데이터 마이닝을 이용한 한의비만변증 설문지 재평가: 실제 임상에서 수집한 설문응답 기반으로

오지홍 · 왕징화¹ · 최선미^{2,3} · 김호준

동국대학교 한의과대학 한방재활의학교실, ¹대전대학교 동서생명과학연구원, ²한국한의학연구원 한의약데이터부, ³과학기술연합대학원대학교 한의융합의학과

Re-evaluation of Obesity Syndrome Differentiation Questionnaire Based on Real-world Survey Data Using Data Mining

Jihong Oh, Jing-Hua Wang¹, Sun-Mi Choi^{2,3}, Hojun Kim

Department of Korean Medicine Rehabilitation, College of Korean Medicine, Dongguk University, ¹Institute of Bioscience & Integrative Medicine, Daejeon University, ²KM Data Division, Korea Institute of Oriental Medicine, ³Korean Convergence Medicine, University of Science & Technology

Received: November 5, 2021
Revised: December 8, 2021
Accepted: December 13, 2021

Correspondence to: Hojun Kim
Department of Korean Medicine
Rehabilitation, Dongguk University
Ilsan Oriental Hospital, 27
Dongguk-ro, Ilsandong-gu, Goyang
10326, Korea
Tel: +82-31-961-9111
Fax: +82-31-961-9009
E-mail: kimklar@dongguk.ac.kr

Copyright © 2021 by The Society of Korean
Medicine for Obesity Research

Objectives: The purpose of this study is to re-evaluate the importance of questions of obesity syndrome differentiation (OSD) questionnaire based on real-world survey and to explore the possibility of simplifying OSD types.

Methods: The OSD frequency was identified, and variance threshold feature selection was performed to filter the questions. Filtered questions were clustered by K-means clustering and hierarchical clustering. After principal component analysis (PCA), the distribution patterns of the subjects were identified and the differences in the syndrome distribution were compared.

Results: The frequency of OSD in spleen deficiency, phlegm (PH), and blood stasis (BS) was lower than in food retention (FR), liver qi stagnation (LS), and yang deficiency. We excluded 13 questions with low variance, 7 of which were related to BS. Filtered questions were clustered into 3 groups by K-means clustering; Cluster 1 (17 questions) mainly related to PH, BS syndromes; Cluster 2 (11 questions) related to swelling, and indigestion; Cluster 3 (11 questions) related to overeating or emotional symptoms. After PCA, significant different patterns of subjects were observed in the FR, LS, and other obesity syndromes. The questions that mainly affect the FR distribution were digestive symptoms. And emotional symptoms mainly affect the distribution of LS subjects. And other obesity syndrome was partially affected by both digestive and emotional symptoms, and also affected by symptoms related to poor circulation.

Conclusions: In-depth data mining analysis identified relatively low importance questions and the potential to simplify OSD types.

Key Words: Syndrome differentiation, Syndrome differentiation questionnaire, pattern identification, Korean traditional medicine, Data mining, Principal component analysis

서론

비만은 에너지 불균형 상태로 체 내에 지방이 과잉 축적된 상태이다. 2019년 국민건강통계에서 만 19세 이상

인구 중 체질량지수(body mass index, BMI)가 25 kg/m² 이상인 인구 비율이 33.8%로 파악되었고, 2010년 비만을 30.9%와 비교했을 때 약 3% 증가했다¹⁾. 비만은 유전이나 내분비계통의 질환 외에도 과도한 열량 섭취 및 불규칙한

식습관, 부족한 신체활동, 정신적 요인 등 다양한 요인으로 유발된다.

한의임상에서는 비만과 동반하여 나타나는 환자의 다양한 증상을 바탕으로 변증을 통해 비만을 치료한다. 비만치료 임상경험이 있는 73명의 한의사를 대상으로 비만의 주요 진단지표를 조사했을 때 변증이 가장 높은 빈도의 응답을 보였다²⁾. 여러 변증 중에서도 한의학연구원에서 제시한 간울(liver qi stagnation), 식적(food retention), 양허(yang deficiency), 비허(spleen deficiency), 담음(phlegm), 어혈(blood stasis) 등의 6개 변증 유형이 다빈도로 사용되고 연구되어 왔다³⁻⁷⁾. 2009년 한방비만학회와 한의학연구원이 공동연구로 한방비만변증 설문지를 개발했는데 설문문은 총 52문항으로 전신증상, 정서/성격, 소화기능, 순환기능에 대한 질문으로 구성된다³⁾. 6개의 변증 별로 10개 문항의 점수를 합산하고, 대상자는 합산 점수가 가장 높은 값을 가지는 변증으로 판별된다.

그러나 기존 연구에서 한의사가 6가지 비만변증에서 주증에 해당한다고 선택한 항목들과 비만환자를 대상으로 수집한 설문에서 변증 별로 증상 빈도가 높게 나타난 항목 순위에는 차이가 있었다²⁾. 또 다른 연구에서 한의비만변증 설문지에서 판단된 변증 빈도는 간울, 식적, 양허, 비허, 담음, 어혈 순으로 나타났고, 특히 담음 변증은 7.6%, 어혈 변증은 3.77%로 전체 변증에서 차지하는 비율이 낮았다⁸⁾. 한의사의 변증과 환자의 설문항목 응답이 주관적이기 때문에 단일 변증이 아닌 여러 변증에 해당하는 증상이 혼재될 수 있으며 실제 임상에서 일관되지 않은 결과가 나타날 수 있다. 따라서 실제 비만환자를 대상으로 수집한 자료를 바탕으로 기존의 비만변증 설문지 항목의 중요도를 재평가하고, 비만변증 유형도 실제 임상 현장에 맞게 개편할 필요성이 있다.

여러 분석 방법 중에서도 본 연구에서는 자료에 내재되어 있는 유용한 특성을 파악하여 의미 있는 지식이나 암묵적인 지식을 파악하는 도구로서 데이터 마이닝을 이용했다⁹⁾. 데이터 마이닝 절차는 크게 자료 수집, 전처리, 탐색적 자료 분석, 알고리즘을 이용한 데이터 마이닝, 결과 해석 및 활용으로 구분할 수 있다. 데이터 전처리 과정에서 결측치나 이상치를 발견해 데이터를 다듬고, 탐색적 자료 분석 단계에서는 특정한 가설 없이 자료의 통계치를 비교해보고, 변수들 간의 관련성을 파악하거나 학습에 사용할 특징(feature) 혹은 변수(variable)를 선택한다. 이후

분석 목적에 따라 적절한 수학적 모형이나 알고리즘을 이용해 데이터에서 유용한 정보를 추출한다. 이 단계에서 실제 데이터를 기반으로 데이터의 구조와 특성을 파악하기 위해서 데이터의 결과값에 해당하는 레이블(label) 혹은 타겟(target)을 참조하지 않는 비지도학습(unsupervised learning) 방법을 활용할 수 있다. 대표적으로 비슷한 특성들끼리 묶어 군집화(clustering)하거나 다차원 데이터를 축소하여 데이터의 분포를 파악하는 차원 축소 등이 있다. 마지막으로 데이터 마이닝 분석 결과를 바탕으로 결과를 해석하고 활용한다.

본 연구에서는 비만환자를 대상으로 수집한 한의비만변증 설문 응답을 바탕으로 데이터 마이닝 분석을 통해 문항의 중요도를 재평가하며 문항 축소 및 비만변증 유형 효율화 가능성을 탐색하고자 한다. 실제 설문자료에서 변증 결과 빈도를 파악하여 기존 연구와의 일관성을 확인하고, 탐색적 자료 분석을 통해 축소 가능한 문항을 선별한다. 또한 변증 결과인 레이블을 고려하지 않고 비슷한 응답 패턴을 보이는 문항끼리 군집화하여 자료에 내재되어 있는 특성을 파악하고, 전체 대상자들의 응답에 따라 2차원 공간에서 분포를 확인한 뒤 분포에 영향을 주는 문항들의 특성을 분석하여 비만변증 개편 가능성을 확인한다.

재료 및 방법

1. 한의비만변증 설문조사 자료 수집

한방비만학회와 한의학연구원에서 개발한 한방비만변증 설문지는 총 52문항으로 전신증상 8문항, 정서/성격 8문항, 소화기능 18문항, 순환기능 18문항으로 구성되어 있다. 대상자는 각 문항에 대해 5점 척도(① 거의 없다, ② 가끔 혹은 조금 있다, ③ 보통 정도로 증상이 있다, ④ 다소 심하다, ⑤ 아주 심하다)로 답변한다(Table 1). 동국대학교 일산한방병원에서 비만환자를 대상으로 2010년 11월 29일, 2011년 2월 15일에 수집한 한방비만변증 설문지 자료를 환자 개인정보를 제외하고 고유코드를 부여하여 한국한의학연구원에 전송했고, 변증 별 문항 합산 점수에 따라 판별 결과를 받아 분석에 활용했다.

2010년 11월 29일에 수집한 설문은 18세 이상 65세 미만의 BMI 25 kg/m² 이상이며 연구 참여 전 3개월 간 체중이 3 kg 이상 변동이 없으며 1달간 체중감량을 하지 않고, 운동량의 변화가 없는 남녀를 대상으로 하였다. 또한 심

Table 1. Obesity Syndrome Differentiation Questionnaire

	Abbreviation	Related syndrome differentiation
General condition (8)		
G1. Feeling tired as usual	Tired	SD
G2. Feeling body heavy or hard to move	Body_heavy	YD, SD, PH
G3. Feeling dizziness	Dizziness	PH
G4. Feeling head heavy or woozy	Woozy	PH
G5. Pale face	Pale	YD
G6. Cold sweats	Cold_sweats	YD
G7. Hating cold weather	Hate_cold	YD
G8. Cold body or cold limbs	Cold_body	YD
Emotional symptoms (8)		
E1. Introverted or sensitive	Sensitive	LS
E2. Getting angry easily	Angry	LS
E3. Worrying a lot as usual	Worry	LS
E4. Feeling heavy in the chest	Heavy_chest	LS
E5. Feeling depressed frequently	Depressed	LS
E6. Worsen pain after psychological stress	Stress_pain	LS
E7. Pain during compression of the sternum	Sternum_pain	LS
E8. Feeling flank pain	Flank_pain	LS
Digestive symptoms (18)		
D1. Tendency to poor digestion	Poor_digestion	FR, SD
D2. Worsen dyspepsia after psychological stress	Stress_dyspepsia	LS
D3. Tendency to overeat	Overeat	FR
D4. Tendency to eat even when full	Greedy_eat	FR
D5. Tendency to tasteless	Tasteless	SD
D6. Eating less	Eat_less	PH (multiply by 0.5)
D7. Feeling bloated stomach.	Bloated	FR, SD, PH (multiply by 0.5)
D8. Frequent belching	Belching	FR
D9. Gassy stomach	Gassy	FR
D10. Frequent indigestion	Indigestion	FR
D11. Feeling food stagnation	Food_stagnation	FR
D12. Frequent upset stomach	Upset_stomach	FR
D13. Frequent nausea or vomiting	Nausea	SD, PH
D14. Loose or watery stools	Loose_stool	SD, YD
D15. Frequent diarrhea and bowel sound	Diarrhea_sound	PH
D16. Stomach growling as usual	Stomach_growl	SD
D17. Difficult defecation as usual	Difficult_defecation	FR
D18. Difficult voiding	Difficult_voiding	YD
Circulatory symptoms (18)		
C1. Swelling whole body	Swelling_body	SD
C2. Swelling face or eyelids	Swelling_face	YD
C3. Swelling limbs.	Swelling_limbs	SD, YD
C4. Swelling tendency of the lower limbs	Swelling_legs	YD
C5. Feeling sharp pain of body	Sharp_pain	BS
C6. Feeling fixed pain in the body	Fixed_pain	BS
C7. Nocturnal pain	Nocturnal_pain	BS

Table 1. Continued

	Abbreviation	Related syndrome differentiation
C8. Physical trauma history	Trauma	BS
C9. Bleeding symptoms or sign in body	Bleeding	BS
C10. Bruise, purpura or visible capillaries on skin	Bruise	BS
C11. Dark eyelids or lips	Dark_eye_lips	BS
C12. Scaly skin	Scaly_skin	BS
C13. Lump in skin or abdomen	Lump_skin_abdomen	PH
C14. Hard lump under the skin	Hard_lump	BS
C15. Lower abdominal pain	Belly_pain	LS, BS
C16. Coughing up with sputum	Sputum	PH
C17. Globus pharyngeus	Globus	PH
C18. Feeling heavy and pain in the chest	Heavy_pain_chest	PH

Obesity syndrome differentiation questionnaire with 5 answers: (① Never ② Rarely ③ Sometimes ④ Often ⑤ Always). Abbreviation indicates concise expression of the question for the convenience of visualization and explanation. The questions that contribute to obesity syndrome differentiation are described in the related syndrome differentiation.

FR: food retention, YD: yang deficiency, LS: liver qi stagnation, PH: phlegm, BS: blood stasis, SD: spleen deficiency.

질환, 신장질환, 간질환, 악성종양 등의 병력이 있거나 체 성분이나 대사량에 영향을 줄 수 있는 약물을 복용하거나, 약물에 과민반응이 있거나, 임신 중인 여성은 연구대상에서 제외했다. 2011년 2월 15일에 수집한 설문에 포함된 대상자는 만 19세 이상 65세 이하의 BMI 25 kg/m² 이상이며 복부둘레가 85 cm 이상인 여성으로 내분비질환, 심질환, 조절되지 않는 고혈압, 악성종양 또는 폐질환, 담석증, 신기능 장애, 간기능 장애, 공복혈당이 140 mg/dL 이상인 인슐린 비의존성 당뇨병, 신경정신질환, 협우각 녹내장, 뇌졸중이나 허혈성 심장마비 병력, 취식기능 이상 병력이 있는 자는 연구대상에서 제외했다. 또한 3개월 이내 체중에 영향을 미칠 수 있는 약물을 복용한 경우, 3개월 이내 고혈압 치료제로서 베타차단제 또는 이뇨제를 복용한 경우, 중추신경계용 약 또는 중추성 체중감량제를 투여받고 있는 경우, 체중감량을 위해 외과적 수술을 받은 경험이 있거나 임신 상태이거나 수유 중, 혹은 임신 계획이 있는 여성, 연구 시작 전 1개월 이내에 다른 시험약을 복용한 경험이 있거나, 6개월 이내에 이전 체중의 10% 이상 감량한 자, 흡연자는 연구 대상에서 제외했다.

본 연구를 개시하기 전 동국대학교 일산한방병원 기관생명윤리위원회로부터 연구진행에 대한 동의 취득 면제 및 심사 면제 승인을 받았다.

2. 자료 전처리

한의비만변증 설문지, BMI, 성별에 대한 정보 중 결측

Fig. 1. Flow chart of study.

치가 있는 대상자는 분석에서 제외했으며, 모든 설문 항목에 1 또는 2점으로 응답했거나 한 개의 문항만 3점이고, 나머지 51개 문항에서 1 또는 2로 답한 대상자는 분석에서 제외했다(Fig. 1).

3. 탐색적 데이터 분석

1) 통계 분석

통계 검증 및 시각화는 Python 3.7 (python) 컴퓨터 언어를 통해 SciPy, Pandas, Numpy, Matplotlib, Seaborn 라이

브러리를 이용했다¹⁰⁾. 성별에 따른 BMI 차이 비교 시 맨 휘트니 U 검정(Mann-Whitney U test)을 시행했다. 6개의 변증 그룹 간의 차이 비교 시 샤피로 검정(Shapiro test)에서 정규분포를 따르지 않아 비모수적 방법인 크루스칼 월리스 검정(Kruskal-wallis test)을 시행했다. 사후분석은 던 검정(Dunn test)을 이용했고, 본페로니 교정(Bonferroni correction)을 시행했으며, 유의수준 0.05 하에서 검정을 시행했다.

2) 설문지 문항 축소를 위한 특징 선택(feature selection)

한의비만변증 설문지에서는 변증 별로 10개 문항에 대한 점수를 합산하여 최고점에 해당하는 변증으로 판별하였다. 변증 레이블을 배제한 상태에서 문항 점수만 고려하여 상대적으로 중요한 문항만 선별하기 위해 분산 임계값 특징 선택(variance threshold feature selection)을 시행했다. 분산이 너무 작은 특징은 레이블 예측에 미치는 영향이 적다고 판단해서 필터링하는 방법으로, 문항에 대한 답변 값이 대상자에 따라 크게 변하지 않는 경우 변증 예측에 있어서 변별력이 낮은 문항이라고 볼 수 있다. 특징 선택을 위한 분산 임계값은 문항 별 응답의 분산 분포에서 1사분위값(25th percentile)으로 설정하고 이 값보다 작은 분산을 보이는 문항은 추가적인 기계학습에서 제외했다.

4. 데이터 마이닝 분석

1) K-평균 군집(K-means clustering): 문항 군집화

유사한 응답패턴을 보이는 문항을 동일한 군으로 분류하여 특성을 파악하기 위해 데이터 마이닝 기법 중 대표적인 군집 방법인 K-평균 군집분석을 시행했다. K-평균 군집은 K개의 군집 중심(centroid)의 위치를 임의로 선택한 뒤 그 중심들을 기준으로 군집을 구성하고, 군집 별 평균군으로 각 군집의 중심 위치를 이동시킨다. 더 이상 군집의 중심이 이동하지 않을 때까지 위의 과정을 반복하는 것으로 데이터가 분포하는 공간 내에서 거리가 가까운 개체끼리 동일한 군으로 분류한다. 거리 척도는 다양한데, 본 연구에서는 군 간 유사도 척도로 유클리드 거리(euclidean distance)를 사용했다.

설문의 변증 결과를 고려하지 않고 분산 임계값에 따라 선택한 39개의 문항을 대상으로 군집화했고, 적절한 군집 수를 결정하기 위해 클러스터 내의 총 변동을 설명하는 그룹 내 분산(within cluster sum of squares, WCSS)

을 이용해 군집 개수 K에 따라 WCSS 값이 가장 작아지는 K값을 구했다. 최적의 군집수 설정을 위해 Yellowbrick 패키지를 이용했고, K-평균 군집 분석은 Scikit-learn 라이브러리에서 제공하는 KMeans 클래스를 이용했다¹¹⁾. 최적의 K값은 3, 초기 centroid 선정 방법은 KMeans++, 최대 반복횟수는 1,000, 재현성을 위해 난수 초기값은 200으로 설정했다.

2) 계층적 군집화(hierarchical clustering)

군집의 수를 설정하지 않은 상태에서 문항 사이의 구조적 관계를 파악하고 K-평균 군집과의 일관성을 확인하고자 계층적 군집화를 시행했다. 문항이 속한 변증에 대해 고려하지 않고, 응답값 사이의 유클리드 거리를 이용하여 군집 결과를 나무 형태로 나타내는 덴드로그램(dendrogram)으로 시각화했다. Seaborn 라이브러리의 clustermap 함수를 이용했다.

3) 주성분 분석(principal component analysis, PCA)을 이용한 차원 축소(dimensionality reduction)

PCA는 다변량 자료를 분석할 때 변수가 보이는 변화(variation)를 최대한 반영하면서 자료의 차원을 줄이는 방법이다. 분산을 극대화하는 순서대로 주성분을 뽑아내는 데, 본 연구에서는 우선 K-평균 군집분석의 결과를 시각화하기 위해 2개의 주성분을 추출하는 PCA를 시행했다. 두 번째로는 대상자들의 응답에 따른 분포를 파악하기 위해 2개의 주성분을 추출하여 산점도로 나타내고, 대상자들 분포에 대한 문항의 영향을 주성분인자 행렬도(principal component factor biplot)로 시각화했다. Scikit-learn 라이브러리의 PCA 클래스를 이용했고, 시각화에는 Matplotlib, Seaborn 시각화 툴을 이용했다. 식적, 간울, 및 기타 변증 대상자의 분포가 통계적으로 유의하게 구분되는지 파악하기 위해 R vegan package의 Adonis (analysis of variance using distance matrices) 함수로 분석했고, 비유사성 거리는 Bray-Curtis 지수를 사용했다.

결과

1. 한의비만변증 설문지 변증 분포

219명의 응답자 중 결측치가 있거나(n=4), 모든 문항에 대한 응답이 1점 혹은 2점이거나(n=6), 한 가지 문항에 대

Fig. 2. Histogram of frequency distribution of obesity syndrome differentiation. The x-axis shows obesity syndrome differentiation and the y-axis is the frequency. The legend in the upper right corner describes the gender with the number. FR: food retention, YD: yang deficiency, LS: liver qi stagnation, PH: phlegm, BS: blood stasis, SD: spleen deficiency.

한 응답만 3점이고 나머지 51개 문항에 대한 응답이 1점 혹은 2점인 대상자(n=6)를 제외한 203명을 대상으로 추가 분석을 수행했다(Fig. 1). 전처리 이후 분석에 포함된 대상자 별 응답 평균값의 분포는 평균 2.21, 표준편차 0.60 이었고, 최대값은 4.35, 최소값은 1.2였다.

전체 대상자의 성별은 여성 182명, 남성 21명이고, 체질량지수(BMI) 평균 및 표준편차는 여성에서는 $28.75 \pm 2.70 \text{ kg/m}^2$, 남성에서는 $29.34 \pm 2.71 \text{ kg/m}^2$ 이었다. 맨휘트니-U 검정에서 유의수준 0.05 하에서 통계적으로 성별에 따른 BMI의 유의한 차이는 없었다($P=0.28$). 식적, 간울, 양허, 비허, 담음, 어혈로 구성된 6가지 비만변증의 분포 상 간울이 33%, 식적이 27%, 양허가 24%로 세 가지 변증이 전체 대상자의 84%를 차지했다. 이외에 비허는 7%, 담음은 5%, 어혈은 3%로 상대적으로 빈도가 적게 나타났다(Fig. 2). 남성 21명 중에서 식적이 10명, 간울이 9명, 비허가 2명이었고 양허, 담음, 어혈 변증은 0명으로, 남성 중 약 47%는 식적, 약 43%는 간울 변증으로 분류되었다.

2. 변증에 따른 체질량지수(BMI) 비교

6가지 변증 별 BMI 분포의 차이를 확인하기 위해 우선 정규성 평가를 위해 사피로 검정을 시행했고, 정규분포를 가정할 수 없어 비모수적 방법인 크루스칼 월리스 검정을 시행했다. 유의수준 0.05 하에서 P값 0.03으로 통계적으로 변증 간 BMI 평균에 유의한 차이를 보였다. 사후분석

Fig. 3. Boxplot of BMI by obesity syndrome. The central line in each box indicates the median, lower and upper limits of the box indicate 25th and 75th percentiles, and whiskers indicate minimum and maximum of observed values. BMI: body mass index, FR: food retention, YD: yang deficiency, LS: liver qi stagnation, PH: phlegm, BS: blood stasis, SD: spleen deficiency. The difference between subjects by obesity syndrome differentiation were significant in Kruskal Wallis test ($P=0.03$). ** refers significant difference between two groups after using Dunn post hoc test ($P=0.015$, with adjusted P-value using Bonferroni).

으로 던 검정을 통해 개별 그룹 간 비교에서 양허와 비허 변증 간의 BMI가 본페로니 방법으로 조정된 P값 0.015로 통계적으로 유의한 차이가 있었으며, 기타 그룹 간 통계적으로 유의한 차이는 없었다(Fig. 3).

3. 특징 선택: 설문 문항 축소

특징 선택을 위해 분산 임계값을 설정하기 위해 문항에 대한 응답 값의 분산 분포를 파악했다(Appendix 1). 문항 별 응답의 분산 분포에서 1사분위값(25th percentile)인 0.807로 임계값을 설정했고, 0.807보다 낮은 분산을 가지는 문항은 13개였다(Table 2). 제외된 13가지 문항 중 7개 문항(야간통, 고정된 부위의 통증, 찌르는 듯한 통증, 검은 눈 주위나 입술, 외상력, 몸의 고착화된 덩어리, 출혈 증상)은 어혈 변증에 기여하는 문항이었다. 이외에 간울(옆구리 부근 팽팽한 통증), 식적(자주 체함), 담음(머리 어지럼증, 피부나 뱃속의 덩어리, 가슴이 답답하고 통증 있음), 양허(창백한 얼굴색)에 기여하는 문항도 낮은 분산을 갖는 것으로 파악됐다. 분산이 낮은 13개 문항의 평균 점수는 모두 2 미만으로 문항에 대한 점수가 대상자에 따라 크게 변하지 않으면서 응답 평균값도 낮아 비만변증에

Table 2. Variance of Questions and Feature Selection by Variance Threshold

Selected questions (39 features)						Excluded questions (13 features)	
Question	Variance	Question	Variance	Question	Variance	Question	Variance
Cold_body	1.716	Cold_sweats	1.302	Body_heavy	1.138	Nocturnal_pain	0.806
Swelling_legs	1.652	Angry	1.295	Worry	1.121	Dizziness	0.804
Gassy	1.646	Difficult_defecation	1.291	Heavy_chest	1.092	Flank_pain	0.756
Stress_dyspepsia	1.612	Poor_digestion	1.286	Nausea	1.065	Fixed_pain	0.756
Swelling_limbs	1.585	Sensitive	1.271	Difficult_voiding	1.042	Upset_stomach	0.746
Hate_cold	1.579	Belly_pain	1.213	Depressed	0.992	Heavy_pain_chest	0.729
Swelling_body	1.528	Loose_stool	1.190	Sputum	0.968	Sharp_pain	0.707
Stress_pain	1.470	Overeat	1.173	Globus	0.964	Lump_skin_abdomen	0.691
Indigestion	1.463	Eat_less	1.170	Woozy	0.961	Dark_eye_lips	0.687
Swelling_face	1.419	Bruise	1.163	Diarrhea_sound	0.878	Trauma	0.625
Food_stagnation	1.345	Belching	1.163	Stomach_growl	0.842	Hard_lump	0.611
Sternum_pain	1.309	Bloated	1.159	Scaly_skin	0.834	Pale	0.415
Greedy_eat	1.302	Tired	1.143	Tasteless	0.808	Bleeding	0.287

Variance threshold was implemented for question (feature) selection. We assume that variables with higher variance may contain more useful information for obesity syndrome differentiation. Variable threshold was set at <0.807 (the value of 25th percentile of the variance).

Table 3. Statistical Characteristics of Excluded Variables

Variables	Nocturnal_pain	Dizziness	Flank_pain	Fixed_pain	Upset_stomach	Heavy_pain_chest	Sharp_pain	Lump_skin_abdomen	Dark_eye_lips	Trauma	Hard_lump	Pale	Bleeding
Mean	1.512	1.946	1.488	1.488	1.522	1.665	1.502	1.340	1.498	1.350	1.310	1.291	1.123
Standard deviation	0.898	0.897	0.870	0.870	0.864	0.854	0.841	0.831	0.829	0.790	0.782	0.644	0.535

있어서 중요도가 낮은 문항이라고 판단하고 추가 분석에서 제외했다(Table 3).

4. 설문 응답에 따른 설문항목 군집화

특징 선택 이후 39개의 문항에 대한 답변을 비지도학습법인 K-평균 군집분석을 이용해 군집화한 결과, 세 가지 군집으로 구분되었고 주성분 분석으로 차원을 축소하여 2차원 공간에 시각화했다(Fig. 4). 군집 1 (cluster 1)은 17개 문항으로 구성되며 잦은 트림, 속이 거북하고 구역감 느낌, 위장의 물소리, 자주 설사하며 배에서 소리가 남, 무른 대변 혹은 설사 경향, 배변이 힘든 편, 소변이 시원하지 않음, 입맛이 없음, 식사량이 적음, 정신이 흐릿하거나 머리가 무거움, 헛 땀, 목의 이물감, 기침이 나고 가래가 많음, 흉골을 눌렀을 때 통증을 느낌, 아랫배 통증, 피부에 멍·자반·실핏줄이 드러남, 피부가 비늘모양으로 갈라짐과 같은 증상이 포함됐다. 군집 2 (Cluster 2)는 11개 문항으로 스트레스 시 소화불량 심해짐, 배가 더부

룩함, 몸이 차거나 손발이 차가움, 배에 가스가 많이 차는 경향, 소화가 잘 안되는 경향, 자주 체함, 음식이 위장에 정체된 느낌 외에 팔다리가 부음, 하체가 더 부음, 온 몸이 전체적으로 부음, 얼굴이나 눈이 부음과 같은 신체 부종과 관련된 문항들이 포함되었다. 군집 3 (Cluster 3)은 11개 문항으로 구성되며 과식 경향, 배가 불러도 음식을 먹는 경향 외에 몸이 무거움, 힘이 없고 피곤함, 추위를 싫어함, 민감한 성격, 고민·걱정이 많음, 쉽게 화가 남, 자주 우울함, 스트레스받으면 통증 심화됨, 가슴이 답답함과 같은 정서 심리적인 증상과 관련된 문항이 포함됐다. 각각의 군집은 변증에 기여하는 항목 별로 뚜렷하게 구분되지는 않았다.

군집 1에는 특성 선택 이후 남아있던 어혈 변증에 기여하는 3가지 문항과 담음 변증에 기여하는 6가지 문항이 포함되었다. 이 외에도 식적, 비허, 양허, 간울에 기여하는 항목들이 포함되어 있었는데 배변이나 배뇨의 어려움, 변이 무른 경향이나, 헛 땀을 흘리는 경향과 같은 배설과

Fig. 4. K-means clustering of 39 selected questions after principal component analysis (PCA). The first two principal components were plotted and the number of clusters was set to K=3. Each data point indicates the question of obesity syndrome questionnaire. The red x indicates the centroid of the cluster. Different colors of datapoints are labeled according to their clusters by K-means clustering; green circle (Cluster 1); orange circle (Cluster 2); blue circle (Cluster 3). Different shapes of datapoints are labeled according to the syndrome to which the questions contribute. FR: food retention, YD: yang deficiency, LS: liver qi stagnation, PH: phlegm, BS: blood stasis, SD: spleen deficiency.

관련된 증상이 특징적이었다. 군집 2의 경우 비허, 양허와 관련된 신체의 붓는 증상이나 음식이 위장에 정체된 양상으로 나타나는 소화불량 증상 위주의 식적 관련된 증상으로 구성됐지만 간울, 답음에 기여하는 문항도 있었다. 군집 3은 정서·심리적인 항목들로 간울에 기여하는 항목이 위주이긴 했으나 비허, 식적, 양허, 답음에 기여하는 항목들이 포함되었다. 이처럼 변증 결과를 배제한 상태에서 대상자들의 응답을 기반으로 분석했을 때 특정 변증에 기여하는 증상들끼리 비슷한 응답을 보이기보다는 여러 변증에 관여하는 증상들이 혼재되어 나타나는 것을 확인할 수 있었다.

이어서 계층적 군집화를 통해 문항 사이의 구조적 관계를 파악하고 K-평균 군집분석 결과와의 일관성을 확인

했다(Fig. 5). 텐드로그램 상에서 가까운 가지일수록 아랫단에서 합쳐지고, 위로 올라갈수록 다른 가지와 합쳐진다. 가장 뚜렷하게 구분되는 문항은 군집 3에 속하는 과식 경향, 배가 불러도 음식을 먹는 경향이었다. 이 외에 군집 3에 속하는 문항은 비교적 가까운 위치에서 군집화되었다. 군집 3에 속하는 추위를 싫어함과 군집 2에 속하는 몸이 차거나 손발이 차가움이 가까운 군집으로 분류되었다. 군집 1도 식사량이 적음, 흉골을 눌렀을 때 통증을 느낌 외에는 모두 아랫단에서 가깝게 합쳐졌다. 군집 2에 속하는 문항은 텐드로그램 상에서 크게 두 갈래로 갈라져서 나타났다. 한 묶음은 소화와 관련된 문항으로 구성됐고, 다른 그룹은 신체가 붓는 증상 및 배에 가스가 차는 증상으로 구성되었다.

Fig. 5. Heatmap with dendrogram of 39 selected questions based on the Euclidean distance. The y-axis represents 39 selected questions and x-axis represents each subject. The colors in the heatmap represent answers range from 1 to 5 (shown in color bar). The circles in the x-axis indicate different clusters resulted from K-means clustering; GREen circle (Cluster 1); orange circle (Cluster 2); blue circle (Cluster 3).

2에서 배에 가스가 많이 차는 경향은 식적에 기여하는 문항인데 비허 변증 비판환자 절반 이상이 4점으로 답했다. 특히 군집 3에서 몸이 무겁게 느껴진다는 문항에 대해서는 비허 변증의 73%가 4, 5점으로 답했다. 몸이 무겁게 느껴진다는 문항은 양허, 비허, 담음 모두에 기여하는 문항으로 특히 비허 변증 대상자에서 4, 5점 응답 비율이 높았다.

담음 변증은 간울 변증과 비슷하게 군집 3에서의 4, 5점 응답 비율이 32%로 제일 높고, 이어서 군집 2에서 평균 23%, 군집 1에서는 14%로 나타났다. 군집 1에 담음 변증에 기여하는 문항이 6가지 포함되어 있는데, 오히려 군집 1에서의 4, 5점 응답 비율은 낮았다. 어혈 변증은 군집 별 4, 5점 평균 응답 비율이 각각 군집 1에서 29%, 군집 2에서 39%, 군집 3에서 29%였다. 어혈 변증에 기여하는 문항 외에도 간울 변증에 기여하는 흉골 눌렀을 때의 통증, 스트레스 시 소화가 잘 안됨, 가슴이 답답함, 스트레스 시 통증이 심해지는 항목의 응답 평균이 3 이상이고, 식적 변증에 기여하는 가스가 잘 참, 잘 체함과 같은 항목의 응답 평균도 3이 넘었다. 이외에도 양허 변증에 기여하는 얼굴, 사지가 붓는 증상의 응답 평균도 3점이 넘었다.

문항 별로 살펴보면 과식이나 배불러도 먹는 경향에 대한 문항은 어혈, 비허 변증을 제외하고는 기타 변증에서 4, 5점의 비율이 30% 이상으로 나타났다. 이외에도 격

정이 많음, 추위를 싫어함, 가스가 잘 참, 몸이 무거움, 피곤함, 팔다리가 부음과 같은 증상에서는 변증에 상관없이 30% 이상의 대상자가 4, 5점으로 답변했다.

6. 차원 축소를 이용한 대상자의 분포 파악 및 대상자와 변수 사이의 관계

대상자들의 분포와 대상자와 변수 사이의 관계를 파악하기 위해 문항에 대한 응답을 바탕으로 PCA를 수행하여 2개의 주성분을 추출해 2차원 그래프로 시각화했다(Fig. 7a). 주성분 1 (principal component 1, PC1)은 전체 분산의 31.3%를 설명하고, 주성분 2 (principal component 2, PC2)는 전체 분산의 6.68%를 설명했다. 다빈도 변증인 식적과 간울과 기타 변증 대상자 간의 분포가 구분되었다(Fig. 7a). PC2를 기준으로 식적 변증 대상자는 양의 방향에, 간울 변증 대상자는 음의 방향에 주로 분포하고 있었으며, 나머지 4가지 변증으로 분류된 대상자는 간울 변증과 식적 변증 대상자 사이에 주로 분포했다. 간울, 식적, 기타 변증 대상자들의 분포는 Adonis 분석 결과 통계적으로 유의하게 차이가 있었다.

대상자들의 분포에 영향을 주는 문항의 기여도를 파악하기 위한 PCA 행렬도에서 PC2를 기준으로 대상자들이 양의 방향, 수평 방향, 음의 방향으로 분포하는 데 영향을 주는 문항들의 특성에 차이가 관찰됐다(Fig. 7b). 양의 방

Fig. 7. PCA and PCA biplot of subjects (a) The first two principal components were plotted. Each data point indicates subject. Different colors of datapoints are labeled according to their obesity syndrome differentiation; Orange circle (FR); sky blue circle (LS); grey circle (other obesity syndromes except FR, LS). (b) PCA biplot of the subjects showing the loading of 20 questions which mainly affect the distribution (arrows). The length of arrows approximates the variance of the questions, whereas the angles between the questions approximate their correlations. Closed data points have similar scores on the PCA components. PCA: principal component analysis, FR: food retention syndrome, LS: liver qi stagnation syndrome.

향으로는 주로 소화 관련 증상(스트레스 시 소화불량 심해짐, 소화 잘 안되는 경향, 속이 더부룩함)이 영향을 주고 있었고, 수평 방향으로로는 식사를 적게 하거나, 추위를 싫어하고, 신체부위가 붓는 증상 등이 주요한 영향을 주었다. 음의 방향으로로는 주로 배불리도 먹거나, 과식하는 경향 외에 민감한 성격, 걱정이 많고, 쉽게 화가 나거나 스트레스 시 통증이 심해지는 등의 정서적인 증상과 관련된 항목들이 주로 영향을 주는 것으로 파악됐다.

고찰

본 연구에서는 데이터 마이닝 기법을 이용해 비만환자를 대상으로 수집한 한의비만변증 설문 조사 결과를 심층 분석했다. 설문 문항의 중요도를 재평가하고, 실제 임상 현장을 반영해 기존에 6개로 분류하던 한의비만변증의 효율화 가능성을 확인했다. 대상자의 비만변증 분포를 확인한 결과 간음, 식적, 양허 순으로 많았고 전체 대상자의 84%를 차지했으며, 이어서 비허, 담음, 어혈 순으로 분포했다. 이러한 경향성은 기존의 연구 결과와 일치한다⁸⁾.

분석에 포함된 203명 중 여성이 182명, 남성이 21명으로 성별에 따른 BMI 차이는 없었으나 변증의 경우 남성에서는 양허, 담음, 비허 변증에 해당하는 대상자가 없었다. 성별에 상관없이 전체 대상자 중 비허, 어혈, 담음 변증이 각 20명 미만으로 표본 수가 충분하지 않았던 점을 미루어 보아 추후 대규모 데이터를 대상으로 본 연구에서의 분석 방식을 적용해 결과의 일관성을 탐구할 필요가 있다. 또한 변증 별로 BMI를 비교했을 때 양허 변증과 비허 변증 사이에 유의한 차이가 관찰되었다. 비허 변증은 15명, 양허 변증은 49명으로 비허 변증 대상자가 상대적으로 표본 숫자가 적었기 때문에 허증 변증으로 분류되는 두 변증 간의 BMI 차이가 일관되게 나타나는지 대규모 집단에서의 데이터를 분석해볼 수 있다.

비만변증에 있어서 변별력이 적은 문항을 축소하여 임상현장에서 활용도 및 편의성을 높이고자 특성 선택을 수행했다. 대상자에 따른 응답 점수의 차이가 작고, 평균 응답 점수가 2점 미만으로 낮은 13개의 문항 중 7개의 문항이 어혈 변증에 기여하는 문항으로 파악되었다. 어혈 변증은 전체 대상자의 3%로 빈도가 낮고 비만환자를 대상으로 조사한 응답 결과에서도 어혈 변증에 기여하는 문항은 4, 5점의 응답비율이 저조했다. 기존에 비만을 치료한

경험이 있는 한의사들이 1,254명의 과체중 환자를 대상으로 시행한 비만 설문지 응답에 따라 변증을 시행했던 연구에서도 어혈 변증의 비율이 약 5%로 가장 낮게 나타났다¹²⁾. 또한 39개의 문항에 대한 군집 분석 시 어혈 변증으로 판별된 대상자들은 어혈과 관련이 없는 문항에 대해서도 대체적으로 높은 점수로 답하는 경향이 있었다. 따라서 어혈 변증을 별도의 변증으로 구별하기보다는 전신의 순환, 소화 및 정서기능 등과 관련된 문제를 여러 가지 겹하고 있는 상태로 파악하고, 비만변증을 개편한다면 어혈 변증 제의를 고려해볼 필요가 있을 것으로 생각된다. 그러나 비만환자에서 어혈 변증의 비율이 낮더라도 비만 치료에 있어서 임상적 중요도가 높을 수 있으므로 비만에서 어혈 변증의 중요성에 대해 한의비만 전문가의 평가 및 논의가 필요하다.

비지도학습 방법인 K-평균 군집화 방법을 이용해 변증 결과를 고려하지 않은 상태에서 비슷한 점수 패턴을 보이는 문항을 군집화했을 때 크게 세 그룹으로 구분할 수 있었다. 우선 소화기 증상 중에서도 물소리나 대소변 배설 관련 문제 및 입맛이 없고 식사량이 적으며 어혈 증상에 관련된 아랫배 통증, 비늘모양으로 갈라지는 피부, 피부에 멍·자반·실핏줄이 잘 나타나는 등의 증상, 담음과 관련된 머리가 무겁고 목의 이물감이 있으며 기침이나 가래 많은 증상 등이 군집 1로 분류됐다. 군집 2는 정체된 경향성을 보이는 소화기 증상(자주 제함, 더부룩함, 가스가 많이 차는 경향 등)과 신체가 붓는 증상을 위주로 분류되었다. 마지막으로 군집 3은 과식 경향성과 정서 및 심리적인 증상 관련된 문항으로 구성된다. 계층적 군집화를 시행한 덴드로그램 상에서 문항 사이의 구조적 관계를 확인할 수 있었고, 추후 설문 문항 중에서 응답 패턴이 비슷해서 변증과 관련된 정보가 중복되는 경우 전문가와의 합의를 통해 문항을 간소화하여 임상에서 활용도를 높일 수 있을 것이다.

변증에 따라 세 가지 군집 별로 응답 중 4, 5점의 비율을 기준으로 증상 심화도를 파악했다. 식적 변증은 과식 경향, 배불리도 먹는 경향 외에는 모든 문항에서 4, 5점 비율이 30% 미만이었었고, 평균 응답 점수도 두 문항 외에는 3 미만으로 나타났다. 식적 변증 대상자는 비만 설문에서 조사하는 전신 증상, 소화기 증상, 정서 증상, 순환기 증상을 겹치는 비율이 적다고 해석할 수 있다. 양허와 비허 변증 대상자는 군집 2의 문항과 관련된 증상에서의

심화도가 높았는데 양허 변증에서는 추위에 관련된 문항에서 4, 5점 비율이 높았고, 비허 변증은 기존에 식적에 기여하는 것으로 분류된 가스가 잘 차는 증상 및 음식이 정제된 느낌에 대한 4, 5점 응답 비율도 높았다. 담음과 간을 변증의 경우 군집 3 문항과 관련된 증상 심화도 비율이 높았다. 군집 3에 속한 문항 대부분이 간을 변증에 기여하기 때문에 간을 변증에서의 비율이 높다는 것은 기존 지식과 일관된 결과이다. 그러나 담음에 기여하는 변증 중 6가지는 군집 1에 포함되어 있는데 담음 변증 대상자의 군집 1 문항의 4, 5점 응답 비율은 낮았고, 오히려 정서와 관련된 문항에서의 4, 5점 비율이 상대적으로 높았다. 어혈 변증 대상자는 다른 변증에 기여하는 문항에서도 응답 평균이 3 이상인 경우가 많아 전신증상, 순환, 소화, 정서 관련 증상을 여러 가지 검하고 있는 것으로 파악됐다.

설문 응답에 따른 대상자들의 분포가 변증 별로 뚜렷하게 구분되는지를 파악하기 위해 차원 축소 기법 중 PCA를 이용했고, 다빈도 변증인 식적, 간울과 기타 변증으로 크게 세 그룹으로 구분되어 분포하는 것을 확인할 수 있었다. 차원 축소 후 식적 변증과 간울 변증 및 기타 변증 대상자의 분포가 유의하게 구분되는 양상을 보였는데, 기존 연구에서도 비만 환자의 간울, 식적, 양허 변증에 따라 방풍통성산과 방기황기탕 투여 후의 치료 효과 차이 및 부작용 발생 빈도의 차이가 관찰된 바 있다¹³⁾.

PCA 결과에서 각 문항들이 대상자의 분포에 미치는 영향을 파악한 결과, 대상자의 주관성이 관여하는 설문의 특성 및 여러 요인이 관여하여 발생하는 비만의 특성 상 여러 변증과 관련된 증상들이 혼재되어 나타났다. 큰 틀에서는 소화기 위주의 증상을 호소하는 유형과 식욕조절 이상 및 정서적인 증상을 위주로 호소하는 유형, 추위를 타거나 몸이 붓는 등의 대사 기능 저하를 나타내는 유형으로 분류할 수 있었다. 그러나 한의학적 지식을 바탕으로 판단했을 때 기존의 변증 체계 내에서 경계가 명확하게 구분되지는 않았다. 또한 PCA 분석 결과 PC1이 전체 자료 분산의 약 31.3% 정도만 설명할 수 있었던 점을 미루어 보아 설문 외에도 비만환자의 특성을 반영할 수 있는 객관적인 지표 탐색이 필요하다. 또한 연구에 포함된 어혈, 담음, 비허 변증 대상자의 표본 수가 각 20명 미만이었기 때문에 대규모 데이터에서 데이터 마이닝 분석 방법을 활용해 결과의 일관성을 확인해볼 필요가 있다. 또

한 주관적인 평가인 설문조사 외에도 설진, 맥진과 같은 한의 검진 및 생화학 검사, 장내미생물, 유전체다형성 등의 객관적인 지표를 발굴하여 데이터 마이닝 알고리즘을 통해 한의비만변증 체계를 객관화, 고도화하는 데 활용할 수 있을 것이다.

결론

본 연구에서는 비만자를 대상으로 수집한 실제 설문 응답을 기반으로 탐색적인 데이터 분석 및 데이터 마이닝을 통한 심층 분석을 통해 설문지 문항을 재평가하고 한의비만변증 효율화 가능성을 확인했다. 응답 분산에 기반하여 축소시킨 39개의 문항을 선택하여 K-평균 군집화 및 계층적 군집화를 통해 변증을 배제했을 때 비슷한 응답패턴을 보이는 문항들의 특성을 파악했다. 결과에서 6가지의 변증 패턴으로 구분되기보다는 문항 특성에 따라 소화기 증상, 정서 심리적인 증상, 대사 기능 저하 증상으로 구분할 수 있었다. 또한 대상자의 응답에 따라 차원 축소한 후 분포에 영향을 미치는 문항을 파악했을 때 소화기 증상, 정서 심리적인 증상, 대사 기능 저하 증상이 각기 다른 방향으로의 분포 패턴을 만드는 데 영향을 주는 것으로 나타났다.

비만변증 설문지 문항을 실제 설문 응답을 바탕으로 재평가하기 위해 통계적 방법을 이용해 실제 비만환자가 많이 겪지 않는 증상을 파악했다. 또한 변증을 배제한 상태에서 데이터 마이닝 분석을 통해 기존의 6가지 변증의 단순화 가능성을 확인했다. 어혈 변증의 경우 기존 연구들과 본 연구에서 모두 전체 대상자에서의 비율이 일관되게 낮았다. 또한 어혈 변증과 관련된 설문 문항 중 7개의 문항이 전반적으로 낮은 평균 점수 및 분산을 보였다. 그러나 비만환자에서 어혈 변증의 비율이 낮다고 하더라도 비만 치료에 있어서 임상적 중요도가 높을 수 있으므로 한의비만 전문가의 평가 및 논의가 필요하다. 또한 어혈, 담음, 비허 변증 대상자의 표본 수가 각 20명 미만이었기 때문에 추후 대규모 설문자료에서 결과의 일관성을 확인해볼 수 있다.

앞으로 본 연구에서 활용한 데이터 마이닝 방식을 이용해 비만변증 설문 응답을 재평가하고, 주관성이 관여하는 설문 외에도 설진, 맥진과 같은 한의 검진 및 생화학 검사, 장내미생물, 유전체다형성 등의 객관적인 지표를

발굴하여 한의비만 전문가와의 합의 및 임상연구를 통해 한의비만변증 체계를 객관화, 고도화하는 데 활용할 수 있을 것이다.

감사의 글

본 연구는 한국연구재단의 중견연구자지원사업(NRF-2021R1A2C1011087)과 보건산업진흥원의 한의약혁신기술 개발사업(HF20C0020)의 연구비 지원으로 이루어졌음.

References

1. Korea Disease Control and Prevention Agency. Korean national health and nutrition examination survey: 2019 national health statistics. Cheongju : Korea Disease Control and Prevention Agency. 2020 : 153-5.
2. Moon JS, Kang BG, Choi SM. A study of syndrome index differentiation in obesity. *Journal of Korean Medicine for Obesity Research*. 2007 ; 7(1) : 55-69.
3. Kang BG, Moon JS, Choi SM. A reliability analysis of syndrome differentiation questionnaire for obesity. *Korean Journal of Oriental Medicine*. 2007 ; 13(1) : 109-14.
4. Moon JS, Kang BG, Kang KW, Choi SM. Weighting method based on experts opinions for obesity syndrome differentiation questionnaire. *Journal of Society of Korean Medicine for Obesity Research*. 2008 ; 8(1) : 53-61.
5. Chung WS, Hwang MJ, Lee AR, Moon JS, Choi SM, Song MY. The difference of syndrome differentiation patterns between premenopausal and climacteric obese Korean women. *Journal of Society of Korean Medicine for Obesity Research*. 2008 ; 8(2) : 37-47.
6. Cho YJ, Lee AR, Hwang MJ, Song MY. Relationship between oriental obesity pattern, life habitual factors and psychological factors in Korean obese and overweight women. *Journal of Society of Korean Medicine for Obesity Research*. 2011 ; 11(2) : 15-24.
7. Kim EJ, Lee AR, Hwang MJ, Cho JH, Choi SM, Chung SH, et al. Relationship between visceral adipose tissue and oriental obesity pattern identification in obese Korean women. *Journal of Korean Medicine Rehabilitation*. 2011 ; 21(2) : 279-88.
8. Kang KW, Moon JS, Kang BG, Kim BY, Choi SM. The comparison of pattern identification diagnosis according to symptom scale based on obesity pattern identification questionnaire. *Journal of Society of Korean Medicine for Obesity Research*. 2009 ; 9(1) : 37-44.
9. Agarwal S. Data mining: data mining concepts and techniques. 2013 International Conference on Machine Intelligence and Research Advancement. 2013 : 203-7.
10. Van Rossum G, Drake FL. Python 3 reference manual. Scotts Valley, CA : CreateSpace. 2009.
11. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*. 2011 ; 12 : 2825-30.
12. Kang KW, Moon JS, Kang BG, Kim BY, Kim NS, Yoo JH, et al. The discrimination model for the pattern identification diagnosis of overweight patients. *Korea Journal of Oriental Medicine*. 2008 ; 14(2) : 41-6.
13. Park J, Lee M, Kim H, Hong S, Lee D, Yoo J, et al. Efficacy and adverse events of Bangpungdongseong-san (Bofutsusho-san) and Bangkihwangki-tang (Boiogiot-tang) by oriental obesity pattern identification on obese subjects: randomized, double blind, placebo-controlled trial. *J Korean Med Obes Res*. 2011 ; 12(2) : 265-78.

Appendix 1. Box plot of variance of 52 questions (features). The central line in each box indicates the median, lower and upper limits of the box indicate 25th and 75th percentiles, and whiskers indicate minimum and maximum of the variance of features.