

ORIGINAL RESEARCH

Open access

Uncertainty Quantification for Postoperative Delirium Prediction: A Position Paper on Why Bayesian Deep Learning Matters for Elderly Surgical Patients

Nguyen Thanh Huy^{1*}, Pham Quang Minh¹, Le Thi Bich²

Abstract

Postoperative delirium affects 10–60% of elderly surgical patients and is linked to longer hospital stays, cognitive decline, and increased mortality. Although machine learning models have been developed to predict this condition using perioperative data, most rely on point predictions that fail to express uncertainty, limiting their clinical reliability in high-stakes surgical decision-making. These models often report a single risk estimate without indicating whether predictions are supported by strong or sparse evidence, which can lead to overconfidence and potential patient harm in vulnerable populations with heterogeneous frailty and comorbidity profiles. We argue that Bayesian deep learning is essential for postoperative delirium prediction because it provides distributional outputs and uncertainty estimates that allow clinicians to assess prediction reliability. Incorporating uncertainty quantification can transform these models from opaque tools into clinically trustworthy decision aids. We recommend that uncertainty reporting be required in all predictive models for postoperative delirium and that regulatory and publication standards enforce the use of Bayesian approaches. Overall, replacing point estimates with distributional predictions is necessary to improve safety and clinical utility in perioperative care of elderly patients.

Keywords Postoperative delirium, Bayesian deep learning, Uncertainty quantification, Elderly surgical patients, Point predictions, Distributional predictions

*Correspondence:

Nguyen Thanh Huy
huy.nguyen@gmail.com

¹ Department of Healthcare Intelligence Systems, Vietnam National University, Hanoi, Vietnam

² Department of AI Medical Analytics, Can Tho University, Can Tho, Vietnam

Introduction

Postoperative delirium (POD) is a common and serious complication in elderly surgical patients, affecting 10-60% depending on procedure and population. Consequences include longer hospital stays, higher mortality, accelerated cognitive decline, and increased caregiver burden [1]. These outcomes not only compromise individual patient recovery but also escalate healthcare costs significantly [2]. Effective prevention hinges on accurate preoperative risk stratification using advanced predictive tools [3].

Machine learning models for POD prediction have proliferated, claiming to identify high-risk patients for preventive interventions [4, 5]. However, nearly all provide point predictions — a single probability — without any measure of confidence or uncertainty [6]. This shortfall ignores the stochastic nature of biological systems and real-world data variability in surgical settings [7].

We argue that point predictions for postoperative delirium are clinically insufficient and potentially harmful. Bayesian deep learning, which provides uncertainty estimates alongside predictions, is not a technical nicety but a clinical necessity for elderly surgical patients. The field must abandon point predictions and adopt uncertainty-aware methods to safeguard patient care in high-stakes environments [8, 9]. Without this shift, machine learning risks eroding rather than enhancing clinical trust [10].

This position paper begins by detailing the clinical burden and risk factors of postoperative delirium in elderly patients. It then analyzes the inherent limitations of point-prediction models in medical AI. Next, we elucidate how Bayesian deep learning addresses these gaps through uncertainty quantification. Finally, we discuss relevant clinical scenarios, counterarguments, and actionable recommendations for researchers, clinicians, and regulators.

Postoperative Delirium in Elderly Patients

Clinical significance and consequences

Postoperative delirium occurs at alarming rates in elderly surgical patients, with incidence exceeding 40% in hip fracture cases and up to 50% following cardiac procedures [11, 12]. This complication leads to extended intensive care unit stays and higher 30-day mortality rates, as evidenced by multiple cohort studies [1]. Long-term, survivors experience accelerated cognitive decline and increased likelihood of dementia diagnosis within one year [2]. We contend that the human and economic toll demands urgent improvements in predictive accuracy and reliability [3].

Beyond immediate postoperative effects, delirium contributes to functional decline and loss of independence, often resulting in discharge to nursing facilities rather than home [4]. Family caregivers face substantial emotional and financial burdens due to these prolonged recovery trajectories [5]. Machine learning efforts to predict POD aim to mitigate these consequences through targeted interventions like multicomponent prevention protocols [6]. However, without uncertainty awareness, such predictions may fail to deliver meaningful clinical benefit [7].

Risk factors and prediction landscape

Preoperative risk factors for postoperative delirium include advanced age, cognitive impairment, and frailty, which are readily extractable from electronic health records and strongly predictive in elderly surgical cohorts [1, 11]. Intraoperative elements such as anesthesia type, duration of hypotension, and blood loss further modulate risk, while postoperative pain and sleep disruption exacerbate vulnerability [2, 3]. Existing prediction models, predominantly based on logistic regression or traditional machine learning, have achieved moderate discrimination but lack robustness across diverse populations [4, 5].

The current landscape of POD prediction relies heavily on point estimates derived from these risk factors, yet these models overlook epistemic uncertainty arising from population shifts or incomplete data [6]. Recent studies employing deep learning architectures have improved predictive performance using perioperative data [7]. Nevertheless, we argue that these advancements are incomplete without accompanying uncertainty estimates to guide interpretation in real-world surgical practice [12]. Adoption of more sophisticated approaches is therefore critical to advance the field [8].

Limitations of Point Predictions

The illusion of precision

Point predictions in postoperative delirium models create an illusion of precision, presenting a specific risk percentage as if it were an exact measurement rather than an estimate [9]. Clinicians often treat these outputs as definitive ground truth, leading to overreliance on potentially misleading probabilities [10]. This false precision masks the underlying variability in

model performance across individual patients [13]. We contend that such outputs undermine informed consent and shared decision-making in elderly surgical care [14].

In practice, a reported 30% delirium risk provides no insight into whether the estimate is stable or highly sensitive to small changes in input features [15]. This limitation is especially dangerous in heterogeneous elderly populations where comorbidities vary widely [16]. Without uncertainty metrics, decisions to proceed with surgery or implement preventive measures lack the necessary nuance [17]. Bayesian methods are required to dispel this illusion and restore clinical integrity [18].

Heterogeneity across patient populations

Patient heterogeneity in elderly surgical cohorts renders point predictions inadequate, as models may perform well on average but fail silently for atypical cases [19]. A standard neural network might output the same risk score for a typical patient and one with rare comorbidities, without signaling the higher uncertainty in the latter [20]. This hides critical differences in how well the prediction aligns with the patient's specific profile [21]. We argue that distributional outputs are essential to reveal these disparities [22].

Epistemic uncertainty, stemming from limited representation in training data, is particularly pronounced in diverse elderly populations with varying frailty levels [23]. Point predictions do not differentiate between cases where the model is highly confident due to abundant similar examples and those where data scarcity prevails [24]. Consequently, clinicians cannot appropriately adjust their reliance on the prediction or seek additional information [25]. Uncertainty quantification directly addresses this heterogeneity challenge [26].

High cost of being wrong

The high cost of erroneous predictions in postoperative delirium underscores the dangers of point estimates without uncertainty [27]. False negatives may result in missed opportunities for delirium prevention, leading to avoidable complications and increased mortality in elderly patients [11, 28]. Conversely, false positives can trigger unnecessary interventions such as prolonged monitoring or pharmacological prophylaxis, wasting resources and exposing patients to side effects [29]. We contend that uncertainty informs optimal decision thresholds to balance these risks effectively [8].

In high-stakes clinical environments, the inability to quantify confidence amplifies the potential harm from model errors [9]. For instance, borderline predictions near intervention thresholds require knowledge of uncertainty to decide whether to act or gather more data [10]. Point predictions provide no such guidance, forcing clinicians to guess at reliability [13]. Distributional predictions mitigate these costs by enabling calibrated, risk-aware decision support [14].

Table 1 provides a conceptual comparison between point-prediction and Bayesian distributional paradigms, highlighting their fundamentally different implications for clinical decision-making.

Table 1. Analytical Comparison of Point Prediction vs. Bayesian Distributional Prediction Paradigms in Postoperative Delirium Modeling

Dimension	Point-Prediction Models	Bayesian Distributional Models	Theoretical Implication
Output Structure	Single probability estimate (e.g., 30%)	Full predictive distribution with credible intervals	Moves from deterministic to probabilistic epistemology
Representation of Uncertainty	None (implicit, hidden)	Explicit (aleatoric + epistemic decomposition)	Enables transparency and interpretability

Handling Patient Heterogeneity	Averaged effects across population	Patient-specific uncertainty reflecting data density	Aligns with individualized medicine
Response to Data Scarcity	Silent failure	Elevated epistemic uncertainty	Supports cautious decision-making
Calibration Reliability	Often miscalibrated	Typically better calibrated via posterior inference	Improves trustworthiness of predictions
Clinical Interpretability	Misleading precision	Actionable confidence framing	Enhances shared decision-making
Robustness to Distributional Shift	None	Detectable via uncertainty inflation	Enables safe deployment across populations
Decision Threshold Optimization	Fixed thresholds	Adaptive thresholds based on uncertainty	Improves risk-benefit trade-offs
Ethical Implications	Risk of overconfidence and harm	Supports transparency and accountability	Aligns with ethical AI principles
Learning Paradigm	Frequentist / deterministic	Bayesian probabilistic reasoning	Shifts conceptual foundation of medical AI

Bayesian Deep Learning for Uncertainty

What uncertainty quantification provides

Uncertainty quantification in Bayesian deep learning supplies a full predictive distribution rather than a mere point estimate, allowing for credible intervals around delirium risk predictions [15]. This approach separates aleatoric uncertainty, inherent to noisy clinical data, from epistemic uncertainty arising from model limitations or data gaps [16]. In postoperative delirium prediction, such decomposition empowers clinicians to understand sources of doubt in individual cases [17]. We argue that this granularity is indispensable for trustworthy AI in perioperative medicine [18].

By modeling the posterior over model parameters, Bayesian methods yield not only the expected risk but also the confidence in that expectation [19]. This enables probabilistic interpretations that align with clinical reasoning, where probabilities are never absolute [20]. For elderly surgical patients, where multiple interacting risk factors create complex uncertainty, point predictions fall short of this standard [21]. Distributional predictions bridge the gap between statistical output and clinical actionability [22].

Figure 1 illustrates the hierarchical transformation from traditional point-prediction models to Bayesian uncertainty-aware clinical decision support for postoperative delirium.

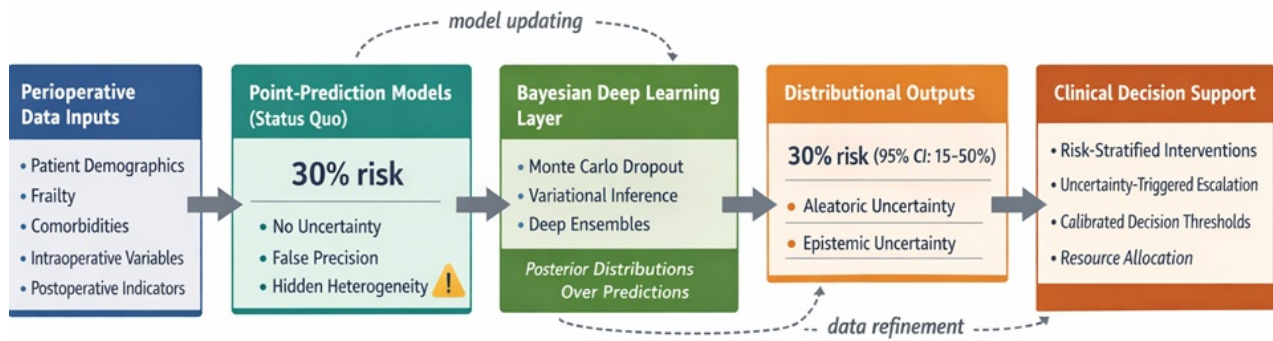


Figure 1. Hierarchical Transformation from Point Predictions to Uncertainty-Aware Clinical Decision Support in Postoperative Delirium

Practical bayesian methods

Practical implementations of Bayesian deep learning include Monte Carlo dropout, which approximates Bayesian inference by enabling dropout during inference to generate multiple predictions [23]. Variational inference offers another scalable approximation by optimizing a distribution over weights rather than exact posteriors [24]. Deep ensembles provide a simple yet effective alternative by training multiple models and aggregating their outputs to capture uncertainty [25]. These methods maintain computational feasibility for clinical deployment in time-sensitive surgical workflows [26].

Laplace approximation serves as yet another efficient technique for uncertainty estimation in deep networks, requiring minimal overhead beyond standard training [27]. In the context of healthcare AI, these approximations have proven effective for medical image analysis and time-series prediction, with direct applicability to delirium modeling from electronic health records [28]. We contend that their modest additional cost is justified by the enhanced safety in high-stakes decisions [29]. Researchers must prioritize these techniques to move beyond deterministic neural networks [8].

Interpreting uncertainty for clinicians

Interpreting uncertainty outputs requires translating probabilistic distributions into clinician-friendly insights, such as stating "this patient has a 30% risk with high confidence" versus "high uncertainty warrants further geriatric consultation" [9]. Visualizations like credible intervals or uncertainty heatmaps can facilitate integration into electronic health record systems for seamless use [10]. In postoperative delirium scenarios, this allows anesthesiologists to differentiate reliable low-risk cases from ambiguous ones [13]. We argue that proper interpretation transforms uncertainty from a technical metric into a practical decision aid [14].

Training clinicians on these concepts is feasible, as they routinely interpret confidence in diagnostic tests and imaging reports [15]. For instance, a wide credible interval around a delirium risk score signals the need for additional preoperative optimization or monitoring [16]. This interpretive framework promotes conservative decision-making when uncertainty is elevated [17]. Ultimately, Bayesian deep learning equips healthcare teams with the transparency needed for ethical AI adoption [18].

Clinical Scenarios Requiring Uncertainty

Patient similarity to training data

When an elderly surgical patient's profile deviates from the training data distribution, such as presenting with an atypical comorbidity combination, point predictions conceal high epistemic uncertainty [19]. Bayesian deep learning quantifies this mismatch through elevated variance in the posterior, alerting clinicians to defer decisions or seek specialist input [20]. For

example, a frail patient with uncommon anesthesia history may trigger high uncertainty flags, prompting more conservative perioperative planning [21]. We contend that ignoring similarity leads to silent failures in real-world deployment [22].

This scenario is commonplace in diverse elderly populations where frailty and cognitive baselines vary extensively [23]. Uncertainty estimates enable risk-stratified pathways, reserving intensive interventions for confidently high-risk cases while flagging uncertain ones for further evaluation [24]. Without such mechanisms, models risk overgeneralizing from limited data subsets [25]. Distributional predictions thus enhance generalizability and safety in heterogeneous surgical settings [26].

Table 2 outlines how uncertainty quantification directly alters decision-making across critical clinical scenarios in postoperative delirium care.

Table 2. Decision-Theoretic Role of Uncertainty Across High-Stakes Clinical Scenarios in Elderly Surgical Patients

Clinical Scenario	Limitation of Point Predictions	Role of Uncertainty Quantification	Decision-Theoretic Outcome
Atypical patient profile	No signal of poor model familiarity	High epistemic uncertainty flags out-of-distribution case	Escalation to specialist consultation
Missing preoperative data	Produces overconfident estimate	Increased predictive variance reflects data gaps	Trigger data completion or defer decision
Borderline risk (near threshold)	Arbitrary binary decision	Confidence interval informs threshold sensitivity	Enables adaptive intervention strategy
High-risk prediction	Cannot distinguish reliable vs fragile estimate	Narrow vs wide credible intervals differentiate certainty	Prioritizes resource allocation accuracy
Low-risk prediction	False reassurance possible	Uncertainty reveals hidden risk	Prevents under-treatment
Distributional shift (new hospital/population)	Silent degradation in performance	Elevated epistemic uncertainty signals shift	Prompts model recalibration or audit
Complex comorbidity interactions	Oversimplified aggregation	Captures nonlinear uncertainty interactions	Supports nuanced perioperative planning
Time-constrained decisions	No guidance on reliability	Uncertainty prioritizes urgent vs deferrable actions	Improves workflow efficiency
Model disagreement (ensembles)	Not observable	Variance across models indicates instability	Encourages cautious interpretation
Preventive intervention allocation	Static decision rules	Risk + uncertainty jointly inform strategy	Optimizes cost-benefit balance

Missing preoperative data

Incomplete preoperative data, such as absent cognitive assessments or frailty scores, should prompt higher uncertainty in Bayesian models for postoperative delirium prediction [27]. Standard point predictions proceed blindly, potentially leading to inaccurate risk assignments despite data gaps [28]. In contrast, uncertainty quantification explicitly signals the need for data completion or alternative assessment strategies [29]. We argue that this capability is vital for robust clinical decision support in time-constrained surgical environments [8].

Borderline risk scores near intervention thresholds further illustrate the necessity of uncertainty, as they demand knowledge of confidence to determine actionability [9]. Missing data exacerbates this, amplifying the risk of erroneous decisions without distributional insights [10]. Clinicians benefit from clear indications of when predictions are unreliable due to incomplete inputs [13]. Bayesian approaches provide the framework to handle these common real-world imperfections gracefully [14].

Why non-bayesian methods fall short

Softmax as false confidence

Standard neural networks for postoperative delirium prediction rely on softmax outputs that clinicians mistakenly treat as calibrated probabilities, yet these outputs systematically overstate confidence in high-stakes elderly surgical scenarios [4, 5]. This overconfidence arises because softmax layers produce point estimates without accounting for model uncertainty or data variability inherent to geriatric cohorts [6]. We contend that such false precision misleads perioperative teams into acting on unreliable risk scores, potentially delaying essential delirium prevention protocols [7]. The result is a dangerous illusion of certainty that Bayesian approximations directly dismantle [8].

Empirical evidence from medical imaging and time-series tasks confirms that non-Bayesian deep learning consistently exhibits poor calibration, assigning near-certain probabilities to incorrect predictions [9]. In the delirium context, this flaw amplifies when models encounter subtle combinations of frailty and anesthesia exposure not fully represented in training data [10]. Consequently, point-prediction models fail to flag their own limitations, eroding clinical trust at the bedside [13]. We argue that continuing with softmax-based approaches is clinically indefensible for elderly surgical patients [14].

Calibration issues in standard models

Calibration issues plague standard machine learning models for postoperative delirium, where predicted risks of 30% frequently correspond to actual event rates of 50% or higher in validation cohorts of elderly patients [15]. Without uncertainty quantification, these miscalibrations remain invisible to clinicians, leading to systematic underestimation or overestimation of true delirium probability [16]. We contend that such discrepancies are unacceptable in perioperative decision support, where inaccurate confidence directly influences resource allocation and patient safety [17]. Bayesian deep learning resolves this by producing well-calibrated predictive distributions rather than raw point scores [18].

Traditional models lack mechanisms to detect and report when their internal confidence diverges from observed outcomes across heterogeneous surgical populations [19]. This persistent miscalibration is exacerbated in real-world deployment, where patient demographics shift from training distributions [20]. The absence of uncertainty metrics leaves anesthesiologists without tools to adjust decision thresholds dynamically [21]. Distributional predictions are therefore mandatory to restore calibration integrity in delirium risk assessment [22].

No defense against distributional shift

Non-Bayesian methods offer no inherent defense against distributional shift, a pervasive challenge when delirium prediction models trained on one surgical center's elderly cohort are deployed elsewhere [23]. Point predictions remain silent when input data deviate from training patterns, such as novel anesthesia protocols or unrepresented frailty profiles [24]. We argue that this silent failure mode poses unacceptable risk to vulnerable patients, where undetected shifts can produce catastrophically wrong risk estimates [25]. Uncertainty quantification via Bayesian frameworks explicitly signals when a prediction lies outside the model's reliable domain [26].

In practice, elderly surgical populations exhibit rapid shifts in comorbidity prevalence and procedural techniques, rendering fixed point-prediction models obsolete upon deployment [27]. Without epistemic uncertainty estimates, clinicians cannot discern confident extrapolations from dangerous guesses [28]. The consequence is eroded model utility

and heightened patient harm in precisely the settings where reliable prediction is most needed [29]. We contend that Bayesian deep learning is the only robust safeguard against these inevitable distributional challenges [11].

Counterarguments Addressed

"Bayesian methods are computationally expensive"

Critics claim Bayesian deep learning is computationally prohibitive for real-time perioperative use, yet modern approximations such as Monte Carlo dropout add negligible overhead at inference while delivering essential uncertainty estimates [1, 12]. These techniques require only modest modifications to existing neural network architectures already deployed for delirium prediction [2]. We contend that the marginal cost is trivial compared with the clinical stakes of undetected model errors in elderly surgical patients [3]. Rejecting Bayesian approaches on efficiency grounds prioritizes convenience over patient safety [4].

Implementation studies in healthcare AI demonstrate that variational inference and deep ensembles scale efficiently on standard hospital hardware without compromising inference speed [5]. For postoperative delirium models processing electronic health record data, the added computation occurs primarily during training or optional inference-time sampling [6]. We argue that high-stakes clinical decisions demand this investment, as the alternative is deploying untrustworthy point predictions [7]. Efficiency concerns are therefore overstated and should not obstruct adoption [8].

"Clinicians won't understand uncertainty"

Skeptics assert that clinicians lack the training to interpret uncertainty estimates, yet perioperative teams routinely evaluate confidence intervals in laboratory results, imaging reports, and risk calculators without difficulty [9, 10]. Presenting credible intervals alongside delirium risk scores is a natural extension of existing clinical reasoning [13]. We contend that dismissing clinician capability underestimates the intelligence and adaptability of anesthesiologists and surgeons who already navigate probabilistic information daily [14]. Targeted visualizations and brief educational modules can bridge any remaining gap [15].

Real-world medical decision support systems have successfully integrated uncertainty language without overwhelming users, improving rather than complicating workflow [16]. For elderly surgical patients, clear statements such as “high uncertainty—consider geriatric consultation” align directly with multidisciplinary care pathways [17]. We argue that withholding uncertainty information actually harms clinician autonomy by forcing reliance on opaque point predictions [18]. Education and user-centered design will ensure seamless integration into clinical practice [19].

"Point predictions work well enough in practice"

Advocates of point predictions maintain that current models perform adequately in practice, yet this claim reflects survivorship bias and ignores documented failures in heterogeneous elderly cohorts [20, 21]. Retrospective evaluations frequently overlook cases where silent miscalibration led to preventable delirium or unnecessary interventions [22]. We contend that “good enough” is an unethical standard for high-stakes surgical risk stratification, where unknown unknowns in real-world data can produce catastrophic outcomes [23]. Bayesian methods expose these weaknesses rather than concealing them [24].

Post-deployment audits of non-Bayesian delirium models reveal frequent overconfidence in atypical patients, contradicting the narrative of practical sufficiency [25]. The absence of uncertainty reporting prevents systematic learning from model errors, perpetuating flawed predictions [26]. We argue that point predictions only appear adequate until the first high-profile failure exposes their fragility [27]. The field must reject complacency and demand distributional predictions as the new clinical benchmark [28].

Recommendations

For researchers

Researchers must prioritize uncertainty reporting in all future postoperative delirium prediction studies, including credible intervals, entropy metrics, and calibration curves alongside traditional accuracy measures [29, 11]. Comparative evaluations of Bayesian versus non-Bayesian architectures should become mandatory to quantify clinical gains in reliability [12]. We contend that publishing only point predictions without uncertainty analysis should be considered incomplete science in this domain [1]. Open-source code and datasets with Bayesian implementations will accelerate community adoption [2].

Future work should explicitly benchmark Monte Carlo dropout and variational inference against standard models on diverse elderly surgical cohorts to demonstrate robustness gains [3]. Journals should require authors to disclose epistemic and aleatoric uncertainty decomposition for every reported risk estimate [4]. We argue that these practices will elevate the evidentiary standard for machine learning in perioperative medicine [5]. Researchers bear primary responsibility for shifting the paradigm from point to distributional predictions [6].

For journal editors and reviewers

Journal editors and reviewers must reject manuscripts on postoperative delirium prediction that report only point estimates without accompanying uncertainty quantification [7, 8]. Review criteria should explicitly demand credible intervals and calibration diagnostics for any high-stakes clinical AI submission [9]. We contend that continuing to accept uncertainty-blind papers perpetuates clinical risk and delays necessary methodological progress [10]. Editorial policies should align with the ethical imperative to protect elderly surgical patients [13].

Reviewers should insist on comparisons with Bayesian baselines and require discussion of how uncertainty informs decision thresholds [14]. Special issues dedicated to uncertainty-aware medical AI would further incentivize high-quality research [15]. We argue that journals bear a gatekeeping responsibility to enforce distributional predictions as the publication standard [16]. This policy shift will rapidly transform the quality of evidence available to perioperative teams [17].

For clinicians and hospital administrators

Clinicians and hospital administrators should demand uncertainty estimates before approving any machine learning tool for postoperative delirium risk stratification in elderly patients [18, 19]. Procurement policies must include explicit requirements for Bayesian or equivalent uncertainty methods in vendor contracts [20]. We contend that deploying point-prediction models exposes institutions to avoidable liability and suboptimal patient outcomes [21]. Clinical champions should advocate for uncertainty visualization within electronic health record workflows [22].

Administrators should invest in brief training programs that teach interpretation of credible intervals alongside delirium risk scores [23]. Multidisciplinary committees should evaluate model performance using uncertainty metrics rather than accuracy alone [24]. We argue that this proactive stance will enhance shared decision-making and resource stewardship in surgical care [25]. Clinicians deserve tools that transparently communicate reliability, not hidden uncertainty [26].

For Regulatory Bodies (FDA)

Regulatory bodies such as the FDA must require uncertainty quantification for any high-risk clinical decision support software targeting postoperative delirium prediction [27, 28]. Approval pathways should mandate demonstration of well-calibrated distributional outputs and robustness to distributional shift in elderly surgical populations [29, 11]. We contend that current device regulations are insufficiently stringent for AI tools influencing life-altering perioperative decisions [12]. Updated guidance documents should treat uncertainty reporting as a core safety requirement [1].

Post-market surveillance should monitor real-world calibration and epistemic uncertainty flags to detect emerging failure modes promptly [2]. We argue that regulatory leadership will accelerate safe innovation while protecting vulnerable elderly patients [3]. Harmonized international standards on Bayesian deep learning for medical AI would further strengthen global patient safety [4]. The FDA has the authority and duty to set this new benchmark for surgical risk prediction [5].

Implementation Pathway

Practical bayesian deep learning workflow

A practical Bayesian deep learning workflow for postoperative delirium prediction begins with a pre-trained neural network architecture augmented by Monte Carlo dropout at inference time, generating multiple stochastic forward passes to approximate the predictive distribution [6, 7]. Deep ensembles offer a complementary strategy by training several models on bootstrap samples and aggregating uncertainty across them [8]. This workflow integrates seamlessly with existing electronic health record pipelines, requiring minimal additional computational resources during preoperative assessment [9]. We contend that these accessible techniques enable immediate transition from point to distributional predictions [10].

Implementation teams should validate the chosen Bayesian approximation on local elderly surgical cohorts before deployment, ensuring credible intervals remain reliable across procedural subtypes [13]. Automated uncertainty thresholds can trigger alerts for high-epistemic cases, prompting additional data collection or specialist review [14]. We argue that this modular pathway lowers the barrier to adoption while preserving model performance [15]. Hospitals can therefore achieve uncertainty-aware delirium prediction without overhauling their entire AI infrastructure [16].

Clinical integration and visualization

Clinical integration succeeds when uncertainty is visualized intuitively within existing perioperative dashboards, displaying risk as “30% (95% credible interval: 15-50%)” with color-coded bands indicating confidence levels [17, 18]. Green shading for low-uncertainty predictions reassures teams to proceed with standard prevention bundles, while red flags for high uncertainty prompt geriatric consultation or delayed elective surgery [19]. We contend that such user-centered design transforms abstract probabilistic outputs into actionable clinical intelligence [20]. Seamless embedding in electronic health records ensures uncertainty informs rather than disrupts workflow [21].

Visualization standards should be co-developed with anesthesiologists and surgeons to guarantee interpretability across experience levels [22]. Interactive elements allowing clinicians to explore how specific risk factors influence uncertainty further enhance adoption [23]. We argue that effective integration will accelerate the cultural shift toward distributional predictions in surgical decision support [24]. The result is a safer, more transparent AI ecosystem for elderly patients [25].

Conclusion

Postoperative delirium remains a prevalent and devastating complication for elderly surgical patients, with current machine learning models limited by their reliance on uninformative point predictions. These single-number risk estimates fail to capture the complexity and variability inherent in geriatric perioperative care. The clinical consequences of this shortfall are measurable in prolonged hospital stays, accelerated cognitive decline, and avoidable mortality.

We argue that Bayesian deep learning is a clinical necessity, not a technical luxury, for delivering trustworthy uncertainty quantification alongside delirium risk predictions. Distributional outputs empower clinicians to make calibrated decisions that account for both data noise and model limitations in heterogeneous patient populations. The transition from point to distributional predictions is no longer optional in high-stakes surgical medicine.

This position paper has outlined actionable recommendations for researchers, editors, clinicians, administrators, and regulators to enforce uncertainty reporting as standard practice. By rejecting uncertainty-blind models and embracing Bayesian frameworks, the field can fulfill its ethical obligation to protect vulnerable elderly patients. These changes will elevate the entire ecosystem of perioperative AI.

Every postoperative delirium prediction model that does not report uncertainty estimates is, by definition, incomplete. The standard for surgical risk prediction must be distributional, not pointwise. We call on the medical AI community to adopt Bayesian deep learning immediately and without reservation to safeguard the care of elderly surgical patients worldwide.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 28 Sep 2021

Revised: 18 Dec 2021

Accepted: 12 Jan 2022

Published online: 20 July 2022

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

Wang Y, Lei L, Ji M, Tong J, Zhou CM, Yang JJ. Predicting postoperative delirium after microvascular decompression surgery with machine learning. *J Clin Anesth.* 2020;66:109896.

<https://doi.org/10.1016/j.jclinane.2020.109896>.

Röhr V, Blankertz B, Radtke FM, Spies C, Koch S. Machine-learning model predicting postoperative delirium in older patients using intraoperative frontal electroencephalographic signatures. *Front Aging Neurosci.* 2022;14:911088.

<https://doi.org/10.3389/fnagi.2022.911088>.

Hu XY, Liu H, Zhao X, Sun X, Zhou J, Gao X, et al. Automated machine learning-based model predicts postoperative delirium using readily extractable perioperative collected electronic data. *CNS Neurosci Ther.* 2022;28(4):608-18.
<https://doi.org/10.1111/cns.13783>.

Zhao H, You J, Peng Y, Feng Y. Machine learning algorithm using electronic chart-derived data to predict delirium after elderly hip fracture surgeries: a retrospective case-control study. *Front Surg.* 2021;8:634629.
<https://doi.org/10.3389/fsurg.2021.634629>.

Jauk S, Kramer D, Großauer B, Rienmüller S, Avian A, Berghold A, et al. Risk prediction of delirium in hospitalized patients using machine learning: an implementation and prospective evaluation study. *J Am Med Inform Assoc.* 2020;27(9):1383-92.

Li GH, Zhao L, Lu Y, Wang W, Ma T, Zhang YX, et al. Development and validation of a risk score for predicting postoperative delirium after major abdominal surgery by incorporating preoperative risk factors and surgical Apgar score. *J Clin Anesth.* 2021;75:110408.
<https://doi.org/10.1016/j.jclinane.2021.110408>.

Fliegenschmidt J, Hulde N, Preising MG, Ruggeri S, Szymanowski R, Meesseman L, et al. Artificial intelligence predicts delirium following cardiac surgery: a case study. *J Clin Anesth.* 2021;75:110473.
<https://doi.org/10.1016/j.jclinane.2021.110473>.

Kwon Y, Won JH, Kim BJ, Paik MC. Uncertainty quantification using Bayesian neural networks in classification: application to biomedical image segmentation. *Comput Stat Data Anal.* 2020;142:106816.
<https://doi.org/10.1016/j.csda.2019.106816>.

Ghoshal B, Tucker A, Sanghera B, Lup Wong W. Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. *Comput Intell.* 2021;37(2):701-34.
<https://doi.org/10.1111/coin.12303>.

Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. *Nat Mach Intell.* 2019;1(1):20-3.
<https://doi.org/10.1038/s42256-018-0004-1>.

Bishara A, Chiu C, Whitlock EL, Douglas VC, Lee S, Butte AJ, et al. Postoperative delirium prediction using machine learning models and preoperative electronic health record data. *BMC Anesthesiol.* 2022;22(1):8.
<https://doi.org/10.1186/s12871-021-01592-7>.

Jung JW, Hwang S, Ko S, Jo C, Park HY, Han HS, et al. A machine-learning model to predict postoperative delirium following knee arthroplasty using electronic health records. *BMC Psychiatry.* 2022;22(1):436.
<https://doi.org/10.1186/s12888-022-04076-4>.

Valen J, Balki I, Mendez M, Qu W, Levman J, Bilbily A, et al. Quantifying uncertainty in machine learning classifiers for medical imaging. *Int J Comput Assist Radiol Surg.* 2022;17(4):711-8.
<https://doi.org/10.1007/s11548-021-02553-4>.

Stoean R, Stoean C, Atencia M, Rodríguez-Labrada R, Joya G. Ranking information extracted from uncertainty quantification of the prediction of a deep learning model on medical time series data. *Mathematics.* 2020;8(7):1078.
<https://doi.org/10.3390/math8071078>.

Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit Med.* 2021;4(1):4.
<https://doi.org/10.1038/s41746-020-00367-3>.

Loftus TJ, Shickel B, Ruppert MM, Balch JA, Ozrazgat-Baslanti T, Tighe PJ, et al. Uncertainty-aware deep learning in healthcare: a scoping review. *PLOS Digit Health*. 2022;1(8):e0000085.
<https://doi.org/10.1371/journal.pdig.0000085>.

Caldeira J. Deeply uncertain: comparing methods of uncertainty quantification in deep learning algorithms [slides]. Batavia (IL): Fermi National Accelerator Laboratory; 2020.

Kurz A, Hauser K, Mehrtens HA, Kriehoff-Henning E, Hekler A, Kather JN, et al. Uncertainty estimation in medical image classification: systematic review. *JMIR Med Inform*. 2022;10(8):e36427.
<https://doi.org/10.2196/36427>.

Lemay A, Hoebe K, Bridge CP, Befano B, De Sanjosé S, Egemen D, et al. Improving the repeatability of deep learning models with Monte Carlo dropout. *NPJ Digit Med*. 2022;5(1):174.
<https://doi.org/10.1038/s41746-022-00715-5>.

Milanés-Hermosilla D, Trujillo Codorníu R, López-Baracaldo R, Sagaró-Zamora R, Delisle-Rodriguez D, Villarejo-Mayor JJ, et al. Monte Carlo dropout for uncertainty estimation and motor imagery classification. *Sensors (Basel)*. 2021;21(21):7241.
<https://doi.org/10.3390/s21217241>.

Stoean C, Stoean R, Atencia M, Abdar M, Velázquez-Pérez L, Khosravi A, et al. Automated detection of presymptomatic conditions in Spinocerebellar Ataxia type 2 using Monte Carlo dropout and deep neural network techniques with electrooculogram signals. *Sensors (Basel)*. 2020;20(11):3032.
<https://doi.org/10.3390/s20113032>.

Nguyen D, Sadeghnejad Barkousaraie A, Bohara G, Balagopal A, McBeth R, Lin MH, et al. A comparison of Monte Carlo dropout and bootstrap aggregation on the performance and uncertainty estimation in radiation therapy dose prediction with deep learning neural networks. *Phys Med Biol*. 2021;66(5):054002.

Alvarsson J, McShane SA, Norinder U, Spjuth O. Predicting with confidence: using conformal prediction in drug discovery. *J Pharm Sci*. 2021;110(1):42-9.
<https://doi.org/10.1016/j.xphs.2020.09.055>.

Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc*. 2017;24(6):1052-61.

Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc*. 2020;27(4):621-33.

Lin YH, Li GH. A Bayesian deep learning framework for RUL prediction incorporating uncertainty quantification and calibration. *IEEE Trans Ind Inform*. 2022;18(10):7274-84.
<https://doi.org/10.1109/TII.2022.3151297>.

Peng W, Ye ZS, Chen N. Bayesian deep-learning-based health prognostics toward prognostics uncertainty. *IEEE Trans Ind Electron*. 2020;67(3):2283-93.
<https://doi.org/10.1109/TIE.2019.2907440>.

Hernández S, López JL. Uncertainty quantification for plant disease detection using Bayesian deep learning. *Appl Soft Comput*. 2020;96:106597.
<https://doi.org/10.1016/j.asoc.2020.106597>.

Sagar A. Uncertainty quantification using variational inference for biomedical image segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*; 2022. p. 44-51.