1. Introduction

Purely data-driven approaches are currently accepted as the primary, if not the only, way of developing predictive models. Because of the impressive results achieved with such approaches by organizations like Amazon and Google, it is often assumed that this success is repeatable in other domains as long as a large enough amount of data is available. However, a purely data-driven approach can only predict the type of values recorded in a dataset, such as measurements made, decisions taken or outcomes recorded. Even when large volumes of data exist, purely data driven machine learning methods may not provide either accurate predictions or the insights required for improved decision-making. In this paper, we consider the common real-world situation in which successful decision making depends on inferring underlying or latent information that is not – and can never be – part of the data. In such a situation a predictive model for decision support will contain latent variables representing this underlying state but the values of these variables will not be present in the data. We therefore need to depend on domain expertise to identify the important latent variables and to model relations between them and the observed variables.

Domain experts do not just substitute guesswork for data. They may have access to information that is not machine-readable and they should back up any judgements by reference to published research whenever possible. Yet, such expert knowledge is usually avoided in data-driven approaches using arguments such as ‘avoiding subjectivity’ and ‘using facts based on the data’ [1,2]. The use of latent variables is also limited: some data-driven approaches, such as regression modelling, do not include latent variables at all. Other approaches contain latent variables but these are estimated only from data values, so that the use of latent variables in these methods does not escape the limits of the data. The objectivity of data-driven approaches holds only so far as the prediction of observed values really serves the needs of users. When this is not the case, erroneous results may follow. In this paper, we show some examples of these errors and how they are avoided by appropriate and rigorous use of domain knowledge.

We propose a pragmatic methodology to develop Bayesian network (BN) models with latent variables. Our method integrates domain expertise with the available data to develop and refine the model systematically through a series of expert reviews. We illustrate the application and results of this method with a clinical case study of a problem for which purely data-driven approaches have been tried but have not been considered to be successful by clinicians. Our case study shows some possible reasons for these past failures. It is beyond of the scope of the paper to provide full details of the developed model, but the details can be found at the ATC BN website [3].
The remainder of this paper is as follows: Section 2 presents the overview of our methodology. The case study is introduced in Section 3 and developed further in Sections 4 (learning and review) and 5 (model refinement). We present our conclusion in Section 6.

2. Method overview

The limitations of data for making predictions useful to a decision-maker can be summarised in three points:

1. **Measurement errors**: a dataset contains measurements of variables, but measurement errors mean that the true state of each variable differs from the measured data. In some domains, including clinical diagnosis, this introduces significant uncertainty about the true value, so that a data-driven model cannot accurately predict the underlying state even if it can accurately predict the associated measurement values.

2. **Sub-optimal decisions**: the objective of a decision-support model is to enable a decision-maker to determine the optimal decisions given the observed situation. A dataset may contain a ‘decision’ variable, that is, one that reflects the decision made (e.g. a treatment given by a clinician). A model that predicts the value of a decision variable can be useful if all the past decision-makers had similar utilities and they were completely rational in evaluating utilities with their underlying uncertainties. However, there is usually no information about the utilities involved in past decisions, and the data may have records of some decisions that were incorrect at the time or, although correct at the time, were made on outdated understanding. A model that predicts the value of a decision variable is therefore limited in its performance even if the prediction is highly accurate. Moreover, a model can only be used for ‘what if’ analysis – exploring the consequences of decision alternatives – if it is causal; choosing one of the decision alternatives erases the factors that influenced past decisions [4]. Although these problems are well known, models that are developed to fit past decisions are common in scientific literature (see Section 3.1).

3. **Causes of outcomes**: an ‘outcome’ variable records what happened. But outcomes can have many causes, only some of which may be recorded in the dataset (for example, in medical applications not all interventions and treatments are recorded). A prediction based on only some causes may be useful – the missing causes simply add uncertainty – but understanding of the scope of the causes included is important to the correct application of the model. A purely data-driven approach does not resolve this problem; only an expert can detect if the data omits known causes of the outcome. If omitted causes can be identified, this information can be used either to improve the model or to clarify its scope and to assess its performance within the scope of the causes modelled.

The main aim of our method (illustrated in the flow diagram in Fig. 1) is to overcome these limitations. We show how to develop BNs that predict and reason with latent variables using a training dataset including measurements of these variables, but not including their true state. Domain expertise is used both at the start of the development to discover latent variables and then later to refine the model in a series of expert reviews; it is during these reviews that discrepancies between knowledge and data are revealed. Expert knowledge can be used in various degrees when deriving the structure of a BN [5]. In our method, the structure of the BN is developed with domain experts by using small BN fragments for commonly occurring reasoning types as building-blocks to form the complete BN structure [6]. The advantage of experts deriving the model’s structure, rather than learning it from data, is to ensure causal coherence: latent variables influence measurements and decision variables influence outcomes. Hybrid approaches that combine expert knowledge and data can also be used at this stage for deriving the BN structure [7,8]. Moreover, structure learning methods can be used as a complementary approach to evaluate and refine a BN structure developed by experts [9]. Of course, all causal assumptions need to be supported by the best available evidence, such as experimental results or expert consensus. Lack of knowledge of true causal relationship is a problem and affects both expert and data-led modelling (aside from the limited capabilities of algorithms such as inductive causation [IC] [10]) alike. Equally, not all causal relationships are uncertain: it is clear that an object’s temperature causes the thermometer reading rather than the other way around.

The next step is to label the latent variables in the training dataset, overcoming the problem that their values are unknown. The first label is derived from measurement data using deterministic (but not necessarily complete) rules defined by domain experts; the second uses data clustering. The experts’ rules can be of any form, but are typically derived from current practice. For example, if the related measurements are continuous, these rules are threshold values for the measurements. For clustering, we use the standard Expectation–Maximisation (EM) for BNs with known structure [11]. EM is an iterative algorithm that is used for learning the parameters of a BN from a dataset with missing values. Each iteration of EM has two steps: the E-step completes the data by calculating the expected values of unobservable variables based on the current set of parameters; the M-step learns a new set of parameters from the maximum likelihood estimate of this completed data. When EM is used for parameter learning, the M-step of the final iteration calculates the BN parameters. When it is used for clustering, the unobserved variables are labelled according to the values in the E-step of the final iteration. In our method, all of the values of the unobserved variable are missing from the dataset and we are using EM for clustering the unobserved values. Although EM can also be used for structure learning [12,13] this is not required in our method as the BN structure is developed with domain experts. Extensions of EM that builds upon the information bottleneck [14], variational Bayesian [15] and hierarchical [16] frameworks have been proposed for learning latent variables. Van der Gaag et al. [17] presents a similar approach to labelling with expert rules where they represent combinations of multiple observations with latent variables.

We now have two labels for each latent variable: one from clinical measurements and the experts’ rules, the other from EM clustering. A final label is achieved by combining the two labels in cases where the labels are the same and by expert review of cases where there is a difference between the two labelling methods. We prepare a list of cases where the labels differ. Domain experts then decide the final label for each data record in this list. The experts can review other data including information that is not machine-readable and cite relevant research to support this decision. We also include a random subset of cases that were labelled consistently in the review to assess the experts’ consistency with the labelling by measurements and clustering approaches. This combination of expert review and data has a number of advantages. It allows for the possibility of errors in measurement, and it uses the experts efficiently. Expert review is a costly resource and using it for every single case in the data is usually not feasible, especially if the dataset is large. Therefore, our method aims to use it only for ambiguous cases, where the labels from measurements conflict with the results of the clustering on our data.
After the expert review, we use the original dataset to which the latent variable labels have been added, to learn the BN’s parameters and to evaluate its performance. We again use the EM algorithm but this time to learn the parameters, since the dataset may still contain missing values of other variables for some patients. The second use of the EM algorithm in this step should not be confused with the previous use of the same EM algorithm to label latent variables in the step 2A (see Fig. 1). Expert constraints [18], in the form of parameter orders, can also be used if the available data is too small for learning a part of the parameters. We use $k$-fold cross-validation to evaluate the performance of the BN. In this approach, the data is divided into $k$ equal sized groups. One of the $k$ groups is used as test data while the remaining $k-1$ groups are used for training the BN. The learning and testing continues iteratively until the model is validated with all of the $k$ groups.

The inaccurate predictions of a predictive model offer useful lessons for improving the model and are the focus of the next stage of review. The BN modelling approach is well-suited for this kind of review since it concretely represents separate medical pathways leading to its predictions [19,20]. When the BN model’s prediction in the cross-validation differs from the value recorded in the data, the domain experts investigate the reasons for this difference to look for potential improvements to the model or clarify its scope. The domain experts look at cases where the recorded values are what is expected in their experience even though it is different to what was predicted by the model. In some cases, the domain experts may agree with the prediction of the model, and they

![Fig. 1. Method for learning BN with latent variables.](image-url)
may consider the value recorded in the data as an error or simply as an unexpected outcome. For example, in a medical decision-support scenario, survival of a patient with a severe injury burden and high blood loss might be considered to be an unexpected outcome. The expert review can also clarify scope of the model: if the recorded outcome is explained by factors that have been excluded from the model then this should be made clear to the model’s users. For example, the experts might note that patient who unexpectedly survived did so as a result of a particular pre-hospital treatment, and the model could not identify this as pre-hospital interventions were out of the scope of the model. Alternatively, additional latent variables and relations that are important for the predictions can be discovered and added to the model. Since the BN’s structure represents domain knowledge, any modifications must be supported by evidence.

Differences between the available data and the target subpopulation of the BN must be examined as the knowledge from these two sources is combined in our method. Correlations, caused by these differences, must be analysed and modelled in the BN structure in order to avoid developing erroneous models [21].

3. Case study: trauma care

We illustrate our method with a clinical case study on acute traumatic coagulopathy (ATC), a pathophysiological disorder affecting the body’s ability to form a stable blood clot that occurs following traumatic injuries and worsens outcome. The case study was done in collaboration with Royal London Hospital (RLH), a world leader in trauma care. A group of domain experts, including the second and fourth authors, and a large hospital dataset were used to develop the model with our methodology.

3.1. Data-driven models in trauma care

A common approach used for data-driven predictive models, especially in medicine, is multivariate logistic regression. These models are functions from several predictor variables and a single outcome variable. The regression function is estimated from a dataset of these variables. Variables that are considered to be related with the outcome should be selected as candidate predictors. Afterwards, a series of statistical tests are often applied to determine the least predictive candidates and remove them from the model [22]. If the aim of the model is to make predictions, the number and identity of the included variables are not considered to be important [23].

Several data-driven prediction models have been developed for decision support in trauma care but with little impact in clinical practice due to the limitations discussed in Section 2:

1. Measurement errors: in the previous models to predict ATC, the presence of ATC is identified with a threshold value on a blood test called international normalised ratio (INR) [24] despite the fact that this test has known limitations at identifying this condition. This approach has been criticised as it fails to produce useful clinical results [25].

2. Sub-optimal decisions: models that predict the decision values in data have been developed in other areas of trauma care. One example is decision support for injured extremities which encompasses knowledge of the presence of ATC. Several data-driven models have been developed to predict amputation decisions in this domain [26–28]. Although some of these models have been used as research or evaluation tools, none of them have been recommended as a decision support tool in clinical practice [28]. The output of these models shows the percentage of clinicians that made amputation decision in similar circumstances. However, recommending an amputation without relating it with patient outcomes makes it difficult to assess the model or to understand its reasoning. Moreover, recent advances in trauma care may have made some of the decisions in the training data inappropriate for current use. A more useful prediction for the decision-maker would be to compare the function expected from a salvaged versus an amputated extremity, given the characteristics of the injury factors.

3. Causes of outcomes: permanent nerve dysfunction may be an indication for amputation since an extremity cannot function without nerve supply. The state of the nerve function during hospital care is available in many clinical datasets, and this variable has been considered as an important factor for amputation in data-driven models [26,27]. However, nerve dysfunction has several causes, some of which can recover [29]. Since data on nerve recovery and causes of nerve dysfunction were not available, these factors were ignored in the data-driven models. Considering the irreversible outcomes of amputation decisions, all relevant factors should be examined.

3.2. Acute traumatic coagulopathy

Up to a quarter of trauma patients develop an acute traumatic coagulopathy (ATC) soon after their injury. These patients have a considerably higher risk of bleeding and death since the body’s protective mechanisms to limit bleeding are deranged. Several effective treatment options are available if ATC can be identified early. Immediate treatment is most effective however; standard laboratory tests to identify ATC take over an hour to produce results. The primary aim of the BN model is therefore to predict ATC with the information normally available within the first 10 min of care. The methodology we have described is relevant to this problem: the values of both ATC and of its causal factors are measured but none of the measurements are perfectly accurate.

3.3. ATC bayesian network structure

The initial structure of the BN, shown in Fig. 2, was developed with domain experts using the AgenaRisk software [30]. The BN structure contains two latent variables: ATC and Hypoperfusion. In addition, several other variables are available in the training dataset but are usually unobserved in the first 10 min of treatment when the model is designed to be used. Each of these unobserved and latent variables is modelled with their measurements as naive BN fragments or ‘measurement idioms’ [6]. These naïve BN fragments were used as building blocks to form the BN structure, connected using causal relations elicited from experts. Table 1 shows the variables modelled with measurement idioms. A detailed description of the model can be found in the ATC BN website [3].

The model is divided into four components, corresponding to the four boxes shown in Fig. 2. The remainder of this section explains the variables and relations in each of these components briefly:

- Coagulopathy: The ATC variable has two states: ‘Present’ and ‘Absent’, and it can be estimated from 5 measurements with continuous values. None of these measurements are available within the first 10 min but the variables are useful for model development. The main drivers of ATC are the degree of tissue injury and hypoperfusion. This may be aggravated by the infusion of large volumes of fluid (PreHosp).
Shock: The hypoperfusion variable represents inadequate oxygen delivery to tissues as a result of blood loss, and it has three states: ‘None’, ‘Compensated’ and ‘Uncompensated’.

Injury: The degree of overall tissue injury may not be known at the early stages of care. Overall tissue injury is estimated from the mechanism and energy of injury, and the number of severely injured body regions in the BN.

Death: The model predicts death caused by physiological derangements. Age is an established independent predictor of death and has important effects on the physiological response to injury.

A dataset of 600 trauma patients from RLH was used to learn the parameters of this BN. In addition to latent variables ‘ATC’ and ‘Hypoperfusion’, 3% of the values are missing in the RLH data, mainly due to recording errors or missing laboratory tests. Inference on continuous variables was calculated by the dynamic discretisation algorithm in AgenaRisk [30].

### 3.4. Issues with ATC measurements

The true state of ATC, which is the main outcome of our model and a crucial factor in trauma care, cannot be directly observed in practice, even after all the laboratory measurements have been completed. The ATC state is estimated using laboratory measurements such as the clotting time of a blood sample. However, none of these measurements can estimate the underlying ATC state with complete certainty. One measurement is the INR which is the normalised ratio of the clotting time of a patient’s blood plasma to the clotting time of a healthy person. INR, and its clinically interchangeable measure prothrombin ratio (PTR), are the clinically accepted standard for diagnosing ATC [31]. A normal INR value is 1, meaning that a patient has the same clotting time as a healthy person, and higher INR values indicate coagulation problems. However, there is not a clear borderline to distinguish normal coagulation from coagulopathy:

1. INR only tests blood plasma, disregarding other components essential to clotting such as the contribution made by platelets and the blood vessel wall.
2. INR does not measure the strength of a formed clot, the primary abnormality in ATC. It only measures the time it takes to form a clot.
3. INR is designed to monitor the effects of the drug Warfarin; it is not specifically designed for trauma.
Developing and validating a model that predicts INR values is convenient, but predicting INR is quite different from predicting the underlying coagulopathy state. For example, Mitra et al. [24] used an INR of 1.5 as a threshold value for classifying ATC. However, a patient with an INR of 1.3 may have serious coagulation problems.

Consequently, the true underlying coagulopathic state of some patients cannot be known with certainty until a completely accurate way of measuring coagulopathy is discovered. Until then, clinicians will continue to estimate coagulopathy using their clinical judgement together with available measurements and observations. These clinical judgements are not recorded in the hospital database. Only the data about INR and similar measurements are recorded in the dataset. The situation is similar for ‘Hypoperfusion’ which is the other latent variable in our model.

4. Learning and review

4.1. Initial labelling by expert thresholds and clustering

The latent variables were labelled twice using two different methods: first using measurement thresholds that reflect current clinical understanding [31–33], and then by clustering using the EM algorithm (step 2A and 2B of Fig. 1). The thresholds used for labelling the ATC and Hypoperfusion variables are shown in Table 2. As a result of missing data, a number of patients could not be labelled. The labelling criteria for Hypoperfusion (see Table 2) are not complete so this state could not be labelled for several patients. Clustering was performed using the EM algorithm on the BN structure shown in Fig. 2. EM uses all of the observed variables and the BN structure to classify the data into coherent groups based on the maximum likelihood estimate of the latent variables. We used EM to classify the data into two coherent ATC states and three coherent hypoperfusion states.

4.2. Expert review of the labelling differences

We compared the labels given by the measurement threshold and clustering approaches and prepared a list of the patients with differing labels, no label and a random subset of other cases. Three domain experts independently reviewed these cases and provided an expert label. All clinical information was available to the experts to assist labelling. The experts were blind to the labels assigned by the measurement threshold and EM clustering methods. The consensus between the experts' labels was assigned as the final label. Table 3 shows the number of cases reviewed for the two latent variables.

This method required the domain experts to review 188 (31%) and 54 (9%) of the 600 cases respectively to label the hypoperfusion and ATC categories. Tables 4 and 5 show the number of measurement threshold labels changed after the review: for example Table 4 shows that 6 patients classified as coagulopathic on the basis of the INR threshold were re-classified to non-coagulopathic by the expert review. At the end of this step, each latent variable had a single set of labels that were obtained from the combination of measurement threshold and clustering approaches, and the expert review of the differing labels.

4.3. Learning and cross-validation

The result of the expert review (step 3 of Fig. 1) is a dataset now including values for the latent variables for almost all patients. The ATC value of 4 patients and Hypoperfusion value of 5 patients remained unlabelled after the expert review because the expert was not confident about the correct value. We used the standard EM algorithm to learn the parameters of the model. The performance of the model trained on the RLH data was tested by 10-fold cross validation. Only the variables that can be observed in the first 10 min of treatment are instantiated for generating the predictions in 10-fold cross validation.

Performance of a model can be measured in terms of its discrimination, calibration and accuracy. Discrimination measures whether the model can distinguish the patients with the event. A model that has well discriminatory performance gives higher probabilities to the patients with the event, and lower probabilities to the patients without the event. Calibration measures whether the predicted probability represents the correct probability on average. For example, when a model predicts 10% chance of survival for a group of patients, 10% of these patients are expected to survive if the model is well calibrated. Accuracy measures whether the predicted outcomes are close to the actual outcomes by combining features of discrimination and calibration. Medlock et al. [34] recommends using multiple performance measures to quantify different aspects of model performance.

We used multiple performance measures to assess the discrimination, accuracy and calibration of the ATC BN as recommended by the Medlock framework [34]. The discrimination of the ATC BN was evaluated with receiver operating characteristic (ROC) curves, sensitivity and specificity values. The area under the ROC curve (AUROC) is 0.90 (95% confidence interval (CI); 0.86–0.94) and 0.81 (95% CI: 0.75–0.86) for the prediction of ATC and death respectively. Brohi [25] argues that a useful prediction model for coagulopathy must operate with at least 90% sensitivity: the BN achieves specificities of 71% (95% CI: 0.67–0.74) for ATC and 44% (95% CI: 0.39–0.48) for death when operating with 90% sensitivity. The initial performance of the model on the cross-validation dataset can be seen in Table 6.

The accuracy of the model was evaluated with the Brier score (BS) and Brier skill score (BSS) [36,37]. BS is the mean squared

---

Table 2
Criteria for labelling ATC and hypoperfusion from measurements in data.

<table>
<thead>
<tr>
<th>ATC</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>INR ≤ 1.2</td>
<td>No</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hypoperfusion</th>
<th>Compensated</th>
<th>Uncompensated</th>
</tr>
</thead>
<tbody>
<tr>
<td>BE &gt; –2 &amp; Lactate ≤ 2 &amp; SI &lt; 0.9</td>
<td>SBP &lt; 90 &amp; BT &gt; 4</td>
<td>BE &lt; –4 &amp; Lactate &gt; 4 &amp; BT &gt; 0</td>
</tr>
<tr>
<td>Alive &amp; BT in 12 h = 0</td>
<td>–4 ≤ BE &lt; –2 &amp; Lactate ≤ 4 &amp; BT &gt; 0</td>
<td>Pre-hospital cardiac arrest</td>
</tr>
<tr>
<td>Died in ≥ 48 h &amp; BT = 0</td>
<td></td>
<td>Death from haemorrhage</td>
</tr>
</tbody>
</table>

BT = Blood Transfusion in 12 h, SI = Shock Index, BE = Base excess, SBP = Systolic blood pressure.

---

1 Please see Fenton and Neil [35] Chapters 1 and 10 for a discussion of the limitations of confidence intervals and test of significance in this context.
The score can take values between 0 and 1; 0 indicates a perfect model and 1 is the worst score achievable. BSS measures the improvement of the model’s prediction relative to a reference prediction which is often the average probability of the event in the data. BSS can take values between negative infinity and 1; a negative value indicates a worse prediction than the average probability and 1 indicates a perfect model. The BN has BS of 0.06 (95% CI: 0.05–0.07) and BSS of 0.32 (95% CI: 0.21–0.43) for ATC predictions, BS of 0.09 (95% CI: 0.07–0.11) and BSS of 0.15 (95% CI: 0.04–0.26) for death predictions.

The calibration of the BN was assessed with the Hosmer–Lemeshow test\[38\]. This test divides the data into 10 subgroups, and calculates a chi-square statistic comparing the observed outcomes to the outcomes expected by the model in each subgroup. Low \(p\)-values indicate a lack of calibration. Hosmer–Lemeshow test is strongly influenced by the sample size. In large datasets, small differences between the expected and observed outcomes can lead to low \(p\)-values but the visual representation of this test provides a concise summary of the model calibration.

The BN was well calibrated for both ATC and death predictions with Hosmer–Lemeshow statistics of 9.7 (\(p = 0.29\)) and 6.7 (\(p = 0.57\)) respectively. Fig. 3 is a visual representation of the model’s calibration for ATC predictions. The similarity between the expected and true outcomes in each subgroup shows that the model was well calibrated.

### 4.4. Inaccurate predictions and unexpected clinical outcomes

After the learning and cross-validation steps, we reviewed the inaccurate predictions of the model with the domain experts (step 5 of Fig. 1). We divided the predictions, given by cross validation of the model, into ten bins according to the predicted probability, and prepared a contingency table that compares the predictions of the model to the outcome values in data for each bin as shown in Table 7.

The negative outcomes with ATC prediction over 0.1, and the positive outcomes with ATC prediction less than 0.1 (shown in bold in Table 7) were considered as the possibly inaccurate predictions since 10% of the patients were initially labelled with ATC and thus 0.1 was our prior probability. A clinician reviewed the data and
Inaccurate predictions and expert review.

causes of each unexpected prediction:

1. ATC may develop in some patients soon after the blood test used for INR and other measurements was taken.
2. Some of the deaths recorded in the dataset were most likely due to conditions other than ATC.
3. There are mechanisms of coagulopathy other than ATC that may be occurring in patients in the dataset.

Table 8 gives a summary of this review: the domain experts agreed with about a third of the apparently inaccurate predictions. During the review, domain experts explained why they agreed with the individual predictions or recorded outcomes which led to a number of refinements to the model and to the clinicians’ understanding of the data. Death predictions were also reviewed by the same approach. We describe these issues and the way the model was refined in the following section.

5. Model refinement

Three issues were found from the review of inaccurate predictions:

1. ATC may develop in some patients soon after the blood test used for INR and other measurements was taken.
2. Some of the deaths recorded in the dataset were most likely due to conditions other than ATC.
3. There are mechanisms of coagulopathy other than ATC that may be occurring in patients in the dataset.

These issues all challenge the supposed objectivity of data and reinforce the need to combine data with expert review. The following sections describe these issues in more detail.

5.1. Incipient coagulopathy

A group of patients who had normal values for their initial ATC measurements (see Table 1) showed significant signs of ATC in a second set of measurements that were conducted soon after. Moreover, these patients had severe injury burden and poor perfusion; they were therefore at high clinical risk of developing ATC. The ATC model predicts high risk of coagulopathy for these patients but the value in the data is negative since only the initial measurements were considered while labelling the ATC state of patients with measurement thresholds and clustering approaches.

Coagulopathy is a dynamic phenomenon that develops in time, so the results of measurements are dependent on the time they are carried out. Variations in the interval between the injury and the arrival at the hospital add further uncertainty. Therefore, the domain experts considered the prediction of those patients with ‘incipient coagulopathy’ as a clinically useful feature of the BN.

We re-learnt the ATC BN and recalculated its performance in a cross validation when patients with incipient ATC were also considered as positive outcomes. The structure of the ATC BN was not changed in this analysis. Fig. 4 compares the ROC curves for ATC prediction based on only the initial measurements with the one for patients with incipient ATC. The AUROC is 0.92 (95% CI: 0.89–0.95), and the model achieves specificity of 79% (95% CI: 0.76–0.82) at the sensitivity 90% level, BS of 0.06 (95% CI: 0.05–0.07) and BSS of 0.39 (95% CI: 0.27–0.50).

Prediction of incipient coagulopathy shows the difference between the clinically useful models and the models that predict measurements in data well. The patients with incipient coagulopathy would count as incorrect predictions for a purely data-based approach, and such an approach would try to change the parameters to ‘correct’ these predictions. In contrast, the expert was able to explain the apparent anomaly and show that predicting incipient coagulopathy was useful; this was not obvious at the beginning of the model development.

5.2. Other causes of death

The review revealed that a large proportion of deaths that could not be predicted by the BN were the result of head injuries and thus these deaths were expected by the domain experts. The ATC model is designed to predict risk of death relevant to bleeding and coagulopathy, so the initial model does not predict deaths related to head injuries. However, the model structure is easily modified to predict these deaths since we already have a head injury variable in the model which is used to estimate the overall tissue injury burden of patients. By adding an arc between head injury and death variable, we increase the accuracy of the model for death prediction. Although death might be considered to be the least ambiguous outcome in a clinical dataset, our experience
shows that this is not the case when there is a mismatch between the modelled and actual cause of death.

This simple modification increased the accuracy of death predictions significantly. The AUROC increased from 0.81 (95% CI: 0.75–0.86) to 0.88 (95% CI: 0.84–0.92) as shown in Fig. 5. The specificity of the BN is increased from 44% (95% CI: 0.39–0.48) to 72% (95% CI: 0.67–0.76) when it is operated at 90% sensitivity level. BS and BSS also indicated an increased accuracy in the death predictions: BS of the BN decreased from 0.09 (95% CI: 0.07–0.11) to 0.08 (95% CI: 0.06–0.10); and BSS increased from 0.15 (95% CI: 0.04–0.26) to 0.23 (95% CI: 0.13–0.33). This change had no impact on ATC prediction.

5.3. Unmodelled mechanisms of coagulopathy

The aim of this BN is to predict ATC, which is driven by a combination of the degree of tissue injury and hypoperfusion following traumatic injuries. The scope of the model has to be clearly defined since other forms of coagulopathy exist. For example, the anticoagulation medicine Warfarin makes a person coagulopathic without any traumatic injury, and predicting drug induced coagulopathy is out of scope for this model.

Another important cause of traumatic coagulopathy is a catastrophic brain injury. These injuries seem to effect coagulation via a different mechanism to ATC. The review of unexpected predictions showed that 9 of the 12 coagulopathic patients that the BN model could not accurately predict had severe head injuries, and in 7 of these patients brain injury was fatal. It is likely that these patients were suffering from a coagulopathy caused by brain injury (BIC) rather than ATC.

BIC is now documented as being outside the scope of our BN. This issue was not clear at the beginning of the model development even though the clinicians were aware of the phenomenon; it was identified as a result of the review of inaccurate predictions with the domain expert. If prediction of the BIC is required by the users, the model structure can be adapted accordingly by adding two variables ‘BIC’ and ‘Coagulopathy’ as shown in Fig. 6. In this model fragment, ‘Coagulopathy’ variable represents the overall coagulopathy risk that sums the risk of ‘BIC’ and ‘ATC’ variables.

6. Conclusion

This paper proposed a method for developing and refining BNs with latent clinical conditions, using a combination of expert knowledge and data. The method is successfully applied to a clinical case study about the prediction of ATC in trauma care. Our method addresses the problems related to measurement errors and causes of outcomes by:
1. Making a clear distinction between a latent variable that we wish to predict for decision support and any measurements of this variable that may be recorded in a dataset; both latent and observed variables are represented explicitly in the BN model.

2. Using iterative expert review of the model to refine the model and to understand the relationship between the data and the real decision problem.

Our methodology systematically integrated domain expertise into model development at two stages. Firstly, the ‘true’ but unobserved state was added to a dataset by combining labelling by observed measurements with data clustering in an expert-elicited BN structure. Focussing the detailed expert review on the cases labelled differently in these two steps saves time compared to a review of all cases. Secondly, the experts examined differences between the model’s predictions and the data.

In our case study, this examination revealed several issues initially neglected by our experts and emphasised the difference between useful predictions for the decision-maker and an accurate prediction of measurements in data. Other latent and observed causes of predicted outcomes, which were not clear at the beginning, were modelled during the review. These issues were resolved either by refining the model or by acknowledging the scope of its applicability, which were not obvious at the initial stages of model development.

The case study demonstrates significant improvements in predictions from the iterative expert reviews and refinements. Identifying and including the other causes of death increased the specificity for death predictions from 44% (95% CI: 0.39–0.48) to 72% (95% CI: 0.67–0.76) when the model is operated at 90% sensitivity. Similarly, identifying the clinically important patient subgroup with incipient coagulopathy increased the specificity for ATC predictions from 71% (95% CI: 0.67–0.74) to 79% (95% CI: 0.76–0.82) at 90% sensitivity.

As further research, we are examining practical approaches that use domain expertise to resolve problems related to predicting decisions. In addition, the dialogue between the expert and the model has so far focussed exclusively on the global performance of the model. Since the model’s structure corresponds to the expert’s understanding of the domain, we are investigating how to compare the internal operation of the model with the clinical reasoning applied to a particular patient. Moreover, a graphical interface that assists the expert in reviewing of inaccurate predictions would also be useful.

References