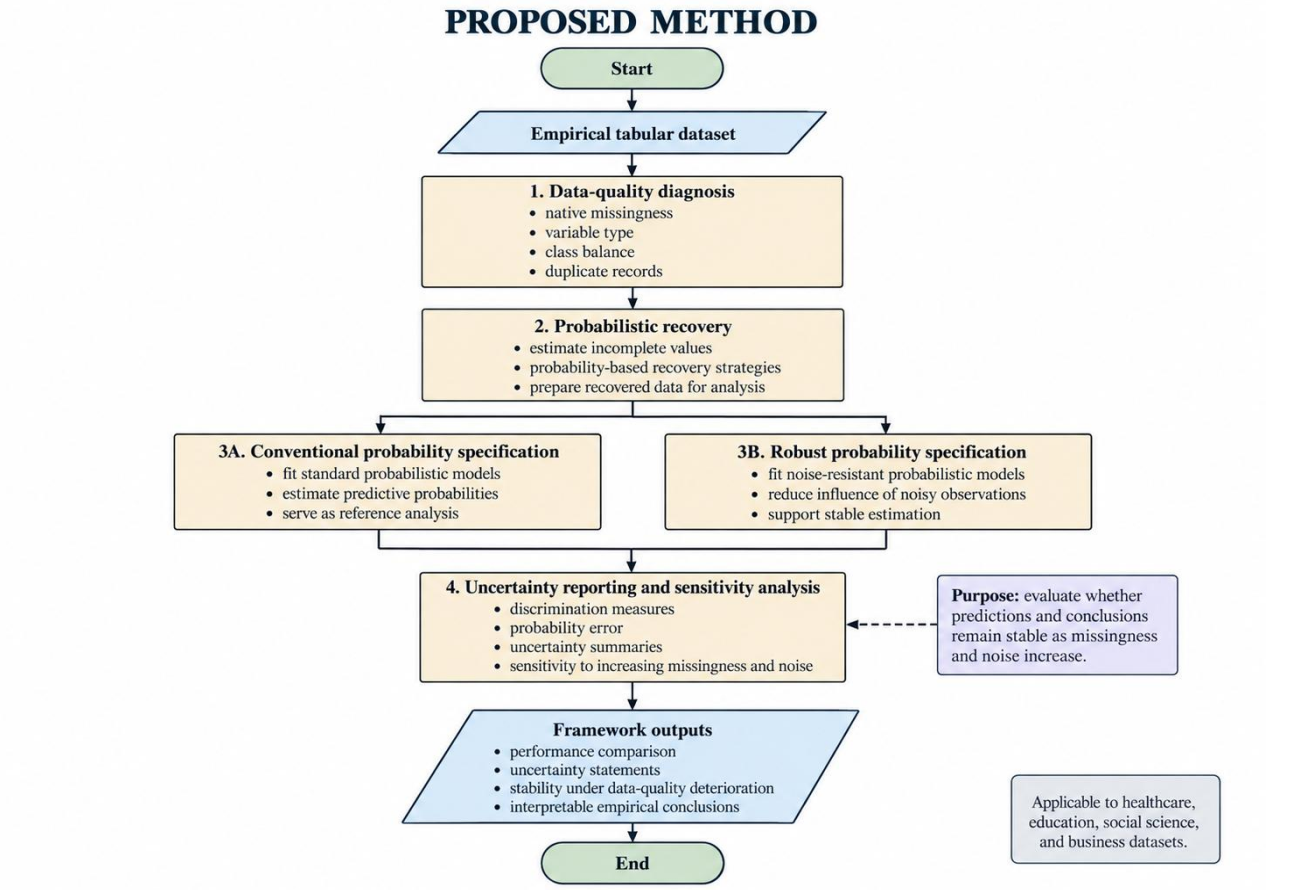


FIGURE 1. Proposed robust probabilistic framework for missing and noisy empirical data.

1. PROBABILITY FORMULATION

Let Y represent the empirical outcome, X represent observed predictors, and R represent the missingness indicator, where R = 1 denotes an observed value and R = 0 denotes a missing value. Missingness is represented as a probability mechanism rather than as a purely mechanical defect:

$$P(R \mid Y_{obs}, Y_{mis}, X, \theta_r). \qquad (1)$$

The predictive distribution of missing values is then estimated conditional on observed information and model uncertainty:

$$p(Y_{mis} \mid Y_{obs}, X) = \int p(Y_{mis} \mid Y_{obs}, X, \theta)\, p(\theta \mid Y_{obs}, X)\, d\theta. \qquad (2)$$

For binary empirical outcomes, the likelihood-based classification model can be written as:

$$L(\beta) = \prod_i [\pi_i(\beta)]^{y_i} [1 - \pi_i(\beta)]^{1-y_i}, \quad \pi_i(\beta) = P(Y_i = 1 \mid X_i, \beta). \qquad (3)$$

To reduce the influence of noisy records, the robust probability model applies observation weights derived from standardized deviation scores d_i:

$$w_i = \frac{1}{1 + \lambda d_i}, \quad 0 < w_i \leq 1. \qquad (4)$$

The robust objective minimizes weighted probability loss rather than treating every observation as equally reliable:

$$\hat{\beta}_r = \arg\min_{\beta} \sum_i w_i \, \ell(y_i, \pi_i). \qquad (5)$$

Predictive uncertainty is summarized using binary predictive entropy:

$$H(p_i) = -p_i \log(p_i) - (1 - p_i)\log(1 - p_i). \qquad (6)$$

VI. METHODOLOGY

1. RESEARCH DESIGN

The study used a secondary quantitative empirical design with controlled data-degradation experiments. The empirical strategy compared four missing-data and noise-handling approaches across native and degraded versions of two real-world datasets. The healthcare dataset served as the main case because it contains native missingness in clinically meaningful variables and a binary risk outcome. The business dataset served as a robustness case because it contains a larger number of records, mixed categorical and numerical predictors, and a binary churn outcome. We did not merge the datasets because they belong to different empirical domains and do not share common identifiers, units, or measurement structures.

2. DATASET SCREENING AND SELECTION

We screened candidate datasets from healthcare, education, and business domains based on sample size, outcome clarity, native missingness, variable diversity, duplicate records, and relevance to the study title. We selected the healthcare and business datasets because they allow meaningful evaluation of probabilistic recovery and robust probability estimation under missingness and noise. Education and smaller medical datasets were not selected as main cases because they offered too little native missingness or insufficient empirical complexity for the present design.

Table 2. Characteristics of the selected empirical datasets.

Dataset	Rows	Columns	Outcome	Positive share	Missing cells	Variables missing	Duplicates
Framingham healthcare	4240	16	TenYearCHD	15.19%	645	7	0
Telco business robustness	7043	21	Churn	26.54%	8261	7	0

3. VARIABLES AND OUTCOME MEASURES

In the healthcare case, the dependent variable was ten-year coronary heart disease risk. Predictors included demographic, behavioral, and clinical measurements. In the business robustness case, the dependent variable was customer churn. Predictors included customer characteristics, service attributes, contract information, and billing variables. The modeling pipeline encoded categorical variables, and we retained numerical variables as continuous predictors after imputation and scaling where appropriate.

4. MISSINGNESS AND NOISE SCENARIOS

The empirical experiment used native missingness and three controlled degradation scenarios. We introduced missingness and numeric noise only into the training data, while keeping the test data fixed. This

design keeps the evaluation target constant across scenarios and lets us compare how each modeling strategy responds to degraded training evidence.

Table 3. Experimental missingness and noise scenarios.

Code	Scenario	Design	Purpose
S0	Native data	Original missingness only	Observed data-quality condition
S1	10% missing + 5% noise	Moderate additional missingness and low numeric noise	Mild empirical degradation
S2	20% missing + 10% noise	Substantial additional missingness and medium numeric noise	Moderate degradation
S3	30% missing + 20% noise	Severe additional missingness and high numeric noise	Stress-test condition

5. COMPARATIVE MODELS

Four comparative methods were evaluated. The complete-case model used only records without missing values after degradation. The median/mode probability model replaced missing numerical variables with medians and missing categorical variables with modes before fitting a probabilistic classifier. The iterative likelihood model used iterative imputation for numerical predictors and mode-based treatment for categorical predictors before fitting the same probability model. The robust probability model used imputed predictors but applied weights that reduced the influence of highly deviant observations.

The comparison was designed to separate three empirical questions: how much evidence is lost when incomplete records are removed; whether simple probabilistic recovery is sufficient to preserve discrimination; and whether robust probability weighting improves probability quality when missingness and noise occur together.

6. EVALUATION METRICS

We evaluated model performance using AUC, accuracy, F1 score, Brier score, log loss, predictive entropy, and retained training sample size. AUC measures ranking discrimination; accuracy and F1 summarize classification behavior; the Brier score and log loss assess probability quality; entropy summarizes predictive uncertainty; and retained training sample size indicates how much evidence remains after each missing-data strategy. Lower Brier score and log loss indicate better probability quality, while higher AUC and F1 indicate stronger discrimination and classification balance.

7. VALIDITY, RELIABILITY, AND REPRODUCIBILITY CONTROLS

The analysis used a fixed train-test split within each dataset and applied the same scenarios consistently across models. We did not degrade the test set, preserving a stable evaluation target. We controlled missingness and noise at predefined levels to make scenario comparisons interpretable. Using two datasets from different empirical domains provides a robustness check without forcing incompatible datasets into a merged structure.

VII. RESULTS

1. NATIVE MISSINGNESS PROFILES

The selected datasets showed different missingness structures. In the healthcare case, missingness was concentrated mainly in glucose, education, medication status, total cholesterol, smoking intensity, body mass index, and heart rate. In the business robustness case, missingness was larger and concentrated in tenure, monthly charges, streaming service status, internet service, partner status, gender, and total charges. These profiles show why the two datasets are useful for the study: the healthcare case tests clinically meaningful missingness, whereas the business case tests a broader and heavier missingness structure.

Table 4. Observed variable-level missingness in the selected datasets.

Dataset	Variable	Missing values	Missing share
Healthcare	glucose	388	9.15%
Healthcare	education	105	2.48%
Healthcare	BPMeds	53	1.25%
Healthcare	totChol	50	1.18%
Healthcare	cigsPerDay	29	0.68%
Healthcare	BMI	19	0.45%
Healthcare	heartRate	1	0.02%
Business robustness	tenure	2500	35.50%
Business robustness	MonthlyCharges	1500	21.30%
Business robustness	StreamingTV	1500	21.30%
Business robustness	InternetService	1000	14.20%
Business robustness	Partner	1000	14.20%
Business robustness	gender	750	10.65%
Business robustness	TotalCharges	11	0.16%

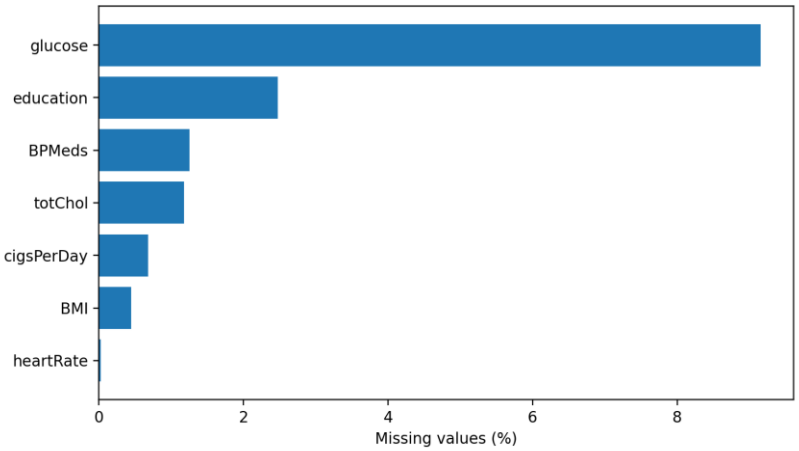


FIGURE 2. Observed missingness profile in the healthcare dataset.

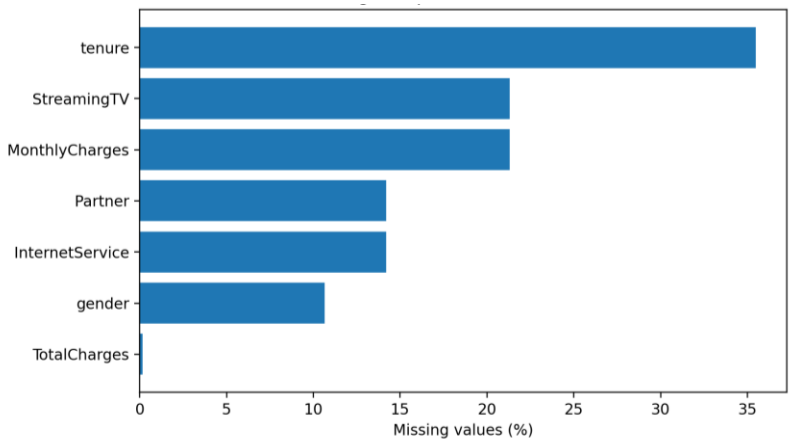


FIGURE 3. Observed missingness profile in the business robustness dataset.

2. HEALTHCARE CASE: FRAMINGHAM PREDICTION RESULTS

Table 5. shows the performance of the four models in the healthcare case. The complete-case model performed reasonably under native and mild degradation conditions, but it became unstable when missingness increased. Under the strongest degradation setting, only 11 complete training records remained, and AUC fell to 0.497, indicating near-random discrimination. Probability-based methods preserved the full training sample and maintained AUC values close to 0.70 across all scenarios. The iterative likelihood model produced better Brier scores than the simple probability model under 20% and 30% degradation, while the robust probability model produced the strongest F1 score under the most severe condition.

Table 5. Healthcare case performance under native and degraded data conditions.

Scenario	Method	Training N	AUC	Accuracy	F1	Brier	Log loss
Native data	Complete-case	2547	0.699	0.671	0.350	0.209	0.606
Native data	Median/mode probability	2968	0.700	0.672	0.361	0.211	0.611
Native data	Iterative likelihood	2968	0.700	0.672	0.361	0.211	0.611
Native data	Robust probability	2968	0.700	0.647	0.361	0.223	0.639
10% missing + 5% noise	Complete-case	500	0.699	0.697	0.362	0.198	0.584
10% missing + 5% noise	Median/mode probability	2968	0.702	0.651	0.355	0.221	0.637
10% missing + 5% noise	Iterative likelihood	2968	0.700	0.669	0.363	0.212	0.614
10% missing + 5% noise	Robust probability	2968	0.700	0.644	0.361	0.224	0.643
20% missing + 10% noise	Complete-case	99	0.548	0.663	0.241	0.238	0.818
20% missing + 10% noise	Median/mode probability	2968	0.704	0.638	0.366	0.234	0.669
20% missing + 10% noise	Iterative likelihood	2968	0.699	0.656	0.360	0.215	0.621
20% missing + 10% noise	Robust probability	2968	0.699	0.634	0.362	0.227	0.649
30% missing + 20% noise	Complete-case	11	0.497	0.714	0.195	0.210	0.788
30% missing + 20% noise	Median/mode probability	2968	0.701	0.608	0.358	0.245	0.698
30% missing + 20% noise	Iterative likelihood	2968	0.698	0.654	0.364	0.220	0.634
30% missing + 20% noise	Robust probability	2968	0.700	0.638	0.369	0.229	0.654

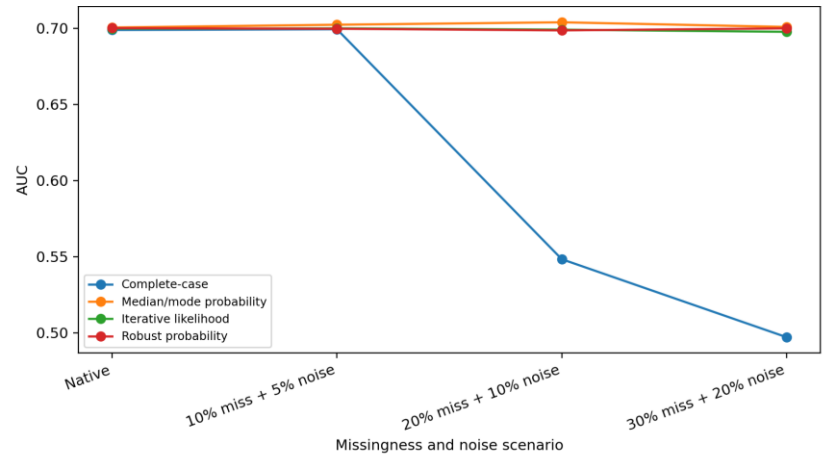


FIGURE 4. AUC stability in the healthcare case under increasing missingness and noise.

3. BUSINESS ROBUSTNESS CASE: TELCO PREDICTION RESULTS

Table 6. presents the business robustness results. In this case, the robust probability model produced the strongest overall evidence. It had the highest AUC and the lowest Brier score across all scenarios. Complete-case analysis again became fragile as degradation increased. At 20% missingness and 10% noise, the complete-case model retained only 45 training records and AUC fell to 0.757. Under the strongest degradation setting, no stable complete-case model was available. In contrast, probability-based methods retained the full training sample and kept AUC values around 0.83.

Table 6. Business robustness case performance under native and degraded data conditions.

Scenario	Method	Training N	AUC	Accuracy	F1	Brier	Log loss
Native data	Complete-case	3150	0.808	0.708	0.588	0.186	0.541
Native data	Median/mode probability	4930	0.830	0.727	0.604	0.174	0.512
Native data	Iterative likelihood	4930	0.834	0.730	0.610	0.172	0.507
Native data	Robust probability	4930	0.835	0.747	0.614	0.161	0.478
10% missing + 5% noise	Complete-case	443	0.817	0.725	0.606	0.181	0.537
10% missing + 5% noise	Median/mode probability	4930	0.830	0.739	0.611	0.167	0.493
10% missing + 5% noise	Iterative likelihood	4930	0.833	0.742	0.615	0.166	0.490
10% missing + 5% noise	Robust probability	4930	0.833	0.757	0.618	0.156	0.465
20% missing + 10% noise	Complete-case	45	0.757	0.719	0.543	0.190	0.571
20% missing + 10% noise	Median/mode probability	4930	0.828	0.745	0.606	0.162	0.479
20% missing + 10% noise	Iterative likelihood	4930	0.829	0.744	0.607	0.162	0.480
20% missing + 10% noise	Robust probability	4930	0.829	0.765	0.613	0.153	0.458
30% missing + 20% noise	Median/mode probability	4930	0.828	0.761	0.618	0.157	0.469
30% missing + 20% noise	Iterative likelihood	4930	0.829	0.760	0.616	0.157	0.467

Scenario	Method	Training N	AUC	Accuracy	F1	Brier	Log loss
30% missing + 20% noise	Robust probability	4930	0.829	0.771	0.614	0.150	0.450

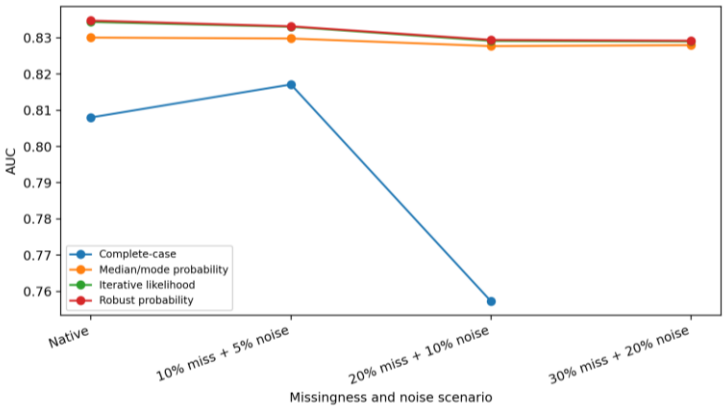


FIGURE 5. AUC stability in the business robustness case under increasing missingness and noise.

4. SUMMARY OF EMPIRICAL FINDINGS

- Result 1. Complete-case analysis was the least stable strategy as degradation increased. It reduced the effective training sample sharply and became unusable or nearly random under severe conditions.
- Result 2. Probability-based methods preserved training evidence and maintained more stable discrimination than record deletion. This pattern appeared in both empirical domains.
- Result 3. In the healthcare case, iterative likelihood modeling improved probability quality under moderate and severe degradation when compared with simple median/mode probability modeling.
- Result 4. In the business robustness case, the robust probability model produced the best Brier score and highest AUC across all scenarios, showing strong calibration and discrimination under missingness and noise.
- Result 5. The two-case design supports the framework's usefulness across different empirical structures without merging incompatible datasets. The framework is therefore better interpreted as a transferable analytical procedure than as a dataset-specific model.

VIII. DISCUSSION

The findings support the study's central argument: missingness and noise should be modeled as sources of uncertainty rather than treated only as routine preprocessing defects. Complete-case analysis appeared acceptable under relatively mild conditions but became unstable as we introduced additional missingness and noise. This instability was not merely a small performance drop; it changed the amount of available evidence. In the healthcare case, the complete-case model retained only 11 training records under the strongest degradation condition. In the business robustness case, the corresponding complete-case model could not be fitted as a stable comparison. These results show why record deletion can mislead empirical research. It may produce a model, but the model may no longer represent the original evidence base.

Probability-based recovery methods produced more stable results because they preserved the analytic sample and allowed the models to learn from partially observed records. In the healthcare case, simple probability modeling, iterative likelihood imputation, and robust probability modeling all maintained AUC values close to 0.70 across degradation levels. This does not mean that the models were unaffected by degradation. Probability error, log loss, and classification behavior changed across scenarios. However, the core discrimination signal was preserved much better than under complete-case analysis. This matters for empirical researchers because it shows that uncertainty-aware recovery can preserve useful evidence even when the dataset is imperfect.

The healthcare and business cases also showed that no single method dominates every empirical situation. In the healthcare case, the simple probability model produced the highest AUC across scenarios, while the iterative likelihood model produced stronger Brier evidence under moderate and severe

degradation. The robust probability model improved F1 under the strongest degradation condition but did not uniformly improve Brier score in this case. This pattern suggests that robust weighting is not automatically superior; its value depends on the structure of contamination, the outcome distribution, and the relationship between noisy values and the decision boundary. A good empirical workflow should therefore compare several probability-based strategies instead of choosing one method by default.

The business robustness case gave stronger support to robust probability modeling. The robust probability model achieved the highest AUC and lowest Brier score in all business scenarios. This result is meaningful because the business dataset had heavier native missingness and a larger mixture of categorical and numerical variables. Robust probability weighting appears to have reduced the influence of unstable observations while preserving enough information for discrimination and probability calibration. The result does not prove universal superiority, but it shows that robust probability modeling can be highly useful when missingness and noise occur together in a heterogeneous empirical dataset.

The Brier-score results in the healthcare case require careful interpretation. In some healthcare scenarios, the complete-case model produced apparently favorable Brier scores despite weak evidence or poor discrimination. This is a known risk in imbalanced classification settings: a probability score may look acceptable when the model predicts close to the majority class, even though the model is not strongly discriminating the outcome. For this reason, the study interprets the Brier score alongside AUC, F1 score, log loss, entropy, and the retained training sample size. The combined evidence is more informative than any single metric.

The study also demonstrates the importance of using separate validation cases rather than merging incompatible datasets. The selected healthcare and business datasets differ in sample size, variable type, outcome prevalence, and missingness structure. Merging them would produce a technically artificial dataset with no meaningful shared unit of analysis. Treating them as separate empirical cases is more scientifically defensible and strengthens the study by showing whether the framework transfers across domains. The evidence suggests that the framework is transferable, even though the best-performing method may differ by dataset.

For empirical research practice, the main lesson is that data-quality treatment should be transparent, comparative, and uncertainty-aware. A complete analysis should report what was missing, how missingness was handled, how noisy observations were treated, how much evidence was retained, and whether conclusions remained stable under degradation. This form of reporting allows readers to assess the credibility of the findings rather than simply trusting that the dataset was cleaned. It also supports better decision-making in applied fields where conclusions may influence healthcare risk assessment, educational interventions, business retention strategies, or social policy interpretation.

Overall, the results support the proposed framework as a practical approach for robust empirical modeling. The framework does not replace field-specific judgment or study design, but it provides a structured method for evaluating incomplete and noisy datasets. Its value lies in connecting missing-data treatment, noise resistance, probability calibration, and sensitivity analysis within a single workflow.

IX. PRACTICAL AND POLICY IMPLICATIONS

For empirical researchers, the framework provides a clear workflow for analyzing imperfect datasets. Researchers should first profile missingness, identify variables most affected by incompleteness, and examine whether noise may influence model estimates. They should then compare deletion, simple probabilistic recovery, likelihood-based recovery, and robust probability modeling rather than relying on a single default method. Reporting should include retained sample size, performance stability, and probability quality.

For healthcare analysts, the results show that removing incomplete clinical variables can weaken risk prediction if you don't account for the loss of evidence. For business researchers, the robustness case shows that probability weighting can improve churn prediction under heavy missingness and mixed variable structures. For education and social science researchers, the same framework can be adapted to survey, assessment, and administrative datasets where incomplete responses and measurement noise are common.

For policy-oriented empirical studies, the framework encourages cautious interpretation. Decisions based on incomplete or noisy data should include evidence about sensitivity to missingness, probability calibration,

and uncertainty. This is especially important when empirical findings inform resource allocation, risk-group identification, or intervention design.

X. LIMITATIONS AND FUTURE RESEARCH

The study has several limitations. First, the empirical analysis used two main datasets, which supports cross-domain robustness but does not cover every empirical field. Second, the controlled degradation scenarios focused on missing-completely-at-random mechanisms and numeric noise injection. Future research should extend the framework to missing-at-random and missing-not-at-random mechanisms, as well as to categorical misclassification and systematic measurement error. Third, the robust probability model uses a transparent weighting strategy, but alternative robust loss functions and Bayesian robust models may further improve performance.

Future studies can expand the framework in three directions. First, the methodological direction involves comparing additional imputation models, posterior predictive methods, and robust estimators. The second direction is empirical: apply the framework to education, social survey, public health, and financial datasets. The third direction is reporting-oriented, by developing a standardized checklist for uncertainty-aware treatment of missing and noisy data in empirical manuscripts.

XI. CONCLUSION

This study developed and empirically tested a robust probabilistic framework for handling missing and noisy data in empirical research. The framework treats missingness and noise as sources of uncertainty that must be diagnosed, modeled, and reported, rather than as minor preprocessing defects. Using a healthcare dataset as the main case and a business dataset as a robustness case, the study compared complete-case analysis, median/mode probability modeling, iterative likelihood imputation, and robust probability modeling under native and degraded data conditions. The results showed that complete-case analysis becomes unstable when data degradation increases, mainly because it discards substantial evidence and may leave too few records for reliable estimation. Probability-based methods preserved training evidence and maintained more stable discrimination across scenarios. In the healthcare case, iterative likelihood modeling improved probability quality under stronger degradation, while in the business robustness case, the robust probability model produced the best AUC and Brier score across all scenarios. The findings support the value of an integrated workflow that combines imputation, likelihood-based estimation, robust weighting, and uncertainty reporting. The study concludes that empirical researchers should not rely on deletion or single deterministic repair methods when data are incomplete and noisy. Instead, they should evaluate multiple probability-based strategies, report retained evidence, and interpret findings through both performance and uncertainty metrics. This approach improves the credibility, transparency, and robustness of empirical conclusions drawn from imperfect real-world data.

Author Contributions

The author conducted the conceptualization, methodology, data analysis, investigation, writing, review, editing, and final approval of the manuscript.

Funding

This research received no external funding.

Data Availability

The dataset will be available from the author upon reasonable request.

Conflicts of Interest

The author declares no conflict of interest.

REFERENCES

[1]. Agostinelli, C., Leung, A., Yohai, V. J., & Zamar, R. H. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *TEST*, *24*, 441-461.

[2]. Audigier, V., White, I. R., Jolani, S., Debray, T. P. A., Quartagno, M., Carpenter, J., van Buuren, S., & Resche-Rigon, M. (2018). Multiple imputation for multilevel data with continuous and binary variables. *Statistical Science*, *33*(2), 160-183.

[3]. Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462-487.

[4]. Beesley, L. J., Bartlett, J. W., Wolf, G. T., & Taylor, J. M. G. (2021). Multiple imputation with missing data indicators. *Statistical Methods in Medical Research*, *30*(12), 2685-2700.

[5]. Carpenter, J. R., & Smuk, M. (2021). Missing data: A statistical framework for practice. *Biometrical Journal*, *63*(5), 915-947.

[6]. Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis: The TRIPOD statement. *BMC Medicine*, *13*, 1.

[7]. Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*, 140.

[8]. Ismail, A. R., Abidin, N. Z., & Maen, M. K. (2022). Systematic review on missing data imputation techniques with machine learning algorithms for healthcare. *Journal of Robotics and Control*, *3*(2), 143-152.

[9]. Jolani, S., Debray, T. P. A., Koffijberg, H., van Buuren, S., & Moons, K. G. M. (2015). Imputation of systematically missing predictors in an individual participant data meta-analysis: A generalized approach using MICE. *Statistics in Medicine*, *34*(11), 1841-1863.

[10]. Kang, H. (2013). Preventing and handling missing data. *Korean Journal of Anesthesiology*, *64*(5), 402-406.

[11]. Kleinke, K. (2017). Multiple imputation under violated distributional assumptions: A systematic evaluation of the assumed robustness of predictive mean matching. *Journal of Educational and Behavioral Statistics*, *42*(4), 371-404.

[12]. Lee, K. J., Tilling, K., Cornish, R. P., Little, R. J. A., Bell, M. L., Goetghebeur, E., Hogan, J. W., & Carpenter, J. R. (2021). Framework for the treatment and reporting of missing data in observational studies: The TARMOS framework. *Journal of Clinical Epidemiology*, *134*, 79-88.

[13]. Linero, A. R., & Daniels, M. J. (2018). Bayesian approaches for missing not at random outcome data: The role of identifying restrictions. *Statistical Science*, *33*(2), 198-213.

[14]. Luo, Y. (2022). Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics*, *23*(1), bbab489.

[15]. Ma, Z., & Chen, G. (2018). Bayesian methods for dealing with missing data problems. *Journal of the Korean Statistical Society*, *47*(3), 297-313.

[16]. McNeish, D. (2017). Missing data methods for arbitrary missingness with small samples. *Journal of Applied Statistics*, *44*(1), 24-39.

[17]. Moons, K. G. M., Altman, D. G., Reitsma, J. B., Ioannidis, J. P. A., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., & Collins, G. S. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis: Explanation and elaboration. *Annals of Internal Medicine*, *162*(1), W1-W73.

[18]. Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, *14*, 75.

[19]. Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, *17*(4), 372-411.

[20]. Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, *9*, 157-166.

[21]. Resche-Rigon, M., & White, I. R. (2018). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*, *27*(6), 1634-1649.

[22]. Rousseeuw, P. J., & Hubert, M. (2018). Anomaly detection by robust statistics. *WIREs Data Mining and Knowledge Discovery*, *8*(2), e1236.

[23]. Rousseeuw, P. J., & Van den Bossche, W. (2018). Detecting deviating data cells. *Technometrics*, *60*(2), 135-145.

[24]. Thomas, T., & Rajabi, E. (2021). A systematic review of machine learning-based missing value imputation techniques. *Data Technologies and Applications*, *55*(4), 558-585.

[25]. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, *17*, 230.

[26]. Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., & Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: From utopia to empirical data. *Journal of Clinical Epidemiology*, *74*, 167-176.

XII. APPENDICES

Appendix A. Dataset Screening Matrix

Table A1. Screening decision for all uploaded candidate datasets.

Candidate dataset	Size	Missingness	Outcome value	Decision
Healthcare cohort	4,240 x 16	Yes; 645 missing cells	Binary disease-risk outcome	Main empirical case
Business churn dataset	7,043 x 21	Yes; 8,261 missing cells after numerical conversion	Binary churn outcome	Robustness case
BRFSS diabetes 50/50 split	70,692 x 22	No native missingness	Binary diabetes outcome	Optional sensitivity only
BRFSS diabetes full binary	253,680 x 22	No native missingness	Binary diabetes outcome	Optional large-scale sensitivity only
BRFSS three-class diabetes	253,680 x 22	No native missingness	Three-class diabetes status	Optional only
Pima diabetes	768 x 9	No direct missingness; zero-coded clinical values	Binary diabetes outcome	Not selected as main case
Student performance	1,000 x 8	Only four missing values	Education score outcomes	Too weak for this title

Appendix B. Metric Definitions

Table B1. Evaluation metrics used in the empirical comparison.

Metric	Meaning	Interpretation
AUC	Ranking discrimination between positive and negative outcome cases	Higher is better
Accuracy	Share of correctly classified cases at the selected probability threshold	Higher is better
F1 score	Balance between precision and recall for the positive class	Higher is better
Brier score	Mean squared error of predicted probabilities	Lower is better
Log loss	Penalty for poorly calibrated predicted probabilities	Lower is better
Predictive entropy	Average uncertainty in predicted probabilities	Interpreted with calibration and discrimination
Training N	Number of records retained for model fitting	Higher indicates less evidence loss