A Causal Modeling Framework for Generating Clinical Practice Guidelines from Data

Subramani Mani and Constantin Aliferis

Vanderbilt University, Nashville TN 37232, USA,
{subramani.mani, constantin.aliferis}@vanderbilt.edu

Abstract. The practice of medicine is becoming increasingly evidence-based and clinical practice guidelines (CPGs) are necessary for advancing evidence-based medicine (EBM). We hypothesize that machine learning methods can play an important role in learning CPGs automatically from data . Automatically induced CPGs can then be used for further manual refinement and deployment, for automated guideline compliance checking, for better understanding of disease processes, and for improved physician education. We discuss why learning CPGs is a special form of computational causal discovery and why simply predictive (i.e., non-causal) methods may not be appropriate for this task.

1 Introduction and background

Clinical practice guidelines (CPGs) can be broadly classified as predictive guidelines or prevention/intervention guidelines based on their goals. While predictive guidelines may be sufficient for diagnosis or assessing prognosis, we need a cause and effect interpretability for prevention and intervention. In this paper we develop a framework using the representation of causal Bayesian networks (CBNs) for automatic generation of guidelines from data with application to prevention and treatment of disease. The generated guidelines can be evaluated by experts, tested and improved before they are adopted by professional societies or hospital managements for deployment. We discuss why learning CPGs is a special form of computational causal discovery and why simply predictive (i.e., non-causal) methods may not be appropriate for this task.

A causal Bayesian network is a Bayesian network in which each arc is interpreted as a direct causal influence between a parent node (variable) and a child node, relative to the other nodes in the network [1]. We define the *causal influence* of a variable A on variable B using the *manipulation criterion* [2, 3]. The manipulation criterion states that if we had a way of setting just the values of A and then measuring B, the causal influence of A on B will be reflected as a change in the conditional distribution of B. That is, there exist values a_1 and a_2 of A such that $P(B|\text{ set } A = a_1) \neq P(B|\text{ set } A = a_2)$. For a philosophical approach to causality in the health sciences see for example [4]. The two basic assumptions that are necessary for our causal discovery framework are the causal Markov condition (CMC) and the causal faithfulness condition (CFC) [2, 3].

The rest of the paper is organized as follows. In Section 2 we discuss the categories of causal and predictive guidelines. In Section 3 we develop the framework for learning clinical practice guidelines using CBNs and point to some preliminary results. In Section 4 we highlight the strengths of the causal modeling framework for guideline induction, point out a limitation and give direction for future research.

2 Causal versus Predictive Guidelines

Consider the hypothetical CBN structure for the Malaria domain shown in Figure 2. Assume we have a dataset $\mathcal{D}_1$ that is faithful to the structure in Figure 1. A plausible predictive model for *Malaria* using for example, a decision tree or a rule learner is given below:

If *Intermittent Fever* and *Enlarged Spleen* Then *Malaria* Else Normal.

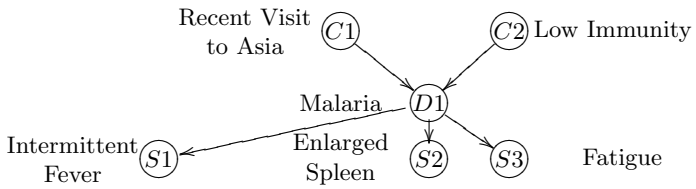

Fig. 1. A hypothetical causal Bayesian network structure for the Malaria domain

Let us call this rule the Fever Spleen (FS) rule. The FS rule would be a good predictor of Malaria. However, it is clear that guidelines to treat fever with aspirin or paracetamol will not have any effect on the distribution of Malaria.

On the other hand consider a causal relationship such as the following that may be induced from $\mathcal{D}_1$ by a causal discovery algorithm:

Recent Visit to Africa causally influences *Malaria*.

Based on this cause and effect relationship, we could propose a travel advisory warning people against traveling to Africa resulting in a reduction in Malaria in the population. While causal discovery methods focus on causal factors for proposing guidelines, predictive machine learning algorithms are known to select relevant but non-causal variables based on predictive accuracy. Moreover, the selected relevant variables may not be in the local neighborhood (direct predictors) of the class variable [5].

3 Framework for Clinical Practice Guidelines

In this section we describe a causal modeling framework using the representation of CBNs for three CPG specific tasks.

3.1 Learn the causal model of the domain from data and propose new CPGs

Using the model shown in Figure 1 (assuming it is learned from $\mathcal{D}_1$ using a causal discovery algorithm), we could propose the following Malaria CPG:

> If *Recent Visit to Africa* and *Intermittent Fever* and *Enlarged Spleen*
> Then diagnose *Malaria*.

3.2 Learn the causal model from practice data and recognize the likely CPGs being followed

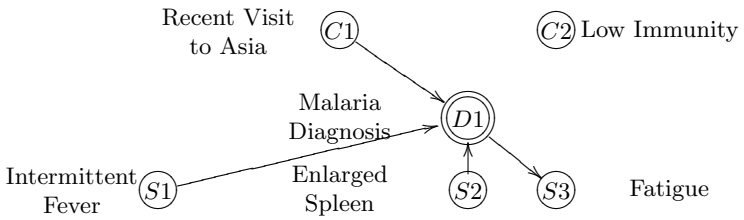

Fig. 2. A causal Bayesian network structure for Malaria diagnosis using guideline

Assume we also have data (dataset $\mathcal{D}_2$) from a hospital that is following the Malaria CPG given in Section 3.1. A plausible causal model for $\mathcal{D}_2$ is shown in Figure 2. The variable $D1$ represents *Malaria Diagnosis* which is a deterministic node and functionally dependent on the three variables $C1$, $S1$ and $S2$. Note that the arc from $C2$ to $D1$ has been removed because of the functional dependency of $D1$ on $C1$, $S1$ and $S2$[1]. From this model it is possible to recognize the guideline being followed for the diagnosis of Malaria. It is known from medical domain knowledge that $S1$ (*Intermittent Fever*) and $S2$ (*Enlarged Spleen*) are symptoms (effects) of Malaria and not causes for Malaria. However, the model tells us that it is "causal" for the diagnosis of Malaria because of the Malaria CPG.

It is not clear that a predictive rule or decision tree learner would identify the Malaria guideline because their goal is not to construct a model of the domain but maximize classification accuracy. Typical statistical and machine learning models seek to maximize predictive accuracy and will not in general generate the right causal model (see for example, [2, chapter 8] for related limitations of regression and [5] for causal limitations of SVMs). We note that causal validity is of the essence for the task of guideline compliance checking as discussed in Section 3.3.

Figure 3 shows the Malaria domain shown in Figure 1 augmented with the Malaria CPG. In Figure 3 node $D1$ denotes the disease Malaria and node $D2$ denotes Malaria diagnosis.

[1] On the other hand, if a variable X is manipulated (set randomly to one of its states), all the incoming arcs of X will be removed. The manipulation model is applicable to a randomized controlled trial (RCT).

3.3 Perform compliance checks for CPGs and quantify the degree of compliance

Compliance check of the Malaria CPG can be done using the CBN in Figure 2. By instantiating the nodes $C1$, $S1$ and $S2$ and propagating the evidence through the network we can ascertain the probability of the disease given the evidence, that is, $\mathrm{P}(D1|C1, S1, S2)$. See [1, 6] for a discussion of the algorithms for evidence propagation in a CBN. The quantification of compliance can be read from the conditional probability table for $D1$ in Figure 2 as the variables $C1, S1$ and $S2$ are the parents of $D1$. The estimated $\mathrm{P}(D1|C1, S1, S2)$ from $\mathcal{D}_2$ will give the degree of compliance for the Malaria CPG for the hospital from which $\mathcal{D}_2$ was obtained.

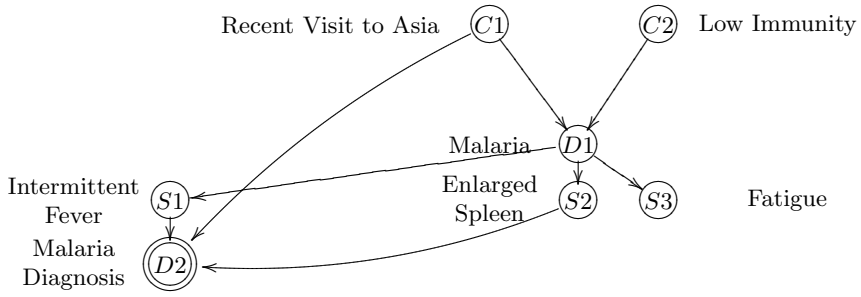

Fig. 3. An augmented causal Bayesian network structure for the Malaria diagnosis using guideline

3.4 Preliminary results

Preliminary results using the causal discovery algorithm FCI [2, Chapter 6] on a population-based dataset for a high blood pressure (HBP) study output a causal model that incorporated the guideline used in the study:

> If *Outpatient Blood pressure* = 1 or *Blood Pressure medication* = 1, HBP = H; else HBP = N.

The causal model output by FCI for the HBP domain had directed arcs from *Outpatient Blood pressure* and *Blood Pressure medication* to HBP. See [7] for additional details.

4 Discussion and Conclusion

In this paper we presented a framework based on causal Bayesian networks for (1) *de novo* generation of cause and effect clinical practice guidelines from data, (2) recognition of a CPG from clinical data and (3) compliance checking and quantification of the degree of compliance of a known guideline. Note that traditional machine learning algorithms such as decision trees and rules will generate

useful predictive models based on accuracy for the class variable (for example, diagnosis). However, they may not be robust under violations of iid (independent and identically distributed) and this may affect generalizability to other populations. Moreover, they will not be appropriate for prevention and intervention as they can neither model nor infer the causal interactions in the domain correctly. Predictive modeling techniques also cannot identify unobserved (hidden) variables or confounding.

Many practical problems exist before causal discovery methods such as the one discussed here will be able to routinely handle guideline discovery and compliance checking. An open problem is to understand how deterministic variables that result from application of clinical guidelines will impact guideline recognition from practice data using a CBN framework. Deterministic variables can result in violations of the faithfulness assumption that is typically required for causal discovery [8, 9]. There are also situations where the guidelines being followed may be implicit and the focus in learning such implicit guidelines is to understand and record what causes decision makers to arrive at certain decisions. In [10] it is shown that physicians are often non-compliant to the gold-standard guideline when they believe they are actually implementing them correctly.

References

1. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. 2nd edn. Morgan Kaufmann, San Francisco, California (1991)
2. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. 2nd edn. MIT Press, Cambridge, MA (2000)
3. Glymour, C., Cooper, G.F., eds.: Computation, Causation, and Discovery. MIT Press, Cambridge, MA (1999)
4. Russo, F., Williamson, J.: Interpreting causality in the health sciences. International Studies in the Philosophy of Science (2007) In Press.
5. Statnikov, A., Hardin, D., Aliferis, C.: Using SVM weight-based methods to identify causally relevant and non-causally relevant variables (2006) NIPS workshop poster.
6. Lauritzen, S., Spiegelhalter, D.: Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society Series B **50** (1988) 157–224
7. Mani, S., Aliferis, C., Krishnaswami, S., Kotchen, T.: Learning causal and predictive clinical practice guidelines from data. In: Proceedings of MedInfo, IOS Press (2007) In Press.
8. Pearl, J., Geiger, D., Verma, T.: The logic of influence diagrams. In Oliver, R.M., Smith, J.Q., eds.: Influence diagrams, belief nets and decision analysis. Wiley, New York (1990) 67–87
9. Cooper, G.F.: An Overview of the Representation and Discovery of Causal Relationships Using Bayesian Networks. In Glymour, C., Cooper, G.F., eds.: Computation, Causation, and Discovery. MIT Press, Cambridge, MA (1999) 3–62
10. Sboner, A., Aliferis, C.: Modeling clinical judgment and implicit guideline compliance in the diagnosis of melanomas using machine learning. In: Proceedings of the AMIA fall symposium. (2005)