데이터사이언스 R 데이터 조별과제

COVID19 연관 질병 ICD코드로 분석한 환자 증상 분석 (Clinical Characterization of COVID-19 by EHR)

제출/작성자 B조 정대욱, 홍상표, 조성민
담당 정유채 교수님

요약

코로나 바이너스로 사회적 거리두기가 지속하고 있다. 코로나 환자의 증가세가 국내외로 많은 파장을 일으키며 세상의 변화가 왔으나 회복자의 증가세 역시 높기 때문에 코로나가 모든 사람에게 치명적이지 않다는 생각을 하게 된다. 여기서 코로나 환자의 진단결과를 이용하여 코로나 환자에게 발생하는 질병증세가 어떠한 것이 있는지를 분석하고 주요한 증세가 어떻게 나타나는지를 EHR기록 분석으로 확인한다. 코로나 환자의 판정은 ICD코드로 기록되는데 ICD코드는 증세에 대한 압축적인 표현이기 때문에 실제 증상에 대한 통계적 발현 증세를 확인하기 어렵다.

여기서 B조팀은 코로나 환자 기록을 증세로 데이터 처리를 하여 어떠한 증세가 주요한 증세인지, 순차적으로 발현 증세를 살펴보고 어떠한 인사이트를 얻을 수 있을지를 분석한다.

1.서론

코로나 바이러스 2019(COVID-19) 대유행병이 전 세계의 경계를 허무는 가운데 전 세계의 생활 방식, 경제, 의료 서비스 등이 재편되고 있다. 바이러스의 강력함과 투과성이 두드러진다. 코로나 2019(COVID-19) 유행은 많은 나라의 건강 시스템, 공중 보건 인프라, 경제에 극심한 변종을 일으켰다. 증가하는 문헌에 따르면 폐, 심장, 면역, 응고, 간, 신장 기능 장애의 주요 실험실 및 임상 표지가 불리한 결과와 관련이 있는 것으로 확인되었다.

이번 분석의 기반으로 코로나바이러스병 2019(COVID-19)에 대한 임상 및 역학 질문을 가진 전자 건강 기록 데이터를 활용했다. 코로나 분석데이터는 5개국 96개 병원 국제컨소시엄에서 생성된 자료 사례이다. EHR 의 진단 데이터 중 국가별 코로나 확진 환자에 질환증세를 ICD로 분류한 자료에 초점을 맞추어 국가별 증상, 증상의 특이점에 대해 분석하고자 한다. EHR의 .csv 포맷양식을 데이터 분석 툴인 R 뿐만 아니라 python의 강력한 데이터 기능으로 분석 작업을 시행하였다.

2. 분석방법 Preprocessing 처리

코로나 환자의 질환에 대한 의미 있는 분석하기 위한 데이터는 여러 개의 코로나 트렌트 csv자료 중 [1]Diagnosis Data를 기반으로 하였으며, 참조한 진단데이터의 구조는 국가코드, 환자수, ICD코드, 사이트(개인정보를 위해 마스킹 처리한)으로 4의 데이터로 간략하게 구성되어 있다.

	A	B	C	D	E
1	siteid	icd_code	icd_version	num_patients	unmasked_sites_num_patients
2	France	Z29.0	10	1232	1
3	France	J12.8	10	1123	1
4	France	J96.0	10	620	1
5	France	I10	10	585	1
6	France	R06.0	10	309	1
7	France	J80	10	306	1
8	France	R50.9	10	255	1

<그림1 . 국가별 코로나 환자 진단 데이터>

국가정보는 유럽(독일,프랑드,이탈리아,영국)+미국 이며 환자 명수 등 문자열 형태로 제공받아 문자열기준으로 처리하였다.

2.1 ICD/KoICD 코드

ICD는 질병 및 관련 건강 문제의 국제 통계 분류라고 하며, 흔히 국제질병분류(International Classification of Diseases, ICD)는 사람의 질병 및 사망 원인에 관한 표준 분류 규정으로 세계 보건 기구에서 발표하는 자료이다. 과거의 정식명칭을 번역한 국제질병사인분류로 부르기도 한다. ICD 코드는 알파벳+숫자형식으로 구성되어 상세질병은 . + 숫자로 상세 분류한다. ICD는 영문기반이기 때문에 증상 분석을 위해 한글로 변환하였다. 이를 위해 [2]KoICD질병분류센터에서 KOICD의 데이터를 참조하였는데, KOICD는 대한민국 실정에 맞도록 이를 반영하여 작성된 코드이나 세계기준과 특정기준에서의 차이가 있어 유의해야 한다. KOICD 코드로 한글자료의 경우 대、중、소、세、세세분류의 단계적 분류체계로 구성되어 대분류 22개, 중분류 267개, 소분류 2,093개, 세분류 12,603개, 세세분류 6,335개(그림1 참

1) https://covidclinical.net/data/index.html
2) http://www.koicd.kr/

조)가 있다.

<그림2. KOICD에 따른 대분류 예시>

대분류상 기준으로 질환의 병태를 참조하여 코로나 환자가 증세를 분석하기 위하여 KOICD의 질환을 기준으로 KOICD <-> 발현 증세 연결 작업을 하기 위해서 ICD와 KOICD간의 데이터셋 맵핑을 하였다. 먼저 KOICD의 코드별 추출 작업을 파이션의 웹서비스 호출 기능으로 구현하였고 코드<->KOICD 한글질환명을 csv로 저장하였다.

2.2 질병질환 증세 추출

한글로 질환증세를 제공하는 대표적인 서비스로 3)서울아산병원에서 질환명으로 검색할 수 있으며 이를 이용하여 KOICD의 질환명 <-> 증세추출이 가능하다. 질환코드로 증세를 제공하는 서비스 중에서 문자열 추출이 4)연속적으로 가능한 서비스로 선택하였다. 이를 Dictionary 사전으로 구성하고 누락되는 질환명에 대해 검색조건이 없기 때문에 KOICD에서 대분류/중분류 지준으로하여 의미상 연관된 증세를 추출하였다.

ICD 코드	질환명(eng_text)	질환명(kor_text)	keyword
B97	Viral agents as the cause of diseases classified to other chapters	다른 장에서 분류된 질환의 원인으로서의 바이러스감염체	병원 감염성 폐렴
B99	Other infectious diseases	기타 감염성 질환	병원 감염성 폐렴

3) 서울아산병원 http://www.amc.seoul.kr/asan/healthinfo/disease/diseaseList.do?searchKeyword=
4) 네이버 및 영문으로 질환 증상 검색이 가능하나 추출 기능에서 적절하지 않으며 주요점은 KOICD <-> 질환명 <-> 증상을 문자열로 분석가능한 형태로 결과가 나와야 한다. 이점에 있어서 질환 증세 연결에 인의적인 면을 최대한 제한하기에는 무리가 있다.

C18	Malignant neoplasm of colon	결장의 악성 신생물	위막성 대장염
C22	Malignant neoplasm of liver and intrahepatic bile ducts	간 및 간내 담관의 악성 신생물	간의 양성 신생물

<표1. KOICD 질환명을 증상 검색 Keyword로 매핑>

질환명에 대한 증상은 연속된 문자열로 확인 할 수 있고 한 질환에 대해 다수의 증상 케이스 또는 단건으로 환자의 증상을 유추할 있다.

철결핍빈혈	스푼형 손톱||창백||심부전||어지러움||빈혈||구각염||설염||피부 긴장도 저하||호흡곤란||빈맥

<표2. KOICD 질환명 - 증상 매핑 결과>

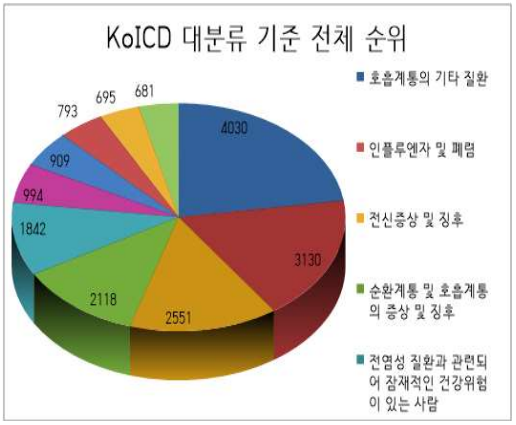
3. COVID 질환/증상 분석

코드 맵핑으로 처리된 csv파일로 하여 의미있는 분석을 하기위해 1) KOICD의 대분류/중분류 매핑, 2) 대분류/중분류로 코드 맹핑값을 재분류처리 3) 환자수 합산 및 순위 처리를 하였다.

3.1 국가별 코로나 증상 순위

환자의 ICD 작성은 상세 분류에서는 유의미한 의미를 갖지 못하였음을 분류분석에 따라 확인 할 수 있는데, 이는 ICD 진단이 정확한 질환으로 기록하기 위해서 사용했다는 것보다는 [5]환자의 증세를 현장에서 분류하기 위해 긴급하게 사용했다는 것을 의미한다. 예를 들어 환장의 정신상태를 표시하기 위해서 정신 대분류 F00-F99 정신 및 행동장해 로 분류하였는데 이 중 우울증, 심한 스트레스 반응, 뇌질환 행동장애등의 코드로 진단한 것으로 보면 이른 환자의 상태를 표시하기 위해 ICD 분류를 했다는 것을 의미한다. 따라서 코로나 증상이 집중된 병태를 알아보기 위해 환자의 장상을 대분류 기준으로 순위 분석을 하였다.

KoICD 대분류 기준	인원
호흡계통의 기타 질환	4030
인플루엔자 및 폐렴	3130
전신증상 및 징후	2551
순환계통 및 호흡계통의 증상 및 징후	2118
전염성 질환과 관련되어 잠재적인 건강위험이 있는 사람	1842
병인이 불확실한 신종질환의 임시적 지정이나 응급사용	994
고혈압성 질환	909
급성 상기도감염	793
당뇨병	695
세균, 바이러스 및 기타 감염체	681

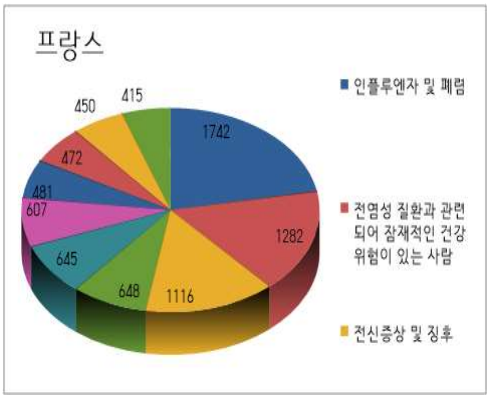


5) ICD 코드중 Z00-Z99 는 보건서비스 접촉에 영향으로 응급처치와 관련되어있다. 긴급조치, 긴급방역등에 해당한다.

<표3. 전체 대상 코로나 질병코드 10위 순위>

증상의 절반이상이 호흡계 계통의 증상을 보이고 있으며 응급처리를 위한 코드가 25%이상 사용되고 있음을 확인할 수 있다. 일반적인 호흡계 계통 증상을 제외하고는 고혈압, 당뇨등의 증세가 나타나고 있다. 이 분류는 대부분 프랑스, 이탈리아 국가에서만 주요원인으로 분석되고 있으며 독일, 미국의 경우 고혈압/당뇨 코드로 등록된 케이스는 적게 나타난다.

프랑스 / KoICD 대분로 순위	인원
인플루엔자 및 폐렴	1742
전염성 질환과 관련되어 잠재적인 건강위험이 있는 사람	1282
전신증상 및 징후	1116
순환계통 및 호흡계통의 증상 및 징후	648
호흡계통의 기타 질환	645
고혈압성 질환	607
기타 상황에서 보건서비스와 접하고 있는 사람	481
당뇨병	472
특정 처치 및 건강관리를 위하여 보건서비스와 접하고 있는 사람	450
가족 및 개인 기왕력과 건강상태에 영향을 주는 특정 병태에 관련된 잠재적 건강위험을 가진 사람	415



<표4. 프랑스 질병코드 10위>

3.2 COVID환자 수 증상 가중치

질환을 증상(질환에 대한 증상 설명으로 1:n)을 전체 환자에 대하여 주요 순위를 추출하였다. 코로나 환자에 대한 ICD코드를 바탕으로 하여 어떤 증세가 우세하게 발생할 것을 표시하기 위해서 환자수를 범위별로 가중치로 한 것과 비교하여 구하였다. 1,2,순위는 동일하게 고열과 호흡곤란이 우세하였으며 가중치에 따라 복부통증, 구토의 소화기관의 증세가 다른 순위가 나오는 것을 알 수 있다.

순위	증상20순위 (가중치)	가중치 점수
1	**열**	218
2	**호흡곤란**	193
3	**기침**	158
4	**피로감**	122
5	**가래**	111
6	**두통**	111
7	**근육통**	106
8	**복부 통증**	99
9	**구토**	95
10	**오한**	91
11	**의식 저하**	87
12	**가슴 통증**	78
13	**설사**	76
14	**청색증**	69
15	**식욕부진**	69

순위	증상 20순위	점수
0	**열**	108
1	**호흡곤란**	81
2	**복부 통증**	72
3	**구토**	69
4	**피로감**	59
5	**기침**	58
6	**두통**	58
7	**식욕부진**	53
8	**설사**	45
9	**오심**	41
10	**체중감소**	39
11	**근육통**	38
12	**가래**	36
13	**환부 부종**	34
14	**무증상**	33

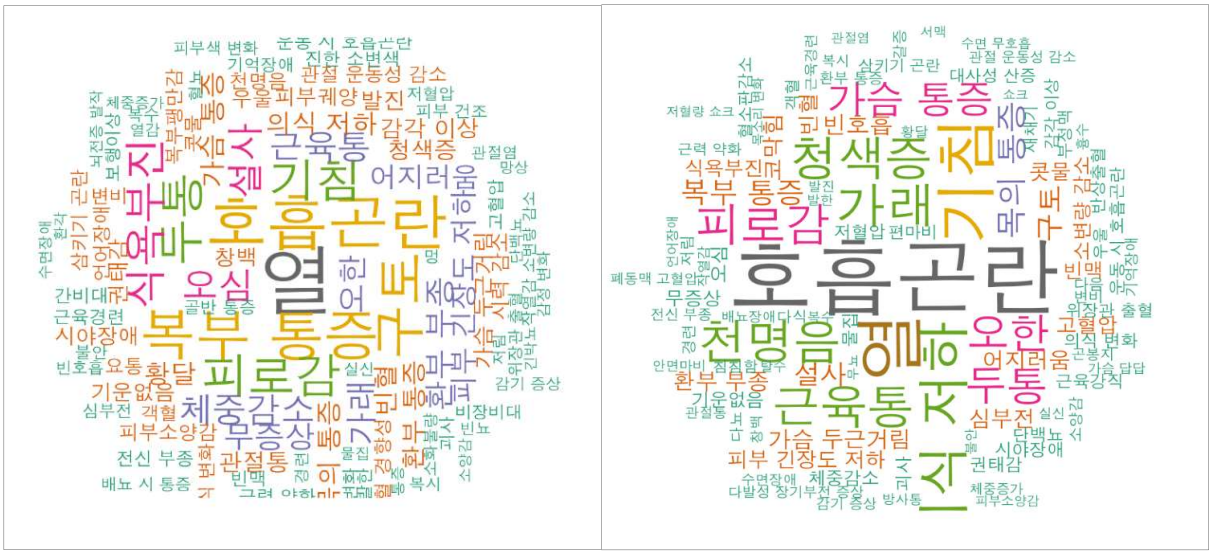
16	천명음	66	15	오한	32
17	목의 통증	64	16	피부 긴장도 저하	29
18	환부 부종	59	17	어지러움	28
19	오심	53	18	가슴 통증	27
20	체중감소	48	19	황달	26

<표5. 20순위의 주된 증상>

증상자체로 보면 호흡기관, 소화기관에 증세가 집중되어 있다는 것을 알 수 있으며, 코로나의 주된 증상의 발현은 열이라는 사실을 여기서 확인할 수 있다.

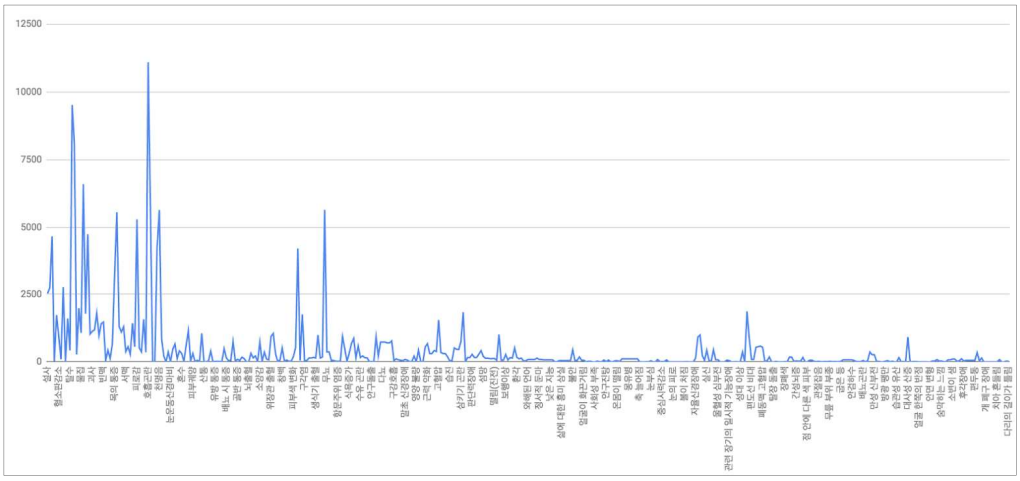
3.3 COVID환자의 증상 시각화

분석한 코로나-증상 분석을 증상 빈도에 따른 인사이트를 얻기 위하여 1) WordCount 2) 증상빈도차트 3) 증상별 연관성을 시각화하였다. 코로나 증상의 특이점을 분석하기 위해 증상 가중치별로 문자열을 카운트하였다(KOICD/질환증상 => 증상 * 환자인원수) 증상순위와 동일하게 열/호흡곤란이 주요 증세임을 확인가능하다.

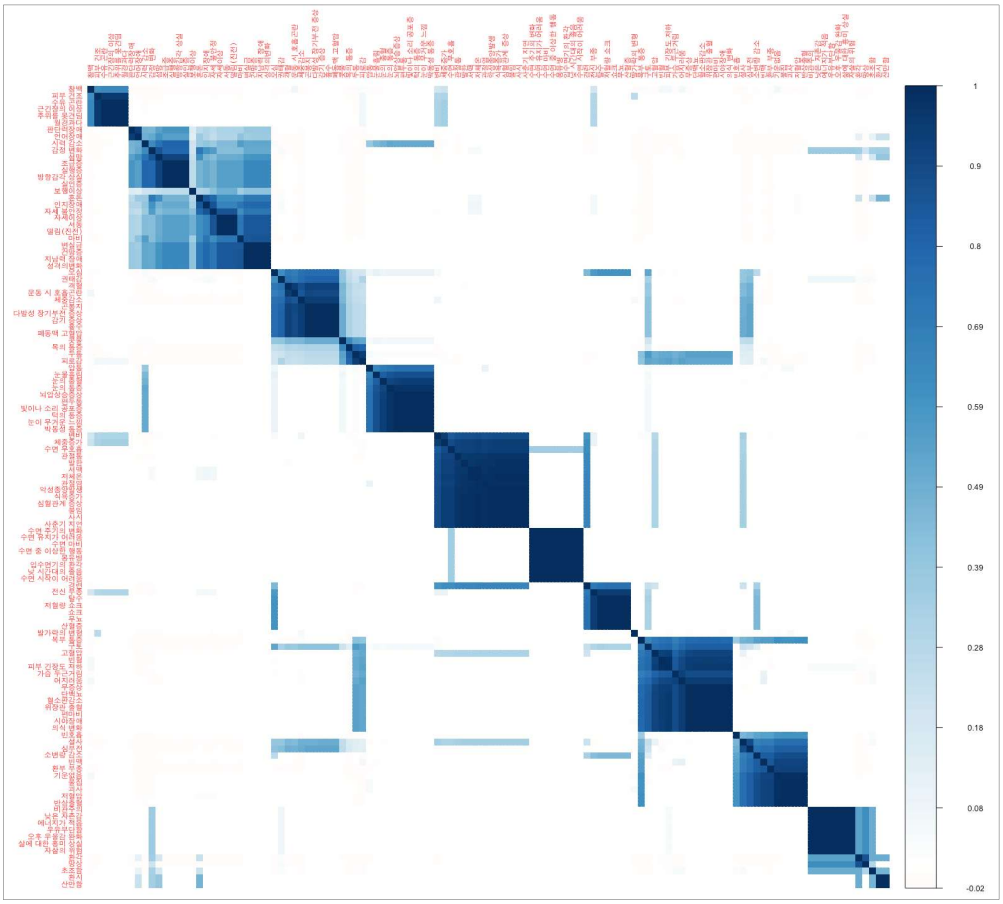


<그림3. 환자/증상 WordCount, 왼쪽 환자,오른쪽 증상>

다음은 증상 단어를 축으로 가중치로 표시한 차트이다. 환자수 가중치로 처리한 데이터로 위의 시각화한 내역과 동일하며 주된 증세의 비율이 매우 압도적인데 상위10가지 증세가 주요하게 장악하고 있다.



<그림4. 자주 발생하는 증상 단어에 대한 크기 그래프>



<그림5. 증상별 연관성 Correlation>

4. 결론

우리 연구의 흥미로운 측면은 코로나에서 보이는 증상들이 일반적으로 알려진 증세를 재확인할 수 있었다. 일반적으로 알려진 폐질환, 인플루엔자와 ICD 코드를 증

상 단어로 변화하면서 주된 증상이 이와 같이 나올 가능성이 있어 실세 ICD의 기록에서 발생하는 정확하지 않은 분류, 단어 맴핑시 증상 사전이 정교하지 못하다는 점에 있어 몇몇 증상이 dominate하다. 코로나의 증상이 주요한 발현은 열이라는 점을 이 분석에서 다시 확인하였으며 기관계, 소화기관 순으로 증상이 나타난다.

환자기록 EHV의 통계적 메타 데이터로 질환-증상 단어 맵핑을 통한 코로나 형태를 이 케이스를 통해 분석하였다. 추가적으로 국가별 차이에 대해서 분석할 여지가 남아있으며, 국가별 환자 발생 현황과 교차하여 증상의 발현시기 등에 대해 추론할 연구 여지가 남아있다.