Journal of the American Medical Informatics Association, 2026, **33(6)**, 1100–1111
https://doi.org/10.1093/jamia/ocag034
Advance access publication 31 March 2026
Research and Applications

Research and Applications

Predictive performance precision analysis in medicine: identification of low-confidence predictions at patient and profile levels (MED3pa I)

Olivier Lefebvre, BSc[1], Félix Camirand Lemyre, PhD[2], Jean-François Ethier, MDCM, PhD[3], Lyna Hiba Chikouche, MSc[4], Ludmila Amriou, MSc[4], Dan Poenaru, MD, MHPE, MA, PhD[5,6], Martin Vallières, PhD[1,7,*]

[1]Department of Computer Science, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada, [2]Department of Mathematics, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada, [3]Department of Medicine, Université de Sherbrooke, Sherbrooke, QC J1H 5N4, Canada, [4]Department of Higher Studies (CS), École nationale Supérieure d'Informatique, Alger, 16309, Algeria, [5]Department of Pediatric Surgery, Montreal Children's Hospital, Montreal, QC H4A 3J1, Canada, [6]Centre for Outcomes Research and Evaluation, Research Institute of the McGill University Health Centre, Montreal, QC H4A 3J1, Canada, [7]Medical Physics Unit, Department of Oncology, McGill University, Montreal, QC H3A 0G4, Canada

*Corresponding author: Martin Vallières, PhD, Medical Physics Unit, Cedars Cancer Centre - MUHC Glen Site, Room DS1.7137, 1001 boul. Décarie, Montreal, QC H4A 3J1, Canada (martin.vallieres@mcgill.ca)

Abstract

Objectives: Artificial Intelligence models are increasingly used in health care, yet global performance metrics can mask variations in reliability across individual patients or subgroups with shared attributes, called patient profiles. This study introduces predictive performance precision analysis in medicine (MED3pa), a method that identifies when models are less reliable, allowing clinicians to better assess model limitations.

Materials and Methods: We propose a framework that estimates predictive confidence using 3 combined approaches: individualized (IPC), aggregated (APC), and mixed predictive confidence (MPC). Individualized predictive confidence estimates confidence for each patient, APC assesses it across profiles, and MPC combines both. We evaluate our method on 4 datasets: 1 simulated, 2 public, and 1 private clinical dataset. Metrics by declaration rate curves show how performance changes when retaining only the most confident predictions, while interpretable decision trees reveal profiles with higher or lower model confidence.

Results: We demonstrate our method in internal, temporal, and external validation settings, as well as through a clinical example. In internal validation, limiting predictions to the 93% most confident cases improved sensitivity by 14.3% and the area under the receiver operating characteristic curve by 5.1%. In the clinical example, MED3pa identified a patient profile with high misclassification risk, demonstrating its potential for safer deployment.

Discussion: By identifying low-confidence predictions, our framework improves model reliability in clinical settings. It can be integrated into decision support systems to help clinicians make more informed decisions. Confidence thresholds help balance model performance with the proportion of patients for whom predictions are considered reliable.

Conclusion: Better leveraging confidence in model predictions could improve reliability and trustworthiness, supporting safer and more effective use in health care.

Key words: machine learning; predictive confidence; model reliability; model performance evaluation.

Introduction

Artificial intelligence (AI) has the potential to transform health care, with clinical models playing a key role in advancing precision medicine by helping with diagnosis and prognosis.[1–3] These models typically undergo internal training and validation, followed by external validation across diverse datasets.[1,4–6] However, as highlighted by multiple studies,[7–11] this standard pipeline does not guarantee generalizability due to variations in populations and contexts. Machine learning (ML) methods often assume congruence between training and application distributions; this assumption, however, is not always met in real-world applications. This discord is often attributed to dataset shifting, where the joint distribution of inputs and outputs differs between training and testing phases.[12] Another deficiency in AI models is their focus on estimating local averages by mapping the predictor space onto a latent dimension where their relationship with the desired outcome is accentuated. Although this approach optimizes global performance metrics, it may overlook underrepresented subpopulations and fail to account for individual variability.[13] As a result, predictions may be unreliable for some patients, meaning that the model's output is less likely

Received: September 2, 2025; **Accepted:** February 26, 2026

to accurately reflect the patient's true outcome. Moreover, predictive performance may vary between patients even if the model is well calibrated. Some patient profiles—defined as subgroups of patients with shared features such as age or comorbidities—may systematically exhibit higher prediction errors. Although patient profiles are defined by rule-based patient characteristics, model performance is used to identify which profiles are associated with systematically higher or lower predictive performance. It then becomes imperative to conduct a comprehensive analysis to identify and characterize such low-confidence predictions. Here, confidence reflects the expected reliability of the model's output for each patient, with lower confidence indicating a higher risk of incorrect prediction. This analysis should be performed prior to model deployment and maintained continuously as clinical contexts and data distributions evolve.

Multiple factors can cause a model to make erroneous predictions. Even when deployment data follow the same distribution as the training data, some cases lack discriminative information, and the inherent variability between outcomes and predictors cannot be reduced, which is known as aleatoric uncertainty. Epistemic uncertainty can also lead to incorrect prediction, arising from limitations within the model itself, including misspecification or insufficient representation of certain patient subgroups. As a result, the model may be correctly specified for the dominant subpopulation but misspecified for varying underrepresented subgroups or individuals. Even with correct model specification, predictions may remain uncertain for certain patient profiles, for example, logistic regression when predicted probabilities are near 0.5.

Figure 1 illustrates how these sources of uncertainty can manifest in practice. Figure 1A shows examples within the training data where the model $\widehat{P}_A$ is less accurate due to limited discriminative information. Figure 1B highlights how temporal or external datasets can diverge from the training distribution, demonstrating the importance of contextual validation. Even if the model $\widehat{P}_A$ is correctly specified for the training distribution (set A), its predictions are less reliable for patients with higher values of $\mathbf{x}_1$ (red rectangle) due to subgroup underepresentation. Temporal datasets refer to future data collected from the same source, while external datasets originate from different institutions or populations. Models developed in a specific context may degrade or fail to generalize, as seen in Vela et al.,[14] who observed temporal performance declines even in stable environments, reinforcing the need for continuous monitoring.

Several approaches have been proposed to improve predictive model reliability and interpretability, particularly under distribution shifts and performance disparities. Ginsberg et al.[12] introduced *Detectron*, a hypothesis test detecting harmful covariate shifts by analyzing an ensemble of constrained disagreement classifiers. Shashikumar et al.[15] proposed *COMPOSER*, a sepsis model using conformal prediction to flag unfamiliar cases as indeterminate, improving model deployment across diverse populations. Conformal prediction provides regions containing the true outcome with a specified confidence level.[16] Recent works extended conformal prediction to handle covariate shifting, enabling confident predictions even when testing data differ from training.[17–19] Conformal prediction provides guarantees on predictions by constructing calibrated prediction regions from past data, but it focuses on producing valid prediction sets rather than characterizing specific patients associated with higher uncertainty.

In contrast, uncertainty quantification (UQ) techniques are specifically designed to estimate predictive uncertainty, aiming to provide reliable predictions.[20–24] These methods typically estimate epistemic and/or aleatoric uncertainty, often during model training, and rely on assumptions on the underlying learning

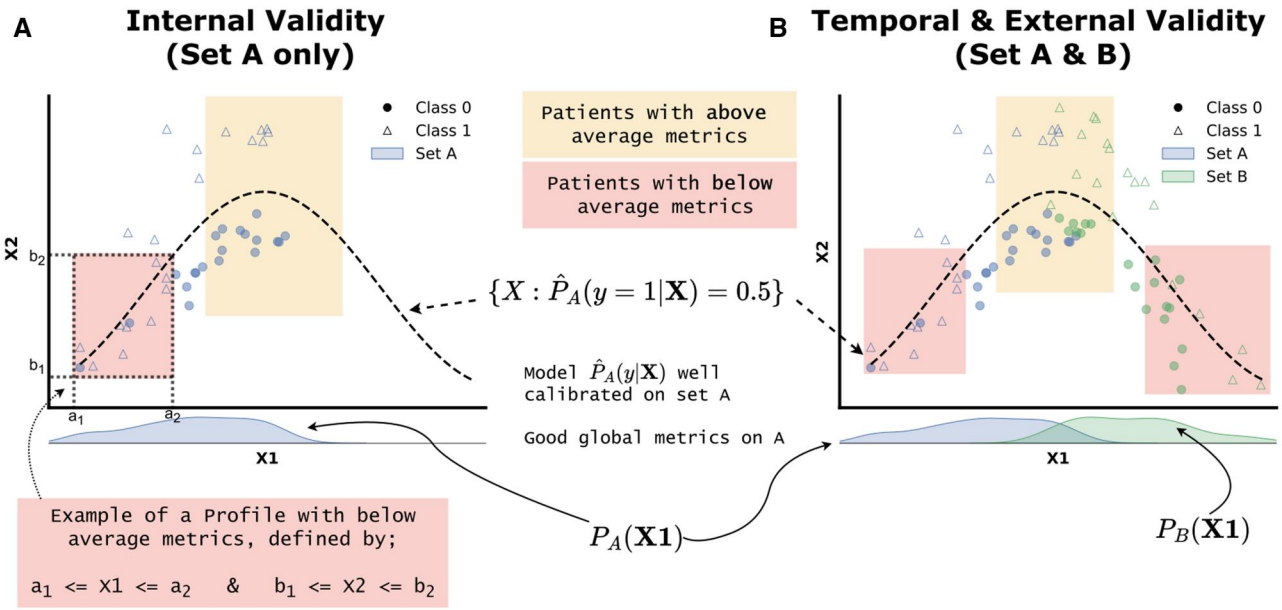

Figure 1. Illustration of model reliability across internal and temporal/external contexts. (A) Internal validity: scatter plot showing a classification model trained on 2 features (*X*1 and *X*2) to separate blue triangles and circles. The dashed curve represents the discriminant boundary of the model, defined by $\{X : \widehat{P}_A(y = 1 \mid X) = 0.5\}$, where $\widehat{P}_A(y \mid X)$ is trained and calibrated on set A. In the central region (yellow rectangle), classes are well separated and the model is more reliable. On the left (red rectangle), class overlap leads to lower model reliability. (B) Temporal and external validity: additional data points from set B (green) reveal a new region on the right (red rectangle) where the model becomes unreliable, due to an overlap between the 2 classes not present in the original data. The marginal distributions $P_A(X1)$ and $P_B(X1)$ illustrate a shift in feature *X*1 between sets A and B, contributing to changes in performance across profiles.

algorithm. As a result, many are not model-agnostic and primarily characterize uncertainty within the training data distribution. Furthermore, those uncertainty estimates do not necessarily reflect whether a trained model will produce an incorrect prediction on a given test instance,[25–27] and the quality of uncertainty estimation degrades under distribution shifts.[28,29]

An alternative line of work reframes uncertainty estimation as a prediction task, aiming to forecast if a model's prediction is likely to be erroneous on new inputs. These methods focus on predicting model performance rather than uncertainty intrinsic to the model or data. Most approaches formulate this task as a binary classification problem, predicting whether a model's output is correct for a given instance.[25,30–34] Framing error prediction as a binary outcome however ignores the magnitude of prediction errors. Using continuous error measures can instead emphasize highly erroneous high probability predictions, assigning lower confidence to predictions that the initial model makes with high confidence but are likely to be wrong.

The effectiveness of these methods is often evaluated using accuracy-rejection curves,[30,33,35,36] which measures the accuracy of a base classifier after progressively rejecting predictions with lowest confidence. Accuracy-based evaluation may however be misleading in certain settings, such as highly imbalanced datasets. Alternative rejection curves have been proposed based on other performance metrics, such as precision, recall or AUROC.[31,37,38]

Complementary to UQ, the hidden stratification literature investigates performance disparities across subgroups, allowing a more interpretable explanation of model limitations. Some approaches rely on manually defined subgroups,[39,40] which may restrain the efficiency of subgroups discovery. Several automatic methods have been proposed without requiring predefined rules; however, many of these are based on clustering representations in the latent space of neural networks,[41–45] restricting their applicability to neural network-based models. To address this limitation, QUEST[46] was proposed as a model-agnostic hidden stratification method that uses decision tree induction to identify easily interpretable subgroups associated with discrete levels of uncertainty. Discretizing uncertainty into a small number of bins may limit the granularity of uncertainty estimates, potentially resulting in less precise subgroup characterization.

In this work, we introduce the predictive performance precision analysis in medicine (MED3pa) method. Predictive performance precision analysis in medicine is a model-agnostic framework for the post hoc identification and characterization of potential failures of a preexisting predictive model, referred to as the BaseModel, which represents a well-trained and calibrated binary classification model. Predictive performance precision analysis in medicine identifies patient profiles for which the BaseModel exhibits reduced performance, supporting both predeployment evaluation and postdeployment monitoring. This article presents the first phase of MED3pa, focusing on identifying when and for whom a BaseModel is less reliable by detecting low-confidence predictions and revealing the most affected patient profiles. Future work will explore the underlying causes of these reliability gaps and explore strategies to improve reliability across patient profiles.

Objective

The overarching goal of MED3pa is not to exclude patients from care, but to improve the inclusivity and reliability of model predictions for *all* patients. A prediction may have a high probability of being incorrect not necessarily because the model is unreliable, but due to inherent uncertainty or variability in the data such as underrepresentation of certain profiles or high outcome variability. By identifying patients for whom a model may be less reliable, we aim to better understand and address its limitations. However, in this first work, we focus on identifying these patients and, for now, withholding predictions for them to explore how excluding uncertain cases can balance performance improvements with patient inclusion. Specifically, we aim to:

- develop a method to detect and characterize patient profiles associated with a high likelihood of inaccurate predictions, to provide insights on prediction reliability in clinical models;
- demonstrate the method on a simulated dataset, to validate its functionality in a controlled setting; and
- apply the method to real clinical datasets, to assess its effectiveness in a clinical context.

Intended uses

The MED3pa method offers a range of applications aimed at improving our understanding and utilization of a preexisting BaseModel, both for initial validation and continuous monitoring. While not exhaustive, key applications include:

1) *assessing applicability and reliability*: evaluate if the BaseModel is suitable for specific clinical contexts;
2) *investigating patient profiles*: identify and analyze patient profiles for which the BaseModel demonstrates reduced performance; and
3) *daily clinical application*: evaluate whether the BaseModel can be used for a given patient, ensuring predictive capabilities align with patient care objectives.

Methods

Predictive performance precision analysis

We consider a generic binary classification model that has been previously trained, calibrated, and validated, denoted as the BaseModel. Our goal is to evaluate one such model on new, unseen data from the deployment context, referred to as evaluation sets. These sets may also include additional patient characteristics, such as sociodemographics, which can later be used to explore hypotheses about performance disparities. While global metrics such as the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity are typically used to assess performance in new contexts, they may overlook patient-level variability. To address this, we propose an approach assessing predictive confidence at both individual patient and customizable patient profile levels within these evaluation sets, identifying where the model is less reliable. This allows exploration of performance improvements by selectively withholding predictions, while balancing metric gains and patient inclusion for which the BaseModel can be more safely applied.

To illustrate our method, we first use a simulated dataset to visualize key concepts (data generation process is described in Material S1A). Figure 2A shows the generated training/testing data and the classification regions produced by the BaseModel, which is a random forest classifier.

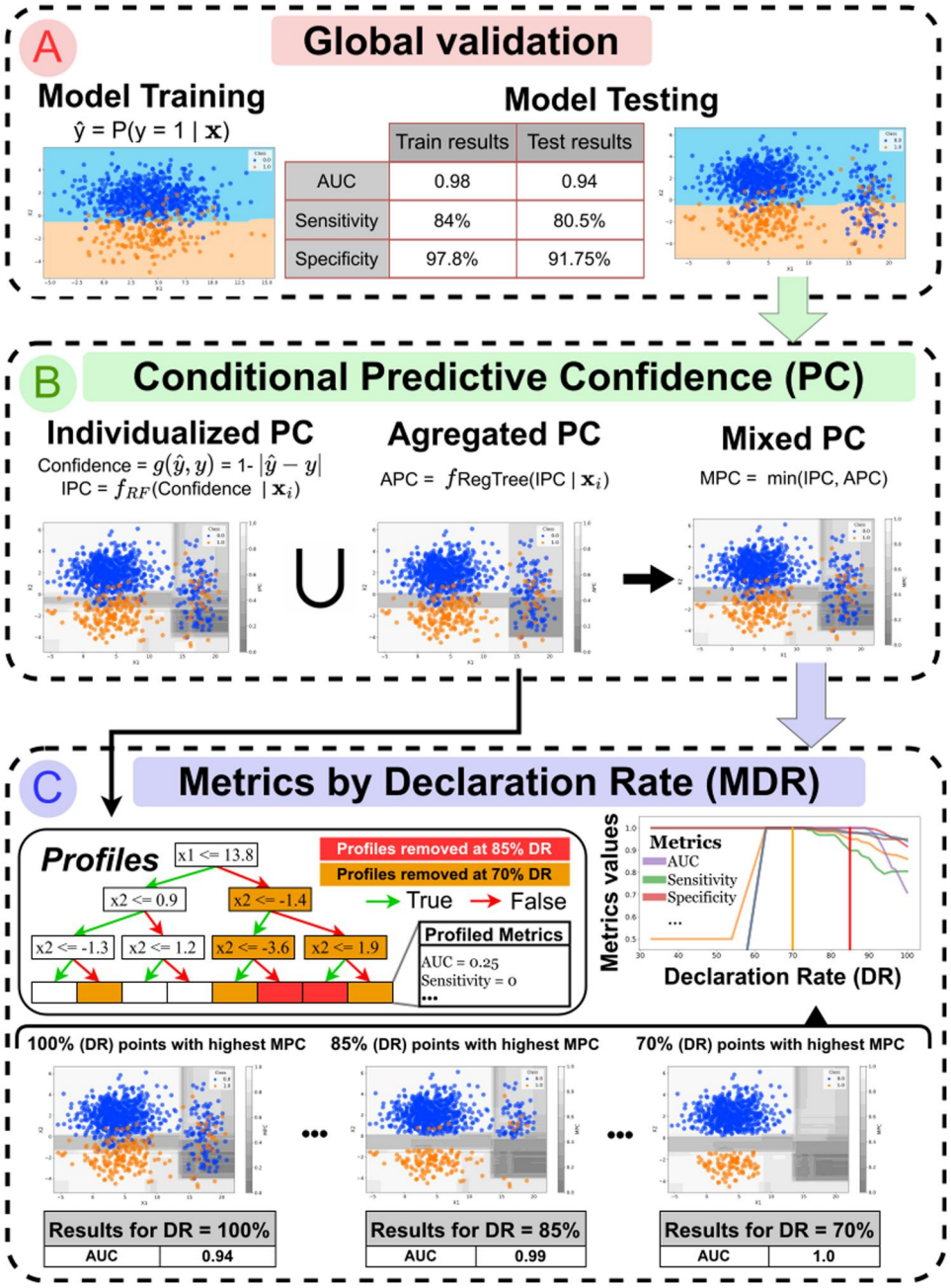

Figure 2. Overview of the proposed MED3pa method for identifying and managing predictive confidence using a simulated dataset. (A) Conventional global validation: a BaseModel is trained on a training dataset and evaluated on a separate test dataset, with performance reported as global metrics. (B) Conditional predictive confidence: example confidence estimates on the test set using the individualized predictive confidence (IPC), aggregated predictive confidence (APC), and mixed predictive confidence (MPC) models, each highlighting different aspects of predictive confidence. (C) Metrics by declaration rate (MDR): MDR curves visualize how performance metrics evolve as increasingly uncertain predictions (based on the MPC) are removed. The APC model is also used to identify interpretable patient profiles associated with low confidence. The bottom panel illustrates 3 examples of DR on the testing data, with DR = 100% where no patient is excluded, and DR = 70% where 30% of patients with lowest confidence are excluded from the use of the BaseModel.

Defining confidence

We define the predictive confidence of a BaseModel for patient i as $c_i = c(\mathbf{x}_i, y_i, \widehat{y}_i)$, where $\mathbf{x}_i$ denotes the patient's characteristics, y_i the true outcome, and $\widehat{y}_i$ the prediction generated by the BaseModel. This confidence metric quantifies the predictive performance of the BaseModel given patient characteristics. This confidence metric can be binary (eg, $c_i = \mathbb{1}_{\widehat{y}_i = y_i}$) or continuous (eg, $c_i = 1 - \|\widehat{y}_i - y_i\|$), where $\widehat{y}_i = \widehat{p}(\mathbf{x}_i)$ is the BaseModel's prediction and y_i is the true label. It may also be the predicted probability itself when the true outcome is unknown. However, our objective is to predict systematic model failures across patients, whereas the predicted probability primarily reflects the BaseModel scoring function and is less effective for failure prediction.[20,25] Empirical results supporting this observation are presented in Material S1B. Asymmetrical metrics can also be used to reflect classification thresholds or misclassification costs. The choice of confidence measure depends on the deployment context.

Individualized predictive confidence model

We then construct the individualized predictive confidence (IPC) model, which predicts an estimate $\hat{c}(\mathbf{x}_i)$ of the confidence metric for each new patient using features $\mathbf{x}_i$. To train IPC, we first apply the BaseModel to a subset of the evaluation set and compute for each patient i a confidence value $c(x_i)$ using our chosen confidence metric. The IPC model is then trained to predict the confidence using patient features. Individualized predictive confidence can be any classification model if the confidence metric is discrete, or a regression model if continuous. It is advisable to choose a model with similar or less capacity than the BaseModel to avoid overfitting confidence predictions. At deployment, IPC takes only features x_i to provide an estimate of the confidence value for each patient before applying the BaseModel

$$IPC(\mathbf{x}_i) = \hat{c}(\mathbf{x}_i),$$

where $\hat{c}(\mathbf{x}_i)$ is the predicted confidence. Figure 2B shows an example using a random forest regressor as IPC to predict a confidence score defined as $1 - \|\hat{y}_i - y_i\|$.

Aggregated predictive confidence model

Although the IPC model provides confidence estimates for individual patients, it does not provide any mechanism to identify groups of patients with similar confidence patterns. Identifying low-confidence profiles highlights disadvantaged profiles and helps investigate the sources of these discrepancies. To address this, we introduce the aggregated predictive confidence (APC) model. The APC model builds confidence profiles using supervised learning, where the target is the confidence values defined by the IPC model. A CART model (classification or regression depending on the confidence metric) is trained on the same subset of the evaluation set as the IPC, partitioning patients into profiles (represented by the leaves of the tree) that share similar expected confidence levels. For any patient profile S_i, APC outputs the average IPC-predicted confidence within that profile:

$$APC(S_i) = \frac{1}{|S_i|} \sum_{j \in S_i} IPC(\mathbf{x}_j).$$

For an individual patient, APC outputs the average confidence associated with the smallest profile containing that patient, corresponding to a leaf of the induced tree. The generation of these confidence profiles follows a rule-based, tree-structured approach similar to *QUEST*.[46] In APC, each profile is however associated with an overall confidence value, which is used to rank profiles by confidence rather than discretizing uncertainty into categories such as high vs low. Tree-based models are particularly valuable for producing interpretable profiles and computing APC values. Profile granularity is adjusted by changing tree depth and should be chosen contextually, reflecting the desired level of detail. Similar to QUEST,[46] it can also be determined by identifying the optimal tradeoff between the number of profiles and the accuracy of predicted confidence. Aggregated predictive confidence profiles are constructed using the patient characteristics in the evaluation set, which may include the features used by the BaseModel as well as additional patient variables. Figure 2B illustrates a regression tree-based APC, while Figure 2C shows the resulting profiles.

Mixed predictive confidence model

The IPC model provides patient-specific confidence estimates, but these can vary widely between patients. In contrast, the APC model offers more consistent estimates by grouping similar patients into profiles, although it may miss important individual differences. By combining both models, we can achieve a better balance between individual precision and overall stability in estimating prediction reliability for each patient. We propose the mixed predictive confidence (MPC) model, defined as

$$MPC(\mathbf{x}_i) = f(IPC(\mathbf{x}_i), APC(\mathbf{x}_i)),$$

where $f(IPC(\mathbf{x}_i), APC(\mathbf{x}_i))$ combines IPC and APC predictions. A simple approach is taking the minimum confidence, but other methods such as averaging could be used depending on context. Mixed predictive confidence balances individual estimates with broader profile-level trends. Figure 2B illustrates an MPC model using the minimum strategy, highlighting both low-confidence profiles and individual points.

Training procedures for IPC, APC, and MPC depend on intended use, whether to capture individual, profiled, or combined confidence. Further details are provided in Material S1C.

Metrics by declaration rate

Once confidence estimates are available and low-confidence patients are identified in the evaluation subset, it remains necessary to determine minimum acceptable confidence levels before deploying the BaseModel. To guide this, we introduce metrics by declaration rate (MDR), a generalization of the *accuracy-rejection curves*.[35,30] Metrics by declaration rate is computed exclusively on the remaining portion of the evaluation set not used for IPC/APC training (further detailed in Material S1C and D), ensuring that no instances used to train the confidence models are included in the reported metrics. It assesses predictive performance by gradually excluding the least confident predictions. The declaration rate (DR) refers to the proportion of patients retained. For each DR, we compute metrics (eg, AUC, accuracy). For example, at DR = 60%, 40% of patients with lowest predicted confidence are excluded, and the BaseModel is evaluated on the remaining 60%.

To construct MDR curves, patients are ranked by predicted confidence (via IPC, APC, or MPC). For each DR $\in [0, 1]$, patients with lowest confidence are incrementally removed, and metrics such as AUC, sensitivity, and specificity are computed in the remaining subset. Plotting metrics against DR yields MDR curves, visualizing how excluding the least confident predictions affects performance.

Ideally, well-calibrated confidence estimates produce monotonically increasing MDR curves as erroneous predictions are excluded. In practice, confidence models often identify some, but not all, uncertain predictions. Metrics by declaration rate curves may then increase, peak, and decline as further exclusion offers no benefit or harms performance. The optimal DR depends on the objective and should balance performance gains and patient inclusion.

Figure 2C shows MDR curves increasing up to DR = 60%, then decrease. For example, if the goal is to increase AUC, DR = 90% may be optimal as it marks where performance plateaus. It also illustrates how different DR values translate into removing the least confident predictions from the evaluation set, where predictive confidence is defined using the

MPC. It is computed as the minimum between the IPC and the APC, thereby enabling a tradeoff between individual-level and profile-level confidence. At DR = 100%, the scatter plot shows all data points, with the background color indicating predictive confidence (dark gray being the least confident and white the most confident). At DR = 85%, the 15% least confident predictions are removed, including the 2 patient profiles highlighted in red. At DR = 70%, the 30% least confident are removed, including profiles highlighted in red and orange.

Experimental setup

We applied our method to 2 classification problems: in-hospital mortality (IHM) using public datasets, and 1-year mortality (OYM) using a private dataset. Further experimental details and algorithmic components are provided in Material S1D.

In-Hospital mortality prediction

The first task predicts IHM for ICU patients 24 h after admission, using 17 predictors from the simplified acute physiology score II.[47] We used MIMIC-IV[48] and eICU-CRD,[49] 2 public datasets from US hospitals. MIMIC-IV contains electronic health records (EHRs) from Beth Israel Deaconess Medical Center (2008-2019), while eICU-CRD includes 208 hospitals (2014-2015). After applying preprocessing and exclusion criteria following Zeng et al.,[50] the final MIMIC-IV dataset included 16 271 admissions. In eICU-CRD, hospitals with less than 225 admissions were excluded to ensure sufficient sample size, resulting in 2583 admissions across 9 hospitals.

The IHM BaseModel is an XGBoost classifier trained on half of the 2008-2013 MIMIC-IV cohort ($n = 4476$; 16.8% positive rate). It was calibrated, with the classification threshold selected by maximizing Youden's J-index.[51] The confidence metric used is a sigmoidal function (Material S1E). Mixed predictive confidence was defined as the minimum confidence between IPC (random forest regressor) and APC (regression tree). The BaseModel was applied to the remaining 2008-2013 patients ($n = 4476$; 17.4% positive rate) for internal validation, the 2014-2019 cohort ($n = 7319$; 18.8% positive rate) for temporal validation, and to each eICU hospital ($n = 2583$; 18.1% positive rate) for external validation. Further details on data processing, variable transformations, exclusion criteria, evaluation sets construction, and descriptive statistics are provided in Material S1F.

One-year mortality prediction

The second task predicts OYM risk upon admission, using data from Laribi et al.,[52] comprising EHRs and administrative data from an integrated university hospital network in Sherbrooke, Quebec, Canada. We used the "AdmDemoDx" dataset, which includes 2 demographic variables, 10 admission characteristics, 85 comorbidities, and 147 admission diagnoses.[52]

Following Laribi et al.,[52] the BaseModel was trained on 2011-2017 admissions ($n = 148\ 587$; 13.9% positive rate) and validated on 2017-2021 admissions ($n = 61\ 310$; 12.5% positive rate). The BaseModel is a calibrated random forest classifier, with threshold optimization via Youden's index.[51] The same confidence metric as IHM was applied, with random forest regressor for IPC, regression tree for APC, and minimum confidence strategy for MPC. Additional dataset details and descriptive statistics are provided in Material S1G.

Results

All methods presented in this study are implemented in the public Python package MED3pa, which provides tools for estimating predictive confidence, evaluating model performance using MDR curves, and identifying error-prone patient profiles. Predictive performance precision analysis in medicine is available at: https://pypi.org/project/MED3pa/. The package includes a simple example based on synthetic data[53] generated from the Laribi et al. dataset. Additionally, the code used to generate all results and figures in this study is available at: https://github.com/MEDomics-UdeS/study_3pa. We evaluated the performance of MED3pa across all experimental settings considered in this study, assessing its ability to detect BaseModel prediction errors using AUC. Detailed results are provided in Material S1B.

Demonstration of the method via IHM prediction

We applied MED3pa to a BaseModel developed to predict IHM using MIMIC-IV[48] and eICU-CRD[49] datasets. First, we assessed internal validity on an MIMIC-IV subset sampled from the same distribution as training data. Next, we evaluated temporal validity using data from a later period, and finally tested external validity on the eICU-CRD dataset. While MDR curves can include any metric, we focused on 5 to simplify analysis and visualization. Accuracy is not shown in these curves, but as detailed in Material S1H, it improved consistently across scenarios. We therefore prioritized other metrics to better capture performance nuances.

Internal validation

Internal validation assesses whether a model is valid within the same context as training data. As shown in the MDR curves in Figure 3, removing patients with lowest predicted confidence led to improvements in all performance metrics except positive predictive value (PPV). Specificity remained stable for DR above 93% with a relative increase of 1.2% compared to DR = 100% ($\Delta_{rel}\%=1.2$), then improved, plateauing near 100% at DR=30%. Negative predictive value (NPV) behaved similarly, mostly increasing for DRs above 93% ($\Delta_{rel}\%=4.4$). However, the PPV continually decreased as DR was reduced, suggesting that maximizing PPV favors a DR of 100%.

Sensitivity and AUC peaked at DR = 93%, with relative improvements of 14.3% and 5.1% respectively, before declining at lower DRs. Excluding the 7% least confident predictions would enhance all metrics except PPV, while more aggressive filtering would benefit only specificity and NPV.

Temporal validation

After deployment, model performance can degrade due to factors such as changes in patient populations or clinical workflows. Temporal validation ensures the model remains reliable as hospital characteristics evolve.

As shown in Figure 4, specificity and NPV improved consistently by removing low-confidence patients, while other metrics tended to decline. For example, if the clinical objective is to maintain the same specificity as internal validation ($\text{specificity}^{IV}_{DR=93\%}=0.838$), this would be achieved with a DR of 91%. However, for other objectives, retaining all patients may be preferable.

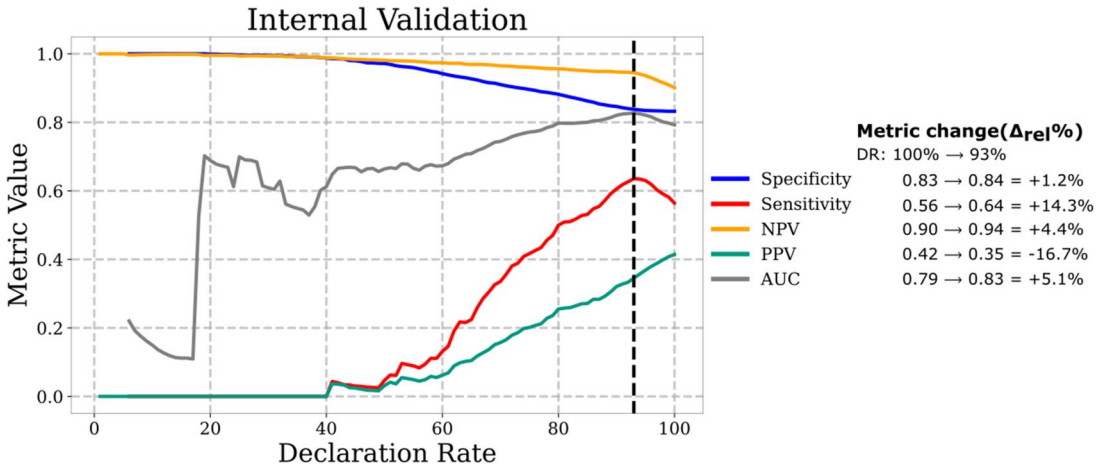

Figure 3. Internal validation of the BaseModel. MDR curves for the internal test set show that removing 7% of the least confident predictions (DR = 93%, black dotted line) leads to notable improvements in sensitivity, AUC, specificity, and NPV, while PPV decreases. Performance metrics at DR = 100% and DR = 93% are reported alongside their relative percentage change (Δ_{rel}%).

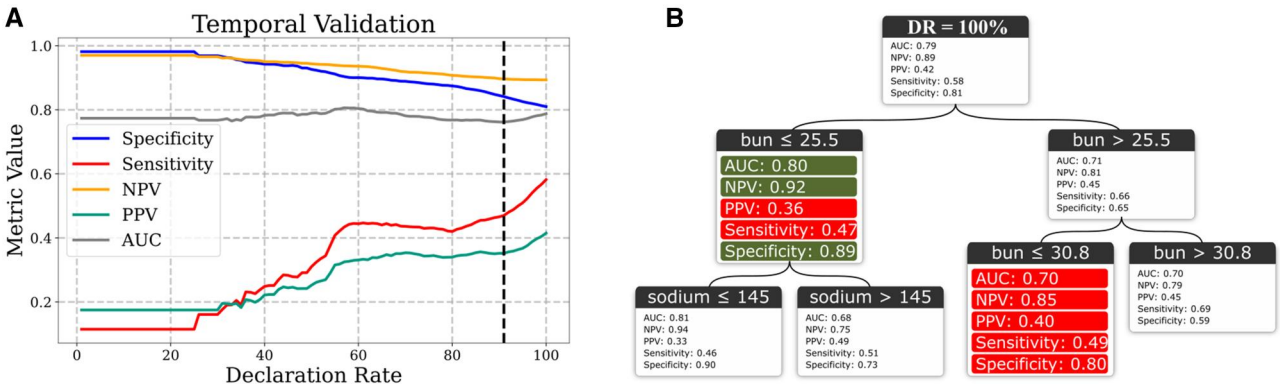

Figure 4. Temporal validation of the BaseModel. (a) MDR curves for the temporal test set show that at DR=91% (black dotted line), specificity and NPV improve while other metrics tend to decline. (b) Confidence profiles identified by the APC model at DR = 100%. Highlighted patient profiles, such as those with BUN between 25.5 and 30.8 mg/dL, exhibit lower performance across all metrics, while other profiles (e.g., BUN ≤ 25.5 mg/dL) show improved specificity and NPV but reduced sensitivity and PPV.

Even when predicted confidences do not improve most metrics, confidence profiles still provide insights into model limitations and improvement targets. In Figure 4B, patients with a BUN between 25.5 and 30.8 mg dL^{-1} showed decreased performance across all metrics. Those with BUN≤25.5 saw improved AUC, specificity, and NPV, but reduced sensitivity and PPV. This suggests the BaseModel is more confident in negative predictions for this profile, at the cost of positive prediction performance. Analyzing confidence profiles can support various goals depending on the context.

External validation

External validation simulates deploying the BaseModel in a context different from training data. Figure 5 shows MDR curves for 3 hospitals from the eICU-CRD database[49] demonstrating metric improvements. For each hospital, we proposed an optimal DR and summarized metric changes compared to DR = 100% in Table 1.

For Hospital #167, specificity slightly improved and PPV remained stable. Other metrics peaked at DR = 83%, where they plateaued. A suggested DR of 83% would increase sensitivity (Δ_{rel}%=31.7), NPV (Δ_{rel}%=10.6), and AUC (Δ_{rel}%=17.9). For Hospital #199, specificity slightly increased, while sensitivity, NPV, and AUC peaked at DR = 94%. In a real-life scenario, this DR could be proposed to maximize all 3 metrics while maintaining a good patient proportion, improving sensitivity (Δ_{rel}%=14.3), NPV (Δ_{rel}%=4.4), AUC (Δ_{rel}%=6.4), and specificity (Δ_{rel}%=1.5), though PPV would relatively decrease by 3%.

Hospital #188 showed a similar pattern where specificity slightly increased, PPV decreased, and other metrics plateaued at DR = 90%. Figure 5C highlights patient profiles at DR = 90% in Hospital 188, where patients with Glasgow Coma Scale (GCS) score under 7.5 and temperature over 38.1 °C were excluded. Given the BaseModel's unreliability for this profile, it would be advisable not to rely on its predictions for clinical decisions.

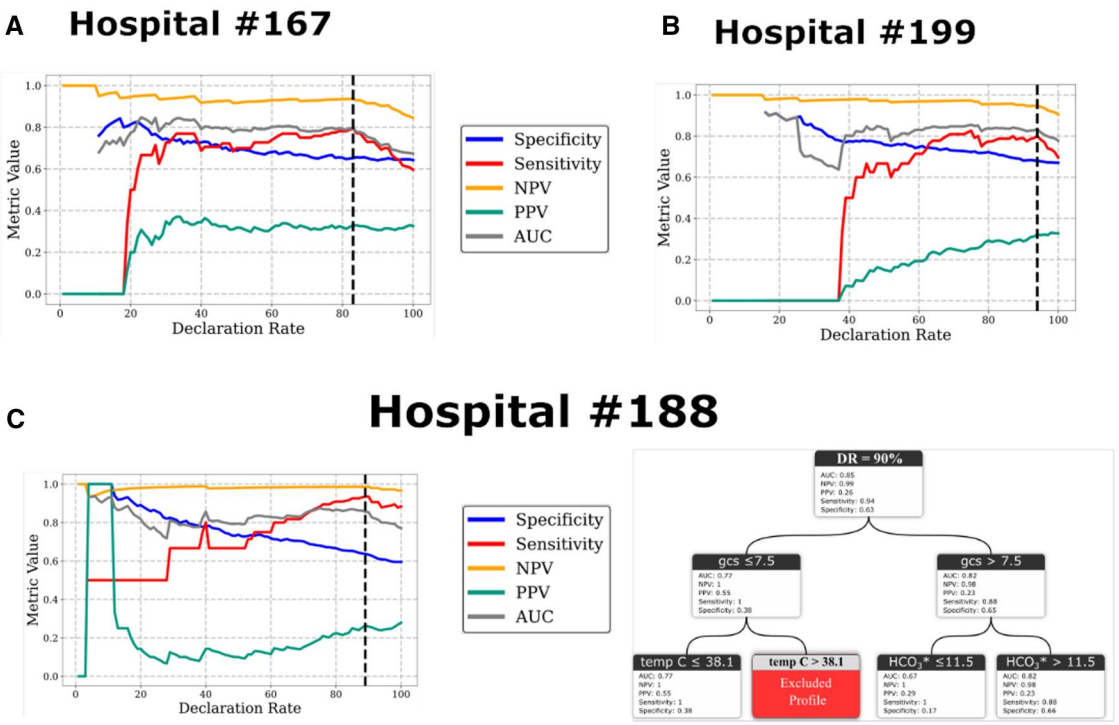

Figure 5. External validation of the BaseModel across 3 hospitals. (A-C) MDR curves for Hospitals #167, #199, and #188 show how model performance changes as the least confident predictions are excluded using MPC-based predictions. Most metrics demonstrate improvements up to a certain DR beyond which gains plateau or decline, except for PPV, which tends to decrease. Abbreviations: DR, declaration rate; MDR, metrics by declaration rate; MPC, mixed predictive confidence; PPV, positive predictive value.

Table 1. External validation results across 3 eICU-CRD hospitals.

Hospital ID	167		199		188	
DR	100	83	100	94	100	90
Specificity	0.64	0.65	0.67	0.68	0.6	0.63
Δ_{rel}%	1.6		1.5		5	
Sensitivity	0.6	0.79	0.7	0.8	0.88	0.94
Δ_{rel}%	31.7		14.3		6.8	
NPV	0.85	0.94	0.91	0.95	0.97	0.99
Δ_{rel}%	10.6		4.4		2.1	
PPV	0.33	0.33	0.33	0.32	0.28	0.26
Δ_{rel}%	0		−3		−7.1	
AUC	0.67	0.79	0.78	0.83	0.77	0.85
Δ_{rel}%	17.9		6.4		10.4	

For each hospital, we report model performance at DR = 100% (all patients included) and at the suggested DR, as well as the relative percentage change (Δ_{rel}%) in metrics by excluding least confident predictions. In all 3 hospitals, all metrics improve except for PPV, which remains stable or decreases. Abbreviations: DR, declaration rate; PPV, positive predictive value.

Clinical illustration using one-year mortality prediction

Predictive performance precision analysis in medicine supports both predeployment evaluation and postdeployment monitoring of a BaseModel. Here, we illustrate predeployment use (Figure 6) with data from Laribi et al.,[52] where the BaseModel predicts OYM upon admission. Predictive performance precision analysis in medicine helps refine decision-making by identifying areas of low predictive confidence.

Before deploying the BaseModel with MED3pa in clinical workflows, an acceptable risk threshold should be determined with regulatory authorities. This begins by analyzing overall performance and determining the optimal DR using MDR curves (Figure 6A). In this example, our goal is to improve performance without prioritizing any specific metric. Most metrics improved up to DR = 95%, after which they remained stable or declined. Based on this, we set DR at 95%.

We then examined retained patient profiles at this DR (Figure 6B). Although no profile was completely excluded, some subgroups showed significant deviations from global metrics. One such profile, highlighted in red, consisted of patients aged >59 years, admitted urgently with a cancer diagnosis. The BaseModel consistently predicted mortality (positive outcome) for this profile, despite an actual mortality rate of 50%, resulting in specificity and NPV of 0.

Once DR is set and high-risk profiles identified, the BaseModel with MED3pa could be deployed clinically. To illustrate practical use, we present a sample table of clinical data from 3 patients (Figure 6C) and show how final predictions are made (Figure 6D). For each patient, the table includes the BaseModel prediction, MED3pa's predicted confidence, and a potential real-life recommendation.

The prediction for patient 1 would be rejected due to low predicted confidence (below the 95% threshold). Patient 2 would receive a warning, as they belong to a high-risk profile (yellow profile in Figure 6B). Patient 3 would be approved as a reliable BaseModel prediction, as their predicted confidence exceeded the threshold and they were not part of a high-risk profile. This example shows how MED3pa highlights model limitations, supports nuanced decisions, and promotes reliable deployment.

Discussion

In this study, we introduced MED3pa, a confidence-based method to identify and mitigate prediction errors in clinical

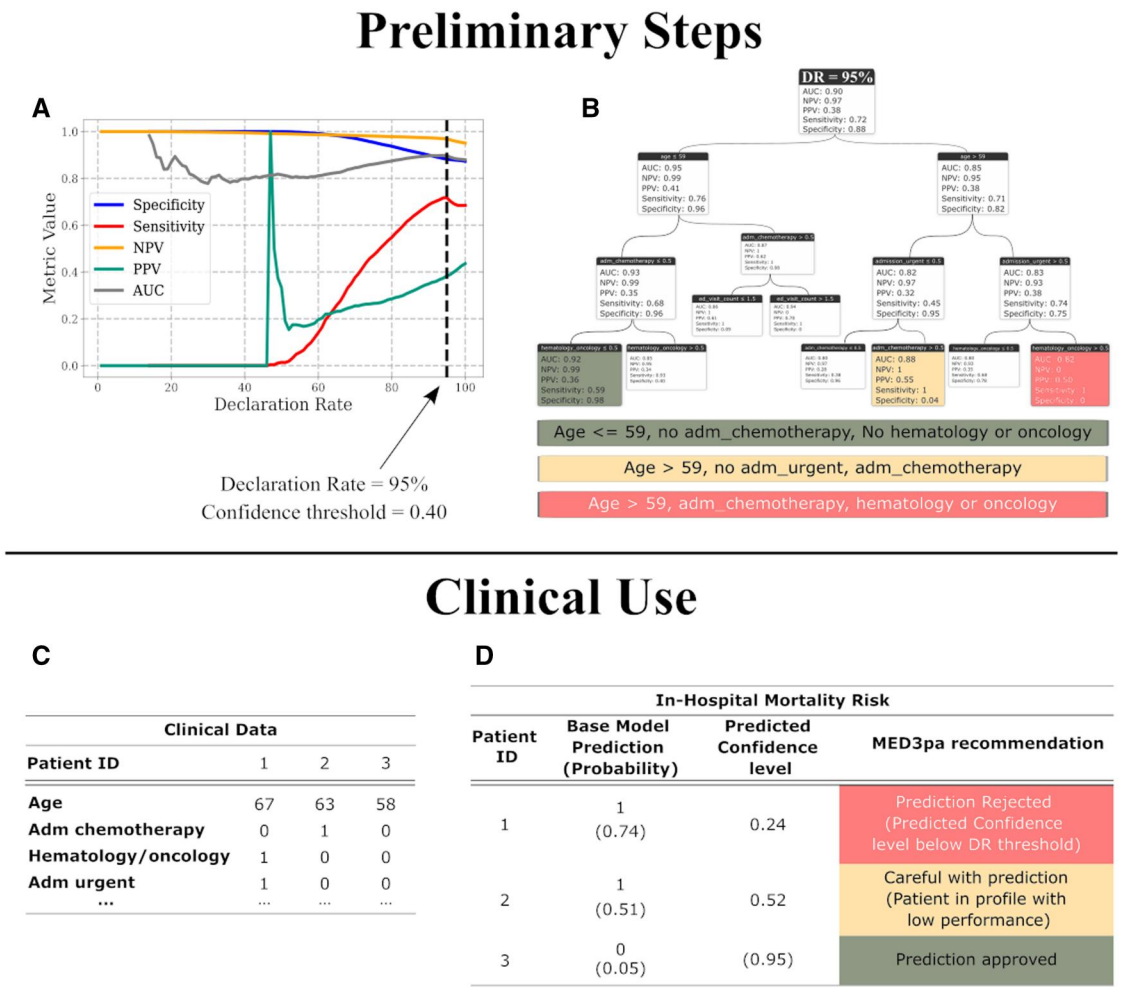

Figure 6. Clinical illustration of the MED3pa method for 1-year mortality prediction. (A and B) Preliminary steps: this step would be done by regulatory authorities before deploying into clinical workflows. The MDR curve (A) is first used to determine an appropriate DR. A suggestion here would be DR = 95%, where most metrics peak before declining, resulting in a minimum confidence threshold of 0.4. We then observe patient profiles identified by the APC model (B). The profile of patients over 59 years old, admitted urgently and with hematology/oncology diagnoses is highlighted in red due to consistently predicting mortality despite a 50% survival rate. (C and D) Example of clinical use for 3 representative patients examples: Once the BaseModel with the MED3pa method is deployed, the BaseModel's predictions are to be interpreted using MED3pa recommendations. For each new patient (C), the confidence score and APC profile membership are first determined with an appropriate recommendation (D): rejection of low-confidence predictions (patient #1), caution for patients in high-risk profiles (patient #2), and acceptance for confident predictions outside such profiles (patient #3). Abbreviations: APC, aggregated predictive confidence; DR, declaration rate; MDR, metrics by declaration rate; MED3pa, predictive performance precision analysis in medicine.

models. Using IPC for patient-specific confidence, APC for profile-level confidence, or their combination via MPC, our approach helps determine when and where a BaseModel is more likely to fail during deployment.

We applied MED3pa to 2 classification tasks: IHM using MIMIC-IV and eICU datasets, and OYM using a private dataset. For both tasks, we trained a BaseModel and evaluated its reliability. Using metrics by declaration rate MDR curves, we identified thresholds where model performance increased while excluding few patients. For example, in internal validation for IHM task, limiting predictions to the top 93% most confident cases via IPC estimates led to a 14.3% increase in sensitivity. We then used the APC model to detect patient profiles more affected by prediction errors. For instance, in OYM, patients over 59 years of age, admitted urgently with cancer (Figure 6B) had 50% actual mortality rate, yet the BaseModel predicted all as positive, resulting in poor specificity and NPV. Predictive performance precision analysis in medicine helps uncover such model limitations unnoticed in global evaluations.

In all experiments, MED3pa improved some performance metrics while excluding only a small fraction of patients from the BaseModel's applicability. Metrics by declaration rate curves showed clear tradeoffs between performance and patient coverage, and APC highlighted interpretable profiles showing BaseModel limitations. These findings emphasize MED3pa's utility in exposing limitations and improving clinical safety for reliable deployment.

Our method identifies potential predictive errors and helps determine acceptable risk tolerance. Integrating MED3pa into predictive models raises awareness of limitations and may reduce systematic errors. It also supports fairness assessments by examining how confidence varies across variables such as age, sex, or ethnicity. While fairness was not the focus here, MED3pa could help detect discrimination in BaseModel predictions. The method is not limited to ML and could

be applied to any decision-making system. It also aligns with FUTURE-AI guidelines for trustworthy, deployable AI in health care, including robustness, explainability, and fairness.[54]

The goal of this work is to introduce predictive confidence analysis for individual patients and profiles. While initial results are promising, several improvements remain necessary. First, the confidence metric could better account for context, such as class imbalance. Although we used a sigmoidal function (Material S1E) that considers the BaseModel's classification threshold, it is symmetrical. Incorporating asymmetry could better reflect class imbalance. The performance of MED3pa also depends on the IPC and APC models, so optimizing these components is important. While tree-based APC models produce interpretable patient profiles, they may not always reflect clinically meaningful subgroups. Exploring alternative clustering methods could help detect more granular, high-risk profiles. It would also be useful to adapt MED3pa to prioritize metrics such as sensitivity or precision, depending on clinical needs. As illustrated in Material S1H, MED3pa consistently improves accuracy across all scenarios explored, but future work could adapt the method to optimize other metrics.

This is the first phase of MED3pa, our long-term goal being to better integrate all patients, not exclude them from the BaseModel's applicability. Once we identify patients for whom the BaseModel underperforms, we aim to explain the performance gaps (eg, covariate shifting) and refine the BaseModel accordingly. Other alternatives include comparing confidence profiles over time, or using unlabeled data to rapidly detect changes in real-time monitoring.

Conclusion

In this study, we introduced MED3pa , a novel method that evaluates the reliability of clinical prediction models by identifying patients for whom the predictions are most reliable. Predictive performance precision analysis in medicine provides a structured approach to identify and characterize model confidence, improving model trustworthiness, and supporting reliable clinical decision-making. Further research will focus on explaining the sources of low confidence in model predictions, refining confidence estimation, and adapting the model to improve performance for patient profiles in which it currently underperforms.

Acknowledgments

The authors thank Ryeyan Taseen, MD, MSc, for data collection. The authors also thank Hakima Laribi, PhD student at Université de Sherbrooke, for data processing and for her helpful comments and suggestions throughout the project.

Author contributions

Olivier Lefebvre (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing—original draft, Writing—review & editing), Félix Camirand Lemyre (Conceptualization, Methodology, Project administration, Supervision, Writing—review & editing), Jean-François Ethier (Conceptualization, Methodology, Project administration, Supervision, Writing—review & editing), Lyna Hiba Chikouche (Investigation, Software, Visualization, Writing—review & editing), Ludmila Amriou (Investigation, Software, Visualization, Writing—review & editing), Dan Poenaru (Methodology, Project administration, Supervision, Validation, Writing—review & editing), and Martin Vallières (Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing—review & editing)

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This study was supported by the Canada CIFAR AI Chair, Mila; the Natural Sciences and Engineering Research Council of Canada (NSERC), discovery grants program (RGPIN-2021-03996); and the Fonds de recherche du Québec—Nature et technologies, programme relève professorale (312290). The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Conflicts of interest

None declared.

Data availability

Software code allowing to run the experiments used to produce the results presented in this work is freely shared under the GNU General Public License v3.0 on the GitHub website at: https://github.com/MEDomics-UdeS/study_3pa. The MED3pa package developed in this study is freely shared under the GNU General Public License v3.0 at: https://pypi.org/project/MED3pa/. The data underlying this article come from both public and private sources. Publicly available data were obtained from the MIMIC-IV and eICU Collaborative Research Databases, which can be accessed upon completion of the appropriate data use agreements via PhysioNet (https://physionet.org/). The hospitalization data used for the clinical illustration via 1 year mortality prediction are not publicly available for confidentiality purposes overseen by the Institutional Review Board of the CIUSSS de l'Estrie—CHUS Nagano #2022-4409. However, a synthetic dataset[53] generated using the AVATAR method[55] in partnership with Octopize (https://www.octopize.io) is publicly shared on the Zenodo website at: https://zenodo.org/records/12954673. The publication of this synthetic dataset has been approved under project #2022-4409 with form #F2H-60510.

References

1. Bhinder B, Gilvary C, Madhukar NS, et al. Artificial intelligence in cancer research and precision medicine. *Cancer Discov*. 2021;11:900-915.
2. Chen ZH, Lin L, Wu CF, et al. Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine. *Cancer Commun (Lond)*. 2021;41:1100-1115.
3. Pinton P. Impact of artificial intelligence on prognosis, shared decision-making, and precision medicine for patients with

inflammatory bowel disease: a perspective and expert opinion. *Ann Med*. 2023;55:2300670.
4. Konikoff T, Loebl N, Yanai H, et al. Precision medicine: externally validated explainable AI support tool for predicting sustainability of infliximab and vedolizumab in ulcerative colitis. *Dig Liver Dis*. 2024;56:2069-2076.
5. Pirracchio R, Petersen ML, Carone M, et al. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. *Lancet Respir Med*. 2015;3:42-52.
6. Hu Q, Rizvi AA, Schau G, et al. Development and validation of a deep learning-based microsatellite instability predictor from prostate cancer whole-slide images. *NPJ Precis Oncol*. 2024;8:88.
7. Youssef A, Pencina M, Thakur A, et al. External validation of AI models in health should be replaced with recurring local validation. *Nat Med*. 2023;29:2686-2687.
8. Van Calster B, Steyerberg EW, Wynants L, et al. There is no such thing as a validated prediction model. *BMC Med*. 2023;21:70.
9. Cabitza F, Campagner A, Soares F, et al. The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Comput Methods Programs Biomed*. 2021;208:106288.
10. Sperrin M, Riley RD, Collins GS, et al. Targeted validation: validating clinical prediction models in their intended population and setting. *Diagn Progn Res*. 2022;6:24.
11. Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ*. 2024;384:e074820.
12. Ginsberg T, Liang Z, Krishnan RG. A learning based hypothesis test for harmful covariate shift. In: *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*. OpenReview. 2023. Kigali, Rwanda.
13. Banerji CRS, Chakraborti T, Harbron C, et al. Clinical AI tools must convey predictive uncertainty for each individual patient. *Nat Med*. 2023;29:2996-2998.
14. Vela D, Sharp A, Zhang R, et al. Temporal quality degradation in AI models. *Sci Rep*. 2022;12:11654.
15. Shashikumar SP, Wardi G, Malhotra A, et al. Artificial intelligence sepsis prediction algorithm learns to say "I don't know". *NPJ Digit Med*. 2021;4:1-9.
16. Shafer G, Vovk V. A tutorial on conformal prediction. *J Mach Learn Res*. 2008;9:371-421.
17. Laghuvarapu S, Lin Z, Sun J. CoDrug: conformal drug property prediction with density estimation under covariate shift. In: Oh A, Naumann T, Globerson A, et al., eds. *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc.; 2023:37728-37747.
18. Fannjiang C, Bates S, Angelopoulos AN, et al. Conformal prediction under feedback covariate shift for biomolecular design. *Proc Natl Acad Sci*. 2022;119:e2204569119.
19. Tibshirani RJ, Foygel Barber R, Candes E, et al. Conformal prediction under covariate shift. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.; 2019: 2530–2540.
20. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning—Volume 48ICML'16*. 2016:1050-1059.
21. Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? In: *Proceedings of the 31st International Conference on Neural Information Processing SystemsNIPS'17*. Curran Associates Inc.; 2017:5580-5590. Long Beach, CA, USA.
22. Malinin A, Prokhorenkova L, Ustimenko A. Uncertainty in gradient boosting via ensembles. In: *International Conference on Learning Representations*. 2021. Virtual conference.
23. Blundell C, Cornebise J, Kavukcuoglu K, et al. Weight uncertainty in neural network. In: Bach F, Blei D., eds. *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. PMLR; 2015:1613-1622.
24. Depeweg S, Hernandez-Lobato JM, Doshi-Velez F, et al. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In: Dy J, Krause A., eds. *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80 PMLR; 2018:1184-1193.
25. Lahoti P, Gummadi K, Weikum G. Responsible model deployment via model-agnostic uncertainty learning. *Mach Learn*. 2023;112: 939-970.
26. Rasmussen MH, Duan C, Kulik HJ, et al. Uncertain of uncertainties? A comparison of uncertainty quantification metrics for chemical data sets. *J Cheminform*. 2023;15:121.
27. Lind SK, Xiong Z, Forssén PE, et al. Uncertainty quantification metrics for deep regression. *Pattern Recognit Lett*. 2024;186:91-97.
28. Valdenegro-Toro M, Mori DS. A deeper look into aleatoric and epistemic uncertainty disentanglement. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2022:1508-1516. New Orleans, LA, USA.
29. Ovadia Y, Fertig E, Ren J, et al. Can you trust your model' s uncertainty? Evaluating predictive uncertainty under dataset shift. In: *Advances in Neural Information Processing Systems*. Wallach H, Larochelle H, Beygelzimer A, et al., eds. Vol. 32. Curran Associates, Inc.; 2019.
30. Zhang P, Wang J, Farhadi A, et al. Predicting failures of vision systems. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014:3566-3573.
31. Chen T, Navratil J, Iyengar V, et al. Confidence scoring using whitebox meta-models with linear classifier probes. In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. PMLR; 2019:1467-1475.
32. Zhou S, Blanchart P, Crucianu M, et al. Why is the prediction wrong? Towards underfitting case explanation via meta-classification. In: *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*. 2022:1-9. Shenzhen, China.
33. Seedat N, Crabbé J, Bica I, et al. Data-IQ: characterizing subgroups with heterogeneous outcomes in tabular data. In: Koyejo S, Mohamed S, Agarwal A, et al., eds. *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc.; 2022: 23660-23674.
34. König M, Hoos HH, Rijn JN. Towards algorithm-agnostic uncertainty estimation: predicting classification error in an automated machine learning setting. In: *Proceedings of the 7th International Conference on Machine Learning Workshop on Automated Machine Learning (AutoML)*. July 13-18, 2020.
35. Nadeem MSA, Zucker JD, Hanczar B. Accuracy-rejection curves (ARCs) for comparing classification methods with a reject option. In: Džeroski S, Guerts P, Rousu J, eds. *Proceedings of the Third International Workshop on Machine Learning in Systems Biology*. Vol. 8. PMLR; 2009:65-81.
36. Brown KE, Talbert S, Talbert DA. Derivation and experimental performance of standard and novel uncertainty calibration techniques. *AMIA Annu Symp Proc*. 2024;2024:212-221.
37. Subbaswamy A, Sahiner B, Petrick N, et al. A data-driven framework for identifying patient subgroups on which an AI/machine learning model may underperform. *NPJ Digit Med*. 2024;7:334.
38. Leibig C, Allken V, Ayhan MS, et al. Leveraging uncertainty information from deep neural networks for disease detection. *Sci Rep*. 2017;7:17816.
39. Tosh CJ, Hsu D. Simple and near-optimal algorithms for hidden stratification and multi-group learning. In: Chaudhuri K, Jegelka S, Song L, et al., eds. *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. PMLR; 2022:21633-21657.
40. Seah J, Tang C, Buchlak QD, et al. Do comprehensive deep learning algorithms suffer from hidden stratification? A retrospective study on pneumothorax detection in chest radiography. *BMJ Open*. 2021;11:e053024.
41. Sohoni N, Dunnmon J, Angus G, et al. Subclass left behind: fine-grained robustness in coarse-grained classification problems. In:

Larochelle H, Ranzato M, Hadsell R, et al., eds. *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc.; 2020:19339-19352.
42. Oakden-Rayner L, Dunnmon J, Carneiro G, et al. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: *Proceedings of the ACM Conference on Health, Inference, and LearningCHIL '20*. Association for Computing Machinery; 2020:151-159.
43. d'Eon G, d'Eon J, Wright JR, et al. The spotlight: a general method for discovering systematic errors in deep learning models. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and TransparencyFAccT '22*. Association for Computing Machinery; 2022:1962-1981.
44. Bansal A, Farhadi A, Parikh D. Towards transparent systems: semantic characterization of failure modes. In: Fleet D, Pajdla T, Schiele B, et al., eds. *Computer Vision—ECCV 2014*. Springer International Publishing; 2014:366-381.
45. Bissoto A, Hoang TD, Flühmann T, et al. Subgroup performance analysis in hidden stratifications. In: Gee JC, Alexander DC, Hong J, et al., eds. *Medical Image Computing and Computer Assisted Intervention—MICCAI 2025*. Springer Nature Switzerland; 2025:594-603.
46. Brown KE, Talbert S, Talbert DA. A QUEST for model assessment: identifying difficult subgroups via epistemic uncertainty quantification. *AMIA Ann Sympos Proc*. 2023;2023:854-863.
47. Le Gall JR, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *JAMA*. 1993;270:2957-2963.
48. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10:219.
49. Pollard TJ, Johnson AEW, Raffa JD, et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data*. 2018;5:180178.
50. Zeng Z, Yao S, Zheng J, et al. Development and validation of a novel blending machine learning model for hospital mortality prediction in ICU patients with Sepsis. *BioData Min*. 2021;14:40.
51. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3:32-35.
52. Laribi H, Raymond N, Taseen R, et al. Leveraging patients' longitudinal data to improve the hospital one-year mortality risk. *Health Inf Sci Syst*. 2025;13:23.
53. Laribi H, Raymond N, Taseen R, et al. *Synthetic Data for Accessible Learning in Healthcare: Improving Mortality Prediction with Longitudinal Data*. Research Square; 2024.
54. Lekadir K, Frangi AF, Porras AR, et al.; FUTURE-AI Consortium. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*.2025;388:e081554.
55. Guillaudeux M, Rousseau O, Petot J, et al. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *NPJ Digit Med*. 2023;6:37-10.

Journal of the American Medical Informatics Association, 2026, 33, 1100–1111
https://doi.org/10.1093/jamia/ocag034
Research and Applications