

# Mining Medical Data for Causal and Temporal Patterns

Ahmed Y. Tawfik<sup>1</sup> and Krista Strickland<sup>2</sup>

<sup>1</sup> School of Computer Science, University of Windsor, Windsor,  
Ontario N9B 3P4, Canada  
atawfik@uwindsor.ca

<sup>2</sup> Department of Mathematics and Computer Science, University of Prince Edward Island,  
Charlottetown, PEI C1A 3P4, Canada  
kstrickland@upei.ca

**Abstract.** The present work presents some causal and temporal patterns found in clinical data collected in Chiba University Hospital, Japan for patients suffering from collagen diseases. To extract causal patterns, statistical techniques based on covariance analysis are used. The extraction of temporal patterns uses survival analysis techniques. The results obtained using these techniques are analyzed and discussed.

## 1 Introduction

Probabilistic causality is playing an increasingly important role in intelligent data mining and evidence-based reasoning. Probabilistic causality [6, 3] extends the established notion of conditional dependence to causal interpretation. Bayesian Belief Networks [4] rely on a causal interpretation of conditional dependence to perform probabilistic reasoning based on evidence and causation. Data analysis methods have been developed to extract causal patterns from data including Tetrad [5] which is used in the present study, learning belief nets [2], and structural equation models.

Probabilistic causality aims at developing methodologies that allow us to deduce that X is a probable cause of Y when the probability of X given Y is different from the probability of X. The techniques differ however in how they try to eliminate spurious correlation, identify when a correlation can be attributed to hidden common causes, and in other implicit assumptions they make about the data (whether it is discrete or continuous for example) and the data distribution. The degree to which causal analysis techniques allow the integration of background knowledge also differs from one technique to another.

The causal analysis performed here uses Tetrad [5] to build structural graph representing causal patterns implied by the data. A structural graph consists of a set of nodes (V) and a set of edges (E) such that each node represents an entity (a data field) and an edge is directed from the cause to the effect. Undirected edges suggest a possible common hidden cause.

Medicine has traditionally used probabilistic causality to measure the effectiveness of a treatment. In such studies, a sample of patients is drawn at random and is then divided into a treatment group and a control group. Statistically significant

cant differences between the treatment and the control group are causally attributed to the treatment. This approach to probabilistic causality cannot be used to study disease causing factor or symptoms associated with diseases for obvious ethical and health problems. Methods of probabilistic causality similar to the one used here allow such studies to be conducted. These methods use the data available from infected individuals only (i.e. without control) by analyzing the covariance of different observed quantities.

In addition to causal analysis, this study considers some temporal aspects related to thrombosis. The time between being diagnosed by an auto-immune disease and developing thrombosis is studied. Here we assume that the description time is the time of diagnosis. If this date is not available, we consider the field labeled “first date”. This may seem counter-intuitive but this choice produced more logical and acceptable results than using the “first date” as the time of first diagnosis.

The temporal analysis utilizes survival analysis techniques. All dates are converted to days. The date of diagnosis is considered as the time origin for each individual, and the date of thrombosis is used as the event of interest. Patients who did not develop thrombosis are considered censored.

Following this introduction, Section 2 describes the data preparation stage. Section 3 presents the results of analyzing the medical data for the collagen disease patients. Section 4 presents the results of a temporal analysis of the data. Section 5 concludes with a discussion of the limitations of causal analysis observed in the results.

## 2 Data Preparation

The data preparation stage involved two steps: data-cleaning and data transformation. The data-cleaning step eliminated unprintable characters, and inconsistent data. Inconsistencies of two forms have been processed in this step as needed:

1. Temporal inconsistencies: for example, some patients have test results predating their date of birth.
2. Data type inconsistencies: for example, some numeric fields contain alphanumeric characters. In many cases, additional characters indicating the units used followed a test result.

Records containing temporal inconsistencies were discarded only if the inconsistency is relevant to the quantities analyzed. For example, we ignored the records corresponding to tests predating birth when the time since infection is considered.

Following the cleaning step, transformation and extraction operations are performed. The transformations aim at converting all the data fields to numeric fields as required by Tetrad input module. It is also necessary to extract a subset of fields and records that do not contain missing values because Tetrad does not allow missing values. Therefore, the number of records decreases as the number of fields increases.

### 3 Causal Graphs

Tetrad starts with a complete graph and uses statistical tests to verify that two data fields are independent or conditionally independent. Edges are eliminated from the graph to reflect the independencies detected. The remaining edges form a graph with three type of edges:

1. Directed edges ( $\rightarrow$ ) indicating a cause-effect relationship: the head of the arrow points to the effect and its tail points to the cause.
2. Bidirected edges ( $\leftrightarrow$ ) indicating a hidden common cause.
3. Directed edges with a small circle at its tail ( $\circ \rightarrow$ ) indicate a cause-effect or common cause.
4. Edge with circles at both ends ( $\circ - \circ$ ) indicate that either could be causing the other or that there is a common cause.

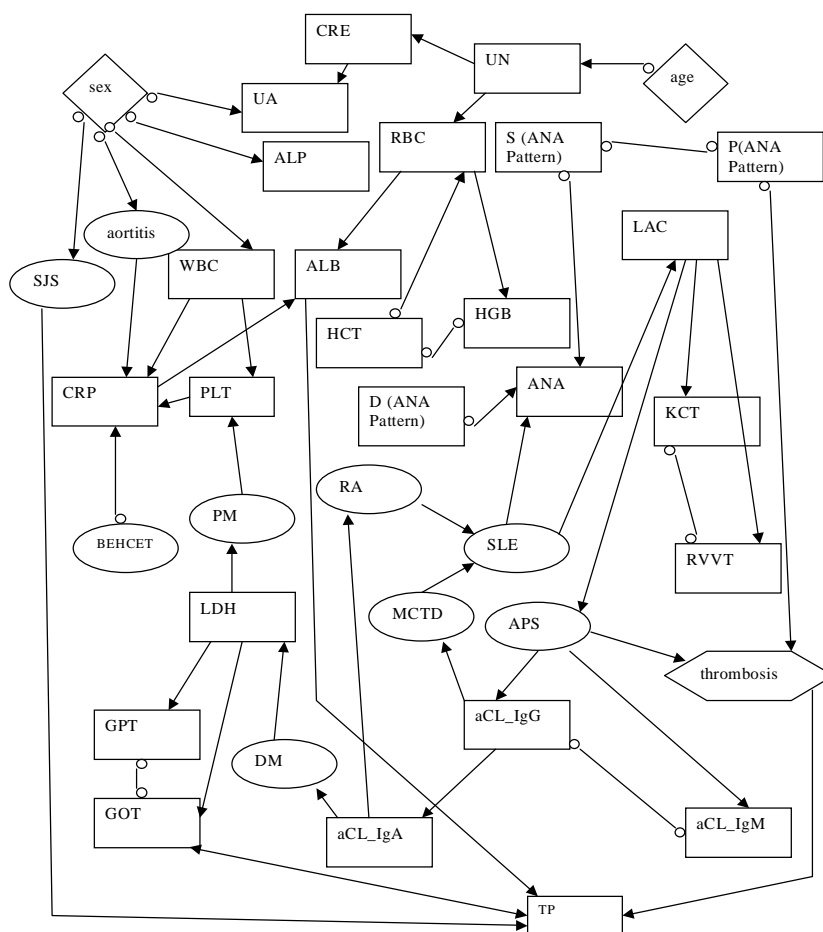
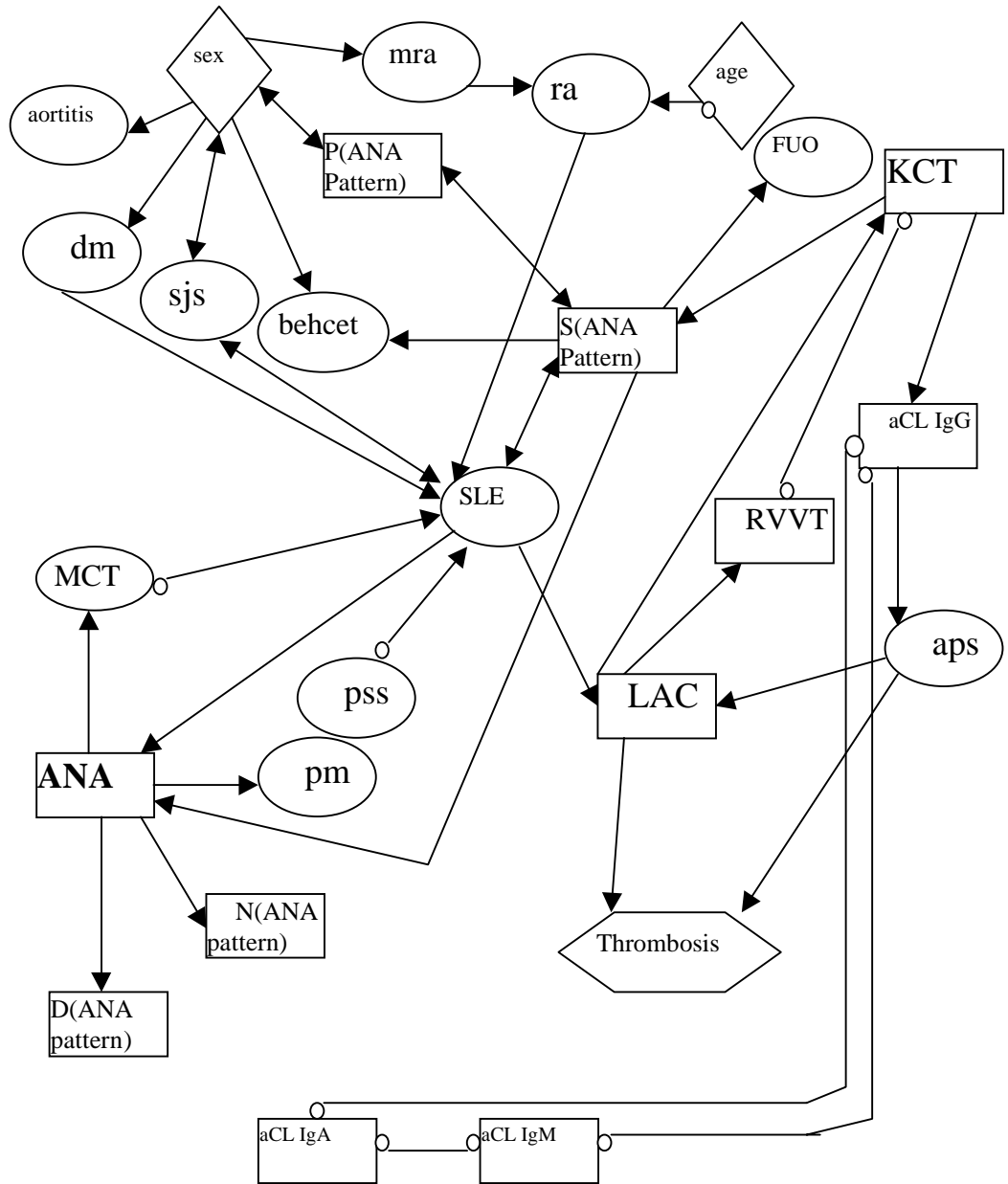


Figure 1. Structural graph (120 patients, tables A, B, and C)



**Figure 2. Structural graph (419 patients from tables A and B)**

Note that the influence of a measured variable on another can be propagated along a path in the graph. Tetrad allows some background knowledge to be incorporated in a number of ways. In the set of results described here we used temporal constraints to avoid having gender and age be caused by other variable. The

variables that temporally precede all others (i.e. cannot be caused by them) are shown as diamonds in the figures. For convenience, we distinguish in the graphs between test results and medical conditions such as diagnoses by using a box for a node corresponding to a test and an ellipse for a node corresponding to a medical condition.

It is apparent from figures 1 and 2 that some causal patterns remain unchanged but others changed as the attributes and the sample size change. For example, both graphs imply that APS is a cause for thrombosis. However, the causal arc connecting APS and aCL IgG got reversed. Moreover, in some situations, the program could not find a consistent orientation to satisfy the causal patterns deduced from the data. This can be a limitation as Tetrad generates acyclic directed graphs and some causal patterns are not acyclic. It is worth noting that APS, identified by Tetrad as a cause for thrombosis, is also chosen as the root of a decision tree generated by C5.0.

#### 4 Temporal Analysis

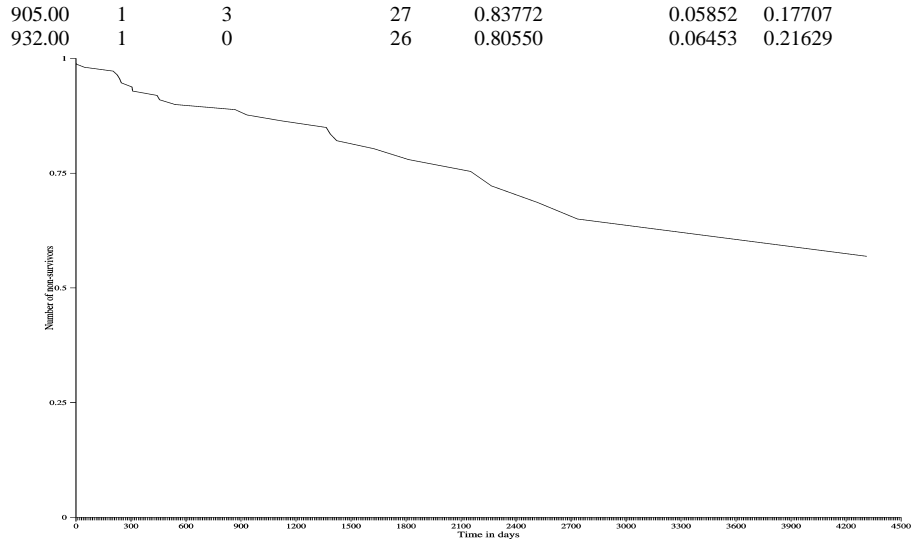
The temporal data in Table C is rather sparse and weakly related to thrombosis and we have not been able to extract useful temporal patterns showing signs or immediate consequences of thrombosis. Instead, we considered the time between infection with an auto-immune disease and developing thrombosis. This study uses survival analysis (or event history analysis) [1] to represent the evolution over time or the rate of occurrence of thrombosis. Initially at the time of diagnosis all the patients diagnosed with an auto-immune disease are considered at risk. As time evolves some patients develop thrombosis, some are known not to have developed (i.e. negative thrombosis results), and for the remaining patients, we do not know (censored).

This temporal analysis has shown that the rate of occurrence of type 1 thrombosis is rather constant over time. Figure 3 shows the survival function for this type. Type 2 results are inclusive and rather problematic. Type 3 thrombosis seems to be rare during the first few years after thrombosis but increases after seven years.

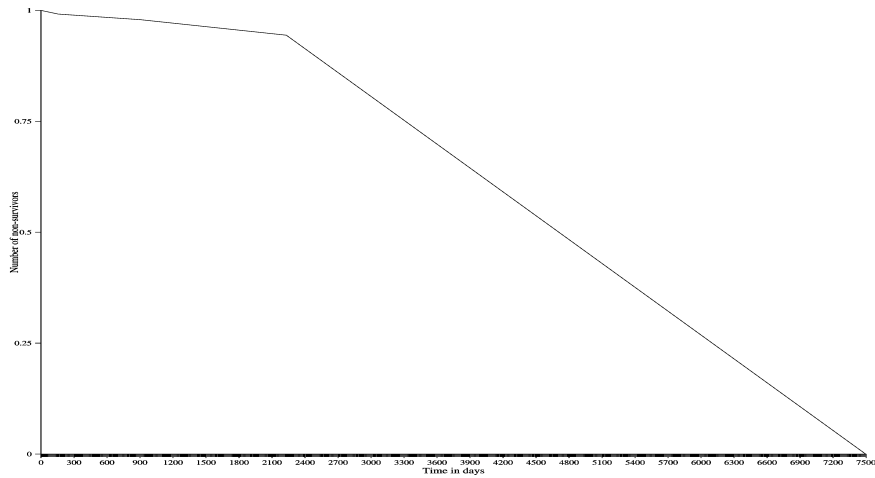
Another set of temporal analysis tried to find a relationship between the rate of occurrence of thrombosis of any type following the diagnosis of each auto-immune disease. For many diseases, the number of patients who developed thrombosis was so small that it was not possible to extract a pattern.

**Table 1. Time to thrombosis for SJS patients**

Time	Events	Number Censored	Number Exposed to Risk	Survivor Function	Std. Error	Cum. Rate
0.00	0	0	113	1.00000	0.00000	0.00000
1.00	3	0	113	0.97345	0.01512	0.02691
46.00	1	65	45	0.95182	0.02600	0.04938
238.00	1	8	36	0.92538	0.03632	0.07755
247.00	1	0	35	0.89894	0.04386	0.10654
570.00	1	3	31	0.86994	0.05114	0.13933



**Fig. 3. Survival curve for type 1 thrombosis (Number of patients who survived versus time in days)**



**Fig. 4. Survival curve for type 3 thrombosis (Number of patients who survived versus time in days)**

Table 1 is a sample of the survival data obtained for the risk of thrombosis over time following a particular diagnosis. It is clear that the rate of occurrence increases gradually as well as the standard error as the population decreases over time due to censoring and the occurrence of the event. Note that the survivor function in the

table represents a cumulative probability distribution over time not the rate of occurrence of the event usually referred to as the hazard rate.

## 5 Conclusions

Structure equation models can be a useful tool for exploratory data analysis and data mining. Here, we used Tetrad to perform a causal analysis of a medical data set. The results presented show that using such tool may be useful. Tetrad assumes linear normally distributed data. Such assumptions do not necessarily hold here. Moreover, as the data lacks information about interventions on the date of thrombosis, we are concerned that results may be reflective of some interventions rather than the normal progression of the condition. In addition to the causal analysis, this work presented some initial finding in a temporal data analysis of this data set. Both the causal and temporal analysis would have benefited from a larger data set.

## References

1. Allison, P. (1984). *Event History Analysis*. Sage, Beverly Hills,CA.
2. Cooper, G. and Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, Vol. 9, 308-347.
3. Eells,E.{1991}. *Probabilistic Causality*. Cambridge Studies in Probability, Induction and Decision Theory. Cambridge University Press, Cambridge, MA.
4. Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman, San Mateo, California.
5. Sprites, P., Glymour,C. and Schienes, R. (1993) Causation, Prediction, and Search. Springer- Verlag Lecture Notes in Statistics 81, Springer-Verlag, NY.
6. Suppes, P. (1970) A probabilistic Theory of Causality, North-Holland, Amsterdam.