Statistical Validation of TCM Syndrome Postulates in the Context of Depressive Patients

Yan Zhao, Nevin L. Zhang, Tianfang Wang, Qingguo Wang, and Tengfei Liu

Abstract

Objective: Traditional Chinese medicine (TCM) has many postulates that explain the 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 contents in them or they are pure subjective notions. We investigate the issue in the context of depressive patients.

Design: In the past, researchers have tried to show that TCM syndrome factors correspond to real entities by means of laboratory tests, but there has been little success. An alternative approach called latent tree analysis has recently been proposed. The idea is to discover latent variables behind unlabeled symptom data based on statistical principles and compare them with TCM syndrome factors. If there is good match, then one obtains statistical evidence in support of the validity of the relevant TCM postulates. We use latent tree analysis in our investigation.

Setting: TCM symptom data of 604 depressive patients were collected from nine hospitals from several regions of China in 2005–2006.

Results: Latent tree analysis of the data yielded a model with 29 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 depressive patients. In other words, they show that TCM postulates are applicable to depressive patients. This is significant because it is a precondition for the TCM treatment of those patients.

Y. Zhao • T. Wang • Q. Wang
Beijing University of Chinese Medicine, Beijing, People's Republic of China

N.L. Zhang (✉) • T. Liu
The Hong Kong University of Science and Technology, Hong Kong, People's Republic of China
e-mail: lzhang@cse.ust.hk

J. Poon and S.K. Poon (eds.), *Data Analytics for Traditional Chinese Medicine Research*,
DOI 10.1007/978-3-319-03801-8_6, © Springer International Publishing Switzerland 2014

1 Introduction

TCM diagnosis starts with an overall observation of symptoms (including signs) using four diagnostic methods, namely inspection, listening, inquiry, and palpation. Based on the information collected, patients are classified into various categories that are collectively known as *ZHENG* (WHO 2007). 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 come from TCM postulates where they are used to explain the 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 of warming and nourishing the body respectively. Deficiency of YANG QI can lead to the co-occurrence of, among others, 'fear of cold' and 'cold limbs'. Hence patients with those symptoms are often classified into the YANG DEFICIENCY class. Similarly, deficiency of YIN FLUID may lead to the co-occurrence of, among others, 'dry mouth and throat' and 'heat in the palms and soles'. Hence patients with those symptoms are often classified into 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. Do they correspond to real-world entities or are they pure subjective notions? Is TCM syndrome differentiation a completely subjective matter or can it be based on objective evidence? For more than half a century, researchers have been seeking answers to those questions by means of laboratory tests (Wang and Xu 1999; Feng et al. 2004). However, the questions still remain open today (Liang et al. 1998).

A different approach has recently been proposed by Zhang et al. (2008a, b). 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 types such as YANG DEFICIENCY and YIN DEFICIENCY, on the other hand, cannot be directly observed and must be indirectly determined based on symptoms. Hence they are called *latent factors*.

Zhang et al. conjecture that specific syndrome notions such as YANG DEFICIENCY and YIN DEFICIENCY originated from patterns of symptom co-occurrence observed in clinic practice. They propose a new approach to TCM syndrome research where a researcher: (1) Collects data about the occurrence of symptoms on patients while excluding the diagnostic judgments by doctors, and (2) tries to, from the unlabelled data collected, re-extract 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 using a new class of probabilistic models called latent tree models that they have developed specifically for TCM syndrome research (Zhang et al. 2008b). As such the approach is known as latent tree analysis.

Zhang et al. (2008a, b) have tested the latent tree analysis method on a KYDNEY DEFICIENCY data set. The latent variables they discovered do match the relevant TCM latent factors well.[1] This provides statistical validation to the relevant TCM postulates and is a breakthrough. Although they have not proved that TCM syndrome classes correspond to real entities, their results have confirmed that the symptom co-occurrence patterns implied by TCM syndromes do exist in data. It is a breakthrough. A similar study has recently been carried out on patients of cardiovascular disease (Xu et al. 2013).

In this work we use latent tree analysis to study a data set of 604 depressive patients, henceforth referred to as the depression data set. The latent variables we discovered also match TCM latent factors well. This provides evidence for the validity of the relevant TCM postulates in the context depressive patients, and consequently offers justifications for dividing depressive patients according into TCM syndrome classes.

2 Methods

2.1 Data Collection

The data were collected in 2005–2006. The subjects were inpatients or outpatients aged between 19 and 69 from nine hospitals from several regions of China. They were selected using the Chinese classification of mental disorder clinic guideline CCMD-3 (Chen 2002). CCMD-3 is similar in structure and categorization to the ICD and DSM manuals, though includes some variations on their main diagnoses and around 40 culturally related diagnoses.

Excluded from the study were subjects who took anti-depression drugs within 2 weeks prior to the survey, women in the gestational and suckling periods, patients suffering from other mental disorders such as mania, and those suffering from other severe diseases or having had operations recently.

The symptoms (and signs) were extracted from the TCM literature on depression between 1994 and 2004. We searched with the phrase “抑郁 and 证” (Depression and ZHENG) on the CNKI (China National Knowledge Infrastructure) database. Among the articles returned by CNKI, we kept only those on studies where patients were selected using the ICD-9, ICD-10, CCMD-2, or CCMD-3 guidelines. This resulted in 65 articles and they contain totally 198 distinct symptoms. The symptoms that appear only one time or two times were removed. We finally ended up with 143 symptoms.

[1] Note that we use ‘latent factors’ to refer to unobserved factors in TCM, and ‘latent variables’ to refer to unobserved variables in statistical models.

An epidemiologic survey was conducted on the 143 symptoms. Six hundred and four patient cases were collected. Each patient case contains information about which symptoms occurred on the patient and which ones did not. Various measures were taken to ensure data quality. Examples include staff training, site visit by principal investigators, and dual data entry.

In the 604 patient cases, 57 symptoms occur fewer than 10 times. They were removed from the data set and the remaining 86 symptoms were included in further analysis.

2.2 Latent Tree Analysis

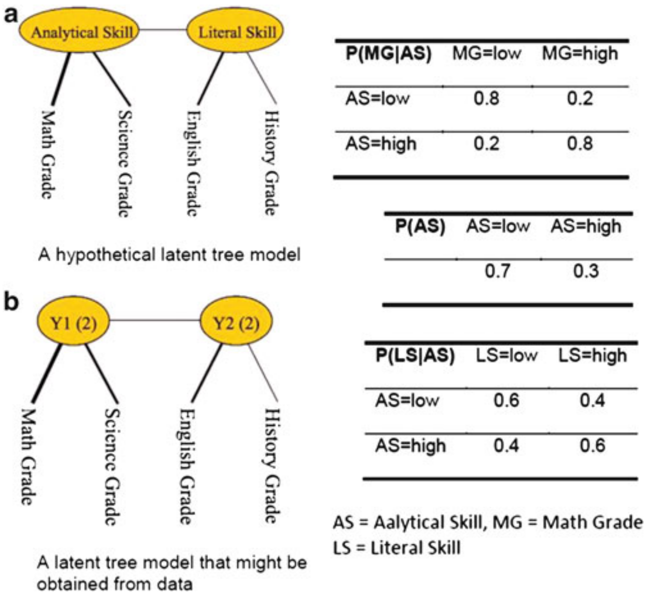
The data were analyzed using latent tree analysis. In the following, we first briefly review latent tree analysis and explain how and in what sense it can provide statistical validation to postulates about latent factors.

Latent tree analysis refers to the analysis of data using latent tree models. An example latent tree model is shown in Fig. 1a. It asserts that a student's Math grade (MG) and Science grade (SG) are influenced by his analytical skill (AS); his English grade (EG) and History grade (HG) are influenced by his literal skill (LS); and the two skills are correlated. Here, the grades are *observed variables*, while the skills are *latent variables*.

For simplicity, assume all the variables have two possible values 'low' and 'high'. The dependence of MG on AS is characterized by the conditional distribution P(MG|AS), which is also shown in Fig. 1. It says that a student with high AS tends to get high MG and a student with low AS is tends to get low MG. 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. They are not shown to save space. The quantitative relationships between AS and LS are described by the distributions P(AS) and (LS|AS). Alternatively, they might also be described by P(LS) and P(AS|LS).

In Fig. 1, correlation strength between variables is visually shown as 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 latent tree analysis is a table where each column represents 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 latent tree models 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. Latent tree analysis uses the Bayes information criterion (BIC) (Schwarz 1978) for this purpose. The BIC 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, while the penalty term ensures that the model is not overly complicated.



P(MG|AS)	MG=low	MG=high
AS=low	0.8	0.2
AS=high	0.2	0.8

P(AS)	AS=low	AS=high
	0.7	0.3

P(LS|AS)	LS=low	LS=high
AS=low	0.6	0.4
AS=high	0.4	0.6

AS = Aalytical Skill, MG = Math Grade
LS = Literal Skill

Fig. 1 The first subfigure and the tables illustrate the concept of latent tree models using an example that involves two latent variables (the skill variables) and four observed variables (the grade variables). The second subfigure shows a model that might be obtained from data on the four observed variables. The numbers next to the latent variables *Y1* and *Y2* indicate that they both have two possible values

There usually are too many possible latent tree models to enumerate exhaustively. An algorithm called Expand-Adjust-Simplify-Termination (EAST) (Chen et al. 2011) is used to deal with this computational difficulty.[2] 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.

Now assume that we want to provide, with respect to a student population, statistical validation for the following postulates:

(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 latent tree analysis on the survey data. Suppose, in the data, 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, SG} and {EG, HG} is not as strong as those between the group members. Then latent tree analysis is likely to yield the model shown in Fig. 1b. If this turns out to be the case, we draw this conclusion: It fits to the data to

[2] A Java implementation of EAST that is available at http://www.cse.ust.hk/~lzhang/ltm/index.htm.

hypothesize that there is latent factor which influences MG and SG, and there is another latent factor that influences EG and HG. In this sense, we have provided statistical evidence that supports the two postulates. Although we have not proved that AS and LS correspond real entities, we have shown that postulating the existence of AS and LS would explain the correlations among four grade variables well.

3 Results

The result of the analysis is a latent tree model, which will be referred to as the depression model. The structure of the model is shown in 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 symptoms. 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 patterns in the data. There is an integer next to each latent variable. It is the number of possible values of that latent variable.

The edges in the model represent probabilistic dependence. Each edge is characterized by a conditional probability distribution. The widths of the edges denote the strength of correlations between variables. For example, Y29 is strongly correlated with 'cold limbs', moderately correlated with 'fear of cold', and weakly correlated with 'rough pulse'. In this paper, we will mainly focus 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. We will discuss them in future work.

4 Discussions

4.1 Latent Variables as Evidence for TCM Postulates

In a latent tree model, the collection of observed variables directly connected to a particular latent variable is called a *sibling cluster*. The sibling cluster together with the latent variable forms a *family*. For example, the three symptom variables under Y28 'aching lumbus', 'lumbar pain liking pressure' and 'lumbar pain liking warmth' make up a sibling cluster. Together with Y28, they form a *family*, which is said to be *headed* by Y28.

Why some symptom variables are grouped to form sibling clusters during latent tree analysis? Why are latent variables introduced? An examination of the model (both qualitative and quantitative information) reveals that there are three cases with regard to this question. First, some symptom variables are grouped into one sibling cluster because they tend to co-occur. One example is 'baking heat' and 'heat in

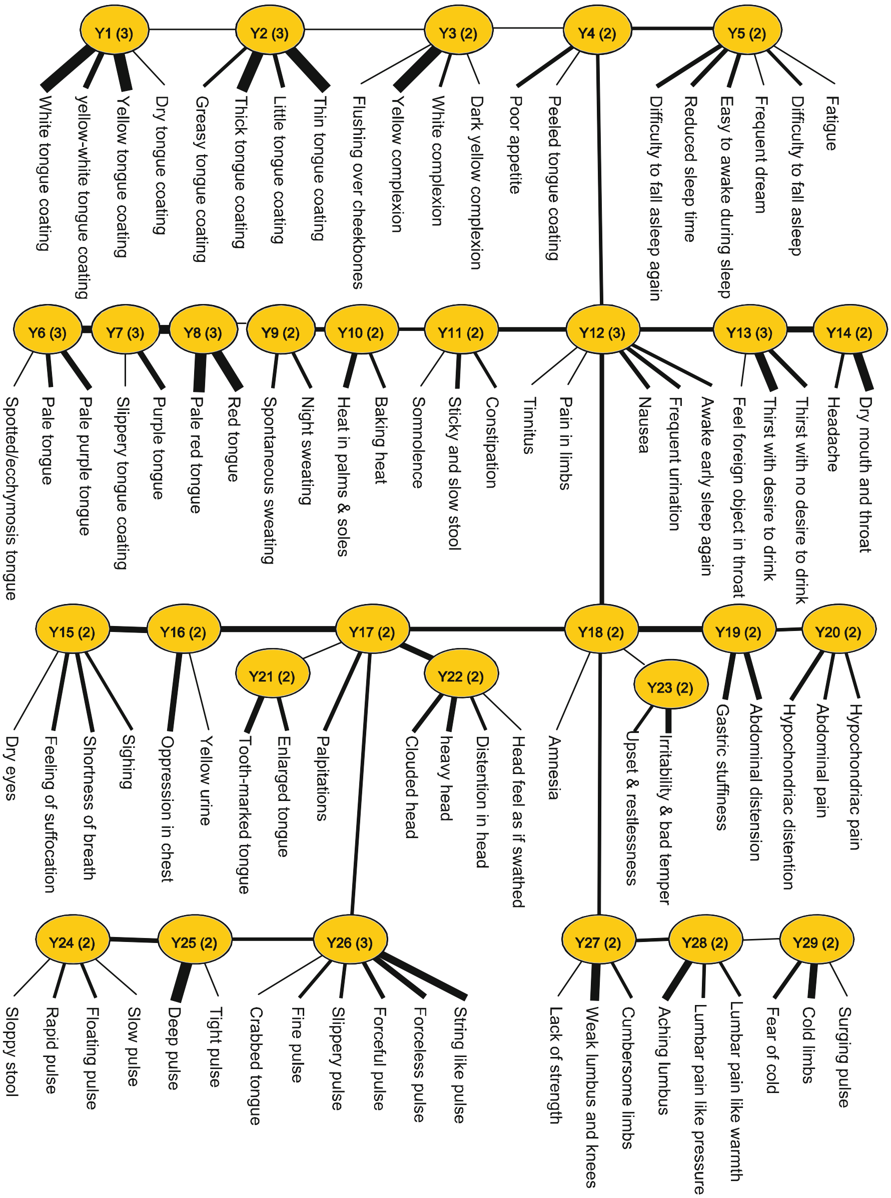


Fig. 2 The structure of the latent tree model learned from the depression data set

palms & soles' in the family headed by Y10. In this case, the latent variable is introduced to explain the co-occurrence of the symptoms. Second, some symptom variables are grouped into one sibling cluster because they are mutually exclusive. One example is 'red tongue' and 'pale red tongue' in the family headed by Y8. In this case, the latent variable is introduced to represent a partition of the patients based on those symptoms. The third case is a mixture of the first two cases. One example is in the family headed by Y26, where ‘forceful pulse’ and ‘forceless’ are mutually exclusive, whereas they both co-occur with ‘string like pulse’.

Latent variables of the first case are evidence for the validity of TCM postulates. For example, TCM posits that YIN DEFICIENCY may lead to ‘baking heat’ and ‘heat in palms & soles’. An implication of this postulate is that the two symptoms would tend to co-occur in clinic practice. The introduction of Y10 during latent tree analysis has confirmed that ‘baking heat’ and ‘heat in palms & soles’ indeed tend to co-occur in the data. In other words, it has verified an implication of the postulate. In this sense, it provides support for the TCM postulate.

Note that several symptom variables in the model seem to be out of place. They include ‘somnolence’ under Y11, ‘tinnitus’ and ‘pain in limbs’ under Y12, ‘dry eyes’ under Y15, ‘yellow urine’ under Y16, ‘sloppy stool’ under Y24, and ‘rough pulse’ under Y29. 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 symptoms variables are only weakly related to the latent variables to which they are directly connected. We will ignore those variables in subsequent discussions.

4.2 Evidence for the Validity of TCM Postulates

Having explained how latent tree analysis can provide evidence for the validity of TCM postulates, we now set out to systematically examine the depression model for such evidence.

Starting from the bottom right corner of the model, we notice that ‘fear of cold’ and ‘cold limbs’ are grouped under Y29. This and relevant quantitative information indicate that the two symptoms tend to co-occur in the data. On the other hand, the co-occurrence of the two symptoms is an implication of the postulate that YANG DEFICIENCY leads to ‘fear of cold’ and ‘cold limbs’. So, Y29 has verified the implication and is hence evidence for the validity of the postulate.

The family headed by Y28 indicates the three symptoms ‘aching lumbus’, ‘lumbar pain like pressure’ and ‘lumbar pain like warmth’ tends to co-occur in the data. It is evidence in support of the postulate that KINDNEY DEPRIVED OF NOURISHMENT may lead to ‘aching lumbus’, ‘lumbar pain like pressure’ and ‘lumbar pain like warmth’. The family headed by Y27 indicates that ‘weak lumbus and knees’ and ‘cumbersome limbs’ tends to co-occur in the data. It is evidence in support of the postulate that KINDNEY DEFICIENCY may lead to ‘weak lumbus and knees’ and ‘cumbersome limbs’. The close proximity of Y27 and Y28 to each other is consistent with the postulate that KINDNEY DEPRIVED OF NOURISHMENT and KINDNEY DEFICIENCY are closely related.

The family headed by Y23 indicates that ‘upset and restlessness’ and ‘irritability & bad temper’ tend to co-occur in the data. It is evidence in support of the postulate

STAGNANT QI TRANSFORMING INTO FIRE leads to ‘upset and restlessness’ and ‘irritability & bad temper’. The family headed by Y21 indicates that ‘enlarged tongue’ and ‘tooth-marked tongue’ tend to co-occur in the data. It is evidence in support of the postulate SPLEEN DEFICIENCY AND INTERNAL ACCUMULATION OF EXCESSIVE DAMPNESS leads to ‘enlarged tongue’ and ‘tooth-marked tongue’.

The family headed by Y20 indicates that the three symptoms ‘hypochondriac distention’, ‘hypochondriac pain’ and ‘abdominal pain’ tend to co-occur in the data. It is evidence in support of the postulate that LIVER QI FAILING TO FLOW FREELY leads to ‘hypochondriac distention’, ‘hypochondriac pain’ and ‘abdominal pain’. The family headed by Y19 indicates that ‘gastric stuffiness’ and ‘abdominal distention’ tends to co-occur in the data. It is evidence in support of the postulate that IMPAIRED HARMONIOUS DOWNBEARING OF STOMACH leads ‘gastric stuffiness’ and ‘abdominal distention’. The close proximity of Y20 and Y19 to each other is consistent with the TCM notion of LIVER-STOMACH DISHARMONY.

The latent variables Y15, Y16 and Y17 are strongly correlated with each other. Collectively, they are strongly correlated with ‘feeling of suffocation’, ‘shortness of breath’, ‘sighing’, ‘oppression in chest’ and ‘palpitation’. Those support the postulate that STAGNATION OF QI ACTIVITY IN THE CHEST leads to those five symptoms.

The family headed by Y22 indicates that the three symptoms ‘clouded head’, ‘heavy head’ and ‘distention in head’ tend to co-occur in the data. It is evidence in support of the postulate that QI STAGNATION IN HEAD leads to ‘clouded head’, ‘heavy head’ and ‘distention in head’.

The family headed by Y11 indicates that ‘constipation’ and ‘sticky and slow stool’ tend to co-occur in the data. It is evidence in support of the postulate that SPLEEN DEFICIENCY AND INTERNAL ACCUMULATION OF EXCESSIVE DAMPNESS leads to ‘constipation’ and ‘sticky and slow stool’. The family headed by Y10 indicates that ‘baking heat’ and ‘heat in palms & soles’ tend to co-occur in the data. It is evidence in support of the postulate that YIN DEFICIENCY leads to ‘baking heat’ and ‘heat in palms & soles’. Finally, the family headed by Y9 indicates that ‘spontaneous sweating’ and ‘night sweating’ tend to co-occur in data. It is evidence in support of the postulate that QI-YIN BOTH IN DEFICIENCY can lead to ‘spontaneous sweating’ and ‘night sweating’.

A summary of the foregoing discussion is given in Table 1. We see that latent variables Y23, Y22, Y20, Y17, Y16, Y15 are all related to QI-STAGNATION. They capture different aspects of the TCM syndrome factor. Similarly, Y21 and Y11 capture different aspects of SPLEEN DEFICIENCY AND INTERNAL ACCUMULATION OF EXCESSIVE DAMPNESS.

The other sibling clusters in the model are also clearly meaningful. The variables under Y1 are mostly about the color of tongue coating; The variables under Y2 are mostly about the thickness of tongue coating; The variables under Y3 are mostly about facial complexions; The variables under Y5 are mostly about sleep disorders; The variables under Y6, Y7 and Y8 are mostly about the color of tongue; The variables under Y13 are about thirst; The variables under Y24, Y25 and Y26 are mostly about

Table 1 Summary of evidence for the validity of TCM postulates: The first column shows TCM syndrome factors. The second column shows some of the symptoms that the syndrome factor might bring about according to TCM postulates. An implication of the postulates is that the symptoms tend to co-occur in clinic practice. The third column shows that latent variables that confirm the co-occurrence patterns in data. They are evidence in support of the postulates

TCM syndrome factors	Symptoms	Latent variables
YANG DEFICIENCY	fear of cold, cold limbs	Y29
KINDNEY DEPRIVED OF NOURISHMENT	aching lumbus, lumbar pain like pressure, lumbar pain like warmth	Y28
KINDNEY DEFICIENCY	weak lumbus and knees, cumbersome limbs	Y27
STAGNANT QI TRANSFORMING INTO FIRE	upset and restlessness, irritability & bad temper	Y23
QI STAGNATION IN HEAD	clouded head, heavy head, distention in head	Y22
SPLEEN DEFICIENCY AND INTERNAL ACCUMULATION OF EXCESSIVE DAMPNESS	enlarged tongue, tooth-marked tongue	Y21
LIVER QI FAILING TO FLOW FREELY	hypochondriac distention, hypochondriac pain, abdominal pain	Y20
IMPAIRED HARMONIOUS DOWNBEARING OF STOMACH	gastric stuffiness, abdominal distention	Y19
STAGNATION OF QI ACTIVITY IN THE CHEST	feeling of suffocation, shortness of breath, sighing, oppression in chest, palpitation	Y17, Y16, Y15
SPLEEN DEFICIENCY AND INTERNAL ACCUMULATION OF EXCESSIVE DAMPNESS	constipation, sticky and slow stool	Y11
YIN DEFICIENCY	baking heat, heat in palms & soles	Y10
QI-YIN BOTH IN DEFICIENCY	spontaneous sweating, night sweating	Y9

pulse. Like the latent variables discussed above, those latent variables identify patterns in the data. However, those patterns do not correspond to TCM latent factors. Rather they represent simple partitions of patients. For example, Y1 represents a partition of the patients into three groups: ‘white tongue coating’, ‘yellow tongue coating’ and ‘yellow-white tongue coating’.

5 Conclusions

We have performed latent tree analysis on the symptom data of 604 depressive patients. We present the resulting model and explain how to understand and appreciate the qualitative aspect of the model. In particular we discuss how and in what sense the data analysis provides evidence in support of TCM postulates. We identify all the evidence contained the model in support of the relevant TCM postulates through a systematic examination of the model. This work has shown that TCM postulates are applicable to depressive patients. Consequently, it is justified to classify depressive patients using TCM syndrome factors.

Acknowledgments Research on this paper was supported by China National Basic Research 973 Program under project No. 2011CB505101, 2011CB505105, Guangzhou HKUST Fok Ying Tung Research Institute, Innovative Team Project of Beijing University of Chinese Medicine (2011-CXTD-08), and Research Base Development Project of Beijing University of Chinese Medicine (2011-JDJS-09). We thank the anonymous reviewers for useful comments.

References

Y.F. Chen, Chinese classification of mental disorders (CCMD-3): towards integration in international classification. Psychopathology **35**(2–3), 171–175 (2002)

T. Chen, N.L. Zhang, T.F. Liu, Y. Wang, K.M. Poon, Model-based multidimensional clustering of categorical data. Artif. Intell. **176**(1), 2246–2269 (2011)

Y. Feng, Z. Wu, X. Zhou, Z. Zhou, W. Fan, Knowledge discovery in traditional Chinese medicine: state of the art and perspectives. Artif. Intell. Med. **38**, 219–236 (2004)

M.X. Liang, J. Liu, Z.P. Hong, Y.Y. Xu, *Perplexity of TCM Syndrome Research and Countermeasures* (People's Health Press, Beijing, 1998)

G. Schwarz, Estimating the dimension of model. Ann. Stat. **6**(2), 461–464 (1978)

H.X. Wang, Y.L. Xu, *The Current State and Future of Basic Theoretical Research on Traditional Chinese Medicine* (Military Medical Sciences Press, Beijing, 1999)

WHO, *WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region* (World Health Organization, Manila, 2007)

Z.X. Xu, N.L. Zhang, Y.Q. Wang, G.P. Liu, J. Xu, T.F. Liu, A.H. Liu, Statistical validation of traditional Chinese medicine syndrome postulates in the context of patients with cardiovascular disease. J. Altern. Complement. Med. **18**, 1–6 (2013)

N.L. Zhang, S.H. Yuan, Y. Chen, Y. Wang, Latent tree models and diagnosis in traditional Chinese medicine. Artif. Intell. Med. **42**, 229–245 (2008a)

N.L. Zhang, S.H. Yuan, Y. Chen, Y. Wang, Statistical validation of TCM theories. J. Altern. Complement. Med. **14**(5), 583–587 (2008b)