빅데이터를 활용한 미세먼지와 질병 간의 상관관계 분석

남경윤, 문소영, 김현희
동덕여자대학교 정보통계학과

20180799@dongduk.ac.kr, 20190959@dongduk.ac.kr, heekim@dongduk.ac.kr

Analysis of the Correlation between Fine Dust and Disease Using Big Data

Kyeongyoon Nam, Soyoung Moon, Hyon Hee Kim
Department of Statistics and Information Science, Dongduk Women's University

요 약

WHO 산하의 국제암연구소는 2013 년부터 미세먼지를 1 급 발암 물질로 분류하고 있으며 미세먼지 노출에 대한 질병 발생의 심각성은 점점 수면 위로 드러나고 있는 추세다. 본 연구에서는 국민건강보험공단의 진료 내역 정보 데이터와 2015 년부터 2021 년까지의 미세먼지 및 초미세먼지 월 평균 농도 데이터를 이용하여 미세먼지 및 초미세먼지 농도와 순환기계와 호흡기계 질병 간의 상관 관계를 보이고, 연관성있는 질병을 찾아내었다. 이를 위해 시계열분석, 상관분석, 빈도분석을 시행하였으며 실험 결과 호흡기질환에서는 급성 부비동염, 코의 농양 등의 질병과 순환기질환에서는 상세불명의 원발성 고혈압, 폐색전증이 상관관계가 높은 질병으로 판명되었다.

Ⅰ. 서론

미세먼지는 여러 만성질환의 위험을 높이고, 순환기계 및 호흡기계 질환을 일으키는데 유의한 관련성이 있는 것으로 알려져 있다. 질병관리본부에 따르면 미세먼지(PM10) 농도가 공기 1 m3 당 10 μg 증가할 때마다 '만성 폐쇄성 폐 질환(COPD)'으로 인한 입원율은 2.7%, 사망률은 1.1% 증가했다[1]. 현재 국내에서는 미세먼지의 심각성에 비해 미세먼지 관련 질환에 대한 연구가 부족하다. 대부분의 연구들이 소수 병원의 응급실 방문이나 입원을 후향적으로 분석하거나 관련 질환을 천식, 만성 폐쇄성 폐질환이나 알레르기 질환 등에 국한된 연구들이 많고, 미세먼지 농도의 증가가 의료비용에 미치는 영향에 관한 연구는 거의 없어 호흡기계, 심혈관계에의 미세먼지 영향을 체계적으로 평가하는 것에 한계가 있다[2,3]. 따라서 미세먼지와 관련된 질병을 찾아내고, 그에 대한 관심과 해결에 대한 노력이 필요하다.

이에 본 연구는 호흡기계, 순환기계 외래·입원 환자 데이터와 미세먼지 데이터를 이용하여 미세먼지와 호흡기계·순환기계 질병의 상관관계를 분석하고자 시행되었다. 국민건강보험공단에서 제공하는 100 만 코호트를 분석함으로써 소수 병원의 데이터로부터 오는 데이터 부족의 제약점을 극복하고 인구 수준 분석을 실시하였다.

본 연구에서는 먼저 미세먼지 농도의 시계열 분석을 통해 미세먼지의 주기성을 밝혀내었고, 이를 통해 미세먼지 농도를 예측하였다. 또한 지역별로 미세먼지와 초미세먼지 농도가 WHO 기준 '나쁨'에 해당하는 달의 빈도 분석하여 비교하였다. 다음으로 순환기계 및 호흡기계 질병들과 미세먼지 농도의 상관분석을 통해 연관성이 있는 질병들을 선별하였다.

마지막으로 선별된 질병들의 월별 빈도 분석을 통해 미세먼지의 농도가 높아지는 시기와 관련 질병들의 빈도가 높아지는 시기가 유사하다는 점을 확인하여 미세먼지와의 연관성을 밝혔다. 본 연구를 통해 미세먼지와 관련된 질병의 대상을 넓혀 다양한 질병들을 대상으로 연구하는 것에 기여하고, 관련 질병에 대한 보건 정책을 통해 보다 효과적인 국민 건강 관리에 활용할 수 있을 것으로 기대된다.

논문의 구성은 다음과 같다. 2 장에서는 데이터 수집 및 탐색에 대해 다루고, 3 장에서는 연구 결과에 대해 설명한다. 마지막 4 장에서는 본 연구의 결론과 시사점을 제시한다.

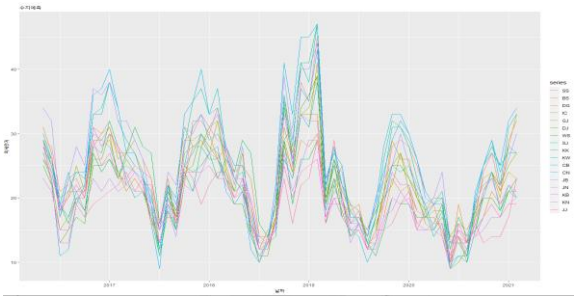
Ⅱ. 데이터 수집 및 탐색

본 연구는 2016 년부터 2021 년까지의 대기오염 자료, 진료내역정보 자료를 이용하여 미세먼지와 관련 질병에 대해 분석한 연구이다.

대기오염자료는 국가통계포털 KOSIS 에서 2015 년 1 월부터 2020 년 12 월까지 월별 평균 미세먼지(PM10)농도와 초미세먼지(PM2.5) 농도 자료를 사용하였다. 진료내역정보자료는 국민건강보험가입자 중 해당 년도에 요양(병/의원)기관으로부터의 진료내역이 1 건 이상 있는 가입자 100 만 명을 무작위로 선별하고, 항목 선정 과정을 거쳐 선정된 해당 가입자의 기본정보와 진료정보를 추출한 뒤 구간분포비율을 최대한 유지한 상태에서 데이터 정제를 거쳐 최종적인 데이터셋을 구성한 국가중점 개방데이터를 사용하였다.

미세먼지 수치 데이터는 열에서는 시도를 세분화한 두 번째 열에서 도평균만 남겼고, 결측치나 특수문자가 들어간 데이터는 제거하였다. 정제된 미세먼지 데이터는 2016 년 5 월부터 2021 년 4 월까지 데이터를 사용하였다. 진료내역정보 자료는 기존 연구에서 미세먼지와 관련이 있다고 밝혀진 순환기계·호흡기계 질환인 I 코드와 J 코드만 주진단명 또는 부진단명으로 뽑아 내었다.

WHO 기준으로 미세먼지(PM10)는 농도 50μg, 초미세먼지(PM2.5)는 농도 25μg 를 초과하면 '나쁨'에 해당한다. 본 연구에서도 이를 기준으로 월별 빈도분석과 관련 질병과의 상관분석을 시행하였다.



<그림 1> PM2.5 농도 변화 그래프

각 연도별 데이터의 요양개시일자에서 월만 추출하여 month 열을 지정하였고, 호흡기 및 순환기 질환 질병코드와 month 의 상관분석을 진행하였다. 두 변수 모두 범주형 변수이기 때문에 크래머 V 계수를 사용하였다. <그림 1>은 PM 2.5 농도 변화 그래프를 나타낸다. 그림에서 볼 수 있는 바와 같이 계층적 시계열 분석을 통해 미세먼지 농도의 증감은 주기성을 갖고 있는 것을 알수 있다.

Ⅲ. 연구결과

분석을 위해 R 언어를 사용하였으며 계층적이거나 그룹화된 시계열 예측을 위한 R 패키지인 hts 를 사용하였다.

순환기 질병코드와 month 사이의 크래머 상관계수는 약 0.706 이었고 같은 방법으로 상관분석을 진행한 결과, 호흡기 질병코드와 month 사이의 크래머 상관계수는 0.705 였다. 이를 통해 month 와 순환기 및 호흡기 질환에 강한 상관관계가 있음을 밝혀냈다.

월별 순환기 및 호흡기 질환의 발병 빈도를 나타낸 데이터 프레임을 만든 후, 지역별 초미세먼지의 월별 평균 농도를 나타낸 PM2.5 데이터 프레임, 지역별 미세먼지의 월별 평균 농도를 나타낸 PM10 데이터 프레임과 각각 상관분석을 진행하였다. 이때 사용한 상관계수는 피어슨 상관계수를 사용하였다. 피어슨 계수가 0.5 이상인 질병만 추출한 결과

호흡기 질환에서는 급성 부비동염(J0180), 코의 농양이나 종(J340), 편향된 비중격(J342), 점액화농성 만성 기관지염(J411)이, 순환기 질환에서는 상세불명의 원발성 고혈압(I109), 폐색전증(I269), 울혈성 심부전(I5004), 뇌경색증 후유증(I693)이 추출되었다.

	J0180	J340	J342	J411	I109	I269	I5004	I693
서울	0.54	0.57	0.73	0.45	0.56	0.53	0.55	0.69
부산	0.51	0.51	0.64	0.36	0.44	0.62	0.55	0.67
대구	0.56	0.61	0.73	0.48	0.55	0.56	0.48	0.65
인천	0.53	0.54	0.71	0.4	0.53	0.47	0.58	0.71
광주	0.45	0.41	0.5	0.23	0.29	0.51	0.45	0.5
대전	0.5	0.55	0.66	0.4	0.48	0.51	0.45	0.6
울산	0.46	0.43	0.6	0.28	0.38	0.61	0.6	0.66
세종	0.52	0.57	0.68	0.44	0.51	0.55	0.44	0.6
평균	0.51	0.52	0.66	0.38	0.47	0.54	0.51	0.64

<표 2> 상위 8 개 미세먼지와
호흡기질환,순환기질환 상관분석표

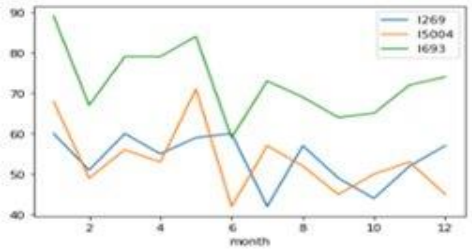
	J0180	J340	J342	J411	I109	I269	I5004	I693
서울	0.41	0.43	0.63	0.62	0.44	0.54	0.44	0.61
부산	0.27	0.32	0.55	0.53	0.37	0.66	0.45	0.59
대구	0.41	0.49	0.63	0.66	0.46	0.55	0.29	0.51
인천	0.38	0.39	0.6	0.56	0.41	0.5	0.45	0.61
광주	0.27	0.3	0.4	0.49	0.24	0.51	0.24	0.36
대전	0.36	0.4	0.55	0.59	0.35	0.48	0.3	0.46
울산	0.17	0.17	0.44	0.41	0.23	0.65	0.47	0.54
세종	0.44	0.49	0.6	0.64	0.43	0.51	0.29	0.48
평균	0.34	0.37	0.55	0.56	0.37	0.55	0.37	0.52

<표 3> 상위 8 개 초미세먼지와 호흡기질환,
순환기질환 상관분석표

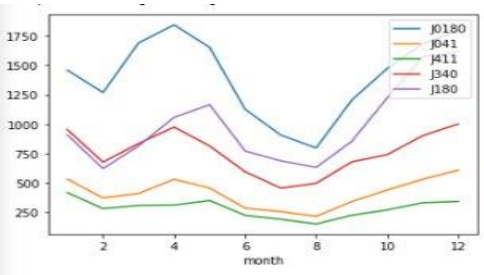
<표 2>와 <표 3>은 서울 및 세종 특별시와 광역시 지역의 미세먼지 농도와 관련 질환의 상관분석을 진행하여 추출한 상위 8 개 질병을 나타낸 표이다. 먼저 <표 2>에 따르면 상관계수 평균이 가장 높은 질병인 편향된 비중격(J342)이 0.66 이었고, 다음으로는 0.64 로 뇌경색증 후유증(I693)이었다. <표 3>에서는 점액화농성 만성 기관지염(J411)이 0.56 으로 가장 높았고, 편향된 비중격(J342)와 폐색전증(I269)이 0.55 로 두 번째로 높은 평균 상관계수를 보였다.

다음으로 2015 년부터 2021 년까지의 미세먼지 농도 '나쁨'에 대한 월별 빈도분석을 진행하였고, 미세먼지 기준으로 3 월은 14 번, 2 월은 13 번, 1 월은 12 번, 5 월은 4 번, 초미세먼지 기준으로 3 월에 17 번, 2 월에 15 번, 1 월에 14 번 5 월과 12 월은 6 번, 11 월에 2 번임을 확인하였다. 또한 6~10 월은 월 평균 미세먼지 수치가 '나쁨'인 적이 없었는데 이를 통해 미세먼지가 1 월에서 3 월이 가장 심한 것을 확인할 수 있었다.

마지막으로 상관분석을 통해 추려낸 질병의 발병 빈도를 월별로 빈도분석을 진행하였다.



<그림 2> 순환기계 질병코드 빈도분석 그래프



<그림 3> 호흡기계 질병코드 빈도분석 그래프

순환기계 질병의 월별 빈도는 <그림 2>에서 볼 수 있듯이 여름에 감소하고 겨울에 증가하는 추세를 보였다. 마찬가지로 호흡기계 질병의 월별 빈도 또한 <그림 3>에서 여름에 감소하고 겨울에 증가하는 추세를 보였다.

Ⅳ. 결론

본 연구에서는 미세먼지(PM10) 농도와 초미세먼지(PM2.5) 농도와 순환기계 및 호흡기계 질병코드를 통해 분석한 결과, 기존 연구들에서 나타나지 않은 질병들을 찾아내는 결과가 있었다. 호흡기질환에서는 편향된 비중격, 점액 화농성 만성기관지염 등이 있었고, 순환기질환에서는 원발성 고혈압, 폐색전증 등이 있었다. 이를 통해 향후 해당 질병들과 미세먼지의 인과관계를 밝히는 연구에 가이드라인을 제시할 수 있을 것으로 보인다.

본 연구에서는 월 평균 미세먼지 농도를 변수로 하여 분석하였기에 미세먼지의 단기간 영향을 정확하게 측정하지 못하였다. 또한 미세먼지 농도 외의 다른 영향요인을 배제하지 못하였음으로 연구의 한계가 있다.

이러한 한계점에도 불구하고, 기존 연구 결과에서 미세먼지와 관련이 있다고 나온 질병들 외에 새로운 질병을 발견하였다는 것에 의미를 가진다. 위의 질병들에 대하여 본 연구에서의 제한적인 부분들을 보완하고, 미세먼지뿐만 아니라 온도, 습도 등의 다른 기상조건과 종합적인 비교를 하는 연구가 필요할 것으로 보인다.

참고 문헌

[1] 함형서, "미세먼지의 계절이 찾아왔다", 금강일보, 2022.04.05, 1 면, http://www.ggilbo.com/news/articleView.html?idxno=905162

[2] Lee, H. S. (2016) "Hospital Visits, Admissions and Hospital Costs among Patients with Respiratory and Cardiovascular Diseases according to Particulate Matter in Seoul," Korean Journal of Environmental Health Sciences. Korean Society of Environmental Health. doi: 10.5668/jehs.2016.42.5.324.

[3] 이승복(2019). 미세먼지가 인체에 미치는 영향에 관한 연구 동향. BRIC View 2019-T26 Available from https://www.ibric.org/myboard/read.php?Board=report&id=3330 (Oct 10, 2019)

[4] KIM, H.-L. and MOON, T.-H. (2021) "Machine learning-based Fine Dust Prediction Model using Meteorological data and Fine Dust data," *Journal of the Korean Association of Geographic Information Studies*. 한국지리정보학회, 24(1), pp. 92-111. doi: 10.11108/KAGIS.2021.24.1.092.

[5] 황수희. "도시지역의 초미세먼지(PM2.5) 농도와 호흡기계 및 순환기계 질환의 영향 연구." 국내석사학위논문 연세대학교 보건대학원, 2015. 서울