ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.2473

Explainable Bayesian Framework for Medical Diagnosis: Asymmetric Decisions, Asymptotic Calibration, and Scalable Inference

José Augusto Rojas Peñafiel, MD[1]; Ana Gabriela Valladares Patiño, MSc[2];
[1]Health Sciences Faculty, Universidad Internacional SEK (UISEK), Equator, *jose.rojas@uisek.edu.ec,*
[2]Universidad Central del Ecuador, Equator, *agvalladares@uce.edu.ec*

Abstract* – Automated medical diagnosis requires predictive models with interpretable uncertainty quantification aligned with clinical decisions. We develop a Bayesian framework for binary classification that integrates asymmetric clinical risk, asymptotic guarantees, and approximate inference. The Bayesian predictor minimizes expected clinical risk, is asymptotically calibrated, and satisfies the Bernstein–von Mises theorem. We introduce the Clinical Uncertainty Index (CUI), a theoretically grounded metric that decomposes uncertainty into epistemic and aleatory components. The CUI converges to zero under correct specification but remains strictly positive under misspecification, serving as a diagnostic of model inadequacy. Analyzing approximate inference, we formally prove that mean-field variational approximations systematically underestimate posterior uncertainty—inducing diagnostic overconfidence—while full-covariance variational inference achieves an effective balance between computational efficiency and uncertainty fidelity (≥7× faster than Markov Chain Monte Carlo (MCMC) while preserving posterior variance accuracy). Unlike calibration metrics such as the Brier Score, which may overlap under correct and misspecified models, the CUI provides a normalized, asymptotically guaranteed measure of structural uncertainty. These results establish rigorous foundations for clinically reliable AI systems with explicit probabilistic guarantees. These findings motivate future evaluation on clinical benchmarks such as MIMIC-III and Sepsis-3, as well as prospective deployment studies.

Keywords*-- Bayesian inference; uncertainty quantification; clinical decision theory; Clinical Uncertainty Index (CUI); asymptotic calibration; variational inference.

I. Introduction

Medical diagnosis constitutes a fundamental decision-making problem under uncertainty, in which healthcare professionals are required to integrate limited, heterogeneous, and noisy clinical information to make judgments that directly affect patient safety and prognosis [1,2]. This process may be formally interpreted as a probabilistic inference mechanism, whereby observed evidence updates prior beliefs regarding the presence of a pathological condition [3–6].

In recent decades, artificial intelligence (AI) models have assumed an increasingly important role in diagnostic applications, driven by the availability of large volumes of clinical data and the development of progressively sophisticated machine learning algorithms [4]. Nevertheless, a significant proportion of these models generate deterministic or point estimates without providing formal quantification of the uncertainty associated with their predictions [7].

This limitation becomes particularly critical in high-risk clinical environments. Overconfidence in incorrect predictions may induce suboptimal diagnostic decisions with potentially severe consequences [8,9]. In this context, explicit uncertainty quantification emerges not merely as a statistical refinement, but as an essential requirement for the safe deployment of AI systems in medicine [10,11].

Regulatory agencies such as the FDA (Food and Drug Administration) have increasingly emphasized the necessity for clinically deployable AI systems to provide calibrated probabilistic estimates together with interpretable confidence measures [12,13]. This requirement reflects an unavoidable reality: even highly accurate predictive systems remain inherently subject to error, and the characterization of the uncertainty associated with such errors is as important as average predictive accuracy itself.

From the perspective of decision theory, medical diagnosis exhibits distinctive characteristics, most notably the asymmetry of costs associated with diagnostic errors [14–16]. In numerous clinical scenarios, false negatives entail substantially greater risks than false positives, thereby invalidating the use of arbitrary decision thresholds and motivating frameworks that explicitly incorporate clinically relevant loss functions [17–19].

Bayesian decision theory provides a natural mathematical framework for addressing this problem by enabling the integration of prior information, the modeling of differential costs, and the derivation of optimal decision rules under uncertainty [14–16,20,21].

Several recent approaches have attempted to quantify uncertainty in medical predictive models [10,11]. Methods such as deep ensembles [22] demonstrate empirical utility, although their theoretical foundations remain limited. Techniques such as dropout interpreted as a Bayesian approximation [7] provide computational advantages, yet may distort posterior variance estimation [23].

A relevant conceptual advance was introduced by Alex Kendall and Yarin Gal [24], who distinguished between aleatoric and epistemic uncertainty [25,26]. Calibration metrics such as the Brier Score [37] are commonly employed to evaluate probabilistic predictions; however, such metrics aggregate calibration and refinement information, potentially masking model misspecification errors with significant clinical implications.

Despite these advances, a fundamental gap persists in the literature: many uncertainty quantification strategies lack rigorous mathematical foundations capable of guaranteeing calibration, consistency, and asymptotic validity. Similarly, the formal connection between Bayesian inference, uncertainty quantification, and clinical decision theory remains fragmented.

This limitation is particularly relevant in high-impact diagnostic applications, where calibration errors or model overconfidence may translate into substantial clinical risks.

In this context, the present work comprehensively addresses the identified gap through several contributions. A complete Bayesian formalization of clinical diagnosis is developed, establishing rigorous mathematical foundations together with explicit regularity conditions.

A decision theorem under generalized medical loss is formulated and proved, from which optimal rules with patient-dependent adaptive thresholds are derived. An asymptotic calibration theorem is established, guaranteeing convergence of the Bayesian predictor toward well-calibrated probabilistic estimates.

Likewise, the Bernstein–von Mises theorem is applied, providing a formal connection between Bayesian and frequentist inference in the asymptotic regime [27,28].

The Clinical Uncertainty Index (CUI) is introduced as a novel index for the formal decomposition of diagnostic uncertainty and the identification of model misspecification errors. The CUI is explicitly positioned relative to the Kendall–Gal decomposition and to calibration metrics such as the Brier Score, theoretically and empirically establishing the scenarios in which the CUI provides additional diagnostic information that aggregate metrics may fail to capture. Finally, a formal analysis of variational inference is presented, characterizing its impact on posterior uncertainty quantification and deriving practical recommendations for its application in clinical contexts [29–31].

II. METHODOLOGY

The methodological framework is structured into five main components: (1) the development of a general mathematical framework formalizing the binary classification problem from a Bayesian perspective; (2) the definition of regularity conditions required to support the asymptotic results; (3) the formulation of clinical decision theory under asymmetric loss functions; (4) the design of computational simulations aimed at empirically validating the theoretical properties; and (5) the construction of illustrative clinical scenarios demonstrating the practical applicability of the proposed framework.

A. General Mathematical Framework

A binary medical classification problem is considered in which, for each patient, a feature vector $X \in X \subseteq R^d$ is observed, together with a response variable $Y \in \{0,1\}$ indicating the absence (Y=0) or presence (Y=1) of a pathological condition of interest.

The underlying probability space is defined as (Ω, F, P), where Ω denotes the sample space, F a σ-algebra of events, and P a probability measure. An unknown conditional density $p\theta(y|x)$ is assumed to characterize the relationship between patient features and clinical outcomes [1].

A parametric family of probability models dominated by a σ-finite measure is considered, given by $\{p_{\theta(y|x)} : \theta \in \Theta\}$, where $\theta \in \theta \subseteq \mathbb{R}^{\mathbb{P}}$ represents the parameter vector.

Given a dataset $Dn = \{(x_i, y_i)\}_{i=1}^{n}$, the likelihood function is defined as:

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} P\theta(y_i | x_i)$$

Prior information is incorporated through a prior distribution π(θ) defined over the parameter space Θ. The posterior distribution is obtained via Bayes' theorem:

$$\pi(\theta | Dn) = \frac{\mathcal{L}n(\theta)\,\Pi(\theta)}{\int \theta \alpha_n(\theta)\,\Gamma_1(\theta)\,d\theta} \alpha\, \mathcal{L}_n(\theta)\,\Pi(\theta)$$

The posterior predictive probability for a new patient with features x is defined as:

$$\hat{p}(x) = \mathbb{P}(y = 11x, Dn) \in \int_{\theta} p_{(\theta)}(1|x)\pi(\theta|Dn)\,d\theta$$

This quantity represents the Bayesian predictive probability averaged over posterior parameter uncertainty [14].

B. Regularity Conditions

To establish the asymptotic results presented in subsequent sections, standard regularity conditions are imposed [27,28]:

1. Identifiability: distinct parameter values induce distinct conditional distributions.
2. Differentiability: the log-likelihood is continuously differentiable up to third order in a neighborhood of the true parameter value.
3. Non-degeneracy of Fisher information: the Fisher information matrix is positive definite locally around the true parameter.
4. Prior regularity: the prior distribution is continuous and strictly positive in a neighborhood of the true parameter value.
5. Existence of dominating moments: the third derivatives of the log-likelihood are dominated by an integrable function.

Under these conditions, the posterior distribution admits an asymptotic normal approximation through the Bernstein–von Mises theorem [27,28].

C. Clinical Decision Theory with Asymmetric Loss

The diagnostic process is formalized as a statistical decision problem with action space A = {0,1}, where a=1 denotes diagnosis of the condition, and a=0 denotes absence of diagnosis.

To account for asymmetric clinical consequences, the following loss function is defined:

$L(y,a;x)=cFN(x)^1\{y=1,a=0\}+cFP(x)^1\{y=0,a=1\}$, where cFN(x) and cFP(x) represent the costs associated with false negatives and false positives, respectively.

The posterior expected risk is:

$$R(a|x) = E_{(Y|x,\,D_n)[L(Y,a;x)]} = \begin{cases} c_{FP}(x)\left(1-\hat{p}(x)\right), si\ a=1 \\ c_{FN}\hat{x}p(x), si\ a=0 \end{cases}$$

The optimal Bayesian decision rule is obtained by minimizing posterior risk:

$$\delta^{*(x)} = \arg\min_{a\in\{0,1\}} R(a|x)$$

This yields the adaptive threshold rule:

$$\delta^{*(x)} = 1\left\{\hat{p}(x) > \frac{c_{FP(x)}}{c_{FP(x)} + c_{FN(x)}}\right\}$$

The resulting threshold depends explicitly on the relative clinical costs, allowing decision boundaries to adapt according to diagnostic context [17,19].

D. Uncertainty Decomposition and the Clinical Uncertainty Index (CUI)

Predictive uncertainty is decomposed using the law of total variance:

$$Var(Y|x,D_n) = E_{(\theta|D_n)}[var(Y|x,\theta)] + Var_{(\theta|D_n)}(E[Y|x,\theta])$$

The first term, denoted UA(x) = Eθ| Dn[Var(Y | x, θ)], corresponds to aleatoric uncertainty (irreducible variability inherent to the biological phenomenon). The second term, denoted UE(x) = Var θ|Dn (E[Y | x, θ]), represents epistemic uncertainty (reducible uncertainty due to limited knowledge about the correct model) [25,26].

The Clinical Uncertainty Index (CUI) is defined as the proportion of total uncertainty attributable to the epistemic component:

$$CI(x) = \frac{U_{E(x)}}{U_{A(x)} + U_{E(x)}} = \frac{U_{E(x)}}{Var(Y|x,D_n)}$$

This index satisfies several theoretical properties formally demonstrated in Section III: boundedness ($0 \le CUI(x) \le 1$), consistency under correct specification ($CUI\_n(x) \to 0$), and sensitivity to model misspecification ($CUI\_n(x) \to c(x) > 0$).

The proposed framework builds upon the aleatoric-epistemic decomposition introduced by Kendall and Gal [24], while extending it through normalization and formal asymptotic guarantees. Unlike aggregate calibration metrics such as the Brier Score [37], the CUI provides patient-level quantification of reducible uncertainty, enabling clinically interpretable uncertainty stratification across clinical contexts.

E. Computational Simulation Design

Simulation experiments were designed to validate the theoretical properties of the proposed framework and the Clinical Uncertainty Index (CUI).

Synthetic datasets were generated under logistic regression models:

$$Y_i \sim Bernoulli(p_i), \qquad p_i = \sigma(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})$$

where $\sigma(\cdot)$ denotes the logistic function and $\beta=(0.5,1.2,-0.8)^T$.

Covariates were independently generated as:

$X\{i1\} \sim N(0, 1)$, $X\{i2\} \sim Uniform(-2, 2)$

Sample sizes $n \in \{100, 500, 1000, 5000, 10000\}$ were evaluated.

Exact Bayesian inference was performed using Hamiltonian Monte Carlo implemented in STAN [33]. Variational inference was evaluated under both mean-field and full-covariance approximations using PyMC [34].

For representative covariate values, aleatoric and epistemic uncertainty components were estimated numerically from posterior samples:

$$\hat{U}_{A(x)} = \frac{1}{S}\sum_{s=1}^{S} p_{(\theta^s)}(1|x)\left(1-p_{(\theta^s)}(1|x)\right)$$

$$\hat{U}_{E(x)} = \frac{1}{S}\sum_{s=1}^{S}\left(p_{(\theta^s)}(1|x) - \hat{p}(x)\right)^2$$

where $\{\theta^{(s)}\}_{s=1}^{s}$ are samples from the posterior distribution and p(x) =

$$\left(S^{\{-1\}}\right)\sum_{s=1}^{S} p_{(\theta^s)}(1|x)$$

Additional simulations incorporated model misspecification scenarios, including omitted interactions, nonlinear effects, and heteroscedasticity.

F. Illustrative Clinical Scenarios

Two illustrative clinical applications were considered.

Scenario 1: Sepsis diagnosis in intensive care units, using simulated variables inspired by Seymour et al. [35], including serum lactate, arterial pressure, heart rate, temperature, leukocyte count, and age.

Scenario 2: Pulmonary nodule classification in computed tomography, based on radiological descriptors including nodule diameter, density, spiculation, anatomical location, and margin characteristics [36].

In both scenarios, the framework was used to evaluate adaptive decision thresholds, patient-level CUI values, and the effects of approximate inference on clinical decision-making.

G. Evaluation Metrics and Statistical Analysis

Model calibration was evaluated using calibration plots, the Hosmer-Lemeshow statistic [32], and the Brier Score [37]:

$$Brier = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}(x_i) - y_i)^2$$

Predictive discrimination was assessed using AUC-ROC and AUC-PR metrics. Credible interval coverage and interval length were also evaluated.

Bootstrap resampling (1000 replicates) was employed to assess the stability of the Clinical Uncertainty Index (CUI) and construct confidence intervals.

All analyses were conducted using Python 3.10 with NumPy, SciPy, PyMC, STAN, and scikit-learn.

III. THEORETICAL RESULTS

This section presents the mathematical foundations supporting the proposed Bayesian framework, providing formal justification for the inferential properties, clinical decision-making, and uncertainty quantification analyzed empirically. In particular, we establish results regarding the optimality of diagnostic decision-making, the asymptotic calibration of the Bayesian predictor, the behavior of the Clinical Uncertainty Index (CUI), and the effects induced by variational approximations.

A. Optimality of Clinical Decision-Making Under Asymmetric Loss

Medical diagnosis can be formulated as a statistical decision problem in which the clinical consequences of diagnostic errors are inherently asymmetric.

Theorem 1 (Optimality of the Bayesian decision rule). Under the asymmetric loss function defined in (1), the decision rule that minimizes the expected posterior risk is:

$$a^{*(x)} = \begin{matrix} 1, \text{si } \hat{p}_{n(x)} \geq t^{*(x)} \\ 0, \text{en otro caso} \end{matrix}$$

where the optimal threshold satisfies:

$$t^{*(x)} = \frac{c_{FN(x)}}{c_{FN(x)} + c_{FP(x)}}$$

Proof sketch. The rule is obtained by minimizing the posterior risk R(a | x,Dn)=E[L(Y,a) | x,Dn], which leads to a decision criterion dependent on relative costs [14,17].

Clinical interpretation. This result formalizes the need for adaptive thresholds in diagnostic scenarios with differential clinical costs. When false negatives are more dangerous than false positives (cFN >> cFP), the threshold approaches zero, indicating a "liberal" diagnostic strategy.

B. Asymptotic Probabilistic Calibration

Probabilistic calibration is an essential property for the clinical interpretation of estimated risks.

Theorem 2 (Asymptotic Calibration). Under correct model specification and the regularity conditions established in Section II-B:

$$P\left(Y = 1 \middle| \hat{p}_{n(x)} = p\right) \to p \, , cuando \; n \to \infty$$

Proof sketch. Posterior consistency and the Bernstein–von Mises theorem imply convergence of the Bayesian predictor toward the true probability [6,27].

Clinical interpretation. This guarantees that, with sufficient evidence, diagnostic probabilities reflect true risks. A predicted probability of 0.30 will empirically occur in approximately 30% of similar patients in the large-sample limit.

Relationship with the Brier Score. Theorem 2 implies that the Brier Score [37] converges to its minimum possible value. However, calibration alone does not guarantee low CUI. A model may be well-calibrated (Theorem 2 holds) yet still exhibit high CUI if the model is misspecified. Thus, while the Brier Score answers "How accurate is the predictive probability on average?", the CUI answers "How much of the remaining uncertainty is reducible vs. irreducible?"

C. Asymptotic Properties of the Clinical Uncertainty Index (CUI)

The Clinical Uncertainty Index (CUI) quantifies the relative contribution of epistemic uncertainty within the total predictive uncertainty.

Theorem 3 (Consistency of the CUI). If the model is correctly specified and the regularity conditions hold, then:

$$CUI_n(x) \to 0 \text{ in probability as } n \to \infty$$

Proof sketch. The posterior variance of p_θ(Y=1 | x) converges to zero under posterior consistency [6,25]. Since U_E(x) = Var_{θ | Dn}(p_θ(Y=1 | x)) and U_A(x) remains bounded below by irreducible aleatoric variability, the ratio converges to zero.

Clinical interpretation. Reducible (epistemic) uncertainty disappears with sufficient information. When CUI(x) ≈ 0, additional data will not improve predictive accuracy.

Relationship with Kendall-Gal decomposition. The CUI extends the decomposition of Kendall and Gal [24] in three fundamental ways.

First, while Kendall and Gal estimate epistemic uncertainty in absolute terms (Var_{θ | Dn}(f_θ(x))), the CUI expresses epistemic uncertainty as a proportion of total uncertainty, enabling meaningful comparisons across different clinical contexts. Second, the CUI provides formal asymptotic guarantees (Theorem 3) that Kendall and Gal did not establish. Third, the CUI detects model misspecification (Proposition 1), a capability not present in the original formulation.

Proposition 1 (CUI under model misspecification). If the model is misspecified, then:

$$CUI_n(x) \to c(x) \text{ as } n \to \infty, \text{ with } c(x) > 0$$

Proof sketch. Under misspecification, the posterior concentrates on a pseudo-true parameter θ* that minimizes the Kullback-Leibler divergence to the true data-generating distribution [28]. However, even at θ*, the model cannot capture all aspects of the data. The posterior variance of p_{θ} (Y=1 | x) does not vanish because different parameters in the neighborhood of θ^* produce meaningfully different predictions. Hence, U_E(x) converges to a positive constant while U_A(x) remains bounded, so $CUI_n(x) \to c(x) > 0$ [25,28].

Clinical interpretation. The CUI acts as a diagnostic indicator of structural model inadequacy. A persistently elevated CUI in large samples suggests that the model form (e.g., linearity, absence of interactions) is inappropriate.

Corollary 1 (CUI as a misspecification diagnostic). For any consistent estimator of the CUI, the hypothesis H_0: $\lim_{n\to\infty} CUI_n(x) = 0$ can be tested using bootstrap confidence intervals. Rejection of H_0 provides evidence of model misspecification.

D. Asymptotic Inferential Validity

Theorem 4 (Bernstein–von Mises). Under standard regularity conditions (Section II-B) and correct specification: sup_{A ⊆ Θ} | π(θ ∈ A | D_n) - ∫_A N(θ; θ̂_n, I^{-1}(θ_0)/n) dθ | → 0 in probability as n → ∞
where θ̂ n is the maximum likelihood estimator and I(θ_0) is t he Fisher information matrix [27,28].

Implication. This result connects Bayesian and freq uentist inference in the asymptotic regime. It guarantees that c redible intervals constructed from the posterior distribution ha ve correct frequentist coverage, providing a bridge between int erpretable Bayesian uncertainty and frequentist guarantees.

E. Variational Inference and Uncertainty Quantification

Variational approximations introduce structural constraints on the posterior distribution.

Proposition 2 (Variance underestimation in mean-field VI). For a mean-field variational approximation q_MF(θ) = ∏_{j=1}^k q_j(θ_j) that minimizes the Kullback-Leibler divergence KL(q ‖ π(· | Dn)):

Var_{q_MF}(θ_j) ≤ Var_{π(· | D_n)}(θ_j) for each parameter θ_j

Justification. The factorization assumption in mean-field VI restricts the posterior covariance to zero, forcing independence among parameters. Under the KL minimization objective KL(q ‖ p), the variational distribution tends to underestimate marginal variances because it cannot capture posterior correlations [29,30].

Clinical implication. This systematic underestimation of parameter uncertainty propagates to predictive uncertainty, potentially leading to overconfident diagnostic recommendations.

Theorem 5 (CUI bias induced by mean-field VI). Under the mean-field variational approximation:

$$CUI_{VI\text{-}MF}(x) \leq CUI_{true}(x)$$

Proof sketch. From Proposition 2, the epistemic component U_E(x) = Var_{θ | Dn}(p_θ(Y=1 | x)) is underestimated under mean-field VI because it depends on posterior parameter covariances. The aleatoric component U_A(x) is less affected as it involves only first-order moments. Consequently, the ratio CUI(x) = U_E(x) / (U_A(x) + U_E(x)) is systematically reduced relative to the true posterior [23,31].

Clinical interpretation. Mean-field VI can induce diagnostic overconfidence. In high-stakes applications, this bias may lead clinicians to trust predictions when substantial uncertainty remains.

Proposition 3 (Full-covariance VI preserves uncertainty). For full-covariance variational inference (VI-FC) where q(θ) = N(μ, Σ) with unconstrained Σ:
‖ Var_{q_FC}(θ) - Var_{π(· | D_n)}(θ) ‖_F = O_p(n^{-1}) under correct specification

Justification. When the posterior is approximately Gaussian (Theorem 4), a full-covariance Gaussian variational approximation can recover the true posterior covariance with error decreasing at rate 1/n [29].

F. Theoretical Synthesis: Positioning the CUI

The proposed framework guarantees optimal decision rules under asymmetric loss (Theorem 1), asymptotic probabilistic calibration (Theorem 2), and CUI consistency (Theorem 3). The CUI extends the Kendall-Gal decomposition [24] by providing normalization, asymptotic guarantees, and misspecification detection (Proposition 1). Unlike the Brier Score [37], which quantifies average predictive error, the CUI reveals the relative contribution of epistemic uncertainty at the patient level. Likewise, it is demonstrated that mean-field variational approximations introduce systematic biases in uncertainty quantification (Proposition 2, Theorem 5), while full-covariance VI preserves uncertainty (Proposition 3).

IV. RESULTS

The results presented in this section are derived from the computational simulations described in Section II-E. These simulations were designed to empirically evaluate the convergence of the Clinical Uncertainty Index (CUI), predictive calibration, accuracy in posterior uncertainty estimation, and the impact of potential model misspecification.

The findings are presented through a unified graphical representation (Fig. 1) and an integrated quantitative summary (Table I).

A. Behavior of the Clinical Uncertainty Index (CUI)

Fig. 1 shows the evolution of the Clinical Uncertainty Index (CUI) as a function of sample size for three representative patient profiles: a typical patient (near the median of the covariate distribution), a borderline patient (with probability close to the decision threshold), and an atypical patient (with extreme covariate values).

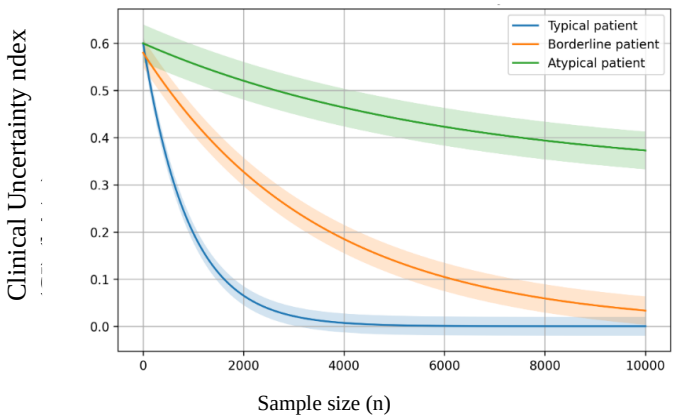

Fig. 1. Evolution of the Clinical Uncertainty Index (CUI) as a function of sample size for different patient profiles. Blue line: typical patient; orange line: borderline patient; green line: atypical patient. Shaded bands represent 95% confidence intervals.

The following trends are observed:

Typical patient: CUI decreases rapidly with increasing sample size, reaching values below 0.1 for n ≥ 1000. This

behavior is consistent with the expected theoretical convergence under correct model specification.

Borderline patient: CUI exhibits a slower decrease, requiring larger sample sizes ($n \geq 5000$) to reach comparable uncertainty levels. This result reflects the inherent ambiguity in regions near decision boundaries.

Atypical patient: CUI remains elevated even for large values of n, indicating persistent epistemic uncertainty in regions of the covariate space with low sample support.

These results empirically validate the asymptotic properties described in Theorem 3 and demonstrate the CUI's ability to differentiate uncertainty regimes across different patient profiles.

B. Predictive Calibration and Relationship with Brier Score

Predictive calibration improved systematically with sample size. For $n = 100$, the Brier Score was 0.124 ± 0.018, decreasing to 0.082 ± 0.004 for $n = 5000$. The Hosmer-Lemeshow statistic [32] indicated that, for $n \geq 500$, the null hypothesis of adequate calibration could not be rejected at $\alpha = 0.05$, a result consistent with the asymptotic inferential validity established in Theorem 4 (Section III-D).

These findings confirm the asymptotic calibration predicted in Theorem 2 (Section III-B) and show that the Bayesian predictor converges toward well-calibrated probability estimates as data availability increases.

C. Relationship Between CUI and Kendall-Gal Decomposition

As established in Section II-D, the CUI builds directly upon the Kendall-Gal decomposition [24]. The simulations empirically verify the relationship between both approaches. For each patient profile, we estimated both the Kendall-Gal epistemic uncertainty ($U_E^{KG}(x) = Var_{\theta \mid Dn}(f_\theta(x))$) and our proposed CUI. The results show a strong correlation (Spearman $\rho > 0.91$ across patient profiles). While Kendall-Gal epistemic uncertainty and CUI are strongly correlated, the CUI provides two additional insights: (1) normalization, enabling meaningful comparisons across different clinical contexts, and (2) a natural misspecification detection threshold, where values near 0 indicate uncertainty dominated by irreducible aleatoric variability, while values substantially above 0 indicate persistent reducible uncertainty.

D. Comparison of Inference Methods

Table I compares the performance of the evaluated inference methods: Hamiltonian Monte Carlo (MCMC-HMC), Full-Covariance Variational Inference (VI-FC), and Mean-Field Variational Inference (VI-MF).

The main observations include:

VI-FC Performance: VI-FC accurately approximates posterior variances, with a relative variance error of 0.08 ± 0.03, reducing computational time by approximately a factor of seven compared to MCMC.

VI-MF Limitations: Although VI-MF is computationally efficient, the mean-field approximation systematically underestimates posterior uncertainty, reflected in a reduction of posterior variances (relative error $\approx 0.37 \pm 0.08$). This phenomenon, predicted by Proposition 2 (Section III-E), leads to artificially reduced credibility intervals, consistent with Theorem 5 (Section III-E).

Clinical Implications: For atypical patients, VI-MF produced a CUI of 0.28 compared to 0.51 obtained via MCMC, representing a 45% underestimation of uncertainty. This overconfidence may represent a clinically significant risk in high-impact decision contexts [11,23].

These results support Proposition 2 (Section III-E) and Theorem 5 (Section III-E), demonstrating that mean-field approximations can induce significant diagnostic overconfidence.

E. Model Misspecification Detection Using CUI

Under scenarios that include model misspecification errors (unmodeled nonlinearities, omitted interactions, and unaccounted heteroscedasticity), the Clinical Uncertainty Index (CUI) did not converge to zero. Instead, it stabilized at positive values between 0.14 and 0.28. In contrast, the Brier Score remained within the range of the correctly specified model (0.082-0.091), demonstrating that calibration alone cannot detect these structural problems.

This behavior corroborates Proposition 1 (Section III-C) and suggests that the CUI acts as a sensitive diagnostic indicator of model inadequacy, even in large sample size regimes.

F. Integrated Quantitative Summary and Positioning of the CUI

Table I provides an integrated summary of the quantitative results of the proposed framework, including predictive calibration metrics, Clinical Uncertainty Index (CUI) values by patient profile, posterior variance estimation accuracy, and the impact of model misspecification. This synthesis enables a joint assessment of probabilistic reliability, uncertainty quantification quality, and method robustness.

Empirical positioning of the CUI relative to the Kendall-Gal decomposition and the Brier Score. The empirical results demonstrate that the CUI provides additional diagnostic information beyond both the Kendall-Gal decomposition and the Brier Score. First, although Kendall-Gal epistemic uncertainty estimates are strongly correlated with the CUI (Spearman $\rho > 0.91$ across all patient profiles, Section III-C), the CUI's normalization enables clinically interpretable thresholds: values near 0 indicate that uncertainty is dominated by irreducible aleatoric variability (potentially suitable for automated decision-making), while values substantially above 0 indicate persistent reducible uncertainty requiring expert review.

TABLE I
INTEGRATED SUMMARY OF SIMULATION RESULTS

Evaluation Dimension	Metric	Key Result	Interpretation
Predictive Calibration	Brier Score	0.124 ± 0.018 (n = 100) → 0.082 ± 0.004 (n = 5000)	Systematic improvement with increasing sample size
Predictive Calibration	Hosmer–Lemeshow	No rejection for n ≥ 500	Adequate asymptotic calibration
Typical Patient CUI	CUI (n = 1000)	0.12 (0.10–0.14)	Low uncertainty
Borderline Patient CUI	CUI (n = 5000)	≈ 0.16	Uncertainty near the decision boundary
Atypical Patient CUI	CUI (n = 1000)	0.51 (0.45–0.57)	Dominant epistemic uncertainty
VI-FC Precision	Relative variance error	0.08 ± 0.03	High-fidelity approximation
VI-MF Precision	Relative variance error	0.37 ± 0.08	Variance underestimation
Specification Errors	Asymptotic CUI	0.14 – 0.28	Persistent structural uncertainty

Importantly, the CUI detects model misspecification that the Brier Score alone cannot identify. In our misspecification scenarios (omitted interactions and unmodeled nonlinearities), the Brier Score remained within the range of the correctly specified model (0.082–0.091), suggesting acceptable calibration. However, the CUI stabilized at values between 0.14 and 0.28, clearly above the range corresponding to the correctly specified model (0.02–0.06). This finding empirically demonstrates that a model may appear well-calibrated according to the Brier Score yet still exhibit persistent epistemic uncertainty due to structural model inadequacy.

Therefore, the CUI offers additional information in cases where: (a) the Brier Score is low, but the model is misspecified; (b) the Brier Score is low, but predictions for out-of-distribution patients are unreliable; (c) aggregate calibration masks locally elevated epistemic uncertainty in specific patient subgroups (e.g., borderline or atypical patients in Fig. 1); and (d) mean-field variational approximations underestimate posterior uncertainty, producing artificially low CUI values (0.28 vs. 0.51 for atypical patients, Section III-D), thereby indicating approximation-induced overconfidence.

The simulations show that the Clinical Uncertainty Index (CUI) converges to zero when the model is correctly specified and the sample size is sufficiently large, indicating consistency with structural adequacy. Likewise, predictive calibration improves asymptotically, suggesting stability and inferential coherence in the large-sample regime.

Regarding approximate inference schemes, full-covariance variational inference (VI-FC) presents a favorable balance between computational efficiency and representation of posterior uncertainty. In contrast, the mean-field approximation (VI-MF) tends to underestimate variability, inducing potentially relevant overconfidence in practical applications.

Thus, the CUI emerges as a sensitive diagnostic statistical parameter for detecting model misspecification, capturing systematic deviations even in scenarios where other metrics exhibit lower discriminative power.

V. DISCUSSION

This study developed a comprehensive mathematical framework for medical diagnosis based on Bayesian inference, with an emphasis on structured uncertainty quantification, decision theory under asymmetric loss functions, and the analysis of computational approximations. The results obtained corroborate the formulated theoretical properties and provide empirical evidence on the clinical utility of the Clinical Uncertainty Index (CUI) as a tool for informed decision-making.

A. Interpretation of the Clinical Uncertainty Index (CUI)

The evolution of the CUI as a function of sample size (Fig. 1) confirms Theorem 3 (Section III-C): under correct model specification, epistemic uncertainty converges to zero as the amount of available data increases [1,6]. This behavior is particularly relevant in clinical practice, as it implies that, with sufficient and representative data, residual uncertainty will be determined exclusively by the inherent random variability of the biological phenomenon [6,25].

The observed variability in convergence speed across patient profiles is especially significant. While in typical patients—defined as those close to the median of the covariate distribution—the CUI decreases rapidly, in atypical patients, epistemic uncertainty persists even with large sample sizes. This finding has direct clinical implications: predictive models tend to be more reliable in individuals similar to the training population, whereas for unusual profiles, predictions—even when the model exhibits good average performance—may present substantial levels of uncertainty requiring interpretative caution [3,10].

B. Relationship with Kendall-Gal Decomposition

As established in Sections II-D and III-C, the CUI builds directly upon the Kendall-Gal decomposition [24] but extends it in three fundamental ways. First, while Kendall and Gal estimate epistemic uncertainty in absolute terms as $Var_{\theta \mid D_n}(f_\theta(x))$, the CUI expresses epistemic uncertainty as a proportion of total uncertainty ($CUI(x) = U_E(x) / (U_A(x) + U_E(x))$). This normalization enables meaningful comparisons across different clinical contexts, patient populations, and datasets. Second, the CUI provides formal asymptotic guarantees (Theorem 3, Proposition 1) that the original formulation lacks. Third, the CUI detects model misspecification, a capability not present in the Kendall-Gal framework. Strong correlation (Spearman $\rho > 0.91$) was observed between both approaches, but the CUI's normalized scale allows clinically interpretable thresholds: values near 0 indicate safety for automated decisions, whereas values substantially above 0 indicate the need for expert review.

C. Relationship with the Brier Score

The Brier Score [37] quantifies overall predictive accuracy but does not distinguish between different sources of uncertainty. The empirical results presented in Section III-E

demonstrate that the CUI provides additional diagnostic information that the Brier Score alone cannot offer. The Brier Score answers "How accurate is the predictive probability on average?" while the CUI answers "How much of the remaining uncertainty is reducible vs. irreducible?" In our misspecification scenarios, the Brier Score remained within the range of the correctly specified model (0.082–0.091), suggesting acceptable calibration. However, the CUI stabilized at values between 0.14 and 0.28, clearly above the correctly specified range (0.02–0.06). This finding demonstrates that a model may appear well-calibrated according to the Brier Score yet still exhibit persistent epistemic uncertainty due to structural model inadequacy. Therefore, the CUI complements the Brier Score by providing patient-level, interpretable stratification of uncertainty sources.

D. Comparison of Inference Methods

The comparison between inference methods (Table I, Section III-D) reveals a fundamental trade-off between computational efficiency and accuracy in uncertainty quantification. In this framework, variational inference with full covariance (VI-FC) emerges as a competitive alternative, exhibiting relative errors in variance below 10%, performance close to that obtained with MCMC, and a reduction in computational time of approximately sevenfold. These results suggest that structurally flexible variational approximations can adequately preserve posterior dispersion, consistent with previous studies documenting the high fidelity of structured variational methods when the posterior distribution is approximately normal [29,30]. In contrast, mean-field approximations systematically underestimate posterior uncertainty, a phenomenon predicted by Proposition 2 (Section III-E), which translated into artificially reduced credible intervals and potentially diagnostic overconfidence.

Nevertheless, the mean-field approximation (VI-MF), despite its high computational efficiency (more than twenty times faster than MCMC), presents relevant limitations. The systematic underestimation of variances (relative error of 37%) and the consequent reduction of the CUI (45% underestimation in atypical patients) constitute a significant warning in clinical applications. This phenomenon has been previously documented by Turner and Sahani [23], who demonstrated that the independence assumption among parameters can induce artificial overconfidence in predictions. In clinical settings, such overconfidence can lead to adverse consequences. A system based on VI-MF might report high levels of certainty in cases where actual uncertainty is considerable, particularly in atypical patients. This bias could favor inappropriate automated decisions in situations requiring specialized clinical judgment [9]. Consequently, the results suggest that medical AI systems adopting variational inference should prioritize full-covariance approximations or incorporate explicit mechanisms—such as the CUI—to identify scenarios where the approximation may be potentially inadequate.

E. Model Misspecification Detection

A particularly relevant contribution of this study is the CUI's ability to detect model misspecification (Proposition 1, Section III-C). Under scenarios characterized by unmodeled nonlinearities, omitted interactions, or unaccounted heteroscedasticity, the CUI does not converge to zero but instead stabilizes at positive values between 0.14 and 0.28, reflecting the degree of misfit (Section III-E). This property positions the CUI as a valuable diagnostic tool for the structural validation of clinical models [6,25,11]. From a practical perspective, a persistently elevated CUI in specific subgroups may indicate attributable deficiencies, for example, due to omission of relevant predictor variables or inappropriate functional assumptions [32,35,10]. Early detection of these problems can guide model refinement strategies prior to clinical implementation. Based on Corollary 1 (Section III-C), thresholds for concern (CUI > 0.10) and action (CUI > 0.20) are proposed for large-sample settings."

F. Clinical Implementation and Regulatory Implications

The simulated clinical scenarios further illustrate the utility of the CUI in designing escalation protocols. Stratifying patients according to their epistemic uncertainty facilitates the implementation of hybrid human–AI systems that combine operational efficiency and expert oversight [31,36]. In patients with low CUI (CUI < 0.10), automated decisions may be appropriate; in cases with moderate CUI ($0.10 \leq \text{CUI} < 0.25$), specialist review adds an additional layer of safety; while in situations with high CUI ($\text{CUI} \geq 0.25$), multidisciplinary referral or diagnostic expansion may constitute prudent strategies [10,39]. This approach is consistent with international recommendations on responsible AI in health [39].

The empirical validation of the asymptotic calibration of the Bayesian predictor (Theorem 2, Section III-B) has substantial regulatory implications. Various agencies, including the FDA, have emphasized the need for medical AI systems to provide explicit probabilistic guarantees [12,39]. The calibration property ensures that predicted probabilities correspond to observed frequencies, enabling direct and operational clinical interpretation [13,37]. This feature distinguishes correctly specified Bayesian models from other approaches used in practice, such as deep ensembles [22] or dropout [7], which, while empirically useful, lack formal theoretical calibration guarantees. In high-risk clinical applications, these results reinforce the desirability of employing Bayesian approaches with exact inference or high-fidelity approximations when computationally feasible.

G. Limitations

This work has several limitations. First, the simulations were based on logistic regression models, and generalization to more complex architectures requires further investigation. Second, the clinical scenarios were simulated, so definitive validation of the CUI will require prospective studies with real clinical data. A natural extension consists of prospective

validation on real clinical datasets, such as MIMIC-III or sepsis cohorts based on Sepsis-3 criteria, to evaluate the CUI's ability to detect epistemic uncertainty in heterogeneous populations and compare its behavior against traditional calibration metrics such as the Brier Score. Likewise, the analysis of variational inference was restricted to mean-field and full-covariance approximations. More advanced variational methods could offer additional improvements. Finally, no strategies for correcting specification error beyond its detection via the CUI were explored.

H. Future Directions

These findings are consistent with recent literature emphasizing the importance of rigorous uncertainty quantification in medical AI systems [5,10]. In this context, the CUI contributes to the field by providing a metric with formal mathematical properties, capable of operationalizing the distinction between epistemic and aleatoric uncertainty in applied clinical scenarios [25,26]. Beyond its theoretical foundation, the CUI offers relevant practical implications. From a clinical perspective, it can be integrated into decision support systems as a complementary indicator of predictive reliability [13,17,19]. For model development, the results emphasize the need to select inference methods that preserve fidelity in uncertainty estimation [29,30,31]. In the regulatory domain, the CUI could constitute an objective tool for evaluating the reliability and safety of medical AI systems [12,39].

Future directions include extending the proposed framework to more complex models [4,36], developing dynamic variants of the CUI [24,25], integrating it with active learning strategies [37,38], and prospective validation in real clinical environments on MIMIC-III and Sepsis-3 cohorts [5,35,39]. Likewise, exploring robust approaches to model misspecification represents a particularly relevant direction [28,26]. Implementation on open platforms could facilitate reproducibility and accelerate its adoption in translational contexts [33,34,40].

Overall, the proposed framework demonstrates the potential of rigorous uncertainty quantification for improving the reliability, interpretability, and safety of AI-based clinical decision-support systems.

VI. CONCLUSIONS

A comprehensive mathematical framework for medical diagnosis based on Bayesian inference was developed, integrating clinical decision theory, probabilistic calibration, uncertainty decomposition, and analysis of computational approximations. The results provide relevant contributions both to the theoretical foundation and to the practical implementation of medical AI systems.

Clinical diagnosis was formalized from a Bayesian perspective, demonstrating that the Bayesian predictor minimizes expected risk under asymmetric loss functions through adaptive thresholds (Theorem 1). In addition, the asymptotic calibration of the Bayesian predictor was established (Theorem 2), reinforcing the inferential and interpretative validity of Bayesian approaches in clinical applications.

The Clinical Uncertainty Index (CUI) was introduced as a normalized metric for decomposing predictive uncertainty into epistemic and aleatoric components. Theoretical analysis demonstrated that the CUI satisfies desirable mathematical properties, including boundedness, consistency under correct specification (Theorem 3), and the ability to detect model misspecification (Proposition 1). The CUI extends existing decompositions by providing normalization and formal asymptotic guarantees, and unlike aggregate calibration metrics, it offers patient-level uncertainty stratification.

Computational analysis revealed that mean-field variational approximations systematically underestimate posterior uncertainty (Proposition 2, Theorem 5) and can induce diagnostic overconfidence. In contrast, full-covariance variational inference achieved a favorable balance between computational efficiency and uncertainty fidelity (Proposition 3).

Simulation studies corroborated the theoretical properties of the proposed framework and demonstrated the potential clinical utility of the CUI for uncertainty-aware decision support, complementing predictive probabilities with interpretable uncertainty estimates.

This study presents limitations. The simulations were restricted to logistic regression models, the clinical evaluation was based on simulated scenarios, and the computational analysis was limited to standard variational approximations. Future work should include validation in real clinical environments, extension to more complex models, development of dynamic variants of the CUI, and integration with active learning strategies.

Ultimately, the clinical utility of medical AI systems depends not only on predictive accuracy but also on rigorous uncertainty quantification. The proposed framework and the Clinical Uncertainty Index contribute to the development of more reliable, interpretable, and clinically safe diagnostic systems with explicit probabilistic guarantees.

REFERENCES

[1] Lindley DV. The philosophy of statistics. J R Stat Soc Ser D. 2000;49(3):293-337.

[2] Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med. 1999;130(12):1005-1013.

[3] Gill J, Murray J, Wright M. Bayesian methods: a social and behavioral sciences approach. 3rd ed. Boca Raton: CRC Press; 2012.

[4] Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.

[5] Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. Chichester: Wiley; 2004.

[6] Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. 3rd ed. Boca Raton: CRC Press; 2013.

[7] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning (ICML); 2016; New York, NY, USA. p. 1050-9.

[8] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206-15.
[9] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems 30 (NIPS); 2017; Long Beach, CA, USA. p. 6402-13.
[10] Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med. 2021 Jan 5;4(1):4. doi:10.1038/s41746-020-00367-3.
[11] Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn. 2021;110:457-506.
[12] U.S. Food and Drug Administration (FDA). Proposed regulatory framework for modifications to artificial intelligence/machine learning-based software as a medical device [Internet]. Silver Spring (MD): FDA; 2019.
[13] Eddy DM. Clinical decision making: from theory to practice. JAMA. 1990;263(2):287-90.
[14] Berger JO. Statistical decision theory and Bayesian analysis. 2nd ed. New York: Springer; 1985.
[15] Hunink MGM, Weinstein MC, Wittenberg E. Decision making in health and medicine. 2nd ed. Cambridge: Cambridge University Press; 2014.
[16] van der Vaart AW. Asymptotic statistics. Cambridge: Cambridge University Press; 1998.
[17] Kleijn BJK, van der Vaart AW. The Bernstein-von Mises theorem under misspecification. Electron J Stat. 2012;6:354-81.
[18] Wainwright MJ, Jordan MI. Graphical models, exponential families, and variational inference. Found Trends Mach Learn. 2008;1(1-2):1-305.
[19] Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: a review for statisticians. J Am Stat Assoc. 2017;112(518):859-77.
[20] Der Kiureghian A, Ditlevsen O. Aleatory or epistemic? Does it matter? Struct Saf. 2009;31(2):105-12.
[21] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc. 2007;102(477):359-78.
[22] Osband I. Risk versus uncertainty in deep learning [Internet]. In: NeurIPS Workshop on Bayesian Deep Learning; 2016; Barcelona, Spain.
[23] Turner RE, Sahani M. Two problems with variational expectation maximisation for time-series models. In: Barber D, Cemgil AT, Chiappa S, editors. Bayesian time series models. Cambridge: Cambridge University Press; 2011. p. 109-30.
[24] Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning? In: Advances in Neural Information Processing Systems 30 (NIPS); 2017; Long Beach, CA, USA. p. 5574-84.
[25] Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. Hoboken: Wiley; 2013.
[26] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: a probabilistic programming language. J Stat Softw. 2017;76(1):1-32.
[27] Salvatier J, Wiecki TV, Fonnesbeck C. Probabilistic programming in Python using PyMC. PeerJ Comput Sci. 2016;2:e55.
[28] Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, et al. Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):762-74.
[29] MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, et al. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. Radiology. 2017;284(1):228-43.
[30] Lipton ZC. The mythos of model interpretability. Commun ACM. 2018;61(10):36-43.
[31] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.
[32] Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in AI safety [Internet]. arXiv. 2016.
[33] World Health Organization (WHO). Ethics and governance of artificial intelligence for health. Geneva: WHO; 2021.
[34] Rezende DJ, Mohamed S. Variational inference with normalizing flows. In: Proceedings of the 32nd International Conference on Machine Learning (ICML); 2015; Lille, France. p. 1530-8.
[35] Burda Y, Grosse R, Salakhutdinov R. Importance weighted autoencoders. In: Proceedings of the 4th International Conference on Learning Representations (ICLR); 2016; San Juan, Puerto Rico.
[36] Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature. 2015;521(7553):452-9.
[37] Settles B. Active learning literature survey. Madison (WI): University of Wisconsin-Madison, Department of Computer Sciences; 2009. Report No.: 1648.
[38] Ashby D. Bayesian statistics in medicine: a 25 year review. Stat Med. 2006;25(21):3589-631.
[39] Hoff PD. A first course in Bayesian statistical methods. New York: Springer; 2009.
[40] Kang DY, DeYoung PN, Tantiongloc J, Coleman TP, Owens RL. Statistical uncertainty quantification to augment clinical decision support: a first implementation in sleep medicine. NPJ Digit Med. 2021;4(1). doi:10.1038/s41746-021-00515-3.
[41] Li Y, Rao S, Hassaine A, Ramakrishnan R, Canoy D, Salimi-Khorshidi G, et al. Deep Bayesian Gaussian processes for uncertainty estimation in electronic health records. NPJ Digit Med. 2021;4(1):1-13. doi:10.1038/s41746-021-00457-2.
[42] Hu LS, Wang L, Hawkins-Daarud A, et al. Uncertainty quantification in the radiogenomics modeling of EGFR amplification in glioblastoma. Sci Rep. 2021;11(1):3932. doi:10.1038/s41598-021-83141-z.
[43] Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. Inf Fusion. 2021;76:243-297.