순차패턴을 이용한

고혈압관리 의사결정지원시스템 모형 개발

연세대학교 보건대학원

보건정보관리학과

김 화 영

순차패턴을 이용한

고혈압관리 의사결정지원시스템 모형 개발

지도 채 영 문 교수

이 논문을 보건학 석사학위 논문으로 제출함

2004년 12월 일

연세대학교 보건대학원

보건정보관리학과

김 화 영

김화영의 보건학석사 학위논문을 인준함

심사위원_____인

심사위원_____인

심사위원_____인

연세대학교 보건대학원

2004년 12월 일

감사의 글

논문이 완성될 때까지 훌륭한 논문은 아니더라도 부끄러운 논문은 되지 말자고 스스로를 위로했지만 결국 부족한 논문이 되어 부끄럽습니다.

보다 많은 것을 배우고자하는 의욕으로부터 출발하였지만 처음 가졌던 마음만큼 채우지 못했다는 아쉬움이 남습니다. 하지만 대학원 생활동안 부족한 저에게 지도말씀을 아끼지 않으시고 끝까지 이끌어주신 채영문 교수님께 머리 숙여 감사의 뜻을 올립니다. 소중한 시간을 할애하여 세심하게 지도하여 주신 이병화 교수님께 감사드립니다. 또한 논문의 기술적인 부분을 세심히 살펴주신 이동하 박사님께 진심으로 감사를 드립니다.

대학원 생활에 많은 도움을 주었던 이윤희 언니, 윤진희 언니, 유승윤 언니, 이선미, 이진에게 고마움을 전합니다. 또한 비슷한 주제로 논문을 쓰면서 고생을 함께한 혜영이와 작은 기쁨을 함께하고 싶습니다. 힘들때면 항상 정신적으로 편안함을 주었던 친구 진희, 애진, 보영에게 고마움을 전합니다.

언제나 제 삶의 버팀목이 되어주시는 친정 부모님, 공부하는 며느리를 위하여 정성어린 배려를 하여주신 시부모님, 항상 든든한 지원을 해 주는 미라 아가씨, 논문을 쓰는 동안 뱃속에서 함께 고생한 이제 100일이 되는 아들 우현이, 그리고 마지막으로 부족한 아내에게 한결같은 사랑과 관심으로 도와주고 격려해 준 사랑하는 남편에게 다시 한번 감사한 마음을 전하며 이 기쁨을 함께 나누고 싶습니다.

2004년 12월

김 화 영 올림

목 차

국문요약 ····· v

Ⅰ. 서 론 ····· 1

1. 연구배경 및 의의 ····· 1

2. 연구의 목적 ····· 5

Ⅱ. 이론적 배경 ····· 6

1. 데이터 마트 ····· 6

2. 데이터 마이닝 ····· 8

3. 만성질환 시스템 구축에 관한 연구 ····· 15

4. 고혈압 위험요인에 관한 연구 ····· 16

Ⅲ. 연구방법 ····· 20

1. 연구대상 및 범위 ····· 20

2. 연구의 틀 ····· 21

3. 연구분석 및 방법 ····· 22

Ⅳ. 결 과 ····· 32

1. 데이터 마트 구축 ····· 32

2. 연구 대상자의 특성 ····· 34

3. 회귀분석 결과 ·····35
4. 데이터 마이닝 기법을 통해 도출된 고혈압 위험요인 예측을 결정짓는 규칙 ·····38
5. 고혈압 위험요인 예측모형의 타당성 검증 ·····45
6. 순차패턴을 이용한 고혈압관리 조언 시스템 모형 개발 ·····46

Ⅴ. 고 찰 ·····48

Ⅵ. 결론 ·····53

참고문헌 ·····55

부록 ·····59

Abstract ·····64

표 목 차

표 1. 연구대상자 ····· 20
표 2. 데이터 항목 ····· 23
표 3. 알코올 섭취 심각도 분류 ····· 26
표 4. 연구에 사용된 변수 ····· 27
표 5. 연구대상자의 일반적 특성 ····· 34
표 6. 로지스틱 회귀분석에 따른 기간별 생활습관 (남자) ····· 36
표 7. 로지스틱 회귀분석에 따른 기간별 생활습관 (여자) ····· 36
표 8. 로지스틱 회귀분석에 따른 기간별 생활습관 (전체) ····· 37
표 9. 순차패턴 결과 (남자 가족력 유무) ····· 43
표 10. 순차패턴 결과 (여자 가족력 유무) ····· 44
표 11. 순차패턴 결과 (전체 대상자) ····· 44
표 12. 의사결정나무와 로지스틱 회귀분석의 비교 ····· 45
표 13. 의사결정나무와 로지스틱 회귀분석의 오분류률 비교 ····· 45

그 림 목 차

그림 1. 시스템 구조도 ····· 31
그림 2. 고혈압관리 데이터 마트 ····· 33
그림 3. 의사결정나무 분석 결과 (성별포함) ····· 40
그림 4. 의사결정나무 분석(성별 제외 ····· 40
그림 5. 건강검진결과 Review ····· 46
그림 6. 고혈압 고위험군 예측정모 및 진단 ····· 47
그림 7. 생활습관의 개선을 위한 건강증진교육 정보 ····· 47

국 문 요 약

고혈압 위험집단 관리를 위한 조언시스템을 개발하고자 데이터 마트를 구축하였으며, 데이터 마이닝 기법을 적용하여 고혈압 위험요인 예측모형을 제시하고 이를 활용한 고혈압 관리 조언 시스템을 개발하였다. 이에 따라 국민건강보험공단의 데이터베이스에 있는 4년간의 건강검진 자료중 1994-1998년도 고혈압 증상이 없고 2000년에 비고혈압이거나 고혈압인 집단 92,354명을 연구대상으로 하였다. 본 연구에서는 로지스틱 회귀분석과 의사결정나무분석으로 고혈압 위험인자를 알아보았고 두가지 방법의 성능을 비교하였다. 로지스틱 회귀분석으로 고혈압 위험인자를 알아본 결과 남자의 경우 연령이 증가할수록, 비만기간이 길수록, 과음기간이 길수록 ,흡연기간이 길수록 고혈압에 걸릴 확률이 높았으며 통계적으로 유의하였다. 의사결정나무 분석을 이용하여 고혈압에 미치는 위험요인을 설명하는 주요 변수로는 성별 변수를 제외한 경우는 흡연, 음주, 비만이였고, 성별 변수를 포함한 경우에 남자들의 고혈압 위험요인은 비만, 음주 변수였고 여자들은 나이, 비만 변수였다. 의사결정나무가 로지스틱 회귀분석보다 더 낮은 정확도(75.17%)와 감수성(73.38%)을 보인 반면 특이도(74.35%)는 더 높았다. 또한, 본 연구는 순차패턴을 이용하여 시간의 경과에 따라 고혈압에 영향을 미치는 각 위험요인의 모든 조합에 대해 영향도를 계산하여 다수의 위험요소가 나타나는 패턴을 찾아내었다. 그 결과 위험요인을 가진 잠재적인 고혈압 환자를 예측할 수 있는 순차적인 패턴을 증명하였다.

이와 같이 데이터 마이닝 기법으로부터 도출된 규칙을 활용하여 고혈압

관리를 위한 조언시스템을 개발하였으며, 시스템의 구성은 Physical exam Review, Prediction, Recommendation으로 이루어져 있다.

본 연구의 결과를 통해 고혈압관리 조언 시스템을 확대 적용하여 고혈압 환자의 조기발견과 더불어 일차예방을 시도하다면 환자 발생을 줄이는데 큰 도움을 줄 수 있으리라 생각된다.

Ⅰ. 서 론

1. 연구배경 및 의의

순환기계 질환은 우리나라 사람의 사망원인 가운데 24.8%를 차지하고 있으며(통계청,2001), 고혈압은 뇌졸중이나 허혈성심질환 같은 사망률이 높은 순환기계질환의 주요 위험요인으로서 중요한 의미를 가진다(김현창 등,1999). 고혈압은 선진국, 후진국의 구별없이 유병율이 약 10-20%로 가장 흔한 질환일 뿐 아니라 뇌혈관계, 심장 그리고 신장에 까지 합병증을 초래하는 만성퇴행성 질환으로 전세계적으로 중요한 보건학적 관심사가 되고 있다. 우리나라에서도 경제발전에 따른 생활수준의 향상과 의료기술의 발달에 의해 사망양상의 변화가 초래되어 주요사인이 비감염성 질환으로 바뀌면서 고혈압관리가 암의 관리와 함께 중요한 국가적인 관심사가 되고 있다.

고혈압은 발생시점이 불분명하므로 환자의 조기발견과 더불어 일차 예방을 시도하는 것이 환자발생을 줄이는데 큰 도움을 줄 수 있다. 또한 고혈압은 유전인자와 환경인자에 따라 영향을 많이 받는데 특히 환경인자들 중 주요부분을 차지하는 비만도, 음주, 흡연, 운동량, 염분섭취 등의 생활습관을 개선함으로서 일차예방에 크게 기여할 수 있다. 미국 Joint National Committee의 제 7차 보고서에도 고혈압 예방과 치료를 위해 일차적으로 생활습관변화를 들고 있다.

기존의 고혈압의 발생 요인과 관련된 연구(Reed 등, 1985; Kornitzer등, 1999)에서 추론되는 변수로는 연령, 성별, 인종, 고혈압의 가족력 등의 인

구학적·유전적 요인과 혈중 지질, 체중등 생리적 요인들과 환경적 요인 및 이들의 복합적인 작용에 의한 것으로 가장 개선 가능한 환경적 위험요인으로는 높은 비만도, 음주, 적은 운동량, 염분섭취량 등이었다. Kuller등(2001)이 연구한 바에 따르면 혈압 조절 여부와 관련성이 있는 생활습관은 체중감소, 과도한 음주 절제, 금연, 저식염섭취, 신체활동의 증가 등이었다. Ockene 등(1990)의 연구에서도 고혈압의 합병증인 심혈관계질환 역시 비만, 음주, 흡연, 운동 등의 생활습관과 밀접한 관련성이 있으며 생활습관들의 긍정적인 변화가 심혈관계질환의 발생률이나 사망률을 낮추는데 중요한 역할을 한다고 하였다. 또한 1990년에 실시한 우리나라의 전국 혈압조사에 따르면 연령, 고혈압 가족력, 비만, 음주등이 혈압과 강한 연관성을 보여 고혈압의 위험요인 중에서도 생활습관이 매우 중요 하였다(김정순, 1991). 이러한 연구는 고혈압의 치료보다는 위험요인을 규명하고 이를 제거함으로서 고혈압을 관리하는데 유용한 자료를 제시할 수 있을 뿐 아니라 혈압을 낮추거나 합병증의 예방적 관리에도 중요하다.

역학 연구를 통하여 위험요인을 밝히고 구명된 위험요인에 대한 국가적 예방 및 관리를 체계적으로 수행하기 위해 미국 국립보건원에서 1998년도부터 6개년 계획으로 심혈관질환 국가감시체계를 개발 중에 있다(박종구 등,2000). 최근 우리나라에서도 고혈압관리사업이 보건소의 주요사업으로 대두되어 2000년부터는 국민건강증진기금을 재원으로 전국예방사업의 일환으로 주민교육을 실시하고 있다. 하지만 표준화된 고혈압관리지침과 국가수준의 고혈압관련 통계자료의 부재로 고혈압관리사업의 내용 역시 단지 내소하는 환자 중심의 치료위주사업으로 환자조기발견과 위험요인 예방에 대한 홍보 및 효율적인 보건교육사업은 대단히 저조하다. 즉 고혈압 환자

의 치료 및 예방은 여러 가지 복합적 요인이 작용하여 이루어지는 만큼 이러한 요소들을 고려하여, 다양한 차이를 나타내는 각 개인을 예방, 치료하는데 있어 복합적 정보를 제공할 수 있는 지침 및 연구가 필요하다. 즉 고혈압의 예방및 치료에 관여되는 요인들이 매우 많기 때문에 이들 정보를 종합하여 고혈압 위험군 관리 정책을 수립하는데 도움을 줄 수 있는 새로운 정보체계가 필요하다. 이러한 고혈압관리 정책수립에 있어서 「국민건강보험공단」에서 보유하고 있는 건강검진자료의 중요성은 매우 크다고 할 수 있다. 건강검진자료는 매년 수백만 건의 개인별 의료정보를 가지고 있어 질병의 유형을 분석하고 국가 의료정책을 수립할 수 있는 중요한 기초자료임에도, 그동안 건강검진자료의 활용도는 미흡하였다.

최근 의료분야에서도 컴퓨터가 단순한 환자 데이터의 관리에서부터 진단 및 치료의 의사결정에까지 광범위하게 활용되고 있다. 의료정보자원의 효율적 활동을 위한 지식기반 경영이 요구되고 있는 실정이다. 의료분야의 지식경영은 국민의료복지에 관한 정책 수립시 결정적 자료가 되는 건강검진 자료에서의 지식경영과 환자 개개인의 임상적 정보를 포함하고 연구, 교육, 진료에 활용할 수 있는 자료가 되는 병원에서의 지식경영으로 나눌 수 있다(지원철 등, 1999). 이 중 건강 검진자료는 보건의료 자원의 부족과 같은 문제점을 해결하기 위해서도 이에 상응하는 정보의 생산 및 활용이 필요하다. 그러나 이와 같은 정보관리의 중요성에도 불구하고 전국에서 수집되는 방대한 양의 정보가 단순히 사례관리 등에만 사용되고 주요 정책이나 연구에 도움이 되는 생산에는 거의 활용이 안 되고 사장되어오고 있다. 이에 따라 건강검진자료와 같은 다양하고 방대한 양의 자료를 효과적으로 다루기 위해서 데이터 마트(Data Mart)를 구축하는 기술이 필요하다.

고혈압관리를 위해 건강검진자료의 데이터 마트를 구축하면 다양하고 대량의 자료로부터 분석에 필요한 자료를 신속하고 효과적으로 추출할 수 있다는 이점이 있다. 방대한 양의 데이터로부터 함축적이며 잠재적 유용성이 있는 정보를 발견할 수 있는 지식 탐사 방법인 데이터 마이닝(Data Mining)에 대한 중요도가 더욱 증대되고 있다(Fayyad 등, 1996). 데이터 마이닝은 패턴인식 기술이나 통계기법, 수학적 알고리즘을 이용하여 의미있는 새로운 상관관계, 패턴, 추세 등을 발견하는 과정으로, 의료분야에서도 다른 분야와 마찬가지로 의료정보자원의 효율적 활용을 위한 지식기반 경영과 이의 요소정보기술로써 데이터 마트와 데이터 마이닝의 중요성이 부각되고 있다.

본 연구에서는 의료분야의 의사결정지원을 위한 지식경영체제 도입을 촉진시키기 위한 방안으로써 데이터 마이닝 기법을 적용하여 대규모 데이터 내에 존재하지만 숨겨져 있는 상호관련성과 패턴에 대한 탐색을 통해 유용한 지식을 이끌어 내고자 하였다. 데이터 마이닝과 같은 새로운 기법이 질환의 예방과 관리에 적극 활용될 필요가 있다. 데이터 마이닝 기법을 적용하여 고혈압에 어떠한 위험요인이 얼마나 중요한지를 유추해 내고 고혈압군과 비고혈압군의 평소생활습관의 특성을 분류, 구명함으로써 고혈압 진단을 예측하고자 하였다. 또한 여러 데이터 마이닝 기법들의 결과에 대한 성과를 비교 분석하고 향후 고혈압 위험군 관리 정책수립에 있어서의 최신 정보기술을 활용하여 의사결정 기능을 향상함으로써 우리나라의 고혈압관리 국민 건강서비스를 향상에 기여하고자 한다.

2. 연구의 목적

본 연구는 국민건강보험공단의 데이터베이스에 있는 4년간의 건강검진 자료를 이용하여 데이터 마이닝 기법을 적용하고 고혈압 위험집단을 관리하기 위한 조언 시스템 모형 구축하기 위해 시도되었다.

구체적인 목적은 다음과 같다.

첫째, 국민건강검진 데이터베이스를 이용하여 고혈압관리 데이터 마트를 구축한다.

둘째, 데이터 마이닝 기법 중 순차패턴(Sequence Pattern)을 적용하여 고혈압관리 정책수립에 필요한 주요 규칙과 패턴을 도출하고 고혈압 위험요인 예측모형을 제시한다.

셋째, 데이터 마이닝 모형에 대한 타당성을 검증한다.

넷째, 이상의 과정을 통하여 고혈압 위험집단 관리를 위한 조언 시스템 모형을 구축한다.

Ⅱ. 이론적 배경

1. 데이터마트

가. 데이터 마트의 개념 및 필요성

데이터 마트는 운영데이터나 기타 다른 원천으로부터 수집된 데이터 저장소로서, 특정그룹의 지식노동자들을 지원하기 위해 설계된다. 데이터는 전사적인 데이터베이스, 또는 데이터 웨어하우스 , 그리고 보다 특화된 것으로부터 끌어낼 수도 있다. 데이터 마트의 중요성은 분석이나, 내용물, 표현 및 사용의 용이성 등의 측면에서 특정 그룹의 지식노동자의 명확한 요구에 부합되는 데에 있다.

데이터 마트의 설계는 사용자요구분석으로부터 시작하고, 데이터 웨어하우스는 이미 존재하는 데이터가 어떤 것인지와, 그러한 것들이 어떻게 수집될 수 있는지에 대한 분석으로부터 출발하는 경향이 있다. 데이터 웨어하우스는 데이터의 중앙 집합체이고, 데이터 마트는 데이터 웨어하우스 등으로부터 유도될 수 있는 데이터의 저장소로서 특별히 계획된 목적을 위해 접근의 용이성과 유용성을 강조한 것이다. 일반적으로, 데이터 웨어하우스는 전략적이지만 다소 덜 다듬어진 개념이며, 데이터 마트는 전술적이며 당장의 요구에 부합하는 데에 목표를 두는 경향이 있다.

나. 의료정보 데이터베이스의 국내외 연구 동향

New Orleans 가정의학과에서는 장기간의 병력과 다면적인 원인이 관

계된 만성질환관리 규칙을 발견하기 위해 30,383명의 당뇨병 환자자료를 가지고 데이터 웨어하우스를 구축하고 flat file로 데이터 마이닝 분석을 적용한 연구가 있었다(Breault, 2002). 미국의 California-Davis대학에서는 생애 주기(Life Cycle)를 주제로 데이터 웨어하우스를 구축하고 생애관리를 위한 스타 스키마(Star schema)를 제시하였다(Verma, 2001).

또한 일본 Osaka대학에서는 인생주기 안에 의료행위 발생(medical event)를 관찰할 수 있는 시간 지향성 데이터 웨어하우스(Time-oriented Data warehouse)의 모형을 오라클을 사용하여 구축하고 의미 있는 의료사건을 감지할 수 있는 효과와 효율성을 측정 보완하였다(Yamamoto, 2002).

2. 데이터 마이닝

가. 데이터 마이닝의 개념 및 필요성

데이터 마이닝은 일반적으로 데이터베이스에 있는 방대한 양의 정보에 숨어있는 지식을 자동적으로 추출하는 과정이다. 데이터 마이닝은 방대한 데이터베이스로부터 숨어있는 예측정보의 추출이라고 설명할 수 있다. 매우 방대한 데이터 속에 숨어있는 고급정보를 추출해서 의사결정, 예측, 예보에 응용하고자 하는 기법이다.

나. 데이터 마이닝의 국내외 연구 동향

고혈압 위험요인을 코호트 연구, 환자-대조군 연구 방법을 활용하여 파악하는 선행연구들은 꾸준히 수행되어져 왔다. 그러나 이 연구는 데이터 마이닝 기법을 활용하여 위험요인을 발견하고 파악하고자 한 것이므로 데이터 마이닝 기법을 활용한 연구자료에 대해서만 비교·분석하였다.

채영문 등(2000)의 「Data Mining for Knowledge Management in a health insurance Domain」은 건강증진 프로그램의 예로서 고혈압관리 프로그램에 필요한 지식도출을 위해서 데이터 마이닝 기법을 이용하여 국민건강보험공단 데이터베이스 사이에 묻혀 있는 패턴을 발견하고 규칙을 추론하고자 한 연구로, 국민건강보험공단에서 시행하는 1998년도 건강검진에 참여한 127,886명의 피보험자로부터 무작위 표본추출 하였다. 데이터 마이닝 결과, CHAID((Chi-Squared Automatic Interaction Detection)) 알고리즘

이 64.04%로 가장 예측력이 높았고 다음은 로지스틱 회귀분석이 63.84%였고 C5.0은 59.22%였다. 민감도 또한 CHAID 알고리즘이 76.30%로 가장 높았고 다음은 로지스틱 회귀분석이 64.36%였고 C5.0은 59.34%였다. 그러나 특이성은 CHAID 알고리즘이 52.3%로 가장 낮았고 로지스틱 회귀분석이 63.33%로 가장 높았다.

호승희 등(2000)의 「Development of decision support system for the management of hypertension using data mining technology」은 데이터 마이닝 기법을 이용하여 임상적으로 중요한 위치를 차지하고 있는 고혈압 환자의 특성과 치료에 따른 예후를 예측할 수 있는 지식을 발굴하고 이를 활용한 의사결정지원시스템을 개발하고 이의 임상적용의 타당성을 검증하고자 한 연구로, 그 결과 기존 임상이론의 처방규칙에 의한 시스템의 처방 시 보다 데이터 마이닝 기법을 통해 도출된 처방규칙을 혼합한 복합처방규칙에 의한 시스템의 처방 시 혈압조절군의 비율이 보다 높게 나타남을 알 수 있었다.

Stephen등의 「Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance」패턴들의 가치를 확인할 수 있는 연관규칙을 기반으로 새로운 감시 처리를 정의하여 자동적으로 실세계의 데이터 셀에서 흥미있고 예기치 않은 패턴들을 발견할 수 있는 프로세스를 기반으로 한 자동 패턴 발견 시스템을 개발하였다.

다. 순차패턴의 개념 및 필요성

순차패턴 마이닝(Mining Sequential Patterns)은 시퀀스로 구성된 주어

진 데이터 베이스에서, 빈번히 나타나는 시퀀스를 찾는 데이터 마이닝의 한 기법이다 (11th International Conference on Data Engineering, 1995). 다시 말하자면, 시간적인 순서에 따른 페이지의 집합을 나열했을 때 세션 간의 아이템의 연관성을 찾아 주는 데이터 마이닝 기술이다. 이는 한 요인에 영향을 미치는 선행 요인을 찾아내는 것이다. 일정기간 동안의 트랜잭션의 데이터베이스가 주어지면 어떤 항목 집합의 존재가 다른 항목 집합의 다음에 오는가와 같은 트랜잭션간의 패턴을 찾는 것이다. 연속적 또는 일시적인 패턴 기능들은 추세를 식별하기 위해서 오랜 시간에 걸친 기록들의 모음(collection of records)을 분석한다. 연속적인 패턴 기능은 관계된 레코드 집합을 분석할 것이고 시간을 두고 자주 발생하는 패턴을 발견할 것이다.

순차 패턴을 찾는 문제는 사용자의 구매 정보 데이터베이스가 주어졌을 때 사용자가 정의한 최소 지지도(minimal support)를 만족하는 모든 시퀀스(sequence)를 찾아냄으로써, 시간의 경과에 따라 발생하는 사건이나 트랜잭션의 빈번한 패턴을 찾아내는 것이다.

순차 패턴 추출 알고리즘에서 사용하는 기본적인 용어와 전제 조건 및 표기 방법에 대해서 정리하면 다음과 같다.

· 항목 집합(itemset):

단위 데이터의 집합으로 $i = (i^1, i^2, .. i^m)$, $1 <= j <= m$ 인 ij 는 단위 데이터로 표현된다.

· 시퀀스(sequence):

항목 집합의 리스트로 $s = < s^1, s^2, s^n>$, $1 <= j <= m$ 인 sj 는 항목

집합으로 표현된다.

· 시퀀스의 길이:

시퀀스 내의 항목집합의 개수로 k-시퀀스는 리스트의 길이가 k 인 시퀀스를 뜻한다.

· 시퀀스의 포함 관계:

주어진 시퀀스 = $\langle a^1, a^2, a^n \rangle$ 와 = $\langle b^1, b^2, b^m \rangle$ 가 있을 때, 만일 1= $i^1 < i^2 < \ < i^m = m$ 이면서 $a^1 \subseteq bi^1$, $a^2 \subseteq bi^{2,}$, $a^n \subseteq bi^m$ 가 되는 정수 i^1, i^2,.. i^m 이 존재하면 α 는 β에 포함되었다고 한다. 이때, 시퀀스의 포함관계는 $\alpha \subseteq \beta$ 로 표현된다.

· 시퀀스의 지지도(Support):

주어진 데이터베이스에서 시퀀스를 포함된 시퀀스의 비율을 의미한다. 시퀀스 s 의 지지도는 sup(s)로 표현된다.

· 최소 지지도(minimum support):

순차 패턴 마이닝의 조건이다. 주어진 데이터베이스에서 최소 지지도 이상을 가진 시퀀스를 모두 추출하게 된다. 최소지지도의 지정에 따라 결과로 얻는 순차 패턴의 개수와 수행 시간이 결정된다.

· 시퀀스의 신뢰도(Confidence):

시퀀스 지지도를 나머지 시퀀스의 지지로 나눈 것을 신뢰도라고 한다.

신뢰도는 나머지 시퀀스 부분이 전체 시퀀스를 유도하는 강도를 의미한다. 시퀀스 s의 신뢰도는 conf(s)= sup(s)/ sup(remain(s,o))로 표현한다.

본 연구에서는 고혈압 위험요인을 분석하기 위한 것으로서, 순차패턴의 마지막 항목 집합에 특별한 의미가 있다. 따라서 연관규칙의 정의를[1] 근거로 하여 다음과 같이 다시 정의한다.

· 시퀀스의 목표 항목(object item):

연구의 대상이 되는 고혈압 여부에 대한 항목(item)을 목표 항목 이라고 한다.

· 나머지 시퀀스(remain sequence):

주어진 시퀀스 s = < s^1, s^2, s^n> 에서 마지막 항목 집합에 해당하는 시퀀스 <s^n>중 목표 항목 o 이 있는 경우, 나머지 시퀀스는 다음과 같다. remain(s,o) = < s^1, s^2, s^n {o}> 만일, s 내의 항목집합이 o 뿐이었다면, 나머지 시퀀스는 < s^1, s^2, s^{n-1}> 이 된다.

· 테일(tail):

마지막 항목 집합을 테일(tail(s) = <s^n >)이라고 한다.

· 헤드(head):

테일을 제외한 부분은 헤드(head(s) = < s^1, s^2, s^{n-1}> 라고 한다.

---

1) 지지도=Pr(a∩b), 신뢰도=Pr(a∩b)/P(a), 향상도=Pr(a∩b)/Pr(a)Pr(b)

· 시퀀스의 향상도(lift):

시퀀스 지지도를 시퀀스 목표 항목의 지지도와 나머지 시퀀스의 지지도를 곱한 것으로 나눈 것을 의미한다. 향상도는 Head가 Tail에게 영향을 주는 수치를 계산한 것으로 Lift(S) = Support(S)/(Support(Head(S)) * Support(Tail(S)))으로 표현된다. 향상도는 1 이상인 경우 양의 상관관계가 존재하는 것을 의미한다. 예를 들어, 주어진 시퀀스 집합이 다음과 같이 주어졌다고 한다.

s1 = (1 ,3 , 6) -> (2 , 4) -> (4)
s2 = (1 , 6) -> (2, 4) -> (4)
s3 = (1, 2, 6) -> (4, 6)

이 경우 주어진 시퀀스 (1, 6) -> (4) -> (4) 는 두개의 시퀀스에 나타나므로 2/3 = 67% 의 지지도를 갖는다. 이 경우 향상도는 (1,6) -> (4) 가 세개의 시퀀스에 나타나므로, 향상도는 100% 이다. 또 S^1의 헤드는 (1 ,3 , 6) -> (2 , 4) 이고 테일은 (4) 이다. 위와 같은 대상 데이터베이스가 주어졌을 때, 길이가 4 이상이면서 최소 지지도 66%이상인 그 외의 시퀀스 집합은 다음과 같다.

(1, 6) -> (2) -> (4)
(6) -> (2, 4) -> (4)
(1, 6) -> (2, 4)
(1, 6) -> (2, 4) -> (4)

고혈압은 장기간 연령과 더불어 증가하는 질병으로 다년간 축적된 건강검진 자료에서 순차패턴 기법의 활용이 가능하다. 이 기법은 시간적인 흐름과 각각의 과거검진시의 위험요인들의 조합을 고려하여 향후 검진시의

고혈압에 대한 영향도를 판단할 수 있다. 개개의 위험요인과 해당 고혈압에 대한 영향도는 상관관계 분석 등으로 얻을 수 있지만, 시간의 흐름에 따른 각 요인의 모든 조합에 대한 영향도를 계산하기에는 어려움이 있어서 순차패턴 기법 연구가 필요하다. 그러나 아직 기술상의 문제로 국내에서는 보건분야에서는 거의 활용이 되고 있지 않으므로 이 연구에서 우리나라 특성에 맞는 순차패턴 기법을 개발하여 고혈압의 시계열적 역학연구 (time dependent epidemiologic study)에 활용할 필요가 있다.

라. 순차패턴의 국내외 연구 동향

Fu-ren Lin등(2001)의 「Mining time dependency patterns in clinical pathways」은 brain stoke를 관리할 주 진료 경로의 개발을 위해서 데이터 마이닝 기법을 이용하여 환자 데이터와 임상 로그 데이터에 묻혀 있는 패턴을 발견하고자 중요한 cases의 활동들 사이에 의존 관계를 확인하고 프로세스 수행 순서의 패턴을 발견하기 위해 time dependency pattern을 활용하였고 주 진료 경로 결정의 예측도를 평가하였다.

3. 만성질환 시스템 구축에 관한 연구

Larsen(2003)은 통합 건강관리 네트워크의 종합건강관리시스템 내에 당뇨병관리 시스템을 구축하고 이를 통해 당뇨병 환자의 혈압 및 합병증 관리의 효과를 측정한 결과 1998년부터 2002년 사이에 매년 15%이상의 위험지표 감소, 예방적 검사 행위의 10%이상 증가 결과를 보고한 바 있다. 이 시스템은 의원에서의 1차 의료서비스 관리자, 의료 전문가, 환자, 행정가, 당뇨교육자 등의 의료행위가 향상되도록 고안된 것으로 만성질환과 관련된 개인소인, 환경적 요인을 조절하는 다면적 접근방법을 구현하여 4년 동안 합병증 위험소인을 가진 당뇨병 환자의 위험요인을 지속적으로 감소시키는 효과를 성취하였다. 또한 American Cancer Society에서는 1차 의료서비스 제공자들의 예방적 건강관리를 증진하기 위하여 computerized health maintenance tracking system을 개발하여 의원급 의사들에게는 요약 정보(summary report)를 제공하고 환자들에게는 이행력(compliance), 방문시기(visiting status)의 정보를 제공하는 조언 시스템을 개발하였다.

MacIntyre등 (2003)은 65세 이상 노인들의 폐렴, 감기 등의 백신 효과를 관리하기 위한, Community-based Reminder system을 개발하고 타당성 검증을 진행하고 있다.

4. 고혈압 위험요인에 관한 연구

고혈압의 발생원인에 대・해서는 그동안 주로 역학적 연구방법론에 의해 고혈압의 분포 뿐 아니라 유전적 요인이나 가족력, 성별, 연령, 비만도, 염분섭취 등의 관련요인에 대해서 많은 연구가 이루어져 왔다.

고혈압 발생에는 유전적인 요인보다 체중, 음주, 운동, 염분섭취정도, 스트레스 등의 요인이 고혈압 발생에 영향을 미치는 대표적인 위험요인으로 제시되고 있다(Hollengerg 등, 1981; Reed 등 1985). 또한 MaO 등(1998),감신 등(1991)의 연구에서도 고혈압의 여러 가지 위험요인 즉, 유전적 요인과 연령, 성, 고혈압 가족력, 운동량, 흡연, 음주, 비만도, 염분섭취정도, 스트레스등의 사회심리적 요인등 환경요인이 복합적으로 관여한다고 하였다.

Muller등(1998)에 의하면 많은 고혈압 환자들이 생활습관(life style)의 변화로 혈압을 낮추고 관상동맥성 질환의 위험을 감소시킬 것으로 기대할 수 있다고 한다. Labarthe(1998), Izzo 등(1999)의 연구에서 혈압을 낮추면 뇌졸중, 심근경색증, 심부전 및 신부전과 같은 순환기 질환에 의한 사망률과 이환율을 줄일 수 있기 때문에 항고혈압제의 복용과 생활습관 개선의 병행을 통해 혈압을 조절해야 한다고 하였다.

가. 비만

많은 연구들이 비만과 고혈압이 상관관계가 있고 연령이 증가하면서 혈압이 증가하는 경향이 있음을 나타내주고 있다.

Elliot 등(1989)은 여러 환경요인에 관한 회귀분석에서 비만도와 알콜섭

취가 유의한 영향을 미친다고 하였으며 또한 혈압과 관련된 인자로 성, 연령, 나트륨, 칼륨, 알콜의 영향을 조정한 후에 비만도가 유의하게 혈압상승과 관련이 있다고 하였다. He 등(2000) 연구에서는 신체질량지수가 높았거나 신체질량지수가 증가하면 혈압이 증가하고, 반대로 신체질량지수가 감소하면 혈압이 감소하였다.

나. 음주

일본 오사카 의과대학원 연구팀이 교외에 거주하는 일본인 남성 1천명을 대상으로 지난 10년간 음주와 혈압간의 상관관계를 조사한 결과 하루 한 두 잔의 소량 음주도 혈압을 상승시킬 수 있다고 하였다. 연구팀은 하루 한 두 잔의 술을 마시는 남성도 고혈압 위험이 20-30% 높아지는 것으로 측정됐으며, 음주량·빈도와 상관없이 술을 마시는 사람이 그렇지 않은 사람보다 고혈압에 걸릴 위험이 높았다고 밝혔다. 연구팀은 또 알코올을 많이 마시면 마실수록 고혈압 위험은 더 높아지는 것으로 나타났다. 반면에, 미국 브리검 부인병원의 라비 타다니 박사는 의학전문지 '내과학 회보' 최신호에 발표한 연구보고서에서 1979년 터 간호사 건강조사에 참여하고 있는 간호사 7만여명(조사시작 당시 25-42세)의 건강자료를 분석한 결과 중년이하 여성들의 경우 일주일에 2-3잔 술을 마시면 전혀 마시지 않는 여성들에 비해 나중에 고혈압 환자가 될 위험이 낮다는 사실이 밝혀졌다는 연구결과가 나왔다.

다. 흡연

Hyman 등(2001)과 He 등(2002)이 NHNES(National Health and Nutrition Examination Survey) 자료를 이용해 분석한 바에 따르면 고혈압과 흡연 여부와의 관련성이 없었다고 한 반면, Shah 등(2001)은 고혈압 조절이 잘 되는 사람들이 조절이 되지 않는 사람에 비해서 흡연율이 낮다고 하였다

라. 운동

Izzo 등(1999)이 운동을 하게 되면 심폐기능이 향상될 뿐만 아니라 혈청지질과 체중이 감소하고 인슐린 감수성이 증가하여 관상동맥질환을 예방할 수 있다고 하였다. O'Sullivan (2000)은 고혈압은 조직에서의 혈액공급과 관련이 있고 운동을 실천하면 근육의 혈액 순환이 활발해져서 말초 순환의 저항성을 줄일 수 있다고 하였다.

Rigotti 등(1983)은 운동량이 적은 사람들이 흡연하는 경향이 많으며, 체중이 과한 경향도 있을 뿐 아니라 혈압도 더 높고 혈중 콜레스테롤도 더 높은 경향을 갖고 있다고 하였다.

마. 염분 섭취

He 등(2000)의 연구에 의하면 환자 교육을 통해 소금 섭취를 제한할 경우 수축기 혈압과 이완기 혈압이 제한하지 않은 군에 비해 각각

5.8mmHg, 3.2mmHg 더 많이 감소하였다. Jeffrey(1991)등은 실험적 연구를 통해 정상 혈압인 에서나 고혈압인에서 모두 식염섭취 제한을 통해 통계적으로 유의한 혈압 강하효과를 얻을 수 있었다고 보고하였다.

Ⅲ. 연구방법

1. 연구대상 및 범위

이 재료는 국민건강보험공단의 데이터베이스에 있는 4년간의 건강검진 자료 중 고혈압관리 모형에 이용할 수 있는 92,354건의 데이터를 대상으로 하였다. 즉, 고혈압 위험인자를 분석하기 위해서 1994년도 건강검진 901,266건, 1996년 1,031,269건, 1998년 1,032,050건, 그리고 2000년 건강검진 871,222건 중에서 4년 모두 건강검진을 시행하고 자료가 존재하는 집단 중 1994-1998년도 고혈압 증상이 없고 2000년에 비고혈압이거나 고혈압인 집단 92,354명을 모델링의 시험데이터로 활용하였다(표 1). 고혈압 판정기준은 JNC 7(Joint National Committee)에 설정된 기준에 따라 수축기 혈압과 확장기 혈압이 각각 120mmHg미만이고 80mmHg미만 범위로 하였다.

표 1. 연구대상자 (1994-2000년) 단위 : 명(%)

구분	남	여	계
비고혈압 집단	20,754 (22.47)	39,974 (43.28)	60,728 (65.76)
고혈압 집단	16,783 (18.17)	14,843 (16.07)	31,626 (34.24)
계	37,537 (40.64)	54,817 (59.36)	92,354 (100)

2. 연구의 틀

고혈압관리 데이터 마트 구축

-진료지침의 자료, 건강검진자료를 추출, 변환,정제

-고혈압관리 테이터 마트 구축

↓

고혈압 위험요인 예측 모형 도출 및 타당성 검증

-의사결정나무와 순차패턴 기법을 활용하영 고혈압 위험요인 예측 모형 도출

-도출된 규칙에 대한 타당성 검증

↓

고혈압관리 조언 시스템 개발

-데이터 마이닝 기법으로부터 도출된 규칙을 활용한 의사결정지원시스템 개발

-건강검진결과 Review, 고혈압 고위험군 예측 정보 및 진단, 생활습관의 개선을 위한 건강증진교육정보 제공 모듈 개발

3. 연구분석 및 방법

가. 데이터 마트 모형 구축

고혈압 위험요인을 결정짓는 규칙 도출을 위해 데이터 마이닝을 수행하기 위해서는 기존의 운영시스템에서 발생하는 검진 데이터, 문진 데이터, 과거력 데이터를 정기적으로 추출하여 필요에 맞게 재구성하였다. 본 연구에서 사용되는 데이터는 서로 다른 형태로 존재하고 있어 먼저 정제과정을 거쳐서 오류를 최소화하고 변환과정을 거쳐서 의미 불일치를 조정한 후 데이터 마트을 구축하였다. 고혈압관리 데이터마트에 저장하고 나면 주기적으로 그 내용을 갱신하여 데이터 원천들의 갱신된 내용을 반영해 주어야 한다. 고혈압관리 데이터 마트는 별도의 통합된 저장 공간을 구축함으로써 데이터 마이닝을 위한 데이터 원천이 될 뿐만 아니라 고혈압관리 정책수립을 위한 여러 가지 방안에도 접근·활용할 수 있게 될 것이다.

본 연구에서 사용되는 데이터는 서로 다른 형태로 존재하고 있어 먼저 정제(cleansing)과정을 거쳐서 오류를 최소화하고 변환(transform) 과정을 거쳐서 의미 불일치를 조정한 후 데이터마트에 저장된다 (그림 1). 고혈압관리 데이터 마트에 저장하고 나면 주기적으로 그 내용을 갱신하여 데이터 원천들의 갱신된 내용을 반영해 주어야 한다. 고혈압관리 데이터 마트는 별도의 통합된 저장공간을 구축함으로써 데이터 마이닝을 수행할 수 있을 뿐만 아니라 고혈압관리 조언 시스템의 지식경영시스템에도 접근·활용할 수 있는 데이터 원천이 되었다.

1) 원본 데이터의 구성

원본 데이터는 국민건강보험공단에서 건강검진 및 문진 데이터베이스와 과거력 데이터베이스와 같은 운영계 데이터베이스로부터 데이터를 추출하하였다. 고혈압관리 데이터 마트를 구축하기 위해서는 최소 3회 이상의 시계열적인 데이터패턴이 필요하여, 이 연구에서는 3개년이상 연속적으로 건강검진한 사람을 분석 대상으로 결정하고 운영계 시스템으로부터 수집된 4회의 건강검진중 3회 이상 건강검진을 받은 검진자를 대상으로 한 결과 3,835,807건이었다. 분석에 적용된 데이터베이스 및 데이터 항목은 다음과 같다.

표 2. 데이터 항목

구분	내용	레코드수	필드
검진데이터	건강검진결과	3,835,807	79
과거력데이터	과거병력	3,835,807	15
문진데이터	문진 결과	3,835,807	12

2) 데이터 추출 및 변환

원본 파일로부터 데이터 마트에 저장될 데이터를 추출하는 과정으로 보통 기존의 시스템에 저장된 모든 데이터가 필요한 것이 아니고 주제 영역에 적합한 일부만을 추출하였다. 또한 건강검진 관련 원본 파일에서 추출

된 데이터는 질적인 문제로 인하여 데이터 마트에서 그대로 사용할 수 없기 때문에 데이터 정제(cleansing) 기법을 통하여 데이터 품질이 개선해야 한다. 따라서 다음과 같은 방법을 사용하여 원본 데이터를 데이터 마트에 적합한 형태로 변환 (transformation)하였다.

가) 분리

분리는 하나의 데이터 값을 여러 필드로 구분하는 것을 의미한다. citizen_no(주민등록번호)의 경우 (년도 - 두 번째 자리) =pat_ age(나이)이고, 주민등록번호의 일곱 번째 자리 수는 pat_sex(성별)을 구분할 수 있는 필드의 값으로 구분할 수 있다.

나) 일치

각 테이블은 unit_no (일련번호) 와 Year (사업 년도), Date (검진일자)로 일치시키도록 하였다. 또한 데이터 웨어하우스에서는 성별인 pat_sex 필드에서 남자는 1, 여자는 0으로 일치시키도록 하였다.

나. 데이터 마이닝 기법을 활용하여 고혈압 고위험군 예측 규칙 도출

1) 분석항목의 선정 및 측정 수준

예측변수(predictor variable)의 설정을 위해 고혈압 환자의 특성 및 치료에서 중요한 특성으로 파악되어지는 항목들을 선별한다. 이 과정에서 JNC7, 임상의의 견해, 실제 환자 데이터에 대한 기술통계량(descriptive

statistics) 측정을 통한 데이터의 특성 파악 등을 통해 이루어진다. 측정 변수의 범주화 분류 기준과 내용을 살펴보면 다음과 같다.

가) 고혈압 여부

세계고혈압학회가 발표한 고혈압 치료 지침에 의하면 JNC-7(Joint National Committee)의 정의와 분류를 수용하여 고혈압을 수축기 혈압 120mmHg이상, 이완기 혈압 80mmHg이상으로 정의하고 있다. 또한 수축기와 이완기 혈압이 120mmHg미만, 80mmHg미만을 정상으로 규정하고 있다. 고혈압은 다시 수축기와 이완기 혈압이 각각 120mmHg이상-140mmHg미만, 이완기 혈압 80mmHg이상-90mmHg미만일 때를 초기 고혈압, 140mmHg이상-160mmHg미만, 이완기 혈압 90mmHg이상-100mmHg미만일 때를 1도 고혈압, 160mmHg이상-180mmHg미만, 이완기 혈압 100mmHg이상-110mmHg미만일 때를 2도 고혈압으로 분류한다. 이 기준에 따라 1994, 1996, 1998년도에는 정상이고, 2000년도에 고혈압으로 나타나는 31,626명과 1994, 1996, 1998, 2000년도 모두 정상으로 나타나는 60,728명을 분석 대상으로 하였다.

나) 환자 기초 정보

성별과 연령은 측정 변수의 원시 형태를 그대로 각각 명목형과 연속형으로 설정한다.

다) 이학적 검진(physical examination)

신체질량지수 (body mass index; BMI)의지표로 사용하는 신체질량지수는 '체중(kg)/신장(m)^{2}' ‘으로 산출하였으며, 이 산출치를 아시아 성인에서 BMI에 의한 비만판정 기준(한국비만학회)에 의해 BMI가 18.5kg/m^{2}미만인 경우 저체중, 18.5-22.9kg/m^{2}인 경우 정상, 23-24.9kg/m^{2}인 경우 과체중, 25kg/m^{2}이상을 비만으로 범주화한 후 JNC7 보고서에 따라 저체중, 정상, 과체중을 정상으로 분류하였다.

라) 생활습관

음주습관은 만성손상에 대해 순수알코올을 기준으로 해서 한번 마실때 〔술의 양 × 빈도 × 0.79〕을 계산해서 위험수준을 표3과 같이 범주화한 후 비음주군과 알코올 섭취 심각도 ‘저’인 경우를 ‘비과음’으로 알코올 섭취 심각도 ‘중’과 ‘고‘를 ‘과음함‘으로 분류하였다.

표 3. 알코올 섭취 심각도 분류 (English et al, 1995)

성	위험수준		
	저	중	고
남자	1-40g	41-60g	61+g
여자	1-20g	21-40g	41+g

흡연은 흡연과 비흡연으로 구분하였고 금연자를 비흡연으로 분류하였다.

운동은 비지속군과 지속군으로 구분하였다. 주간 운동량이 3-4회 이하로 운동시간이 30분 미만인 집단을 비지속군으로 하고 주간 운동량이 5-6회 이상이고 운동시간이 30분 이상인 집단을 지속군으로 분류하였다.

염분섭취는 저염식과 고염식으로 구분하였고 음식을 좀 싱겁게 먹는편과 보통을 저염식으로 원래 맵거나 짜게 먹는다는 고염식으로 분류하였다.

마) 고혈압 가족력

부모, 형제, 자매가 고혈압을 앓았거나 질환으로 사망한 경우가 있는지 여부에 따라 가족력 유, 무로 구분하였다.

표 4. 연구에 사용된 변수

변수	세부내용	측정수준	
독립변수 응답자의 인구사회학적특성	성	1)남	2)여
	연령		
이학적 검사	신체질량지수	1)정상	2)비만
병력	고혈압 가족력	1)유	2)무
생활습관	과음여부	1)유	2)무
	흡연정도	1)비흡연군	2)흡연군
	지속적인운동	1)안한다	2)한다
	염분섭취	1)저염식	2)고염식
종속변수 고혈압 예방	고혈압	1)유	2)무

다. 의사결정나무를 이용한 고혈압 위험요인 예측 및 대상군 분류

의사결정나무분석은 예측과 분류를 위해 나무구조로 규칙이 표현된다. 따라서 어떤 적용에서는 얼마나 잘 분류하거나 예측하느냐 만이 문제화되기도 한다. 의사결정나무분석의 나무형성 알고리즘은 다양하지만, 가장 보편적인 것으로 CART(Classification And Regression Trees)와 CHAID이고, 좀 더 새로운 알고리즘은 C4.5을 들 수 있다. 본 연구에서는 이러한 알고리즘을 구현할 수 있는 의사결정나무 분석 솔루션 중의 하나인 SAS사의 E-Miner 프로그램 8.0을 사용하였고, 사용한 CHAID는 카이제곱-검정(이산형 목표변수) 또는 F-검정(연속형 목표변수)을 이용하여 다지분리(Multiway Split)를 수행하는 알고리즘이다. CHAID는 각 설명변수의 범주들이 자료를 반응변수의 각 범주들로 구분하는 판별력의 크기에 따라 설명변수의 범주들을 이용하여 나무구조를 만드는 분석방법으로 전체 자료를 둘 이상의 하위노드 (Child Node)로 반복적으로 분할한다. 이 과정에서 설명변수의 범주의 쌍에 대한 반응변수의 유의한 차이가 없으면 설명변수의 범주들을 병합하며, 유의적이지 않은 쌍들이 없을 때까지 과정을 계속한다. 각 설명변수에 대한 최고의 분할을 찾고, 모든 설명변수에 대한 유의성을 조사하여 가장 유의적인 설명변수를 선택한다. 선택된 설명변수의 범주들의 그룹을 사용해 자료를 상호 배반인 부분집합으로 분할하며 각 부분집합에서 정지규칙중의 하나가 만족될 때까지 이 과정을 독립적으로 순환, 반복한다. 데이터 마이닝의 선행작업으로 설명변수와 목표변수의 관계를 파악하기 위하여 카이제곱 분석을 시행하였다. 그 결과 고혈압 군의 년도별 패턴을 알 수 있었다. (부록1)

라. 순차패턴을 이용한 고혈압 위험요인 예측 및 대상군 분류

많은 연구들이 단일 위험 요소가 고혈압 발생에 어떠한 영향을 미치는지 활발하게 연구를 수행하고 있는데 비해 다수의 위험요소를 갖고 있는 경우 시간의 경과에 따라 발생하는 빈번한 패턴을 찾아 고혈압이 발생할 수 있는 위험도가 어느 정도 변할 것인가에 대한 명확한 근거를 제시하고 있는 연구는 드문 실정이다. 따라서 이 연구에서는 순차패턴을 이용하여 시간의 경과에 따라 고혈압에 영향을 미치는 각 요인의 모든 조합에 대해 영향도를 계산하여 다수의 위험요소가 나타나는 패턴을 찾아내었다.

고혈압이 걸릴 '가능성' 과 '가능성 상승 정도' 를 예측하기 위해서 대상 집단을 성별과 가족력 여부에 따라 4가지 경우로 구분한 후 각각의 대상 집단에 대해 순차패턴을 추출하고 지지도, 신뢰도, 향상도에 따라 순차패턴을 선택하였다.

마. 고혈압 위험요인 예측 모형의 타당성 검증

고혈압 위험요인 예측 모형을 선정 한 후 검사 자료에 의해 모형을 평가하였다. 이때 모형이 얼마나 타당성을 가지고 만들어졌는지를 평가하는 것이 중요한데, 민감도(Sensitivity), 특이도(Specificity), 정확도(Accuracy), 그리고 오분류 확률(misclassification probability)을 가지고 평가하였다. 또한 고혈압 위험요인 예측 모형을 산출하기 위한 우선 작업으로 로지스틱 회귀분석 방법으로 고혈압 위험요인과 관련된 영향변수를 파악하였다. 이때 모형의 정확성을 수리적으로 측정했을 뿐 아니라 선택된 모형에 의해

얻을 수 있는 것이 얼마나 되는지 경영 전략적 측면에서도 평가하였다. 이에 따라 여러개의 모형을 평가하여 최종 모형을 선택하고 고혈압 예측을 결정짓는 규칙을 실제 환자 데이터에 적응시킴으로써 모형의 예측력을 평가하였다. 이는 실제 환자의 데이터 값과 데이터 마이닝 모형에 의해 내려진 결정이 일치하는 가를 평가하기 위한 방법이다.

바. 순차패턴을 이용한 고혈압관리 조언 시스템 모형 개발

고혈압관리 조언 시스템은 데이터 마이닝 기법으로부터 도출된 규칙을 활용하여 개발하였다. 의사결정지원시스템 개발도구는 Microsoft Visual Basic 6.0을 이용한다. 의사결정지원시스템 모듈은 건강검진 결과를 제공하는 건강 검진결과 Review, 고혈압 고위험군 예측 정보 및 진단, 생활습관의 개선을 통한 예방과 치료 효과 정보 , 생활습관의 개선을 위한 건강증진교육정보 제공을 제공한다.

고혈압관리 조언 시스템은 순차패턴의 데일에 고혈압 발병 또는 치유 여부를 적용하여, 개별 환자에게 조언해 줄 수 있는 시스템이다. 시스템의 원리는 다음과 같다.

① 대상 데이터베이스를 시퀀스로 변형한 후 그 이후의 데이터를 마지막 시퀀스로 추가하였다.

② 대상 데이터베이스에서 최소 지지도를 바꾸어 가면서 적당한 양의 순차패턴을 추출하였다.

③ 결과 순차패턴에서 신뢰도와 향상도가 높은 유용한 패턴을 추출하였다.

④ 개별 대상자의 패턴을 포함하는 결과 순차 패턴을 추출하여 차기 연도까지

고쳐야 하는 습관에 대해 우선순위를 제시해 주었다.

그림 1은 시스템의 구조도이다. 고혈압관리 조언 시스템은 고혈압관리 데이터마트로부터 순차패턴 데이터베이스를 구축하여, 파티션 관리기로 대상 데이터를 파티션 하여 규칙을 생성하고 그 적용도 파티션에 따라 수행되도록 해주는 모듈로 순차패턴의 정확도를 높여줄 수 있다. 패턴 관리기는 고혈압관리 데이터마트의 내용이 바뀌거나, 새로운 순차패턴 데이터베이스가 교체되었을 때 수행해주는 모듈이다. 사용자 관리기는 데이터마트, 패턴 데이터베이스 및 개별 하위 시스템의 운용 및 관리 기능에 대한 제한적 접근을 관리할 수 있는 모듈이다.

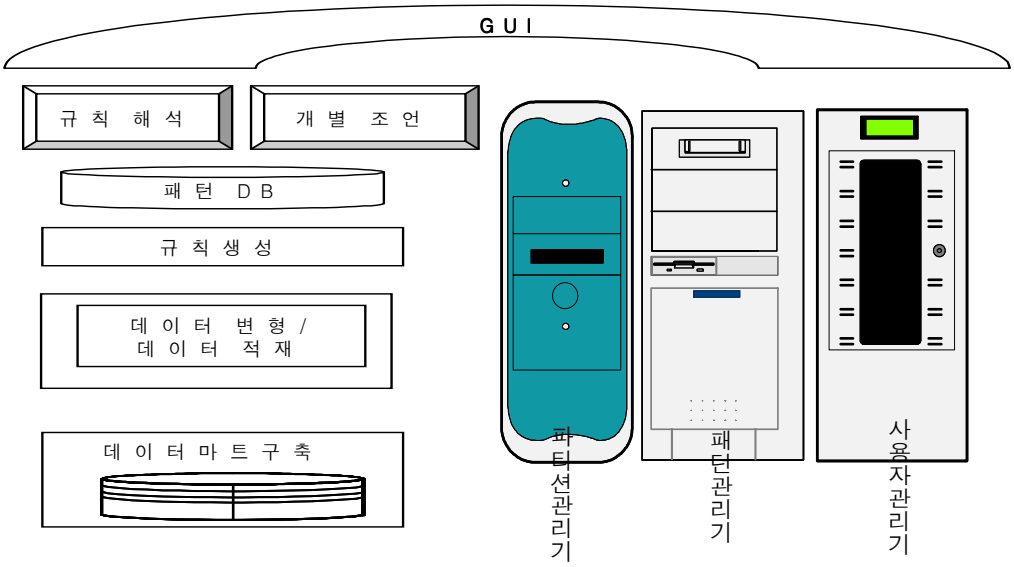

그림 1. 시스템 구조도

Ⅳ. 결 과

1. 고혈압관리 데이터 마트 구축

고혈압관리 데이터 마트 모델은 국민건강보험공단의 운영계 시스템에서 검진 데이터베이스, 과거력 데이터베이스, 문진 데이터베이스로부터 데이터를 수집하여 설계하였다. 고혈압관리 데이터마트는 1994-2000년 사이에 실시된 4회의 검사중 3회 이상 검진을 받은 검진자를 대상으로 하여 건강검진을 3회이상 연속 검진한 사람을 분석 대상으로 결정하여 구성하였다. 고혈압관리 데이터마트는 고혈압 환자 생활습관과 관련된 변수로 구성되며 이들 테이블에 소수의 칼럼으로 구성되고 테이블간에 그림 2와 같은 관계를 이루고 있음을 알 수 있다.

그림 2는 고혈압관리 마트를 구성하는 스타(star-shaped)스키마의 일부를 보여주고 있다. 이 스키마는 실측값을 지닌 1개의 사실테이블을 중심으로 요약정보를 지닌 9개의 차원데이블이 별모양으로 연결된 개념도이다(그림 4). 그림에서 제시한 차원 데이블의 데이터 정의과정을 설명하면 비만도는 정상, 비만으로, 과음여부는 과음함, 비과음으로, 흡연정도는 흡연, 비흡연으로, 염분섭취 정도는 저염도, 고염도로 나누어 분류하였다. 또한 가족력여부, 지속적인 운동여부, 고혈압유무는 예, 아니오로 분류하였다. 비만년, 과음년, 흡연년, 가족력년, 염분섭취년, 비지속적인 운동년은 1994년, 1996년, 1998년중 몇해 동안 위험요소가 있었는가에 따라 3년,2년,1년으로 나누어 분류하였다.

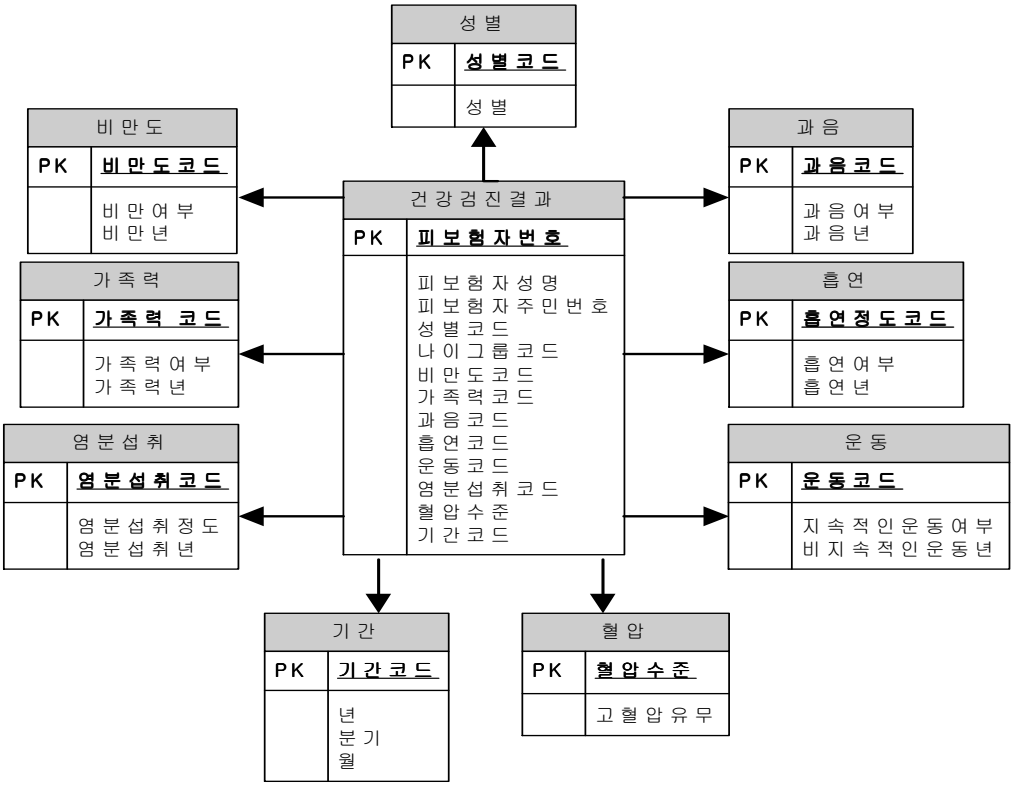

그림 2 . 고혈압관리 데이터 마트 (스타 스키마)

2. 연구 대상자의 특성

이 연구에서의 고혈압군, 비고혈압군의 일반적 특성에 대해서 살펴본 결과 성별, 연령별, 가족력에 있어서 통계적으로 유의한 차이가 있었다. 대상자들의 59.36%가 여성이었고 고혈압군은 남성이 18.17%, 여성이 16.07%를 차지했다. 연령별로는 대상자가 30-39세에서 49.41%로 가장 많았고, 그 다음으로는 40-49세가 36.09%였다. 고혈압 가족력은 고혈압군 중 고혈압 가족력이 있는 경우가 5.29%였고, 비고혈압군 중 고혈압 가족력이 있는 경우는 9.60%였다.

표 5. 연구대상자의 일반적 특성 단위 : 명(%)

항목	내 용	고혈압군	비고혈압군	계	$X2$-값
성별	남 자	16,783(18.17)	20,754(22.47)	37,537(40.64)	3075.73***
	여 자	14,843(16.07)	39,974(43.28)	54,817(59.36)	
연령	20 - 29세	872(0.94)	2,582(2.80)	3,454(3.74)	1049.70***
	30 - 39세	13,794(14.94)	31,835(34.47)	45,629(49.41)	
	40 - 49세	12,583(13.62)	20,751(22.47)	33,334(36.09)	
	50 - 59세	4,144(4.49)	5,308(5.75)	9,452(10.23)	
	60 - 69세	233(0.25)	251(0.27)	484(0.52)	
	70세 이상	0(0.00)	1(0.00)	1(0.00)	
가족력	유	4,888(5.29)	8,865(9.60)	13,753(14.89)	12.00***
	무	26,738(28.95)	51,863(56.16)	78,601(85.11)	
계		31,626(34.24)	60,728(65.76)	92,354(100)	

*: p<0.05, **: p<0.01, ***: p<0.001

3. 로지스틱 회귀분석 결과

표 6에서 남자의 경우 고혈압군과 비고혈압군에 대한 기간별 생활습관 변수들의 로지스틱 회귀분석(logistic regression)의 결과를 나타낸다. 다른 조건이 일정할 때, 연령이 증가할수록 고혈압에 걸릴 확률이 1.006배 통계적으로 유의하게 높았다. 정상인 경우에 비해 1년간 비만인 경우에 1.385배, 2년간 비만인 경우에 1.566배, 3년간 비만인 경우에 1.850배 고혈압에 걸릴 확률이 통계적으로 유의하게 높았다. 그리고 과음하지 않은 군에 비해 2년간 과음한 경우에 1.217배 더 높았고, 3년간 과음한 경우에 1.246배 고혈압에 걸릴 확률이 통계적으로 유의하게 높았다. 하지만, 비흡연군에 비해 2년간 흡연한 경우 0.895배 더 낮았고, 3년간 흡연한 경우 0.918배 고혈압에 걸릴 확률이 통계적으로 유의하게 낮았다.

표 7에서 여자의 경우 고혈압군과 비고혈압군에 대한 기간별 생활습관 변수들의 로지스틱 회귀분석(logistic regression)의 결과를 나타낸다. 다른 조건이 일정할 때, 연령이 증가할수록 고혈압에 걸릴 확률이 1.025배 통계적으로 유의하게 높았다. 정상인 경우에 비해 1년간 비만인 경우에 1.363배, 2년간 비만인 경우에 1.572배, 3년간 비만인 경우에 1.742배 고혈압에 걸릴 확률이 통계적으로 유의하게 높았다. 그리고 과음하지 않은 군에 비해 1년간 과음한 경우에 0.979배 더 낮았고, 2년간 과음한 경우에 1.073배 더 높았고, 3년간 과음한 경우에 1.034배 고혈압에 걸릴 확률이 높았으나 통계적으로 유의하지 않았다.

전체 대상자를 대상으로 한 표 8에서 염분섭취를 제외한 나이와 모든 생활습관 변수가 모두 통계적으로 유의하였다.

표 6. 로지스틱 회귀분석에 따른 기간별 생활습관 (남자)

위험인자	기간[1)]	OR	95% Wald Confidence Limits
연령		1.006	1.003 - 1.009***
비만	1년	1.385	1.253 - 1.532***
	2년	1.566	1.374 - 1.785***
	3년	1.850	1.631 - 2.098***
과음	1년	0.994	0.914 - 1.082
	2년	1.217	0.130 - 1.310***
	3년	1.246	1.189 - 1.305***
흡연	1년	1.011	0.923 - 1.107
	2년	0.895	0.821 - 1.975*
	3년	0.918	0.877 - 0.961***
운동	1년	0.824	0.660 - 1.027
	2년	0.947	0.785 - 1.142
	3년	0.855	0.722- 1.012
염분섭취	1년	1.097	1.013-1.188*
	2년	1.068	0.966 - 1.180
	3년	1.005	0.841 -1.073

주: 1) 건강검진시 위험인자가 나타난 년 수

표 7. 로지스틱 회귀분석에 따른 기간별 생활습관 (여자)

위험인자	기간	OR	95% Wald Confidence Limits
연령		1.025	1.022 - 1.028***
비만	1년	1.363	1.257 - 1.478***
	2년	1.572	1.392 - 1.775***
	3년	1.742	1.518 - 2.000***
과음	1년	0.979	0.914 - 1.049
	2년	1.073	0.990 - 1.164
	3년	1.034	0.967 - 1.104
흡연	1년	1.116	0.855 - 1.456
	2년	0.512	0.145 - 1.814
	3년	2.141	0.922 - 4.969
운동	1년	0.932	0.687 - 1.266
	2년	0.947	0.731 - 1.228
	3년	0.911	0.712 - 1.166
염분섭취	1년	0.993	0.917 -1.076
	2년	0.987	0.891 - 1.095
	3년	1.051	0.983 -1.124

표 8. 로지스틱 회귀분석에 따른 기간별 생활습관 (전체)

위험인자	기간	OR	95% Wald Confidence Limits
연령		1.026	1.024 - 1.029***
비만	1년	1.348	1.266 - 1.435***
	2년	1.556	1.424 - 1.700***
	3년	1.822	1.662 - 1.998***
과음	1년	1.015	0.963 - 1.070
	2년	1.250	0.185 - 1.319***
	3년	1.346	1.299 - 1.395***
흡연	1년	1.493	1.374 - 1.621***
	2년	1.402	1.292 - 1.520***
	3년	1.431	1.378 - 1.485***
운동	1년	0.837	0.701 - 1.000*
	2년	0.890	0.766 - 1.033
	3년	0.815	0.710- 0.937**
염분섭취	1년	1.033	0.977-1.093
	2년	1.011	0.942 - 1.086
	3년	0.995	0.950 -1.043

4. 데이터 마이닝 기법을 통해 도출된 고혈압 위험요인 예측을 결정짓는 규칙

고혈압에 영향을 미치는 위험요인을 규명하기 위한 예측의 과정을 나무구조에 의한 추론규칙에 의해 표현하였다. 그림 5의 의사결정나무 분석에서 사용된 설명변수는 1994-19988년도 비만도, 음주, 고혈압 가족력여부, 흡연, 운동, 염분섭취와 2000년도 연령과 성이였으며 목표변수는 1994-2000년 모두 데이터가 존재하는 집단 중, 1994-1998년도에는 고혈압 증상이 없는 집단이였다.

가. 의사결정나무 분석

의사결정나무 분석을 이용하여 고혈압에 영향을 미치는 각 각의 위험요인 중 어떤 변수가 가장 영향을 많이 미치는 지를 알아보았다. 설명변수 비만은 1994년, 1996년, 1998년중 몇해 동안 비만이였는가에 따라 3년간비만, 2년간비만, 1년간비만, 체중정상으로 측정수준을 나누었다. 또한 과음, 흡연, 운동, 염분섭취도 1994년, 1996년, 1998년중 몇해 동안 위험요소가 있었는가에 따라 측정수준을 나누었다.

고혈압 위험요인 예측을 위해 1994, 1996, 1998년, 2000년 모두 데이터가 존재하는 집단 중, 1994-1998년도에는 JNC 7에 설정된 기준에 따라 수축기 혈압과 확장기 혈압이 각각 120mmHg미만이고 80mmHg미만 범위인 경우인 고혈압 증상이 없는 집단 92,354명에서 2000년도에 비고혈압 군 60,726명으로 전체 65.8%,고혈압이 발생한 군 31,626명의 전체 34.2%로 나타났다.

고혈압 위험요인에 가장 영향을 미치는 변수는 성별이였으며, 그 다음으로는 연령, 비만이였고, 마지막은 비만, 과음이였다. 성별에서 여자가 고혈압 발생 확률은 27.1%였고, 남자는 44.7%였다. 여자들이 41.5세 이하인 경우 고혈압 발생확률은 25.4%였으며, 정상인 군은 74.6%였다. 41.5세 이하인 군에서는 다시 3년간 비만인 군은 고혈압 발생 확률이 32.9%였으며, 정상인 군은 24.9%로 분류되었다(그림 3).

혈압 위험요인 예측에 가장 영향을 미치는 변수는 흡연이였으며, 그 다음으로는 과음, 연령, 비만이였다. 흡연정도에서는 흡연을 1년간 한 군의 고혈압 발생 확률은 44.2%였고, 비흡연군은 30.7%였다. 흡연을 1년간 한 군 중에 과음를 3년 연속 한 군의 고혈압 발생확률은 46.4%였으며, 음주를 1년간 한 군의 고혈압 발생확률은 40.8%였다. 과음를 3년간 한 군에서는 다시 비만여부에 의해 규칙이 도출되었는데 3년간 비만인 군은 고혈압 발생 확률이 57.4%였으며, 정상인 군은 45.2%로 분류되었다(그림 4).

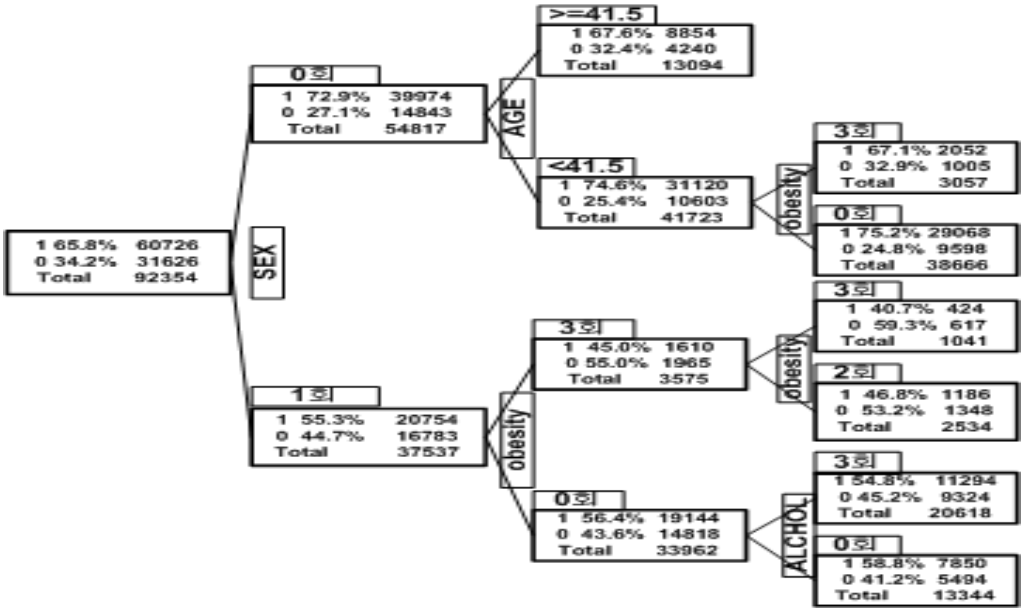

그림 3 . 의사결정나무 분석 결과(성별포함)
(0: 고혈압군 1: 비고혈압군)

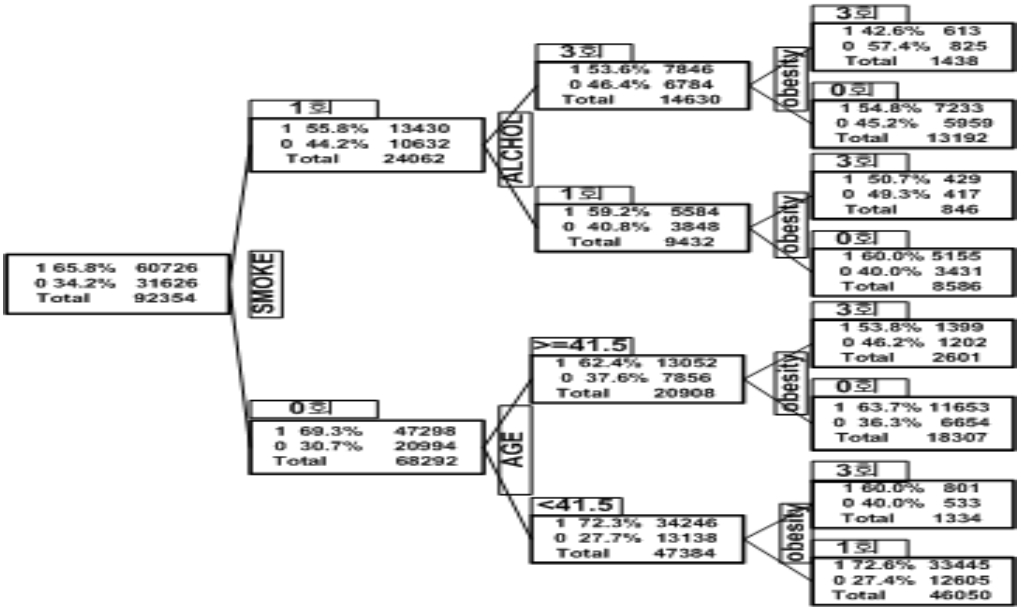

그림 4 . 의사결정나무 분석 결과(성별제외)

나. 순차패턴

순차패턴을 이용하여 시간의 경과에 따라 고혈압에 영향을 미치는 각 위험요인의 모든 조합에 대해 영향도를 계산하여 다수의 위험요소가 나타나는 패턴을 찾아내었다. 3개년에 걸친 사업연도(1994, 1996, 1998)동안 혈압이 정상 수치에 있다가 2000년도에 고혈압이 발생한 집단에 대하여 생활습관에 대한 순차패턴 추출을 시행하여 지지도, 신뢰도, 향상도를 비교하였다.

표 9의 규칙 1,2,3은 남자를 대상자로 하였고 규칙 4,5,6은 남자 대상자 중에 가족력이 있는 군이고 7,8,9는 가족력이 없는 군의 순차패턴 결과이다. 규칙 1에 의하면 전체 대상자 중에 3회 중에 1회만 과음 한 비율은 28.308%였고 1회 사업연도 동안 과음을 하고 고혈압이 발생한 비율은 46.373%였으며 비과음자보다 고혈압이 발생할 확률은 10.37%(= (lift - 1) x 100)높았다. 규칙 2에 의하면 전체 대상자 중에 3개년 에 걸친 사업연도 중에 2회 동안 과음을 하고 2회 중 1회 비만인 비율은 2.850%였고 고혈압이 발생한 비율은 57.744%였으며 과음겸험이 없고 체중이 정상인 대상자보다 고혈압이 발생할 확률은 29.1% 더 높았다. 규칙 1과 규칙 2를 비교해 보면 1회 사업연도만 과음한 경우보다 3회 사업연도 중에 2회 동안 과음을 하고 2회 중 1회는 비만인 경우가 고혈압 발생 확률이 25.4% 더 높음을 알 수 있다. 또한 규칙 6과 규칙 9를 비교해 보면 고혈압 가족력이 있는 남자의 경우 3회 사업연도 동안 과음을 지속적으로 하고 3회 중 1회 비만인 경우가 고혈압 가족력이 없이 3회 사업연도 동안 과음을 지속적으로 하고 3회 중 1회 비만인 경우보다 고혈압 발생 확률이 2.8% 더 높음을 알

수 있다.

표 10의 규칙 1,2,3은 여자를 대상자로 하였고 규칙 4,5,6은 여자 대상자 중에 가족력이 있는 군이고 7,8,9는 가족력이 없는 군의 순차패턴 결과이다. 규칙 1에 의하면 전체 대상자 중에 3회에 걸친 사업연도 중에 1회 사업연도 동안 지속적으로 운동하지 않고 비만인 비율은 3.096%였고 1회 사업연도 동안 지속적으로 운동하지 않고 비만인 대상자가 고혈압이 발생한 비율은 35.892%였으며 지속적으로 운동하고 체중이 정상인 대상자보다 고혈압이 발생할 확률은 32.5%(= (lift - 1) x 100)높았다. 규칙 2에 의하면 전체 대상자 중에 3회 사업연도 중에 2회 동안 지속적으로 운동하지 않고 2회 중 1회 비만인 비율은 2.705%였고 고혈압이 발생한 비율은 36.8264%였으며 지속적으로 운동하고 체중이 정상인 대상자보다 고혈압이 발생할 확률은 36.0% 더 높았다. 규칙 1과 규칙 2를 비교해 보면 1회 사업연도 동안 지속적으로 운동하지 않고 비만인 경우보다 3회 사업연도 중에 2회 동안 지속적으로 운동하지 않고 2회 중 1회은 비만인 경우가 고혈압 발생 확률이 4.5% 더 높음을 알 수 있다. 또한 규칙 6과 규칙 9를 비교해 보면 고혈압 가족력이 있는 여자의 경우 3회 사업연도 동안 운동을 지속적으로 하지 않고 3회 중 1위험요인은 비만인 경우가 고혈압 가족력이 없이 3회 사업연도 동안 운동을 지속적 하지 않고 3회 중 1회 비만인 대상자 보다 고혈압 발생 확률이 1.6% 더 높음을 알 수 있다.

표 11의 규칙 1에 의하면 전체 대상자 중에 3회 사업연도 중에 1회만 과음을 한 비율은 15.312%였고 1회 사업연도 동안 과음을 하고 고혈압이 발생한 비율은 39.111%였으며 과음경험이 없는 대상자보다 고혈압이 발생할 확률은 14.2%(= (lift - 1) x 100)높았다. 규칙 2에 의하면 전체 대상자

중에 3회 사업연도 중에 2회동안 과음을 한 비율은 12.774%였고 2회 사업연도 동안 과음을 하고 고혈압이 발생한 비율은 41.131%였으며 비과음 대상자보다 고혈압이 발생할 확률은 20%높았다. 규칙 1과 규칙 2를 비교해 보면 1회 사업연도만 과음한 경우보다 2회 연속 과음한 경우가 5.8%정도 고혈압 발생 확률이 더 높음을 알 수 있다. 또한 규칙 5에 의하면 3개년 사업연도 동안 과음과 흡연을 지속할 경우에 고혈압 발생 확률은 35.5%였으며 규칙1과 규칙 5를 비교해 보면 1회 사업연도만 음주한 경우보다 음주와 흡연을 지속적으로 할 경우에 21%정도 고혈압 발생 확률이 더 높음을 알 수 있다.

표 9. 순차패턴 결과 (남자 가족력 유무)

구분	규칙번호	지지도	신뢰도	향상도	순차패턴
남자 전체	1	28.308	46.373	1.037	(Drinking , Hypertension)
	2	2.850	57.744	1.291	(Drinking , Drinking & Obesity, Hypertension)
	3	2.096	58.038	1.298	(Drinking ,Drinking , Drinking & Obesity, Hypertension)
가족력 유	4	29.756	48.825	1.040	(Drinking , Hypertension)
	5	3.288	61.947	1.320	(Drinking , Drinking & Obesity, Hypertension)
	6	2.419	62.048	1.322	(Drinking ,Drinking , Drinking & Obesity, Hypertension)
가족력 무	7	28.123	46.060	1.037	(Drinking , Hypertension)
	8	2.795	57.160	1.287	(Drinking , Drinking & Obesity, Hypertension)
	9	2.055	57.479	1.294	(Drinking ,Drinking , Drinking & Obesity, Hypertension)

(Drinking:알코올 섭취 심각도 중이상 Obesity: 8.5kg/m^2이상)

표 10. 순차패턴 결과 (여자 가족력 유무)

가족력	규칙번호	지지도	신뢰도	향상도	순차패턴
여자 전체	1	3.096	35.892	1.325	(Exercise & Obesity , Hypertension)
	2	2.705	36.826	1.360	(Exercise , Exercise & Obesity , Hypertension)
	3	2.131	37.726	1.393	(Exercise ,Exercise , Exercise & Obesity , Hypertension)
가족력 유	4	3.676	39.038	1.282	(Exercise & Obesity , Hypertension)
	5	3.275	40.181	1.320	(Exercise , Exercise & Obesity , Hypertension)
	6	2.717	42.644	1.401	(Exercise ,Exercise , Exercise & Obesity , Hypertension)
가족력 무	7	2.974	35.159	1.333	(Exercise & Obesity , Hypertension)
	8	2.586	36.028	1.366	(Exercise , Exercise & Obesity , Hypertension)
	9	2.008	36.531	1.385	(Exercise ,Exercise , Exercise & Obesity , Hypertension)

(Exercise: 주간 3-4회 이하로 운동시간이 30분미만)

표 11. 순차패턴 결과 (전체 대상자)

규칙번호	지지도	신뢰도	향상도	순차패턴
1	15.312	39.111	1.142	(Drinking , Hypertension)
2	12.774	41.131	1.200	(Drinking , Drinking, Hypertension)
3	10.312	44.169	1.289	(Smoking, Smoking, Hypertension)
4	6.691	46.380	1.354	(Drinking & Smoking, Drinking & Smoking, Hypertension)
5	5.200	46.426	1.355	(Drinking & Smoking, Drinking & Smoking, Drinking & Smoking, Hypertension)
6	5.008	46.456	1.356	(Drinking & Smoking, Drinking & Smoking, Drinking & Smoking & Exercise, Hypertension)

5. 고혈압 위험요인 예측모형의 타당성 검증

표 12에서 의사결정나무와 로지스틱 회귀분석의 민감도(Sensitivity),특이도(Specificity), 정확도(Accuracy)를 나타내었다. 의사결정나무는 로지스틱 회귀분석보다 낮은 정확도와 감수성를 나타내었고 특이도는 더 높았다. 또한 의사결정나무 오분류률 또한 높았다(표 13).

표 12 . 의사결정나무와 로지스틱 회귀분석의 비교

Method	Sensitivity(%)	Specificity (%)	Accuracy (%)
Decision tree	38.38	74.35	75.17
Logistic regression	39.68	73.97	76.54

표 13 . 의사결정나무와 로지스틱 회귀분석의 오분류률 비교

Method	Root ASE	Misclassification
Decision tree	0.4673	0.3410
Logistic regression	0.4673	0.3388

(Root ASE: Root Average Squared Error)

6. 순차패턴을 이용한 고혈압관리 조언 시스템 모형 개발

고혈압관리 데이터마트로부터 고혈압관리를 위한 주요 패턴을 추출하여 고혈압 발병 위험요인을 순차패턴 결과를 이용하여 분석한 후 주요 패턴을 도출하며, 개별 환자에게 적절한 조언을 해줄 수 있는 시스템을 개발하였다. 고혈압관리 조언시스템은 건강검진결과를 제공하는 건강검진결과 Review, 고혈압 고위험군 예측 정보 및 진단, 생활습관의 개선을 위한 건강증진교육정보 제공으로 시스템을 구성하였다. 건강검진결과 Review 화면은 검진자의 건강검진결과를 모두 보여주고, 고혈압 고위험군 예측 정보 및 진단 화면은 순차패턴 기법을 활용하여 검진자의 고혈압 발생확률을 알려주고, 생활습관의 개선을 위한 건강증진교육정보 제공 화면은 검진자가 개선해야 하는 습관에 대한 정보를 제공해 주는 화면이다.

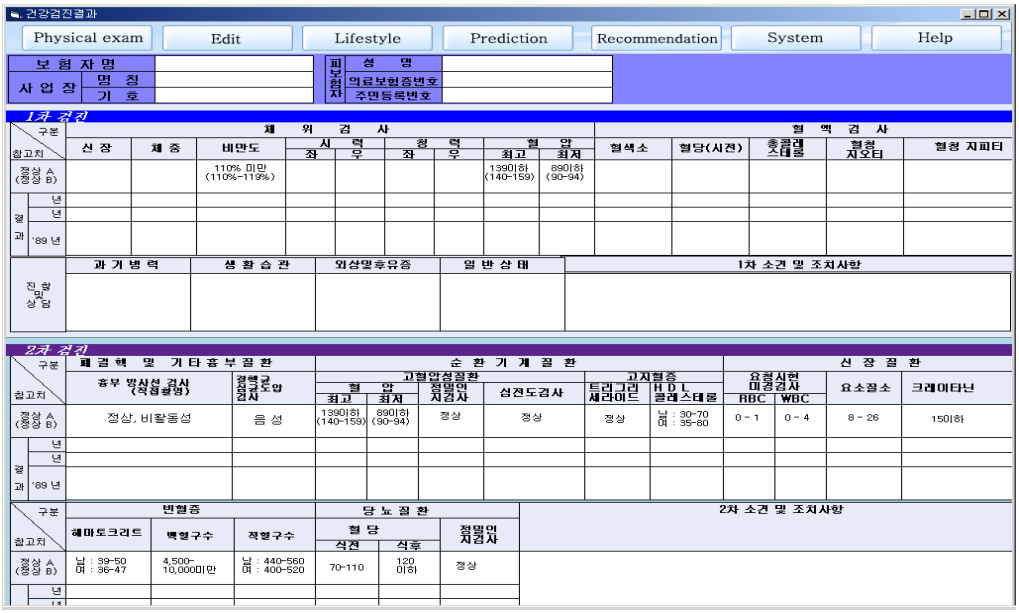

그림 5 . 건강검진결과 Review(Physical exam)

그림 6 . 고혈압 고위험군 예측 정보 및 진단(Prediction)

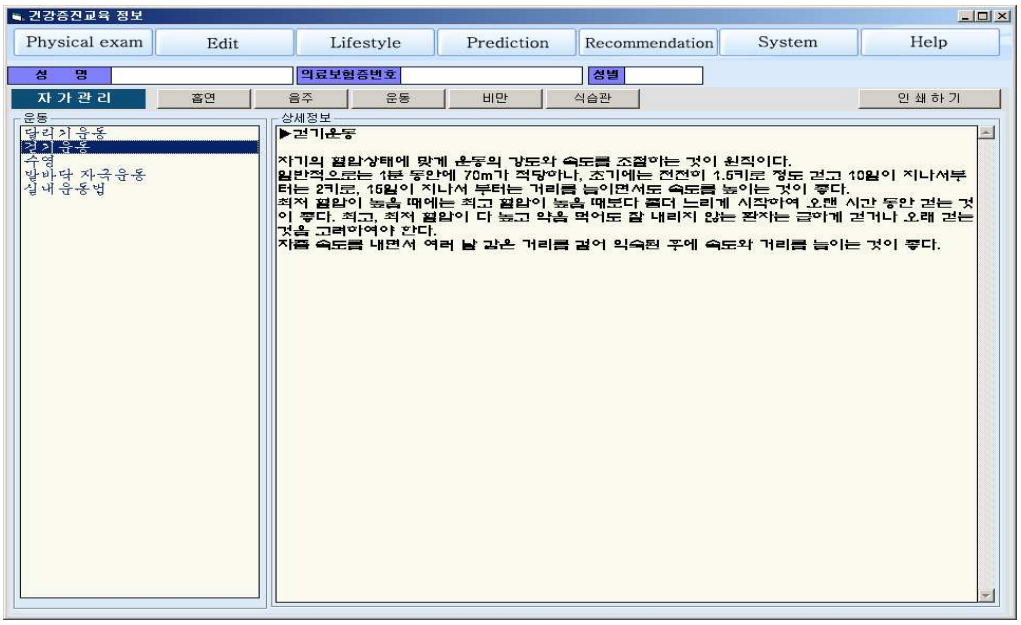

그림 7 . 생활습관의 개선을 위한 건강증진교육 정보(Recommendation)

Ⅴ. 고찰

1. 연구자료 및 방법에 대한 고찰

본 연구는 고혈압 발생의 위험요인 파악과 고혈압 환자의 예측을 위하여 정기적인 집단 건강검진을 실시하고 있는 국민건강보험공단의 피보험자 건강진단 결과를 이용하였다. 이 연구의 제한점을 구체적으로 보면 다음과 같다.

첫째, 연구대상 집단이 공무원 및 사립학교 교직원들로 제한되어 있어 전체의 고혈압 환자를 대표할 수 없다는 제한점은 있으나 1994년-2000년까지 계속 관찰하였으며 건강검진결과, 문진표 등의 다양한 자료를 구체적으로 파악할 수 있는 장점을 가지고 있다.

둘째, 사용된 변수 중 건강검진 결과는 한 병원에서 실시된 것이 아니고 전국 여러 의료기관에서 수집된 자료이다. 혈압은 측정자의 방법, 혈압계의 종류 등에 따랄 결과의 차이가 있으므로 이 자료도 타당성에 문제를 제기할 수 있다. 특히 본 연구에서 사용된 혈압자료는 처음부터 연구를 위하여 측정된 자료가 아니므로 집단검진에서 나타날 수 있는 끝수의 선호도(digital preference)와 같은 측정오차가 개입되었을 가능성이 높다.

셋째, 연구대상자의 일반적 특성 중 연령을 2000년 기준으로 보았으므로 위험요인이 존재했을 당시의 연령과 기준연령에는 차이가 있다.

넷째, 연구 자료 기간이 1994년-2000년 까지 격년으로 4개년을 대상으로 하였는데 그 중 1994년에는 문진 자료가 없어서 1996, 1998, 2000년 데

이터를 기반으로 2회 사업년도 이상 같은 문진 결과가 나온 값을 1994년 문진결과 값으로 대체하여 분석결과의 신뢰성에 문제가 있을 수 있다.

다섯째, 과거력은 고혈압과 강한 연관성을 지니고 있어 본 연구결과에 영향을 미칠 수 있다. 그러나 과거력 변수가 과거 검사 시점의 진단과 중복될 수 있어서 판단기준이 모호하여 이를 제외함으로써 해석시 이에 대한 고려가 필요하다.

여섯째, 순차패턴의 결과로 위험요인의 발생 년도까지 추정할 수는 없어서 위험요인이 연속적으로 있는 경우와 비연속적으로 있는 경우와의 고혈압 발생 확률의 차이를 알 수 없었다.

2. 연구결과에 대한 고찰

고혈압은 사망률이 높은 순환기계질환의 주요 위험요인으로서 중요한 의미를 가지고 있어 최근 우리나라에서도 보건소의 주요사업으로 고혈압 관리사업을 실시하고 있다. 하지만 표준화된 고혈압관리지침과 국가수준의 고혈압관련 통계자료의 부재로 환자 중심의 치료위주사업으로 환자조기발견과 위험요인 예방에 대한 홍보 및 효율적인 보건교육사업은 대단히 저조하다. 이 문제의 해결방안으로 고혈압 환자 각 개인을 예방하는데 있어 복합적 정보를 제공할 수 있는 지침 및 연구를 해야 하는 시점이 된 것이다.

이에 따라 본 연구에서는 데이터마이닝 기법 중 순차패턴 탐색을 이용하여 시계열 분석을 함으로써 기존의 조사나 연구에서 발견할 수 없었던 의미 있는 새로운 상관관계, 패턴, 추세를 밝혀내고 이와 같은 유용한 정보를 집적하고 체계화하여 고혈압의 예방을 위한 고혈압관리 조언시스템을

개발하였다.

대부분의 선행연구들은 고혈압 위험요인을 환자-대조군 연구, 코호트 연구방법을 활용하여 수행되어져 왔다.

감신 등(1991)이 1988년에 대구 근로자중 330명의 고혈압 환자군과 1,336명의 대조군을 대상으로 고혈압의 위험요인(직업,비만,음주,흡연,우유섭취,염분섭취,가족력)에 대한 환자-대조군 연구를 하였다. 그 결과 남자에서는 음주(OR=3.23), 비만(OR=2.31), 염분섭취(OR=1.75)가, 여자에서는 음주(OR=16.49), 가족력 (OR=3.70), 비만(OR=1.74), 염분섭취(OR=1.73)가 고혈압 발생과 유의한 관련이 있는 위험요인이었다. 손행미 등(1999)이 연구한 '본태성 고혈압 환자의 고혈압 발생 위험요인예측'은 대학병원의 110명 환자를 대상으로 하여 다중회귀분석 방법으로 분석한 결과 비만과 과거력 기간이 고혈압 발생과 유의한 관련이 있는 위험요인이었다. 오희숙 등(2000)은 농촌지역 성인들에서 고혈압 발생과 관련이 있는 위험요인을 조사하기 위하여 1999년 20세이상 고혈압이 없는 1,781명을 대상으로 추적조사 하였다. 다중회귀분석 결과 고혈압 발생과 관련된 위험요인으로는 여자에서는 연령, 높은 정상혈압 그리고 폐경이 고혈압 발생 위험요인이였고 남자에서는 연령, 고혈압 가족력 그리고 허리-엉덩이 둘레비 그리고 높은 정상혈압이었다.

본 연구는 고혈압 예방측면에 중점을 두어 고혈압 위험요인변수를 생활습관변수로 제한하였고 위험요인의 기간이 증가함에 따른 고혈압 발생과의 관련을 로지스틱 회귀분석으로 알아본 결과 남자의 경우 연령이 증가할수록, 비만기간이 길수록, 과음기간이 길수록 고혈압에 걸릴 확률이 높았으며 통계적으로 유의하였다. 여자의 경우는 연령이 증가할수록, 비만기간이

길수록 고혈압에 걸릴 확률이 높았으며 통계적으로 유의하였다.

본 연구는 기존의 통계적 방법 외에 데이터 마이닝 기법을 활용하였고 통계적 방법인 로지스틱 회귀분석과 데이터 마이닝 방법인 의사결정나무 분석의 성능을 비교하였다. 그 결과, 의사결정나무는 로지스틱 회귀분석보다 낮은 정확도와 감수성를 나타내었고 특이도는 더 높았다. 하지만, 채영문 등 (2000)의 연구에서는 고혈압 환자 예측을 위해 국민건강보험공단의 1998년 건강검진 데이터베이스에서 생의학 변수, 생활습관 변수, 인구사회학적 변수로 로지스틱 회귀분석, 의사결정나무 분석 알고리즘의 C4.5와 CHAID의 Performance를 비교하였다. 이 결과 CHAID 알고리즘이 64.04%로 가장 예측력이 높았고 다음은 로지스틱 회귀분석이 63.84%였고 C5.0은 59.22%였다. 민감도 또한 CHAID 알고리즘이 76.30%로 가장 높았고 다음은 로지스틱 회귀분석이 64.36%였고 C5.0은 59.34%였다. 그러나 특이성은 CHAID 알고리즘이 52.3%로 가장 낮았고 로지스틱 회귀분석이 63.33%로 가장 높았다.

또한 채영문 등 (2000)의 연구에서는 연관규칙 기법을 활용하여 다수의 위험 요인과 고혈압과의 연관성을 제시하였다. 본 연구는 더 나아가 순차패턴 기법을 활용하여 시간의 경과에 따라 고혈압이 발생하는 다수의 위험요소를 찾아 고혈압 발생 위험도를 제시하였다. 순차패턴의 결과에 따르면 1회의 사업 연도 동안 과음한 경우보다 2회의 사업 연도동안 과음과 흡연을 하고 3회의 사업 연도에는 과음, 흡연, 비지속적인 운동을 한 경우에 약 21%정도 고혈압 발생 확률이 더 높음을 알 수 있었다. 또한 남자, 여자 모두 가족력이 있는 경우가 가족력이 없는 경우보다 고혈압 발생 확률이 높음을 알 수 있었다.

데이터 마이닝 기법을 활용하여 의사결정지원시스템을 개발한 예를 들면 호승희 등(2000)은 고혈압 데이터마트를 구축한 후 데이터마이닝 기법 중 의사결정나무를 적용하여 치료결과에 어떠한 요인이 얼마나 중요한 지를 유추해 내고 치료후 혈압조절군과 비조절군의 특성을 분류, 규명함으로써 새로운 치료 전략을 도출했으며, 이를 이용한 의사결정지원시스템을 개발하였다. 본 연구는 고혈압관리 데이터마트로부터 데이터 마이닝의 순차패턴 기법을 적용하여 순차 패턴 DB를 구축하여 고혈압 위험집단에게 적절한 조언을 해 줄수 있는 시스템을 개발하였다.

본 연구의 결과는 고혈압 환자 각 개인을 예방하는데 있어 복합적 정보를 제공할수 있을 것으로 판단되며, 의사결정지원시스템의 개발을 통해 실제 임상에 적용해 봄으로써 그 효과와 실증적 가치를 창출할 수 있을 것이다.

Ⅵ. 결론

본 연구는 과거의 축적된 환자 데이터를 이용하여 고혈압 위험요인을 규명하기 위해 데이터 마이닝 기법 중 순차패턴 탐색을 이용하여 시계열 분석을 함으로써 기존의 조사나 연구에서 발견할 수 없었던 의미 있는 새로운 상관관계, 패턴, 추세를 밝혀내고 이와 같은 유용한 정보를 집적하고 체계화하여 고혈압 관리를 위한 의사결정지원시스템의 개발했다는데 그 의의가 있다고 할 수 있다. 연구결과는 다음과 같다.

첫째, 로지스틱 회귀분석으로 고혈압 위험인자를 알아본 결과 남자의 경우 연령이 증가할수록, 비만기간이 길수록, 과음기간이 길수록 고혈압에 걸릴 확률이 높았으며 통계적으로 유의하였다. 반면에 ,흡연기간은 짧을수록 고혈압에 걸릴 확률이 높았으나 이는 정부의 금연교육 효과가 아닐까 생각된다. 여자의 경우는 연령이 증가할수록, 비만기간이 길수록 고혈압에 걸릴 확률이 높았으며 통계적으로 유의하였다. 둘째, 의사결정나무 분석을 이용하여 고혈압에 미치는 위험요인을 설명하는 주요 변수로는 성별 변수를 제외한 경우는 흡연, 과음, 비만이였고, 성별 변수를 포함한 경우에 남자들의 고혈압 위험요인은 비만, 과음 변수였고 여자들은 나이, 비만 변수였다. 셋째, 순차패턴을 이용하여 시간의 경과에 따라 고혈압에 영향을 미치는 각 위험요인의 모든 조합에 대해 영향도를 계산하여 다수의 위험요소가 나타나는 패턴을 찾아내었다. 이 결과에서 시간이 고혈압 결과에 영향을 미치는 중요한 요인임을 알 수 있었다. 순차패턴의 결과에 따르면 1회의 사업 연도 동안 과음한 경우보다 2회의 사업 연도동안 과음과 흡연을

하고 3회의 사업 연도에는 과음, 흡연, 비지속적인 운동을 한 경우에 약 21%정도 고혈압 발생 확률이 더 높음을 알 수 있었다. 또한 남자의 경우 고혈압 위험요인은 주로 과음과 비만으로 나타났으며 가족력이 있는 경우가 가족력이 없는 경우보다 과음하고 비만할 경우 고혈압 발생 확률이 약 3%정도 높았다. 여자의 경우는 고혈압 위험요인은 주로 비지속적인 운동과 비만으로 나타났으며 가족력이 있는 경우가 가족력이 없는 경우보다 비지속적으로 운동하고 비만할 경우 고혈압 발생 확률이 약 2-4%정도 높았다. 넷째, 고혈압관리 데이터마트로부터 데이터 마이닝의 순차 패턴 기법을 적용하여 순차 패턴 DB를 구축하였고 이 과정에서 대상 데이터를 선별하고, 가공하여 순차패턴을 추출할 수 있는 형태로 처리하여 고혈압 위험집단에게 적절한 조언을 해 줄수 있는 시스템을 개발하였다.

본 연구의 결과를 통해 고혈압관리 조언 시스템을 확대 적용하여 고혈압 환자의 조기발견과 더불어 일차예방을 시도한다면 환자 발생을 줄이는데 큰 도움을 줄 수 있을 것이다.

향후 연구에서는 고혈압 외에 만성퇴행성 질환인 당뇨질환등의 관리를 위해 본 연구에서 활용한 순차패턴기법 등의 방법으로 만성질환을 예측할 수 있다. 또한 본 연구결과를 기반으로 하여 신경망과 같은 데이터 마이닝 기법을 적용해 볼 수 있으며 생활습관에서 의미있는 패턴들을 찾기 위해 기간을 증가하여 순차패턴을 함으로써 더 좋은 결과를 이끌어 낼수 있을 것이다.

참고문헌

감신, 예민혜, 이성국, 천병렬. 고혈압의 위험요인에 대한 환자-대조군 연구 예방의학회지 1991;24:221-231

김정순. 역학원론, 제1판,신광출판사, 서울,1994,125-138

김현창, 서일, 지선하, 이강희, 김창수, 남정모. 강화지역 성인남녀의 12년간 고혈압 발생률과 위험요인: 강화연구. 예방의학회지 1999;32(4): 435-442

박종구, 김기순, 김춘배 등. 의료보험청구자료중 뇌혈관질환 상병기호의 정확도에 관한 연구. 예방의학회지 2000;33(1):76- 82

지원철. 신경망을 이용한 시계열 분석 : M1-Competition Data에 대한 예측성과 분석. 한국전문가시스템학회지 1995.1

채영문, 호승희, 조경원 등. Data Mining for Knowledge Management in a health insurance Domain.한국지능정보시스템학회 논문지 2000 6(1)

Breault JL, Goodall CR, Fos PJ. Data mining a diabetic data warehouse. Artificial Intelligence in medicine 2002; 26(1-2): 37-54.

C.R.MacIntyre, M.A.Kainer, G.V.Brown. A randomised, clinical trial comparing the effectiveness of hospital and community based reminder systems for increasing uptake of influenza and pneumococcal vaccine in hospitalised patients aged 65 years and over. 2003; 49; 33-40.

Ellit Svide, et al. Relation of alchol and cigarette consumtion to blood pressure and serum cretinine levels. J Chrons Dis 1994;37(8):617-623.

Fayyad U, Piatetsky-Shapiro G and Smyth P, Advances in Knowledge Discovery and Data Minin. AAAI Press, 1996

Frame PS. Computerized health maintenance tracking systems: a clinician's guide to necessary and optional features. A report from the American Cancer Society Advisory Group on Preventive Health Care Reminder Systems. J Am Board Fam Pract. 1995 May-Jun; 8(3):221-9

Graham A, MacGregor. Sodium is more important than calcium in essential hypertenseion. Hypertension 1985;7:628-637

He J, Whelton PK, Appel LJ,Charleston J, Klag MJ.Long-term effects of weight loss and dietary sodium reduction on incidence of hypertension. Hypertension 2000;35(2):544-9

Hollenberg NK, et al. Essential hypertension,abnormal renal vescular and endocrin response to a mild psychological stimulus. Hypertension 1981;3:11-17

Hyman DJ, Pavlik VN. Charateristics of patients with uncontrolled hypertension in the United States. N Engl J Med 2001;345(7):479-86

Ho SH, Chae YM, Cho SY. et al. Development of decision support system for the management of hypertension using data mining

technology. Proceedings of the conference the Korean Society of Intelligent information System 2000:271-282

Izzo JL, Black HR; editors. Hypertension Primer. 2nd ed. Baltimore: Lippincott Williams & wilkins;1999

Jeffrey AC, Elliott P.Suh I. An overview of randomized trials of sodium reduction and blood pressure. Hypertension 1991;17(Supple I):I-27-I-33

Kornitzer M, Dramaix M, De Backer G. Epicemiology of risk factors for hypertension: implication and therapy.Drugs 1999:57(5);692-712

Larsen DL, Cannon W, Tower S. Diabetes care management system in an integrated health network. Journal of management care pharmacy 2003 Nov-Dec; 9(6): 552-8

Lin F, Chou S, Pan S, and Chen Y. Mining time dependency patterns in clinical pathways. Int J Med Informatics 2001: 62: 11-25

MaO Y, et al. Risk factors for hypertension: results from cross-sectional survey. J of Clin Epid 1988;41:411-415

Mishima N, Iwata N, Shimizu T. Creating notifications tailored for individual workers by using a database: a conceptual framework and its application. Ind Health 2000 Jul; 38(3):331-7

Muller JF, Franz IW.Significance of a life style change in arterial hypertension.Fortschritte der Medizin 1998:116(34);20-5

Ockene JK,Kuller LH, Svendsen KH,Meilahn E. The relationship of

smoking cessation to coronary heart disease and lung cancer in the Mutiole Risk Factor intervention Trial. Am J Public Health, 1990,80(8), 954-958

Reed D, McGee D, Yano K, Hankin J. Diet, blood pressure, and multicollinearity. Hypertension 1985,7,405-410

Verma R, Harper J. Life cycle of a data warehousing project in healthcare. Journal of healthcare information management 2001; 15(2): 107-17.

Yamamoto Y, Namikawa H, Inamura K. Development of a Time-Oriented Data Warehouse Based on Medical Information Event Model. Japanese Journal of Medical physics. 2002; 22(4): 327-33

http://www.aistudy.co.kr

<부록 1>

표 1. 고혈압 유무와 비만도 (남자) 단위 : 명(%)

년도	구분	고혈압군	비고혈압군	$X2$-값
1994	정상	15,876(94.63)	20,061(96.68)	96.06***
	비만	901(5.37)	688(3.32)	
	계	16,777(100)	20,749(100)	
1996	정상	15,526(92.52)	19,795(95.52)	137.49***
	비만	1,255(7.48)	956(4.61)	
	계	16,781(100)	20,715(100)	
1998	정상	15,226(90.73)	19,516(94.05)	148.57***
	비만	1,556(9.27)	1,234(5.95)	
	계	16,782(100)	20,750(100)	
2000	정상	14,612(87.10)	19,058(91.88)	229.76***
	비만	2,164(5.77)	1,684(4.49)	
	계	16,776(100)	20,742(100)	

표 2. 고혈압 유무와 비만도 (여자) 단위 : 명(%)

년도	구분	고혈압군	비고혈압군	$X2$-값
1994	정상	14,112(95.11)	38,710(96.91)	100.16***
	비만	725(4.89)	1,236(3.09)	
	계	14.837(100)	39.946(100)	
1996	정상	13,942(93.96)	38.432(96.18)	126,32***
	비만	897(6.04)	1526(3.82)	
	계	14,839(100)	39,958(100)	
1998	정상	13,568(91.44)	37,875(94.77)	208.20***
	비만	1,270(8.56)	2,089(5.23)	
	계	14,838(100)	39,964(100)	
2000	정상	13,225(89.15)	37,347(93.47)	284.58***
	비만	1,610(10.85)	2,607(6.53)	
	계	14,835(100)	39,954(100)	

표 3. 고혈압 유무와 과음여부(남자) 단위 : 명(%)

년도	구분	고혈압군	비고혈압군	$X2$-값
1994	비과음	9,517(56.71)	10,749(51.79)	89.99***
	과음	7,266(43.29)	10,005(48.21)	
	계	16,783(100)	20,754(100)	
1996	비과음	9,156(55.02)	10,518(51.09)	56.79***
	과음	7,486(44.98)	10,069(48.91)	
	계	16,642(100)	20,587(100)	
1998	비과음	9,304(55.86)	10,589(51.38)	74.36***
	과음	7,351(44.14)	10,022(48.62)	
	계	16,655(100)	20,611(100)	
2000	비과음	9,214(55.71)	10,356(53.87)	81.75***
	과음	6,481(41.29)	8,867(46.13)	
	계	15,695(44.95)	19,223(100)	

표 4. 고혈압 유무와 과음여부(여자) 단위 : 명(%)

년도	구분	고혈압군	비고혈압군	$X2$-값
1994	비과음	2,277(15.34)	33,839(84.65)	0.00
	과음	12,566(84.66)	33,839(84.65)	
	계	14,843(100)	39,974(100)	
1996	비과음	2,289(15.85)	6,536(16.72)	5.75**
	과음	12,152(84.15)	12,152(84.15)	
	계	14,441(100)	6,536(16.72)	
1998	비과음	2,614(17.95)	7,006(17.82)	0.11
	과음	11,946(82.05)	32,299(82.18)	
	계	14,560(100)	7,006(17.82)	
2000	비과음	2,782(20.24	7,221(19.60)	2.55
	과음	10,962(79.76)	29,620(80.40)	
	계	13,744(100)	36,841(100)	

표 5. 고혈압 유무와 흡연정도(남자) 단위 : 명(%)

년도	구분	고혈압군	비고혈압군	$X2$-값
1994	비흡연	7,274(43.34)	8,736(42.09)	5.86**
	흡연	9,509(56.66)	12,018(57.91)	
	계	16,783(100)	20,754(100)	
1996	비흡연	6,793(41.40)	8,336(40.99)	0.62
	흡연	9,616(58.60)	12,003(59.01)	
	계	16,409(100)	20,339(100)	
1998	비흡연	7,133(43.40)	8,517(42.12)	6.09**
	흡연	9,301(56.60)	11,779(57.88)	
	계	16,434(100)	20,350(100)	
2000	비흡연	7,508(48.26)	8,889(46.62)	9.12***
	흡연	8,051(51.74)	10,178(53.38)	
	계	15,559(100)	19,067(100)	

표 6. 고혈압 유무와 흡연정도(여자) 단위 : 명(%)

년도	구분	고혈압군	비고혈압군	$X2$-값
1994	비흡연	14,829(99.91)	39,950(99.94)	1.37
	흡연	14(0.09)	24(0.06)	
	계	14,843(100)	54,817(100)	
1996	비흡연	11,344(99.54)	31,148(99.67)	3.56*
	흡연	52(0.46)	102(0.33)	
	계	11,396(100)	31,250(100)	
1998	비흡연	12,134(99.57)	33,083(99.65)	1.13
	흡연	52(0.43)	117(0.35)	
	계	12,186(100)	33,200(100)	
2000	비흡연	12,186(100)	33,200(100)	7.46**
	흡연	46(0.36)	73(0.16)	
	계	12,669(100)	34,050(100)	

표 7. 고혈압 유무와 지속적인 운동(남자) 단위 : 명(%)

년도	구분	고혈압군	비고혈압군	$X2$-값
1994	지속적인운동	611(3.64)	704(3.39)	1.62
	비지속적인운동	16,172(96.36)	20,050(96.61)	
	계	16,783(100)	20,754(100)	
1996	지속적인운동	857(5.16)	900(4.38)	12.20***
	비지속적인운동	15,752(94.84)	900(4.38)	
	계	16,609(100)	19,645(95.62)	
1998	지속적인운동	684(4.12)	801(3.90)	1.11
	비지속적인운동	15,905(95.88)	19,727(96.10)	
	계	16,589(100)	20,528(100)	
2000	지속적인운동	806(5.15)	951(4.97)	0.60
	비지속적인운동	14,831(94.85)	18,203(95.03)	
	계	15,637(100)	19,154(100)	

표 8. 고혈압 유무와 지속적인 운동(여자) 단위 : 명(%)

년도	구분	고혈압군	비고혈압군	$X2$-값
1994	지속적인운동	259(1.74)	608(1.52)	3.34*
	비지속적인운동	14,584(98.26)	39,366(98.48)	
	계	14,843(100)	39,974(100)	
1996	지속적인운동	266(1.84)	645(1.65)	2.09
	비지속적인운동	14,215(98.16)	38,440(98.35)	
	계	14,481(100)	39,085(72.97)	
1998	지속적인운동	217(1.49)	599(1.42)	0.29
	비지속적인운동	14,327(98.51)	38,689(98.58)	
	계	14,544(100)	39,248(100)	
2000	지속적인운동	291(2.14)	853(2.33)	1.54
	비지속적인운동	13,326(97.86)	35,783(97.67)	
	계	13,617(100)	36,636(100)	

표 9. 고혈압 유무와 염분섭취((남자) 단위 : 명(%)

년도	구분	고혈압군	비고혈압군	$X2$-값
1994	저염식	14,088(83.94)	17,533(84.48)	1.98
	고염식	2,695(7.18)	3,221(8.58)	
	계	16,783(100)	20,754(100)	
1996	저염식	13,649(81.97)	17,063(82.81)	4.46**
	고염식	3,002(18.03)	3,541(17.19)	
	계	16,651(100)	20,604(100)	
1998	저염식	13,763(82.63)	17,178(83.30)	2.91*
	고염식	2,894(17.37	3,444(16.70)	
	계	16,657(100)	20,622(100)	
2000	저염식	12,972(82.59)	15,990(83.10)	1.56
	고염식	2,735(17.41)	3,252(16.90)	
	계	15,707(100)	19,242(100)	

표 10. 고혈압 유무와 염분섭취(여자) 단위 : 명(%)

년도	구분	고혈압군	비고혈압군	$X2$-값
1994	저염식	12,975(87.41)	35,051(87.59)	0.30
	고염식	1,868(12.59)	4,959(12.41)	
	계	14,843(100)	39,974(100)	
1996	저염식	12,629(85.68)	34,221(86.16)	2.03
	고염식	2,111(14.32)	5,497(13.94)	
	계	14,740(100)	39,718(100)	
1998	저염식	12,760(86.46)	34,309(86.29)	0.24
	고염식	1,999(13.54)	5,497(13.94)	
	계	14,759(100)	39,762(100)	
2000	저염식	11,934(86.27)	32,022(62.95)	0.34
	고염식	1,900(13.73)	5,010(13.53)	
	계	13,834(100)	37,032(100)	

Abstract

Development of Decision Support System Model for Hypertension Management Using Sequential Patterns

Hwa Young Kim
Graduate School of
Public Health
Yonsei University

(Directed by professor Young Moon Chae, Ph.D.)

Data Mart was established in order to develop reminder system for managing hypertension risk group. Hypertension management reminder system was developed using hypertension risk factor prediction prototype presented by applying data mining. Consequently, 92,354 people who had no symptoms of hypertension from 1994- to 1998 who either had or did not have hypertension in 2000 were selected as researchsubjects out of a 4 year period medical record database of the National Health Insurance Corporation. The

current research determined patients at risk of hypertension through logistic regression analysis and decision tree analysis and compared the effectiveness of the two methods. As a result of determining patients at risk of hypertension through logistic regression analysis, males who were older, had a longer period of obesity, had a longer period of binge drinking, and had smoked for a longer period of time had a more statistically significant and higher probability of suffering from hypertension. Main factors explaining risk factors that affect hypertension through decision making tree analysis were smoking, drinking, obesity when gender variable was excluded, and when gender was taken into account, obesity and drinking were risk factors for hypertension in men, and age and obesity for women. Decision tree analysis had lower accuracy (75.17%) and sensitivity (73.38%) than logistic regression analysis while having higher particularity (74.35%). Also, the current research has found a pattern where many risk factors appear by using sequential patterns and calculating the degree of influence on all combinations of each risk factor that affect hypertension with the passage of time. As a result, a sequential pattern able to predict potential hypertension patient with risk factors was proven.

In this manner, reminder system for hypertension management was developed using rules deduced from data mining. The system organization is made up of Physical exam Review, Prediction, and Recommendation.

It is the belief of the current research that its findings can greatly contribute in reducing the occurrences of hypertension patients if efforts for their early detection and primary preventive measures are taken by expanding and applying reminder system for hypertension managements.