

# 머신러닝을 이용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형 개발

**박 종 호**

(계명대학교 동산의료원)

**강 성 흥\***

(인제대학교)

본 연구는 머신러닝을 이용하여 동반상병 보정 방법에 따른 중증도 보정 재원일수 예측 모형을 개발하고 이를 평가하여 중증도 보정 재원일수 예측 모형 개발의 알고리즘을 제시하기 위해 수행되었다. 본 연구를 위해 2006년부터 2015년까지 10년간의 질병관리본부 퇴원손상심층조사 자료를 수집하였으며, 재원일수 관리가 시급한 신경계통의 질환을 대상으로 중증도 보정 재원일수 예측 모형을 개발하였다. 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형 개발 시 동반상병 보정 방법은 CCI, ECI, CCS 진단군 분류 기준 등 3가지, 머신러닝 분석기법으로는 회귀분석, 의사결정나무, 랜덤 포레스트, 서포트 벡터 회귀분석, 신경망 등 5가지를 적용하여 모형을 개발하고 개발된 모형을 평가하였다. 모형 평가 결과 CCS 진단군 분류 기준 동반상병 보정 방법 및 신경망을 이용하여 개발한 중증도 보정 예측 모형의 모형 설명력(R-square)이 가장 높았으며, 모형의 예측력이 가장 우수한 것으로 나타났다. 따라서 중증도 보정 재원일수 예측 모형 개발 시 CCS 진단군 분류 변수를 이용한 동반상병 보정 방법을 이용하여 중증도 보정 예측 모형을 개발하는 것이 필요하며, 머신러닝의 다양한 분석 기법 등을 이용하여 예측력 높은 중증도 보정 예측 모형을 개발하여 재원일수 변이요인 파악 등 재원일수 관리를 위해 활용하는 것이 필요하다.

주요 용어: 재원일수, 신경계통의 질환, 동반상병 보정, 머신러닝, 중증도 보정 예측 모형

\* 교신저자: 강성흥, 인제대학교(hcmkang@inje.ac.kr)

■ 투고일: 2019.1.14    ■ 수정일: 2019.3.20    ■ 게재확정일: 2019.3.21

## I. 서론

입원환자의 재원일수는 제공되는 의료서비스의 질을 측정하는 지표이자 의료기관의 자원이 얼마나 효율적으로 사용되는지를 보여주는 중요한 지표이다(OECD, 2014). 우리나라 입원환자의 평균재원일수는 2015년 기준 16.1일로 경제협력개발기구(Organization for Economic Cooperation and Development, 이하 OECD) 평균보다 약 2배 높았으며, OECD 국가 중 일본에 이어 2번째로 높았다. 특히 우리나라의 경우 입원환자의 평균재원일수가 2002년 14.7일에서 2015년 16.1일로 증가하여 재원일수 관리 문제가 심각하다고 할 수 있다(OECD, 2017). 국가 차원 및 의료기관 차원의 효율적 재원일수 관리 방안을 제시하기 위해 선행연구에서는 중증도 보정 재원일수 예측 모형을 개발(강성홍, 석향숙, 김원중, 2013; 최병관, 함승우, 김축환, 서정숙, 박명화, 강성홍, 2018)하고, 이를 통한 재원일수 변이요인을 규명하였으며, 개발된 중증도 보정 재원일수 예측 모형을 이용하여 의료기관 간 재원일수 비교가 가능한 중증도 보정 재원일수 벤치마킹 시스템(질병관리본부, 2012; 질병관리본부, 2014)을 개발하였다. 정확하게 재원일수 변이요인을 규명하고, 의료기관이 재원일수 벤치마킹 시스템 효과적으로 이용하여 재원일수를 줄이는 노력을 기울이게 하려면 개발된 중증도 보정 재원일수 예측 모형의 예측 정확도가 보장되어야 한다. 이를 위해서는 중증도 보정 재원일수 예측 모형 개발 시 입원환자의 중증도가 반영되어야 하며, 다음의 사항이 반영되어야 한다.

첫째, 입원환자의 중증도 반영이다. 중증 환자가 경증 환자보다 상대적으로 재원일수가 높다고 할 수 있기 때문에 재원일수 예측 모형을 개발하기 위해서는 입원환자에 대한 중증도 보정이 우선되어야 한다(송태민 등, 2013; 홍성욱, 김영택, 박종호, 강성홍, 2015). 우리나라 뿐 아니라 미국, 캐나다, 호주 등에서는 진료비 지불 등 행정 및 통계 산출 목적으로 국가 차원에서 전체병원을 대상으로 수집한 행정자료를 이용하여 중증도 보정 방법 및 중증도 보정 재원일수 모형 개발에 대한 연구가 활발히 진행되고 있다(강성홍, 석향숙, 김원중, 2013). 행정자료를 이용한 중증도 보정 방법에는 부상병 즉 동반상병을 이용하여 입원환자의 중증도를 보정하는 방법이 있다. 동반상병을 이용하여 입원환자의 중증도를 보정하는 대표적인 방법에는 Charlson comorbidity index(이하 CCI), Elixhauser comorbidity index(이하 ECI)가 있다. 하지만 CCI의 경우 1984년에 개발(Charlson, Pompei, Ales & MacKenzie, 1987) 되었고, ECI는 1998년에 개발

(Elixhauser, Steiner, Harris & Coffey, 1998)되었음에 따라 동반질환 기여도가 변화했을 가능성이 있다(Quan et. al., 2011). 이에 외국에서는 ECI와 CCI 대신 각 질환마다 자국의 퇴원환자 자료로 산출한 별도의 중증도 보정 방법을 개발하여 동반질환을 보정하였으며, 우리나라의 선행연구에서도 CCS(Clinical classification software) 진단군 분류(이하 CCS 진단군 분류)를 이용하여 부상병을 분류하고 ECI 방법처럼 각각의 질환 동반 여부를 이분형 형태로 중증도 보정 모형에 입력하여 환자의 중증도를 보정하였다(질병관리본부, 2012; 질병관리본부, 2014). 따라서 재원일수 예측모형에 입원환자의 중증도를 정확히 반영하기 위해서는 CCI, ECI를 이용한 동반상병 보정 뿐 아니라 CCS 진단군을 이용한 새로운 동반상병 보정 또한 고려하여야 할 필요가 있다.

둘째, 예측모형의 정확도 향상을 위한 최신 분석기법의 반영이다. 2016년 1월 다보스 세계경제포럼 이후 모든 기술의 자동화에 기반한 소비자 맞춤형 경제를 의미하는 4차 산업혁명이 이슈가 되고 있으며, 현재 4차 산업혁명이 사회 전 분야로 확산되고 있다(World Economic Forum, 2016; 이강운, 김준혁, 2016). 4차 산업혁명은 인공지능(Artificial Intelligence) 기술을 통하여 세상을 이해하는 시대이다. 인공지능이란 빅데이터의 시각화(Visualization), 머신러닝(Machine learning), 딥러닝(Deep learning)의 기술을 통해 자료를 분석하고 더 나은 답을 찾아내는 기술을 의미한다. 인공지능 기술 중 머신러닝은 수많은 자료를 읽고 알고리즘을 찾아내어 변화를 예측하는 예측기법을 말한다(이강운, 김준혁, 2016). 보건의료산업에서도 진단의학, 신약 개발 시스템 등에 머신러닝 기법을 적용하여 활용하는 사례가 많이 나타나고 있으며(이관용, 김진희, 김현철, 2016), 전통적 통계기법과 머신러닝 기법을 적용한 모형의 예측력을 비교한 결과 머신러닝 기법을 적용한 모형의 예측력이 더 높았다는 연구결과가 지속적으로 발표되고 있다(김현호, 양승범, 강연석, 박영배, 김재효, 2016; 고은해, 강희양, 김용식, 정성훈, 2017; Rajkumar et. al. 2018). 따라서 머신러닝을 이용하여 중증도 보정 재원일수 예측모형을 개발하고 예측력을 비교할 필요가 있다.

이에 본 연구에서는 머신러닝을 이용하여 동반상병 보정 방법에 따른 중증도 보정 재원일수 예측 모형 개발하고, 예측력 비교를 근거로 재원일수의 효율적 관리를 위한 중증도 보정 재원일수 모형 개발의 알고리즘을 제시하고자 한다.

## II. 이론적 배경

### 1. OECD 국가의 퇴원환자 평균재원일수

OECD에서는 OECD 회원국의 퇴원환자 평균재원일수 비교를 위해 진단범주별로 각 회원국의 퇴원환자 평균재원일수를 조사하고 있으며, 비교 결과를 OECD.Stat 홈페이지(<https://stats.oecd.org>)에 공지하고 있다. OECD.Stat 홈페이지에 공지된 OECD 각 회원국의 퇴원환자 평균재원일수를 파악한 결과 우리나라의 경우 정신 및 행동 장애(F00-F99), 신경계통의 질환(G00-G99), 순환계통의 질환(I00-I99)의 순으로 퇴원환자의 평균재원일수가 높았다. 평균재원일수가 높은 3개 단위질환에 대해 평균재원일수가 가장 낮은 OECD 회원국의 퇴원환자 평균재원일수와 비교한 결과 정신 및 행동 장애(F00-F99)의 재원일수는 우리나라 145.7일, 벨기에 9.4일, 신경계통의 질환(G00-G99)의 재원일수는 우리나라 48.5일, 네덜란드 3.3일, 순환계통의 질환(I00-I99)의 재원일수는 우리나라 35.6일, 덴마크 4.8일로 30일 이상 차이를 보여 정신 및 행동 장애(F00-F99), 신경계통의 질환(G00-G99), 순환계통의 질환(I00-I99)의 재원일수 관리가 필요함을 알 수 있었다. 선행연구에 따르면 정신 및 행동 장애(F00-F99)는 진단명보다 정신과적 증상, 기능상태, 사회적지지 정도, 일반 대중의 정신장애인에 대한 인식의 차이 등이 재원기간에 더 큰 영향을 미친다고 보고하고 있다(최정규, 장홍석, 이명수, 홍진표, 박종익, 2013). 따라서 정신 및 행동 장애(F00-F99)의 재원일수 관리는 본 연구의 연구목적에 부합하지 않음에 따라 본 연구에서는 우선적으로 신경계통의 질환(G00-G99)을 대상으로 중증도 보정 재원일수 예측 모형을 개발하고 예측력을 평가하여 중증도 보정 재원일수 예측 모형 개발의 알고리즘을 제시하고자 한다.

### 2. 머신러닝 분석 기법

#### 가. 머신러닝 분석 기법

머신러닝은 기존 데이터를 최대한 활용하여 환경에 따른 데이터의 특징을 추출하고 테스트하여 최적화 또는 자가 발전을 하는 일련의 과정이다(이호현, 정승현, 최은정,

2016; 이길홍 등, 2017). 머신러닝은 학습 데이터(training data)의 기대 출력값인 종속 변수 또는 반응변수 포함 유무에 따라 지도학습(Supervised Learning), 자율학습(Unsupervised Learning) 등으로 구분된다(이호현, 정승현, 최은정, 2016). 지도학습은 정답을 알고 있는 샘플 데이터를 이용하여 머신러닝 모델을 만들어서 새로운 데이터가 주어졌을 때 예상치를 맞추거나 종류를 판별하는 등의 과정이다(이호현, 정승현, 최은정, 2016; 이길홍 등, 2017). 본 연구는 머신러닝을 이용하여 중증도 보정 재원일수 예측 모형 개발하는 것이며, 본 연구를 위해 수집된 자료에는 기대 출력값인 재원일수 자료가 포함되어 있음에 따라 머신러닝의 지도학습에서 사용되는 모델링 기법을 선택하여 분석을 수행하고자 한다. 지도학습에 사용되는 모델링 기법에는 매우 다양한 방법이 있지만(김현호, 양승범, 강연석, 박영배, 김재효, 2016) 본 연구에서는 외국 및 국내 선행연구(Hachesu, Ahmadi, Alizadeh & Sadoughi, 2013; Morton, Marzban, Giannoulis & Patel, 2014; 홍성욱, 김영택, 박종호, 강성홍, 2015; 최병관, 함승우, 김축환, 서정숙, 박명화, 강성홍, 2018)에서 중증도 보정 재원일수 예측 모형 개발을 위해 사용한 분석 기법인 회귀분석, 의사결정나무, 랜덤 포레스트, 서포트 벡터 회귀분석, 신경망 기법을 이용하여 분석하였다.

#### 나. 회귀분석(Regression Analysis)

회귀분석(Regression Analysis)은 주어진 데이터와 선택된 학습 모델에 의해 얻어진 예측값 간의 오차를 최소화하기 위한 반복적인 과정을 수행하면서 데이터들간의 관계를 모델링하는 분석기법을 의미한다(이근영, 2015). 회귀분석 알고리즘 중 선형회귀(linear regression)는 종속변수와 한 개 이상의 독립변수의 선형관계를 분석하는 기법으로 종속 변수가 연속형 변수인 경우 선형회귀를 이용하여 분석하며, 선형회귀는 모델링에 사용되는 독립변수의 개수에 따라 단순선형회귀(Simple linear regression), 다중선형회귀(Multiple linear regression) 기법으로 분류된다(정성호, 이대엽, 이경상, 2018). 선형회귀를 이용하여 모델링하는 경우 회귀계수의 추정방법에는 최소제곱법(least squares method), 최대우도추정법(maximum likelihood estimation) 등이 있으나 일반적으로 최소제곱법을 사용해 회귀 모델을 만들고 예측 결과를 분석한다(임현일, 2018). 최소제곱법은 회귀계수 추정을 위해 구성되는 모델의 오차(모델의 추론값과 실제값의 차이)를

최소화하기 위해서 오차의 제곱의 합을 최소화하는 방법으로 최소제곱법에서는 오차의 제곱을 최소화하는 모델을 분석함으로써 추론 결과의 오차가 실제 결과와 근사한 값으로 추론될 수 있는 모델을 찾는 것을 목표로 한다(임현일, 2018).

#### 다. 의사결정나무(Decision Tree)

의사결정나무(Decision Tree)는 데이터의 속성에 따라 나무 형태의 의사결정 학습모델을 만들고, 반복을 통해 주어진 문제에 대한 최종 결정을 도출하는 방법이다(이근영, 2015). 의사결정나무(Decision Tree)는 분석과정이 나무 형태의 그래프로 표현됨에 따라 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있으며(최종후, 서두성, 1999), 이와 같은 장점으로 인해 지도학습 문제에서 최종모형의 해석력이 더 중요한 경우 주로 의사결정나무를 활용한다(오미애, 최현수, 김수현, 장준혁, 진재현, 천미경, 2017). 의사결정나무는 종속변수의 범주형 또는 연속형 여부에 따라 분류나무(classification trees)와 회귀나무(regression trees) 모형으로 구분된다(오미애, 최현수, 김수현, 장준혁, 진재현, 천미경, 2017). 의사결정나무를 형성시키는 알고리즘은 대표적으로 CHAID(Chi-squared Automatic Interaction Detection), CART(Classification And Regression Trees) 등이 있다(박정선, 고영경, 2003).

#### 라. 랜덤 포레스트(Random Forest)

랜덤 포레스트(Random Forest)는 말 그대로 의사결정나무들이 많이 있는 모형이다. 랜덤 포레스트의 목적은 의사결정나무 모형을 다수 만들어 더 정확한 예측을 하는데 있다(유진은, 2015). 랜덤 포레스트는 학습 데이터로부터 부트스트랩(주어진 자료로부터 동일한 크기의 표본을 랜덤 복원 추출)을 통해 무작위 표본을 다수 생성하여 각 표본에 대한 독립적인 의사결정나무를 학습하고, 그 결과를 종합하여 최종 모형을 결정하는 의사결정나무 기반의 앙상블 방법(ensemble methods)이다(유진은, 2015; 오미애, 최현수, 김수현, 장준혁, 진재현, 천미경, 2017; 김보찬, 김유현, 김민정, 이종석, 2018). 의사결정나무 등 머신러닝 기법에서 모형의 타당화(model validation) 방법으로 자료를 훈련용(training)과 시험용(test)로 나누어 모형을 개발하고 평가하는 것과 달리 랜덤 포레

스트는 굳이 훈련용과 시험용으로 나누어 모형 타당화를 시도할 필요 없이 OOB(out-of-bag) 분석을 할 수 있다(유진은, 2015). 랜덤 포레스트에서 각 의사결정나무의 학습은 학습 데이터로부터 부트스트랩 표본을 추출하여 수행되므로, 시험용 자료 대신 부트스트랩 표본에 포함되지 않은 OOB 데이터를 활용하여 모형의 정확도를 계산할 수 있다(김보찬, 김유현, 김민정, 이종석, 2018). 랜덤 포레스트는 의사결정나무 모형에서와 같이 나무형태의 그래프가 최종모형으로 도출되지 않음에 따라 최종 결과에 대한 해석이 어렵다는 단점이 있다. 랜덤 포레스트에서는 이러한 문제를 해결하기 위한 방안으로 독립변수에 대한 상대적 중요도 지수(variable of importance index)를 제공함으로써 종속변수에 대한 독립변수의 중요도(영향력)를 숫자 또는 그래프로 알아볼 수 있도록 한다(유진은, 2015; 오미애, 최현수, 김수현, 장준혁, 진재현, 천미경, 2017).

#### 마. 서포트 벡터 회귀분석(Support Vector Regression)

서포트 벡터 회귀분석(Support Vector Regression)은 서포트 벡터 머신(Support Vector Machine)에 기반한 회귀 모형이다(Smola & Scholkopf, 2004; 김보찬, 김유현, 김민정, 이종석, 2018). 서포트 벡터 머신은 커널 함수를 이용하여 데이터를 고차원의 벡터 공간에 표현함으로써 입력 데이터들을 분류하는 거리(margin)를 최대화하여 가장 적합한 클래스로 분류하는 최적의 초평면(hyperplane)을 찾는다(장은나, 임정호, 하성현, 이상균, 박영규, 2016; 김보찬, 김유현, 김민정, 이종석, 2018). 서포트 벡터 회귀분석은 이러한 서포트 벡터 머신을 회귀분석에 적용할 수 있도록 확장한 것으로  $\epsilon$ -무감도 손실함수를 도입하여 임의의 실수값을 예측하는데 사용되며(서기성, 2017), 임의의 실수값들을 회귀함수를 사용하여 예측할 수 있다(김성진, 유은정, 정민규, 김재경, 안현철, 2012). 서포트 벡터 회귀분석은 비선형성 또는 복잡한 패턴의 데이터에 대해서도 높은 설명력을 가지는 장점이 있지만 계산의 복잡도가 높아 학습시간이 길고 독립변수와 종속변수의 직접적인 연관성을 해석할 수 없음에 따라 모형에 대한 해석이 쉽지 않다는 단점을 가진다(김보찬, 김유현, 김민정, 이종석, 2018). 서포트 벡터 회귀분석에서 데이터를 원래의 차원에 고차원으로 변환시켜 초평면을 보다 수월하게 찾기 위해 사용하는 커널함수(kernel function)는 일반적으로 선형(linear), 다항식(polynomial), 방사상 기반 함수(Radial Basis Function, RBF)를 사용한다(장은나, 임정호, 하성현, 이상균, 박영

규, 2016; 김선웅, 최홍식, 2017; 김보찬, 김유현, 김민정, 이종석, 2018).

## 바. 신경망(Neural Networks)

신경망(Neural Networks)은 인간의 뇌 신경세포의 특성을 수학적으로 표현한 모델링이다(정광훈 등, 2015). 신경망은 현재 많은 관심을 받고 있는 심층학습인 딥러닝의 기초가 된다. 신경망은 데이터의 비선형성을 잘 설명하며 복잡하고 불규칙적인 패턴에 대해서도 높은 설명력을 보여주는 장점이 있지만 종속변수와 독립변수의 직접적인 상관관계나 영향력의 크기를 측정하기 어려운 단점을 가지고 있다(김보찬, 김유현, 김민정, 이종석, 2018). 신경망은 외부 데이터가 입력층(input layer)에 들어오면 은닉층(hidden layer)을 거쳐 출력층(output layer)에서 결과값을 출력하게 되며(정광훈 등, 2015), 은닉층의 노드들은 이전 층으로부터 받은 수치들과 각각의 가중치(weight)들을 가중합(weighted sum)으로 입력받아 활성화함수(activation function)를 통과시켜 다음 층으로 보낸다. 이때 가중치는 여러 회귀모형과 마찬가지로 기본적으로 오차, 즉 실제 값과 신경망의 결과값 차이를 최소화하도록 결정된다(김보찬, 김유현, 김민정, 이종석, 2018). 신경망 학습의 주요 알고리즘에는 입력층과 출력층 사이에 한 개의 은닉층이 있는 단층 퍼셉트론(Single-Layer Perceptron)과 입력층과 출력층 사이에 다수의 은닉층이 있는 다층 퍼셉트론(Multi-Layer Perceptron) 등이 있다. 신경망 학습 알고리즘 중 가장 보편적으로 이용되는 알고리즘은 다층 퍼셉트론을 이용한 역전파 알고리즘(backpropagation algorithm)이다. 다층 퍼셉트론에 의한 학습은 역전파 알고리즘을 이용하여 수행되는데 외부의 입력값에 원하는 출력결과를 학습시키고자 할 때 사용하며, 출력층에서 발생된 출력오차를 역으로 전파하면서 연결링크의 가중치를 통한 수정 오차를 최소화시키는 것이다(정광훈 등, 2015).



### III. 연구방법

#### 1. 자료수집 및 분석대상

본 연구를 위해 질병관리본부로부터 2006년부터 2015년까지 10년간의 퇴원손상심층조사 자료 2,804,004건을 수집하였다. 퇴원손상심층조사는 미국의 퇴원환자조사 방법론을 근거로 100병상 이상의 급성기 병원 중 표본병원 약 150개를 대상으로 조사되고 있는 전국단위 행정자료이며(최병관, 함승우, 김축환, 서정숙, 박명화, 강성홍, 2018; 홍성욱, 김영택, 박종호, 강성홍, 2015), 퇴원손상심층조사의 조사항목에는 성, 연령, 입원경로, 주진단, 부진단 등 환자의 주요 진료내역 정보가 포함되어 있음에 따라 중증도 보정 재원일수 등을 개발하기 위한 가장 적합한 자료이다(홍성욱, 김영택, 박종호, 강성홍, 2015). OECD 국가별 퇴원환자 평균재원일수 진단범주 정의(<https://stats.oecd.org>) 및 세계 보건 기구(WHO)의 국제질병사인분류(The International Statistical Classification of Diseases and Related Health Problems, 이하 ICD) 10차 버전에서 정의한 진단군 대분류 기준에 따라 주진단이 G00~G99인 퇴원환자를 신경계통의 질환 퇴원환자로 정의하였다. 수집된 퇴원손상심층조사 자료 2,804,004건 중 주진단이 G00~G99인 신경계통의 질환 퇴원환자 자료 50,228건을 추출하였으며, 이 중 입원경로가 기타 또는 미상인 퇴원환자 자료 168건을 제외한 50,060건을 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형 개발을 위한 분석대상자로 정의하였다.

#### 2. 변수정의

##### 가. 변수정의

신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형의 독립변수는 문헌고찰 결과와 퇴원손상심층조사 자료에서 활용할 수 있는 변수를 고려하여 부진단을 이용한 동반상병 보정 방법에 따른 동반상병 보정 결과, 주진단, 성, 연령, 진료비 지불방법, 입원경로, 수술유무, 사망유무 등을 독립변수로 정의하였다. 신경계통의 질환 퇴원환자

의 주진단 범주는 OECD 국가별 퇴원환자 평균재원일수 진단범주 정의 (<https://stats.oecd.org>)에 따라 알츠하이머병(퇴행성 뇌질환)(G30), 다발성경화증(G35), 간질(G40, G41), 일과성 대뇌허혈성 발작 및 관련증후군(G45), 기타 신경계 질환(기타 G00-G99)으로 정의하였다. 성, 연령, 진료비 지불방법, 입원경로, 수술유무, 사망유무의 정의는 선행연구의 연구방법(홍성욱, 김영택, 박종호, 강성홍, 2015; 최병관, 함승우, 김축환, 서정숙, 박명화, 강성홍, 2018)에 따라 정의하였다. 연령의 경우 29세 이하, 30세에서 49세, 50세에서 64세, 65세 이상, 진료비 지불방법의 경우 국민건강보험, 의료급여, 기타로 그룹화하여 분석에 활용하였으며, 입원경로의 경우 응급, 외래, 수술유무의 경우 수술 시행여부에 따라 수술 유와 무, 사망유무 경우 퇴원환자의 사망유무에 따라 사망 유와 무로 구분하여 정의하였다.

#### 나. 동반상병 보정 방법의 정의

신경계통의 질환 퇴원환자의 재원일수에 대한 맞춤형 동반상병을 보정방법을 개발하고, 이의 타당성을 입증하기 위해 선행연구(강성홍, 석향숙, 김원중, 2013; 홍성욱, 김영택, 박종호, 강성홍, 2015)의 연구방법에 따라 CCI, ECI와 CCS 진단군 분류 변수를 이용하여 각기 신경계통의 질환 퇴원환자의 증증도를 보정하였다. CCI는 뇌혈관 질환, 울혈성 심부전 등 퇴원환자의 진료결과에 영향을 미친다고 정의한 17개의 동반 상병에 대해 1점부터 6점까지 가중치를 부여한 뒤 이 가중치의 합(comorbidity score)을 이용하여 퇴원환자의 증증도를 보정하는 방법(Charlson, Pompei, Ales & MacKenzie, 1987)이고, ECI는 울혈성 심부전, 심장 부정맥 등 퇴원환자의 진료결과에 영향을 미친다고 정의한 31개의 동반상병에 대해 상병의 동반 유무 즉, 이분형 변수로 처리하여 퇴원환자의 증증도를 보정하는 방법(Elixhauser, Steiner, Harris & Coffey, 1998)이다. CCS 진단군 분류 변수를 이용한 동반상병 보정은 미국 AHRQ(The Agency for Healthcare Research and Quality)에서 개발한 모든 질병을 임상학적으로 근집화한 방법인 CCS 진단군을 이용하여 259개의 동반상병을 분류하고 ECI 방법처럼 각각의 259개 상병의 동반 여부를 동반 유무 즉, 이분형 변수로 처리하여 환자의 증증도를 보정한 방법이다(질병관리본부, 2012; 질병관리본부, 2014; 홍성욱, 김영택, 박종호, 강성홍, 2015). 기존에 개발된 CCI, ECI를 이용한 동반상병 보정의 경우 퇴원환자의 진료결과에 영향을

미친다고 정의한 일부 동반상병만 동반상병 보정에 고려되는 것에 비해 CCS 진단군 변수를 이용한 동반상병 보정은 퇴원환자의 모든 동반상병이 동반상병 보정에 고려됨에 따라 신경계통의 질환 퇴원환자의 재원일수에 영향을 미치는 동반상병이 동반상병 보정 방법에 있어 제외되는 경우는 없다.

### 3. 분석방법

본 연구는 오픈소스 머신러닝 분석 도구인 Knime 3.5.3을 이용하여 분석자료의 전처리, 일반적 특성 및 중증도 보정 재원일수 예측 모형 개발 및 평가를 실시하였다.

독립변수는 변수정의에 따라 독립변수를 그룹화하여 일반적 특성 분석에 활용하였으며, 중증도 보정 재원일수 예측 모형 개발 시에는 더미형으로 변환 후 분석에 활용하였다. 단, 의사결정나무, 랜덤 포레스트의 경우 더미형으로 변환하지 않아도 모형을 개발하고 평가하는 것이 가능함에 따라 그룹화된 독립변수 그대로를 분석에 활용하였다. 종속변수인 재원일수의 경우 일반적 특성 분석에는 그대로 활용한 반면, 중증도 보정 재원일수 예측 모형을 개발하고 평가할 시에는 선행연구의 연구방법에 따라 전반적인 모형의 성능 향상과 해석상의 용이함을 위해 로그형을 취한 뒤 활용하였다(김보찬, 김유현, 김민정, 이종석, 2018). 중증도 보정 재원일수 예측 모형 개발 및 평가를 위해 선행연구의 연구방법(박일수, 용왕식, 김유미, 강성홍, 한준태, 2008; 김보찬, 김유현, 김민정, 이종석, 2018)에 따라 2006년, 2008년, 2010년, 2012년, 2014년 자료는 모형개발 및 내적검증을 위한 자료로 2007년, 2009년, 2011년, 2013년, 2015년 자료는 외적검증을 위한 자료로 활용하였으며, 모형개발 및 내적검증을 위한 자료는 모형개발을 위해 60%, 내적검증을 위해 40%로 구분하여 분석에 활용하였다.

분석대상자의 일반적 특성을 파악하기 위해 빈도분석을 실시하였다. 중증도 보정 재원일수 예측 모형 개발 시 신경계통의 질환 퇴원환자의 중증도는 CCI, ECI, CCS 진단군 분류 변수를 이용한 동반상병 보정 방법을 통해 보정하였다. 맞춤형 동반상병 보정방법의 타당성을 제시하기 위해 중증도 보정 재원일수 예측 모형을 개발하고 평가할 시 동반상병 보정방법 뿐 아니라 퇴원환자의 주진단, 성, 연령, 진료비 지불방법, 입원경로, 수술 유무, 사망유무를 단계적(1단계 : 동반상병 보정방법, 2단계: 동반상병 보정방법, 주진단, 3단계: 동반상병 보정방법, 주진단, 성, 연령, 4단계: 동반상병 보정방법, 주진단,

성, 연령, 진료비 지불방법, 입원경로, 수술유무, 사망유무)으로 고려하여 증증도 보정 재원일수 예측 모형을 개발하고 평가하였다. 머신러닝 분석 기법은 회귀분석, 의사결정 나무, 랜덤 포레스트, 서포트 벡터 회귀분석, 신경망 기법을 활용하였다. 머신러닝 기법 중 회귀분석은 최소제곱법을 이용한 회귀계수 추정 방법을 적용하여 다중선형회귀 모델을 구성하였으며, 의사결정나무는 회귀나무 모형에서 ANOVA (analysis of variance)의 F-검증을 이용하여 분리(split)를 수행하는 알고리즘(최종후, 서두성, 1999; 신혜정, 이윤동, 이은경, 2016)인 CHAID 알고리즘을 이용하여 의사결정나무 모형을 개발하였다. 정지규칙으로는 최대나무깊이는 10, 최소 케이스 수는 parent node 5, child node 1로 설정하였으며 가지치기(pruning)를 병행하였다. 랜덤 포레스트는 의사결정나무 모델 개발의 수를 100으로 하였으며, 독립변수에 대한 상대적 중요도 지수를 산출하였다. 서포트 벡터 회귀분석은 커널함수 중 가장 기본이 되는 선형 커널함수를 이용해서 분석하였으며, 회귀모형의 일반화를 결정짓는 모수인  $c$ 는 15.0 정밀 모수인  $\epsilon$ -무감도 손실함수의 값은 0.001로 설정하여 분석하였다. 신경망은 모형 개발 시 다층 퍼셉트론을 이용한 역전파 알고리즘을 사용하며, 은닉층 개수를 3개, 뉴런(연결선)의 개수를 10개로 설정하였다. 동반상병 보정방법 3가지, 단계적 적용 4단계, 머신러닝 분석 기법 5가지에 따라 본 연구에서 개발하고 평가한 신경계통의 질환 퇴원환자의 증증도 보정 재원일수 예측 모형은 총 60개이다. 본 연구에서 개발된 신경계통의 질환 퇴원환자의 증증도 보정 재원일수 예측 모형의 타당성과 예측력을 평가하고 검증하기 위해 모형 설명력을 나타내는 R-square와 모형의 예측력을 평가하는 Mean Absolute Error, Root Mean Squared Error 값을 이용하였다.

## IV. 연구결과

### 1. 분석대상자의 일반적 특성

#### 가. 일반적 특성 분석을 위한 KNIME 워크플로우

신경계통의 질환 퇴원환자의 일반적 특성 분석을 위한 본 연구의 분석도구인 KNIME의 일반적 특성 분석 워크플로우는 [그림 1]과 같다. 일반적 특성 분석을 위해 퇴원손상 심층조사의 신경계통의 퇴원환자 데이터, CCI, ECI, CCS 진단군 분류 기준 동반상병 보정 방법에 따른 동반상병 보정 결과 데이터를 불러온 다음 각 데이터의 조인, 독립변수 그룹화 등 분석을 위한 데이터 전처리를 실시하였으며, 최종 빈도분석을 실시하여 신경계통의 질환 퇴원환자의 일반적 특성 결과를 확인하였다.

#### 나. 신경계통의 질환 퇴원환자의 일반적 특성

본 연구의 분석대상인 신경계통의 질환 퇴원환자의 일반적 특성을 분석한 결과는 표 1과 같다. 일반적 특성을 분석한 결과 모형개발 및 모형평가(Internal), 모형평가(External) 모두 성별로는 남자보다 여자의 비율이 높았으며, 연령별로는 29세 이하의 비율이 가장 높았다. 진료비 지불방법별로는 국민건강보험의 비율이 각각 87.0%로 가장 높았으며, 입원경로별로는 응급 각각 36.1%, 34.0%, 외래 각각 63.9%, 66.0%로 나타났다. 수술환자의 비율은 각각 16.2%, 16.9%였으며, 사망환자의 비율은 각각 0.9%로 조사되었다. 주진단별로는 기타 신경계 질환(기타 G00-G99)이 각각 74.7%, 74.5%로 가장 높았으며, 간질(G40, G41) 각각 12.9%, 12.4%, 일과성 대뇌허혈성 발작 및 관련증후군(G45) 각각 10.2%, 10.5% 등의 순으로 높은 것으로 나타났다.

그림 1. 신경계통의 질환 퇴원환자의 일반적 특성 분석 KNIME WorkFlow

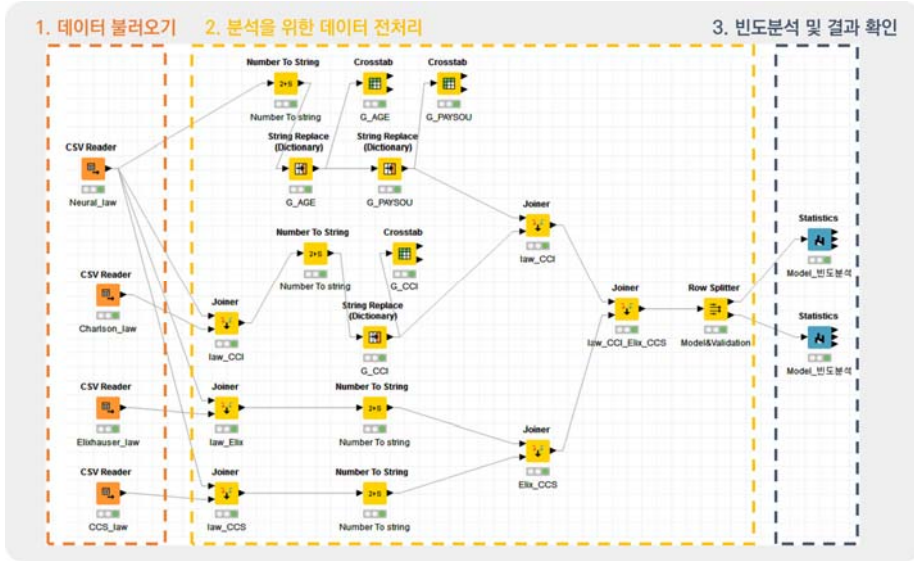


표 1. 신경계통의 질환 퇴원환자의 일반적 특성

(단위: 건, %)

구분		모형개발 및 모형평가(Internal)		모형평가(External)	
		건수	비율	건수	비율
성별	남자	12,537	49.5	12,016	48.6
	여자	12,810	50.5	12,697	51.4
연령	29세 이하	7,747	30.6	6,858	27.8
	30~49세	5,391	21.3	5,337	21.6
	50~64세	5,866	23.1	5,949	24.1
	65세 이상	6,343	25.0	6,569	26.6
진료비 지불방법	국민건강보험	22,047	87.0	21,500	87.0
	의료급여	2,499	9.9	2,402	9.7
	기타	801	3.2	811	3.3
입원경로	응급	9,161	36.1	8,401	34.0
	외래	16,186	63.9	16,312	66.0
수술유무	수술 무	21,253	83.8	20,545	83.1
	수술 유	4,094	16.2	4,168	16.9

구분		모형개발 및 모형평가(Internal)		모형평가(External)	
		건수	비율	건수	비율
사망유무	사망 무	25,115	99.1	24,487	99.1
	사망 유	232	0.9	226	0.9
주진단	알츠하이머병(퇴행성 뇌질환)(G30)	400	1.6	475	1.9
	다발성경화증(G35)	179	0.7	164	0.7
	간질(G40, G41)	3,260	12.9	3,072	12.4
	일과성 대뇌허혈성 발작 및 관련 증후군(G45)	2,580	10.2	2,600	10.5
	기타 신경계 질환(기타 G00-G99)	18,928	74.7	18,402	74.5
전체		25,347	100.0	24,713	100.0

#### 다. 신경계통의 질환 퇴원환자의 동반상병 특성

##### 1) CCI

CCI 동반상병 보정 방법을 이용하여 신경계통의 질환 퇴원환자의 중증도를 보정한 결과는 표 2와 같다. 중증도 보정 결과 모형개발 및 모형평가(Internal), 모형평가(External) 모두 0점이 각각 76.2%, 75.3%로 가장 높게 나타났으며, 1점 각각 15.5%, 16.1%, 2점 각각 5.7%, 5.8%, 3점 이상 각각 2.7%, 2.8%의 순으로 높게 조사되었다.

표 2. 신경계통의 질환 퇴원환자의 CCI 분포

(단위: 건, %)

구분	모형개발 및 모형평가(Internal)		모형평가(External)	
	건수	비율	건수	비율
0점	19,303	76.2	18,617	75.3
1점	3,919	15.5	3,980	16.1
2점	1,436	5.7	1,422	5.8
3점 이상	689	2.7	694	2.8
전체	25,347	100.0	24,713	100.0

2) ECI

ECI 동반상병 보정 방법을 이용하여 신경계통의 질환 퇴원환자의 중증도를 보정한 결과는 표 3과 같다. 중증도 보정 결과 모형개발 및 모형평가(Internal), 모형평가(External) 모두 비합병증성 고혈압의 동반 유가 각각 16.1%, 16.3%로 가장 높았으며, 비합병증성 당뇨병의 동반 유 각각 8.0%, 8.4%, 기타 신경 장애의 동반 유 각각 5.1%, 5.0%, 우울증의 동반 유 각각 2.1%, 2.5%, 심장 부정맥의 동반 유 각각 1.5%, 1.7% 등의 순으로 동반 유의 비율이 높게 조사되었다.

표 3. 신경계통의 질환 퇴원환자의 ECI 분포

(단위: 건, %)

구분	모형개발 및 모형평가(Internal)			모형평가(External)		
	동반 유		전체 건수	동반 유		전체 건수
	건수	비율		건수	비율	
심장 부정맥	368	1.5	25,347	427	1.7	24,713
울혈성 심부전	164	0.6	25,347	175	0.7	24,713
만성 폐질환	252	1.0	25,347	296	1.2	24,713
합병증성 당뇨병	297	1.2	25,347	283	1.1	24,713
우울증	530	2.1	25,347	610	2.5	24,713
비합병증성 당뇨병	2,030	8.0	25,347	2,071	8.4	24,713
체액 및 전해질 이상	247	1.0	25,347	220	0.9	24,713
갑상선 기능 저하증	165	0.7	25,347	184	0.7	24,713
비합병증성 고혈압	4,074	16.1	25,347	4,037	16.3	24,713
간 질환	306	1.2	25,347	308	1.2	24,713
기타 신경 장애	1,294	5.1	25,347	1,248	5.0	24,713
마비	305	1.2	25,347	319	1.3	24,713
신부전	241	1.0	25,347	233	0.9	24,713
비전이성 고형 압	260	1.0	25,347	248	1.0	24,713

주: 전체 31개 질환 중 모형개발 및 모형평가(Internal & External) 질환의 유가 100건 이상인 질환을 주요 동반질환으로 하여 일반적 특성을 제시함. 모형개발 시에는 전체 31개 질환 모두를 포함함



### 3) CCS 진단군 분류 기준

CCS 진단군 분류 기준 동반상병 보정 방법을 이용하여 신경계통의 질환 퇴원환자의 중증도를 보정한 결과는 <표 4>와 같다. 중증도 보정 결과 모형개발 및 모형평가 (Internal), 모형평가(External) 모두 본태성 고혈압의 동반 유가 각각 16.1%, 16.3%로 가장 높았으며, 합병증이 없는 당뇨병의 동반 유가 각각 8.1%, 8.5%, 기타 신경계 장애의 동반 유가 각각 6.0%, 6.5%, 뇌혈관 질환의 후유증의 동반 유가 각각 5.5%, 5.6%, 노인성 장애 및 기질적 정신장애의 동반 유가 각각 4.2%, 5.0%, 급성 뇌혈관 질환의 동반 유가 각각 3.9%, 3.7%등의 순으로 동반 유의 비율이 높게 조사되었다.

표 4. 신경계통의 질환 퇴원환자의 CCS 진단군 분류 기준 동반상병 분포

(단위: 건, %)

구분	모형개발 및 모형평가(InternaI)		모형평가(External)			
	동반 유		전체 건수	동반 유		전체 건수
	건수	비율		건수	비율	
합병증이 없는 당뇨병	2,057	8.1	25,347	2,091	8.5	24,713
지질 대사 장애	735	2.9	25,347	897	3.6	24,713
노인성 장애 및 기질적 정신장애	1,055	4.2	25,347	1,241	5.0	24,713
정동장애	543	2.1	25,347	626	2.5	24,713
불안, 신체형, 해리성, 성격장애	364	1.4	25,347	425	1.7	24,713
기타 유전성 및 퇴행성 신경계 조건	456	1.8	25,347	446	1.8	24,713
마비	398	1.6	25,347	395	1.6	24,713
간질; 경련	679	2.7	25,347	656	2.7	24,713
두통; 편두통 포함	461	1.8	25,347	493	2.0	24,713
어지럼증이나 현기증과 관련된 증상	477	1.9	25,347	504	2.0	24,713
기타 신경계 장애	1,515	6.0	25,347	1,615	6.5	24,713
본태성고혈압	4,074	16.1	25,347	4,037	16.3	24,713
심장 부정맥	361	1.4	25,347	422	1.7	24,713
급성 뇌혈관 질환	977	3.9	25,347	906	3.7	24,713
뇌혈관 질환의 후유증	1,395	5.5	25,347	1,383	5.6	24,713
폐렴(결핵이나 성병에 의한 경우 제외)	534	2.1	25,347	535	2.2	24,713
급성 및 만성 편도염	352	1.4	25,347	316	1.3	24,713

구분	모형개발 및 모형평가(Internal)			모형평가(External)		
	동반 유		전체	동반 유		전체
	건수	비율	건수	건수	비율	건수
기타 상부 호흡기 감염	666	2.6	25,347	560	2.3	24,713
기타 상부 호흡기 질환	483	1.9	25,347	480	1.9	24,713
식도 장애	384	1.5	25,347	427	1.7	24,713
위염 및 십이지장염	712	2.8	25,347	708	2.9	24,713
기타 위장 장애	615	2.4	25,347	701	2.8	24,713
요로 감염	546	2.2	25,347	498	2.0	24,713
방광과 요도의 다른 질병	431	1.7	25,347	397	1.6	24,713
척추증, 추간판질환, 기타 등 문제	759	3.0	25,347	875	3.5	24,713
기타 결합 조직 질환	693	2.7	25,347	730	3.0	24,713

주: 전체 259개 질환 중 모형개발 및 모형평가(Internal & External) 질환의 수가 300건 이상인 질환을 주요 동반질환으로 하여 일반적 특성을 제시함. 모형개발 시에는 전체 259개 질환 모두를 포함함

## 2. 중증도 보정 재원일수 예측 모형 개발 및 평가

### 가. 머신러닝을 이용한 중증도 보정 재원일수 예측 모형 개발 및 평가 워크플로우

머신러닝 분석 기법인 회귀분석, 의사결정나무, 랜덤 포레스트, 서포트 벡터 회귀분석, 신경망 및 동반상병 보정방법인 CCI, ECI, CCS 진단군 분류 기준을 이용하여 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형을 개발하고, 개발된 모형의 예측력을 평가하기 위한 본 연구의 분석도구인 KNIME의 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형 개발 및 평가의 워크플로우는 그림 2와 같다. 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형을 개발하고, 개발된 모형의 예측력을 평가하기 위해 퇴원손상심층조사의 신경계통의 퇴원환자 데이터, CCI, ECI, CCS 진단군 분류 기준 동반상병 보정 방법에 따른 동반상병 보정 결과 데이터를 불러온 다음 각 데이터의 조인, 독립변수 그룹화, 독립변수의 더미(dummy) 변수 처리 등 분석을 위한 데이터 전처리를 실시하였으며, 회귀분석, 의사결정나무, 랜덤 포레스트, 서포트 벡터 회귀분석, 신경망을 이용하여 신경계통의 질환 퇴원환자의 중증도 보정 재원일수

예측 모형을 개발하고, 개발된 모형의 예측력을 평가한 후 그 결과를 최종 확인하였다. 동반상병 보정방법별로 신경계통의 퇴원환자 데이터와 CCI, 신경계통의 퇴원환자 데이터와 ECI, 신경계통의 퇴원환자 데이터와 CCS 진단군 분류 기준 동반상병 보정 방법을 각각 조인하여 데이터 전처리를 실시하였으며, 머신러닝 분석 기법을 이용하여 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형을 개발하고, 개발된 모형의 예측력을 평가한 후 그 결과를 최종 확인하였다.

그림 2. 머신러닝을 이용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형 개발 및 평가의 KNIME WorkFlow



## 나. 회귀분석을 이용한 중증도 보정 재원일수 예측 모형 개발 및 평가

### 1) 모형 개발 및 모형의 예측력 평가

회귀분석을 이용하여 동반상병 보정 방법 및 독립변수의 단계적 고려에 따라 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형을 개발하고 평가한 결과는 표 5와 같다. 개발된 모형의 모형 설명력(R-square) 및 Mean Absolute Error 값을 평가한 결과 1단계, 2단계, 3단계, 4단계 모두 CCS 진단군 분류 기준 동반상병 보정 방법을

이용하여 개발된 모형이 CCI 및 ECI 동반상병 보정 방법을 이용하여 개발된 모형보다 내적(Internal), 외적(External) 모형평가 모두에서 모형 설명력(R-square)이 높았으며, Mean Absolute Error 값, Root Mean Squared Error 값이 낮은 것으로 나타났다. 본 연구에서 회귀분석을 이용하여 동반상병 보정 방법 3가지 및 독립변수의 4단계의 고려에 따라 개발된 12개 모형 중 CCS 진단군 분류 기준 동반상병 보정 방법 및 주진단, 성, 연령, 진료비 지불방법, 입원경로, 수술유무, 사망유무 모두를 이용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형의 내적(Internal), 외적(External) 모형평가의 모형 설명력(R-square)이 각각 29.3%, 30.7%로 가장 높았으며, Mean Absolute Error는 각각 0.079, 0.080, Root Mean Squared Error는 각각 0.105, 0.106으로 가장 낮은 것으로 나타나 모형의 예측력이 가장 우수한 것으로 조사되었다.

표 5. 회귀분석을 이용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형의 예측력 평가

구분	내적(Internal) 모형평가			외적(External) 모형평가		
	R square	MAE	RMSE	R square	MAE	RMSE
1 CCI	0.105	0.113	0.143	0.104	0.115	0.143
단 ECI	0.086	0.113	0.142	0.096	0.115	0.144
계 CCS	0.261	0.082	0.108	0.257	0.083	0.109
2 CCI + 주진단	0.125	0.112	0.141	0.123	0.114	0.142
단 ECI + 주진단	0.103	0.111	0.141	0.115	0.114	0.143
계 CCS + 주진단	0.272	0.081	0.107	0.265	0.082	0.108
3 CCI + 주진단 + 성 + 연령	0.132	0.111	0.140	0.130	0.113	0.141
단 ECI + 주진단 + 성 + 연령	0.117	0.110	0.140	0.130	0.113	0.141
계 CCS + 주진단 + 성 + 연령	0.275	0.080	0.106	0.268	0.080	0.107
CCI + 주진단 + 성 + 연령 + 진료비 지불방법 + 입원경로 + 수술유무 + 사망유무	0.175	0.088	0.117	0.170	0.090	0.118
4 ECI + 주진단 + 성 + 연령 + 진료비 지불방법 + 입원경로 + 수술유무 + 사망유무	0.157	0.088	0.117	0.168	0.090	0.118
계 CCS + 주진단 + 성 + 연령 + 진료비 지불방법 + 입원경로 + 수술유무 + 사망유무	0.293	0.079	0.105	0.307	0.080	0.106

주: MAE는 Mean Absolute Error, RMSE는 Root Mean Squared Error를 의미함

2) 모형의 결과

CCS 진단군 분류 기준 동반상병 보정 방법 및 주진단, 성, 연령, 진료비 지불방법, 입원경로, 수술유무, 사망유무 모두를 독립변수로 활용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 회귀모형의 결과는 표 6과 같다. 회귀모형 결과 신경계통의 질환 퇴원환자의 재원일수에 통계적으로 유의한 영향을 미치는 요인은 성별, 주진단, 입원경로, 수술유무, 사망유무, 연령, 진료비 지불방법, 세균 감염, 지정되지 않은 부위의 동반 유무, 진균증의 동반 유무, 전립선 암의 동반 유무 등으로 나타났다. 성별로는 여자 보다 남자의 재원일수가 높았으며, 주진단별로는 알츠하이머병(퇴행성 뇌질환)(G30)의 재원일수가 가장 높았다. 입원경로별로는 외래보다 응급의 재원일수가 높았으며, 수술 유무별, 사망유무별로는 각각 수술을 시행한 퇴원환자, 사망한 퇴원환자의 재원일수가 높은 것으로 나타났다. 진료비 지불방법별로는 기타의 재원일수가 가장 높았으며, 세균 감염, 지정되지 않은 부위, 진균증, 전립선 암 등 동반상병의 유무별로는 동반상병이 있는 퇴원환자의 재원일수가 동반상병이 없는 퇴원환자의 재원일수보다 높은 것으로 조사되었다.

표 6. 회귀분석을 이용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형의 결과

변수		회귀 계수	표준 오차	t값	p값
상수		11.018	2.131	5.171	0.000
성별(기준: 남)	여	-1.234	0.453	-2.727	0.006
	다발성경화증(G35)	-5.208	3.295	-1.580	0.114
	간질(G40, G41)	-7.988	2.152	-3.713	0.000
주진단(기준: 알츠하이머병 (퇴행성 뇌질환)(G30))	일과성 대뇌 허혈성 발작 및 관련 증후군(G45)	-10.342	2.160	-4.788	0.000
	기타 신경계 질환 (기타 G00-G99)	-4.189	2.050	-2.043	0.041
입원경로(기준: 응급)	외래	-2.017	0.492	-4.101	0.000
수술유무(기준: 수술 무)	수술 유	3.984	0.662	6.015	0.000
사망유무(기준: 사망 무)	사망 유	6.533	2.405	2.716	0.007
연령(기준: 29세 이하)	30세~49세	1.550	0.656	2.365	0.018

변수		회귀 계수	표준 오차	t값	p값
	50세~64세	1.961	0.677	2.896	0.004
	65세 이상	0.889	0.733	1.212	0.225
진료비 지불방법(기준 : 국민건강보험공단)	의료급여	4.766	0.736	6.476	0.000
	기타	10.336	1.353	7.640	0.000
세균 감염, 지정되지 않은 부위(기준: 동반 무)	동반 유	19.123	2.219	8.617	0.000
진균증(기준: 동반 무)	동반 유	25.559	2.878	8.881	0.000
전립선암(기준: 동반 무)	동반 유	13.593	6.776	2.006	0.045
뇌 및 신경계 암(기준: 동반 무)	동반 유	19.705	5.573	3.536	0.000
기타의 정신 상태(기준: 동반 무)	동반 유	16.315	5.144	3.171	0.002
기타 중추 신경계 감염 및 소아마비(기준: 동반 무)	동반 유	11.456	5.245	2.184	0.029
마비(기준: 동반 무)	동반 유	12.163	1.749	6.953	0.000
혼수, 혼미, 뇌손상(기준: 동반 무)	동반 유	43.702	3.484	12.543	0.000
급성 심근경색(기준: 동반 무)	동반 유	51.188	6.954	7.361	0.000
혈전증(기준: 동반 무)	동반 유	22.936	5.571	4.117	0.000
흉막염, 기흉, 폐붕괴(기준: 동반 무)	동반 유	23.670	3.395	6.973	0.000
항문 및 직장 상태(기준: 동반 무)	동반 유	60.270	7.034	8.568	0.000
복막염과 장 농양(기준: 동반 무)	동반 유	105.653	12.599	8.386	0.000
요로감염(기준: 동반 무)	동반 유	15.900	1.595	9.967	0.000
피부의 만성 궤양(기준: 동반 무)	동반 유	27.443	2.371	11.572	0.000

## 다. 의사결정나무를 이용한 중증도 보정 재원일수 예측 모형 개발 및 평가

### 1) 모형 개발 및 모형의 예측력 평가

의사결정나무를 이용하여 동반상병 보정 방법 및 독립변수의 단계적 고려에 따라 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형을 개발하고 평가한 결과는 표 7과 같다. 개발된 모형의 모형 설명력(R-square) 및 Mean Absolute Error 값을 평가한 결과 1단계, 2단계, 3단계, 4단계 모두 CCS 진단군 분류 기준 동반상병 보정 방법을 이용하여 개발된 모형이 CCI 및 ECI 동반상병 보정 방법을 이용하여 개발된

모형보다 내적(Internal), 외적(External) 모형평가 모두에서 모형 설명력(R-square)이 높았으며, Mean Absolute Error 값, Root Mean Squared Error 값이 낮은 것으로 나타났다. 본 연구에서 의사결정나무를 이용하여 동반상병 보정 방법 3가지 및 독립변수의 4단계의 고려에 따라 개발된 12개 모형 중 CCS 진단군 분류 기준 동반상병 보정 방법 및 주진단, 성, 연령, 진료비 지불방법, 입원경로, 수술유무, 사망유무 모두를 이용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형의 내적(Internal), 외적(External) 모형평가의 모형 설명력(R-square)이 각각 18.8%, 16.4%로 가장 높았으며, Mean Absolute Error는 각각 0.086, 0.089, Root Mean Squared Error는 각각 0.116, 0.119로 가장 낮은 것으로 나타나 모형의 예측력이 가장 우수한 것으로 조사되었다.

표 7. 의사결정나무를 이용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형의 예측력 평가

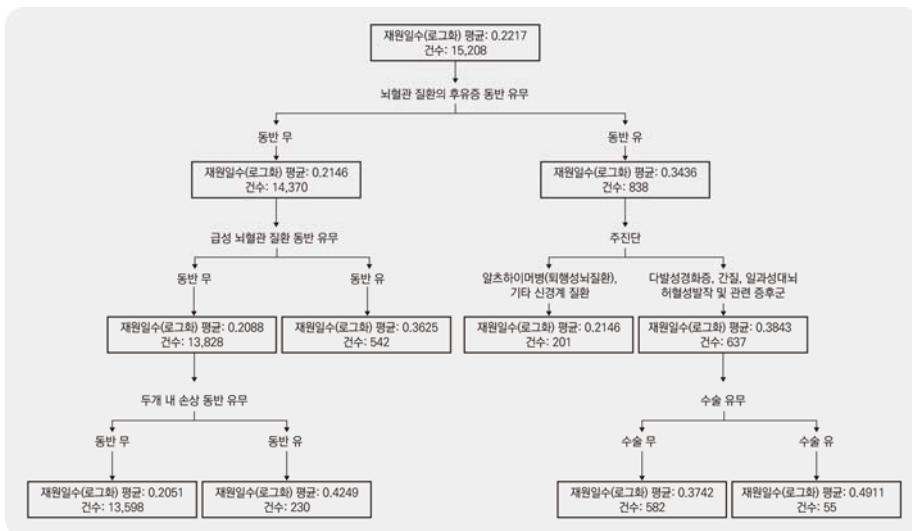
구분	내적(Internal) 모형평가			외적(External) 모형평가			
	R square	MAE	RMSE	R square	MAE	RMSE	
1 단 계	CCI	0.105	0.113	0.143	0.104	0.115	0.143
	ECI	0.075	0.113	0.143	0.083	0.116	0.145
	CCS	0.137	0.088	0.119	0.132	0.090	0.120
2 단 계	CCI + 주진단	0.135	0.111	0.140	0.131	0.113	0.141
	ECI + 주진단	0.087	0.112	0.142	0.104	0.114	0.143
	CCS + 주진단	0.141	0.088	0.118	0.138	0.089	0.119
3 단 계	CCI + 주진단 + 성 + 연령	0.094	0.090	0.122	0.075	0.092	0.124
	ECI + 주진단 + 성 + 연령	0.068	0.113	0.144	0.084	0.115	0.145
	CCS + 주진단 + 성 + 연령	0.147	0.088	0.109	0.136	0.088	0.110
4 단 계	CCI + 주진단 + 성 + 연령 + 진료비 지불방법 + 입원경로 + 수술유무 + 사망유무	0.093	0.090	0.122	0.099	0.093	0.124
	ECI + 주진단 + 성 + 연령 + 진료비 지불방법 + 입원경로 + 수술유무 + 사망유무	0.065	0.091	0.124	0.089	0.092	0.125
	CCS + 주진단 + 성 + 연령 + 진료비 지불방법 + 입원경로 + 수술유무 + 사망유무	0.188	0.086	0.116	0.164	0.089	0.119

주: MAE는 Mean Absolute Error, RMSE는 Root Mean Squared Error를 의미함

2) 모형의 결과

CCS 진단군 분류 기준 동반상병 보정 방법 및 주진단, 성, 연령, 진료비 지불방법, 입원경로, 수술유무, 사망유무 모두를 독립변수로 활용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 의사결정나무 모형의 결과는 그림 3과 같다. 의사결정나무 모형 결과 신경계통의 질환 퇴원환자의 재원일수에 가장 큰 영향을 미치는 요인은 뇌혈관 질환의 후유증 동반 유무로 나타났으며, 뇌혈관 질환의 후유증 동반 유무 뿐 아니라 급성 뇌혈관 질환 동반 유무, 주진단, 두 개 내 손상 동반 유무, 수술 유무 등이 신경계통의 질환 퇴원환자의 재원일수에 영향을 미치는 요인으로 나타났다. 뇌혈관 질환의 후유증을 동반한 퇴원환자의 경우 동반하지 않은 환자보다 재원일수가 높았으며, 뇌혈관 질환의 후유증을 동반하지 않은 퇴원환자군에서는 급성 뇌혈관 질환, 두 개 내 손상 동반한 퇴원환자의 경우 동반하지 않은 환자보다 재원일수가 높았다. 뇌혈관 질환의 후유증을 동반한 퇴원환자군에서는 주진단이 다발성경화증(G35), 간질(G40, G41), 일과성 대뇌허혈성 발작 및 관련증후군(G45)인 퇴원환자의 재원일수가 높은 것으로 나타났으며, 다발성경화증(G35), 간질(G40, G41), 일과성 대뇌허혈성 발작 및 관련증후

그림 3. 의사결정나무를 이용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형 결과





군(G45)인 퇴원환자군에서는 수술을 시행한 퇴원환자보다 수술을 시행하지 않은 퇴원환자의 재원일수가 높은 것으로 조사되었다.

## 라. 랜덤포레스트를 이용한 중증도 보정 재원일수 예측 모형 개발 및 평가

### 1) 모형 개발 및 모형의 예측력 평가

랜덤포레스트를 이용하여 동반상병 보정 방법 및 독립변수의 단계적 고려에 따라 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형을 개발하고 평가한 결과는 <표 8>과 같다. 개발된 모형의 모형 설명력(R-square) 및 Mean Absolute Error 값을 평가한 결과 1단계, 2단계, 3단계, 4단계 모두 CCS 진단군 분류 기준 동반상병 보정 방법을 이용하여 개발된 모형이 CCI 및 ECI 동반상병 보정 방법을 이용하여 개발된 모형보다 내적(Internal), 외적(External) 모형평가 모두에서 모형 설명력(R-square)이 높았으며, Mean Absolute Error 값, Root Mean Squared Error 값이 낮은 것으로 나타났다. 본 연구에서 랜덤포레스트를 이용하여 동반상병 보정 방법 3가지 및 독립변수의 4단계의 고려에 따라 개발된 12개 모형 중 CCS 진단군 분류 기준 동반상병 보정 방법 및 주진단, 성, 연령, 진료비 지불방법, 입원경로, 수술유무, 사망유무 모두를 이용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형의 내적(Internal), 외적(External) 모형평가의 모형 설명력(R-square)이 각각 26.8%, 26.9%로 가장 높았으며, Mean Absolute Error는 각각 0.081, 0.083, Root Mean Squared Error는 각각 0.107, 0.110으로 가장 낮은 것으로 나타나 모형의 예측력이 가장 우수한 것으로 조사되었다.

### 2) 모형의 결과

랜덤포레스트 결과에 있어서 CCS 진단군 분류 기준 동반상병 보정 방법 및 주진단, 성, 연령, 진료비 지불방법, 입원경로, 수술유무, 사망유무 등 독립변수의 상대적 중요도를 살펴본 결과는 그림 4와 같다. 독립변수의 상대적 중요도 지수를 파악한 결과 최장의 질환 동반유무의 상대적 중요도 지수가 10으로 가장 높아 신경계통의 질환 퇴원환자의 재원일수에 가장 중요한 요인인 것으로 나타났으며, 대퇴골의 골절 동반유무 9, 항문과 직장 상태 동반유무, 감염성 관절염 및 골수염 동반유무 7, 정맥과 림프종의 질병

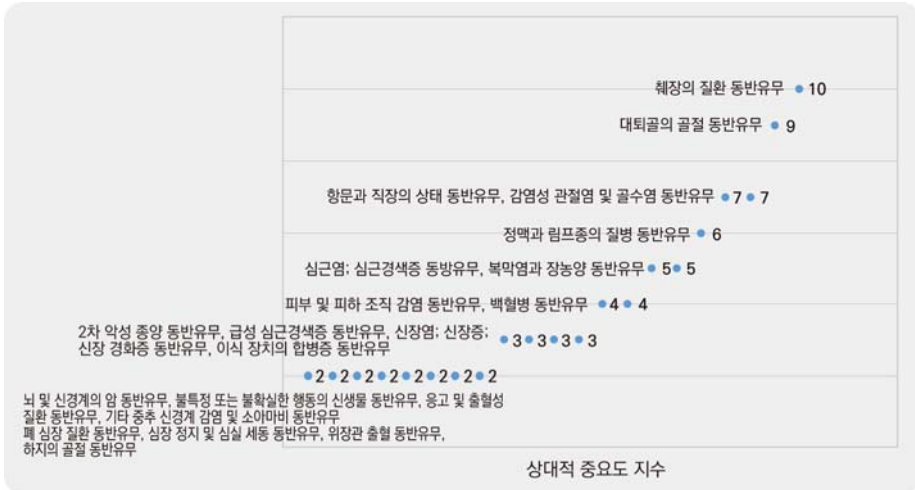
동반유무 6, 심근염, 심근경색증 동반유무, 복막염과 장농양 동반유무 5등의 순으로 상대적 중요도 지수가 높은 것으로 조사되었다.

표 8. 랜덤포레스트를 이용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형의 예측력 평가

구분		내적(Internal) 모형평가			외적(External) 모형평가		
		R square	MAE	RMSE	R square	MAE	RMSE
1 단 계	CCI	0.105	0.113	0.143	0.104	0.115	0.143
	ECI	0.074	0.113	0.143	0.080	0.116	0.145
	CCS	0.226	0.084	0.111	0.214	0.086	0.113
2 단 계	CCI + 주진단	0.116	0.112	0.142	0.117	0.114	0.142
	ECI + 주진단	0.079	0.113	0.143	0.086	0.116	0.145
	CCS + 주진단	0.232	0.084	0.110	0.221	0.085	0.112
3 단 계	CCI + 주진단 + 성 + 연령	0.139	0.111	0.140	0.134	0.112	0.141
	ECI + 주진단 + 성 + 연령	0.091	0.112	0.142	0.100	0.115	0.144
	CCS + 주진단 + 성 + 연령	0.238	0.083	0.110	0.227	0.085	0.112
4 단 계	CCI + 주진단 + 성 + 연령 + 진료비 지불방법 + 입원경로 + 수술유무 + 사망유무	0.175	0.088	0.117	0.171	0.090	0.118
	ECI + 주진단 + 성 + 연령 + 진료비 지불방법 + 입원경로 + 수술유무 + 사망유무	0.146	0.089	0.118	0.153	0.091	0.119
	CCS + 주진단 + 성 + 연령 + 진료비 지불방법 + 입원경로 + 수술유무 + 사망유무	0.268	0.081	0.107	0.269	0.083	0.110

주: MAE는 Mean Absolute Error, RMSE는 Root Mean Squared Error를 의미함

그림 4. 랜덤 포레스트의 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형 독립변수의 상대적 중요도 지수 결과



#### 마. 서포트 벡터 회귀분석을 이용한 중증도 보정 재원일수 예측 모형 개발 및 평가

서포트 벡터 회귀분석을 이용하여 동반상병 보정 방법 및 독립변수의 단계적 고려에 따라 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형을 개발하고 평가한 결과는 <표 9>와 같다. 개발된 모형의 모형 설명력(R-square) 및 Mean Absolute Error 값을 평가한 결과 1단계, 2단계, 3단계, 4단계 모두 CCS 진단군 분류 기준 동반상병 보정 방법을 이용하여 개발된 모형이 CCI 및 ECI 동반상병 보정 방법을 이용하여 개발된 모형보다 내적(Internal), 외적(External) 모형평가 모두에서 모형 설명력(R-square)이 높았으며, Mean Absolute Error 값, Root Mean Squared Error 값이 낮은 것으로 나타났다. 본 연구에서 서포트 벡터 회귀분석을 이용하여 동반상병 보정 방법 3가지 및 독립변수의 4단계의 고려에 따라 개발된 12개 모형 중 CCS 진단군 분류 기준 동반상병 보정 방법 및 주진단, 성, 연령, 진료비 지불방법, 입원경로, 수술유무, 사망유무 모두를 이용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형의 내적(Internal), 외적(External) 모형평가의 모형 설명력(R-square)이 각각 28.8%, 30.3%로

가장 높았으며, Mean Absolute Error는 각각 0.080, 0.081, Root Mean Squared Error는 각각 0.106으로 가장 낮은 것으로 나타나 모형의 예측력이 가장 우수한 것으로 조사되었다.

표 9. 서포트 벡터 회귀분석을 이용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형의 예측력 평가

구분	내적(Internal) 모형평가			외적(External) 모형평가			
	R square	MAE	RMSE	R square	MAE	RMSE	
1	CCI	0.088	0.113	0.144	0.088	0.115	0.145
단	ECI	0.077	0.112	0.143	0.089	0.115	0.145
계	CCS	0.248	0.081	0.109	0.245	0.083	0.110
2	CCI + 주진단	0.121	0.112	0.141	0.119	0.114	0.142
단	ECI + 주진단	0.098	0.112	0.141	0.111	0.114	0.143
계	CCS + 주진단	0.267	0.081	0.107	0.260	0.083	0.109
3	CCI + 주진단 + 성 + 연령	0.127	0.112	0.141	0.125	0.114	0.142
단	ECI + 주진단 + 성 + 연령	0.114	0.110	0.140	0.127	0.113	0.142
계	CCS + 주진단 + 성 + 연령	0.272	0.081	0.107	0.265	0.082	0.108
	CCI + 주진단 + 성 + 연령 + 진료비 지불방법 + 입원경로 + 수술유무 + 사망유무	0.172	0.089	0.117	0.167	0.090	0.118
4	ECI + 주진단 + 성 + 연령 + 진료비 지불방법 + 입원경로 + 수술유무 + 사망유무	0.149	0.088	0.117	0.163	0.090	0.119
계	CCS + 주진단 + 성 + 연령 + 진료비 지불방법 + 입원경로 + 수술유무 + 사망유무	0.288	0.08	0.106	0.303	0.081	0.106

주: MAE는 Mean Absolute Error, RMSE는 Root Mean Squared Error를 의미함

## 바. 신경망을 이용한 중증도 보정 재원일수 예측 모형 개발 및 평가

신경망을 이용하여 동반상병 보정 방법 및 독립변수의 단계적 고려에 따라 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형을 개발하고 평가한 결과는 표 10과 같다. 개발된 모형의 모형 설명력(R-square) 및 Mean Absolute Error 값을 평가한 결과 1단계, 2단계, 3단계, 4단계 모두 CCS 진단군 분류 기준 동반상병 보정 방법을

이용하여 개발된 모형이 CCI 및 ECI 동반상병 보정 방법을 이용하여 개발된 모형보다 내적(Internal), 외적(External) 모형평가 모두에서 모형 설명력(R-square)이 높았으며, Mean Absolute Error 값, Root Mean Squared Error 값이 낮은 것으로 나타났다. 본 연구에서 신경망을 이용하여 동반상병 보정 방법 3가지 및 독립변수의 4단계의 고려에 따라 개발된 12개 모형 중 CCS 진단군 분류 기준 동반상병 보정 방법 및 주진단, 성, 연령, 진료비 지불방법, 입원경로, 수술유무, 사망유무 모두를 이용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형의 내적(Internal), 외적(External) 모형평가의 모형 설명력(R-square)이 각각 29.9%, 31.3%로 가장 높았으며, Mean Absolute Error는 각각 0.075, 0.077, Root Mean Squared Error는 각각 0.103, 0.104로 가장 낮은 것으로 나타나 모형의 예측력이 가장 우수한 것으로 조사되었다.

표 10. 신경망 이용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형의 예측력 평가

구분	내적(Internal) 모형평가			외적(External) 모형평가			
	R square	MAE	RMSE	R square	MAE	RMSE	
1	CCI	0.105	0.113	0.143	0.104	0.115	0.143
단	ECI	0.085	0.113	0.142	0.094	0.115	0.144
계	CCS	0.255	0.082	0.108	0.246	0.083	0.110
2	CCI + 주진단	0.135	0.111	0.140	0.131	0.113	0.141
단	ECI + 주진단	0.106	0.111	0.141	0.119	0.113	0.142
계	CCS + 주진단	0.278	0.080	0.106	0.273	0.081	0.107
3	CCI + 주진단 + 성 + 연령	0.149	0.110	0.139	0.141	0.112	0.140
단	ECI + 주진단 + 성 + 연령	0.121	0.110	0.140	0.133	0.112	0.141
계	CCS + 주진단 + 성 + 연령	0.274	0.080	0.107	0.274	0.081	0.107
	CCI + 주진단 + 성 + 연령 + 진료비 지불방법 + 입원경로 + 수술유무 + 사망유무	0.219	0.085	0.113	0.203	0.087	0.115
4	ECI + 주진단 + 성 + 연령 + 진료비 지불방법 + 입원경로 + 수술유무 + 사망유무	0.195	0.085	0.114	0.198	0.088	0.116
계	CCS + 주진단 + 성 + 연령 + 진료비 지불방법 + 입원경로 + 수술유무 + 사망유무	0.299	0.075	0.103	0.313	0.077	0.104

주: MAE는 Mean Absolute Error, RMSE는 Root Mean Squared Error를 의미함

## V. 고찰

본 연구는 머신러닝을 이용하여 동반상병 보정 방법에 따른 중증도 보정 재원일수 예측 모형을 개발하고, 예측력 비교를 근거로 재원일수의 효율적 관리를 위한 중증도 보정 재원일수 예측 모형 개발의 알고리즘을 제시하고자 수행되었다.

CCI, ECI, CCS 진단군 분류 기준 동반상병 보정 방법에 따른 중증도 보정 재원일수 예측 모형의 예측력을 파악하기 위해 1단계 동반상병 보정 방법만 고려, 2단계 동반상병 보정 방법 및 주진단만 고려, 3단계 동반상병 보정 방법 및 주진단, 성, 연령만 고려, 4단계 동반상병 보정 방법 및 주진단, 성, 연령, 진료비 지불방법, 입원경로, 수술유무, 사망유무 모두를 고려하여 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형을 개발하고 평가한 결과 1단계, 2단계, 3단계, 4단계 모두 CCS 진단군 분류 기준 동반상병 보정 방법을 포함한 모형이 CCI 또는 ECI를 포함한 모형보다 모형 설명력 (R-square)이 높았으며, Mean Absolute Error, Root Mean Squared Error가 낮은 것으로 나타나 모형의 예측력이 우수한 것으로 조사되었다. 이는 본 연구에서 분석기법으로 고려한 머신러닝 분석 기법인 회귀분석, 의사결정나무, 랜덤 포레스트, 서포트 벡터 회귀분석, 신경망 기법 모두 동일하게 나타났다. 이와 같은 본 연구의 연구결과는 급성심근경색증, 허혈성심증질환, 급성뇌졸중의 중증도 보정 재원일수 예측 모형 개발 및 평가를 위해 CCI, ECI, CCS 진단군 분류 기준 동반상병 보정 방법 및 회귀분석, 의사결정나무, 신경망 분석을 이용하여 중증도 보정 재원일수 예측 모형을 개발하고 평가한 결과 회귀분석, 의사결정나무, 신경망 분석 모두에서 CCI, ECI 보다 CCS 진단군 분류 기준 동반상병 보정 방법의 Root Average squared error의 값이 낮았다는 선행연구의 연구결과(질병관리본부, 2012; 강성홍, 석향숙, 김원중, 2013; 질병관리본부, 2014)와 같은 맥락을 취한다고 할 수 있다. 선행연구에서는 중증도 보정 재원일수 예측 모형을 개발하기 위해 전체 독립변수만 고려하여 동반상병 보정 방법에 따라 중증도 보정 재원일수 예측 모형을 개발하고 평가한 결과를 제시한 반면, 본 연구에서는 4단계에 따라 단계별로 동반상병 보정 방법별 중증도 보정 재원일수 예측 모형을 개발하고, 개발된 모형의 평가 결과를 제시함으로써 중증도 보정을 위한 맞춤형 동반상병 보정 방법에 대한 타당성의 근거를 확보한 것이 의미가 있다고 할 수 있다.

머신러닝 분석 기법인 회귀분석, 의사결정나무, 랜덤 포레스트, 서포트 벡터 회귀분

석, 신경망 기법을 이용하여 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형을 개발하고 평가하였다. 그 결과 5가지 머신러닝 분석 기법 모두 CCS 진단군 분류 기준 동반상병 보정 방법 및 주진단, 성, 연령, 진료비 지불방법, 입원경로, 수술유무, 사망유무 모두를 이용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형의 모형 설명력(R-square)이 가장 높았으며, Mean Absolute Error, Root Mean Squared Error가 가장 낮아 모형의 예측력이 가장 우수한 것으로 나타났다. 5가지 머신러닝 분석 기법별로 모형의 설명력이 가장 높고, 예측력이 가장 우수한 최종 모형의 내적(Internal) 모형평가 및 외적(External) 모형평가 결과를 살펴보면 의사결정나무의 경우 모형 설명력(R-square)은 각각 18.8%, 16.4%였으며, Mean Absolute Error는 각각 0.086, 0.089, Root Mean Squared Error는 각각 0.116, 0.119로 5가지 머신러닝 분석 기법 중 모형 설명력(R-square)이 가장 낮고, Mean Absolute Error, Root Mean Squared Error가 가장 높아 예측력이 가장 낮은 것으로 나타났다. 이와 반대로 신경망 기법의 경우 모형 설명력(R-square)은 각각 29.9%, 31.3%였으며, Mean Absolute Error는 각각 0.075, 0.077, Root Mean Squared Error는 각각 0.103, 0.104로 5가지 머신러닝 기법 중 모형 설명력(R-square)이 가장 높았으며, Mean Absolute Error, Root Mean Squared Error가 가장 낮아 모형의 예측력이 가장 우수한 것으로 조사되었다.

의사결정나무 기법을 이용한 최종 모형의 설명력이 가장 낮고, 예측력이 가장 높았다는 본 연구의 연구결과는 의사결정나무는 반응변수가 연속형인 회귀모형에서 그 예측력이 떨어진다는(오미애, 최현수, 김수현, 장준혁, 진재현, 천미경, 2017)는 의사결정나무(Decision Tree)의 단점을 반영하는 결과이나 급성심근경색증, 허혈성심증질환, 급성뇌졸중, 목의 손상에 대한 중증도 보정 재원일수 예측 모형을 개발하고 평가한 결과 의사결정나무 및 CCS 진단군 분류 기준 동반상병 보정 방법을 이용하여 개발된 예측 모형의 Root Average squared error의 값이 가장 낮아 가장 우수한 모형이었다는 국내 선행연구의 연구결과(질병관리본부, 2012; 강성홍, 석향숙, 김원중, 2013; 질병관리본부, 2014; 홍성욱, 김영택, 박종호, 강성홍, 2015)와 일치하지 않는 결과이다. 국내 선행연구의 연구결과와 본 연구의 연구결과가 일치하지 않는 이유는 분석도구의 제한점으로 인한 의사결정나무의 분석방법에 대한 차이 때문인 것으로 판단된다. 본 연구에서 의사결정나무는 회귀나무 모형의 CHAID 알고리즘을 이용하여 의사결정나무 모형을 개발한 반면 국내 선행연구(질병관리본부, 2012; 강성홍, 석향숙, 김원중, 2013; 질병관리본부,

2014; 홍성욱, 김영택, 박중호, 강성홍, 2015)에서 의사결정나무는 회귀나무 모형 알고리즘 기반의 대화식(interactive) 방식을 이용하여 분석하였다. 의사결정나무의 대화식 방식은 데이터 자체에만 의존하는 기존 의사결정나무 모형에 비해 연구자 또는 전문가 중요하다고 생각하는 변수, 전문적인 지식을 모형에 반영할 수 있다는 장점이 있다(황규연 등, 2016). 따라서 국내 선행연구에서 개발한 의사결정나무 모형에는 연구자의 지식, 전문가의 자문 결과 등이 반영됨에 따라 의사결정나무 모형의 예측력이 더 우수하게 나타난 것으로 판단된다. 이는 데이터 자체에만 의존하는 머신러닝 분석 기법의 단점을 보완하기 위해 전문가, 연구자 등 인간의 선행된 지식, 경험 등이 필요함을 시사한다고 할 수 있다. 신경망 기법을 이용한 최종 모형의 설명력이 가장 높았으며, 예측력이 가장 우수한 것으로 조사되었다는 본 연구의 연구결과는 관상동맥질환, 당뇨병, 화상 등의 단위 질환에 대한 증증도 보정 재원일수 모형 개발 시 서포트 벡터 회귀분석, 랜덤 포레스트 등의 분석방법을 이용한 모형의 예측력이 가장 우수하였다는 외국의 선행연구(Yang, Wei, Yuan, & Schoung, 2010; Hachesu, Ahmadi, Alizadeh & Sadoughi, 2013; Morton, Marzban, Giannoulis & Patel, 2014) 결과와 일치하지 않았으나 급성뇌졸중에 대한 증증도 보정 재원일수 모형 개발 시 신경망 기법의 모형 예측력이 가장 우수하였다는 국내 선행연구(최병관, 함승우, 김축환, 서정숙, 박명화, 강성홍, 2018) 결과와 일치하는 결과였다. 증증도 보정 재원일수 예측 모형 개발 시 딥러닝의 기반 기술인 신경망 기법이 가장 우수하였다는 본 연구 및 국내 선행연구(최병관, 함승우, 김축환, 서정숙, 박명화, 강성홍, 2018)의 결과는 우리나라 퇴원환자의 재원일수 관리를 위해 국가 차원의 질 관리가 이루어진 더 많은 퇴원환자 빅데이터 자료를 수집하고, 딥러닝 기법의 도입을 통해 예측력이 향상된 증증도 보정 재원일수 예측 모형 개발하고, 이를 활용하는 것이 필요하다는 기초자료를 제공하였다는 것에 의의를 갖는다고 할 수 있다.

이와 같은 의의에도 불구하고, 머신러닝 분석 기법인 회귀분석, 의사결정나무, 랜덤 포레스트, 서포트 벡터 회귀분석, 신경망 기법을 이용하여 개발된 신경계통의 질환 퇴원환자의 증증도 보정 재원일수 예측 모형의 정확성을 비교하기 위해 Resampling 등의 방법을 이용하여 정확도의 Confidence interval을 제시하고, Confidence interval의 중복 여부를 살펴봄으로써 특정 모델의 우수성을 평가하는 평가 방법을 제시하지 못한 것은 본 연구의 제한점이라 할 수 있다.



## VI. 결론

동반상병 보정 방법 및 머신러닝 분석기법에 따라 개발된 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형의 설명력 및 예측력을 근거로 다음과 같이 중증도 보정 재원일수 예측 모형 개발의 알고리즘을 제시하고자 한다.

첫째, 중증도 보정 시 CCS 진단군 분류 변수를 이용한 동반상병 보정 방법을 사용할 필요가 있다.

CCS 진단군 분류는 모든 질병이 임상학적으로 군집화되어 있음에 따라 모든 동반상병을 중증도 보정을 위해 고려할 수 있다. 이는 신경계통의 질환 이외 다른 질환에도 맞춤형 동반상병 보정이 가능하며, 재원일수 뿐 아니라 사망 등의 다른 진료결과를 관리하기 위한 중증도 보정 모형 개발에도 맞춤형 동반상병 보정이 가능하다. 맞춤형 동반상병 보정은 중증도 보정 예측 모형의 예측력 향상을 가져 올 수 있다.

둘째, 중증도 보정 재원일수 예측 모형 개발 시 재원일수에 영향을 미치는 요인을 단계적으로 고려할 필요가 있다. 국내 중증도 보정 재원일수 예측 모형 개발의 선행연구는 재원일수에 영향을 미치는 요인을 정의하고, 요인 모두를 중증도 보정 재원일수 예측 모형 개발 시 고려하였다. 예측 모형의 예측력 향상을 위해서는 재원일수에 영향을 미치는 요인을 단계적으로 고려하고, 이의 예측력을 평가하여 가장 우수한 모형을 개발하고 활용할 필요가 있다.

셋째, 중증도 보정 재원일수 예측 모형 개발 시 머신러닝의 다양한 분석 기법, 딥러닝 분석 기법 등을 활용할 필요가 있다. 다양한 머신러닝 분석 기법을 활용하여 중증도 보정 재원일수 예측 모형을 개발하고 모형의 예측력을 평가하여 가장 우수한 모형을 활용하여야 한다. 또한 재원일수 관리를 위해 국가 차원에서 질 관리가 이루어진 퇴원환자 빅데이터 자료를 수집하고, 딥러닝을 이용하여 중증도 보정 재원일수 예측 모형을 개발하고 활용하는 것이 필요하다.

## 참고문헌

- 강성훈, 석향숙, 김원중. (2013). 급성 뇌졸중 환자의 중증도 보정 재원일수 변이에 관한 연구. *디지털정책연구*, 11(6), pp.221-233.
- 고은해, 강희양, 김용식, 정성훈. (2017). 조현병 감별진단에 대한 머신 러닝 기법의 적용: WAIS-IV의 진단 예측 역량. *신경정신의학*, 56(3), pp.103-110.
- 김보찬, 김유현, 김민정, 이종석. (2018). 데이터마이닝을 활용한 서울 주요 대학가 주거용 부동산 임대료 모형 수립에 관한 연구. *대한산업공학회지*, 44(4), pp.259-271.
- 김선웅, 최흥식. (2017). Support Vector Regression을 이용한 GARCH 모형의 추정과 투자전략의 성과분석. *지능정보연구*, 23(2), pp.107-122.
- 김성진, 유은정, 정민규, 김재경, 안현철. (2012). 감정예측모형의 성과개선을 위한 Support Vector Regression 응용. *지능정보연구*, 18(3), pp.185-202.
- 김현호, 양승범, 강연석, 박영배, 김재효. (2016). 기계학습을 적용한 자기보고 증상 기반의 어혈 변증 모델 구축. *경락경혈학회지*, 33(3), pp.102-113.
- 박일수, 용왕식, 김유미, 강성훈, 한준태. (2008). 데이터마이닝 기법을 활용한 맞춤형 고혈압 사후관리 모형 개발. *응용통계연구*, 21(4), pp.639-647.
- 박정선, 고영경. (2003). A Decision Tree Algorithm using Genetic Programming. *한국통계학회논문집*, 10(3), pp.845-857.
- 서기성. (2017). SVR을 사용한 데이터 학습 기반의 풍속 예측 모델 생성. *한국지능시스템학회논문지*, 27(6), pp.493-498.
- 송태민, 오영호, 정영호, 도새록, 고숙자, 최정수 등. (2013). 2012년도 환자조사 심층분석. 서울: 한국보건사회연구원.
- 신혜정, 이윤동, 이은경. (2016). 임상자료를 이용한 나무구조 분류모형의 성능 비교. *보건정보통계학회지*, 41(1), pp.135-146.
- 오미애, 최현수, 김수현, 장준혁, 진재현, 천미경. (2017). 기계학습(Machine Learning) 기반 사회보장 빅데이터 분석 및 예측모형 연구. 세종: 한국보건사회연구원.
- 유진은. (2015). 랜덤 포레스트 의사결정나무의 대안으로서의 데이터 마이닝 기법. *교육평가연구*, 28(2), pp.427-448.

- 이강윤, 김준혁. (2016). 인공지능 왓슨 기술과 보건의료의 적용. *의학교육논단*, 18(2), pp.51-57.
- 이관용, 김진희, 김현철. (2016). 의료 인공지능 현황 및 과제. *보건산업브리프*, 219, pp.1-28.
- 이근영. (2015). 머신러닝을 활용한 스마트 서비스와 금융. *전자금융과 금융보안*, 7, pp.31-66.
- 이길홍, 김경신, 신무철, 최경수, 김종민, 박동현 등. (2017). 머신러닝 기반 악성코드 분석 알고리즘 적합성 연구. 나주: 한국인터넷진흥원.
- 이호현, 정승현, 최은정. (2016). 기계학습 응용 및 학습 알고리즘 성능 개선방안 사례연구. *디지털융복합연구*, 14(2), pp.245-258.
- 임현일. (2018). 코드 분포의 선형 회귀를 이용한 프로그램 유사성 분석. *디지털콘텐츠학 회논문지*, 19(7), pp.1357-1363.
- 장은나, 임정호, 하성현, 이상균, 박영규. (2016). GOCI 위성영상과 기계학습을 이용한 한반도 연안 수질평가지수 추정. *대한원격탐사학회지*, 32(3), pp.221-234.
- 정광훈, 김가희, 박진원, 이해성, 김형중, 최홍용 등. (2015). 의료 데이터 매시업과 빅데이터 기법 활용을 통한 환자의 재입원 가능성 예측과 원인 분석. *Entrue Journal of Information Technology*, 13(3), pp.133-149.
- 정성호, 이대업, 이경상. (2018). 딥러닝 오픈 라이브러리를 이용한 하천수위 예측. *한국방재학회논문집*, 18(1), pp.1-11.
- 질병관리본부. (2012). 퇴원손상심층조사 표본 병원 의료정보 질 평가를 위한 중증보정 병원통계 생성모형 개발. 오송: 동 기관.
- 질병관리본부. (2014). 중증도 보정 병원통계 프로그램 시범적용 및 평가. 오송: 동 기관.
- 최명관, 함승우, 김축환, 서정숙, 박명화, 강성홍. (2018). 인공지능을 이용한 급성 뇌졸중 환자의 재입원일수 예측모형 개발. *디지털융복합연구*, 16(1), pp.231-242.
- 최정규, 장홍석, 이명수, 홍진표, 박종익. (2013). 국내 정신보건시설의 재원기간과 관련된 정신사회적 요인. *신경정신의학*, 52, pp.98-107.
- 최종후, 서두성. (1999). 데이터마이닝 의사결정나무의 응용. *통계분석연구*, 4(1), pp.61-83.
- 홍성욱, 김영택, 박종호, 강성홍. (2015). 목의 손상 입원환자의 중증도 보정 재원일수

- 변이요인. *보건사회연구*, 35(2), pp.561-583.
- 황규연, 이은숙, 김고원, 홍성욱, 박정선, 광미숙 등. (2016). 대화식 의사결정나무를 이용한 보건의료 데이터 질 관리 알고리즘 개발: 당뇨병환자의 고혈압 동반을 중심으로. *보건의료산업학회지*, 10(3), pp.63-74.
- Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., et al. (2018). Scalable and accurate deep learning for electronic health records. *npj Digital Medicine*, 18(1), pp.1-10.
- Morton, A., Marzban, E., Giannoulis, G., Patel, A. (2014). *A Comparison of Supervised Machine Learning Techniques for Predicting Short-Term In-Hospital Length of Stay Among Diabetic Patients*. Paper presented at the 2014 13th International Conference on Machine Learning and Applications.
- Charlson, M. E., Pompei, P., Ales, K. L., MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*, 40, pp.373-383.
- Yang, C. S., Wei, C. P., Yuan, C. C., Schoung, J. Y. (2010). Predicting the length of hospital stay of burn patients: Comparisons of prediction accuracy among different clinical stages. *Decision Support Systems*, 50, pp.325-335.
- Elixhauser, A., Steiner, C., Harris, D. R., Coffey, R. M. (1998). Comorbidity measures for use with administrative data. *Med Care*, 36, pp.8-27.
- Quan, H., Li, B., Couris, C. M., Fushimi, K., Graham, P., Hider, P., Januel JM. et al. (2011). Updating and Validating the Charlson Comorbidity Index and Score for Risk Adjustment in Hospital Discharge Abstracts Using Data From 6 Countries. *American Journal of Epidemiology*, 173(6), pp.676-682.
- OECD. (2014). *Society at a Glance: Asia/Pacific 2014*. Paris: author.
- OECD. (2017). *Health at a Glance 2017*. Paris: author.
- Hachesu, P. R., Ahmadi, M., Alizadeh, S., Sadoughi, F. (2013). Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients. *Healthc Inform Res*, 19(2), pp.121-129.
- World Economic Forum. (2016). *The fourth industrial revolution Klaus Schwab*.

Smola, A. J., Scholkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3), pp.199-222.

OECD.Stat. (2018. Nov. 8). *Hospital average length of stay by diagnostic categories*.  
<https://stats.oecd.org/index.aspx?queryid=30165>에서 2018.11.18. 인출

박중호는 계명대학교에서 의료경영학 석사를, 인제대학교에서 보건행정학 박사학위를 받고, 현재 계명대학교 동산의료원 기획팀에서 부주사로 재직 중이다. 주요 관심분야는 의료정보, 병원경영 등이다. 현재 머신러닝, 딥러닝, 의료서비스 최적화에 대해서 연구 중이다.

(E-mail: jh8283p@naver.com)

강성홍은 서울대학교에서 보건학 석사를, 인제대학교에서 보건학 박사학위를 받고, 현재 인제대학교 보건행정학과 교수로 재직 중이다. 주요 관심분야는 의료정보, 의료의 질관리 등이다. 현재 인공지능, 의료의 질 향상 등에 대해서 연구 중이다.

(E-mail: hcmkang@inje.ac.kr)

# A Study on the Development of Predictive Model for Severity-Adjusted Length of Stay in Nervous System Patients Using Machine Learning

**Park, Jongho**

(Keimyung University  
Dongsan Medical Center)

**Kang, Sunghong**

(Inje University)

---

This study aims to develop a severity-adjusted length of stay predictive model according to comorbidity index by using machine learning and propose a algorithm of severity-adjusted length of stay (LOS) predictive model. The dataset was taken from Korea Centers for Disease Control and Prevention database of the hospital discharge survey from 2006 to 2015 and the severity-adjusted length of stay predictive model was developed for the nervous system patients to need a urgent management for length of stay. when it comes to the severity-adjusted length of stay predictive model about nervous system discharging patients, three tools were used for the severity-adjustment of comorbidity: the CCI, the ECI, and the CCS. The models using Regression, Decision Tree, Random Forest, Support Vector Regression, Neural Network as a Machine learning analysis methods were developed and then evaluate. As a result, Severity-adjusted predictive model using CCS as the severity-adjustment of comorbidity and Neural Network method has the highest R-square and has the most excellent prediction capability. In conclusion, there is a need to develop a severity-adjusted predictive model using CCS as the severity-adjustment of comorbidity and make use of severity-adjusted predictive model to has high prediction capability by using various machine-learning analytics.

---

**Keywords:** Length of Stay, Diseases of the Nervous System, Comorbidity Index, Machine Learning, Severity-Adjusted Predictive Model