데이터 마이닝을 이용한 대장암 환자의 진료비 예측 모형

이승미[a], 강진오[b], 서용무[c]

[a] 고려대학교 대학원 경영학과
136-701, 서울특별시 성북구 안암동 5가 1
Tel: +82-2-3290-1945, E-mail:seungmi@korea.ac.kr

[b] 경희대학교 의과대학 의학과
130-701, 서울특별시 동대문구 회기동 1
Tel: +82-2-958-8664, E-mail:kangjino@khmc.or.kr

[c] 고려대학교 경영대학 경영학과
136-701, 서울특별시 성북구 안암동5가 1
Tel: +82-2-3290-1945, E-mail:ymsuh@korea.ac.kr

Abstract

병원에서 환자의 진료비 예측 및 분석은 한정된 의료 자원을 효율적으로 분배하고 의료 정책을 수립할 때 기본 근거가 되는 매우 중요한 자료이다. 그러나 환자의 진료비 예측 및 분석과 관련한 대부분의 기존 연구들은 통계적 분석 기법을 바탕으로 하였기 때문에 다양한 형식의 데이터로 구성된 의료 데이터의 극히 일부분만을 사용할 수 밖에 없었다. 데이터 마이닝은 통계적 분석 기법을 적용하기 어려운 대량의 의료 데이터를 효율적으로 분석하여 숨겨진 지식을 발견하는데 있어서 적합한 방법이다. 본 연구에서는 대장암 환자의 진료비 예측 모형을 생성하기 위하여 데이터 마이닝 기법인 신경망과 의사 결정 나무 기법를 이용하였으며, 기존 연구에서 주로 사용한 통계적 분석 방식인 회귀 분석 모델과 비교 분석하였다. 그 결과, 기존의 통계적 분석 방식에 비하여 데이터 마이닝을 이용한 방식이 대장암 환자의 진료비를 예측하는 데에 있어서 더 좋은 성과를 얻을 수 있음을 밝혔다.

Keywords:

진료비 예측, 의료 데이터, 신경망, 의사결정나무, 회귀분석

1. 서론

정보 기술의 발달로 환자의 상태와 그 변화를 기록하기 위한 장비들과 다양한 검사 장비들이 등장하였으며, 이로 인한 환자 기록의 증가와 멀티미디어 기술의 발달로 인하여 각 주제 별로 대량의 의료데이터를 저장하기 위한 데이터베이스의 개발이 중요해졌다. 또한 이로부터 의료 연구 및 병원 운영을 위한 중요한 정보들을 추출하기 위한 의료 정보시스템이 점차 부각되고 있다.

데이터 마이닝은 통계적인 분석 기법을 적용하기 어려운, 축적된 대량의 의료 데이터를 효율적으로 분석하여 숨겨진 지식을 발견하는데 매우 적합한 방법이다[1]. 그러나 대부분의 의료 데이터 마이닝과 관련한 연구들은 주로 환자의 질병 진단이나 예후를 예측하는 데에 초점을 두고 있으며, 환자의 진료비 분석 및 예측에는 널리 활용되지 못하는 실정이다. 환자의 진료비 산정은 환자에 대한 진료 정보를 바탕으로 하며, 이러한 정보들은 의료 정보시스템이 등장한 이후 꾸준히 축적되고 있다. 이러한 의료 정보들을 바탕으로 데이터 마이닝을 이용하여 환자의 진료비 예측 모형을 생성하면 한정된 의료 자원을 효율적이고 효과적으로 분배할 수 있고, 환자의 진료비를 바탕으로 불필요한 치료나 검사가 있었는지를 추적할 수 있다. 또한 환자에게도 미리 진료비에 대한 계획을 수립할 기회를 부여할 수 있으며, 의료 정책을 수립할 때 기본 근거 자료로 이용할 수 있다 [2, 3].

본 연구는 K대학병원의 1999년부터 2002년까지의 입원기록이 있는 대장암 환자에 대한 검사 및 진단, 치료 정보를 바탕으로 새로운 대장암 환자의 진료비를 예측하는 데 활용할 수 있는 모형을 만드는 것을 목적으로 하며, 신경망 기법과 의사 결정 나무 기법을 이용하여 예측 모형을 구축하였다. 본 연구는 이러한 데이터 마이닝 기법들을 이용하여 기존의 통계적 방법 기반의 예측모형보다 더욱 정확한 진료비 예측이 가능한 모형을 생성하였다.

본 논문의 구성은 다음과 같다. 2장에서는 의료 데이터 마이닝을 위하여 고려해야 할 의료 데이터의 특성과 기존의 환자의 진료비와 관련한 연구들을 살펴보고, 3장에서는 대장암 환자의 진료비를 예측하기 위하여 회귀 분석, 신경망 그리고 의사 결정 나무 기법을 이용한 모형 설계에 대하여 설명하였다. 4장에서는 3장의 각 기법에 따른 모형

생성의 결과들을 비교하여 기술하고, 5장에서 연구의 결론 및 향후 연구 과제에 대하여 논하였다.

2. 의료 데이터 마이닝

2.1. 의료 데이터 마이닝의 특성

데이터 마이닝을 수행할 때, 인간의 의료 데이터는 모든 생물학적 데이터 중에서 가장 가치 있고 분석하기 어려운 분야이다[6]. 의료 데이터는 데이터가 수집되는 원천과 방식이 다양하고, 한 명의 환자에 대한 기록이 장기간에 걸쳐 축적되는 경향이 있다. 따라서 의료 데이터 마이닝을 수행하고자 할 때에는 이러한 의료 데이터의 특성을 충분히 고려한 후에 수행해야 한다. 의료 데이터를 분석하는데 있어서 이에 대한 특성 및 주의점에 관한 많은 연구들이 있었는데[6, 7, 8, 9], 특히 Cios와 Moore의 연구는 의료 데이터 마이닝을 수행할 때 주의해야 할 의료 데이터의 특성에 대하여 논하고 있다[6]. 그들은 의료 데이터의 특성을 크게 의료 데이터의 이질성, 윤리적·법적·사회적인 문제점들, 통계적 관점, 의학의 특별한 지위 등의 측면에서 기술하였다. Tsumoto는 의료 데이터의 특성으로 시계열 측면에서의 비동질성, 1년 이상 입원하는 환자의 경우 1000개 이상의 필드 발생 등을 언급하였다[8]. Groselj는 의료 데이터 마이닝을 수행할 때, 환자와 관련된 많은 문서 정보들을 전산화하는 작업의 어려움과 전산화 하는 과정에서 주의할 점에 대하여 언급하였다[7]. Berman은 의료 데이터를 대상으로 데이터 마이닝을 수행하는 연구자들에게 의료 데이터의 특성상 환자에 대한 기밀성의 유지가 매우 중요함을 강조하였다[9]. 이러한 여러 연구에서 언급하였던 의료 데이터의 특성들을 데이터를 실제 분석함에 있어서의 실용성 측면을 강조하여 종합하면 다음과 같은 특성들로 요약할 수 있다.

- 방대한 분량과 이질성
- 시차별 데이터
- 상대적으로 많은 null 값

2.1.1. 방대한 분량과 이질성

의료 데이터는 이미지, 숫자, 동영상 등 이질적인 대량의 데이터로 구성되어 있어서 다른 분야의 데이터에 비하여 그 분석이 복잡하다. 따라서 대량의 이질적인 의료 데이터에 데이터 마이닝을 적용하는 경우에는 첫째, 다양한 형태의 데이터를 어떻게 종합적으로 분석할 것인가의 문제와 둘째, 더 효율적인 분석을 위하여 수많은 속성들 중에서 중요한 속성들을 어떻게 선택할 것인가의 문제가 발생하게 된다. 첫 번째 문제의 해결을 위한 일환으로 이미지 데이터를 효율적으로 분석하기 위한 기법들에 대한 연구들이 활발하게 진행되고 있으며[10, 11, 12, 13], 두 번째 문제의 경우에는 속성 선택(feature selection) 문제로 연구되고 있다[14, 15, 16, 23]. 특히 의료 데이터 마이닝에서의 속성 선택의 문제는 좋은 성능의 분류자를 생성하는데 매우 중요할 뿐만이 아니라[15, 16], 어떤 속성이 질병을 진단할 때 중요한가를 결정하는 데에도 많은 영향을 미치기 때문에[14] 신중하게 수행되어야 한다.

2.1.2. 시차별 데이터

의료 데이터는 한 명의 환자에 대하여 시간의 흐름에 따라 데이터가 축적되는 경향이 있다. 따라서 데이터를 분석할 때 이러한 시간적 특성을 고려해야 한다. 의료 데이터에서 발생하는 시간적 간격은 일정하지 않은 경우가 많은데, 이는 환자의 발병이나 병원 방문이 불규칙하기 때문이다. 이러한 시간적 특성은 다른 데이터에 비하여 상대적으로 많은 null 값의 빈도에도 큰 영향을 준다[8].

2.1.3. 상대적으로 많은 null 값

의료 데이터는 환자의 증상이나 발병의 정도, 부위에 따라 검사 및 치료 정보를 표현하는 속성들이 차이가 나기 마련이다. 따라서 의료 데이터 마이닝을 수행함에 있어서 이러한 이질적인 속성들을 모두 고려하게 되면 null 값의 발생이 많아지게 된다. 또한 앞에서 언급한 바와 같이 데이터가 불규칙적으로 발생하기 때문에 시간적인 속성을 고려한다면 이러한 null 값은 더욱 많이 발생할 수 밖에 없다. 그리고 null 값은 이와 같이 불가피하게 생성되는 경우도 있지만 데이터의 처리 과정에서 오류 때문에 누락되어 발생하게 되는 결측값도 포함한다. 따라서 의료 데이터 마이닝을 수행할 때에는 이러한 null 값이 예측 모형을 생성할거나 규칙을 추론할 때 큰 영향을 미치지 않도록 주의해야 할 것이다. Grzymala-Busse et al.(2000), Grzymala-Busse et al.(2001)는 의료 데이터에서 결측값을 어떻게 처리할 것인지에 대하여 연구를 수행한 바 있다[17, 18].

2.2. 진료비와 관련한 연구들

진료비와 관련한 연구는 진료비를 직접적인 연구 대상으로 다루는 경우와 간접적으로 고려하는 경우로 크게 구분할 수 있다. 진료비를 간접적으로 고려하는 경우는 진료비 산정에 큰 영향을 주는 항목의 변동을 예측하는 연구들이 많다. Goss와 Vozikis는 효율적인 의료 자원 사용을 위해서 중환자실에서 회복 가능성이 높은 환자를 예측하는 모형을 신경망 기법을 이용하여 구성하고 그 결과를 기존의 통계 기반의 소프트웨어 제품인 APACHE와 비교 분석한 바 있으며[19], Marshall 등은 노인병 환자의 입원 기간을 예측하는 모형을 생성하는데 신경망 방식을 사용하였으며, Barthel grade(일상생활

동작지수)와 퇴원 후 목적지가 입원 기간에 중요한 영향을 미치는 요인임을 밝혀낸 바 있다[20].

직접적인 연구 대상으로서의 진료비에 대한 연구들은 주로 회귀 모형과 같은 통계적 기법들을 기반으로 하였는데, 이와 같은 연구들은 상대적으로 그 예측률이 매우 낮을 뿐만 아니라, 의료 데이터베이스에 저장되어 있는 다양한 정보들의 매우 일부분만을 분석에 이용할 수 밖에 없었다. Penberthy 등은 유방암, 대장암, 폐암, 전립선암 환자들 중에서 특히 65세 이상의 환자들을 대상으로 진료비를 종속변수로 하는 회귀 모형을 만들었는데, 회귀식의 설명력이 38%~49%로 미약하게 나타나 이러한 회귀 모형을 바탕으로 진료비 예측을 하기에는 부족한 점이 많았다[5]. 이와 같은 통계적 분석 방식은 환자군별 진료비의 차이가 있는지의 여부 정도만을 판별하는 경우가 많았다[2, 3, 4].

신경망 기법을 이용하여 진료비 예측을 시도한 연구에서, Ismael 등은 신경망을 이용한 급성 관동맥 증후군 환자의 진료비 예측 모형을 생성하고, 이를 비교하는 연구를 수행하였다[21]. 총 16개의 입력변수를 이용하였으며, 환자의 특성을 나타내는 변수들과 합병증과 관련된 변수들, 병원 자원의 이용과 관련된 변수들로 구성하였다. 진료비를 네 개의 범주로 구분하여 예측률을 평가하였는데, 네 개의 모형 중에서 두 개의 모형이 79%의 분류 예측률을 보였다. 또한 예측된 진료비와 실제 진료비를 비교하기 위하여 회귀 모형의 R-square 값을 이용한 비교를 하였는데, 네 개의 모형 중에서 가장 높은 R^2값은 0.72였다. 대체적으로 진료비에 대한 연구는 다른 연구에 비하여 미미한 실정이며, 특히 데이터 마이닝 기법을 이용하여 진료비를 예측하는 연구는 그다지 많지 않았다.

3. 대장암 환자의 진료비 예측 모형

3.1. 대장암 환자 데이터

본 연구에서 진료비 예측의 대상은 대장암 환자들이다. 현재 우리나라에서 사망 원인 1위는 '암'이다. 통계청(2002)이 발표한 자료에 의하면 2001년 우리나라 전체 사망자는 약 24만 3천 여 명인데 이중 암으로 인한 사망자는 약 5만 9천 여 명으로 전체 사망자 4명 중 1명이 암으로 사망한 것으로 보고되고 있다[22]. 사망 원인의 1위라는 점과 함께 암이 주목 받고 있는 또 하나의 이유는 암으로 인한 사망률이 10년 전과 비교해 볼 때 매우 급격히 증가하고 있다는 점이다. 이러한 암으로 인한 사망의 급격한 증가에 대한 대책으로 정부에서는 지난 1996년부터 암 정복 10개년 계획을 수립하여 추진하고 있으며, 2001년 6월 국립 암 센터가 개원하여 국가 암 관리 사업을 추진하고 있다. 정부는 2005년까지 5대 암(위암, 간암, 대장암, 유방암, 자궁 경부암)에 대한 검진체계를 구축할 계획이며 그 일환으로 올해부터 건강보험 대상자중 저소득 하위 20%를 대상으로 위암과 유방암에 대한 무료검진사업을 추진하고 있다[22].

따라서 5대 암 중의 하나인 대장암의 진료비를 예측하고자 하는 본 연구는 국가 정책적 측면에서 볼 때, 향후 암 관리 사업에 대하여 그 기여도가 높을 것이며, 한정된 의료 자원의 효율적 분배에도 도움이 될 것이다. 또한 환자의 진료비 모형을 바탕으로 불필요한 치료나 검사가 있었는지를 추적할 수 있도록 하며, 환자에게도 미리 진료비에 대한 계획을 세울 수 있는 기회를 부여할 수 있고 새로운 치료 방법과 기존의 치료 방법의 효과와 비용을 비교할 때 기초적인 자료로 이용될 수도 있다.

3.2. 데이터 설명 및 속성 선택

본 연구에서 분석 대상으로 하는 데이터는 K 대학 병원의 1999년부터 2002년까지 발생한 병원에 입원하여 치료를 받은 대장암 환자의 진료기록을 바탕으로 하며, 492명의 환자들을 대상으로 1022개의 레코드, 154개의 속성들로 구성되어 있다. 1022개의 레코드 중 진료비가 다른 환자에 비하여 지나치게 크게 발생한 환자가 한 명 있었는데, 도메인 전문가와 의논하여 이를 삭제하고 1021개의 레코드를 바탕으로 분석을 수행하였다. 레코드는 환자의 병록번호를 식별자로 하며, 진료 및 치료와 관련된 속성들과 환자의 인구 통계학적인 특성들을 나타내는 속성들, 발생한 진료비 속성들로 구성되어 있다.

가장 많이 입원한 환자는 4년간 14번 입원하였으며 전체 데이터에서 환자 별로 평균 2.079회 입원하였다. 대장암 환자의 남녀 구성비는 각각 62%, 38%로 남성의 비율이 매우 높은 것을 알 수 있다. 환자의 평균 연령은 58.86세로 상당히 고령의 환자들이 많으며 평균 입원 기간은 11.38일이었다.

진료비 항목들은 보험 지급 항목들과 비보험 지급 항목들, 특진비 항목들 등의 세 집단으로 이루어져 있으며, 이러한 진료비 항목들 중 특히 보험 지급 항목과 비보험 지급 항목의 구성비가 전체 진료비에서 차지하는 비중이 높다. 진료비에 대한 기초 통계량은 <표 1>에 자세히 나타내었다.

본 연구에서는 보험 지급 항목 총계와 비보험 지급 항목 총계를 예측하는 모형을 만들고자 하는데, 기존의 진료비 연구는 대부분 보험회사의 자료를 근간으로 하였기 때문에 환자가 부담해야 하는 부분에 대한 설명은 부족한 측면이 있었다 [2, 3, 4, 5]. 따라서 이 두 개의 항목을 구분하여 진료비를 예측하는 모형을 생성하게 되면 이러한 점을 다소 보완할 수 있으리라 생각한다.

의료 데이터 마이닝에서 가장 신중해야 하는 단계들 중 하나가 속성 선택이다. 모형을 생성할 때,

표1: 진료비 구성항목의 기초통계량 (단위: 원)

진료비 구성항목	최소값	최대값	평균	분산
보험 지급항목 총계	5,353	14,037,226	2,312,199.56	1665172.75
비보험 지급 항목 총계	0	6,495,387	688,617.35	805124.27
특진비 항목 총계	0	2,344,019	329,946.55	443004.87
총 진료비	10,589	16,525,256	3,330,763.46	2527158.40

예측률을 높이기 위해서는 모든 속성들을 이용하여 모형을 생성하는 것보다 적절한 속성들을 선택하여 모형을 생성하는 것이 더 바람직하다 [14]. 그러나 본 연구에서는 이러한 속성 선택을 전적으로 도메인 전문가에 일임하였는데, 그 이유는 해당 의료 분야에서 오랜 경험을 가지고 있는 전문가들은 많은 속성들 중에서 진료비 예측과 무관한 속성들을 알고 있으며, 또한 중요한 속성들이 제외되는 것을 피할 수 있기 때문이다. 실제로 Demšar 등의 연구에서는 RELIEFF 기법에서 선택된 속성들에 대하여 도메인 전문가들이 중요한 속성인데도 불구하고 속성선택과정에서 배제된 속성들을 사후에 추가하거나 더 중요한 속성으로 대체하였다[14].

본 연구에서 결측값의 처리 역시 도메인 전문가의 도움을 받아 해결하였는데, 전문가들이 선택한 속성들로부터 생성한 유도 변수들을 통하여 결측값을 최소화하고 입력 변수로 사용할 필드의 숫자를 줄일 수 있었다. 예를 들어, 현재 분석하고자 하는 데이터에는 수술1부터 수술10까지 필드 값으로 수술을 담당한 의사코드와 수술횟수가 결합된 값이 입력되어 있는데, 수술9와 수술10 필드의 경우에는 대부분 null 값으로 구성되어 있다. 이러한 경우 수술 횟수를 모두 합한 필드 하나를 생성한다면 null값을 상당히 감소시킬 수 있고, 필드의 개수를 줄일 수 있다. 또한 기타 질병 코드는 다시 그룹화하여 23개의 필드를 생성하였으며, 그룹별 질병 코드의 보유 개수를 표시하도록 하였다. 필드 operation1~operation10에서 표시하고 있는 ICD9 코드[1] 들도 추출하여 도메인 전문가에 의뢰하여 그룹으로 묶어 수술적 치료, 방사선 검사, 방사선 치료, 항암제 치료 관련 술식의 네 개의 필드를 생성하여 각 그룹 별 ICD9 코드의 보유개수를 표시하도록 하였다.

이러한 전처리 과정은 앞에서 언급한 의료 데이터의 특성을 고려하여 많은 필드들로 구성되어 있고, null 값이 많은 데이터로부터 도메인 전문가에 의하여 분석에 필요한 속성들을 선택하고, 유도된 속성들을 생성하여 null 값을 줄이고자 하였다. 또한 시간적 특성을 고려하여 표본 추출에 있어서 어느 특정 기간에 치우친 훈련데이터 및 검증데이터를 생성하지 않도록 층화추출을 사용하였다. 이러한 전처리 과정을 거쳐 생성된 필드는 총 51개이다. 전처리 과정은 주로 MS EXCEL VBA과 ACCESS를 이용하였다.

3.3. 진료비 예측 모형

본 연구에서는 진료비와 관련한 기존의 다른 연구들에서 주로 사용한 통계적 방법인 회귀 분석과 데이터 마이닝 기법인 신경망 방식과 의사 결정 나무 방식을 통하여 생성된 모형의 결과를 비교하고자 하였다.

3.3.1. 회귀 분석을 이용한 진료비 예측 모형

본 연구에서는 Penberthy등의 연구를 참조하여 회귀 분석을 적용하였다[5]. 회귀 모형은 진료비 항목에서 보험 지급항목 총계를 종속변수로 하는 경우와 비보험 지급항목 총계를 종속변수로 하는 경우 두 개의 모형을 생성하였다. 두 모형에서 독립변수는 동일하게 구성하였는데, 독립변수는 나이, 입원기간, 중환자실 입실횟수, 전과횟수, 협진횟수, 수술횟수, 보유하고 있는 암 이외의 기타 질병의 개수, 수술적 치료, 방사선 검사, 방사선 치료, 항암제 치료 관련 술식 11개이다. 여기서 나이를 제외하고는 모두 환자의 진료와 직접적으로 관련된 변수들이다. 이것은 Penberthy 등[5]의 연구에서 회귀 모형을 생성할 때, 총 13개의 독립변수들 중에서 진료와 직, 간접적인 관련이 있는 변수는 6개였다는 점을 고려할 때, 더 좋은 예측력을 가지는 회귀 모형을 생성할 수 있음을 시사한다.

종속 변수인 진료비는 보험 지급 항목 총계와 비보험 지급 항목 총계 모두 그 분포가 한쪽으로 치우쳐져 있으므로 자연로그를 취하였다[5]. 이 과정에서 비보험 지급 항목 총계의 경우 로그를 취한 값이 매우 작은 경우에는 결측치 처리가 되었는데, 이러한 결측치는 모두 0으로 처리하였다. 회귀 분석은 SPSS 10을 이용하였으며, 두 회귀식 중 비보험 지급 항목 총계를 종속변수로 하는 회귀 모형은 Durbin-Watson 검정 결과 자기 상관이 있는 것으로 나타났기 때문에 본 논문에서는 그 결과를 설명하지 않았다. 구성된 회귀식은 다음과 같다.

[1] ICD code는 World Health Organization(WHO)에서 만든 질병, 건강 상태, 시술에 대한 분류 체계로, 질병에 대한 숫자로 구성된 코드와 라벨에 대한 국제 기준을 표시한다. ICD-8, ICD-9, ICD-9-CM (Clinical Modifications), ICD-O (Oncology), ICD-10 등 여러 가지 버전이 있다.

$$\ln(Y_1) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + b_8x_8 + b_9x_9 + b_{10}x_{10} + b_{11}x_{11}$$

$$\ln(Y_2) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + b_8x_8 + b_9x_9 + b_{10}x_{10} + b_{11}x_{11}$$

Y_1 : 총 진료비 중 보험 지급 항목 총계
Y_2 : 총 진료비 중 비보험 지급 항목 총계
x_1 : 나이
x_2 : 입원기간
x_3 : 중환자실 입실 횟수
x_4 : 전과 횟수
x_5 : 협진 횟수
x_6 : 수술 횟수
x_7 : 기타 질병 진단명 보유개수
x_8 : 수술적 치료
x_9 : 방사선 검사
x_{10} : 방사선 치료
x_{11} : 항암제 치료 관련 술식

3.3.2. 신경망을 이용한 진료비 예측 모형

본 연구에서는 진료비를 예측하는 모형을 만들기 위하여 Clementine 7.1에서 지원하는 신경망기법(feed-forward backpropagation network)을 사용하였으며, 두 개의 은닉 층에 대하여 다양한 위상 구조를 가진 모형을 훈련시켜 가장 예측력이 좋은 최적화된 모형을 도출하도록 하였다. 신경망 모형은 출력 변수가 보험 지급항목 총계인 것과 비보험 지급항목 총계인 것, 그리고 두 변수를 모두 출력 변수로 하는 세 가지 모형을 생성하였다. 입력 변수는 회귀모형에서 사용한 독립변수 11개에 기타질병 그룹1~기타질병그룹23, 성별, 원내감염여부, 환자 구분을 더한 총 37개이다. 이러한 추가된 변수들은 대부분 회귀 분석에서 사용할 수 없는 명목 척도들이다. 훈련 데이터와 검증 데이터는 2:1의 비율로 구성하였으며, 샘플링 기법 중 층화 추출방식을 이용하여 검증데이터를 추출하였다. 훈련 데이터 집합은 681개의 레코드로 구성하였으며, 검증 데이터 집합은 340개의 레코드로 이루어졌다.

3.3.3. 의사 결정 나무를 이용한 진료비 예측 모형

본 연구에서 사용한 의사 결정 나무 방식은 연속형 변수를 목표변수로 설정할 수 있는 CART를 이용하였다. 진료비 예측 모형은 출력 변수를 보험 지급항목 총계로 하는 경우와 비보험 지급항목 총계로 하는 경우 두 가지로 생성하였으며, 입력 변수는 신경망에서 사용한 37개의 변수에 환자의 주진단을 더하여 총 38개의 변수들로 구성하였다. 훈련 데이터 집합과 검증 데이터 집합은 신경망 기법을 사용한 경우와 동일하게 층화추출을 이용하여 2:1의 비율로 구성하였다. CART 역시 Clementine 7.1을 이용하여 모형을 생성하였다.

4. 세 가지 진료비 모형의 결과 비교

4.1. 회귀 모형

앞에서 언급한 바와 같이 비보험 지급항목 총계를 종속변수로 하는 회귀모형의 경우는 자기상관이 있는 것으로 나타나 본 논문에서 그 결과에 대하여 설명하지 않도록 한다. 보험 지급항목 총계를 종속변수로 하는 회귀모형은 자기상관과 다중 공선성이 없는 것으로 나타났다. 회귀 분석 결과 11개의 독립변수들 중에서 회귀계수가 5%의 유의수준에서 유의한 것으로 나타난 변수들은 입원기간, 협진횟수, 수술횟수, 항암제 치료 관련 술식, 방사선 치료였다.

표2:보험지급항목총계를 종속변수로 하는 회귀 모형

R	R 제곱	수정된 R 제곱	추정값의 표준오차	통계량 변화량					Durbin -Watson
				R 제곱 변화량	F 변화량	자유도 1	자유도 2	유의확률 F 변화량	
0.69	**0.478**	0.473	0.6448	0.478	84.135	11	1009	0.00000	1.9827

<표2>에서 회귀모형의 R-square 값은 0.478로 Penberthy 등의 논문에서 대장암 환자와 관련한 회귀 모형의 R-square값이 0.46인 것이 비하여 약간 높은 것을 볼 수 있다[5]. 본 연구가 Penberthy 등의 연구보다 더 적은 수의 독립 변수를 분석에 사용했음을 고려한다면 비교적 좋은 결과를 보여주고 있다고 말할 수 있다.

표3: 입력방식으로 생성된 회귀 모형의 회귀 계수

독립 변수	회귀계수	t-통계량	유의확률
(상수)	**13.28358**	**117.58023**	**0.00000**
나이	0.00225	1.30697	0.19152
입원기간	**0.03708**	**12.18255**	**0.00000**
중환자실 입실 횟수	0.11236	1.23348	0.21768
전과 횟수	0.04983	0.71359	0.47564
협진 횟수	**-0.06426**	**-3.46442**	**0.00055**
수술 횟수	**0.58643**	**8.88692**	**0.00000**
기타 질병명 보유개수	0.01664	0.98670	0.32403
수술적 치료	0.02130	0.98227	0.32620
방사선 검사	0.04637	1.86235	0.06284
항암제 치료 관련 술식	**0.53045**	**10.84594**	**0.00000**
방사선 치료	**-0.19389**	**-2.32885**	**0.02006**

입력 방식으로 생성된 회귀모형의 회귀계수는 <표3>과 같다. <표3>중 굵은 글씨로 표시된 항목은 유의수준 5%에서 유의한 회귀계수를 나타낸다. 방사선 검사의 경우 유의수준 5%에서는 유의확률이 0.063으로 나타났으나, 단계 투입법으로 모형을 생성하였을 때에는 5% 유의수준에서 유의한 회귀계수로 나타났다.

회귀 분석 결과, 진료비에 유의한 영향을 미치는 변수는 입원기간, 수술횟수, 항암제 치료 관련 술식, 협진횟수, 방사선 검사, 방사선 치료의 순으로 나타났으며, 이러한 변수들이 진료비 결정에 중요한 영향을 미치는 변수임을 알 수 있다.

표 4: 세 가지 신경망 모형의 변수의 상대적 중요도

신경망 모형 A (목표변수:보험 지급 항목 총계)		신경망 모형 B (목표변수:비보험 지급 항목 총계)		신경망 모형 C (목표변수:보험 지급 항목 총계, 비보험 지급 항목 총계)	
변수명	상대적 중요도	변수명	상대적 중요도	변수명	상대적 중요도
hospital_stay	0.69522	hospital_stay	0.25352	hospital_stay	0.37410
icu_count	0.20111	consult_count	0.11959	operation_count	0.15797
operation_count	0.17215	방사선 검사	0.10060	기타질병그룹 12(세균감염증)	0.08796
기타질병그룹 4(기생충질환)	0.11613	operation_count	0.09491	기타질병그룹 4(기생충질환)	0.08364
항암제 치료 관련 술식	0.10384	수술적 치료	0.08795	기타질병그룹 2(결합조직질환)	0.07965
기타질병그룹 5 (기타 보건상태에 대한 영향)	0.06118	기타질병그룹 16(심혈관계질환)	0.08696	기타질병그룹 5 (기타 보건 상태에 대한 영향)	0.07307
consult_count	0.05887	기타질병그룹 3(근골격계질환)	0.06403	patient_diff	0.06947
korean_age	0.05330	patient_diff	0.06031	수술적 치료	0.06569
기타질병그룹 10 (비뇨생식기계질환)	0.05028	transfer_count	0.05626	방사선 검사	0.05912
기타질병그룹 1(간담도계질환)	0.04980	기타질병그룹 13(소화기질환)	0.04969	기타질병그룹 16(심혈관계질환)	0.05717

4.2. 신경망 모형

본 절에서 세 가지 신경망 모형의 예측 결과를 설명하고자 하며, 설명의 편의를 위해서 보험 지급 항목 총계를 목표 변수로 하는 신경망 모형을 신경망 모형A, 비보험 지급항목 총계를 목표 변수로 하는 신경망 모형을 신경망 모형B, 보험 지급항목 총계와 비보험 지급항목 총계를 목표변수로 하는 신경망 모형을 신경망 모형C라고 지칭하도록 한다.

신경망 모형은 민감도 분석을 통하여 진료비 예측에 있어서 상대적으로 중요한 변수들을 파악할 수 있도록 하는데, 각 모형 별로 상대적으로 중요한 변수들이 상이하였다. 세 모형 모두 입원일수가 진료비 예측에 있어서 가장 중요한 변수로 나타났는데, 특히 보험 지급항목 총계를 목표변수로 하는 신경망 모형A에서 그 상대적 중요도가 가장 높게 나타났다. 또한 세 모형 공히 수술횟수가 진료비 예측에 있어 상대적으로 중요도가 매우 높은 변수로 나타났다. 신경망 모형A에서는 입원일수, 중환자실 입실횟수, 수술횟수, 기타질병그룹4 등이 중요한 변수였지만, 신경망 모형B에서는 입원일수, 협진 횟수, 방사선 검사, 수술횟수의 순으로 그 중요도가 서로 달랐다. 신경망 모형C는 입원일수, 수술횟수, 기타질병그룹12, 기타질병그룹4의 순으로 상대적으로 중요한 변수로 판명되었다. <표4>는 세 모형에서 변수의 상대적 중요도를 비교하고 있다.

각 모형 별로 예측치와 관측치의 차이가 어느 정도 되는지를 살펴보고, 그 둘의 선형 상관관계를 살펴보았는데, 회귀 모형에 비하여 모두 높은 상관관계를 나타내었다. 특히 신경망 모형A와 신경망 모형C에서 보험 지급 항목 총액을 예측하는 경우에는 선형 상관관계가 0.8이상으로 매우 높게 나타났다. 이것은 회귀 모형에서 예측치와 관측치의 선형 상관관계가 0.692임과 비교할 때 상당히 높다고 말할 수 있다 (<표5> 참조).

표 5: 신경망 모형의 예측 오차 및 선형 상관관계

	측정항목	training dataset	test dataset
신경망 모형 A	Mean Error	-10030.191	-96131.95
	Mean Absolute Error	602515.492	693106.097
	Linear Correlation	0.885	0.845
신경망 모형 B	Mean Error	8713.662	-61076.544
	Mean Absolute Error	323603.46	376597.691
	Linear Correlation	0.782	0.661
신경망 모형 C	목표변수 1: 보험 지급 항목 총계		
	Mean Error	9323.347	1546.188
	Mean Absolute Error	713946.592	759673.535
	Linear Correlation	0.833	0.805
	목표변수 2: 비보험 지급 항목 총계		
	Mean Error	387.624	-32772.841
	Mean Absolute Error	332760.32	353895.882
	Linear Correlation	0.748	0.711

4.3. CART 모형

본 연구에서 CART를 이용하여 보험 지급항목 총계와 비보험 지급항목 총계를 각각 목표 변수로 하는 의사결정나무 모형을 생성한 결과를 설명하고자 한다. 신경망 모형의 경우와 마찬가지로 설명의 편의를 위하여 보험 지급항목 총계를 목표 변수로 하는 CART 모형을 CART 모형A라고 하고 비보험 지급항목 총계를 목표 변수로 하는 CART 모형을 CART 모형B라고 하도록 한다.

CART 모형A의 생성 결과로 나온 의사 결정 나무는 7단계로 구성되어 있고, 가지를 분화할 때 불순도 측정 방식은 Gini Index를 사용하였다. 생성된 의사 결정 나무를 살펴 보았을 때, 뿌리 노드에서 맨 처음 가지를 분화하는 기준은 입원일수였다. 의사결정나무에서 가지를 분화하는 기준으로 사용된 속성들은 입원일수, 주진단, 수술횟수, 성별, 수술적 치료였으며 규칙집합(ruleset)을 구성하는 규칙들은 모두 11개로 진료비가 높은 영역을 잘 세분화하는 규칙들이 생성되었다. 규칙집합(ruleset)에서 특히

'operation_count'과 '수술적 치료'필드는 진료비가 상대적으로 높은 환자들을 가늠하는 중요한 세부 기준으로 사용되었다.

CART 모형B는 CART 모형A와 동일한 입력 변수를 사용하였고, 나무의 크기는 초기에는 7단계로 동일하게 구성하였는데, 생성된 의사결정 나무가 진료비가 적은 부분에 대하여 세분화되는 경향이 나타나 가지치기를 하고 4단계로 구성하였다. 가지를 분화하는 기준으로 사용하는 불순도 측정 방식 역시 Gini Index를 사용하였다. 가지를 분화하는 기준으로 사용된 속성들은 입원일수, 주진단, 방사선 치료, 방사선 검사, 기타 질병 그룹 3(근골격계 질환)이었으며, 규칙집합을 구성하는 규칙의 개수는 10개였다. 규칙집합을 살펴보면 입원일수가 11일 이상 30일미만인 경우에 방사선 검사 횟수가 예측된 진료비에 큰 차이를 주는 가지 분화의 기준이었다.

CART모형의 예측 오차와 선형 상관계수를 조사한 결과, 예측 오차의 경우 CART 모형A가 신경망 모형 A보다 약간 큰 것으로 나타났으며, 선형 상관계수는 신경망 모형A보다 작게 나타났다. 그리고 CART 모형B의 경우에는 훈련데이터에서는 신경망 모형 B보다 선형 상관계수가 낮게 나타났으나, 검증 데이터에서는 선형 상관계수가 보다 높게 나타났다. 전반적으로 CART 모형이 신경망 모형에 비하여 데이터에 대하여 과잉 적합(overfitting)이 덜 나타났다. CART 모형에 대한 예측 오차 및 선형 상관계수는 <표6>에 나타내었다.

표6: CART 모형의 예측 오차 및 선형 상관관계

	측정항목	Training dataset	Test dataset
CART 모형A	Mean Error	0.598	28377.009
	Mean Absolute Error	753055.907	679692.174
	Linear Correlation	0.724	0.796
CART 모형 B	Mean Error	0.345	-1995.118
	Mean Absolute Error	310534.295	317274.641
	Linear Correlation	0.768	0.767

4.4. 세 가지 방식의 진료비 예측 모형 비교

본 연구에서는 회귀 분석과 신경망, CART 세 가지 방식을 이용하여 진료비를 예측하는 모형을 생성하였다. 회귀 모형을 포함하여 본 연구에서 생성한 진료비 예측 모형은 모두 여섯 개이며, 회귀 분석의 경우는 모형의 설명력을 R^2값으로 설명하고 있지만 나머지 모형들은 모두 연속형 변수를 목표변수로 하고 있어 예측 정확도(prediction accuracy)를 측정할 수가 없었다. 그래서 예측값과 실제값의 선형 상관관계를 계산하여 어느 정도의 선형적 상관관계가 있는지를 조사함으로써 모형을 평가하였다.

<표7>을 살펴보면 신경망 모형들이 선형 상관관계가 가장 강하고, 다음이 CART 모형들이며, 회귀 모형의 선형 상관계수가 가장 작게 나타남을 알 수 있다. 이것은 진료비를 예측하는 데 있어서 기존의 통계적 방법에 비하여 데이터 마이닝 방법이 더 효과적임을 나타낸다. 데이터 집합에서는 보험 지급 항목 총계를 예측하는 경우가 비보험 지급 항목 총계를 예측하는 경우보다 선형 상관계수가 더 높게 나타났는데, 특히 신경망 모형A와 신경망 모형C의 경우에는 선형 상관 계수가 0.8이상으로 매우 높게 나타났다. 또한 훈련 데이터 집합과 검증 데이터 집합의 선형 상관계수를 비교해보면, 신경망 모형에 비하여 CART 모형이 데이터에 대하여 보다 절 과잉적합 되는 것을 알 수 있다. 이것은 CART 모형이 보다 일반화된 모형을 생성함을 의미한다.

표7: 모형 별 선형 상관계수

모형		피어슨 상관계수	
		Training dataset	Test dataset
회귀모형 (보험지급항목 총계)		0.692 (Total dataset)	
신경망모형 A (보험지급항목 총계)		0.885	0.845
신경망모형 B (비보험지급항목 총계)		0.782	0.661
신경망모형 C	보험지급항목 총계	0.833	0.805
	비보험지급항목 총계	0.748	0.711
CART 모형 A (보험지급항목 총계)		0.724	0.796
CART 모형 B (비보험지급항목 총계)		0.768	0.767

이러한 결과들을 종합해 볼 때, 진료비 예측에 있어서 기존의 통계적인 방법보다 데이터 마이닝을 이용한 방법이 보다 효과적이라고 말할 수 있다.

5. 결론

본 연구는 K대학병원의 대장암 환자들의 진료비 예측을 위한 예측모형을 생성하였다. 기존의 진료비 연구에서 주로 사용하였던 회귀 분석과 데이터 마이닝 기법인 신경망과 의사결정나무를 사용하여 대장암 환자의 진료비 예측 모형을 생성하고 비교하였는데, 그 결과 통계적 방식인 회귀 분석에 비하여 데이터 마이닝 기법인 신경망과 의사 결정 나무가 더 좋은 예측결과를 보여주었다.

그러나 본 연구에서 진료비 예측모형을 생성하기 위한 데이터에는 환자의 모든 검사기록과 치료 기록이 포함되어 있지 않은데, 이것은 K 병원의 진료 기록이 일부분만 전산 처리되어 있기 때문이었고, 따라서 한정된 환자 진료 기록을 바탕으로 진료비 예측모형을 생성할 수 밖에 없었다. 특히 암 환자의 진료비와 관련된 많은 연구들이 암의 진행 단계에 대한 정보를 매우 중요하게 고려하였는데[4, 5], 본 연구에서는 이러한 속성을 연구에 포함하여 모형을 생성할 수 없었다는 한계점이 있다.

하지만 현재 K대학병원은 2003년부터 모든 검사 결과와 진단 및 치료 기록을 일원화하고 자동 기록하는 체계를 구축하여 환자 기록을 수집하고

있다고 한다. 따라서 향후 이러한 기록들이 축적되어 이를 바탕으로 진료비 예측 모형을 생성한다면 보다 정확한 예측력을 가지는 모형을 생성할 수 있을 것이라 기대된다. 그리고 본 연구에서는 연구대상을 대장암 환자의 진료비 예측에 국한하였는데, 5대 암으로 연구대상을 확대한다면 각 암 별 환자의 진료비를 비교할 수 있고, 어떤 속성이 특정 암의 진료비 산정에 더 큰 영향을 미치는지를 파악할 수 있을 것이다. 또한 5대 암에 대한 진료비 연구는 암 정복 사업과 같은 국가 정책을 세우는데 그 근간이 되는 자료로 사용될 수도 있을 것이다.

참고 문헌

[1] Jenn-Lung Su, Guo-Zhen Wu and I-Pin Chao (2001), "The Approach of Data Mining Methods for Medical Database," *Engineering in Medicine and Biology Society, Proceedings of the 23rd Annual International Conference of the IEEE*, Vol. 4, pp.3824 -3826.

[2] Roche, K., Paul, N., Smuck, B., Whitehead, M., Zee, B., Pater, J., Hiatt, M.A., and Walker, H. (2002), "Factors Affecting Workload of Cancer Clinical Trials: Results of a Multicenter Study of the National Cancer Institute of Canada Clinical Trials Group," *Journal of Clinical Oncology*, Vol. 20, No. 2, pp.545-556.

[3] Fireman, B.H., Fehrenbacher, L., Gruskin E.P., and Ray, G.T. (2000), "Cost of Cancer for Patients in Cancer Clinical Trials," *Journal of the National Cancer Institute*, Vol. 92, No. 2, pp.136-142.

[4] Tollestrup, K., Frost, F.J., Stidley, C.A., Bedrick, E., McMillan, G., Kunde, T., and Petersen, H.V. (2001), "The excess costs of breast cancer health care in Hispanic and non-Hispanic female members of a managed care organization," *Breast Cancer Research and Treatment 66*, pp.25-31.

[5] Penberthy, L., Retchin, S.M., McDonald, M. K., McClish, D.K., Desch, C.E., Riley, G.F., Smith, T.J., Hillner, B.E. and Newschaffer, C.J. (1999), "Predictors of Medicare costs in elderly beneficiaries with breast, colorectal, lung, or prostate cancer," *Health Care Management Science 2*, pp.146-160.

[6] Cios, K.J. and Moore, G.W. (2002), "Uniqueness of Medical Data Mining," *Artificial Intelligence in Medicine*, Vol. 26, pp.1-24.

[7] Groselj, C. (2002), "Data Mining Problems in Medicine," *the 15th IEEE Symposium on Computer-Based Medical Systems*, pp.377 -380.

[8] Tsumoto, S. (2001), "Temporal knowledge discovery in time-series medical databases based on fuzzy-rough reasoning," *IFSA World Congress and 20th NAFIPS International Conference*, Vol.4, pp.1973 -1978.

[9] Berman, J.J. (2002), "Confidentiality issues for medical data miners," *Artificial Intelligence in Medicine*, Vol. 26, pp. 25-36.

[10] Miin-Shen Yang, Yu-Jen Hu, Karen Chia-Ren Lin and Charles Chia-Lee Lin (2002), "Segmentation techniques for tissue differentiation in MRI of Ophthalmology using fuzzy clustering algorithms," *Magnetic Resonance Imaging*, Vol. 20, pp. 173-179.

[11] Barra, V. and Boire, J.Y. (2001), "A General Framework for the Fusion of Anatomical and Functional Medical Images," *NeuroImage*, Vol. 13, pp. 410-424.

[12] Masulli, F. and Schenone, A. (1999), "A fuzzy clustering based segmentation system as support to diagnosis in medical imaging," *Artificial Intelligence in Medicine*, Vol. 16, pp. 129-147.

[13] Markey, M.K., Lo, J.Y., Tourassi, G.D. and Floyd, C.E., Jr. (2003), "Self-organizing map for cluster analysis of a breast cancer database," *Artificial Intelligence in Medicine* Vol. 27, pp. 113-127.

[14] Demšar, J., Zupan, B., Aoki, N., Wall, M.J., Granchi, T.H. and Beck, J.R. (2001), "Feature mining and predictive model construction from severe trauma patient's data," *International Journal of Medical Informatics*, Vol. 63, pp.41-50.

[15] Skrypnik, I., Terziyan, V., Puuronen, S. and Tsymbal, A.(1999), "Learning Feature Selection for Medical Databases," *Computer-Based Medical Systems, 1999. Proceedings. 12th IEEE Symposium on*, pp.53 -58.

[16] Puuronen, S., Tsymbal, A., and Skrypnyk, I.(2000), "Advanced Local Feature Selection in Medical Diagnostics," *Computer-Based Medical Systems, Proceedings of 13th IEEE Symposium on*, pp. 25 -30.

[17] Grzymala-Busse, J. W. and Hu., M. (2000), "A comparison of Several Approaches to Missing Attribute Values in Data Mining, " *Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing RSCTC'2000*, pp. 340 - 347.

[18] Grzymala-Busse, J.W., Grzymala-Busse W.J., and Goodwin, L.K. (2001), "Coping With Missing Attribute Values Based on Closest Fit in Preterm Birth Data: A Rough Set Approach," *Computational Intelligence*, Vol. 17, Issue 3, pp. 425-434.

[19] Goss, E.P. and Vozikis, G. S. (2002), "Improving Health Care Organizational Management Through Neural Network Learning," *Health Care Management Science 5(3)* pp.221-227.

[20] Marshall, A.H., McClean, S.I., Shapcott, C.M. and Millard, P.H. (2002), "Modelling Patient Duration of Stay to Facilitate Resource Management of Geriatric Hospitals," *Health Care Management Science 5(4)* pp.313-319.

[21] Ismael, M.B., Eisenstein, E.L and Hammond, W.E.(1998), "A comparison of neural network models for the prediction of the cost of care for acute coronary syndrome patients, " *Proceedings of AMIA Symposium*, pp. 533-537.

[22] 김창보, 김기영 (2002), 2001년 건강보험 암 환자의 진료실태, 연구자료 2002-14, 국민 건강보험 공단.

[23] Kohavi, R. and John, G. H. (1997), "Wrappers for feature subset selection," *Artificial Intelligence*, Vol. 97, pp. 273-324