

대기오염 장기 노출의 건강 영향 평가를 위한 표본코HORT 자료의 탐색과 처리 방법 제안: 사망을 중심으로

김옥진¹, 김선영², 권혜영³, 김 호^{1,2}

¹서울대학교 보건대학원, ²서울대학교 보건환경연구소, ³목원대학교 의생명보건학부

Data Issues and Suggestions in the National Health Insurance Service-National Sample Cohort for Assessing the Long-term Health Effects of Air Pollution Focusing on Mortality

Okjin Kim¹, Sun-Young Kim², Hye-Young Kwon³, Ho Kim^{1,2}

¹Graduate School of Public Health, Seoul National University, Seoul; ²Institute of Health & Environment, Seoul National University, Seoul; ³Division of Biology & Public Health, Mokwon University, Daejeon, Korea

Objectives: National-scale cohort studies provided evidence of the association between long-term exposure to air pollution and mortality. Recently, South Korean government created the National Health Insurance Service-National Sample Cohort (NHIS-NSC) which included about one million subjects in 2002 and followed up every year through 2013. We aimed to explore major data issues of NHIS-NSC and to provide plausible suggestions in data processing for cohort studies of long-term air pollution and mortality. **Methods:** We specified four data issues: tracking failure, address inconsistency, limited address information, and absence of risk factors. Our exploratory analyses clarified data characteristics related to four issues and led us to provide the best options out of several suggestions in data processing to avoid inaccurate risk estimates in subsequent health analyses. **Results:** Twelve percent of the cohort was lost to follow up for at least one year. We proposed to restrict the study population to cohort members with more than 50% of follow-ups. Although the NHIS-NSC report mentioned some address changes from workplaces to residences in 2005, our exploratory analysis provided little evidence of the change. We concluded that it is not suggestive to consider the change in data processing. The incomplete address information limited to the district level prevents us from estimating individual exposures to air pollution. However, we could compute population-representative district-level averages by using exposure prediction models. Important individual-level risk factors such as smoking were not available in NHIS-NSC. Our suggestion was to use area-level information from external data or to restrict the study population to a subset who carried out national health examinations. **Conclusions:** Our extensive investigation of the NHIS-NSC data and constructive suggestions of data handling focusing on cohort studies of air pollution and mortality will help assess the association between long-term exposures to air pollution and health in South Korea.

Key words: Air pollution, Cohort study, Data issue, Data processing, Long-term exposure, Mortality

Corresponding author: Sun-Young Kim

1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-70-4320-8860, Email: puha0@snu.ac.kr

Received: January 19, 2017 Revised: February 24, 2017 Accepted: February 26, 2017

*This study was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A6A3A04059017) and the National Institute of Environmental Research Project (project number: 2014001310008) funded by the Korea Ministry of Environment. This study used NHIS-NSC data (NHIS-2016-2-230), made by National Health Insurance Service (NHIS). The authors alone are responsible for the content and writing of the paper.

No potential conflict of interest relevant to this article was reported.

How to cite this article:

Kim O, Kim SY, Kwon HY, Kim H. Data issues and suggestions in the national health insurance service-national sample cohort for assessing the long-term health effects of air pollution focusing on mortality. J Health Info Stat 2017;42(1):89-99. Doi: <https://doi.org/10.21032/jhis.2017.42.1.89>

© It is identical to the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0>) which permit unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2017 Journal of Health Informatics and Statistics

서론

코호트 자료를 이용한 대기오염 장기 노출 연구들은 오랜 기간 높은 대기오염 농도에 노출되는 것이 사망 및 질병 이환의 위험을 높이는 것으로 보고하였다[1-6]. 국내에서는 대기오염 장기 노출의 건강 영향을 평가한 연구가 많이 보고되지 않았는데, 이는 이런 연구에 이용할 수 있는 장기 추적 코호트 자료가 없었기 때문이다.

최근 건강보험공단은 2002년 국민건강보험 및 의료급여 대상자 중 일부를 대한민국 인구 집단에 대한 대표성을 갖는 표본으로 추출하여 '표본코호트(National Health Insurance Service-National Sample Cohort, NHIS-NSC)'를 구성하였다[7]. 코호트 구성 초기 해인 2002년 기준으로 총 1,205,340명이 표본코호트 대상자가 되었으며, 이들을 2013년까지 총 12년간 추적하였다. 추적된 정보는 개인별 주소, 연령, 성별, 소득, 사망 관련 정보, 질병 이환 및 의료 이용, 국가건강검진 결과 등이다. 모든 정보는 거의 매년 재수집되어 시간에 따른 변화를 파악할 수 있다.

표본코호트 자료를 이용하면 대기오염 장기 노출의 건강 영향을 대한민국 전체 인구에 대한 대표성을 가지고 평가할 수 있다. 또한 매해 수집된 주소와 개인별 정보가 주소지 이동을 고려한 대기오염 개인 노출 및 관심 요인들의 시간 변이 효과 보정 등을 가능하게 하여 역학 연구 분야에서 연구 가치가 높다. 하지만 표본코호트 자료 구성에 이용된 원시 자료가 연구를 위해 설계 및 수집된 것이 아니므로 자료 처리 과정 없이 통계 분석에 이용할 경우 부정확한 연구결과가 도출될 수 있다.

대기오염의 건강 영향을 평가한 코호트 연구는 확실성과 위중도가 큰 사망을 우선적으로 평가되어야 할 건강 결과로 고려하였다[8]. 국내에서도 대기오염 장기 노출의 건강 영향을 평가하는 초기 단계로 다른 건강 결과보다는 사망에 초점을 맞춰 대기오염 장기 노출의 영향을 평가하는 연구가 필요하다. 이런 이유로 본 연구는 주소지 이동을 고려한 대기오염 장기 노출과 사망 간의 연관성을 파악하기 위한 연구와 이를 위한 최종 분석 자료를 선정하는 것을 가정하였다. 이를 위해 표본코호트 자료를 탐색하고, 그 결과를 바탕으로 한 자료 처리 방법을 제시하는 것이 본 연구의 목적이다. 본 연구에서 제시되는 자료 처리 방법은 대기오염 역학 연구를 위한 자료 분석 시 개인 노출 평가와 건강 영향 추정에 큰 영향을 미치는 요소들 중 네 가지 특성에 초점을 맞추고자 한다. 본 연구에서 제시할 자료 특성은 코호트 대상자에 대한 추적 여부, 주소지 부여 기준의 일관성, 주소 자료의 완전성, 이용 가능한 위험 요인이다.

대기오염과 사망의 연관성을 평가한 코호트 연구 방법론

개인 노출 평가 방법

북미와 유럽의 대기오염 역학 연구에서는 대기오염과 건강의 인과

적인 연관성을 평가하고자 대규모 코호트 자료를 이용하였는데, 그 결과 높은 농도의 대기오염에 대한 장기 노출은 사망의 위험을 높이는 것으로 나타났다[1-5,9,10]. 최근 이런 연구들은 아시아 지역으로도 확대되고 있다[11-13].

이들 연구에서 대기오염 개인 노출은 대기오염 측정소에서 측정된 대기오염 농도 자료로 평가되었다. 단순하게는 개인별 거주지에서 가장 가까운 측정소의 측정치 혹은 근접한 여러 개의 측정소 측정치들의 가중 평균, 거주지가 속한 행정 구역 혹은 도시 내 관측소 측정치들의 평균을 이용했다[1,2,11-14]. 이러한 단순한 방법을 이용하면 측정소와 개인 위치의 측정값 간의 차이에서 기인한 측정 오차(measurement error)가 포함될 수 있어 이후 건강 영향 분석에서 편향된 결과가 야기될 수 있다. 측정 오차를 줄이기 위해 측정소 가까이 거주하는 개인들만 분석에 포함하는 경우, 측정소가 설치되지 않은 지역 거주민은 연구에 포함시킬 수 없게 된다. 기존의 설치된 측정소 또한 개인의 위치와 이들이 노출되는 대기오염 농도를 반영하기에 수적으로 충분하지 않아 개인별 노출의 변이가 충분히 고려되지 못하는 문제도 있다.

최근 이러한 제한점을 극복하고자 통계 모형 혹은 대기질 모형을 기반으로 한 대기오염 노출 예측 모형을 개발하고, 이를 통해 예측된 대기오염 농도를 개인 노출 평가에 이용하고 있다[15-19]. 대기오염 노출 예측 모형을 대기오염 측정소가 설치되지 않은 지점의 대기오염 농도를 예측할 수 있어 다양한 대상 및 지역에 대한 연구를 수행할 수 있다. 이용 가능한 개인 주소가 있으면 이를 좌표화하는 지오코딩을 통해 각 지점에서 예측된 대기오염 농도로 개인 노출을 평가하였다[3,9,20,21]. 개인의 제한된 주소, 즉 우편 번호 지역 혹은 행정 구역 정보만 이용할 수 있는 경우, 지역 중심점과 같이 인구 분포를 대표할 수 있는 임의의 지점에서 예측된 대기오염 농도를 이용하여 개인별 노출을 평가하였다[10,22].

개인의 대기오염 노출 평가는 거주지 이동을 포함하는 방향으로 변화되었다. 초기 연구에서는 주소 이동이 없다는 가정 하에 코호트 등록 당시의 주소에서 측정 혹은 예측된 대기오염 농도를 개인의 장기 노출 대신 이용하였다. 후속 연구들에서 주소 이동을 고려하지 않은 노출 평가에는 측정 오류가 포함된 점이 지적되었다[1,23]. 최근 연구에서는 이를 바탕으로 거주지 이동에 따른 주소 자료를 확보하고, 이를 고려한 개인 노출이 분석에 이용되었다[3,10,21,22,24].

통계 분석 방법

기존 연구에서 대기오염 장기 노출과 사망 간의 연관성을 평가하기 위해서 보정된 변수는 개인 수준의 성별, 연령, 인종, 소득, 결혼상태, 흡연, 음주, 체질량 지수(body mass index, BMI), 기저 질환, 미세먼지 혹은 위해 요인에 대한 직업적 노출 여부, 식이 행태 등이었고, 지역 수

준의 저소득자 비율, 고등학교 이수자 비율, 이민자 비율 등이었다.

일반적으로 대기오염 농도 증가에 따른 사망의 위험비(hazard ratio)는 콕스의 비례 위험 모형(Cox proportional hazard model)을 이용한 생존 분석으로 추정되었다. 개인별 주소가 정기적으로 재 수집되어 시간에 따라 다른 농도를 고려할 수 있는 경우 시간 가변 콕스 비례 위험 모형(time-varying Cox proportional hazard model)을 이용하기도 했다 [3,24].

연구 방법

표본코호트 자료

본 연구에서 활용하고자 하는 표본코호트 자료는 대한민국 전체 인구 집단에 대한 대표성을 갖도록 설계 및 추출된 자료다. 대상자는 약 백만 명으로, 2002년 전국의 국민건강보험 및 의료급여 대상자 중 성, 연령, 소득 분위(건강보험료), 자격에 따라 구분된 총 1,476개 층으로부터 임의 추출되었고, 대상자들에 대한 자료는 2002년부터 2013년까지 총 12년간 추적 수집되었다[7]. 이때 자격은 가입한 보험 종류와 지위에 따라 지역세대주, 지역세대원, 직장가입자, 직장피부양자, 의료급여 세대주, 의료급여세대원으로 구분된다. 사망 혹은 자격 박탈 등의 이유로 감소된 대상자 수는 다음해 출생한 신생아 중 추가적으로 표본을 선택하여 보충하도록 설계되었다. 모든 사망 관련 정보는 통계청의 국가 사망 통계 자료를 통해 확인되었으며, 사망 원인은 질병 및 관련 건강 문제의 국제 통계 분류 10차 개정판(International Statistical Classification of Disease and Related Health Problems 10th Revision, ICD-10)으로 표기되었다. 표본코호트의 대상자 수와 추적 기간은 미국의 대규모 코호트인 미국 암협회 코호트(the American Cancer Society cohort)와 비슷하다[1].

표본코호트 자료는 대상자들의 다양한 정보를 포함하고 있다. 이 자료들은 국민건강정보 data base (DB)에서 추출되었는데, 이는 국민건강보험 및 의료급여 대상자의 관리, 의료 비용 청구 및 지급을 목적으로 수집된 자료이다. 국민건강정보 DB는 자격, 진료, 건강검진의 총 3가지 DB로 구성되어있다. 자격 DB는 대상자들의 성별, 연령, 소득 분위, 자격, 장애 등급, 사망 관련 정보, 거주지 등을 포함하고 있는데, 이 정보는 대상자 관리를 위해 매년 재 수집된다. 진료 DB는 대상자가 국민건강보험에서 보장하는 의료 이용을 한 경우 수집되는 정보로, 의료 이용이 발생할 때마다 수집된다. 이 DB에는 환자 정보와 진료 내역 및 명세서, 이용한 요양 기관의 정보 등이 포함되어 있다. 건강검진 DB는 국민건강보험공단에서 시행하는 국가검진결과가 기록되어 있다. 국가건강검진 대상자는 지역세대주와 직장가입자는 연령에 관계없이 받을 수 있고, 지역세대원, 직장피부양자, 의료급여 대상자는 만 40세 이상

부터 받을 수 있다. 검진은 일반적으로 격년으로 시행되는데, 비사무직 직장가입자의 경우 매해 건강 검진을 받아야 한다. 또한 2008년부터는 자격에 관계없이 만 40세, 만 66세의 대상자에게 생애전환기 건강 검진이 실시되고 있다.

표본코호트 자료는 보건학 및 의학적 연구 가치가 높다. 코호트 대상자들은 전체 인구 집단에 대한 대표성을 가지고, 인구학적, 사회경제적 정보뿐만 아니라 질병 및 사망 관련 정보를 포함하고 있기 때문이다. 전체 대상자 중 국가건강검진 수신자의 경우에는 건강 행태 및 신체 계측을 포함한 기초 검사 결과 또한 포함되어 있다. 대상자들의 자료가 재 수집되는 것도 큰 장점인데, 시간에 따라 변화되는 요인들의 영향을 고려할 수 있기 때문이다. 하지만 이 자료를 구성하는 데 이용된 원시 자료가 대기오염 역학연구를 위해 수집된 자료가 아니기 때문에, 자료의 특성을 파악하고 이를 반영한 자료 처리 방법이 필요하다.

4가지 특성 및 자료 탐색

본 연구에서는 표본코호트 자료를 활용하여 개인별 주소 이동을 고려한 대기오염 장기 노출이 사망에 미치는 영향을 평가하기 위한 분석을 가정하였다. 즉, 2002년 표본코호트 대상자인 1,024,350명을 2013년까지 추적 관찰한 자료를 이용하여 대기오염 노출과 사망과의 연관성을 평가하는 분석을 가정하였다. 선행연구를 바탕으로 자격 DB에서 성별, 연령, 소득 분위, 장애 등급, 검진 DB로부터 흡연, 음주, 건강 행태, 신체 계측 정보를 보정해야 하는 위험요인으로 이용하는 것을 가정하였다. 이런 분석 방법을 가정하였을 때 대기오염의 개인 노출 평가 및 건강 영향 평가에 영향을 줄 수 있고 고려되어야 할 표본코호트의 4가지 특성은 다음과 같다.

연구 대상자에 대한 추적 여부

연구 대상자에 대한 추적 여부는 매년 수집되는 자격 자료의 수집 여부에 따라 결정된다. 국민건강정보 DB중 자격 DB는 국민건강보험 및 의료급여 대상자 관리 및 제공된 의료 서비스 지불 등을 목적으로 수집된 정보로, 일시적인 해외 거주 등의 이유로 자격이 상실된 사람들의 정보는 수집되지 않는다. 연구 대상자 중 일시적인 자격 상실 등의 이유로 추적되지 않아 자료가 수집되지 않은 사람들의 분포를 확인하기 위해 추적 실패 빈도와 기간 등에 대한 기술 통계량을 작성하였다.

주소지 부여 기준의 일관성

표본코호트 자료에 포함된 일부 변수는 당시 행정 기준 등에 따라 수집되기 때문에 시기에 따라 수집된 변수와 그 내용이 달라진 경우가 있다. 대상자들의 주소가 대표적인데, Lee et al. [25]은 초기 직장보험 대상자, 즉 가입자와 피부양자 모두의 주소가 가입자의 직장 주소

로 수집되었으며, 2006년 이후 가입자의 거주지 주소로 변경되어 수집되었다고 했다. 반면 지역보험 및 의료급여 대상자는 주소지 부여 기준의 변화가 없었다고 했다. 일반적으로 개인별 대기오염 장기 노출은 거주지 혹은 직장 소재지 등 한 곳을 중심으로 평가되는데, 거주지와 직장의 주소지 혼용은 건강 영향 분석 결과에 의도하지 않은 영향을 끼칠 수 있다. Lee et al. [25]의 보고서에서 언급된 주소지 기준 변화의 정도 및 시기를 확인하기 위해 두 가지 분석을 실시하였다. 먼저 국민건강보험 대상자를 중심으로 자격별 주소 변화율을 계산하였다. 주소 변화율은 연속된 두 해 동안 자격 변동이 없는 사람들 중 주소가 바뀐 사람들의 비율로 계산하였다. 이때 연도별로 거주지 이동 비율이 크게 변화하지 않는다고 가정하면, 주소지 기준이 변경되었다고 언급된 2006년에는 다른 자격에 비해 직장보험 대상자들의 주소 변화율이 높게 나타날 것이다. 이때 의료급여 대상자는 행정구역별로 관리되어 왔기 때문에 주소 부여 기준과 무관하여 분석에서 제외하였다. 추가적으로 서울에 거주하는 50세 이상 직장가입자의 구별 분포를 계산하였다. 50세 이상의 서울 거주자는 거주지 이동 빈도가 높지 않은 반면 직장 소재지와 거주지가 동일 시군구에 위치할 가능성이 높지 않다고 보았다. 따라서 2006년 전후로 표본코호트의 주소가 직장 소재지에서 거주지 주소로 변화되었다면 이들의 구별 분포 경향이 다를 것이라 가정하였다. 이런 가설 하에 50세 이상의 서울 거주자 중 직장가입자를 대상으로 해서 2002, 2007, 2013년의 구별 주소 분포를 살펴보았다.

주소 자료의 완전성

표본코호트는 개인 정보 보호 차원에서 대상자의 주소를 시군구로 제한해서 제공한다. 따라서, 주소를 기반으로 평가되는 대기오염 노출은 동일한 시군구에 거주하는 모든 대상자들이 동일한 값을 가진다. 시군구는 지리적으로 매우 큰 단위로 개인별 노출을 평가하는 데 적합하지 않을 수 있다. 주소 자료의 불완전성으로 인해 개인별 노출을 시군구 대표값으로 이용할 때 얼마나 많은 대상자들이 동일한 노출로 평가되는지 확인하기 위해 각 시군구 연구 대상자의 수를 시도에 따라 구분하여 계산하였다.

대기오염 노출 평가를 위해서는 대기오염 측정소의 자료를 이용하는 것이 일반적이다. 그런데 2010년 기준으로 전국의 251개 시군구 중 61%에만 총 243개의 측정소가 설치되어 있다[26]. 관측된 대기오염 농도만 이용하는 것으로는 전국에 거주하는 인구 집단의 대기오염 노출을 대표하기 어렵다[26,27]. 또한 측정소가 설치된 지역과 그렇지 않은 지역에 거주하는 인구가 다른 성향을 가질 수 있다. 따라서, 연구 기간 동안 대기오염 측정망 측정소가 설치여부에 따라 연구에 포함되거나 제외된 대상자들의 특성을 비교하였다.

이용 가능한 위험요인

대기오염 장기 노출이 사망에 미치는 영향을 평가하기 위해서 대기오염 외에 사망에 영향을 미치는 위험 요인들을 보정해야 한다. 따라서 선행 연구에서 고려하였던 변수들을 정리하고, 이 중 표본코호트 자료에서 이용할 수 있는 변수들을 확인하였다.

자료 처리 방법 제시

연구 대상자에 대한 추적 여부

자격 상실로 일부 기간의 자료가 수집되지 않은 대상자들에 대한 자료 처리 방법은 다음과 같다. 일반적으로 추적 실패로 인해 노출 혹은 위험 요인 자료를 이용할 수 없는 대상자는 연구에서 제외된다[1]. 표본코호트 자료를 이용한 연구에서도 연구 대상자에 대한 추적 실패로 자료가 수집되지 않은 대상자를 분석에서 제외시킬 수 있다. 즉, 전체 추적 기간 동안 자료 손실이 전혀 발생하지 않은 대상자들만 분석에 포함시키는 방법이다. 이때 제외되는 대상자 수가 많아 분석 대상자 수가 크게 감소되면 통계적 검정력이 낮아지게 된다. 다른 방법으로는 한시적으로 추적에 실패한 대상자에 대해 추적되지 않은 일부 시기의 정보만 제외하고, 나머지 정보는 분석에 이용할 수 있다[10]. 마지막으로, 손실된 정보를 다른 측정값으로 대체하는 방법을 고려할 수 있다. 예를 들면 자료가 손실된 시기와 가장 가까운 시기의 자료로 손실된 정보를 대체하는 방법이다.

주소지 부여 기준의 일관성

개인 주소가 거주지가 아닌 직장 주소로 수집된 자료를 처리하는 방법으로 다음 세 가지를 고려할 수 있다. 먼저 직장 주소가 수집된 대상자, 즉 2002년부터 2005년 사이에 직장보험 대상자였던 사람들을 모두 연구에서 제외시킬 수 있다. 다른 방법으로는 수집된 직장 주소를 다른 해에 수집된 동일 대상자의 거주지 주소로 대체하여 이용할 수 있다. 이 방법은 대상자들의 거주지 이동 빈도가 빈번하지 않다는 가정 하에 고려할 수 있는 방법으로, 예를 들어 직장 주소로 주소 정보가 수집되었던 4년 내내 직장보험 대상자라면 이 기간의 주소는 2006년의 거주지 주소로 대체될 수 있다. 따라서 이 기간 내에 실제 거주지가 변경되었다면 전혀 다른 지역이 거주지로 선택될 수도 있다. 마지막으로 표본코호트 자료에서 주소 부여 기준 변화가 건강 영향 분석에 미치는 영향이 미미할 것이라는 가정 하에 자료 수정 없이 이용하는 방법도 있다.

주소 자료의 완전성

시군구로 제한된 주소 사용과 대기오염 노출 평가에 이용되는 대기오염 측정소의 설치 지역과 수가 제한적이라는 점을 고려할 때 제한할

수 있는 개인별 대기오염 노출 평가 방법은 다음과 같다. 먼저 개인 노출을 시군구 단위로 평가하되, 측정소가 설치된 지역 거주민만 연구 대상으로 선정하는 방법이다. 이때 전국 39%의 시군구 거주자가 연구에서 제외되어 전체 인구 집단에 대한 대표성은 가지지 못할 수 있다 [26]. 이런 점을 극복하기 위해서 최근 대기오염 장기 노출 연구에서 이용되고 있는 노출 예측 모형을 이용하는 방법을 고려할 수 있다. 시군구별 대기오염 농도에 대한 대표성을 갖도록 개발된 대기오염 예측 모형을 이용하여, 이로부터 얻은 시군구별 연평균 대기오염 농도를 개인 노출에 이용할 수 있다[26,27].

이용 가능한 위험요인

대기오염 장기 노출이 사망에 미치는 영향을 정확하게 파악하기 위해서는 개인 혹은 지역 수준의 변수 효과가 보정되어야 한다. 표본코호트 자료를 구성하는 3가지 DB는 각기 다른 변수에 대한 정보를 포함하고 있다. 자격 DB는 전체 대상자의 성별, 연령, 소득 수준 등의 정보를 포함하고 있다. 진료 DB는 전체 대상자의 의료 이용 중 국민건강보험에서 비용을 부담하는 의료 이용에 대한 정보를 포함하고 있다. 진료 DB에 포함된 의료 이용 정보가 청구 자료라는 한계점에도 불구하고, 질병 코드 및 약제 처방 자료를 이용하여 혼란 요인으로 작용될

수 있는 기저 질환 등의 효과를 보정할 수 있다는 강점이 있다. 건강검진 DB는 건강보험공단에서 제공하는 검진을 수진자의 건강 행태, 기저 질환, 생활 습관 등에 대한 설문 결과와 신체 계측 및 기초 검사 결과를 포함한다. 각 DB가 포함하고 있는 정보들을 이용하여 대기오염 장기 노출과 사망 간의 연관성을 평가할 때 위험 요인을 보정하는 방법은 다음과 같다. 먼저 표본코호트 대상자를 모두 연구 대상으로 선정하되, 이용할 수 있는 최대한의 개인 수준의 변수를 보정할 수 있다. 이때 선행 연구들에서 보정된 변수들 중 이용 가능한 개인 수준의 변수가 없다면 외부 자료, 예를 들면 총인구조사 자료 혹은 지역사회건강조사 등의 자료를 이용하여 지역 수준의 변수로 대신 이용할 수 있다. 국가건강검진 대상자의 경우 전체 코호트 대상자보다 많은 개인 변수를 포함하고 있다. 따라서 보다 많은 개인 수준의 변수 사용을 위해서는 연구 대상자를 건강검진 대상으로 한정하여 분석하는 방법도 있다.

연구 결과

표본코호트

전체 연구 대상자 총 추적 기간 동안 총 55,798건(5.4%)의 사망이 발생하였고, 사망 외의 이유로 중도 탈락된 사람들은 총 57,387명(5.6%)이었

Table 1. Summaries of characteristics of study population of NHIS-NSC by 2002 and 2013

Characteristics	Year (%)			
	2002 (N = 1,025,340, 100.0%)	2013 (N = 917,254, 89.5%)		
Sex	Male	50.1	49.8	
	Female	49.9	50.2	
Age (y)	0-19	27.2	12.6	
	20-64	65.1	73.4	
	≥64	7.8	14.0	
Area	Seoul	21.3	20.0	
	7 metropolitan cities ¹	26.4	26.2	
	Gyeonggi province	20.6	23.5	
	8 provinces ²	31.7	30.3	
Eligibility of NHI	Self-employed insured	Householder	17.4	14.5
		House member	31.2	15.7
	Employee insured	Insured	16.7	30.6
		Dependent	31.7	36.1
	Medical aid beneficiaries	Householder	1.6	2.2
		House member	1.4	1.0
Income ³ (%)	Medical aid beneficiaries	3.0	3.1	
	0-20	12.1	14.2	
	21-50	24.7	24.0	
	51-90	47.0	44.2	
	91-100	13.2	14.6	

NHI, the national health insurance.

¹All metroplitan cities except Seoul: Pusan, Deagu, Incheon, Kwangju, Deajeon, Ulsan, and Sejong.

²All provinces except Gyeonggi: Gangwon, Chungcheongbuk, Chungcheongnam, Jeollabuk, Jeollanam, Gyeongsangbuk, Gyeongsangnam, and Jeju.

³NHIS-NSC provided income as percentiles ("The manual for User of the National Health Insurance Service of National Sample Cohort Database") without original continuous income data.

다. 코호트 등록 당시 서울과 경기도에 거주하는 성인의 비율이 높았고, 지역과 직장보험의 피부양자이면서 가구 소득이 50-90% 사이인 대상자의 비율이 높았다. 연구 종료 시기인 2013년에는 지역별 거주 분포와 가구 소득분포는 2002년과 비슷했던 반면, 직장가입자인 성인, 특히 노인 여성의 비율이 증가되었고 지역세대원 비율은 감소되었다(Table 1).

자료 탐색

연구 대상자에 대한 추적 여부

전체 대상자 1,025,340명 중 119,477명(11.7%)의 자료가 추적 실패로 인해 누락되었다. 이 중 96,421명(80.7%)은 1년치의 자료만 누락된 것으로 확인되었다(Table 2). 개인별 추적 실패 기간의 평균은 1.3년, 표준편차는 0.8년이었고, 최대 7년의 자료가 누락된 경우가 있었다.

주소지 부여 기준의 일관성

자격별 주소 변화율을 살펴본 결과는 다음과 같다. 전체 직장보험

Table 2. Numbers of study subjects of NHIS-NSC across years of tracking failure

Total tracking failure periods (y)	n	%
1	96,421	80.7
2	15,323	12.8
3	4,078	3.4
4	2,097	1.8
5	918	0.8
6	460	0.4
7	180	0.2
Total	119,477	100.0

NHIS-NSC, national health insurance service-national sample cohort.

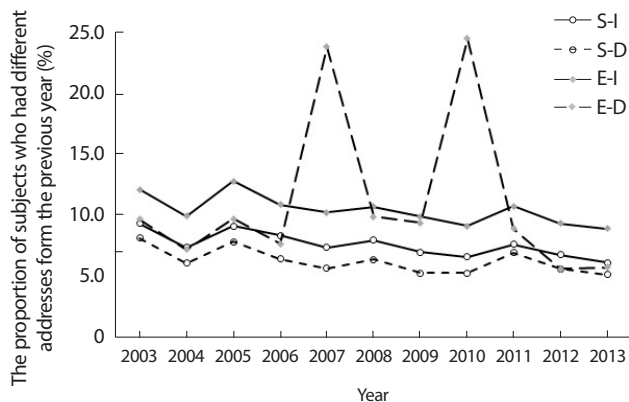


Figure 1. Temporal trend of proportions of NHIS-NSC subjects who had different addresses from the previous year by insurance type and eligibility. Insurance type: S, self-employed; E, employee insured; eligibility. NHIS-NSC, national health insurance service-national sample cohort. Insurance (or householder); D, dependent (or house-member)] from 2003 through 2013.

대상자 중 직장가입자의 주소는 연도에 따라 큰 차이를 보이지 않았고, 직장피부양자의 주소가 2006-2007년 사이와 2009-2010년 사이에 크게 변동된 것으로 나타났다. 지역보험대상자에서는 이런 경향이 나타나지 않았다(Figure 1). 서울에 거주하는 50세 이상의 직장가입자들의 구별 인구 분포를 살펴본 결과, 대상자들의 구별 분포가 연도별로 차이를 보이지 않았다. 특히 직장 주소로 표기되었다고 보고된 2007년 이전의 인구 분포가 다른 해의 분포와 다르지 않았으며, 다수의 사업장이 있는 강남구, 영등포구, 서초구, 중구, 마포구 등에서의 비율이 특별히 높은 경향을 보이지 않았다(Figure 2).

주소 자료의 완전성

2002년 기준으로 대상자들은 전국 244개 시군구에 거주하며, 한 시군구 당 평균 4,203명(표준편차, standard deviation [SD]: 3,218)이 거주하는 것으로 나타났다. 시도별로 살펴보면 서울시의 구별 평균 연구 대상자 수는 8,720명(SD: 2,587)으로 가장 많았고, 대구, 대전, 광주, 인천 순으로 시군구별 평균 연구 대상자 수가 많았다. 강원도는 한 시군구 당 평균 1,776명(SD: 1,706)의 연구 대상자가 포함되어 가장 적은 값을 나타냈다. 그 뒤로는 전남, 경북, 제주 순으로 시군구별 연구 대상자 수가 적은 것으로 나타났다.

전체 연구 기간동안 측정소가 설치된 지역에 거주하는 대상자는 시도에 고르게 거주하는 중장년층 대상자인 반면, 측정소가 설치되지 않은 지역에 거주하는 대상자는 시를 제외한 도지역에 거주하는 중장년 및 노인이 다수를 차지하였다. 따라서 측정소가 있는 시군구만 포함하면 이들은 연구에서 제외되게 된다(Table 3).

이용 가능한 위험 요인

대기오염의 건강영향을 평가한 선형 코호트 연구에서는 성별, 연령,

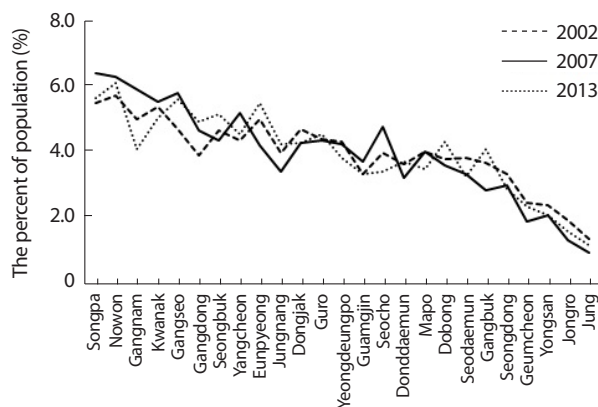


Figure 2. Percents of the employee-insured (>=50 years) across 25 districts in Seoul from NHIS-NSC in 2002, 2007, and 2013. NHIS-NSC, national health insurance service-national sample cohort.

Table 3. Individual characteristics of NHIS-NSC subjects between the areas with and without regulatory air pollution monitoring sites for 2002-2013

Characteristics		Area with monitoring sites (n = 743,106, 72.5%)	Areas without monitoring sites (n = 282,234, 27.5%)	
Sex	Male	50.1	50.0	
	Female	49.9	50.0	
Age (y)	0-19	27.7	25.7	
	20-64	65.9	62.9	
	≥ 64	6.4	11.4	
Area	Seoul	28.8	1.4	
	7 metropolitan cities ¹	27.9	22.7	
	Gyeonggi province	20.0	22.2	
	8 provinces ²	23.4	53.7	
Eligibility of NHI	Self-employed insured	Insured	17.1	18.2
		Dependent	30.5	32.9
	Employee insured	Insured	17.7	14.1
		Dependent	32.2	30.4
	Medical-aid beneficiaries	Householder	1.3	2.6
		Member of household	1.2	1.8
Income ³ (%)	Medical-aid beneficiaries	2.5	4.4	
	0-20	11.5	13.7	
	21-50	23.9	26.8	
	51-90	47.8	45.0	
	91-100	14.4	10.1	

NHI, the national health insurance.

¹All metropolitan cities except Seoul: Pusan, Deagu, Incheon, Kwangju, Deajeon, Ulsan, and Sejong.

²All provinces except Gyeonggi: Gangwon, Chungcheongbuk, Chungcheongnam, Jeollabuk, Jeollanam, Gyeongsangbuk, Gyeongsangnam, and Jeju.

³NHIS-NSC provided income as percentiles ("The manual for User of the National Health Insurance Service of National Sample Cohort Database") without original continuous income data.

소득, 교육수준, 결혼 상태, 흡연, 음주, 체질량 지수, 식이, 과거 질병력, 직업적인 노출 정보 등을 보정하였다. 선행 연구에서 보정된 변수들 중 전체 연구 대상자에 대해 이용할 수 있는 변수는 성별, 연령, 소득이다. 더 많은 개인 수준의 변수를 이용하기 위해서는 의료 이용 정보와 국가건강검진 결과 자료를 이용할 수 있다. 의료 이용 정보를 이용하여 과거 및 기저 질환 등의 효과를 보정할 수 있으며, 건강 검진 결과를 통해 흡연 및 음주와 건강 행태, 체질량 지수, 설문을 통해 얻은 과거 질병력 등을 보정할 수 있다. 교육수준과 결혼 상태, 직업적인 노출 여부에 대한 자료는 표본코호트 자료를 이용해서는 보정하기 어렵다.

자료 처리 방법에 대한 제안

연구 대상자에 대한 추적 여부

자료 손실이 발생한 대상자는 전체 대상자의 11.7%로 이들을 모두 분석에서 제외하면 전체 인구에 대한 대표성을 잃을 수 있다. 또한 표본 수가 감소되어 통계적 검정력이 감소될 수 있다. 이런 한계점을 극복하기 위해 손실된 자료, 특히 노출 계산에 이용되는 대상자의 주소를 대상자들의 거주지 이동이 빈번하지 않다는 가정 하에 가장 가까운 해의 값으로 대체하여 사용할 수 있다. 이때 대체하여 사용된 정보와 손실된 자료 간의 차이에 따른 노출의 오분류 오류가 발생할 수 있

고, 오류의 방향성을 예측하기 어렵다는 한계점이 있다. 따라서 본 연구에서는 전체 연구 대상자를 분석에 포함시키되, 손실된 자료는 결측 처리하여 이용하는 것을 제안한다. 이때 개인별 노출 계산에 이용되는 주소가 전체 추적 기간 중 50% 이상 결측값으로 기록된 대상자는 연구에서 제외하는 것을 추가로 제안한다. 이를 통해 자료로 계산되는 개인별 노출이 전체 추적 기간에 대한 대표성을 갖는 자료만 연구에 포함할 수 있다.

주소지 부여 기준의 일관성

연구 대상자의 자격별로 주소 변화율의 추이를 확인한 결과, Lee et al. [25] 저자들의 보고서 내용에서 제시된 내용을 확인할 수 없었다. Lee et al. [25] 저자들에 따르면 2006년 이전 직장보험 가입자의 주소는 직장 주소지이며, 2006년 거주지 주소로 변경되었기 때문에 2005년 대비 2006년 직장가입자의 주소 변화율이 다른 해에 비해 크게 증가되어야 한다. 하지만 2006년 직장보험 가입자의 주소 변화율은 다른 해와 크게 차이가 없어 주소지 기준이 변경되었다고 보기 어렵다(Figures 1, 2). 반면, 직장피부양자의 주소 변화율은 2006-2007년 사이와 2009-2010년 사이에 증가되는 경향을 보였다(Figure 1). 그러나 직장피부양자의 주소지 부여 기준 변경에 대해서는 보고서나 건강보험공단

Table 4. Individual characteristics of NHIS-NSC subjects from the entire cohort for 2002-2013 and the national health screening examinees in 2002

Characteristics		Study pop. (N = 1,025,340, 100.0%)	NHE examinee in 2002 (N = 113,641, 11.1%)	
Sex	Male	50.1	63.5	
	Female	49.9	36.5	
Age (y)	0-19	27.2	0.3	
	20-64	65.1	93.5	
	≥ 64	7.8	6.2	
Residential district	Seoul	21.3	18.0	
	7 metropolitan cities ¹	26.4	28.2	
	Gyeonggi province	20.6	19.2	
	8 provinces ²	31.7	34.7	
Eligibilities of NHI	Self-employed insured	Insured	17.4	11.8
		Dependent	31.2	7.4
	Employee insured	Insured	16.7	71.1
		Dependent	31.7	9.6
	Medical-aid beneficiaries	Householder	1.6	0.0
		Member of household	1.4	0.0
Income ³ (%)	Medical-aid beneficiaries	3.0	0.1	
	0-20	12.1	13.2	
	21-50	24.7	24.1	
	51-90	47.0	49.8	
	91-100	13.2	12.8	

NHE, the national health examination; NHI, the national health insurance.

¹All metropolitan cities except Seoul: Pusan, Daegu, Incheon, Kwangju, Daejeon, Ulsan, and Sejong.

²All provinces except Gyeonggi: Gangwon, Chungcheongbuk, Chungcheongnam, Jeollabuk, Jeollanam, Gyeongsangbuk, Gyeongsangnam, and Jeju.

³NHIS-NSC provided income as percentiles ("The manual for User of the National Health Insurance Service of National Sample Cohort Database") without original continuous income data.

에의 문의를 통해서도 변경 요인에 대한 정보를 찾을 수 없었다. 또한 이 결과를 표본코호트 자료 제공처인 국민건강보험공단 빅데이터 운영실에 문의하여 본 결과, 이런 현상을 설명할 만한 어떠한 근거도 제공받을 수 없었다. 따라서 직장피부양자의 주소 변화율이 관찰되었음에도 이를 설명할만한 근거 부족으로 표본코호트 자료의 주소를 다른 처리 없이 거주지 주소로 가정하고 이용하는 것을 제안한다.

주소 자료의 완전성

표본코호트 대상자의 주소가 시군구로 제한되어 제공되기 때문에 개인 노출은 시군구 대푯값을 이용하여 평가될 수 밖에 없으며, 평균적으로 약 4천 명의 노출이 동일한 값으로 평가된다. 개인별 노출 평가를 고려할 때, 시군구라는 지리적인 단위는 상대적으로 넓은 단위로 동일한 지역 내 거주민 간의 대기오염의 노출의 차이가 발생할 가능성이 높다. 하지만 시군구로 제한된 주소 제공은 개인정보 보호를 위한 결정이기 때문에, 현재로서는 읍면동 등 세부적인 주소 정보 획득이 어려운 실정이다. 대기오염측정망 측정소의 관측값이 있는 시군구만 연구에 포함하는 경우, 2002년부터 2013년까지 전체 연구기간 동안 측정소가 설치된 시군구에 거주하여 연구에 포함될 수 있는 사람들은 총 743,106명(72.5%)이다. 또한 측정소가 설치된 지역에 거주하는 대상자

들의 특성이 전체 대상자의 특성과는 차이를 보이기 때문에 대기오염 노출 효과 외의 다른 요인에 의한 결과를 얻을 수도 있다(Table 3). 따라서 본 연구에서는 대기오염 노출 예측 모형을 이용하여 각 시군구를 대표할 수 있는 대기오염 농도를 산출하고, 이를 바탕으로 개인별 노출을 평가하는 것을 제안한다. 최근 국내에서도 통계 모형이나 대기질 모형을 기반으로 대기오염 농도를 예측하는 모형을 개발하는 연구들이 보고되고 있다[28-30]. 이 방법을 이용하면 전체 인구 집단에 대한 대표성을 확보할 수 있으며, 전국의 대기오염 분포를 고려할 수 있게 된다. 연구에 적용된 노출 예측 모형으로 산출된 농도 예측값의 타당성 평가를 위해서 측정소에서 측정된 대기오염 농도와 비교가 필요하다.

이용 가능한 위험요인

표본코호트 자료를 이용하여 대기오염의 장기 노출과 사망 간의 연관성 평가를 위해 변수를 보정하는 방법은 다음과 같다. 전체 표본코호트 대상자를 연구 대상으로 선정하는 경우 개인 수준의 성별, 연령, 소득을 보정하고, 그 외의 변수들은 외부 자료로부터 얻은 지역 수준의 대푯값을 이용하여 보정할 수 있다. 예를 들어 총인구조사 및 지역 사회건강조사 자료 등을 이용하면 시군구별 고등학교 졸업자 비율, 음주율, 흡연율, 비만율 등의 효과를 보정할 수 있는데[31,32], 이때 이용

되는 지역 수준의 변수 효과는 개인 수준의 그것과는 다른 경향을 보일 수도 있다. 다른 방법은 국가건강검진 대상자를 연구 대상으로 선정하여 자료에 포함된 개인 수준의 위험 요인을 보정하는 것이다. 그런데 국가건강검진 대상자는 직장가입자와 지역세대주, 혹은 40세 이상의 성인으로 한정되어 있어 전체 인구에 대한 대표성을 갖지 못한다는 제한점이 있다. 두 집단 간의 차이를 비교하기 위해 전체 연구 대상자와 2002년 국가건강검진 수진자의 인구학적, 사회경제적 변수의 분포를 비교하였다(Table 4). 국가건강검진 수진자는 113,641명(11.1%)으로 전체 연구 대상자에 비해 청장년 층의 남성 직장가입자 비율이 높았고, 소득 및 지역에 따른 차이는 크지 않았다. 따라서 다양한 보정 변수를 포함하기 위해서는 국가건강검진 자료를 이용하는 것을 제안하지만, 전체 자료와의 인구학적 특성 차이를 고려할 때 두 자료를 이용한 분석 결과 비교 또한 필요하다. 이 모든 자료를 이용하는 경우에도 국내에서는 독성 물질 혹은 분진에 대한 직업적인 노출 관련하여 이용 가능한 정보가 없어 고려하기 어려운 실정이다.

표본코호트 자료의 강점 중 하나가 대상자들의 의료 이용 정보를 이용할 수 있다는 점이다. 이 정보는 대기오염의 건강 영향 평가를 위한 연구 수행 시 기저 질환에 대한 효과를 혼란 변수로 여겨 보정하거나, 효과 교정 인자(effect modifier)로 이용할 수 있다. 그런데 표본코호트 자료에 포함된 의료 이용 정보는 건강보험에서 의료비용을 지급하는 치료에 국한된다는 제한점이 있다. 또한 의료 이용 자료를 분석을 위해서는 반복 처방, 일괄 청구, 다중 질병 코드 등의 처리 방법 등에 대한 충분한 고려가 필요하기 때문에 자료 가공에 많은 노력이 필요하다. 따라서 의학적인 진단과 건강보험 청구 자료에 대한 전문적인 지식에 근거한 충분한 연구가 뒷받침되지 않는 한, 기저 질환 관련 변수는 정확도가 조금 떨어지더라도 국가건강검진 결과에 포함된 현재 앓고 있는 질환에 관한 설문 결과를 이용할 것을 제안한다.

고찰 및 결론

본 연구는 표본코호트 자료를 활용하여 대기오염 장기 노출이 사망에 미치는 영향을 평가하기 위한 역학 연구를 수행할 때 고려해야 하는 자료 특성을 탐색하고 자료 처리 방법을 제안하기 위해 수행되었다.

본 연구에서 제시한 4가지 자료 특성은 연구 대상자에 대한 추적 여부, 주소지 부여 기준의 일관성, 주소 자료의 완전성, 이용 가능한 위험 요인이었다. 표본코호트 전체 대상자 중 약 11.7%에 대한 자료가 연구 기간 동안 일부 수집되지 않았으며, 전체 연구 기간 동안 대상자들의 주소 부여 기준의 변화가 명확하지 않은 것으로 파악되었다. 개인 정보 보호 차원에서 시군구 단위로 제공되는 거주지 정보를 이용할 경우 평균적으로 약 4천 명의 대기오염 노출이 동일하게 평가되는 것

로 나타났고, 전체 표본코호트 대상자보다는 국가검진대상자들을 연구 대상으로 하는 경우 고려할 수 있는 개인 수준의 위험 요인이 많은 것으로 확인되었다. 이러한 결과를 바탕으로 추적 실패로 인해 손실된 자료는 결측으로 처리하되, 그 기간이 개인별 전체 추적 기간 중 50% 이하인 대상자만 연구에 포함하는 것을 제안하였다. 또한 개인별 대기오염 장기 노출은 표본코호트 자료에서 제공하는 시군구 정보를 이용하여 평가하되, 대기오염 노출 예측 모형에 의해 추정된 대기오염 농도를 이용하도록 제안하였다. 표본코호트 대상자를 분석에 포함시키기 위해 선행 연구에서 고려하였지만 자격DB에 포함되지 않아 이용할 수 없는 변수들은 외부 자료의 지역 수준의 변수로 보정하는 것을 제안하였으며, 보다 정확한 개인 수준의 위험 요인 보정을 위해서 국가건강검진 대상자 자료를 이용한 추가 분석을 수행할 것을 제안하였다.

본 연구 결과는 대기오염 역학 연구를 위해 중요한 자료인 표본코호트의 활용 방법을 제시함으로써, 우리나라에서 대기오염의 건강 악영향 규명에 기여할 수 있다. 북미와 유럽에서는 대규모 코호트 연구를 통해 대기오염 장기 노출과 사망 및 만성 질환의 연관성이 보고되고 있지만[1-5,9,10], 우리나라에서는 미세 먼지 등 대기오염의 건강 영향에 대한 관심이 증가하고 있음에도 불구하고 활용할 수 있는 코호트 자료의 부족으로 연구 결과가 거의 발표되지 않았다. 표본코호트 자료는 국내 유일의 전국민에 대한 대표성을 갖는 대규모 장기 코호트로, 대기오염의 장기 노출을 평가하는 데 필요한 개인별 주소 및 위험 요인들에 대한 정보를 포함하고 있다. 특히 매년 갱신되는 주소 정보를 이용하면 주소지 이동을 고려한 대기오염 장기 노출 평가를 할 수 있어 한 시점의 주소 정보를 이용할 때 발생할 수 있는 측정 오차를 감소시킬 수 있다. 표본코호트 자료가 가지는 이런 강점에도 불구하고 본 연구에서 제시된 4가지 자료 특성이 고려되지 않으면 개인의 노출 평가 및 건강 영향 평가 결과에서 예측하지 못한 오차나 편이가 발생할 수 있다. 본 연구 결과를 통해 표본코호트 자료를 활용한 대기오염 역학 연구 특히 대기오염 장기 노출 연구를 수행할 때 자료 탐색 시간을 단축하고, 보다 명확하게 자료를 파악할 수 있으며, 나아가 추후 연구에서 보다 정확하게 건강 영향을 평가할 수 있을 것으로 기대된다.

연구 대상자의 추적 실패로 발생한 자료 손실에 대한 자료 처리 방법은 제외되는 대상자 수를 최소화하여 통계 분석의 검정력을 잃지 않으면서 개인별 노출 기간을 충분히 고려하고, 실 거주지 자료를 이용함으로써 측정 오차가 적게 발생할 수 있도록 한다. 최근 대기오염 장기 노출 연구에서는 거주지 이동을 고려하여 개인 노출을 평가하고 있는데, 이런 연구는 표본코호트와 같이 대상자의 주소가 코호트 등록 후 재 수집되기 때문에 적용 가능한 방법이다. 코호트 등록 초기의 정보만 이용할 수 있는 연구에서는 대상자가 연구 기간 내내 동일한 거주지에 살았으며 시간에 따른 지역별 대기오염 농도의 순위는 바뀌지 않

는다는 가정하에 수행된다. 따라서 사망 신고 시의 거주지가 연구 초기와 상이하게 다르면 대상자를 연구에서 제외시킬 수밖에 없어 상당 수 자료를 제외해야 하는 제한점이 있었다[1].

표본코호트 자료를 기반으로 환경 역학 연구를 수행하는 경우 개인의 환경 노출은 시군구 단위로 평가된다. 최근 해외의 대기오염 장기 노출 코호트 연구에서 개인 노출은 우편번호 지역 혹은 집계구 등 상대적으로 좁은 지역을 단위로 평가되거나[1,10,22], 실제 주소를 기반으로 평가되었다[3,9,20,21,24,33]. 이에 반해 표본코호트 자료의 분석 단위인 시군구는 상대적으로 넓은 지역으로 동일한 지역 내 개인 간에 있을 수 있는 노출 차이를 반영할 수 없다는 한계점이 있다. 향후 연구에서는 시군구 단위의 개인 노출을 이용한 대기오염 건강 영향 분석 결과와 개인 주소 단위 노출을 이용한 결과를 비교함으로써, 공간적인 오분류로 인해 발생할 수 있는 오류에 대한 탐구가 필요하다.

향후 연구로 본 연구에서 제안된 자료 처리 방법이 실제 분석 결과에 미치는 영향을 정량적으로 평가하는 연구가 필요하다. 본 연구는 표본코호트 자료를 탐색한 결과를 바탕으로 별도의 가공 없이 건강 영향 분석을 수행하는 경우 발생할 수 있는 오류들을 제시하였다. 따라서 제안된 자료 처리 방법의 적용 전 후의 개인 노출과 건강 영향 분석 결과 비교를 통한 오류의 정도가 적량적으로 평가되어야 한다.

또한 표본코호트 자료를 활용하여 사망 외의 질병 이환 등 다른 건강 결과를 이용하여 대기오염의 건강 영향을 평가할 때 고려되어야 할 특성 및 자료 처리를 제시하는 연구가 필요하다. 표본코호트 자료의 가장 큰 장점 중 하나가 본 연구에서 가정한 사망 외에도 대상자들의 질병 이환을 포함한 의료 이용 등의 정보가 이용 가능하다는 점이다. 이 정보를 이용하여 대기오염 장기 노출의 영향을 사망뿐만 아니라 다양한 건강 결과에 대해 평가할 수 있다. 그러나 명확한 건강 결과인 사망과 달리, 질병 발생이나 이환은 건강보험 청구자료에서 정확도의 차이를 보일 수 있다. 또한, 표본코호트에 수록된 질병 이환 및 의료 이용 자료는 국민건강보험 청구 자료로 보험에서 보장하는 치료 내용만 기록되며, 자격 상실 시기의 자료는 수집되지 않는 등의 특이점이 있다. 따라서 이후 연구에서 사망 외의 질환 이환을 관심 건강 결과로 고려하는 경우에 적용할 수 있는 자료 처리 방법을 독립적으로 제시할 필요가 있다. 그 결과가 바탕이 된다면 다양한 건강 영향을 평가할 수 있을 것이다.

본 연구는 표본코호트를 기반으로 거주지 이동을 고려한 대기오염 장기 노출이 사망에 영향을 주는지를 평가하는 경우, 정확한 건강 영향 평가를 위해 적용할 수 있는 자료 처리 방법을 제시하였다. 이 결과는 대기오염과 사망의 연관성에 대한 연구에 구체적인 지침을 제공하는 한편, 유사한 주제의 연구를 계획하는 연구자들에게 자료의 특성 파악, 자료 가공, 분석방법 적용에 대한 방향을 제시할 수 있다. 궁극적

으로는 우리나라에서 대기오염의 장기간 노출이 건강에 미치는 악영향을 정확하게 평가하는 데 기여할 것으로 기대된다.

RERERENCES

1. Pope CA, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, et al. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *J Am Med Assoc* 2002;287(9):1132-1141.
2. Laden F, Schwartz J, Speizer FE, Dockery DW. Reduction in fine particulate air pollution and mortality extended follow-up of the Harvard Six Cities study. *Am J Respir Crit Care Med* 2006;173:667-672.
3. Hart JE, Liao X, Hong B, Puett RC, Yanosky JD, Suh H, et al. The association of long-term exposure to PM2.5 on all-cause mortality in the Nurses' Health Study and the impact of measurement-error correction. *Environ Health* 2015;14:38.
4. Beelen B, Stafoggia M, Raaschou-Nielsen O, Andersen ZJ, Xun WW, Katsouyanni K, et al. Long-term exposure to air pollution and cardiovascular mortality an analysis of 22 European cohorts. *Epidemiol* 2014;25(3):368-378.
5. Beelen R, Raaschou-Nielsen O, Stafoggia M, Andersen ZJ, Weinmayr G, Hoffmann B, et al. Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicenter ESCAPE projects. *The LANCET* 2014;383(9919):785-795.
6. Lipfert FW. A critical review of the ESCAPE project for estimating long-term health effects of air pollution. *Environ Int* 2017;99:87-96.
7. Lee JY, Lee JS, Park SH, Shin SA, Kim KW. Cohort profile: The National Health Insurance Service-National Sample Cohort (NHIS-NSC), South Korea. *Int J Epidemiol* 2015;0(0):1-8.
8. Pope CA, Dockery DW. Health effects of fine particulate air pollution: lines that connect. *J Air Waste Manag Assoc* 2006;56:709-742.
9. Beelen R, Hoek G, Brandt PA, Goldbohm RA, Fischer P, Schouten LJ, et al. Long-term effects of traffic-related air pollution on mortality in a Dutch cohort (NLCS-AIR Study). *Environ Health Perspect* 2008;116(2):196-202.
10. Crouse DL, Peters PA, Hystad P, Brook JR, Donkelaar AV, Martin RV, et al. Ambient PM2.5, O3, NO2 exposures and associations with mortality over 16 years of follow-up in the Canadian Census Health and Environment Cohort (CanCHEC). *Environ Health Perspect* 2015; 123(11):1180-1186.
11. Ueda K, Nagasawa S, Nitta H, Miura K, Ueshima H Exposure to par-

- ticulate matter and long-term risk of cardiovascular mortality in Japan: NIPPON DATA80. *J Atheroscler Thromb* 2012;19:246-254.
12. Tseng E, Ho WC, Lin MH, Cheng TJ, Chen PC, Chen PC, et al. Chronic exposure to particulate matter and risk of cardiovascular mortality: cohort study from Taiwan. *Bio Med Central Public Health* 2015;15:936-944.
 13. Zhang P, Dong G, Sun B, Zhang L, Chen X, Ma N, et al. Long-term exposure to ambient air pollution and mortality due to cardiovascular disease and cerebrovascular disease in Shenyang, China. *PLoS One* 2011;6(6):e20827.
 14. Ostro B, Lipsett M, Reynolds P, Goldberg D, Hertz A, Garcia C, et al. Long-term exposure to constituents of fine particulate air pollution and mortality: results from the California teachers study. *Environ Health Perspect* 2011;118(3):363-369.
 15. Paciorek CJ, Yanosky JD, Puett RC, Laden G, Suh HH. Practical large-scale spatio-temporal modeling of particulate matter concentrations. *Ann Appl Stat* 2009;3(1):370-397.
 16. Sampson PD, Richards M, Szpiro AA, Bergen S, Sheppard L, Larson TV, et al. A regionalized national universal kriging model using partial least squares regression for estimating annual PM2.5 concentrations in epidemiology. *Atmos Environ* 2013;75:383-392.
 17. Isakov V, Irwin JS, Ching J. Using CMAQ for exposure modeling and characterizing the subgrid variability for exposure estimates. *J Appl Meteorol Clim* 2007;46:1354-1371.
 18. Jerrett M, Burnett, Kanaroglou P, Eyles J, Finkelstein N, Giovis C, Brook JR. A GIS-environmental justice analysis of particulate air pollution in Hamilton, Canada. *Environ Plann A* 2001;33:955-973.
 19. Jerrett M, Arain A, Kanaroglou P, Beckerman B, Dimitri P, et al. A review and evaluation of intraurban air pollution exposure models. *J Expo Anal Environ Epidemiol* 2005;15:185-204.
 20. Hart JE, Garshick E, Dockery DW, Smith TJ, Ryan L, Laden F. Long-term ambient multipollutant exposures and mortality. *Am J Respir Crit Care Med* 2011;183:73-78.
 21. Cesaroni G, Badaloni C, Gariazzo C, Stafoggia M, Sotgiu R, Davoli M, et al. Long-term exposure to urban air pollution and mortality in a cohort of more than a million adults in Rome. *Environ Health Perspect* 2013;121(3):324-331.
 22. Zeger SL, Dominici F, McDermott A, Samet JM. Mortality in the Medicare population and chronic exposure to fine particulate air pollution in urban centers (2000-20005). *Environ Health Perspect* 2008;116(12):1614-1619.
 23. Kim SY, Olives C, Sheppard L, Sampson PD, Larson TV, Keller JP, et al. Historical prediction modeling approach for estimating long-term concentrations of PM2.5 in cohort studies before the 1999 implementation of widespread monitoring. *Environ Health Perspect* 2017;125(1):38-46.
 24. Puett RC, Schwartz J, Hart JE, Yanosky JD, Speizer FE, Suh H, et al. Chronic particulate exposure, mortality, and coronary heart disease in the Nurses' Health Study. *Am J Epidemiol* 2008;168:1161-1168.
 25. Lee JY, Kim KH, Lee JS. Construction of an appropriate sampling design and a sample database using the National Health Information Database. Seoul: National Health Insurance Service; 2012, p. 1-51 (Korean).
 26. Song IS, Kim SY. Estimation of area representative PM10 concentration in South Korea. The Korean Society of Environmental Health and Toxicology Conference 2015(Abtract) (Korean).
 27. Kim SY. National exposure prediction approach for PM10 and NO2 in South Korea. Conference of International Society for Environmental Epidemiology and International Society of Exposure Science – Asia Chapter 2016 (Abstract) (Korean).
 28. Song IS, Kim SY. Estimation of representative area-level concentration of particulate matter (PM10) in Seoul, Korea. *J Korean Assoc Geogr Inf Stud* 2016;19(4):118-129 (Korean).
 29. Kim and Song. National scale exposure prediction for long-term concentrations of particulate matter and nitrogen dioxide in South Korea. 2017 (submitted).
 30. Park RJ, Kim SW. Air quality modeling in East Asia: present issues and future directions. *Asia-Pac J Atmos Sci* 2014;50(1):105-120.
 31. KOSIS. Statistical database: population census. Available at http://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1JC1018&conn_path=I2
 32. KOSIS. Statistical database: Community Health Survey. Available at http://kosis.kr/statHtml/statHtml.do?orgId=117&tblId=DT_HSM&conn_path=I2
 33. Huss A, Ppoerri A, Egger M, Roosli. Aircraft noise, air pollution, and mortality from myocardial infraction. *Epidemiol* 2010;21(6):829-836.