질병 예측 모델 개발을 위한 지리정보시스템(GIS)기반 데이터베이스 구축

장우영*, 우창우*, 송하림*, 손호선*, 류근호*, 김영규*
*충북대학교 의생명과학경영융합대학원
email : Wooyeong@dblab.chungbuk.ac.kr

GIS-based Database for Development of Disease Prediction Model

Wooyeong Jang*, Changwoo Woo*, Harim Song*, Ho Sun Shon*, Keun Ho Ryu*, YoungGyu Kim*
*Graduate School of Health Science Business Convergence, Chungbuk National University

요 약

국민소득 증가와 고령화 사회에 진입하면서 다양한 의료소비가 의료서비스산업에 영향을 미치고 있으며, 이러한 사회 구조 및 환경변화에 따라 새로운 질병에 대한 대응 또한 절실히 요구되고 있다. 질병 예측을 위한 연구는 기후변화와 질병, 건강행태와 질병, 사회적 위치와 질병 등 질병에 영향을 주는 많은 요인들이 있다. 그러나 이러한 요인들을 통합, 분석 활용하는 데는 해결해야 할 문제들이 많이 있다. '정부3.0 공공데이터 개방 정책'을 통해 질병에 관련된 자료가 공개 되면서 본 연구에서는 2010년부터 2012년까지의 질병에 영향을 주는 공공데이터를 연도별로 통합하여 지리정보시스템(GIS)기반 데이터베이스를 구축하고 활용 할 수 있게 하였다. 향후 기후변화에 민감한 질병을 찾기 위해 해당기관의 자료를 활용하여 월별로 데이터베이스를 구축하고, 이를 기반으로 의료서비스의 활성화 및 효율성에 기여 할 수 있다.

1. 서론

현대 사회는 데이터가 경제적 자산이 되는 시대로 공공부문, 산업(서비스), 개인 등의 데이터가 축척되면서 천문학적인 규모의 빅 데이터가 형성되고 있다. 이러한 빅 데이터의 공통적인 정의는 "현재 기술로 감당하기 어려운 규모로 양이 엄청나며, 데이터의 유형이 다양하기 때문에 처리 및 관리 기법이 필요한 데이터"이다[1].

정부는 공공정보를 적극적으로 개방하고 공유하는 '정부3.0 공공데이터 개방 정책'을 발표하고 적극적인 지원에 따라 많은 데이터들을 공개하고 있으며 이러한 공공데이터에 대한 필요성과 활용에 대한 방법론은 많이 제시되었지만 반정형화나 비정형화되어 있어 실제 활용은 일부 산업에 국한되어 있다[2].

2000년 이후 국민소득이 증가함에 따라 다양한 소비자 욕구가 의료서비스산업에서도 표면화되기 시작했다. 또한 고령화 사회의 도래, 급속한 사회적 변화에서 유발되는 스트레스의 증가, 식이생활 변화 등으로 기존 주요 질병들에서 새로운 질병으로 의료수요가 변화하고 있다. 이에 따라 새로운 질병에 대한 대응이 필요하다[3][4].

기존 연구를 살펴보면 기후변화에 따라 질병에 영향을 미치는 것을 알 수 있다[5][6][7]. 하지만 2007년 IPCC(Intergovernmental Panel on Climate Change) 보고서에 의하면 기후변화가 인간 환경에 미치는 다른 요인과의 식별이 쉽지 않고, 질병은 기후변화 뿐 아니라 건강 행태, 사회적 특성에 따라 차이가 있다는 연구도 많이 발표 되었다[8][9][10].

질병에 영향을 주는 요인이 이런 다양함에도 불구하고 많은 연구에서 한 가지 측면에서만 접근했기 때문에 정확한 예측이 어렵다.

본 논문에서는 이러한 문제점을 해결하기 위해 질병에 영향을 주는 여러 요인을 통합하여 지리정보시스템(GIS)기반으로 데이터베이스를 구축하고 활용할 수 있게 하였다.

2. 본론

2-1. 데이터 수집

<표 1>과 같이 질병에 영향을 주는 요인을 포함한 공공데이터를 우리나라의 시·도 행정구역별로 2010년에서 2012년 자료를 수집하였다.

<표 1> 데이터 수집 기관과 수집 정보

기관	수집정보	사이트
질병관리본부 -국민건강영양조사	건강상태, 사회적 위치	https://knhanes.cdc.go.kr
질병관리본부 -지역사회건강조사	건강행태, 사회적 위치	https://chs.cdc.go.kr
통계청 -통계지리정보서비스	지리	http://sgis.kostat.go.kr
기상청 -지상기상관측	기후	http://sts.kma.go.kr
통계청 -인구 주택 총 조사	인구	http://www.index.go.kr
건강보험심사평가원 -질병·행위통계	질병	http://www.hira.or.kr

국민건강영양조사는 제1기(1998)부터 제3기(2005)까지 3년 주기 2~3개월 단기조사체계로 실시하였으나 제4기(2007~2009)부터 연중조사 체계로 개편됨에 따라 3개년도가 각기 독립적인 3개의 순환표본으로 대표하는 순환표본조사방식을 도입했다.

제5기(2010~2012)는 제4기 표본설계와 같이 순환표본조사 방식을 적용하였으나 일반지역과 아파트지역을 분리하여 두 개의 표본추출틀을 사용했고, 층화변수로는 시도(16개)를 이용했다. 국민건강영양조사 자료는 회원가입 없이 자료를 받을 수 있다[11].

지역사회건강조사는 2008년 시작되어 목표 모집단은 대한민국 영토 내에 거주하는 만 19세 이상의 모든 성인 이며 주민등록주소자료와 국토해양부 주택유형 정보와 연계하여 표본추출틀을 사용했다. 그리고 층화변수로는 동·읍·면, 주택유형을 이용했다. 지역사회건강조사 자료는 자료이용계획서를 작성하고 검토 한 후 자료를 받을 수 있다[12].

센서스용 행정구역경계는 시·도 경계가 많은 곳에 공개되어있으며 읍·동·면 경계는 통계지리정보서비스에서 요청공문을 통해 자료를 받을 수 있다[13].

지역별 기상 관측지점은 전국 72개로 수도권 6개, 강원권 12개, 충청권 11개, 전라권 17개, 경상권 22개, 제주권 4개 이다. 시·도별 대표 지점은 <표 2>와 같이 선택 하였으며, 지점의 선택은 해당 시·도청의 위치에 근접한 정도를 기준으로 하였다. 기상청의 기상 관측 자료는 회원가입 없이 자료를 받을 수 있다[14].

<표 2> 시도별 관측지점

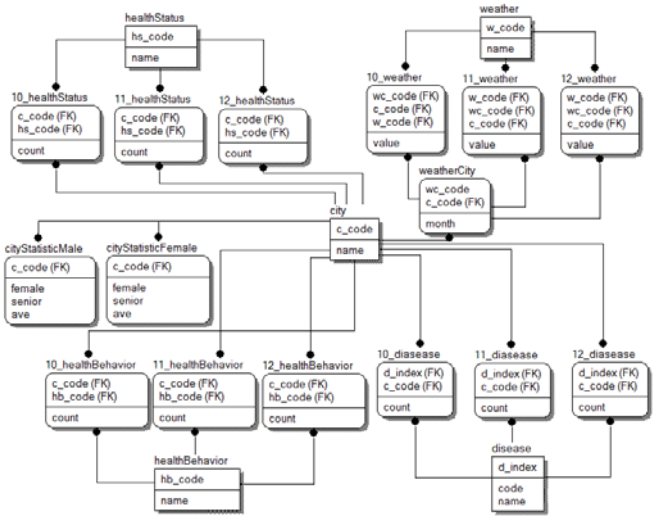
시·도	지점	시·도	지점
강원도	춘천	서울특별시	서울
경기도	수원	울산광역시	울산
경상남도	창원	인천광역시	인천
경상북도	상주	전라남도	목포
광주광역시	광주	전라북도	전주
대구광역시	대구	제주특별자치도	제주
대전광역시	대전	충청남도	서산
부산광역시	부산	충청북도	청주

인구 주택 총 조사는 5년 마다 실시되는 전국단위의 대규모 조사로 대한민국에 상주하는 모든 내·외국인들의 성별, 나이, 직업, 거주기간 등 인구 동향 등을 파악하기 위해 실시되는 조사이고 이러한 정보는 e-나라지표에서 받을 수 있다[15].

건강보험심사평가원에서는 최근 5년간 건강보험으로 발생한 모든 질병의 총 환자 수, 진료비 등을 질병 단위(소분류(3단 상병)) 별로 검색하고 자료를 받을 수 있다.

2-2. 데이터 통합

(그림 1)은 각 기관의 데이터를 통합한 통합데이터베이스 스키마이다.



(그림 1) 통합데이터베이스 스키마

지표들의 통계 처리는 SAS 9.3을 이용하였으며 오라클 11g 버전을 사용하여 데이터베이스를 구축하였다.

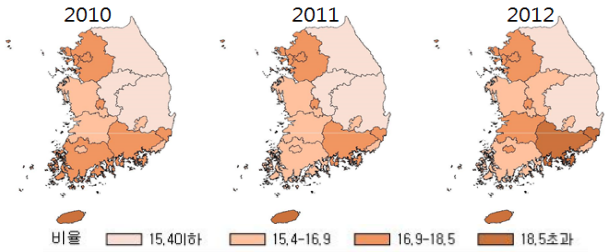
2-3. 데이터를 활용한 분석

<표 3>은 환경성질환의 질병코드 및 명칭을 나타내었다.

<표 3> 환경성질환의 질병코드 및 명칭

항목	질병코드 및 명칭	
환경성질환	J30	혈관운동성 및 알레르기성 비염
	J45	천식
	L20	아토피 피부염

(그림 2)는 연도별 환경성질환의 분포를 지역별로 표시하였으며, 충청북도와 경상북도의 환경성질환의 환자비율이 증가하는 것을 가시적으로 확인 할 수 있다.



(그림 2) 연도별 환경성질환의 분포

<표 4>는 환경성질환의 회귀분석을 위한 독립변수와 종속변수를 표현하였다.

<표 4> 독립변수와 종속변수

변수	정보	요인
독립변수	기후	온도, 강수량, 습도, 풍속
	건강행태	흡연, 비만율
	인구	인구형태
종속변수	질병	환경성질환

구축한 데이터베이스를 활용하여 환경성질환이 기후, 건강행태 그리고 인구형태에 따라서 어떤 영향을 미치는지 분석하였다.

<표 5>는 구축한 데이터베이스를 활용하여 회귀분석 결과를 보여주고 있다.

<표 5> 환경질환에 따른 회귀분석

Variable	Parameter Estimate	Standard Error	t-value	Pr > |t|
Intercept	-0.04131	0.05301	-0.78	0.4403
온도	0.01264	0.00298	4.24	0.0001
강수량	4.62E-06	7.59E-06	0.61	0.5458
풍속	-0.0082	0.00539	-1.52	0.1358
습도	0.00085	0.0004	2.12	0.0399
흡연	-0.00502	0.1419	-0.04	0.9719
비만율	0.08339	0.1113	0.75	0.4581
인구형태	-0.02184	0.00899	-2.43	0.0197

변수 선택과 모델 설정은 임의로 설정하였으며, 환경성질환의 경우 유의수준 0.05에서 온도, 습도, 인구형태가 유의한 영향을 미치는 것으로 나타났다.

3. 결론 및 향후 연구

본 연구에서는 지리정보시스템(GIS)기반 데이터베이스 구축을 통해 지역별 건강상태, 건강행태, 기후변화를 가시화 하였다. 또한 독립변수에 따른 질병의 변화를 분석하기 위한 데이터 수집 단계가 생략되어 향후 의료 데이터를 활용하여 보건의료서비스 관련 산업에 많은 도움을 줄 수 있다.

본 연구의 향후 연구로는 질병 발생에 대한 정보가 연별로 되어 있어 기후변화에 민감한 질병을 찾기 어렵기 때문에 공공기관의 원시자료를 요청하여 월별로 데이터베이스를 구축 할 것이다. 또한 정부 3.0을 통해 개방된 공공 자료 중 질병에 관한 자료를 모두 수집하여 빅 데이터 분석에 활용 가능한 데이터베이스를 구축할 것이다.

사사표기

본 논문은 2013년도 미래창조과학부의 재원으로 과학벨트기능지구지원사업(2013K001552), 한국연구재단의 기본연구지원사업(2013R1A1A206518)과 PTERC(2008-0062611)지원을 받아 수행된 연구임

참고문헌

[1] 조영임, "빅 데이터의 이해와 주요 이슈들", 한국지역정보화학회지, 제16권 제3호, pp.43-65, 2013.

[2] 이만재, "빅 데이터와 공공 데이터 활용", 한국인터넷진흥원, 제2권 제2호, pp.47-64, 2011

[3] 이견직, "의료서비스산업", 보건복지포럼, 통권 제35호, pp.9-24, 1999.

[4] 정기택, "의료서비스산업의 2020 비전과 전략", 산업연구원, 2007.

[5] 박윤형, “기후변화에 의한 전염병 발생영향 통합관리체계 구축”, 한국건강증진재단 건강증진 연구사업, 2006.

[6] 홍윤철, "Climate Change and Human Health", 대한의사협회지, 제51권 제8호, pp.764-769, 2008.

[7] 김동진, "기후변화가 설사병 발생에 미치는 영향 분석", 보건사회연구, 제32권 제1호, pp.281-297, 2012.

[8] IPCC, "Climate Change 2007 - Synthesis

Report", IPCC 사무국, 2008
[9] Marilyn A. Wnkleby,"Socioeconomic Status and Health:How Education, Income, and Occupation Contribute to Risk Factors for Cardiovascular Disease", American Journal of Public Health, Vol.82, No.6, pp.816-820, Jun. 1992.
[10] 전종덕, "도시와 농촌지역 노인의 건강행태 및 건강수준 비교", 한국농촌의학지역보건학회, 제38권 제3호, pp.182-194, 2013.
[11] 국민건강영양조사, http://knhanes.cdc.go.kr
[12] 지역사회건강조사, http://chs.cdc.go.kr
[13] 통계지리정보서비스, http://sgis.kostat.go.kr
[14] 기상청, http://sts.kma.go.kr
[15] e-나라지표, http://www.index.go.kr