MOA-CDM : Medical record Observation and Assessment for drug safety – Common Data Model

- 식품의약품안전처 산하 한국의 약품안전관리원에서 능동적 약물감시를 위해 의료기관별로 상 이한 전자건강기록의 자료 구조와 용어 체계를 공통의 형식으 로 변환하여 구축한 분산형 데이터베이스 시스템

**□ 사업 최종목표**

- MOA-CDM 기반 주요 의약품 부작용(신독성, 간독성 등) 실마리 탐지 알고리즘 적용 가능성 검토 및 의약품 부작용 실마리 탐지 알고리즘 개발 및 다기관 통합 분석을 통해 알고리즘 타당성 검증

- MOA-CDM 기반 딥러닝 알고리즘 적용 가능성 검토 후 주요 의약품 부작용 예측 목적의 머신러닝/딥러닝 알고리즘 적용 및 국내외 다기관 모델 성능 및 유용성 검증 및 대규모 외부 검증

- 인공지능 알고리즘 기반 연합 학습 알고리즘을 통한 확산 적용 및 외부 검증 활용 목표로 합성 데이터 구축 및 공개

**□ 1차년도**

1) Transformer 기반 설명가능한 의약품 부작용 예측 모델

- Transformer는 최초 자연어 처리 문제를 해결하기 위해 제안되었음. 현재는 컴퓨터 비전 등 다양한 도메인에서 기존의 RNN, CNN 기반 접근 모델들에 비해 더 높은 성능을 보여줌.

- Transformer는 RNN의 Recurrent Connection을 이용하지 않는 대신 Poisitional Encoding과 Self-Attention의 방법을 이용하여 시퀀스 데이터를 처리함.

- 해당 방법의 이점은 RNN 계열 모델의 고질적인 문제인 학습되는 타임스텝의 초기단계의 내용을 크게 잃어버리지 않는 다는 점과 계산 과정 상 기울기 소실 문제를 회피할 수 있다는 점에 있음.

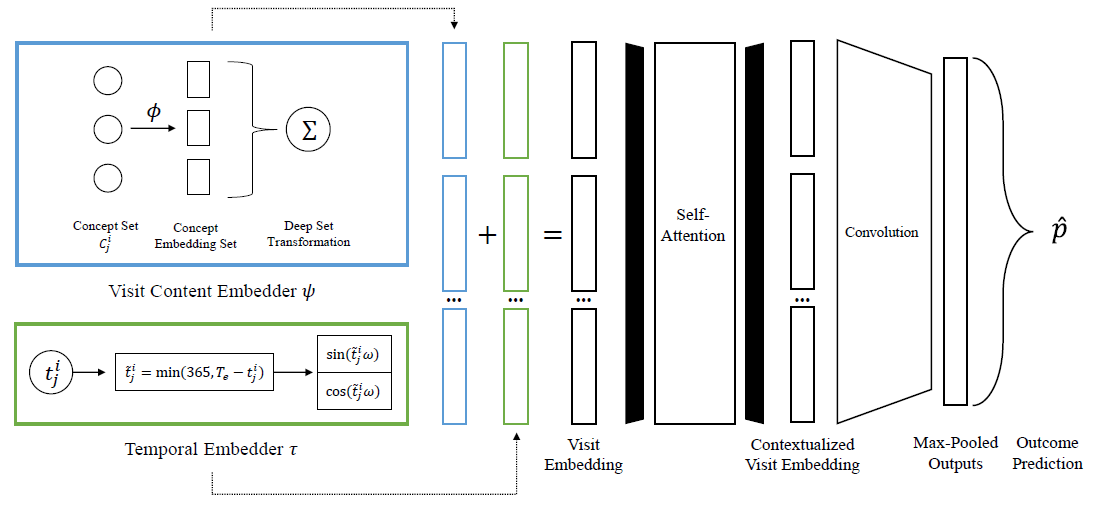


그림. SARD 알고리즘 개요도

- Self-Attention 기법을 통해 시점과 변수에 대한 기여도 또한 분석할 수 있음.

- 본 연구에서는 딥러닝 알고리즘으로 high-performing linear model로 초기화하여 역추출(reverse distillation)을 시도하는 Self Attention with Reverse Distillation (SARD) 알고리즘을 MOA-CDM에 구현할 예정

2) RNN을 기반 설명가능한 실시간 의약품 부작용 예측 모델

- 어텐션(Attention) 메커니즘은 특정 시점에 가중치를 부가하여 RNN 모델이 보다 긴 시계열 데이터의 분석을 가능하도록함.

- 위에서 언급된 RNN과 어텐션의 특징은 긴 기간의 시계열 데이터의 특징을 갖고 있는 전자의무기록과 공통데이터모델에 적합함

- EHAN (EHR History-based prediction using Attention Network) 모델은 어텐션 스코어와 Grad-CAM(gradient-weighted class activation map)을 통해 전자의무기록 에서 특정 시점의 특정 변수에 대한 기여도를 평가할 수 있는 모델임.

- 해당 방법론은 전자의무기록의 진단, 약물의 각 도메인의 특징적인 패턴을 학습하고 이들은 병합하기 때문에 CDM의 진단, 약물, 혈액검사, 처치의 도메인으로 확장하여 구현할 예정임

- 본 과제에서는 타겟 약물 투여 이후 4 주간의 데이터를 이용하여 특정 기간(2주, 1개월) 이내 신독성, 간독성의 부작용이 발생하는지 예측하는 실시간 예측 모델을 개발할 예정임

그림. EHAN 모델의 개요

3) RNN을 이용한 설명가능한 의약품 부작용 발생 및 발생 시간 예측 모델

- Weibull distribution은 분포의 형태 유연하여 Normal distribution 및 Exponential distribution 등의 다양한 통계적 분포를 흉내낼 수 있어 산업 현장의 생존 분석에서 자주 이용되는 분포임. WTTE-RNN(Weibull Time To Event RNN) 모델은 RNN모델을 이용하여 Weibull 분포를 추정하는 방식으로 이벤트에 대한 발생과 발생 시간을 예측하기위해 개발됨.

- 본 과제에서는 WTTE-RNN에 Attention 방법을 이용하여 변수의 기여도를 측정할 수 있는 모델을 구현예정임

- 타겟 약물의 투여 이전 CDM에 기록된 환자의 과거 데이터를 이용하여 투여 이후 치료기간 동안 신독성 및 간독성 관련 부작용이 발생 여부와 발생까지의 기간을 예측할 수 있는 모델을 적용예정임.

◌ 의약품 부작용 실마리 탐지 알고리즘 및 의약품 부작용 예측 딥러닝 알고리즘 참여 기관 검증을 통한 모델 성능 및 유용성 규명

- 본 과제의 참여 기관인 신촌 세브란스, 강남 세브란스, 건양대병원에서 목표로 하는 신독성 관련하여 수집 환자는 각 19,605 명, 6,479 명, 3,387 명으로 딥러닝 모델을 적용 가능한 것으로 검토함.

- 데이터 편향성에 대한 이슈 극복하기 위해 MOA CDM을 활용한 의약품 부작용 실마리 탐지 알고리즘인 트리 기반 통계량 (Tree-based scan statistic)을 공유해서 기관별 결과 분석 수행

- 기관별 결과는 메타분석을 적용하여 유의한 시그널을 판정하고, 임상적 판단과 해석을 덧붙여 실질적인 의약품과 부작용의 인과성을 평가할 예정임

◌ MOA-CDM 기반 의약품 부작용 실마리 탐지 / 인공지능 결과 분석 및

평가지표 마련

- 간독성과 신독성을 대상으로 개발한 용어 세트에 기존에 의약품 부작용 실마리 탐지에 많이 사용되는 불균형 분석은 물론, 트리기반 통계 분석 및 drug combination effect을 계산하기 위한 mixture drug-count response 모델을 적용

- 실마리 탐지 평가지표: 불균형 분석을 통해서는 PRR, ROR, IC, chi-square statistics 알고리즘 특성을 사용하고, 트리 기반 통계분석에서는 로그가능도함수(Log Likelihood Ratio, LLR)를 계산한 후 귀무가설 가정하에 검정통계량 분포를 파악하고 p-value를 도출함. 또한 복합 약물 위험 분석에서는 mixture model을 기반으로 likelihood의 값을 구하고 그에 따라 모수 추정 후 해당 모수를 대입한 LFDR을 기준으로 평가지표로 사용함

- 인공지능 예측 모델 평가지표: 발생 예측 여부 (이진 분류)에 대해 accuracy, f-1 score, AUROC, AUPRC 제시할수 있음. 발생 시점 예측의 평가지표로서 RMSE를 제시할 예정.

- 또한, 설명가능한 인공지능 모델의 결과물로서 분석된 변수 기여도의 랭크를 제시할 예정.

- 추후, 각 알고리즘의 특성을 상호 보완하여 의약품 부작용 실마리 탐지에 특화된 방법론을 개발할 수 있을 것으로 기대됨

1) 다기관 통합 모델(Weighted Based Integrated Model; WIM) 기반 연합학습 적용 및 검증

- 연합학습에서 주요 문제 중 하나는 기관 – 중앙 서버 사이의 Communication Cost (통신 비용) 줄이는 것임, Weight based Integrated model(WIM) 알고리즘은 가중치 기반으로 모델을 통합 시 일반적인 연합학습 모델에 비해 절대적인 Communication 수가 적음.

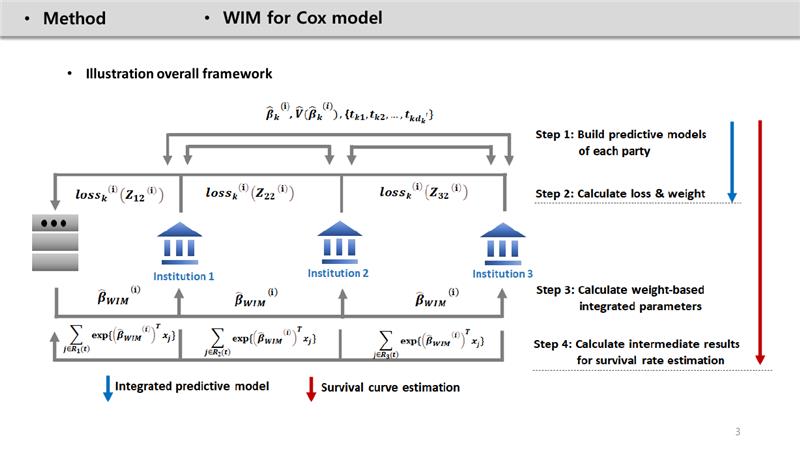


그림. Weight based Integrated Model 개요

- 본 과제를 통해 딥러닝에 적용한 모델을 개발하고 이를 다기관 의약품 약물 부작용 모델 통합 과정에 적용할 예정임.

- 1차년도에서 개발된 설명가능한 의약품 부작용 발생 및 발생 시간 예측 모델의 다기관 확장을 위해 본 알고리즘 적용 예정이며,

- 신독성 및 간독성의 부작용 발생과 발생 시간 예측모델을 다기관의 데이터를 이용해 확장시킬 예정임.

2) 기관 특이적 자료를 학습 지원을 위한 개인화된 연합학습(Personalized Federated Learning) 개발 및 적용

- 다기관 약물부작용 예측 모델 개발 연구에서는 일반적으로 기관에 공통적인 변수를 기반으로 연구를 진행하게 됨

- 연합학습(FedAvg)의 기본 가정은 분산된 환경의 데이터 분포가 독립적이고 동일하다는 i.i.d(independent and identical distribution) 가정임. 하지만, 현실 세계의 분산된 이용자 혹은 기관이 보유한 데이터 분포는 다양할 수 있으며, 이 경우 연합학습 모델 성능은 저하됨.

- 한편으로, 연합학습 모델의 디자인은 수평 연합학습과 수직 연합학습으로 나눌 수 있음. 수평 연합학습은 데이터의 변수가 동일한 채 샘플이 다른 경우를 의미하며, 수직 연합학습은 동일한 샘플 데이터에서 서로 다른 변수가 분산된 환경에 존재할 때의 접근 방법을 의미함. 수평 연합학습은 특정 사용자에 연합학습에서 이용되지 못한 수직 분할 데이터를 이용하지 못하며, 수직 연합학습은 많은 양의 동일한 샘플이 겹쳐야만 이용할 수 있음.

- 본 과제에서 개발 예정인 PPFL(Personalized Progressive Federated Learning) 모델은 앞선 두 문제를 해결하기 위함임. (Kim. et. al, In preparation)



그림. PPFL 알고리즘의 개요

- 이 때, 공통 연합학습 네트워크 내부만의 파라메터 (파란색 점선)은 학습되지 않도록 설정하여, 기 구축된 연합학습 모델의 지식을 보존할 수 있도록함. 즉, 연합학습을 학습한 통해 공통 변수의 일반화된 지식과 사용자에 특징적이게 학습된 파라메터 사이의 균형을 조절할 수 있도록 만듬. CDM 파라미터 후 특이적 파라미터 더 학습

- 본 과제에서는 1차 년도에서 개발한 RNN 기반의 설명가능한 실시간 부작용 예측 모델을 연합학습을 통해 학습시킬 예정임

**□ 2차년도**

. MOA-CDM 활용 의약품 부작용 실마리 탐지 알고리즘 고도화 작업 진행

- 1차년도 실마리 탐지 결과를 기반으로, 기존에 의약품 부작용 실마리 탐지에 많이 사용되는 불균형 분석을 추가로 수행하여 각 알고리즘의 결과를 객관적으로 비교

- 알고리즘 특성에 따라 접근 방식과 결과가 달라지기 때문에, 각각의 알고리즘에서 유의한 시그널을 확인하고, 알고리즘 간 공통적으로 탐지된 시그널에 대해서도 분석

- 각 알고리즘의 특성을 상호 보완하여 의약품 부작용 실마리 탐지에 특화된 방법론으로 고도화

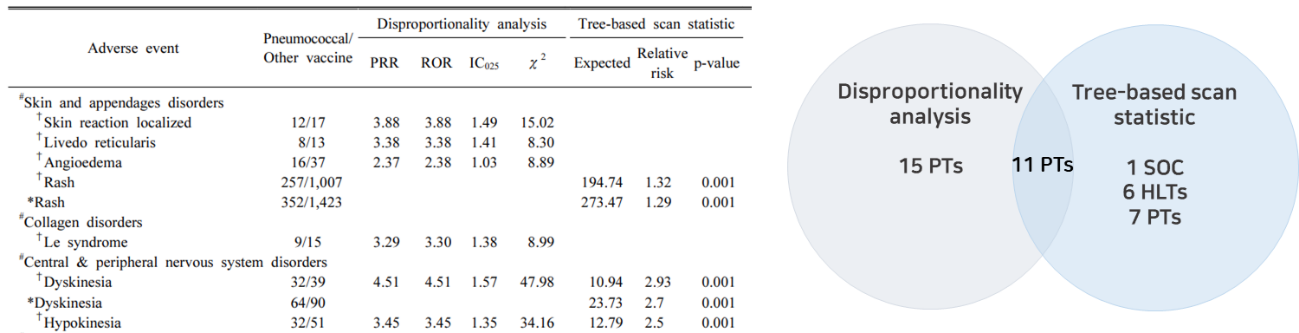


그림. 폐렴구균 백신과 부작용에 대해 불균형 분석과 트리 기반 통계분석 결과를 함께 제시한 선행연구

\* 불균형 분석 (Disproportionality analysis)

- 의약품 부작용 실마리 탐지를 위한 통계분석 중 하나로, 데이터베이스 내의 다른 의약품 및 부작용을 대조군으로 하여 상대적인 보고분율을 비교하는 방법

- 비교 대상이 되는 데이터베이스 구성에 영향을 받고, 데이터베이스에 따라 결과가 달라지는 특성을 가짐. 대표적인 지표로 PRR, ROR이 있음

|  |  |  |  |
| --- | --- | --- | --- |
|  | **관심 의약품** | **그 외 의약품** | **총합** |
| **관심 부작용** | A | B | A+B |
| **그 외 부작용** | C | D | C+D |
|  | A+C | B+D | A+B+C+D |

표. 불균형 분석 계산 표

- PRR은 관심 의약품에 의해 발생한 부작용 사례 중 관심 부작용이 차지하는 분율을, 그 외 의약품에 의한 부작용 사례 중 관심 부작용이 차지하는 분율로 나눈 값

- ROR은 관심 의약품에 의해 발생한 관심 부작용의 오즈를 그 외 의약품에 의해 발생한 관심 부작용 오즈로 나눈 값

- IC는 관심 의약품으로 인한 관심 부작용이 발생할 확률을 관심 의약품의 복용 확률과 관심 부작용의 발생 확률의 곱으로 나누어 로그를 취한 값

2) 기관 특이적 자료를 학습 지원을 위한 개인화된 연합학습(Personalized

Federated Learning) 개발 및 적용

- 다기관 약물부작용 예측 모델 개발 연구에서는 일반적으로 기관에 공통적인 변수를 기반으로 연구를 진행하게 됨

- 연합학습(FedAvg)의 기본 가정은 분산된 환경의 데이터 분포가 독립적이고 동일하다는 i.i.d(independent and identical distribution) 가정임. 하지만, 현실 세계의 분산된 이용자 혹은 기관이 보유한 데이터 분포는 다양할 수 있으며, 이 경우 연합학습 모델 성능은 저하됨.

- 한편으로, 연합학습 모델의 디자인은 수평 연합학습과 수직 연합학습으로 나눌 수 있음. 수평 연합학습은 데이터의 변수가 동일한 채 샘플이 다른 경우를 의미하며, 수직 연합학습은 동일한 샘플 데이터에서 서로 다른 변수가 분산된 환경에 존재할 때의 접근 방법을 의미함. 수평 연합학습은 특정 사용자에 연합학습에서 이용되지 못한 수직 분할 데이터를 이용하지 못하며, 수직 연합학습은 많은 양의 동일한 샘플이 겹쳐야만 이용할 수 있음.

- 본 과제에서 개발 예정인 PPFL(Personalized Progressive Federated Learning) 모델은 앞선 두 문제를 해결하기 위함임. (Kim. et. al, In preparation)



그림. PPFL 알고리즘의 개요

- 이 때, 공통 연합학습 네트워크 내부만의 파라메터 (파란색 점선)은 학습되지 않도록 설정하여, 기 구축된 연합학습 모델의 지식을 보존할 수 있도록함. 즉, 연합학습을 학습한 통해 공통 변수의 일반화된 지식과 사용자에 특징적이게 학습 된 파라메터 사이의 균형을 조절할 수 있도록 만듬. CDM 파라미터 후 특이적 파라미터 더 학습

- 본 과제에서는 1차 년도에서 개발한 RNN 기반의 설명가능한 실시간 부작용 예측 모델을 연합학습을 통해 학습시킬 예정임.

◌ MOA-CDM 데이터 외부 활용을 위한 합성 데이터 생성 및 공유

- 구축된 MOA-CDM의 데이터를 바탕으로 분산학습(FL)과 차분 프라이버시(Differential Privacy, DP) 방법론을 연계한 모든 기관들의 데이터의 분포와 변수 간 연관성을 반영한 합성 데이터(Synthetic Data)를 생산할 수 있음

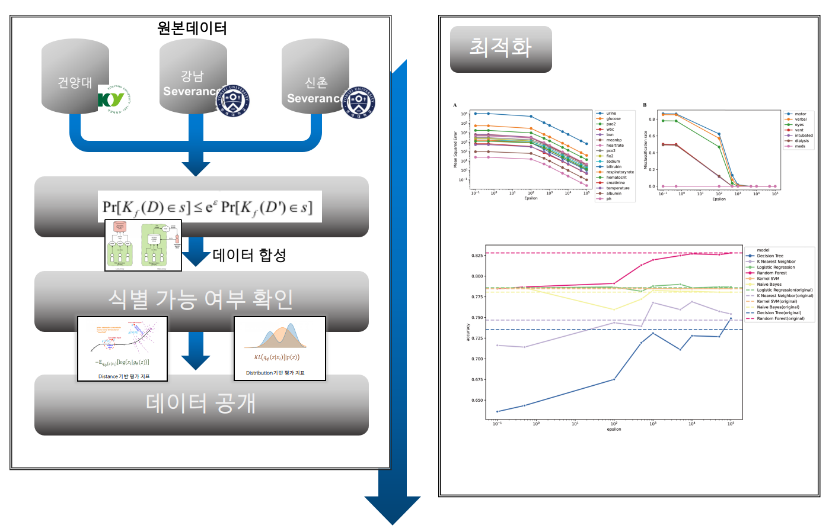
- 참여하는 기관의 간독성, 신독성 환자는 대략 건양대학교(간독성 : 565명, 신독성 : 3,387명. 2012-2019년도 기준), 강남 세브란스(간독성 : 1,588명, 신독성 : 6,479명. 2003-2021년도 기준), 신촌 세브란스(간독성 : 4,955명, 신독성 : 19,605명, 2003-2021년도 기준)임. 모든 기관의 데이터 중 30% 정도에 DP 알고리즘을 적용.

- DP 알고리즘 적용시 연속형 변수와 이산형(명목형) 변수들을 구분하여 Laplace mechanism을 이용한 적절한 epsilon noise를 적용함.

- 적용된 DP에 의해 각 기관에서 생산된 합성 데이터와 원본 데이터와 일치하는 환자가 있는지 cosine similarity를 통해 확인 및 개인 식별 가능 여부를 확인(데이터간 일치하는 데이터 여부 확인).

- 적용된 합성 데이터의 분포가 각 기관의 데이터와 유사한지 Kullback Leibular Divergence 방법과 원본 데이터와 합성 데이터 간의 PCA(차원축소) 결과, 약물 부작용 예측 알고리즘 적용 결과 비교 등을 통해서 utitlity를 비교하여 검증.

- 합성된 데이터의 개인정보보호 확인과 타 알고리즘 적용 시 차이가 없음을 확인 후 기관별 합성된 데이터를 모두 합쳐서 공개할 수 있음

****

◌ 국내외 다기관 기반 의약품 부작용 실마리 탐지 / 인공지능 알고리즘의 다기관 검증

- 의약품 부작용 실마리 탐지 알고리즘의 데이터 편향성에 대한 이슈를 극복과 타당성 평가를 위해, 참여 의료기관에 코드를공유하여 기관별 결과에 대해서 살펴보고자 함. 다기관 분산 분석이 가능한 모아넷 네트워크를 활용하여 검증을 수행할 예정



- 기관별 결과는 메타분석을 적용하여 유의한 시그널을 판정하고, 임상적 판단과 해석을 덧붙여 실질적인 의약품과 부작용의 인과성을 평가

- 궁극적으로, MOA-CDM을 활용한 다기관 의약품 부작용 실마리 탐지 분석 결과를 시각화하여 효율적으로 공유하고, 다기관 검증 시 발견된 오류 사항을 별도로 관리하여 지속적인 알고리즘 유지·보수를 하고자 함

- 1차년도 국내 협력 병원 및 기관들을 통한 유용성이 검증된 후, 국제 공통자료모델 커뮤니티 기반 유용성 검증 진행 계획

- 의약품 부작용 실마리 탐지 알고리즘의 다기관 통합분석을 통해 해당 알고리즘의 유용성과 활용도를 증진시킬 계획

- CDM 기반 국제 커뮤니티 검증은 UCSD(University of California San Diego) 오노 마차도 연구팀에서 본 연구과제에서 제안한 인공지능 예측모델 기반으로 검증 예정

- 신독성·간독성 등 의약품-부작용 쌍(pair) 20개 이상에 대해 개발한 인공지능 알고리즘을 이용한 개발한 인공지능 알고리즘을 검증할 수 있는 검증 패키지 작성 예정

- 인공지능 검증 결과를 체계적으로 수집할 수 있도록 CDM prediction model

library를 개발

|  |
| --- |
|  |
| 그림. 약물부작용 예측 인공지능 모델 검증 결과를 저장하는 prediction model library 데이터베이스 구조 |

- 5개 이상의 기관 (세브란스 병원, 건양대 병원 포함)에서 검증을 시도하여 총 100개 이상의 검증 결과 대외 공개는 검토 후 협의 하에 진행 예정