NAIST-IS-DD1661020

Doctoral Dissertation

Investigation of CCMQ-J system based on multivariate analysis

Akihiro Yamamori

December, 2017

Department of Bioinformatics and Genomics

Graduate School of Information Science

Nara Institute of Science and Technology

Japan

A Doctoral Dissertation

Submitted to the Graduate School of Information Science,

Nara Institute of Science and Technology

In partial fulfillment of the requirements for the degree of

Doctor of Engineering

Thesis Committee;

Professor Shigehiko Kanaya	(Supervisor)
Professor Keiichi Yasumoto	(Co-supervisor)
Associate professor Md. Altaf-Ul-Amin	(Co-supervisor)
Assistant professor Naoaki Ono	(Co-supervisor)

Investigation of CCMQ-J system based on multivariate analysis[1]

Akihiro Yamamori

Abstract

Although Eastern medicine is based on thousands of years of human experience, the subjective nature of the summaries of this experience has led to doubts regarding the validity of Eastern medicine. To understand Eastern medicine objectively, especially the concept of body constitutions, this thesis presents a multivariate analysis of the Japanese version of the Constitution in Chinese Medicine Questionnaire (CCMQ-J). This questionnaire consists of 60 questions that comprise nine subscales (representing nine body constitutions) labeled “gentleness,” “qi deficiency,” “yang deficiency,” “yin deficiency,” “phlegm wetness,” “wet heat,” “blood stasis,” “qi depression,” and “special diathesis.” Each question is answered using a Likert scale that ranges from 1 to 5, corresponding to “never”, “rarely”, “sometimes”, “often”, and “always”, respectively. First, to understand overall picture of research field in Chinese Medicine questionnaire objectively, 5469 abstracts of research paper ware classified based on a text-mining method. The position of the research topic in this thesis in relation to previous research was elucidated and similar research was identified objectively. Second, the relationships between the 60 questions were examined, based on the CCMQ-J scores of 597 respondents. The 60 questions were tentatively classified into 12 clusters, using Ward’s hierarchical clustering method, and the similarity between these clusters and the nine subscales is discussed. Third, using a partial least square model, I found that Body Mass Index (BMI) and age can be estimated based on the scores of the CCMQ-J questions. The correlation coefficient between the real and estimated

values for BMI was 0.81 for male and 0.82 for female, and that for age was 0.82 for male and 0.83 for female. These results indicate that the 60 questions reflect aspects of aging and BMI and that the CCMQ-J could be used as an indicator of body constitution for evaluating an individual's aging process. Finally, a simplified version of the CCMQ-J which consists of 13 questions and was found to have averaged >80% accuracy was developed.

Keywords; *Multivariate statistical analysis, Body constitution, BMI, age, Japanese version of Constitution in Chinese Medicine Questionnaire (CCMQ-J)*

---

[1]Doctorial Dissertation. Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1661020

Acknowledgements

First of all, I wish to express my great gratitude to **Professor Shigehiko Kanaya**, who has been the advisor for the research described in this dissertation. Regardless of his busy schedule, he helped for the accomplishment of this dissertation. I never forget his warm assistance for me.

I would also like to thank **Professor Keiichi Yasumoto** for his valuable suggestions and comments, and reviewing my papers and this dissertation, to **Associate professor Md. Altaf-UI-Amin**, **Assistant professor Naoaki Ono** for their suggestions.

I'm highly thankful to the member of my work place, **Dr. Takeo Kitayama**, **Dr. Tomoyuki Watanabe**, **Dr. Kazunori Ono and Dr. Yukio Tominaga,** who recommended me to obtain doctoral degree and supported me to study big data science while working at company.

Finally, I would like to offer special thanks to my family. **My parents** and **my ground-mother** support school fee and travel cost from Tokyo to Nara. And **Tomomi** encouraged me many time.

I never forget the appreciation to these supporters.

February 2017

Akihiro Yamamori

Contents

List of abbreviations

BMI	Body mass index
CCMQ	Constitution in Chinese Medicine Questionnaire
CCMQ-J	Japanese version of the CCMQ
CMQ	Chinese medicine questionnaires
IDF	Inverse document frequency
LDA	Latent Dirichlet allocation
PLS	Partial least square
RMSEP	Root mean square error prediction
TF	Term frequency

List of figures

List of tables

Chapter 1 Introduction

1.1 Previous studies on Eastern medicine that involved multivariate analysis

Various forms of Eastern medicine, for example jamu (Indonesian traditional medicine), Ayurveda (traditional medicine with historical roots in the Indian subcontinent), and traditional Chinese medicine, are based on summarizing thousands of years of experience of treating disease. However, the method of summarizing these experiences is often not objective. This has led to doubt regarding the validity of Eastern medicine.

Advances in information technology have led to the accumulation of vast amounts of digital data (so-called "big data"), and the data processing speed has increased rapidly. Multivariate analysis involving correlation, clustering, and regression analyses can be used to rapidly and objectively analyze this big data.

Multivariate analysis of the accumulated knowledge and data can allow Eastern medicine (including the clinical diagnostic methods and the medicinal plants prescribed) to be investigated in scientific manner. For example, statistical models of the ratio of natural medicine and functions have been used to investigate jamu [Afendi et al, 2013]. In addition, the "kyo" (emptiness) and "jitsu" (fullness) diagnoses in Kampo (a traditional Japanese therapeutic system) have been modeled by the ratio of natural medicine [Okada et al, 2012].

1.2. Previous studies on the relationship between body constitution and disease

When the natural world is classified hierarchically, the human body can be said to involve a lower level (consisting of the "individual," "organs and tissues," "cellular," and "molecular" aspects) and an upper level (consisting of the "biological," "earth," and "universe" aspects). The individual level is linked to the cellular and molecular aspects. However, an individual has its own rule, and the individual is not fully explained by its link to the lower level [Tada, 2004].

In traditional medicine, it is believed that disease is derived from an individual. Therefore, prescriptions are based on an individual's body constitution and condition. For example, depending on the individual, prescriptions may involve personalized cures, health promotion, lifestyle, and anti-aging methods [Kamibaba, 2004]. The purpose of traditional medicine is not only to cure the individual, but also to maintain and promote good health by using preventive medicine, and to promote anti-aging by activating the body's natural healing ability. Maintaining good health (by taking into account an individual's body constitution and condition) is a critical issue for individual-level health care.

In Western medicine, research about the relationship between individuals' body constitutions and disease was started by Léon Rostan in 1828. He proposed that body constitutions can be classified into three types: "digestive" (thickset and round), "cerebral-respiratory" (thin and elongated), and "muscular" (broad and muscular). In addition, Ernst Kretschmer believed that body constitution was related to mental

health. For example, he believed that individuals who are “pyknic” (fat) are more likely to have manic depression, those who are “asthenic” (thin) are more likely to have schizophrenia, and those who are “athletic” (muscular) are more likely to have epilepsy. Furthermore, in 1947, the relationship between body constitution and mental health was investigated by Rees in a multivariate analysis [Rees, 1947].

In Eastern medicine, disease has historically been associated with body constitution. In India, body constitution is thought to be the result of the inherent balance of the body fluid. In contrast, in China, body constitution is thought to be the result of both inherent and acquired factors. In Eastern medicine, understanding the relationship between body constitution and disease leads to the practice of individualized medicine.

In this study, the Constitution in Chinese Medicine Questionnaire (CCMQ) was analyzed. The aim was to understand overall picture of research in Chinese Medicine Questionnaire objectively and the system (i.e., the questions, subscales, and their relationships with age and BMI) based on respondents’ answerer to 60 questions on their body constitution by using a statistical approach and qualitatively assessing the results.

1.3. Outline of thesis

This thesis is organized as follows:

In Chapter 2, a review of the research on Chinese medicine questionnaires (CMQ) is presented, and the position of the research in this thesis in relation to previous research is elucidated.

Chapter 3 presents the correlations between respondents' answers to the 60 questions in the Japanese version of the CCMQ (CCMQ-J). Subsequently, the questions were grouped based on their correlations with each other. The resultant clusters were compared to the nine subscales of the questionnaire (which were constructed based on the experience that underlies traditional Chinese medicine).

Chapter 4 presents a partial least square (PLS) regression analysis of the relationship between the CCMQ-J scores and two variables, age and body mass index (BMI), in order to investigate whether CCMQ-J scores can be used to estimate these objective variables.

In Chapter 5, a simplified version of CCMQ-J is presented. The construction of the simplified version involved removing some of the questions in the original CCMQ-J based on the results of a machine learning method. It was found that a questionnaire involving 14 questions can be used to accurately judge body constitution.

Finally, Chapter 6 outlines the possibility of applying the methods used in this thesis to other questionnaires mentioned in Chapter 2, and it sets out the concluding remarks.

Chapter 2 Overview of research into Chinese medicine questionnaires

2.1 Overview of the CCMQ

The CCMQ is a traditional method of evaluating body constitution in China that has been adopted as the national standard, and it is frequently used part of a process to promote health and prevent disease. The CCMQ is used to classify individuals according to nine body constitutions known as “gentleness,” “qi deficiency,” “yang deficiency,” “yin deficiency,” “phlegm wetness,” “wet heat,” “blood stasis,” “qi-depression,” and “special diathesis.” The questionnaire consists of 60 questions with nine subscales (corresponding to the nine body constitutions). Gentleness represents health, while the other body constitutions represent a lack of health. Each question is answered using a Likert scale that ranges from 1 to 5, corresponding to “never,” “rarely,” “sometimes,” “often,” and “always,” respectively [Zhu et al, 2006].

2.2 Objectives

There are large amounts of data in multiple types of documents, and particularly in theses and patents. The US National Institutes of Health's PubMed search engine (which primarily accesses the MEDLINE database of references and abstracts on life sciences and biomedical topics) includes approximately 27 million theses, many of which are on medicine. Elsevier's Scopus bibliographic database of academic abstracts and citations includes approximately 57 million theses, including theses on the fields of science, mathematics, engineering, technology, health and medicine, social sciences, and arts and humanities [Elsevier, 2016]. The Thomson Innovation database includes approximately 23 million patents. The abstracts of all these documents can be used for research. However, it is difficult for humans to read thousands of documents and summarize them objectively in a short period, especially regarding documents on unfamiliar fields.

Advances in information technology in the last 10 years have led to a rapid increase in information processing speeds. In addition, text-mining methods have recently enabled researchers to rapidly classify thousands of documents and summarize them objectively.

The position of the research topic in this thesis in relation to previous research was elucidated by reviewing CMQ of literature. Research similar to that presented in this thesis was identified, and it is clear that the analysis technique used in this thesis could be applied to other questionnaires.

2.3 Data

To investigate research involving CMQ, documents accessed via PubMed were used because PubMed provides access to a large number of medical documents. On 20 September 2016, the search string “Chinese medicine questionnaire” was used, and 5,469 abstracts were retrieved. These abstracts also included bibliographic data such as titles and publication years. My abstract of this research [achievement1] is added to these abstracts. The 5,470 abstracts were used for the subsequent analysis.

2.4 Identifying topics in the literature using Latent Dirichlet allocation

2.4.1 Latent Dirichlet allocation

Various text-mining methods can be used to classify documents. A commonly used method for extracting keywords is the term frequency–inverse document frequency (TF–IDF) method, which can be used to create vector representations of documents (each component of the vector corresponds to the TF–IDF value of a particular word). Document similarity can then be assessed using the cosine similarity measure (which is calculated based on the inner product of two vectors from two documents divided by the product of their vector lengths. This method is quite simple, but when one is attempting to extract low-frequency keywords, documents containing these keywords are often missed.

A topic consists of a group of keywords. A topic model is a type of statistical model for identifying the topics in a collection of documents that is used as a text-mining tool. One topic model is probabilistic latent semantic indexing (PLSI). In PLSI, each document has one topic. However, a limitation of this topic model is that one document often involves multiple topics.

Another topic model, Latent Dirichlet allocation (LDA), models each topic as a finite mixture over an underlying set of topics. In the context of text modeling, topic probabilities provide a representation of a document. [Blei et al, 2003; Griffiths and Steyvers, 2004].

2.4.2 Preprocessing

To prevent trivial and overlapping words from being extracted, the abstracts were preprocessed using the following process:

(i) Remove symbols

(ii) Remove punctuation

(iii) Remove digits

(iv) Remove whitespaces

(v) Stem document

(vi) Remove stopwords

Stopwords can be set by researchers with the aid of Snowball [Porter, 2015], a processing language designed for creating stemming algorithms for use in information retrieval. The following words were removed, except for words from Snowball: “can,” “say,” “one,” “way,” “use,” “also,” “howev,” “tell,” “will,” “much,” “need,” “take,” “tend,” “even,” “like,” “particular,” “rather,” “said,” “get,” “well,” “make,” “ask,” “come,” “end,” “first,” “two,” “help,” “often,” “may,” “might,” “see,” “someth,” “thing,” “point,” “post,” “look,” “right,” “now,” “think,” “ve,” “re,” “anoth,” “put,” “set,” “new,” “good,” “want,” “sure,” “kind,” “larg,” “yes,” “day,” “etc,” “quit,” “sinc,” “attempt,” “lack,” “seen,” “awar,” “littl,” “ever,” “moreov,” “though,” “found,” “abl,” “enough,” “far,” “earli,” “away,” “achiev,” “draw,” “last,” “never,” “brief,” “bit,” “entir,” “brief,” “great,” and “lot.”

When this preprocessing was complete, a document term matrix was created, which served as the input for the LDA analysis.

2.4.3 LDA analysis condition

For the LDA analysis, the number of topic was set so that perplexity (which measures how well a probability distribution predicts an observed sample) would be minimized [Blei et al, 2003].

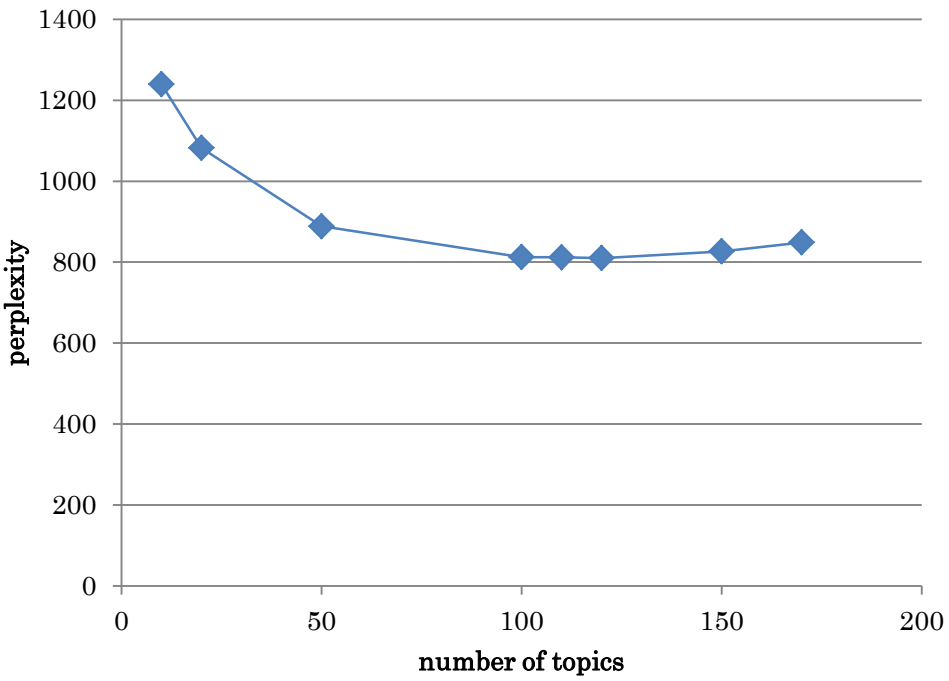


Figure 2.1: Graph of perplexity against the number of topics

Based on the minimization of perplexity, 120 topics were extracted from the abstract data (Figure). The probability of generating each topic in a document was calculated, and the main topic was determined according to the highest topic probability in the document.

2.4.4 Top three topics

The top three topics were selected according to the number documents associated with each of the topics (**Figure 2.2**).

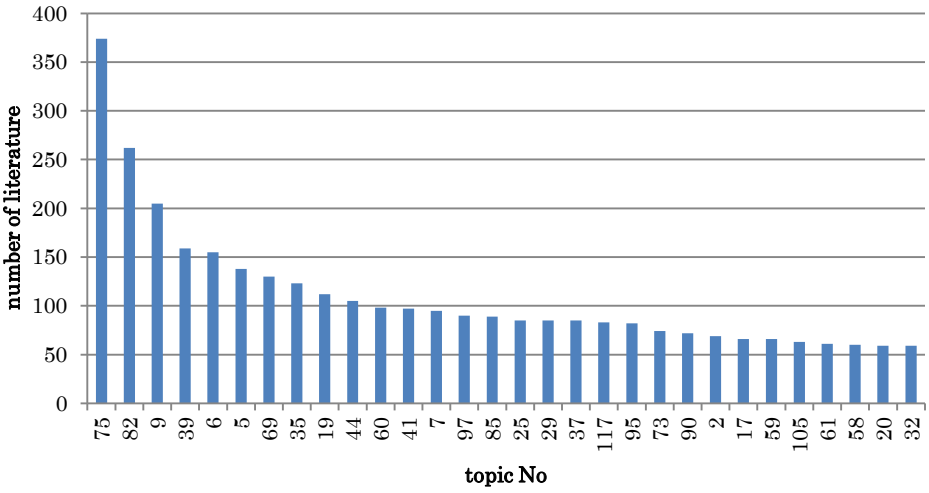


Figure 1.2: Top 30 topics in the literature on Chinese medicine questionnaires

These topics were labeled according to the words that constituted the topics (**Table 2.1**). The top three topics were as follows:

(i) Topic no. 75, which was labeled "statistical analysis of questionnaires" because it included words and word stems such as "correl," "coeffici," and "cronbach."

(ii) Topic no. 82, which was labeled "clinical trial" because it included words such as "clinic," "trial," and "placebo."

(iii) Topic no. 9, which was labeled "nutrition" because it included words and word stems such as "nutri," "food," and "intak."

Table 2.1: Words and word stems included in of the top three topics

Topic 75: Statistical analysis of questionnaires	Topic 82: Clinical trail	Topic 9: Nutrition
Valid	Acupuncture	Intak
Chines	Week	Dietary
Reliabl	Treatment	Food
Scale	Effect	Diet
Item	Random	Pattern
Version	Trial	Consumpt
Correl	Studi	Veget
Test	Control	Frequenc
Factor	Outcom	Soy
Assess	Intervent	Protein
Coeffici	Group	Fruit
Questionnaire	Improv	Chines
Consist	Placebo	Nutrient
Intern	Measur	Total
Measure	Receiv	Energi
Instrument	Clinic	Studi
Analysi	Efficaci	Vitamin
Cronbach	Baselin	Nutrit
Retest	Follow	Meat
Translat	Evalu	Consum

The number of studies on CMQ is increasing; the frequency of the three top topics is also increasing, especially regarding the "statistical analysis of questionnaires" and "clinical trial" topics (**Figure 2.3**). The "statistical analysis of questionnaires" topic is currently the most frequent topic in the CMQ literature, and 6% of all the theses on CMQ include this topic.

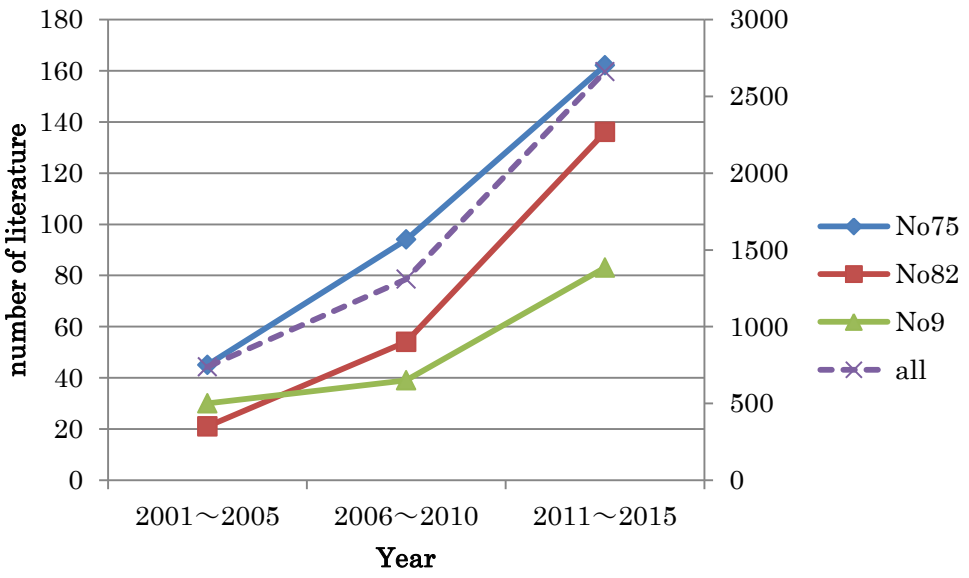


Figure 2.3: Time-course analysis of the number of documents on the top three topics

The research presented in this thesis belongs to the "statistical analysis of questionnaires" category, based on the highest topic probability calculated using LDA.

Regarding the "statistical analysis of questionnaires" topic, other questionnaires that were found during the review of the CMQ literature included questionnaires on the following factors: quality of life [Su et al, 2016], sleep [Thai et al, 2016], chronic pain acceptance [Liu et al, 2016], reactive-proactive aggression [Tuvblad et al, 2016], physical activity among older children [Wang et al, 2016], self-efficacy among patients who had undergone a stroke [Lo et al, 2016], mandibular function impairment [Xu et al, 2016], and eating behavior [Chong et al, 2016].

A variety of questionnaires on Chinese medicine was found. However, many publications described how to apply a non-Chinese questionnaire to Chinese medicine (by translating and validating the non-Chinese questionnaire) rather than investigations of questionnaires (for example, the relationships between the questions in questionnaires). When adapting a questionnaire for use in a new country, revising the questionnaire is important because the respondents in the new country will be different from those in the country in which the questionnaire was originally developed (and, for many questionnaires, validated).

The research presented in this thesis focused on understanding the questionnaire system (i.e., the questions and subscales) by using a statistical approach and qualitatively assessing the results. The methods used in this thesis could be applied to the analysis of other questionnaires.

Chapter 3 Multivariate analysis of CCMQ-J questions and subscales

3.1 Japanese version of the Constitution in Chinese Medicine Questionnaire

The CCMQ-J was developed by Zhu et al., who translated the Chinese version to Japanese and showed that the CCMQ-J was correlated with the 36-Item Short Form Survey (SF-36), a widely used health survey [Zhu et al, 2006].

There are 60 questions in the CCMQ-J (**Table 3.1**), which takes 8±4.2 min to complete [Zhu et al, 2014]. When there are too many questions to answer, respondents may feel inconvenienced, and some may not answer accurately. The 60 questions were devised based on the years of experience that underlie traditional Chinese medicine. To determine whether any questions could be removed from the questionnaire while maintaining its accuracy, the relationships between the questions was investigated.

Table 3.1: Questions in the CCMQ-J

No.	Question Contents
1	Are you active?
2	Do you get tired easily?
3	Do you get out of breath?
4	Do you have heart palpitations?
5	Do you feel dizzy on standing up?
6	Do you prefer quiet, and find it troublesome to have a chat?
7	Do you talk in a quiet voice?
8	Are you forgetful?
9	Do you feel blue or depressed?
10	Do you feel nervous or irritated?
11	Do you feel sentimental and get moved to tears easily?
12	Are you easily upset and uneasy?
13	Do you feel pain under your arms or swellings in your breasts?
14	Do you have a tight chest or a swollen abdomen?
15	Do you often sigh unconsciously?
16	Do you feel tired in your body or arms and legs?
17	Do your palms or soles get hot?
18	Do you have cold hands and feet in summer?
19	Do you feel cold in your back, belly, and knees?
20	Are you sensitive to the cold and do you wear more clothes than others?
21	Does your face or body feel hot?
22	Are you sensitive to cold (cold in the winter, or air-conditioner in summer)?

(Continued)

No.	Question Contents
23	Do you catch colds more easily than others?
24	Do you sneeze even when you don't have a cold?
25	Do you have a stuffy nose even when you don't have a cold?
26	Do you feel that it is stifling at the turn of seasons, due to unstable temperatures, or due to unpleasant odors?
27	Do you sweat without doing anything?
28	Do you have a greasy forehead?
29	Do you have dry skin or lips?
30	Are you sensitive to medicine, foods, pollen, seasons, climate, etc.?
31	Do you often have nettle rash?
32	Do you have spotty or mottled red-purple congestive marks on your skin?
33	Do you get blue bruises without knowing how they occurred?
34	When scratched, does your skin turn red or do marks remain on your skin?
35	Are your lips redder than others'?
36	Do you have broken capillaries on your cheeks?
37	Are you in pain?
38	Do you have red cheeks and feel hot?
39	Do you have an oily, shiny nose?
40	Do you have sallow skin and often get dark spots on your face?
41	Do you often get pimples?
42	Are your upper eyelids easily swollen?
43	Do you often get dark under-eye circles?

(Continued)

No.	Question Contents
44	Do you have dry eyes?
45	Are your lips blue?
46	Do you feel thirsty or have a dry mouth?
47	Do you have an uncomfortable feeling as if something is stuck in your throat?
48	Do you have bad breath or a sour taste in your mouth?
49	Does your mouth feel sticky?
50	Is your tongue thickly covered with plaque?
51	Do you always have a lot of phlegm?
52	Do you feel that your health conditions get worse after having something cold?
53	Can you easily adapt to changes in your surroundings, including your social environment?
54	Do you have trouble sleeping?
55	Do you have diarrhea after having something cold?
56	Do you feel as if your bowels are not completely empty after having a bowel movement because your stools are sticky?
57	Do you often have constipation because your stools are hard?
58	Do you have a fat flabby belly?
59	During urination, do you feel burning in the urethra and have dark urine?
60a	(For women) Is your vaginal discharge yellow?
60b	(For male) Do you feel dampness around your scrota?

A score for each of the nine subscales is calculated using the formulae shown in **Table 3.2** [Zhu et al, 2006]. To investigate the classification of the subscales, a qualitative analysis was necessary.

Table 3.2: Formulae for calculating scores for each body constitution

Body constitution	Characteristic
Gentleness 平和質	Ideal health (good physical and mental health) $\frac{100}{32}(Q_1 + Q_{53} - Q_2 - Q_7 - Q_8 - Q_9 - Q_{22} - Q_{54} + 28)$
Qi deficiency 気虚質	Short of "気" (life force; impaired physical function) $\frac{100}{32}(Q_2 + Q_3 + Q_4 + Q_5 + Q_6 + Q_7 + Q_{23} + Q_{27} - 8)$
Yang deficiency 陽虚質	Short of "陽" (energy; often cold) $\frac{100}{28}(Q_{18} + Q_{19} + Q_{20} + Q_{22} + Q_{23} + Q_{52} + Q_{55} - 7)$
Yin deficiency 陰虚質	Short of water (often thirsty and irritated) $\frac{100}{32}(Q_{17} + Q_{21} + Q_{29} + Q_{35} + Q_{38} + Q_{44} + Q_{46} + Q_{57} - 8)$
Phlegm wetness 痰湿質	Stop of water (accumulated metabolites) $\frac{100}{32}(Q_{14} + Q_{16} + Q_{28} + Q_{42} + Q_{49} + Q_{50} + Q_{51} + Q_{58} - 8)$
Wet heat 湿熱質	Stop of wet and heat (lack of discharge) $\frac{100}{24}(Q_{39} + Q_{41} + Q_{48} + Q_{56} + Q_{59} + Q_{60} - 6)$
Blood stasis 血瘀質	Stop of blood (bad circulation and often stain) $\frac{100}{28}(Q_8 + Q_{33} + Q_{36} + Q_{37} + Q_{40} + Q_{43} + Q_{45} - 7)$
Qi depression 気鬱質	Stop of "気" (often doom-filled) $\frac{100}{28}(Q_9 + Q_{10} + Q_{11} + Q_{12} + Q_{13} + Q_{15} + Q_{47} - 7)$
Special diathesis 特禀質	Sensitive (quick to react external environmental stimulation) $\frac{100}{28}(Q_{24} + Q_{25} + Q_{26} + Q_{30} + Q_{31} + Q_{32} + Q_{34} - 7)$

3.2 Subjects

The relationships between scores based on 597 respondents' answers to the 60 questions were examined.

3.3 Multivariate analysis

The following section describes the multivariate analysis that was carried out in order to scientifically explore the responses to the 60 questions.

3.3.1 Dataset

The explanatory variable in the dataset, which had 60 dimensions is shown in **equation (1).**

$$\boldsymbol{x}_i = \left(x_{i1}, x_{i2}, \ldots, x_{iM}\right) \qquad \textbf{(1)}$$

where x^{ij}, is the score for the j^{th} answer of the i^{th} individual.

The data can be expressed using the matrix in **equation (2)**, where N is the number of respondents and M is the number of questions in the questionnaire. A data matrix with 60 columns and N rows was constructed for evaluating the similarity between the questions. A matrix with nine columns and N rows was constructed for evaluating the similarity between the scores of the nine subscales. **Equation (2)** represents the dataset used as the input dataset.

$$\begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1M} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{iM} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{N1} & \cdots & x_{N1} & \cdots & x_{NM} \end{pmatrix} \quad (2)$$

3.3.2 Preprocessing

When the dataset being analyzed had a null number, no correlations were calculated, and the relevant column was removed from the dataset.

3.3.3 Clustering method

The relationships between the 60 questions was investigated using the Pearson correlation coefficient (**equation (3)**).

$$r_{st} = \frac{\sum_{i=1}^{N}(x_{is} - \bar{x}_s)(x_{it} - \bar{x}_t)}{\sqrt{\sum_{i=1}^{N}(x_{is} - \bar{x}_s)^2 \sum_{i=1}^{N}(x_{it} - \bar{x}_t)^2}} \quad (3)$$

where $\bar{x}_s$ is defined as $\frac{\sum_{i=1}^{N} x_{is}}{N}$

A hierarchical clustering analysis was used to investigate the relationships (or independence) between the answers to the 60 questions. Ward's hierarchical clustering method [Ward et al, 1963] was selected. Initially, distance was defined as in **equation (4)**, based on the correlation coefficients of the 60 questions.

$$d_{st} = 1 - r_{st} \quad (4)$$

According to Ward's method, the increase in the sum of squares needs to be minimized when selecting two clusters to merge. This method is highly suitable for clustering because it avoids the "chaining" effect whereby poorly separated but distinct clusters are merged at an early stage. The method used to evaluate the distance between the clusters is described below.

The sum of squares of cluster A was calculated as in **equation (5)** and that of cluster B was calculated using a corresponding formula.

$$S_A = \sum_{i=1}^{N} \sum_{j(A)=1}^{n_A} (x_{ij(A)}^A - \bar{x}_i^A)^2 \quad (5)$$

where $x_{ij(A)}^A$ is the score of the jth answer of the ith individual in cluster A, n_A is the number of questions in cluster A, and $\bar{x}_i^A$ is defined as $\frac{1}{n_A} \sum_{j(A)=1}^{n_A} x_{ij(A)}^A$.

When cluster C is created by merging clusters A and B, the sum of squares of cluster C is defined as in **equation (6)**

$$S_C = \sum_{i=1}^{N} \sum_{j(C)=1}^{n_C} (x_{ij(C)}^C - \bar{x}_i^C)^2 = \sum_{i=1}^{N} \left(\sum_{j(A)=1}^{n_A} (x_{ij(A)}^A - \bar{x}_i^C)^2 + \sum_{j(B)=1}^{n_B} (x_{ij(B)}^B - \bar{x}_i^C)^2 \right) \quad (6)$$

where $\bar{x}_i^C$ is defined as in **equation (7)**

$$\frac{1}{n_C} \sum_{j(C)=1}^{n_C} x_{ij(C)}^C = \frac{1}{n_A + n_B} \left(\sum_{j(A)=1}^{n_A} x_{ij(A)}^A + \sum_{j(B)=1}^{n_B} x_{ij(B)}^B \right) = \frac{n_A \bar{x}_i^A + n_B \bar{x}_i^B}{n_A + n_B} \quad (7)$$

When $\bar{x}_i^C$ in **equation (6)** is substituted based on **equation (7), equation (8)** is created.

$$S_C = S_A + S_B + \frac{n_A n_B}{n_A + n_B} \sum_{i=1}^{N} (\bar{x}_i^A - \bar{x}_i^B)^2 \quad (8)$$

Therefore, when merging clusters A and B, the sum of squares is equal to the sum square of clusters A and B with an additional element, as expressed in **equation (9)**

$$\Delta S_{AB} = \frac{n_A n_B}{n_A + n_B} \sum_{i=1}^{N} (\bar{x}_i^A - \bar{x}_i^B)^2 \quad (9)$$

When cluster C is merged with cluster U, the additional element is expressed as in **equation (10)**.

$$\begin{aligned} \Delta S_{CU} &= \frac{n_C n_U}{n_C + n_U} \sum_{i=1}^{N} (\bar{x}_i^C - \bar{x}_i^U)^2 \\ &= \frac{n_A + n_U}{n_C + n_U} \Delta S_{AU} + \frac{n_B + n_U}{n_C + n_U} \Delta S_{BU} - \frac{n_U}{n_C + n_U} \Delta S_{AB} \end{aligned} \quad (10)$$

Therefore, the cluster distance (d_{CU}) is defined as in **equation (11)**

$$d_{CU} = \frac{n_A + n_U}{n_C + n_U} d_{AU} + \frac{n_B + n_U}{n_C + n_U} d_{BU} - \frac{n_U}{n_C + n_U} d_{AB} \quad (11)$$

In Ward's method, the clusters are created so as to minimize the d_{CU}. A dendrogram was constructed to repeatedly calculate the distance between the clusters, and merge the two clusters with the minimum distance [Blei et al, 2003].

These analyses were carried out using the open-source software, R [R Core Team, 2013].

3.4 Results and discussion

3.4.1 Correlations between the 60 questions

A histogram of the correlation coefficients between all 60 questions is displayed in **Figure 3.1**a. to show the relationships between the questions and to allow the meaning of the questions to be interpreted. The mean correlation coefficient between the questions was positive but small (0.19). In contrast, the mean correlation coefficient between Q_1 (Q_1, "Are you active?") and the other questions was negative (-0.16) (**Figure 3.1**b). This difference can be explained by the fact that Q_1 is about good health while the other questions are about ill health. The mean correlation coefficient between Q_{53} (Q_{53}, "Can you easily adapt to changes in your surroundings, including your social environment?") and the other question is close to 0 (0.01) (**Figure 3.1**c). Q_{53} is about adaptation to changes in the natural and social environment. Q_1 and Q_{53} ask about how good the respondent's health condition is from bad awareness. Therefore, the scores of Q_{53} and Q_1 are very different compared to those of the other questions.

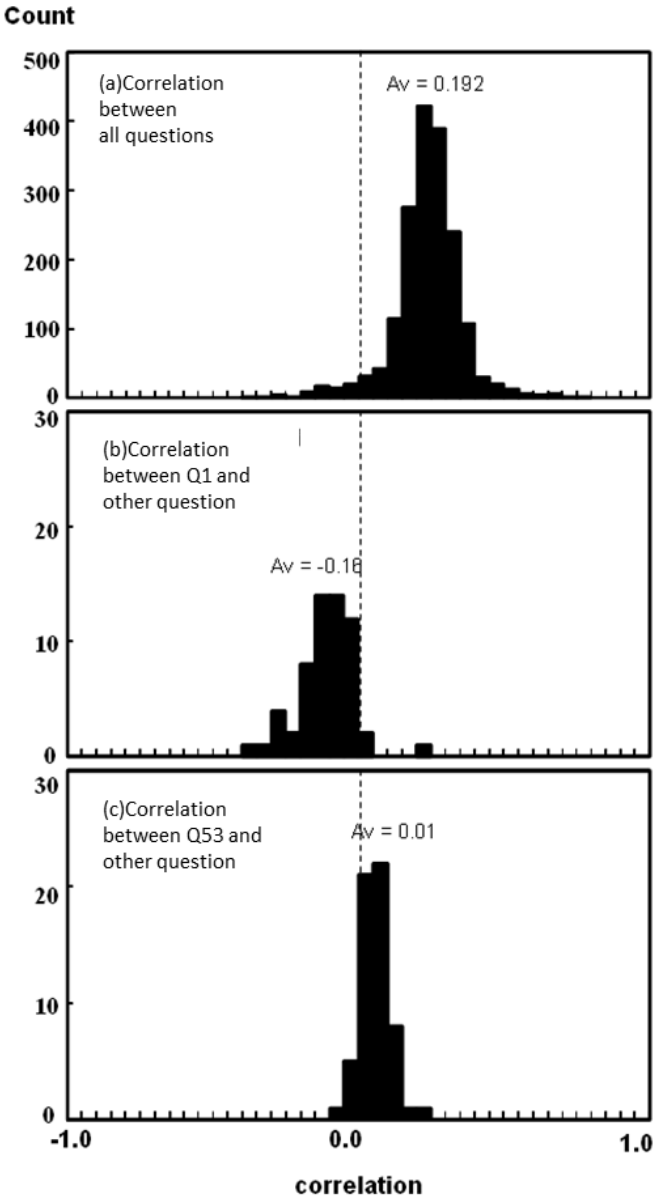


Figure 3.1: Histogram of the Pearson correlation coefficient between questions. (A) Correlation between all the questions. (B) Correlation between Q_1 and the other questions. (C) Correlation between Q_{53} and the other question

The World Health Organization defined “health” as a state of complete physical, mental, and social well-being, and not merely the absence of disease or infirmity [Masumoto, 2000]. The CCMQ-J evaluates health by taking into account satisfaction with one’s physical and spiritual side (Q_1, “Are you active?”) and satisfaction with social and environment (Q_{53}, “Can you easily adapt to changes in your surroundings, including your social environment?”).

Several of the correlation coefficients between pairs of questions were >0.6. This suggests that these pairs of questions concern similar conditions, for example, greasy skin (Q_{28} and Q_{39}, r = 0.68), mouth health (Q_{48} and Q_{49}, $r = 0.65$), anxiety and depression (Q_9 and Q_{10}, $r = 0.65$), breath condition (Q_3 and Q_4, $r = 0.64$), oral health (Q_{49} and Q_{50}, $r = 0.63$), and cold-like symptoms (Q_{24} and Q_{25}, $r = 0.61$). However, it is difficult to predict the scores for one question based on those of another question (**Figure 3.2**).

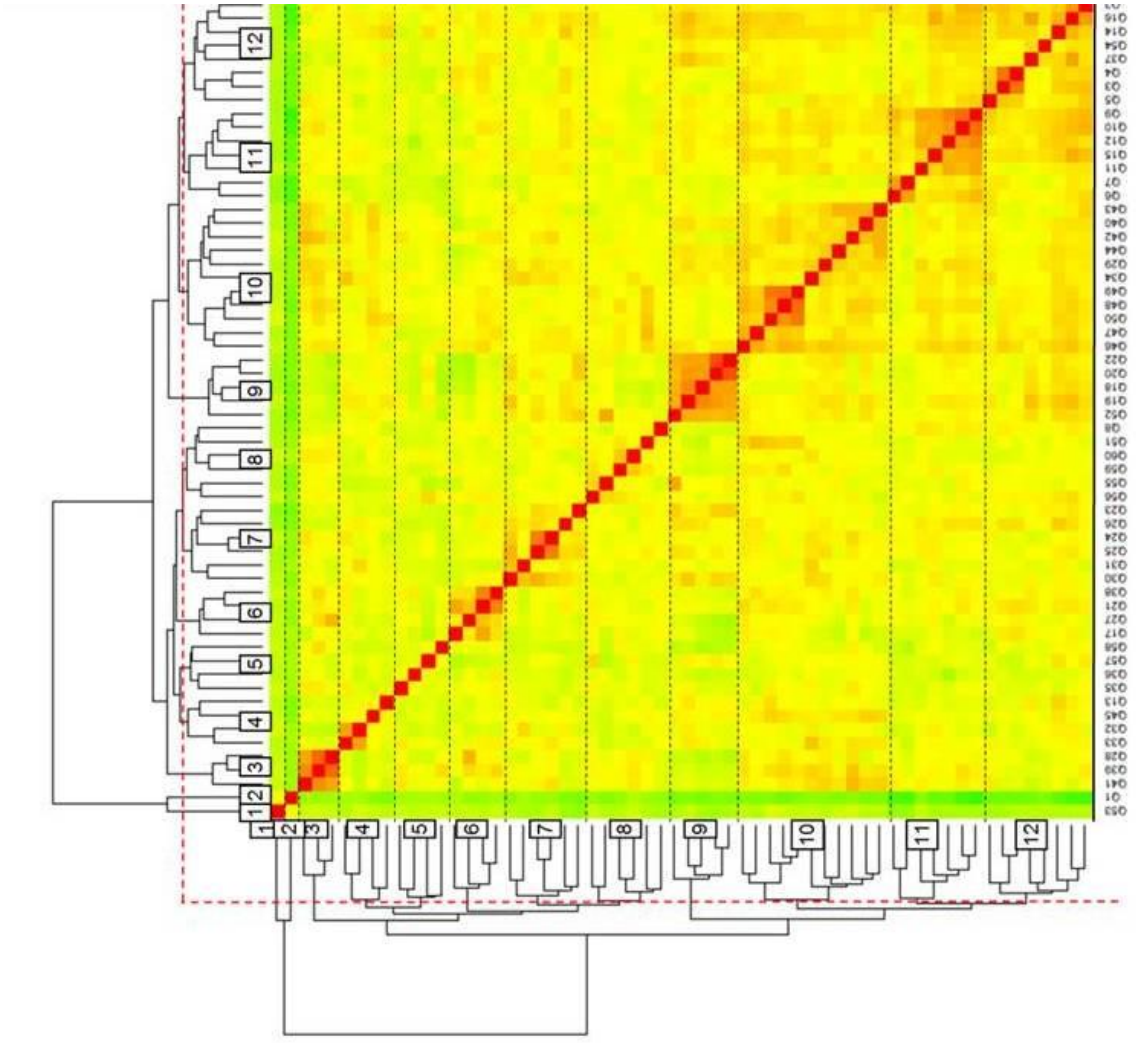


Figure 3.2: Heat map and dendrogram showing the similarities between the 60 questions (Red represents a correlation of 1, yellow represents 0, and green represents -1)

3.4.2 Correlations between the nine subscales

The correlations between the nine subscales were assessed to identified the relationships between the body constitutions, and to qualitatively validate the concepts underlying the design of the subscales (**Figure 3.3**).

The gentleness subscale was negatively correlated with the other subscales, clearly indicating that it is different from the other subscales. In contrast, the following pairs of subscales were correlated: special diathesis and yang deficiency; phlegm wetness and wet heat; yin deficiency and qi depression; and qi deficiency and blood stasis. In Chinese medicine, the biased subscales are known as "henpisitsu," (i.e., the eight subscales minus the gentleness subscale). Thus, the results of the statistical analysis were in line with the concepts underlying the design of the nine subscales)

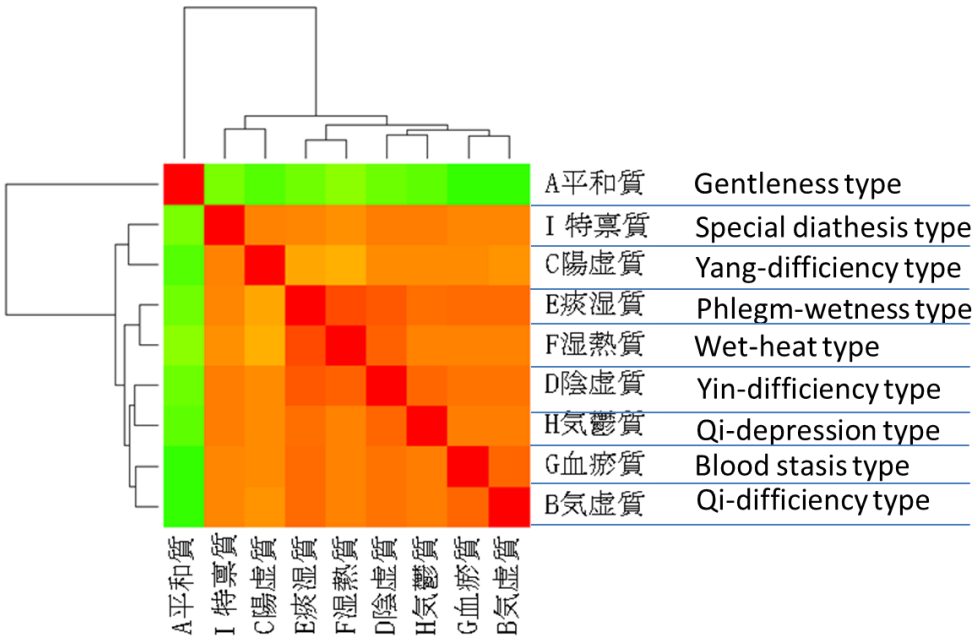


Figure 3.3: Heat map and dendrogram showing the similarities between the nine subscale scores (Red represents a correlation of 1, yellow represents 0, and green represents -1.)

3.4.3 Grouping the 60 questions into clusters

Using Ward's method, the 60 questions were grouped into 12 clusters. Most of these 12 clusters could be easily summarized based on the subscales associated with the questions in the cluster (**Table 3.3**).

Table 3.3: Questions in the 12 clusters

*Subscale occupies more than half of the cluster.

Cluster	Questions	Subscale
1	**Health**: Adaptation to natural and social environment (Q_{53}, gentleness)	Gentleness*
2	**Health**: Cheerfulness (Q_{1}, gentleness)	Gentleness*
3	**Sebaceous glands**: forehead (Q_{28}, phlegm wetness), nose (Q_{39}, wet heat), acne or pustules (Q_{41}, wet heat)	Wet heat*, phlegm wetness
4	**Mammary glands**: pain under arms or swellings in breasts (Q_{13}, qi-depression) **Skin**: spotty or mottled red-purple congestive marks (Q_{32}, special diathesis), blue bruises (Q_{33}, blood stasis), blue lips (Q_{45}, blood stasis)	Blood stasis*, Qi depression, special diathesis
5	**Blood color**: redness of lips (Q_{35}, yin deficiency), broken capillaries on cheeks (Q_{36}, blood stasis) **Stools**: constipation (Q_{57}, yin deficiency) **Obesity**: flabby belly (Q_{58}, phlegm wetness)	Yin deficiency*, blood stasis, phlegm wetness

(Continued)

Cluster	Questions	Subscale
6	**Hot**: palms or soles (Q_{17}, yin deficiency), face or body (Q_{2}, yin deficiency), cheeks (Q_{38}, yin deficiency), sweat (Q_{27}, qi depression) **Colds**: often get colds (Q_{23}, qi depression, yang deficiency)	Yin deficiency*, qi-depression
7	**Allergies:** sneezing (Q_{24}, special diathesis), stuffy noise (Q_{25}, special diathesis), feels stifling at turn of seasons (Q_{26}, special diathesis), sensitive (Q_{30}, special diathesis), nettle rash (Q_{31}, special diathesis)	Special diathesis*, yang deficiency, Qi depression
8	**Brain activity**: forgetful (Q_{8}, gentleness, blood stasis) **Discharge**: phlegm (Q_{51}, phlegm wetness), diarrhea (Q_{55}, yang deficiency), sticky stools (Q_{56}, wet heat), dark urine (Q_{59}, wet heat), yellow vaginal discharge or dampness around scrota (Q_{60}, wet heat)	Wet heat*, gentleness, blood stasis, phlegm wetness, yang deficiency
9	**Sensitivity to cold**: hands and feet (Q_{18}, yang deficiency), back, belly, and knees (Q_{19}, yang deficiency), sensitive to cold and need to wear more clothes (Q_{20}, yang deficiency), sensitive to cold in winter or due to air conditioning (Q_{22}, gentleness, yang deficiency), worsen after having something cold (Q_{52}, yang deficiency)	Yang deficiency*, gentleness

(Continued)

Cluster	Questions	Subscale
	Oral health: thirsty (Q_{46}, yin deficiency), dull feeling in throat (Q_{47}, qi-depression), bad breath (Q_{48}, wet heat), sticky mouth (Q_{49}, phlegm wetness), tongue plaque (Q_{50}, phlegm wetness)	Yin deficiency qi depression, wet heat, phlegm wetness, special diathesis, blood stasis
10	**Blood circulation**: dry skin and lips (Q_{29}, yin deficiency), marks when scratching skin (Q_{34}, special diathesis), pigmented spots (Q_{40}, blood stasis), puffy eyes (Q_{42}, phlegm wetness), under-eye shadows (Q_{43}, blood stasis), dry eyes (Q_{44}, yin deficiency)	
11	**Autonomic nerves**: find it troublesome to chat (Q_6, qi deficiency), quiet voice (Q_7, gentleness, qi deficiency), melancholic (Q_9, gentleness, qi-depression), irritated (Q_{10}, qi depression), easily moved (Q_{11}, qi depression), uneasy (Q_{12}, qi-depression), sighing (Q_{15}, qi deficiency)	Qi depression*, qi deficiency, gentleness
12	**Stress**: easily tired (Q_2, gentleness, qi deficiency), shortness of breath (Q_3, qi deficiency), palpitation (Q_4, qi deficiency), dizziness (Q_5, qi deficiency), tightness in chest (Q_{14}, phlegm wetness), tired body or arms and legs (Q_{16}, phlegm wetness), pain (Q_{37}, blood stasis), trouble sleeping (Q_{54}, gentleness)	Qi deficiency*, phlegm wetness, blood stasis, gentleness

Clusters 1 and 2 comprised questions from the gentleness subscale, and thus they can be interpreted as elements of gentleness. Based on the statistical clustering analysis, clusters 1 and 2 were differentiated from each other. In contrast, CCMQ system defines gentleness as a mixture of clusters 1 and 2. This difference is a characteristic of the CCMQ system.

The majority of questions in cluster 3 were from the wet-heat subscale, so cluster 3 can be interpreted as the wet-heat cluster. In contrast, cluster 10 is made up of questions from multiple subscales. For example, the questions in cluster 10 on oral health (regarding feeling thirsty and having a dull feeling in the throat, bad breath, a sticky paste in the mouth, and tongue plaque) belonged to four different subscales. Furthermore, the questions in cluster 10 on blood circulation (regarding dry skin and lips, marks when scratching, pigmented spots, puffy eyes, dark under-eye circles, and dry eyes) also belonged to four subscales. Therefore, it is difficult to explain cluster 10 using the CCMQ subscales. However, according to the statistical analysis, the 10 questions in cluster 10 belonged to the same cluster because the scores were correlated. The accuracy of the classification into the nine body compositions (reflected by the nine subscales) could be improved by understanding the differences between the clusters and making the questions easier to answer.

None of the clusters was composed mainly of questions in the phlegm-wetness subscale because these were distributed among four clusters. While the other subscale are composed of questions based on similar symptoms, the phlegm-wetness subscale is composed of questions based on various different symptoms, namely, skin (Q_{28} in cluster 3), obesity (Q_{58} in cluster 5), discharge (Q_{51} in cluster 8), oral

health (Q_{50} in cluster 10), blood circulation (Q_{42} in cluster 10), and stress (Q_{14} in cluster 12) symptoms. These questions were not strongly correlated with each other. In the CCMQ-J, phlegm wetness is defined as having a “foreign” characteristic, and this subscale is composed of questions on a large range of symptoms.

Chapter 4 Relationships between CCMQ-J and age and BMI

To elucidate how CCMQ considers body constitution changing every day, the relationships between body constitution and age and BMI were investigated by constructing statistical models.

4.1 BMI data

BMI (kg/m^2) is calculated by dividing an individual's mass (kg) by the square of their height (m). BMI is a physiologically important parameter as many diseases (such as hypertension, hyperglycemia, hepatic damage, and impaired glucose tolerance) are less common when BMI is 22 kg/m^2.

4.2 Subjects

There were 385 respondents (163 men and 222 women) with data on CCMQ-J scores and age, and 214 respondents (123 men and 91 women) with data on CCMQ-J scores and BMI.

4.3 Partial least square regression analysis

PLS regression was used to investigate the relationships between the scores and age and BMI. The PLS method is widely used in the chemo- and bio-informatics fields because PLS models can be constructed even if there are more variables than observations. In addition, this method is useful when there is multi-collinearity between the independent variables.

The PLS method was used to explain an objective variable, Y (i.e., either BMI or age) based on the scores of the 60 questions (X_1, X_2, …, X_M) using a linear model, as shown in **equation (12)**, for men, and a separate model for women.

$$Y = a_0 + a_1 X_1 + ... + a_j X_j + ... + a_M X_M \quad (12)$$

The PLS model is shown in **equations (13)** and **(14).**

$$\mathbf{y} = \bar{\mathbf{y}} + \sum_{k=1}^{A} \mathbf{t}_k q_k + \mathbf{e} = \bar{\mathbf{y}} + \mathbf{T} \cdot \mathbf{q} + \mathbf{e} \quad (13)$$

$$\mathbf{X} = \bar{\mathbf{X}} + \sum_{k=1}^{A} \mathbf{t}_k \mathbf{p}_k{}^T + \mathbf{E} = \bar{\mathbf{X}} + \mathbf{T} \cdot \mathbf{P}^T + \mathbf{E} \quad (14)$$

where q_k is the coefficient of y for the k^{th} component, p_k is the loading vector of X, A is the number of components, and t_k is a score vector for the k^{th} component. The residual matrix and vector are represented by E(M×N) and e(M×1), respectively. **Equations (13)** and **(14)** can be combined to create **equation (15)**.

$$\mathbf{Y} = \bar{\mathbf{y}} - \bar{\mathbf{X}}^T \mathbf{W}(\mathbf{P}^T \mathbf{W})^{-1} \mathbf{q} + \mathbf{X}^T \mathbf{W}(\mathbf{P}^T \mathbf{W})^{-1} \mathbf{q} \quad \mathbf{(15)}$$

The number of PLS components was determined by minimizing the root mean square error of prediction (RMSEP) value, which was calculated by leaving out one cross-validation for each component, as shown in **equation (16)**.

$$RMSEP = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{i(obs)} - y_{i(pred)})^2} \qquad (16)$$

where $y_{(obs)}$ is the experimental y value and $y_{(pred)}$ is the predicted y value.

A PLS package in R was used for this analysis [Mevik and Wehrens, 2008].

4.4 Results and discussion

4.4.1 Construction of the PLS model

The PLS results indicated that BMI and age can be predicted by the scores of the 60 questions. The optimum number of components for the age estimation model was 15 and 9 for men and women, respectively. The optimum number of components for the BMI estimation model was 15 for both men and women. The models were constructed using these numbers of components (**Figure 4.1**). The coefficients are shown in **Table 4.1**. Large correlation coefficients between the actual and estimated values were obtained for BMI (0.81 for men and 0.82 for women) and age (0.82 for men and 0.83 for women) (**Figures 4.1 and 4.2**). These results indicate that the 60 questions correlate with individuals' BMIs and ages. The CCMQ-J could be used as an indicator of body constitution for evaluating individuals as they age.

(A) (B)

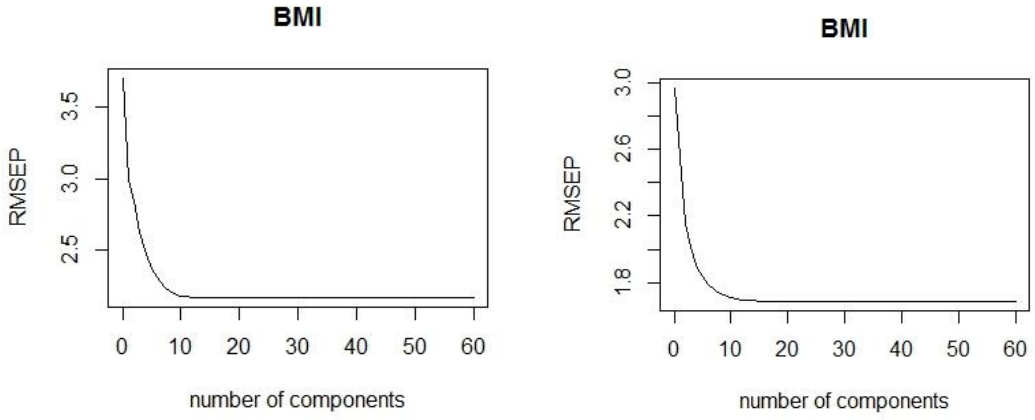


Figure 4.1: Relationship between the number of components and RMSEP for (A) male and (B) female.

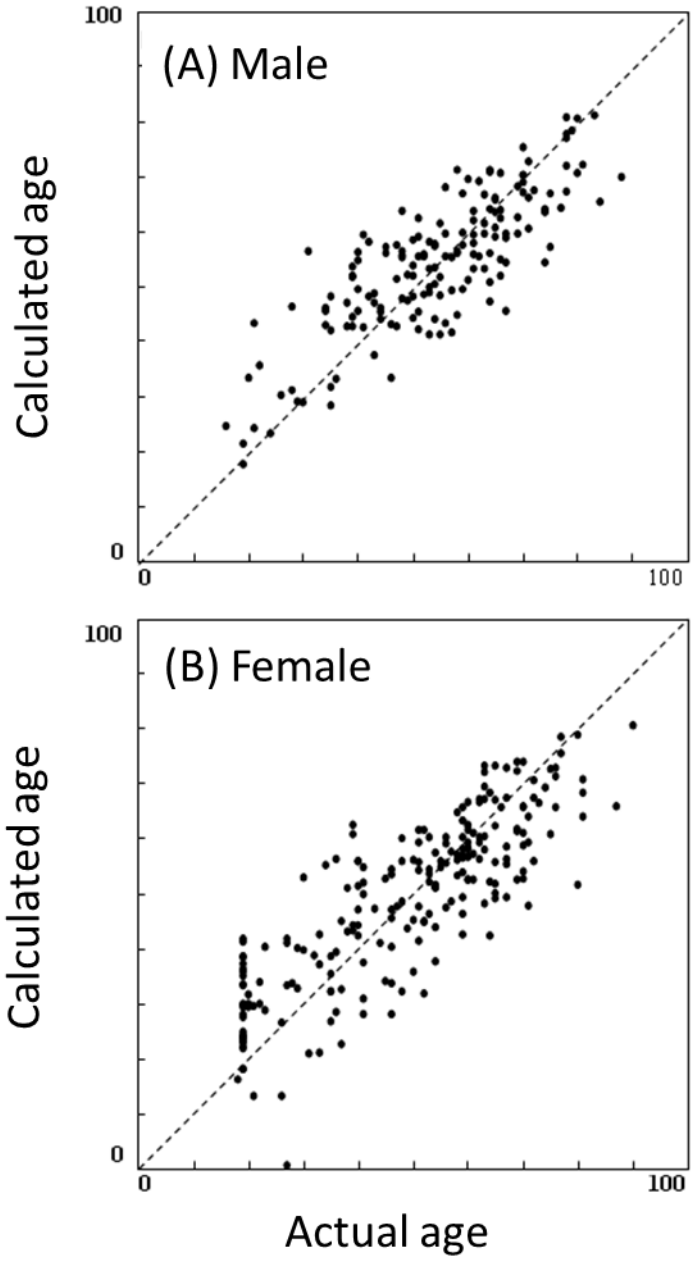


Figure 4.2: Relationship between calculated and actual age for (A) men and (B) women.

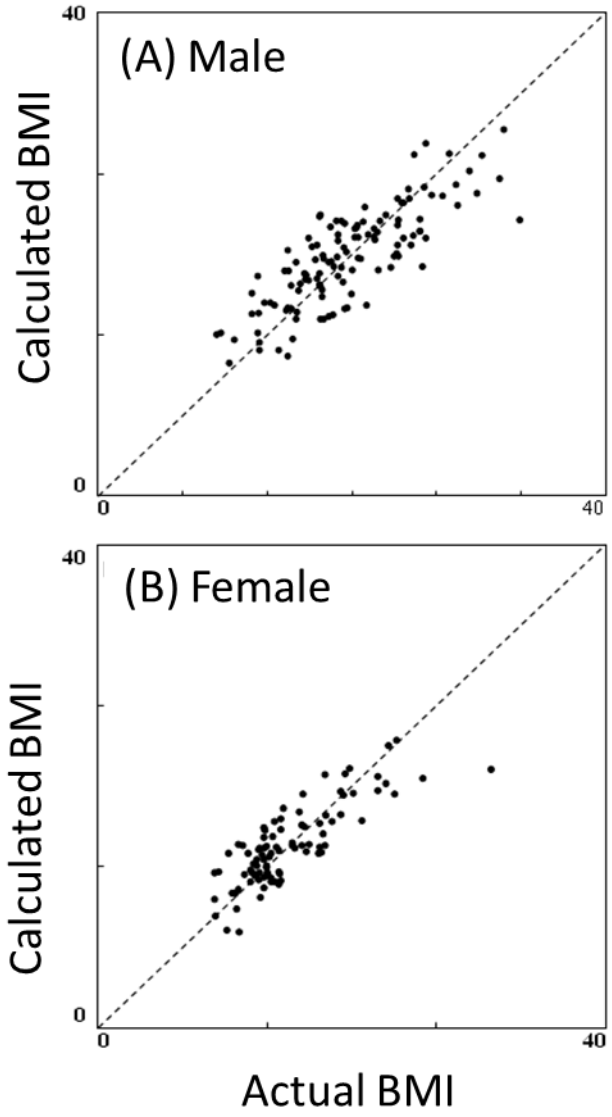


Figure 4.3: Relationship between calculated and actual BMI for (A) male and (B) female

Table 4.1: Correlation coefficients from the partial least square regression model

Question number	Age		Body mass index	
	Men	Women	Men	Women
0	60.19	55.84	25.07	18.09
1	-0.30	1.00	-0.02	0.28
2	0.03	-2.14	-0.12	0.52
3	1.79	0.56	1.11	-0.58
4	0.54	3.51	-0.48	0.72
5	2.08	-3.77	1.49	-0.38
6	0.94	-0.12	0.26	-0.14
7	-2.77	-0.84	-0.84	0.72
8	1.15	3.83	-0.96	0.06
9	-0.07	-0.43	-0.87	0.61
10	-0.83	0.09	-0.58	-0.16
11	2.16	0.13	0.88	-0.18
12	-1.92	0.17	-0.05	-0.56
13	-4.94	1.45	0.69	-0.63
14	-3.16	-1.33	0.84	0.40
15	-2.55	-1.34	-0.62	-0.60
16	3.37	2.88	0.46	0.07
17	-0.81	-0.55	0.15	1.26
18	1.21	-0.41	0.92	-0.09
19	1.74	2.54	-0.39	-0.01
20	0.74	0.80	-0.94	0.00
21	-2.06	0.83	-0.32	-0.54
22	-2.11	-2.01	-0.02	0.37
23	1.43	0.71	-0.30	0.14
24	0.28	-0.10	1.38	-0.37
25	-0.58	-0.05	-0.90	0.49
26	-2.49	-1.54	0.37	0.29
27	-0.59	-1.38	1.44	0.63
28	0.59	-1.43	-0.73	-0.92
29	-0.42	-2.72	-0.14	-0.45
30	0.16	1.45	0.15	-0.38
31	2.63	-0.89	-0.61	0.45
32	4.32	1.19	0.26	-1.31
33	2.79	-1.93	0.50	-0.19
34	-4.33	-0.28	-0.72	0.61
35	0.65	-2.78	1.06	0.24
36	0.12	0.46	0.26	-0.34
37	0.45	1.27	-0.05	-0.38
38	1.34	-1.30	0.57	0.08
39	-1.98	-0.70	0.68	0.60
40	1.84	5.37	-0.28	-0.51
41	-2.79	-3.66	-0.78	-0.62

(Continued)

Question number	Age		Body mass index	
	Men	Women	Men	Women
42	-0.85	-1.62	-0.11	0.01
43	-1.69	-2.03	-0.40	-0.63
44	-0.71	-1.02	-0.18	0.56
45	1.77	1.39	-1.60	0.34
46	0.40	1.30	0.14	-0.45
47	0.92	-0.80	-0.16	-0.34
48	-0.80	-1.28	-0.76	-0.16
49	-3.90	3.09	0.03	0.58
50	4.45	0.07	-0.01	-0.41
51	3.56	1.33	0.42	0.39
52	1.42	1.37	0.38	-0.15
53	-0.37	0.16	-0.49	0.32
54	-1.07	0.37	-0.87	0.36
55	-3.80	-2.31	-0.03	-0.49
56	2.26	3.10	-0.90	0.11
57	0.64	-1.46	-0.31	0.06
58	-0.11	0.57	1.27	1.09
59	1.67	0.30	1.35	0.10
60	-0.10	-2.02	0.39	-0.19

4.4.2 Comparison of regression coefficients in models of age and body mass index

The differences between the regression coefficients for male and female was carefully studied, and many differences were identified.

The age estimation model included 21 questions with correlation coefficients that had the opposite signs in the regression for men compared to that for women (**Figure 4.4**). This indicates that the contribution of the 60 questions to the estimation of age is different for male and female. The answers to the questions that contributed substantially to the estimation of age in the models for both men and women were investigated.

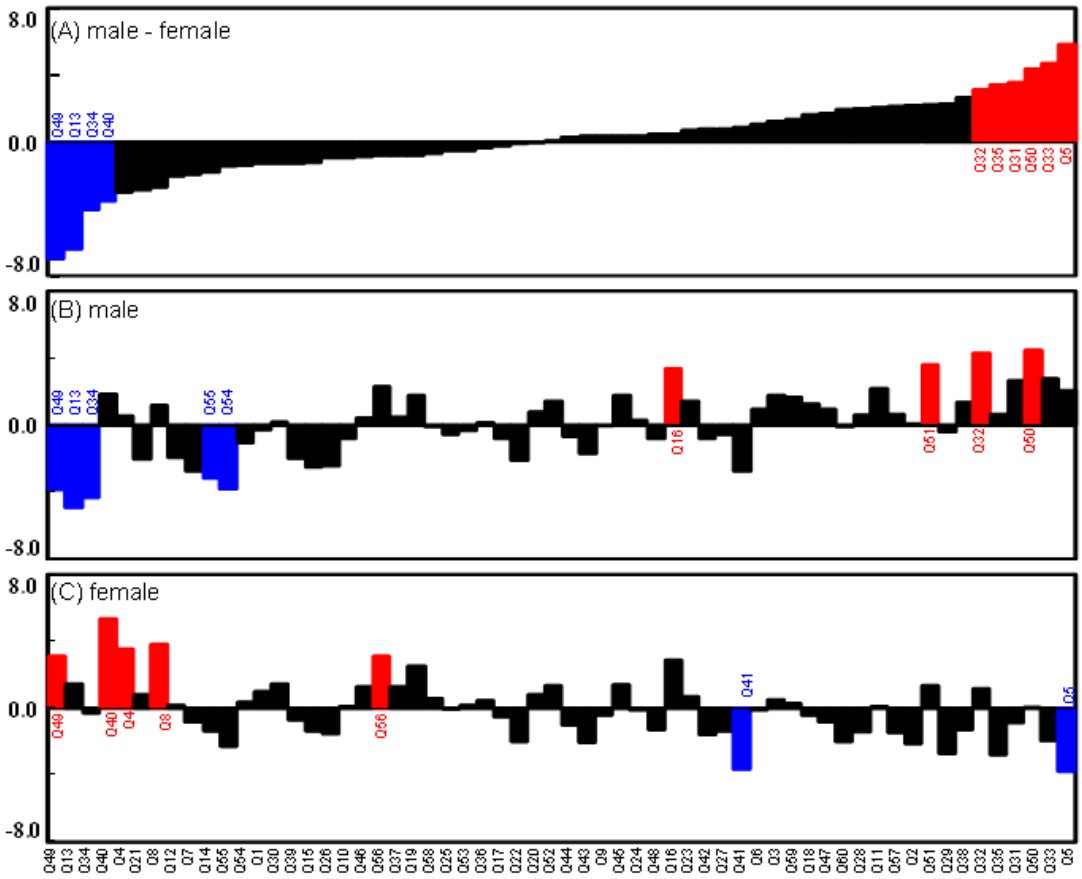


Figure 4.4: Correlation coefficients of the 60 questions in the age estimation model. (A) Differences between male and female. (B) Correlation coefficients for male. (C) Correlation coefficients for female. Red questions represent (M-F)≥3, and blue questions represents (M-F)≤-3, where M is the regression for male and F is the regression for female.)

Table 4.2: Questions that contributed substantially to the estimation of age

(Red question numbers represent (M-F)≥3, and blue question numbers represents (M-W)≤-3, where M is the regression for male and F is the regression for female. The red ♂ or ♀ symbols represent ≥3 for M and F, respectively. The blue ♂ or ♀ symbols represent ≤-3 for M and F, respectively.)

ID	Questions
3	**Sebaceous glands:** acne or pustules (Q_{41}♀, wet heat)
	Mammary glands: sighing (Q_{13}♂, qi-depression)
4	**Skin:** congestive marks (Q_{32}♂, special diathesis), blue bruises (Q_{33}, blood stasis)
5	**Coloring:** redness of lips (Q_{35}, yin deficiency)
7	**Cold:** hives (Q_{31}, special diathesis)
	Brain activity: forgetfulness (Q_{8}♀, gentleness, blood stasis)
8	**Discharge:** phlegm (Q_{51}♂, phlegm wetness), diarrhea (Q_{55}♂, yang deficiency), sticky stools (Q_{56}♀, wet heat)
	Oral health: sticky mouth (Q_{49}♀♂, phlegm wetness), tongue plaque (Q_{50}♂, phlegm wetness)
10	**Blood circulation:** sallow skin and dark spots (Q_{40}♀, blood stasis), marks when scratching skin (Q_{34}♂, special diathesis)
12	**Stress:** racing pulse (Q_{4}♀, qi deficiency), dizziness (Q_{5}♀), feeling of heaviness (Q_{16}♂, phlegm wetness), trouble sleeping (Q_{54}♂, gentleness)

The estimated age was correlated with the questions regarding “sebaceous glands,” “mammary glands,” “skin,” “coloring,” “cold,” “brain activity,” “discharge,” “oral health,” “blood circulation,” and “stress”.

The regression coefficient of Q_{49} (“Does your mouth feel sticky?) in the age estimation model is interesting. This question positively correlated with age in women but negatively correlates with age in men.

4.4.3 Comparison of regression coefficients in the BMI models between men and women

As for age, models of BMI for male and female were constructed (**Figure 4.5**).

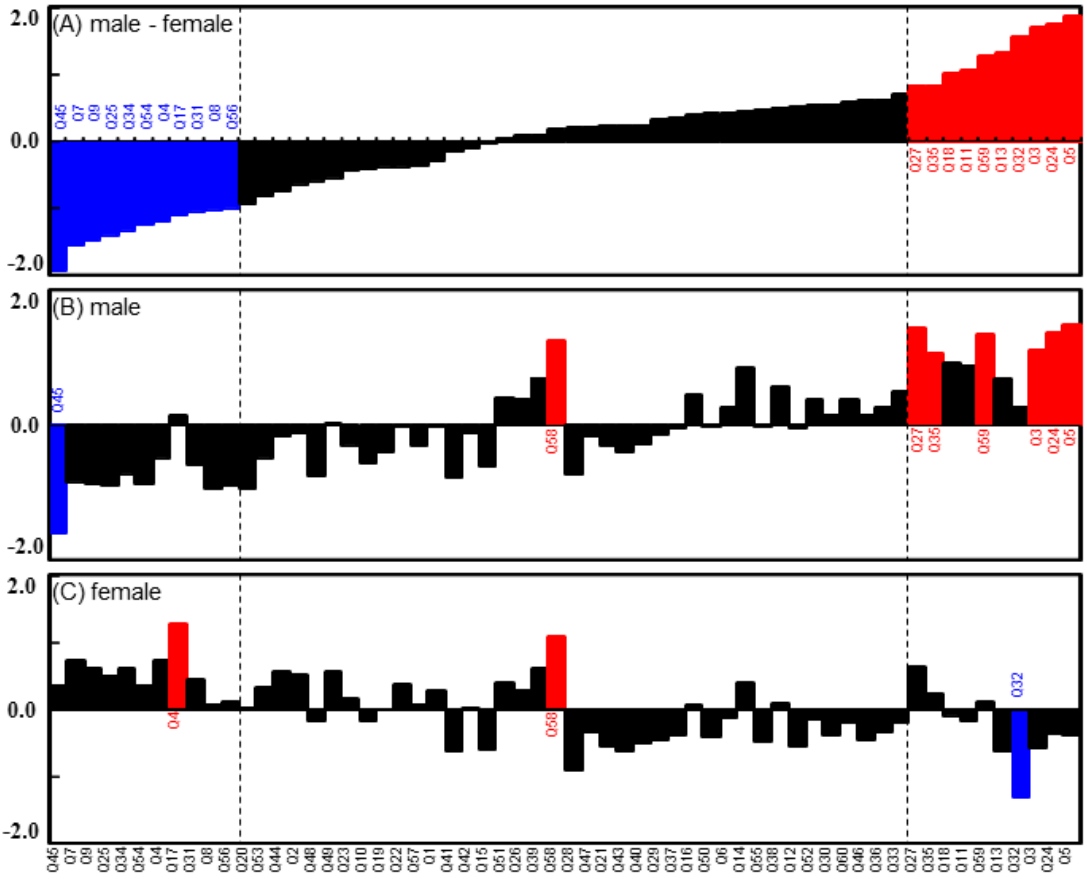


Figure 4.5: Correlation coefficients of the 60 questions in the BMI estimation models. ((A) Differences between male and female. (B) Correlation coefficients for male. (C) Correlation coefficients for female. Red questions represent (M-F)≥5, and blue questions represents (M-F)≤-5, where M is the regression for male and F is the regression for female.)

Table 4.3: Questions that contributed substantially to the estimation of BMI

(Red question numbers represent (M-W)≥5, and blue question numbers represents (M-W)≤-5, where M is the regression for male and F is the regression for female. The red ♂ or ♀ symbols represent ≥5 for M and F, respectively. The blue ♂ or ♀ symbols represent ≤-5 for M and F, respectively.)

ID	Questions
	Mammary glands: breast swellings (Q_{13}♂, Qi-depression)
4	**Skin:** congestive marks (Q_{32}♀, special diathesis), dark lips (Q_{45}♂, blood stasis)
5	**Coloring:** redness of lips (Q_{35}♂, yin deficiency), obesity (Q_{58}♂♀, phlegm wetness)
6	**Heat:** palms or soles (Q_{17}, yin deficiency), sweaty (Q_{27}♂, qi deficiency)
7	**Allergies:** sneezing (Q_{24}♂, special diathesis), runny nose (Q_{25}, special diathesis), hives (Q_{31}, special diathesis)
	Brain activity: forgetfulness (Q_{8}, gentleness, blood stasis)
8	**Discharge:** sticky stools (Q_{56}, wet heat), burning dark urine (Q_{59}♂, wet heat)
9	**Excessive sensitivity to cold:** hands and feet (Q_{18}, yang deficiency)
	Oral health: sticky mouth (Q_{49}, phlegm wetness)
10	**Blood circulation:** marks when scratching skin (Q_{34}, special diathesis)
11	**Autonomic nerves:** quiet voice (Q_{7}, gentleness, qi deficiency), depression (Q_{9}, gentleness, qi-depression), sentimental (Q_{11}♂, qi-depression)
12	**Stress:** breathlessness (Q_{3}♂, qi deficiency), racing pulse (Q_{4}♀, qi deficiency), dizziness (Q_{5}♂, qi deficiency), trouble sleep (Q_{54}, gentleness)

The estimated BMI was correlated with questions regarding “mammary glands,” “skin,” “coloring,” “obesity,” “heat,” “allergies,” “brain activity,” “discharge,” “excessive sensitivity to cold,” “oral health,” “autonomic nerves,” and “stress.”

There were many differences in the correlation coefficients in the BMI estimation models for men and women.

Di et al. investigated the relationship between the principle component of CCMQ scores and age based on 20,000 respondents [Di et al, 2014]. The research presented here on the construction of linear models for estimating age and BMI helps to understand how responses to the CCMQ-J are associated with aging and BMI.

Chapter 5 Simplification of the CCMQ-J

5.1 Purpose

The purpose of this study was to simplify the CCMQ-J in order to make the questionnaire easier to use to judge whether an individual has gentleness. Individuals with gentleness do not require treatment. However, individuals with body constitutions other than gentleness require treatment. A modified version of the CCMQ-J that contains fewer questions but still has good accuracy is proposed in this thesis.

The conventional way to judge whether an individual has gentleness is based on whether the following conditions are fulfilled after answering the 60 questions: gentleness subscale score >60 and other subscale scores <40. The conventional way to judge whether an individual has any of the other body constitutions is based on whether the following condition is fulfilled: subscale score >30.

The key CCMQ-J questions that are needed to accurately judge body composition were elucidated using an ensemble machine learning method and a dataset of CCMQ-J scores and body constitutions.

5.2 Method

(i) Judgment regarding whether an individual has gentleness

Datasets and variables

Training dataset: Sets of CCMQ-J answers from 419 individuals

Test dataset: Sets of CCMQ-J answers from 100 individuals

Explanatory variables: Answers to 60 questions

Dependent variable: Judgement regarding body constitution (gentleness or no gentleness)

Calculation method

The random forest method was used [Breiman, 2001], which involves ensemble machine learning. Data were selected from the training data set using random sampling, and decision trees were subsequently built. Classification conditions were voted for and selected. The degree that a question contributed to the classification was calculated in terms of the Gini coefficient, which is a statistical measure of the degree of variation represented in a set of values. Thus, the key questions were selected based on the Gini coefficients. New training and test data were constructed using these important questions. A model was constructed using the training data. The model was applied to the test data, and the accuracy rate was measured. Questions were removed until the accuracy rate was <80%.

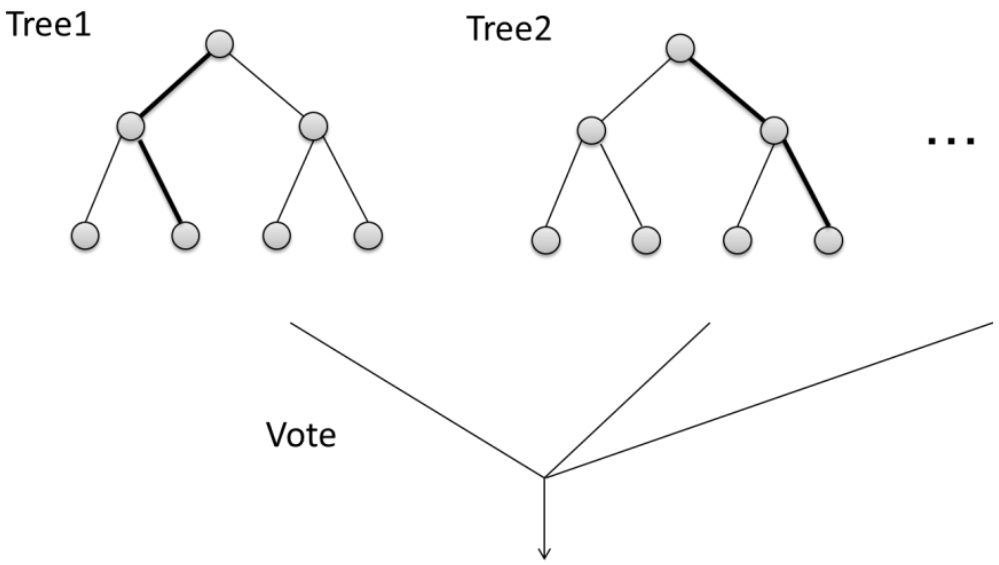


Figure 5.1: Schematic of the random forest method

(ii) Judgment regarding whether an individual has any of the eight other body constitutions

Datasets and variables

Training dataset: Sets of CCMQ-J answers from 244 individuals

Test dataset: Sets of CCMQ-J answers from 55 individuals

Explanatory variables: Answers to the 60 questions

Dependent variable: Judgements regarding each of the eight other body constitutions

Calculated method

As for the gentleness analysis, the random forest method was used.

5.3 Results

(i) Judgment regarding whether an individual has gentleness or not

The judgment regarding whether an individual has gentleness (**Table 5.1**) was 90% accurate when all 60 questions were used.

Table 5.1: Comparison between calculated and actual data in the test dataset

		Calculated data	
		Gentleness	Not gentleness
Actual data	Gentleness	48	6
	Not gentleness	4	42

Key questions were extracted based on the Gini coefficients (**Table 5.2**).

Table 5.2: Gini coefficients of questions used to classify whether an individual has gentleness or not

Question number	Gini coefficient
10	11.6
19	8.9
2	8.8
20	7.4
46	6.5
16	6.3
9	6.2
52	6.0
22	6.0
6	5.4

A graph of the number of questions against the accuracy rate of the judgment regarding whether an individual has gentleness is shown in **Figure 5.2**. The conditions used to make the judgement are shown in **Table 5.3.** When seven questions were used, the accuracy rate was >80%, which is a sufficient accuracy rate.

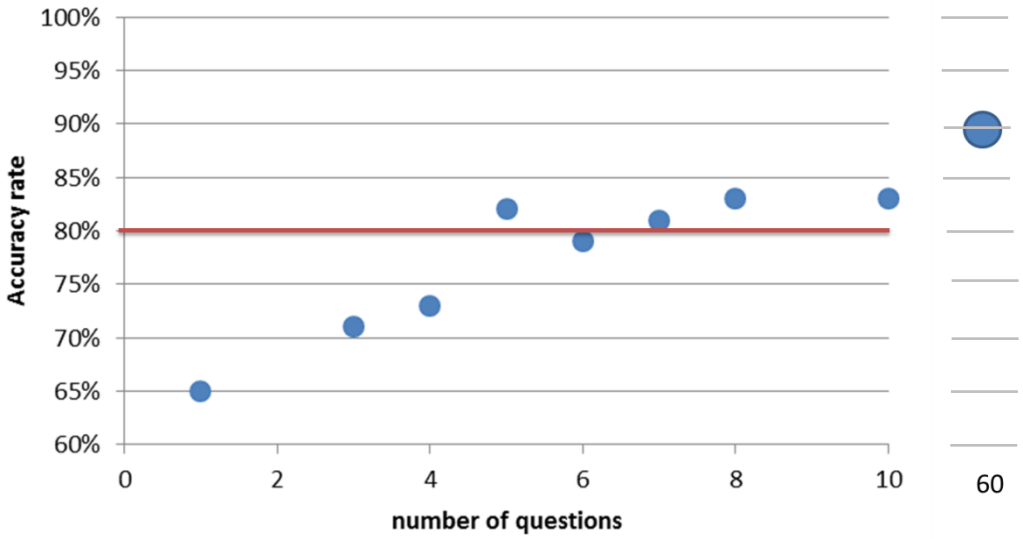


Figure 5.2: Accuracy rate of judgement regarding gentleness by number of questions answered

Table 5.3: Conditions used to classify gentleness and no gentleness

Condition	Prediction
Q_{19}>2.5, Q_{20}>2.5, and Q_{9}>1.5	No gentleness
Q_{2}≤2.5, Q_{20}≤4.5, Q_{46}≤2.5, Q_{16}≤2.5, and Q_{9}≤2.5	Gentleness
Q_{2}>2.5, Q_{46}≤2.5, Q_{46}>1.5, and Q_{9}>2.5	No gentleness
Q_{19}≤1.5, Q_{20}≤3.5, Q_{20}>1.5, Q_{46}≤3.5, and Q_{9}≤3.5	Gentleness
Q_{19}>1.5 and Q_{46}>2.5	No gentleness
Q_{46}>2.5 and Q_{9}≤1.5	No gentleness
Q_{10}≤2.5, Q_{20}≤2.5, and Q_{16}≤3.5	Gentleness
Other	Not gentleness

Therefore, judgments regarding gentleness can be made using the seven questions shown in **Table 5.4**. If an individual is judged to have gentleness, using the seven questions, he/she does not require treatment, and is not required to fill out any further questions in the questionnaire.

Table 5.4: Seven selected questions

Question number	Question
10	Do you feel nervous or irritated?
19	Do you feel cold in your back, belly, and knees?
2	Do you get tired easily?
20	Are you sensitive to cold and do you wear more clothes than others?
46	Do you feel thirsty or have a dry mouth?
16	Do you feel tired in your body or arms and legs?
9	Do you feel blue or depressed?

(ii) Judgments regarding whether an individual has any of the eight other body constitutions

If an individual is judged had not having gentleness after answering the first seven questions, they should answer a further seven questions because the applicability of some body conditions (for example, special diathesis) is difficult to judge based on the first seven questions (**Table 5.5**). The six further questions were selected based on their Gini coefficients to improve judgment accuracy to body constitution in minimum accuracy until mean accuracy is >80% (**Figure 5.3**, **Table 5.6**). The final set of 13 questions in shown in **Table 5.7.**

Table 5.5 : Accuracy rate of judgment regarding the eight other body constitutions using fewer questions ((a)Number of questions and accuracy rate of judgment The final selection of questions (involving 13 questions) is highlighted in yellow. The minimum of accuracy rate of judgment is highlighted in blue in each number of questions. (b):)

(a)

Number of question	Accuracy rate of judgment (%) YI	QF	BS	QD	WH	SD	PW	YA	Mean (%)
7	63.6	69.1	67.3	90.9	72.7	47.3	67.3	78.2	69.6
8	65.5	65.5	61.8	87.3	60.0	63.6	63.6	78.2	68.2
9	63.6	69.1	60.0	90.9	87.3	65.5	67.3	78.2	72.7
10	63.6	70.9	70.9	90.9	89.1	67.3	70.9	76.4	75.0
11	78.2	72.7	69.1	90.9	90.9	67.3	69.1	76.4	76.8
12	78.2	70.9	74.5	89.1	90.9	80.0	65.5	78.2	78.4
13	76.4	72.7	74.5	94.5	92.7	80.0	76.4	80.0	80.9
14	76.4	83.6	76.4	92.7	92.7	80.0	76.4	80.0	82.3
15	78.2	85.5	78.2	92.7	90.9	78.2	85.5	78.2	83.4
16	87.3	85.5	76.4	92.7	90.9	78.2	85.5	76.4	84.1
17	85.5	85.5	78.2	90.9	92.7	80.0	83.6	85.5	85.2
18	89.1	85.5	76.4	92.7	90.9	78.2	83.6	83.6	85.0
60	94.5	92.7	94.5	96.4	94.5	89.1	90.9	89.1	92.7

YI: yin deficiency, QD: qi deficiency, BS: blood stasis, QD: qi depression, WH: wet heat, SD: special diathesis, PW: phlegm wetness, YA: yang deficiency.

(b)

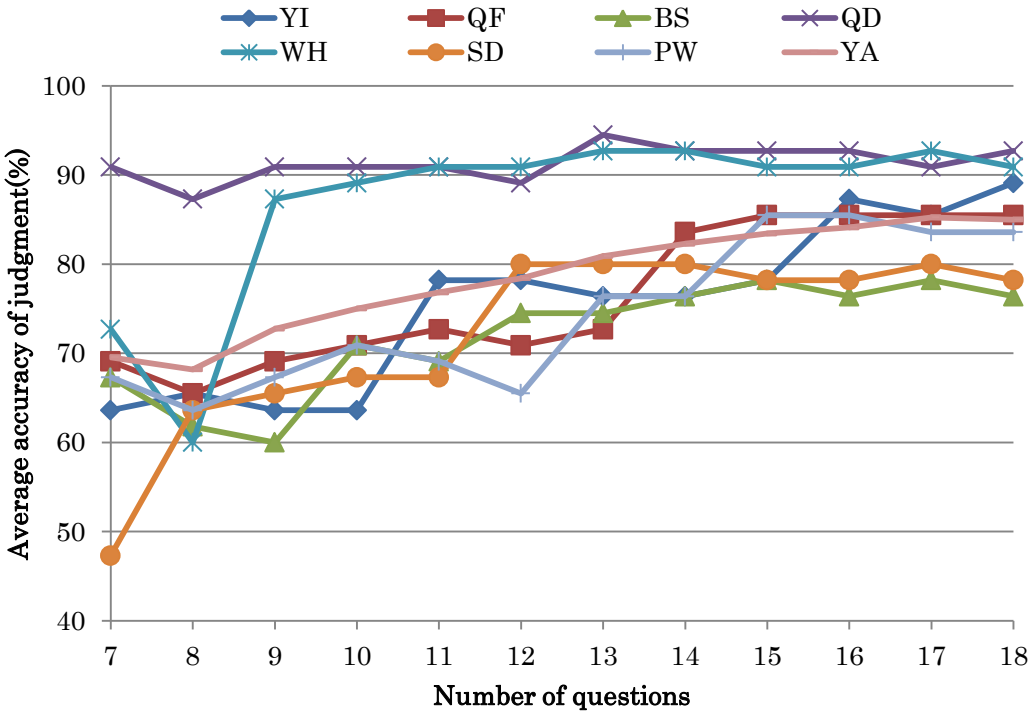
Number of question	Items of questions
7	Q2, Q9, Q10, Q16, Q19, Q20, Q46
8	Q2, Q9, Q10, Q16, Q19, Q20, Q25, Q46
9	Q2, Q9, Q10, Q16, Q19, Q20, Q25, Q39, Q46
10	Q2, Q9, Q10, Q16, Q19, Q20, Q25, Q39, Q40, Q46
11	Q2, Q9, Q10, Q16, Q19, Q20, Q25, Q38, Q39, Q40, Q46
12	Q2, Q9, Q10, Q16, Q19, Q20, Q25, Q30, Q38, Q39, Q40, Q46
13	Q2, Q9, Q10, Q16, Q19, Q20, Q25, Q30, Q38, Q39, Q40, Q46, Q58
14	Q2, Q3, Q9, Q10, Q16, Q19, Q20, Q25, Q30, Q38, Q39, Q40, Q46, Q58
15	Q2, Q3, Q9, Q10, Q16, Q19, Q20, Q25, Q30, Q38, Q39, Q40, Q46, Q50, Q58
16	Q2, Q3, Q9, Q10, Q16, Q19, Q20, Q25, Q30, Q38, Q39, Q40, Q44 Q46, Q50, Q58
17	Q2, Q3, Q9, Q10, Q16, Q19, Q20, Q22, Q25, Q30, Q38, Q39, Q40, Q44, Q46, Q50, Q58
18	Q2, Q3, Q9, Q10, Q16, Q19, Q20, Q22, Q25, Q30, Q38, Q39, Q40, Q43, Q44 Q46, Q50, Q58

Table 5.6: Gini coefficients of the questions

Subscale	Question	Gini coefficient
Yin deficiency	38	6419
	44	5601
	21	4537
	46	4186
	29	3581
Qi deficiency	3	8162
	5	4927
	4	4844
	6	4305
	27	2595
Blood stasis	40	10282
	43	8487
	45	6505
	37	2937
	33	2487
Qi depression	12	11527
	10	10784
	8	8130
	15	5028
	11	3469
Wet heat	39	14018
	41	9491
	60	8288
	48	5121
	56	5102

(Continued)

Subscale	Question	Gini coefficient
	25	12342
	30	11545
Special diathesis	24	7729
	34	6070
	31	5919
	58	9170
	50	5923
Phlegm wetness	49	5307
	42	4673
	48	4354
	20	21738
	22	18684
Yang deficiency	52	12132
	18	11127
	19	9830



YI: yin deficiency, QD: qi deficiency, BS: blood stasis, QD: qi depression, WH: wet heat, SD: special diathesis, PW: phlegm wetness, YA: yang deficiency.

Figure 5.3: Accuracy rate of judgment regarding the eight other body constitutions using fewer questions

When many questions are used, the judgments are more accurate, but answering the questionnaire is more inconvenient. When 13questions are used, the eight body constitutions can be judged with a good degree of accuracy (**Table 5.5**).

Thus, the following simplified version of the CCMQ-J is proposed:

(i) First, seven key questions should be answered

(ii) For individuals who are judged to have gentleness, no further questions should be answered.

(iii) For individuals who are judged to not have gentleness, six more questions should be answered.

(iv) Body composition subscale is assigned.

In the simplified version, the number of questions used to judge an individual's body composition is greatly decreased.

Table 5.7: Fourteen selected questions in the simplified CCMQ-J

Question number	Question
10	Do you feel nervous or irritated?
19	Do you feel cold in your back, belly and knees?
2	Do you get tired easily?
20	Are you sensitive to cold and wear more clothes compared to others?
46	Do you feel thirsty or have a dry mouth?
16	Do you feel tired in your body or arms and legs?
9	Do you feel blue or depressed?
38	Do you have red cheeks and feel hot?
25	Do you have a stuffy nose even when you don't have a cold?
30	Are you sensitive to medicine, foods, pollen, seasons, climate etc.?
3	Do you get out of breath?
40	Do you have sallow skin and easily get dark spots on your face?
39	Do you have an oily, shiny nose?
58	Do you have a fat flabby belly?

Chapter 6 Conclusion

In this thesis, first, to understand overall picture of research field in Chinese Medicine questionnaire objectively, 5469 abstracts of research paper ware classified based on a text-mining method. The position of the research topic in this thesis in relation to previous research was elucidated and similar research was identified objectively. Second, the relationships between questions of CCMQ-J were examined. The 60 questions were tentatively classified into 12 clusters and the similarity between these clusters and the nine subscales is discussed. Third, using a partial least square model, I found that BMI and age can be estimated based on the scores of the CCMQ-J questions. The correlation coefficient between the real and estimated values was over 0.8. Finally, a simplified version of the CCMQ-J which consists of 13 questions and was found to have averaged >80% accuracy was developed.

In Section 2.4, the top three topics in the CMQ literature were found to be statistical analysis of questionnaires, clinical trials, and nutrition. The research presented in this thesis was classified as a statistical analysis of a questionnaire. The research involved investigating the CCMQ-J, and the analysis techniques could be applied to other questionnaires related to “invisible” health conditions. Regarding the topic of nutrition, the effect of nutrition on health (i.e., blood pressure, quality of life and body constitution) has been explored using a questionnaire [Hamadate et al, 2014].

Body constitution was found to be correlated with age and BMI. Body constitution can be improved by considering the five principles of body constitution: eating, mental health, lifestyle, exercise, and self-care. These principles have been studied in order to develop ways to maintain health and prevent a serious disease from developing. This is highly necessary for individuals with a condition called "mi-byo," which is a condition that is not disease, but is developing into a disease. Accumulated data on the five principles and questionnaire response data can be merged to develop "mi-byo" systems. Adequate advice based on questionnaire responses contributes not only to reducing the growth of medical costs by testing for and treating individuals before they develop serious diseases, but also to maintaining health without needing to go to a hospital. Japan's Ministry of Economy, Trade and Industry and the Tokyo Stock Exchange strategically select companies to manage employee health as "health management stock" [Ministry of Economy, Trade and Industry, 2016]. The CCMQ can be used to easily check employee health as part of this process.

A food selection system that takes individuals' body constituents into consideration has been developed [KNApSAcK Family, 2016]. Moreover, there is a plan to develop and test a smartphone application based on the CCMQ to prevent disease. In this application, suitable foods for particular body constitutions will be proposed. Further research should focus on classifying individuals' body constitutions in order to treat them effectively (thereby lengthening their lives by reducing the rate of three key illnesses). I hope this thesis contributes to accelerate future medicine "mi-byo" system.

References

1) Afendi FM, Darusman LK, Morita AH, Altaf-Ul-Amin M, Takahashi H, Nakamura K, Tanaka K, Kanaya S. Efficacy prediction of jamu formulations by PLS modeling. Curr Comput Aided Drug Des. 2013;9(1):46-59.

2) Mevik BM, Wehrens R. The pls package: principal component and partial least squares regression in R. J Stat Softw. 2007;18(2):1-23.

3) Tuvblad C, Dhamija D, Berntsen L, Raine A, Liu J. Cross-cultural validation of the Reactive-Proactive Aggression Questionnaire (RPQ) using four large samples from the US, Hong Kong, and China. J Psychopathol Behav Assess. 2016;38(1):48-55.

4) Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3:993-1022.

5) Di J, Zhu YB, Wang Q, Wang, YY. Correspondence analysis of Chinese medical constitution features in different ages population. Zhongguo Zhong Xi Yi Jie He Za Zhi. 2014;34(5):627-630.

6) Elsevier. Content [Internet]. Elsevier; 2016 [cited Sept 20 2016]. Available from: http://jp.elsevier.com/online-tools/scopus/content-overview

7) Hamadate N, Kyo H, Matsumoto Y, Shikura M, Mizukami C, Seto K, Uebaba K, Suzuki N. Effects of dietary supplement containing kurozu concentrate on blood pressure, QOL and TCM constitution. Japanese Journal of Complementary and Alternative Medicine. 2014;11(1):95-102.

8) Wang JJ, Baranowski T, Lau WP, Chen TA, Pitkethly AJ. Validation of the Physical Activity Questionnaire for Older Children (PAQ-C) among Chinese children. Biomed Environ Sci. 2016;29(3):177-186.

9) Kamibaba K. 伝統医学の可能性. Japanese Journal of Complementary and Alternative Medicine. 2004;1:63-76.

10) KNApSAcK Family. Yakuzen database [Internet]. KNApSAcK Family [cited Sept 20 2016]. Available from: http://kanaya.naist.jp/YAKUZEN/top.jsp

11) Breiman L. Random forest. Mach Learn. 2001; 45(1);5-32.

12) Xu L, Cai B, Fang Z. Translation and validation of a Chinese version of the Mandibular Function Impairment Questionnaire. J Oral Rehabil. 2016;43(8):608-614.

13) Chong MF, Ayob MN, Chong KJ, Tai ES, Khoo CM, Leow MK, Lee YS, Tham KW, Venkataraman K, Meaney MJ, Wee HL, Khoo EY. Psychometric analysis of an eating behaviour questionnaire for an overweight and obese Chinese population in Singapore. Appetite. 2016;101:119-124.

14) Ministry of Economy, Trade and Industry. Health management stock 2016 [Internet]. Ministry of Economy, Trade and Industry (Japan); 2016 [cited

Sept 20 2016]. Available from: http:
//www.meti.go.jp/policy/mono_info_service/healthcare/
downloadfiles/meigara2016report.pdf.

15) Okada et al 漢方薬のインフォマティクス 多変量解析による漢方薬と「証」の相関解析. 132th Annual Meeting of the Pharmaceutical Society in Japan. 2012;3.

16) Rees L. The physical constitution and mental illness. Eugen Rev. 1947;39(2):50-55.

17) Lo SH, Chang AM, Chau JP. Translation and validation of a Chinese version of the Stroke Self-Efficacy Questionnaire in Community-Dwelling Stroke Survivors. Top Stroke Rehabil. 2016;23(3):163-169.

18) Masumoto T.「健康」概念に関する一考察. 立命館産業社会論集. 2000;36:123-139.

19) Thai SY, Shun SC, Lee PL, Lee CN, Weaver TE. Validation of the Chinese version of the Functional Outcomes of Sleep Questionnaire-10 in pregnant women. Res Nurs Health. 2016;39(6):463-471.

20) Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci U S A. 2004;101(Suppl 1):5228–5235.

21) Tada T. 補完代替医療の理念 Japanese Journal of Complementary and Alternative Medicine. 2004;1:1-3.

22) Ward JH, Jr. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963;58(301):236-244.

23) Liu Y, Wang L, Wei Y, Wang X, Xu T, Sun J. Validation of a Chinese version of the Chronic Pain Acceptance Questionnaire (CAPQ) and CPAQ-8 in chronic pain patients. Medicine (Baltimore). 2016;95(33):e4339.

24) Zhu Y, et al. 中医体質学. 中医臨床. 2006;27:8-15.

25) Zhu Y, et al. 健康体質づくり-スマートライフの実現に向けて. 未病体質研究会. 2014. Kanazawa.

26) R Core Team. R: A language and environment for statistical computing [Software]. R Foundation for Statistical Computing; 2016 [cited Sept 20 2016]. Available from: http://www.R-project.org/.

27) Porter M. Snowball [Software]. 2015 [cited Sept 20 2016]. Available from: http://snowballstem.org.

Achievements

Publication

1. Akihiro Yamamori[1, 2] , Feng Hao Xu[3], Tomoyuki Watanabe[2], Kou Mei[1], Naohiro Ono[1], Tetsuo Sato[1], Katsushi Kawabata[3], Uebaba Kazuo[4], Imanishi Keihou[5], Md. Altaf-Ul-Amin[1] , Suzuki Nobutaka[3], Shigehiko Kanaya[1]

   [1]Nara Institute of Science and Tehcnology, Graduate School of Information Science, Computational Systems Biology Lab, [2] Sumitomo Chemical Technology & Management Planning Division, [3] Kanazawa Univsersit, Graduate School of Medical Sciences, CAM Research Room, [4] Science Research Center of Alternative Medicine, [5]e-Growth Co., Ltd.

Japanese Journal of Complementary and Alternative Medicine 2016 vol13 No2 p43-56

International conference

1. Akihiro Yamamori, 2016, Relationship between physical constitution as assessed using the Chinese Medicine Questionnaire in Japan (CCMQ-J) and physiological conditions (BMI and aging) , 2nd International Conference on Movement and Nutrition in Health and Disease University of Regensburg, Lecture Hall H24, 93040 Regensburg, Germany.

Supplemental information

S1. Method used to calculate the correlations

The Pearson correlation coefficient (a parametric method) was selected to calculate the correlations between the questions and subscales because it is easy to use and the purpose of the research was not to obtain accurate correlation statistics but simply to analyze the CCMQ-J scientifically.

The CCMQ data are discrete, based on answers using a Likert scale ranging from 1 to 5. There have been long-running debates over whether to treat this type of data as ordinal or interval [Carifio and Perla, 2008]. If the data are treated as ordinal, a non-parametric method should be used. However non-parametric methods are less powerful compared to parametric methods, and they may miss out important findings.

The Pearson correlation coefficients were verified using Spearman's rank correlation coefficients, which involve treating data as ordinal. The correlation coefficients of the 60 answers (Chapter 3) calculated using the Pearson and Spearman methods were compared. The mean Pearson and Spearman's rank correlation coefficients were 0.192 and 0.196, respectively (**Figure S1.1**). The mean difference in the correlation coefficients (Pearson correlation coefficient -

Spearman's rank correlation coefficient) was 0.04 (**Figure S1.2**). The difference between the two methods was deemed to be small.

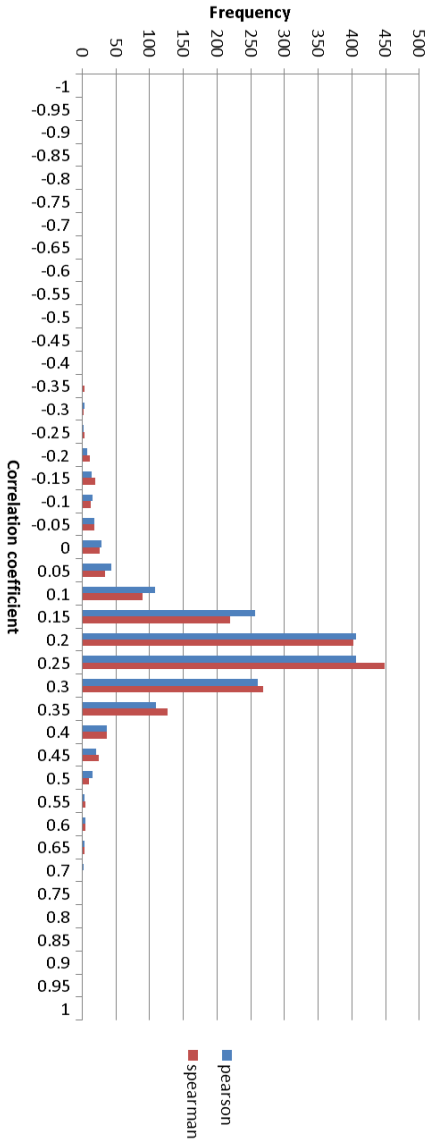


Figure S1.1: Histogram of Pearson and Spearman's rank correlation coefficients

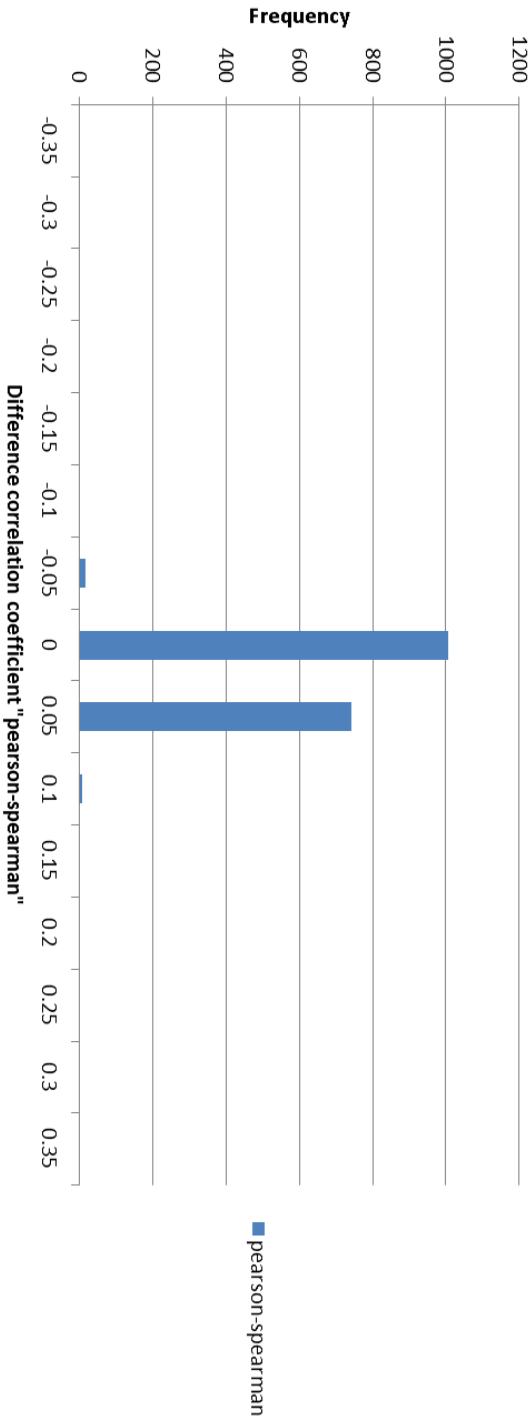


Figure S1.2: Histogram of differences in the Pearson and Spearman's rank correlation coefficients

S1 Reference

Carifio J, Perla R. Resolving the 50-year debate around using and misusing Likert scales. Med Educ. 2008;42(12):1150-1152.

S2. PLS regression using random answers to 60 questions

It was found that the correlation coefficients between the real and estimated values for BMI (0.81 for men and 0.82 for women) and age (0.82 for men and 0.83 for women) were large. To investigate this high level of correlation between the values calculated using PLS and the actual values for BMI and age for men, random answers to 60 questions were selected, and correlation coefficients was calculated.

Data: Random answers to the 60 questions and BMI data for men.

Method: PLS regression

Result: The correlation coefficient between the predicted and actual BMIs was 0.11 when random answers were used. The correlation coefficient for men's BMI (R=0.81) was comparatively high.

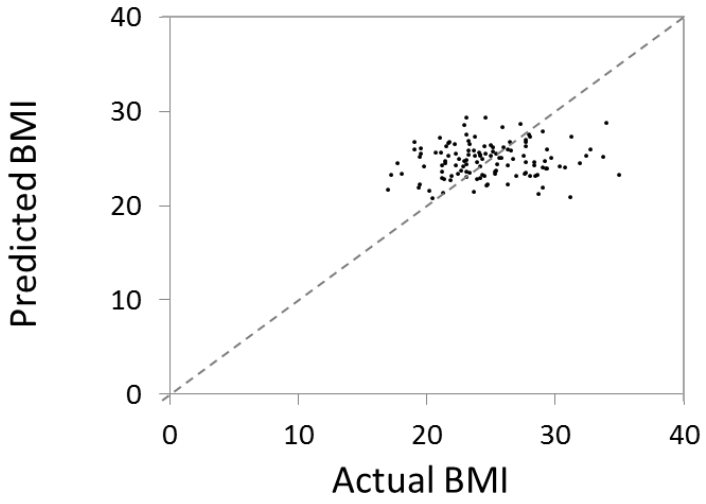


Figure S2.1: Relationship between the calculated BMI (using a random answer) and actual BMI