Learning Causal and Predictive Clinical Practice Guidelines from Data

Subramani Mani[a], Constantin Aliferis[a], Shanthi Krishnaswami[b,c] and Theodore Kotchen[c]

subramani.mani@vanderbilt.edu constantin.aliferis@vanderbilt.edu shanthi.krishnas@vanderbilt.edu tkotchen@mcw.edu

[a] Department of Biomedical Informatics, Vanderbilt University, Nashville TN 37232

[b] Department of Rheumatology, Vanderbilt University, Nashville TN

[c] Department of Medicine, Medical College of Wisconsin, Milwaukee WI

Abstract

Clinical practice guidelines (CPG) propose preventive, diagnostic and treatment strategies based on the best available evidence. CPG enable practice of evidence-based medicine and bring about standardization of healthcare delivery in a given hospital, region, country or the whole world. This study explores generation of guidelines from data using machine learning, causal discovery methods and the domain of high blood pressure as an example.

Keywords: *Clinical Practice Guidelines, Machine Learning, Prediction, Compliance, Causal Discovery, High Blood Pressure*

Introduction and Background

Clinical practice guidelines (CPG) have been developed to streamline practice of medicine in a complex healthcare environment. Guidelines are the foundation of evidence-based medicine (EBM) and provider compliance with evidence-based CPG facilitates the practice of evidence-based medicine.

Typically guidelines are developed by a committee of domain experts specifically constituted for the purpose. After a rigorous review and analysis of published evidence over a period of time, the committee publishes its set of guidelines. However, this process is laborious, time consuming and expensive. See for the example the joint national committee report on high blood pressure [1].

Based on published evaluations of CPG, Grimshaw and Russell showed that practice guidelines streamlined the process of care and contributed to better outcomes in general [2]. There has been considerable interest in guideline representation for making them computable. Many different frameworks for guideline implementation such as PRODIGY [3], Arden Syntax [4], GLIF [5], PROforma [6], Asbru [7] and EON [8] have been proposed.

The fields of machine learning (ML), data mining and knowledge discovery including methods for learning cause and effect relationships have matured over the years with applications to clinical and biological data. There have only been limited attempts at using machine learning and knowledge discovery methods for generating practice guidelines from data [9-12]. The goal of Abston et al. [9] was to discover the variation between the American College of Cardiology guidelines for management of acute myocardial infarction and documented practices in a tertiary care facility in Salt Lake city. The study did not address generation of new guidelines from data. Mani et al. [10] presented a two-stage machine learning model as a data mining method to develop clinical practice guidelines and showed its value in dementia staging. It modeled the methodology used by clinicians by deriving intermediate concepts in the first phase and using the intermediate concepts for dementia staging. However, it is not clear if the method is generalizable across different domains. Morik et al. [11] used a combination of prior knowledge from experts and learning from data for clinical protocol generation and validation. Sboner et al. [12] used machine learning techniques to model the decision making of dermatologists in melanoma diagnosis. Svatek et al. [13] describe a data mining approach based on association rule learning for checking guideline adherence. However, the clinical validity of the approach could not be ascertained due to the small sample size of the study.

Machine learning techniques have typically been explored for disease screening [14], differential diagnosis [15], and other outcome measures [16]. A method based on inductive logic programming for learning qualitative physiological models from clinical data has also been described [17]. However, to the best of our knowledge causal discovery methods have not been explored for guideline generation from data.

This paper explores automated generation and compliance checking of guidelines. Guidelines can be broadly classified as predictive guidelines or prevention/intervention guidelines based on their goals. Note that predictive guidelines are sufficient for diagnosis (diagnose disease D1 from symptoms S1 and S2). For prevention and intervention we need a cause and effect interpretation for the guidelines.

We now introduce a formal notion of causality. We define the *causal influence* of a variable *A* on a variable *B* using the *manipulation criterion* [18, 19]. 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 $P(B \mid \text{set } A = a_1) \neq P(B \mid \text{set } A = a_2)$. A causal influence of variable *A* on variable *B* is represented as an arc from *A* to *B* i.e. $A \rightarrow B$. We say that variable *A* causally influences variable *B* if and only if *A* and *B* satisfy the manipulation criterion.

A causal influence of a variable *A* on a variable *B* is said to be *unconfounded* if and only if there is *no* measured or unmeasured variable *C* that is a common cause of variables *A* and *B*.

Materials and Methods

Algorithms

For learning predictive guidelines we selected two machine learning algorithms with the following properties.

1. Perform classification (prediction) tasks well.
2. The generated models are comprehensible to humans.
3. The models can be easily implemented as computerized guidelines.

The C4.5 algorithm that uses the decision tree representation formalism [20] and RIPPER [21] that has the format of an *If ... Then rule* were selected. Decision trees and rules generate clear descriptions of how the ML method arrives at a particular classification.

For checking the cause and effect interpretation of the guideline we used the FCI algorithm [18]. The FCI algorithm takes as input a dataset D and outputs a graphical model consisting of edges between them that have a cause and effect interpretation. The FCI algorithm can handle hidden (unmeasured) variables and sample selection bias that are likely to be present in real-world datasets. There are other causal discovery algorithms (for example, PC [18]) that output a causal Bayesian network (CBN) model [22, 23], incorporating all the variables represented in a dataset. However, PC assumes that all the variables in a domain are observed and there are no unobserved variables. There are also causal discovery algorithms that take a local approach and output causal relationships of the form "variable *A* causally influences variable *B*". LCD [24] and BLCD [25] are two such algorithms. The local causal discovery algorithms are particularly suitable for large datasets.

In this study we apply C4.5, RIPPER and FCI to the high blood pressure (HBP) dataset that is described below.

Dataset

The prevlanence of high blood pressure in the US is approaching 30% and the rate of prevalence is also showing an increasing trend [26]. The dataset used in this work is part of an ongoing NIH funded study with Dr. Kotchen, T as the principal investigator and its goal is to ascertain genetic determinants and other causal factors of high blood pressure. The HBP dataset is a population based dataset consisting of data collected from consenting African Americans between the ages of 18 and 55 years in Milwaukee and neighboring areas. Anthropometric measurements included height, weight, waist, hip, arm circumferences and skinfold thickness measured at different sites. Consenting subjects who satisfied inclusion and exclusion criteria were admitted for a 2 day inpatient protocol to obtain additional hemodynamic and renal measurements under standardized controlled conditions. Exclusion criteria included secondary hypertension, diabetes, creatinine >2.2 mg/dl, body mass index (BMI) >35, recent stroke or myocardial infarction, malignancy and substance abuse including alcohol. Currently 369 people are enrolled (202 hypertensives and 167 normotensives). 47.3% of the normotensives and 53% of hypertensives were females. The average blood pressure of normotensive subjects was 114/74 vs. 147/96 in hypertensive subjects. 31% of the subjects were on antihypertensive medication. We selected 23 variables after excluding patient identifiers and redundant variables (variables derived from other variables present in the dataset). See Table 1 for the list of variables selected. The variables were categorized based on either the established risk levels of each variable for cardiovascular diseases or the study-specific cutpoints using the 90th / 10th percentile levels of values in normotensive subjects. The outcome (class) variable was coded H for hypertensive and N for normotensive based on the following guideline. If the outpatient blood pressure (OP-BP) was greater than or equal to 140/90 or the subject was on BP medication (BP-MED), the outcome variable (HBP) was coded H, otherwise it was coded N. All the independent variables were categorized as 0 and 1 (one representing risk for high blood pressure). Thus the guideline used for creating the outcome variable was our target hypothesis for learning which is given below.

If OP-BP = 1 or BP-MED = 1, HBP = H; **else** HBP = N.

We created two datasets DS1 and DS2 as follows. DS1 had 21 variables after excluding OP-BP and BP-MED that were used for generating the outcome HBP. The purpose of creating DS1 was to ascertain the causal factors for the outcome variable as a baseline before generating guidelines. DS2 included all the 23 variables. A third dataset DS3 was created from DS2 by toggling the value of the outcome variable for a randomly selected 10% of the subjects i.e. if the value was H it was changed to N and vice versa. This was to artificially create a set of 37 patients for whom the guideline was violated. DS3 was used for verification of guideline compliance. A fourth dataset DS4 was also created with data on this set of 37 patients for whom the class label was manipulated. C4.5 and Ripper were run using datasets DS2, DS3 and DS4. FCI was run on datasets DS1 and DS2.

The FCI program is available from the Tetrad project site at Carnegie Mellon (www.phil.cmu.edu/tetrad/tetrad4.html). We used the Java implementation of C4.5 and RIPPER available in the Weka machine learning software package [27]. We performed a ten fold cross-validation for C4.5 and RIPPER and report the results based on the test cases that were not used in model building. Note that for causal discovery cross-

validation is not relevant because we are not performing classification or regression.

Table 1: Variables used in the study

Plasma Aldo/ plasma Renin ratio
Age
Cardiac output baseline
Creatinine clearance
Gender
High density lipoprotein
heart rate at baseline
Insulin Resistance
Potassium excretion
Calculated Low density lipoprotein
Sodium excretion
Baseline renal Blood flow
Baseline Renal vascular resistance
Systemic Vascular resistance Index baseline
Stroke Volume baseline
Serum Triglycerides
Urine 24hrs Microalbumin
Waist circumference risk
Outpatient Hypertension yes/no
Glucose risk
Screening Height
Outpatient high BP /Normal BP
On antihypertensive medication

Results

Figures 1 and 2 present the results of the application of C4.5 and Ripper to the HBP DS2 dataset.

```
OP-BP = 0
|   BP-MED = 0: N
|   BP-MED = 1: H
OP-BP = 1: H

Number of Leaves  :     3
Size of the tree :      5
```

Figure 1: A C4.5 tree from DS2

The C4.5 tree classified all the 369 instances correctly with an accuracy of 100%. The precision and recall were 1 for the H and N classes.
RIPPER

DS2 Ripper rules:

(OP-BP = 0) and (BP-MED = 0) => HBP=N
=> HBP = H

Figure 2: A RIPPER rule set from DS2

The RIPPER classified all the 369 instances correctly with an accuracy of 100%.

Results of C4.5 and Ripper on DS3

Both C4.5 and Ripper misclassified the 37 instances for which the class assignments had been changed. All the other instances were classified correctly. Both the algorithms generated the same models shown in Figures 1 and 2.

Results of C4.5 and Ripper on DS4

The results of application of C4.5 and Ripper to the DS4 dataset are shown in Figure 3 and Figure 4 respectively. Recall that the DS4 dataset has just the 37 instances with labels manipulated.

```
OP-BP = 0
|   BP-MED = 0: H
|   BP-MED = 1: N
OP-BP = 1: N

Number of Leaves  :     3
Size of the tree :      5
```

Figure 3: A C4.5 tree from DS4

The DS4 C4.5 tree classified 35 out of the 37 instances correctly with an accuracy of 95%. The precision for class N was 1 and recall for class H was 1.

RIPPER

DS4 Ripper rules:

(OP-BP = 0) and (BP-MED = 0) => HBP=H
=> HBP = N

Figure 4: A RIPPER rule set from DS4

The RIPPER rule set classified 35 out of the 37 instances correctly with an accuracy of 95%.

FCI Results

When applied to the HBP DS1 dataset, FCI output four possible causal factors for high blood pressure. Table 2 enumerates them. Note that when the relationship is categorized as "O—>", there could be a common cause or a feedback loop.

Discussion

The machine learning algorithms C4.5 and Ripper recovered the study guideline that was used for assigning labels to the outcome variable when the guideline was followed in all the cases (DS2). The same guideline was also generated from the dataset when the guideline was not followed in 10% of the subjects (DS3). The subjects for whom the guideline was not followed were also identified as misclassified instances by C4.5 and Ripper. This shows that ML methods can be used for generating simple guidelines and guideline compliance checking.

Table 2: The output of FCI from DS1

Aldo/Renin ratio O—> HBP
Age O—> HBP
Renal vascular resistance <—> HBP
Waist risk <—> HBP

A O—> B Means there is definitely an arrowhead at B. There may or may not be an arrowhead at A.
A <—> B denotes that there is a common cause for A and B.

Table 3: The output of FCI from DS2

OP-BP → HBP
BP-MED → HBP
HDL Cholesterol → Insulin resistance
Insulin resistance → Glucose risk

The results shown in Figure 3 and Figure 4 using the dataset DS4 with the 37 misclassified cases from DS3 shows the alternate guideline model generated for those cases. This is the alternate guideline that was followed for these instances. This shows that when a specified guideline is not followed, ML can be used to identify any alternate guideline that might have been used instead. Note though that the instances would be misclassified only if the application of the alternate guideline changed the outcome variable.

We also used a causal discovery algorithm (FCI) to ascertain the cause and effect basis of the guideline. Using DS1 FCI output four causal influences shown in Table 2. When DS2 was used OP-BP and BP-MED were output by FCI as unconfounded causal factors for HBP. Note that OP-BP and BP-MED are causal factors for HBP based on the manipulation criterion for causality. However, they are also causal based on the definition of the BP study guideline. Two other unconfounded causal relationships from the domain were also output by FCI (see Table 3).

Decision tree models and If … Then rules are expressive and easily interpretable by humans. Moreover, the tree and rule formats are also suitable for computerized guidelines and hence useful for incorporation as decision support tools in electronic medical record systems.

Our study addresses the question of generating new practice guidelines in a data driven way and explores the role of causal discovery along with traditional machine learning approaches for guideline generation from data.

Two relevant issues that come up in guideline application are generalizability and customization. A guideline developed in one institution or organization may not be exactly applicable in another practice setting. Likewise, a guideline developed by a committee of national or international experts might need to be customized to a local setting. Fridsma et al. have developed a knowledge-based approach to customization based on the separation of site specific and site independent factors that can be identified from the knowledge of the organization and understanding of its workflow [28]. We believe that data driven machine learning approaches could be a useful tool in the overall effort to make guidelines generalizable and customizable.

Limitations

The dataset that we used in this study was from a population-based study of high blood pressure. There were only a small number of variables in the dataset. Only two machine learning and one causal discovery algorithm were used in this study. The guidelines were also very simple.

Conclusions and Future Work

In this paper we presented a machine learning approach to generate guidelines from data, check for guideline compliance and if non-compliant for a set of patients, generate the alternate guideline used. We also provided a method for ascertaining whether the guideline has a causal semantics using a causal discovery algorithm.

In future we plan to apply machine learning and causal discovery algorithms to different medical datasets involving more complex guidelines for further evaluation of our approach.

References

[1] Chobanian, A, Bakris, G, Black, H, Cushman, W, Green, L, Izzo, J, Jones, D, Materson, B, Oparil, S, Wright, J, Roccella, E and NHBPEPCC. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. Hypertension l2003;42: 1206-1252.

[2] Grimshaw, JM and Russell, IT. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. Lancet l1993;342: 1317-22.

[3] Purves, I, Sugden, B, Booth, N and Sowerby, M. The PRODIGY project--the iterative development of the release one model. In: AMIA fall symposium; 1999. p. 359-63.

[4] Clayton, P, Pryor, T, Wigertz, O and Hripcsak, G. Issues and structures for sharing knowledge among decision-making systems: The 1989 Arden Homestead Retreat. In: K. LC, editor. Thirteenth Annual Symposium on Computer Applications in Medical Care: IEEE Computer Society Press; 1989. p. 116-21.

[5] Ohno-Machado, L, Gennari, J, Murphy, S, Jain, N, Tu, S, Oliver, D, Pattison-Gordon, E, Greenes, R, Shortliffe, E and Barnett, G. The guideline interchange format: a model for representing guidelines. JAMIA l1998;5: 357-72.

[6] Fox, J, Johns, N and Rahmanzadeh, A. Disseminating medical knowledge: The PROforma approach. Artiff Intell Med l1998;14: 157-81.

[7] Miksch, S, Shahar, Y and Johnson, P. Asbru: a task-specific, intention-based, and time-oriented language for representing skeletal plans. In: Seventh Workshop on Knowledge Engineering Methods and Languages (KEML-97): Milton Keynes, UK; 1997.

[8] Musen, M, Tu, S, Das, A and Shahar, Y. EON: a component-based approach to automation protocol-directed therapy. JAMIA l1996;3: 367-88.

[9] Abston, K, Pryor, T, Haug, P and Anderson, J. Inducing practice guidelines from a hospital database. In: JAMIA Supplement; 1997. p. 168--172.

[10] Mani, S, Shankle, WR, Dick, MB and Pazzani, MJ. Two-Stage Machine Learning Model for Guideline Development. In: Artificial Intelligence in Medicine; 1998.

[11] Morik, K, Imboff, M, Brockhausen, P, Joachims, T and Gather, U. Knowledge discovery and knowledge validation in intensive care. Artiff Intell Med l2000;19: 225-249.

[12] Sboner, A and Aliferis, CF. Modeling clinical judgment and implicit guideline compliance in the diagnosis of melanomas using machine learning. In: AMIA fall symposium; 2005. p. 664-668.

[13] Svatek, V, Riha, A, Peleska, J and Rauch, J. Analysis of guideline compliance---a data mining approach. In: K. Kaiser, S. Miksch and S. W. Tu, editors. Symposium on Computerized Guidelines and Protocols: IOS Press; 2004. p. 157-161.

[14] Shankle, W, Mani, S, Pazzani, M and Smyth, P. Detecting Very Early Stages of Dementia from Normal Aging with Machine Learning Methods. In: E. Keravnou, C. Garbay, R. Baud and J. Wyatt, editors. Lecture Notes in Artificial Intelligence: Artificial Intelligence in Medicine, AIME97: Springer; 1997. p. 73--85.

[15] Ohmann, C, Yang, Q, Moustakis, V, Lang, K and Elk, van PJ. Machine Learning Techniques Applied to the Diagnosis of Acute Abdominal Pain. In: P. Barahona and M. Stefanelli, editors. Lecture Notes in Artificial Intelligence: Artificial Intelligence in Medicine, AIME95: Springer; 1995. p. 276--281.

[16] Cooper, G, Aliferis, C and Ambrosino, R. An evaluation of machine learning methods for predicting pneumonia mortality. Artificial Intelligence in Medicine l1997;9: 107--138.

[17] Hau, DT and Coiera, EW. Learning qualitative models of dynamic systems. Machine Learning l1993;26: 177-211.

[18] Spirtes, P, Glymour, C and Scheines, R. Causation, Prediction, and Search. In. Cambridge, MA: MIT Press; 2000.

[19] Cooper, GF. An overview of the representation and discovery of causal relationships using Bayesian networks. In: C. Glymour and G. F. Cooper, editors. Computation, Causation, and Discovery. Cambridge, MA: MIT Press; 1999. p. 3--62.

[20] Quinlan, J. C4.5: Programs for Machine Learning. In. Los Altos, California: Morgan Kaufmann; 1993.

[21] Cohen, WW. Fast effective rule induction. In: ICML; 1995. p. 115-123.

[22] Pearl, J. Probabilistic Reasoning in Intelligent Systems. In. San Francisco, California: Morgan Kaufmann; 1991.

[23] Neapolitan, RE. Probabilistic Reasoning in Expert Systems: Theory and Algorithms. In. New York: John Wiley and Sons; 1990.

[24] Cooper, GF. A Simple Constraint-Based Algorithm for Efficiently Mining Observational Databases for Causal Relationships. Data Mining and Knowledge Discovery l1997;1: 203--224.

[25] Mani, S. A Bayesian Local Causal Discovery Framework (doctoral dissertation). In: Intelligent Systems Program: University of Pittsburgh; 2005.

[26] Hajjar, I and Kotchen, TA. Trends in prevalence, awareness, treatment, and control of hypertension in the United States, 1988 - 2000. JAMA l2003;290: 199-206.

[27] Witten, IH and Frank, E. Data Mining: Practical machine learning tools and techniques. In. 2nd ed: Morgan Kaufmann, San Francisco; 2005.

[28] Fridsma, DB, Gennari, J and Musen, M. Making generic guidelines site-specific. In: AMIA fall symposium; 1996.