Predicting Drug Side Effects and Treatment Failure of Tuberculosis

2019. 09. 22 (일)

Team 12

[Korea Clinical Datathon 2019]

Team 12 :)

의사

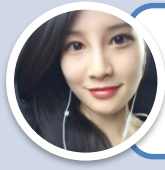
김동윤
가톨릭대학교/내과/내과전문의
- 팀 리더
- 연구 방법 설계

지성환
서울아산병원/인턴
- 연구 방법 설계
- 의학 지식 설명

최성욱
국군고양병원/여성의학/산부인과전문의
- 의학 지식 설명
- 데이터 분석

Data Scientist

강영현
Medtronic/대외협력부/사원
- 데이터 분석
- 데이터 예측

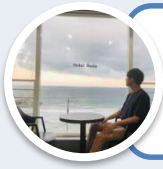


김동민
가톨릭대학교/의료정보학/박사과정
- Concept Id 매핑
- 데이터 분석

정재균
KAIST/의과학대학원/한의사
- 데이터 클렌징
- 데이터 분석

SQL 전문가

김동환
HDJunction/개발팀/CTO
- 데이터 추출
- 데이터 클렌징



이지수
SK C&C / 품질 OG / 선임
- 데이터 추출
- 데이터 클렌징

IT 전문가 & 학생

류한나
생명정보연구원/IT전문가
- 데이터 추출
- 데이터 클렌징

민향숙
국민대학교/학생
- 데이터 분석
- 데이터 시각화

박주용
숭실대학교/산업공학/학생
- 데이터 추출
- 데이터 클렌징

석종일
EA/FIFA/SW 개발자
- 데이터 시각화
- 어플리케이션 데모 작성

우현수
연세대/디지털애널리틱스/석사과정
- 데이터 분석
- 데이터 예측

Contents

Team 12

I. Introduction

- 배경
- 연구주제

배경) 한국, "결핵 관리 후진국"

한국, 여전한 '결핵 관리 후진국'

맹미선 기자 입력 2018년 3월 23일 11:48

우리나라가 경제협력개발기구(OECD) 회원국 가운데 결핵 발생률, 사망률 꼴찌를 기록했다.

국가명	발생률	사망률
한국	77	5.2
라트비아	37	2.8
멕시코	22	2.3
포루투갈	20	2.5
폴란드	18	1.3
터키	18	0.62
일본	16	2.4
칠레	16	2.3
에스토니아	16	1.5

OECD 가입국 결핵 지표

(단위=10만명당 명)

순위	발생률	사망률
1위	한국(70)	한국(5)
2위	라트비아(32)	라트비아(3.7)
3위	멕시코(22)	포르투갈(2.2)
OECD 평균	11.1	0.9
전 세계 평균	133	21

※2017년 기준. 자료=세계보건기구(WHO)

세계보건기구(WHO)의 2017년 국제 결핵 보고서 통계에 따르면, 2016년 한국의 결핵 발생률은 인구 10만 명당 77명으로 OECD 회원국 가운데 가장 높다. 이는 OEDC 평균인 인구 10만 명당 11.7명에 7배 수준이다. 결핵 발생률 2위를 기록한 라트비아는 인구 10만 명당 37명, 3위 멕시코는 인구 10만 명당 22명으로 나타났다.

연구주제

기존

데이터셋	• 아주대학교 Common Data Model (CDM) 데이터
표본	• 결핵 표준 초치료를 받는 환자
설명변수	• 성별, 나이, 진단 당시 검사결과, 기저질환, 약물복용, 흡연력, 음주력 등
의존변수	• 추적 AST / ALT (또는 ALP, bilirubin, Prothrombin time…)
분석방법	• Python, R ➔ 딥러닝, 머신러닝, 혹은 다른 통계분석 • ATLAS 툴 ➔ Estimation, Prediction

변경

데이터셋	• 아주대학교 Common Data Model (CDM) 데이터
표본	• 폐결핵에 해당되는 상병코드(15종류) **and** 결핵 1차 치료약제 처방
설명변수	• 나이, 성별, GFR, Creatinine, ALT, AST, HbA1c, ALP, 약제개수, 약물중단일수
의존변수	• 치료실패(2차치료제로 변경) 예측, 치료제 간독성 예측
분석방법	• Logistic regression , random forest, neural network, XGB(Extreme Gradient Boosting)


김동윤
가톨릭대학교/내과/내과전문의


지성환
서울아산병원/인턴


최성욱
고양병원/여성의학/산부인과전문의

II. Methods

- Concept id 추출
- 데이터 추출

Concept id 추출

진단 (Conditions)

Concept id	Condtions (SNOMED)
253954	Pulmonary tuberculosis
261495	Tuberculosis of pleura
255454	Tuberculosis of lung, confirmed by sputum microscopy with or without culture
256896	Tuberculosis of bronchus
256018	Tuberculosis of lung, confirmed histologically
261774	Respiratory tuberculosis, not confirmed bacteriologically or histologically
434557	Tuberculosis
434559	Miliary tuberculosis
4091167	Tuberculosis of lung, bacteriologically and histologically negative
4088075	Tuberculosis of larynx, trachea and bronchus, confirmed bacteriologically and histologically
253121	Tuberculosis of lung, confirmed by culture only
4091166	Primary respiratory tuberculosis, confirmed bacteriologically and histologically
260630	Tuberculosis of lung, bacteriological and histological examination not done
434416	Maternal tuberculosis during pregnancy, childbirth and the puerperium
256622	Pulmonary disease due to Mycobacteria

검사 (Measurement)

Concept id	Measurement
4102154	Skin test for tuberculosis, Tine test
3011588	Microscopic observation [Identifier] in Sputum by Acid fast stain
3016485	Microscopic observation [Identifier] in Bronchial specimen by Acid fast stain
4057167	Microscopy, culture and sensitivities
4268450	Respiratory microscopy, culture and sensitivities
40771922	Glomerular filtration rate/1.73 sq M.predicted [Volume Rate/Area] in Serum, Plasma or Blood
40761553	Hepatitis B virus surface Ag [Units/volume] in Serum
3007685	Hepatitis C virus Ab [Units/volume] in Serum
3006923	Alanine aminotransferase serum/plasma
3013721	Aspartate aminotransferase serum/plasma
3035995	Alkaline phosphatase serum/plasma
3028833	Bilirubin.total [Mass/volume] in Blood
40758583	Hemoglobin A1c in Blood
Concept id	**Conditions (SNOMED)**
439727	Human immunodeficiency virus infection

약제 (Drug Exposure)

1차 약제

Concept id	Drug Exposures (RxNorm)
1782573	isoniazid 100 MG Oral Tablet
19022105	Rifampin 600 MG Oral Tablet
19019700	Rifampin 150 MG Oral Capsule
19022342	Rifampin 450 MG Oral Tablet
40222616	Ethambutol Hydrochloride 400 MG Oral Tablet
1759456	Pyrazinamide 500 MG Oral Tablet

2차 약제

2군

Concept id	Drug Exposures (RxNorm)
1836193	Streptomycin 1000 MG Injection
40173560	Amikacin Sulfate 5 MG/ML Injectable Solution
45892658	Amikacin 250 MG/ML Injectable Solution

3군

Concept id	Drug Exposures (RxNorm)
19107185	Levofloxacin 100 MG Oral Tablet
1742255	Levofloxacin 500 MG Oral Tablet
1742254	Levofloxacin 250 MG Oral Tablet
1716905	moxifloxacin 400 MG Oral Tablet
19077575	gatifloxacin 200 MG Oral Tablet

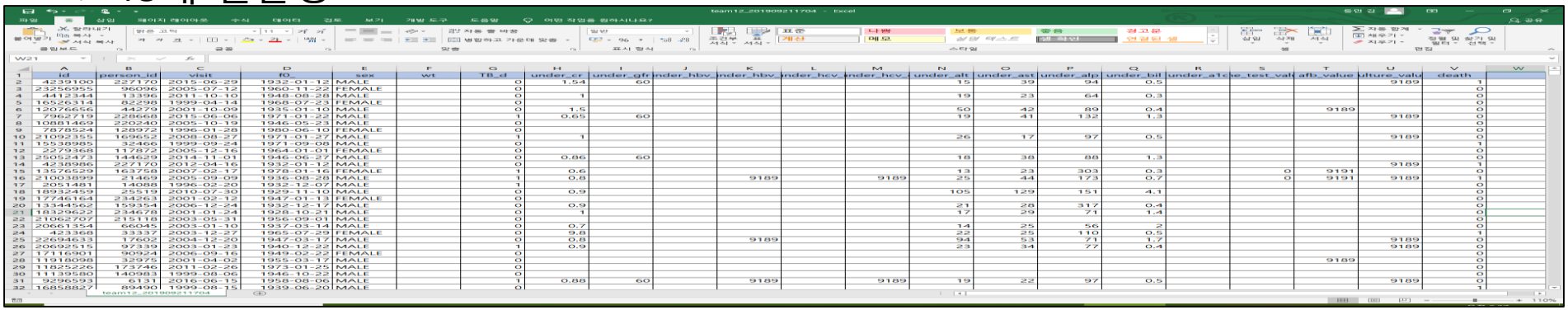
4군

Concept id	Drug Exposures (RxNorm)
1710447	Cycloserine 250 MG Oral Capsule

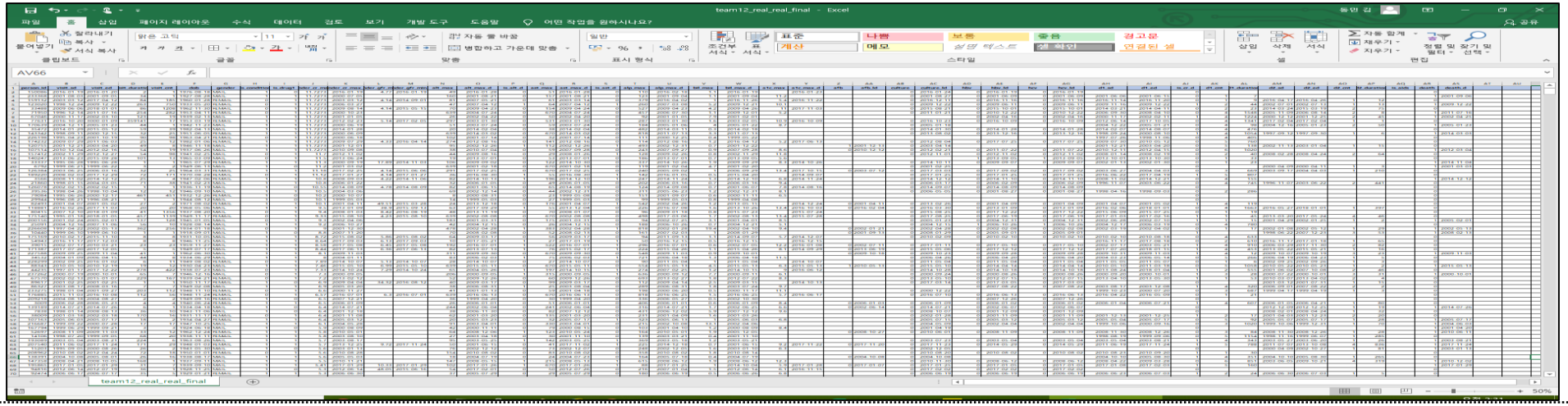
김동민 가톨릭대학교의료정보학/박사과정
정재균 KAIST/의과학대학원/한의사
지성환 서울아산병원/인턴
최성욱 고양병원/여성의학/산부인과전문의

데이터 추출

- **1차 추출 (n = 2,465)**
  - 15개 진단명



- **n차 추출 (n = 1,892)**
  - 15개 진단명 + 1차 초치료약제 복용력이 있는 환자




류한나
생명정보연구원/IT전문가


김동환
HDJunction/개발팀/CTO


이지수
SK C&C / 품질 OG / 선임


박주용
숭실대학교/산업공학/학생

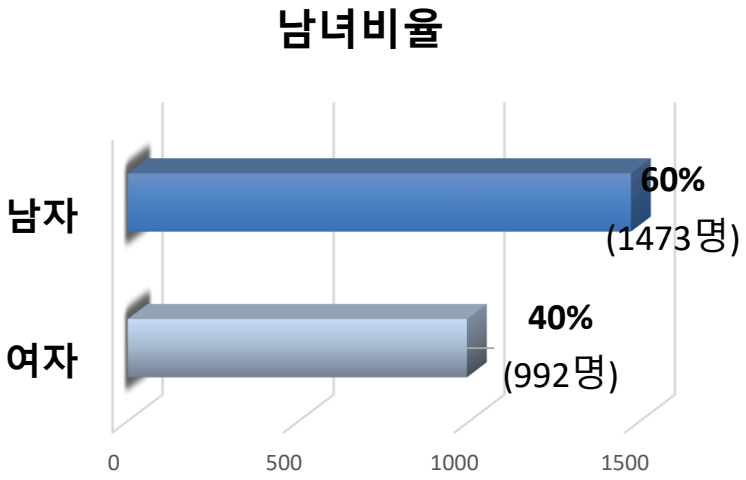
Team 12

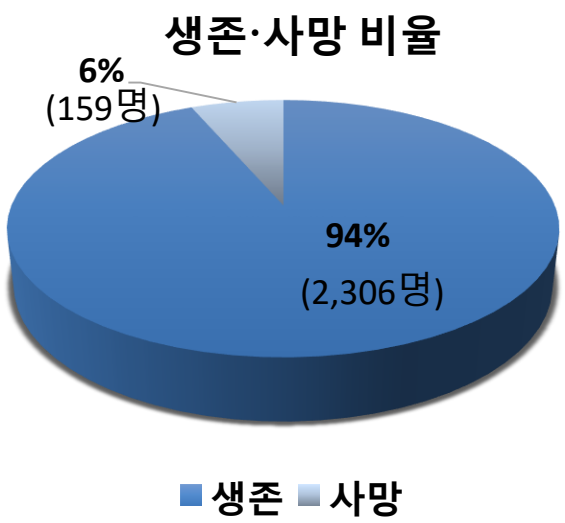
III. Results

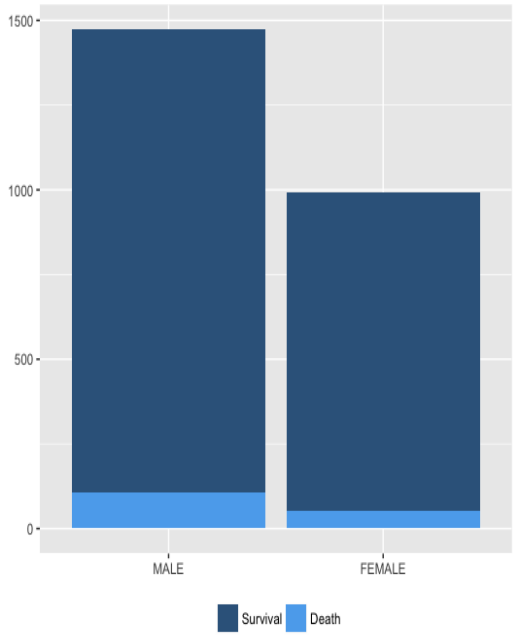
- 탐색적 분석
- 단일 모델 결과
- 스태킹 모델 결과
- 시각화

탐색적 분석

기초통계 (n = 2,465)







	Creatinine	AST	ALT	ALP	Bilirubin
평균	1.2	54.7	66.2	116.2	1.2
표준오차	0.1	4.2	4.9	3.9	0.1
중앙값	0.9	24.0	28.0	85.0	0.7
최빈값	0.8	17.0	18.0	84.0	0.4
표준편차	1.4	106.9	124.8	99.4	2.1
분산	2.0	11431.5	15575.5	9879.0	4.4
최소값	0.3	5.0	10.0	29.0	0.2
최대값	11.7	639.0	670.0	818.0	19.4

단일 모델 결과

Logistic Regression

- **Table**

	0	1
0	206	54
1	42	77

- **Measure**

ACC	0.7467
AUC	0.7197

- **Results**

```
Confusion Matrix and Statistics


pred_glm1   0   1
        0 206  54
        1  42  77

               Accuracy : 0.7467
                 95% CI : (0.6998, 0.7897)
    No Information Rate : 0.6544
    P-Value [Acc > NIR] : 6.826e-05

                  Kappa : 0.4277

 Mcnemar's Test P-Value : 0.2616

            Sensitivity : 0.8306
            Specificity : 0.5878
         Pos Pred Value : 0.7923
         Neg Pred Value : 0.6471
             Prevalence : 0.6544
         Detection Rate : 0.5435
   Detection Prevalence : 0.6860
      Balanced Accuracy : 0.7092

       'Positive' Class : 0
```

Random Forest

- **Table**

	0	1
0	211	54
1	37	77

- **Measure**

ACC	0.7599
AUC	0.7358

- **Results**

```
Confusion Matrix and Statistics


pred_rf2   0   1
        0 211  54
        1  37  77

               Accuracy : 0.7599
                 95% CI : (0.7136, 0.802)
    No Information Rate : 0.6544
    P-Value [Acc > NIR] : 5.73e-06

                  Kappa : 0.4524

 Mcnemar's Test P-Value : 0.09349

            Sensitivity : 0.8508
            Specificity : 0.5878
         Pos Pred Value : 0.7962
         Neg Pred Value : 0.6754
             Prevalence : 0.6544
         Detection Rate : 0.5567
   Detection Prevalence : 0.6992
      Balanced Accuracy : 0.7193

       'Positive' Class : 0
```

XG Boost

- **Table**

	0	1
0	223	67
1	25	64

- **Measure**

ACC	0.7573
AUC	0.744

- **Results**

```
Confusion Matrix and Statistics


pred_xgb1   0   1
        0 223  67
        1  25  64

               Accuracy : 0.7573
                 95% CI : (0.7109, 0.7996)
    No Information Rate : 0.6544
    P-Value [Acc > NIR] : 9.665e-06

                  Kappa : 0.4195

 Mcnemar's Test P-Value : 1.915e-05

            Sensitivity : 0.8992
            Specificity : 0.4885
         Pos Pred Value : 0.7690
         Neg Pred Value : 0.7191
             Prevalence : 0.6544
         Detection Rate : 0.5884
   Detection Prevalence : 0.7652
      Balanced Accuracy : 0.6939

       'Positive' Class : 0
```

Cat Boost

- **Table**

	0	1
0	210	52
1	38	79

- **Measure**

ACC	**0.7625**
AUC	0.7384

- **Results**

```
Confusion Matrix and Statistics


pred_cat1   0   1
        0 210  52
        1  38  79

               Accuracy : 0.7625
                 95% CI : (0.7164, 0.8045)
    No Information Rate : 0.6544
    P-Value [Acc > NIR] : 3.35e-06

                  Kappa : 0.4615

 Mcnemar's Test P-Value : 0.1706

            Sensitivity : 0.8468
            Specificity : 0.6031
         Pos Pred Value : 0.8015
         Neg Pred Value : 0.6752
             Prevalence : 0.6544
         Detection Rate : 0.5541
   Detection Prevalence : 0.6913
      Balanced Accuracy : 0.7249

       'Positive' Class : 0
```

스태킹 모델 결과

Stacking GLM

- **Table**

	0	1
0	198	62
1	50	69

- **Measure**

ACC	0.7045
AUC	0.6707

- **Results**

```
Confusion Matrix and Statistics


pred_glm1   0   1
        0 198  62
        1  50  69

               Accuracy : 0.7045
                 95% CI : (0.6558, 0.75)
    No Information Rate : 0.6544
    P-Value [Acc > NIR] : 0.02186

                  Kappa : 0.3323

 Mcnemar's Test P-Value : 0.29862

            Sensitivity : 0.7984
            Specificity : 0.5267
         Pos Pred Value : 0.7615
         Neg Pred Value : 0.5798
             Prevalence : 0.6544
         Detection Rate : 0.5224
   Detection Prevalence : 0.6860
      Balanced Accuracy : 0.6626

       'Positive' Class : 0
```

Stacking RF

- **Table**

	0	1
0	193	53
1	55	78

- **Measure**

ACC	0.715
AUC	0.6855

- **Results**

```
Confusion Matrix and Statistics


pred_rf2   0   1
        0 193  53
        1  55  78

               Accuracy : 0.715
                 95% CI : (0.6667, 0.76)
    No Information Rate : 0.6544
    P-Value [Acc > NIR] : 0.006919

                  Kappa : 0.3723

 Mcnemar's Test P-Value : 0.923342

            Sensitivity : 0.7782
            Specificity : 0.5954
         Pos Pred Value : 0.7846
         Neg Pred Value : 0.5865
             Prevalence : 0.6544
         Detection Rate : 0.5092
   Detection Prevalence : 0.6491
      Balanced Accuracy : 0.6868

       'Positive' Class : 0
```

Stacking XGB

- **Table**

	0	1
0	220	63
1	28	68

- **Measure**

ACC	0.7599
AUC	**0.7429**

- **Results**

```
Confusion Matrix and Statistics


pred_xgb1   0   1
        0 220  63
        1  28  68

               Accuracy : 0.7599
                 95% CI : (0.7136, 0.802)
    No Information Rate : 0.6544
    P-Value [Acc > NIR] : 5.73e-06

                  Kappa : 0.4335

 Mcnemar's Test P-Value : 0.000365

            Sensitivity : 0.8871
            Specificity : 0.5191
         Pos Pred Value : 0.7774
         Neg Pred Value : 0.7083
             Prevalence : 0.6544
         Detection Rate : 0.5805
   Detection Prevalence : 0.7467
      Balanced Accuracy : 0.7031

       'Positive' Class : 0
```


강영현 Medtronic/대외협력부/사원
민향숙 국민대학교/학생
우현수 연세대/디지털애널리틱스/석사

Interactive Visualization + R Shiny

결핵환자의 초치료실패 간독성, 신독성 예측 비율

	step	step_true	dtree	dtree_true	ann	ann_true	svm	svm_true
failure	0.71	0	0.65	0	0.68	1	0.71	0
is_alt_d	0.57	0	0.54	0	0.55	0	0.58	0
is_ast_d	0.57	0	0.54	0	0.57	0	0.57	0
is_cr_d	0.64	0	0.59	1	0.6	1	0.64	0

Web with Python & JS

- 치료실패 예측

석종일 EA/FIFA/SW 개발자
민향숙 국민대학교/학생

Team 12

IV. Conclusions

- 결론

결론

- **Why should we care?**
  - 결핵은 초치료에 실패할 경우,
    효과는 감소하고, 부작용이 많은 약을 사용해야 함
  - 후진국형 질병이므로, 의료 선진국 연구 역량 투입이 부족
- **Who Cares if you succeed?**
  - 결핵환자 ➔ 예상되는 부작용에 대해 환자에게 교육기회를 제공
  - 국민보건 ➔ 국민의 경각심 제고 및 전염병 감소
  - 의료진 ➔ 치료 경과를 예측함으로써 맞춤 치료 제공하고 주의 깊게 대응
- **Novelty of Method**
  - 최신 머신러닝 방법론 (앙상블 등)
  - 실시간 웹기반 시각화
- **Interpretation of Findings**
  - 환자 Profile 로부터 치료실패 및 간독성, 신독성 가능성을 예측

Q & A

Thank you :)