딥러닝을 이용하여 국내 노인인구의 COPD 사망 위험 추정

2018. 7. 7
한국환경정책평가연구원
사회환경연구부 강선아

딥러닝 이용 국내 노인인구 호흡기 질환 사망 위험 추정

- 목적: 딥러닝을 활용한 예측 의학 성과를 환경성 질환 분석에 적용
  - 예) 머신러닝 알고리듬으로 파악한 심혈관 발병 인자를 이용하여 추정한 발병 위험 추정치가 기존 학회 제공 발병 인자 이용 추정치보다 더 정확함을 확인(Stephen et al. 2017)
  - 실시간으로 갱신되는 데이터를 반영하여 결과를 update 할 수 있는 딥러닝의 장점 활용

- 내용: 만성폐쇄성 폐질환 사망 위험을 딥러닝을 이용하여 추정
  - 연구 대상 :65세 이상 만성폐쇄성폐질환(COPD) 환자
    - 2009년 현재 치료 중 환자 192,496명/ 2010년 전체 사망원인 중 7위에 해당
  - 자료 : 건강보험 맞춤형연구 DB , 2006-2015년 건강보험 코호트 DB version 2.0, 인구, 기후, 대기오염도 및 대기오염물질 배출량 자료를 연계
    - 맞춤형연구 DB: 만성폐쇄성 폐질환 질병에 영향을 끼치는 요인 분석
    - 건강보험 코호트 DB: 인구 특성 (성별, 연령) , 건강 관련 특성(병력, 식전혈당..), 진료기록
    - 기후: 기상청 제공 시군구별 기후 데이터
    - 환경자료: 대기오염물질 오염도, 대기오염물질 배출량

- 방법론: 딥러닝과 일반적인 호흡기 질환 사망위험 예측 모델링의 예측 정확도 비교
  - 머신러닝 방법론: Lag 변수를 변인(feature)으로 포함하는 ANN/시계열 분석이 가능한 RNN 적용 점검
  - 일반적으로 알려진 위험인자: 대한결핵 및 호흡기 학회/WHO 제공

관련문헌분석

Can machine-learning improve cardiovascular risk prediction using routine clinical data? (Stephen et al, 2017)

- 연구목적: 인공 신경망 등 네 가지 기계학습 알고리즘을 통해 환자의 진료기록을 분석하여 심혈관 질환의 발병 과 관련된 패턴 파악
- 데이터: 2005~2010년 30세에서 84세의 378,256명의 코호트 데이터

관련문헌분석

- 연구 방법론:

1. ACC(미국 심장병 학회)/AHA(미국 심장 협회)에서 만든 가이드라인에서 제시한 심혈관 질환 위험 인자와 코호트 데이터를 기반으로 logistic regression, random forest, gradient boosting, neural network 결과 심혈관 질환 위험 인자 비교

2. 위험인자 분석 결과를 바탕으로 심혈환 질환 발병 예측

- 연구 결과

머신러닝을 이용한 예측의 정확도가 높은 것으로 나타남(neural network가 가장 높음)

Table 4. Performance of the machine-learning (ML) algorithms predicting 10-year cardiovascular disease (CVD) risk derived from applying training algorithms on the validation cohort of 82,989 patients. Higher c-statistics results in better algorithm discrimination. The baseline (BL) ACC/AHA 10-year risk prediction algorithm is provided for comparative purposes.

Algorithms	AUC c-statistic	Standard Error*	95% Confidence Interval		Absolute Change from Baseline
			LCL	UCL	
BL: ACC/AHA	0.728	0.002	0.723	0.735	—
ML: Random Forest	0.745	0.003	0.739	0.750	+1.7%
ML: Logistic Regression	0.760	0.003	0.755	0.766	+3.2%
ML: Gradient Boosting Machines	0.761	0.002	0.755	0.766	+3.3%
ML: Neural Networks	0.764	0.002	0.759	0.769	+3.6%

딥러닝을 이용하여
국내 노인인구의 COPD 사망 위험 추정

Study framework

Data collection

1. 대기오염 농도 데이터 수집(시간단위: 2006년~2015년 일단위, 공간단위: 측정소)
2. 기상데이터(시간단위: 2006년~2015년 일단위, 공간단위: 측정소)
3. 맞춤형DB(건보 심의완료/ 자료구축 중,
   수집요청 데이터: 호흡계통 질환 중 만성 하부호흡기 질환(J40-J46))
4. 코호트DB(향후 신청 예정)

딥러닝을 이용하여
국내 노인인구의 COPD 사망 위험 추정

Study framework

Data preprocessing

1. 대기오염 농도 데이터: 공간 해상도를 일치시키기 위해 측정소별 데이터를 평균내어 시군구 단위로 전처리
2. 기상데이터(온도, 습도, 강수량만 취급): 시군구 단위로 데이터를 보간하기 위해 크리깅 기법 사용
3. 맞춤형 DB+대기오염 농도 데이터+기상 데이터 결합(key: 시군구 코드)
4. 코호트 DB+대기오염 농도 데이터+기상 데이터 결합(key: 시군구 코드)

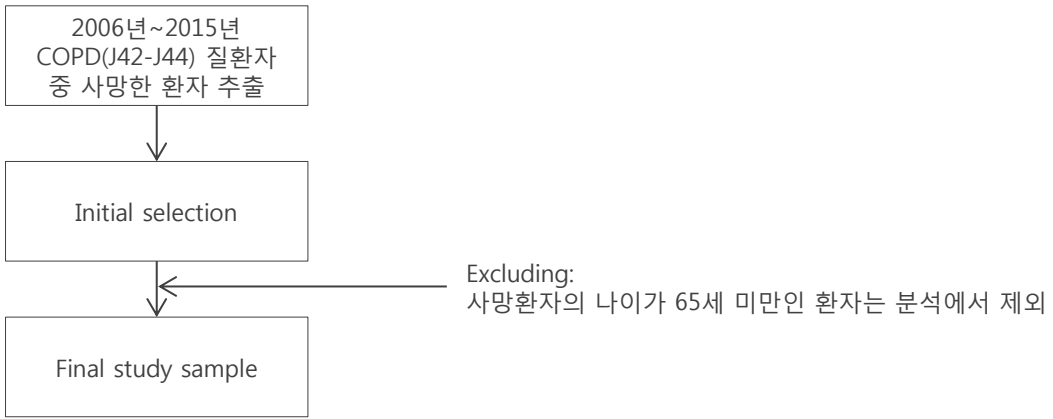
딥러닝을 이용하여
국내 노인인구의 COPD 사망 위험 추정

Study framework

Data collection → **Data preprocessing** → Effect analysis → prediction

Data preprocessing

질병 데이터 추출



딥러닝을 이용하여
국내 노인인구의 COPD 사망 위험 추정

Study framework

Data preprocessing

질병 데이터 추출

- 연도: 2006-2015년
- 상병내역: J42, J43, J44
- 서식코드: T1(의과_보건기관) - 의과입원(02), 의과외래(03)

딥러닝을 이용하여
국내 노인인구의 COPD 사망 위험 추정

Study framework

Effect analysis(변수를 고정)

- 분석기간: 2006년~2015년(연구기간은 환경 데이터와 환자 데이터 매핑 가능성 여부에 따라 변동)
- 방법론: Logistic Regression, GAM
- 대상질병: COPD 질병
- 변수: 대기오염 데이터(PM10, O3), 기후데이터(평균온도, 평균습도), 흡연유무, 흡연기간, 나이, 성별, 소득, COPD 중증도, comorbidity, 거주지역(metropolitan, urban, rural)

1. Logistic Regression, GAM analysis: COPD 질환자의 사망에 영향을 미치는 변수 추출

2. HEAT package(서울대학교 임연희 교수, 2015)
: GLM(generalized linear model)을 통해서 변곡점을 찾고 그 전후의 기울기 방향과 크기를 확인
데이터의 lag을 설정하는 파라미터를 가지고 있음

딥러닝을 이용하여
국내 노인인구의 COPD 사망 위험 추정

Study framework

Effect analysis

1. 고려사항
   Confounding variable: 개인정보, 흡연여부, 요일, 시간, 계절, 날씨 요인 통제(더미변수)

2. 노출기간: 단기-day1, day2, day3, ave(Ren et al, 2017)
   장기-1년 간의 노출, 사망 전 5년간의 평균 노출, 사망 당해연도의 노출(Stockfelt et al, 2017)

3. 오염정도: the daily maximum 8-hour concentration of air pollution variables(Di et al, 2017))

미세먼지가 안전 수치임에도 노인인구의 사망률을 높일 수 있다는 연구결과(Di et al, 2017)가 있었으므로
대기오염의 농도와 노출기간을 모두 고려하는 것이 바람직함.

딥러닝을 이용하여
국내 노인인구의 COPD 사망 위험 추정

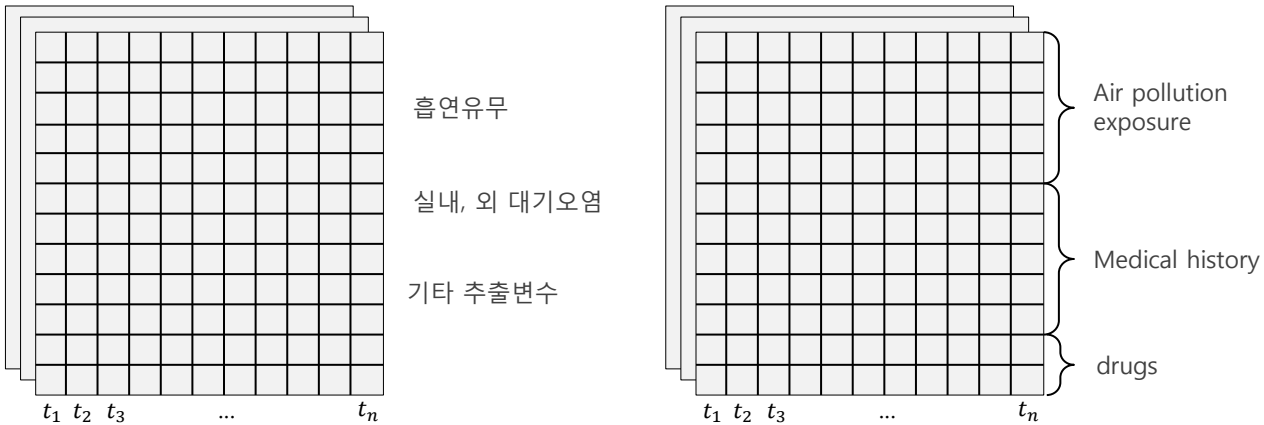
Study framework

Effect analysis(머신러닝 기법을 통한 변수 추출)

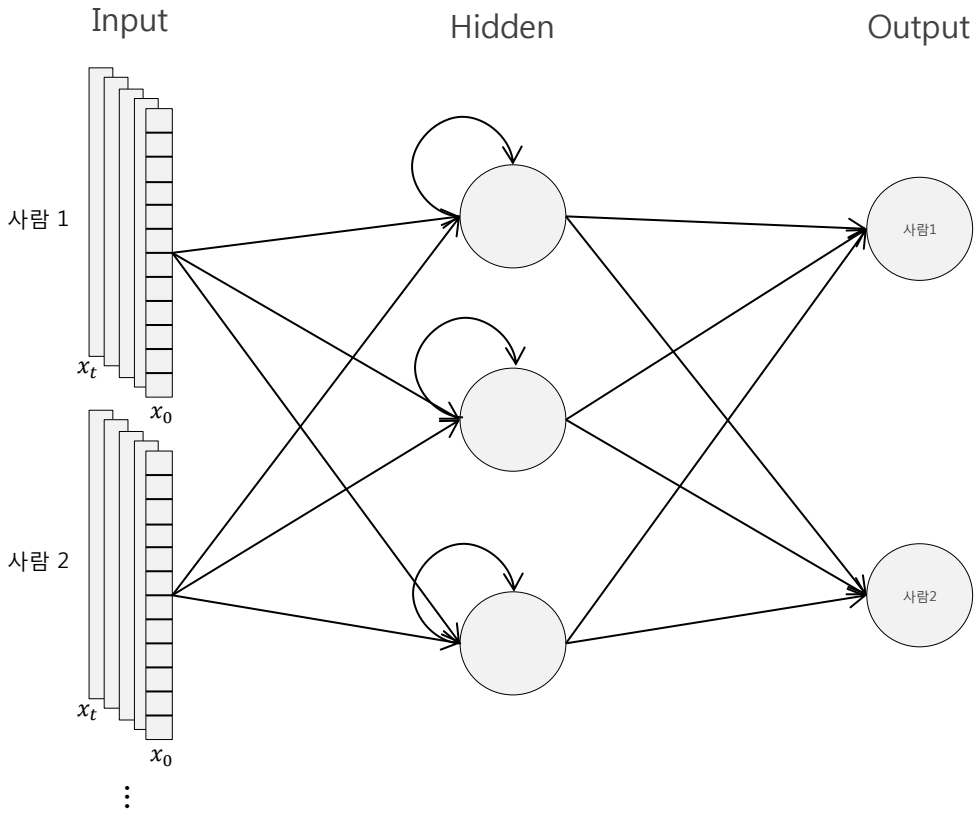
- 분석기간: 2006년~2015년(연구기간은 환경 데이터와 환자 데이터 매핑 가능성 여부에 따라 변동)
- 방법론: 머신러닝 기법(OLS, random forest, boosting)
- 대상질병: COPD 질병
- 변수: 대기오염 데이터(PM10, O3), 기후데이터(평균온도, 평균습도), 맞춤형 DB 데이터

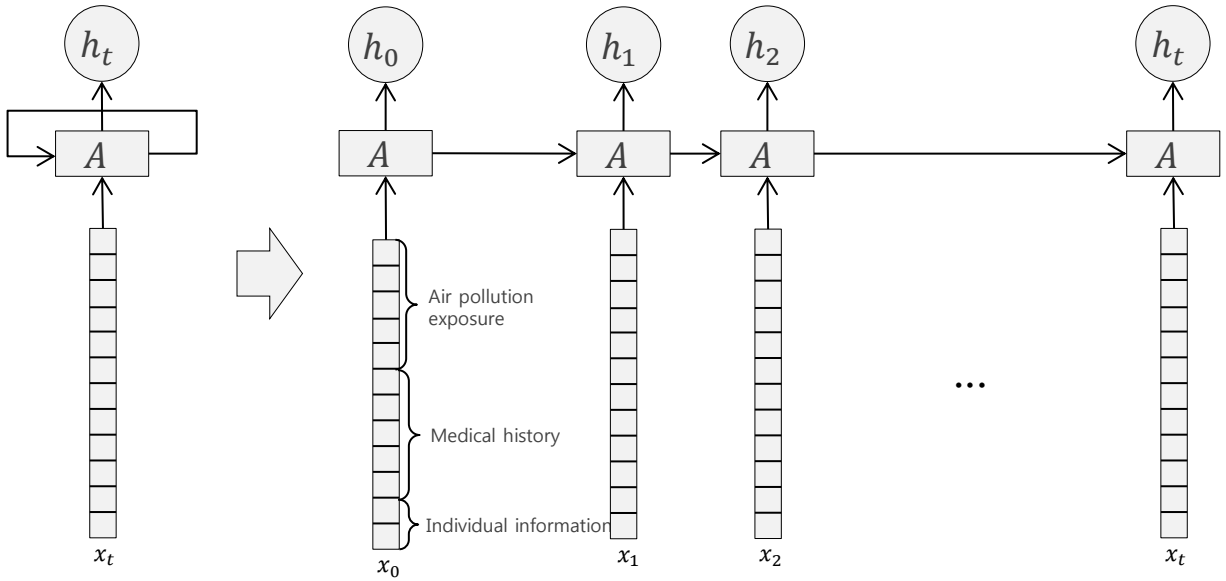
예측 데이터: 표본 코호트 2.0
예측 대상: 65세 이상 COPD 환자의 사망위험

예측 전 데이터 전처리 작업을 통해 환자 기반 진료 기록, 대기오염, 기후 데이터를 matrix 형태로 변형(Aczon et al, 2017)



Data collection → Data preprocessing → Effect analysis → **prediction**




h_t
A
x_t
h_0
h_1
h_2
h_t
Air pollution exposure
Medical history
Individual information
x_0
x_1
x_2
x_t

Reference

Aczon, M., Ledbetter, D., Ho, L., Gunny, A., Flynn, A., Williams, J., & Wetzel, R. (2017). Dynamic Mortality Risk Predictions in Pediatric Critical Care Using Recurrent Neural Networks. *arXiv preprint arXiv:1701.06675*.

Di, Q., Dai, L., Wang, Y., Zanobetti, A., Choirat, C., Schwartz, J. D., & Dominici, F. (2017). Association of short-term exposure to air pollution with mortality in older adults. Jama, 318(24), 2446-2456.

Løkke, A., Lange, P., Scharling, H., Fabricius, P., & Vestbo, J. (2006). Developing COPD: a 25 year follow up study of the general population. *Thorax*, *61*(11), 935-939.

Ren, M., Li, N., Wang, Z., Liu, Y., Chen, X., Chu, Y., ... & Xiang, H. (2017). The short-term effects of air pollutants on respiratory disease mortality in Wuhan, China: comparison of time-series and case-crossover analyses. Scientific reports, 7, 40482.

Stockfelt, L., Andersson, E. M., Molnár, P., Gidhagen, L., Segersson, D., Rosengren, A., ... & Sallsten, G. (2017). Long-term effects of total and source-specific particulate air pollution on incident cardiovascular disease in Gothenburg, Sweden. *Environmental research*, *158*, 61-71.