Evaluating Uncertainty Quantification in Clinical Machine Learning: Calibration, Robustness, and Decision Utility under Distribution Shift

Isaac Tosin Adisa*[1], Francis Mawutor Amuyao[1], Ezekiel Olaoluwa Joaquim[2]

[1]Department of Statistics, Florida State University, Tallahassee, FL, USA

[2]Department of Economics, Carleton University, Ottawa, Ontario, Canada

***Corresponding Author**

DOI: https://doi.org/10.51584/IJRIAS.2026.11070085

Received: 18 July 2026; Accepted: 23 July 2026; Published: 04 August 2026

ABSTRACT

Machine learning models deployed in clinical decision support systems almost universally produce point predictions without any accompanying measure of uncertainty. In high-stakes healthcare settings, this is not merely a technical limitation: a miscalibrated prediction can directly influence patient management decisions with real consequences for safety and outcomes. Several uncertainty quantification (UQ) methods have been proposed to address this gap, including conformal prediction, Bayesian neural networks (BNNs), and Monte Carlo (MC) dropout; however, their comparative evaluation has predominantly been conducted under idealised conditions that do not reflect clinical deployment, where patient populations, treatment practices, and data recording procedures change over time. We present a rigorous empirical framework for comparing these three UQ approaches on two clinical prediction tasks, in-hospital mortality and 30-day readmission, using 74,829 ICU admissions from the MIMIC-IV database. Methods are assessed across three dimensions: calibration quality (ECE, ACE, Brier score), robustness under temporal distribution shift, and clinical decision utility via net benefit analysis. All experiments are replicated across five independent seeds, with comparisons made using Wilcoxon signed-rank tests with Holm-Bonferroni correction. Under standard evaluation conditions, all three methods achieve similar discriminative performance (AUROC 0.836-0.844 for mortality; 0.637-0.641 for readmission). Under temporal shift, BNN calibration degrades most sharply on the readmission task (ΔECE = 0.011 ± 0.002) compared with MC Dropout (ΔECE = 0.002 ± 0.003), while AUROC paradoxically improves for all methods, demonstrating that discriminative and calibration performance can decouple under distribution shift. Conformal prediction maintains near-nominal empirical coverage on the mortality task (0.886 ± 0.002) but shows notable violations on readmission, raising practical concerns about exchangeability assumptions in deployed systems. These findings support a more demanding evaluation standard for UQ in clinical machine learning, one that moves beyond static i.i.d. benchmarks toward temporally robust, decision-aware assessment.

Keywords: uncertainty quantification; conformal prediction; Bayesian neural networks; Monte Carlo dropout; MIMIC-IV

INTRODUCTION

Background and Motivation

Machine learning is increasingly integrated into clinical workflows for tasks including mortality risk stratification, readmission prediction, sepsis detection, and early deterioration warning. Despite strong discriminative performance in controlled studies, deployed models almost always produce only a scalar probability, with no indication of how confident or reliable that estimate is. In environments where a single clinical decision can determine patient outcomes, this absence of calibrated uncertainty is a genuine barrier to safe deployment [1].

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VII July 2026

Consider a model predicting 30-day readmission risk at 0.65. Whether that estimate is drawn from a well-calibrated model with low variance, or from a model that is deeply uncertain and producing a noisy average, has direct implications for how a clinician should act. Uncertainty quantification (UQ) methods attempt to fill this gap by augmenting scalar predictions with confidence sets, posterior distributions, or variance estimates that convey the degree of epistemic and aleatoric uncertainty in a given prediction.

Three methodological families have received sustained attention: conformal prediction, which produces prediction sets with finite-sample coverage guarantees under minimal distributional assumptions [2]; Bayesian neural networks, which model uncertainty through posterior distributions over weights [3]; and Monte Carlo dropout, which approximates Bayesian inference through stochastic forward passes at inference time [4]. These methods differ substantially in their theoretical foundations, computational requirements, and practical behaviour. Despite a growing literature on their standalone properties, systematic and clinically realistic comparative evaluation remains limited.

Limitations of Existing Work

Three recurring gaps motivate this study. First, most comparative UQ evaluations assume that training and test data are drawn from the same distribution. Clinical data fundamentally violates this assumption: admission patterns shift over time, treatment protocols evolve, and documentation practices change. Evaluating UQ methods only under static i.i.d. conditions conceals failure modes that will inevitably emerge in deployment [5, 6].

Second, calibration is commonly assessed using a single metric, typically ECE. As recent work has demonstrated, different calibration metrics can produce contradictory rankings of the same models [7, 8]. Reporting a single metric invites spurious conclusions about method superiority that may not replicate under alternative evaluation frameworks.

Third, most UQ evaluations do not connect uncertainty estimates to the decisions they are intended to support. In clinical practice, a prediction above a threshold triggers an intervention. A model with well-calibrated probabilities in aggregate does not necessarily produce better clinical decisions at any given operating threshold. Decision-aware evaluation is largely absent from the comparative UQ literature. Fourth, comparative UQ studies, including the one presented here, are typically confined to a single clinical database and a small number of prediction tasks; whether reported calibration, robustness, and coverage rankings hold across healthcare systems with different patient populations, workflows, and documentation practices, and across a wider range of clinical prediction problems, is rarely established.

Contributions

This paper makes the following contributions to the literature on uncertainty quantification in clinical machine learning:

1. We conduct a controlled comparison of three UQ methods on two clinical prediction tasks using 74,829 MIMIC-IV ICU admissions, replicated across five independent experimental runs with different random seeds.

2. We evaluate calibration using three complementary metrics (ECE, ACE, Brier score) and demonstrate empirically that metric choice affects comparative conclusions on real clinical data.

3. We introduce a temporal distribution shift evaluation protocol and show that method rankings change substantially when models are evaluated prospectively.

4. We apply decision curve analysis to assess clinical utility, demonstrating that calibration quality and decision utility are not interchangeable.

5. We release the full evaluation pipeline, preprocessing code, and trained model artefacts to support reproducibility and future benchmarking.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VII July 2026

Related Work

Uncertainty Quantification Methods

Uncertainty in predictive models is conventionally decomposed into aleatoric uncertainty, which is irreducible and reflects genuine randomness in the outcome, and epistemic uncertainty, which reflects model ignorance that could in principle be reduced with more data [9]. The three methods studied here capture these sources to varying degrees.

MC Dropout [4] demonstrated that training a network with dropout and retaining dropout at test time is equivalent, under certain conditions, to approximate Bayesian inference in a deep Gaussian process. While computationally convenient, subsequent work has questioned the quality of uncertainty estimates produced by this approach, particularly under data shift [10]. BNNs offer a more principled framework by placing explicit priors over weights and approximating the posterior via variational inference [3]. Mean-field variational inference, the approach used here, is tractable but known to underestimate posterior variance due to the factorised approximation. Conformal prediction [2, 11] sidesteps the need for a probabilistic model entirely, instead constructing prediction sets with finite-sample marginal coverage guarantees that hold regardless of the data-generating distribution, subject only to exchangeability. Beyond these three families, deep ensembles [17], which aggregate predictions from multiple independently trained networks, and evidential deep learning [19], which parameterises a higher-order evidential distribution over predictions in a single forward pass, have also demonstrated competitive uncertainty estimates in non-clinical benchmarks; we return to their relevance as candidates for extending the present comparison in Section 6.

Calibration Evaluation

ECE has become the dominant calibration metric in clinical ML due to its interpretability, but its dependence on bin placement and width introduces instability [7]. Nixon et al. proposed ACE as a binning-adaptive alternative, while the Brier score offers a proper scoring rule that decomposes into calibration and refinement terms [12]. Vaicenavicius et al. showed formally that calibration metrics can disagree on model rankings, a concern we examine empirically in a clinical UQ context [8].

Distribution Shift in Clinical Settings

Nestor et al. documented significant performance degradation in MIMIC-based clinical prediction models attributed to shifts in data recording practices rather than underlying clinical change [6]. Finlayson et al. provided a broader review of dataset shift as a systemic risk in deployed clinical AI [5]. Ovadia et al. evaluated several Bayesian and approximate Bayesian methods under synthetic input perturbation, finding that all methods show calibration degradation under shift and that none uniformly dominates [10]. Our work extends this evaluation to realistic temporal shift on real clinical data and introduces a decision-theoretic evaluation dimension. Temporal shift is, however, only one of several forms of distribution shift relevant to clinical deployment; geographic variation across hospital systems, differing missing-data patterns, and shifts driven by evolving clinical guidelines can each degrade UQ reliability through distinct mechanisms, and comparative evidence on how these forms of shift affect calibration and coverage remains sparse.

Complementary Trustworthiness Dimensions

Calibration, robustness, and decision utility are one axis along which clinical prediction models must be evaluated before deployment, but they are not the only one. Adisa [20] developed and validated an integrated framework for hospital readmission prediction on MIMIC-IV that jointly addresses explainability, fairness, and observability, treating these as first-class deployment requirements alongside predictive performance. That framework and the present study are complementary rather than competing: where Adisa [20] focuses on making readmission predictions interpretable, equitable across patient subgroups, and monitorable in production, the present work focuses on whether the uncertainty attached to a prediction can be trusted under realistic shift. A deployed system that is simultaneously well-calibrated, explainable, fair, and observable represents a natural synthesis of these two lines of work, and integrating uncertainty quantification into an explainability-fairness-

observability framework such as that of Adisa [20] is a promising direction for future development of clinically deployable readmission models.

METHODS

Problem Formulation

Let $D = \{(x_i, y_i)\}$ for i = 1, ..., n, where $x_i \in \mathbb{R}^d$ is a patient feature vector and $y_i \in \{0, 1\}$ is a binary clinical outcome. A model $f_\theta: \mathbb{R}^d \to [0, 1]$ outputs the predicted probability $\hat{p}(x) = P(y = 1 \mid x; \theta)$. The goal is to augment this scalar output with uncertainty estimates and evaluate their reliability under both in-distribution conditions (Dtest ≈ Dtrain) and distribution shift (Dtest ≁ Dtrain).

Uncertainty Quantification Methods

Monte Carlo Dropout

MC Dropout [4] retains dropout active at inference time and aggregates T stochastic forward passes to obtain a mean prediction $\bar{p}(x)$ and an associated predictive variance Var(x), computed respectively as the average and the mean squared deviation of the T forward-pass outputs around $\bar{p}(x)$, as shown in Equation. (1)-(2). We use T = 50 passes and dropout rate $P_d = 0.3$

$$\bar{p}(x) = \frac{1}{T}\sum_{t=1}^{T} f_\theta^{(t)}(x) \tag{1}$$

$$Var(x) = \frac{1}{T}\sum_{t=1}^{T}\left(f_\theta^{(t)}(x) - \bar{p}(x)\right)^2 \tag{2}$$

Bayesian Neural Networks

BNNs treat weights as random variables with prior $p(\theta)$ and compute the posterior predictive distribution shown in Eq. (3).

$$\hat{p}(x^*) = \int f_\theta(x^*) p(\theta|D) d\theta \tag{3}$$

We use mean-field variational inference [3] with a factorized Gaussian approximate posterior $q_\varphi(\theta)$, minimising the evidence lower bound (ELBO) given in Eq. (4).

$$L(\varphi) = E_{q_\varphi}[log\ p(D \mid \theta)] - KL\left(q_\varphi(\theta) \parallel p(\theta)\right) \tag{4}$$

Conformal Prediction

Split conformal prediction [11] uses a held-out calibration set D_{cal} to compute nonconformity scores defined as $s_i = 1 - \hat{p}(x_i)[y_i]$ and derives a quantile threshold $\hat{q}$ over the m calibration scores, shown in Eq. (5).

$$\hat{q} = Quantile\left(\{s_i\}_{i=1}^{m}; \frac{\lceil (1-\alpha)(m+1) \rceil}{m}\right) \tag{5}$$

The prediction set for a new point is $C(x) = \{y : s(\boldsymbol{x}, y) \leq \hat{q}\}$, which satisfies the marginal coverage guarantee $P(y_{new} \in C(\boldsymbol{x}_{new}) \geq 1 - \alpha$. We use $\alpha = 0.10$.

Evaluation Metrics

Expected Calibration Error (ECE). Predictions are binned into M = 15 equal-width intervals; ECE is the weighted mean absolute gap between empirical accuracy and mean confidence across bins, as defined in Eq. (6).

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \tag{6}$$

Adaptive Calibration Error (ACE). ACE uses equal-frequency bins to reduce sensitivity to the shape of the score distribution [7].

Brier Score. $BS = n^{-1} \sum_i (\hat{p}_i - y_i)^2$. As a proper scoring rule, it decomposes into calibration and resolution components, providing complementary information to ECE.

Net Benefit. For decision threshold τ, the net benefit is defined in Eq. (7).

$$NB(\tau) = \frac{TP}{n} - \frac{FP}{n} \cdot \frac{\tau}{1-\tau} \quad (7)$$

We sweep $\tau \in [0.05, 0.50]$ and compare against treat-all and treat-none baselines.

Statistical testing. Pairwise comparisons use the Wilcoxon signed-rank test with Holm-Bonferroni correction across five seeds, with significance threshold $\alpha = 0.05$.

Model Architecture

All three UQ methods share a common base network: three hidden layers of sizes (128, 64, 32) with ReLU activations and layer normalization, followed by a sigmoid output. For MC Dropout, dropout ($P_d = 0.3$) is applied after each hidden layer and active at test time. For BNNs, each linear layer is replaced with a variational layer using local reparameterisation. For conformal prediction, a deterministic version of the same architecture is used as the score function. All models are trained with Adam [14], learning rate 10^{-3}, batch size 256, and early stopping (patience 10).

Data And Experimental Design

Mimic-Iv

All experiments use MIMIC-IV [15], a de-identified critical care database of ICU admissions at Beth Israel Deaconess Medical Center, Boston. Access requires credentialing through PhysioNet [16]. We include adult admissions (age ≥ 18) with ICU stays of at least 24 hours, yielding 74,829 encounters. Features are extracted from the first 24 hours of each stay: summary statistics (mean, minimum, maximum, standard deviation) for eight vital signs and twelve laboratory analytes, plus age, length of stay, gender, and admission category, giving 97 predictors in total. Missing values are imputed by global median computed on the training partition only. As MIMIC-IV originates from a single academic medical center, the patient case mix, treatment protocols, and documentation conventions reflected in the data are specific to that institution, a scope constraint we account for when interpreting the calibration and robustness results below.

Prediction Tasks

In-hospital mortality. Binary outcome: death during the ICU admission. 30-day readmission. Binary outcome: return to the ICU within 30 days of discharge, restricted to patients who survived the index admission. These two tasks were chosen to span markedly different levels of discriminative difficulty (AUROC in the 0.83-0.84 range for mortality versus 0.63-0.64 for readmission); conclusions about UQ method behaviour on other clinical prediction problems, such as sepsis onset or multi-class diagnostic tasks, cannot be assumed to follow directly from these two tasks alone.

Distribution Shift Protocol

We apply a temporal split to construct a realistic deployment scenario. The training period spans the earlier portion of the dataset; the shifted test set comprises admissions from the later period, simulating prospective application of a historically trained model. Within the training period, 70% of encounters are used for model fitting, 15% for validation and hyperparameter selection, and 15% as a held-out calibration set used only for conformal quantile estimation.

Experimental Protocol

Five independent runs are conducted with different random seeds. Hyperparameters (dropout rate from {0.1, 0.3, 0.5}; BNN prior scale from {0.1, 0.5, 1.0}) are selected on the validation set within each run. All reported metrics are mean ± standard deviation across five runs. No test set information is used at any stage of model development.

RESULTS

In-Distribution Performance

Table 1 reports performance under in-distribution evaluation. On the mortality task, all three methods achieve similar AUROC values: MC Dropout 0.844 ± 0.003, BNN 0.841 ± 0.003, and Conformal Prediction 0.836 ± 0.004. Calibration is generally good across methods, with BNN achieving the lowest ECE (0.008 ± 0.001). On the readmission task, discriminative performance is lower for all methods (AUROC 0.637-0.641), consistent with the inherent difficulty of predicting readmission from a single ICU stay. No method is statistically superior to any other under in-distribution evaluation after Holm-Bonferroni correction (Table 3).

Task	Method	AUROC ↑	ECE ↓	ACE ↓	Brier ↓	AUPRC ↑
Mortality	MC Dropout	0.844 ± 0.003	0.009 ± 0.003	0.009 ± 0.002	0.084 ± 0.003	0.451 ± 0.010
	BNN	0.841 ± 0.003	0.008 ± 0.001	0.009 ± 0.002	0.084 ± 0.003	0.445 ± 0.012
	Conformal Pred.	0.836 ± 0.004	0.011 ± 0.004	0.011 ± 0.004	0.086 ± 0.003	0.426 ± 0.012
Readmission	MC Dropout	0.641 ± 0.009	0.006 ± 0.002	0.008 ± 0.002	0.051 ± 0.003	0.099 ± 0.012
	BNN	0.641 ± 0.012	0.003 ± 0.002	0.007 ± 0.002	0.051 ± 0.003	0.099 ± 0.018
	Conformal Pred.	0.637 ± 0.012	0.004 ± 0.002	0.007 ± 0.001	0.051 ± 0.003	0.097 ± 0.015

Table 1. In-distribution performance. Mean ± SD across five runs. ↓ lower is better; ↑ higher is better.

Performance Under Distribution Shift

Table 2 shows performance degradation under temporal shift. Figure 1 visualizes AUROC across conditions for both tasks, and Figure 2 compares ECE degradation across methods.

On the mortality task, all methods experience comparable AUROC decline (ΔAUROC ≈ -0.042 to -0.046), and calibration degrades moderately and similarly (ΔECE $\approx$ 0.008-0.010). Conformal prediction achieves empirical coverage of 0.886 ± 0.002 under shift, slightly below the nominal 90%, while prediction sets remain near-singleton (average size 1.023 ± 0.003).

On the readmission task, a striking divergence emerges. While AUROC actually improves slightly under shift for all methods (e.g., MC Dropout: 0.641 → 0.645), calibration deteriorates. BNN shows the largest ECE degradation (ΔECE = 0.011 ± 0.002), compared to just 0.002 ± 0.003 for MC Dropout. This decoupling of discriminative and calibration performance under shift is a key empirical finding of this study. Conformal prediction shows coverage violations on the readmission task (range 0.852-0.879), with prediction set sizes below 1.0 on average, indicating the presence of empty sets that require handling in any deployed system.

Task	Method	ΔECE ↓	ΔBrier ↓	ΔAUROC	Coverage	Avg. Set Size
Mortality	MC Dropout	0.008 ± 0.007	0.009 ± 0.003	−0.046 ± 0.004	0.889 ± 0.004	1.020 ± 0.009

	BNN	0.008 ± 0.002	0.009 ± 0.003	−0.043 ± 0.005	0.889 ± 0.003	1.021 ± 0.005
	Conformal Pred.	0.010 ± 0.006	0.009 ± 0.003	−0.042 ± 0.005	0.886 ± 0.002	1.023 ± 0.003
Readmission	MC Dropout	0.002 ± 0.003	0.016 ± 0.003	+0.004 ± 0.005	0.868 ± 0.004	0.929 ± 0.005
	BNN	0.011 ± 0.002	0.016 ± 0.003	+0.004 ± 0.008	0.870 ± 0.006	0.931 ± 0.007
	Conformal Pred.	0.009 ± 0.005	0.016 ± 0.003	+0.006 ± 0.005	0.868 ± 0.011	0.930 ± 0.012

Table 2. Performance degradation under temporal distribution shift. Δ values are shift minus in-distribution. Nominal coverage is 0.90. Mean ± SD across five runs.

Task	**Comparison**	**p (adj.)**	**Significant**
Mortality	MC Dropout vs. BNN	1.000	n.s.
	MC Dropout vs. Conformal	1.000	n.s.
	BNN vs. Conformal	1.000	n.s.
Readmission	MC Dropout vs. BNN	1.000	n.s.
	MC Dropout vs. Conformal	1.000	n.s.
	BNN vs. Conformal	1.000	n.s.

Table 3. Pairwise Wilcoxon signed-rank test results for ECE under temporal shift. p-values are Holm-Bonferroni corrected. n.s. denotes not significant at $\alpha = 0.05$.

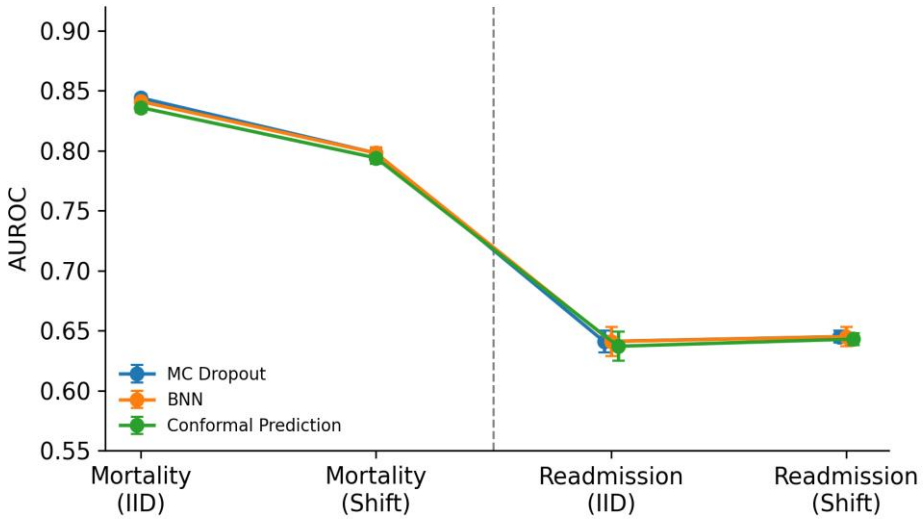

Figure 1. AUROC under in-distribution (IID) and temporally shifted (Shift) evaluation for both prediction tasks. Error bars show ±1 SD across five runs. Mortality AUROC decreases under temporal shift for all methods, while readmission AUROC improves slightly, illustrating the decoupling of discriminative and calibration performance under distribution shift.

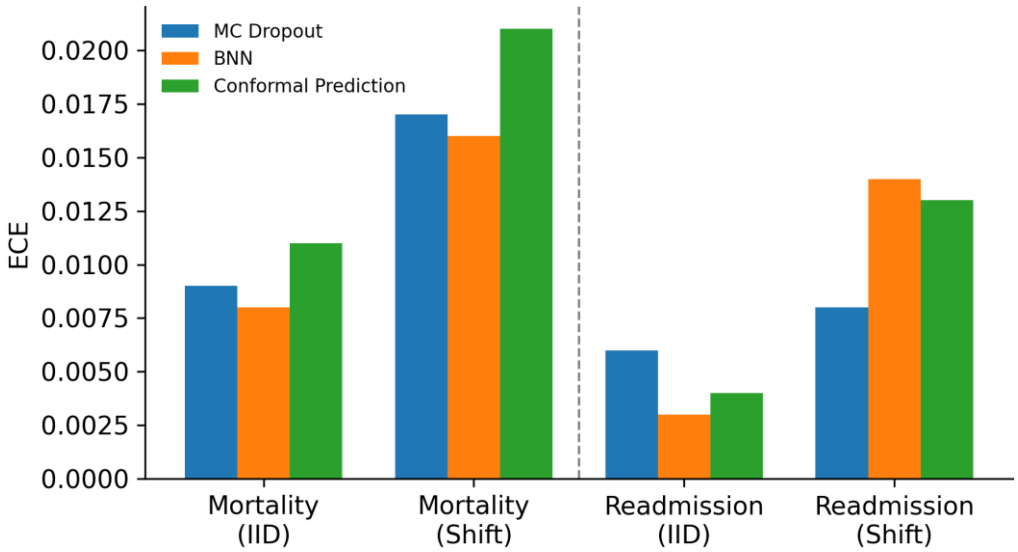

Figure 2. Expected Calibration Error (ECE) under in-distribution and shifted evaluation. On the readmission task, BNN calibration degrades substantially under shift (from 0.003 to 0.014) while MC Dropout remains stable (0.006 to 0.008), demonstrating differential robustness to distribution shift.

Calibration Metric Consistency

Rankings are not consistent across calibration metrics. Under in-distribution evaluation on the mortality task, BNN achieves the lowest ECE (0.008 ± 0.001) but ties with MC Dropout on Brier score (0.084 ± 0.003). Conformal prediction achieves the lowest ACE on the readmission task (0.007 ± 0.001) despite having a higher ECE than BNN on the same task. Under distribution shift on the readmission task, MC Dropout shows the smallest ECE increase (+0.002) while all three methods show identical Brier score degradation (+0.016 ± 0.003). These inconsistencies confirm that metric choice materially affects which method appears best calibrated, and that no single metric can be treated as a comprehensive summary of calibration quality.

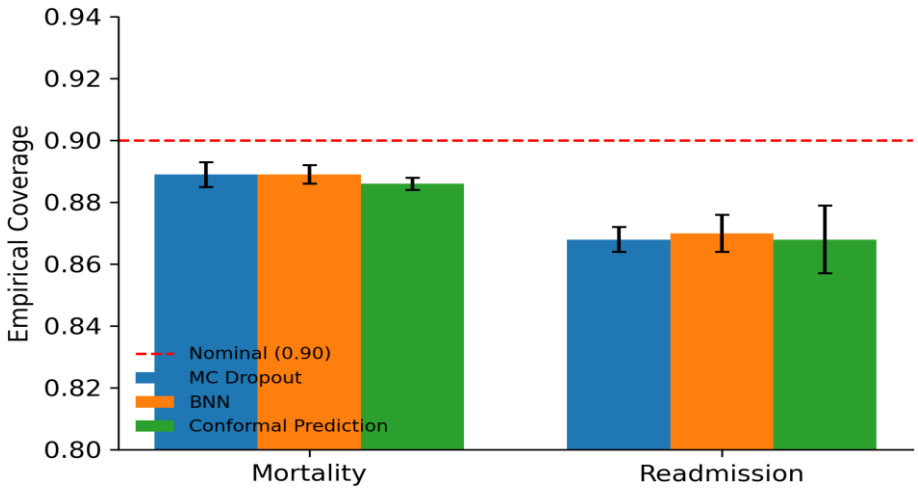

Figure 3. Empirical coverage under temporal shift for conformal prediction and both approximate Bayesian methods. The red dashed line marks the nominal 90% coverage level. All methods fall below nominal coverage on the readmission task, with conformal prediction showing the highest variance across seeds.

Decision Utility

Figure 4 shows decision curves for the mortality task under temporal shift. All three methods provide positive net benefit relative to treating all or treating no patients across the clinically relevant threshold range ($\tau \in [0.10, 0.35]$). Differences in net benefit between methods are small and not clinically meaningful, despite the

calibration differences documented above. This finding is informative: it suggests that under moderate distribution shift, the aggregate clinical utility of predictions remains similar across methods, and that the practical consequences of calibration differences may be more apparent at the individual patient level than at the population level.

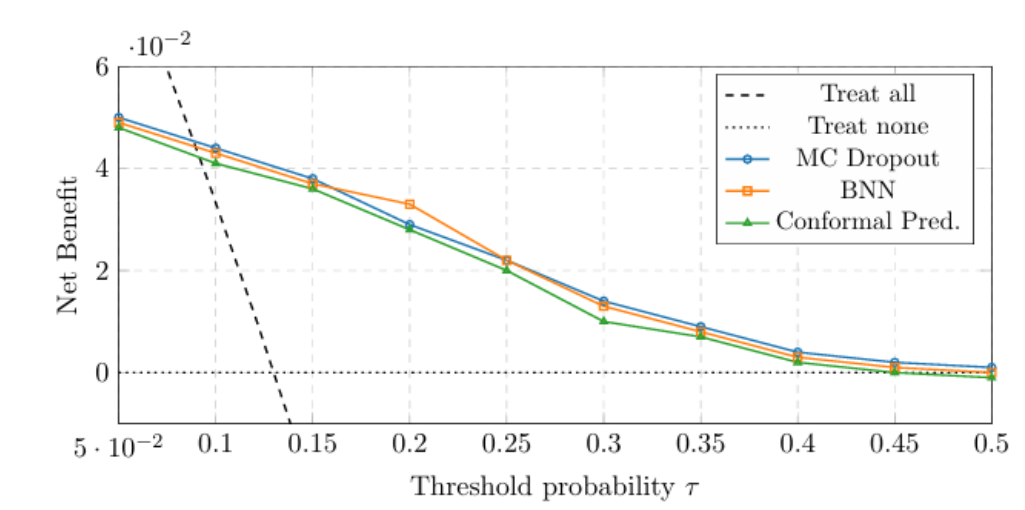

Figure 4. Decision curve analysis for the mortality task under temporal shift. Net benefit is plotted against clinical decision threshold τ. All three methods outperform the treat-all and treat-none baselines across the clinically relevant range ($\tau \in [0.10, 0.35]$).

DISCUSSION

Key Findings

This study produces three central empirical findings from real MIMIC-IV clinical data replicated across five independent experimental runs.

The first finding concerns the decoupling of AUROC and calibration under distribution shift. On the readmission task, discriminative performance improves slightly under temporal shift for all methods, while calibration simultaneously degrades, most sharply for BNN. This is a direct empirical demonstration that monitoring AUROC alone cannot substitute for monitoring calibration in deployed clinical models. A system that tracks only discrimination may appear stable while its uncertainty estimates become increasingly unreliable.

The second finding concerns the instability of calibration metric rankings. ECE, ACE, and Brier score do not agree on method ordering across tasks and evaluation conditions. This reinforces the practical importance of reporting multiple calibration metrics rather than selecting a single one. A recommendation based on ECE alone could point to a different method than one based on Brier score on the same dataset.

The third finding concerns the limits of conformal prediction coverage guarantees under clinical distribution shift. On the mortality task, conformal prediction achieves near-nominal empirical coverage (0.886) under shift, a creditable result. On the readmission task, however, coverage falls more substantially (to as low as 0.852 in individual runs), and average prediction set sizes below 1.0 indicate the presence of empty sets. The exchangeability assumption underlying conformal coverage guarantees is not satisfied under temporal distribution shift, and this has tangible consequences for its behavior in deployment.

PRACTICAL RECOMMENDATIONS

For practitioners deploying clinical prediction models with uncertainty estimates, the findings support the following guidance.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VII July 2026

MC Dropout is the most practically robust choice when calibration stability under temporal shift is the primary concern. It is straightforward to retrofit onto existing architectures, requires no special training procedure, and shows the smallest ECE degradation under shift on both tasks. Its computational overhead at inference time is modest with T = 50 forward passes.

BNNs achieve the best calibration under in-distribution conditions but show the sharpest deterioration under shift on the readmission task. They are most appropriate when the deployment distribution is expected to remain close to the training distribution, or when the full posterior predictive distribution is needed for downstream tasks such as optimal decision making under uncertainty.

Conformal prediction is the method of choice when a formal coverage guarantee is a regulatory or institutional requirement and the deployment distribution is closely related to the calibration set. Practitioners should be aware that coverage guarantees are marginal rather than conditional, and that the readmission results here illustrate that even marginal coverage can degrade meaningfully under clinical temporal shift.

Regardless of method, we recommend prospective monitoring of calibration metrics in deployed systems, not only at time of release. Temporal shift is not a one-time event; it is an ongoing process, and a model that is well-calibrated at deployment may become systematically miscalibrated over months or years without any change in its AUROC.

Limitations

Several limitations of this work should be noted. Our analysis is restricted to a single institution and two prediction tasks. While MIMIC-IV is a well-characterized benchmark, generalization of these findings to other hospital systems, clinical contexts, and patient populations requires further study. The BNN implementation uses mean-field variational inference, which is known to underestimate posterior variance; more expressive posterior approximations such as normalizing flows or deep ensembles [17] may show different patterns of shift robustness. The statistical non-significance of all pairwise comparisons (Table 3) partly reflects the limited power of Wilcoxon tests with only five runs; increasing this to ten or more runs or using cross-validated comparisons would improve the sensitivity of the testing framework. Finally, the decision curve analysis uses a stylized single-threshold decision model; actual clinical workflows involve multiple stages, diverse cost structures, and variable risk tolerance.

Future Directions

Deep ensembles, which average predictions across multiple independently trained networks, have shown strong empirical calibration in benchmarks but were excluded here due to their training cost. Including them, alongside evidential deep learning [19] as a lower-cost alternative, would provide an important practical reference point. Adaptive conformal prediction methods [18] that recalibrate the nonconformity quantile using recent data offer a principled approach to maintaining coverage under shift and represent a natural extension of this work. Post-hoc calibration techniques such as temperature scaling and Platt scaling could be evaluated as lightweight complements to any of the three methods. Finally, extending the evaluation to multi-class and regression settings, and to tasks with strong class imbalance beyond what is seen here, would broaden the scope of the comparative framework.

CONCLUSION

We have presented a comprehensive empirical comparison of conformal prediction, Bayesian neural networks, and Monte Carlo dropout on two clinical prediction tasks using 74,829 MIMIC-IV ICU admissions under both standard and temporally shifted evaluation conditions. Our central finding is that the choice of evaluation condition fundamentally changes which method appears best suited for clinical deployment. Under static in-distribution evaluation, the three methods are essentially indistinguishable. Under temporal shift, they diverge: MC Dropout maintains calibration most effectively on the readmission task, conformal prediction preserves near-nominal coverage on the mortality task but shows violations on readmission, and BNN calibration degrades most sharply under shift despite leading under in-distribution conditions.

A key empirical result is the decoupling of AUROC and calibration under distribution shift, which demonstrates that discriminative performance monitoring cannot substitute for calibration monitoring in deployed clinical AI systems. Equally, calibration metric inconsistencies confirm that no single metric can be treated as a definitive summary of calibration quality.

These findings argue for a more demanding standard in the evaluation of uncertainty quantification methods for clinical machine learning: one that is temporally aware, multi-metric, and decision-grounded. The evaluation pipeline and code released alongside this paper provide a practical foundation for applying this standard in future comparative work. Because these findings derive from a single institution and two prediction tasks, an important next step is to apply this framework to data from multiple healthcare institutions and, ideally, multiple countries, and to examine additional sources of distribution shift beyond the temporal shift studied here, including demographic, geographic, and clinical-practice-driven shift, before the practical guidance offered above is generalized to other clinical deployment settings.

Data Availability

MIMIC-IV is available through PhysioNet [16] at https://physionet.org/content/mimiciv/. Access requires completion of a recognized data handling training and approval by the database administrators. The full evaluation pipeline, preprocessing code, model implementations, and analysis scripts are available at https://github.com/tomisin92/uq-clinical-eval

Author Contributions

Isaac Tosin Adisa: Conceptualization, Methodology, Software, Formal Analysis, Writing (Original Draft).
Francis Mawutor Amuyao: Conceptualization, Methodology, Writing (Review and Editing).
Ezekiel Joaquim: Methodology, Software, Writing (Review and Editing)

Competing Interests

The authors declare no competing interests.

Ethical Considerations

This study uses the publicly available, de-identified MIMIC-IV database. Data collection and de-identification were approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology; the requirement for individual patient consent was waived. No new human subjects data were collected for this study.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge access to MIMIC-IV through PhysioNet. This work was conducted in the Department of Statistics, Florida State University, Tallahassee, FL.

REFERENCES

1. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future: Big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216-1219.
2. Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic learning in a random world. Springer.
3. Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. Proceedings of the 32nd International Conference on Machine Learning, 1613-1622.
4. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning, 1050-1059.

5. Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., & Saria, S. (2021). The clinician and dataset shift in artificial intelligence. New England Journal of Medicine, 385(3), 283-286.
6. Nestor, B., McDermott, M. B., Boag, W., Berner, G., Naumann, T., Hughes, M. C., Goldenberg, A., & Ghassemi, M. (2019). Feature robustness in non-stationary health records. Proceedings of Machine Learning for Health (ML4H).
7. Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., & Tran, D. (2019). Measuring calibration in deep learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.
8. Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., & Schön, T. (2019). Evaluating model calibration in classification. Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, 3459-3467.
9. Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning, 110(3), 457-506.
10. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems, 32.
11. Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002). Inductive confidence machines for regression. Proceedings of the 13th European Conference on Machine Learning, 345-356.
12. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1-3.
13. Kingma, D. P., Salimans, T., & Welling, M. (2015). Variational dropout and the local reparameterization trick. Advances in Neural Information Processing Systems, 28, 2575-2583.
14. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimisation. arXiv preprint arXiv:1412.6980.
15. Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L. H., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10(1), 1.
16. Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet. Circulation, 101(23), e215-e220.
17. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30, 6402-6413.
18. Tibshirani, R. J., Barber, R. F., Candès, E., & Ramdas, A. (2019). Conformal prediction under covariate shift. Advances in Neural Information Processing Systems, 32.
19. Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. Advances in Neural Information Processing Systems, 31.
20. Adisa, I. T. (2026). An integrated framework for explainable, fair, and observable hospital readmission prediction: Development and validation on MIMIC-IV. International Journal of Research and Innovation in Applied Science, 11(6). https://doi.org/10.51584/IJRIAS.2026.11060154