Appendix 1 Tree Augmented Naïve Bayesian Network (TAN)

Classification refers to a task that assigns corresponding labels to outcomes. Supervised classification finds the suitable functions or rules to achieve the known outcomes using the evidence. Unsupervised classification clusters object into groups or categories when the outcomes are unknown. Let Y = {0,1} or {–1,1} be the binary indicating the outcome of the diagnosis, and n random variables **X**= {X1, X2, ..., Xn} represent the information needed to diagnose the patient's condition, known as variables or features. Applying the Bayes rules, we can derive the conditional probability of an outcome given attributes and the supervised classification problem is then described as:

$$argmax_Y\left(P\left(Y\left|\boldsymbol{X}\right.\right)\right)=argmax_Y\left(\frac{P\left(Y\right)P\left(\boldsymbol{X}\left|Y\right.\right)}{P\left(\boldsymbol{X}\right)}\right) \quad [1]$$

There is no independence assumption among variables in the general Bayesian classifier, leading to a complete network with complex and costly calculations. This type of network is only suitable for problems with few variables. Depending on the processing power, this number of variables can be increased. The more variables, the more it will tax the training time and resources. Here, we review some alternatives of Bayesian classifiers, including Naive Bayes, Semi-Naive Bayesian, and Tree Augmented Naive Bayes. The main difference between them is the assumption of conditional independence among the variables.

If the general Bayesian classifier considers no independence and the Naive Bayesian considers all independence, TAN is the model in between. It allows some dependence between variables, allowing information to flow better into the graph while maintaining a simple network structure and low computational costs. This dependence is measured using mutual information, a pairwise score based on information theory:

$$MI\left(X_i,X_j\left|Y\right.\right)=\sum_{k=1}^{N_{Xi}}\sum_{h=1}^{N_{Xj}}\sum_{l=1}^{N_Y}P\left(k,h,l\right)log\frac{P\left(k,h\left|l\right.\right)}{P\left(k\left|l\right.\right)P\left(h\left|l\right.\right)} \quad [2]$$

Tree Augmented Naïve Bayes construction algorithm

TAN, compared to other BN classifiers, is superior in capturing the associations of variables by utilizing the mutual information (MI) between them (21). This MI is calculated by the following equation:

$$MI\left(X_i,X_j\left|Y\right.\right)=\sum_{k=1}^{N_{Xi}}\sum_{h=1}^{N_{Xj}}\sum_{l=1}^{N_Y}P\left(k,h,l\right)log\frac{P\left(k,h\left|l\right.\right)}{P\left(k\left|l\right.\right)P\left(h\left|l\right.\right)} \quad [3]$$

where X_i, X_j are the pair of variables (e.g., gender, age), Y is sepsis outcome, N_{Xi}, N_{Xj}, N_Y are the number of values in X_i, X_j, and Y.

TAN network for sepsis diagnosis is built using the following steps (21):

Step 1: Build the complete undirected graphs with the variables. Assign the MI of each pair of variables as the weight of the respective edge between them.

Step 2: Find the spanning tree that has the maximum weight.

Step 3: Transform the maximum weighted spanning tree in step 2 to a directed graph with the arbitrary or chosen root node.

Step 4: Attach the sepsis node to the network with the arrows coming from sepsis node to other nodes.

https://dx.doi.org/10.21037/tp-22-510

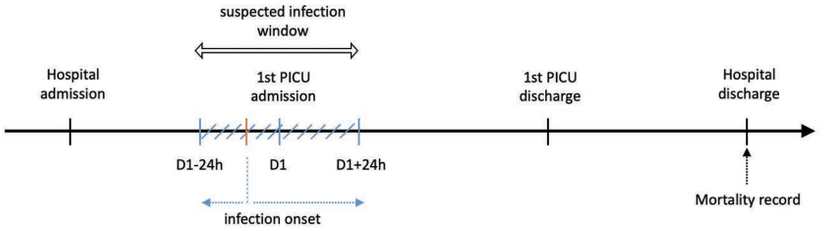

Figure S1 Definition of patients with suspected infection. D1 is the first day of the first PICU admission of hospital admission. Suspected infection was defined as the combination of culture drawn and antibiotic administration within a specific time interval (15). The suspected infection window is between 24 hours prior to and 24 hours after PICU admission. The onset of infection was taken to be either the time of cultures drawn or the time of antibiotic administration, whichever occurred first. Mortality was recorded at hospital discharge within the given hospitalization. PICU, pediatric intensive care unit.

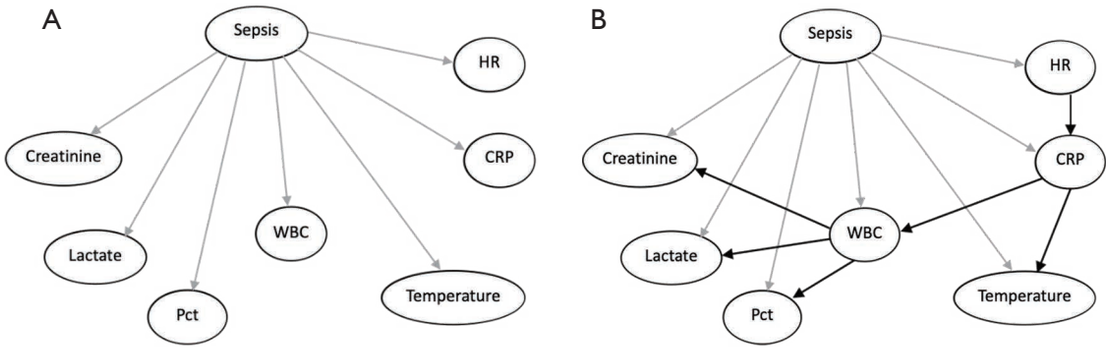

Figure S2 Example of Naïve Bayesian Network (NB) and Tree Augmented Naïve Bayes Network (TAN) to diagnose sepsis. The outcome of interest is the sepsis node. The prediction performed on the sepsis node is calculated based on the given state of seven other variables in the graph. NB classifier assumes that all the variables are independent, which offers a simple graphical representation and calculation (A). However, it may not be practical in some real-world scenarios. TAN is an extension of NB that allows additional dependency of one variable to one another, which enables the information to flow more naturally in the graph and improves performance (B). CRP, C-reactive protein; HR, heart rate; Pct, procalcitonin; WBC, white blood cell count.

Table S1 List of vasoactive agents for shock

Vasoactive agents
Adrenaline
Amrinone
Dobutamine
Dopamine
Dopexamine
Ephedrine
Epinephrine
Erlipressin
Isoprenaline
Levosimendan
Metaraminol
Milrinone
Noradrenaline
Norepinephrine
Phenylephrine
Vasopressin

https://dx.doi.org/10.21037/tp-22-510