의약품 처방 데이터 기반의 질병 발생 예측에 관한 연구

장정현[1,0] 김영재[1] 최종혁[1] 김창수[2] 나스리디노프 아지즈[1]
[1]충북대학교 소프트웨어학과
[2]배재대학교 컴퓨터공학과
sky456139@chungbuk.ac.kr, juk1413@naver.com, leopard@chungbuk.ac.kr, ddoja@pcu.ac.kr, aziz@chungbuk.ac.kr

A Study on Medicine Prescription Data-based Disease Occurrence Predictions

Jeong-Hyeon Chang[1,O] Young-Jae Kim[1] Jong-Hyeok Choi[1] Chang-Soo Kim[2] Nasridinov Aziz[1]
[1]Department of Software, ChungBuk National University
[2]Department of Computer Engineering, Pai Chai University

요 약

현재 공공 보건 의료 빅데이터의 개방 정책이 확대됨에 따라 이전에는 접근이 어려워 연구에 난항을 겪었던 보건 의료 분야의 데이터 분석이 주목 받고 있으며, 다양한 형태로 활용되어 서비스가 제공되고 있다. 특히 국민건강보험공단의 경우, 현재 국민건강알림서비스란 이름의 질병 예측 서비스를 제공하고 있지만 해당 서비스의 경우, 제한된 종류의 질병에 대한 단기의 예측만을 제공하는 문제를 지니고 있다. 본 연구에서는 기존 서비스의 질병 종류 및 기간의 한정으로 인한 문제를 해결하기 위해 의약품 사용정보 서비스, 인구 통계정보 서비스, 보건의료 빅데이터 서비스를 통해 데이터를 수집한 후, 분석을 통해 각 질병의 지역별, 연월별 환자 수 예측 모델을 제안한다. 또한, 예측 모델의 효용성을 파악하기 위해 여러 질병 중 감기에 대한 예측 모델을 생성한 후 예측 결과와 실제 환자 수와의 비교를 통해 오차율을 분석한 결과, 제안된 예측 모델의 오차율은 약 4%로 매우 높은 정확도의 예측을 수행함을 알 수 있었다.

1. 서 론

빅데이터는 현재 많은 분야에서 활용되고 있으며, 특히나 최근에는 보건의료 분야에서의 빅데이터 활용이 활발해지고 있다. 예를 들어 현재 싱가포르의 경우, 보건의료 빅데이터를 이용한 전염성 질병 예방 모델인 RAHS를 구축하여 활용하고 있으며, 영국의 경우는 Foresight Horizon Scanning 센터를 설립하여 보건의료 빅데이터를 이용한 전염성 질병 대응을 실시하고 있다[1]. 최근 우리나라에서도 보건의료 분야의 빅데이터를 이용한 서비스를 개발하고 있으며, 이에 따른 다양한 활용 사례와 수요가 늘어나고 있다[4]. 이들 서비스의 대표적 예로서 현재 국민건강보험공단의 '국민 건강 알람 서비스'의 경우, SNS 및 여러 의료 데이터를 활용하여 감기, 눈병, 식중독, 천식, 피부염 등 5가지 질병에 대해 발생 및 전염 위험에 대한 안내 서비스를 제공하고 있다. 하지만 해당 서비스의 경우 한정된 5가지의 질병, 그리고 3일 이내의 단기예측만 제공한다는 문제점을 갖고 있으며, 각 질병별 예측식이 선형형태로 이루어지기 때문에 계절성 질병에 대한 낮은 예측력을 갖는 문제가 있다.

본 논문에서는 이와 같은 기존의 문제점을 해결하기 위해 의약품 처방 데이터와 보건의료 빅데이터, 인구통계정보를 활용하여 다양한 질병에 대한 지역별, 연월별 예상 질병 환자수를 예측하는 새로운 질병 예측 모델을 제안한다. 또한 연월별 실제 환자수를 바탕으로 예측 모델의 오차율 검증을 진행하여 예측력에 대한 신뢰성을 확인한다.

2. 관련 연구

"빅데이터 분석을 통한 개인 형 맞춤 의료 대책 방안 연구"[1]에서는 2010~2014년도까지의 1군~3군까지의 법정 전염병 데이터를 분석하였다. 하지만 해당 논문에서는 전염병 데이터 분석과정에서 연도별, 월별은 고려하였으나 기후, 환경, 성별, 나이, 지리적 조건 등 다양한 요소들을 고려하지 못한 한계점을 시사하고 있다.

"기상 기후 및 질병 빅데이터 기반의 질병 예측 및 건강 정보 어플리케이션 구현"[2]에서는 기상 기후 정보와 질병에 대해 다중회귀분석을 통해 각 질병별 발병 가능성을 예측하는 연구를 진행했다. 하지만 다양한 요인을 고려하지 않고 기상 기후만을 고려한 점과 각 실험결과와 실제 발병 환자 수와의 정확성 검증이 이루어지지 않아 연구 결과에 대한 신뢰성을 보장하지 못했다.

"의료정보 빅데이터 분석을 통한 개인 맞춤형 유의 질병 및 병원정보 앱 서비스 개발"[3]에서는 보건의료 빅데이터와 총 60여 종의 관계형 텍스트 활용하여 대용량 데이터 분석 모델을 개발하였다. 해당 연구에서의'유의 질병 확인' 기능은 사용자와 진료 및 처방 패턴을 바탕으로 사용자가 걸릴 가능성이 높은 질병을 안내한다. 이러한 패턴 기반의 방법은 환자간의 진료 및 처방 패턴이 유사하다는 관점에서만 접근하였으며 여러 의학적인 요소들을 고려하지 못했다. 이는 환자의 현재 상태나 여러 의학적 상황에 따라 예민하게 반응하는 질병은 처방 정보가 규칙적인 패턴을 가지지 않으므로 서비스를 이용하는 사용자의 혼란을 초래할 수 있는 문제를 지니고 있다.

3. 제안 방법

3.1 데이터 수집

본 연구에서는 질병 예측을 위해 국민건강심사평가원에서 수집하고 관리되는 의약품 사용정보[5]와 보건의료 빅데이터 개방시스템을 통해 제공되는 국민관심 질병통계정보[6], 통계청에서 관리되는 통계지리 정보[7]를 사용하였다. 의약품 사용정보는 연월별 도시별, 의약품 분류 국제 표준인 ATC 코드별

의약품 처방 건수 정보를 제공하고 있다. 통계지리정보의 경우 각 도시별 인구 정보를 2000년을 기준을 5년 주기로 정보를 제공하고 있으며, 국민관심 질병 통계 정보는 전체 인구를 대상으로 연월별 관심질병별 환자수를 집계하여 제공하고 있다.

본 연구에서는 상기 서술한 데이터를 2010~2017년까지의 데이터를 수집하여 활용하였으며, 통계 지리 정보의 경우 2000년도부터 5년 주기로 제공되는 데이터를 사용하였다.

3.2 데이터 전처리

분석을 위해 수행된 데이터 전처리는 다음과 같다.

1) 통계지리정보의 행정구역 코드 및 의약품 사용 정보의 행정구역 코드의 맵핑 작업 실시
2) 수집된 의약품 사용 데이터에서 중복 또는 누락 데이터 제거
3) 연월별, 시도 및 시군구별로 수집된 의약품 사용 데이터로부터 구 이하에 대한 의약품 사용 데이터 제거

3.3 분석 절차

본 논문에서 제안하는 의약품 사용 정보 기반의 질병 예측 방법의 분석 절차는 다음과 같다.

1) 의약품 사용 정보에서 연월별 질병에 따른 의약품의 총 처방 건수를 수집
2) 보건의료 빅데이터 서비스로부터 연월별 질병에 따른 실제 총환자수 정보를 수집
3) 수집된 두 데이터를 대상으로 상관분석을 통한 상관계수 및 상관관계확인
4) 약품 처방건수와 실제 총환자수를 대상으로 회귀분석 실시
5) 회귀분석 결과를 바탕으로 약품 처방 수에 따른 실체 총 환자수를 예측하는 선형회귀식 산출
6) 질병에 따른 의약품 총 처방 건수를 상기 선형회귀식에 대입하여 해당 질병의 예상 총 환자수를 예측
7) 예측된 총환자수와 실제 총 환자 수를 비교분석하여 결과 검증

또한 도시별 인구 구성비의 차이를 반영하기 위한 인구 대비 질병 발생 가중치 산출 방법은 다음과 같은 순서로 진행된다.

1) 연월별, 도시별 질병에 따른 의약품 사용 정보 수집
2) 연월별, 도시별로 의약품 처방건수를 합산하여 총 처방 수를 계산
3) 연월별, 도시별 총 의약품 처방 건수와 해당 도시의 인구수 대비 의약품 처방 가중치를 산출
4) 해당 연월별 전체 도시의 가중치를 합하여 월별 전체 가중치 값 산출
5) 연월별 해당 도시의 가중치 값과 같은 연월별 전체 가중치 값을 통해 각 도시별 가중치 퍼센트 값을 산출

그리고 일부 질병이 갖는 계절특성을 반영하기 위한 연월별 질병별 도시별 예상 총환자수 예측 방법은 다음과 같은 순서로 예측된다.

1) 연월별, 질병별 의약품 처방 건수와 실제 총 환자수를 통해 도출된 회귀식을 이용하여 해당 질병의 연월별 예상 총 환자수의 예측을 실시
2) 연월별 해당 질병의 총 환자 수를 같은 연월의 각 도시의 가중치 퍼센트를 이용하여, 각 도시별 예상 환자수의 예측을 실시

3.4 제안 수식

본 절에서는 본 논문을 통해 제안되는 질병별 환자 수 예측 모델에서 사용되는 다양한 수식들을 나타낸다.

본 논문을 통해 제안되는 질병별 환자 수 예측 모델에서는 질병별 월간 의약품 처방건수를 바탕으로 해당 질병의 예상 총 환자 수를 예측하며, 이를 위한 환자 수 예측 계산식은 다음의 [수식 1]과 같다.

$$predictPatient = A + B \times monthlyMedicineAmount$$

[수식 1] 처방건수에 따른 환자 수 예측(A, B는 회귀상수)

이때 의약품 처방건수는 해당 도시의 인구수에 비례하여 증가하는 특성을 갖고 있으며, 인구수 대비 높은 처방량은 해당 질병의 높은 위험도를 보인다 할 수 있다. 따라서 각 시별 인구수 대비 의약품 처방의 편차를 고려하는 경우 지역별 위험성을 도출 할 수 있으며 이를 위한 월별 시별 총 인구 대비 질병 발생 가중치 계산식은 [수식 2]와 같다.

$$Weight = Round[(monthlyMedicineAmount \div population) \times 100,1]$$

[수식 2] 월별 시별 총 인구 대비 질병 발생 가중치

그리고 각 도시별 가중치를 퍼센트로 변환하여 지역별 의약품 사용 편차에 활용하기 위해 월별 각 도시의 전체 가중치 합 계산이 필요하며 이를 위한 월별 전체 가중치 합 계산식은 [수식 3]과 같다.

$$total\,Weight = Round[(\sum weight\ of\ per\ month),1]$$

[수식 3] 월별 가중치 합

또한 앞서 계산된 월별 가중치 합을 바탕으로 각 도시별 예상 총 환자수를 계산하기 위한 각 시별 가중치 퍼센트 계산식은 다음 [수식 4]와 같다.

$$weightPercent = Round[(weight\ of\ per\ City \div total\,Weight) \times 100,3]$$

[수식 4] 가중치 퍼센트

연월별 각 시별 예상 총 환자수를 용이하게 계산하기 위해 예상 총 환자수로부터 1%에 해당하는 환자 수만을

추출한다.

$$patient\,Amount\,Of\,Percent = Round[predictPatient \div 100,1]$$

[수식 5] 예상 총 환자수의 1%당 환자 수

의약품 처방 건수에 따른 전체 예상 총 환자수를 계산하는 [수식 5]과 각 도시별 의약품 처방 편차 가중치 값을 퍼센트로 바꾼 [수식 4]를 이용하여 각 시별 예상 총 환자 수를 계산하는 시별 예상 환자 총 환자 수 계산식은 [수식 6]과 같다.

$$predictPatientOfCity = Round[(weightpercent \times patient\,Amount\,Of\,Percent),1]$$

[수식 6] 연월별 각 시별 예상 총 환자 수

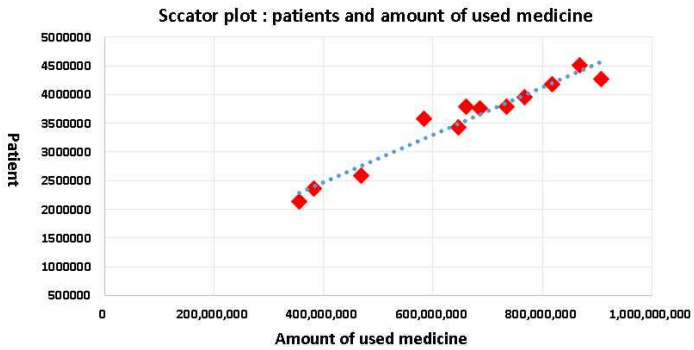
마지막으로 의약품 처방 수 와 실제 환자수를 바탕으로 [수식 1]을 통해 예측한 연월별 전체 총 예상 환자수와 [수식 4]를 이용하여 연월별 각 도시별 예상 총 환자수의 예측 결과를 실제 결과와 비교하여 오차를 계산하기 위한 계산식은 [수식 7]과 같다.

$$Average\,error\,Rate = AVG[(|realPatient - predictPatient| \div realPatient) \times 100]$$

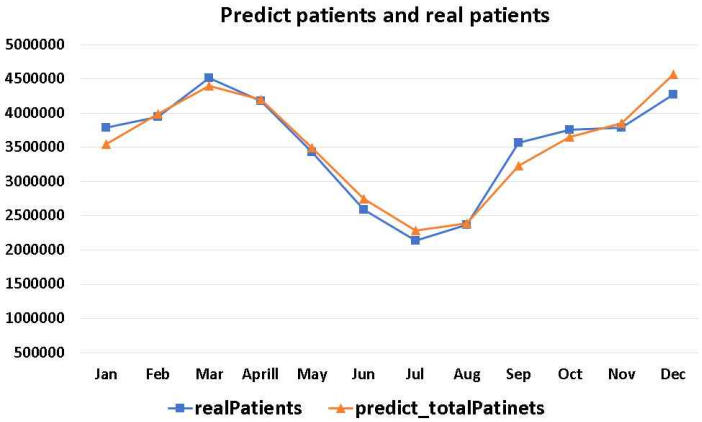
[수식 7] 예측 오차율

4. 실험 조건 및 분석방법

본 논문에서 제안하는 방법을 통해 의약품 처방 건수에 따른 지역별 환자수를 예측하기 위해 진행한 실험에서는 예측 모델의 효용성을 알기 위해 발병 대상이 일반적이고 높은 유행성을 띄는'감기'질병을 대상으로 진행하였다. 실험에는 연월별'감기'관련 의약품 처방 데이터 6,6291건, 도시별 인구통계 정보 83건, 보건의료 빅데이터 12건이 사용되었으며 모두 2015년도에 수집된 데이터를 사용하였다.



[그림 1] 의약품 처방건수와 환자 수 사이의 상관도



[그림 2] 감기 질병 예측 및 실제 환자 수 비교 그래프

[표 1] 월별 감기 예상환자 수와 실제 환자 수 및 오차율

연월	실제 환자 수	예상 환자 수	오차율
2015-01	3,792,925	3,544,762	6.5%
2015-02	3,949,827	3,985,774	0.9%
2015-03	4,512,010	4,401,728	2.4%
2015-04	4,177,475	4,197,525	0.5%
2015-05	3,433,044	3,489,382	1.6%
2015-06	2,585,334	2,749,958	6.4%
2015-07	2,139,986	2,287,730	6.9%
2015-08	2,365,389	2,394,181	1.2%
2015-09	3,570,879	3,225,298	9.7%
2015-10	3,754,740	3,650,522	2.8%
2015-11	3,788,978	3,848,913	1.6%
2015-12	4,274,347	4,568,075	6.9%

[표 2] 각 시별 예상 환자 수 및 실제 환자 수와 오차율

연월	도시	실제 환자수	예상 환자수	오차율
2015-01	청주시	46,274	43,246	6.5%
2015-02	청주시	44,238	44,641	0.9%
2015-03	청주시	57,303	55,902	2.4%
2015-04	청주시	52,636	52,889	0.5%
2015-05	청주시	41,197	41,873	1.6%
2015-06	청주시	29,731	31,625	6.4%
2015-07	청주시	26,108	27,910	6.9%
2015-08	청주시	29,567	29,927	1.2%
2015-09	청주시	47,850	43,219	9.7%
2015-10	청주시	48,436	47,092	2.8%
2015-11	청주시	48,499	49,266	1.6%
2015-12	청주시	54,284	58,015	6.9%

5. 실험 결과

[그림 1]은 의약품 처방건수와 환자 수 사이의 상관도를 그림으로 나타낸 결과다. 해당 실험을 통해 우리는 의약품 처방건수와 실제 환자 수 사이에는 높은 관련이 있음을 알 수 있다. 또한 이들 사이의 상관계수는 0.97로 강한 양의 상관관계를 가지고 있음을 알 수 있었다.

다음의 [표 1]은 월별 감기 예상 총 환자 수를 계산하기 위한 [수식 1]을 적용한 후 이에 대한 오차율을 판단하기 위해 오차율 계산식인 [수식 7]을 적용한 환자 수 예측 결과다. 이러한 예측 결과, 본 논문에서 제안된 환자 수 예측 모델의 경우 매우 근사한 환자수를 예측함을 알 수 있었으며, 실제 환자수와의 오차율은 최소 0.5%에서 최대 9.7%, 평균 3.9%로 매우 적은 오차를 보임을 알 수 있다. 또한 다음의 [그림 2]는 이와 같은 감기에 대한 제안 모델의 예측 환자 수와 실제 환자 수의 차이를 그래프를 통해 나타낸 그림으로, 예측 환자 수 그래프와 실제 환자 수 그래프는 매우 유사한 형태를 가지는 것을 확인할 수 있다. 이와 같은 결과는 본 논문을 통해 제안된 환자 수 예측 모델의 높은 정확성을 보이는 결과라 하겠다.

[표 2]의 실험 결과는 앞서 제안된 [수식 6]을 적용하여 계산한 청주시의 예상 감기 환자수와 함께 해당 오차율을 나타낸다. [표 2]의 분석 결과는, 본 논문을 통해 제안된 질병별 환자 수 예측 모델이 전국 단위뿐만 아니라 지역 단위에서도 실제 환자 수와 매우 유사한 예측을 수행할 수 있음을 보이는 결과라 할 수 있다.

6. 결론

본 논문에서는 기존 시스템이 특정 질병에 대하여 단기적인 예측을 제공한다는 문제점을 해결하기 위해 의약품 처방 데이터와 인구통계정보, 보건의료 빅데이터 정보를 활용한 질병 발생 예측 방법에 대한 연구를 통해 새로운 예측 모델을 제안하였다. 또한 제안된 예측 모델의 효용성을 평가하기 위해 2015년도 기준 '감기'질병을 대상으로 예측을 수행한 결과 평균 3.9%의 매우 낮은 오차, 즉 96.1%의 높은 정확도를 보였다. 이러한 결과는 본 논문을 통해 제안된 질병별 환자수 예측 모델을 통해 연월별 시별 환자 수를 예측할 수 있음을 보이는 결과라고 하겠다.

본 논문에서 제안된 모델은 '감기' 질병 환자수에 대하여 예측을 수행한 결과가 실제 환자수와 매우 적은 오차를 보임을 알 수 있었으며, 향후 보다 다양한 질병을 활용한 실험 및 보완을 통해 질병별 환자 수 예측 모델로서의 연구를 계속 하고자 한다. 또한 향후의 연구에서는 제안된 질병 예측 모델의 예측 결과가 시계열 형태의 데이터인 점에 착안하여 시계열 데이터 기반의 예측 모델인 ARIMA 모델을 활용하여 다양한 질병에 대하여 보다 장기적인 예측이 가능한 모델로 확장하고자 한다.

ACKNOWLEDGEMENT

"본 논문은 교육부가 지원하고 충북대학교가 수행하는 지역선도대학육성사업의 지원을 받아서 수행되었습니다."

참고 문헌

[1] 허준구. 빅데이터 분석을 통한 개인 형 맞춤 의료 대책 방안 연구. 숭실대학교 정보과학대학원 석사학위 논문. pp.1-60. 2015

[2] 김기연, 김언찬, 노상현, 임동욱, 정진우. 기상 기후 및 질병 빅데이터 기반의 질병 예측 및 건강 정보 어플리케이션의 구현. KIIT Summer Conference. 496-497(2page). 2017

[3] 김성현, 황현석. 의료정보 빅데이터 분석을 통한 개인 맞춤형 유의질병 및 병원정보 제공 앱 서비스 개발. Entrue Journal of Information Technology. Vol.15 . No.2 . pp.7-16. 2016

[4] 송태민. 우리나라 보건복지 빅데이터 동향 및 활용 방안. 과학기술정책. 56-73(18page). 2013

[5] 의약품 사용정보, 보건의료 빅데이터 개방 시스템, http://opendata.hira.or.kr/op/opc/selectOpnsApi.do?sno=1406

[6] 국민관심 질병통계정보, 보건의료 빅데이터 개방 시스템, http://opendata.hira.or.kr/op/opc/olapMfrnIntrsIlnsInfo.do

[7] 통계지리 정보, SGIS PLUS 개발지원 센터, https://sgis.kostat.go.kr/developer/html/openApi/api/data.html