Original Article

J Sasang Constitut Med 2017;29(3):242-247
https://doi.org/10.7730/JSCM.2017.29.3.242

한의임상정보은행 활용도 제고를 위한 교육용 데이터 개발

백영화 · 이시우*

한국한의학연구원 미병연구단

Abstract

Development of Korean Medicine Data Center(KDC) Teaching Dataset to Enhance Utilization of KDC

Younghwa Baek · Siwoo Lee*

Mibyeong Research Center, Korea Institute of Oriental Medicine

Objective

Korean medicine Data Center (KDC) has established large-scale biological and clinical data based on Korean medicine to demonstrate and validate its theory. The aim of this study was to develop KDC teaching dataset and user guideline to improve utilization of the KDC.

Method

KDC teaching dataset were selected using stratified random sampling according to the Sasang constitution (SC). This dataset included 72 variables of 500 sample subjects. The user guideline described how to conducted eight statistical analysis methods using the teaching dataset.

Results

The KDC teaching dataset was sampled from 200(40%) Taeeumin, 125(25%) Soeumin, and 175(35%) Soyanain. It was consisted of questionnaire (basic, habit, disease, symptom), physical exam (body measurement, blood pressure), blood exam, and expert` SC diagnosis. The usage guidelines provided instruction for users to perform several statistical analysis step by step with KDC teaching dataset.

Conclusion

We hope that our results will contribute to enhancing KDC utilization and understanding.

Key Words : Korean medicine Data center (KDC), teaching dataset, user guideline

Received September 14, 2017 **Revised** September 14, 2017 **Accepted** September 21, 2017

Corresponding author Siwoo Lee
Mibyeong Research Center, Korea Institute of Oriental Medicine, 1672 Yuseongdaero, Yuseong-gu, Daejeon, South Korea
Tel : +82-42-868-9555 Fax : +82-42-869-9388 E-mail : bfree@kiom.re.kr

Ⅰ. 緖論

대규모의 임상정보는 만성질환의 원인지표와 질병-유전, 질병-환경 간의 상호작용을 파악하여 개인 감수성을 고려한 개별화된 건강관리 방법을 제시하고 질병의 관리 정책 수립에 중요한 역할을 한다[1]. 또한 바이오뱅크의 활성화는 Clinical trial 비용 감소, 진단·치료 비용 감소 등 경제적 이익 효과를 보이고 있으며[2], 세계 바이오뱅크 관련 산업은 2013-2018년간 연평균 성장률이 약 12%로, 향후 2023년에는 약 38조 원 규모로 예상되고 있다[3]. 이에 주요 선진국은 국가 주도형의 바이오뱅크 사업이 진행하고 있으며[1], 초창기에는 단순히 제한된 연구 목적으로 자료를 수집하였으나, 현재는 자료의 양적, 질적 확장시대에서 더 나아가 유사 정보를 연계와 호환을 통해 공유와 활용이 강조되고 있다[4]. 이를 위해 사용자가 다각도로 복잡한 자료의 흐름을 쉽게 이해할 수 있도록 사용지침서와 함께 교육용 데이터를 개발하여 제공하고 있다. 대표적인 예로 미국의 BioLINCC teaching dataset, 영국의 Health Survey for England 2011 teaching dataset, 우리나라의 경우 KoGES teaching dataset이 있다[5-7]. 이러한 교육용 데이터는 사용자로 하여금 자료에 대한 친숙도와 활용 가능성에 대한 동기 부여를 제공한다.

한의약 연구 분야에서도 국가적으로 지원을 바탕으로 연구 투자가 꾸준히 증가하는 경향을 보이고 있다[8]. 또한 한의학 관련 치료기술의 임상적 안전성과 효과성 검증 등을 목적으로 다양한 임상연구가 진행되어[9] 이를 통해 수많은 임상자료가 축적되고 있다. 그러나 단일 연구의 적은 대상자 수는 항상 연구의 제한점으로 인지되고 있으며 임상적 근거 수준을 높이기 위해서는 정보 통합으로 대규모의 자료 구축이 필요하다[10]. 그러나 현재로서는 한의계 내의 이런 연구 흐름이 부족한 상태로 한국한의학연구원의 한의임상정보은행(Korean medicine data center, 이하 KDC)이 유일하며, 진 등[11]의 연구에서 체질임상자료를 중심으로 구축된 임상정보의 현황에 대해 발표한 바 있다. KDC는 2007년부터 여러 한의임상연구를 통해 수집된 자료로 체질, 한열, 미병을 중심으로 설문임상자료, 기기자료, 혈액검사, 인체유래물 등을 포함한 한의계 최대의 임상정보은행이다. 체계적인 자료 관리 및 통합으로 약 22,000여명의 임상정보가 확보되어 있다. KDC는 신규 연구의 기초자료로 활용되어 연구기간의 단축의 효과가 예상되며, 대규모 임상자료를 활용하여 높은 수준의 연구 결과 도출 등 이차적 자원으로 활용 가능성이 높다. 그러나 KDC 인지도가 부족하고, 사용자의 자료 특성에 대한 이해가 필요하다.

따라서 본 연구는 사용자의 KDC 자료 활용과 이해를 높이기 위한 목적으로 체질임상정보를 중심으로 구성된 KDC 교육용 데이터의 개발 과정과 사용지침서를 소개하고자 한다.

Ⅱ. 硏究 方法

1. KDC 교육용 데이터의 정의

KDC 교육용 데이터는 사용자가 KDC 자료 활용 전에 자료의 특성과 성격을 파악할 수 있게, KDC 전체 자료 중에서 일부 자료를 표본 추출하여 재가공한 테스트 데이터이다.

2. 교육용 데이터 자료원

KDC 교육용 데이터는 KDC 내의 체질확진자 3,891명의 자료를 활용하였다. 체질확진자는 일정 기준의 체질처방을 복약하여 증상의 호전 정도가 평가되고, 이를 근거를 사상의학 전문의 또는 사상체질 임상경력 5년 이상의 한의사에게 체질진단을 명확히 받은 대상자를 의미한다[11,12]. 체질확진자 자료는 2007년부터 2013년까지 국내 대학 및 한방병원 등 26개 한방의료기관에서 체계적으로 수집, 관리되었으며, 다양한 체질임상정보, 계측정보, 혈액검사 정보를 포함하고 있다.

3. 교육용 데이터 표본 추출방법

KDC 교육용 데이터의 표본 추출은 모집단인 체질확진자의 체질정보 특성을 고려하여 진행하였다. 체질확진자의 체질 분포는 태음인 1,539명 (39.6%), 소음인 1,004명(25.8%), 소양인 1,267명(32.6%), 태양인 81명(2.0%)로 구성되어 있다.

표본 추출은 체질별 층화 무작위 추출법을 사용하였다. 층화 무작위 추출법은 모집단을 먼저 중복되지 않도록 층을 나누 다음 각 층에서 표본을 추출하는 방법이다. 층을 나눌 때 층내는 동질적, 층간은 이질적 특성을 가지도록 하며, 전체 모집단뿐만 아니라 각 층의 특성에 대한 추정도 할 수 있다는 장점이 있다[13]. 본 연구에서는 모집단인 체질확진자의 체질정보에 따라 3개의 층으로 나누고, 모집단의 체질 크기를 반영하여 태음인 4 : 소음인 2.5 : 소양인 3.5 비율로 배분하여 각 층별 무작위추출 방법으로 표본을 추출하였다. 태양인은 수가 적어 표본 추출에서 제외하였다. KDC 교육용 데이터 표본수는 통계적 활용 측면을 고려하여 500명으로 하였다. 5개의 표본을 생성하여, 각 표본의 성별, 나이, 신장, 체중 등의 특성을 모집단의 특성과 비교하여 가장 유사한 표본을 교육용 데이터로 선정하였다. Figure 1은 모집단과 교육용 데이터의 체질별 성별 분포의 비교이다.

4. 교육용 데이터 변수 선정

교육용 데이터의 변수는 전체 594개 변수 중에서 72개를 선정하였다. 일반사항 5개, 생활습관 2개, 성격설문 15개, 소증설문 11개, 한열설문 8개, 질병력 3개, 자가건강수준 및 증상 9개, 전문가 체질진단 1개, 체형 11개, 혈압 2개, 혈액검사 5개로 구성되어 있다. 상세내용은 Table 1과 같다.

Ⅲ. 硏究 結果

1. KDC 교육용 데이터의 일반적 특성

KDC 교육용 데이터 500명의 일반적 특성은 다음과 같다. 체질 분포는 태음인 200명(40%), 소음인 125명(25%), 소양인 175명(35%)이었으며, 여자가 67.2%로 남자 32.8% 보다 많았고 이는 체질별 성별 분포에서도 비슷한 경향을 보였다. 평균 나이는 48.07세(±16.26)이

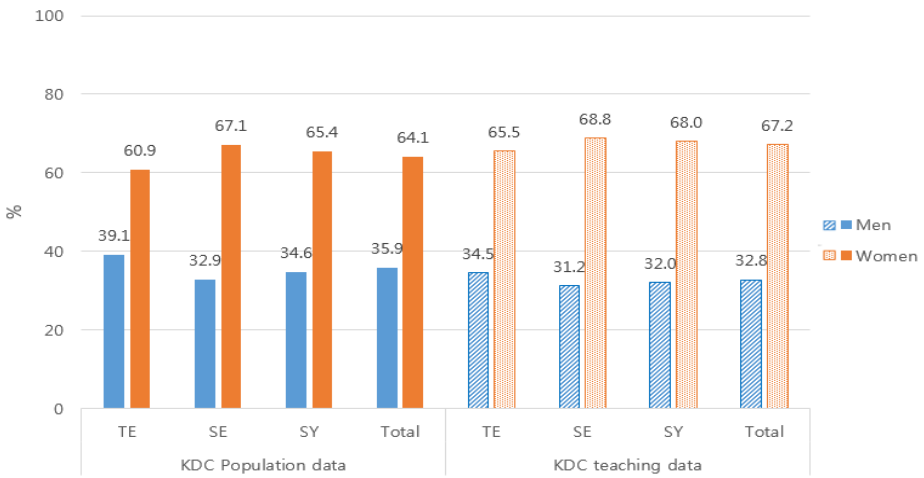

Figure 1. Comparison of sex distribution according to sasang constitution in Korean medicine Data Center(KDC) population and teaching data

며, 50대가 23.6%로 가장 많았고, 40대 19.4%, 60대 17.8% 순이었다. 직업은 기타직을 제외하고 사무직이 26%로 가장 많았고, 교육 정도는 대학교 졸업이 32.6%, 결혼상태는 기혼이 74.4%로 높은 비율을 차지하였다. 체질별 일반적 특성은 Table 2와 같다.

2. KDC 교육용 데이트 활용한 통계 사용지침서 개발

본 연구에서는 교육용 데이터 개발과 함께 KDC 교육용 데이터를 이용하여 시행 가능한 통계 분석법

Table 1. Variables Structure in Korean medicine Data Center(KDC) Teaching Dataset

Domain (N of variables)	Variables
Basic (5)	sex, age, job, education, marruage
Lifestyle (2)	drinking, smoking
Personality questionnaire (15)	personality (ex; broad-minded vs. narrow-minded, quickly vs. slowly)
Symptom questionnaire (11)	diet(amount, speed), digestion(Y/N, appetite), sweat(amount, feeling), stool(habit, hardness, discomfort), urine(night)
Heat and cold questionnaire (8)	heat and cold(sensitivity, hand, foot, abdomen), drinking water(amount, temperature), heat and cold classification(score, group)
Disease (3)	hypertension, diabetes, dyslipidemia
Self-rated health (9)	general health status, sleep(time), fatigue(amount, time)
Expert` SC diagnosis (1)	Sasang constitution(SC) diagnosis by KMD
Body measurement (11)	height, weight, BMI, 8 body circumference
Blood pressure (2)	systolic BP, diastolic BP
Blood exam (5)	glucose, total cholesterol, triglyceride, HDL cholesterol, LDL cholesterol

Table 2. General Characteristics in Korean medicine Data Center(KDC) Teaching Dataset

Variables		Taeeumin (N=200)	Soeumin (N=125)	Soyangin (N=175)	Total (N=500)
Sex	men	69(34.5)	39(31.2)	56(32.0)	164(32.8)
	women	131(65.5)	86(68.8)	119(68.0)	336(67.2)
Age	mean±s.d.	50.31±16.49	44.29±16.37	48.21±15.49	48.07±16.26
	<20	10(5.0)	10(8.0)	8(4.6)	28(5.6)
	20-29	19(9.5)	19(15.2)	18(10.3)	56(11.2)
	30-39	23(11.5)	22(17.6)	28(16.0)	73(14.6)
	40-49	37(18.5)	21(16.8)	39(22.3)	97(19.4)
	50-59	46(23.0)	33(26.4)	39(22.3)	118(23.6)
	60-69	47(23.5)	12(9.6)	30(17.1)	89(17.8)
	≥70	18(9.0)	8(6.4)	13(7.4)	39(7.8)
Job	official	45(22.5)	43(34.4)	42(24)	130(26)
	service	22(11)	8(6.4)	20(11.4)	50(10)
	labor	36(18)	11(8.8)	26(14.9)	73(14.6)
	others	97(48.5)	63(50.4)	87(49.7)	247(49.4)
Education	none	18(9.0)	9(7.2)	11(6.3)	38(7.6)
	elementary school graduate	37(18.5)	10(8.0)	24(13.7)	71(14.2)
	middle school graduate	39(19.5)	15(12.0)	23(13.1)	77(15.4)
	high school graduate	42(21.0)	24(19.2)	44(25.1)	110(22.0)
	university graduate	52(26.0)	54(43.2)	57(32.6)	163(32.6)
	over university	12(6.0)	13(10.4)	16(9.1)	41(8.2)
Marriage	never married	42(21.0)	34(27.2)	31(17.7)	107(21.4)
	married	150(75.0)	87(69.6)	135(77.1)	372(74.4)
	divorced	2(1.0)	1(0.8)	3(1.7)	6(1.2)
	widowed	6(3.0)	3(2.4)	6(3.4)	15(3.0)

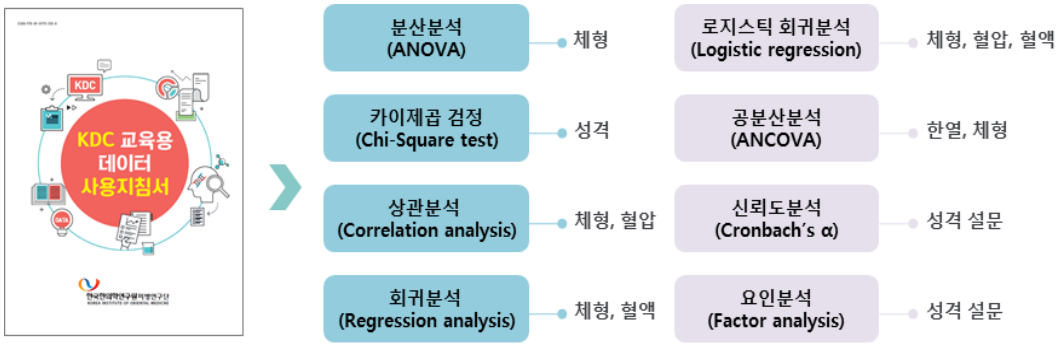

Figure 2. User guidelines of Korean medicine Data Center(KDC) teaching dataset

의 과정을 제시한 사용지침서를 개발하였다. 사용지침서는 분산분석, 카이제곱 검정, 상관분석, 회귀분석, 로지스틱 회귀분석, 공분산분석, 신뢰도분석, 요인분석 등 8개의 통계 분석법을 포함하였다. 각 통계 분석별 한의학적 특성을 고려한 연구가설을 설정하고, 통계 프로그램을 이용한 통계 분석 절차와 분석 결과 해석 방법을 상세히 기술하였다 (Figure 2).

3. KDC 교육용 데이터 활용 방법

KDC 교육용 데이터는 SPSS, EXCEL, CSV 3가지 형태로 제공되며, 사용지침서와 함께 KDC 포탈시스템 (https://kdc.kiom.re.kr) > KDC 활용에서 누구나 다운로드 받아 사용할 수 있다.

IV. 考察 및 結論

본 연구는 KDC 활용의 활성화 제고와 자료 접근성과 이해도를 높이기 위해 KDC 교육용 데이터와 사용지침서를 개발하였다.

KDC 교육용 데이터는 KDC 자료 중에서 체질 정보를 중심으로 다기관에서 수집된 체질확진자 자료원을 활용하였다. 체질확진자 자료는 선행 연구에서 다양한 연구 주제로 활용되어, 체질 진단이나 특성 파악, 체질과 질병 등 건강상태와의 연관성 등 수준 높은 연구 결과로 발표되었다. KDC 교육용 데이터는 연구용으로 활용도가 높은 자료원을 이용하였으며, 자료 특성인 체질정보를 반영한 표본 추출 방법을 실시하였다. 또한 사용지침서는 다양한 예시를 통해 시각적으로 통계 분석 절차를 설명하였으며, 사용자가 단계적으로 통계 분석이 가능하도록 활용도를 높였으며, 다양한 분야에서 통계 교육 프로그램으로 이용할 수 있다.

유전체 분석기술과 빅데이터 분석기술 발달에 따라 대규모 임상정보 분석이 가능하게 되어 대량의 인제자원을 전문적으로 수집, 관리하는 바이오뱅크의 중요성 부각, 국가별 투자가 증대되고 관련 산업도 성장하고 있다[1,8]. 국내의 경우 대표적으로 인체자원은행사업을 중심으로 2015년 기준 약 70만 명의 인체자원을 확보하는 등 국가적으로 종합 관리를 시행하고 있으며[14], 구축된 자료를 선도적으로 공개하고 인체자원을 활용한 연구결과의 선순환체계 활성화 등 다양한 시도를 하고 있다. 그러나 일반적으로 자료 특성에 대한 충분한 이해 없이 통계 분석을 진행하는 경우에는 결과 해석에서 심각한 오류가 발생될 위험이 있어 자료 이용을 돕기 위한 자료 활용 컨텐츠 개발의 필요성이 제안되었다[15].

한국한의학연구원의 KDC는 국내 한의계 최대의 임상정보은행으로 임상정보와 함께 기기정보, 인체자원 등이 확보되어 있어 다양한 연구목적으로 활용될 수 있다. 이를 위해 국내외 한의학 관련 융합연구 활성

화를 위해 자료를 활용하여 연구 수행이 가능하도록 체계적인 자료 관리와 함께 정보 공개를 위해 임상정보 코드북 제공, KDC 포탈 홈페이지 운영, 자료분양 시스템을 운영하는 등 다양한 노력을 기울이고 있다. KDC 교육용 데이터는 사용자가 자료에 접근하여 직접 활용해 볼 수 있도록 능동적인 접근 방법으로 개발되어 사용자 활용도 측면을 고려하였다.

다만, KDC 교육용 데이터의 분석 결과는 전체 자료 중에서 일부를 추출하여 재구성한 것으로 KDC의 대표값으로 해석하는 것은 적합하지 않으며, 교육용으로 개발된 것으로 목적 외 논문 작성 등에 활용되어서는 안된다.

현재 한의계에서 한의 빅데이터 구축을 위한 다양한 노력과 접근을 시도하고 있다. 한의계 전반에 흩어져 있는 임상자료의 체계적인 자료통합과 활용이 가능하다며, 한의학의 우수한 성과 창출에 견인 역할을 할 것으로 기대한다. 이를 위해 연구 자료의 양과 질의 확대가 필요하며 관련 규정의 정비와 연구자들의 관심과 노력이 수반되어야 한다.

V. Achnowledgement

본 연구는 한국한의학연구원 기관주요사업인 '한의 유전체 역학 인프라 구축'(K17091) 과제의 지원을 받아 수행되었음.

VI. References

1. Kang B, Park J, Cho S, Lee M, Kim N, Min H, Han B. Current status, challenges, policies, and bioethics of biobanks. Genomics & informatics. 2013; 11(4):211-217.
2. Rogers J, Carolin T, Vaught J, & Compton C. Biobankonomics: a taxonomy for evaluating the economic benefits of standardized centralized human biobanking for translational research. Journal of the National Cancer Institute Monographs. 2011;42:32-38.
3. Biobanking for medicine: technology, industry and market 2014-2024. 2014.
4. Simeon-Dubach D, Watson P. Biobanking 3.0: evidence based and customer focused biobanking. Clinical biochemistry. 2014;47(4):300-308.
5. http://ukdataservice.ac.uk/
6. https://biolincc.nhlbi.nih.gov/static/studies/teaching/framdoc.pdf.
7. http://kbn.cdc.go.kr.
8. Korea Health Industry development institute. 2014 Health Industry White Paper. 2015. p.448-450. (Korean)
9. Korea Institute of Oriental Medicine. 2014 Year book of traditional Korean medicine. 2015. p116-117. (Korean)
10. Korea Centers for Disease Control and Prevention. Development of system for performance analysis utilizing the National Biobank of Korea(NBK)'s bioresource. 2015. (Korean)
11. Jin HJ, Baek Y, Kim HS, Ryu J, Lee S. Constitutional multicenter bank linked to Sasang constitutional phenotypic data. BMC complementary and alternative medicine. 2015;15:46.
12. Baek YH, Jin HJ, Kim HS, Jang ES, Lee SW. An Overview on the Construction of Korea Constitutional Multicenter Bank for Sasang Constitutional Medicine. J of Sasang Constitutional Medicine. 2012;24(2):47-53. (Korean)
13. Lee HY, Lee PY. Introduction to sample survey. Kyowoo. 2002. p133. (Korean)
14. Korea Centers for Disease Control and Prevention. 2016-2020 The 3rd Korea Biobank Project Plan. 2016. (Korean)
15. Ko SB et al. Investigating unmet research needs based on data of the Korean Genome and Epidemiology Study. 2015. (Korean)