Bayesian Averaging over Decision Tree Models for Trauma Severity Scoring

V. Schetinin*, L. Jakaite

University of Bedfordshire

W. Krzanowski

University of Exeter

Abstract

Health care practitioners analyse possible risks of misleading decisions and need to estimate and quantify uncertainty in predictions. We have examined the "gold" standard of screening a patient's conditions for predicting survival probability, based on logistic regression modelling, which is used in trauma care for clinical purposes and quality audit. This methodology is based on theoretical assumptions about data and uncertainties. Models induced within such an approach have exposed a number of problems, providing unexplained fluctuation of predicted survival and low accuracy of estimating uncertainty intervals within which predictions are made. Bayesian method, which in theory is capable of providing accurate predictions and uncertainty estimates, has been adopted in our study using Decision Tree models. Our approach has been tested on a large set of patients registered in the US National Trauma Data Bank and has outperformed the standard method in terms of prediction accuracy, thereby providing practitioners with accurate estimates of the predictive posterior densities of interest that are required for making risk-aware decisions.

Keywords: Bayesian method, Decision Tree, Predictive posterior distribution, Injury Severity Scoring

1. Introduction

Health care systems based on Machine Learning (ML) technologies are increasingly demanded for the prevention of lifestyle-related and chronic diseases as well as for emergency care and life support, see e.g. [1, 2]. This interest is explained by the need for efficient access to data related to patients' conditions in

*Corresponding author
Email address: v.schetinin@beds.ac.uk (V. Schetinin)

ambulance, hospital or home environments. To assist health care practitioners with decision making, these data can be analysed using various ML approaches.

Million people worldwide are injured and admitted to hospitals for emergency treatment. Just in 2014, around 40 million people were treated in the US and 192,945 of them obtained fatal injuries and died [3]. Reliable, accurate, and timely information about a patient's condition is therefore of critical importance for improving trauma care outcomes.

For evaluation of injury severity and prediction of survival, practitioners exploit a logistic regression model known as the Trauma and Injury Severity Score (TRISS). The TRISS model predicts the probability of survival for a patient on arrival at a hospital, see e.g. [4, 5, 6, 7]. The prediction model combines screening information and physiological parameters recorded by a paramedic at an accident scene. For some patients, data that are obtained with medical devices, such as blood pressure and heart rate, can be missing at the moment of the examination.

Uncertainties that exist in data as well as in the prediction model will affect the results and might lead to fatal errors or inadequate treatment. For this reason, practitioners have raised a concern about the ability of TRISS to provide reliable and accurate predictions and estimates of uncertainty [8, 5].

The accuracy of predictions is compared against actual survival during model calibration. A relationship between the predicted and actual probabilities can be visualised as a calibration curve [9]. In this sense, the TRISS calibration curve has drifted away from the ideal curve, see e.g. [10, 6, 8].

In [5], it has been found that the accuracy of TRISS predictions is acceptable when the types and severities of patients injuries are typical. However, for patients with four or more injuries as well as those with atypical combinations of injuries, the accuracy has to be improved.In practice, it is critically important to accurately estimate the uncertainty in a predicted survival probability. The uncertainty estimates are required in order to minimise risks of fatal errors. Uncertainty can be represented by confidence intervals. These intervals are reliably estimated when the density of predicted probabilities is fully tractable, which is achievable only in trivial cases. Thus TRISS methodology that is based on theoretical assumptions cannot realistically estimate the uncertainty [11, 12].

To tackle the above problems, we employ the Bayesian approach to learn prediction models from data. This methodology in theory provides the most accurate predictions and uncertainty estimation, see e.g. [13, 14, 15, 16]. This approach, however, requires intensive computations, see e.g. [17, 18, 19].

In our approach we use Bayesian averaging over Decision Tree (DT) models, also known as Classification and Regression Trees, which are well-known for their ability to select input variables that maximally improve the performance [20, 21, 22]. DT models split the given data along input variables recursively, which is relatively simple to compute. The strategy, however, cannot provide a global view on the entire data. At the same time, the partitions which are made along variables are transparent, and when the number of the partitions is reasonably small then DT models can assist users with new insights into the data, see e.g. [23].

We analyse existing approaches and describe our approach based on Bayesian averaging over DT models. Then we test and compare the proposed and TRISS methods on the main trauma data benchmark, the US National Trauma Data Bank (NTDB) [24]. The comparison of the methods is made in terms of the Area Under the receiver operating characteristic Curve (AUC), which is a summary measure of the accuracy of a quantitative diagnostic test, see e.g. [25]. Finally, we discuss a DT model that can be used for purposes of interpretation with the maximum predictive ability.

2. Logistic Regression Model for Predicting Survival Probability

Logistic regression modelling is a way of calculating probabilities of survival for given predictors, see e.g. [4, 9]. As such, the TRISS model includes both continuous and categorical tests. The former include: age, systolic blood pressure, and respiratory rate, while the latter include: severity scores of injuries that a patient can obtain, the Glasgow Coma Scale (GCS), and the type of injury. Screening tests are evaluated on the patient's arrival by a trained scorer, see e.g. [6].

The above screening tests form two aggregated predictors: Injury Severity Score (ISS), and Revised Trauma Score (RTS). However, practitioners have found that such an aggregation causes unexplained fluctuations of the ISS over observed probabilities of survival, which affects the prediction accuracy, see e.g. [5, 26]. The calculation of survival probabilities has been made available online as a TRISS Calculator [27].

The current standard TRISS allows for up to three of the most severe injuries that a patient can obtain in six regions of the body: head, face, chest, abdomen, extremities, and external (skin, subcutaneous tissue and burns).

Within this methodology, a density of predicted values is assumed to be a Gaussian distribution, $N(\mu, \sigma^2)$, where μ and σ^2 are defined by the parameters b and by the regression error, respectively. As follows from [28], under such an assumption, the uncertainty interval for a prediction cannot be realistically estimated for a patient.

3. Methodology

In most practical cases, any given model is incapable of fully explaining the real-world data, which means that a single "true" model does not exist. The method of Bayesian averaging over models, adopted in our study, assumes that different models can be mixed together so that their average under certain conditions will approximate the true model of interest. The averaging strategy is often more efficient than model selection in real-world applications when the predictive ability (or fitness function) is not unimodal, see e.g. [28, 16].

The use of DT models within Bayesian method gives us the following advantages [20, 29]. In comparison with other Machine Learning methods, the DT technique is directly applied to the given data without time-consuming data

preprocessing or careful tuning of the learning algorithm, and so the DT technique is often called "off-the-shelf". The following advantages could be also important when dealing with real-world problems. DT models are robust to outliers in the given data. When a domain problem is represented by a mix of numerical and categorical variables, the DT technique naturally captures the relationships between them.

DTs perform internal feature selection as an integral part of the learning procedure, and so the use of general data transformations, such as Principal Component Analysis, is not required, see e.g. [30].

Models learnt from given data are calibrated and their accuracy is statistically evaluated by goodness-of-fit tests. In the medical domain the calibration is usually assessed via the Hosmer-Lemeshow (HL) statistic [9, 8]. HL statistics are typically calculated for 10 intervals of predicted values. Under certain conditions, the larger the HL statistic, the worse is the calibration. The HL-test, however, is statistically significant in 100% of models when the number of patients is 50,000 or more. So this test has to be analysed along with the overall number of patients, see e.g. [31], which has been taken into account in our experiments.

The HL-test of goodness-of-fit is typically used along with others metrics of medical decision-making models, such as sensitivity and specificity, True Positive (TP) and False Positive (FP) rates. We also compare the diagnostic potentials of the proposed and existing methods in terms of AUC as discussed in Section 1.

4. Data

For comparison of the proposed and standard TRISS methods, we use a set of patient records from the US NTDB, the major source of data about injured patients admitted to hospitals and emergency units [24]. The data include patient age, gender, type and regions of injuries along with some clinical and background information about patient state. The NTDB also includes the TRISS prediction and the outcome of care, alive or died, for each patient.

Table 1 shows the screening tests (or predictors) that are used by the TRISS method based on the NTDB. The variables *Age*, *Blood pressure*, and *Respiration rate* are continuous, and the remaining variables are categorical. The patient outcome is the *discharge status*, $y \in \{0, 1\}$, where 0 is alive, and 1 is died. The table also shows the minimal and maximal values of each test.

For our study, we selected records of patients with 1-20 injuries. After exclusion of missing values, the number of records was 571,148. Approximately 11% of the data were missing not at random. For example, the respiration rates were undefined for intubated patients. However, both our approach and TRISS method consider only complete cases, see e.g. [4]. Thus the bias caused by excluding patient data with missing values not at random is not investigated.

The distribution of the records over 3 groups of injuries was as follows: (1) 174,647 with 1 injury, (2) 381,137 with 2-10 injuries, and (3) 15,364 with 11-20 injuries. Survival rates in these groups were 0.977, 0.953, and 0.831, respectively.

Table 1: Screening tests and ranges of NTDB

#	*Name*	*min*	*max*
1	Age	0	100
2	Gender	0 female	1 male
3	Injury type	0 penetrating	1 blunt
4	Blood pressure	0	300
5	Respiration rate	0	200
6	GCS Eye	1	4
7	GCS Verbal	1	5
8	GCS Motor	1	6
9	Head severity	0	6
10	Face severity	0	4
11	Neck severity	0	6
12	Thorax severity	0	6
13	Abdomen severity	0	6
14	Spine severity	0	6
15	Upper extremity severity	0	6
16	Lower extremity severity	0	6
17	External severity	0	6

Table 2 shows statistics of the screening tests A to E_S (listed in Table 1) in the groups of patients who obtained 1-20, 1, 2-10, and 11-20 injuries. The statistics are represented by values of the mean, standard deviation, median, and quartiles.

The records of patients with the largest numbers of injuries in group 3 were equally split into 2 subsets, one for training and the other for validating the model. The model was learnt from the training subset, and its ability to predict new data was analysed on the remaining data including groups 1, 2, and the validation subset of group 3. The model calibration was analysed on all groups.

Table 2: Statistics of the screening tests over four patient groups with the following number of injuries: 1-20,1,2-10, and 11-20

		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}
Mean	1-20	39.3	0.7	0.9	134.6	19.1	3.7	4.6	5.6	0.9	0.3	0.0	0.7	0.4	0.4	0.6	0.8	0.1
	1	41.3	0.6	0.8	136.4	19.3	3.9	4.8	5.8	0.4	0.1	0.0	0.2	0.2	0.2	0.3	0.8	0.1
	2-10	38.4	0.7	0.9	134.3	19.0	3.7	4.5	5.5	1.1	0.4	0.0	0.8	0.4	0.4	0.7	0.8	0.1
	11-20	37.7	0.7	1.0	120.8	17.3	2.9	3.3	4.3	2.5	1.2	0.1	2.5	1.4	1.3	1.4	1.8	0.2
Standard deviation	1-20	22.6	0.5	0.3	29.7	7.1	0.8	1.1	1.3	1.5	0.7	0.2	1.3	1.0	0.9	0.9	1.2	0.3
	1	25.4	0.5	0.4	28.8	5.7	0.6	0.8	0.8	1.0	0.3	0.2	0.7	0.7	0.6	0.7	1.2	0.3
	2-10	21.2	0.5	0.3	29.6	7.3	0.9	1.2	1.3	1.5	0.7	0.3	1.4	1.0	1.0	1.0	1.1	0.4
	11-20	18.3	0.5	0.2	36.9	13.0	1.4	1.8	2.2	1.8	1.0	0.5	1.6	1.5	1.3	1.1	1.3	0.4
Median	1-20	36	1	1	135	20	4	5	6	0	0	0	0	0	0	0	0	0
	1	37	1	1	136	19	4	5	6	0	0	0	0	0	0	0	0	0
	2-10	35	1	1	135	20	4	5	6	0	0	0	0	0	0	0	0	0
	11-20	35	1	1	124	18	4	4	6	3	1	0	3	1	2	2	2	0
1^{st} quartile	1-20	21	0	1	119	16	4	5	6	0	0	0	0	0	0	0	0	0
	1	20	0	1	120	16	4	5	6	0	0	0	0	0	0	0	0	0
	2-10	21	0	1	120	16	4	5	6	0	0	0	0	0	0	0	0	0
	11-20	22	0	1	104	12	1	1	1	1	0	0	1	0	0	0	1	0
3^{rd} quartile	1-20	54	1	1	150	22	4	5	6	2	1	0	1	0	0	1	2	0
	1	60	1	1	152	20	4	5	6	0	0	0	0	0	0	0	2	0
	2-10	52	1	1	150	22	4	5	6	2	1	0	1	0	0	1	2	0
	11-20	49	1	1	143	24	4	5	6	4	2	0	4	2	2	2	3	0

5. Bayesian Averaging over DT Models

Bayesian method is analytically tractable for trivial cases when the likelihood function and probability distributions of given data and model parameters are known. In practice, when the distributions are unknown, the Bayesian method can be feasibly approximated with Markov chain Monte Carlo (MCMC) methods, see e.g. [28]. Bayesian averaging over DT models has been implemented with MCMC in [32, 17].

MCMC aims to explore a posterior density of model parameters by making random walk proposals. The desired density is approximated by drawing samples from areas of high posterior density, so-called areas of interest. Then samples of model parameters are used for calculating probabilities of predicted outcomes that are distributed according to a predictive posterior distribution function of interest.

The MCMC approximation can be outlined as follows. First we define parameters, Θ, of a prediction model that can be learnt from a labelled data set, $\mathbf{D} = (x^{(i)}, y^{(i)})_{i=1}^{n}$, where $x = (x_1, \ldots, x_m)$ are the m-dimensional input vectors, y are the model outcomes, and n is the numbers of instances in the data $\mathbf{D}$. Given an input x, the predicted outcome y is assigned to one of the given classes, $y \in \{1, C\}$. In our case of survival prediction, $C = 2$, and the outcomes are: $y = 0$ if a patient is survived, and $y = 1$ if died.

The predictive posterior distribution of interest, $p(y|x, \mathbf{D})$, is calculated as an integral over model parameters Θ as follows:

$$p(y|x, \mathbf{D}) = \int_{\Theta} p(y|x, \Theta) p(\Theta|\mathbf{D}) \mathrm{d}\Theta, \tag{1}$$

were $p(y|x, \Theta)$ is the posterior predictive density given input x and model parameters Θ, and $p(\Theta|\mathbf{D})$ is the posterior density of Θ given data $\mathbf{D}$.

The above integral is analytically tractable only in trivial cases when the distribution $p(\Theta|\mathbf{D})$ is known. In practice, we can generate N samples, $\left(\Theta^{(i)}\right)_{i=1}^{N}$, distributed with a density function, $\hat{p}(\Theta|\mathbf{D})$, that under certain conditions can be accurately simulated with MCMC:

$$\Theta^{(i)} \sim \hat{p}(\Theta|\mathbf{D}). \tag{2}$$

The desired approximation is achieved when MCMC generates a random sequence with a stationary probability distribution. Thus we can draw samples $\Theta^{(i)}$ defined in Eq. 2 and then calculate the predictive density of interest as follows:

$$p(y|x, \mathbf{D}) \approx \sum_{i=1}^{N} p(y|x, \Theta^{(i)}, \mathbf{D}) p(\Theta^{(i)}|\mathbf{D}) = \frac{1}{N} \sum_{i=1}^{N} p(y|x, \Theta^{(i)}, \mathbf{D}). \tag{3}$$

From Eq. 2, the required model parameters $\Theta^{(i)}$ are drawn from a posterior distribution simulated by MCMC. The collected samples are then used in Eq. 3 to calculate the posterior predictive probabilities.

In practice, the model parameter space can be very large, so the desired approximation of $p(\Theta|D)$ is achieved with the Reversible Jump (RJ) extension of MCMC [33]. The approximation becomes accurate when RJ MCMC algorithm can explore all areas of high posterior density. However, when posterior density functions are multimodal, the detailed exploration of the areas of interest cannot be achieved in a reasonable time, see e.g. [28]. In order to mitigate the above problems, we proposed a new RJ MCMC strategy in our previous work [34]. This strategy has delivered more accurate results than the existing MCMC techniques that employ the restarting [32] and restricting [17] strategies, as well as randomised DT strategy [35].

DT models are grown to be fitted to given data by the RJ MCMC sampler. The sampler can search model parameters of variable dimensionality by making the following types of moves:

1. *Birth*. To randomly split the data points falling in one of the terminal nodes by adding a new splitting node with a variable and rule drawn from a given prior.
2. *Death*. To randomly pick a DT splitting node with two terminal nodes to be assigned a single terminal node with the merged data points.
3. *Change-split*. To randomly pick a splitting node and assign it a new splitting variable and rule drawn from a given prior.
4. *Change-rule*. To randomly pick a splitting node and assign it a new rule drawn from a given prior.

The birth and death moves are reversible and change the dimensionality of the model parameters. The change moves are required to search the parameters within the current dimensionality of the model.

In the next section we describe our experiments with these methods for the prediction of survival probabilities and assessment of uncertainty.

6. Results

The proposed method was tested and compared with the standard TRISS method on the US NTDB data outlined in Section 4. The comparison was made in terms of classification accuracy and goodness-of-fit (or calibration) using the HL statistic as discussed in Section 4.

According to the HL-test, calibration curves were calculated for 10 intervals equidistantly distributed over survival probabilities $[0, 1]$. The curves calculated for the TRISS and proposed Bayesian DT (BDT) methods are shown in Fig. 1. Here the confidence intervals, calculated as 25^{th} and 75^{th} percentiles, are shown by the lines, and the actual (observed) survival probabilities falling within the intervals are denoted by the filled circles.

There is a significant difference in the plotted curves in terms of accuracy of estimating the confidence intervals. The TRISS calibration on the left-hand plot in Fig. 1 shows that most of the observed survival probabilities (marked by the circles) lie outside the confidence interval (solid lines). By contrast for

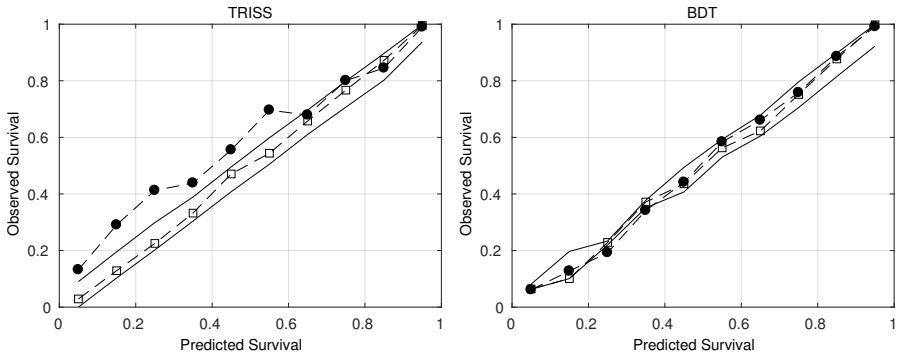

Figure 1: Calibration curves and confidence intervals for TRISS (left) and proposed BDT (right) methods for patients with 1-20 injuries. Observed survivals are denoted by the filled circles, predicted survivals are the squares, and confidence intervals are the solid lines.

our method, all the observed survival probabilities shown in the right-hand plot lie within the confidence interval, which is evidence of more accurate estimation. The values of the HL statistics were 3,893.4 and 557.1 for the TRISS and Bayesian methods, respectively.

We see that the second result is just one-sixth of the first, and thus the Bayesian method is considerably better in terms of the HL statistic. A high value of the HL statistic for the Bayesian method is explained by the large number of patients, as we discussed in Section 4.

The HL-test shows a significant improvement of goodness-of-fit of the Bayesian method. In terms of uncertainty, 2σ intervals were reduced from 0.628 to 0.569 for patients with 2-10 injuries and from 1.227 to 0.930 for patients with 11-20 injuries. Fisher's F-test shows that these improvements are statistically significant (p-value < 0.005).

The proposed and TRISS methods were also compared in terms of classification accuracy. The patients with predicted survival ≥ 0.5 were assigned to class "alive", and the others with smaller probability to class "dead". Table 3 shows the classification accuracy, sensitivity, and specificity along with AUC values and TP rates given for FP=0.1. We can see that the accuracy and AUC of the Bayesian method are slightly higher in all injury groups. Fig. 2 shows the ROC curves for both BDT and TRISS methods.

Table 3: Classification accuracy of the TRISS (TRS) and BDT methods in the injury groups: 1-20, 1, 2-10, and 11-20.

	Accuracy		*Sensitivity*		*Specificity*		*AUC*		*TP(FP=0.1)*	
Inj.	*TRS*	*BDT*	*TRS*	*BDT*	*TRS*	*BDT*	*TRS*	*BDT*	*TRS*	*BDT*
1-20	0.968	0.971	0.528	0.474	0.988	0.994	0.948	**0.954**	0.855	0.858
1	0.986	0.987	0.487	0.518	0.998	0.998	0.945	**0.951**	0.815	0.824
2-10	0.964	0.968	0.517	0.464	0.987	0.993	0.946	**0.954**	0.849	0.864
11-20	0.838	0.875	0.664	0.475	0.874	0.956	0.882	**0.894**	0.621	0.660

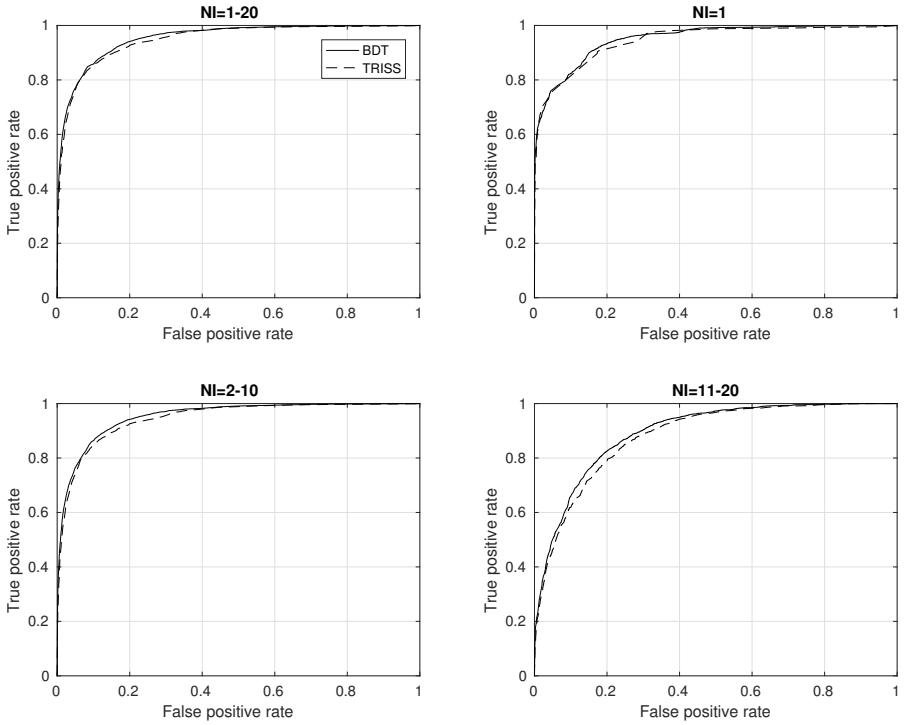

Figure 2: ROC curves for TRISS and proposed methods in the injury groups: 1-20, 1, 2-10, and 11-20. The solid line denotes ROC for BDT and the dashed line denotes ROC for TRISS. In all groups AUC for BDT is larger than for TRISS.

We applied the McNemar statistical test in order to estimate the changes in proportions of outcomes that are predicted by the Bayesian and TRISS methods on patients in the three groups. The results show that the proposed Bayesian method has improved the prediction accuracy by 0.04%, 0.36%, and 3.64% (p-value < 0.05) for patients with 1, 2-10, and 11-20 injuries, respectively.

7. Discussion

In this section we discuss the main findings and work related to applications of the Bayesian method.

7.1. Main Findings

Practitioners are guided by the TRISS methodology of evaluating injury severity and prediction of survival probabilities. TRISS is the current standard for evaluation of patients on arrival at a hospital as discussed in Section 1. The TRISS method outlined in Section 2 analyses the screening tests described in Section 4.

The results of the TRISS evaluation are affected by uncertainties that exist in both data and prediction model. These uncertainties increase risks of making fatal errors. Practitioners are unsatisfied with the ability of TRISS to make reliable predictions when errors affect patient outcomes [8, 5], they are concerned that goodness-of-fit of the TRISS model is not ideal [8], and they have also found that the accuracy of TRISS predictions for patients with four or more injuries as well as with atypical injuries needs to be improved [5].

Uncertainty intervals within which predictions are distributed can be reliably estimated if the distribution function of predicted probabilities is known. The TRISS methodology is based on theoretical assumptions about distributions of probabilities, and so cannot reliably estimate intervals of interest [5].

In Section 5 a Bayesian approach to the above problems has been described. Bayesian method is well known for accurate modelling and estimation of uncertainty, which however require intensive computations. The methodology applied to DT models is made feasible with MCMC, which under certain conditions can accurately approximate a parameter density function of interest [28, 17].

In our experiments on the US NTDB described in Section 6, we found that the Hosmer-Lemeshow statistical test shows a significant improvement of calibration for the new model. As a result, uncertainty intervals shown in Fig. 1 were reduced from 0.628 to 0.569 for patients with 2-10 injuries and from 1.227 to 0.930 for patients with 11-20 injuries. The improvement of model calibration is the key factor for increasing the accuracy of predictions and classification.

Our contribution is that the accuracy of the proposed method has been improved by 0.36% for patients with 2-10 injuries and by 3.64% for patients with 11-20 injuries. Further improvements of evaluation accuracy could be achieved with new variables, such as the number of injuries, added to the standard screening tests. The current TRISS method that is based on these tests should then be modified in order to provide trauma care practitioners and researchers with a new "gold" standard.

7.2. Related Work on Bayesian Applications

In terms of usability for reasoning, DT models are complementary to Bayesian Networks (BNs) that are graphical models where the nodes represent variables and the arcs represent probabilistic dependencies between these variables. Such networks are called Directed Acyclic Graphs (DAGs), see e.g. [16, 36].

There are two main types of methods for learning BN structures from given data: (i) methods based on scoring functions plus search, and (ii) methods based on conditional independence tests, also known as constraint-based methods. The first type of methods aims to find a graph that maximises a score function which is designed to measure fitness of the graph to the given data. The scoring function and search method are designed so as to define a space of feasible solutions [16]. The second type of methods aims to find a BN that explains the dependencies between variables in the best way. However the result can be sensitive to failures in individual independence tests when BNs are large [16].

It is important to note that the above score-based and constraint based methods employ model selection, whereas our approach is based on Bayesian averaging over models.

BNs have been studied to model the uncertainty of factors that influence the performance of emergency medical service at a hospital [37]. Different algorithms for learning of BNs have been applied and compared in the specific case of the emergency service in order to develop a management-oriented decision support system aiming to improve the health service. The best result has been obtained with a scoring-based algorithm using a local search in the space of DAGs. Starting from an initial DAG the algorithm makes step-by-step local changes that maximise the gain until a local maximum is reached. This strategy is implemented by adding or deleting an arch in the BN.

In [38] the BNs and DTs have been learned to predict access to renal transplant waiting lists on a data set including 809 patients. It is interesting that both models have provided the same prediction accuracy, using the variable age that makes the greatest contribution. Using the BN, physicians have a global view of the relationships between variables, while the DT was more easily interpretable.

A new approach to building BNs from given data has been recently proposed in [39] within a generalised framework capable of providing a repeatable method for building BN models from patient questionnaires and interviews containing contradictory responses. Learning of the BN model structure and parameters from such data is often unsatisfactory. The proposed framework has demonstrated an improvement in predictive performance in terms of AUC scores. Besides the improvement, the resultant BN models were useful for medical decision support.

In study [40], BNs have outperformed DTs in terms of accuracy and ability to model the complexity of the underlying decision-making. However BNs were limited in terms of interpretation and efficiency of rules derived from the BN, while rules derived from DTs have a simple and direct interpretation. The idea of combining DTs and BNs has been shown to be capable of maintaining the potential advantages of both techniques.

It is important to note that the above methods of learning BNs provide practitioners with a global view on clinical decision-making, while DTs are local models which are limited in this capacity, see e.g. [39]. This advantage, however, is difficult to achieve without an expert's knowledge and theoretical insights. At the same time, DT models provide a direct interpretation of rules learnt from given data, see e.g. [38].

However, when a domain problem is represented by observations only, and experts have no knowledge to share, a DT model can be learnt from the data more efficiently than a BN. This is because the space of possible DT structures is smaller than that for possible BN structures. The use of heuristic-based search methods does not guarantee finding the best result. However when experts share their knowledge, BN can outperform DT models. The BN framework is more convenient than the DT one in terms of the ability to formulate and incorporate the expert's knowledge [36, 16].

From this point of view, BNs and DTs are complementary to each other. Despite the differences, BN models can be used within the MCMC methodology, see e.g. [16], to approximate predictive posterior distribution when estimates of uncertainty intervals for making risk-aware decisions are needed.

8. Conclusion

We analysed the standard logistic regression model, known as the Trauma and Injury Severity Score (TRISS), that is currently used for predicting survival of injured patients and found areas where TRISS methodology can be improved in terms of accuracy of prediction and uncertainty estimation. The main findings are as follows.

The TRISS methodology does not support the estimation of predictive survival probability density that is required for evaluating an individual confidence interval for a patient in order to assist practitioners with making risk-aware decisions. Trauma care practitioners found unexplained deviations in the TRISS calibration curve, which can lead to inadequate decisions. They also found that the accuracy of predicting outcomes of patients with multiple injuries has to be improved.

To improve the evaluation accuracy, we proposed a Bayesian method for prediction and uncertainty modelling. The proposed method was compared with TRISS on the large set of records included in the US NTDB, the main data repository in trauma care research.

In our experiments we found that the goodness-of-fit of the Bayesian method is superior to that evaluated for the TRISS method. The proposed method was shown to be capable of reducing uncertainty intervals and increasing the prediction accuracy in all groups of patients, especially in the group with multiple injuries. The achieved improvements were statistically significant.

Acknowledgement

This research was partly supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant GR/R24357/01 "Critical Systems and Data-Driven Technology".

References

[1] C. Free, G. Phillips, L. Watson, L. Galli, L. Felix, P. Edwards, V. Patel, A. Haines, The effectiveness of mobile-health technologies to improve health care service delivery processes: A systematic review and meta-analysis, PLoS Med 10 (1) (2013) 1–26. doi:10.1371/journal.pmed.1001363.

[2] B. Dinesen, B. Nonnecke, D. Lindeman, E. Toft, K. Kidholm, K. Jethwani, M. H. Young, H. Spindler, U. C. Oestergaard, A. J. Southard, M. Gutierrez, N. Anderson, M. N. Albert, J. J. Han, T. Nesbitt, Personalized telehealth in the future: A global research agenda, J Med Internet Res 18 (3) (2016) e53. doi:10.2196/jmir.5257.

[3] National Center for Health Statistics, National health interview survey, http://www.cdc.gov/nchs/fastats/injury.htm (2014).

[4] C. R. Boyd, M. A. Tolson, W. S. Copes, Evaluating trauma care: The TRISS method, Journal of Trauma 27 (1984) 370–378.

[5] P. Kilgo, J. Meredith, T. Osler, Injury severity scoring and outcomes research, in: D. V. Feliciano, K. L. Mattox, E. E. Moore (Eds.), Trauma (6th ed), New York, McGraw-Hill, 2008, pp. 223–230.

[6] O. Bouamra, A. Wrotchford, S. Hollis, A. Vail, M. Woodford, F. Lecky, A new approach to outcome prediction in trauma: A comparison with the TRISS model, Journal of Trauma 61 (3) (2006) 701–710.

[7] R. Lefering, S. Huber-Wagner, U. Nienaber, M. Maegele, B. Bouillon, Update of the trauma risk adjustment model of the traumaregister dgu^{TM}: the revised injury severity classification, version ii, Critical Care 18 (5) (2014) 476. doi:10.1186/s13054-014-0476-2.

[8] F. Rogers, T. Osler, M. Krasne, A. Rogers, E. Bradburn, J. Lee, D. Wu, N. McWilliams, M. Horst, Has TRISS become an anachronism? A comparison of mortality between the National Trauma Data Bank and major trauma outcome study databases, Journal of Trauma and Acute Care Surgery 73 (2) (2012) 326–331.

[9] E. Steyerberg, A. Vickers, N. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. Pencina, M. Kattan, Assessing the performance of prediction models: A framework for traditional and novel measures, Epidemiology 21 (1) (2010) 128–138.

[10] D. Becalick, T. Coats, Comparison of artificial intelligence techniques with UKTRISS for estimating probability of survival after trauma. UK Trauma and Injury Severity Score, Journal of Trauma 51 (1) (2001) 123–133.

[11] T. Bailey, R. Everson, J. Fieldsend, W. Krzanowski, D. Partridge, V. Schetinin, Representing classifier confidence in the safety critical domain an illustration from mortality prediction in trauma cases, Neural Computing and Applications 16 (3) (2007) 1–10.

[12] V. Schetinin, L. Jakaite, J. Jakaitis, W. Krzanowski, Bayesian decision trees for predicting survival of patients: a study on the US National Trauma Data Bank, Computer Methods and Programs in Biomedicine 111 (3). doi:10.1016/j.cmpb.2013.05.015.

[13] P. Magni, G. Sparacino, R. Bellazzi, G. M. Toffolo, C. Cobelli, Insulin minimal model indexes and secretion: Proper handling of uncertainty by a Bayesian approach, Annals of Biomedical Engineering 32 (7) (2004) 1027–1037. doi:10.1023/B:ABME.0000032465.75888.91.

[14] Z. Schechner, G. Luo, J. J. Kaufman, R. S. Siffert, A poisson process model for hip fracture risk, Medical & Biological Engineering & Computing 48 (8) (2010) 799–810. doi:10.1007/s11517-010-0638-6.

[15] A. Achilleos, C. Loizides, M. Hadjiandreou, T. Stylianopoulos, G. D. Mitsis, Multiprocess dynamic modeling of tumor evolution with Bayesian tumor-specific predictions, Annals of Biomedical Engineering 42 (5) (2014) 1095–1111. doi:10.1007/s10439-014-0975-y.

[16] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning, The MIT Press, 2009.

[17] D. Denison, C. Holmes, B. Mallick, A. Smith, Bayesian Methods for Nonlinear Classification and Regression, Wiley, 2002.

[18] W. J. Krzanowski, T. C. Bailey, D. Partridge, J. E. Fieldsend, R. M. Everson, V. Schetinin, Confidence in classification: A bayesian approach, Journal of Classification 23 (2) (2006) 199–220.

[19] M. A. Negrin, J. Nam, A. H. Briggs, Bayesian solutions for handling uncertainty in survival extrapolation, Medical Decision Making.

[20] V. Schetinin, L. Jakaite, Classification of newborn EEG maturity with Bayesian averaging over decision trees, Expert Systems with Applications 39 (10) (2012) 9340–9347.

[21] L. Jakaite, V. Schetinin, C. Maple, Bayesian assessment of newborn brain maturity from two-channel sleep electroencephalograms, Comp. Math. Methods in Medicine 2012 (2012) 629654:1–629654:7. doi:10.1155/2012/629654.
URL https://doi.org/10.1155/2012/629654

[22] V. Schetinin, L. Jakaite, Extraction of features from sleep eeg for bayesian assessment of brain development, PLoS ONE 2 (3). doi:10.1371/journal.pone.0174027.

[23] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Chapman and Hall, 1984.

[24] The American College of Surgeons, National Trauma Data Bank, https://www.ntdbdatacenter.com (2014).

[25] W. J. Krzanowski, D. J. Hand, ROC Curves for Continuous Data, 1st Edition, Chapman & Hall/CRC, 2009.

[26] T. Osler, L. Glance, J. Buzas, D. Mukamel, J. Wagner, A. Dick, A trauma mortality prediction model based on the anatomic injury scale, Annals of Surgery 247 (6) (2008) 1041–1048.

[27] K. Brohi, TRISS - Overview and desktop calculator, http://www.trauma.org/index.php/main/article/387/ (2012).

[28] C. Robert, G. Casella, Monte Carlo Statistical Methods, Springer Texts in Statistics, Springer, 2004.

[29] L. Jakaite, V. Schetinin, Feature selection for Bayesian evaluation of trauma death risk, in: The 14th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics, Springer, 2008, pp. 123–126.

[30] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference and prediction, 2nd Edition, Springer, 2009.

[31] A. A. Kramer, J. E. Zimmerman, Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited., Critical Care Medicine 35 (9) (2007) 2052–2056.

[32] H. Chipman, E. George, R. McCullock, Bayesian CART model search, Journal of American Statistics 93 (1998) 935–960.

[33] P. J. Green, Reversible jump Markov chain Monte Carlo and Bayesian model determination, Biometrika 82 (1995) 711–732.

[34] V. Schetinin, J. E. Fieldsend, D. Partridge, W. J. Krzanowski, R. M. Everson, T. C. Bailey, A. Hernandez, Comparison of the Bayesian and randomized decision tree ensembles within an uncertainty envelope technique, Journal of Mathematical Modelling and Algorithms 5 (2006) 397–416.

[35] T. G. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization, Machine Learning 40 (2) (2000) 139–157.

[36] K. B. Korb, A. E. Nicholson, Bayesian Artificial Intelligence, Second Edition, 2nd Edition, CRC Press, Inc., Boca Raton, FL, USA, 2010.

[37] S. Acid, L. M. de Campos, J. M. Fernndez-Luna, S. Rodrguez, J. M. Rodrguez, J. L. Salcedo, A comparison of learning algorithms for bayesian networks: a case study based on data from an emergency medical service, Artificial Intelligence in Medicine 30 (3) (2004) 215 – 232, bayesian Networks in Biomedicince and Health-Care. doi:http://dx.doi.org/10.1016/j.artmed.2003.11.002.

[38] S. Bayat, M. Cuggia, D. Rossille, M. Kessler, L. Frimat, Comparison of bayesian network and decision tree methods for predicting access to the renal transplant waiting list, in: Medical Informatics in a United and Healthy Europe - Proceedings of MIE 2009, The XXIInd International Congress of the European Federation for Medical Informatics, Sarajevo, Bosnia and Herzegovina, Agust 30 - September 2, 2009, 2009, pp. 600–604. doi:10.3233/978-1-60750-044-5-600.

[39] A. C. Constantinou, N. Fenton, W. Marsh, L. Radlinski, From complex questionnaire and interviewing data to intelligent bayesian network models for medical decision support, Artif. Intell. Med. 67 (C) (2016) 75–93. doi:10.1016/j.artmed.2016.01.002.

[40] D. Janssens, G. Wets, T. Brijs, K. Vanhoof, T. Arentze, H. Timmermans, Integrating bayesian networks and decision trees in a sequential rule-based transportation model, European Journal of Operational Research 175 (1) (2006) 16 – 34. doi:10.1016/j.ejor.2005.03.022.