ISSN 2093-5986(Print)
ISSN 2288-0666(Online)

The Korean Journal of Health Service Management
Vol.10 No.3 pp.63-74 September 2016
http://dx.doi.org/10.12811/kshsm.2016.10.3.063

대화식 의사결정나무를 이용한 보건의료 데이터 질 관리 알고리즘 개발: 당뇨환자의 고혈압 동반을 중심으로

황규연[1], 이은숙[1], 김고원[1], 홍성옥[2], 박정선[3], 곽미숙[3], 이예진[3], 임채혁[4] , 박태현[4] , 박종호[4] , 강성홍[4] ‡
[1]부산대학교병원, [2]질병관리본부, [3]한국보건산업진흥원, [4] 인제대학교 보건행정학과

Development of Healthcare Data Quality Control Algorithm Using Interactive Decision Tree: Focusing on Hypertension in Diabetes Mellitus Patients

Kyu-Yeon Hwang[1], Eun-Sook Lee[1], Go-Won Kim[1], Seong-Ok Hong[2], Jung-Sun Park[3], Mi-Sook Kwak[3], Ye-Jin Lee[3], Chae-Hyeok Lim[4] , Tae-Hyun Park[4] , Jong-Ho Park[4] , Sung-Hong Kang[4] ‡
[1]Pusan National University Hospital, [2]Department of Korea Centers for Disease Control & Prevention, [3]Korea Health Industry Development Institute, [4] Department of Health Administration, Inje University

<Abstract>

Objectives : There is a need to develop a data quality management algorithm to improve the quality of healthcare data using a data quality management system. In this study, we developed a data quality control algorithms associated with diseases related to hypertension in patients with diabetes mellitus. **Methods :** To make a data quality algorithm, we extracted the 2011 and 2012 discharge damage survey data from diabetes mellitus patients. Derived variables were created using the primary diagnosis, diagnostic unit, primary surgery and treatment, minor surgery and treatment items. **Results :** Significant factors in diabetes mellitus patients with hypertension were sex, age, ischemic heart disease, and diagnostic ultrasound of the heart. Depending on the decision tree results, we found four groups with extreme values for diabetes accompanying hypertension patients. **Conclusions** : There is a need to check the actual data contained in the Outlier (extreme value) groups to improve the quality of the data.

Key Words : Data Mining, Data Quality Management Algorithm, Outlier Detection Method, Diabetes Mellitus, Hypertension

* 본 연구는 보건복지부의 재원으로 한국보건산업진흥원의 첨단의료기술개발사업 지원에 의하여 이루어진 것임(과제고유번호 : HI14C2756).
‡ Corresponding author : Sung-Hong Kang(hcmkang@inje.ac.kr) Department of Health Administration, Inje University
• Received : Jun 22, 2016 • Revised : Jul 22, 2016 • Accepted : Sep 20, 2016

Ⅰ. 서론

건강검진자료, 질병자료, 전자의무기록자료, 수술용 로봇 데이터, 유전체 분석 데이터, 바이오 센싱, 의료 영상 데이터의 생성이 급증함에 따라 보건의료 빅데이터 대한 관심이 크게 증가하고 있다. 보건의료 빅데이터를 효율적으로 활용할 경우 질병예방에 따른 의료비 절감, 의료기관의 운영비용 절감, 오류 및 부정에 따른 손실비용 절감 등의 경제적 효과와 보건의료 정책결정 강화, 보건의료 R&D 투자확대, 질병예방·맞춤형 의료서비스 제공을 통한 국민행복 증진 등의 사회적 효과가 창출될 것으로 전망된다. 이와 같은 기대효과 때문에 보건의료 빅데이터 산업도 크게 성장할 것으로 전망된다. 전문가 조사에 따르면 연평균 25% 이상 성장할 것으로 전망된다[1]. 보건의료 빅데이터가 효율적으로 활용되기 위해서는 이를 지원하는 기술 분야도 함께 발전을 해야 한다. 보건의료 빅데이터에 관련된 기술 분야는 크게 대용량 데이터의 저장, 대용량 데이터의 빠른 처리, 분석 및 추론 분야가 강조되고 있다[2]. 그러나 데이터 관리의 가장 중요한 요소인 신뢰할 수 있는 데이터 질 관리 분야에 대한 관심은 낮은 편이다. “쓰레기가 들어가면 쓰레기가 나온다.”라는 말처럼 보건의료 빅데이터 시스템이 그 역할을 제대로 수행하기 위해서는 무엇보다도 보건의료 데이터의 질적 수준의 향상이 필요하다. 보건의료 데이터의 낮은 질적 수준은 환자안전(Patient Safety), 공공안전(Public Safety), 진료의 연속성(Continuity of Patient Care), 임상연구 및 결과관리(Clinical Research and Outcomes) 측면에서 많은 문제를 야기할 뿐만 아니라 막대한 경제적 손실을 유발한다[3]. 미국 전체적으로 데이터의 낮은 질적 수준으로 인한 경제적 손실 규모는 연간 660조 정도로 추정된다[4]. 보건의료분야의 낮은 데이터의 질적 수준으로 인한 경제적 손실 규모는 의료기관 매출액의 15% 정도 되는 것으로 추정된다[5].

우리나라 보건의료 데이터의 질적 수준은 낮은 것으로 나타났다. 2010년 한국 데이터베이스 진흥원에서 발표한 연구결과에 따르면 공공, 금융, 통신, 미디어, 유통/서비스. 제조, 의료 등 6개 산업군 317개 공공기관 및 민간 기업을 대상으로 조사한 결과 의료부분의 데이터 질 관리 수준은 전체 평균에 밑돌고 있으며, 이는 데이터 질 관리에 있어 가장 기초적인 도입단계에도 이르지 못하는 수준을 의미한다[6]. 또한 국내에서 의료 정보 수준이 가장 잘 되어있다고 평가받는 병원의 전자의무기록의 데이터에 유용성에 대한 연구결과 전체 데이터의 88.6%는 형태나 의미적 측면에서 적합성을 갖추고 있으나 나머지 11.4%는 적합성을 갖추지 못한 것으로 알려져 다른 대학병원, 중소병원 및 의원의 전자의무기록의 데이터의 질적 수준은 이보다 낮은 것으로 추정된다[7]. 따라서 우리나라에서 보건의료 빅데이터 산업이 활성화 위해서는 보건의료 데이터 질 관리 사업이 효율적으로 이루어져야 한다.

보건의료 데이터 질 관리 사업은 보건의료 정보의 양이 방대해짐 따라 사람이 눈으로 일일이 확인을 하는 방식은 많은 인력을 필요로 하고 그로 인한 인건비의 증가 및 시간적 손실이 발생하므로 IT 기술을 효율적으로 활용할 필요성이 증대되고 있다. 즉, 보건의료 데이터 질 관리 검증시스템을 개발하고, 이러한 전산시스템을 이용하여 오류를 검증할 필요성이 있다. 보건의료 데이터 질 관리 검증시스템을 개발하기 위해서는 먼저 질 관리의 오류를 검증할 수 있는 로직을 개발하는 것이 필요하다. 이에 대한 알고리즘은 크게 4가지로 분류할 수 있는데, 첫째는 각 변수에는 제한된 값만을 가져야 한다는 알고리즘이다. 예를 들어, 성별(sex)이라는 변수에는 남자나 여자라는 값만 올 수 있으며, 다른 값은 올 수 없음을 나타내는 알고리즘을 말한다. 두 번째 알고리즘은 변수들 간의 연관

성에 근거하여 명확한 오류를 검증하는 알고리즘이다. 이는 전립선 비대증이라는 상병에는 성별(sex)은 반드시 남자여야 하는 경우를 말한다. 세 번째 알고리즘은 데이터 입력이 누락되는 것을 변수들 간의 연관성을 기반으로 찾아 주는 알고리즘이다. 예를 들어 당뇨환자로서 고혈압을 동반한 환자인데 고혈압 상병이 입력되어있지 않는 것을 찾아주는 알고리즘이다. 네 번째 알고리즘은 변수간의 연관성에 근거하여 입력이 잘못되어진 것을 찾아주는 알고리즘이다. 첫 번째 알고리즘은 데이터를 수집하는 데이터베이스를 구축하는 시기에 이를 반영하여 근본적으로 데이터 오류를 막아줄 수 있지만 그 외 나머지 질 관리의 오류를 검증할 수 있는 알고리즘은 환자마다 케이스가 너무나 다양하기 때문에 파악하기가 매우 어렵다. 그러나 이러한 알고리즘이 개발되어야 보다 정확한 데이터 질 관리가 이루어질 수 있다. 최근에 이러한 필요성에 부응하여 외국에서는 Outlier Detection Method 방법에 근거하여 데이터 질 관리 알고리즘을 개발하여 활용하고 있음에 따라 우리나라에서도 이러한 연구를 시도할 필요가 있다. Outlier Detection Method는 보건의료 데이터 간의 관계에서 발생할 경우를 확률화하여 확률이 매우 낮은 경우를 Outlier라 정의하고 이러한 케이스에 대해서 데이터 질 관리 전문가가 내용을 확인하여 오류를 보완하도록 하는 방법으로 외국에서는 암 환자 자료 질 관리 등에서 이를 활용하고 있다[8][9]. 이에 본 연구에서는 당뇨환자 중 고혈압을 동반하는지에 대해서 Outlier Detection Method 방법에 근거하여 데이터 질 관리 알고리즘을 개발하고자 한다. 당뇨환자 중 고혈압 동반을 연구 대상으로 잡은 이유는 우리나라의 고혈압과 당뇨의 유병률이 고혈압 25.5%, 당뇨 10.2%이며, 연간 고혈압 입원환자는 43,387명이고, 당뇨 입원환자는 66,606명으로 전체 입원 상병 중 각각 41위와 22위를 차지할 정도로 많은 사람들이 고혈압과 당뇨로 인해 진료를 받고 있기 때문이다[10]. 2011년~2012년 퇴원손상심층조사 자료에 의하면 전체 당뇨환자 중 고혈압을 동반한 환자는 53.8%로 나타나 당뇨환자가 동반상병으로 고혈압을 절반 이상이 가지고 있다는 것을 알 수 있다. 따라서 당뇨환자의 고혈압 상병의 동반에 관련된 데이터 질 관리 알고리즘을 개발하면 활용성이 높을 것으로 판단됨에 따라 이에 대해서 우선적으로 개발할 필요가 있다.

Ⅱ. 연구방법

1. 연구자료

본 연구는 질병관리본부의 퇴원손상심층조사 자료를 이용하였다. 퇴원손상심층조사 자료는 의료기관의 퇴원환자 의무기록 자료를 활용하여 주요 만성질환 및 손상에 대한 지속적이고 체계적인 보건의료 통계를 생산하여 국민 건강증진 및 보건의료 정책에 필요한 기초 자료로 활용하기 위해 실시되는 조사이며, 자료 수집 및 데이터베이스 구축 과정에서 철저한 검증 절차를 거침에 따라 데이터 정제가 잘되어 있는 대표적인 입원환자에 관한 데이터베이스이다. 조사의 주요 항목은 환자 번호, 성별, 연령, 주 진단, 부 진단, 주 수술 및 처치, 부 수술 및 처치 등 총 71개의 항목으로 구성이 되어 있다[11][12]. 본 연구를 위하여 2011년, 2012년 퇴원손상심층조사 자료를 수집하였다. 2년 치 자료를 사용한 이유는 2011년부터 새로운 질병분류체계가 도입됨에 따라서 동일한 질병분류코드를 사용하기 위해서 2011년 자료부터 사용하기로 하였다. 또한 현재 질병관리본부에서 공개하는 자료가 2012년까지 자료임에 따라 2011년, 2012년 자료를 사용하였다. 2011년, 2012년 퇴원손상심층조사 자료 중 분석대상자인 당뇨 상병을 가지는 환자만을 추출하였다. 당뇨 상병을 가지는 케이스란 KCD-6기준으로 퇴원손상심층조사에서 주 진단과 부 진단 중

E10~E14.9에 해당하는 상병을 가진 환자를 말한다.

2. 변수정의

2011년, 2012년 퇴원손상심층조사 자료에서 수집 가능한 성, 연령, 주진단, 부진단, 주수술 및 처치, 부수술 및 처치의 변수를 이용하여 분석에 필요한 파생변수를 생성하였다. 질병 유무 및 수술처치 유무 파생변수는 KCD-6와 ICD-9CM을 이용한 Clinical Classifications Software(C.C.S) 기준과 문헌고찰, 내과 전문의, 외과 전문의 등 전문가의 자문을 통하여 선정하였다. C.C.S란 미국 The Agency for Healthcare Research and Quality)에 의해 개발된 질병 군별 분류 지표이다. 분석에 사용된 주요 변수는 성별, 연령, 고혈압, 허혈성 심장질환, 심장의 진단적 초음파 등이다.

<Table 1> Definition of variables

variables	Definition
SEX	Man, Woman
AGE	Univariate analysis : less than 60 / between 60 and 69 / 70 and over Multivariate analysis : continuous variable
Malignant neoplasm of stomach	Anyone who has C16(ICD-6) code in data Yes / No
Malignant neoplasm of breast	Anyone who has C50(ICD-6)code in data Yes / No
Malignant neoplasm of corpus uteri	Anyone who has C54(ICD-6)code in data Yes / No
Malignant neoplasm of ovary	Anyone who has C56(ICD-6)code in data Yes / No
Myeloid leukemia	Anyone who has C92(ICD-6)code in data Yes / No
Dementia in Alzheimer's disease	Anyone who has F00(ICD-6)code in data Yes / No
Nerve root and plexus compressions in diseases classified elsewhere	Anyone who has G55(ICD-6)code in data Yes / No
Other disorders of nervous system in diseases classified elsewhere	Anyone who has G99(ICD-6)code in data Yes / No
Pneumonia, unspecified	Anyone who has J189(ICD-6)code in data Yes / No
Other spondylopathies	Anyone who has M48(ICD-6)code in data Yes / No
Other intervertebral disc disorders	Anyone who has M51(ICD-6)code in data Yes / No
Hypertension	Anyone who has I10~I15.9(ICD-6)code in data Yes / No
Ischemic heart disease	Anyone who has I20~I25.9(ICD-6)code in data Yes / No
Complete thyroidectomy	Anyone who has 06.4(ICD-9-CM)code in data Yes / No
Ophthalmoscopy	Anyone who has 16.21(ICD-9-CM)code in data Yes / No
Other bronchoscopy	Anyone who has 33.23(ICD-9-CM)code in data Yes / No
Biopsy of bone marrow	Anyone who has 41.31(ICD-9-CM)code in data Yes / No
Diagnostic ultrasound of heart	Anyone who has 88.72(ICD-9-CM)code in data Yes / No
Other diagnostic ultrasound	Anyone who has 88.79(ICD-9-CM)code in data Yes / No

3. 분석방법

분석 대상자의 일반적 특성에 대해서는 빈도분석을 실시하였고, 제 특성에 따른 당뇨유무에 대해서는 교차분석을 실시하였다. 고혈압 환자의 당뇨유무에 관련된 모형을 대화식 의사결정나무 기법을 이용하여 분석을 하였다. 대화식 의사결정나무의 모형평가는 지지도(Lift)값을 이용하였다. 의사결정나무(decision tree) 모형은 특정한 분류 기준에 따라 목표변수와 가장 관련성이 높은 독립변수를 선정한 후 의사결정규칙(decision rule)을 몇 개의 소집단으로 분류하여 나무구조로 표현하는 것으로 요인의 규명, 분류, 예측에 유용하다. 또한 의사결정나무모형은 독립변수간의 관계를 도식화하여 보여주기 때문에 연구자가 분석과정을 쉽게 이해하고 설명할 수 있는 장점을 가지고 있다. 특히 의사결정나무의 대화식(interactive) 방식은 연구자가 중요하다고 생각하는 변수를 모형에 반영할 수 있다는 장점이 있기 때문에 이를 이용하여서 분석을 하였다[13][14].

분석 툴은 통계처리를 위해서는 SAS Enterprise Guide 9.4 ver을 사용하였고 데이터마이닝 분석을 위해서는 SAS Enterprise Miner 9.4 ver을 사용하였다.

Ⅲ. 연구결과

1. 분석대상자의 일반적 특성

빈도분석을 통해 본 연구의 분석 대상 당뇨환자 42,292건에 대한 일반적 특성을 살펴본 결과 성별로는 남자 54.7%, 여자 45.3%로 여자보다 남자가 높은 것으로 나타났다. 연령별로는 70세 이상이 36.5%로 가장 높았으며, 60세 미만 34.0%, 60~69세 29.4% 순으로 높게 나타났다. 당뇨환자의 상병별 동반율을 살펴보면 고혈압이 53.8%로 가장 높았으며, 허혈성 심장질환 12.8%, 상세불명의 폐렴 4.0%, 위의 악성 신생물 2.9%, 기타 척추병증 1.8%, 기타 추간판 장애와 유방의 악성 신생물이 각각 1.0% 순으로 높은 것으로 조사되었다. 당뇨환자의 수술 및 처치별 시행률을 살펴보면 심장의 진단적 초음파가 11.1%로 가장 높게 나타났고, 기타 진단적 초음파 1.1%, 기타 기관지경 검사 0.6%, 완전 갑상샘 절제술 0.3%, 안구검사 0.1% 순으로 높았다<Table 2>.

2. 제 특성에 따른 고혈압 유무

당뇨환자의 일반적 특성에 따른 고혈압 유무를 살펴보기 위해 교차분석을 실시한 결과 성별에 따른 고혈압 동반율 남자 50.1%, 여자 58.2%로 남자보다 여자가 높았으며, 연령별 고혈압 동반율은 70세 이상 64.3%, 60~69세는 57.3%, 60세 이하 39.3%의 순으로 높게 나타났다. 당뇨환자의 동반상병별 고혈압 동반율은 허혈성심질환이 동반된 당뇨환자의 고혈압 동반율은 70.2%로 가장 높았으며, 알츠하이머병에서의 치매가 동반된 당뇨환자의 고혈압 동반율 63.3%, 기타 추간판 장애가 동반된 당뇨환자의 고혈압 동반율 49.3%, 유방의 악성 신생물이 동반된 당뇨환자의 고혈압 동반율 48.9% 등으로 조사되었다. 당뇨환자의 수술 및 처치 유무별 고혈압 동반율은 심장의 진단적 초음파를 시행한 당뇨환자의 고혈압 동반율은 69.2%, 완전 갑상샘 절제술을 시행한 당뇨환자의 고혈압 동반율은 60.3%, 안구 검사를 시행한 당뇨환자의 고혈압 동반율은 59.6%, 기타 진단적 초음파를 시행한 당뇨환자의 고혈압 동반율은 53.8%, 기타 기관지경 검사를 시행한 당뇨환자의 고혈압 동반율은 52.9%, 골수의 생검을 시행한 당뇨환자의 고혈압 동반율은 51.3%로 나타났다<Table 3>.

<Table 2> Characteristics of Subjects

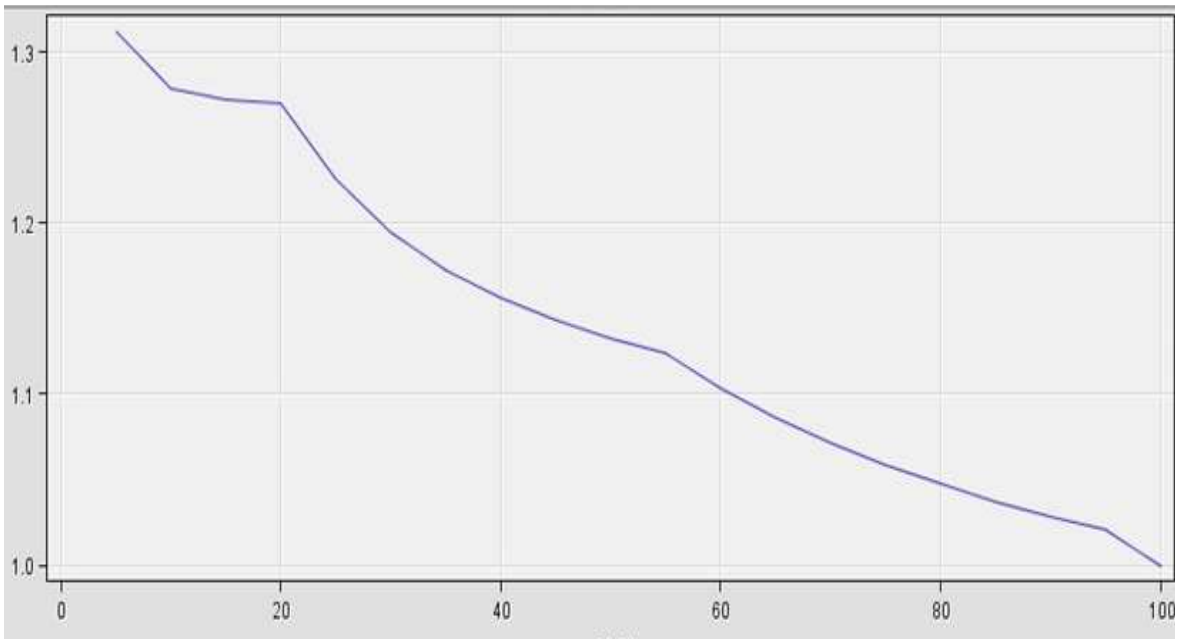
	variables	Value	N	%
Demography	SEX	Man	23,138	54.7
		Woman	19,154	45.3
	AGE (years)	less than 60	14,395	34.0
		between 60 and 69	12,440	29.4
		70 and over	15,457	36.5
Disease Prevalence	Malignant neoplasm of the stomach	YES	1,230	2.9
		NO	41,062	97.1
	Malignant neoplasm of the breast	YES	413	1.0
		NO	41,879	99.0
	Malignant neoplasm of the corpus uteri	YES	63	0.1
		NO	42,229	99.9
	Malignant neoplasm of the ovary	YES	153	0.4
		NO	42,139	99.6
	Myeloid leukemia	YES	107	0.3
		NO	42,185	99.7
	Dementia in Alzheimer's disease	YES	147	0.3
		NO	42,145	99.7
	Nerve root and plexus compressions in diseases classified elsewhere	YES	57	0.1
		NO	42,235	99.9
	Other disorders of the nervous system in diseases classified elsewhere	YES	143	0.3
		NO	42,149	99.7
	Pneumonia, unspecified	YES	1,710	4.0
		NO	40,582	96.0
	Other spondylopathies	YES	759	1.8
		NO	41,533	98.2
	Other intervertebral disc disorders	YES	438	1.0
		NO	41,854	99.0
	Hypertension	YES	22,740	53.8
		NO	19,552	46.2
	Ischemic heart disease	YES	5,415	12.8
		NO	36,877	87.2
Surgery & Treatment	Complete thyroidectomy	YES	116	0.3
		NO	42,176	99.7
	Ophthalmoscopy	YES	52	0.1
		NO	42,240	99.9
	Other bronchoscopy	YES	261	0.6
		NO	42,031	99.4
	Biopsy of bone marrow	YES	158	0.4
		NO	42,134	99.6
	Diagnostic ultrasound of the heart	YES	4,704	11.1
		NO	37,588	88.9
	Other diagnostic ultrasounds	YES	465	1.1
		NO	41,827	98.9
	Total		42,292	100.0

<Table 3> Hypertension Prevalence according to the Characteristics of Subjects

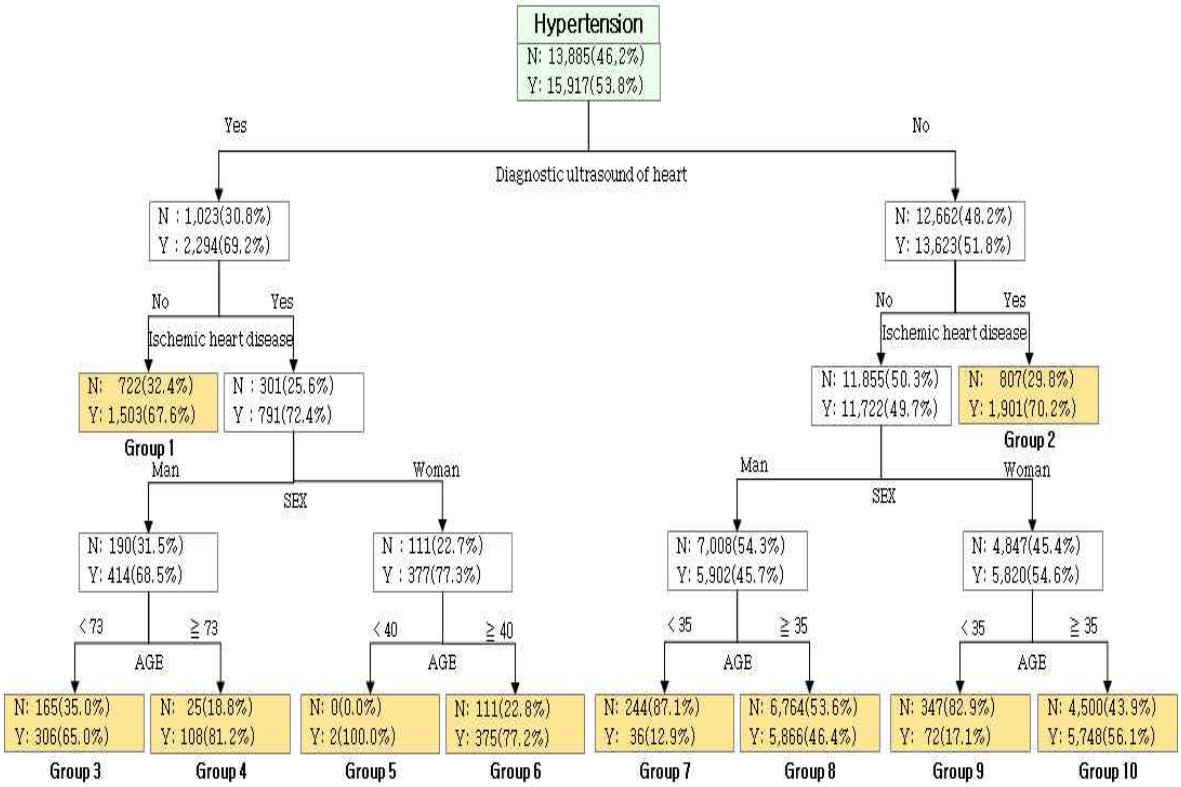
	Variables	Value	Hypertension YES	Hypertension NO	Total	p value
			N(%)	N(%)	N(%)	
Demography	SEX	Man	11,587(50.1)	11,551(49.9)	23,138(100.0)	0.000
		Woman	11,153(58.2)	8,001(41.8)	19,154(100.0)	
	AGE	less than low 60	5,663(39.3)	8,732(60.7)	14395(100.0)	0.000
		between 60 and 69	7,131(57.3)	5,309(42.7)	12440(100.0)	
		70 and over	9,946(64.3)	5,511(35.7)	15457(100.0)	
Disease Prevalence	Malignant neoplasm of the stomach	YES	655(53.3)	575(46.7)	1,230(100.0)	0.367
		NO	22,085(53.8)	18,977(46.2)	41,062(100.0)	
	Malignant neoplasm of the breast	YES	202(48.9)	211(51.1)	413(100.0)	0.026
		NO	22,538(53.8)	19,341(46.2)	41,879(100.0)	
	Malignant neoplasm of the corpus uteri	YES	33(52.4)	30(47.6)	63(100.0)	0.462
		NO	22,707(53.8)	19,522(46.2)	42,229(100.0)	
	Malignant neoplasm of the ovary	YES	81(52.9)	72(47.1)	153(100.0)	0.450
		NO	22,659(53.8)	19,480(46.2)	42,139(100.0)	
	Myeloid leukemia	YES	49(45.8)	58(54.2)	107(100.0)	0.059
		NO	22,691(53.8)	19,494(46.2)	42,185(100.0)	
	Dementia in Alzheimer's disease	YES	93(63.3)	54(36.7)	147(100.0)	0.013
		NO	22,647(53.7)	19,498(46.3)	42,145(100.0)	
	Nerve root and plexus compressions in diseases classified elsewhere	YES	31(54.4)	26(45.6)	57(100.0)	0.516
		NO	22,709(53.8)	19,526(46.2)	42,235(100.0)	
	Other disorders of nervous system in diseases classified elsewhere	YES	77(53.8)	66(46.2)	143(100.0)	0.526
		NO	22,663(53.8)	19,486(46.2)	42,149(100.0)	
	Pneumonia, unspecified	YES	929(54.3)	781(45.7)	1,710(100.0)	0.327
		NO	21,811(53.7)	18,771(46.3)	40,582(100.0)	
	Other spondylopathies	YES	428(56.4)	331(43.6)	759(100.0)	0.077
		NO	22,312(53.7)	19,221(46.3)	41,533(100.0)	
	Other intervertebral disc disorders	YES	216(49.3)	222(50.7)	438(100.0)	0.034
		NO	22,524(53.8)	19,330(46.2)	41,854(100.0)	
	Ischemic heart disease	YES	3,799(70.2)	1,616(29.8)	5,415(100.0)	0.000
		NO	18,941(51.4)	17,936(48.6)	36,877(100.0)	
Surgery & Treatment	Complete thyroidectomy	YES	70(60.3)	46(39.7)	116(100.0)	0.092
		NO	22,670(53.8)	19,506(46.2)	42,176(100.0)	
	Ophthalmoscopy	YES	31(59.6)	21(40.4)	52(100.0)	0.240
		NO	22,709(53.8)	19,531(46.2)	42,240(100.0)	
	Other bronchoscopy	YES	138(52.9)	123(47.1)	261(100.0)	0.410
		NO	22,602(53.8)	19,429(46.2)	42,031(100.0)	
	Biopsy of bone marrow	YES	81(51.3)	77(48.7)	158(100.0)	0.290
		NO	22,659(53.8)	19,475(46.2)	42,134(100.0)	
	Diagnostic ultrasound of the heart	YES	3,257(69.2)	1,447(30.8)	4,704(100.0)	0.000
		NO	19,483(51.8)	18,105(48.2)	37,588(100.0)	
	Other diagnostic ultrasounds	YES	250(53.8)	215(46.2)	465(100.0)	0.518
		NO	22,490(53.8)	19,337(46.2)	41,827(100.0)	

3. 대화식 의사결정나무 기법 모형의 개발 및 평가

당뇨환자의 고혈압 동반 모형을 개발하기 위해 대화식 의사결정나무 분석을 실시하였다. 본 연구의 목적이 Outlier Detection Method 방법에 근거하여 극단적인 경우를 찾아내고, 이를 확인, 수정할 수 있도록 데이터 질 관리 알고리즘을 개발하는 것임에 따라 당뇨환자가 고혈압을 동반할 확률이 아주 낮거나 높은 특정 집단을 찾아내는 것에 주안점을 두고 모형을 개발하였다. 교차분석에서 당뇨환자의 고혈압 유무에 통계적으로 유의한 변수였던 성, 연령, 심장의 진단적 초음파 검사, 허혈성 심장질환 유무 등을 중심으로 모형을 개발하였다. 개발된 모형의 성능을 지지도(Lift)를 이용하여 평가한 결과 <Figure 1>과 같이 우수한 모형임을 알 수 있었고 평균제곱오차 또한 0.25로 나타났다. 개발된 당뇨환자의 고혈압 동반 의사결정나무 모형에서 Outlier로 볼 수 있는 집단은 4개 집단으로 나타났다. 4개 집단을 선정한 기준은 고혈압을 동반상병으로 가질 확률 값이 80% 이상이거나, 20% 이하인 것을 기준으로 하였다. 여기에 해당되는 집단군은 그룹 4, 그룹 5, 그룹 7, 그룹 9로 나타났다, 그룹 4의 특성은 심장의 진단적 초음파를 시행하고 허혈성심질환이 있으면서 성별이 남자이고 연령이 73세 이상인 집단군으로 고혈압을 동반상병으로 가질 확률 값이 81.2%로 나타났고, 그룹 5의 특성은 심장의 진단적 초음파를 시행하였고 허혈성심질환이 있으면서 성별이 여자이고 연령이 40세 미만인 집단군으로 확률 값이 100.0%로 나타났다. 그룹 7의 특성은 심장의 진단적 초음파를 시행하지 않았고, 허혈성심질환이 없으면서 성별이 남자이고 연령이 35세 미만인 집단군으로 고혈압을 동반상병으로 가질 확률 값이 12.9%로 나타났으며, 그룹 9의 특성은 심장의 진단적 초음파를 시행하지 않고 허혈성심질환이 없으면서 성별이 여자이고 연령이 35세 미만인 집단군으로 확률 값이 17.2%로 나타났다<Figure 2><Table 4>.



<Figure 1> Model evaluation of lift value



<Figure2> Interactive Decision Tree Model

<Table 4> Rules of Finding Healthcare Data Outlier based on decision tree

Node	Character of node(Anyone who have HTN in DM patient)	HTN N(%)
1	Diagnostic ultrasound of the heart Yes & Ischemic heart disease No	1,503(67.6)
2	Diagnostic ultrasound of the heart No & Ischemic heart disease Yes	1,901(70.2)
3	Diagnostic ultrasound of the heart Yes & Ischemic heart disease Yes & SEX Man & AGE<73	306(65.0)
4	Diagnostic ultrasound of the heart Yes & Ischemic heart disease Yes & SEX Man & AGE≧73	108(81.2)
5	Diagnostic ultrasound of the heart Yes & Ischemic heart disease Yes & SEX Woman & AGE<40	2(100.0)
6	Diagnostic ultrasound of the heart Yes & Ischemic heart disease Yes & SEX Woman & AGE≧40	375(77.2)
7	Diagnostic ultrasound of the heart No & Ischemic heart disease No & SEX Man & AGE<35	36(12.9)
8	Diagnostic ultrasound of the heart No & Ischemic heart disease No & SEX Man & AGE≧35	5,866(46.4)
9	Diagnostic ultrasound of the heart No & Ischemic heart disease No & SEX Woman & AGE<35	72(17.1)
10	Diagnostic ultrasound of the heart No & Ischemic heart disease No & SEX Woman & AGE≧35	5,748(56.1)

Ⅳ. 고찰

우리나라 보건의료 데이터의 질적 수준은 낮다. 보건의료 데이터의 낮은 질적 수준은 의료기관, 환자, 국가 모두에게 많은 문제를 야기할 뿐만 아니라 막대한 경제적 손실을 유발한다[3]. 실제 보건의료분야의 낮은 데이터의 질적 수준으로 인해 의료기관 매출액의 15% 정도가 경제적으로 손실되는 것으로 추정되고 있다[5]. 보건의료 데이터의 질적 수준 향상을 위해서는 보건의료 데이터의 질 관리가 필요하다. 보건의료 데이터의 질 관리를 위해서는 IT 기술을 이용하여 질 관리 오류 검증 시스템을 개발하고, 이를 실제 현장에 적용하는 것이 필요하며, 질 관리 오류 검증 시스템 개발 전에 데이터 질 관리 알고리즘에 대한 개발이 선행되어야 한다. 외국에서는 암 환자 자료의 질 관리 등에 Outlier Detection Method 방법을 활용하고 있다[8][9]. Outlier Detection Method는 데이터 간의 발생 관계를 확률화하여 확률이 매우 낮거나 매우 높은 경우를 Outlier(극단적 케이스)라 정의하고, 이러한 극단적 케이스에 대해서 데이터 질 관리 전문가를 통해 내용을 확인하고 수정, 보완하게 하는 방법이다. Outlier Detection Method 방법은 극단적인 케이스의 원인을 찾아 주지 못하여 이에 대한 요인은 추가적인 분석을 통해서 확인하여야 하는 단점이 있지만 데이터 질 관리를 위해 모든 케이스를 확인하고, 수정, 보완해야 하는 부담, 시간, 비용을 절감시켜 준다는 장점이 있다[8][9]. 또한 고혈압은 인구의 고령화와 함께 이로 인한 심혈관 질환의 이환율과 사망률이 지속적으로 증가하고 있고, 당뇨병 역시 흔한 만성질환임과 동시에 신부전증 등의 발병의 가장 중요한 원인이기도 하다.[18] 이에 본 연구에서는 질 관리 오류 검증 방법인 Outlier Detection Method 방법에 따라 당뇨환자의 고혈압 동반에 대한 Outlier(극단적 케이스)를 찾을 수 있는 데이터 질 관리 알고리즘을 개발하고자 하였다. 본 연구에서 당뇨환자의 고혈압 동반에 대한 Outlier(극단적 케이스)를 찾는 알고리즘을 개발하기 위해 질병관리본부의 퇴원손상심층조사 자료를 수집하여 대화식 의사결정나무 분석을 실시하였다. 의사결정나무(decision tree) 모형은 특정한 분류 기준에 따라 목표변수와 가장 관련성이 높은 독립변수를 선정한 후 의사결정규칙(decision rule)을 나무 구조로 도표화하여 관심의 대상이 되는 집단에 대한 요인의 규명, 분류, 예측이 유용하며, 독립변수 간의 관계를 도식화하여 보여주기 때문에 연구자가 분석 과정을 쉽게 이해하고 설명할 수 있는 장점을 가지고 있다[11][12][15]. 특히 의사결정나무의 대화식(Interactive) 방식은 데이터 자체에만 의존하는 기존 의사결정나무 모형에 비해 전문적인 지식을 모형에 반영할 수 있다는 장점이 있기 때문에 최근 데이터마이닝을 이용한 요인 분석 연구에서 많이 활용되고 있다[15][16][17]. 따라서 본 연구에서는 대화식 의사결정나무 모형을 이용하여 고혈압 동반에 대한 Outlier(극단적 케이스)를 찾는 알고리즘을 개발하고자 한 분석방법에는 문제가 없을 것으로 판단된다.

대화식 의사결정나무 기법을 이용하여 당뇨환자의 고혈압 동반 모형을 개발한 결과 당뇨환자의 고혈압 동반에 대한 Outlier(극단적 케이스)로 볼 수 있는 집단은 4개 집단으로 나타났다. 4개 집단을 선정한 기준은 고혈압을 동반상병으로 가질 확률 값이 80%이상이거나, 20%이하인 것을 기준으로 하였다. 여기에 해당되는 집단군은 그룹 4, 그룹 5, 그룹 7, 그룹 9로 나타났다. 각 그룹의 특성은 다음과 같다. 그룹 4는 심장의 진단적 초음파를 시행하고 허혈성심질환이 있으면서 성별이 남자이고 연령이 73세 이상인 집단군으로 고혈압을 동반상병으로 가질 확률 값이 81.2%로 나타났고, 그룹 5는 심장의 진단적 초음파를 시행하였고 허혈성심질환이 있으면서 성별이 여자이고 연령이 40세 미만인 집단군으로 확률 값이 100.0%로 나타났다. 그

리고 그룹 7은 심장의 진단적 초음파를 시행하지 않았고, 허혈성심질환이 없으면서 성별이 남자이고 연령이 35세 미만인 집단군으로 고혈압을 동반상병으로 가질 확률 값이 12.9%로 나타났으며, 그룹 9는 심장의 진단적 초음파를 시행하지 않고 허혈성심질환이 없으면서 성별이 여자이고 연령이 35세 미만인 집단군으로 확률 값이 17.2%로 나타났다. 이와 같은 연구결과를 통해 당뇨환자의 고혈압 동반에 대한 Outlier(극단적 케이스)의 임상적 요인을 명확하게 파악하는 것은 불가능하다. 그 이유는 본 연구는 당뇨환자의 고혈압 동반에 대한 Outlier(극단적 케이스)를 찾을 수 있는 데이터 질 관리 알고리즘을 개발하고자 하였고, 이에 따라 극단적인 케이스를 가장 잘 찾아 주는 것에 주안점을 두고 모형을 개발하였기 때문에 본 모형에서 개발된 특성과 임상적 특성을 연계시키는 것은 연구목적과는 부합되지 않기 때문이다.

본 연구에서 수집한 2011년, 2012년 퇴원손상심층조사 자료 중 당뇨환자의 고혈압 동반 데이터 질 관리 대상은 총 22,740건이다. 본 연구의 연구결과인 당뇨환자의 고혈압 동반 Outlier(극단적 케이스)로 볼 수 있는 4개 집단에 포함되는 자료 즉, 데이터 질 관리 알고리즘 기반의 당뇨환자의 고혈압 동반 데이터 질 관리 대상은 그룹 4 108건, 그룹 5 2건, 그룹 7 36건, 그룹 9 72건 등 총 218건이다. 질 관리 대상이 되는 22,740건 모두에 대해 자료를 확인하고, 수정 보완하여 데이터의 질을 향상시키고, 관리하는 것이 가장 확실하고, 이상적인 방법이다. 하지만 전체 자료 모두를 질 관리한다는 것은 인력적, 비용적, 시간적 등을 고려할 때 불가능한 것이 현실이다. 따라서 당뇨환자의 고혈압 동반 Outlier(극단적 케이스)에 대해 자료를 우선 점검 대상으로 하고, 고혈압 동반 유, 무에 대한 내용을 점검, 보완, 관리하여 데이터의 질적 수준을 향상시킬 필요가 있다. 따라서 본 연구는 데이터의 질적 수준을 향상시키기 위해 당뇨환자의 고혈압 동반 데이터 질 관리 알고리즘을 개발하고, 이를 제시하였다는데 의미가 있다고 할 수 있다.

Ⅴ. 결론

본 연구에서는 외국에서 보건의료 데이터 질 관리를 위해 사용하고 있는 Outlier Detection Method 방법에 따라 대화식 의사결정나무 모형을 이용하여 당뇨환자의 고혈압 동반 데이터 질 관리 알고리즘을 개발하여 보건의료 데이터 질 관리를 위한 기초자료로 제공하고자 하였다. 본 연구는 보건의료 데이터 질 관리 필요성 제시 및 연구 목적에 따라 데이터 질 관리 알고리즘을 개발하고, 이를 통해 효율적 데이터 질 관리 방법론 제시하는 기초연구로서 의미가 있다고 할 수 있다. 하지만 본 연구에서 개발된 데이터 질 관리 알고리즘이 의미가 있는지 실제 현장에 적용하여 검증하지 못한 것은 본 연구의 제한점이라 할 수 있다. 따라서 데이터 질 관리 알고리즘 개발에 대한 실제 적용 및 실제 적용을 통한 검사 결과, 투약 등 임상적 요인 반영 등 모형 고도화에 대한 후속 연구가 필요하다.

우리나라의 보건의료 데이터 질 향상을 위해서는 우선적으로 보건의료 데이터 질 향상에 대한 국가적, 국민적 공감대 형성이 필요하다. 보건의료 데이터의 낮은 질적 수준은 국민, 의료기관, 국가 차원의 경제적 손실을 야기하고 있음에 따라 보건의료 데이터 질 향상에 대한 필요성 등 공감대 형성이 절실하며, 보건의료 데이터 질 향상의 필요성에 따라 IT 기술을 활용한 보건의료 데이터 질 관리 오류 검증 시스템 개발 등 국가적 차원의 보건의료 데이터 질 향상 방안 마련이 필요하다.

REFERENCES

1. M.G. Kim, Y.H. Cho, J.H. Park, ETRI(2013), Healthcare Big Data Industry Forecast and competitiveness directions, IT issue report, Vol.27;2-4.
2. T.M. Song(2013), Big data Trends and Utilization of Health & Welfare in Korea, Science and Technology Policy Institute, Vol.23(3);56-73.
3. Medical Record Institute(2002,) HEALTHCARE DOCUMENTATION: A REPORT ON INFORMATION CAPTURE AND REPORT GENERATION, pp.11-14.
4. W.W. Eckerson(2002), Data Quality and the Bottom Line: Achieving Business Success Through a Commitment to High Quality Data. The Data Warehousing Institute Report Series, Chatsworth, USA, pp.11-15.
5. http://www.informationweek.com/healthcare/clinical-information-systems/poor-data-management-costs-healthcare-providers/d/d-id/1105481?
6. Korea Database Agency(2010), 2010 Data quality management maturity level research report, p.13.
7. I.S. Cho(2009), Assessing the Quality of Structured Data Entry for the Secondary Use of Electronic Medical Records, Med Informatics, Vol.15(4);423-431.
8. Juliano(2011), A Systemic Review Of Outlier Detection Techniques In Medical Data: Preminary Data, In Proceedings of the International Conference on Health Informatics (HEALTHINF-2011), pp.575-782.
9. K. Suganya, S. Dhamodharan(2014), Assessment of Data Quality in healthcare Using Association Rules, International Journal of Engineering and Advanced Technology, Vol.3(4);36-37.
10. http://kosis.kr/statHtml/statHtml.do?orgId=350&tblId=DT_35001_A071111&conn_path=I2
11. S.H. Kang, H.S. Seok, W.J. Kim(2013), The Variation of Factors of severity-adjusted length of stay(LOS) in acute stroke patients. The Journal of Digital Policy & Management, Vol.11(6);221-233.
12. Y.M. Kim(2011), A study on analysis of factors on in-hospital mortality for community-acquired pneumonia, Journal of the Korean Data & Information Science Society, Vol.22(3);389-400.
13. Y.M. Kim, D.G. Cho, S.O. Hong, E.J. Kim, S.H. Kang(2014), Analysis on Geographical Variations of the Prevalence of Hypertension Using Multi-year Data, The Korean Geographical Society, Vol.49(6);935-948.
14. M. Ankerst, C. Elsen, M. Ester, H.P. Kriegel(1999), Visual Classication: An Interactive Approach to Decision Tree Construction, KDD '99 Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and datamining, pp.392-396.
15. E.H. Jee(2015), Study on Optimization of Customer Satisfaction hospital, Doctoral thesis Health Administration, Inje university Graduate school, p.4.
16. J.H. Lim, S.H. Kang(2015), Convergence-based analysis on geographical variations of the smoking rates, Journal of Digital Convergence, Vol.13(8);375-385.
17. I.S. Park, E.J. Kim, Y.M. Kim, S.O. Hong, Y.T. Kim, S.H. Kang(2015), A Study on Regional Variations for Disease-specific Cardiac Arrest, Journal of Digital Convergence, Vol.13(1);353-366.
18. S.H. Park, B.D. Hwang(2013), The Effect of Their Sense of Depression and Suicidal Thinking for Managerial Characteristics in Hypertense and Diabetic Patients, The Journal of health Service Management, Vol.7(4):221-232.