Statistical Validation of TCM Syndrome Postulates in the Context of Patients with Cardiovascular Disease

Zhaoxia Xu, MD,1 Nevin L. Zhang, PhD,2 Yiqin Wang, MD,1 Guoping Liu, MD,1 Jin Xu, MD,1 Tengfei Liu, BS,2 and April Hua Liu, BS2

Abstract

Objective: Traditional Chinese Medicine (TCM) has many postulates that explain the occurrence and co-occurrence of symptoms using syndrome factors such as yang deficiency and yin deficiency. A fundamental question is whether the syndrome factors have verifiable scientific content or are purely subjective notions. This analysis investigated the issue in the context of patients with cardiovascular disease (CVD).

Design: In the past, researchers have tried to show that TCM syndrome factors correspond to real entities by means of laboratory tests, with little success. An alternative approach, called latent tree analysis, has recently been proposed. The idea is to discover latent variables behind symptom variables by analyzing symptom data and comparing them with TCM syndrome factors. If there is a good match, then statistical evidence supports the validity of the relevant TCM postulates. This study used latent tree analysis.

Setting: TCM symptom data of 3021 patients with CVD were collected from the cardiology departments of four hospitals in Shanghai, China, between January 2008 and June 2010.

Results: Latent tree analysis of the data yielded a model with 34 latent variables. Many of them correspond to TCM syndrome factors.

Conclusions: The results provide statistical evidence for the validity of TCM postulates in the context of patients with CVD; in other words, they show that TCM postulates are applicable to such patients. This finding is important because it is a precondition for the TCM treatment of those patients.

Introduction

Traditional Chinese Medicine (TCM) diagnosis starts with an overall observation of symptoms (including signs) using four diagnostic methods: inspection, listening, inquiry, and palpation. On the basis of the information collected, patients are classified into various categories that are collectively known as zheng.1 The Chinese term zheng is usually translated as "TCM syndrome." The process of classifying patients into various syndrome classes is known as syndrome differentiation.

TCM syndrome classes such as yang deficiency and yin deficiency are TCM postulates used to explain the occurrence and co-occurrence of signs and symptoms. For example, TCM asserts that yang qi and yin fluid are essential materials of human body and have the functions, among others, of warming and nourishing the body, respectively. Deficiency of yang qi can lead to cold manifestations, such as fear of cold and cold limbs. Hence, patients with those symptoms are often classified in the yang deficiency class. Similarly, deficiency of yin fluid may lead to deficient fire symptoms, such as dry mouth and throat and heat in the palms and soles. Thus, patients with those symptoms are often classified in the yin deficiency class.

Western medicine divides patients into various classes according to disease types or subtypes and treats them accordingly. In contrast, TCM divides patients into various classes according to syndrome types and treats them accordingly. Syndrome-oriented treatment, rather than disease-oriented treatment, is regarded as the key characteristic and strength of TCM.

Two fundamental questions are often asked of TCM syndrome classes: (1) Do they correspond to real-world entities or are they pure subjective notions? (2) Is TCM syndrome differentiation a completely subjective matter, or can it be based on evidence? For more than half a century, researchers have been seeking answers to those questions through laboratory tests. However, the questions still remain unanswered.

1Shanghai University of Traditional Chinese Medicine, Shanghai, China.
2The Hong Kong University of Science and Technology, Hong Kong, China.
Zhang et al. recently proposed a different approach. They distinguish between two kinds of variables in TCM. Symptoms such as fear of cold and dry mouth and throat can be directly observed clinically and hence are called observed variables. Syndrome factors such as yang deficiency and yin deficiency, on the other hand, cannot be directly observed and must be indirectly determined on the basis of symptoms. Hence they are latent factors.

Zhang et al. conjecture that specific syndrome notions, such as yang deficiency and yin deficiency, originated from regularities about symptom occurrence and co-occurrence that were observed in ancient times. They propose a new approach to TCM syndrome research wherein researchers (1) collect data about the occurrence of symptoms on patients while excluding the diagnostic judgments by doctors and (2) re-extract, from the unlabelled data collected, the latent factors postulated in TCM. Diagnosis results are not collected in the first step because the very purpose of the method is to provide objective evidence for TCM diagnosis. The second step is done by using a new class of probabilistic models, called latent tree models, that they have developed specifically for TCM syndrome research. The approach is known as latent tree analysis (LTA).

Zhang et al. have tested the LTA method on a kidney deficiency data set. The latent variables they discovered match the relevant TCM latent factors well. (Note that in this article the term 'latent factors' refers to unobserved factors in TCM, and 'latent variables' refers to unobserved variables in statistical models.) This provides statistical validation to the relevant TCM postulates. Although Zhang et al. have not proved that TCM syndrome classes correspond to real entities, they have shown that postulating the existence of TCM syndromes would explain the regularities in symptom occurrence observed in the data. It is a breakthrough.

The current analysis used LTA to study a data set of 3021 patients with cardiovascular disease (CVD). The data were collected in Shanghai and hence are referred to as the Shanghai CVD data set. The latent variables discovered herein also match TCM latent factors well. This provides evidence for the validity of the relevant TCM postulates in the context of patients with CVD and helps justify dividing these patients into TCM syndrome classes.

Research Method

This section briefly reviews the LTA method, explains how and in what sense it can statistically validate postulates about latent factors, and describes data collection and data preparation for this work.

Review of LTA

LTA refers to the analysis of data using latent tree models (LTMs). An example LTM is shown in Figure 1A. It asserts that a student’s math grade (MG) and science grade (SG) are influenced by the student’s analytical skill (AS), that the English grade (EG) and history grade (HG) are influenced by literal skill (LS), and that the two skills are correlated. Here, the grades are observed variables and the skills are latent variables.

For simplicity, assume all the variables have two possible values: low and high. The dependence of math grade on analytical skill is characterized by the conditional distribution \( P(MG|AS) \) (Fig. 1). It says that a student with high AS tends to get high MGs and a student with low AS tends to get low MGs. Similarly, the dependence of other grade variables on the skill variables are characterized by \( P(SG|AS) \), \( P(EG|LS) \), and \( P(HG|LS) \), respectively (not shown). The quantitative relationships between AS and LS are described by the distributions \( P(AS) \) and \( P(LS) \). Alternatively, they might also be described by \( P(AS|LS) \) and \( P(LS|AS) \).

In Figure 1, correlation strength between variables is visually shown as the edge (line) width. For example, the dependence of MG on AS is stronger than that of SG on AS, and the dependence of EG on LS is stronger than that of HG on LS. Technically, the width of an edge represents the mutual information between the two variables that it connects. The mutual information is computed from the probability distributions of the model.

The input to LTA is a table in which each column corresponds to an observed variable and each row consists of the values of the observed variables for an individual. It does not contain values for latent variables. Many different LTMs can be constructed for the observed variables that appear in the data. A model selection criterion is used to pick one of the models as the output. LTA uses the Bayes information criterion for this purpose. The Bayes information criterion score consists of two terms: a likelihood term and a penalty term. The likelihood term requires that the model fits the data as closely as possible, and the penalty term ensures that the model is not overly complicated.

There usually are too many possible LTMs to enumerate them exhaustively. An algorithm called expand-adjust-simplify-termination (EAST) is used to deal with this computational difficulty. (An implementation of EAST is available at http://www.cse.ust.hk/faculty/lzhang/ltm/index.htm.) It has empirically been shown to be efficient enough to handle data with up to 100 observed variables and is able to find high-quality models. It is the state-of-the-art algorithm for learning LTMs from data.

Now assume that, with respect to a student population, statistical validation for the following postulates is desired:

1. MG and SG are influenced by the latent factor AS, and
2. EG and HG are influenced by the latent factor LS.

The first step would be to sample a subset of students and survey their grades on the four subjects. The next step would be to perform LTA on the survey data. Suppose, in the data, that high MG is frequently accompanied by high SG, while high EG is frequently accompanied by high HG. Further suppose that the correlation between the two groups (MG and SG, EG and HG) is not as strong as correlations between the group members. Then LTA is likely to yield the model shown in Figure 1B. If this turns out to be the case, we draw this conclusion: It fits to the data to hypothesize that a latent factor influences MG and SG, and another latent factor influences EG and HG. In this sense, the LTA provides statistical evidence that supports the two postulates. Although it is unproven that AS and LS correspond to real entities, postulating the existence of AS and LS would explain the correlations among four grade variables well.

Data collection and preparation

The Shanghai CVD data were collected between January 2008 and June 2010. The participants were hospitalized patients.
from the cardiology departments of four hospitals in Shanghai. The data set is not about one particular disease. Rather, it involves multiple heart diseases. Patients with concurrent severe diseases affecting other organs were excluded, as were those who had communication difficulties or incomplete medical records and those who declined to cooperate.

The symptom variables used in the survey are those that a TCM doctor would consider when treating patients with CVD. Each data collection team was led by a person with the title of attending physician or higher. Members of the team were trained beforehand to enforce consistency in judging the presence or absence of symptoms.

The data set contains, for each patient, information collected by using all four diagnostic methods. For simplicity, only 81 inquiry symptom variables are used in this analysis. The data set does not contain diagnosis conclusions. Thus, presented here is unsupervised analysis rather than supervised analysis.

Results

The data set was analyzed by using the EAST algorithm. The resulting latent tree model is referred to as the Shanghai CVD model (Fig. 2). In the model, the nodes labeled with English phrases represent symptom variables. Each of them has two possible values, representing the presence or absence of the symptom. The symptom variables come from the data set. The nodes labeled with the capital letter Y and integer subscripts are the latent variables. They are not from the data set; rather, they were introduced during data analysis to explain regularities in the data.

As explained in the previous section, each edge in the model is associated with a conditional distribution. The widths of the edges indicate strengths of correlation between variables. They are computed from the probability distributions. This report focuses on the links between variables and the strength of those links. The conditional probability distributions contain quantitative information that can be used as evidence for syndrome differentiation. This will be discussed in future work.

In a latent tree model, the collection of observed variables connected to a particular latent variable is called a sibling cluster. The sibling cluster, together with the latent variable, forms a family. In Figure 2, "chest oppression and shortness of breath" makes up a sibling cluster, which forms a family together with Y2. We say that the family is headed by Y2.

To appreciate the Shanghai CVD model, consider why some symptom variables are grouped to form sibling clusters. An examination of the model reveals three cases. First, some symptom variables are grouped into one sibling cluster because they tend to co-occur. One example is chest oppression and shortness of breath in the family headed by Y2. Second, some symptom variables are grouped into one sibling cluster because they are mutually exclusive. One example is white phlegm, yellow phlegm, and grey/black phlegm in the family headed by Y14. The third case is a mixture of the first two cases. One example is in the family headed by Y1, in which the symptoms dull pain, gripping pain, and distending pain are mutually exclusive but all co-occur with chest pain. Among the three cases, the first one is the most common in the Shanghai CVD models.
Note that several symptom variables in the model seem to be out of place: morning diarrhea under Y10, excessive eating-hunger under Y15, irregular sloppy-bound stool under Y18, burning urination under Y26, aching pain/heavy body under Y30, and constipation under Y34. Those symptoms occur rarely in the data and hence there is not sufficient information to determine appropriate locations for them in the model. As a matter of fact, those symptom variables are only weakly related to the latent variables to which they are connected. Those symptoms will be ignored in subsequent discussions.

FIG. 2. The structure of the latent tree model learned from the Shanghai cardiovascular disease data set. Some of the variable names are abbreviated: for example, distending pain chest = distending pain in chest and hypochondrium; worse w emotion disorder = symptoms worsening with emotional disorder; sto...pain like warm pressing = stomach pain easing with warm pressing.
Discussion

The Shanghai CVD model contains abundant evidence that supports relevant TCM syndrome postulates as discussed earlier. For example, TCM asserts that the occurrence of chest oppression and shortness of breath is influenced by one common latent factor called qi disorder (stagnation and deficiency) in chest.\(^3\) In the CVD model, the two symptom variables are connected to the latent variable Y2. This means that it fits the data for the postulate that latent factor influences the two symptom variables. In other words, regularities in the data suggest that a common latent factor should lie behind chest oppression and shortness of breath. Hence, the family headed by Y2 provides evidence supporting the postulate that qi disorder in chest leads to chest oppression and shortness of breath.

Furthermore, although lack of strength is not placed under Y2, it is close to Y2. This is consistent with the postulate that qi deficiency leads to lack of strength. The family headed by Y12 provides evidence to support the postulate that qi deficiency is exacerbated by physical exertion and relieved by rest. Y2 and Y12 represent two aspects of qi deficiency.\(^3\)

The family headed by Y9 provides evidence supporting the postulate that yin deficiency leads to cold manifestations, such as fear of cold and cold limbs. The family headed by Y10 provides evidence to support the postulate that yin deficiency leads to deficient fire in the interior, resulting in manifestations such as spontaneous sweating, tidal fever, heat in the palms and soles, and night sweating.\(^3\)

The family headed by Y11 provides evidence in support of the postulate that liver qi depression leads to frequent sighing and distending pain in chest and hypochondrium. The fact that “worse with emotion disorder” is in close proximity to Y11 is apparently reasonable according to TCM.\(^3\)

The family headed by Y25 provides evidence supporting the postulate that heat harassing the heart spirit leads to chest oppression and shortness of breath. A closer look at the numeric information reveals that the partition has two clusters. Patients in the first cluster have much higher probabilities of exhibiting qi disorder in chest symptoms than patients in the second cluster. Such clusters are natural clusters in the data that were identified by unsupervised data analysis, and they are medically meaningful. As such, they can be used as evidence to establish syndrome differentiation standards. This matter will be discussed in depth in future work.

Conclusions

LTA was performed on the symptom data of 3021 patients with CVD. This article presents the model obtained and explains how to understand and appreciate the qualitative aspect of the model. In particular, this report discusses how and in what sense the data analysis provides evidence in support of TCM syndrome postulates. All such evidence was identified through a systematic examination of the model.

This analysis has shown that according to data, it is reasonable to assume that the occurrence of chest oppression and shortness of breath is influenced by a common latent variable. This supports the TCM postulate that qi disorder (stagnation and deficiency) in chest leads to chest oppression and shortness of breath. In the same manner, this work provides statistical validation to TCM postulates regarding yin deficiency, yin deficiency, liver qi depression, heat harassing the heart spirit, heart fire flaming upward, fluid-humor depletion, kidney qi insecurity, and kidney yin deficiency. It should be noted, however, that this analysis has not shown that the relevant TCM latent factors correspond to real entities.

Acknowledgments

Research on this paper was supported by China National Basic Research 973 program project no. 2011CB505101, Shanghai Third Leading Academic Discipline project no. S30302, China National Natural Science Foundation project no. 81173199, and Guangzhou HKUST Fok Ying Tung Research Institute.

Disclosure Statement

No competing financial interests exist.

References


Address correspondence to:
Nevin L. Zhang
The Hong Kong University of Science & Technology
Clear Water Bay Road
Kowloon, Hong Kong, China
E-mail: lzhang@cse.ust.hk