HIRA RESEARCH 2021;1(2):196-207 | pISSN 2765-6764 eISSN 2765-7353
https://doi.org/10.52937/hira.21.1.2.196

Original Article

보건의료 빅데이터 기반 인공지능 활용 전략

조상아[1], 김한상[2]

[1]건강보험심사평가원 심사평가연구소, [2]국민건강보험공단 의료기관자원실

Healthcare Big Data-Based Artificial Intelligence Utilization Strategy

Sang-A Cho[1], Hansang Kim[2]

[1]Health Insurance Review and Assessment Research Institute, Health Insurance Review and Assessment Service; [2]Department of Healthcare Institution Support, National Health Insurance Service, Wonju, Korea

Correspondence to:
Hansang Kim
Department of Healthcare Institution Support, National Health Insurance Service, 32 Geongang-ro, Wonju 26464, Korea
Tel: +82-33-736-4421
Fax: +82-33-749-6395
E-mail: yoonkim0423@gmail.com

Received: October 22, 2021
Revised: November 15, 2021
Accepted after revision: November 16, 2021

Background: Healthcare studies mainly use statistical methodologies, but recently, with the development of artificial intelligence (AI) technology, studies to solve the limitations of some existing methodologies are being published. This study aims to suggest a strategy for using AI technology in the future by reviewing existing research and research cases of the last 10 years using artificial intelligence technology.
Methods: We selected research papers and domestic and international journal papers by the Health Insurance Review and Assessment Service since 2010. Health insurance claim data were divided into 'cross-sectional study', 'association and comparative analysis of specific factors', and 'longitudinal cross-sectional, time series analysis study'.
Results: Four situations requiring AI technology were defined, and AI methodologies applicable to each situation were presented.
Conclusion: It is judged that it is necessary to actively discuss more specific situations and various methodologies for the use of big data in health care.

Keywords: Big data; Healthcare; Health Insurance Review & Assessment Service; Artificial intelligence

서 론

보건의료 빅데이터는 급변하는 환경에서 사회적 문제 및 변화에 대한 예측과 해결방안 마련을 위한 근거자료 제공 등 다양한 목적으로 활용되고 있다. 대표적인 보건의료 빅데이터로는 건강보험 청구자료가 있는데, 이는 전 국민이[1] 의료기관을 방문하여 제공 받은 의료서비스에 대한 내역과

1 건강보험 가입자

금액을 요양기관이 보험자에게 청구할 때 발생되는 데이터로 여기에는 전체 인구 중 약 98%의 의료이용 정보가 비식별화되어 축적되어 있다[1].

건강보험 청구데이터에 포함 정보는 기본적인 환자와 요양기관을 식별할 수 있는 고유번호와 환자의 진단명, 진료과목, 진료일(진료시작일, 입 · 내원일수 등), 진료비 등에 대한 정보 및 상세적인 진료, 검사, 시술, 수술, 처치 등의 진료내역이 포함되어 있다. 청구데이터는 표본자료가 아닌 전수 자료로 제한된 환경이 아닌 현실적인 보건의료환경을 반영하고, 단면적 연구와 코호트 연구가 모두 가능한 데이터이다. 현재 보건의료정책 근거자료 생성, 보건의료분야의 다양한 연구 및 의료계 및 산업계의 연구개발(research and development, R&D)에 이용되고 있다. 그간 청구데이터를 이용한 분석연구는 크게 단면적 분석 중심의 현황 및 추세분석연구[2,3], 특정 요인에 대한 설명(영향) 및 비교분석연구[4–6], 과거 정보를 고려한 종단면, 시계열 분석[7–9]으로 구분할 수 있고 주로 통계적 방법이 이용되었다.

통계적 방법은 현상을 선형적 관계 기반으로 간결하게 설명하는 데 강점이 있어 주로 사용되었지만, 비선형적 복잡 정보의 설명 및 예측에는 제한이 존재한다[10–12]. 최근 이런 통계적 방법론들의 한계를 보완하기 위해 보건의료 빅데이터에 인공지능(artificial intelligence) 기술을 활용하는 연구들이 국외에서 발표되고 있다[13,14]. 기존 방법론의 주요 제한점으로는 비선형적 복잡 정보에 대한 모델링 제한, 텍스트 정보 활용 시 발생되는 고차원화 문제, 불규칙한 과거 정보의 반영 등이 있다. Goldstein 등[11]은 보건의료 데이터 기반 예측연구들에서 선형적 관계 중심의 전통적 방법론보다 머신러닝(machine learning) 방법이 상대적으로 많은 정보를 활용하였고 모델 성능도 향상되는 것으로 나타났다. 그리고 최근 인공지능 기법 기반 자연어처리기법을 적용하여 고차원 문제를 해결하는 연구들도 소개되고 있다[15,16]. 보건의료 데이터에서 진료내역(진단명, 처치, 약제 등)은 주요 정보 중 하나이지만, 고차원의 범주(코드) 정보로 기존 방법론으로는 이를 모델에 반영하기엔 제한점이 존재했다.[2] 인공지능 방법론 중 하나인 딥러닝 기반 단어 임베딩 기법을 활용하여 고차원 문제를 해결하고자 하였고, 실제 발표된 여러 연구에서 이를 적용하여 예측모델의 성능이 향상되는 것으로 나타났다[16,17]. 또한 과거 불규칙한 진료 정보의 활용에 있어서도 다양한 딥러닝 방법론을 적용한 시도들이 발표되고 있다[18]. 따라서 이 연구는 보건의료 빅데이터를 활용한 기존 국내 연구들을 분석하여 활용범위 및 제한점을 검토하고 이를 보완할 수 있는 인공지능 기술 활용 전략을 제안하고자 하였다.

방 법

이 연구는 최근 10년간 건강보험 청구자료를 이용한 27개의 국내외 연구보고서와 논문을 고찰하였다. 자료는 2010년 이후 발표된 건강보험심사평가원 연구보고서와 분석방법이 비교적 상세하게 기록된 국내외 논문을 중심으로 선정하였다. 이를 건강보험 청구자료의 분석관점으로 분류하고자 '단면적 현황 및 추세 연구', '특정 요인에 대한 설명(영향) 및 비교분석연구', '과거 정보를 고려한 종단면, 시계열 분석연구'로 구분하였다[2–9,19–37] (표 1).

2 차원의 저주(curse of dimensionality)

표 1. 주요 연구사례 선정

구분	주요 분석기법	주요 이용정보	연구사례(27개)
단면적 자료 중심 현황 및 추세분석	• 현황 및 추세분석: 빈도분석; 유병률, 발병률 산출	종별 구분, 환자의 주진단명, 진료과목, 산정특례코드, 입원 · 외래 구분, DRG 분류군, 입 · 내원일수, 진료비, 진료내역(행위, 약제 등), 약효분류 구분, 이용량 등	최지숙 등[19] (2018), Kim 등[2] (2017), 한승진 등[3] (2020), 김한상 등[20] (2020), 오동관 등[21] (2015), Park 등[22] (2020), Kim 등[23] (2021),
특정 요인에 대한 설명(영향) 및 비교분석연구	• 요인분석: 상관분석, 선형회귀분석, 로지스틱회귀분석, 음이항회귀분석 등 • 비교 분석: *t*-test, chi-square test 등	종별 구분, 환자의 주진단명, 진료과목, 산정특례코드, 입원 · 외래 구분, DRG 분류군, 입 · 내원일수, 진료비,진료내역(행위, 약제 등), 약효분류구분, 이용량 등	김동숙 등[4] (2017), Kang 등[5] (2021), 이성우 등[6] (2018), 김동숙 등[24] (2017), 박효성 등[25] (2017), Kim 등[26] (2018), An 등[27] (2020), Lee 등[28] (2021), Ko 등[29] (2021), Kim 등[30] (2021)
과거 진료이력 등을 반영한 종단면, 시계열 분석연구	• 시간을 기준으로 전·후 비교분석: 이중차이분석 • 환자 이동 추적분석: 빈도분석 • 시간을 고려한 예측 및 추세분석: 시계열분석, 포아송 회귀분석, 생존분석 등	종별 구분, 환자의 주진단명, 진료과목, 산정특례코드, 입원 · 외래 구분, DRG 분류군, 입 · 내원일수, 진료비, 진료내역(행위, 약제 등), 약효분류 구분, 이용량 등	민인순 등[31] (2017), 김지애 등[32] (2020), 오주연 등[33] (2020), 이도경 등[7] (2020), 김한상 등[34] (2020), Ryu 등[9] (2021), 박찬미 등[36] (2010), 신민선 등[8] (2020), Lee 등[35] (2017), Lee 등[37] (2020)

DRG, diagnosis-related group.

결 과

1. 보건의료 빅데이터 활용 분석사례 및 제한점

1) 연구사례

단면적 분석 중심 연구를 살펴보면, 문제점에 대한 해결점을 찾기 위한 현황분석 중심으로 이뤄졌다. 2020년 한승진 등[3]은 의료이용 불균형의 현상문제를 파악하고, 이를 해결하기 위한 근거자료 마련을 목적으로 2008년에서 2019년 동안 연도별 의료이용현황을 분석하였다. 연도별 의료기관 종별 입 · 내원일수, 진료비, 중증환자 구성비율, 산정특례환자 수, 신규환자 수 등의 추이를 연도별 증감률 및 연도 내 점유율을 분석하여 의료이용 불균형에 대한 현상을 보여주었다. 2017년 Kim 등[2]은 급성호흡기계 질환에서 항생제 처방의 경향 및 사용량을 파악하고자 2005년부터 2008년까지 호흡기계 질환을 가진 환자를 대상으로 연도별로 세부 상병에 따른 항생제 계열별 사용량에 대해 분석하였다.

그리고 특정 요인에 대한 설명(영향) 및 비교분석 연구는 청구자료 분석을 통해 특정 요인과 다른 요인 간의 관련성 및 영향을 미치는 요인들을 찾기 위한 목적으로 주로 수행되었다. 2017년 김동숙 등[4]은 항생제 사용에 영향을 미치는 요인을 찾아내기 위해 2011년부터 2015년까지 병원 규모별로 청구자료를 분석하였고, 다중선형회귀분석을 적용하여 내원일수와 항생제 사용의 관련성을 보여주었다. 2021년 Kang 등[5]은 알레르기 비염환자를 대상으로 의과 · 한의과 간 차이를 비교하기 위해 치료기간 및 치료금액 등을 비교분석하였으며, 이성우 등[6]도 그룹 간의 비교분석과 요인 간의 관련성을 파악하기 위해 회귀분석을 적용하였다.

끝으로 과거 정보를 고려한 종단면, 시계열 분석 연구는 환자별 과거 진료이력 정보 등을 활용하여 추적분석과 특정(정책 시행) 시점 전, 후 차이분석

및 시간에 따른 변화 추이를 예측하는 등의 목적으로 수행되었다. 이도경 등[7]은 상급종합병원의 회송환자의 이동경로를 추적하여 회송시범사업의 효과성을 파악하기 위해 2018년 회송환자 관리료가 청구된 환자를 대상으로 2019년 회송 후 상병별, 의료기관 종별 의료이용에 대한 추적결과를 발표하였고, Ryu 등[9]은 정신분열증 환자의 의료이용 추이를 예측하기 위해 2010년부터 2019년까지의 자료로 시계열분석을 수행하였다. 또한 신민선 등[8]은 코로나19 유행기간의 사망과 의료이용량의 변화를 예측하기 위해 2010년부터 2020년 9월까지의 청구자료로 포아송 회귀모형을 이용하여 사망자 수 또는 의료이용의 변화를 예측하여 발표하였다.

2) 분석 제한점

최근 현상의 이해 중심의 분석방법론뿐만 아니라 빅데이터 기반 복잡한 정보들을 학습하고, 이를 통해 특정 현상을 예측하는 방법들에 대한 방법론에 대해서도 많은 연구가 이루어지고 있다. 예측관점에서 통계적 방법론 중심 건강보험 청구자료 분석은 비선형적인 복잡한 정보, 텍스트 정보[3], 불규칙한 시계열 정보를 반영하기 어렵다는 제한이 존재한다.

앞서 고찰한 연구들의 분석모델을 살펴보면 주로 20여 개의 적은 정보 중심으로 특정 현상 및 문제를 설명하였다[2–9,19–33,35–37] (표 2). 비교적 복잡하지 않은 현상 및 문제를 설명하고자 할 때 설명이 쉬운 통계적 방법론은 좋은 선택일 수 있지만 보다 복잡한 상황에서는 선형성 등 기본가정을 만족시키기 어렵고, 실제 비선형적 복합관계를 가지므로

표 2. 청구데이터 내 정보 활용현황

연구사례	보유 정보	이용현황	연구사례
기본 정보 (약 230개)	• 일자 정보: 요양개시일자, 요양종료일자 등 • 요양기관 정보: 요양기관기호 및 종별 코드 등 • 수진자 정보: 수진자 개인식별번호, 보험자 구분코드 등 • 상병 정보: 주상병코드, 부상병코드 등 • 기타 코드정보: 지급구분코드, 서식구분코드 등 • 지표 정보: 명세서 CI • 일수 정보: 내원일수, 원외처방일수 등 • 이용량: 원외처방전건수, 원내처방약품수 등 • 금액 정보: 심사결정요양급여비용총액금액 등 • 기타 정보: 심사부서코드, 적재일시 등	약 20개	한승진 등[3] (2020), 김한상 등[20] (2020), 최지숙 등[19] (2018), 박찬미 등[36] (2010), Kang 등[5] (2021), 김동숙 등[4] (2017), 오동관 등[21] (2015), 김지애 등[32] (2020), 이성우 등[6] (2018), 김지애 등[2] (2017), 민인순 등[31] (2017), 오주연 등[33] (2020), 신민선 등[8] (2020), 이도경 등[7] (2020), Ryu 등[9] (2021), Kim 등[30] (2021), Park 등[22] (2020), Kim 등[23] (2021), 박효성 등[25] (2017), Lee 등[35] (2017), Kim 등[26] (2018), An 등[27] (2020), Lee 등[28] (2021), Ko 등[29] (2021)
상세 정보 (약 200개)	• 상세진료내역: 서식구분코드, 항목코드, 통합분류코드, 약효분류번호구분코드 등 • 세부이용량: 총투여일수 실시횟수, 총사용량 실시횟수 등 • 세부금액 정보: 산출단가, 기본담가, 인정금액, 조정금액 등 • 기타 정보: 약품규격명, 약효분류번호, 적재일시 등 • 상병 정보(상세): 주상병코드, 부상병코드, 상병기호 등 • 특정 내역 정보: 특정 내역코드(상세) 등	약 20개	박찬미 등[36] (2010), 김지애 등[2] (2017), 김동숙 등[4] (2017), 오동관 등[21] (2015), 이성우 등[6] (2018), 오주연 등[33] (2020), 이도경 등[7] (2020), Lee 등[37] (2020), Kim 등[23] (2021), 박효성 등[25] (2017), Lee 등[35] (2017), Kim 등[26] (2018), An 등[27] (2020), Lee 등[28] (2021), Park 등[22] (2020), Kim 등[23] (2021), Ko 등[29] (2021), 박찬미 등[36] (2010), 김지애 등[2] (2017), 김동숙 등[24] (2017), Lee 등[35] (2017), 김지애 등[32] (2020), 신민선 등[8] (2020)

3 진료내역(행위, 약제 등) 코드 정보 등

설명(예측)모델의 성능을 담보하기 어렵다.

그리고 대부분 연구에서 설명변수로 연속형 변수와 범주가 작은 이산형 변수[4]를 이용하였다. 해당 정보가 특정 요인을 대부분 설명(예측)한다면 문제가 되지 않지만, 주요 정보를 통계적 모델링 제약으로 인해 반영하지 못할 경우 모델 성능에 영향을 미친다. 대표적인 예로 고차원의 범주형 정보가 있다. 범주형 정보 처리의 경우 대표적 방법으로 더미변수의 활용이 있다. 그러나 N-차원의 범주형 정보를 더미변수로 취환 시 (N-1)개의 설명변수가 만들어지게 되는데, 차원이 커질수록 '차원의 저주(curse of dimensionality)'의 문제가 발생된다. 환자의 진료 정보 중 주요 정보인 진단명, 처치 및 수술내역, 약제 처방내역 등의 정보가 대표적 고차원의 범주형(텍스트) 정보[5]이다.

건강보험 청구자료의 정보는 기본적으로 환자가 요양기관을 방문해 진료를 받을 때 발생된다. 즉 정보의 발생시점이 불규칙하다. 이런 불규칙 시점 정보와 복잡한 정보들의 반영 등의 어려움으로 인해 국내 연구에서 종단면 및 시계열 분석 활용성이 적었는데, 이런 경향은 해외에서도 비슷하게 나타났다. Goldstein 등[11]은 전자건강기록(electronic health record, EHR) 데이터를 이용한 분석연구들을 검토하였는데, 여기서도 대부분의 연구들이 시간(혹은 반복측정)이 고려되지 않은 것으로 나타났다.

2. 딥러닝 방법론 및 보건의료분야 활용사례

앞서 언급된 통계적 방법론의 특정 제한점을 개선하고자 다양한 인공지능 방법들이 연구되었다. 본 장에서는 인공지능 방법론에 대해 간단히 소개하고 활용사례를 정리하였다.

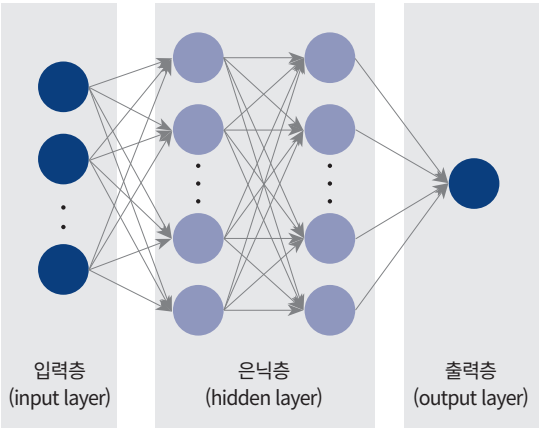


그림 1. 딥러닝 분석 기본 프로세스.

1) 인공지능 방법론

(1) 심층신경망

딥러닝은 기계학습의 인공신경망(artificial neural network)을 다중(심층)으로 연결한 모델로 하나의 인공신경망은 여러 정보들을[6] 이용해 특정 정보를[7] 설명하는 과정에서 선형식과 비선형식의 조합을 이용해 보다 복잡한 정보들을 구분(설명)할 수 있다. 그리고 이를 다중(심층)으로 연결할수록 더 복잡한 구조의 정보도 모델링이 가능해진다. 즉 앞서 언급한 보건의료 정보의 복잡성 및 불규칙성의 반영이 가능하다. 딥러닝 모델은 입력층(input layer)과 다수의 은닉층(hidden layer) 그리고 출력층(output layer)으로 구성되어있고, 은닉층이 많을수록 더 깊은(deep) 분석모형이 된다(그림 1).

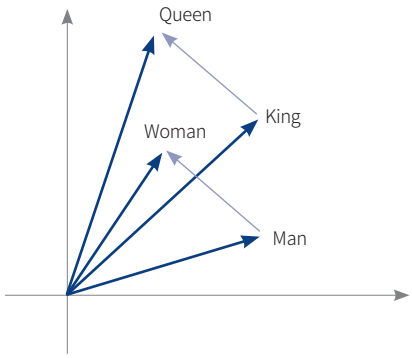
4 기존 분석연구사례에서 범주형 변수의 경우 2 – 9개 분류를 설명변수로 반영함
5 진단명 약 1,400개(3단 기준), 행위수가코드 약 5만 개
6 독립, 설명변수, input, feature 값 등
7 종속, 결과변수, output, label, target 값 등

(2) 단어 임베딩

딥러닝 기술이 발전함에 따라 자연어처리(natural language processing) 분야에서도 이에 기반한 연구가 활발히 진행되고 있고, 많은 성과를 보이고 있다[38]. 특히 단어 임베딩(word embedding)에서 큰 성과를 보였는데, 이는 주변 단어들의 분포에 기반하여 단어들의 유사도를 계산하고, 이를 n차원 벡터로 매핑시키는 기법이다. 이는 "유사한 분포를 가지는 언어항목(linguistic items)은 유사한 의미를 나타내는 경향이 있다"라는 언어학의 'distributional hypothesis'에 기반한다[39]. 단어 임베딩 기법을 통해 고차원의 텍스트 정보를 저차원으로 벡터화할 수 있고, 벡터 간의 수학적 연산을 통해 유사도를 계산할 수 있다[40] (그림 2). 이를 데이터 기반 모델링 관점으로 확장하면, 고차원화의 문제로 인해 적용할 수 없었던 범주형 변수를[8] 저차원의 벡터로 변환하여 설명변수로 활용할 수 있다는 의미이다.[9] 보건의료분야에서도 이를 활용한 연구들이 발표되고 있는데, 기본 아이디어는 다음과 같다. 우선 진료 정보[10] 코드들을 하나의 단어로 간주하고 이런 단어들을 묶어 문장(sentence)을 구성[11]한다. 그리고 문장 내 특정 코드가 주변 코드들과 자주 발생되면 유사 의료상황에서 나타날 가능성이 높을 것이라는 가정하에 문장을 학습하고, 각 코드들이 저차원 벡터값으로 매칭한다(그림 3). 최종적으로 매칭된 벡터값을 설명변수에 추가하여 모델의 예측(설명) 정확도를 검토한다.

Word	Encoding
Woman	[1, 0, 0, 0]
Queen	[0, 1, 0, 0]
Man	[0, 0, 1, 0,]
King	[0, 0, 0, 1]

A. 원핫인코딩(one-hot-encoding)



B. 단어 임베딩(word embedding)

그림 2. 원핫인코딩과 단어입베딩의 차이. (A) 범주 개수 4개, 차원(변수) 4개, 범주 간 유사도 계산 불가. (B) 범주 개수 4개, 차원(변수) 2개, 범주 간 유사도 계산 가능. 자료: Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space [Internet]. Ithaca (NY): arXiv.org; 2013 [cited 2019 Sep 15]. Available from: https://arxiv.org/abs/1301.3781 [40].

8 범주의 개수가 많은 변수

9 기존엔 주로 더미변수(dummy variable), 원핫인코딩(one-hot-encoding)을 사용하였지만, 범주의 개수가 많은 경우 차원의 저주(curse of dimensionality) 문제가 발생되고 각 범주별로 유사도 등을 계산할 수 없었음

10 진단명, 약제, 처치 및 검사내역 등

11 연구에 목적에 따라 문장은 다양하게 정의될 수 있고(예, 환자의 한 방문에서 발생된 진료 정보 코드 묶음), 문장 내 구성은 진단상병, 약제, 처치내역 등의 다양한 조합으로 이루어짐

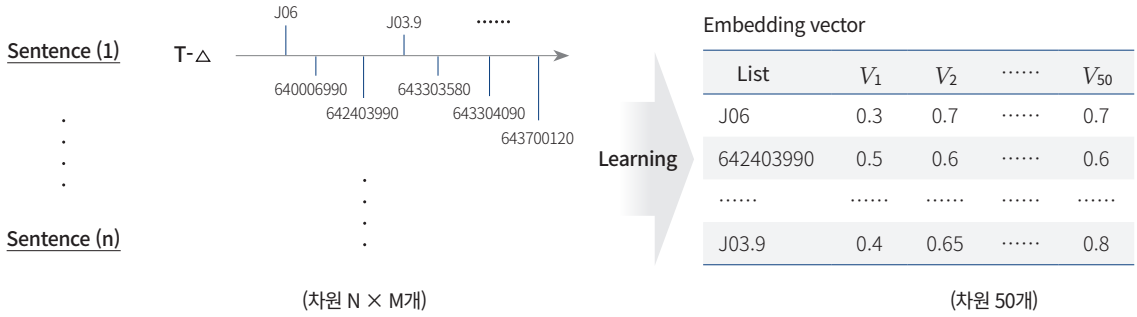


그림 3. 단어 임베딩 예시. 주: 예제로서 임의의 상황을 가정함.

(3) 순환신경망

순환신경망(recurrent neural network)은 과거 및 시퀀스(sequence) 정보 등을 학습하기 위한 인공신경망이다. 입력층, 은닉층, 출력층 외 메모리 셀이 있어 이후 시점에 과거 정보를 반영하는 역할을 하게 된다[41] (그림 4). 그리고 순환신경망의 장기 의존성 문제[12]를 보완한 LSTM (long short-term memory), GRU (gated recurrent unit) 방법이 발표되었고, 최근 성능이 개선된 트랜스포머(transformer) 등의 방법론이 추가 개발되고 있다.

(4) SHAP

Shapley additive explanation (SHAP)은 복잡한 예측모형에서의 원인에 대한 설명 등에 대한 제한(black-box)을 해결하기 위해 개발된 방법이다. SHAP을 통해 모든 독립변수들의 조합들이 모델에 미치는 영향을 수치화하여 예측결과에 대한 각 특성들의 기여도를 계산할 수 있다.

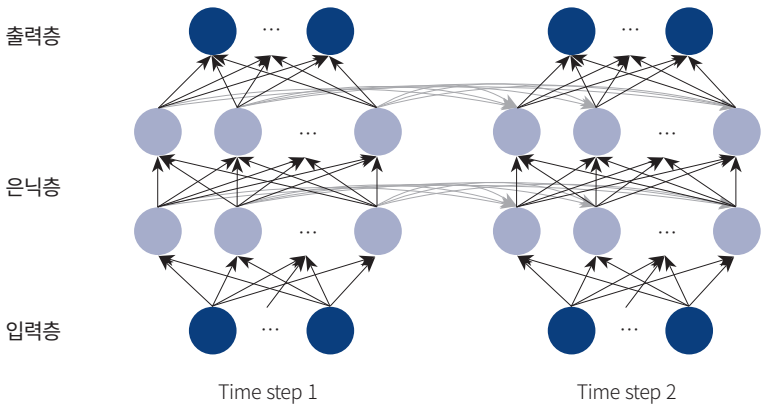


그림 4. Recurrent neural network. 자료: 유원준. 딥 러닝을 이용한 자연어 처리 입문. [발행지 불명]: 위키독스; 2021 [41].

12 더 먼 과거 정보 혹은 시퀀스일수록 현재 시점 값에 영향을 주지 못하는 문제

2) 해외 보건의료분야 인공지능 방법론 활용사례

최근 보건의료분야에서 다양한 딥러닝 분석기법을 활용한 연구가 발표되었는데, 이 중 청구자료 및 EHR 데이터를 이용하여 심층, 순환신경망 및 딥러닝 기반 단어 임베딩 기법을 사용한 연구들이 많은 비중을 차지하고 있다.

2016년 Choi 등[14]은 1년 동안 한 환자에게서 발생되는 행위코드, 약제코드, 국제질병사인분류코드, 진단검사코드를 하나의 문장으로 정의하여 단어 임베딩 기법을 적용하였고, 도출된 벡터값에 기반하여 코드 간의 유사도를 계산하였다. Choi 등[14], Che 등[16], Nagata 등[15], Jin 등[17]도 EHR 및 청구자료, 건강검진 자료 내 행위, 약제 코드, 진단검사코드의 조합들을 문장으로 정의하였고, 단어 임베딩 기법을 이용해 100-200개의 벡터로 매칭하였다. 그리고 이를 질환 발생(심부전, 당뇨) 예측 모델링의 설명변수로 활용하여 예측의 정확도를 향상시켰다. Zhang 등[18]은 EHR 자료에서 환자의 진료에서 발생된 코드 내역들을 이용해 환자단위 시계열 예측모델인 Patient2Vec을 제안하였고, 이를 이용해 조기 재입원을 예측한 결과 기존 시계열모형보다 예측정확도가 높게 나타났다.

3. 인공지능 방법론 활용 전략

통계적 모델링 방법은 선형성 가정 기반 간결하게 현상을 설명하고 해석하는 데 강점이 있지만, 정보가 복잡할수록 설명(예측)모델의 적합에 문제가 발생된다. 특히 예측의 관점에서는 정확도를 위해 복잡하고, 충분한 정보의 활용이 필요하여 활용이 어렵다. 반면, 인공지능 기반 모델링 방법은 복잡하고 방대한 정보에서 패턴을 찾는 것에 강점이 있어 예측분석에 주로 활용된다. 그러나 통계적 모델링보다 현상 등 결과의 해석이 어렵다는 제한점[13]을 가진다. 따라서 연구(분석)의 내용에 따라 적절한 방법론에 대한 선정이 필요하다. 본 장에서는 건강보험 청구자료 기반 분석 시 인공지능 방법론 적용 필요 검토단계를 설명하였다.

첫 번째, 모델의 목적이 예측이라면 인공지능 방법론 적용 검토가 필요하다. 일반적으로 예측의 경우 많은 정보의 조합을 통해 패턴을 학습하게 되므로 선형관계 중심인 통계적 방법론보다 일반적으로 좋은 성능을 보인다. 두 번째, 예측(또는 설명)하고자 하는 종속변수가 복잡한 인과관계를 가질수록 인공지능 방법론의 적용 검토가 필요하다. 예를 들어, 질병예측모델을 만든다고 가정했을 때 타겟이 특정 외래 경증질환일 때 보다 중증질환일 때 과거 진료내역, 환자상태 등 더 많은 요인에 대한 복합적 검토(학습)가 필요하게 된다. 세 번째, 많은 범주를 가진 진료내역 코드 정보를 설명변수로 활용하고자 할 때 인공지능 방법론 적용 검토가 필요하다. 진단명, 처치 및 수술, 검사, 약제처방 등의 진료내역 정보는 텍스트(코드) 정보로, 최소 1,000여 개의 카테고리로 구분된다. 이를 앞서 소개한 인공지능 기반 단어 임베딩 방법을 통해 축소된 N차원 벡터로 매칭이 가능하고, 생성된 벡터값을 통해 각 범주 간 유사도도 계산이 가능해진다. 네 번째, 과거(혹은 이전 시퀀스) 정보의 반영 여부 그리고 이 정보의 활용방법에 따른 검토가 필요하다. 과거 정보를 모델의 설명변수로 활용할 수 있는 방법은 과거 이력을 하나의 정보로 축약하거나 시점별 과거 이력 정보를 있는 그대로 모델에 반영하는 두 가지로 크게 구분된다. 전자의 대표적인 예로 중증도 점수(Charlson comor-

13 앞서 소개한 SHAP 등 모델의 해석력을 보완한 방법론들이 개발되고 있지만, 상대적으로 선형관계 중심 모델보다 해석이 복잡함

bidity index 등)를 들 수 있다. 만약 반영할 과거 정보의 패턴이 복잡하거나 축약이 불가능할 경우 인공지능 방법론 적용 검토가 필요하다. 더욱이 앞서 언급한 것처럼 건강보험 청구자료의 정보 발생시점이 불규칙하여 종단면(혹은 시계열) 분석 시 더욱 복잡한 모델링 방법이 필요하다.

고 찰

보건의료 빅데이터인 건강보험 청구자료는 전체 인구 중 약 98%의 의료이용 정보가 축적되어 있고 [1], 이는 보건의료정책 근거자료 생성, 보건의료분야의 다양한 연구 및 의료계 및 산업계의 R&D 개발에 활용되고 있다. 건강보험 청구자료를 이용한 연구들을 정리하면 단면적 분석 중심의 현황 및 추세 분석연구, 특정 요인에 대한 설명(영향) 및 비교분석 연구, 과거 정보를 고려한 종단면, 시계열 분석연구로 구분할 수 있고, 주로 통계적 방법이 이용되었다. 그러나 복잡 정보, 텍스트 정보, 불규칙 시계열 정보 등의 활용에 있어 선형성, 간결성에 기반한 전통적 모델링 방법으로는 이런 문제점들을 해결하기 어려워 최근 인공지능 기술을 활용하여 이런 문제들을 극복하려는 시도가 지속적으로 이루어지고 있다. 이에 본 연구에서는 기존 국내 연구들을 분석하여 활용범위 및 분석 제한점을 검토하고, 이를 보완할 수 있는 인공지능 기술 활용 전략을 네 가지 관점에서 제안하였다. 이는 국내 보건의료 빅데이터의 활용범위를 확장시키고 4차 산업의 주요 기술인 인공지능 기술 관련 연구의 활성화에 기여할 수 있을 것으로 기대한다.

끝으로, 현재까지 데이터 분석을 위한 다양한 방법론이 개발되었고, 각각의 장점과 제한점이 존재한다. 따라서 연구내용 및 상황에 따라 적정 방법론 선정을 위한 검토는 매우 중요하다. 최근 4차 산업의 주요 기술인 인공지능 기술에 대한 관심이 높아졌는데, 신중한 검토 없이 새로운 기술 도입 시 모델의 성능 측면이나 효율성 측면에서 안 좋은 결과가 발생될 수 있다. 본 연구에서는 큰 틀에서 인공지능 분석기법이 필요한 상황에 대해 정리하였고, 향후 세부적인 필요 분야와 건강보험 청구자료 내 주요 설명변수 및 적정 활용방법 등에 대한 연구 및 논의가 필요하다.

ORCID

Sang-A Cho: https://orcid.org/0000-0002-0772-3039

Hansang Kim: https://orcid.org/0000-0001-7347-7342

참고문헌

1. Kim JA, Yoon S, Kim LY, Kim DS. Towards actualizing the value potential of Korea Health Insurance Review and Assessment (HIRA) data as a resource for health research: strengths, limitations, applications, and strategies for optimal use of HIRA data. J Korean Med Sci. 2017;32(5):718-28. DOI: https://doi.org/10.3346/jkms.2017.32.5.718.
2. Kim JA, Park J, Kim BY, Kim DS. The trend of acute respiratory tract infections and antibiotic prescription rates in outpatient settings using health insurance data. Korean J Clin Pharm. 2017;27(3):186-94. DOI: https://doi.org/10.24304/kjcp.2017.27.3.186.
3. 한승진, 이근정, 조도연, 조상아, 박다혜, 엄혜은 등. 의료이용 추이 모니터링 고도화 연구. 원주: 건강보험심사평가원; 2020.
4. 김동숙, 박주희, 이근우, 최지원 등. 항생제 사용량 심층분석 및 내성정보 연계방안 검토. 원주: 건강보험심사평가원; 2017.

5. Kang CY, Kim HJ, Kim JH, Hwang JS, Lee DH. Outcomes analysis for Western medicine and Korean medicine using the propensity score matching in allergic rhinitis: data from the Health Insurance Review and Assessment Service. J Korean Med Ophthalmol Otolaryngol Dermatol. 2021;34(2):53-69. DOI: https://doi.org/10.6114/jkood.2021.34.2.053.
6. 이성우, 한승진, 안보령, 진다빈, 박진관. 치매환자의 의료이용 분석. 원주: 건강보험심사평가원; 2018.
7. 이도경, 유혜림, 김지우, 조상아. 상급종합병원 회송 환자 의료이용 분석 및 개선방안. 원주: 건강보험심사평가원; 2020.
8. 신민선, 이풍훈, 장원모. 코로나19 유행 시기의 사망과 의료이용 변화에 대한 탐색적 연구. 원주: 건강보험심사평가원; 2021.
9. Ryu S, Nam HJ, Kim JM, Kim SW. Current and future trends in hospital utilization of patients with schizophrenia in Korea: a time series analysis using national health insurance data. Psychiatry Investig. 2021;18(8):795-800. DOI: https://doi.org/10.30773/pi.2021.0071.
10. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1:18. DOI: https://doi.org/10.1038/s41746-018-0029-1.
11. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2017;24(1):198-208. DOI: https://doi.org/10.1093/jamia/ocw042.
12. Cai X, Gao J, Ngiam KY, Ooi BC, Zhang Y, Yuan X. Medical concept embedding with time-aware attention [Internet]. Ithaca (NY): arXiv.org; 2018 [cited 2019 Dec 4]. Available from: https://arxiv.org/abs/1806.02873.
13. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236-46. DOI: https://doi.org/10.1093/bib/bbx044.
14. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: predicting clinical events via recurrent neural networks [Internet]. Ithaca (NY): arXiv.org; 2016 [cited 2019 Sep 15]. Available from: https://arxiv.org/abs/1511.05942.
15. Nagata M, Takai K, Yasuda K, Heracleous P, Yoneyama A. Prediction models for risk of type-2 diabetes using health claims. Proceedings of the BioNLP 2018 Workshop; 2018 Jul 19; Melbourne, Australia. Stroudsburg (PA): Association for Computational Linguistics; 2018. pp. 172-6.
16. Che Z, Cheng Y, Sun Z, Liu Y. Exploiting convolutional neural network for risk prediction with medical feature embedding [Internet]. Ithaca (NY): arXiv.org; 2017 [cited 2019 Sep 15]. Available from: https://arxiv.org/abs/1701.07474v1.
17. Jin B, Che C, Liu Z, Zhang S, Yin X, Wei X. Predicting the risk of heart failure with EHR sequential data modeling. IEEE Access. 2018;6:9256-61. DOI: https://doi.org/10.1109/ACCESS.2017.2789324.
18. Zhang J, Kowsari K, Harrison JH, Lobo JM, Barnes LE. Patient2vec: a personalized interpretable deep representation of the longitudinal electronic health record. IEEE Access. 2018;6:65333-46. DOI: https://doi.org/10.1109/ACCESS.2018.2875677.
19. 최지숙, 채송이, 박기찬. 2018년 기준 의료이용 현황 연구. 원주: 건강보험심사평가원; 2018.
20. 김한상, 조상아, 김수민. 대형병원 중증도 중심 의료이용 분석 및 모니터링 체계 마련 연구. 원주: 건강보험심사평가원; 2020.
21. 오동관, 손강주. 한방병원 입원일수 관리지표 개발 및 관

리방안 연구. 원주: 건강보험심사평가원; 2015.

22. Park JM, Kang CD, Lee JC, Hwang JH, Kim J. Recent 5-year trend of endoscopic retrograde cholangiography in Korea using National Health Insurance Review and Assessment Service open data. Gut Liver. 2020;14(6):833-41. DOI: https://doi.org/10.5009/gnl19249.
23. Kim HY, Chang SA, Kim KH, Kim JY, Seo WK, Kim H, et al. Epidemiology of venous thromboembolism and treatment pattern of oral anticoagulation in Korea, 2009-2016: a nationwide study based on the National Health Insurance Service Database. J Cardiovasc Imaging. 2021;29(3):265-78. DOI: https://doi.org/10.4250/jcvi.2021.0014.
24. 김동숙, 조현민, 박주희, 조도연, 문경준, 변지혜 등. 암환자 사용약제 보장성 강화정책 효과분석. 원주: 건강보험심사평가원; 2017.
25. 박효성, 엄태웅, 김나권. 얼굴마비 환자의 의·한의 협진 의료이용 연구: 건강보험심사평가원 환자표본 데이터를 이용. 한국데이터정보과학회지. 2017;28(1):75-86. DOI: https://doi.org/10.7465/jkdi.2017.28.1.75.
26. Kim JW, Lee JH, Kim TG, Kim YH, Chung KJ. Breast reconstruction statistics in Korea from the Big Data Hub of the Health Insurance Review and Assessment Service. Arch Plast Surg. 2018;45(5):441-8. DOI: https://doi.org/10.5999/aps.2018.00220.
27. An SH, Youn MK, Kim IY. Effect of laparoscopic surgery on the risk for surgical site infections in colorectal resection: results from the Health Insurance Review & Assessment Service Database. Ann Surg Treat Res. 2020;98(6):315-23. DOI: https://doi.org/10.4174/astr.2020.98.6.315.
28. Lee Y, Jo M, Kim T, Yun K. Analysis of high-intensity care in intensive care units and its cost at the end of life among older people in South Korea between 2016 and 2019: a cross-sectional study of the Health Insurance Review and Assessment Service National Patient Sample Database. BMJ Open. 2021;11(8):e049711. DOI: https://doi.org/10.1136/bmjopen-2021-049711.
29. Ko YR, Lee SR, Kim SH, Chae HD. Pelvic organ prolapse is associated with osteoporosis in Korean women: analysis of the Health Insurance Review and Assessment Service National Patient Sample. J Clin Med. 2021;10(16):3751. DOI: https://doi.org/10.3390/jcm10163751.
30. Kim DH, Ha DJ, Lee YS, Chun MJ, Kwon YS. Benign convulsions with mild rotavirus and norovirus gastroenteritis: nationwide data from the Health Insurance Review and Assessment Service in South Korea. Children (Basel). 2021;8(4):263. DOI: https://doi.org/10.3390/children8040263.
31. 민인순, 김선정, 함명일, 이윤노, 김복미, 김동준 등. 전문병원 지정 및 평가체계 개선방안연구. 원주: 건강보험심사평가원, 순천향대산학협력단; 2017.
32. 김지애, 김수민, 김한상, 유혜림. COVID19 대응을 위해 한시적으로 허용된 전화상담·처방 효과 분석. 원주: 건강보험심사평가원; 2020.
33. 오주연, 최효정, 박다혜, 신양준, 임재우, 이다희 등. 호스피스 완화의료서비스 제도 개선 방안: 유형별 연계 강화 및 환자중심의 통합적 이용활성화 방안을 중심으로. 원주: 건강보험심사평가원; 2020.
34. 김한상, 박다혜, 안보령, 박주희, 박진관. 의료이용 모니터링 지표개발 연구. 원주: 건강보험심사평가원; 2020.
35. Lee N, Lee JD, Lee HY, Kang DR, Ye YM. Epidemiology of chronic urticaria in Korea using the Korean Health Insurance Database, 2010-2014. Allergy Asthma Immunol Res. 2017;9(5):438-45. DOI: https://doi.org/10.4168/aair.2017.9.5.438.
36. 박찬미, 장수현, 장선미. 치료지속성에 따른 의료이용 및 건강결과 분석. 원주: 건강보험심사평가원; 2010.

37. Lee GI, Lim DH, Chi SA, Kim SW, Shin DW, Chung TY. Risk factors for intraocular lens dislocation after phacoemulsification: a nationwide population-based cohort study. Am J Ophthalmol. 2020;214:86-96. DOI: https://doi.org/10.1016/j.ajo.2020.03.012.
38. Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. IEEE Comput Intell Mag. 2018;13(3):55-75. DOI: https://doi.org/10.1109/MCI.2018.2840738.
39. Harris ZS. Distributional structure. Word 1954;10(2-3):146-162. DOI: https://doi.org/10.1080/00437956.1954.11659520.
40. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space [Internet]. Ithaca (NY): arXiv.org; 2013 [cited 2019 Sep 15]. Available from: https://arxiv.org/abs/1301.3781.
41. 유원준. 딥 러닝을 이용한 자연어 처리 입문. [발행지 불명]: 위키독스; 2021.