

Statistical Validation of Traditional Chinese Medicine Theories

NEVIN L. ZHANG, Ph.D.,¹ SHIHONG YUAN, M.S.,² TAO CHEN, Ph.D. (Cand.),¹
and YI WANG, Ph.D. (Cand.),¹

ABSTRACT

The theories of Traditional Chinese Medicine (TCM) originated from experiences doctors had with patients in ancient times. We ask the question whether aspects of TCM theories can be reconstructed through data analysis. To answer the question, we have developed a data analysis method called latent tree models and have used it to analyze several TCM data sets. This paper reports the results we obtained on one of the data sets and explains how they provide statistical validation to the relevant TCM theories.

INTRODUCTION

Traditional Chinese Medicine (TCM) is an important avenue for disease prevention and treatment for ethnic Chinese and is gaining popularity among non-Chinese. However, it suffers a serious credibility problem, especially in the West. One reason is the lack of rigorous randomized trials in support of the efficacy of TCM herb treatments.¹ Another equally important reason, on which this paper focuses, is the lack of validation for TCM theories.

How could TCM theories be possibly validated? In what sense? To answer those questions, we first present a view on the nature of TCM diagnosis and TCM theories.

TCM diagnosis starts with collection of patient information through an overall observation of symptoms rather than micro-level laboratory tests. The conclusion of TCM diagnosis is called *syndrome* and the process of reaching a conclusion from symptoms is called *syndrome differentiation*. There are several syndrome differentiation systems, each focusing on a different perspective of the human body and with its own theory. To understand the nature of those theories, consider the following excerpt from a Chinese–English bilingual textbook:²

Kidney *Yang* is the basis of all *Yang* in the body. When Kidney *Yang* is in deficiency, it cannot warm the body and the patient feels cold, resulting in intolerance to

cold, cold limbs, and cold lumbus and back. Deficiency of Kidney *Yang* also leads to spleen disorders, resulting in loose stool and indigested grain in stool.

Here factors such as Kidney *Yang* failing to warm the body and spleen disorders due to Kidney *Yang* deficiency (KYD) are not directly observed. They are similar in nature to concepts such as “intelligence” and are indirectly measured through their manifestations. Hence we call them “TCM latent factors.” In contrast, symptoms such as “cold limbs” and “loose stool” are directly observed. TCM theories involve a large number of latent factors and symptoms. Abstractly speaking, they describe relationships among the TCM latent factors, and between the latent factors and the symptoms. In this sense, we say that they are “latent structures” and will sometimes refer to them as the “TCM latent structures.”

A conceptually straightforward way to validate TCM theories is to (1) somehow directly measure TCM latent factors, (2) reveal the relationships between them and symptoms, and (3) compare the relationships with those described in TCM theories. Researchers in China have been trying to find ways to directly measure TCM latent factors for more than half a century. Such efforts are known as research on the essence of TCM syndromes (RES). Unfortunately, no breakthrough has been achieved. When reflecting on RES, Zhao³ lamented that “RES has accumulated a lot of experimental data in the past 40 years. Against the original re-

¹Department of Computer Science & Engineering, The Hong Kong University of Science & Technology, Kowloon, Hong Kong, China.

²Department of Chinese Medicine Diagnostics, Beijing University of Traditional Chinese Medicine, Beijing, China.

search objectives, all the data have brought us are confusion and perplexity.”

Is it possible to provide some kind of validation to TCM theories without directly measuring the TCM latent factors? The answer is in theory positive because latent factors, as are well known, induce correlations among observed variables and statements about correlations can be verified or falsified using data. So, one could start from the TCM theories, write down the correlation relationships they imply among observed variables, and check those relationships against data. If the relationships are verified by data, then one would have validated implications of TCM theories and hence would have provided indirect validation to TCM theories themselves.

Unfortunately, this indirect validation approach is difficult to carry out in practice. The reason is that TCM theories are described in vague and qualitative terms. It is not possible to obtain from them clear and precise hypotheses that can be statistically tested.

The basic idea behind the indirect validation approach is to establish a connection between TCM theories and data. It is difficult to do so starting from TCM theories. So we choose to start from the other end, namely, data. Given a data set, we first obtain the statistical latent structure model, and then compare it with the relevant TCM theory. The comparison is possible because, as was pointed out earlier, TCM theories are themselves latent structures described in natural language. If the TCM theory matches aspects of the model well, then we would have shown that it can be viewed as part of the statistical model that best explains the data. In this sense, we would have provided indirect validation to the TCM theory. This idea is schematically shown in Figure 1.

The indirect validation that our method can provide is clearly weaker than the direct validation mentioned earlier, but it might be the best that we could hope for as long as TCM latent factors escape direct measurement.

LATENT TREE MODELS

Latent tree models are the simplest latent structure models. They were previously known as “hierarchical latent class models.”⁴ A “latent tree model” is a tree-structured Bayesian network⁵ where variables at leaf nodes are observed and are hence called “manifest variables,” whereas variables at internal nodes are hidden and hence are called “latent variables.” All variables are assumed to be discrete. Arrows represent direct probabilistic dependence. In the model shown on the left of Figure 2, there is an arrow from variable X_1 to variable X_2 . This means that X_2 depends on X_1 directly. The dependence is characterized by a conditional distribution $P(X_2|X_1)$, which gives a distribution for X_2 for each value of X_1 .

As an illustrative example, consider high school students who take several subjects (e.g., mathematics, science, literature, and history). The grades they obtain on those subjects are manifest variables. The grades are influenced by latent factors Analytic Skill and Literal Skill, which in turn are influenced by general Intelligence. The relationships constitute a latent tree structure, shown on the right of Figure 2.

The question of interest to this paper is how to induce latent tree models from data. This can be divided into two subquestions. First, among all the possible models, which one is the best? This is the model selection problem. Zhang⁴ empirically examined several criteria, namely, the Bayesian

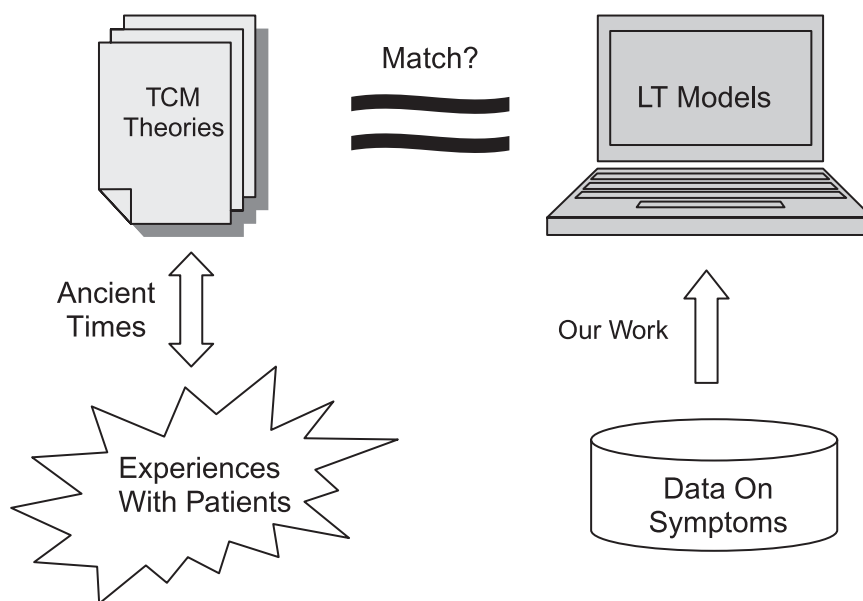


FIG. 1. Schematic illustration of the idea behind the work.

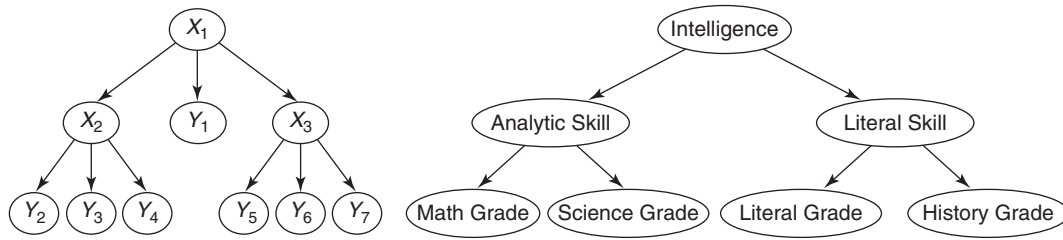


FIG. 2. **Left:** The structure of a latent tree model. The X_i 's are latent variables and the Y_j 's are manifest variables. **Right:** An illustrative latent tree structure.

information criterion (BIC) score,⁶ the Akaike information criterion (AIC) score,⁷ the Cheeseman-Stutz score,⁸ and holdout-likelihood.⁹ The Bayesian information criterion (BIC) score turns out to be the most appropriate for the task. The BIC score of a latent tree model m is given by:

$$BIC(m|\mathbf{D}) = \log P(\mathbf{D}|m, \theta^*) - \frac{d(m)}{2} \log N$$

where \mathbf{D} is the data set, θ^* is the maximum likelihood estimate of the model parameters, $d(m)$ is the number of free parameters, and N is the sample size. Note that this definition of the BIC score is used in the machine learning community, whereas researchers in social sciences usually use its negation.

The second subquestion is how to find the model with the highest BIC score in the space of all possible models. The first algorithm for this task was published in Zhang.⁴ It can handle data sets with only about a half dozen manifest variables. As such, it is a concept-testing algorithm. The second algorithm is called SHC.¹⁰ It can handle about one dozen manifest variables. It is a stepping-stone to a more efficient algorithm called HSHC,¹⁰ which can handle about 50 manifest variables. HSHC has been tested on synthetic data and

real-world data sets from market research¹¹ and from a social survey. It found interesting latent structures in all cases.

ANALYSIS OF A TCM DATA SET

HSHC has also been used to analyze several TCM data sets. In the following, we report the results of one of the data sets. The data set consists of 35 symptom variables and 2600 records. The symptom variables were selected in consultation with the China national standards on clinic terminology of TCM syndromes¹² and some textbooks on TCM diagnosis.² Those variables are the most important factors that a TCM doctor would consider when determining whether a patient has an illness condition called Kidney Deficiency and if so, which subtype. Hence, we refer to the data set as the Kidney data. Each symptom variable has four possible values, namely, “no,” “light,” “medium,” and “severe,” representing four severity levels. An operational standard¹³ was adopted during data collection to enforce uniformity in determining the severity levels of the symptoms.

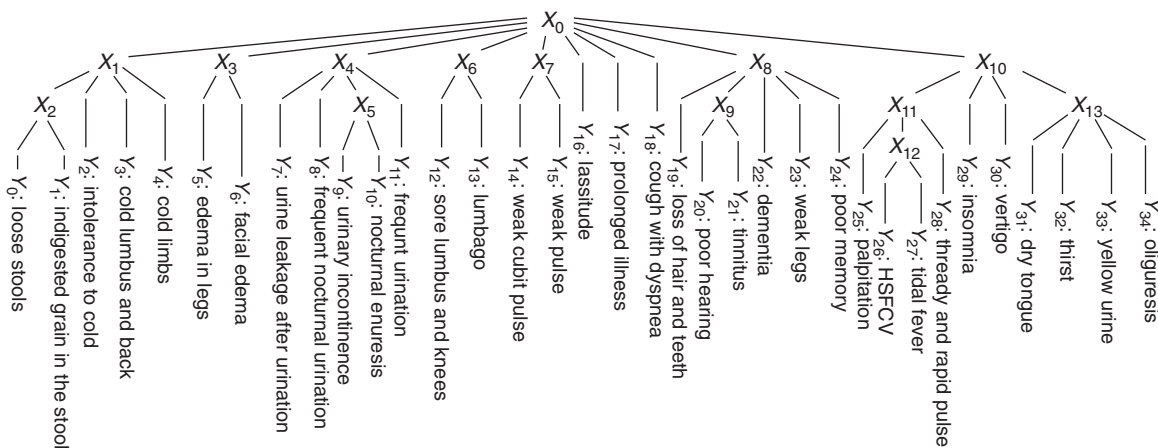


FIG. 3. The structure of the best model M found for Kidney data. The abbreviation HSFCV stands for Hot Sensation in Five Centers with Vexation, where the five centers refer to the centers of two palms, the centers of two feet, and the heart.

The data were collected from communities in several regions in China, and all the subjects were at or above the age of 60 years. Hence, all the conclusions to be drawn are with respect to this population.

We analyzed the Kidney data using the HSHC algorithm. The best model that we obtained is denoted by M. Its BIC score is $-73,860$ and its structure is shown in Figure 3. In the model, Y_0 to Y_{34} are the manifest variables that appear in the data, while X_0 to X_{13} are the latent variables introduced in the process of data analysis. The model also has quantitative contents, which we do not examine in this paper.

COMPARING MODEL M AND TCM THEORY

The relevant TCM theory that explains the occurrence of the 35 symptom variables will be referred to as the TCM Kidney Theory. As mentioned earlier, it is a latent structure described in natural language. In this section, we compare this TCM latent structure with the structure of model M to see whether and how well they match.

According to the semantics of latent tree models, the left-most part of model M states that there is a latent variable X_1 that is (1) directly related to the symptoms intolerance to cold (Y_2), cold lumbus and back (Y_3), and cold limbs (Y_4); and (2) through another latent variable X_2 indirectly related to loose stools (Y_0) and indigested grain in the stool (Y_1). On the other hand, the TCM Kidney Theory asserts that when Kidney *Yang* is in deficiency, it cannot warm the body and the patient feels cold, resulting in manifestations such as cold lumbus and back, intolerance to cold, and cold limbs. Deficiency of Kidney *Yang* also leads to Spleen disorders, resulting in symptoms such as loose stools and indigested grain in the stool.

Here model M and the TCM Kidney Theory both mention the same 5 symptom variables; they both describe how two latent variables are related to those five symptoms; and the relationships described in the 2 cases share the same structure. The only difference is that the latter has named the two latent variables and given an explanation to the relationships, while model M simply states the relationships. Therefore, we have a good match between model M and the TCM Kidney Theory here. The latent variable X_1 can be interpreted as Kidney *Yang* Failing to Warm the Body, whereas X_2 can be interpreted as Spleen Disorders due to KYD.

To the right of X_1 , model M states that there is a latent variable X_3 that is directly related to the symptoms of edema in legs (Y_5) and facial edema (Y_6). On the other hand, the TCM kidney theory asserts that when Kidney *Yang* is in deficiency, it cannot control Water, which overflows to the surface of the face and the legs, resulting in facial edema and edema in legs. Here we see another good match. The latent variable X_3 can be interpreted as Edema due to KYD.

To the right of X_3 , model M states that there is a latent variable X_4 that is (1) directly related to the symptoms of urine leakage after urination (Y_7), frequent nocturnal urination (Y_8), and frequent urination (day) (Y_{11}); and (2) through another latent variable X_5 indirectly related to urinary incontinence (day) (Y_9) and nocturnal enuresis (Y_{10}). On the other hand, the TCM Kidney theory asserts that when Kidney fails to control the Urinary Bladder, one would observe clinical manifestations such as frequent urination, urine leakage after urination, frequent nocturnal urination, and in severe cases urinary incontinence and nocturnal enuresis. Once again, there is a good match between this part of M and the relevant aspect of the TCM Kidney Theory. The latent variable X_4 can be interpreted as Kidney Failing to Control UB, where UB stands for the Urinary Bladder.

According to the TCM Kidney Theory, clinical manifestations of the Kidney Essence Insufficiency Syndrome includes premature baldness, tinnitus, deafness, poor memory, trance, declination of intelligence, fatigue, weakness, and so on. Those match the symptom variables in model M that are located under X_8 fairly well. The clinical manifestations of the Kidney *Yin* Deficiency syndrome includes dry throat, tidal fever or hectic fever, fidgeting, hot sensation in the five centers, insomnia, yellow urine, rapid and thready pulse, and so on. Those match the symptom variables under X_{10} fairly well. The symptoms under X_{11} match those caused by Deficiency Fire, whereas the symptoms under X_{13} match those caused by Kidney Water Deficiency.

Finally, the TCM Kidney Theory asserts that Kidney Deficiency can be caused by prolonged illness and manifests as one or more sub-Kidney Deficiency syndromes such as KYD, Kidney Failing to Control UB, Kidney Essence Deficiency, and Kidney *Yin* Deficiency. Moreover, patients suffering from Kidney Deficiency usually share common symptoms such as lumbago, sore and weak lumbus and knees, and mental and physical fatigue. Those and the structure of M suggest that X_0 should be interpreted as Kidney Deficiency.

There are also aspects of M that do not match TCM theory well. One mismatch is in the scope: model M involves fewer symptom variables than all those TCM associates with Kidney. Another kind of mismatch is more technical. Consider the symptoms insomnia and vertigo. According to the TCM Kidney Theory, they can be caused both by Kidney *Yin* Deficiency and Kidney Essence Insufficiency. However, in M they are directly connected only to X_{10} , and not to X_8 . Such mismatches are caused by one limitation of tree models, namely, a manifest variable can be connected to only one latent variable.

DISCUSSION

It is very interesting that the TCM Kidney Theory matches the structure of model M. If we disregard the pos-

sibility of local maximum, M is the latent tree model that explains the data the best according to the BIC criterion. Because of the match, the TCM Kidney Theory can be viewed as part of M . Consequently, one can view the TCM Kidney Theory as part of the latent tree model that best explains the data. In this sense, we have provided statistical validation to the TCM Kidney Theory.

There is another perspective to looking at this matter. As mentioned in the previous section, TCM asserts that there exists a latent factor that causes the three cold symptoms and influences a second, a latent factor, which in turn causes the two stool symptoms. This assertion matches the leftmost part of model M . Because of the relationship between M and the data, we can say that the assertion is consistent with the data. Similarly, other assertions in the TCM Kidney Theory that match model M are also consistent with the data.

ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers for valuable feedback on an earlier version of the paper. Research on this work was supported by Hong Kong Grants Council Grants 622105 and 622307, and The National Basic Research Program (also known as the 973 Program) under project No. 2003CB517106.

REFERENCES

1. Normile D. The new face of traditional Chinese Medicine. *Science* 2003;299:188–190.
2. Yang W, Meng F, Jiang Y. *Diagnostics of Traditional Chinese Medicine*. Beijing: Academy Press, 1998.
3. Zhao GQ. Perplexity and outlet of TCM syndrome essence's study. *Med Philosophy* 2001;22:9–11.
4. Zhang NL. Hierarchical latent class models for cluster analysis. *J Machine Learning Res* 2004;5:697–723.
5. Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Palo Alto, CA: Morgan Kaufmann, 1988.
6. Schwarz G. Estimating the dimension of a model. *Ann Statistics* 1978;6:461–464.
7. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Contr* 1974;19:716–723.
8. Cheeseman P, Stutz J. Bayesian classification (AutoClass): Theory and results. In: Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, eds. *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press, 1996:153–180.
9. Cowell RG, Dawid AP, Lauritzen SL, Spiegelhalter DJ. *Probabilistic networks and expert systems*. New York: Springer Verlag, 1999.
10. Zhang NL, Kocka T. Efficient learning of hierarchical latent class models. In: *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, Boca Raton, Florida, 2004.
11. Zhang NL. Discovery of latent structures: Experience with the CoIL Challenge 2000 Data Set. In: *Proceedings of the 7th International Conference on Computational Science (ICCS 2007)*. Beijing, May 27–30, 2007.
12. China State Bureau of Technical Supervision. *National standards on clinic terminology of traditional Chinese Medicinal diagnosis and treatment—Syndromes, GB/T 16751.2-1997*. China Standards Press, Beijing, 1997.
13. Yan SL, Zhang LW, Wang MQ, Yuan SH. Operational standards for determining the severity levels of kidney deficiency symptoms. *J Chengdu Univ Chin Med* 2001;24:56–59.

Address reprint requests to:

Nevin L. Zhang, Ph.D.

*Department of Computer Science & Engineering
The Hong Kong University of Science & Technology
Clear Water Bay Road
Kowloon, Hong Kong, China*

E-mail: lzhang@cse.ust.hk