A Bayesian approach to predictive uncertainty in chemotherapy patients at risk of acute care utilization

Claudio Fanconi,[a,b] Anne de Hond,[b,c] Dylan Peterson,[b] Angelo Capodici,[b,d] and Tina Hernandez-Boussard[b,*]

[a]Department of Information Technology and Electrical Engineering, ETH Zürich, Zürich, Switzerland
[b]Department of Medicine (Biomedical Informatics), Stanford University, Stanford, USA
[c]Clinical AI Implementation and Research Lab, Leiden University Medical Centre, Leiden, the Netherlands
[d]Department of Biomedical and Neuromotor Science, University of Bologna, Bologna, Italy

eBioMedicine
2023;92: 104632
Published Online 1 June 2023
https://doi.org/10.1016/j.ebiom.2023.104632

Summary

Background Machine learning (ML) predictions are becoming increasingly integrated into medical practice. One commonly used method, ℓ_1-penalised logistic regression (LASSO), can estimate patient risk for disease outcomes but is limited by only providing point estimates. Instead, Bayesian logistic LASSO regression (BLLR) models provide distributions for risk predictions, giving clinicians a better understanding of predictive uncertainty, but they are not commonly implemented.

Methods This study evaluates the predictive performance of different BLLRs compared to standard logistic LASSO regression, using real-world, high-dimensional, structured electronic health record (EHR) data from cancer patients initiating chemotherapy at a comprehensive cancer centre. Multiple BLLR models were compared against a LASSO model using an 80–20 random split using 10-fold cross-validation to predict the risk of acute care utilization (ACU) after starting chemotherapy.

Findings This study included 8439 patients. The LASSO model predicted ACU with an area under the receiver operating characteristic curve (AUROC) of 0.806 (95% CI: 0.775–0.834). BLLR with a Horseshoe+ prior and a posterior approximated by Metropolis–Hastings sampling showed similar performance: 0.807 (95% CI: 0.780–0.834) and offers the advantage of uncertainty estimation for each prediction. In addition, BLLR could identify predictions too uncertain to be automatically classified. BLLR uncertainties were stratified by different patient subgroups, demonstrating that predictive uncertainties significantly differ across race, cancer type, and stage.

Interpretation BLLRs are a promising yet underutilised tool that increases explainability by providing risk estimates while offering a similar level of performance to standard LASSO-based models. Additionally, these models can identify patient subgroups with higher uncertainty, which can augment clinical decision-making.

Funding This work was supported in part by the National Library Of Medicine of the National Institutes of Health under Award Number R01LM013362. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Keywords: Bayesian logistic LASSO regression; Predictive uncertainty; Acute care utilization; Chemotherapy

Introduction

Machine learning (ML) is becoming increasingly common in healthcare. It can perform various tasks, from classifying skin lesions to identifying breast cancer in mammograms and providing risk estimates for hospital readmission and mortality.[1–4] In general, the majority of models that have been brought into clinical practice provide fixed point estimates (e.g., 68% chance of readmission). However, these predictions rarely quantify their uncertainty, which could alter clinical decisions made using the prediction. For instance, low uncertainty estimates could have automated interventions, while high uncertainty estimates could be triaged to a provider for a more thorough review.[5]

There are several approaches to quantifying uncertainty in ML models, including the Bayesian framework

*Corresponding author. Stanford University, 453 Quarry Road, Palo Alto, CA, 94304, USA.
E-mail addresses: boussard@stanford.edu (T. Hernandez-Boussard), fanconic@ethz.ch (C. Fanconi), a.a.h.de_hond@lumc.nl (A. de Hond), dpeters9@stanford.edu (D. Peterson), angelo.capodici@studio.unibo.it (A. Capodici).
@Boussard (T. Hernandez-Boussard).

Research in context

Evidence before this study

There are a growing number of studies that use machine learning algorithms to predict the risk of various medical problems. However, most of these machine learning models provide point estimates of risk and do not incorporate uncertainty. To evaluate the evidence before this study, we searched PubMed for articles from the start of the database until 30 June 2022 using the search terms ("bayesian machine learning" OR "uncertainty quantification" OR "uncertainty estimation") AND ("medicine" OR "oncology" OR "acute care utilization"), with no date or language restrictions. Among the studies found that do use uncertainty, they generally apply it to improve classification accuracy, incorporate prior information, or detect samples outside the distribution. A single study was found that quantifies the uncertainty of its model's predictions, but did not compare different Bayesian models or determine the best model for an entire dataset. The study also does not highlight other advantages of using a Bayesian model or compare predictive uncertainty based on patient groups.

Added value of this study

This study compares methods of quantifying uncertainty for risk estimation using Bayesian logistic LASSO regression in health informatics. The results show that the Bayesian logistic LASSO regression (BLLR) models perform well compared to frequentist models, even with high-dimensional data. The study also visualized features that are 95% credibly correlated with the outcome and extended the examination of algorithmic bias to uncertainty estimation, identifying subgroups where bias is prevalent.

Implications of all the available evidence

The developed method for estimating uncertainty provides information about the certainty of the ACU risk score and model weights. It allows for Data scientists and clinicians to consider not only about the risk probability of an event, but also about the acceptable uncertainty of predictions. The range of uncertainty may vary depending on the prediction problem and resources available to medical institutions. For example, if an ML model is developed for patient triage, predictions with an uncertainty range that exceeds the decision threshold are likely to be misclassified and thus uncertain. We argue that clinicians should be aware of these cases rather than relying solely on point estimates of probabilities. Furthermore, Users should be aware of potential biases in predicted uncertainty and their implications.

(Supplementary materials D).[5] Compared to the frequentist framework (e.g., LASSO models), the Bayesian approach aims to estimate the full posterior distribution of the model parameters as opposed to point estimates of parameters that minimise the prediction error or maximise a penalised likelihood. This means that the parameters are not just single-point estimates, but are sampled from a distribution. Consequently, the risk predictions produced from Bayesian models are distributions. From these, one can quantify uncertainty by looking at the dispersion of the prediction distribution: high dispersion indicates high uncertainty, while low dispersion conversely indicates low uncertainty in the prediction.

Although research on uncertainty estimation in medical informatics is still in its infancy, there is progress using Bayesian models, including models to forecast metabolic control in type II diabetes, estimate drug efficacy, image classification for diagnosis, and disease detection based on structured health data.[6–9] A study evaluated Bayesian and frequentist statistical methods for comparing multiple medical treatments and found that while Bayesian methods are more flexible and more clinically interpretable, they are also more challenging to develop.[10] As uncertainty estimations through Bayesian modelling can improve interpretability and thereby increase provider confidence in risk predictions, this study sought to test the performance of Bayesian logistic LASSO regression (BLLR) model against a previously developed LASSO model that leveraged dense, electronic health record (EHR) data.[11] Specifically, we investigated performance in predicting the risk of emergency department visits or hospital admissions after starting chemotherapy, as defined by the Centre for Medicare and Medicaid Services (CMS) OP-35 quality measure.[12] This measure is ideal for ML-based risk predictions, as early outpatient interventions can prevent up to 20–50% of all acute care utilization (ACU), which accounts for nearly half of the costs associated with oncology care in the United States.[13–16] In the present study, after comparing the predictive performance of BLLR models, we sought to demonstrate the value that uncertainty estimates can provide for clinicians deciding how to interpret and intervene upon these predictions.

Methods

Setting and ethical approval

This retrospective cohort study was performed at a Comprehensive Cancer Centre comprised of an academic medical centre and affiliated community practices. This study was approved by the university institutional review board at Stanford University with a waiver of informed consent. The study was approved by the Stanford University ethics committee (protocol #47644). It followed the Minimum Information for Medical AI Reporting (MINIMAR, Supplementary Table S1) and Transparent Reporting of a Multivariable

Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guidelines.[17,18]

Data set

To compare the previously published LASSO model by Peterson et al. against Bayesian approaches, we leveraged the same data set.[11] In brief, this data set contains all adult cancer patients treated at the Comprehensive Cancer Centre on January 1, 2013, and July 10, 2019. All patients met inclusion criteria for CMS' OP-35 measure based upon the 2019 Chemotherapy Measure Updates and Specifications Report who had adequate follow-up and EHR data to make predictions.[12]

The same 760 EHR-derived variables were used for model training, which included social and demographic variables, procedures, diagnoses, medications, laboratory values, vital signs, cancer-specific data and healthcare utilization. Imputation of missing data followed the same procedures reported by Peterson et al. As with the prior study, all data were restricted to EHR data generated prior to the initiation of chemotherapy to limit data leakage. The same random split of the cohort into an 80% training cohort and 20% testing cohort from the original paper was maintained to enable comparison across the models. Missing vitals values were mean imputed, while missing laboratory values were mean imputed on the ten nearest neighbours (KNN algorithm) in the training dataset. Outcome labels (ACU within 30 days of chemotherapy initiation) were defined using the aforementioned CMS definition. A detailed description of how the patient cohort was extracted, the inclusion and exclusion criteria for the OP-35 metric, and a complete list of features can be found in the original paper.[11] The EHR dataset generated during or analysed in this study is not publicly available due to restrictions by privacy laws.

Model development

We compared the previously published *Frequentist LASSO* model against three additional Bayesian models: *Laplace-VI*, *Laplace-MH*, and *Horseshoe-MH*. These models leveraged either Laplacian or Horseshoe priors with posteriors estimated by meanfield variational inference or Metropolis–Hastings sampling. All were modelled with the Bernoulli likelihood probability distribution. Model training specifics, including the reasoning for the choices of prior and posterior probability distributions, can be found in the supplemental methods (Supplementary materials A).

Model evaluation

Discrimination

The models were compared using the Area Under the Receiver Operator Characteristic Curve (AUROC, or C-statistic), the Area Under the Precision-Recall Curve (AUPRC), and log-loss (cross-entropy loss). As Bayesian models provided a predictive distribution, the mean of the distribution was used as the risk prediction for performance evaluation. Confidence intervals were calculated using 1000-fold bootstrapping on the test dataset.

Calibration and bias assessments

Expected calibration error (ECE) and flexible calibration curves, fit with flexible, non-linear calibration curves with locally weighted running line smoothers (LOESS) were used to compare model calibration.[19–22] To investigate algorithmic bias, we analysed the differences in prediction uncertainty for demographic and clinical subgroups using box and whisker plots.[23]

Uncertainty evaluation

To better characterise uncertainty (i.e., the standard deviation of the distribution of the prediction, σ, unless otherwise specified) across the different Bayesian models at a cohort-wide level, risk predictions were plotted against their corresponding uncertainties. To demonstrate uncertainty an individual level, the probability density for models' predictions on three patients (low risk, medium risk, high risk) were produced with KDE plots. Additionally, to investigate the uncertainty of different parameters, we examined the posterior distributions of the Horseshoe-MH for parameters with 95%-credible intervals that did not cross zero (e.g., those with a 95% probability of having a credible effect).

Clinical utility

Finally, we assessed the initial clinical utility of these four models through a Decision Curve Analysis (DCA) by plotting the Net Benefit across a range of decision thresholds and quantifying the number of true positives penalised by false positives.[24] To illustrate how these models could be deployed for automatic treatment classification, we also performed a series of experiments in which we iterated over various decision thresholds for treatment (predicted risk greater than 10, 16, 30, and 50%). We then compared the coverage (the proportion of certain classifications) against the classification performance metrics (F1 score, recall, precision) to determine which model could automatically classify patient labels while maintaining good classification performance. Additional details on these experiments can be found in the supplemental methods (Supplementary materials A.3.3).

Statistics

All analyses were implemented in Python, using the scikit-learn library for the metrics and classical LASSO model, and PyMC3 for the Bayesian models.[25,26] The code for the Bayesian logistic LASSOs and our analysis is publicly available on GitHub (https://github.com/suboussard-lab/acu-uncertainty-estimation).

We compared the bias assessment with the Kruskal–Wallis test to examine if the medians of the group distributions are significantly different from each other.[27]

Sample size determination, randomisation, blinding, inclusion and exclusion criteria can be found in the original paper.[11]

For further learning on this topic, we provide a mathematical overview of Bayesian machine learning, details on the choice of prior and likelihood, and methods for approximating posterior distributions, in the Supplementary materials D.

Role of funders

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.

Results

The study cohort included 8439 patients, with a mean age at the start of chemotherapy of 60.4 (±14.5). 50.4% were female. 1306 patients (15.5%) met the primary criteria of having at least one OP-35 event within the first 30 days of starting chemotherapy. The majority of patients in the cohort were White (n = 4630; 54.9%), followed by Asian patients (n = 1897; 22.5%), and the least represented were Black patients (n = 233; 2.8%). The most common cancer types were breast (n = 1383; 16.4%), lymphoma (n = 1175; 13.9%), and pancreas (n = 980; 11.6%). ACU events were most common for lymphoma (n = 364; 26.5%) and least common for prostate cancer (n = 12; 0.9%). Most chemotherapy patients had a stage IV tumour (n = 2318; 27.5%) and were insured by Medicare (n = 3236; 38.3%) or private health insurance (n = 3049; 36.1%). Cohort characteristics are summarised in Table 1.

Discriminative performance and calibration

Table 2 details performance characteristics of the three Bayesian models and the frequentist LASSO. While the model performance was generally similar, the Horseshoe-MH model was the best performing overall. It had the highest AUROC (0.807, 95% CI: 0.780–0.834), lowest negative log-likelihood (0.355, 95% CI: 0.328–0.384), and lowest expected calibration error (0.006, 95% CI: <0.001–0.055). This model was only outperformed by the frequentist LASSO in terms of AUPRC (0.511, 0.95% CI: 0.447–0.579), but the confidence intervals overlapped for this metric.

Calibration curves demonstrated that the Laplace-VI model often overestimates the ACU risk, while the other models had much better and comparable calibration (Supplementary Fig. S3). In addition, the decision curve analysis showed that the net benefit of the Laplace-VI model was consistently lower than the other models, which had a similar performance and higher net benefit scores across a broader range of decision thresholds (Supplementary Fig. S4).

Uncertainty predictions across the cohort

Investigating predictive uncertainty at a cohort-wide level, the uncertainty of all Bayesian models was most prominent when the predicted probability of ACU was near 50% (Supplementary Fig. S5). For a given risk probability, the uncertainty of the Laplace-VI predictions was, in almost all cases, higher than the uncertainties of the MH samples.

To examine the ability of these models to automatically triage downstream interventions, the prediction distributions were compared against a decision threshold set at the ACU event rate of 16%. If the mean ± standard deviation of a prediction overlapped with this threshold, then a prediction was determined to be too uncertain to classify automatically. Fig. 1 shows the sorted risk predictions of the Horseshoe-MH model with their predictive uncertainty. The coverage, or proportion of certain classifications, for this particular use case was 0.72, which means that 72% of the patients could be automatically classified with low uncertainty. When repeating this at a more stringent threshold requiring the 95%-credible interval not to overlap the decision threshold, the coverage was 0.54 (i.e., 46% of patients were too uncertain to be classified; Supplementary Fig. S6).

In the Supplementary materials, further examination of coverage against performance, shows the coverage score of the four models at different decision thresholds and uncertainty against the F1 score, the sensitivity score (recall) and the positive predictive value score (PPV/precision; Supplementary Fig. S7). The Laplace-VI model had the highest F1 and recall scores over the thresholds but had significantly lower coverage than the other models. The Horseshoe-MH had a higher F1 score and recall than the Laplace-MH and frequentist LASSO models at $t \in \{0.1, 0.16\}$, higher precision for $t \in \{0.16, 0.3, 0.5\}$, and the highest coverage for the BLLR models across the thresholds. Frequentist LASSO always had coverage of 1.0, as each of its predictions is a point estimate that cannot cross the decision threshold. These results are also presented in Supplementary Fig. S8.

When analysing the different quantified uncertainties of the Horseshoe-MH predictions, as expected, σ-uncertainty had the highest coverage, followed by a 95%-credible interval, then 2σ and finally 99%-credible. The coverage ranged from 0.28 ($t = 0.1$, 99%-credible interval) to 0.93 ($t = 0.5$, σ). For F1 score and sensitivity, the σ-quantified uncertainty had combined values across all thresholds closest to the optimum in the upper right-hand corner (Supplementary Fig. S9).

Uncertainty prediction for individual patients

To examine risk prediction performance for individual patients, Fig. 2 shows the distributions and the expected values of the Bayesian prediction models for ACU for three patients predicted to be at high, intermediate, and

low risk for ACU by the frequentist LASSO model. While the predictive distribution of the Laplace-VI model was mainly weighted near 0.0 or 1.0, the predictive distributions of the models sampled with Metropolis–Hastings only spanned over a limited range of probabilities that more closely approximated their mean predicted values and the predictions of the frequentist LASSO model. In all three cases, the predictive distribution of the Laplace-MH model had a more extensive spread than the Horseshoe-MH model, indicating increased uncertainty.

Algorithmic bias in uncertainty estimates

Investigating bias in terms of differential uncertainty for patient subgroups, the Horseshoe-MH model had significantly greater uncertainty, and therefore higher expected error, in predictions for Black patients (Median Error: ±6.5%) than for White patients (Median Error: ±3.9%), Asian patients (Median Error: ±4.6%), and other races (Median Error: ±0.5%) (Kruskal–Wallis test: $p < 0.001$) (Fig. 3a). Stage IV patients had significantly higher predictive uncertainty (median = 0.057) compared to lower staged patients (medians < 0.05) (Kruskal–Wallis test: $p < 0.001$) Fig. 3b. By cancer type, prostate cancer had the lowest uncertainty (median = 0.011), while sarcomas had the highest (median = 0.083) (Fig. 3c). Further stratification by patient sex (Kruskal–Wallis: $p = 0.05$), ethnicity (Kruskal–Wallis test: $p < 0.001$), and insurance status (Kruskal–Wallis test: $p < 0.001$) are provided in Supplementary Fig. S11.

Posterior distribution of variables

There were 27 of 760 variables in the Horseshoe-MH model that were credibly correlated with the outcomes, with their 95%-credible intervals not overlapping 0 (example features and distributions are displayed in Supplementary Fig. S10). We see that e.g., Sarcoma cancers (median = 0.086), Non-palliative patients (median = 0.14), and the number of days a patient was previously hospitalised (median = 0.25) were credibly positively correlated with the outcome, while albumin

Patient characteristic	Total cohort (n = 8439)	Patients with OP-35 events (n = 1306, 15.5%)	Patients without OP-35 events (n = 7133, 84.5%)
Age, mean ± sd			
At diagnosis	58.7 ± 14.4	56.23 ± 15.8	59.1 ± 14.1
At first chemotherapy	60.4 ± 14.5	57.9 ± 15.8	60.8 ± 14.2
Sex, No. (%)			
Female	4250 (50.4)	619 (47.4)	3631 (50.9)
Race, No. (%)			
White	4630 (54.9)	653 (50.0)	3977 (55.8)
Asian	1897 (22.5)	299 (22.9)	1598 (22.4)
Black	233 (2.8)	51 (3.9)	182 (2.6)
Other or unknown	1679 (19.9)	303 (23.2)	1376 (19.3)
Ethnicity, No. (%)			
Non-Hispanic or non-Latino	7231 (85.7)	1091 (83.5)	6140 (86.1)
Hispanic or Latino	1094 (13.0)	208 (15.9)	886 (12.4)
Cancer type, No. (%)			
Breast	1383 (16.4)	125 (9.6)	1258 (17.6)
Lymphoma	1175 (13.9)	346 (26.5)	829 (11.6)
Pancreas	980 (11.6)	141 (10.8)	839 (11.8)
Gastrointestinal	949 (11.2)	121 (9.3)	828 (11.6)
Thoracic	825 (9.8)	127 (9.7)	698 (9.8)
Genitourinary	596 (7.1)	99 (7.6)	497 (7.0)
Head and neck	697 (8.3)	100 (7.7)	597 (8.4)
Prostate	569 (6.7)	12 (0.9)	557 (7.8)
Gynecologic	562 (6.7)	80 (6.1)	482 (6.8)
Other	703 (8.3)	155 (11.9)	548 (7.6)
Cancer stage, No. (%)			
Stage I	1432 (17.0)	177 (13.6)	1255 (17.6)
Stage II	1679 (19.9)	175 (13.4)	1504 (21.1)
Stage III	1168 (13.8)	192 (14.7)	976 (13.7)
Stage IV	2318 (27.5)	486 (37.2)	1832 (25.7)
Unknown	1842 (21.8)	276 (21.1)	1566 (22.0)
Insurance, No. (%)			
Medicare	3236 (38.3)	429 (32.8)	2807 (39.4)
Private	3049 (36.1)	512 (39.2)	2537 (35.6)
Medicaid	719 (8.5)	170 (13.0)	549 (7.7)
Other or unknown	1435 (17.0)	195 (14.9)	1240 (17.4)

Information about the complete patient cohort eligible for the OP-35 metric for 30-day prediction. Results reported as "mean ± standard deviation" or as "number (% of group)".

Table 1: **Patient cohort.**

Model	AUROC	AUPRC	Log-loss	ECE
Frequentist LASSO $c = 0.02$	0.806 (0.775, 0.834)	**0.511** (0.447, 0.579)	0.357 (0.332, 0.385)	0.045 (0.030, 0.075)
Laplace-VI $b = 1/\sqrt{2}$	0.776 (0.745, 0.807)	0.437 (0.377, 0.504)	0.539 (0.517, 0.567)	0.242 (0.227, 0.267)
Laplace-MH $b = 1/\sqrt{2}$	0.769 (0.738, 0.800)	0.452 (0.397, 0.517)	0.38 (0.348, 0.417)	0.032 (0.019, 0.064)
Horseshoe-MH $\sigma = 1$	**0.807** (0.780, 0.834)	0.498 (0.443, 0.559)	**0.355** (0.328, 0.384)	**0.006** (<0.001, 0.055)

Resulting metrics on the test set of the frequentist LASSO and the BLLRs. We report the 95%-confidence intervals of the metric estimates that have been calculated with 1000-fold bootstrap in the brackets: (2.5% CI, 97.5% CI). The best-performing metrics for every label type per metric are marked in bold. The inverse regularisation parameter for the LASSO is denoted as c, while the scale parameter of the Laplace prior is denoted as b. For the Horseshoe-MH model we set the variance $\sigma = 1$ for all the Gaussian hyper-prior initialisations. We provide an overview of the bootstrapped distribution as violin plots in Supplementary Fig. S2.

Table 2: **Evaluation of models on predictive performance.**

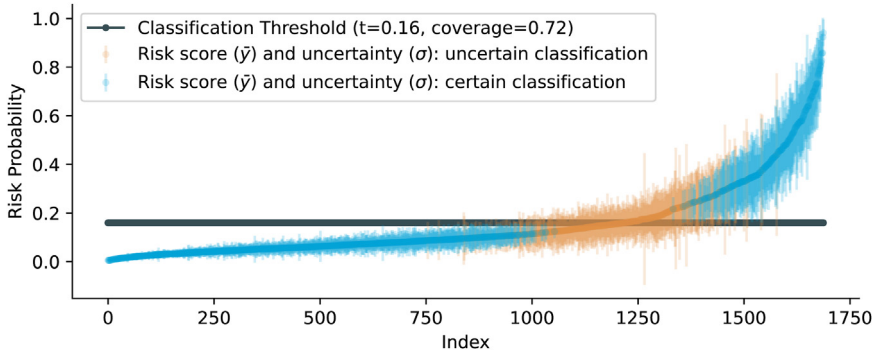

***Fig. 1:* Sorted risk probability estimates with uncertainties.** Sorted final risk predictions (mean of the predictive distribution, $\bar{y}$) with uncertainty range (standard deviation, $\pm\sigma$) for the Horseshoe-MH model. The predictions whose uncertainty does not exceed the decision threshold (certain classifications) are coloured blue, and those that do (uncertain classifications) are coloured orange. The dark grey line is our chosen classification threshold at 0.16, the event rate. The ratio of certain predictions (coverage) is 0.72.

was negatively correlated (median = −0.25). The point estimates of the coefficients from the frequentist LASSO model were within the credible interval in 19 of the 27 cases.

Discussion

In this study of over 8400 patients with cancer receiving chemotherapy treatment within a large healthcare system, found that Bayesian models were robust at predicting patient risk for acute care utilization, with performance metrics similar to common LASSO models. The Bayesian analyses provided additional information regarding the uncertainty of each point estimate. Using the BLLR approach, we can quantify the proportion of predictions in a dataset that are uncertain by identifying estimates where the uncertainty errors cross a set decision threshold. Finally, we demonstrate that BLLR predictive uncertainties can be prone to algorithmic bias, similar to other ML models.

For clinicians and health systems, individual patient predictions demonstrate how BLLRs provide a useful predictive distribution rather than just a point estimate, as these can provide insight into the risk prediction uncertainty. This can benefit providers, since patients predicted with high versus low uncertainty, could be treated differently. For instance, patients predicted with low uncertainty could be triaged to differential downstream interventions using an automated approach based on the algorithm output. Automated ML workflows have already been implemented, such as in the case of the SHIELD-RT trial, which demonstrated a reduction in ACU for high-risk radiotherapy patients assigned to increased levels of care based on risk predictions.[28] Taking this ML-based triage a step further, uncertainty estimates allow the models to alert providers when additional review by a human would be beneficial to determine downstream interventions. With their uncertainty filtering method, a similar approach has been explored by Joshi and Dhar where uncertain predictions were excluded from further ML classification.[29] Frequentist LASSO-based approaches do not provide these outputs and may appear inappropriately confident for some patients. For informaticians, the posterior distribution of the input features in the BLLRs provides insight that can be used for model improvement and feature reduction, reducing deployment and maintenance costs. Based on our findings, BLLRs perform equivalently to ordinary logistic LASSO when deployed at the point of care but can increase trust in and utility of the models.

Data bias and algorithmic fairness are a priority for medical informatics.[30] Using the BLLR approach, one can visualize potential biases in quantified uncertainty stratified by different patient groups and potentially mitigate ML-driven disparities. Our results showed increased uncertainty in some groups, including Black, stage IV, and sarcoma cancer patients. The increased uncertainty for Black and sarcoma patients may be due to our data set's comparably low patient count. However, this is different for stage IV patients. While this uncertainty bias might be reasonable for different tumour stages or types, we believe it should not be the case for demographic values such as race and insurance. In our model, Black patients have a median 67% higher typical error around their risk prediction compared to the median uncertainty for a White patient, which would not have been identified with frequentist approaches. Such analyses can improve patient care by allowing strategies to limit consequences for these patients through model revision or alerting clinicians to potential bias.

For those seeking to implement a Bayesian model at the point of care, the choice of the prior and posterior distribution is important, and we demonstrate essential differences between different BLLRs. Specifically in our case, the Metropolis-Hastings-based

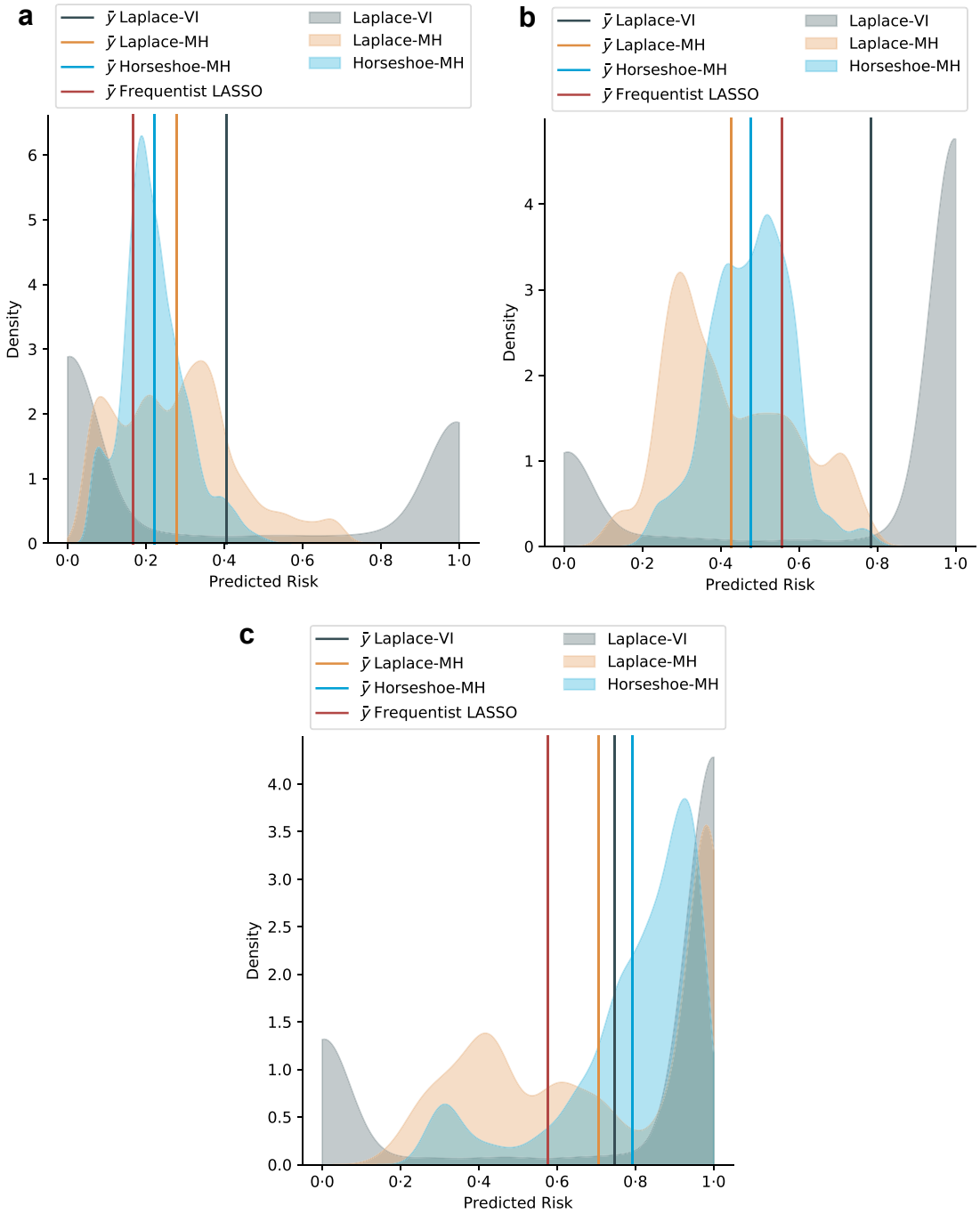

Fig. 2: **Predictive distributions of risk for individual patients.** Predictive distributions in form of probability densities of risk of acute care within 30 days after the start of chemotherapy for three individual patients with beliefs of risk: low-risk (a), mid-risk (b), high-risk (c). The KDE-plot indicate the predictive distributions, while the lines in the respective colours are the distributions' expected values ($\bar{y}$).

models outperformed the meanfield variational inference model in nearly all aspects. This underperformance is potentially because MCMC methods are asymptotically exact, while VI is not, or that a different family of distributions (non-Gaussian) could approximate the true posterior better.[31,32] Furthermore, as demonstrated in the individual patient predictions, the probability mass of the predictive distribution is often at very high and low values. Regardless of the underlying distribution for the predicted outcomes, we

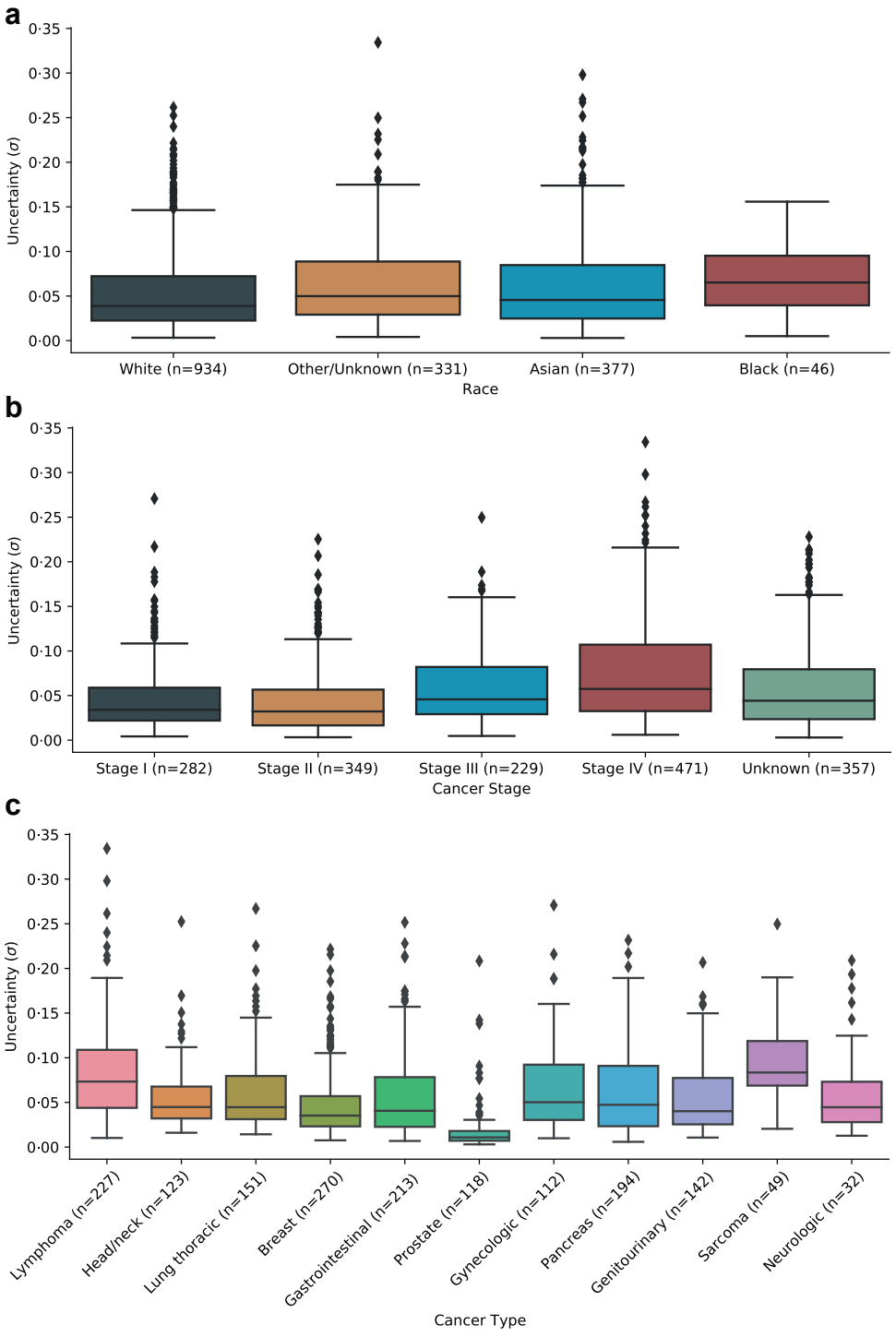

***Fig. 3:* Distribution of quantified uncertainty by patient group.** Distribution of uncertainty (standard deviation, σ) of the test set, stratified by race (a) and cancer stage (b), and cancer type (c), calculated with the Horseshoe-MH model.

suggest that those seeking to implement clinical interventions based on BLLR output leverage the 95%-credible interval for uncertainty estimates, as these have a more intuitive interpretation that clinicians are familiar with. This approach allows data scientists and clinicians to work together to determine the best-performing model, decision threshold, and quantified uncertainty based on their guidelines and available resources.

In this paper, our primary focus was directed towards Bayesian methodologies. However, it is noteworthy to mention that it is feasible to obtain predictive distributions from a frequentist LASSO as well. The approach entails fitting several models to bootstrapped training data and generating predictions, which results in an approximation of a posterior predictive distribution.[33,34] In Appendix E, we present the calibration and discriminative results from testing this approach. Our findings indicate that bootstrapping can provide outcomes comparable to the Horseshoe-MH model, and therefore may be a viable alternative to the Bayesian framework.

This study has several limitations. First, when using Bayesian analysis in practice, results are possibly sensitive to the choice of priors. This was also seen in the results, where the Laplace-MH and Horseshoe-MH demonstrated a different performance. It is, therefore, critical informed decisions about the prior, and test different feasible ones. Second, other, more sophisticated VI approximation or MCMC sampling techniques, such as the NUTS sampler, lead to more stable posterior approximations.[35] However, these techniques require substantially more computation, especially for high-dimensional input features, which may affect the feasibility of model deployment at the point of care. Third, it is difficult to truly compare the quantified uncertainties of models, as there is no ground truth for uncertainty. Different definitions of uncertainty could lead to other results in our experiments. To this end, further research is needed on the clinical utility of uncertainty estimations for ML at the point of care. Fourth, our study was validated only on one dataset for risk prediction of ACU. Testing these experiments on various medical prediction tasks and in other care systems would improve the generalizability of our work and BLLR models for clinical prediction tasks.

Conclusion

Bayesian methods can provide uncertainty estimations that promote the trust and utility of ML models in healthcare settings. Our results demonstrate that BLLR models perform similarly to frequentist logistic LASSO, even with high-dimensional data, and should be considered for future medical application studies. We additionally extend the study of algorithmic bias to uncertainty estimation and identify subgroups where bias is prevalent and should be mitigated. Overall, this work offers a paradigm shift in thinking about and using uncertainty estimates for risk scores in clinical decision support. Accounting for uncertainty increases allows clinicians to use risk predictions in a more informed context. This approach can improve automated decision-making capabilities and identify cases where ML uncertainty is high, allowing for human review. Taken together, these approaches can promote confidence through uncertainty.

Contributors

All authors confirm that they had full access to all the data in the study and accept responsibility to submit for publication. All authors read and approved the final version of the manuscript.

- Study Design: C.F. and T.H-B.
- Literature search: C.F.
- Access and verified underlying data: D.P. and C.F.
- Data analysis: C.F. and A.d.H
- Code writing, model training: C.F.
- Prepared the code repository: C.F.
- Drafting of the manuscript: C.F.
- Reporting checklist: A.C.
- Revisions of a manuscript: C.F., A.d.H, A. C., D.P., and T.H-B.
- Study supervision: T.H-B.

Data sharing statement

The EHR dataset generated during or analysed in this study is not publicly available due to restrictions by privacy laws. Requests for sharing of all data and material should be addressed to the corresponding author within 15 years of the date of publication of this Article and include a scientific proposal. Depending on the specific research proposal, corresponding author will determine when, for how long, for which specific purposes, and under which conditions the requested data can be made available, subject to ethical consent. The code for the Bayesian logistic LASSOs and our analysis is publicly available on GitHub (https://github.com/su-boussard-lab/acu-uncertainty-estimation).

Declaration of interests

The authors declare no competing financial or non-financial interests.

Acknowledgements

This work was supported in part by the National Library Of Medicine of the National Institutes of Health under Award Number R01LM013362.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2023.104632.

References

1 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115–118.
2 Sammut SJ, Crispin-Ortuzar M, Chin SF, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*. 2022;601(7894):623–629.
3 Alsentzer E, Murphy J, Boag W, et al. Publicly available clinical bert embeddings. In: *Proceedings of the 2nd clinical natural language processing workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019:72–78. Available from: https://www.aclweb.org/anthology/W19-1909.

4 Marafino BJ, Park M, Davies JM, et al. Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. *JAMA Netw Open*. 2018;1(8):e185097.
5 Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit Med*. 2021;4(1):4.
6 Dagliati A, Malovini A, Decata P, et al. Hierarchical Bayesian Logistic Regression to forecast metabolic control in type 2 DM patients. *AMIA Annu Symp Proc*. 2016;2016:470–479.
7 Beker W, Wołos A, Szymkuć S, Grzybowski BA. Minimal-uncertainty prediction of general drug-likeness based on Bayesian neural networks. *Nat Mach Intell*. 2020;2(8):457–465.
8 Syrykh C, Abreu A, Amara N, et al. Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning. *NPJ Digit Med*. 2020;3(1):63.
9 Ulmer D, Meijerink L, Cinà G. Trust issues: uncertainty estimation does not enable reliable OOD detection on medical tabular data. In: Alsentzer E, McDermott MBA, Falck F, Sarkar SK, Roy S, Hyland SL, eds. *Proceedings of the machine learning for health NeurIPS workshop. Vol. 136 of proceedings of machine learning research*. PMLR; 2020:341–354. Available from: https://proceedings.mlr.press/v136/ulmer20a.html.
10 Carlin BP, Hong H, Shamliyan TA, Sainfort F, Kane RL. *Case study comparing bayesian and frequentist approaches for multiple treatment comparisons*. Rockville (MD): Agency for Healthcare Research and Quality (US); 2013.
11 Peterson DJ, Ostberg NP, Blayney DW, Brooks JD, Hernandez-Boussard T. Machine learning applied to electronic health records: identification of chemotherapy patients at high risk for preventable emergency department visits and hospital admissions. *JCO Clin Cancer Inform*. 2021;(5):1106–1126. https://doi.org/10.1200/CCI.21.00116. PMID:34752139.
12 Centers for Medicare Medicaid Services. 2019 chemotherapy measure facts admissions and emergency department (ED) visits for patients receiving outpatient chemotherapy hospital outpatient quality reporting (OQR) program (OP-35) (non peer-reviewed). https://qualitynet.cms.gov/files/5dcc6762a3e7610023518e23?filename=CY21_OQRChemoMsr_FactSheet.pdf.
13 Brooks GA, Li L, Uno H, Hassett MJ, Landon BE, Schrag D. Acute hospital care is the chief driver of regional spending variation in Medicare patients with advanced cancer. *Health Aff*. 2014;33(10):1793–1800.
14 Yabroff KR, Lamont EB, Mariotto A, et al. Cost of care for elderly cancer patients in the United States. *J Natl Cancer Inst*. 2008;100(9):630–641.
15 Adelson KB, Dest V, Velji S, Lisitano R, Lilenbaum R. Emergency department (ED) utilization and hospital admission rates among oncology patients at a large academic center and the need for improved urgent care access. *J Clin Oncol*. 2014;32(30_suppl):19. https://doi.org/10.1200/jco.2014.32.30_suppl.19. PMID: 28141471.
16 Uno H, Jacobson JO, Schrag D. Clinician assessment of potentially avoidable hospitalization in patients with cancer. *J Clin Oncol*. 2014;32(30_suppl):4.
17 Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum information for medical AI reporting): developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc*. 2020;27:2011–2015.
18 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350(jan07 4):g7594.
19 Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Precup D, Teh YW, eds. *Proceedings of the 34th international conference on machine learning. Vol. 70 of proceedings of machine learning research*. PMLR; 2017:1321–1330. Available from: https://proceedings.mlr.press/v70/guo17a.html.
20 Kumar A, Liang P, Ma T. Verified uncertainty calibration. In: *Advances in neural information processing systems (NeurIPS)*. 2019.
21 Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167–176.
22 Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med*. 2014;33(3):517–535.
23 McGill R, Tukey JW, Larsen WA. Variations of box plots. *Am Stat*. 1978;32(1):12–16.
24 Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res*. 2019;3:18.
25 Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12(Oct):2825–2830.
26 Salvatier J, Wiecki TV, Fonnesbeck CJ. Probabilistic programming in Python using PyMC3. *PeerJ Comput Sci*. 2016;2:e55.
27 Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc*. 1952;47(260):583–621. Available from: https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441.
28 Hong JC, Eclov NCW, Dalal NH, et al. System for high-intensity evaluation during radiation therapy (SHIELD-RT): a prospective randomized study of machine learning-directed clinical evaluations during radiation and chemoradiation. *J Clin Oncol*. 2020;38(31):3652–3661.
29 Joshi P, Dhar R. EpICC: a Bayesian neural network model with uncertainty correction for a more accurate classification of cancer. *Sci Rep*. 2022;12(1):14628.
30 McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. *Lancet Digit Health*. 2020;2(5):e221–e223.
31 Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: a review for statisticians. *J Am Stat Assoc*. 2016;112:859–877.
32 Salimans T, Kingma D, Welling M. Markov chain Monte Carlo and variational inference: bridging the gap. In: Bach F, Blei D, eds. *Proceedings of the 32nd international conference on machine learning. Vol. 37 of proceedings of machine learning research*. Lille, France: PMLR; 2015:1218–1226. Available from: https://proceedings.mlr.press/v37/salimans15.html.
33 Efron B. Bootstrap methods: another look at the jackknife. *Ann Stat*. 1979;7:1–26.
34 Alfaro M, Zoller S, Lutzoni F. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Mol Biol Evol*. 2003;20(2):255–266.
35 Homan MD, Gelman A. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J Mach Learn Res*. 2014;15(1):1593–1623.