Causal Discovery from Medical Textual Data

Subramani Mani, MBBS, MS, and **Gregory F. Cooper, MD, PhD**
{mani,gfc}@cbmi.upmc.edu
Center for Biomedical Informatics and the Intelligent Systems Program, University of Pittsburgh

Medical records usually incorporate investigative reports, historical notes, patient encounters or discharge summaries as textual data. This study focused on learning causal relationships from intensive care unit (ICU) discharge summaries of 1611 patients. Identification of the causal factors of clinical conditions and outcomes can help us formulate better management, prevention and control strategies for the improvement of health care. For causal discovery we applied the Local Causal Discovery (LCD) algorithm, which uses the framework of causal Bayesian Networks to represent causal relationships among model variables. LCD takes as input a dataset and outputs causes of the form variable Y causally influences variable Z. Using the words that occur in the discharge summaries as attributes for input, LCD output 8 purported causal relationships. The relationships ranked as most probable subjectively appear to be most causally plausible.

Introduction

Text is ubiquitous. In various fields text is still the preferred format for summarizing information. Even though more and more data are being coded and analyzed, textual data often remain the richest source of information about clinical care. For example, in medicine, textual data coexist with coded data in patient records. Patient history, radiological reports, patient encounters and discharge summaries are some areas where textual data are prevalent.

Causality plays a central role in all scientific disciplines. Causal knowledge aids planning and decision making. In the domain of medicine, determining the cause of a disease helps in prevention and treatment.

Learning from textual data is an emerging area of research (see [1] for a recent review). In this work, we report on research to discover causal influences from discharge summaries of patients admitted to the intensive care unit (ICU). Our ultimate goal is to provide an analytic tool that assists clinical researchers to discover interesting and novel causal relationships in medicine.

This paper first introduces the algorithm called LCD [2] that was designed for efficient discovery of possible causal relationships from large observational databases. We have previously applied LCD to a population-based infant birth and death dataset of 41,000 instances and 87 variables [3]. In that research, we obtained nine relationships out of which eight seem plausibly causal. The present study focused on causal discovery from textual data using LCD.

Methods

Assumptions for Causal Discovery

In the research reported here, we use causal Bayesian networks to represent causal relationships among model variables. This section provides a brief introduction to causal Bayesian networks, as well as a description of the assumptions we used to apply these networks for causal discovery.

A causal Bayesian network (or *causal network* for short) is a Bayesian network in which each arc is interpreted as a direct causal influence between a parent node (cause) and a child node (effect) [4]. Figure 1 illustrates the structure of a hypothetical causal Bayesian network, which contains five nodes. A root node (node without parents) is associated with a prior probability distribution, and a non-root node has a conditional probability table quantifying the parent-child probabilistic relationships. The probabilities are not shown in the figure.

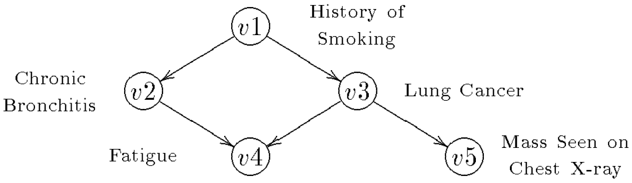

Figure 1: A hypothetical causal Bayesian network

The causal network in Figure 1 indicates, for example, that a *History of Smoking* can causally influence whether *Lung Cancer* is present, which in turn can causally influence whether a patient experiences *Fatigue* and whether the patient presents with a *Mass Seen on Chest X-ray.*

The **causal Markov condition** gives the independence relationships[1] that are specified by a causal Bayesian network:

> A variable is independent of its non-descendants (i.e., non-effects) given just its parents (i.e., its direct causes).

According to the Markov condition, the causal network in Figure 1 is representing that *Mass Seen on Chest X-ray* is not influenced by *History of Smoking,* given that we know whether *Lung Cancer* is fixed at some state (present or absent). While the causal Markov condition specifies independence rela-

[1] We use the terms *independence* and *dependence* in this section in the standard probabilistic sense.

tionships among variables, the **causal faithfulness condition** specifies *dependence* relationships:

> Variables are independent only if their independence is implied by the causal Markov condition.

For the causal network in Figure 1, three examples of the causal faithfulness condition are (1) *History of Smoking* and *Lung Cancer* are dependent, (2) *History of Smoking* and *Mass Seen on Chest X-ray* are dependent, and (3) *Mass Seen on Chest X-ray* and *Fatigue* are dependent. The intuition behind that last example is as follows: a *Mass Seen on Chest X-ray* increases the possibility that there exists *Lung Cancer* which in turn increases the chance of *Fatigue*; thus, the variables *Mass Seen on Chest X-ray* and *Fatigue* are expected to be probabilistically dependent. In other words, the two variables are dependent because of a common cause (i.e., a confounder).

The causal Markov and faithfulness conditions describe *probabilistic* independence and dependence relationships, respectively, that are represented by a causal Bayesian network. In causal discovery, we do not know the probabilistic relationships among variables precisely, because we only have a finite amount of data. Thus, we make the following **statistical testing assumption**:

> A statistical test performed to determine independence (or alternatively dependence) given a finite dataset will be correct relative to independence (dependence) in the joint probability distribution defined by the causal process that is generating the data under study.

That is, we assume our statistical test gives valid results. The greater the number of records in a dataset, the more likely it is that the statistical testing assumption will hold.

LCD Algorithm for Causal Discovery

LCD relies on the following:

> **Assumption 1:** The causal Markov condition
> **Assumption 2:** The causal faithfulness condition
> **Assumption 3:** The statistical testing assumption

In addition, LCD makes the following assumption:

> **Assumption 4:** Given measured variables X, Y, and Z, if Y causes Z, and Y and Z are not confounded, then one of the causal networks in Figure 2 must hold.

Assumption 4 implies that X is not causally influenced by Y or by Z. As we discuss in later sections, in our experiments we chose X so that this implication holds.

Before introducing the LCD algorithm in more detail, we define some terms. Let Independent$_T$(A, B) denote that A and B are independent according to test T applied to our dataset. Let Independent$_T$(A, B given C) denote that A and B are independent given C, according to T. Finally, let Dependent$_T$(A, B) denote that A and B are dependent according to T^2. As explained below, LCD outputs Y as a cause of Z if all the following tests hold:

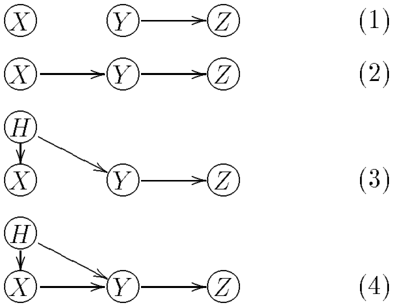

Figure 2: Causal models in which Y causes Z; X and Y are independent (1), X and Y are dependent due to X causing Y (2,4), X and Y dependent being confounded by a hidden variable(s) represented by H (3,4).

Test$_1$. Dependent$_T$(X, Y)
Test$_2$. Dependent$_T$(Y, Z)
Test$_3$. Dependent$_T$(X, Z)
Test$_4$. Independent$_T$(X, Z given Y)

The first network in Figure 2 violates Test$_1$, and thus, LCD is unable to detect that Y causally influences Z in such situations. Under Assumptions 1 through 3, the other three networks in Figure 2 satisfy Test$_1$ through Test$_4$. If Y and Z are confounded, then at least one of the four tests will always be violated [2].

As an example, Figure 3 shows a hypothetical case in which Y and Z are confounded by a hidden (latent) variable H. For this causal network, it follows from Assumptions 1 and 2 that X and Z will be dependent given Y, and thus, Test$_4$ will fail. In this hypothetical model, Gender (X) and Gene (H) are independent. Consider that the gene (H) has two alleles $a1$ and $a2$ and that $a1$ predisposes to (1) increased alcohol consumption and (2) cirrhosis of the liver. This model also assumes that male gender increases alcohol intake. Now assume that a patient with cirrhosis of liver gives a history of increased alcohol intake. If the patient is male the probability of the patient having the allele $a1$ decreases, and if female, the probability of $a1$ increases. When there are two independent causes for a phenomenon, evidence for one of the causes reduces the probability of the other cause. Knowing the state of Y (that a person's alcohol consumption is high) makes Gen-

[2] Although the three tests in this paragraph should technically be distinguished from each other by using separate labels, such as *T1*, *T2*, and *T3*, for simplicity of notation we use a single label T.

der (X) and Gene dependent and hence Gender and Cirrhosis of Liver (Z) dependent.

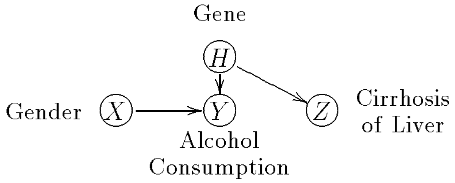

Figure 3: Causal model with hypothetical labels in which X causes Y, and Y and Z are dependent due to confounding by a hidden variable(s) represented by H.

To summarize, under Assumptions 1 through 4, when Y causally influences Z and these two variables are unconfounded, the four tests hold (unless X and Y are independent). Conversely, when Y and Z are confounded (or when X and Y are independent), one or more of the four tests will fail. From these propositions, we can conclude that if the four tests hold, then one of the three causal networks (2,3,4) in Figure 2 must hold, and thus, we can conclude that Y causes Z and the two variables are unconfounded.

The algorithm LCD performs Test$_1$ through Test$_4$ sequentially to evaluate triplets of the form XYZ in the database. Simple variations of the Independence and Dependence tests described in [2] were used. Both tests have $O(m)$ time complexity, where m is the number of records (cases) in the database. If all four tests are passed, LCD outputs that *Y causally influences Z and the two variables are unconfounded* (under Assumptions 1–4), and the probability distribution of Z given Y is displayed.

Time Complexity of LCD We assume here that all the variables are categorical, and the number of levels (states) of any of the variables in V is bounded by some constant. If there are n variables in the database, the time complexity of LCD is $O(mn^2r)$, where n is the number of variables and r is the number of variables of type X. If we restrict the number of variables of type X in the search, so that r is bounded above by some constant, then the time complexity is $O(mn^2)$. If we additionally focus on a bounded number of effects of interest (variable Z), the time complexity reduces further to $O(mn)$. In all variations, the space complexity of LCD is $O(mn)$, which is the size of the database.

Further, the algorithm can be implemented to output the causes as they are discovered. The time complexity of LCD makes it appropriate for exploring possible causal relationships in databases that contain a very large number of records (on the order of hundreds of thousands) and a moderately large number of measured variables per record (on the order of a few thousand). This order of efficiency makes LCD particularly suitable for causal data mining from textual data.

Text Data Mining Text data mining is a nascent field [1]. To our knowledge, the present paper reports the first investigation of automated causal discovery from text. A comprehensive review of work in text data mining is not attempted here. Recent works of interest include the following: Feldman and Dagan describe a methodology for knowledge discovery from text by constructing a concept hierarchy by annotating text articles. Their framework makes use of text categorization for finding new patterns [5]. A novel knowledge discovery approach to text has been reported by Swanson and Smallheiser who have developed a user-interactive specialized knowledge discovery tool. Using Medline abstracts from isolated disciplines that do not cite each other, these researchers have been successful in proposing new and useful relationships of scientific merit in the biomedical domain. Some of these postulates were later verified by experimental work [6].

Experimental Methods

Text dataset The text dataset was 2060 intensive care unit discharge summaries (documents) of patients admitted to two medical ICUs of the University of Pittsburgh Medical Center (UPMC) during the years 1993, 1994, and 1995. Patient names, addresses, physician names and other relevant identifying features were removed from the records before further processing. A dictionary was created by incorporating all words occurring in at least 50 of the summaries. We used this filter (1) to focus on words that occur frequently, and (2) to reduce computation time. Multiple occurrences of the same word in any one document were counted as one. After ignoring stop words, there were 1808 unique words that occurred in 50 or more documents.

Age, gender and race of the patients were extracted from the summaries using a computer program. We could identify these attributes in 1611 records. Each of the 1808 words were coded as 1 if it was present in a particular document and 0 if it was not present. After incorporating age, gender and race we had a dataset with 1811 attributes and 1611 records.

LCD Runs The LCD algorithm was executed as follows. Age, gender and race were used as the acausal X type variables (see Figure 2). They are considered *acausal* as they cannot be caused by any other measured variables. We make the assumption that they are not causally influenced by the attributes used in this study. Since we had 1811 attributes, to reduce computational time we focused on twenty Z variables (possible effects). We looked for medically meaningful conditions (nouns) and selected the following words from the list of attributes:

> Nausea, hypotension, lymphadenopathy, seizures, encephalopathy, effusion, embolism, metastases, thrombocytopenia, acidosis, dizziness, bacteremia, thrombosis, rash, pancreatitis, cirrhosis, hemateme-

sis, dyspnea, hypokalemia, and allergy.

This was done to avoid searching just for the causes of commonly occurring words such as slow, marked, high, same or room. With this restriction LCD would have evaluated 35,760 pairs of words for causal influence. For the statistical tests of dependence and independence, thresholds of 0.9, 0.8, and 0.7 were used with the test reported in [2].

Results

Table 1 summarizes the causal output at the dependence and independence threshold level of 0.9. The first two entries causally linking *alcoholic* and *alcohol* to *cirrhosis* seem plausible; this relationship is well-known in the medical literature. Table 2 gives the probability distribution of *cirrhosis* given *alcoholic*. The third entry in Table 1 seems to be plausible only in the reverse direction, assuming *portal* denotes *portal hypertension*. We are investigating the possibility that this false positive output was due to subtle confounding, which could be eliminated if the confounders (e.g., alcoholic) are considered. Using a threshold of 0.8 and lower gave five relationships that appear unreliable including for example that "ascitis causes cirrhosis". Hence, we did not analyze those results further.

Table 1: LCD output (Y causally influencing Z) at threshold 0.9.

X NODE	Y NODE	Z NODE
GENDER	ALCOHOLIC	CIRRHOSIS
GENDER	ALCOHOL	CIRRHOSIS
GENDER	PORTAL	CIRRHOSIS

Table 2: Conditional probability table of cirrhosis given alcoholic

	Alcoholic	
Cirrhosis	Absent	Present
Absent	0.96*	0.61
Present	0.04	0.39

*The probability that cirrhosis is absent *given* that alcoholic is absent.

Discussion

By our medical judgment, the program output one false positive and one true positive causal relationship. We believe the false positive result can be disregarded based on known causal relationships in medicine. In general one could use time precedence or prior medical knowledge to evaluate the output and eliminate many false positives. Such a filter could also be incorporated in the causal discovery system. Consider the output "Y causes Z". A useful enhancement would be a graphical user interface that highlights the appearance of the terms Y and Z in the original text records; such an interface would help the user judge the clinical plausibility of Y causing Z. We also did not find any novel relationships in this study. We believe the chance of finding novel relationships will improve when we include (1) multi-word phrases, (2) a larger number of records, and (3) multivariate causes of effects. It may also be useful to encode (1) variable-value pairs (e.g., serum sodium = high), (2) the number of occurrences of a phrase in documents, and (3) the location of phrases in documents.

In causal discovery using coded data, the attributes are well defined entities and the values they take are assigned based on a widely accepted protocol. For example, a recording of the blood pressure is done using a sphygmo-manometer and the reading is noted. This might then be recoded based on accepted cut-off values into normal, mild hypertension, moderate hypertension or severe hypertension. In *text*, based on the context, all these categories of hypertension might just be referred to as hypertension or high blood pressure. So the words appearing in medical text capture the real world through a sort of prism of the care provider. In other words, a level of abstraction or interpretation is involved in the creation of textual data.

Related work

Causal Discovery Algorithms Traditional statistical approaches using for example χ^2 tests or logistic regression can establish dependence between variables. Likewise, machine learning algorithms such as decision tree learners (e.g., C4.5 and CART), rule inducers (e.g., C4.5Rules and FOCL) and neural networks can build useful domain models from data and capture the inter-dependence among the variables. But none of these techniques is intended to establish relationships of the form Y *causally* influences Z.

Structural equation models (SEMs) [7], represent causal relationships, thus going beyond correlation and dependence. The emphasis in SEM research is on hypothesis testing of manually specified models, rather than on automated search over the space of models. Typically SEM assumes linear relationships (with statistical noise) among continuous model variables.

For a detailed discussion of the relationship between statistical association and causation, including philosophical issues, see for example [8] and [9]. Earlier research on learning Bayesian networks from data using a Bayesian approach [10, 11] simultaneously modeled all the causal relationships among the model variables. These global approaches have worst-case time complexities that are exponential in the number of measured variables V.

Constraint-based approaches to causal discovery were put forward by Pearl and Verma [12] and by Spirtes, Glymour, and Scheines [13]. The PC and FCI algorithms, for example, take a global approach to causal discovery and output a network with different types of edges between variables that represent for example that X causes Y, X does not cause Y, or the causal direction is undetermined [14]. The FCI can also model latent variables.

LCD is a constraint-based algorithm that limits its search to triplets of variables, and outputs only causes of the form Y causes Z. By searching only for pairwise causal relationships, it trades off completeness for efficiency.

The goal of causal discovery from textual data is to find relationships between words in text *across* documents that are potentially causal. Natural language processing (NLP) can help automate text understanding, for example, it could be used to summarize text. NLP methods could also be used to map words and phrases to concepts which could then be used as attributes for causal discovery.

Limitations

Only words, not phrases were used. The text words take only two values in each record—present or absent. The context in which the word appears is not considered. For example, negation of words are ignored. Hence, *hypertensive* and *not hypertensive* appearing in a text document are coded as *present* for the attribute *hypertensive*. Likewise, synonyms are considered as different attributes.

Conclusion and Future Work

This paper reports early research in causal data mining from medical textual data using a local causal discovery (LCD) approach. It would also be useful to apply this paradigm to other medical textual data like radiological reports and patient encounter summaries. This approach should be generalizable to similar text data in other fields. But this remains to be tested.

We plan to refine our attribute gathering by using stemming of words and also by the use of an extended stop list. Another future direction would be to consider phrases containing up to four contiguous words, rather than just single words. These attributes can also be mapped to a concept library like the UMLS system and refined, thereby incorporating synonymy.

Acknowledgements

We thank Paul Hanbury for de-identifying and preprocessing the records, and Mehmet Kayaalp for helpful discussions and comments on an earlier draft. This work was supported by the National Library of Medicine (training grant LM07059 and grant R01-LM06696) and by the National Science Foundation (grant IIS-9812021).

References

[1] Marti A. Hearst. Untangling text data mining. In *Proceedings of ACL'99*, 1999.

[2] Gregory F. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1:203–224, 1997.

[3] Subramani Mani and Gregory F. Cooper. A study in causal discovery from population-based infant birth and death records. In *Proceedings of the AMIA Annual Fall Symposium*, pages 315–319, Philadelphia, PA, 1999. Hanley & Belfus.

[4] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Francisco, California, 1991.

[5] Ronen Feldman and Ido Dagan. Knowledge discovery in textual databases (KDT). In *Proceedings of the First Annual Conference on Knowledge Discovery and Data Mining (KDD)*, pages 112–117. AAAI Press, 1995.

[6] Don R. Swanson and N.R. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91:183–203, 1997.

[7] Kenneth A. Bollen. *Structural Equations With Latent Variables.* Wiley, New York, 1989.

[8] J.H. Fetzer, editor. *Probability and Causality.* D. Reidel Publishing Company, Boston, 1989.

[9] Wesley C. Salmon. *Causality and Explanation.* Oxford University Press, New York, 1998.

[10] Gregory F. Cooper and Edward Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[11] David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

[12] J. Pearl and T. Verma. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Prepresentation and Reasoning: Proceedings of the Second International Conference*, pages 441–452, San Francisco, CA, 1991. Morgan Kaufmann.

[13] P. Spirtes, C. Glymour, and R. Scheines. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

[14] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* Springer-Verlag, New York, 1993.