딥러닝을 이용하여
국내 노인인구의 호흡기 질환 사망 위험 추정

2018. 2. 28

강선아

딥러닝을 이용하여
국내 노인인구의 호흡기 질환 사망 위험 추정

▪ 연구배경 및 목적

- 환경오염은 인간의 건강에 치명적인 영향을 미침

- Lencet(2015)의 보고에 의하면 2015년 환경 오염으로 인해 사망한 환자는 전세계 사망인구의 15%정도에 해당함
- 인간의 건강을 위협하는 환경 오염원은 대기, 수질, 기온 상승 등의 기후변화, 오존층 파괴 등이 있으며(질병관리본부, 2017), 환경 오염원 중 대기오염으로 인한 사망자가 가장 높았으며, 수질오염이 그 뒤를 이었음(Lencet, 2015).
- 즉, 환경오염이 심화된 지역일수록 인간의 건강에 치명적인 피해를 미치며, 이를 예방할 필요가 있음.

- 최근 딥러닝의 발달은 예측 의학, 예방 의학을 가능케함

- 과거 의사들에 의해 주로 이루어지던 진단의학에서 나아가 딥러닝은 발달은 예측 의학, 예방 의학을 가능케함
- 특히 딥러닝의 경우 데이터가 실시간으로 쏟아지는 상황에서 총체적인 분석을 해야하는 경우 적합한 분석방법론으로 환경오염과 인간의 특성을 종합적으로 고려하여 질병을 예측하는 분야에 적합할 것으로 판단됨

환경성 질환(environmental burden disease)

- 정의

역학조사 등을 통하여 환경유해인자와 상관성이 있다고 인정되는 질환을 말하며, 수질오염에 의한 질환, 유해화학물질에 의한 중독증과 신경계 및 생식계 질환, 석면에 의한 폐질환, 환경오염사고로 인한 건강장애 그리고 대기오염과 관련된 호흡기 및 알레르기 질환을 의미(국립환경과학원, 2009)

- 국제기구 및 각국의 환경성 질환 목록

수인성 질환, 신경계질환(수은원인), **호흡기 질환**, 심장질환이 공통적으로 주로 언급

국제기구	환경성 질환 목록
WHO	설사질환, 신경정신질환, 중독, 하기도, 상기도, 만성폐쇄성 폐질환, 천식, 심혈관 질환, 백내장, 난청, 말라리아, 샤가스병, 회선사상충증, 리슈만편모충증, 뎅기, 일본뇌염, 성병, HIV, B형 C형 간염, 결핵, 주산기문제, 선청성 기형, 암, 영양실조, 육체활동, 차사고, 낙상, 익사, 트라코마, 림프사상충증, 주혈흡충증
UNEP	설사질환, 콜레라, 장티푸스, 발달장애, 내분교란, 천식, 피부결절, 청력저하, 급성독성, 지능저하, 빈혈, 뇌발달저하, 뇌성마비, 청색아증후군, 괴저, 치아불소 침착증, 주혈흡출증, 리슈마니아증, 옴, 피프스, 기니충간염, 수면병, 황열
미국 NIEHS	수인성 질환, 신경계질환, 생식계질환, 납중독, 수은중독, 우라늄중독, 유쇼중독, 아연중독, 천식, 폐기종, 진폐종, 피부염, 색소성건피증, 알레르기질환, 심장질환, 시력문제, 출생결함, 갑상선종, 직업관련질환, 신장질환, 골다공증, 퀸즈랜드열, 일광화상, 충치
미국 CPRC	신경계질환, 신경계발달관련, 생식계장애, 납중독, 만성폐쇄성 폐질환, 천식, 피부염, 심장질환 및 뇌졸중, 신장질환, 당뇨병, 자가면역장애, 다중화학민감증후군, 만성피로증후군
환경 보건법	수인성질환, 신경계질환, 생식계질환, 중독증, 호흡기질환

출처: 정기혜(2009), 환경성 질환 및 어린이 환경유해인자의 관리 동향.

출처: 국립환경과학원(2009), 환경성질환의 이해와 국내 동향.
정기혜(2009), 환경성 질환 및 어린이 환경유해인자의 관리 동향.

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
- 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

Table 3. Top 10 risk factor variables for CVD algorithms listed in descending order of coefficient effect size (ACC/AHA; logistic regression), weighting (neural networks), or selection frequency (random forest, gradient boosting machines). Algorithms were derived from training cohort of 295,267 patients.

ACC/AHA Algorithm		Machine-learning Algorithms			
Men	**Women**	**ML: Logistic Regression**	**ML: Random Forest**	**ML: Gradient Boosting Machines**	**ML: Neural Networks**
Age	Age	Ethnicity	Age	Age	Atrial Fibrillation
Total Cholesterol	HDL Cholesterol	Age	Gender	Gender	Ethnicity
HDL Cholesterol	Total Cholesterol	SES: Townsend Deprivation Index	Ethnicity	Ethnicity	Oral Corticosteroid Prescribed
Smoking	Smoking	Gender	Smoking	Smoking	Age
Age x Total Cholesterol	Age x *HDL Cholesterol*	Smoking	*HDL cholesterol*	*HDL cholesterol*	Severe Mental Illness
Treated Systolic Blood Pressure	Age x Total Cholesterol	Atrial Fibrillation	HbA1c	Triglycerides	SES: Townsend Deprivation Index
Age x Smoking	Treated Systolic Blood Pressure	Chronic Kidney Disease	Triglycerides	Total Cholesterol	Chronic Kidney Disease
Age x *HDL Cholesterol*	Untreated Systolic Blood Pressure	Rheumatoid Arthritis	SES: Townsend Deprivation Index	HbA1c	*BMI missing*
Untreated Systolic Blood Pressure	Age x Smoking	Family history of premature CHD	BMI	Systolic Blood Pressure	Smoking
Diabetes	Diabetes	COPD	Total Cholesterol	SES: Townsend Deprivation Index	Gender

관련문헌분석

- 연구 방법론:

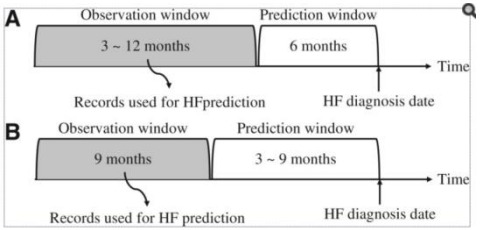
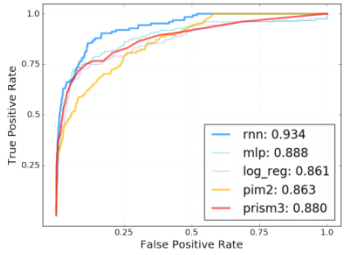
1. ACC(미국 심장병 학회)/AHA(미국 심장 협회)에서 만든 가이드라인에서 제시한 심혈관 질환 위험 인자와 코호트 데이터를 기반으로 logistic regression, random forest, gradient boosting, neural network 결과 심혈관 질환 위험 인자 비교

2. 위험인자 분석 결과를 바탕으로 심혈환 질환 발병 예측

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%

관련문헌분석

구 분			선행연구와의 차별성		
			연구목적	연구방법	주요 연구내용
주요 선행 연구	분석 방법	1	- 논문명: Using recurrent neural network models for early detection of heart failure onset - 연구자(년도): Edward, C. et al(2017), Journal of the American medical informatics association - 연구목적: 전자건강기록을 바탕으로 심부전 초기 진단 예측	- Rnn - MLP - SVM - KNN	- 데이터: 2000년 5월 16일~2013년 5월 23일 Eletronic health record의 심장마비 환자 3,884명과 primary care patients 28,903명의 진료기록 - 연구내용:  - 연구결과: 12개월의 observation window에서는 RNN의 정확도가 0.777로 가장 높았으며, 18개월의 observation window 역시 RNN의 정확도가 0.883로 다른 방법론에 비해 높았음
		2	- 논문명: Dynamic Mortality Risk Predictions in Pediatric Critical Care Using Recurrent Neural Networks - 연구자(년도): Melissa, A. et al(2017) - 연구목적: EHR의 데이터로 RNN을 이용하여 환자의 사망 위험확률을 예측	- RNN - SIM2 - PRISM3 - MLP - 로지스틱 회귀분석	- 데이터: 10년 이상 입원한 환자 1,2000명의 진료기록 - 연구내용: vital, labs, interventions, drugs와 시점 데이터를 이용하여 환자가 특정 시점에 사망할 확률을 RNN을 이용하여 모델링 - 연구결과: 

딥러닝을 이용하여
국내 노인인구의 호흡기 질환 사망 위험 추정

- 연구 대상: 65세 이상의 만성폐쇄성폐질환(COPD)자
- 데이터: 2002년~2012년 코호트 DB version 1.0, 인구 데이터, 대기오염물질 농도, 대기오염물질 배출량

- 개인 data: 성별, 연령, 소득, 장애중증도, 장애유형, 신장, 체중, 혈압, 식전혈당, 콜레스테롤, 요당, 병력, 흡연유무, 음주량
- 진료기록: 입원, 외래진료, 유사질병 입원, 처방약, MRI 진료기록
- 대기오염물질 농도
- 대기오염물질 배출량

딥러닝을 이용하여
국내 노인인구의 호흡기 질환 사망 위험 추정

- 연구 대상: 65세 이상의 만성폐쇄성폐질환(COPD)자
- 데이터: 2002년~2012년 코호트 DB version 1.0, 인구 데이터, 대기오염물질 농도, 대기오염물질 배출량
- 연구 방법론:

1. 머신러닝을 이용한 만성폐쇄성폐질환 위험인자 도출 vs 일반적으로 알려진 위험인자

* 만성폐쇄성폐질환: 나이가 들면서 생기고 장기간 흡연자에게 나타나는 질병으로, 서서히 진행하며 처음에는 가벼운 호흡곤란과 기침이 간혹 나타나지만 병이 진행되면서 호흡곤란이 심해짐, 말기에는 심장기능도 떨어짐

- 2008년 국민건강영양조사에 따르면 한국 COPD 유병률은 13.4%임
- 2009년 심평원 자료에 의하면 COPD 진단으로 치료 중인 환자는 192,496명이며, 평균 연령은 69.3세 남성이 63.1%를 차지함
- COPD 사망자 수가 증가하고 있으며, 2010년에는 전체 사망원인 중 7위.
- 특히 80세 이상 전체 사망원인 중 5위로 10만 명 당 3,732명이 사망

	위험인자
대한결핵 및 호흡기학회	- 흡연 - 숙주인자: 유전자, 노령, 성별, 폐성장, 기도과민반응 - 외부인자: 외부 유해물질(흡연, 직업성 분진과 화학물질, 실내외 대기오염), 사회 경제적 수준, 만성기관지염, 호흡기 감염
WHO	- 흡연 - 실내공기오염, 실외대기오염 - 직업성 분진 및 화학물질 - 어린 시절 잦은 호흡기 감염

딥러닝을 이용하여
국내 노인인구의 호흡기 질환 사망 위험 추정

- 연구 방법론:

2. 예측 모델링 비교

* 딥러닝과 일반적인 호흡기 질환 사망위험 예측 모델링의 예측 정확도 비교((t+1)시점)

개인신상정보 / Medical history

개인코드	연도	성별	연령	소득	장애유무	입원기록	거주지	오염원 및 배출량 데이터		사망
1111	2006									0
1111	2007									0
…										0
1111	2015									1

감사합니다