[별지 제5호 서식]

# 최종보고서 제출양식

겉표지 양식 : (4×6배판(가로19㎝×세로26.5㎝))

| (뒷 면) | (옆면) | (앞 면) |
|---|---|---|
| | 국내AI신약개발데이터공유활성화방안마련을위한연구 과학기술정보통신부 | 2023 - 1<br>국내 AI 신약개발 데이터 공유 활성화 방안 마련을 위한 연구<br>(Research to develop plans for revitalizing domestic AI new drug development data sharing)<br>연구기관 : 한국제약바이오협회<br>연구책임자 : 정원용<br>2023. 5. 31.<br>과 학 기 술 정 보 통 신 부 |

# 안 내 문

본 연구보고서에 기재된 내용들은 연구책임자의 개인적 견해이며 과학기술정보통신부의 공식견해가 아님을 알려드립니다.

과학기술정보통신부 장관 ○ ○ ○

# 제 출 문

과 학 기 술 정 보 통 신 부 장 관 귀하

본 보고서를 “국내 AI 신약개발 데이터 공유 활성화 방안 마련을 위한 연구”의 최종보고서로 제출합니다.

2023. 05. 31.


연구기관명 : 한국제약바이오협회
연구책임자 : 정 원 용
연 구 원 : 여 의 주
연 구 원 : 홍 성 은
연 구 원 : 최 아 연
연 구 원 : 김 관 헌
연 구 원 : 정 소 이
연 구 원 : 이 경 미
외부연구원 : 방 준 일


※ 연구기관 및 연구책임자, 연구원은 실제 연구에 참여한 기관 및 자의 명의임.

## | 요약 |

❑ 연구 개요

❍ (신약개발의 AI 도입) 질병에 대한 마커 발굴, 약물 설계 및 탐색, 약물 효과 예측, 임상 시험환자 모집 등의 신약개발 과정에 중요한 문제들을 효율적으로 해결하기 위해 인공지능을 도입하고 있음

- 38개 인공지능(Artificial Intelligence, AI) 신약개발 기업 등장, 48개 이상 제약바이오기업 AI 활용 중

❍ (데이터 공유 필요성) AI 모델의 활용성은 성능과 직결되며 좋은 성능을 위해서는 다양한 대량의 데이터 필요함. 이러한 데이터를 모으기 위해서는 분산 저장된 데이터 공유가 필요

- 제도·인식·경쟁 등 한계로 기관 또는 기업 간 데이터 공유가 어려워지면서 데이터 편향과 AI 모델 활용성 저하 문제가 발생하였고, 신약개발 분야에서도 같은 문제 발생

❍ 수행기관은 본 과제를 통해 AI 신약개발 생태계 활성화를 위한 국내 신약개발 데이터 공유 활성화 방안 마련 연구를 수행함

❑ 신약개발 데이터

❍ 신약개발은 질병 치료를 위한 새로운 약물 또는 치료제를 식별하는 복잡한 과정으로 약물 발견, 표적 식별, 선도물질 식별 및 최적화, 임상시험이 필요하며, 이 과정에서 화합물, 생물학, 약리학, 임상 등 다분야의 방대한 데이터가 발생

❍ 신약개발 데이터는 공개 여부와 목적에 따라 크게 공공 데이터와 민간 데이터로 구분되며, 이외에도 다양한 속성 차이가 존재함(아래 표 참고)

| 분류 / 속성 | 공공 데이터(Public Data) | | 민간 데이터(Private Data) |
|---|---|---|---|
| 공개 여부 | 공개 | | 비공개(기밀) |
| 목적 | 공공성, 다목적 | | 신규 특허, 특정 목적 |
| 소유 주체 | 정부, 기관 | | 민간 기업 |
| 갱신 주기 | 대규모로 1회(장기간) | | 소규모로 지속적 발생(단기간) |
| 민감 여부 | 개인정보 포함 | | 지식재산권 해당 |
| 데이터 축적 방식 | 정부 연구 성과물로 데이터 수집 | 정부 사업으로 직접 생산 | 신약개발 실험 및 검증과정에서 발생하는 데이터 |
| 데이터 특징 | 다양성의 폭이 넓지만, 저밀도 | | 다양성의 폭은 좁으나 고밀도 |

❍ 본 연구에서는 신약개발 데이터를 공공, 민간 데이터의 두 가지로 분류하고, 각각의 데이터 공유 활성화 방안을 연구했음

< 1. 국내 공공 데이터의 공유 활성화 방안 >

- 공공 데이터 활용 현황을 분석하고 국내 공공 데이터 공유 활성화 방안으

로 △ 수요 기반 공공 데이터 매칭, △ 공공 데이터 기반 경진대회 개최, △ 기존 인공지능 신약개발 플랫폼 사업(KAIDD)과의 연계 방안을 제시

< 2. 협력 사업을 통한 민간 데이터 공유 활성화 방안 >

- 고립된 신약개발 데이터 공유 활성화 체계 구축을 목표로 세계 최초 데이터 기반 제약기업 협력 프로젝트인 유럽의 MELLODDY 사업 분석, 벤치마킹
- 연합학습 기반으로 제약기업, 병원, 공공기관 데이터를 안전 활용하는 민관 협력(PPP, Private Public Partnerships) K-MELLODDY 프로젝트 제안

❑ 신약개발 데이터 공유 현황 분석

❍ (공유 플랫폼 현황) 신약개발 분야의 공공 데이터 구축을 목표로 데이터 생산·수집·구축에 정부 자원 적극 투입 중이나 홍보 부족, 활용 부진, 낮은 품질 신뢰도, 절차의 복잡성 등 해결과제가 많은 것으로 확인됨

❍ (공개 데이터베이스) 국내 데이터베이스(DB) 이용에는 회원가입, 연구자 등록, 데이터 신청, 계획서 제출 등 복잡한 과정이 필요함. 반면 해외 유명 DB는 신청 과정 없이 바로 접근할 수 있어 여러 방면에서 활용되고 있으며 인지도 높은 PDB의 경우 연구자들이 연간 1만 개 이상의 데이터를 기탁하고 있음

❍ (데이터 공유 법적 규제)

- 데이터의 지식재산권 보호에 대한 법적 보호 수단 부재, 민법상 소유권 인정이 어려워 계약법 상 보호만 가능, 데이터 용도 별 법적 보호 역시 데이터 라벨링에 대한 분류체계 기준이 없거나 모호하다면 보호 수단 미존재, 데이터를 보호하는 기술이 필요하며 가명 정보 처리 기술이 유일한 대안
- EU와 영국은 신약개발 관련 데이터의 비식별처리가 가능해 공유·활용에 제약이 없으나, 국내는 비식별 처리할 수 있지 않고 제도적으로 보호받는 가명 처리 가능 유무도 모호

❍ (민간 데이터 공유 사례) EU MELLODDY는 고립된 신약개발 데이터를 AI가 활용할 수 있도록 블록체인, 연합학습, 다중작업 기계학습 기술에 기반한 데이터 기반 협력체계 구축을 목표. 데이터 학습 결과 협력 모델의 성능이 독립 모델보다 우수하며 연구 비밀을 유지하면서 데이터 기반 협력이 가능함을 증명

❍ (연합학습 기술 분석) 연합학습은 데이터 공유를 통한 데이터 유출 위험 없이 모델을 여러 기관과 공유하여 협력 모델을 학습하는 방법으로 데이터 프라이버시 향상, 법적 규제 극복, 데이터 활용 지속성 확보, 기관별 데이터 영향 평가가 가능한 기술로 밝혀짐

❑ 데이터 공유 활성화 방안

❍ 신약개발 공공 데이터 매칭 프로젝트 실행

– 수요기업(제약바이오기업, AI 신약개발기업)들이 AI 신약개발 과정에서 요구하는 공공 데이터를 발굴하여 수요기업과 공급 기관(공공 데이터 보유기관)을 연결하는 AI 신약개발 공공 데이터 매칭 프로젝트를 실행

- AI 신약개발 데이터 맵을 구축하고 신약개발 단계별 필요한 공공 데이터를 선제적으로 수집, 가공, 공급
- 공공 데이터를 AI 모델에 바로 입력할 수 있도록 하는 데이터 전처리 프로토콜 개발(AI 요구 데이터로의 변환)
- 매칭 프로젝트를 실행하여 AI 신약개발에 공공 데이터의 활용도 제고

❍ 공공 데이터 활용 AI 신약개발 경진대회 개최

– 경진대회를 개최하여 공공 데이터의 활용도를 높이고 활용 범위를 확대. 경진대회는 AI 신약개발 분야 창의적 아이디어 발굴, 신규 전문인력을 유인, AI 신약개발에 대한 대국민 관심을 제고하는 부대 효과도 얻을 수 있음

– 경진대회는 AI 신약개발에 사용되는 공공 데이터 확보, 문제 발굴, 예측 모델의 기준(Baseline) 설정, 참가자 모집, 경진대회 웹사이트의 리더보드(Leader-board)와 코드 공유방안 확보 등의 준비 작업이 필요

❍ 기존 AI 신약개발 플랫폼 사업(KAIDD)의 연합학습 활용

– 정부가 구축한 AI 신약개발 플랫폼(KAIDD)의 도구는 AI 모델이며, 이 모델들은 연합학습의 학습 모델로 활용가능함. 연합학습을 통해 다기관, 공공 데이터가 학습된 고성능 AI 신약개발 모델 개발되면 KAIDD 이용자의 만족도는 물론 활용성도 높아질 것으로 예상됨

– 더 나아가 아래에 제안하는 한국형 연합학습 기반 AI 신약개발 플랫폼의 학습 모델로 KAIDD에 탑재된 도구 활용 가능

❑ 한국형 연합학습 기반 AI 신약개발 플랫폼(K-MELLODDY) 과제 기획

❍ 기획 배경

– (경쟁력 열세) 신약 강국 도약이 국가 지상 과제이나 기술 수준과 R&D 투자비의 절대적 열세로 전통적 신약개발 방식으로는 선진국의 추격 요원

– (AI로 열세 극복) 신약개발에 AI 기술을 접목해 R&D 투자 비용을 절감하고 신약개발 기간 단축으로 제약 선진국을 추월하는 퀀텀점프 전략이 필요한 때

– (데이터 공유의 한계) AI 기술은 다량의 다양한 데이터를 활용할수록 성능이 개선되나 각 기관이 보유한 데이터를 연계·활용할 수 있는 체계가 미흡하여 각자 보유하고 있는 데이터를 폐쇄적으로만 활용하거나 특정 기

관 간 1:1 협업에 그침

- 보건의료 데이터는 대부분 개인정보보호, 지식재산권, 연구 비밀 이슈가 있는 민감 데이터이기 때문에 공유-연계-활용이 어려움
- 공공기관, 기업, 병원이 보유한 화합물의 생물 활성 데이터(Bio-activity), 유전체 데이터, 임상데이터, 의료 데이터 등을 다른 기관과 연계하여 활용하지 못하고 있음

❍ 사업 필요성

- (데이터 협업 필요성) 경쟁력 있는 AI 신약개발을 하기 위해서는 정부를 중심으로 기업, 병원, 대학 등 이해관계자들이 보유 데이터의 기밀은 유지하면서 타 기관이 보유한 데이터를 효과적으로 활용할 수 있는 AI 신약개발 협업 프레임워크 필요
- (정부 주도 PPP 필요성) 다양한 기관 간의 협력은 일부 기업이 주도하기 어려우며, 국가 산업 경쟁력 제고를 위한 민관협력(PPP, Private Public Partnerships)이 필요
- (혁신기술 개발 및 활용 필요성) 연합학습(Federated learning) 기술은 ❶ 개인정보보호 및 지식재산권 이슈를 극복하면서 ❷ 다기관 데이터를 연계하는 협력 기술로 ❸ AI 신약개발의 저비용 고효율 효과를 보건의료, 금융 등 다양한 분야에 활용할 수 있는 시장 선도형 기술임
  - 데이터 공개 및 공유 이슈를 해결하는 법적 제도적 장치가 완비되기를 기다리지 않고 AI 신약개발에 필요한 다기관 데이터를 즉시 활용할 수 있는 현실적인 협력 기술
- (저비용 고효율 실증 필요성) 연합학습 기반 ADME/Tox 예측 모델을 기반으로 플랫폼을 구축할 경우, 제약바이오산업의 연간 신약개발 R&D 투자 비용의 20%인 4,600억 원을 절감할 수 있음

❍ 유사사업 분석

- 신약개발에 연합학습을 사용한 외국 사례는 유럽연합에서 추진한 멜로디(MELLODDY) 컨소시엄으로 글로벌 빅파마(GSK, Amagen, Merck, Novatis 등 10개 기관)를 포함한 총 17개 기관이 참여해 기관 간 새로운 협력 모델을 보여줌
- 10억 개 이상의 약물 발견, 1,000만 개 이상 화합물의 약리 활성 실험데이터를 활용하기 위한 연합학습 플랫폼을 구축했고, 활용 결과 연합 학습한 모델이 독립 기관 모델보다 평균 2~4% 성능이 높았음 (2019.06~2022.05, 약 256억 원)

❍ K-MELLODDY(Machine Learning Orchestration for Drug DiscoverY) 사업안

– (목표) 분산된 민간 데이터의 활용 및 데이터 기반 협력이 가능한 한국형 연합학습 기반 AI 신약개발 플랫폼을 구축하고 응용 사례를 제시

- (연구 주제) 다기관 데이터 기반 ADME/Tox Multi-Task 예측 모델 연합학습

– ADME/Tox 분야에도 상세하게 나누면 매우 다양한 분야가 존재 ①물성 예측, ②투과도 예측, ③분포용적 예측, ④수송체 약물 분포 영향력 예측, ⑤버퍼 안정성, ⑥대사 안정성, ⑦심장 독성, ⑧간 독성, ⑨발암성, ⑩생식, 내분비 독성 등

– (연합학습 장점) ADME/Tox는 후보물질의 종류와 관계없이 제약기업이 공통으로 수행하는 신약개발 단계로 여러 기관의 데이터를 연합학습한 ADME/Tox 예측 모델이 개발돼 활용된다면 각 기업이 감수하고 있는 동일한 시행착오와 반복적 실험검증 횟수를 크게 줄임으로써 비용 절감과 기간 단축 효과를 얻을 수 있음

– (실현 가능성) 표적에 상관없이 신약개발 초기에 공통 확인이 필요한 실험으로 국내 기업에도 다수의 실험데이터가 누적(제약사 자체 조사 결과 반영)

❍ 사업 추진 전략

| 비 전 | 인공지능 기반 신약개발 생태계 활성화 |
|---|---|
| 최종목표 | 연합학습 기반 신약개발 가속화 시스템 구축과 성공사례 창출 |
| 사업1 | 연합학습 기반 신약개발 가속화 프로젝트 사업 |

| | | |
|---|---|---|
| 세부목표 | (세부1) 플랫폼 구축·운영 및 사업 관리 | 플랫폼 사업 관리<br>• 과제지원 및 성과 관리 체계 구축 운영(거버넌스(추진위원회)구성 및 운영)<br>• 플랫폼 ISP 수립, SOP 제공<br>• 플랫폼 교육 기획 및 운영<br>• 데이터 품질 및 생산 절차 관리<br>• 플랫폼 활용 가이드라인 개발<br>• 플랫폼 고도화 방안연구<br>• 플랫폼 확산 지원(경진대회, 홍보)<br>• 신약개발 데이터 공동활용성 확대 연구<br>• 플랫폼 활용확산방안 연구 및 성과분석 |
| | | 플랫폼 구축<br>• 플랫폼 설계(플랫폼 구조 및 기능, 인프라 구조, 연합학습)<br>• 플랫폼 구축(인프라 구축, 플랫폼 개발, 연합학습 개발) |

| | | |
|---|---|---|
| | | • 클라우드 활용비, 지속 운영 계획 수립<br>• 연합학습 실행 대응, 결과 보고<br>• 플랫폼 유지보수 및 고도화 |
| | **(세부2) 연합학습 원천 기술개발** | • 연합학습 원천 기술 R&D(보안 강화 기술, 취합알고리즘, 학습 프로토콜, 기여도 평가 지표, 공정성) 5개 주제 |
| | **(세부3) 플랫폼 활용과제** | 데이터 공급 및 연합학습 참여, 실험 검증<br>• 데이터 공급 및 생산<br>• 연합학습 참여 및 예측모델 검증<br>• 실험적 검증, 데이터 큐레이션 및 디지털 전환 |
| | | 표준 데이터 처리 도구 및 모델 개발<br>• 데이터 전처리 도구 개발<br>• 예측 모델 개발<br>• 모델 성능 평가 및 지속 개선 |

❍ 사업 규모(사업 기간 5개년, 총 466억)

표 1. 연차별 예산(안)

(단위 : 억원)

| 구분 | 과제수 | '24 | '25 | '26 | '27 | '28 | 합계 |
|---|---|---|---|---|---|---|---|
| □ 연합학습 기반 신약개발 가속화 프로젝트 | 29 | | | | | | |
| ○ (내역1) 연합학습 기반 신약개발 가속화 프로젝트 사업단 | 29 | 54 | 103 | 103 | 103 | 103 | 466 |
| **– (세부1) 플랫폼 구축·운영 및 사업 관리** | 1 | 13.5 | 22 | 22 | 22 | 22 | 101.5 |
| **– (세부2) 연합학습 원천기술 개발 과제** | 5 | 7.5 | 15 | 15 | 15 | 15 | 67.5 |
| **– (세부3) 플랫폼 활용 과제** | 23 | 33 | 66 | 66 | 66 | 66 | 297 |
| · 데이터 공급 및 연합학습 참여, 실험 검증 | 20 | 30 | 60 | 60 | 60 | 60 | 270 |
| · 전처리도구개발 및 모델 개발 | 3 | 3 | 6 | 6 | 6 | 6 | 27 |

❍ 참여기관 별 역할

– (제약기업, 연구소) 데이터 생산 및 공급을 통한 디지털 전환, 연합학습된 모델의 예측값 실험검증

– (AI 신약개발 기업) 데이터 전처리 도구 개발, AI 모델 개발 및 검증

– (대학) 연합학습의 원천기술 개발

– (IT 기업) 한국형 연합학습 플랫폼의 요구사항 정의와 구축

– (공공기관) 공공 데이터의 공급자 역할로 공공 데이터를 수집, 가공, 서비스

• 예) 한국화학연구원(한국화합물은행), 한국생명공학연구원(국가생명연구자원정보센터), 한국보건의료정보원 등

❍ 참여 혜택

– 공공기관·제약기업·연구소: **공공기관**은 보유한 공공 데이터의 활용성 제고로

부가가치 확보, **제약기업**은 ADME/Tox 예측 모델 확보로 중복 실험과 시행착오 비용 절감, 경쟁 기관 데이터의 간접 활용

- 대학·AI 신약개발 기업: **대학**은 선도 기술 연구 수행으로 기술력 및 지식재산권 확보, **AI 신약개발 기업**은 현장 실험 데이터를 기반으로 모델 개발 및 검증
- **IT 기업**: 선도 기술 개발 이력으로 사업 분야 확장 및 기업 경쟁력 제고
- **산업계 대표단체**: 제약·바이오 산업계 기여 (AI 신약개발 변화 주도, 인력 양성, 협력 네트워크 구축), AI 신약개발 전문가 네트워크 확충

❍ 기존 사업과의 차별성

- 유럽연합의 MELLODDY와 목적, 추진 주체, 데이터 현황, 인프라의 차이와 차별성이 존재함

표 2. 벤치마크 사업과의 차별성

| 비교 항목 | EU MELLODDY | K-MELLODDY |
|---|---|---|
| 목적 | 연합학습 기술에 대한 실증사업 (사업의 활용방안 부분이 부족) | 연합학습 기술에 대한 활용사업 (본사업을 플랫폼화하여 데이터 활용 생태계를 조성하고자 함) |
| 추진 주체 | Owkin(연합학습 솔루션)과 Kubermatic(클라우드 플랫폼)과 같은 인프라 ICT 기업을 중심으로 추진 | 다기관(학교, 공공기관, 연구소, 제약·바이오, AI, IT 기업 등)의 협력이 필요하고, 자국의 신약개발 경쟁력을 확보하기 위하여 산·학·정 협력 필요 |
| 데이터 현황 | 참여 제약사는 글로벌 빅파마로 오랜 기간에 걸친 신약개발 연구 경험과 다양한 신약개발 파이프라인을 보유해 공유 가능 데이터가 많이 존재 | 글로벌 제약사보다 규모가 영세해 공유가능 데이터가 상대적으로 적지만, 약물 발견과정의 실험 결과인 ADME/Tox는 공유 활용이 가능한 수준이며, 부족 데이터는 제약사를 지원하여 생산 |
| ICT·인프라 환경 | NVIDIA, Owkin, Kubermatic의 연합학습, 인프라 전문기업이 참여해 원천기술을 확보한 상태로 사업수행 | 자체 조사에 따르면 국내의 경우, 대학, IT 및 클라우드 기업이 협력하여 연합학습 원천기술 연구개발이 필요 |
| 데이터 공개 범위 | 데이터 공개 없음, 기탁도 없음 (단, 기본 모델 학습에 활용한 전처리된 공개 화합물 데이터는 공개) *동 사업에서 해당 전처리 데이터를 사전 훈련된 모델 개발에 활용 가능 | 기업이 기보유한 데이터는 공개하지 않음 (단, 특허로 공개된 화합물인 경우, 과제비로 생산한 데이터의 경우 공개) 국내 연구자들에게만 공개 |
| 모델 공개 범위 | 모델의 구조와 사용법만 공개 학습된 모델 결과(가중치)는 비공개 (단, 참여기업은 연합학습된 모델 공유) | 모델 연합학습에 기여하는 참여자들에게 공개하는 것을 원칙으로 함(추후 참여자도 협의를 통해 승인받으면 가능) 모델의 구조와 사용법은 공개 |
| 플랫폼 공개 | 개발된 플랫폼 미공개(참여자들이 활용) (Owkin, Substra 연합학습 솔루션만 공개) | 개방형 플랫폼으로 공개 (연합학습 팀 구성, 데이터, 전처리기, 모델, 연합학습 추적 등 기능 탑재 예상) |

❍ 추진 방안

- (K-MELLODDY 추진위원회) 복지부와 과기부를 중심으로 산·학·연 전문가 10인 내외로 추진위원회를 구성하여 사업 전반의 추진계획을 수립·시행
- (데이터 파트너십) 과기부와 복지부가 주도하여 K-MELLODDY 프로젝트에 참여할 데이터 보유 공공기관, 제약기업, 바이오벤처 간의 파트너십 체결

❍ 활용 방안

- (AI 신약개발 데이터 활용 중개 플랫폼) K-MELLODDY로 구축한 연합학습 플랫폼은 사업 종료 후에도 산업계에서 다양한 용도로 활용하며, 기업 간 서로의 미충족 영역을 충족시켜 줄 수 있는 창구로서 역할이 기대됨
- (데이터 중심 협업 도구) AI와 제약·바이오 기업의 협업 진행 시 제약·바이오 기업의 데이터 활용이 프로젝트 성공률을 높일 수 있으나 데이터 제공에 부담. 반면, K-MELLODDY 플랫폼은 안전하게 데이터 협력을 진행할 수 있음
- (사업 확장) 약물 탐색 이외에도 표적 발굴이나 바이오마커 발굴 등 신약개발의 다른 분야로도 사업을 확장하여 활용할 수 있음
- (AI 기반 예측 모델의 유/무료 서비스) 동 사업에서 개발한 연합학습 기반 AI 신약개발 시스템을 참여한 기업들의 신약개발에 활용
- 사업 성과로 도출되는 예측 AI 응용 프로그램을 기본 모델판 버전(무료), 유료화 버전, 기업 내부용 등 다양한 형태로 개발하여 국내 기업, 학계에도 공개하여 사용할 수 있도록 함
- 사례) 일본 “창약 지원 인포매틱스 시스템 구축사업”에서는 ADME, 약동학 예측 모델의 예측 서비스를 유료화

❍ 기대효과

- (R&D 효율화) 연합학습 기반 ADME/Tox 예측모델 개발로 4,000억 원의 직접적 R&D 투자비 절감(전체 투자비의 20%), 인산화효소 활성 저해 예측 모델 등으로 사업 확대 시 국가 및 민간의 신약개발 R&D 비용을 1조 원 이상 절감
  - 2021년 기준 국내 혁신형 제약기업의 신약개발 R&D 비용은 2조 1,193억 원
  - 연구에 따르면 In silico ADME/Tox 분석을 통해 후보 물질의 안정성, 용해도, 독성문제를 각각 2배, 7배, 1.4배 개선하여 신약개발의 성공확률을 높임(’17, Journal of Medicinal Chemistry)
- (공동연구 촉진) 제약바이오산업의 디지털 전환과 AI 기술 도입을 촉진하여 저비용 고효율 AI 신약개발로의 패러다임 전환 가속화
- (AI 기술 혁신) AI 신약개발에서 데이터 활용 문제를 해소하고 신약개발에서 요구되는 AI 기술 수요를 정확히 파악하여 신약개발 분야 AI 기술 혁신 촉진
- (제약산업 글로벌 경쟁력 확보) 연합학습 기반 플랫폼 구축을 통한 국내 제약산업 협력체계를 조성하여 분산된 제약산업의 투자와 방향을 응집함으로써 투자의 결집화, 신약개발 변화혁신 창출로 글로벌 제약기업과의 경쟁

력 강화

- (기술 패권주의에 대응) 글로벌 빅테크 기업의 AI·IT 기술의 승자독식 구조의 심화, 한국의 기술 및 데이터 주권을 확보하기 위해서는 데이터를 안전하게 AI에 활용하는 연합학습 기술의 실용화와 실용화 주도로 글로벌 표준국 도달이 우선 목표

# 목 차

# 제1장 연구개발 목표

## 1.1. 연구개발 과제의 목표

### 1.1.1. 연구 배경

❑ 신약개발 분야의 인공지능(AI: Artificial Intelligence) 도입

- 2019년 12월 COVID-19 발생 이후 현재까지 심각한 사회적 변화를 겪고 있음. 전염병 발생 주기를 추적했을 때, 향후 인류를 위협할 새로운 질병이 점차 자주 등장할 것으로 예상되며 **신약개발의 중요성이 더욱 커짐**
- 평균 10년 이상, 10억 달러 이상의 비용이 소요되는 신약개발은 성공 위험 부담이 크고 임상시험 단계 약물의 12%만이 FDA 승인을 받을 수 있어 **높은 실패율에 대한 부담을 줄이기 위한 다양한 노력이 필요**함[1)]
- 지난 10여 년 동안 인공지능이 이미지 인식, 자연어처리 등에서 큰 성과를 보이며 타 과학 기술개발에 인공지능이 접목되기 시작했고 그 성과가 검증되면서 이론 중심에서 **데이터 중심으로 연구 패러다임이 전환**
- 신약개발의 인공지능 기술 도입은 최근 3~5년 사이에 글로벌 빅파마부터 시작되었고, **최근에는 많은 벤처기업이 투자를 유치하면서 신약개발 분야 인공지능 연구개발이 본격화**되고 있음

❑ 신약개발 분야 인공지능 도입 배경

- **방대한 데이터 축적**: 2,000년대 이후 다양한 프로젝트를 통해 사람과 약물 정보의 디지털화가 급격히 증가함
- **빅데이터 활용 기술 등장**: 2010년 이후 기계학습 기반 인공지능 기술의 성능이 급속도로 향상됨(퍼셉트론, 다층퍼셉트론, 인공신경망, 합성곱 신경망(CNN), 순환신경망(RNN), Attention, Transformer, 그래프 신경망 등 딥러닝 기술의 발전)
- **신약개발 비효율성 개선 수요**: 기존의 신약개발은 막대한 소요 시간과 비용이 들며 성공률이 매우 낮은 고위험 분야였으나 감염병의 대유행으로 신약개발 기간을 급속히 단축해야만 하는 사회적 요구 발생
- **신약개발 분야 인공지능 기술의 발전** : 사람이 하던 일의 상당 부분을 인공지능으로 대체하면서 문헌 검토, 약물 스크리닝, 약물 디자인 등에서 성과를 나타냄 (Benevolent AI 사: 자연어처리 기반 AI 문헌 스크리닝, Insilico Medicine 사: AI 기반 약물 디자인 등의 기술로 후보물질 발굴 시간 단축)

---

1) Research and Development in the Pharmaceutical Industry, Congressional Budget Office, 2021.04

❑ 국내 인공지능 신약개발 동향

❍ 정부에서는 국가 통합 바이오 빅데이터 구축사업과 AI 활용 혁신 신약 발굴 사업 등 27개 사업을 통해 제약바이오산업의 AI 활용을 21년 출범한 국가 신약개발 사업은 2030년까지 국비 1조 4,767억을 지원할 예정[2)]

❍ 국내 제약바이오 기업 중 신약개발에 인공지능을 활용하고 있는 기업은 총 48개이고, AI 신약개발 기업은 38개이며, 2021년 상반기 14개 기업은 총 1,700억 원의 투자를 받았음[3)]

❍ 주요 제약바이오기업들은 글로벌 추세에 맞춰 AI 신약개발 기업과 연구 협력 계약을 체결해 AI 플랫폼과 기술을 활용하여 신약개발 연구 중

표 3. 국내 제약바이오 기업의 활용 중인 인공지능 기술

| 국내 제약사 | 협업 기관 | AI 기술, 플랫폼 |
|---|---|---|
| 한미약품 | 스탠다임 | AI 기반 후보 물질 최적화 플랫폼 |
| 대웅제약 | 에이조스바이오 | AI 기반 신약개발 플랫폼, 'iSTAs' 구축 |
| | 온코크로스 | 유전자 발현 패턴 기반 플랫폼, 'RAPTOR AI' |
| SK 바이오 | twoXAR | AI 기반 약물 설계 플랫폼 |
| | 스탠다임 | 아이클루앤 애스크(iCLUE& ASK) |
| 유한양행 | 사이클리카 | AI 기반의 통합 후보 물질 발굴 플랫폼, 'Ligand Design, Ligand Express' |
| 중외제약 | 신테카바이오 | 개인 유전체 맵 플랫폼, 'PMAP' |
| | | 윈트(Wnt)신호 전달 특화 빅데이터 플랫폼 주얼리 |
| | 디어젠 | AI 신약개발 플랫폼 'DEARGEN iDears' |
| 현대약품 | 파미노젠 | AI 기반 양자화학 소프트웨어와 빅데이터로 구축된 플랫폼, 'LucyNet' |
| SK 케미컬 | 스탠다임 | 신약 재창출 플랫폼, '스탠다임 인사이트(Standigm Insight)' |
| 삼진제약 | 심플렉스 | 신약후보물질 발굴 플랫폼 'CEEK-CURE' |
| 동아ST | 심플렉스 | 신약후보물질 발굴 플랫폼 'CEEK-CURE' |

❍ 스탠다임과 디어젠이 AI 신약개발 선두 기업 Top 33에 선정되었음 (2021 Deep Pharma Intelligence 2020년 AI 신약 바이오마커 개발 및 R&D 환경 시장 보고서[4)]

❍ COVID-19의 대유행으로 백신 및 치료제 수요가 급증하고 신속 개발에 필수적인 비용 및 시간 효율화가 요구되면서 인공지능 신약개발 시장이 급성장하고 있음

2) 이승덕, 국가신약개발사업, 111개 과제로 시작…'투자 적격성 검증' 차별화, 의학신문, 2022.03.16
3) 김양균, "제약사는 인공지능을 모르고, AI기업은 신약개발을 모른다", ZDNet Korea, 2022.03.30
4) 남대열, 걸음마 뗀 국내 AI신약개발 '성장의 싹' 틔워, HITNEWS, 2022.01.05

### 1.1.2. 연구의 필요성

❑ 국내 신약개발 데이터 공유 현황

- ❍ 정부 주도로 신약개발 관련 다수의 공공 데이터 구축사업이 진행되고 있지만, 데이터 활용이 부진하며 활용 성과를 내기 위해서는 **데이터 및 공유 활성화 전략이 필요**함
- ❍ **제약바이오기업**들은 신약개발을 위해 AI 신약개발 기업과 협력하는 오픈이노베이션에 적극적이지만 **데이터의 연계 활용체계가 미흡하여** 보유 데이터를 폐쇄적으로만 활용하거나 협력 기관과의 1:1 협업에 그침
- ❍ AI 신약개발 기업은 제약바이오기업 데이터와 공공 데이터의 활용이 어려워 해외 공개·유료 데이터로 AI 모델을 개발하는 상황
- ❍ 신약개발 분야는 AI 기술 도입에 적극적이나 여러 가지 문제로 **신약개발 데이터를 열어주지** 못하면서 AI 기업은 외부 데이터로 모델 개발만 가능
- ❍ 결국, 수요처(제약바이오기업)의 데이터와 맞지 않는 AI 모델 개발로 수요처에서 모델 활용 시 낮은 성능 체감으로 AI 기술 신뢰도가 저하되는 상황

❑ 신약개발 데이터 공유 저해 요인

- ❍ 우수한 성능의 AI를 위해서는 빅데이터가 핵심 요소이나 이를 구성하는 데이터 대부분은 독립된 사일로(Silo)[5]나 별도의 장소에 수집 저장됨
- ❍ 정부에서는 신약개발 관련 공공 빅데이터를 구축하고자 데이터 표준을 정하고 표준에 맞춰 데이터를 쌓고 있으나, 역동적으로 변화하는 AI 기술과 변화에 유연하지 않은 데이터 표준의 불일치가 존재함
- ❍ 민간 데이터의 융합을 통한 빅데이터 구축을 위해 제약바이오기업 보유 데이터의 공유 활용을 위한 협력이 필요하나 지식재산권, 데이터 유출, 데이터 공유 인식 부족 등의 이슈가 존재함

❑ 신약개발 분야 데이터 공유 활성화 필요

- ❍ 국내 제약산업은 추격 요원으로 글로벌 빅파마와 경쟁하기 위해 AI를 도입한 가속화 전략이 필요하나, 공공 데이터 활용은 어렵고 민간 데이터는 공유가 제한되어 효율성 개선이 가능한 수준의 AI 모델 개발에 난항
- ❍ 따라서, 국내 제약산업 경쟁력 확보 전략인 AI 기술 도입을 위해서는 다양한 대량의 데이터 확보가 필요하며, 공공 데이터의 공유 활성화 방안 필요
- ❍ 제약바이오기업 현장의 민간 데이터를 AI에 활용하는 것은 더 시급하고 중요한 문제로 민간 데이터의 공유·협력 체계가 필요

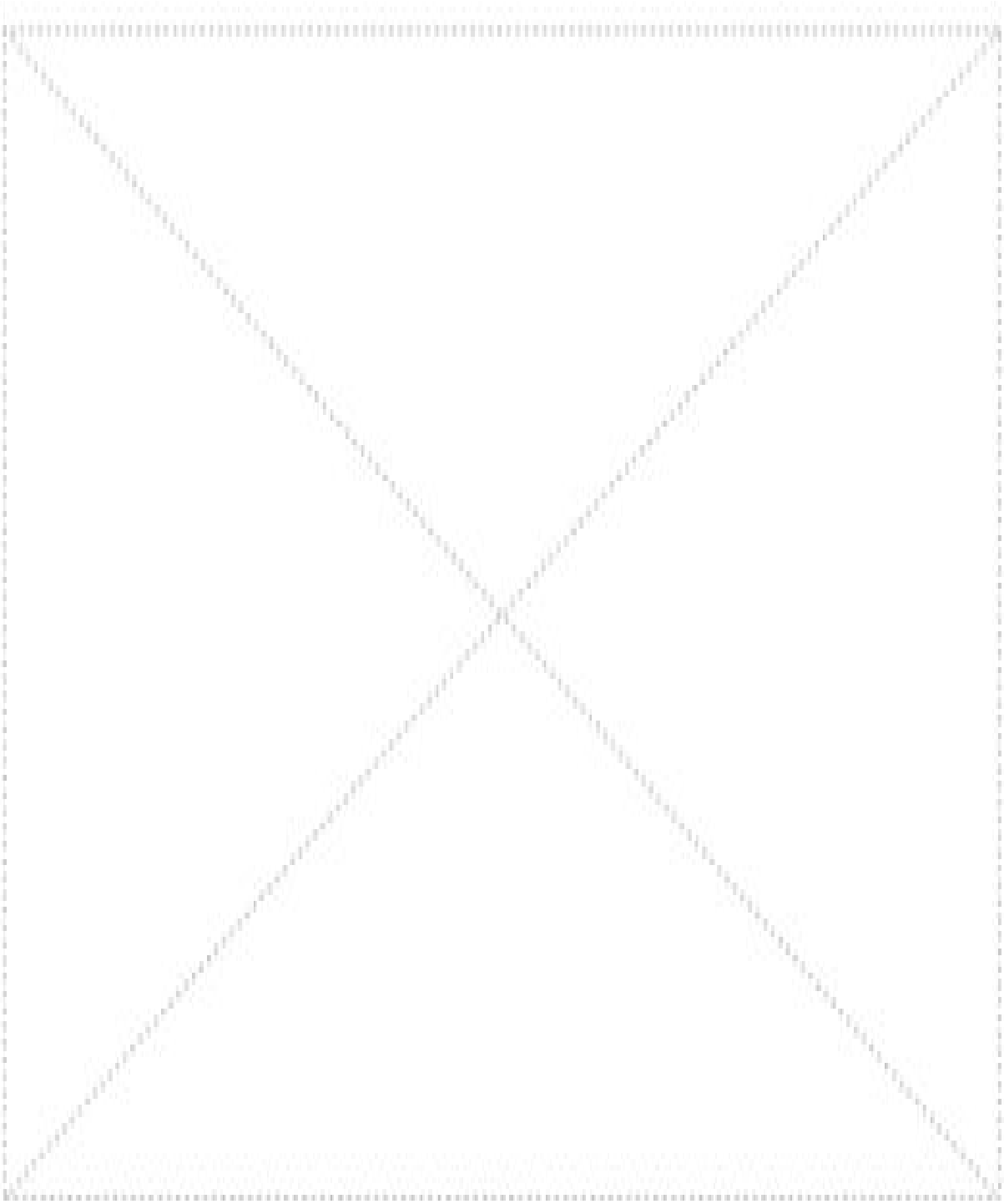
5) 데이터 사일로: 정보가 단일 부서 또는 비즈니스 특정 단위에만 공유되고 조직 전체적으로 쉽게 접근하기 어려운 상태로, 마치 곡물이 사일로(저장탑) 안에 갇히는 것을 비유하는 말 (위키피디아)

### 1.1.3. 연구 목표와 범위

| 구분 | 내용 |
|---|---|
| 최종 목표 | **인공지능 활용 신약개발 생태계 강화를 위한 데이터 공유 활성화 방안을 수립하고 상세 추진전략 및 추진내용 도출** |
| 세부 목표 | • 세부 연구 목표 1: 국내 신약개발 데이터 공유 활성화 방안 수립지원<br>– 국내외 데이터 공유 사례 및 현황 분석<br>– 현황 분석을 통한 시사점 도출<br>– 전문가위원회 구성을 통한 데이터 공유 활성화 방안 마련<br>– 수행 중인 사업과의 적용 활용 방안 제안<br>• 세부 연구 목표 2: 한국형 연합학습 기반 AI 신약개발 플랫폼 (K-MELLODDY) 기획<br>– 국내외 연합학습 사례 및 현황 분석<br>– EU MELLODDY 벤치마킹<br>– K-MELLODDY 구축 전략 수립 |
| 연구 범위 | • 국내외 데이터 공유 사례 및 현황 분석<br>– 국내외 정보보호 법률 분석 및 개선 내용 검토<br>– 국내외 데이터 공유, 구축 과제 및 사업 분석<br>– 신약개발 관련 공개 데이터베이스 조사<br>• 현황 분석을 통한 시사점 도출<br>– 데이터 공유를 위한 저장소와 기술적 방법에 따른 장단점 분석<br>– 국내 인공지능 신약개발을 위한 데이터 공유 현황 분석을 통한 시사점 도출<br>• 전문가위원회 구성을 통한 데이터 공유 활성화 방안 마련<br>– 데이터 공유 실용성에 대한 조사<br>– 데이터 공유 기술에 대한 수요조사 및 평가<br>• 수행 중인 사업과의 적용 활용 방안 제안<br>– 인공지능 신약개발 플랫폼(KAIDD)에 적용·활용 방안 제안<br>• 국내외 연합학습 사례 및 현황 분석<br>– 국내외 연합학습 시장, 사례, 솔루션, 연구 조사 분석<br>– 국내외 보건의료 신약개발 분야의 연합학습 활용 프로젝트 분석<br>• K-MELLODDY 개념 정립<br>– K-MELLODDY 개념과 정의, 기술, 활용 방안 수립<br>– 연합학습을 기반으로 수행 중인 해외 협력 프로젝트 분석 (MELLODDY)<br>• K-MELLODDY 구축 전략 수립<br>– 개인정보 문제와 데이터 공유 문제를 해결하는 방안 제시<br>– 데이터 제공 기관의 데이터 퀄리티 검증 방안 제시<br>– 국내 제약 바이오 현황에 기반한 K-MELLODDY 구축 및 활용성 향상 계획<br>– K-MELLODDY 시범 사업기획 및 FDDP 추진위원회 구성 방안 제시<br>– K-MELLODDY구축 소요 예산 및 근거 마련, 차별성 제시 |

### 1.1.4. 추진 전략

❑ 연구 목표 달성을 위해 현황조사 및 분석, 자문단 운영을 통한 상세 추진 전략 및 추진내용 도출 및 프로젝트 기획



## 1.2. 연구과제의 목표달성도

| 구 분 / 연구내용 | 연 구 기 간 | | | | | | | | | | 진도(율) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 1 | 2 | 3 | |
| 국내외 데이터 공유 사례 현황 분석 | ● | | | | | | | | | | 100% |
| 법률에서 정의한 정보보호 기술 분석 | ● | ● | | | | | | | | | 100% |
| 전문가 위원회 구성으로 데이터 공유 방안 마련 | | ● | ● | ● | ● | ● | ● | ● | | | 100% |
| 수행 중 사업에 적용 활용 방안 마련 | | | ● | ● | ● | | | | | | 100% |
| 국내외 연합학습 사례 및 현황 분석 | | | | | | ● | ● | ● | | | 100% |
| 공공 민간 파트너십 조사 분석 | | | | | | ● | ● | ● | | | 100% |
| 연합형 정보보호 프로젝트 기획 | | | | | | | ● | ● | ● | ● | 100% |
| 총 진 도 율 | | | | | | | | | | | 100% |

# 제2장 신약개발 데이터 공유 현황 분석

## 2.1. 신약개발 데이터 정의

❑ 신약개발 데이터

❍ 신약개발 데이터란 생물학, 화학, 약리학, 임상 영역에서 수집된 신약개발 전 영역에서 활용 가능한 데이터를 의미하며, 각 영역에서 신약개발에 사용되는 데이터를 아래에 나열하였음

❍ 생물학 데이터(Biology data)

– 생물학 데이터는 생물체로부터 유래한 DNA, RNA, 단백질, 세포, 조직 등에 대한 정보를 수집한 데이터로 신약개발의 질병 및 환자 코호트 정의, 질병 치료 표적 단백질의 예측 및 검증, 단백질 구조 및 상호작용 예측, 단백질 치료제 디자인, 치료 효과 식별을 위한 바이오마커 발굴, 약물 신호전달경로 및 작용기전 예측, 약물 상호작용 예측에 활용(데이터 유형)

- 유전체 데이터 : 유전체, 전사체, 단백체, 대사체, SNP 등
- 단백질 데이터 : 단백질 구조, 서열, 도메인, 단백질 간 상호작용 등
- 생물정보 데이터 : 유전자 기능, 신호전달경로, 생체 네트워크 등

❍ 화합물 데이터(Chemistry Compound data)

– 화합물 데이터는 약물이 될 수 있는 화합물의 구조, 물리적 특성, 정보를 포함한 데이터로 신약개발의 신규물질 라이브러리 생성, 화합물-표적 결합력 예측, 유사 구조 약물 검색, 물질 신규성 및 특허성 검증, 약물 합성 경로 분석, 물질 최적화에 활용(데이터 유형)

- 화합물 구조 데이터 : 화합물 3차원 구조, Fragment, 토폴로지(Topology) 표현자(Descriptor), 화합물-단백질 결합구조 및 결합력
- 화합물 물성 데이터 : 용해도, 친유성, 분자량, 수소결합 Acceptor / Donor, Pharmacopore, 양자역학적 에너지 등
- 화합물 정보 데이터 : 화합물 식별자(SMILES, InChI), 중간체, 파생체, 합성경로, 특허 정보, 구매 정보 등

❍ 약리학 데이터(Pharmacology data)

– 약리학 데이터는 다양한 수준의 생물학적 복잡성에서 화합물의 효과를 조사할 수 있는 세포 또는 조직 기반 모델의 단백질 표적에 대한 분석 데이터와 동물 모델에서 테스트 된 화합물 또는 약물에 대한 정보를 제공

– 신약개발의 독성 예측, 약물 동태 예측, 생체 활성 및 저해능 예측, 동물실험 약동·약력학 모델링에 활용(데이터 유형)

표 4. 신약개발 약리학 데이터 유형과 예시

| 모델 | 유형 | 세부 유형 | 예시 |
|---|---|---|---|
| 세포 | 약동학 | 흡수(Absortion) | 단층 세포막(Caco-2 Cell) 투과도<br>인공 지질막(PAMPA) 투과도<br>소장 상피세포(MDCK) 투과도<br>뇌세포(MBEC) 투과도<br>분배 계수 측정 |
| | | 분배(Distribution) | 단백질 결합<br>혈액 및 혈장 단백 결합<br>뇌척수액 및 조직 결합<br>뇌 장벽 투과성(BBB)<br>혈장-혈중 비율(BBP) |
| | | 대사(Metabolism) | 혈액 중 안정성<br>인공위액 및 인공장액 중 약물 안정성<br>간 대사 안정성<br>UGT 효소 저해성<br>글루타티온 접합성 |
| | | 배설(Excretion) | 화합물 및 대사산물 배설량 |
| | 약력학 | 약효 | 표적 단백질(효소, 수용체, 이온채널) 결합 친화도<br>표적 단백질 활성 및 저해성<br>표적 세포 도메인 결합 친화도<br>표적 세포 활성 및 저해성 |
| | | 선택성 | 인산화효소 단백질별 결합 친화도<br>인산화효소 단백질 활성 및 저해성<br>GPCR 단백질 결합 친화도<br>GPCR 단백질 활성 및 저해성<br>이온채널 결합 친화도 |
| | | 독성 및 부작용 | 일반 독성, 간 독성(CYP450 등), 세포독성<br>심장 독성 (hERG, K+, Na+, Ca2+ channel 등) |
| 동물 | 약동학 | | 흡수 속도, 약물 분포용적, 생체이용률, 청소율, 반감기<br>PK/PD 모델링 |
| | 약력학 | | PMPK, 생체 지표 농도<br>독성(생식, 발생, 유전, 항원성, 면역, 발암성 등) |

❍ 임상 데이터(Clinical data)

- 임상 데이터는 환자를 진료하는 과정에서 생성되는 데이터로 환자에게 약물을 사용했을 때 일어나는 반응을 포함한 자료이며, 환자 유전체 데이터와의 결합을 통해 환자 코호트 정의, 질병 치료 표적 단백질의 예측 및

검증, 임상 바이오마커 발굴, 임상시험 디자인 최적화, 임상시험 피험자 선별, 임상시험 데이터 분석에 활용(데이터 유형)

- 환자 인적 데이터 : 키, 시력, 청력, 몸무게, BMI, 나이, 성별, 인종 등
- 검사 데이터 : 생화학 검사, 혈액 검사, 면역 검사, 갑상선 검사, 소변 검사, 초음파 검사, 영상 검사(MRI, CT, X-ray) 등
- 진료 데이터 : 처방, 진단, 생존, 치료 정보 등
- 설문조사 데이터 : 가족력, 생활 습관, 식품 섭취 빈도, 수면 주기 등

※ 수행기관은 환자의 유전체 검사(NGS 패널 등) 데이터의 경우 생물학(유전체) 데이터로 분류

❑ 신약개발 데이터의 구분

❍ 신약개발은 질병 치료를 위한 새로운 약물 또는 치료제를 발굴하는 복잡한 과정으로 질병 치료 표적 발굴, 약물 설계 및 탐색, 약물 효과 예측, 전임상시험, 임상시험 등의 과정이 필요하며 이 과정에서 생물학, 화합물, 약리학, 임상 등 다분야의 방대한 데이터가 발생

❍ 신약개발 데이터는 공개 여부와 목적에 따라 크게 공공 데이터와 민간 데이터로 구분할 수 있으며 다양한 속성 차이를 정리하면 아래의 표와 같음

표 5. 신약개발 데이터 공개 여부에 따른 분류 및 속성

| 구분 / 속성 | 공공 데이터(Public Data) | | 민간 데이터(Private Data) |
|---|---|---|---|
| 공개 여부 | 공개 | | 비공개(기밀) |
| 목적 | 공공성, 다목적 | | 신규 특허, 특정 목적 |
| 소유 주체 | 정부, 기관 | | 개인, 기업 |
| 갱신 주기 | 대규모로 1회(장기간) | | 소규모로 지속적 발생(단기간) |
| 민감 여부 | 개인정보 포함 | | 지적재산권 해당 |
| 데이터 축적 방식 | 정부 연구 성과물로 데이터 수집 | 정부 사업으로 직접 생산 | 신약개발 실험 및 검증과정에서 발생하는 데이터 |
| 데이터 특징 | 다양성의 폭이 넓지만, 밀도가 낮음 | | 다양성의 폭은 좁지만, 밀도가 높음 |

❍ 신약개발 데이터의 공유 활성화 방안 연구 수행을 위해 신약개발 데이터의 공개 여부 속성에 따라 크게 공공과 민간 데이터 2가지로 구분하였고, 이에 따라 공유 활성화 방안 연구의 접근을 달리하였음

❍ 공공 데이터의 공유 활성화 방안 마련을 위해서 국내외 신약개발 데이터가 모이고 있는 공유 플랫폼과 공개 데이터를 분석했고, 데이터 공유와 직접 관련된 법률적 환경 분석을 수행하여 각각의 시사점을 도출하고 이에

따른 공유 활성화 방안을 제시하였음

- ❍ 민간 데이터의 공유 활성화 방안 마련을 위해서 민간 신약개발 데이터의 연구 비밀을 유지하면서 AI에 활용하는 선도 사례와 기술을 분석해 새로운 방안을 제시하였음

## 2.2. 신약개발 데이터 공유 플랫폼 현황 분석

❑ 신약개발 데이터 공유 플랫폼 현황 분석 개요

❍ 국내외 정부 주도 신약개발 데이터 플랫폼 중에서 대표 사례 6개를 선정했고, 이를 분석해 국내 플랫폼의 활성화 전략을 도출하고자 했음

❍ (국내) 국내의 대표적인 신약개발 데이터 플랫폼 사업인 국가 통합 바이오 빅데이터 구축 시범사업, 국가 바이오 데이터 스테이션(K-BDS), 인공지능 신약개발 플랫폼(KAIDD) 사업을 분석 대상으로 선정하였음

❍ (해외) 해외의 대표적인 신약개발 데이터 플랫폼 사업인 All of Us 연구 프로그램, 영국 UK 바이오뱅크, 핀란드 핀젠 연구 프로젝트, 미국 NCBI Database Services, 스위스 Drug Design 사업을 분석 대상으로 선정하였음

표 6. 신약개발 데이터 공유 플랫폼 비교

| 분류 / 속성 | 국내 데이터 플랫폼 | 해외 데이터 플랫폼 |
|---|---|---|
| 임상·유전체 데이터 수집·공유 | 국가 통합 바이오 빅데이터 구축사업 | All of Us (미국)<br>UK 바이오뱅크 (영국)<br>핀젠 연구프로젝트 (핀란드) |
| 바이오 데이터 수집·공유 | 국가 바이오 데이터 스테이션(K-BDS) | NCBI Database Services (미국) |
| 공개 데이터 가공 및 모델 공유 | 인공지능 신약개발 플랫폼 (KAIDD) | SWISS Drug Design (스위스) |

### 2.2.1. 국가 바이오 빅데이터 구축 시범사업

❑ 개요

❍ 100만 명 이상의 한국인 건강·유전 정보를 모으고 디지털화하는 사업

❍ (목표) 국민의 자발적 참여로 한국인의 건강정보와 유전 정보를 모아 안전한 플랫폼으로 관리하며, 자격 있는 연구자들이 정보를 분석하는 체계 구축

- 중증 난치질환자 20만 명, 암 환자 10만 명, 희귀질환자 10만 명, 자발적 참여자 60만 명을 모집하여 2023년부터 2028년까지 6년 동안 한국인 100만 명 이상의 데이터를 구축할 예정임
- 2개년에 걸친 시범 사업(2020년 6월 ~ 2022년 12월)을 통해 누적 20,000명의 데이터 구축이 예정되어 있음

❑ 데이터 구축 방법

❍ 수집하는 참여자의 임상 데이터, 인체 유래물(혈액, 소변, 타액 등) 기반 유전체 서열 및 변이 데이터, 주민등록번호를 수집함

- 참여자가 제공한 임상 데이터와 인체 유래물은 국립중앙인체자원은행에 보관되며 일부분은 연구 활용을 위해 유전체 분석 기관으로 옮겨져 유전체 데이터로 전환됨
- 생성된 유전체 데이터는 국가생명연구자원정보센터로 전달되어 분석을 시행하고, 유전 변이 분석 데이터를 생성하며, 다시 질병관리청으로 전달되어 보관됨



그림 1. 참여자 정보 보관 체계(출처: 질병관리청, 국가 바이오 빅데이터 시범 구축사업 프리스킷)

❑ 데이터의 현황 및 이용

❍ (데이터 현황) 2023.01.04.일 기준 공시된 데이터 현황은 참여자 15,000건, 유전체 데이터 15,010건, 진단 참고용 보고서 13,167건임

표 7. 국가 바이오 빅데이터 구축 시범사업 데이터 현황

| 코호트 | 데이터 구분 | 세부 데이터 |
|---|---|---|
| 울산만명 유전체프로젝트 | 임상 데이터 | 신체 계측, 시력 검사, 청력 검사, 생 검사, 혈액 검사, 면역 검사, 갑상선 검사, 소변 검사, 자궁 경부 검사, X선 검사, 초음파 검사, 종양표지자 검사 데이터 등 116개 변수 |
| | 유전체 데이터 | bam : 정렬된 유전체 서열 정보<br>gvcf : 유전체 변이 정보(SNP, Indel 등) |
| 희귀 질환 | 임상 데이터 | 가족 질환 분류, 가족 관계, 성별, 만 나이, 질환 분류(대/중/소), 질병코드(ICD10) 등 19개 변수 |
| | 유전체 데이터 | Fastq : 기기를 통해 생성된 원시 데이터 |

| | | |
|---|---|---|
| | | bam : 정렬된 유전체 서열 정보<br>gvcf : 유전체 변이 정보(SNP, Indel 등) |
| 자폐<br>스펙트럼 | 임상 데이터 | 환자정보, 사회적 의사소통 수준, 사회성 반응성, 아동기 행동평가척도, 지능검사, 바인랜드 적응 행동 척도, 정밀주의력 검사, 자폐증 진단 관찰 스케쥴, 자폐증 진단 면담지, 아동기 자폐증 평정척도 등 69개 변수 |
| | 유전체 데이터 | bam : 정렬된 유전체 서열 정보<br>gvcf : 유전체 변이 정보(SNP, Indel 등) |
| 대장암 | 임상 데이터 | 성별, 나이, 진단일, 중복암 여부, 암 진단 정보, 생존 정보, 치료 정보 등 88개 변수 |
| | 유전체 데이터 | bam : 정렬된 유전체 서열 정보<br>gvcf : 유전체 변이 정보(SNP, Indel 등) |
| KoGES | 임상 데이터 | 성별, 나이, 질병 과거력, 가족력, 생활습관, 식품섭취 빈도, 임상검, 수면력, 추적통합 등 199개 변수 |
| | 유전체 데이터 | Fastq : 기기를 통해 생성된 원시 데이터<br>bam : 정렬된 유전체 서열 정보<br>gvcf : 유전체 변이 정보(SNP, Indel 등) |

❍ (데이터 이용) 임상·유전체 데이터 활용을 희망하는 연구자는 연구계획서를 제출해야 하고 연구 윤리위원회(IRB) 심의를 통과한 후 비식별 처리된 임상·유전체 데이터를 받을 수 있으며 모든 연구 과정은 폐쇄된 연구 플랫폼 환경 내에서만 진행됨

– 데이터 이용자에게 제공되는 모든 정보는 비식별 처리되어 제공되기 때문에 누구의 정보인지 절대 알 수 없고, 모든 연구 과정을 국가가 감시할 수 있는 폐쇄된 연구 플랫폼 내에서만 진행되며, 정보의 반출은 절대 허용되지 않음

❑ 시사점

❍ 국내에서 처음 시도하는 대규모 바이오 데이터 구축 프로젝트로 성공 시 다양한 참여자의 유전체 빅데이터 확보가 가능하겠으나, 폐쇄된 데이터 활용환경은 인공지능 신약개발 연구에 활용을 저해하는 요인으로 작용할 수 있어 데이터 공유를 위한 정책적 지원 방안 마련이 필요함

### 2.2.2. 인공지능 신약개발 플랫폼(KAIDD)

❑ 개요

❍ 글로벌 신약개발에 필요한 인공지능 플랫폼을 구축하여, 국내 신약개발 연구자를 대상으로 서비스를 제공

❍ (목표) 신약 표적 탐색부터 디자인, 최적화, 검증, 신약 후보 도출까지 신약개발 전 과정을 수행하는 인공지능 기반 통합 플랫폼을 개발

- 신약 후보 물질 발굴을 위한 인공지능 기반 플랫폼 구축
- 약물 재창출을 위한 인공지능 기반 플랫폼 구축
- 스마트 약물감시를 위한 인공지능 기반 플랫폼 구축
- 인공지능 활용 신약개발 플랫폼 구축 지원

❑ 데이터 구축 방법

❍ 스마트 약물감시 및 인공지능 활용 신약개발 플랫폼 구축사업에는 빅데이터 플랫폼과 표준화된 통합 데이터베이스 구축 목표가 있음

❍ 신약개발 데이터 표준화에 대한 가이드라인 문서[6]는 있으나 표준화된 공개된 데이터가 없는 것으로 파악

❍ 표준화 가이드라인에서 예시 또는 연계 계획으로 제시한 데이터는 ChEMBL, SureChEMBL, ZINC 15, PubChem이며, ChEMBL 데이터베이스 테이블 유형을 사용하여 해당 사업에서 생성될 정형 데이터(화합물 구조, 단백질 구조, 활성 값, Physicochemical, ADME/Tox)와의 연계성을 높이겠다고 언급

❑ 플랫폼 현황 및 지원 기능

❍ (플랫폼 현황) 플랫폼 제공 도구에는 AD3: 단백질 구조 기반, CSK Studio: Neuro degenerative 신경 퇴행성, MiLearn: Anticancer Drug 항암신약, AIDrug: 빅데이터/AI 신약, Synbi: 약물 재창출, SmartPV 스마트 약물감시가 있음

❍ (플랫폼별 지원 기능)

- AD3 단백질 구조 기반 플랫폼: 약물 발견의 단계를 Target, Discovery, Development의 3단계를 구분하고 각 단계에 서브 도구들을 제공하여 인공지능 기반 약물 발견을 지원함
- CSK Studio 신경 퇴행성 플랫폼: 스탠다임의 iCLUE&ASK를 사용하여 새로운 표적을 식별할 수 있도록 질병-유전자 연관성에 관한 지속 업데이트되는 데이터베이스를 사용할 수 있는 도구를 제공하여 질병을 검색

6) 인공지능신약개발플랫폼, AI기반 신약개발 플랫폼 구축사업 데이터 표준화 가이드라인, 2021.09.08

하면 질병에서의 우선시 되는 표적을 빠르게 찾을 수 있도록 지원함

- MiLearn 항암 신약 플랫폼: AiCAD(표현형 기반 항암 표적 치료제 스크리닝 모델), AiGPro(다중서열 정렬 기반 표적-항암제 가상 스크리닝 모델), AiKPro(다중서열 정렬 기반 표적-항암제 가상 스크리닝 모델), AiP450(CYP450 기반 독성 예측 모델) CRX4 (Kinase-likeness, GPCR-likeness 예측 모델)을 지원함
- AIDrug 빅데이터/AI 신약 플랫폼: ADMET (예측 AI 모델 12건), Toxicity 등(예측 AI 모델 16건), De novo Design(Scaffold, 물성, 단백질 기반 구조 생성 등 AI 모델 3건), Virtual Screening(약물-단백질 상호작용 예측 등), 기타(빅데이터 검색 등)를 지원함
- Synbi 약물 재창출 플랫폼: Synbi Drug-R(약물 다중 특성 기반 승인 약물의 항암 표적 및 효능 예측)을 지원함
- Smart PV 스마트 약물감시 플랫폼: Smart PV irAE(인공지능 기반 면역 관련 부작용 예측 모델), CDM Based ADR screening tool(환자 정보 및 바이오마커 기반 부작용 예측 모델)을 지원함

❑ 데이터 현황

표 8. KAIDD 데이터 구축 현황

| 플랫폼 | 데이터 구축 현황 | |
|---|---|---|
| AD3 | 원본 데이터 출처 표기 여부 | O |
| | 원본 데이터 수량 표기 여부 | O |
| | 원본 데이터 전처리 기준 및 방법 공개 여부 | X |
| | 가공 데이터 공개 여부 | O |
| CSK Studio | 원본 데이터 출처 표기 여부 | △ |
| | 원본 데이터 수량 표기 여부 | X |
| | 원본 데이터 전처리 기준 및 방법 공개 여부 | X |
| | 가공 데이터 공개 여부 | X |
| Mi-Learn | 원본 데이터 출처 표기 여부 | O |
| | 원본 데이터 수량 표기 여부 | X |
| | 원본 데이터 전처리 기준 및 방법 공개 여부 | X |
| | 가공 데이터 공개 여부 | X |

※ O : 있음, X : 없음, △ : 일부 누락 됨, - : 해당 없음

* : 해외, ** 국내

표 9. KAIDD 플랫폼 별 데이터 유형 및 출처

| 플랫폼 | 모델명(설명) | 데이터 유형 | 데이터 세부 유형 | 데이터 출처 |
|---|---|---|---|---|
| AD3 | GalxyTBM (단백질 구조예측) | 생물학 | 단백질 구조 | Protein Data Bank* |
| | | 생물학 | 단백질 서열 | UniProt*<br>Big Fantastic Database** |
| | Redesigns (후보물질 생성) | 생물학 | 단백질 구조 | Protein Data Bank* |
| | | 화합물 | 화합물 구조 | ZINC* |
| | | 약리학 | 단백질-화합물 결합력 | ChEMBL* |
| | GalaxyDock3 (결합 가능성 예측) | 생물학 | 단백질 구조 | Protein Data Bank* |
| | | 화합물 | 단백질-화합물 결합구조 | Protein Data Bank Bind* |
| | AK-Scores (결합 친화도 예측) | 화합물 | 단백질-화합물 결합구조 | Protein Data Bank Bind* |
| | spica(ADME 예측) | 약리학 | 약동학 ADME | ChEMBL, Moleculenet* |
| CSK Studio | NetExp (표적 예측) | 생물학 | 유전자 기능 | Gene Ontology* |
| | | | 신호전달경로 | KEGG* |
| | | | 인간 전사체 | 데이터 출처가 표기되지 않음 |
| | | | 단백질 도메인 및 기능 | InterPro* |
| | | | 단백질 상호작용 | STRING* |
| CSK Studio | Molecule Generation (화합물 예측) | 화학 | 화합물 구조 | ZINC*<br>MOSES* |
| | | * 상세 표기되지 않은 데이터를 포함 | | |
| | BioActivity Prediction (독성 예측) | 약리학 | 독성시험(심장) | 데이터 출처가 표기되지 않음 |
| | | | 독성시험(간 독성) | |
| | | | 뇌 장벽 투과성 | |
| | Drug Neighbor (유사 약물 검색) | 화합물 | 화합물(약물) 구조 | Drug bank* |
| Mi-Learn | AiCAD (세포주 활성 예측) | 약리학 | 약효 및 선택성(세포 생장 저해) | NCI 60* |
| | AiGpro (저해 활성 예측) | | 약효 및 선택성 (단백질(GPCR) 저해) | GLASS* |
| | AiKpro (저해 활성 예측) | | 약효 및 선택성 (인산화효소 저해) | PKIS* |
| | AiP450 (CYP 억제 예측) | | 간 독성 (CYP 저해) | PubChem* |
| | | | | K-MEDI Hub** |
| | CRX4 (Likeness 예측) | 화합물 | 화합물 구조 (인산화효소 저해제, GPCR 저해제) | ChEMBL* |
| AI- | ADME | 약리학 | 수동 수송 | 데이터 출처가 표기되지 않음 |
| | | | 약동학 ADME | |

| | | | | |
|---|---|---|---|---|
| Drug | (ADME 예측) | | (뇌 장벽 투과성) | |
| | | | 간 독성(CYP 저해) | |
| | | | 대사 안정성 | |
| | | | 독성(심장) | |
| | Toxicity 등 (Toxicity 예측) | 약리학 | 독성 | MoleculeNet Tox21* |
| | | | 약동학 ADME (뇌장벽 투과성) | MoleculeNet BBBP* |
| | | | 약효 및 선택성 (EGFR/TGFR1/VGFR1) | CheMBL* |
| | Denovo Design (신규물질 디자인) | 화합물 | 화합물 구조 | ZINC* |
| | Virtual Screening (상호작용 예측) | 화합물 | 단백질-화합물 결합력 (인산화효소) | 데이터 출처가 표기되지 않음 |
| AI-Drug | 기타 (빅데이터 검색) | 화합물 | 화합물 구조 | ZINC* |
| | | | 물리시험 | MoleculeNet* Esol, FreeSolv* |
| | | 약리학 | 독성시험 | MoleculeNet* BBBP, Tox21, ToxCast, MUV, SIDER, ClinTox, HIV* |
| Synbi | SynbiR (항암효능예측) | 데이터 세부 유형과 출처가 표기되지 않음 | | |
| Smart PV | Smart PV irAE (면역 부작용 예측) | 임상 | 환자 면역 부작용 기록 | 아산병원 외 8개 기관** |
| | | 임상 생물학 | 환자 유전체(WES) 지표 | |
| | CDMBased ADR Screening tool (바이오마커 기반 부작용 예측) | 데이터 출처가 표기되지 않음 | | |

❑ 시사점

❍ 신약개발에 활용할 수 있는 인공지능 도구를 서비스하는 플랫폼 개발이 핵심으로 신약개발 전 과정의 적용 가능한 다양한 모델 및 서비스를 제공함 그러나, 데이터 접근법, API, 라이브러리에 대한 보완이 필요

## 2.2.3. 국가 바이오 데이터 스테이션(K-BDS) 사업

❑ 개요

❍ 국가 바이오 R&D를 통해 생산·활용되는 모든 데이터를 연계하여 수집·제공하는 사업

❍ (목표) 국내 바이오 데이터를 범부처가 협력하여 국가 차원에서 체계적으로 확보·관리하고 연구자들에게 제공하여 연구데이터 활용체계를 마련

- 국가 R&D로 생산된 바이오 연구데이터의 지속 가능한 통합 수집·제공
- 다량의 데이터가 안정적으로 저장·관리될 수 있도록 전산 환경 구축
- 수집된 바이오 연구데이터의 통합 분석·활용 환경 제공
- 빅데이터 분석 기술 개발 및 전문 공공·민간 기관 육성
- 국가 바이오 재난 대응을 위한 데이터의 체계적 확보·공유
- 현장에서 필요로 하는 바이오 연구데이터 분석·활용 전문인력 양성
- 해외 바이오 데이터 기관들과 글로벌 데이터 협력 업무수행

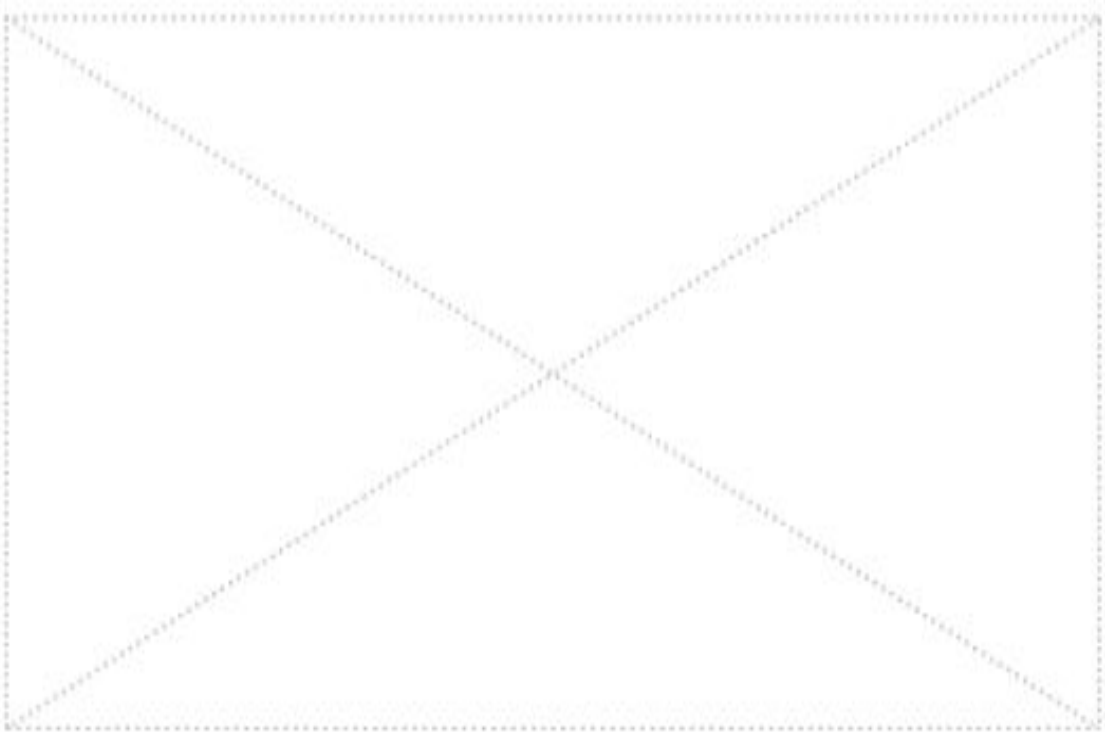


그림 2. 국가 바이오 데이터 스테이션(K-BDS) 개념도(출처: 국가 바이오 데이터 스테이션)

❑ 데이터 구축 방법

❍ 국가 연구개발 사업으로 추진된 바이오 연구데이터(바이오 연구 수행에서 생성된 모든 데이터를 의미하며, 실험을 통해 얻은 실데이터와 이를 설명하기 위한 메타데이터로 구성됨)를 모두 포함

- 국가연구개발사업에서 창출, 파생된 데이터는 표준 등록 양식에 따라 데이터를 등록하고, 품질관리자가 데이터를 검수 등록하고 있음
- 해외에서는 미국 NCBI(National Center for Biotechnology Information), 유럽 생물 정보학 연구소(EMBL-EBI), 일본 국립유전학연구소(DDBJ)와 연계하고 있음



그림 3. K-BDS 연구데이터 등록 절차

❑ 데이터의 현황 및 이용

❍ (데이터 현황) 2023.01.04.일 기준 공시된 데이터 현황은 유전체 1,264,997건, 단백체 5,078건, 대사체 7,111건, 화합물 197,201건, 이미지 10,342건, 기타 93건으로 전체 1,484,822건의 데이터가 구축되어있음

❍ (데이터 이용) 학계 연구자, 정부출연 연구기관 연구원, 병원이나 바이오 계열 산업계, 개인 연구자 모두가 이용할 수 있게 공개되어 있으나, 회원가입 시 국가연구자 번호를 입력해야 하므로 외국 연구자의 데이터 활용에 어려움이 있을 것으로 예상함

❑ 시사점

❍ 국가 단위의 통합된 바이오 연구데이터 공유 포털을 구축한 사업으로 연구데이터 공유로 연구 연속성 확보, 중복연구 가능성 감소, 데이터 기반 연구의 강화 등의 장점이 있음

❍ 그러나, 데이터 활용성이 미흡하여 데이터 공유 및 활용체계 구축이 필요함

## 2.2.4. 미국 All of Us 연구 프로그램(All of us Research Project)

❑ 개요

❍ 정밀 의료 달성을 위해 특정 질병이 아닌 100만 명 이상의 다양한 참여자를 모집하여 다양성 높은 건강 데이터를 구축하는 프로그램

❍ (목표) 미국 국립 보건원(NIH)에서 주도하는 연구로 개인의 생활습관, 환경 및 생물학적 구성이 건강과 질병에 어떻게 영향을 주는지 탐구하고 건강을 유지하는 제일 나은 방법을 알려주는 정밀 의료의 목표 달성을 위해 필요한 100만 명 이상의 데이터를 구축하고 공유하는 프로그램



그림 4. All of us 연구 절차

❑ 데이터 구축 방법

❍ 참여자의 임상 데이터(건강 관련 설문 조사 내용, 전자 건강 기록, 건강검진, 디지털 헬스케어 데이터, 참가자 개인정보)를 유전체 데이터(혈액, 타액, 소변 등 생체시료)와 연계해서 데이터를 구축 중이며, 2023년 1월 기준 57만 명 이상이 참여했음

- (참여 유인책) 참여자에게 유전적 변이 및 약물유전학에 관한 데이터 전달을 우선으로 수행하며 심각한 질병을 유발할 것으로 예상되는 유전적 변이에 대해서도 알려주는 참여 혜택을 제공함
- 2019년 All of us 연구 발표에 따르면 약 30,000명의 참여자에게 유전 분석 결과를 전달했으며, 유전체 상담을 할 수 있도록 시스템을 구축하였음

❑ 데이터의 현황 및 이용

❍ (데이터 현황) 2023.01.04. 기준 공시된 데이터 현황은 578,000명의 341,000건의 전자의료기록(Electronic Health Record, EHR), 418,000건의 생체시료가 구축되어있으며, 다양성 확보를 위해서는 인종, 성적, 민족 소수자의 비중이 50% 이상을 포함

- 데이터 세트를 탐색할 수 있는 공용 브라우저는 2019년 5월에 출시되었

으며 분석을 위한 도구는 2020년 초에 제공하였음

❍ (데이터 이용) 2020. 5. 27. All of us Research Hub 사이트 베타 서비스를 오픈하여 미국 전역의 다양한 참가자로부터 수집된 약 225,000명의 건강 데이터에 접근할 수 있으며, 현재까지 서비스하고 있음

- (데이터 브라우저) 공개적으로 접근할 수 있도록 설문 조사, EHR, 생체지표, 웨어러블 데이터, 약물 노출 데이터를 제공하고 있음
- (연구자 워크벤치) 미국에 기반을 둔 모든 학계, 의료기관, 비영리 단체 등이 연구 프로그램과의 데이터 사용 계약 체결 시 사용할 수 있으며, 플랫폼은 데이터 분석 및 협업 기능을 지원함(데이터 분석 도구 주피터 탑재)



그림 5. Public Data Browser(왼쪽), Researcher workbench (오른쪽)

❑ (시사점) 높은 다양성을 추구하면서 100만 명이라는 대량의 데이터를 확보하는 것이 목표이며, 구축 완료 시 다양한 분야의 연구에서 활용할 수 있는 것으로 예상됨

❍ 데이터 생성 참여자들에게 유전자 분석 서비스 및 개인 맞춤형 의료서비스 제공의 참여 동기 부여로 데이터 생산 전략이 뛰어나며, 사용자에게 데이터 분석 도구를 내장형(Built-in)으로 제공하기에 활용성이 높음

## 2.2.5. 유럽연합 Beyond 1+Million Genomes

❑ 개요

❍ 유럽 23개국이 참여하는 임상 데이터 네트워크 구축으로 100만 개의 시퀀싱 된 유전체 데이터를 축적하고 공유하는 연구 인프라 조성 프로젝트

❍ (목표) 유럽 전역의 생명 과학 자원을 하나로 모으는 ELIXIR 기관 주도 100만 개 이상 유전체 확보, 데이터의 품질, 표준, 기술 인프라 및 ELSI(Ethical, Legal and Social Issues)에 대한 요구사항 정의 및 구현을 목표

- 생명 의학 연구를 위해 임상 데이터를 축적하고, 영구 보관하기 위한 저장소인 EGA(European Genome-Phenom Archive)에 통합함
- EGA 통합은 연합 프레임워크를 사용한 통합을 의미하며, 데이터는 로컬에 위치하면서 메타데이터로 연결되어 분산된 모든 데이터에 접근 가능



그림 6. Federated EGA 연합 데이터 플랫폼 개념

❍ (데이터 구축 방법) 유럽연합 국가들에 구축되어있는 500개 이상의 바이오뱅크에 1억 개 이상의 생체시료를 통해 유전체 데이터(Whole Genome Sequencing, WGS)를 최소 100만 건 이상 구축할 예정임

❑ 데이터의 현황 및 이용

❍ (데이터 현황) 2023.01.05. 기준 현황은 핀란드 513,700명, 네덜란드 222,371명, 이탈리아 144,393명, 영국 98,341명, 스페인 69,851명, 독일 47,385명, 스웨덴 22,859명 등 16개국에서 1,117,148개의 유전체(WGS) 데이터 구축

❍ (이해관계자 협업) 2020.10.19. B1MG(Beyond 1 Million Genomes)에서 시민, 환자, 임상의, 의료전문가, 연구자, 의약품 당국, 자금제공자, 산업계, 국가 정책 입안자 등 유전체 데이터 구축 활용에 관련된 이해관계자들이 의견을 공유하고 조율할 수 있는 포털을 출시하고 운영하고 있음

❑ 시사점

❍ 100만 개의 유전체 데이터가 축적된 연합 데이터 공유 플랫폼을 구축하는

사업으로 로컬 데이터 저장소의 독립성과 데이터 안전성을 유지하면서 전 유럽의 유전체 데이터의 통합 검색을 지원해 활용성이 높을 것으로 예상됨

## 2.2.6. 핀란드 핀젠 연구 프로젝트(FinnGen Research Project)

❑ 개요

❍ 공공-민간 협력 기반 최초의 대규모 정밀 의료 프로젝트

❍ (목표) 50만 명의 핀란드 바이오뱅크 참가자로부터 유전체와 임상 데이터를 수집하고 분석을 통해 질병 원인에 대한 이해를 높이고 질병의 진단, 예방 및 치료법 개발 연구를 촉진할 수 있는 리소스 구축을 목표로 함

- (주요 목표) 임상 정보와 유전체 데이터를 결합하여 의료 혁신을 창출, 핀란드가 생물의학 및 정밀 의료 분야의 선구자가 되도록 지원, 공공-민간의 협력 모델을 생성, 모든 핀란드인을 위한 정밀 의료, 건강 혁신의 초창기 모델 제시



그림 7. FinnGen의 민관협력 구조

❑ 데이터 구축 방법

❍ 유전자 데이터를 기반으로 정보 제공자의 인구 통계 데이터(생년월일, 성별, 사망 원인 등)와 임상 데이터(혈압, 신진대사, 병원 진료기록, 투약 정보 등)는 전자 건강기록부인 칸타 서비스에 결합

- (참여 유인책) 데이터 제공자인 시민의 참여 유도를 위해 페이스북(SNS 캠페인), 유튜브(짧은 애니메이션 영상을 제작 발표), 모바일 게임(Bioholvi 모바일 어드벤처 게임), 설문 조사 등을 수행하였음
- (참여 혜택) 국가 차원의 건강 증진 향상에 이바지한다는 것, 질병의 원인에 대한 추가정보를 제공한다는 점

❑ 데이터의 현황 및 이용

❍ (데이터 현황) 2022.12.01. 기준 연구 결과에서 발표된 데이터 현황에 따르면, 총 표본 342,499명으로 여성 190,879명, 남성 151,620명으로 구성되어 있고, 분석된 변이는 20,175,454개, 질병은 2,202개를 구축하였음

❍ (데이터 이용) 초기부터 글로벌 제약사가 주요 파트너(애브비, 아스트라제네카, 바이오젠, 얀센, 화이자, 노바티스, 사노피 등)로 의사결정에 참여하였으며, 데이터는 공개적으로는 접근할 수 없고 컨소시엄 파트너를 대표하는 연구원만 접근할 수 있게 설계되었음

- 대신 요약 통계나 전장 유전체 연관분석(Genome Wide Association Study, GWAS) 결과는 6개월마다 외부에 공개하고 있음[7)]



그림 8. FinnGen 민간 파트너 협력사

❑ 시사점

❍ 의료 연구 및 신약개발을 목표로 참여 컨소시엄 파트너에게만 데이터를 공개하는 폐쇄적 구조로 데이터의 공공성은 떨어지나, 핀란드의 제약산업의 경쟁력 제고에 기여

7) FinnGen 결과 공유 페이지, https://www.finngen.fi/en/access_results

## 2.2.7. 미국 NCBI Database Service

❑ 개요

❍ NIH 산하 1988년 설립된 NCBI가 제공하는 바이오 데이터베이스

❍ (목표) 건강과 질병에 영향을 끼치는 근본적인 분자와 유전 생물학적 과정의 이해를 돕기 위한 생물학적 정보 시스템 및 데이터베이스 구축

– (주요 목표) 임상, 유전체, 유전자, 단백질, 화합물, 문헌 등 6개의 범주의 데이터를 수집하고 내려받기 위한 인프라, 시각화 및 분석 SW 및 서비스 제공

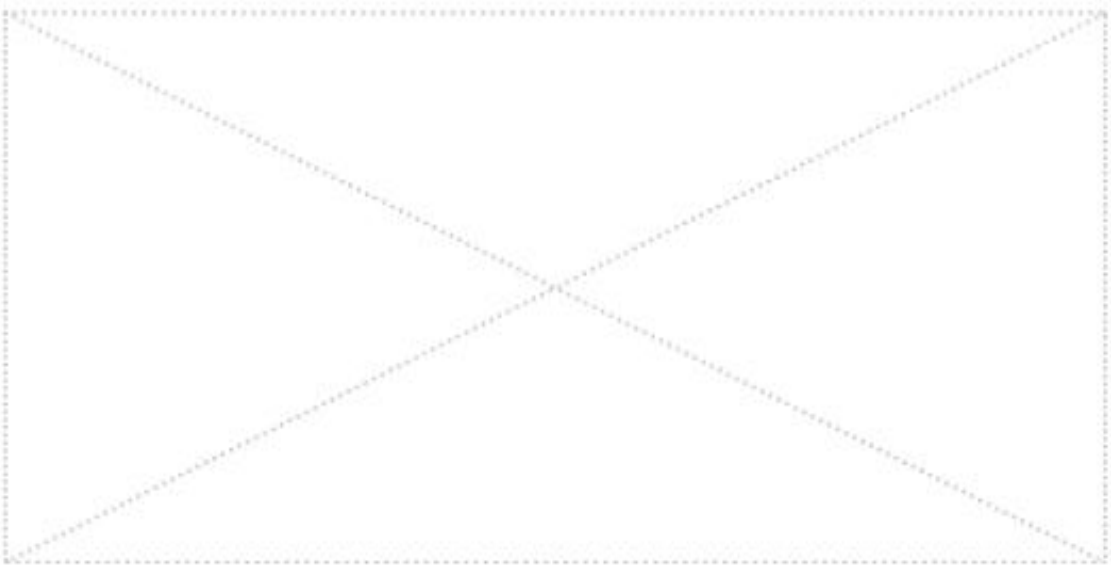


그림 9. NCBI 바이오 시스템 데이터베이스 개념도 (출처 : Nucleic Acids Research)

❍ (데이터 구축 방법) 연구자의 직접적인 제출, 미국 및 국제 협력 및 계약을 통한 데이터 공급자 및 연구 컨소시엄과의 계약, 내부 데이터 큐레이션 등을 통해 데이터를 수집

– 국제 주요 정보기관(ENA/EBI, DDBJ, INSDC, KBDS 등)과의 데이터 파트너링

❑ 데이터의 현황 및 이용

❍ (데이터 현황) 2016년 기준, 문헌 데이터 3,000만 건, 임상 데이터 69만 건, 유전체 데이터 10억 건, 유전자 데이터 2억 건, 단백질 데이터 2억 건, 화합물 데이터 23억 건 이상을 수집하였음

❍ (데이터 이용) 데이터의 검색성, 접근성, 상호운용성, 재활용성을 고려하여 공유하자는 FAIR(Findable, Accessible, Interoperable, Reusable) 원칙을 기반으로 데이터를 정리했고, 교육 및 매뉴얼 제공으로 별도의 인력 및 체계 확보 없이 데이터 활용이 가능함

- DB가 목적에 따라 분류되어 있어 메타데이터(Metadata) 및 식별자(Identifier)를 부여한 검색 및 제어가 쉬워 데이터 검색이 편리함
- 데이터는 즉시 다운로드 가능하며 필요에 따라 대용량 데이터를 선택적으로 내려받을 수 있는 FTP 및 SW를 제공함
- EBI, DDBJ, INSDC 등과 데이터 표준체계를 공유하여 상호운용성이 높음
- 데이터 활용을 위해 필요한 메타데이터 정보를 충분히 제시하며 타 DB의 메타데이터와 호환성이 높음

표 10. NCBI 데이터 유형 및 신약개발 관련 대표 데이터베이스

| 유형 | 데이터베이스 | 설명 |
|---|---|---|
| 문헌 | PubMED | 모든 생명 과학 및 의학 저널에 대한 인용 및 초록 DB |
| | PubMED Central | 무제한 접근이 가능한 저널 및 보고서를 모아놓은 DB |
| 임상 | ClinVar | NIH Genetic Testing Registry에 포함된 변이에 따른 건강 변화에 대한 정보를 공개한 DB |
| | ClinicalTrials | 전 세계의 임상시험 등록 및 결과에 대한 DB(비공개 포함) |
| 유전체 | dbGaP | 유전자형 및 표현형의 상호작용을 조사하여 연구 결과 및 설명을 저장한 DB |
| | BioSample | 실험에 활용된 세포(세포주 및 환자 세포) 정보를 모아놓은 DB |
| | BioProject | 유전체 및 기능 유전체학 관련 연구 결과 DB |
| | GEO | NCBI가 수집한 유전체(NGS, Array) 데이터를 연구목적, 결과와 함께 제공하는 DB |
| | OMIM | 인간 유전자 및 변이에 따른 유전 질환 연관 정보 DB |
| | Consensus CDS | 인간 및 쥐의 유전체 내 단백질 코딩 영역 공통 세트에 대한 정보DB |
| | dbSNP | 짧은 변이(SNV, Miscrosatelites, small Indel) 빈도 및 통계 DB |
| | dbVar | 대규모 변이(large Indel, Translocation, Inversion) 등 정보 DB |
| 유전자 | Assebly | 조립(Assembly)이 완료된 유전체의 구조, 이름, 서열 정보 및 메타데이터 제공 |
| | GenBank | DNA 서열 및 주석을 모아놓은 유전자 서열 DB |
| | RefSeq | 중복되지 않는 유전체, 전사체, 단백질 참조서열(표준) DB |
| 단백질 | Protein Cluster | 유전체 참조서열 기반의 단백질 서열 DB |
| | Protein Database | 다기관에서 수집한 단백질 서열 DB |
| | PFM | 유사한 구조 및 기능의 단백질 패밀리 정보를 제공 |
| | CDD | 진화 과정에 따라 보존된 단백질 도메인 서열 정보를 제공 |
| 화합물 | PubChem | 약리학(BioAssay) 실험 결과로 생성된 화합물의 활성 정보제공 Pubchem에 기탁된 화합물 구조 및 상세 정보 제공 |

❑ 시사점

❍ 신약개발 연구에 필요한 생물학, 화합물 데이터를 체계적으로 제공하여 AI 신약개발 연구에 필수적으로 활용

## 2.2.8. SWISS Drug Design

❑ 개요

❍ 신약개발에 필요한 저분자 약물 설계 데이터 및 프로그램(기술)을 쉽게 활용할 수 있는 플랫폼을 구축하고 공개 무료 운영 중

❍ (목표) 글로벌 신약개발 연구자를 위한 저분자 약물 설계 프로그램(기술)을 개발하고, 글로벌 연구자가 활용할 수 있도록 웹 기반 환경을 구축



그림 10. SWISS Drug Design ADME 예측 도구 (출처 : SWISS ADME)

❑ 플랫폼 및 데이터 현황과 이용

❍ (플랫폼 기술 현황) 5개의 약물 설계 프로그램(기술) 및 데이터베이스 제공

- SwissDock : 화합물-단백질 상호작용 및 결합구조 예측 프로그램
- SwissParam : 화합물 토폴로지 구조 매개변수 제공 프로그램
- SwissTargetPrediction : 화합물 기반 표적 단백질 예측 프로그램
- SwissSimilarity : 합성할 수 있는 유사 분자 생성 가상 스크리닝 프로그램
- SwissADME : 화합물의 물리화학 특성(용해도, 친유성, Drug-likeness 등), 약물 동태 특성(뇌 장벽 투과도, CYP 저해 등), 의약화학적 특성(합성 가능성 등) 예측 프로그램

❍ (플랫폼 DB 현황) 2개의 약물 설계를 위한 데이터베이스 제공

- SwissSidechain : 단백질 설계 및 디자인에 활용하는 단백질 곁사슬(사이드체인) 구조 및 분자 역학 데이터베이스 제공
- SwissBioisostere : 약물 최적화 단계에서 화합물 구조를 치환하기 위한 생동배체(유사한 생물학적 특성을 가진 치환기 그룹)의 빈도 및 영향 정보를 수집한 데이터베이스

표 11. SwissDrugDesign 프로그램별 데이터 유형 및 출처(출처 : SwissDrugDesign)

| 모델명(설명) | 데이터 유형 | 데이터 세부 유형 | 데이터 출처 |
|---|---|---|---|
| SwissDock | 화합물 | 단백질-화합물 결합구조 | Ligand Protein Database(LPDB) |
| SwissParam | 물리화학적 계산값으로, 데이터 해당 사항 없음 | | |
| SwissBioisostere | 화합물 | 화합물 및 동배체 구조 | ChEMBL |
| | 약리학 | 화합물 활성 | |
| SwissTargetPrediction | 약리학 | 화합물-단백질 활성 | ChEMBL |
| | 화합물 | 화합물 구조 | |
| SwissSimilarity | 화합물 | 화합물 구조 | ChEMBL |
| | | 약물 구조 | Drug Bank |
| SwissADME | 생물학 | 단백질 구조<br>단백질 서열 | Metrabase<br>PubChem<br>ZINC |
| SwissSidechain | 생물학 | 단백질 구조 | Protein Data Bank |

❍ (데이터 이용) 활용한 데이터의 출처 및 수량이 표기되어있고 전처리 기준 및 방법을 공개했고 모델에 대한 검증이 완료되어 있음

- 가공 데이터는 웹사이트를 통해 공개하거나, 발표된 논문을 통해 내려받을 수 있음

❑ 시사점

❍ 적용할 수 있는 분야가 약물 탐색으로 한정되어있고 AI 기반이 아닌 프로그램이 포함돼있으나, 원본 데이터를 내려받을 수 있고 웹서비스 기반으로 편리한 접근이 가능하여 활용성이 높음

## 2.3. 국내외 신약개발 공개 데이터 및 데이터베이스 현황 분석

### 2.3.1. 현황 분석 개요

- ❍ AI 신약개발에 화합물 및 유전체 데이터가 가장 많이 활용되는데 전자는 약물 후보물질의 구조정보와 특성 정보를 제공하며, 후자는 유전자 변이 및 발현 데이터 등을 포함함
- ❍ 현시점에서 AI 신약개발은 신규 약물 디자인, 약물 구조 최적화, 약물 합성 경로 분석, 약물 활성 및 독성 예측 등 약물 탐색 및 최적화 단계의 예측 모델 연구개발에 집중되고 있음
- ❍ 본 현황 분석에서는 AI 신약개발에 가장 많이 활용되는 데이터를 획득할 수 있는 국내외 신약개발 공개 데이터베이스는 어떤 것들이 있는지 조사하고, 국내외 대표 사례의 비교 분석을 통해 국내 공개 데이터의 공유 활성화 부족의 근거를 도출했음

❑ 신약개발 단계별 데이터 및 데이터베이스

- ❍ 공개 데이터베이스 현황 분석에 앞서 신약개발 단계별 필요 데이터와 공개 데이터베이스를 정리하면 아래의 그림과 같음
- ❍ 일부 데이터베이스는 다양한 데이터를 포함하고 있어 여러 신약개발 단계에 걸쳐서 사용됨



* 대규모 유전체 사업 데이터(국가 바이오 빅데이터 구축 시범사업, All of US, FinnGen, Genomics England 등)

그림 11. 신약개발 단계별 기술과 필요 데이터 그리고 데이터베이스

## 2.3.2. 대표 사례 비교 분석

❑ (국내) 한국화합물은행 화합물 라이브러리

❍ (개요) 2000년을 시작으로 질병 관련 유전체 또는 단백질 기능 조절 연구 등의 화합물로 활용하거나 신약후보물질 개발 연구의 출발 물질로 사용될 수 있는 유기화합물 및 단일 성분 천연물 약 737,000 여종의 화합물 보유

표 12. 화합물은행 제공 라이브러리 현황

| 라이브러리 종류 | 화합물 수 | 농도(평균) | 설명 |
|---|---|---|---|
| 전체 | 680,000 | 5 mM, 5 uL | 전체 화합물 라이브러리 |
| 대표 | 7,000 | 5 mM, 5 uL | 전체 화합물을 대표하는 라이브러리, 순도 및 분자량 검증(LC/MS) |
| Kinase | 3,000 | 5 mM, 5 uL | 분자 모델링 방법을 적용하여 Kinase 표적 대상 active site에 결합할 가능성이 높은 화합물 |
| 임상화합물 | 3,100 | 5 mM, 5 uL | 임상 I-III상 단계 화합물 및 승인 약물 |
| Fragement | 1,600 | 20 mM, 5 uL | 분자량 300이하 라이브러리, 순도 및 분자량 검증(LC/MS) |
| 천연물 | 1,500 | 5 mM, 5 uL | 단일 성분 천연물 및 천연물 유사골격 구조의 화합물 |
| GPCR | 9,000 | 5 mM, 5 uL | 해외 vendor로부터 선별 구매화합물 |
| PPI | 17,000 | 5 mM, 5 uL | 해외 vendor로부터 선별 구매화합물 |
| PharmaCore | 요청개수 | 5 mM, 5 uL | 요청골격으로 선별한 화합물 또는 가상탐색으로 선별한 화합물 |

❍ (데이터 활용 방식) 공공 COVID-19 데이터를 제외하면 통합데이터플랫폼 사이트를 통해 화합물 활용 신청 및 계약을 진행해야 하며 활용 계약서를 작성하거나 데이터 프로젝트(사용 기관 정보, 프로젝트 책임자 정보, 담당자 정보, 데이터 활용 정보)를 등록해야 함

❍ (활용 결과 권리) 지식재산권에 특별한 관여를 하지 않으나, 한국화합물은행의 라이브러리를 활용한 연구 결과를 논문, 특허 등에 발표 또는 공개할 때는 사사 또는 문구를 기재하도록 하고 있음

❍ (화합물 기탁 현황) 최근 5년간 화합물 기탁 내용은 연구기관 연간 평균 25,855.2건, 학교는 44.6건, 산업체는 22.4건, 해외는 0건으로 화합물 은행의 전체 기탁의 대다수가 연구기관에 집중되어 있음

그림 12. 화합물 연도별 누적 및 기관 현황



그림 13. 기관별 화합물 기탁 현황

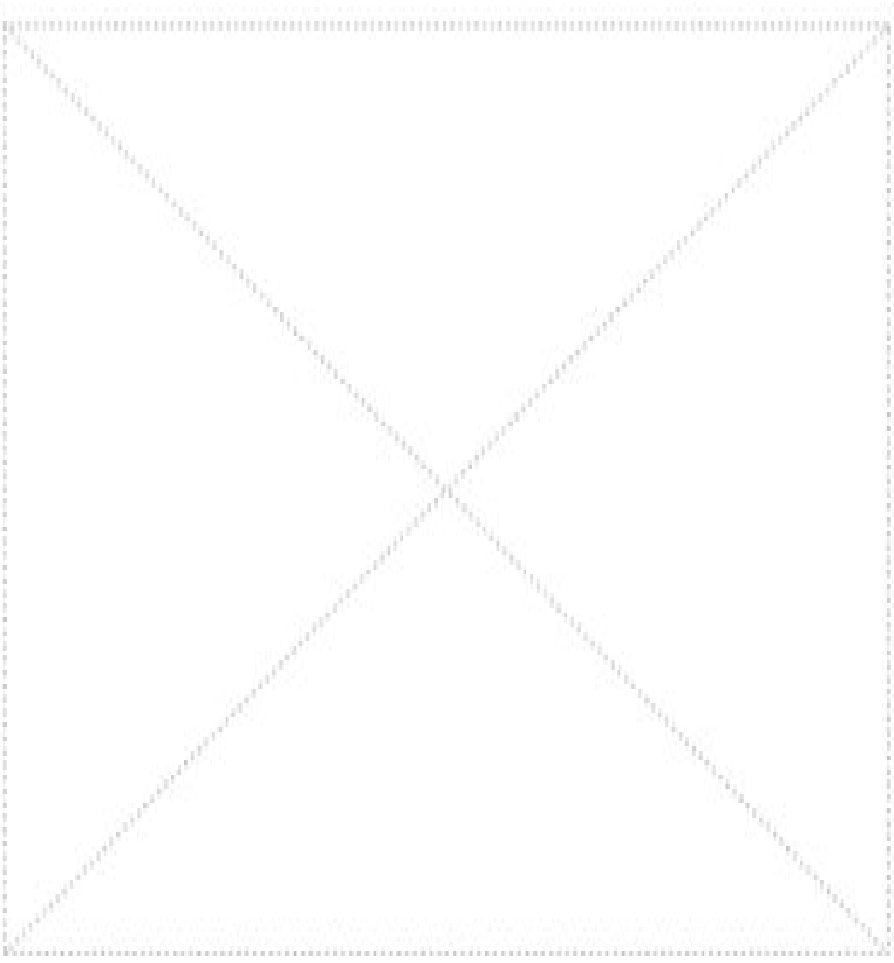


그림 14. 화합물 은행 활용 절차

❍ (시사점) 대량의 화합물 데이터 보유로 활용도가 높을 것으로 예상했으나, 데이터 활용에 요구되는 절차의 복잡성이 활용성 저해 요인이며 접근 폐쇄성으로 데이터 기탁이 필수인 연구기관 이외는 기탁 수가 적다는 한계 존재

❑ (국외) Protein Data Bank, PDB

❍ (개요) 실험적으로 찾아낸 단백질 및 핵산의 3차원 구조를 수록한 데이터베이스로 표준화된 포맷 수립으로 일정한 형태의 데이터가 쌓이고 있으며, 2023년 기준 단백질 구조 200,069개 보유

– 논문 투고 시, 단백질 구조정보를 PDB에 등록해야 하고 논문에 PDB 코드 기재



그림 15. PDB 연간 데이터 기탁 수

❍ (데이터 활용 방식) 별도 가입 절차 없이 바로 웹에서 PDB code를 검색하여 내려받기, ftp 제공, PDB code로 내려받기 링크 확인이 가능

– 단백질 구조 및 기능 분석, 단백질 구조 예측 모델 개발, 단백질 구조 기반 신약개발 등에 활용하고 있음

❍ (활용 결과 권리) CC0 1.0 Universal Public Domain Dedication 라이선스에 따라 PDB에 기탁한 단백질 데이터는 저작권이 없어 PDB에 포함된 데이터는 상업적 목적으로 허가 없이 복사, 수정, 배포가 가능

– 저작권이 없어 전면적인 이용이 가능하지만, PDB가 학술 목적으로 이용되면 DOI 형식으로 학술 발표 내용에 인용하도록 하고 있음

❍ (데이터 기탁 현황) 전 세계 사람들이 접근할 수 있는 단백질에 대한 3D 구조 무료 데이터 구축을 지향하며, 기탁 출처가 명시돼있지 않으나, 최근 5년간 연간 평균 9,229.6건의 기탁 수를 달성했고 성장하는 모습을 보임

❍ (시사점) 데이터 관리 및 포맷이 일괄적이며, 복잡한 사용 절차가 없어 높은 접근성 및 활용성을 갖고 있어 구조 기반 신약개발 연구자들이 필수로 이용하는 데이터로 자리매김하고 있음

❑ 추가 공개 데이터베이스 및 데이터

❍ 아래의 표는 신약개발에 활용되는 공개 데이터베이스를 나열한 자료임

표 13. 신약개발 공개 데이터베이스 및 데이터 목록

| 데이터베이스/데이터 | 분류 | 개요 |
|---|---|---|
| ChEMBL | 화합물<br>약리학 | 생리활성 약물 유사 소분자 데이터베이스로 2차원 구조, 계산된 속성, 추상화된 생체 활성을 포함 (240만 개 이상 화합물) |
| Zinc | 화합물 | 가상 스크리닝에 이상적인 상용 화합물의 무료 DB로 하위집합, drug-like 등을 검색할 수 있음 |
| PubChem | 화합물 | NCBI의 Entrez 정보 검색 시스템 내에서 세 개의 DB를 연결했고, 화합물 유사성 검색, 화합물 문헌에 대한 링크, 물리화학 특성 포함 |
| BindingDB | 약리학 | 6,263개의 단백질 표적과 378,980개의 소분자에 대한 910,836개의 결합 데이터를 포함하는 소분자의 결합 친화성 DB |
| eMolecules | 화합물 | ChemSketch 시각화 패키지로 화합물 구조를 그리고 140개 이상의 화합물 공급업체에서 800만 개 이상의 고유한 화합물 구조를 검색할 수 있음(DrugBank, National Cancer Institute, NIST WebBook, PubChem, EPA 조회) |
| ChemSpider | 화합물 | 수백 만개의 화학구조에 대한 접근 및 다양한 온라인 서비스를 통합 제공하며, 구조 기반 화학 정보의 가장 풍부한 DB로 특성 예측 서비스, 웹 기반 탐색 편의 기능을 제공함 |
| PDB-Bind | 약리학 | 알려진 3차원 구조를 가진 단백질-리간드 복합체에 대한 결합 친화도 DB로 5,671개의 단백질-리간드 복합체를 포함함 |
| DrugBank | 약리학 | 상세 약물(화학적, 약리학적, 약학적) 데이터와 포괄적인 약물 표적(서열, 구조, 경로) 정보를 결합한 DB로 1,448개의 FDA 승인 소분자 약물, 131개의 FDA 승인 단백질/펩타이드 약물, 85개의 약효 식품과 5,080개의 실험 약물 등 총 6,712개의 약물 포함 |
| KEGG | 생물학 | genomic, chemical, systemic 기능 정보를 통합한 DB |
| ChemDB | 화합물 | 500만 화학물질의 3차원 구조, 용융 온도 및 용해도와 같은 예측되거나 실험적으로 결정된 물리화학적 특성을 포함 |
| MMsINC database | 화합물 | 가상 스크리닝 및 화학 정보학 응용 프로그램을 위해 상업적으로 이용할 수 있는 화합물의 무료 웹 기반 DB로 3D 형식의 400만 개 이상의 비 중복 화합물을 보유 |
| DSSTox | 화합물<br>약리학 | 875,000개 이상의 화합물 구조, 텍스트 및 속성 정보를 포함하는 SDF 파일 형식 화합물 독성 정보 DB |
| DisGeNET | 생물학 | 16,000개 이상의 유전자와 13,000개 이상의 질병 간의 380,000개 이상의 연관성(GDA) DB |
| Reframe.db | 화합물 | 약물 용도 변경을 용이성을 제공하기 위해 두 개의 데이터베이스를 결합한 12,000개 분자 스크리닝 DB |
| KLIFS | 화합물 | 촉매적 키나아제 도메인의 실험 구조와 키나아제 억제제의 상호작용을 분석하는 키나아제 DB |

| | | |
|---|---|---|
| Enamine REAL Database | 화합물 | '5의 규칙' 및 Veber 기준을 준수하는 45억 개 이상의 실제 합성할 수 있는 가장 큰 분자 데이터베이스로 SMILES, SDF 포맷으로 데이터를 제공 |
| ChemSpace | 화합물 | 16억 개 이상의 빌딩 블록, 스크리닝 화합물, 시약 및 중간체 DB |
| OTAVA | 화합물 | 600만 개 이상의 스크리닝 화합물 컬렉션을 제공하는 DB |
| Uniprot | 생물학 | 56만 개 이상의 단백질 서열과 기능, 도메인 정보를 제공하는 DB |
| GEO | 생물학 | 미국 NCBI에서 운영하는 기능 유전체학 데이터베이스로 180만 건 이상의 마이크로 어레이, 차세대 염기서열 분석 데이터를 저장 및 제공하는 DB |
| NCI 60 | 생물학<br>약리학 | 3,000개의 화합물에 대해서 60개의 서로 다른 인간 종양 세포주에 대한 성장 억제 시험 결과를 제공하는 DB |
| CCLE | 생물학<br>약리학 | 1,000종 이상의 인간 종양 세포주에 대한 멀티 오믹스(유전체, 전사체, micro RNA, 단백체 등) 데이터와 74만 건 세포주-화합물 저해 농도 시험 결과를 제공하는 DB |
| GDSC | 생물학<br>약리학 | 1,000종 이상의 암세포 주에 대한 멀티 오믹스 데이터(유전체, 전사체, 단백체 등)와 57만 건 세포주-화합물 저해 농도 시험 결과를 제공하는 DB |
| TCGA | 임상<br>생물학 | 33종의 암에 대하여 암 환자 20,000여 명의 임상 데이터와 멀티 오믹스 데이터(유전체, 전사체, 단백체, miRNA, 변이 등)를 제공하는 DB |
| CMAP/L1000 | 생물학<br>약리학 | 50종의 암 세포주에 대한 978개의 Landmark Gene의 전사체 발현 프로파일 1,400,000개의 데이터를 제공하는 DB |
| BFD | 생물학 | 25억 개의 단백질 서열 및 마르코프 모델 및 MSA로 표시되는 6,500만 개의 단백질 패밀리로 구성되는 데이터를 제공하는 DB |
| Moleculenet | 화합물<br>약리학 | 분자 특성의 기계학습 방법을 테스트하기 위해 70만 개의 화합물에 대한 화학(양자역학, 물리화학), 약리학(생물물리학, 생리학) 시험 DB |
| Pfam | 생물학 | Hidden Markov model을 이용하여 단백질의 패밀리의 다중서열 정렬을 모아놓은 DB(총 18,259개의 패밀리와 635개의 클랜을 구축) |
| Gene Ontology | 생물학 | 150만 개의 Gene Product의 Molecular Function, Cellular Component, Biological Process에 대한 주석 DB |
| STRING | 생물학 | 단백질 간 200억 건 이상의 물리적 상호작용 및 기능적, 간접적 상호작용에 대한 정보를 모아놓은 DB |
| Drug Bank | 화합물<br>약리학 | 50만 개 이상의 의약품에 대한 구조정보, 약리학적 정보, 상호작용 정보, 임상시험 정보 등을 모아놓은 DB |
| KLIFS | 화합물<br>약리학 | 318개의 인산화효소(Kinase)에 대한 구조, 모노머, 리간드, 시험 데이터를 모아놓은 DB |
| GTEX | 생물학 | 건강한 대상자 948명의 54개의 조직에 대한 17,382개의 샘플의 유전체, 전사체 데이터를 모아놓은 DB |
| ConnectivityMap | 화합물<br>생물학 | 1,309가지의 생물체 작용 저분자 (bioactive small molecule)를 처리한 human cell을 배양하여 얻은 유전체 수준의 전사 발현 데이터를 모아 놓은 DB |

| | | |
|---|---|---|
| DrugCentral | 생물학<br>약리학 | 활성 성분 화학 물질, 의약품, 약물 작용 방식, 적응증, 약리작용에 대한 정보를 제공 |
| CDM, EMR | 의료<br>임상 | Sentinel CDM(약물감시), OMOP CDM(임상 연구방법론의 적용 및 평가), PCORnet CDM(환자 중심 임상 연구 네트워크) 등에 활용되는 공통 데이터 모델 데이터,<br>전자의무기록(EMR) : 환자의 진료 정보 데이터 |
| HIRA | 의료 | 건강보험심사평가원의 보험 청구 자료 빅데이터 |
| GDC | 생물학<br>임상 | 미국 국립 암 연구소의 연구 프로그램으로 데이터베이스(TCGA, TARGET)를 포함하는 데이터 포털, 종양 변이 정보, 데이터 통합 정보제공, 데이터는 자유로운 이용 'open' 유형과 허가 필요 'controlled' 유형으로 나뉨 |

- ❍ (데이터 활용 방식) 국외 대부분의 화합물 데이터베이스는 사용자에 대한 제약이 주어지지 않고, 누구나 데이터를 내려받을 수 있음
- ❍ (활용 결과 권리) 데이터의 출처를 인용할 것을 권장하지만 법적으로 요구하는 사항은 아님

❑ (시사점) 국내 화합물 데이터베이스에서는 데이터 활용을 위해 국가 과학기술 지식정보서비스(NTIS) 연계가 필요하고, 이후에도 활용 신청 프로세스에서 관련된 여러 가지 정보를 요구하는 반면 국외에서는 대부분 단순 다운로드만으로도 데이터를 확인할 수 있어 접근성에 큰 차이를 보임

## 2.4. 국내 신약개발 데이터 구축·표준화·활용사업 현황 분석

### 2.4.1. 국내 신약개발 데이터 사업 수행 현황 분석

- ❍ 국내 신약개발 데이터 공유 플랫폼 사업 이외에도 이와 관련된 데이터 구축, 표준화, 활용사업을 사업개요와 한계를 아래의 표에 정리함

표 14. 정부 주도 데이터 수집 구축 · 표준화 · 활용사업의 사업개요

| 사업명 | 주관 | 금액<br>(백만원) | 사업 내용[8] | 기간 | 구분 |
|---|---|---|---|---|---|
| **보건의료 빅데이터 플랫폼 구축** | 복지부 | 5,953 | 국민건강보험공단, 건강보험심사평가원, 국립암센터, 질병관리청 4개 공공기관에서 보유한 보건의료 빅데이터 연계, 분석, 활용이 가능한 정보 시스템 구축 | ('18~'21) | 수집<br>구축 |
| **바이오 빅데이터 구축 시범** | 과기부<br>복지부<br>산업부 | 34,551 | 자발적 참여자로부터 데이터의 수집, 생산 및 활용을 체계 통해 바이오·의료 정보 빅데이터 시범적 구축 | ('21-'22) | 수집<br>구축 |
| **국가 암 빅데이터 구축** | 복지부 | 2,453 | 다양한 암 정보 통합, 연구목적 DB화 축적·연계·가공·분석 체계화 | ('21~'25) | 수집<br>구축 |

| | | | | | |
|---|---|---|---|---|---|
| **의료데이터 중심병원 지원** | 복지부 | 9,402 (‘21년도) | 난치성 질환 등의 다중 오믹스 정보 생산 및 분석, 바이오마커 발굴·검증 등을 통해 정밀 의료실용화 기반 마련 | (’20~‘25) | 수집 구축 |
| **암 빅데이터 플랫폼 구축** | 과기부 | 4,417 | 암 임상 데이터의 수집-연계-관리 및 제공하는 암 빅데이터 플랫폼 및 센터별 표준화된 암종별 빅데이터 구축 | (‘20~’21) | 수집 구축 |
| **혁신신약 파이프라인 발굴** | 과기부 | 16,800 | 대학·연구소 등을 대상으로 신약개발 초기 단계를 지원하여 항암제, 당뇨치료제 등 기업에 기술이전이 가능한 유망 후보물질 발굴 | (‘19~’20) | 수집 구축, 활용 |
| **인공지능 신약개발 플랫폼 구축** | 과기부 복지부 | 24,833 | 데이터 확보 및 표준화, 후보물질 발굴, 스마트 약물감시 등 신약개발 단계별 인공지능 플랫폼을 개발하고 신약개발에 적용 | (‘19~’21) | 수집 구축, 활용 |
| **국가 항암 신약개발** | 복지부 | 51,937 | 국산 항암신약 후보물질 개발, 항암제 개발을 가속화하고 국산 글로벌 항암 신약개발 촉진에 기여 | (‘17~’21) | 수집 구축, 활용 |
| **의료기관 진료정보 교류 기반 구축 및 활성화** | 복지부 | 2,303 (‘21년도) | 진료기록을 표준화하여 환자 진료기록 등을 의료기관 간 교류할 수 있도록 제도·기술적 인프라 조성 | (‘20~’20) | 표준화, 정보보호 |
| **의료 데이터 보호· 활용 기술개발** | 복지부 | 31,027 | 보건의료 빅데이터 연계 · 활용을 위한 정책개선 연구 및 정보보호 기술 연구 등 | (‘19~’23) | 정보보호 |
| **국가 보건의료 표준화** | 복지부 | 1,623 (‘21년도) | 보건의료 정보표준(용어, 서식, 기술) 마련 및 인센티브 연계, 헬스케어 빅데이터 활용 중장기 로드맵 수립 등 | (‘21~’25) | 표준화 |
| **다부처 국가 생명 연구자원 선진화** | 과기부 | 23,562 | 부처·사업·연구자별 흩어져 있는 데이터를 통합한 ‘국가 바이오 데이터 스테이션’을 구축, 데이터 기반 바이오 연구 환경 조성(데이터 등록 표준(안), 등록 표준지침 마련) | (‘20~’22) | 수집 구축, 표준화 |
| **AI 정밀 의료솔루션 (닥터앤서2.0)** | 과기부 | 24,697 | AI 정밀 의료솔루션(닥터앤서2.0)은 폐암, 간질환 등 12개 질환의 진단 보조를 지원하는 AI SW를 개발 | (‘21~’23) | 활용 |
| **오믹스 기반 정밀 의료 기술개발** | 과기부 | 28,167 | 오믹스 데이터를 통합적으로 발굴, 분석하여 질환별 바이오마커 및 신약 표적 도출 | (‘19~’23) | 활용 |
| **연구 중심 병원육성(R&D)** | 복지부 | 322,544 | 산·학·연·병 협력 및 R&D 비즈니스 모델 개발 지원, 연구 역량 확보 | (‘14~’23) | 활용 |
| **실사용 데이터** | 복지부 | 5,000 | 전향적 실사용 데이터 레지스트리 구축 | (‘22~’24 | 활용 |

| | | | | | |
|---|---|---|---|---|---|
| **기반의 임상 근거 창출 지원** | | (‘21년도 ) | 및 이를 활용한 임상 근거 창출 연구 지원 | ) | |
| **전자약 기술개발** | 과기부 | 6,300 (‘21년도 ) | 전자약 핵심 원천기술 신규 개발 및 생체 적용 가능성 검증, 핵심 기술 고도화 및 성능 향상 | (‘22~’24 ) | 활용 |
| **국가 신약개발** | 과기부 복지부 산업부 | 98,251 | 국가 신약개발 사업의 공백 영역인 표적 발굴·검증단계 지원과 CAR-T[9], PROTAC[10] 등 혁신적 기작을 이용한 기술, 인공지능 신약개발 플랫폼 고도화 등 연구 혁신을 지원 | (‘21~’23 ) | 활용 |
| **범부처 전주기 신약개발** | 과기부 복지부 산업부 | 231,238 | 약물 개발 글로벌 신약개발 프로젝트 R&D 사업, 포트폴리오의 글로벌 라이센싱 지원 및 산업계 확산을 위한 R&D 사업화 지원, 국내 신약개발 글로벌화 및 선진화를 촉진하기 위한 네트워크 구축, 성과관리 시스템 도입을 통한 성과 중심 운영체제 운용 | (‘12~’22 ) | 활용 |
| **인공지능 활용 혁신 신약 발굴** | 과기부 | 8,453 | 인공지능 신약개발 플랫폼을 고도화하고, 신약 후보 물질을 도출함으로써 인공지능 활용 신약개발의 가시적 성과 창출 | (‘22~’23 ) | 활용 |

❍ 추가로 데이터 구축사업에서 운영 중인 사이트 링크와 제공 정보, 접근 가능 사용자를 조사하여 플랫폼을 활용성을 조사했음(22.08.10일 자 기준)

표 15. 정부 주도 데이터 구축사업과 활용 가능 데이터

| 사업명 | 사업 기간 | 주관기관 | 웹사이트 | 제공 정보 | 접근 가능 사용자 |
|---|---|---|---|---|---|
| **보건의료 빅데이터 플랫폼** | ’15~‘21 | 복지부 보건의료 데이터 진흥과 | 보건의료 빅데이터 플랫폼 | 국민건강영양조사, 검역, 결핵 환자 신고현황, KoGES 기반 통합자료, 예방접종, 자격 및 보험료, 건강검진 대상자 및 문진, 요양기관, 사망 정보, 암 DB, 명세서 및 진료 내역 등 | IRB 승인 연구자 |
| **국가 암** | ‘21~’25 | 국립암센터(병원 ) | 국가 암 데이터센터 | 임상 데이터 유전체, 영상, 공공 | 국민 전체 |

8) 국회예산정책처, 보건의료 데이터 재정사업 분석, 2021.01.08
9) CAR(chimeric antigen receptor)-T :암세포만 표적하여 공격하는 항암 면역세포 치료 기술
10) PROTAC(proteolysis targeting chimera) : 질병 유발 특정 단백질을 제거하는 신개념 플랫폼 기술

| | | | | | |
|---|---|---|---|---|---|
| **빅데이터 구축** | | | | 데이터(준비 중) | |
| **의료 데이터 중심병원 지원사업** | '19~24 | 서울아산병원 삼성서울병원 한림대성심병원 등 (병원) | 의료데이터 중심병원 데이터 포털 | 질환별 특화 임상 데이터 (7개 컨소시엄) | IRB 승인 연구자 |
| **암 빅데이터 플랫폼 구축** | '19~'21 | 국립암센터 (병원) | 국립 암센터 CONNECT | 유방, 갑상선, 난소, 폐, 대장, 신장, 위, 간, 전립선, 췌장, 담도 등 암 데이터 | 의료진, IRB 승인 연구자, 국민 전체 |
| **국가 바이오 빅데이터 구축 시범사업** | '20-'21 | 국립중앙인체자원은행 국가생명연구자원정보센터 한국과학기술정보연구원 | 국가통합 바이오 빅데이터 구축사업 | 희귀 질환, 자폐 스펙트럼 장애, 암, 일반인 등의 임상 정보, 전장 유전체 분석 정보 | IRB 승인 연구자 |
| **인공지능 활용 혁신 신약 발굴 사업** | '22~'22 | 아론티어(기업) 이화여자대학교(대학) 심플렉스(기업) 대구·경북 첨단의료 산업진흥재단 | 인공지능 신약개발플랫폼(KAIDD) | - | - |

❑ 시사점

❍ 정부 주도의 데이터 사업은 대부분의 다양한 목적에 활용가능한 공공 데이터 구축을 목표로 수행되었으며 AI에 활용하기 위해서는 활용목적에 부합하도록 전처리 또는 정제 과정이 추가로 요구됨

- 사용자 접근제한과 절차 복잡성은 공유 활성화 저해 요인으로 판단됨

## 2.5. 데이터 공유 규제 현황 분석

### 2.5.1. 민감·비민감 데이터의 정의

- 신약개발 데이터 중 **개인정보를 포함한 경우를 민감 데이터**, **포함하지 않은 경우를 비민감 데이터**라고 정의했으며 개인정보 대신 **지식재산권**(Intellectual Property Rights)이 **개입**되면 **민간 데이터**라고 함
- 데이터 취득·관리 주체로 구분했을 때 공공기관이 생성·취득해 관리하는 자료와 정보는 **공공 데이터**, **기업·개인 등이 생성·취득해 관리하는 데이터는 민간 데이터**로 정의되며, 민간 데이터의 데이터 소유권은 기업이 갖는 지식재산권(Intellectual Property Rights)이 개입됨
- AI 신약개발에 널리 활용되는 화합물 데이터 관점에서 지식재산권을 보호하며 AI 활용을 위해서는 제도적 기술적 보호 방법의 검토가 필요

### 2.5.2. 분석 개요

- 신약개발 데이터 중 개인정보가 포함된 **민감 데이터**와 지식재산권이 개입된 **민간 데이터**의 AI에 활용을 위한 제도·기술적 보호 방법을 검토했음
- 데이터라는 추상적 객체의 물질적 소유권을 보장하기 어렵기에 현시점에 데이터를 AI에 활용하기 위해서는 법적, 제도적 방법의 마련에는 시간이 필요하며, 즉시 활용가능한 기술적 보호 방법에 대해 주로 다루게 되었음
- 규제 환경 분석에서는 민감 또는 민간 데이터의 보호와 공유를 위해 국내외에서 제도적 장치에는 무엇이 있는지, 규제 수준은 어떻게 다른지 조사분석해 국내의 제도적 한계를 도출했음
- 게다가, 민감 데이터를 보호하는 규제와 데이터를 공유할 수 있게 하는 비식별 처리 가이드라인을 조사하여 국내와 국외의 데이터 공유의 규제 검토 및 공유 기술을 조사했음
- 조사 분석한 대상은 규제에서 허락하는 데이터 공유 기술인 비식별 처리, 가명 처리 기술로 기술에 대한 명확한 이해를 통해 기술의 장단점 그리고 한계를 파악해 개선안을 마련하고자 했음

### 2.5.3. 데이터 지식재산권 보호의 제도적 현황과 방안[11)]

❑ 현시대의 원유라고도 불리는 데이터는 4차 산업혁명의 핵심

- ❍ 시가 총액 가치의 상위를 차지하는 페이스북, 아마존, 애플, 넷플릭스, 구글 이른바 FAANG이라 불리는 빅테크 기업들이 모두 데이터 플랫폼을 기반으로 한다는 것은 데이터가 얼마나 중요한지를 나타냄
- ❍ 충분한 데이터를 확보하지 못하면 정교한 알고리즘이 무용지물이며, 반대로 충분한 데이터가 있어야 정교한 알고리즘이 만들어지기도 함
- ❍ 데이터의 중요성이 대두되고 있는 시기에 데이터에 대한 법적 권리 해석을 명확히 할 필요가 있음
- ❍ 하지만, 여전히 데이터 자체 또는 가공 처리된 데이터 세트에 대한 법적 보호 가능성에는 합의에 도달하지 못하였음
- ❍ 기존 법체계 하에 데이터의 소유권(물권적 권리)을 인정하는 것이 비 마모성, 비 배타성, 실시간성 등 데이터의 특성으로 한계가 있다는 의견이 전 세계적으로 지배적임

❑ 법적 성질에 따른 데이터 유형의 구분과 보호 내용

- ❍ 데이터가 물건의 객체에 해당하지 않아, 데이터에 대해서 전통적인「민법」상 객체에 대한 소유권[12)]이 인정되기 어려우며, 지식재산권법 또는 계약법상 보호 대상만 가능함
- ❍ 현행법상 데이터의 법적 지위는 아래의 표와 같이 다양한데, 화합물에 대한 화학 데이터는 창작성 있는 데이터베이스, 영업비밀인 데이터, 데이터 관련 특허 3가지에 적용될 수 있음

표 16. 현행법상 데이터의 법적 지위

| 구분 | 보호 대상 | 보호 내용 | 관련 제도 |
|---|---|---|---|
| 데이터 보호 | 데이터베이스 | • 데이터베이스 제작자의 복제·배포·방송 또는 전송권(독점·배타적 권리)<br>• 창작시로부터 권리가 발생하여 다음 해부터 가산하여 5년간 보호<br>• 침해 정지·예방, 손해배상청구권<br>• 침해 시 형사처벌(3년/3천만 원, 병과 가능 | 「저작권법」상 데이터베이스 제작자의 권 |

11) 한국법제연구원, "2020년 데이터 지식재산권 보호 방안연구", 2020.12

12) 소유권은 자신의 물건을 직접적·배타적·전면적으로 지배하여 사용·수익·처분할 수 있는 사법상의 권리로서, 물권에서의 기본적인 권리이며 소유권의 객체가 되는 물건에 대한 전면적인 지배 권리로서, 그 객체는 물건에 한한다. 이때, 물건이란 유체물 및 전기 기타 관리할 수 있는 자연력을 말하며(민법 제98조), 구체적으로 공간 일부를 차지하고 지각될 수 있는 유형적 존재인 유체물, 음향·향기·열·빛·원자력·기타 에너지 등의 자연력 등이 이에 해당한다.

| | | | |
|---|---|---|---|
| | 창작성 있는 데이터베이스 | • 저작물로서 저작인격권 및 저작재산권<br>• 창작시로부터 권리가 발생하여 저자 사망 후 다음 해부터 가산하여 70년간 보호<br>• 침해 정지·예방, 손해배상청구권 및 명예회복 등의 청구권<br>• 침해 시 형사처벌(5년/5천만 원, 병과 가능) | 「저작권법」상 편집저작물 |
| | 영업비밀인 데이터 | • 영업비밀로서 보유자의 침해금지·예방, 손해 배상 청구권 및 신용 회복 청구권<br>• 침해 시 형사처벌(해외 유출 15년/15억 원, 국내 유출 10년/5억 원) | 「부정 경쟁 방지 및 영업비밀보호에 관한 법률」상 영업비밀 |
| 데이터 관련 간접 보호 | 데이터 관련 특허 | • 데이터 그 자체에 대한 보호 불가<br>• 특허권으로서 보호(보유자의 침해금지·예방, 손해배상 청구권 및 신용 회복 청구 | 「특허법」상 특허권 |
| | 데이터 관련 계약의 보호 | • 데이터 그 자체에 대한 보호 불가<br>• 당사자 간의 계약에 따른 채권의 보호 | 「민법」 채권편 |

❑ 데이터 용도에 따른 분류에 기반한 법적 보호 수단

❍ 화학 데이터를 AI 모델 학습에 활용하는 경우 소분류 기준 기타 AI 학습용 데이터, AI 학습·데이터 구축용 원천데이터에 속할 것

❍ 하지만, 라벨링, 분류체계의 기준이 없거나 모호하다면 보호 수단 미 존재

❍ 이외에도 데이터 소유권 확정을 위한 Ownership, Co-ownership 등 다양한 소유권 논의가 존재하나, 여전히 협의에 도달하지 못하고 있음

표 17. 현행법상 데이터의 법적 지위

| 대분류 | 소분류 | 특징 | 주요 보호 수단 |
|---|---|---|---|
| 분석용 | 통계 데이터 | • 설문 조사, 통계분석 등의 가공된 데이터 | 저작권, 데이터베이스 제작자의 권리 |
| | 현황 데이터 | • 정보통신망을 통해 수집된 각종 데이터로서, 정형화·가공된 경우와 그렇지 않은 경우로 구분 | 정형화된 데이터의 경우 데이터베이스 제작자의 권리를 검토할 수 있으나, **비정형화 데이터의 경우 적절한 보호 수단 부재** |
| AI 엔진 학습용 | 자연어 데이터 | • 텍스트 데이터를 품사 단위로 라벨링 | 저작권, 데이터베이스 제작자의 권리 |
| | 이미지·영상 데이터 | • 이미지·영상에 대한 라벨링 | |
| | 기타 AI 학습용 데이터 | • 분야별 데이터에 대한 라벨링 | |
| 원천데이터 | AI 학습·데이터 구축용 | • 인공지능 학습을 위한 분야별 원천데이터 | 개별 데이터에 대한 라벨링, 분류 체계화 |

| | | | |
|---|---|---|---|
| | 원천데이터 | | 가 |
| | 특정분야 지식베이스 | • 검색, 챗봇, 통합정보서비스 등에 활용할 수 있는 원천데이터 | 수행되었다면 데이터베이스 제작자의 권리를 검토할 수 있으나 **그렇지 않은 경우, 적절한 보호 수단 부재** (단, 원천데이터 자체에 포함된 저작권은 고려하지 않음) |
| | 온톨로지 모델 | • (온톨로지 모델) 데이터 정의, 클래스·인스턴스 생성 및 의미관계에 따른 모델링 | 저작권, 데이터베이스 제작자의 권리 |

❑ 산업의 디지털 전환 및 지능화 촉진에 관한 법률안(국내 법안 발의)

❍ 2016년 발의되었던 「빅데이터의 이용 및 산업진흥 등에 관한 법률안」은 구체적인 데이터 보호 규정을 마련하지 못하였다는 한계로 보호 방안으로 활용하기 어려움

❍ 「산업의 디지털 전환 및 지능화 촉진에 관한 법안」은 보다 적극적으로 데이터 산업을 육성하고 보호하기 위한 규정을 마련하였다는 의의가 있음

❍ 해당 법안에서 산업 데이터 정의와 산업데이터 생성 및 활용의 개념을 정의하고, 관련 산업진흥을 위한 추진 체계 구축은 물론 구체적인 권리관계, 보호 조항 및 데이터 플랫폼 설립의 근거를 규정하고 있다는 점에서 한걸음 앞선 입법 시도임

❍ 산업 데이터의 폭넓은 정의, 생성자의 영업상 이익을 보호하도록 규정하고 있다는 점, 산업 데이터 권리 등을 정의하고 있음

❍ 국내에서는 20년 12월 8일 데이터 기본법안에서도 유사하게 정의하고 있으며, 향후 지식재산권의 보호의 방향이 소유권을 정의, 경제적 가치를 산정 및 보호하는 등으로 변화할 가능성이 높음을 시사

표 18. 산업의 디지털 전환 및 지능화 촉진에 관한 법안

제2조(정의) 이 법에서 사용하는 용어의 뜻은 다음과 같다.

- 1. “산업 데이터”란 「산업발전법」 제2조에 따른 산업, 「광업법」 제3조 제2호에 따른 광업, 「에너지법」 제2조 제1호에 따른 에너지와 관련한 산업과 「신에너지 및 재생 에너지 개발·이용·보급 촉진법」 제2조 제1호 및 제2호에 따른 신에너지 및 재생 에너지와 관련한 산업의 활동 과정(이하 “산업 활동”이라 한다)에서 생성 또는 활용되는 것으로서 광(光) 또는 전자적 방식으로 처리될 수 있는 모든 종류의 자료 또는 정보를 말한다.
- 2. “산업 데이터 생성”이란 산업 활동 과정에서 인적 또는 물적으로 상당한 투자와 노력을 통하여 기존에 존재하지 아니하였던 산업 데이터가 새롭게 발생하는 것(공정한 상거래 관행이나 경쟁 질서에 반하지 않는 범위 내에서 산업 데이터 활용을 통하여 당 초 산업 데이터와 구분되어 독자성을 인정할 수 있는 산업 데이터가 발생하는 경우를 포함한다)을 말한다.

제9조(산업 데이터 활용 및 권리보호)

- ① 산업 데이터를 생성한 자는 해당 산업 데이터를 가공, 분석, 이용, 제공 등의 방법으로 활용하여 사용·수익할 권리를 가진다.
- ② 산업 데이터를 2인 이상이 공동으로 생성한 경우, 각자 해당 산업 데이터를 가공, 분석, 이용, 제공 등의 방법으로 활용하여 사용·수익할 권리를 가진다. 다만, 당사자 간의 약정이 있는 경우에는 그에 따른다.
- ③ 제1항 또는 제2항에 따라 생성된 산업 데이터가 제삼자에게 제공된 경우, 산업 데이터를 생성한 자와 제삼자 모두 각자 해당 산업 데이터를 가공, 분석, 이용, 제공 등의 방법으로 활용하여 사용·수익할 권리를 가진다. 다만, 당사자 간의 약정이 있는 경우에는 그에 따른다.
- ④ 누구든지 산업 데이터에 대한 제1항부터 제3항까지의 권리를 공정한 상거래 관행이나 경쟁 질서에 반하는 방법으로 침해하여서는 아니 된다. 이 경우 공정한 상거래 관행이나 경쟁 질서에 반하는 방법인지를 판단할 때는 산업 데이터 활용의 목적 및 성격, 산업 데이터의 활용이 그 산업 데이터의 현재 또는 잠재적 가치에 미치는 영향 등을 종합적으로 고려하여야 한다.
- ⑤ 산업 데이터 생성 또는 활용에 관여한 이해관계자들은 산업 데이터의 원활한 활용과 그 결과에 따른 이익의 합리적인 배분 등에 관한 사항을 내용으로 하는 계약을 체결하도록 노력하여야 한다. 이 경우 이해관계자들은 합리적인 이유 없이 그 지위를 이용하여 불공정한 계약을 강요하거나 부당한 이득을 취득하여서는 아니 된다.
- ⑥ 산업 데이터를 사용·수익할 권리를 가지는 자는 산업 데이터의 무결성·신뢰성을 확보하고 산업 데이터가 분실·도난·유출·위조·변조 또는 훼손되지 아니하며, 산업 데이터를 활용한 제품·서비스가 위해를 발생시키지 않도록 노력하여야 한다.
- ⑦ 고의 또는 과실에 의하여 제4항 및 제6항을 위반하여 타인에게 손해를 입힌 자는 손해를 배상할 책임을 진다.

❑ 화합물 활용 가이드북(한국화학연구원, 한국화합물은행)

❍ 화합물 은행의 제공 화합물 활용 결과에 대한 권리관계 규정을 분석했음

❍ 논문 특허에 대한 권리관계 규정에서 논문에 화합물 기탁자의 단순 기탁 이외에 추가적 기여가 없는 경우 기탁자로서 논문의 사사에 표기하는 것을 원칙으로 한다고 명시

❍ 특허에 대한 권리관계 규정에서 기탁자의 추가 기여가 없으며 기탁자가 **물질특허**가 있는 경우, 활용 결과에 대한 “용도특허”는 사용자(발견자)가 취득할 수 있으며, “물질특허”를 소지한 기탁자와의 사전 협의 권장 명시

❍ 한국화합물은행은 화합물 활용에 관한 지식재산권(intellectual property)에 관여하지 않는다고 명시

❑ 결론 및 시사점

❍ 현재 데이터의 소유권과 보호에 대한 법률적 방안은 협의 중이거나 협의가 필요한 상황으로 해당 **법적 근거는 현행법을 기준으로 판단**할 수밖에 없으며, **기준이 모호한 경우 법적 보호를 받지 못함**

❍ 따라서, 데이터를 활용하기 위해서는 데이터 보호 기술의 활용이 필요하며, 데이터 활용에 필요한 지원 사항으로 데이터 보호 기술 중 하나인 가

명 정보 처리 기술을 활용해야 함

❍ 가명 정보 처리 기술은 21년 11월 4차위 데이터특위서도 데이터 보호 기술에 대한 원천기술 확보를 강조한 바 있음[13)]

13) 박수형, “데이터 유통 자유롭게”...데이터 보호 기술 집중개발, https://zdnet.co.kr/view/?no=20211118131826

## 2.5.4. 데이터 지식재산권 보호에 관한 기술적 방법

❑ 비식별화, 가명화, 익명화

❍ (개인정보의 정의) 개인정보는 살아 있는 개인에 관한 정보로 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보(해당 정보만으로 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 것을 포함)라 정의됨

❍ (개인정보의 두 유형) 이름, 주민등록번호, 여권번호 등 그 정보만으로 개인을 식별할 수 있는 정보를 **식별정보**라고 하며 그 정보만으로는 직접 개인을 식별할 수 없으나 다른 정보와 결합하여 개인을 식별할 수 있는 정보를 **식별 가능 정보**라고 함

❍ 개인정보 보호법은 개인정보처리자가 이러한 개인정보를 처리하는 데 처리목적에 필요한 범위에서 최소한의 정보만을 수집, 처리해야 하고, 원칙적으로 정보 주체(개인정보 수집을 당하는 이)의 결정(사전 동의, 개인정보 이용 동의)에 기초하여 처리하여야 한다고 정하고 있음

❍ 개인으로부터 정보를 수집할 때 개인을 식별할 수 있는 개인정보를 수집하지 않는다면 개인정보 보호법의 적용을 받지 않음

❍ 수집할 당시에 개인정보였지만 가공을 거쳐 식별 가능성을 제거했을 때, 더는 법의 적용 제외 대상 정보로 **비식별 정보**라고 함

❍ 비식별 정보 이외에도 **익명 정보**가 있는데 익명 정보는 역시 식별 가능성이 제거된 정보로 법의 적용 제외 대상임

❍ 반면, **가명 정보**는 여전히 개인정보에 속하는 정보로 이름 기타 정보 주체를 직접 특정하는 식별자를 다른, 그 자체로 식별기능을 하지 못하는 기호 등으로 대체한 정보를 말하며 가명화 후에도 곧바로 개인정보가 아니라고 할 수 없음

❍ 비식별 처리, 가명화, 익명화와 같은 기법은 개인정보 보호법에서 주로 언급되고 있는 기술이며, 여기서 언급되지 않은 암호화는 가명화의 하위 범주에 속하는 기술로 원본 데이터 유출을 방지하는 기술적 보호 수단임

❑ 비식별화 (De-identification)

❍ 비식별화는 개인 신원이 데이터와 연결될 위험을 줄이기 위해 개인식별정보[14](Personally Identifiable Information, PII)나 지식재산권 정보의 전부 또는 일부 삭제하거나 대체하는 것을 말함

- 원본 데이터를 식별하기 위해 비식별화된 데이터를 역으로 사용하는 과

14) 개인식별정보(Personally Identifiable Information): 단독으로 또는 다른 변수와 조합하여 합리적인 확실성을 가지고 설문에 참여한 단일 개인을 식별하는 데 사용할 수 있는 변수로 개인의 이름, 주소, 생일, GPS 좌표, 연락처, 주민등록번호, 개인이나 주택의 사진, 성별, 민족, 등급, 급여, 직위 등을 포함

정을 데이터 재식별화(Data re-identification) 혹은 비 익명화(De-anonymization)라고 하는데, 익명 데이터에 속한 원본 데이터를 식별하기 위해 이미 공개된 정보나 보조 데이터와 일치시키는 과정을 말함

- 예) 인간 대상 연구 중 생성된 데이터는 연구 참가자의 개인정보를 보호하기 위해 비식별화될 수 있으며, 임상 및 유전체 데이터는 환자 개인정보 보호법을 정의하고 규정하는 HIPAA(Health Insurance Portability and Accountability Act) 규정을 준수하는 방법으로 비식별화(De-identification) 처리를 할 수 있음

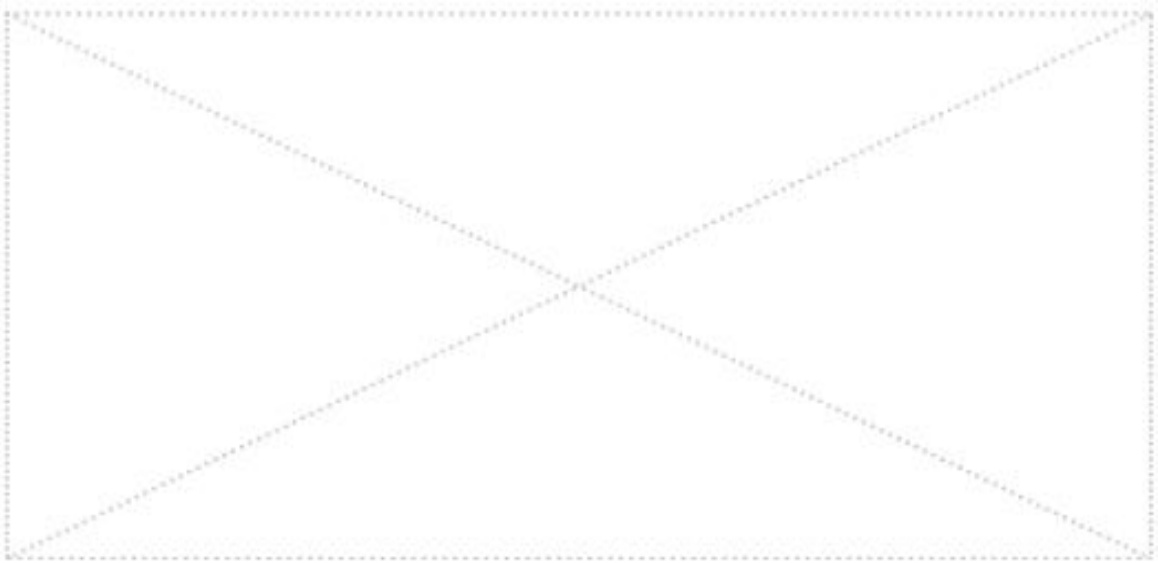


그림 16. 비식별화에 대한 정의(출처: 개인정보와 익명화, 이광춘, Tidyverse-Korea)

❑ 익명화(Anonymization)

❍ 비식별화 조치를 하는 것은 같으나 익명화 처리 이후에는 원본 데이터를 재식별할 수 없게 처리되며, 처리된 정보는 원본 데이터 재식별을 위해 다른 정보와 결합할 수 없어 개인정보와 지식재산 데이터 식별이 불가해져 제도 보호의 범주에 속하지 않음



그림 17. 데이터의 익명화

❍ 익명화 방법에는 데이터 마스킹, 합성 데이터, 데이터 스와핑, 데이터 교란, 가명화 방법이 있음

❍ **데이터 마스킹**(Data Masking)은 원본 데이터를 가짜 데이터 사이에 숨기는 방법으로 같은 데이터의 다른 버전이 무작위로 생성되고 보호 대상인 데이터의 원래 버전과 섞임

- 데이터 마스킹에는 다양한 유형과 기술이 있는데, 결정적 데이터 마스킹, 동적 데이터 마스킹, 즉시 마스킹, 정적 데이터 마스킹 방법이 있음

❍ **재현 데이터**(Synthetic Data)는 실제로 측정된 데이터(Real Data)를 생성하는 모형이 존재한다고 가정하고, 통계적 방법이나 기계학습 방법 등을 이용해 추정된 모형에서 새롭게 생성한 모의 데이터(Simulated Data)를 말함

- 개인의 프라이버시를 보호하면서 민감한 정보를 분석하고자 하는 연구자들에게 데이터를 제공할 수 있는 대안적 개인정보 비식별 조치 기법임
- 재현 데이터를 생성할 때 사용되는 알고리즘은 전통적 통계 및 베이지안 방법, 기계학습 모형 방법, 차등 정보보호에 의한 방법 등이 있음

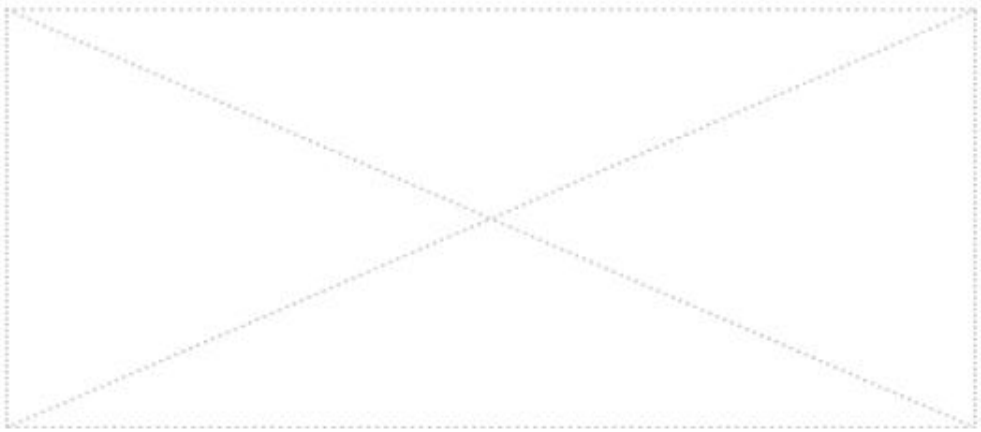


그림 18. 합성 의료 데이터의 생성 및 활용 과정

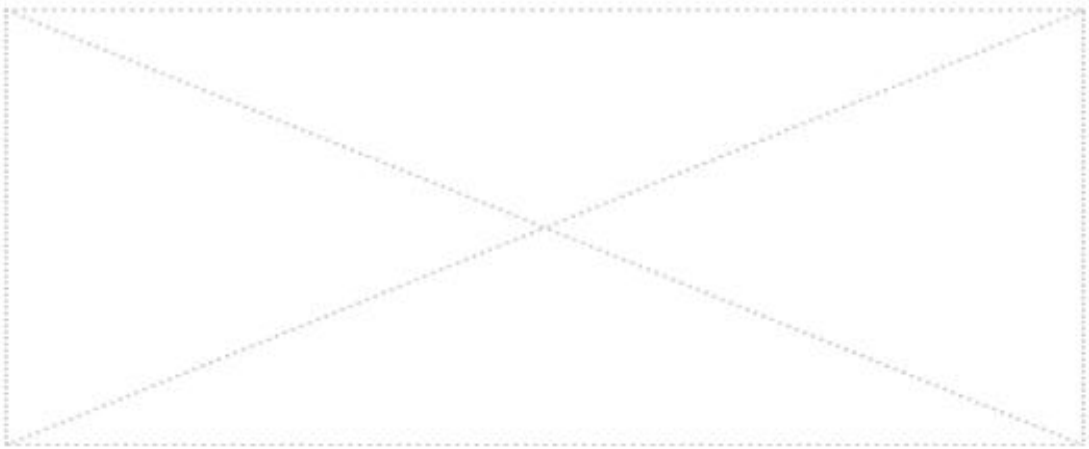


그림 19. 재현 데이터 생성 기법

– 예) AI 신약개발 기업 Insilico Medicine의 딥러닝 기반 재현 데이터 사례

| |
|---|
| AI 신약개발 기업 Insilico Medicine은 46일 이내에 새로운 신약 후보 물질의 설계, 합성, 검증이 가능한 GENTRL을 개발했음<br>GENTRL 시스템을 통해 생성기는 항암 속성을 가진 새로운 분자들을 만들고, 판별기는 기존 치료법을 기반으로 새로 만든 분자가 적절한지를 판별하여 항암 치료 후보물질 탐색의 속도와 성공률을 높였음<br>실험 결과, 7,200만 가지 화학물질에서 판별기를 통해 신약 후보 물질을 제시했고, 이 가운데 특허받은 항암제가 60가지 포함되었음 |

❍ **데이터 스와핑** (Data Swapping)은 데이터를 뒤섞어 재배열하는 방법으로 원본 데이터베이스와 결과 레코드 사이에 유사점이 없게 하는 방법임

– 이 방법은 불일치로 인해 공격자가 익명화를 해제하는데 매우 어려운 것으로 밝혀졌음

❍ **데이터 교란**(Data Perturbation) 방법은 수치형 데이터에 적용할 수 있는 방법으로 특정 값과의 연산으로 데이터를 변형하는 것을 말함

– 일례로 데이터의 모든 값에 8을 더해서 데이터의 원본을 유추하지 못하도록 하는 방법이 있음

❑ 가명화(Pseudonymization)

❍ 가명화는 데이터 레코드 내의 개인식별, 지식재산권 정보 필드를 하나 이상의 인공 식별자 또는 가명으로 대체하는 데이터 관리 및 비식별화 절차임

❍ 가명화된 정보는 추가정보를 활용하면 개인을 식별할 수 있기에 법의 보호를 받지만, 가명 처리라는 안전조치가 적용된 것으로 프라이버시 침해 위험을 감소시켜 일정한 범위에서 활용할 수 있음

❍ **스크램블**링(Scrambling)은 보유하고 있는 데이터를 섞는 방법으로 섞는

방법의 알고리즘을 알고 있다면 원본 데이터로 되돌릴 수 있음

❍ **암호화**(Encryption)는 원본 데이터를 알 수 없게 만드는 방식으로 올바른 암호 해독키를 가지고 있어야만 원본 데이터로 되돌릴 수 있는 방법으로 공개키 암호화 방식이 대표적임

- 공개키 암호화 방식은 암호화에 쓰이는 키와 복호화에 쓰이는 키가 다름
- 암호화에 쓰이는 키는 공개되어 공유하지만, 복호화할 때는 개인이 가지고 있는 키로만 복호화할 수 있도록 만들어 안전한 데이터 통신을 가능하게 하는 방식임
- 최근에는 정보보호와 데이터 분석이 모두 가능하게 할 수 있는 동형암호화 기술이 주목받고 있음

❍ **동형암호**(Homomorphic Encryption)는 데이터를 암호화된 상태에서 연산하는 방법으로, 암호화된 데이터를 이용한 연산의 결과는 새로운 암호문이 되며, 이를 복호화하여 얻은 원본 데이터는 암호화하기 전 원래 데이터의 연산 결과와 같음

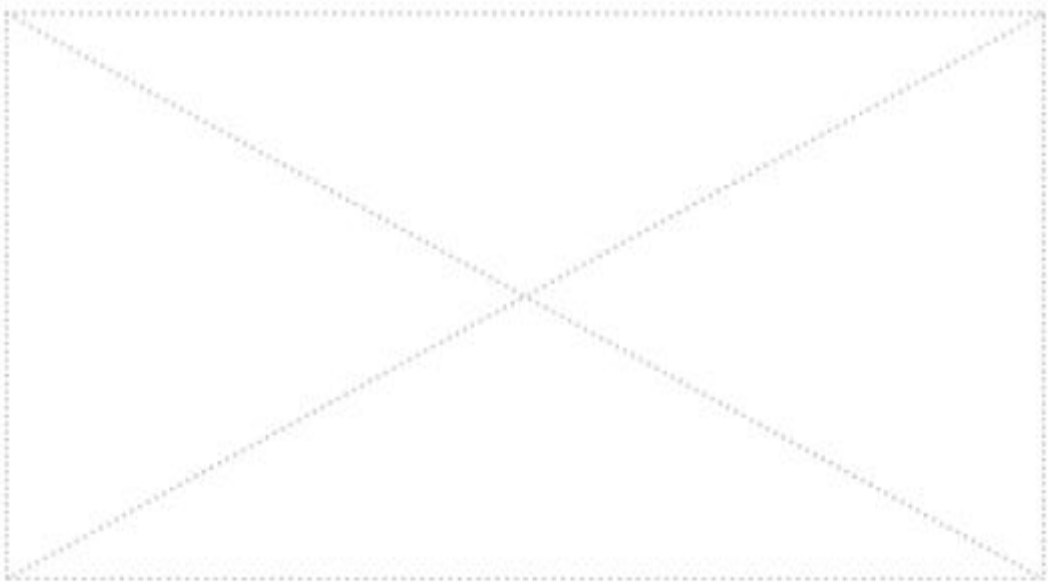


그림 20. 공개키 암호화 방법

- 평문(원본 데이터)에 일정한 암호화 방식(아래 그림에서는 나머지 연산)을 거쳐 변환된 암호문에 10은 (2, 3) 15는 (3, 1)로 암호화됨
- 이후 두 개를 더하면 (5, 4)로 표현되며 여기에 다시 암호화(나머지 연산)를 적용하면 (1, 4)로 변환됨
- 평문의 연산 결과를 같은 암호화 방법으로 암호화하면 (1, 4)로 표현되기에 암호문이 일치하므로 복호화할 수 있음
- 암호문에 원하는 연산 결과를 수행하여도 복호화하면 원본 데이터 연산과 같은 결과를 도출하기에 데이터 분석이나 기계학습, 딥러닝에 이용 가능

그림 21. 동형암호 예시

❍ 토큰화(Tokenization)는 민감한 데이터를 토큰이라고 하는 민감하지 않은 불투명한 값으로 대체하여 데이터를 보호하는 방식임

- 토큰에는 어떠한 의미나 값이 없고 데이터의 길이나 유형을 변경하지 않기에 나중에 길이 및 데이터 유형에 민감한 시스템에서 활용하기 좋음

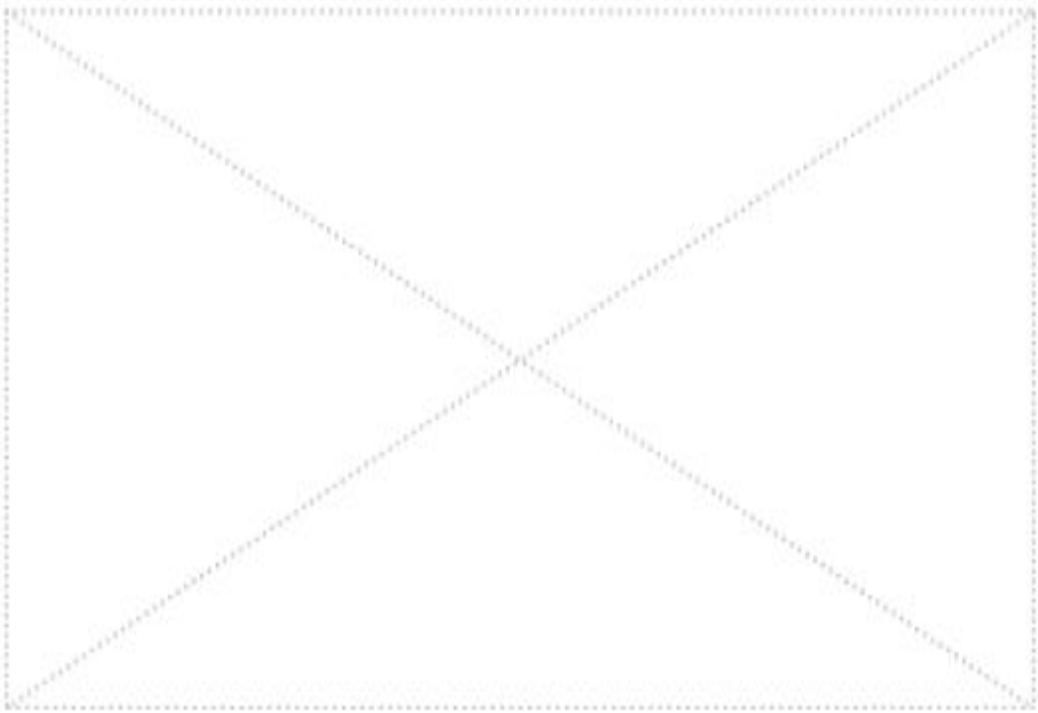


그림 22. 토큰화 기술 개념도

❍ 아래의 표에 비식별 처리 방법의 정의와 법적 보호, 주요 기법과 그 설명을 정리하였음

표 19. 비식별 처리 방법 요약

| 비식별 처리 방법 | 정의 | 법적 보호 | 주요 기법 | 설명 |
|---|---|---|---|---|
| 익명화 (Anonymization) | 개인식별 정보를 제거하는 것을 목표로 하여 익명화된 데이터는 재식별이 불가능함 | 법의 보호를 받지 못하나 자유롭게 활용 가능 | 데이터 마스킹 | 원본 데이터를 가짜 데이터 사이에 숨기는 것으로 같은 데이터의 다른 버전이 무작위로 생성되고 보호 대상인 데이터의 원래 버전과 섞임 |
| | | | 합성 데이터 | 실제로 측정된 데이터(Real Data)를 생성하는 모형이 존재한다고 가정하고, 통계적 방법이나 기계학습 방법 등을 이용해 추정된 모형에서 새롭게 생성한 모의 데이터(Simulated Data)를 말함 |
| | | | 데이터 스와핑 | 데이터를 뒤섞어 재배열하는 방법으로 원본 데이터베이스와 결과 레코드 사이에 유사점이 없게 하는 방법 |
| 가명화 (Pseudonymization) | 개인 식별 정보를 다른 정보로 대체하여 개인을 재식별할 수 있는 정보로 데이터의 분석에 활용할 수 있도록 처리함 | 법의 보호를 받음 | 스크램블링 | 데이터를 섞는 방법으로 섞는 방법의 알고리즘을 알고 있다면 원본 데이터로 되돌릴 수 있음 |
| | | | 암호화 (동형암호) | 데이터를 암호화된 상태에서 연산하는 방법으로, 암호화된 데이터를 이용한 연산의 결과는 새로운 암호문이 되며, 이를 복호화하여 얻은 원본 데이터는 암호화하기 전 원본 데이터 연산 결과와 같음 |
| | | | 토큰화 | 민감한 데이터를 토큰이라고 하는 민감하지 않은 불투명한 값으로 대체하여 데이터를 보호하는 방식 |

## 2.5.5. 국내외 법률과 비식별화 가이드라인 현황 분석

- ❑ 전 세계적으로 다양한 비식별화 가이드라인이 발표되었지만 국가별 비식별화 데이터의 정의와 비식별 조치 기준에는 차이가 있음
- ❑ 유럽연합 (일반 개인정보보호 규정 General Data Protection Regulation, GDPR)
  - ❍ (개요) 유럽연합 규범 체계상 지침(Directive)이 아닌 규정(Regulation)으로 모든 회원국 내에서 직접 효력을 가지며, 유럽연합 인의 개인정보 처리에 대한 법적 규율을 전반적으로 강화했음
    - 개인정보처리자가 준수해야 할 6대 원칙: 적법성·공정성·투명성, 목적 제한, 개인정보 처리의 최소화, 정확성, 보관 기간의 제한, 무결성·기밀성
  - ❍ GDPR은 그 이전의 개인정보보호 지침과 달리 익명화와 가명화에 대한 여러 규정을 두고 있음
    - 익명 정보: 개인정보의 반대되는 개념, GDPR의 적용을 받지 않고, 처음부터 익명인 경우와 사후적으로 익명화되는 경우 모두 해당됨
    - 가명화된 정보: 추가정보를 통하여 간접적으로 개인을 식별할 수 있는 정보로 현재 시점에서 효과적이고 안전한 데이터 보호를 위해 모든 상황에서 사용할 수 있는 기술로 언급
  - ❍ (가명화) 가명화는 직접 식별자를 제거하고 데이터가 간접적으로 식별 가능한 데이터만 포함한 유용한 개인정보 강화 기술임
    - 의무와 유인(incentive)에 관한 규정이 있으며 데이터 처리 시 가명화에 대한 일부 법적 혜택을 제공하여 이를 유도함
    - 데이터가 유출되어도 식별 위험이 크지 않고, 정보 주체에게 그 사실을 통지할 의무가 암호화를 포함한 가명화 처리가 된 경우 완화될 수 있음
  - ❍ GDPR은 가명화의 중요성을 인식하고 데이터 보안 및 보안 처리를 보장하기 위한 기술적 수단으로 언급하고 있음
    - 그럼에도, 가명 정보는 개인정보라고 말하고 있음
  - ❍ (익명화) 개인정보와 관련이 없도록 정보 주체를 더는 식별할 수 없도록 익명으로 처리된 정보
    - 법률적으로 효과적인 익명화 논의가 제기되었지만, 익명화 프로세스에 대한 명확한 규정을 제공하지 않음 (개인정보로 고려되지 않음)
    - 익명화 처리 자체에 대한 조건보다는 익명화 결과에 초점을 두고, 처리자는 재식별 위험을 평가해야 함
  - ❍ (법령 이외) 제29조 작업반(The Article 29 Data Protection Working Party : WP 29)이 2014년에 발표한 Opinion 05/2014에서 완전히 익명화된 정보는 GDPR의 적용 범위가 아님이라고 판단

❑ 영국 (UK Data Protection Act of 2018)

❍ (개요) EU 탈퇴 이후 GDPR에 직접적인 구속력을 받을 필요가 없어져 유럽연합과 같은 수준의 개인정보 규제 유지를 위해 기존의 데이터 보호법(UK Data Protection Act of 1997)을 개정했음

– 개정법률에는 GDPR의 규정에 따른다는 조항이 다수 포함되며 GDPR 내 일부 예외적인 항목에 관해서도 개별 조항을 만들어 반영했음

❍ (비식별 데이터 정의) 개정법률에서는 GDPR에서 규정한 익명, 가명 개념을 특별한 수정 없이 그대로 수용했음

❍ (익명화) 특정 데이터에서 어떤 개인이 식별될 위험성을 무시할 수 있는 수준으로 낮추는 과정으로 비식별 조치를 익명화의 방법의 하나로 고려

– 익명화된 데이터는 GDDR과 같은 법적 보호 범위가 아님

❍ (비식별 조치) 개인의 이름, 주소와 같은 직접 식별자를 제거하는 절차

– Information Commissioner's Office(ICO)의 가이드라인은 법적 구속력이 없지만 강력한 집행력을 가지고 있음

– 허용하는 익명화의 정도를 판단할 때 재식별의 위험성을 기준으로 보지만, 재식별 위험성이 존재하지 않는 수준을 요구하지는 않음

– 식별 위험성이 매우 낮은 수준까지 식별 위험성을 완화할 것을 요구

– 재식별 위험성 평가를 위해서는 데이터 자체뿐만 아니라 데이터 환경에 대한 파악을 전제하여 데이터의 환경적 요소에 따른 재식별 위험성의 변별력을 강조

– 위험성 평가: 데이터 관리 주체, 정보가 공유되는 유형, 재식별하려는 데이터 공격자 측면으로 구분해 위험성을 세밀하게 평가

❍ (가명화) 실제의 정체성을 드러내지 않는 특정 식별자를 사용해서 데이터 세트에 있는 개인을 구분하는 방법으로 익명화 기법임

– 가명화된 데이터는 여전히 개인정보로 법적 보호의 범위에 해당함

❑ 미국 (미국 건강 보험 이동성 및 책임법 Health Insurance Portability and Accountability Act, HIPAA)

❍ (개요) 미국은 개인정보보호를 포괄적으로 규제하는 연방 차원의 법률이 마련되어 있지 않아 분야별로 규율되는 형태(sectoral regulation)로 되어 있으며 그중 보건 의료영역을 규제하는 연방법이 HIPAA임

– 캘리포니아주의 캘리포니아 소비자 프라이버시 법 (The California Consumer Privacy Act of 2018 : CCPA)은 미국 내에서 가장 엄격한 법률로 유럽연합의 GDPR에 버금가는 개인정보보호 원칙들을 일부 포함함

❍ (HIPAA 비식별 데이터 정의) HIPAA 프라이버시 규칙은 비식별 조치된 정보의 개념을 정의하지는 않지만, HIPAA의 적용 대상이 되지 않기 위해 조치할 수 있는 비식별 조치의 기준과 구체적 방법들을 제시함

❍ (HIPAA 비식별 조치) 비식별 조치 방법을 사용하려는 수범 기관이 관련 전문가를 선임해서 전문가가 재식별의 측면에서 해당 데이터에 적용된 비식별 조치를 평가, HIPAA 프라이버시 규칙에서 나열한 18가지의 식별자들을 제거한 데이터는 비식별 조치가 적용된 것으로 간주함

❍ (CCPA 비식별 데이터 정의) 소비자를 합리적으로 식별할 수 없는 정보

❍ (CCPA 비식별 조치) HIPAA 프라이버시와 같은 구체적 비식별 조치 방법을 제시하지 않고, 비식별 조치 개념의 규정을 통해 지침을 제공함

| [1] 재식별을 방지하는 기술적 안전장치 적용 [2] 재식별을 특정적으로 금지하는 절차 마련 [3] 비식별 조치된 정보의 의도치 않은 공개를 예방하는 절차 마련 [4] 재식별 행위의 시도를 하지 않을 것 |
|---|

❍ (가이드라인) HIPAA 비식별화 가이드라인은 2012년 11월에 발표되었으며, HIPAA에서 정의한 18개의 PHI (Protected Health Information)에 대해서 전문가 판단 방법(Expert determination method)이나 완전히 제거하는 방법(Safe harbor method)을 제시함

❍ (법령 이외) 미국 상무부(US Department of Commerce) 산하 기술 표준화 담당 기구인 국가기술표준원(National Institute of Standards and Technology, NIST)에서 발표한 NIST 2015에 비식별 조치와 익명화의 개념이 정의돼있지 않아, 비식별 조치와 익명화의 개념을 정의했음

– 비식별 조치: 일련의 식별데이터와 정보 주체 사이의 연관성을 제거하는 모든 절차에 대한 포괄적인 개념임

– 익명화: 식별 데이터 세트와 정보 주체 사이의 연관성을 제거하는 절차임

– 가명화: 정보 주체와의 연관성을 제거하면서 해당 정보 주체의 특징들과 가명 사이의 연관성을 추가하는 특정 유형의 익명화를 말함

❑ 대한민국 (개인정보 보호법)

❍ (개요) 2020년 개인정보 보호법이 개정되었으며 개인정보보호와 관련된 기본법으로 개인식별정보 비식별화와 관련된 사항도 동 법의 조항을 먼저 살펴보아야 함

❍ (비식별 데이터 정의) 개인정보 보호법에서는 비식별 데이터 정의는 없으나 특정 개인을 알아볼 수 없는 형태라는 표현으로 익명 정보를 정의함

❍ (비식별 조치) 현행법상 비식별 조치를 규율하는 규정은 없으나 2022년 발표된 보건의료 데이터 활용 가이드라인에서 단계별 가명 처리 절차와 개념을 설명함

- 사전 준비: 가명 정보 처리목적을 명확히 하고 가명 처리를 위한 적합성 검토 및 계약서, 개인정보 처리 방침, 내부 관리계획 등 필요한 서류를 작성함
- 위험성 검토: 사전 준비에서 설정한 목적 달성에 필요한 항목을 개인정보 파일에서 선정하고 가명 처리 대상 데이터의 식별 위험성을 분석·평가하여 가명 처리 방법 및 수준에 반영함
- 가명 처리: 식별 위험성 검토 결과를 기반으로 가명 정보 활용목적 달성에 필요한 가명 처리 방법 및 수준을 정하여 항목별 가명 처리 계획을 설정
- 적정성 검토: 1, 2, 3단계의 가명 처리에 대해 결과 적정성을 최종 검토함
- 안전한 관리: 적정성 검토 이후 생성된 가명 정보는 법에 따라 기술적·관리적· 물리적 안전조치 등 사후관리를 이행함

❍ (가이드라인) 보건의료 데이터 활용 가이드라인은 개정된 개인정보 보호법이 시행(20.08.05)됨에 따라 데이터 활용의 핵심인 가명 정보 활용에 대한 법적 근거 마련을 위해 발표된 가이드라인이며 개인정보보호 법령 등에서 정하지 않은 가명 처리, 가명 정보의 처리 및 결합 활용 등을 규정함

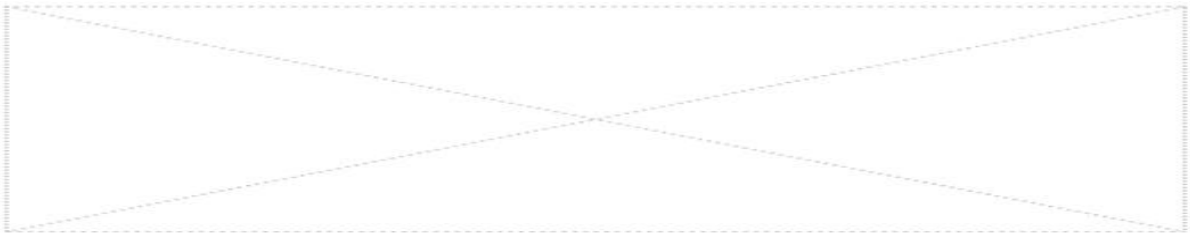


그림 23. 보건의료 데이터 활용 가이드라인 가명 처리 절차도

❍ 해외 규정에서 비식별 데이터의 개념은 데이터만으로 개인을 식별할 수 없는 데이터로 포괄할 수 있으며 국내의 가명, 익명 처리 데이터와 유사하지만, 국내에서는 데이터에 따라 가명·익명 처리 가능 유무를 구분함

❍ 국내 보건의료 데이터 가이드라인에서는 **개인 의료정보를 가명, 익명 처리 후 사용해야 한다고** 하며, **일부 데이터는 가명 처리 가능 유무 결정을 유보하여 예외적인 경우를 제외하고는 데이터 사용이 불가능**하며 대표적으로 유전체 데이터가 해당함

❑ 시사점

❍ 국가별 유전체 데이터의 비식별, 가명 처리 기준이 다르며, 국내에서는 이를 비식별 정보로 취급하지 않고, 가명 처리 여부 역시 보류 중

- (EU) 식별이 안 되게 처리한 유전체 정보는 개인정보로 취급하지 않아 비식별 조치된 유전체 데이터의 수집 및 사용이 가능
- (영국) 비식별 처리된 유전체 정보는 개인정보로 취급하지 않음
- (미국) HIPAA에서 포괄하지 않는 보건의료 데이터 중 소비자 직접 의뢰 유전자 검사(Direct-To-Consumer Genetic Test, DTC-GT)가 포함되면 이를 활용하여 유전체 데이터의 수집 및 사용이 가능
- (대한민국) 유전체 정보는 그 안에 담긴 정보의 내용을 모두 해석해내지 못하고 있고, 부모·조상·형제·자매·자손·친척 등의 제3자 정보를 담고 있을 수 있으므로 적절한 가명 처리 방법이 개발될 때까지는 가명 처리 가능성 유무 결정을 유보하였음[15)]

❍ 화합물 데이터도 신약개발에 활용성이 높으나 이러한 데이터는 개인정보가 포함된 데이터보다는 연구자나 실험한 기관, 업체 등의 지식재산권으로 접근하여 해결해야 할 문제임

15) 보건의료데이터 활용 가이드라인, 2022.12

표 20. 국가별 보건의료 관련 비식별화 가이드라인과 신약개발 데이터 비식별화 가능 유무

| 국가 | 법령 | 기관 | 가이드라인 문서 | 발표일시 | 유전체 비식별 가능 | 화합물 비식별 가능 |
|---|---|---|---|---|---|---|
| EU | GDPR | European Medicines Agency | European medicines agency policy on publication of clinical data for medical pro-ducts for human use | 2014.10 | O | O |
| 영국 | UK Data Protection Act of 2018 | Information Commissioner's Office | Anonymization: managing data protection risk code of practice | 2012.11 | O | O |
| | | UK Anonymisation Network | The anonymization decision-making framework | 2016 | | |
| 미국 | HIPAA | National Institute of Standards and Technology | NIST SP 800-188: De-identification government dataset | 2016.12 | O | O |
| | | Health Information Trust Alliance | De-identification Framework | 2016.4 | | |
| | | National Institute of Standards and Technology | NISIR 8053: De-Identification of Personal Information | 2015.10 | | |
| | | National Academies of Sciences, Engineering, and Medicine | Sharing clinical trial data | 2015.1 | | |
| | | IHE (Integrating the Healthcare Enterprise) | IHE IT Infrastructure Handbook: De-identification | 2014.3 | | |
| | | Department of Health and Human Service | (HIPAA) Privacy Rule | 2012.11 | | |
| 대한민국 | 개인정보 보호법 | 보건복지부 | 보건의료 데이터 활용 가이드라인 | 2022.12 | X | O |

# 제3장 신약개발 공공 데이터 공유 활성화 방안

## 3.1. 일반적인 데이터 공유 활성화 방안

❑ 데이터 공유 문화 확산

❍ 기업과 개인의 데이터 공유에 대한 이해를 높이고 데이터를 공유하는 것이 적극적으로 장려되는 분위기 조성이 필요하며 이를 위해 교육 및 캠페인 등을 통해 데이터 공유 중요성의 홍보가 필요함

❑ 개인정보보호 강화

❍ 데이터 공유가 활성화됨에 따라 개인정보보호 문제도 발생할 수 있으며 이를 방지하기 위해 기존법규를 강화하거나, 새로운 법규를 마련하여 데이터 공유와 개인정보보호를 모두 보장해야 함

❑ 데이터 표준화

❍ 데이터 공유가 원활하게 이루어지기 위해서는 데이터의 표준화가 필요

– 표준화를 통해 데이터의 일관성과 효율성을 높일 수 있으며, 데이터의 재사용성과 가치를 극대화할 수 있음

❑ 데이터 활용 규제 완화

❍ 데이터 공유·활용에 대한 제약이 많기에 데이터 공유가 어려워졌고, 이에 따라 기업과 개인이 데이터를 사일로(Silo)나 저장소에 축적하고 있음

– 이러한 문제를 해결하기 위해 데이터 활용 규제를 완화하고 데이터를 안전하게 보호하면서도 자유롭게 활용할 수 있는 환경을 조성해야 함

❑ 데이터 공유 인프라 구축

❍ 데이터 공유를 위해서는 데이터 공유 인프라가 필요하며 이를 위해 데이터 공유 플랫폼과 데이터 공유 네트워크 등의 인프라를 구축해야 함

❑ 본 연구에서는 데이터 공유 문화의 확산, 데이터 활용 규제 완화, 데이터 공유 인프라 구축에 중점을 두고 공유 활성화 방안을 제시하였음

❑ 프라이버시 보호 강화나 데이터 표준화 측면의 대안을 본 연구에서 제시하기는 어렵고, 두 가지를 모두 충족시킬 수 있는 기술적 방안을 제시에 집중하였음

## 3.2. 데이터 공유 활성화 의견 수렴 (전문가위원회)

❑ 데이터 공유 활성화 의견 수렴 개요

❍ 현황조사로 파악한 국내 데이터 신약개발 데이터 공유 활용의 현황을 공유하여, 전문가들의 현실적인 의견을 수렴하기 위해 2회 자문을 수행했음

❍ 보건의료 데이터 활용 가이드라인의 유전체 정보의 가명 가능 여부의 유보 결정으로 향후 인공지능 신약개발을 위한 데이터 협력 과제 기획에 중심이 될 기술에 대한 조사도 병행하였음

❑ 데이터 공유 방안 마련을 위한 전문가 의견 청취

❍ 데이터 활용 활성화 방안 논의를 위한 AI 신약개발 협의회 회의 (22.10.19)

❍ 참가자 명단

표 21. AI 신약개발 협의회 회의 참가자

| 연번 | 소속 | 직위 |
|---|---|---|
| 1 | 온코크로스 | 회장 |
| 2 | 디어젠 | 부회장 |
| 3 | 스탠다임 | 부회장 |
| 4 | 넷표적 | 위원 |
| 5 | 넷표적 | 참관 |
| 6 | 닥터노아바이오텍 | 위원 |
| 7 | 메디리타 | 위원 |
| 8 | 바이온사이트 | 위원 |
| 9 | 아이겐드럭 | 위원 |
| 10 | 에이조스바이오 | 위원 |
| 11 | 파로스아이바이오 | 위원 |

❍ 회의내용

- (개최목적) AI 연구자들에게 데이터 활용의 어려움을 청취하고 해결방안을 마련하기 위한 논의를 진행
- (주요 내용) AI 연구자들에게 필요한 데이터와 기존의 데이터 구축 및 공

유를 위한 정부 정책 사이에 불일치가 존재한다는 것을 확인

- 예) 민감 정보를 포함한 데이터에서 민감 정보가 필요하지 않은 연구자도 있으므로 민감 정보만 삭제하고 빠르게 데이터를 제공받길 원함
- 예) 가공된 데이터는 가공 방식에 따라 연구목적에 부적합해질 수 있음 이미 알려진 지식을 기반으로 처리되어 가공된 데이터의 경우 새로운 정보를 찾기 위한 연구용으로 부적합하다는 의견이 제기됨

❑ AI 신약개발 데이터 공유방안 논의를 위한 회의(22.11.16)

❍ 참가자

표 22. AI 신약개발 데이터 공유방안 회의 참가자

| 연번 | 소속 | 부서 |
|---|---|---|
| 1 | 대웅제약 | 신약센터 AI신약팀 |
| 2 | 대웅제약 | 신약센터 AI신약팀 |
| 3 | 일동제약 | MC팀 |
| 4 | 일동제약 | - |
| 5 | 동아ST | 신약연구소 |
| 6 | 유한양행 | 합성신약팀 |
| 7 | 유한양행 | 합성신약팀 |
| 8 | 현대약품 | 신약연구소 |
| 9 | 한미약품 | - |
| 10 | JW중외제약 | 데이터사이언스팀 |
| 11 | C&C신약연구소 | 데이터사이언스팀 |
| 12 | LG화학 | 생명과학본부 |
| 13 | LG화학 | 신약연구소 |
| 14 | 디어젠 | - |
| 15 | 스탠다임 | - |

❍ 회의내용

- (개최목적) EU MELLODDY 프로젝트에 대한 정보를 제약업계에 공유하고, 신약개발 데이터 공유 및 활용 방법과 프로젝트의 핵심 기술인 연합학습 접목에 대한 수요조사 수행

- (주요 내용) 제약사 간 AI를 위한 신약개발 데이터 공유 활용을 위한 방법으로 연합학습이 적절하다는 의견을 수렴하였음
- 최근 제약사들의 AI 기술 인식 수준이 높아져 연합학습을 활용한 신약개발 데이터 공유에 대한 거부감이 적어지는 분위기였으며, 몇몇 제약사는 이를 활용하는 사업을 추진하길 원했음
- 현실적으로 국내 제약사들은 글로벌 빅파마에 비해 적은 수의 데이터를 보유하고 있지만, EU MELLODDY와 같게 약물 동태 예측을 위한 데이터는 충분히 공유 가능하다는 의견이 있었음

❍ (결론) 민간 데이터의 안전한 공유 협력은 데이터를 보호와 활용 모두 만족시킬 수 있는 연합학습 기술로 가능할 것으로 예상되며, 제약사는 이 기술로 협력하여 학습한 AI 모델을 신약개발 도구로 활용 가능

❑ 데이터 공유 활성화 의견 수렴 결과

❍ 수요자들이 요구하는 AI 목적의 데이터와 공공 데이터 구축 목적의 불일치가 존재, AI 신약개발의 기술 수준을 파악하고 기술에서 요구하는 데이터의 형태는 무엇인지 파악해야 함

❍ 신약개발 관련 공공 데이터를 정부가 구축하고 있다는 것을 알고는 있지만, 데이터의 위치, 접근 가능성, 데이터 품질, 재가공의 필요 등으로 활용성이 낮아 활용성 증진에 대한 방안이 필요함

❍ 안전하게 신약개발 민간 데이터를 공유하는 수단으로 연합학습 기술이 제시되었고, 이 기술이 활용될 수 있는 연구 주제가 도출되었으며 사업기획 요청으로 연합학습 기반의 민간 데이터 협력 사업의 필요성을 확인

## 3.3. 공공 데이터 공유 활성화 방안

❑ 공공 데이터 공유 활성화 방안 개요

❍ 공공 데이터의 공유와 관련된 정부 주도 사업, 공유 플랫폼, 공공 데이터베이스, 법률적 환경의 현황 분석과 전문위원회로부터의 공유 활성화 의견 수렴 결과 공공 데이터의 공유 활성화 방안 3가지를 도출했음

- 방안 1. AI 신약개발 데이터 수요에 부합하는 공공 데이터를 연결하는 **수요기반 공공 데이터 매칭 프로젝트**
- 방안 2. AI 신약개발 활용사례 발굴을 위한 **공공 데이터 경진대회 개최**
- 방안 3. 여러 소스의 데이터를 안전하게 활용할 수 있는 **연합학습** 기술을 활용한 **AI 신약개발 플랫폼 KAIDD 고도화**

### 3.3.1. AI 신약개발 공공 데이터 매칭 프로젝트

❍ (목적) AI 신약개발의 실질 수요를 파악해 연구개발에 사용할 수 있는 공공 데이터를 매칭하여 공공 데이터의 제약산업 활용성을 높이고자 함

| AS-IS | TO-BE |
|---|---|
| 정부 주도로 다양한 공공 데이터가 구축돼 있으나, 제약산업에 활용할 수 있는 신약개발 데이터가 무엇이고 어디에 있는지 알기 어려움 | AI 기술 수요에 부합하는 필요 데이터를 정의하고 공공 데이터 중 신약개발 데이터가 어디에 있는지를 발굴해 수요자와 공급자를 매칭함 |
| 공공 데이터는 다목적성으로 구축된 데이터로 목적이 명확한 AI 기술 연구개발에서는 데이터를 바로 활용하기 어려움 | AI 신약개발 단계별 데이터의 포맷과 형태를 정의하고, 데이터를 즉시 활용할 수 있도록 표준 데이터 처리 프로토콜을 마련 |
| 공공 데이터 이용의 절차 복잡성, IRB 승인, 데이터 활용 시 특허 문제 등이 존재 | 데이터 처리 프로토콜을 통한 정제, 데이터공유 서비스를 제공할 수 있는 AI 신약개발 특화 데이터 공유 플랫폼 운영이 필요 |

❍ (추진 전략) 프로젝트를 수행하기 위한 5가지의 전략 제시

- 전략 1. 신약개발 단계별 AI 기술과 필요 데이터 맵을 만들어 공공 데이터 발굴
- 전략 2. 발굴된 데이터를 AI가 활용할 수 있도록 통일된 데이터 처리 절차 수립
- 전략 3. 공공 데이터 이용 절차 간소화와 AI 신약개발 특화 데이터 공유 플랫폼 운영을 통한 공공 데이터 매칭
- 전략 4. 인센티브 제도 및 성공사례 발굴
- 전략 5. 공유 촉진을 위한 정책적 지원 방안 마련

❍ (전략 1) AI 신약개발 데이터 맵 (Data Map) 구축

- 신약개발 과정의 단계별 AI 기술과 데이터 매핑이 필요하며, 그 결과를 바탕으로 단계별 기술에 맞는 공공 데이터 발굴하고, 개선방안을 마련
- 데이터와 AI 기술의 비전과 전략 제시로 수요자-공급자 간 상호 소통하는 생태계의 토대를 마련해 신약개발 관련 공공 데이터 공유 활성화에 기여

❍ (전략 2) 데이터 처리 절차 수립

- AI가 공공 데이터를 활용하기 위해서는 공공 데이터를 AI 모델 입력에 맞게 전처리, 가공하는 데이터 처리 절차와 데이터 표준이 필요함
- 데이터 처리 절차 수립은 신약개발 단계별 존재하는 모든 AI가 활용할 수 있는 데이터의 표준을 수립하는 것이 아니라, AI 모델 개발 목적에 맞게 처리하는 데이터 절차와 모델에 입력가능한 데이터 형태를 말함
- AI 신약개발 데이터 표준 마련
  - AI 신약개발 데이터 표준은 신약개발 과정 단계에서 활용할 수 있는 AI 데이터 요소에 대한 명칭, 정의, 형식, 규칙에 대한 원칙을 수립하여 이를 전사적 적용을 의미하며 기존의 표준이 있으면 활용
  - 동일 종류 데이터에 동일 표준을 적용하면 데이터의 호환성, 편의성, 상호운용성 등이 높아져 원활한 데이터 공유가 가능함
  - 아래 표와 같이 오믹스 데이터는 유전체 정보센터 연합체 (INSDC: International Nucleotide Sequence Database Collaboration)에 의해 국제표준이 정해졌으며, 화합물 데이터는 ChEMBL, PubChem의 형식이 표준으로 인정받고 있음

표 23. 바이오 데이터의 국제표준 사례

| 분류 | 데이터 타입 | 국제 표준형식 |
|---|---|---|
| 오믹스 | 차세대 핵산 서열 (Next generation reads) | SRA(Sequence Read Archive) 양식 |
| | 1세대 핵산 서열 (Capillary reads) | Trace Archive 양식 |
| | 주석된 핵산 (Annotated sequences) | GenBank 양식 |
| | 샘플(Samples) | BioSample 양식 |
| | 프로젝트 | BioProject 양식 |
| | 마이크로에레이 | NCBI의 Gene Expression Omnibus(GEO, EMBL-EBI의 ArrayExpress |
| | 단백질 구조 | PDB 양식 |
| 이미지 | Bio Image | Image Data Resource (IDR) 양식 DICOM 형식 |
| 동영상 | 동영상 | sf, avi, wmv, mp4, mov 등 |
| 화합물 | 화합물 | PubChem 양식, ChEMBL 양식 |

- 국내에서는 국제표준을 참고해 각 수요처의 요구사항 및 의견을 반영하고, 표준화 위원회를 통한 데이터 타입별 표준 및 공유 체계를 마련
- 표준화 위원회는 제약사, AI 신약개발 기업, BT·IT 분야 데이터 전문가 등으로 구성하며 표준화할 데이터를 선정하고 표준을 만듦

❍ (전략 3) AI 신약개발 특화 데이터 공유 플랫폼

- AI 신약개발 데이터 맵과 데이터 처리 절차를 수립하였으면, 이를 바탕으로 데이터를 공유할 수 있는 플랫폼이 필요함
- 특화 데이터 플랫폼에서는 신약개발 단계별 필요한 데이터와 이용 목적을 상세하게 파악하고 집중적으로 지원하는 방안을 마련해야 함
- AI 신약개발 단계별 데이터를 공유할 수 있도록 구성한 특화 데이터 플랫폼으로 기술 수요자와 공급자(데이터 수요자)를 연결해주고, 데이터 공유·활용 가능

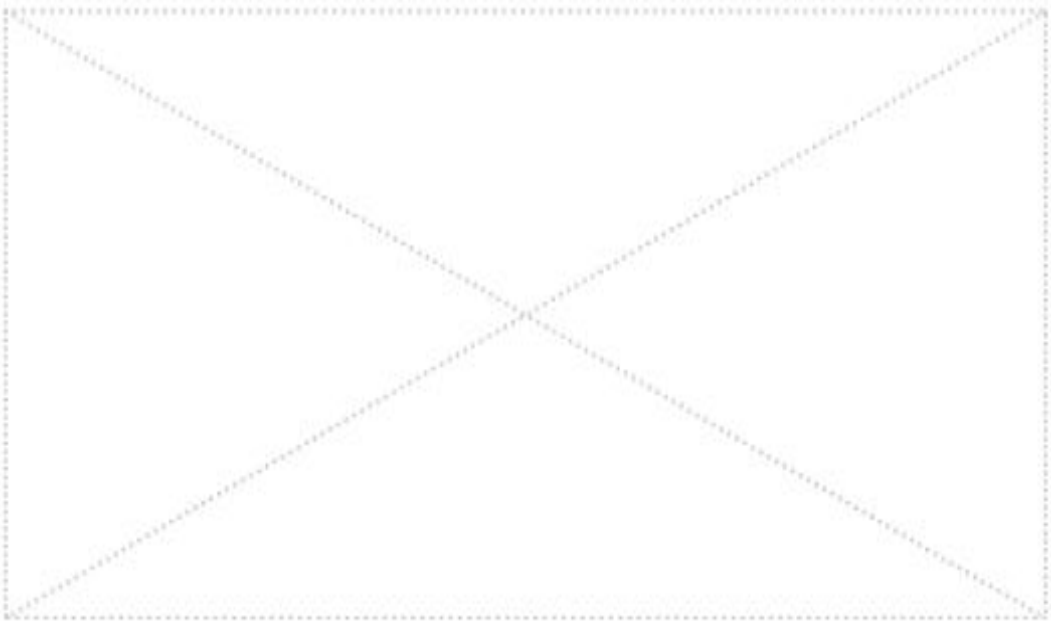


그림 24. AI 신약개발 공공 데이터 매칭 프로젝트 체계(안)

❍ (전략 4) 인센티브 제도 및 성공사례 발굴

- 신약개발 데이터 공유를 위한 여러 분야의 노력을 인정하고 이에 대한 이익을 공유하는 방법으로 다양한 인센티브를 제공하는 방안 마련 필요
- 데이터 공유를 통한 수요자와 제공자 간의 협업 성공사례 발굴로 신약개발 분야의 성장과 공유 생태계 활성화 도모

❍ (전략 5) 공유 촉진을 위한 정책적 지원 방안 마련
- 신약개발 데이터는 종류에 따라 민감 데이터(개인정보, 지식재산권 등)를 포함하고 있어 법적 문제의 해결방안과 안전한 공유 방법 마련이 필요함
- 신약개발 관련 연구 완료 후 수집되는 데이터는 공동 이용의 활성화를 위해 공유되어야 한다는 연구관리 차원의 전략 수립 필요함

❑ 기대효과
❍ 수요 기반 데이터 요구를 청취하여 공공기관에 수집·확보된 데이터에 대해 인공지능 데이터 표준을 통한 지속적인 관리 요청으로 활용할 수 있는 순환 생태계를 구축
❍ 수요 기반 단기 성과형 소규모 다과제 발굴로 AI 신약개발의 단계별 R&D 과제를 수립 및 수행 결과로 표준절차에 따른 수요 기반 데이터들을 쌓아나감으로 고급 공유 데이터 확보 및 활용성 증대
❍ 인센티브 제도와 성공사례 도출을 통한 신약개발의 핵심 데이터 확보를 촉진하고, 폭넓은 신약개발 과정의 단계별, 종류별 표준화된 데이터를 공유할 수 있는 활용 생태계를 조성이 가능
❍ 공유 데이터 활용 촉진을 위한 정책적 지원으로 신약개발에서 편리하고 안전한 공유 환경 정착

### 3.3.2. 공공 데이터 활용 AI 신약개발 경진대회 개최

❑ 목적

❍ 국내 공공 신약개발 데이터의 공유 활용 아이디어 발굴, 신규 서비스 창출, 난제 해결 등을 목표로 개최해 공공 데이터의 활용성 및 인식 향상

| AS-IS | TO-BE |
|---|---|
| 신약개발에 AI 기술을 접목하려는 기업은 점차 증가하고 있으나, 인력은 계속 부족한 상황 | 경진대회 개최로 AI 신약개발의 중요성과 가치를 홍보하고, 젊은 과학자와 연구원의 유인을 통한 안정적 인력 수급 환경을 조성 |
| 인력 문제, 기술 부족, 연구 환경 등 다양한 이유로 적극적으로 AI의 신약개발 도입에 어려움을 겪는 기업이 많음 | 현업의 AI 신약개발 도입의 난제를 경진대회에 제시하여 해결의 실마리를 획득하거나, 직접 해결 가능 |
| 공공 데이터의 접근성, 절차 복잡성, 지식재산권 등 문제로 AI가 활용할 데이터 부족 | 신약개발에 유용한 데이터를 발굴하고, 공공 데이터의 신뢰도 및 활용성의 향상이 가능 |

❑ 추진 전략

❍ AI 신약개발 경진대회 개최 및 활성화 유도, 공공 데이터 활용을 위한 방안을 마련

- 전략1. 공공 데이터로 공개 이전에 새롭게 생산된 데이터를 경진대회에 활용하고자 함(즉, 신규 공공 데이터 활용)
  - 예) 한국화학연구원에서는 매년 기탁받은 화합물 데이터를 공공에 공개, 공개 이전 해당 화합물과 연구원에서의 실험을 통한 값들을 활용(미공개) AI 개발 경진대회를 개최할 예정
- 전략 2. 데이터와 문제 발굴, 모델 기준선(Baseline) 설정, 제한 없는 참가자 모집, 경진대회 사이트 마련, 리더보드(Leader-board)와 코드 공유, 커뮤니티 기능 등의 준비 작업 수행 및 지속 개최방안 마련
  - 데이터 분할 방법: 신약개발 특성상 훈련 데이터와 예측 데이터(테스트)가 상이한 경우가 많으므로 단순 랜덤 샘플링을 통한 데이터 분할이 아닌 Scaffold 기준, 상이한 데이터 분포를 테스트 데이터로 활용
  - 순위 기준: AI 경진대회 특성상 순위 기준에 모델의 성능이 주요하겠지만, 단순 성능 평가만으로는 실용성이 떨어지므로, 상위 10위를 우선선발, 발표 평가를 통해 어떠한 신약개발 관련 지식을 모델 개발에 활용했는지 점검하고, 발표 평가를 기준으로 최종 순위 평가
- 전략 3. 상금의 규모 확대와 수상 시 유용성 제공 방안 마련

그림 25. 경진대회 추진 체계(안)

❑ 기대효과

❍ AI 신약개발의 가치를 널리 알리며, 관련 분야의 인력을 유인하여 안정적인 인력 수급 환경을 조성

❍ 신약개발 현업에서 AI로 해결해줄 수 있는 문제에 도움을 받거나, 난제의 해결 및 방안이나 아이디어를 획득할 수 있는 아웃소싱 장소가 될 수 있음

❍ 경진대회 사용을 위해 공공 데이터를 검토하고 공개함으로써 유용한 데이터를 발굴할 수 있고, 데이터 활용성도 증진할 수 있음

### 3.3.3. 연합학습 활용 AI 신약개발 플랫폼 KAIDD 고도화

❑ 인공지능 신약개발 플랫폼(KAIDD)

❍ (개요) 신약개발에 필요한 인공지능 플랫폼을 구축하여 국내의 신약개발 연구자 대상 AI 서비스를 제공하는 것을 목표

– 신약개발 과정에 활용할 수 있는 AI 도구를 제공해 신약개발에 걸리는 시간과 비용을 대폭 단축하고자 함 (신약개발에 도움이 되는 AI 도구 개발)

❍ (성과) 신약 표적 탐색부터 디자인, 최적화, 검증, 신약 후보 도출까지 신약개발 전 과정을 수행하는 인공지능 기반 통합 플랫폼 구축

– 유전체 기반의 질환 표적 탐색 및 후보 화합물 선별 플랫폼

- 차세대 서열 해독 방법으로 생산된 유전체 정보를 이용하여 유전체 발현 마커를 찾는 방법 개발
- 유전체 발현 마커 기반 신약 후보 물질 및 질환 표적 탐색 방법 개발

– 단백질 구조예측 기반의 신약 디자인, 리간드 기반의 신약 디자인을 통합한 신약 디자인 플랫폼

- 화합물과 표적 단백질 구조 기반 신약 디자인이 가능한 8가지의 신약개발 인공지능 모델 개발

• 신약 초기 화합물 탐색부터 Scaffold 디자인, 화합물 최적화, 구조 기반 후보물질 선별, 특성 분석 및 임상시험 실행 가능성 등 모델 개발

– 신약 성공 가능성 예측을 위한 Insilico 신약 성공 가능성 예측 플랫폼



그림 26. 인공지능 신약개발 플랫폼(KAIDD)에서 제공하는 인공지능 신약개발 도구

❑ 연계 및 활용 방안

❍ (기존 사업의 한계) 신약개발에 사전 학습된 인공지능 모델을 제대로 활용하려면 사용자 데이터에 맞도록 모델의 미세조정이 필요한데 기존 플랫폼에는 이러한 미세조정 기능이 없고 추가 데이터로 학습이 불가함

❍ (해결방안) AI 신약개발 도구의 **지식 확장과 모델 개인화가** 가능하며, 다중 소스의 데이터로 AI 모델을 학습할 수 있는 연합학습 기술 활용을 제안

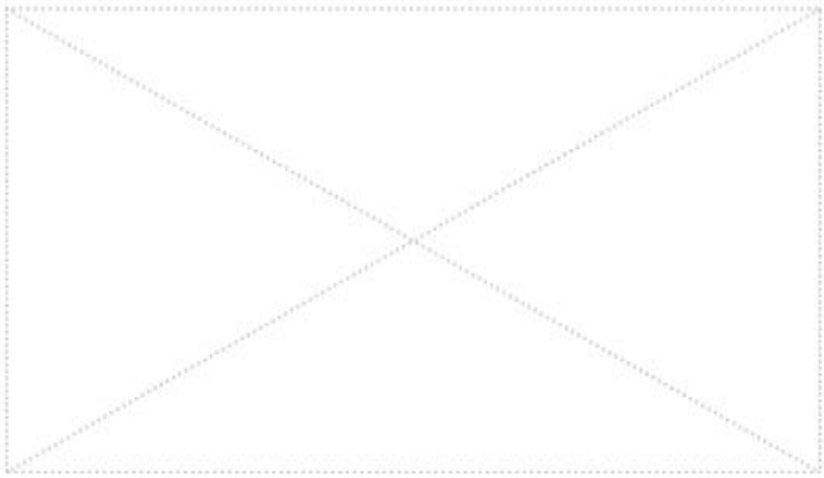


그림 27. 기존 사업 인공지능 모델 활용 계획

❍ (연계 및 활용 방안) KAIDD 플랫폼의 도구를 연합학습의 학습 모델로 활용해 플랫폼을 고도화하고, 도구의 성능 향상으로 활용성 제고

– (기본 모델 및 학습 공개 데이터 개방) KAIDD에서 활용한 AI 모델 구조 및 활용한 데이터를 공개하여 모델 활용성 증대

– (지식 확장 도구 마련) 모델에 사용자가 원하는 민간, 공공 데이터를 입

력하는 **데이터 입력 도구**를 개발하고 전이 학습, 미세 조정 기능 추가

- (도구 개선) 도구 개발에 사용된 공개 데이터(Zinc, BindingDB, ChEMBL, Chemspace, OTAVA, Enamine, DRUGBANK)에 공공 데이터(KOBIC, 한국화합물은행)나 민간 데이터를 추가 학습하여 모델 성능 개선
- (서비스 제공) AI 신약개발 도구의 주 사용자는 화학, 바이오 연구자로 AI 모델 자체의 활용이 어려우므로 손쉽게 이용할 수 있도록 AI 모델을 활용한 예측 서비스를 개발하여 이용성 향상

❑ 기대효과

❍ 국내외 공공 또는 민간 데이터 유출 없이 안전한 데이터 학습이 가능하고, 개별 기관의 데이터와 모델을 갱신할 수 있는 자동화된 시스템 구축이 가능해 가장 최신의 AI 도구를 연구자나 사용자들이 활용할 수 있음

❍ 연합학습은 다양한 소스의 데이터를 학습할 수 있다는 장점이 있어 방대한 데이터 학습이 가능해 사용자에게 더 좋은 성능의 AI 기반 서비스를 받을 수 있고, 정부는 기존 사업의 지속성과 연계성 확보할 수 있음

# 제4장 신약개발 민간 데이터 공유 현황 분석

## 4.1. 신약개발 민간 데이터 공유 개요

❑ 신약개발 민간 데이터 공유 현황

- ❍ 과거에는 각 제약사나 연구기관이 자체적으로 데이터를 생산·수집하여 사용했지만, 최근에는 이를 공유하는 추세임
- ❍ 아스트라제네카는 COVID-19 백신 개발에 사용된 임상시험 데이터, 화이자는 항암제 개발에 사용된 임상시험 데이터를 공개했음
- ❍ 데이터 공유는 신약개발 분야에서 연구 생산성 향상과 비용 절감에 큰 역할을 할 수 있으며 다양한 연구자들이 데이터를 활용하여 창의적인 아이디어를 제시하고, 새로운 진단법이나 치료법을 개발할 수 있음
- ❍ 하지만, 데이터 공유에는 데이터 보안, 개인정보보호, 지식재산권 등의 민감한 문제가 있을 수 있으므로 적절한 보안 및 보호 대책이 필요함

❑ 신약개발의 민관 파트너십(Public-Private Partnership, PPP)

- ❍ 민관 파트너십은 정부와 민간 기업 간의 협력관계를 의미하며 사전에 계획된 협력 구조를 구축하고 자원과 기술을 공유하여 사회적 이익, 사업 주체의 이익을 동시에 추구할 수 있음
- ❍ 2004년부터 다양한 형태의 신약개발 민관 파트너십 이니셔티브(Initiative)가 출범했고 일본 국립암센터와 아스트라제네카의 항암제 개발 협력 등 다양한 사례가 있음

표 24. 신약개발 민관 파트너십 대표 사례

| 단체명 | 목적 |
|---|---|
| Innovative Medicines Initiative (IMI) | 의료 또는 사회적 필요가 충족되지 않은 분야의 건강 연구 및 혁신 자금 지원에 주로 초점을 맞춘 생명 과학 분야의 EU 민관협력 기구 |
| Drugs for Neglected Disease Initiatives (DNDi) | 소외된 질병에 대한 신약개발을 공공 민간 협력으로 해결하려는 기구로 신약개발 가속화를 위한 화학 라이브러리 구축을 위해 제약 파트너와 협력하고 있음 |
| Open Targets | 체계적인 약물 표적 식별 및 우선순위 지정을 위한 인간 유전학 및 유전체학 데이터를 사용하는 혁신적인 대규모 민관협력 기구 |
| ATOM(Accelerating Therapeutics for Opportunities in Medicine) consortium | 신약개발의 느리고 실패율이 높은 프로세스를 신속하고 통합된 환자 중심 모델로 전환하는 것을 목표하며, 주요 업무 중 다양한 생물학적 데이터를 민관이 협력하여 구축하는 것이 있음 |
| Accelerating COVID-19 Therapeutic Interventions and Vaccines (ACTIV) | COVID-19 대유행에 대한 가장 유망한 치료법 및 백신 개발의 우선순위를 정하고 가속화하기 위한 공동연구 전략 도출을 위한 NIH 주도 민관협력 기구 |

❑ 신약개발 민간-민간 파트너십의 첫 사례 등장

❍ 이러한 노력에도 불구하고 신약개발 분야에서 민간 기업 사이의 파트너십을 찾아보기는 어렵고 일대일 협업 수준에 정체되어 있음

❍ 최근, IMI의 주도로 10개 대형제약사의 데이터를 블록체인과 연합학습 기술로 기밀 유출 없이 안전하게 거대 AI 모델을 만드는 MELLODDY(Machine Learning Ledger Orchestration for Drug Discovery) 프로젝트가 실시되었음

## 4.2. EU MELLODDY

❑ 개요

❍ 세계 최초 제약사 간 민간 데이터 협력 프로젝트로 다양한 약물 동태 예측 목적의 AI 모델 학습을 위해 블록체인 연합학습 기술을 활용

❍ (목적) 블록체인, 연합학습 기술로 연구 비밀 노출 없이 협력하는 플랫폼 구축

❍ (목표) 연구 비밀 노출 없이 약물 탐색 관련 모델의 예측성능 향상 및 화학적 적용 가능 도메인을 확장할 수 있다는 가설을 시연

❍ (수행 기간과 자금, 참여기관) 2019.06.01.부터 2022.05.31.까지 Owkin이 주도하는 컨소시엄에서 3년간 수행하는 프로젝트로 초기 1,840만 유로의 자금을 지원받았고 과제 종료까지의 전체 지원금은 1,956만 유로였음

표 25. EU MELLODDY 세부 투자금 내역

| 참여 기관 명 | 기관 유형 | 투자금(IMI, EFPIA) |
|---|---|---|
| Owkin France | 기업(IT 연합학습 솔루션) | IMI: 2,683,635<br>EFPIA: 100,705 |
| KUBERMATIC GMBH | 기업(IT 클라우드 기술) | IMI: 1,716,178<br>EFPIA: 1,128,537 |
| BUDAPESTI MUSZAKI ES GAZDASAGTUDOMANYI EGYETEM | 대학(IT 보안기술) | IMI: 1,253,297<br>EFPIA: 45,505 |
| SUBSTRA | 기업(IT 연합학습 프레임워크) | IMI: 858,482<br>EFPIA: 50,513 |
| IKTOS | 기업(IT 의약화학, 신약개발 AI 모델) | IMI: 343,402<br>EFPIA: 15,191 |
| KATHOLIEKE UNIVERSITEIT LEUVEN | 대학(IT 연합학습 보안 알고리즘) | IMI: 1,145,006<br>EFPIA: 48,834 |
| JANSSEN, ASTRAZENECA, GLAXOSMITHKLINE, MERCK, BOEHRINGER INGELHEIM, BAYER, NOVARTIS, ASTELLAS, SERVIER, AMGEN | 제약사(10개 기관) | IMI: 0<br>EFPIA: 10,054,420 |
| NVIDIA | 기업(IT GPU 인프라) 및 연합학습 | EFPIA: 120,000 |

- 유럽 혁신 의약품 이니셔티브인 IMI (Innovative Medicines Initiative)와 유럽 제약산업협회 EFPIA (European Federation of Pharmaceutical Industries and Associations)의 자금 지원을 받았음
- 제약 기업(10개)과 대학(2개), 중소기업(4개) 및 AI 컴퓨팅회사(1개) 등 총 17개의 파트너가 유럽 전역에서 참여

❍ (세부 목표) 공통의 기계학습 모델 개발의 설득력 확보를 위해 기술의 컨셉을 증명하고, 기술 임곗값을 낮추는 것을 목표로 함
- 1,000만 개 이상의 주석이 있는 소분자 화합물 (공공+민간 데이터 활용)
- 10억 개 이상의 약리학적 검정 활성 레이블 사용(Bio Activity)
- 고 처리량 스크리닝에서 다중 복잡 표현형 사용
- 데이터 및 연합 모델의 정보보호가 가장 중요(블록체인, 연합학습, 보안)

❑ 거버넌스와 조직 구성도

그림 28. EU MELLODDY의 거버넌스와 조직 구성도

❍ WP1: 데이터 전처리

– 약물동태(Absorption, Distribution, Metabolism, Excretion, ADME)와 약물 용량에 따른 반응 분석 데이터와 예측 결과 최적화를 위한 보조 데이터를 포함하여 기계학습 모델의 입력에 사용할 수 있도록 데이터를 전처리

- 데이터의 전처리 방법론을 선택하고, 특징 추출, 차원 축소, 가중 데이터 통합 등을 포함하는 기능을 스크립트로 구현(파이썬 실행 코드)
- 연합학습 아키텍처와 호환되는 방식으로 스크립트 배포
- 데이터를 전처리하고 스크립트를 사용자가 사용할 수 있도록 함

❍ WP2: 연합 및 개인정보 기계학습 알고리즘

– 기계학습 모델의 설계, 최적화, 성능 평가를 기반으로 모델의 정보 유출 위험, 잠재적 공격 및 완화 전략을 고려

- 연합학습에서 동작할 수 있는 기계학습 알고리즘의 프로토타입 개발
- 공개 데이터를 활용한 조기 예측성능 추정
- 공개 데이터를 활용하여 보안 평가를 가능하게 하는 알고리즘 구현

❍ WP3: 정보보호와 성능의 균형 평가, 플랫폼의 예측 성능 평가

– 알고리즘 전반에 걸친 정보보호와 성능 간의 균형 평가, 성능 평가

- 성능과 정보보호 간의 균형 평가(프로토타입에서)
- 집계된 예측 성능 측정 지표 평가(엔터프라이즈 단계에서)
- 소프트웨어 감사(Audit)

❍ WP4: 엔터프라이즈 지원 소프트웨어 구현

– SW 솔루션 업체에서는 연합학습 후 피드백, 엔터프라이즈 지원 SW 구현

- 사업수행에서 발생하는 성능 및 보안 요구사항을 반영하여 개선

❍ WP5: 보안 인프라 및 SW 배포, 산업용 IT 기술 범위 검토

– WP4에서 개발한 SW를 호스팅할 수 있는 보안 컨테이너 인프라 제공

- 성능과 개인정보보호 간의 균형 평가(프로토타입에서)
- 집계된 예측 성능 측정 지표 평가(엔터프라이즈 단계에서)

❍ WP6: 플랫폼 운영 및 모니터링, 유지관리, 서비스 제공

– 연간 실행의 효율성 모니터링, 종료 후 플랫폼 지속 가능성 확보

- 플랫폼 기술문서를 개발
- 플랫폼 연합학습 결과를 홍보
- 상용 서비스 개발을 위한 로드맵 준비
- 수행 기간 이후에도 플랫폼 접근이 가능하도록 지속 가능성 계획

❍ WP7: 전반적인 프로젝트 및 소통 관리

– 프로젝트의 성공적 완수를 위해 전체와 개별 프로젝트 관리

- 보조금 관리
- 전략, 운영, IP 및 재무 관리
- 의사소통(컨소시엄 내부 및 관련 외부 협력자)
- 과학 커뮤니티 및 제약 부문에 과학적 결과 및 연구데이터 보급

❑ 중장기 로드맵

❍ Phase 1 (2019.06~2020.01): 12월까지 준비 과정을 수행, 학습을 위한 데이터 세트를 준비(데이터 확보 및 전처리 기능 개발), 외부 자문 수행 및 보고서 생성(업계의 의견 수렴, 프로젝트 방향성 점검)

❍ Phase 2 (2020.02~2021.01): 자문을 통한 구현과 측정에 대한 승인, 전반적인 준비 상태 확인(인프라 및 플랫폼 개발 완료)

❍ Phase 3 (2021.02~2022.01): 연합학습을 반복 수행 및 결과 분석, 오류 및 문제 사항 등을 수정, 전체 모델의 성능 개선 및 보안과 모델 성능의 절충안 찾기

❍ Phase 4(2022.02~2022.05): 최종 결과 보고

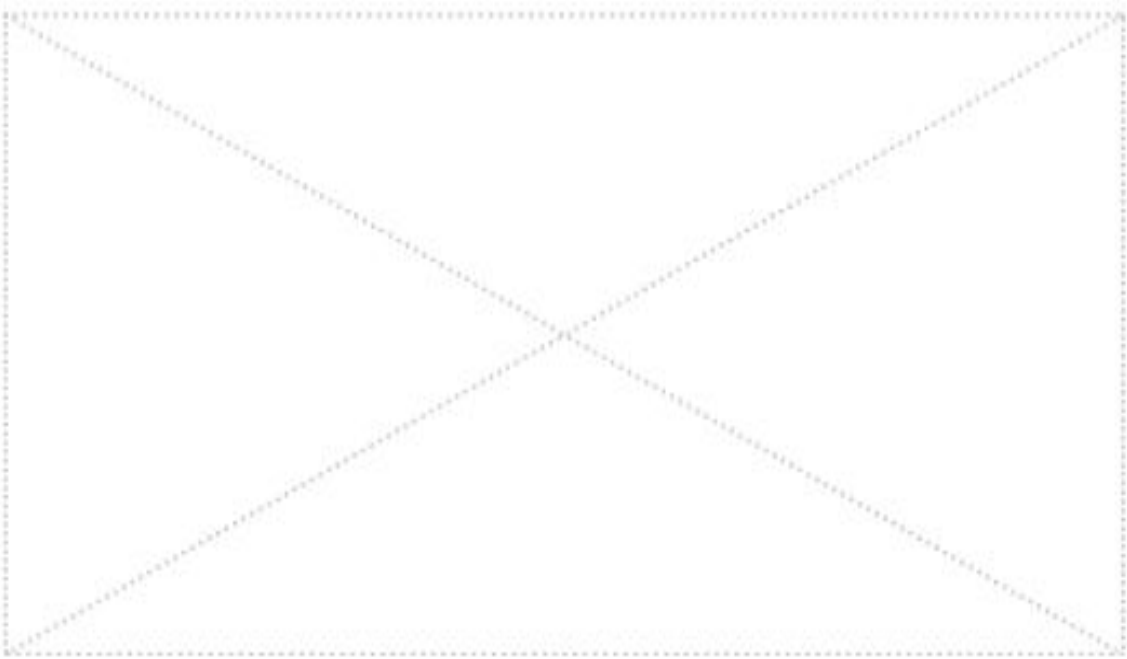


그림 29. MELLODDY 연간 로드맵

❑ 데이터 확보 전략

❍ 산업계 컨소시엄 확정 및 데이터양 추산

– 사업 RFP에 참여 제약기업 명시(상기 기재)

– 산업계 참여자들의 데이터는 동 프로젝트에서 새로 생성하지 않음

– 민간 및 공공 데이터양 추산

- 용량-반응 품질 활성 데이터로 주석이 달린 최소 5백만의 화합물(5 million chemical compounds annotated with dose-response quality activity data)
- 일부 활성에 대한 주석이 달린 최소 1천만 개의 화합물(10 million chemical compounds annotated with some activity)

- 단일 용량으로 수집된 최소 10억 개의 분석 활성 데이터(HTS에서 화합물 당 1~ 몇 개의 숫자 값) 1 billion assay activity data points collected at single dose
- 용량-반응에서 수집된 최소 1억 개의 활성 데이터(추적/2차 선별) 100 million activity data points collected in dose response
- HTS로 수집된 복합적인 활성(표준화된 경우로 100,000개의 화합물, ex. 고해상도 현미경 이미지, 웰당 1,000개의 판독 값이 있는 전사 프로필)
- 솔루션이 개발된 2차 연도부터 산업계 참여자들이 **100,000개 이상**의 전사체, HTS 데이터를 포함해 이익을 얻을 수 있도록 해야 한다고 명시

❑ 위험 관리

❍ 동 사업에서 위험 관리를 WP7(Janssen이 리더, Owkin이 참여)에서 수행

– 프로젝트 목표와 관련된 위험 항목을 정의 및 관리하였으며, 위험에 대한 대응 논의는 운영위원회에서 신규 위험 식별, 조정, 완화 전략 수립

– 위험의 발생과 중요도(Low, Medium, High의)를 3단계로 구분하여 논의하였음

| 프로젝트 목표 | 리스크 항목 |
|---|---|
| 다중 파트너가 기계학습과 멀티 태스크 모델로 예측성능의 개선을 증명 | 다중 파트너 기계학습 모델이 단일 파트너와 비교하여 예측성능 및 화학 공간 개선에 실패하는 경우 |
| 모델의 화학 공간 개선으로 응용 분야 확장을 증명 | 다중 파트너 기계학습 모델이 단일 파트너와 비교하여 예측성능 및 화학 공간 개선에 실패하는 경우 |
| 프라이버시 보호 | 운영 중과 종료 후에도 프라이버시(제약기업의 지식재산권) 보호가 보장되지 않는 경우 |
| 유연하고 확장가능하며, 안전한 FL 및 ML 프레임워크 | 플랫폼이 유연하지 않거나 확장이 불가능한 경우 |
| 플랫폼의 감사, 스트레스 테스트 및 평가 | IT 요구사항 설정과 3차 연도 플랫폼 감사 결과와 관련된 위험, 연간 실행 중 발생하는 운영 관련 위험성 |

❑ 과제 결과

❍ (정량적 지표)

– 화합물 활성 데이터 전처리 도구 MELLODDY TUNER(SW 1건):

ChEMBL 공개 데이터와 제약사에서 제공하는 ADME, 약물-표적 상호작용 예측 관련 활성 데이터를 활용할 수 있도록 입력 데이터는 SMILES을 ECFP6 32 Kbit로 변환했고, 출력 데이터는 화합물의 활성 값을 회귀에 활성 여부는 분류 예측에 사용했음



그림 30. SparseChem 모델의 입력 데이터 형태

- 멀티 태스크 기계학습 접근 알고리즘 SparseChem(SW 1건): 학습 모델의 보안성 강화를 위해 공개되는 부분(Trunk)과 공개되지 않는 부분(head)을 사용하는 Trunk-head 모델을 사용하였음
  - 모델은 개별 영역(private head)과 공통 영역(common trunk)으로 구분되어, MELLODDY 플랫폼에서 연합학습이 실행되는 동안에는 공통 영역의 가중치는 참여 클라이언트들의 학습 결과인 개별 가중치(gradients, weights)를 안전하게 수집하여 학습이 이루어짐
  - 개별 영역의 가중치(private weights)는 비공개로 클라이언트에 남음
  - 연합학습이 종료된 후 생성된 모델은 common trunk(클라이언트 모두 동일)와 private head(클라이언트마다 달라 비공개)로 구성

그림 31. 지식재산권 보안 강화 목적의 학습 모델 네트워크 구조

- 프라이버시 보호 및 연합학습 기반 MELLODDY 플랫폼(1건): 일반적인 연합학습 구조로 학습 결과를 연합하는 중앙 서버와 개별 학습을 수행하는 클라이언트 구조를 사용했음
  - 단순히 연합학습 아키텍처만 활용한 것이 아니라 블록체인(분산 원장 기록)을 통한 추적 기능 탑재



그림 32. MELLODDY 플랫폼 아키텍처

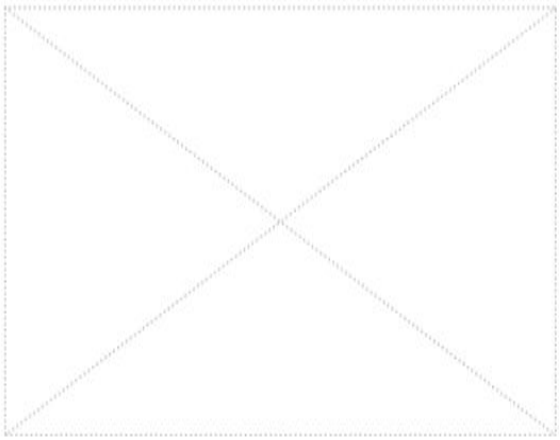


그림 33. MELLODDY 플랫폼 노드 상세

- 논문 총 12건, 저널 9건(MDPI, ACM, arXiv 등), 심사 중 3건(AAAI, Chemical Science, ACS Central Science 등)
- MELLODDY 학습 결과: 개별 제약사의 모델(그림에서 y축 기준 0)보다 연합학습 모델이 회귀 및 분류 예측 모두 성능이 개선됨(회귀 모델 평균 2%, 분류 모델 평균 4% 향상)

그림 34. MELLODDY 학습 결과

❑ 평가 결과

❍ 평가에 관한 자세한 내용은 현재 공개된 자료에서는 찾을 수 없었음

❍ 성공적 프로젝트의 결과로 플랫폼의 상용화 진행 중이라고 함(보고서에 언급은 되어 있으나 상세 내용 확인 불가)

❍ 기금을 제공했던 제약회사들에서 AI 모델을 활용하기로 했다고 함(보고서에 언급은 돼 있으나 상세 내용 확인 불가)

❑ SWOT 분석

❍ MELLODDY 과제에 대한 SWOT 분석

| | |
|---|---|
| 강점(STRENGTHS):<br>– 특별한 컨소시엄 협업(제약, AI, IT, 대학)<br>– 제약사는 높은 가시성, 잠재력을 보유<br>– 기술 기여자(AWS, NVIDIA) 인지도가 높아 많은 잠재 고객을 보유<br>– 프로젝트에는 기술, 연구, 약물 개발의 다양한 전문가가 참여 | 약점(WEAKNESS):<br>– 데이터 과학 및 기술은 여전히 모호한 부분이 존재함(복잡하고, 고도한 기술적 개념이 많음)<br>– 약물 개발 초기 단계에 집중했기에 소수의 관련자에게만 산출물을 제공할 수 있음<br>– 여러 컨소시엄 구성으로 우선순위와 요구사항이 충돌하는 경우가 많음 |
| 기회(OPPORTUNITIES):<br>– AI, ML, FL은 의료 분야에서 화두<br>– 프로젝트가 성공하면 기술 중소기업을 위한 서비스 또는 제품을 구축하는 동시에 다른 제약 파트너로 확장 가능<br>– 프로젝트에서 사용하는 기술은 조직이 GDPR을 준수하는데 핵심적인 유용성을 제공할 수 있음<br>– 정보보호는 대중과 정책 입안자의 관심사 | 위협(THREATS):<br>– 의사소통이 환자에 집중되면 프로젝트의 전임상 초점이 잘못 해석될 수 있음<br>– 프로젝트의 결과를 아직 파악하기 어렵거나 보고할 중요한 결과가 없을 수도 있음<br>– 일반 미디어는 AI, 개인정보보호, 기술 독점 등의 문제를 중점으로 보도하고 있음(민감 데이터의 AI 활용 적대감 조성) |

❑ 시사점

- (산업 및 사회적) EU MELLODDY 사업은 성공적으로 목표하는 바를 정확하게 데모하였으며, 프로젝트의 초기 목표인 제약사의 지식재산권이나 연구 비밀을 보호하면서 제약사 간 협력이 기계학습 모델로 가능하다는 것을 보여주었기에 향후 제약산업의 많은 변화를 주도할 것
- (참여자 별 혜택) ICT 기업에서는 AI와 ICT 기술을 활용하여 분산 데이터 협력 플랫폼이라는 새로운 솔루션을 개발하여 신기술을 확보했고, 제약기업은 산업의 불문율이었던 제약사 간 비밀 유지 협업이 가능하며, 여러 회사의 지식이 축적돼 성능이 향상된 모델(도구)을 획득할 수 있었음
- (기술적 개선) 여러 기관의 공통 데이터를 학습하기 위해 다양한 노력을 하였는데, 화합물의 연관된 활성 값(단일 작업)들을 하나의 모델로 학습하는 Multi Task learning(다중작업 학습) 기술과 연합학습 모델을 공유할 부분과 비밀로 나눠 보안성을 강화한 Trunk-head 모델을 고안함
- (한계) 신약개발 초기 단계의 약물 동태(ADME) 예측 분야에서 개별 기관 데이터로 학습한 모델보다 여러 기관이 협력한 모델의 성능이 더 좋아진다는 시나리오를 증명하였을 뿐, 신약개발 전 단계의 민간 협력 기반 AI 기술 적용 가능성을 보여주는 사례는 아님 (약물 탐색 단계만 시도)

## 4.3. AI를 위한 데이터 협력 기술

### 4.3.1. 연합학습 등장 배경

❑ 데이터와 인공지능이 산업을 주도

❍ 4차 산업 혁명에서 주목받는 기술들은 아래 그림과 같이 모바일 장치, 클라우드 컴퓨팅, AR, IoT 플랫폼, 위치 인식 기술, HCI, 인증과 사기 탐지, 3D 프린팅, 스마트센서, 빅데이터 분석 등이 있음

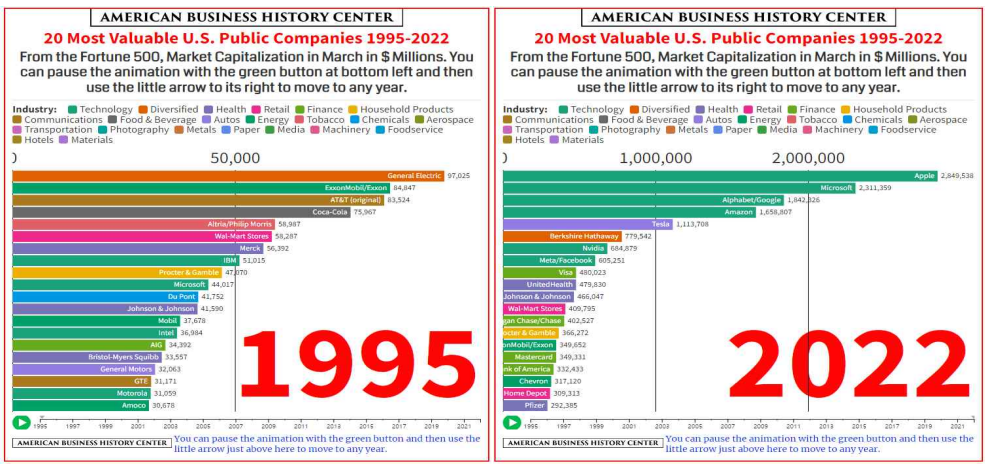
❍ 4차 산업 혁명을 주도하는 것은 결국 데이터라고 언급되며, 과거보다 더욱 데이터의 중요성이 강조되고 있음



(출처: ICTworks, 5 Problems with 4th Industrial Revolution)

그림 35. 4차 산업 혁명의 인포그래픽

❍ 인터넷이라는 기술이 등장하면서 세상의 모든 사람이 인터넷으로 모여들면서 수많은 데이터가 발생하고 새로운 비즈니스가 창출되었음




(출처: American business history center)

그림 36. 1995년에서 2022년 사이의 시가 총액 상위 20위의 변화

❍ 인터넷에 의한 데이터의 힘은 1995년부터 2022년도까지의 전 세계 시가총액 상위 20개 기업의 순위로 확인할 수 있는데, 1995년 제조업 중심에서 2022년에는 IT 기업(빅테크)으로 상위 기업의 업종이 바뀔 정도의 큰 파급효과가 있었음

❑ 데이터 유출 문제

❍ 인공지능은 많은 데이터를 입력으로 발전하는 기술로 추천 결과 생성을 위해 사용자들의 사용 정보, 개인정보 등을 적극적으로 활용하면서 고성능 인공지능을 탄생시켜 편리한 추천 서비스를 제공할 수 있었음

❍ 하지만, 몇몇 서비스 제공자는 데이터를 사용자로부터 무분별하게 수집, 활용하고 있었으며 심지어는 수집한 개인정보를 팔아서 수익을 창출하는 등의 비윤리적인 행동이 있었다는 것이 밝혀졌음

❍ 일련의 사건들로 인해 사람 개인으로부터 수집되는 정보가 엄청난 부가가치를 창출할 수 있으며 이에 대한 권리 보장의 필요성을 인지하게 되었음

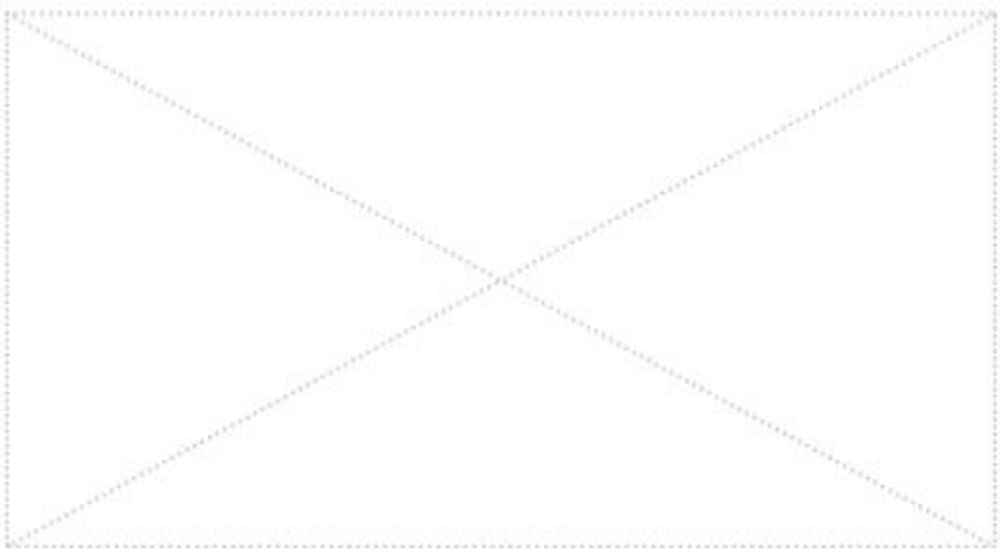


그림 37. 인공지능 기술에 대한 소비자의 우려

❑ 데이터 비즈니스 산업에서 발생한 유출 문제와 개인정보 보호법의 도입

❍ 유럽연합의 GDPR이 제정되기 바로 전 해인 2017년 상반기만 해도 전 세계 918건의 데이터 유출, 19억 개의 데이터 침해 사고가 일어났음

❍ 유출 수준 인덱스의 보고서에 따르면 2013년 이후 2017년 상반기까지 90억 개 이상의 데이터 기록이 유출된 것으로 파악되며, 평균적으로 개인정보가 하루에 1,000만 개가 유출되고 있었음

❍ 대형 개인정보 유출 사건으로 유럽의 GDPR을 필두로 호주의 프라이버시법(데이터 유출 의무 신고), 한국의 개인정보 보호법 등의 개인정보를 보

호하는 법·제도를 구축하였음

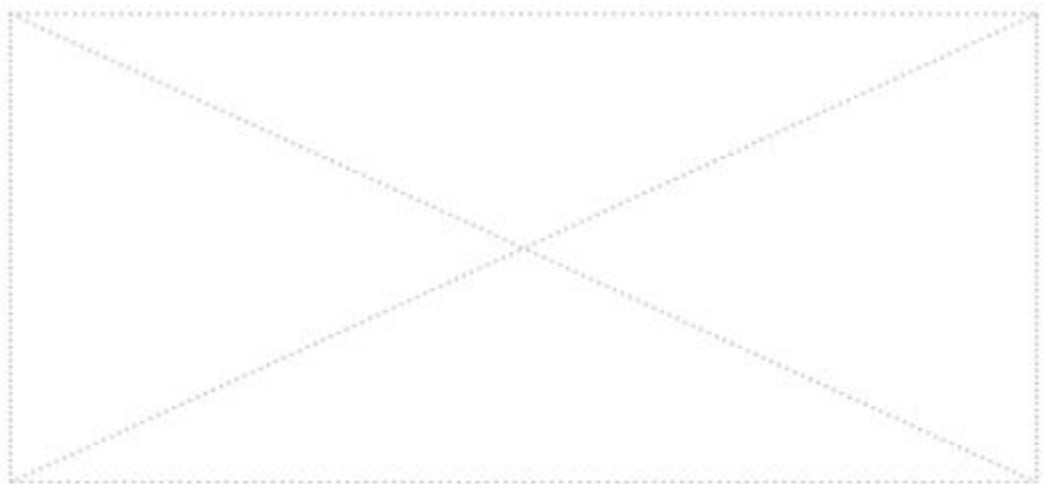
- 수집할 데이터에 개인정보가 포함된 경우, 개인의 동의를 받아야 하며 동의 이외의 목적으로 활용 불가하고, 개인이 원하는 시점 언제든지 개인정보를 삭제하는 기능을 갖춰야 하며, 위반 시 벌금형, 징역형까지도 처벌

❑ 인공지능 기술과 충돌하는 개인정보 보호법

- ❍ 인공지능에는 많은 데이터가 필요하나, 법·제도로 인하여 개인으로부터 많은 데이터 수집이 어려워졌고 보안 처리가 필요해짐(가명화, 익명화)
- ❍ 과도한 정보보호 처리된 데이터를 인공지능 모델의 학습데이터로 활용했을 때 모델 성능 저하 요인이 되었음
- ❍ 인공지능 빅데이터 학습과 개인정보보호 이 두 가지가 양립할 수 있는 방법론이 필요하게 되었음

❑ 연합학습의 등장

- ❍ 개인정보를 보호하면서 인공지능 모델을 학습할 수 있게 하는 방법으로 구글에서 2016년에 처음 제안된 연합학습 기술이 주목받기 시작하였음
- ❍ 초기에는 구글의 지보드(Gboard)라는 스마트폰 키보드 앱의 문장 자동완성 기능을 개선하기 위해 많은 사용자의 장치 데이터를 연합 학습하려는 방법이었음
- ❍ 연합학습은 데이터를 한곳에 모으는 중앙 집중식 학습 방법과 달리 모델과 가중치만을 공유하여 공통 모델을 협력 학습하는 방법으로 원본 데이터가 외부로 나가지 않는다는 장점이 있음



(출처: Kaissis, et al, End-to-end privacy preserving deep learning on multi-institutional medical imaging)

그림 38. 일반적인 중앙 집중식 학습(a)과 연합학습의 차이(b, c)

### 4.3.2. 연합학습 개요

❑ 연합학습

- ❍ 연합학습은 기기나 기관 등 여러 위치에 분산 저장된 데이터를 직접 공유하지 않고, 협력해 인공지능 모델을 학습할 수 있는 분산형 기계학습 기법
- ❍ 일반적인 인공지능 모델은 각 클라이언트(개인 기기, 기관 등)가 보유한 데이터를 중앙 서버에 모아 일괄 학습 과정이 수행되어 완성됨
- ❍ 연합학습은 개별 데이터를 중앙 서버로 전송 또는 공유하지 않고 중앙 서버의 글로벌 모델을 클라이언트로 보내 각각 보유한 데이터로 모델을 훈련하고 학습 결과를 중앙 서버로 보내며, 중앙 서버는 개별 클라이언트에서 받은 학습 결과를 집계하여 글로벌 모델을 갱신함
- ❍ 이 과정을 반복함으로써 중앙 서버의 글로벌 모델은 점점 일반화되고 클라이언트의 로컬 모델 또한 정확도가 향상됨

❑ 연합학습의 특징

- ❍ 연합학습에서 데이터를 공유하지 않음으로써 개인정보를 보호하고 통신 비용을 최소화하며, 원본 데이터를 안전한 상태로 로컬에 유지한 채로 AI 모델 학습이 가능함
- ❍ 연합학습 접근 방식을 사용하면 기밀 데이터를 서로 직접 공유할 필요 없이 여러 조직이 모델 학습을 위해 협력할 수 있으며 연합학습은 의료 및 생명 과학, 제조 및 소매, 약물 발견, IoT 및 비디오 분석과 같은 다양한 분야에서 연구되고 있음
- ❍ 연합학습 완료 이후 학습된 모델을 참여자가 공유해 사용하는 것이 일반적으로 참여자들이 데이터로 모델 학습에 기여한 이익을 공유

❑ 연합학습 동작 절차

- ❍ 글로벌 모델을 한번 학습하는 과정을 라운드라고 함. 첫 라운드에는 아래 1번부터 6번까지의 과정을 수행하고, 2번째 라운드부터는 3번~6번 과정을 반복함
  - ② 클라이언트 선택: 연합학습 참여자(클라이언트)를 선택하고 통신하기 위해 연결
  - ③ 글로벌 모델 배포: 서버는 공동으로 학습하려는 모델의 초기 버전($G_{M_1}$)을 클라이언트에게 배포
  - ④ 클라이언트(로컬) 학습: 각 클라이언트는 전송받은 초기 글로벌 모델($G_{M_1}$)을 로컬 모델($G_{M_1} = L_{M_1}^{1,\ldots,N}, N$: 클라이언트수)로 사용, 보유한 로컬 데이터로 로컬 모델을 학습시킴

- ⑤ 로컬 모델 파라미터 공유: 각 클라이언트는 업데이트된 로컬 모델 파라미터를 서버로 전송
- ⑥ 파라미터 취합: 서버는 각 클라이언트에서 받은 모델 파라미터를 집계($G_{M_2} = AVG(\triangle W_1, \triangle W_2, \triangle W_3)$)하여 글로벌 모델을 갱신

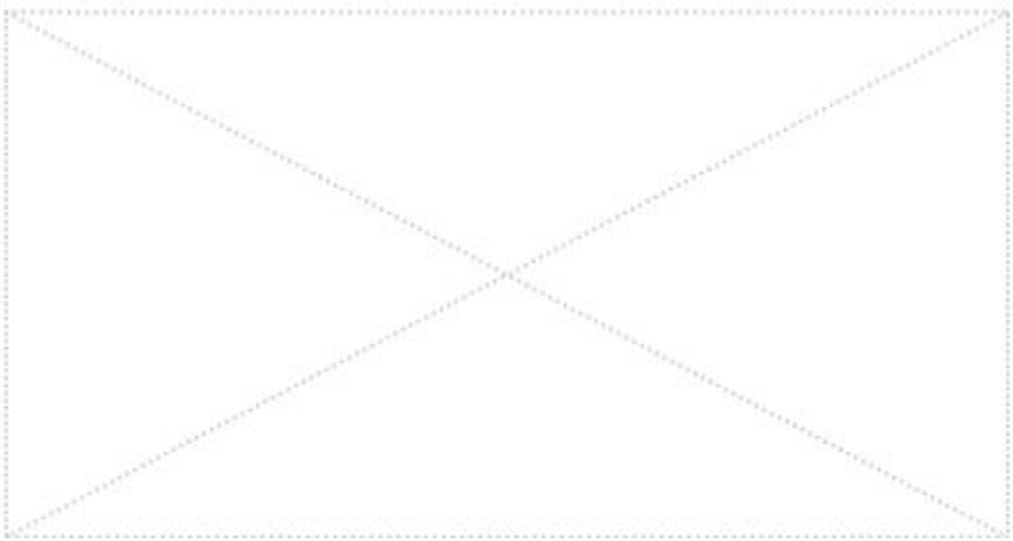


그림 39. 연합학습 절차 예시

### 4.3.3. 연합학습 프라이버시 강화 기술(Privacy Enhancing Technology, PET)

❑ 연합학습에 추가되는 프라이버시 강화 기술

❍ 최근 연구에 따르면, 연합학습 알고리즘만으로는 프라이버시를 온전히 보호하기에는 미흡한 부분이 있다고 언급되고 있음

- 클라이언트에서 서버로 전달하는 가중치 값을 가지고 특정 속성을 가진 샘플이 어느 배치(Batch)에 있는지 확인할 수 있거나, 적대적 생성 네트워크를 통해 학습데이터와 유사한 데이터를 생성할 수 있다는 논문이 발표되었음

❍ 연합학습만으로 해결 불가능한 프라이버시 및 보안 문제를 해결하기 위해 프라이버시 보장형 연합학습(Privacy-Preserving FL)이 연구되고 있음

❍ 프라이버시 보장형 연합학습의 대표 기술에는 차등 프라이버시(Differential Privacy), 동형 암호(Homomorphic Encryption), 안전한 다자간 계산(Secure Multi-Party Computation) 등이 있음

❑ 차등 프라이버시

❍ 2016년 Dework가 제안한 기술로 차등 정보보호는 원본 데이터에 수학적

노이즈를 추가하여 프라이버시 노출 위험을 낮추는 기술임(애플이 2016년 IOS 10에 이 시스템을 적용했다고 발표했을 때부터 유명해짐)

- ❍ 하나의 개인정보가 전체 자료에 추가로 포함될 때 증가하는 노출 위험을 차등 정보보호라고 정의하고 이를 수학적으로 측정하는 방법을 제안하였음
- ❍ 차등 정보보호 기술은 FedSDG, FedAVG 등 연합학습 알고리즘을 사용하되 학습 파라미터에 노이즈(Noise)를 추가해 프라이버시 노출을 방지함

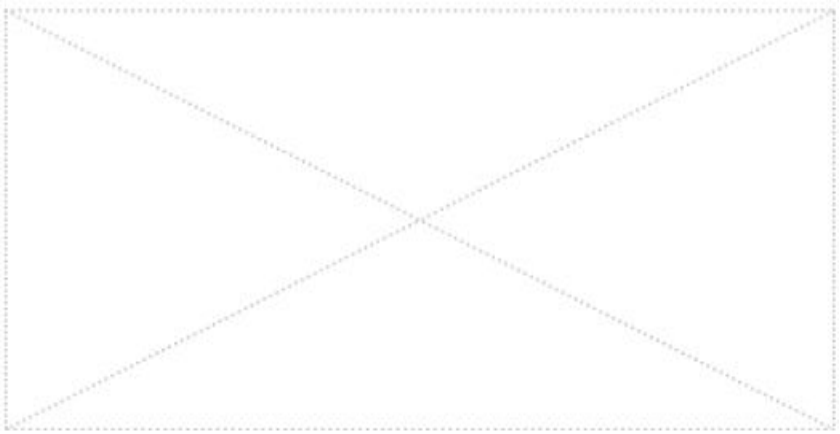


(출처: Azure Machine Learning 차등 프라이버시 정의)

그림 40. 기계학습 과정 중 차등 정보보호 적용

❑ 동형암호

- ❍ 동형암호는 암호화된 데이터를 복호화 없이도 연산할 수 있는 암호 기술임
- ❍ 원본 데이터를 암호화한 상태에서 각종 연산을 했을 때, 그 결과가 암호화하지 않은 상태의 연산 결과와 같게 나오는 암호 알고리즘임
- ❍ 동형암호는 1978년 Rivest, Addleman, Dertouzous 등이 제안하였고, 2009년 Gentry에 의해 동형암호의 안전성이 난제로 환원됨을 증명하였음
- ❍ FedSDG, FedAVG 등 연합학습 알고리즘은 동형암호를 사용함으로써 보안성을 한층 강화할 수 있음 (학습 결과인 파라미터에 동형암호화 적용)



(출처: OpenMined Blog)

그림 41. 동형 암호화 과정

❑ 안전한 다자간 계산(Secure Multi Party Computing)

- ❍ 다자간 계산은 동형암호와 유사하게 각 클라이언트에서 서버로 전달하는 원래의 값을 노출하지 않으면서 전체 합을 알 수 있게 하는 방법임
- ❍ 다수의 사용자가 각자의 비밀 값을 입력값으로 함숫값을 함께 계산하는 기술로 관련 개념은 1982년에 앤드류 야오가 두 명의 백만장자의 재산 대결이라는 문제를 기반으로 양자 간 계산을 제안하였고, 이후 다자간 계산으로 발전하였음
- ❍ 백만장자 A와 B가 MPC 시스템에 각각 자신의 자신을 등록하면 시스템이 자산의 총량을 연산하여 누구의 재산 크기가 더 큰지를 출력해주기에 서로의 재산을 공개하지 않은 채 원하는 결과인 재산의 크기만 얻을 수 있음
- ❍ MPC는 참여자들이 서로의 투입 값을 모르게 하면서, 함수의 연산에 안전하게 참여하면서 민감한 정보를 전달하지 않고 자신의 신원이나 내용을 증명하는 기술로 여러 방면에서 응용이 가능
- ❍ 연합학습에 사용되는 가장 대표적인 다자간 계산 알고리즘으로 Secure Aggregation이 있는데, 보안성과 프라이버시를 강화하면서 서버에 대한 공격에도 강인한 방법을 포함
- ❍ 하지만, SMC와 동형암호화 같은 기술은 상당한 성능 오버헤드를 부과하며 개인 정보보호 딥러닝에 대한 적용은 여전히 미결 문제로 남아 있음



그림 42. 안전한 다자간 계산 방법론 시각화
(출처: Jie Xu et al., 2020)

### 4.3.4. AI 신약개발 연합학습 도입 장점

❑ 신약개발의 데이터 딜레마

❍ 고성능 AI 모델을 만들기 위해서는 대량의 고품질 데이터가 필요한데, 신약개발 분야의 편향된 소규모 데이터의 활용 환경과 대조됨

❍ 신약개발에는 과학적 실험의 측정이 필요하고, 이 과정에서 발생하는 데이터의 편향과 데이터 획득의 어려움은 신약개발에서의 AI 역량을 크게 제한함

❍ 데이터 획득의 어려움은 과학적 실험으로 데이터를 생성하는 과정의 비용 발생, 이러한 연구 비밀을 기밀로 유지해야 한다는 점에서 기인함

- ADME/Tox는 매우 표준화되고 좋은 품질을 가지고 있어 AI 모델 학습 시 고성능 예측이 가능하나, 일부는 기밀성으로 최신 딥러닝 모델에 활용 불가

❍ 과학적 실험 측정에서 발생하는 데이터 편향도 AI 신약개발에 혼란을 초래

- 같은 과학적 실험에서 측정한 특정 분자의 속성값도 데이터 출처가 다르다면 값의 차이가 큰 경우가 존재(연구자, 실험 환경 등의 변인)
- 실험 측정으로 기록된 값은 일반적으로 반복측정 후 기록하기에 분산이 최소화된 값으로 판단. 기록된 값의 차이는 데이터 편향으로 간주

❑ 분산되고 산재한 소규모 데이터를 취합하는 효과

- 연합학습은 다수의 참여자가 데이터 공유 없이 모델을 공유하는 시나리오
- 게다가 연합학습의 개별 참여자는 각각 맞춤형 모델을 가질 수 있도록 공동 학습한 모델의 미세조정(Fine-tuning)이 가능하여 중앙집중 학습에서 발생하는 표적 데이터 불일치 문제를 해결할 수 있음
- 신약개발 데이터의 연구 기밀성과 지식재산권 보호 요구사항을 충족시키면서 소규모 데이터들이 취합된 공동 모델을 만드는 효과가 있음

❑ 편향된 데이터의 딜레마 해결

- 기계학습이나 딥러닝과 같은 AI 기술에서는 데이터 출처의 차이를 평균, 중앙값, 또는 다수 투표와 같은 방식으로 균일하게 만들어 예측하므로 이러한 편향의 딜레마를 해결할 수 있음
- Zhaoping Xiong는 신약개발 데이터(용해도, 인산화효소 활성 저해, hERG 채널 활성 저해)를 사용해 편향된 조건 Non-IID[16]에서 연합학습으로 예측 모델을 학습하고 개별 데이터 학습 모델과 성능을 비교 분석했음
- 개별 참여자의 데이터 수와 분포의 편차에도 참여자 데이터로만 학습한 모델보다 연합 학습한 모델이 대체로 성능이 높았음

16) Xiong, Z., Cheng, Z., Lin, X., Xu, C., Liu, X., Wang, D., ... & Zheng, M. (2021). Facing small and biased data dilemma in drug discovery with enhanced federated learning approaches. Science China Life Sciences, 1-11.

## 4.3.5. 연합학습 응용 분야 및 사례

❑ 신약개발

❍ 심장 독성 예측 모델 연합학습 활용 사례(Effiris)

- Lhasa Limited에서 개발한 FL 솔루션이며, 2차 약리학 예측에 적용
- 연합학습 모델의 평가를 위해서 8개의 대형 제약회사의 실제 데이터를 사용하여 hERG 채널 활성 저해 분류 작업에 대해 평가되었으며, 개별 파트너 모델보다 정확도가 17% 향상되었음



그림 43. Effiris의 민간, 공공 데이터를 활용한 hERG 저해 모델의 성능 비교

❍ 임상시험 환자 선별 모델 연합학습 활용 사례(Bitfount)

- 연합학습 플랫폼과 연령 관련 황반변성에 대한 최첨단 AI 바이오마커 모델을 결합하여 제약사가 운영하는 임상시험 모집을 촉진하는 프로젝트를 수행[17)]
- Moorfields Eye Hospital NHS Foundation Trust에서 정기적으로 수집되는 안구 관련 영상 데이터를 활용하여 황반변성 바이오마커 모델을 개발하고, 병원의 데이터는 외부로 전송할 필요 없이 AI 모델을 학습할 수 있도록 연합학습을 활용
- 서로 다른 위치에 격리된 데이터를 학습한 진단모델 훈련을 연합학습으로 수행하고 모델을 배포하는 것을 목표로 하고 있음

17) Bitfount blog,Bitfount awarded Innovate UK grant to transform clinical trials with distributed data science, 2022.12.17

❑ 신약개발용 kMoL(Elix 사와 교토대학 공동 개발 플랫폼)

❍ kMoL은 신약개발, 바이오 분야에서 분자를 대상으로 한 기계학습 모델 구축을 위한 라이브러리[18]임

– 동 라이브러리는 교토대학 연구팀이 오픈소스로 개발한 신약개발용 AI 라이브러리 kGCN에서 확보한 인사이트를 토대로 업데이트한 것임

– 화합물 데이터의 분자구조, 경로(pathway) 등 바이오 분야에서 광범위하게 사용할 수 있는 그래프 구조를 다룬 그래프 신경망(graph neural network)도 포함되어 있음

❍ kMoL 특징에는 연합학습 기능을 탑재하고 있다는 것으로, kMoL의 기능 중 하나로 연합학습 라이브러리 형태(Elix Mila)로 포함되어 있음

❍ kMoL의 연합학습 특징

– kMoL은 AI 신약개발용으로 공개된 플랫폼 중 유일하게 연합학습 기능을 탑재한 기계학습 라이브러리임

- 동 라이브러리를 이용함으로써 화합물 데이터의 기밀성을 손상하지 않은 채 보다 많은 데이터를 활용하여 학습을 수행할 수 있음
- 또한, 학습에 사용되는 데이터양이 모델 정확도에 영향을 주기에 동 라이브러리에 포함된 모델을 기반으로 모델 정확도를 개선할 수 있음

– kMoL의 가장 큰 특징은 그래프 기반의 예측 모델을 연합학습으로 이용할 수 있다는 점임

- 화합물의 분자구조를 그래프 표현 데이터로 입력할 수 있는 예측 모델은 분자구조 전체 정보를 표현할 수 있어 모델 정확도를 더욱 개선할 수 있을 것으로 예상됨
- 또한, kMoL은 ADME(A: absorption, D: distribution, M: metabolism, E: excretion), 독성, 결합 친화성의 데이터 세트로 검증도 수행 가능

– kMoL은 AMED의 DIIA 지원을 통해 “최첨단 AI 기술을 이용한 멀티표적 예측과 구조발생을 조합한 포괄적인 창약 AI 플랫폼 개발” 과제의 일환으로 개발됨[19]

- kMoL의 멀티모달 신경망에 대해서는 NEDO의 “신약개발 효율화·가속화하는 제제 처방 설계 AI 개발”과제 성과도 포함되어 있음
- 대규모 그래프 신경망 개발에 대해서는 PRISM 프로그램의 “신약 창출을 가속화하는 증례 데이터베이스의 구축·확충·창약 표적 추론 알고리즘 개발”성과로 만들어짐

---

18) kMoL 오픈소스, https://github.com/elix-tech/kmol

19) 最先端のAI技術を用いたマルチターゲット予測と構造発生を組み合わせた包括的な創薬AIプラットフォームの開発, https://research-er.jp/projects/view/1117171

❑ 헬스케어

❍ 네덜란드의 회사 TNO

- TNO는 Lifelines와 협력하여 2세~11세 사이의 아동의 2형 당뇨병 출현을 예측
- 데이터는 실험실에서 테스트한 단백질 데이터와 라이프스타일에 대한 사람들의 설문 조사를 포함하여 다양한 소스에서 가져왔고, 이들을 연합 학습했음
- 사람들의 개인정보를 보호하면서 위험 그룹을 식별하고 가능한 빨리 2형 당뇨가 발병하기 전 조기에 통보하는 모델을 구축

그림 44. TNO 2형 당뇨병 예측 소개

❍ 병원(MRI 연합학습으로 자폐스펙트럼 장애 예측 모델 성능 향상)

- 2020년 li et al은 rs-fMRI 뇌 영상 데이터를 사용하여 자폐 스펙트럼 장애(Autism spectrum disorder)를 식별하는 연구를 수행했음
- 52~167명 환자의 데이터를 4개의 다른 병원에서 가져왔고 이미지 분류 모델을 학습했을 때 단일 병원 모델의 정확도 69.5%, 연합학습을 활용한 다수의 병원 모델은 84.9%로 더 높은 성능을 도출하였음



그림 45. MRI 데이터 기반 자폐증 예측 모델 성능 향상을 위한 연합학습 구조와 그 성능

❍ EXAM(Electronic medical record (EMR) chest X-ray AI model) Initiative

- 2019년부터 지속된 COVID-19 팬데믹 위기에 밀려드는 환자의 예후를 신속하게 예측할 수 있는 AI 기술 수요가 발생
- 전 세계 20개의 병원과 연구소가 협력하여 기관별 축적된 환자 의무기록과 흉부X-ray 사진으로 COVID-19 환자의 산소 요구량을 예측하여 치료 수준 결정에 도움을 주는 EXAM 개발에 다기관 연합학습 기술을 활용
- 개별 모델보다 연합학습 모델의 분류 성능 지표(Area Under Curve, AUC)가 16% 개선되었음

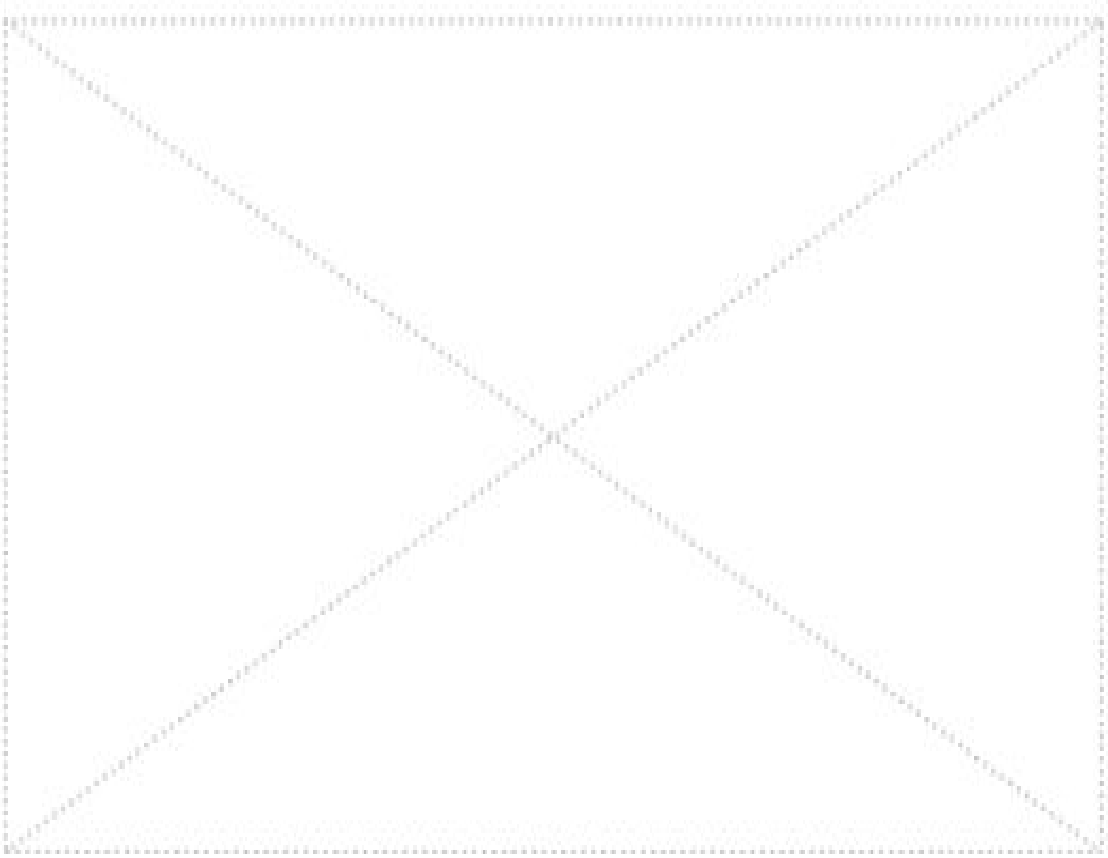


**그림 46.** 20개 기관의 개별 모델(Local)과 연합학습 모델(FL)의 성능 비교

❍ 희소 질환 치료예측 모델 개발을 위한 HealthChain project[20)21)]

- 프랑스의 4개 병원에서 연합학습 프레임워크를 개발하여 배포하는 것을 목적으로 유방암, 흑색종 환자의 치료반응을 예측하는 공통 모델을 개발한 프로젝트임

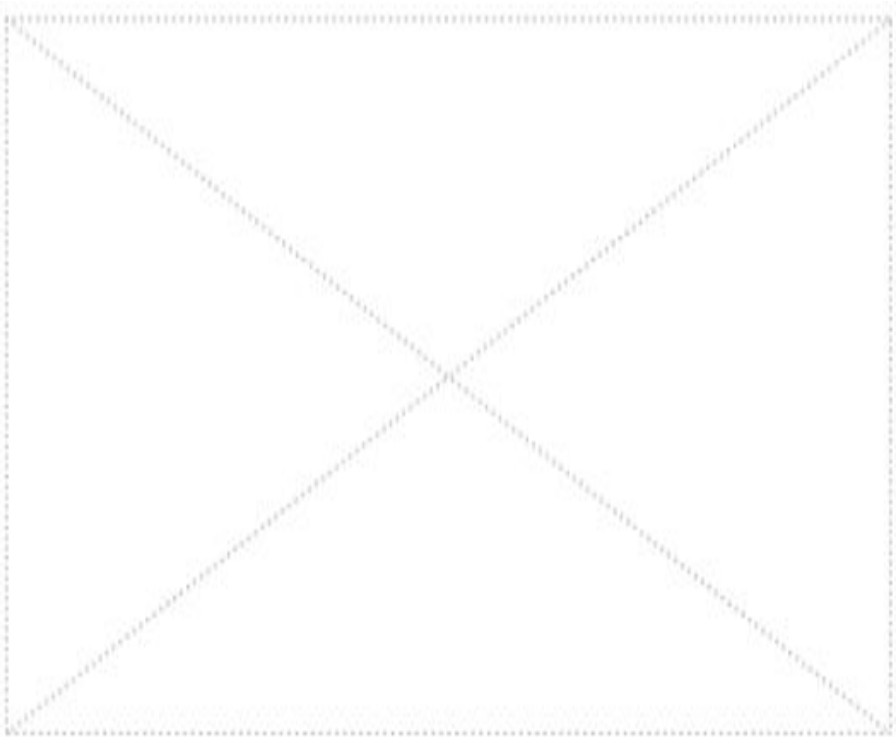
20) Healthchain consortium. https://www.labelia.org/en/healthchain-project
21) R Durga et al., Federated Learning Model for Healthchain System, 2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)

- HealthChain 컨소시엄은 2018년 6월에 출범하여 프랑스 공공부문 투자은행인 BPIFrance로부터 천 만유로 지원금(약 147억)을 받아 진행된 프로젝트임
- 2개의 AI 기업(Owkin, Apricity), 4개 병원(파리병원, 낭트병원 등) Institut Curie, Centre Léon Bérard research centers, 에콜 폴리테크니크, 파리 데카르트 대학, Substra 재단이 참여함

– 유방암[22]과 관련해서는 삼중음성 유방암(TNBC)의 희귀유방암을 대상으로 유방 보존 수술을 목표로 하는 화학요법(NACT)에 대한 민감도 평가임

- 환자마다 다른 조직반응의 이질성에 대해 데이터 부족으로 연구 수행의 어려움이 있었으나, 이를 극복하기 위해 연합학습을 도입함
- 아래 그림 (a)와 같이, ①TNBC 치료과정인 진단(1차 종양 생검 및 조직병리학 분석, MRI 및 CT 스캔) → ②화학요법 NACT 다음의 수술 → ③치료 후 TNM 점수 도출, 조직분석 등으로 치료반응 분석으로 진행
- 그림 (b)에서는 이 치료과정에서 두 개의 코호트 A와 B(두 개의 병원)에서는 그림(a)에서 설명한 변수들과 생검 슬라이드 이미지(WSI)를 이용하여 데이터를 추출하여 NACT 반응 예측을 위한 딥러닝 모델 학습

22) Jean Ogier du Terrail et al., Collaborative Federated Learning behind Hospitals' Firewalls for Predicting Histological Response to Neoadjuvant Chemotherapy in Triple-Negative Breast Cancer, doi: https://doi.org/10.1101/2021.10.27.21264834

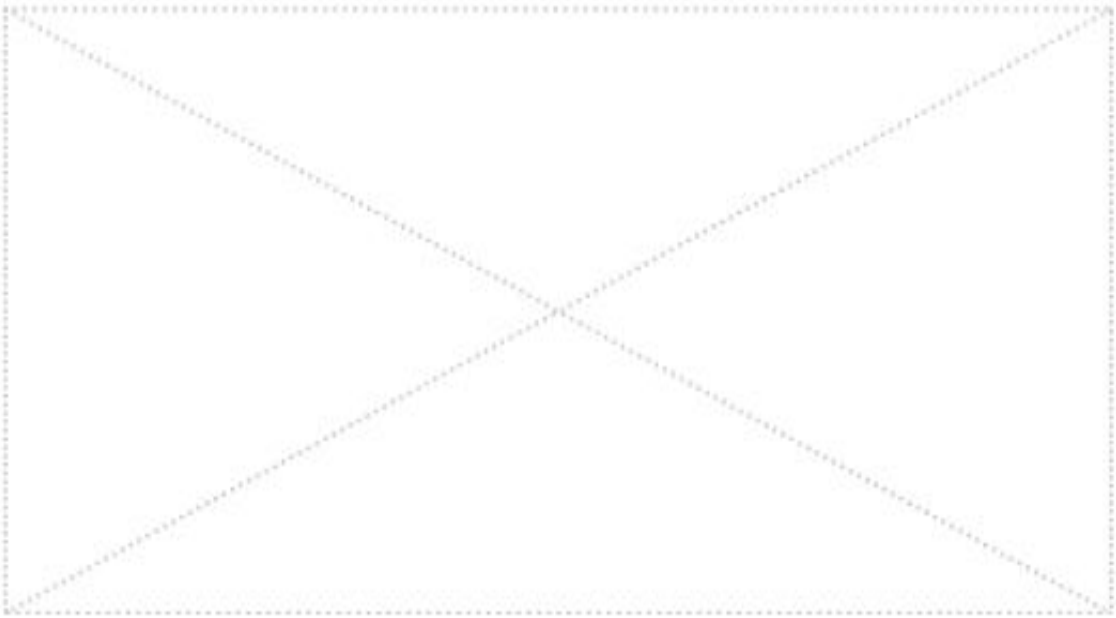


(출처 :ean Ogier du Terrail et al. 2021)
그림 47. TNBC 치료과정 및 병원간 협력 연합학습 연구 개요

❍ Federated Tumour Segmentation(FeTS) Initiative[23)]

- FeTS는 고형암 연구를 위한 연합학습 오픈소스[24)] 개발 프로젝트로, 그래픽 유저 인터페이스가 포함된 오픈소스 연합학습 프레임워크를 개발함
  - 인텔랩과 펜실베이니아 대학 페럴만 의과대학(이하, 펜 메디슨)과 협력하여 29개 국제 보건의료 관련 연구기관이 참여함
  - 동 프로젝트는 NIH 산하 암 연구를 위한 정보 기술학(ITRC) 프로그램을 통해 펜실베이니아 대학 주관으로 3년간 120만 달러를 지원받아 수행됨
- 이를 통해 29개 의료기관의 국제 연합학습이 이루어지고 있고, 동 컨소시엄의 목적은 뇌 신경아교종, 유방암, 간암, 뼈 병변 등 종양 경계 부분 검출을 개선하기 위함임
- 프로젝트는 FeTS 플랫폼 개발(SA1), FL 프레임워크 실행(SA2), FeTS 활용으로 구분되어 진행
  - FeTS 플랫폼 개발에는 직관적인 GUI 개발을 통해 전처리 툴, pre-trained 모델과 융합 전략, 다양한 상호작용 도구, 매뉴얼 고도화 등 수행

23) Sarthak Pati et al., The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research, Phys Med Biol. 2022 Oct 12;67(20):10.1088/1361-6560/ac9449.
24) https://github.com/FETS-AI/Front-End



(출처 : Spyridon (aka Spyros) Bakas 발표자료)
그림 48. FeTS의 FL 프레임워크 실행 부분 UI

- FL 프레임워크 실행에 대해서는 아래의 UI를 통해 각 참여기관은 모델 업데이트만을 공유하고 모델 성능을 향상하는 다양한 추가 데이터에 대한 지식을 얻을 수 있음

– 최근 네이처 발표에서는 초기에 16개 의료기관의 231개 케이스에서 시작하여 71개 의료기관의 6,316개 케이스로 370만 개의 이미지 데이터를 학습하여 아래와 같은 성능 향상 효과를 획득하였음

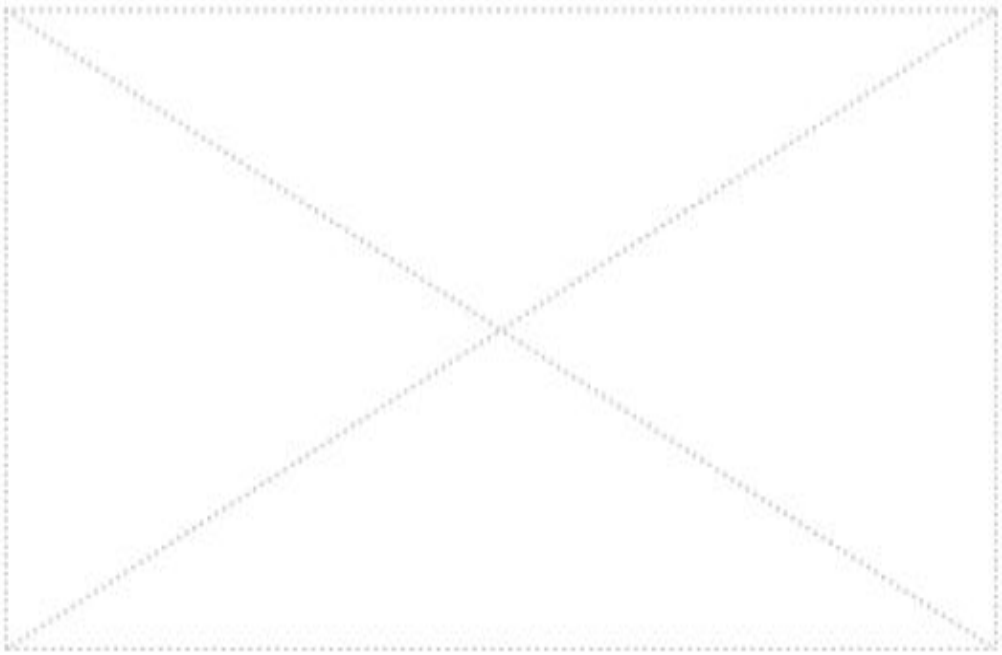
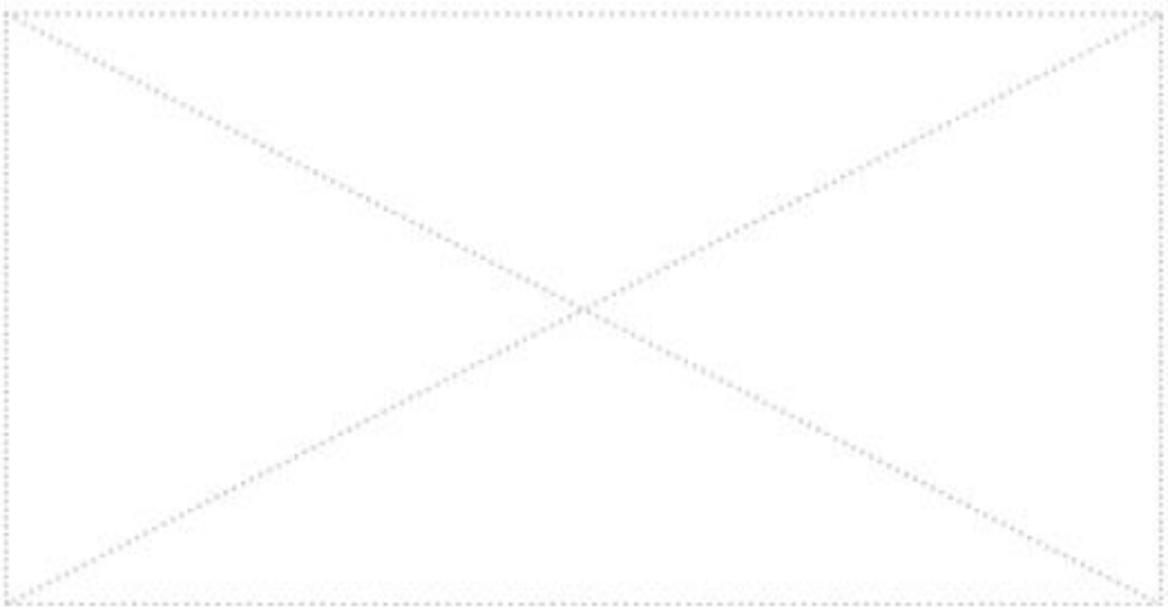


**그림 49.** 연합학습 모델의 성능 향상 도표

❍ 연합학습을 이용한 유방선 평가를 위한 AI 모델 개발 사례[25)]

- NVIDIA는 American College of Radiology, 브라질 이미징 센터 Diagnosticos da America, Partners HealthCare, 오하이오대학, 스탠포드 메디슨의 5개 기관이 협력하여 유방조영술에 나타나는 섬유질, 유방조직 밀도 등을 평가하는 연구 수행
- NVIDIA Clara Federated Learning 소프트웨어를 기반 우선 5개 기관이 각각 2D 유방조영술 데이터를 제공, 총 100,000개의 학습용 데이터 확보
- 마찬가지로 중앙서버에는 글로벌 모델을 보유하고, 각 참여기관의 로컬서버에서는 자체 데이터 세트로 학습할 모델을 배포 받는 형태로 이루어짐
  - 각 기관에서 사전에 합의한 작업을 완료하면 자동으로 로컬 모델의 파라미터를 중앙서버에 다시 보내지고, 중앙서버는 이를 집계하여 업데이트된 파라미터를 다시 각 로컬에 보냄
- 이러한 과정을 몇차례 진행한 후 참여기관에서는 향상된 성능의 모델을 얻게 되며, 실제 연합학습을 통해 기존의 2D 유방조영술 분류 모델을 개선하여 기존 모델보다 더 나은 예측성능을 보임

25) Medical institutions collaborate to improve mammogram assessment ai. https://blogs.nvidia.com/blog/2020/04/15/federated-learning-mammogram-assessment/



(출처: NVIDIA)

그림 50. 5개 기관이 참여한 유방선 평가 AI 모델개발 사례

❑ 금융

❍ MTN(통신사, 고객 이탈률 예측)

- MTN은 2억 7,200만 명 이상의 가입자에게 통신 서비스하는 아프리카 기업으로 550만 명 이상의 사용자를 보유한 메시지전달 앱인 Ayoba와 파트너십을 맺어 고객 이탈률 예측에 연합학습 솔루션(Pysyft)를 사용함
- 개인정보보호 연합 모델로 학습하여 기존 RF 모델 예측성능과의 비교 시 향상 학습 사이클마다 약 1~2%씩 더 높은 성능을 달성

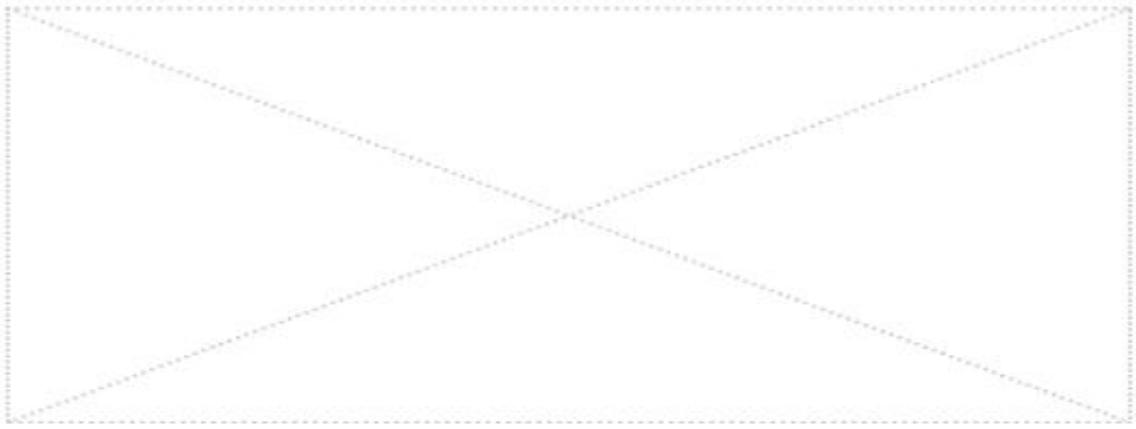


그림 51. MTN 연합학습 프레임워크 구조

❑ 제조업

❍ Apheris는 연합학습 플랫폼 제공 업체로 제조업체와 협력하여 생산 품질

향상, 기계 가동 중지 시간 및 위험 감소 설비 종합 효율 최적화, 공급망 프로세스 최적화 달성을 목표로 수행 중-최근 제품의 이상 탐지 모델에서 실제 프로세스에서 오류의 97%를 식별했음



그림 52. Apheris 제조 연합학습 솔루션 예시

❑ 응용 분야

❍ 서비스, e-커머스: 고객의 취향, 소비 패턴을 파악하여 아이템을 추천하는 데 연합학습 활용 (알리바바 연합학습 이용을 위한 플랫폼 공개)

- 알리바바에서는 포괄적 기능을 제공하는 사용하기 쉬운 연합학습 플랫폼 (Federated Scope) 출시함
- 구글 픽셀(스마트폰)에서 음악 재생 어플 나우플레잉 지원에 활용(디바이스 환경의 음악 DB를 활용해 주변에서 재생 중인 음악을 식별)함

❍ 공공 국방: EU Remap 프로젝트

- 실시간 적응형 항공기 정비를 위한 실시간 상태 예측 모델을 만들기 위해 연합학습을 활용하는 시스템 IFHM: Integrated Fleet Health Management 과제를 수행하고 있음

❍ 스마트 홈: IoT 센서로부터 사용자의 행동을 예측하여 서비스로 연결시키기 위한 연합형 사용자 행동 예측 모델 연구 수행하였음

- ClusterFL: a similarity-aware federated learning system for human activity recognition

❍ 자율 주행: 자동차를 개별 클라이언트로 고려하고, 자동차로부터 수집되는 정보를 연합 학습하여 자율 주행을 위한 예측 모델 개발 연구를 수행 중임

- CoFed-DLAD: 자율 주행을 위한 협업 인식 및 연합 ML 학회

### 4.3.6. 신약개발 분야의 연합학습 사례

- ❑ 최근 5년간 연합학습 기술의 신약개발 분야 적용 가능성을 파악하기 위해 여러 프로젝트가 시작되었고, 제약기업의 데이터를 연합학습에 사용한 프로젝트는 EU MELLODDY와 Effiris 총 2개임
- ❑ 최근 학계 인사의 발표에 따르면, 신약개발 분야의 AI 도입에는 대량의 고품질 데이터가 필요하나, 현재 신약개발 파이프라인에 편향된 소규모 데이터의 활용 환경과 대조되며, 연합학습이 이러한 문제해결의 키라고 언급(Zhaoping Xiong)

표 26. (23년 기준) 최근 5년간 신약개발 분야 연합학습 도입 프로젝트

| 프로젝트 명 | 플랫폼 | 기관 | 응용 | 배포 |
|---|---|---|---|---|
| MELLODDY(Machine Learning Ledger Orchestration Drug DiscoverY) | Substra | Innovative Medicines Initiative (IMI) | 10개 제약기업에서 멀티태스킹 아키텍처를 사용하는 대규모 연합학습에 대한 실제 개념 증명 | 실제 데이터 검증 |
| Effiris | Cronos | Lhasa Limited | 2차 약리학(hERG채널 활성 저해 예측)에 적용되는 협업 모델을 구축하기 위한 실제 플랫폼 | 산업 수준 제품 |
| FL-QSAR | Pytorch, Crypten | Shanghai East Hospital | 15개의 QSAR 작업에 대한 일반 벤치마크 데이터를 활용해 실험했으며 연합학습모델이 개별 파트너 모델보다 성능이 우수함을 증명 | 벤치마크 시뮬레이션 |
| FedChem | FedChem | University of Rochester | 9가지 분자 특성 예측을 함에 있어 데이터 불균형을 문제를 완화하는 FLIT 방법을 평가 | 벤치마크 시뮬레이션 |
| FedGraphNN | FedML | University of Southern California | 15개의 QSAR 작업에 그래프 신경망 기반의 연합학습 사용 시 이점을 평가 | 벤치마크 시뮬레이션 |
| FL-Disco | FL-Disco | University New Mexico | 원하는 특성을 보유한 분자 생성 가능성 평가를 위해 2가지 용해도 데이터 세트를 활용 | 벤치마크 시뮬레이션 |
| kMol | kGCN | Elix | Tox21, Chembl 데이터 세트를 포함한 분자 특성 예측 모델을 평가하였고 FL | SW라이브러리, 벤치마크 시뮬레이션 |

## 4.4. 정책 환경 분석

❑ 국내외 정책 환경

❍ (국내) 첨단바이오와 인공지능을 12대 국가 전략기술로 선정을 비롯해 AI·빅데이터 등 신약개발 분야 R&D 지원 강화

- 제5차 기본계획의 주요 방향으로 전략성 강화와 민간 중심을 설정, 전략2에서는 기업수요 기반의 민간 주도 강조
- 제3차 제약산업 육성·지원 종합계획에서도 R&D 부문에 대한 전략으로 연합학습 기반의 K-MELLODDY 사업에 대한 추진계획 명시
- 민간에서도 AI 신약개발 전문위원회 출범 등 AI 신약개발 분야의 협업 생태계 가속화 예상
- 20년 데이터 3법이 통과되면서 개인정보 활용 분야가 연구·개발 영역으로 제한하여 공유 가능성을 열어두었으나, 기관별로 서로 다른 데이터의 이질성 등의 문제는 남아 있음

❍ (미국) ARPA-H 신설과 AI 부문 강화, NIH, DARPA 등에서 '17년부터의 혁신적인 신약개발 AI R&D 프로그램과 FDA의 규제과학을 통해 지원 지속

- NIH/NIDA는 자체 프로그램을 통해 신약개발 전 주기에 걸쳐 연구를 계속해서 지원하고 있고, DARPA는 17년부터 20년까지 선제적으로 지원하여 도출된 연구 결과들이 AI 기반 신약개발에 계속 영향을 끼치고 있음
- FDA는 AI 신약개발 관련 규제지원을 위해 신약개발 과학지원 사업을 론칭하여 임상시험 단계에서 발생할 수 있는 문제 등을 다룸

❍ (EU) MELODDY 프로젝트를 통해 세계 최초로 산업 규모의 연합학습 기반 신약개발 플랫폼 구축사업 추진

- 동 사업은 10개 제약사의 화학 라이브러리를 기반으로 후보물질 발굴을 위한 AI 모델을 탑재한 플랫폼을 개발하였고, EU의 IMI 펀딩으로 추진된 민간-공공 협력의 프로젝트임

❍ (일본) 일본 정부는 15년부터 신약개발 기반 강화를 위해 의약품과 관련 화합물 통합 DB 구축과 다양한 분석 방법을 개발하는 AI 개발사업 추진

- 공공과 제약기업 협력을 통해 약물 동태와 독성을 비롯해 화합물의 최적화, 분자설계가 가능한 기술개발 등 지원

## 4.5. 국내외 시장 규모·산업 동향

❑ (해외) AI 신약개발 세계시장 규모는 2021년 4억 1,320만 달러로 추정되며 2027년까지 매년 45.7%씩 성장 예상

- ❍ 면역항암제 분야가 전체 시장의 44.5%로 가장 큰 비중을 차지, 이는 암의 높은 유병률과 효과적인 항암제 수요가 계속 증가하고 있기 때문임
- ❍ 글로벌 제약사들의 적극적인 투자로 AI 신약개발 시장의 치열한 경쟁 전망되고, 최근 빅테크 기업까지 AI 신약개발 관심 증대, 시장 성장 전망
- ❍ 연합학습과 관련해서는 Sanofi가 자사의 종양 분야의 파이프라인을 확대하기 위해 연합학습 기반 신약개발 전문기업인 Owkin에 투자

❑ (국내) 국내 AI 신약개발 시장 규모는 아직 미미한 수준이나, 정부와 제약업계의 협업으로 AI 개발 생태계 구축에 노력

- ❍ 글로벌 빅파마에 비해 자본이 적은 국내 제약기업들의 AI 기술 도입이 늦어질 수 있으므로 빅파마 들과의 격차를 좁히기 위한 정부지원과 AI 신약개발 생태계 구축 노력이 필요한 상황
- ❍ 현재 국내 기업들은 유효물질 탐색 단계를 중심으로 서비스가 이루어지고 있고, 약물 최적화 단계에서 AI 활용의 성공사례를 만들기 위해 노력 중

❑ SWOT 분석결과

- ❍ AI 신약개발에 대한 세계적인 열풍으로 국내 제약기업도 AI 역량을 확보하고자 노력하고 있으나, 데이터 부족과 전문인력 부족 등으로 어려운 상황임
- ❍ 연합학습 기반 신약개발 도입은 국내 기업들이 데이터의 공유와 고성능의 AI 모델 확보 기회를 제공함으로써 정체된 AI 신약개발 생태계 활성화

| 강점 | 약점 |
|---|---|
| • AI 활용 신약개발에 대한 정부의 R&D 정책 확대<br>• 신약개발 관련 데이터의 지속적 축적<br>• 국내 제약기업들의 AI 활용에 적극적 | • 국내기업의 자금력 열세로 데이터/ AI 전문인력 부족<br>• 기관별로 분산된 데이터 접근 어려움<br>• 국내 AI 활용 신약개발 시장 정체 |
| **기회** | **위협** |
| • AI 활용 신약개발 세계시장 확대 전망<br>• 연합학습을 통한 기존 AI 모델개발에서 어려운 데이터 공유문제 해결, 성능 우수성 입증 | • 신약개발 등 의료분야에서 주요국의 연합학습을 통한 다양한 기회 모색<br>• 데이터 표본의 한계로 AI 모델개발의 편향성 문제 지속<br>• 챗GPT와 같이 혁명적인 기술이 등장하는 경우<br>• 연합학습 절차의 새로운 기술이 등장하는 경우 |

**제약산업의 국제경쟁력 확보를 위해 국내 기업의 대규모 협력을 통한 연합학습 기반 신약개발의 가속화 수단 필요**

❍ 다만, 연합학습은 이제 막 상용화 단계에 접어들고 있는 기술로, 기술적으로 아직 해결해야 할 과제가 다수 있음

– 데이터 유출 가능성은 없으나, 중앙서버와 로컬서버 간 모델 공유 시 가중치와 모델이 동시 유출 가능성은 남아 있어 보안 문제는 반드시 해결되어야 함 => 기본적으로 암호화 기법을 활용해야 하며, 유출에 대한 부분은 블록체인 분산 원장 기술로 보완해야 함

❍ 위협과 대응 방안

– 데이터 대신에 각 로컬서버에서 학습한 모델(파라미터)을 결합하기 때문에, 재조합 리스크가 존재할 수 있음 => Split Learning을 통한 학습 모델까지도 공개 비밀 방법 활용하여 재조합 리스크 제거

– 챗GPT와 같이 혁명적인 기술이 등장하는 경우 => 개발하려는 연합학습 기반 신약개발 가속화 플랫폼은 분산 데이터를 AI 모델 학습에 활용하는 방법으로 신규 모델 아키텍처가 등장한다면 플랫폼에 활용되는 모델을 바꿔서 연합학습하는 과정을 거치면 됨

• 기술적 접근) GPT에서 활용된 기술인 Transformer 기술을 이미 신약개발 도메인과 화합물 구조 학습에 적극 활용 중이며, 화합물의 구조 정보만을 활용할 수 있는 등 추가 개선의 여지가 많이 존재

❍ 연합학습 절차의 새로운 기술이 등장하는 경우 => 연구 기간 내에 연합학습 절차의 새로운 기술이 등장할 경우, 학습 절차를 수정할 수 있도록 기능을 사전에 구축하면 됨. 네트워크 구성을 변경이 쉽도록 네트워크 config 파일 형식으로 네트워크 절차 구성이 가능하도록 네트워크 기능 구현으로 해결 가능

- 데이터 공유는 어렵지만, 모델 개발을 위해서는 기관별로 데이터가 서로 다르다는 점을 인정하여 이질성 문제를 해결해야 함
- 실제 서비스로 활용되기 위해서는 시스템의 재현성이 필요하며, 이를 위해서는 학습이 진행되는 과정 동안 모든 이력 추적이 필요함
- 다수의 이해관계가 있는 기관이 참여하기 때문에 지속적인 협력관계를 유지하기 위해서는 인센티브 메커니즘 개발 필요

# 제5장 한국형 연합학습 기반 AI 신약개발 플랫폼 사업 기획

## 5.1. 사업추진 배경 및 필요성

### 5.1.1. 신규 사업 기획 배경

❑ 신약개발 민간 데이터 공유 현황 분석

- ❍ 복잡한 신약개발 과정에서 발생하는 다양하고 중요한 데이터가 연구 비밀, 이해관계, 제도적 한계, 정보보호 이슈 등으로 고립됨
- ❍ 데이터의 고립 문제해결을 위해 세계적으로 민간-공공 파트너십의 컨소시엄을 구성하고 다양한 목적으로 데이터 협력을 수행하고 있음
- ❍ 4차 산업혁명과 더불어 신약개발 데이터 공유의 최종 목적지이기도 한 신약개발 분야에 AI 기술을 적용한 성과가 점차 도출되고 있으나, 여전히 공공과 민간 사이의 협력 또는 일대일 협업으로 산업적 파급력이 부족함

❑ 세계 최초의 민간-민간 데이터 기반 협력 사업 등장

- ❍ 유럽연합의 MELLODDY 사업은 민간 기업 사이의 데이터 기반 안전한 협력이 블록체인과 연합학습 기술로 가능하다는 것을 증명하고자 해당 기술들로 구축된 플랫폼을 개발하였음
- ❍ 연합학습 기술로 데이터 기반 협력의 가능성을 보여주기 위해 민간 참여자들이 일정량 이상 보유하고 있어야 하며, 혹시나 모를 데이터 유출 문제에도 덜 민감한 약물 동태(ADME) 및 독성(Toxicity) 데이터를 사용
- ❍ 연합학습을 통해 10개 기관의 민간 데이터를 로컬에 유지한 채로 모델 공유만을 통해 동일 목적의 AI 모델을 만들 수 있었고, 개별 기관 데이터로 학습한 예측 모델보다 연합 학습한 예측 모델의 성능이 전반적으로 높았음
- ❍ 신약개발 초기 단계인 약물 발견과정의 민간-민간 데이터 협력이 가능하고 협력의 효과가 있음을 초기 증명했을 뿐 신약개발의 여러 단계로 확장하거나, 고립된 데이터 협력이 정말 필요한 분야에 적용하지도 못했음

❑ 연합학습을 활용한 데이터 고립 문제 해결이 절실히 필요한 분야

- ❍ 현재 대한민국은 연구, 개인정보, 지식재산권이 포함된 데이터를 외부에서 사용하기 위해서는 매우 복잡한 절차를 거치거나 데이터에 여러 가지 보안 처리를 해야만 하는 데이터 활용에 제약이 많은 구조
- ❍ 데이터의 공개와 협력을 통한 AI로 산업을 가속화 하자는 외침과 반대로 데이터 활용 제약이 늘어나며 데이터는 더욱 고립되는 역행적 상황
- ❍ 과제 수행기관은 이러한 데이터 활용의 역설적 상황을 해결할 수 있는 열쇠가 연합학습이라고 판단하고 있음

## 5.1.2. 사업추진 배경

❑ 신약개발 경쟁력 열세

❍ 보건 안보와 경제성장을 담보하는 신약 강국 도약이 국가 지상 과제이나 기술 수준과 R&D 투자비의 절대적 열세로 전통적 신약개발 방식으로는 제약 선진국의 추격 요원임

❍ (기술 수준) 미국 대비 78% 수준 3.8년의 기술격차[과학기술기획평가원]

❍ (R&D 투자) 글로벌 대비 1.7% 수준(국내 10대 기업 R&D 투자비 1.4조 원, VS 글로벌 10대 기업의 투자비 82조 원)으로 글로벌 빅파마, 블록버스터 신약을 아직 보유하지 못하고 있음

❍ (시장 규모) 의약품 시장 2022년 25조 원 규모로 신약개발 생태계 조성에 부적합하며, 세계의약품 시장은 1,600조 원 규모로 미국-유럽-일본 제약기업이 성장 주도하고 있음

❑ AI 기술로 열세 극복

❍ 천문학적 비용이 소요되는 신약개발에 AI 기술을 접목, R&D 투자 비용을 절감하고 신약개발 기간을 단축시켜 제약 선진국을 추월하는 퀀텀점프 전략을 시도 중

❍ 20개 제약바이오기업, AI 기반 신약개발 스타트업과 총 51건 이상의 공동연구 프로젝트를 진행 중



그림 53. AI 신약개발 국내 동향(공동연구)
(출처: 제안기관 자체 네트워크 기업 대상 설문 조사 응답 2022년, 11개 사)

❍ 29곳 AI 기반 신약개발 스타트업은 총 6천억 원의 투자금을 유치했고, 11곳은 총 105건의 파이프라인 보유(2022.07 기준)

그림 54. AI 신약개발 투자 유치와 파이프라인 국내 현황

❑ 데이터 활용의 한계

❍ AI 기술은 다량의 다양한 데이터를 활용할수록 성능이 개선되나 각 기관이 보유한 데이터를 연계·활용할 수 있는 체계가 미흡하여 각자 보유하고 있는 데이터를 폐쇄적으로만 활용하거나 특정 기관 간 일대일 협업에 그침

❍ 보건의료 데이터는 대부분 개인정보보호, 지식재산권, 연구 비밀 이슈가 있는 민감 데이터이기 때문에 공유-연계-활용이 어려움

❍ 공공기관, 기업, 병원이 보유한 약리 활성 데이터, 유전체 데이터, 임상데이터, 의료 데이터 등을 다른 기관과 연계 활용하지 못하고 있음

## 5.1.3. 추진 필요성

❑ 데이터 연계 협업 필요성

❍ 경쟁력 있는 AI 신약개발을 위해서는 정부를 중심으로 기업, 병원, 대학 등 이해관계자들의 데이터 기밀은 유지하면서 기관별 데이터를 효과적으로 활용할 수 있는 AI 신약개발 협업 플랫폼이 필요함

❍ AI 신약개발을 가속화하는 협업 플랫폼 구축은 민간이 주도할 수 없으며 정부가 추진해야 성공할 수 있음

❍ 주요국에서는 자국의 신약개발 경쟁력을 위한 신약개발 가속화를 목표로 제약회사 협업 컨소시엄 프로젝트인 EU MELLODDY, 미국 MIT MLPDS, 미국 ATOM, 영국 Medicines Discovery Catapult를 진행 중

표 27. 해외 주요국의 신약개발 데이터 파트너십 추진 현황

| 국가 | 주도 | 프로젝트 |
|---|---|---|
| 유럽 연합 | 정부 주도 | • EU MELLODDY: 10개의 바이오 제약회사의 화학 라이브러리를 효과적으로 공유할 수 있게 하려고 만들어진 17개 파트너의 컨소시엄(더 나은 약물 후보를 식별하기 위해 예측 정확성 개선) |
| 미국 | 민간 주도 | • MIT MLPDS(Machine Learning for Pharmaceutical Discovery and Synthesis): 미국 MIT에서 주도하는 13개의 바이오 제약 회사가 참여하는 협업 컨소시엄으로 소분자 화합물 발견 및 합성 자동화를 위한 협력으로 모든 구성원은 지적 재산을 공유하고, 모든 로열티에 접근을 무료로 제공 |
| 미국 | 정부 주도 | • ATOM(Accelerating Therapeutics for Opportunities in Medicine): AI 기반 약물 발견 개발을 위한 미국 기반의 협력 기구로 약물 발견 기간의 단축을 위한 플랫폼을 만들기 위해 고성능 컴퓨팅, 다양한 생물학적 데이터 및 새로운 생명 공학 기술을 통합 |
| 영국 | 정부 주도 | • Medicines Discovery Catapult: 약물 발견을 가속화 하기 위한 협력 기구로 연구 네트워크를 통해 전세계 연구에서 발생한 기존 공개 및 통제 데이터 세트에 접근 및 화학 실험, 데이터 분석 및 알고리즘을 지원 |

❑ 혁신 기술개발 및 활용 필요성

❍ 연합학습(Federated learning) 기술은 ❶개인정보보호 및 지식재산권 이슈를 극복하면서 ❷다기관 데이터를 연계하는 협력 기술로 ❸AI 신약개발의 저비용 고효율 효과를 신약개발 모든 단계로 확장할 수 있는 시장 선도형 기술임

– 데이터 공개 및 공유 이슈를 해결하는 법적 제도적 장치가 완비되기를 기다리지 않고 AI 신약개발에 필요한 다기관 데이터를 즉시 활용할 수 있는 현실적인 협력 기술

❑ 저비용 고효율 실증 필요성

❍ 연합학습 기반 ADME/Tox 예측 모델을 통해 기업 및 국가 R&D 투자비 4,000억 원을 절감하는 실증연구를 수행하여 머뭇거리고 있는 제약바이오 산업의 디지털 전환과 AI 기술 도입을 촉진하고 저비용 고효율 AI 신약개발로의 패러다임 전환 가속화

## 5.2. 추진 전략 및 계획

### 5.2.1. 사업추진 전략

| 비 전 | 인공지능 기반 신약개발 생태계 활성화 |
|---|---|

| 최종목표 | 연합학습 기반 신약개발 가속화 시스템 구축과 성공사례 창출 |
|---|---|

| 사업1 | 연합학습 기반 신약개발 가속화 프로젝트 사업 |
|---|---|

| | | |
|---|---|---|
| 세부목표 | (세부1) 플랫폼 구축·운영 및 사업 관리 | 플랫폼 사업 관리<br>• 과제지원 및 성과 관리 체계 구축 운영(거버넌스(추진위원회)구성 및 운영)<br>• 플랫폼 ISP 수립, SOP 제공<br>• 플랫폼 교육 기획 및 운영<br>• 데이터 품질 및 생산 절차 관리<br>• 플랫폼 활용 가이드라인 개발<br>• 플랫폼 고도화 방안연구<br>• 플랫폼 확산 지원(경진대회, 홍보)<br>• 신약개발 데이터 공동활용성 확대 연구<br>• 플랫폼 활용확산방안 연구 및 성과분석 |
| | | 플랫폼 구축<br>• 플랫폼 설계(플랫폼 구조 및 기능, 인프라 구조, 연합학습)<br>• 플랫폼 구축(인프라 구축, 플랫폼 개발, 연합학습 개발)<br>• 클라우드 활용비, 지속 운영 계획 수립<br>• 연합학습 실행 대응, 결과 보고<br>• 플랫폼 유지보수 및 고도화 |
| | (세부2) 연합학습 원천 기술개발 | • 연합학습 원천 기술R&D(보안 강화 기술, 취합알고리즘, 학습 프로토콜, 기여도 평가 지표, 공정성) 5개 주제 |
| | (세부3) 플랫폼 활용과제 | 데이터 공급 및 연합학습 참여, 실험 검증<br>• 데이터 공급 및 생산<br>• 연합학습 참여 및 예측모델 검증<br>• 실험적 검증<br>• 데이터 큐레이션 및 디지털 전환 |
| | | 표준데이터 처리도구 및 모델 개발<br>• 데이터 전처리 도구 개발<br>• 예측 모델 개발<br>• 모델 성능 평가 및 지속 개선 |

## 5.2.2. 추진 계획

❑ 사업 목적

❍ 분산된 민간-공공 데이터의 활용 및 데이터 기반 협력이 가능한 한국형 연합학습 기반 AI 신약개발 플랫폼(K-MELLODDY, Machine Learning Orchestration for Drug DiscoverY)을 구축하고 응용 사례를 제시하여 국내 제약산업의 인공지능 기반 신약개발 생태계 활성화 도모

- 연구 기간 내 다양한 제약사 참여를 통해 연합학습 기반 인공지능 모델을 고도화하고 연구종료 후 지속적 활용이 가능하도록 개방형 플랫폼 구축

❑ 사업 내용

❍ (내역 1) 연합학습 기반 신약개발 가속화 프로젝트 사업단

- (세부 1) 플랫폼 구축·운영 및 사업 관리
  - 요구사항 정의 및 설계
  - 보안기술 기반의 연합학습 프레임워크 설계 및 개발
  - 개발되는 다양한 알고리즘에 대응할 수 있는 플랫폼 운영 검증
  - 개인정보 보안 리스크 분석 및 대응(risk analysis and mitigation)
  - 제약사의 민감한 정보들이 관여되기 때문에 감사(audit) 수행
  - MELLODDY 사업에서는 독일 보안 서비스 기업 cirosec이 감사를 수행하였고, 감사 후 변경된 설정에 대해 배포 전에 최소한 3개 제약사(참여기관)가 참여하여 검토 프로세스 거침
- 연합학습 플랫폼 및 거버넌스(추진위원회) 운영·관리
  - 전체 전략 방향, 진행 상황, 소통 원칙 결정, 관리 절차 및 플랫폼 품질 보증, 참여기관의 계약, 계획 변경 등 과제 전반에 대한 관리
  - 참여기관 간 의견조정을 위한 협의체, 프로젝트 실무위원회 등 운영
  - 공통의 소통 전략 사전 수립
  - 데이터 처리 원칙 마련, SOP 제공, 모델 성능 평가 및 지속 개선
  - 사업단 총괄 운영, 사업 기획·평가·관리, 성과 확산 및 홍보 등

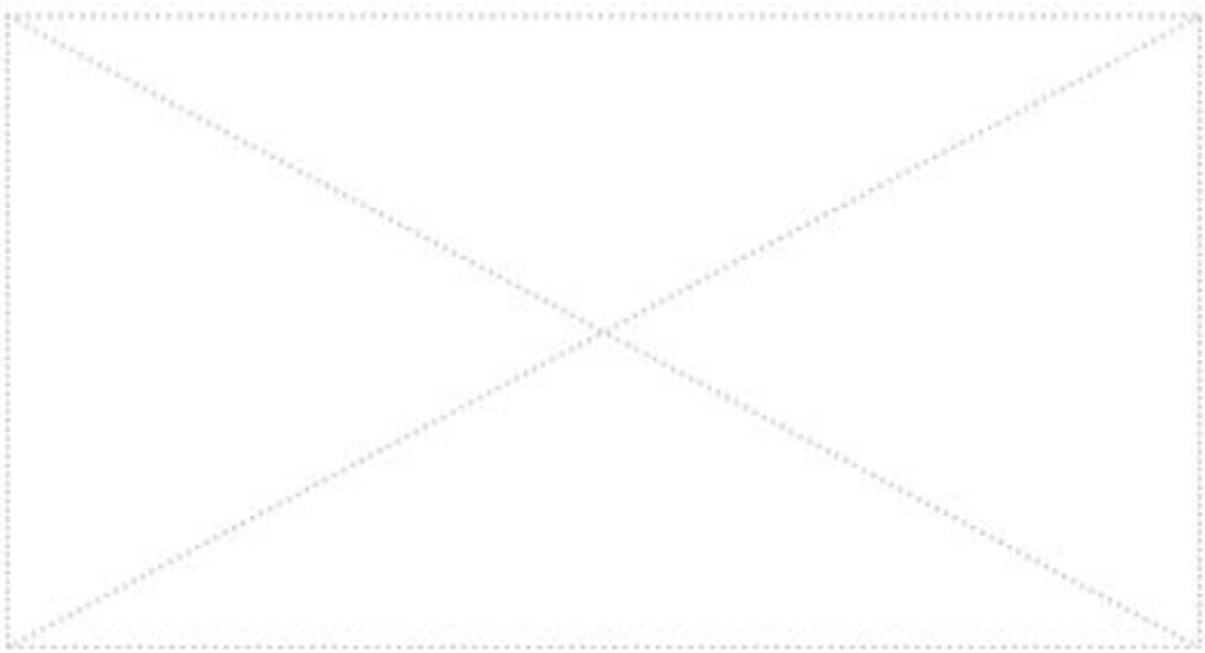


(출처: denhof et al. 2022

그림 55. 연합학습 플랫폼 사례 : EU MELLODDY 프로젝트

❍ (세부 2) 연합학습 원천기술 개발

– 연합학습의 핵심 연구 주제를 키워드로 5개의 기관 간 연합학습 원천기술 연구 주제 선정



(출처: Bingyan Liu, Recent Advances on Federated Learning: A Systematic Survey)
그림 56. 연합학습 연구 주제

표 28. 연합학습 핵심 연구 분야와 세부 연구 주제(안)

| 연구분야 | 연구 주제 |
|---|---|
| 취합 알고리즘 최적화 | • 가중치 수준의 최적화<br>• 특징 수준의 최적화<br>• 네트워크 최적화<br>• 기타 다양한 최적화 방법 |
| 이질성 극복 방안(학습 프로토콜) | • 데이터 이질성 극복(분포, 예측 값)<br>• 모델 이질성 극복 |
| 보안강화 기술 | • 위험 분석<br>• 암호화 알고리즘<br>• 가중치 보호 방안 |
| 공정한 연합학습 | • 데이터 공정성<br>• 모델 공정성<br>• 클라이언트 공정성 |
| 기여도 평가 방법 | • 데이터 가치 평가 방법<br>• 클라이언트 기여도 평가 지표<br>• 이익 공유 방법 |

❍ (세부 3) 플랫폼 활용사업

– (개발 대상) 신약개발 AI 개발 분야* 선정(과제당 5년 지원)

- 연합학습 예측 모델 개발 가능 분야(안): ①물성 예측, ②투과도 예측, ③분포용적 예측, ④수송체 약물 분포 영향력 예측, ⑤버퍼 안정성, ⑥대사 안정성, ⑦심장 독성, ⑧간 독성, ⑨발암성, ⑩생식, 내분비 독성 등

－(연합학습) 분야별 연합학습 기반의 AI 모델 개발

- 사업 초기에 각 참여기관에 배포할 기본 모델(Base Model) 개발
- 검증과 하이퍼파라미터 최적화를 위한 컴퓨팅 계획 수립



(출처: Martijn Oldenhof et al. 2022)
그림 57. AI 모델의 연합학습 개념

- 참여기관(로컬서버)에서는 준비된 데이터를 이용해 배포된 AI 모델을 학습하고, 업데이트된 모델의 파라미터(그림 ΔWi)를 중앙서버에 전달
- 중앙서버에서의 AI 모델 결합(aggregation) 및 성능 진단
- 중앙서버에서는 공유된 파라미터를 결합하여 업데이트된 결과 모델(ΔW) 을 다시 참여기관들에 배포

－(검증) 분산 데이터 환경에서 AI 모델을 검증, 실험을 통한 비교분석 수행

－(디지털 전환)데이터 품질관리 및 최적화 연구

- 분할구조 등 연합학습을 위한 데이터셋 준비 전략 수립
- 데이터 사양(specification), 데이터 준비 시스템 등 참여기관에서 AI 모델 학습시 필요한 데이터를 통일하기 위해 Data Preparation Manual 개발
- 분류 데이터, 회귀데이터, 하이브리드 데이터 등 다양한 데이터를 이용할 수 있도록 MELLODDY Tuner*와 같은 툴과 데이터 준비 매뉴얼 개발을 통해 각 기관에서 준비한 데이터의 일관성과 호환성 유지
- Open-Source Cheminformatics Software RDkit에 구축된 MELLODDY Tuner[26]라는 패키지로 MELLODDY 사업에서 개발되어 데이터 준비에 활용

26) https://github.com/melloddy/MELLODDY-TUNER

- (데이터 생산 및 준비) 참여기관, 공공 DB 등 AI 신약개발 데이터 준비
  - 기관별 보유한 데이터가 부족한 경우, 실험을 통해 추가 데이터 생성
  - 제약기업들의 AI 연합학습 참여

### 5.2.3. 연구 주제 선정

❑ 선정 주제의 중요성

❍ ADME/Tox는 Absorption(흡수), Distribution(분배), Metabolism(대사), Excretion(배출), Toxicity(독성)의 앞 글자를 딴 단어로 약물이 질병 표적에 좋은 활성을 보여도 ADME/Tox에서 문제가 발생하면 임상 단계에서 실패하므로 신약개발 초기 단계에 약물 구조로 ADME/Tox 성질을 잘 예측하는 것이 중요함

- ADME/Tox 분야에도 상세하게 나누면 매우 다양한 분야가 존재 ①물성 예측, ②투과도 예측, ③분포용적 예측, ④수송체 약물 분포 영향력 예측, ⑤버퍼 안정성, ⑥대사 안정성, ⑦심장 독성, ⑧간 독성, ⑨발암성, ⑩생식, 내분비 독성 등

표 29. ADME/Tox 및 약효 연구 주제

| 대분류 | 중분류 | 소분류 |
|---|---|---|
| 약물 흡수 | 물성 | • 용해도 예측 (Water, 인공장액, 인공위액, Intrinsic)<br>• 이온화 상수(pKa) 예측<br>• 친유성, 분배계수 예측 (logP, logD) |
| | 투과도 | • 소장 상피세포 투과도 예측 (Caco-2, MDCK)<br>• 인공지질막 투과도 예측(PAMPA)<br>• 뇌 미세혈관 내피 세포 투과도 예측 (BMeC)<br>• 뇌장벽투과도 예측<br>• 각막 투과도 예측<br>• 피부 투과도 예측 |
| 약물 분포 | 분포용적 | • 혈장 및 혈장단백질 결합 예측(알부민 등)<br>• 혈액 및 혈액단백질 결합 예측<br>• Fraction unbound 비율 예측<br>• 분포용적 값 예측 |
| | 수성체 약물 분포 영향력 | • OAT family(OATP1B1, OATP1B3, OAT1) 결합친화도 및 기질성 예측<br>• OCT 전사인자 or OCT TF(OCT1. OCT2) 결합친화도 및 기질성 예측<br>• BCRP 결합친화도 및 기질성 예측<br>• BSEP 결합저해도 및 기질성 예측<br>• BBB Transporter(P-gp 등) 결합친화도 및 기질성 예측 |
| 약물 대사 및 배설 | 버퍼 안정성 | • 인공 위액, 장액<br>• DMSO, Water<br>• 혈장 |

| | | |
|---|---|---|
| | 대사 안정성 | • 마이크로좀 안정성(청소율, 반감기 예측) (mouse, human, rat)<br>• 간세포 안정성 (Mouse, Human, Rat) |
| 약물 독성 | 심장 독성 | • hERG 채널 저해 예측 |
| | 간 독성 | • CYP450 저해 예측<br>• 간 독성 5종 효소 활성 예측 (AlkPhos, SGOT, SGPT, LDH, GGT) |
| | 발암성 | • Ames Test 결과 예측<br>• TD50 예측 (mouse, rat) |
| | 생식, 내분비 독성 | • 에스트로겐 수용체 결합 예측<br>• 안드로겐 수용체 결합 예측<br>• 생식독성 예측 |
| | 세포 독성 | • 인지질증 예측 |
| 약효 | 인산화효소 (Kinase) 선택성 | • 인산화효소 (Kinase) 선택성 예측 |
| | GPCR 선택성 | • GPCR 선택성 예측 |

❍ 선정 이유

– ADME/Tox는 표적 질환의 차이와는 상관없이 신약개발 초기에 공통으로 확인이 필요한 작업으로 국내 기업에도 실험데이터가 누적되어있음

❑ 후보 학습 방법론

❍ 다중작업 학습(Multi-task learning, MTL) 입력 데이터에서 공통의 특징을 추출하고, 이후에는 단일 작업(Single-Task)을 해결할 수 있도록 분기를 나누어 추가 학습하는 딥러닝 네트워크 구조로 단일 작업이 연관되어 있으면 멀티 작업의 성능 향상이 가능한 학습 방법론임

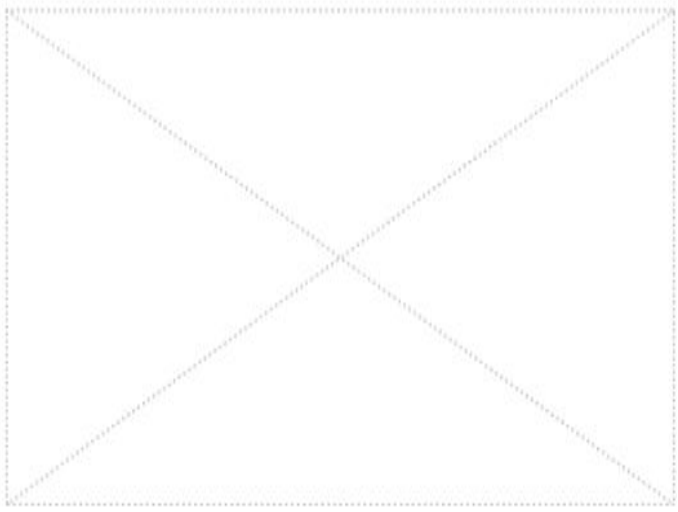


그림 58. 일반적인 Multi-task Learning 모델 구조

❍ (활용 이유) ADME는 서로 같은 분자에 대한 다양한 Assay가 포함되고, single-Task 별 연관성이 높은 것도 존재

- ADME의 single-Task 예측 모델을 각각 만드는 것보다 MTL 방법으로 하나의 약물에 대해 여러 가지 ADME 특성을 예측하는 모델을 개발한다면 모델 개발비의 비용이 절감되고, 성능 향상도 가능할 것으로 예상

표 30. MTL 방법론 활용의 장단점

| | 특징 | 설명 |
|---|---|---|
| 장점 | Knowledge Transfer (EavesDropping) | 하나의 Task를 학습하면서 얻은 유용한 정보(표현)이 다른 Task에도 좋은 영향을 줌 |
| | Over fitting Reduced (Generalized Model) | 여러 Task를 동시에 맞추기 위해 학습을 수행하기에 모델이 특정 데이터에 오버피팅되는 현상을 방지해 모델 일반화 수행 능력이 향상되어 새로운 데이터에 대한 예측 능력이 높아짐 |
| | Computational Efficiency | 예측이 필요한 Task 수만큼 모델을 만들어야하는 번거로움 없이 다양한 Task를 하나의 모델로 학습하고 예측할 수 있다는 장점이 있음 |
| | Real-world Application | 현실에서 매우 다양한 Task가 한번에 요구되는 경우가 많음 ex) 자율 주행차, ChatGPT |
| 단점 | Negative Transfer | 다른 Task에 악영향을 끼치는 Task가 포함된다면 성능저해요인이 될 수 있음 |
| | Task Balancing High | Task 별 학습 난이도가 상이할 경우 학습에 요구되는 하이퍼 파라미터의 조정이 필요한데 Task별 원하는 파라미터를 모두 반영하기 어려워 모델이 수렴하지 않거나 강건하지 않을 수 있음 |

❍ (MTL의 ADME 분야 활용)

- Bayer의 연구자들이 사용할 수 있도록 내부 데이터를 분석해주는 in sillico ADME 플랫폼 개발을 위해 사용 중인 모델을 분석했음
- 2019년 기준 ADME/Tox에 가장 많이 사용되는 모델의 형태는 Random Forest와 이어서 다중작업 네트워크 (Multi-task Neural Network, MTNN) 이었음

그림 59. Bayers in sillico ADME 모델 포트폴리오

❑ GUI 도구의 필요성

❍ (Swiss ADME) 약물 발견 단계에 활용할 수 있는 물리화학적 특성, 약동학, 약물 유사성 및 의약화학 친화성에 대해 분석할 수 있고 접근성이 뛰어난 웹 서비스를 2017년에 개발

– 기존의 연구자들이 고안한 수식이나 ML 모델을 활용해 화합물의 SMILES를 입력받아 실험값의 예측 결과를 출력해줌

❍ (AI 도구의 활용 문제) AI로 구축된 모델은 단순 프로그래밍 언어와 AI용 프로그램 언어를 모두 활용하여 개발되며, AI 전문가가 아닌 사용자가 학습 완료된 AI 모델을 활용하기 위해서는 여러 가지 추가 절차가 필요함

– 제안하는 K-MELLODDY 사업의 목표도 민간 및 공공의 다양한 대량의 데이터로 학습된 AI 기반 ADME/Tox 예측 모델을 개발하는 것임

– 이를 활용하기 위해서는 Swiss ADME와 유사하게 접근성과 사용성이 뛰어난 형태의 서비스 GUI 도구로의 개발이 필요함

그림 60. Swiss ADME 화면

### 5.2.4. 연합학습 플랫폼의 장점

❑ 연합학습 플랫폼의 장점

- ❍ 연합학습은 안전한 인공지능 학습 기술로 데이터 프라이버시 향상, 제도적 한계 극복, 데이터 활용 지속성, 영향 평가가 가능함
- ❍ (데이터 프라이버시 향상) 기존의 데이터를 한곳에 모아 학습하는 데이터 공유 기반 통합학습과 달리, 연합학습은 각 기관(로컬 클라이언트)에 데이터를 그대로 둔 채로 모델과 학습 파라미터 공유를 통한 모델 공유 기반 연합학습방식을 사용하기에 원시 데이터의 직접 이동이 없어 통신 위협, 서버 해킹 위협, 데이터 유출 위협으로부터 안전
- ❍ (제도적 한계 극복) 신약개발 과정에는 다양한 데이터가 필요하고, 이에 일부 데이터(유전체, 임상 정보)는 정보 보호법에 따라 데이터 활용을 위해서 가명처리, IRB 승인 절차가 필요함. 연합학습은 제도적 문제를 회피할 수 있는 현실적인 대안임
- ❍ (데이터 활용 지속성) 연합학습은 자동화 구축이 가능해 기관에 새로운 데이터 발생으로 학습이 필요할 때, 일련의 과정을 재수행하여 사전 학습된 AI 모델에 새로운 지식 축적이 가능하며, 더 나아가 학습 완료된 모델에 기관 특화 데이터로 재학습하는 전이 학습이 가능해 기관별 맞춤 모델 개발이 가능
- ❍ (데이터 영향 평가) 연합학습은 각 기관의 로컬 학습 결과를 공유받기 때문에 AI 모델에 대해 기관 보유 데이터의 영향력을 평가할 수 있어, AI 모델 성능에 기반한 데이터의 신뢰도, 활용성, 보상체계 평가가 가능함

## 5.2.5. 활용과제 개요

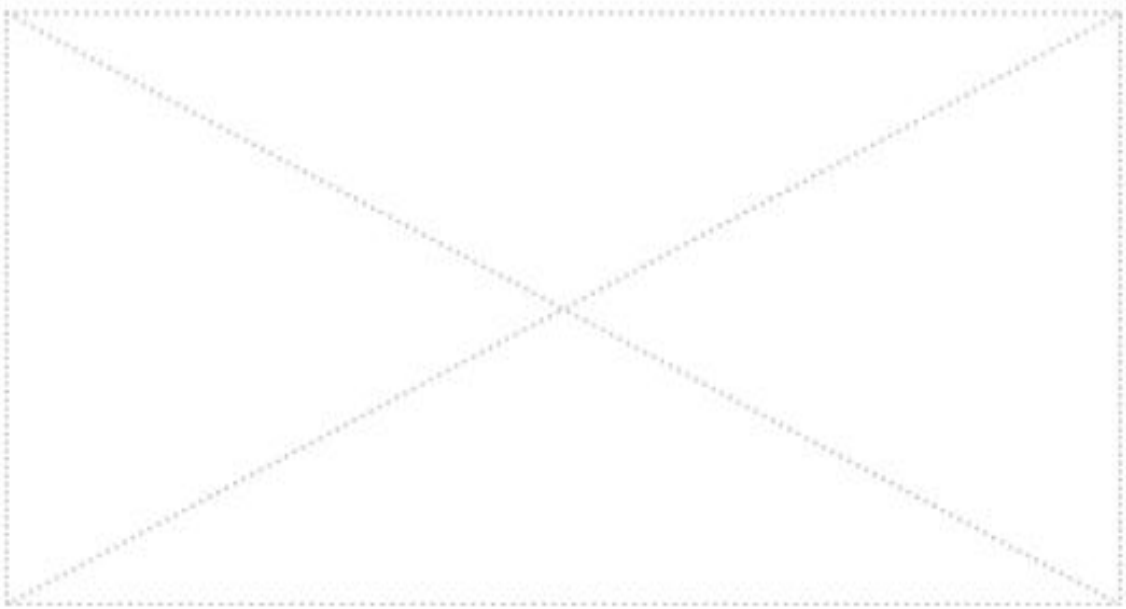


그림 61. 연합학습 모델 활용 사업의 시스템 구조

❑ 활용배경

❍ 신약개발 후보의 실패, 특히 임상시험 중 탈락 비율이 높아지면서 신약발굴과 개발에 필요한 비용이 점점 더 증가. 임상시험 실패의 주요 원인인 약물대사 및 약물동태 문제를 해결하기 위하여 신약개발의 초기단계에서부터 여러 가지 의약 후보 물질의 약효뿐만 아니라 물리화학적 특성과 ADME/Tox(흡수, 분포, 대사, 배설, 독성)를 분석하는 전략이 중요

❍ 기업에서는 초기 약물 탐색 단계에서부터 임상에서 성공 가능성을 높이기 위하여 후보물질의 Drug-likeness(ADME/Tox) 분석을 수행. 미국 NIH 보고서에 따르면 신약개발 비용의 약 20%가 ADME/Tox분석에 소요

❑ 활용전략

❍ 본 사업에서는 약물 최적화 과정(Lead Optimization)에서 필수적이나, 특정질환에 한정되지 않아 여러 제약기업에 보편적으로 분포되어 있는 ADME/Tox데이터를 사용하여 연합학습 활용연구를 수행하고자 함

❑ 실행가능성

❍ ADME/Tox 데이터는 현재 국내 제약기업이 가장 많이 갖고 있는 공유가능 데이터이며, 제약기업들의 수요가 높은 만큼 즉시 실행가능

❑ 기대효과

❍ (비용효과성) 국내 제약기업의 R&D투자비 연간 4,600억원 절감, 소요기간 50%이상 단축 가능(평균 2년→11개월)

❍ (시행착오, 실패비용 최소화) ADME/Tox 분석은 제약기업이 공통 수행하는 단계로 한 기업이 하나의 화학물질을 대상으로 수십 번의 실험을 수행
- 같은 실험을 여러 기업에서 반복 수행하는 불필요 비용 지속 발생. 기업 간 데이터 협업이 가능해지면 시행착오 및 실패비용을 최소화할 것임

❍ (데이터 활용 생태계 활성화) 현재 AI 신약개발 기업들은 민감 데이터에 접근하기 어려워 개별 기업 간의 1:1 계약을 통하여 활용 데이터를 확장하고 있지만 이러한 협업으로는 효과적인 성과를 내기 어려우며, AI 스타트업은 협업의 기회가 더욱 부족

❍ 본 사업을 통하여 AI 신약개발 기업들이 데이터를 효율적으로 활용하는 기반을 마련하고자 하며, 이로써 데이터 활용 생태계 활성화 가능

## 5.2.6. 사업 규모

❑ (사업 기간 및 예산) `24~`28(총 5년 이내), 466억원 지원

❑ (구성 및 연간 과제 수) 1개 사업, 26개 과제 지원

❍ (내역①) (`24) 총 5,400백만원 지원

- (세부① 플랫폼 구축·운영 및 사업 관리) 1개 × 2,700백만원 × 6/12개월 = 1,350백만원
- (세부② 연합학습 원천기술 개발 과제) 5개 × 300백만원 × 6/12개월 = 750백만원
- (세부③ 플랫폼 활용 과제)
  - (데이터 공급 및 연합학습 참여, 실험검증) 20개 × 300백만원 × 6/12개월 = 3,000백만원
  - (전처리 도구 개발 및 모델 개발) 3개 × 200백만원 × 6/12개월 = 300백만원
  - 내역①는 내역②의 사업단 선정 후 3개월 후 추진되기 때문에 '24년에는 6개월로 설정

표 31. 연차별 예산(안)

(단위 : 억원)

| 구분 | 과제수 | '24 | '25 | '26 | '27 | '28 | 합계 |
|---|---|---|---|---|---|---|---|
| □ 연합학습 기반 신약개발 가속화 프로젝트 | 29 | | | | | | |
| ○ (내역1) 연합학습 기반 신약개발 가속화 프로젝트 사업단 | 29 | 54 | 103 | 103 | 103 | 103 | 466 |
| **– (세부1) 플랫폼 구축·운영 및 사업 관리** | 1 | 13.5 | 22 | 22 | 22 | 22 | 101.5 |
| **– (세부2) 연합학습 원천기술 개발 과제** | 5 | 7.5 | 15 | 15 | 15 | 15 | 67.5 |
| **– (세부3) 플랫폼 활용 과제** | 23 | 33 | 66 | 66 | 66 | 66 | 297 |
| · 데이터 공급 및 연합학습 참여, 실험 검증 | 20 | 30 | 60 | 60 | 60 | 60 | 270 |
| · 전처리도구개발 및 모델 개발 | 3 | 3 | 6 | 6 | 6 | 6 | 27 |

❑ 소요 예산 : 총 466억 원 (5개년)

| 사업명 | 세부명 | 세부 연구 내용 | 단가 | 총계(억원) |
|---|---|---|---|---|
| 연합학습 기반 신약개발 가속화 프로젝트 사업단 | 플랫폼 구축 ·운영 및 사업 관리 | **– (대표단체) 플랫폼 운영 및 사업관리** | **47.5** | 101.5 |
| | | · 과제지원 및 성과 관리 체계 구축 운영(거버넌스(추진위원회)구성 및 운영) | 9 | |
| | | · 플랫폼 ISP 수립 | 3 | |
| | | · 데이터 처리 원칙을 마련하고 SOP 제공 | 2 | |
| | | · 플랫폼 교육 기획 및 운영 | 12 | |
| | | · 데이터 품질 및 생산 절차 관리 | 2 | |
| | | · 플랫폼 활용 가이드라인 개발 | 2 | |
| | | · 플랫폼 고도화 방안연구 | 2 | |
| | | · 플랫폼 확산 지원(경진대회, 홍보) | 7 | |
| | | · 신약개발 데이터 공동활용성 확대 연구 | 6 | |
| | | · 플랫폼 활용확산방안 연구 및 성과분석 | 2.5 | |
| | | **– (IT) K-MELLODDY 플랫폼 구축** | **54** | |
| | | · 플랫폼 설계(플랫폼 구조 및 기능, 인프라 구조, 연합학습) | 6 | |
| | | · 플랫폼 구축(인프라 구축, 플랫폼 개발, 연합학습 개발) | 9 | |
| | | · 클라우드 플랫폼 활용비 | 30 | |
| | | · 인프라 지속 운영 계획 수립 | 1 | |
| | | · 연합학습 실행 대응, 결과 보고 | 4 | |
| | | · 플랫폼 유지보수 및 고도화 | 4 | |
| | 연합학습 원천기술개 | **– (대학, 연구소) 연합학습 원천기술 R&D** | **67.5** | 67.5 |
| | | · 연합학습 원천 기술R&D(보안 강화 기술, 취합알고리즘, 학습 프로토콜, 기여도 평가 지 | 67.5 | |

| | | | | |
|---|---|---|---|---|
| | **발** | 표, 공정성) 5개 주제 X 300 | | |
| | **플랫폼 활용과제** | · 20개 기관(후보물질 대사, 독성) x 13.5<br>· 3개 기관 (AI 모델 개발) x 9 | 297 | 297 |
| | | **-(제약기업, 공공기관) 데이터 공급 및 연합학습 참여, 모델 결과 실험 검증** | **13.5** | |
| | | 데이터 공급 및 생산<br>연합학습 참여 및 예측모델 검증<br>실험적 검증<br>데이터 큐레이션 및 디지털 전환 | 7<br>2<br>3<br>1.5 | |
| | | **-(AI) 표준 데이터 처리 도구 및 모델 개발** | **160** | |
| | | - 데이터 전처리 도구 개발<br>- 예측 모델 개발<br>- 모델 성능 평가 및 지속 개선 | 60<br>60<br>40 | |

### 5.2.7. 로드맵

❑ 연합학습과 플랫폼에 대한 원천기술 개발이 필요하고, 실용성 확보 및 데이터 협력 효과 확인을 위한 학습 및 검증에 충분한 시간이 필요해 K-MELLODDY 사업은 EU MELLODDY 기간보다 장기적인 36개월 => 60개월로 사업 기간 장기적으로 설정하였음

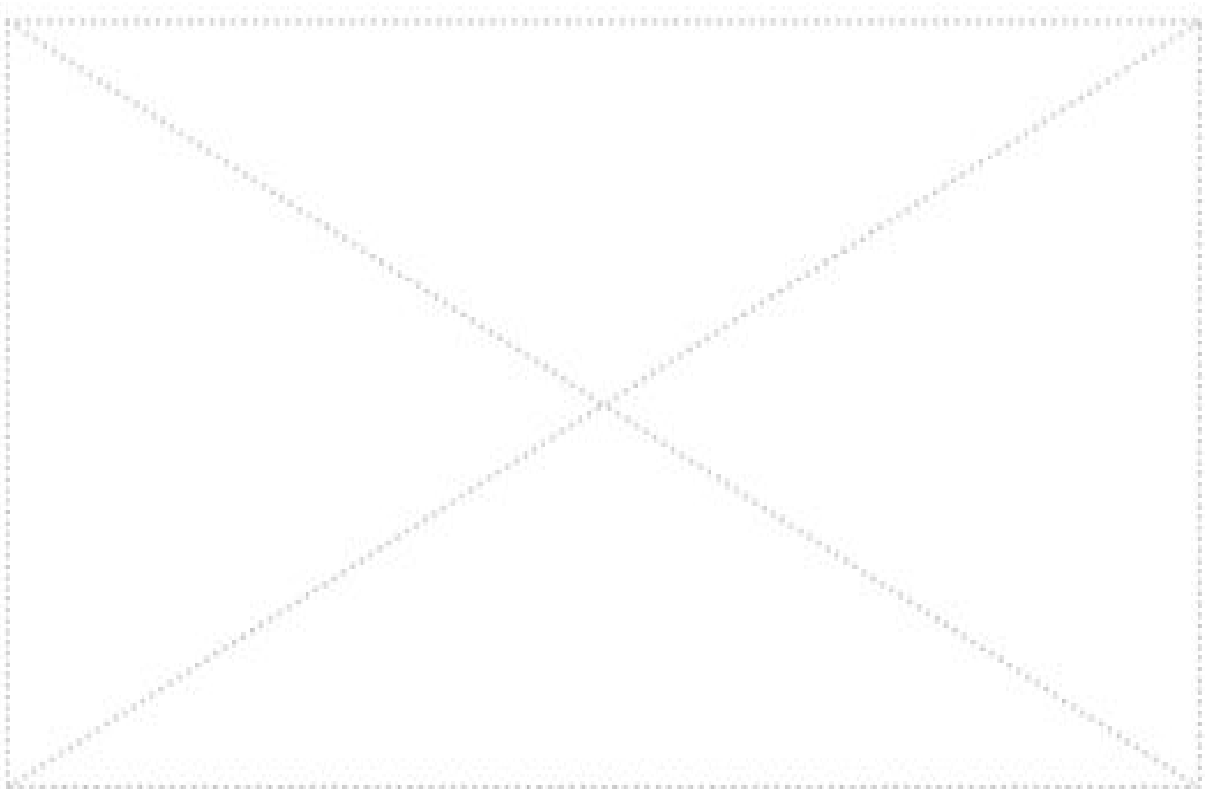


그림 62. K-MELLODDY 사업의 연차별 사업 목표에 따른 5개년 로드맵

## 5.2.8. 세부 과제별 연구개발 내용 및 범위

(총괄 과제, 세부1)

| RFP 번호 | 1-1 |
|---|---|

| 연구 과제명 | 한국형 연합학습 기반 AI 신약개발 플랫폼 구축·운영 및 사업 관리 |
|---|---|

**필요성**

- 데이터 관련 제한 강화로 신약개발 분야의 민간·공공 데이터의 고립이 심화되고 있음
- 한국 제약산업은 추격 요원으로 글로벌과 경쟁하기 위해서는 AI 도입을 통한 퀀텀점프 전략이 필요
- 신약개발 분야 AI 도입을 위해서는 신약개발 민간 데이터의 공유·협력 체계 구축이 필요

**연구 목표**

- 신약개발 분야의 민간·공공 데이터의 고립을 해제하여 데이터 공유를 통한 AI 기술 도입을 촉진
- 고립 해제의 방법으로 연합학습 기술을 활용하고, 기술 기반의 신약개발 데이터 협력 플랫폼 구축
- 개발한 플랫폼의 검증 및 활용성 확보를 위하여 시범 사업 추진
- 연합학습 프레임워크가 탑재되고 단순 프레임워크보다 사용성이 개선된 연합학습 기반 플랫폼 구축

**연구개발 내용**

| | |
|---|---|
| 2024년 | – 참여기관과 데이터 파트너십 협약 및 실제 데이터 파악<br>– 추진위원회 구성 및 정기 개최를 통한 이해관계자 의견 수렴<br>– 추진위원회를 통한 플랫폼 요구사항 수렴<br>– K-MELLODDY 플랫폼 요구사항 정의 및 설계 |
| 2025년 | – 데이터 처리 SOP 수립 및 가이드라인 작성<br>– 추진위원회 정기 개최를 통한 이해관계자 의견 수렴 지속 수렴<br>– K-MELLODDY 플랫폼 개발 및 인프라 구축<br>– 개발 플랫폼의 동작 검증 및 오류 수정(공공 데이터 활용) |
| 2026년 | – 플랫폼 시범 운영 및 수요기관의 플랫폼 검증 및 사용 보고서 제공<br>– 플랫폼의 시범운영 홍보 및 소개, 확산<br>– 실제 데이터 연합학습 수행에서 발생하는 문제점을 취합 플랫폼 안정성 개선 |
| 2027년 | – 플랫폼을 운영하여 시범사업의 실제 데이터와 모델로 연합학습 1차, 2차 수행<br>– 1차 2차의 학습 결과를 검토 및 분석하여 피드백<br>– 플랫폼 운영 보고서 작성 및 개선안 제시<br>– 모델 활용 서비스 개발, 서비스 평가<br>– 연구 개발 또는 최신 기술을 활용하여 플랫폼 고도화 |
| 2028년 | – 3차 연합학습 수행 및 결과 분석, 플랫폼 이용 및 활용 실적 보고<br>– 성과 홍보 및 확산, 최종 보고서 작성<br>– 플랫폼을 활용하기 위한 API 문서, 가이드라인 개발<br>– 플랫폼 응용 분야 도출 |

**예상 성과**

– AI 신약개발 응용 분야 예시 (건)
– 표준 데이터 처리 방법 (건)

- 플랫폼 홍보, 활성화 (회)
- K-MELLODDY 플랫폼 구축 (건)
- 플랫폼 동작 검증 보고서 (건)
- 플랫폼 인증 (건)
- 플랫폼 예상 활용 분야 보고서(건)

**활용 계획**

- 안전한 데이터 협력 기술인 연합학습에 기반한 데이터 협력 장치로 활용
- 제약기업과 AI 신약개발기업, 바이오인포메틱스나 분석 기업 등이 매칭할 수 있는 플랫폼으로 응용
- 연합학습 기반의 협력 플랫폼은 데이터가 고립된 다양한 산업에 적용이 가능
- 국내 최초의 연합학습 기반 데이터 협력 플랫폼으로 활용

**연구기간 및 소요예산**

2024년 ~ 2028년(5개년) /총 101.5억

(2세부 과제)

| RFP 번호 | 1-2 |
|---|---|

| 연구 과제명 | 연합학습 실용화를 위한 원천기술 연구 |
|---|---|

**필요성**

- 세계적으로 주목하고 있는 데이터 협력 기술인 연합학습의 국내 기술 확보 미흡
- 분산되고 고립된 데이터를 해제할 수 있고, 다양한 소스의 데이터를 AI 모델에 입력에 활용 가능
- 신약개발 분야의 AI 도입을 위한 제도적, 사회적, 인지적, 경제적 문제를 해결할 수 있는 열쇠

**연구 목표**

- 국내 연합학습 기술의 성숙도 향상을 위한 연합학습 원천기술 연구개발
- 연합학습 핵심 과제에 따른 연구주제 설정 및 최신 연구의 한계 극복 및 도메인에 적용하여 실증
- 연합학습 기술이 도입된 한국형 연합학습 기반 신약개발 플랫폼에 연구개발한 기술을 탑재

**연구 개발 내용**

| | | |
|---|---|---|
| | 2024년<br>~<br>2028년 | – 핵심 과제 예시(취합 알고리즘 최적화, 이질성 극복 방안, 보안강화 기술, 공정한 연합학습, 기여도 평가 방법, 정보보안 리스크 등) |
| | | – 연합학습 학습 방법론(Federated learning, Split learning, Swarm Learning) 등 다양하나, 의료 데이터에만 적용한 사례가 있을 뿐 신약개발 사례는 없음<br>– 연합학습 기술연구의 성능 평가의 일반적 기준: 동일한 AI 모델을 로컬 데이터로만 학습한 결과와 연합학습한 결과의 비교<br><br>그림 63. Swarm Learning<br><br>그림 64. Split Learning<br>– 서버가 존재하지 않은 Swarm Learning 기반의 Spread GNN 연구 결과의 성능치 ROC-AUC(독성기준 0.666 이상 달성 목표)<br>그림 65. SpreadGNN 아키텍처 |
| | | – 연합학습의 데이터 이질성, 사용자 이질성, 시스템(스펙) 이질성에서 오는 문제를 극복할 수 있는 프레임워크 개발(대표 사례FederatedScope) |

그림 66. FedeartedScope의 다른수준의 프로그래밍 인터페이스 제공 아키텍처

- 전송 가중치의 보안강화 기술 (JointPEQ(Privacy Enhancement Quantization)의 0.68 acc 이상 달성 또는 정확도 손실 보정, 신약개발 도메인 적용 후 개선 등)

그림 67.JointPEQ

- 공정한 연합학습 알고리즘의 성능 손실 개선, 새로운 방법론 연구, 도메인 적용 등(FairFed의 성능 손실 발생, 이질성 강화 시 성능 저하 문제 개선)

그림 68. FairFed 알고리즘

- 모델 기여도 평가 방법론 연구(입력 데이터 퀄리티와 수에 따른 성능 기여도, 자원 사용에 대한 기여도 등 이외에도 연합학습 참여의 다양한 기여도 평가)

그림 69. 데이터 가치 평가

그림 70. 자원 기여 평가

| |
|---|
| **예상 성과** |
| – 연합학습 취합 알고리즘 최적화 기술 논문(건)<br>– 연합학습 이질성 극복 방안 논문(건)<br>– 연합학습 보안 강화 기술 논문(건)<br>– 연합학습 공정한 연합학습 논문(건)<br>– 연합학습 기여도 평가 방법 논문(건)<br>– 연합학습 정보보안 리스크 검토 논문(건)<br>– 연구 주제별 특허 출원 및 등록(건)<br>– 연구개발 기술의 활용사례 확보 (건) |
| **활용 계획** |
| – 한국형 연합학습 기반 신약개발 플랫폼에 활용<br>– 연합학습 기술 성숙과 실용화를 통한 신사업 창출<br>– 연합학습 기술 주도권 확보와 기술 표준 주도 |
| **연구기간 및 소요예산** |
| 2024년 ~ 2028년(5개년) / 5개 과제/ 총 67.5억(과제당 13.5억) |

(3세부 과제, 활용사업 1)

| RFP 번호 | 1-3 |
|---|---|

| 연구 과제명 | 한국형 연합학습 플랫폼 기반 ADME/Tox 연합학습 활용 |
|---|---|

**필요성**

- ADME 분석은 모든 제약기업이 공통으로 수행하는 연구단계로 실험 오차 보정을 위해 반복 실험, 측정으로 불필요한 비용이 지속적 발생
- AI는 다양한 소스의 측정값 오차를 보정하는 효과가 있어 비용 절감 및 시행착오를 최소화할 수 있음
- 민간 기업의 ADME 데이터는 기업의 주요 표적과 관련된 약물에 편향됨. 편향의 완화와 최신의 ADME 데이터를 사용할 수 있도록 민간 데이터 융합이 필요함

**연구 목표**

- 제약기업 또는 공공기관이 보유한 ADME/Tox 데이터를 안전하게 연합학습하여 연합학습 기반 AI 예측 모델을 개발
- 제약기업의 데이터 활용 및 디지털 전환, 부족한 데이터 생산
- 연합학습한 ADME/Tox 예측 모델의 성능 평가를 실제 실험을 통해 진행함으로써 실용성 확보

**연구 개발 내용**

| | | |
|---|---|---|
| | 2024년 | – 제약기업, 공공기관, 대학, 연구소의 데이터 파악(기업별 데이터 부족분 파악)<br>– 물질대사 및 독성 연구 주제 : ①물성 예측, ②투과도 예측, ③분포용적 예측, ④수송체 약물 분포 영향력 예측, ⑤버퍼 안정성, ⑥대사 안정성, ⑦심장 독성, ⑧간 독성, ⑨발암성, ⑩생식, 내분비 독성 등<br>– 정의된 플랫폼 입력 데이터 표준에 따른 데이터 디지털 전환 및 DB화(벤치마크 사업 EU MELLODDY의 데이터 전처리 도구인 Tuner 참고)<br>– EU MELLODDY에서는 화합물 데이터 입력을 SMILES로 받아 32K Fingerprint로 변환하여 사용<br>– 실험값은 편차와 표준편차 오차를 활용하여 스케일링 수행<br><br>그림 71. 벤치마크 사업의 데이터 포맷 |

| | |
|---|---|
| | <br>그림 72. 벤치마크 사업의 데이터 스케일링 |
| 2025년 | – 연합학습을 위한 필요 데이터 생산 및 품질 관리<br>– 데이터 생산 비용 평균 124만원 내외 데이터 생산비 총 7억 지원으로 기관당 5,000개 이상의 데이터 생산 예상<br><br>그림 73. ADME 생산비용 조사(한국제약바이오협회 자체 조사 결과) |
| 2026년 | – 플랫폼 검증 및 사용 보고, 초기 연합학습 수행 및 결과 분석<br>– 개별 기관 데이터 활용 모델과 연합학습한 모델의 성능 평가 비교 분석(평균 성능 벤치마크 사업 이상의 성과 개선 지표 제시)<br><br>그림 74. 벤치마크 사업의 성능 평가(회귀 평균 2%, 분류 평균 4% 향상) |
| 2027년 | – 연합학습 수행 및 성능 평가, 예측 결과의 실험적 검증<br>– 개별 기관 데이터 활용 모델과 연합학습한 모델의 성능 평가 비교 분석 |
| 2028년 | – 3차 연합학습 수행 이후 필요 데이터 추가 생산 및 보정<br>– 연합학습 플랫폼 이용 보고, 모델 활용 방안 도출 |

**예상 성과**

– 데이터 품질 및 생산 보고(기관당 1건 이상)

| |
|---|
| – AI 모델의 화합물 실험 예측값과 실험값 검증 보고(기관당 1건 이상)<br>플랫폼 사용 보고서(건)<br>– 연합학습 결과 보고(건) :로컬 학습 모델의 성능 평가 결과, 개별 기관 모델과의 비교 분석 수행<br>– 모델 활용 방안 보고서(건) |
| **유의사항** |
| – 데이터 확보 전략: 20개의 참여기관으로부터 기관당 최소 5,000개의 화합물(10만 개 데이터 확보 계획) 실험데이터 확보(부족 시 실험 데이터 생성), 공공기관(한국화학연구원의 기탁 화합물 활용), 특허로 공개된 화합물의 실험데이터도 활용(제약기업 의견수용), 실험데이터 확보 및 생산이 가능하다면 대학 및 연구소도 참여 가능 |
| **활용 계획** |
| – 국가연구개발비로 생산된 데이터의 기탁<br>– 기업 민간 데이터의 권리에 따른 연합학습 모델의 유료 서비스<br>– 연합학습 플랫폼의 성공 사례로 활용 |
| **연구기간 및 소요예산** |
| 2024년~2028년(5개년) / 20개 과제/ 총 270억(과제당 13.5억) |

(3세부 과제, 활용사업 2)

| RFP 번호 | 1-4 |
|---|---|

| 연구 과제명 | 한국형 연합학습 플랫폼 기반 ADME/Tox 모델 개발 |
|---|---|

**필요성**

- ADME(물질 대사) 및 독성(Toxicity) 분석은 신약개발 임상시험의 성공 실패를 결정하는 매우 중요한 요소로 제약기업이 공통으로 수행함
- AI 모델이 민관 협력으로 보유한 데이터를 모두 학습하여 예측 모델의 효율성을 향상시켜 제약산업의 공통된 시행착오 비용을 감소시킬 필요가 있음
- AI 모델 개발은 데이터 소유자인 제약기업이 직접 수행하기에는 경제적 인력적 문제가 있으며, 주제에 따른 모델 개발을 AI 신약개발 전문기업이 수행할 필요가 있음

**연구 목표**

- 제약기업이 보유한 물질대사 및 독성 데이터를 안전하게 연합학습하는 AI 기반 예측 모델 개발
- 글로벌, 공공기관의 공개 데이터를 활용하여 주제별 연합학습 초기모델 개발
- 예측 모델의 연합학습 성능 평가를 통해 예측 모델 개선
- 사업 참여에 따른 공개 수준 결정으로 모델의 공개 비공개가 가능한 분할학습 기반의 모델 개발

**연구 개발 내용**

| | | |
|---|---|---|
| | 2024년 | – 제약기업의 데이터 현황 파악(물질대사, 독성의 데이터 타입, 저장 방식, 보유 데이터 수 등), 입력 데이터 표준 데이터 처리 방법 수립<br>– 물질대사 및 독성 연구 주제 : ①물성 예측, ②투과도 예측, ③분포용적 예측, ④수송체 약물 분포 영향력 예측, ⑤버퍼 안정성, ⑥대사 안정성, ⑦심장 독성, ⑧간 독성, ⑨발암성, ⑩생식, 내분비 독성 등<br>– AI 신약개발 기업은 입력 데이터 표준에 따른 데이터 처리 도구를 개발(참고, 벤치마크 사업의 데이터 처리 도구 구조) |

그림 75. 벤치마크 사업의 데이터 처리 도구 구조

<table>
<tr><td>2025년</td><td>– 물질대사, 독성 AI 모델 개발<br>– 공공 데이터 활용 초기모델 개발 (기존 SOTA 모델의 성능 개선, 추후 모델 보안 및 성능 강화를 위한 유연한 모델 구조를 고려해야 함)<br>– 기존의 물질대사 및 독성 예측 모델 연구에서 기계학습을 사용한 경우, 다양한 표현자가 고려되는 반면, 구조적 정보가 부족하고 딥러닝 모델의 경우 구조정보 위주의 실험값을 예측한다는 한계가 있음(하이브리드, 추가 입력정보에 기반한 모델 개선 필요)<br>– 참고) [27]그래프 기반의 GLAM(Graph Learning method for automated molecular interactions) 모델의 성능
<table>
<tr><th>데이터세트</th><th>평가지표</th><th>성능</th></tr>
<tr><td>BBBP</td><td>AUC</td><td>0.932</td></tr>
<tr><td>BingdingDB</td><td>AUC</td><td>0.954</td></tr>
<tr><td>ESOL</td><td>RMSE</td><td>0.592</td></tr>
<tr><td>FreeSolv</td><td>RMSE</td><td>1.319</td></tr>
<tr><td>Lipophilicity</td><td>RMSE</td><td>0.596</td></tr>
<tr><td>SIDER</td><td>AUC</td><td>0.659</td></tr>
<tr><td>Tox21</td><td>AUC</td><td>0.841</td></tr>
<tr><td>ToxCast</td><td>AUC</td><td>0.744</td></tr>
</table>
</td></tr>
<tr><td>2026년</td><td>– 플랫폼 활용 및 연합학습 후 모델 성능 검증 및 평가<br>– 개발 모델의 플랫폼 탑재, 검증 및 모델 고도화</td></tr>
<tr><td>2027년</td><td>– 2차 3차 연합학습 수행 및 모델 성능 및 애로 사항 등 분석<br>– 연합학습을 수행하며 성능 모니터링 및 효과 분석, 모델 개선</td></tr>
<tr><td>2028년</td><td>– 연합학습 3차까지 시도 이후 필요 데이터 추가 생산 및 보정<br>– 연합학습 플랫폼 사용 보고, 모델 개선안 도출, 예측 모델의 실험적 검증<br>– 연합학습 결과 보고, 모델 사용법 제작 및 모델 배포</td></tr>
</table>

**예상 성과**

– 민간 데이터와 공공 데이터로 연합학습한 물질 대사, 독성 저해 예측 모델 (건)
– AI 모델 학습에 이용하기 위한 물질 대사, 독성 표준 데이터 처리 도구 (건)
– 연합학습을 활용한 신약개발 데이터 협업 및 활용사례 (건)
– 연합학습 결과 보고, 모델 사용법 제작, 전이학습 및 개인화 가이드라인 (건)

**유의사항**

**활용 계획**

– 제약기업의 비밀 데이터로 개인화, 전이 학습하여 기업에서 활용
– AI 기반 독성 저해 예측 서비스를 공개 운영하여 신약개발 연구자를 지원하는 도구로 활용

**연구기간 및 소요예산**

2024년~2028년(5개년) / 총 148억
2024년~2028년(5개년) / 3개 과제/ 총 27억(과제당 9억)

27) Paper with code, 성능비교, https://paperswithcode.com/paper/an-adaptive-graph-learning-method-for, 2023.05.31

## 5.3. 추진 방법

### 5.3.1. 추진 체계

❑ 참여자와 역할

❍ 산·학·연·정의 신약개발 다양한 이해관계자가 참여하는 구조로, 제약기업은 데이터 생산·공급 및 연합학습된 모델 예측 결과의 실험검증, AI 신약개발 기업은 데이터 전처리 도구 개발 및 AI 모델 개발·검증, 대학은 연합학습의 원천 기술개발, IT 기업은 한국형 연합학습 플랫폼의 요구사항 정의와 구축, 공공기관은 공공 데이터의 공급자 역할, 산업계 대표단체는 플랫폼의 운영을 담당하며 아래 그림에 추진 체계를 나타내었음

❍ K-MELLODDY 추진위원회를 통해 데이터 파트너십 체결, 플랫폼 운영 지원, 시범사업 추진계획을 수립·시행함



그림 76. K-MELLODDY 추진 체계

❑ 참여혜택

❍ 공공기관, 제약·바이오 기업, 대학, AI 신약개발 기업, IT 기업, 산업계 대표단체가 참여하는 다기관 협력 구조로 각각의 참여 혜택이 다름

❍ (제약·바이오 기업) ADME/Tox 실험 비용의 절감, 보유 데이터의 디지털 전환 지원, ADME/Tox 예측 모델 및 GUI 기반 ADME/Tox 예측 도구 확보

- 연합학습 참여 유인책: 개별 기업이 보유하지 않은 데이터까지 학습된 신약개발 분야 모델의 확보(연구 주제 별 연합학습에 참여한 기업에게만)

- 제약기업의 지적재산권인 데이터로 학습된 예측 모델의 유료 예측 서비스화 (데이터의 기여 가치를 서비스 수익으로 전환)

❍ (대학 / AI 신약개발 기업)

- (AI 신약개발 기업) 사업 분야의 확장(신약개발 분야의 AI 모델 개발 이력), 실제 데이터를 학습한 모델의 범용성 검증 (다양한 데이터에서 좋은 성능이 확보되는지), 제약 ·바이오 기업과의 사업 연계
- (대학) 연합학습 기술의 선도 연구 수행, 개발 기술의 실증 수행 경험, 선도기술 개발로 지식재산권(논문) 확보

❍ (IT 기업) 선도 기술개발 이력, 사업 분야의 확장(보안 클라우드 사용사례), 연합학습의 툴킷(Tool-Kit)화로 클라우드 기능 확보

❍ (산업계 대표단체) 산업계 기여 (AI 신약개발 변화 주도, 인력양성, 협력 네트워크 구축), AI 신약개발 분야 전문가 인력 네트워크 확충

❍ (공공기관) AI의 데이터 수요 파악, 공공 데이터의 활용성 및 지속성 확보

표 32. K-MELLODDY 참여자별 역할과 참여 혜택

| 참여자 | 역할 | 참여 혜택 |
|---|---|---|
| 공공기관·제약·바이오 기업 | - 신약개발 데이터 공급·생산<br>- 데이터 큐레이션 및 디지털 전환<br>- 연합학습 참여 및 검증<br>- AI 모델 예측 결과 실험검증<br>- 모델의 개인화, 미세조정(자사 활용) | - ADME/Tox 실험 비용의 절감<br>- 보유 데이터 디지털 전환 지원<br>- ADME/Tox 예측 모델 및 GUI 기반 ADME/Tox 도구 확보<br>- 공공기관은 구축 데이터의 활용성 확보 |
| 대학·AI기업 | (대학)<br>- 연합학습 프레임워크 개발<br>- 연합학습 원천 기술연구<br>- (AI 신약개발기업)<br>- 목적별 학습 모델개발<br>- 데이터 전처리 도구개발 | (대학)<br>- 연합학습 기술의 선도 연구 수행<br>- 개발 기술의 활용연구 수행 경험<br>- 지적 재산권(논문) 확보<br>(AI 신약개발기업)<br>- 사업 분야의 확장(신약개발 분야의 AI 모델개발 이력), 지적재산권(논문) 확보<br>- 개발 기술의 응용 연구 수행 경험 |
| IT기업 | - 인프라 구축<br>- 보안시스템 구축<br>- 연계 서비스 운영 | - 선도기술 개발 이력<br>- 사업 분야의 확장(보안 클라우드 사용사례)<br>- 연합학습 툴킷화로 클라우드 기능 확보 |
| 산업계 대표 단체 | - 연합학습 플랫폼 기술 제공<br>- 플랫폼 교육 및 인력지원<br>- 사업 관리, 참여자 소통 주도<br>- 성과관리, 사업 홍보, 확산<br>- 공공, 공개 데이터 연계, 확보 | - 제약·바이오 산업계 기여 (AI 신약개발 변화 주도, 인력양성, 협력네트워크 구축)<br>- AI 신약개발 전문가 인력 네트워크 확충 |

※ 데이터 공급자에 속하는 공공기관(예) : 국가생명연구자원정보센터(유전체 데이터), 한국화학연구원(화합물 데이터), 보건의료연구자원정보센터(임상데이터) 등

※ 데이터 공급 기업에게 과제 종료 후 3년까지 민간 데이터 및 정부지원금으로 생산한 데이터를 학습한 ADME/Tox 모델 활용의 우선 권한을 부여하고, 이후에는 정부자금으로 생성한 데이터나 모든 SW를 공개하도록 함

## 5.3.2. 추진 방안

❑ 추진목적

❍ 혁신적 AI 기술 중 하나인 연합학습(Federated learning)을 통해 다기관에 분산된 민감 데이터를 효과적으로 활용, AI 신약개발 협력 프로젝트를 지원·운영, 저비용 고효율 AI 신약개발 가속화

❑ 플랫폼 구축

❍ 과기정통부를 중심으로 IT 기업과 파트너십을 체결하여 신약개발 수요에 부응하는 플랫폼과 인프라를 구축

❑ 기술개발

❍ 과기정통부를 중심으로 대학과 파트너십을 체결하여 연합학습 원천기술을 개발하여 플랫폼의 안정성과 고도화에 기여

❑ 플랫폼 운영

❍ 산업계 대표단체에서 K-MELLODDY 프로젝트를 관리하며 제약바이오 기업과 공공기관이 플랫폼을 지속 활용할 수 있도록 지원하여 연합학습 기반 AI 신약개발 모델의 실용성 제고 및 성능 향상

❍ (K-MELLODDY 추진위원회 구성) 보건복지부와 과기정통부를 중심으로 산·학·연·병 전문가 10인 내외로 추진위원회를 구성하여 데이터 파트너십, K-MELLODDY 플랫폼 구축, 플랫폼 운영, 시범사업 수행 전반의 추진계획을 수립·시행

❍ (데이터 파트너십 체결) 과기정통부와 보건복지부가 주도하여 K-MELLODDY 프로젝트에 참여할 데이터 보유 공공기관, 제약기업, 바이오벤처, 의료기관 간의 파트너십 체결

❑ 활용사업 운영

❍ 보건복지부를 중심으로 공공기관, 제약 바이오 기업이 활용사업에 참여토록 하고, 시범사업으로 축적된 연합학습 기술의 안전성과 기술적 신뢰도를 바탕으로 사업과제 및 참여기관을 지속 확대

❍ (제약산업계 사업 필요성 확인) 한국제약바이오협회 자체 조사 결과 다수의 제약기업이 사업의 필요성에 공감하였으며, 특히 ADME/Tox 예측 모델 개발 의견을 피력함

※ 향후 공공기관이나 의료기관이 데이터 파트너십에 참여하는 경우 사업의 완성도가 더욱 높아질 것으로 기대

## 5.4. 사업 및 성과관리 방안

### 5.4.1. 사업 관리 방안

❑ 목표관리 집중형 상시과제 모니터링 체제 구축을 통한 사업의 성공률 향상 및 목표 달성

❍ 사업 성공률 향상 및 목표 달성을 위해 사업 목표 달성도, 달성 가능성, 위협요인 등에 대한 상시 모니터링 실시

❍ 전문기관은 동 사업의 추진과정에서 과제 선정 이후 연차 또는 단계평가, 중간평가, 최종 평가 및 필요시 수시 검토를 통해 과제 초기와 비교하여 수요 변화와 과제 목표 달성 정도 지속 점검

– 과제의 주요 마일스톤 별로 기존 평가위원을 활용하여 수요 및 외부 연구개발 동향, 목표 달성 가능성 등을 종합 점검하며 필요시 종합진단팀을 구성하여 정밀 점검실시

– 정밀 점검 시 마일스톤 별 점검의 주요 방향에 따라 수요 부처나 수요기관, 외부 기술전문가, 시장 전문가 등 관련 전문가를 확대

❍ 점검을 통해 모니터링 결과를 과제 내에서 반영할 수 있는 경우 해당 과제를 추진하며 성공 가능성이 크지 않으면 과제 추진 방향 변경을 통한 추진 또는 과제 중단 등 후속 조치를 추진

표 33. 상시 모니터링 및 평가 항목

| 구 분 | 항목 |
|---|---|
| 환경 및 수요 | • 정책 및 환경변화에 부합되는 이슈를 반영하고 있는가?<br>• 사업에 영향을 미칠 수 있는 최근 정책 변화는 없는가?<br>• 사업에서 다루고 있는 주요 이슈 중에서 그 중요성이 현저히 떨어진 이슈는 없는가? |
| 연구개발 동향 | • 최근 발표된 중요한 연구개발 성과는 무엇인가?<br>• 주요 선진국 연구기관에서의 최근 중요한 연구 방향의 변화는 없는가? |
| 사업 목표 | • 당초 계획된 사업 목표가 어느 정도 달성되었는가?<br>• 최근 환경변화 및 연구개발 동향 변화에 따라 사업 목표의 수정이 필요한가? |
| 사업 수행 | • 당초 계획된 사업 진도가 차질 없이 진척되고 있는가?<br>• 목표 달성을 위해 현재 수행되는 연구 방법이 적절한가?<br>• 사업 목표와 진도에 따라 연구자원이 적절하게 배분 혹은 집행되고 있는가? |
| 사업 성과 | • 당초 계획된 연구성과가 창출되고 있는가?<br>• 사업을 통해 창출된 성과 정보가 적절하게 관리되고 있는가? |
| 추진 계획 | • 향후 사업 추진 일정이 적절하게 수립되어 있는가?<br>• 추진 일정의 조정이 가능한 프로그램이나 과제는 없는가?<br>• 향후 추진계획에서 환경변화를 적절히 반영하고 있는가? |

❍ 사업 종료 이후에는 성과관리 및 추적평가를 통해 기술개발 결과의 현장 적용도 및 실증화 실적을 지속 점검하여 필요한 조치를 함으로써 성과활용도 향상

- 공공기술의 경우 성과관리 및 추적평가를 통해 종료된 기술개발 과제가 현장에 적용되고 있는지와 효과성 파악
- 실용화·실증화 기술의 경우 기술 개발 성과가 민간 기업으로 이전되어 사업화 가능 여부와 사업화 성과, 성과향상을 위한 니즈 파악
- 현장 적용 또는 사업화가 미흡한 우 필요한 기술개발 니즈를 관련 과제의 기술개발에 반영하거나 후속 과제 기획을 통해 해당 성과 제고
- 성과가 우수한 과제의 성공 요인, 현장 적용이 미흡한 과제의 성공도 향상 요인 분석을 통해 연차별 사업 기획에 반영

❍ 성과분석의 지속성 확보를 위해 관리 주체는 한국보건산업진흥원으로 하여 운영

❍ 한국보건산업진흥원이 사업수행기관으로부터 성과자료를 제출받아 확인, 취합하여 성과자료 DB를 구축

- 사업수행기관의 성과자료 제출 등을 기초로 양적지표에 대한 성과분석
- 확인된 성과자료를 활용하여 성과검증·분석 업무를 수행
- 필요시, 질적 지표에 대한 평가가 필요한 경우 전문가 등 대상으로 조사 업무 수행

❍ 한국보건산업진흥원은 성과분석 결과를 확정하고 지원유지 여부 결정이나 예산조정 반영 등 사업 운영에 피드백

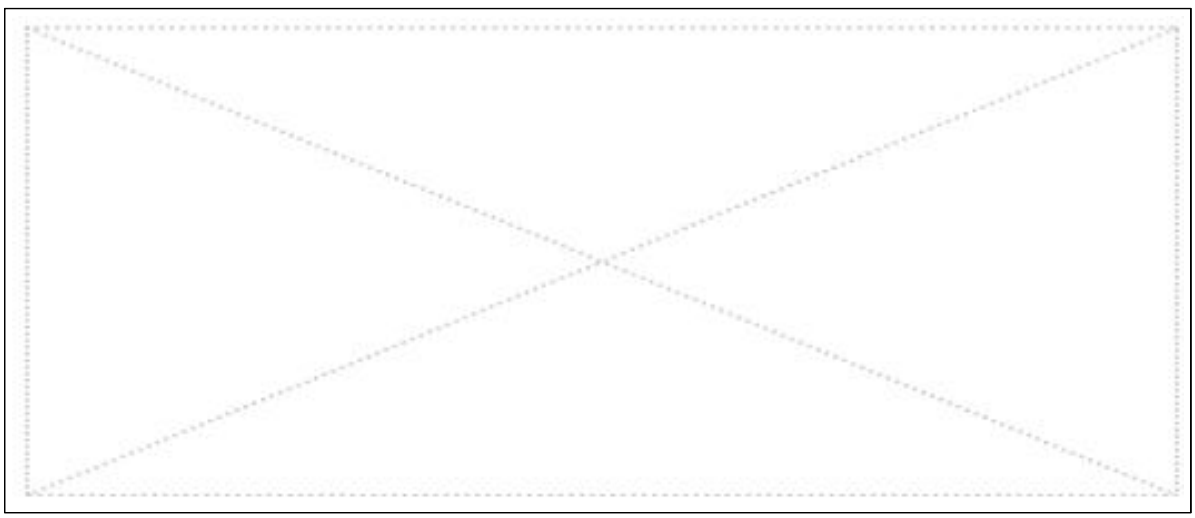


그림 77. 성과관리 체계 및 절차

❑ 사업 별 지표 선정

| 세부 사업명 | 연합학습 기반 신약개발 가속화 프로젝트(K-MELLODDY) | | |
|---|---|---|---|
| 사업 별 성과 목표 | **플랫폼 구축 운영 및 사업 관리** | **플랫폼 활용 사업** | **연합학습 원천기술개발** |
| | 연합학습 기반 신약개발 가속화 프로젝트 사업단<br>총괄과제 및 사업단 사무국 운영 | 데이터 생산 및 공급 연합학습 검증 | 연합학습 기술 구현 및 고도화 |
| | • 연합학습 플랫폼 운영 3건(3년간 연도별 1건씩 누계)<br>• 연합학습 플랫폼 운영관련 SOP 2건<br>• 연합학습 프레임워크 개발 및 검증 보고 1건<br>• 플랫폼 인증 1건<br>• 플랫폼 개발 보고 1건<br>• 플랫폼 검증 보고 1건 | • 데이터 품질 및 생산 보고 20건 (기업별 1건)<br>• AI 모델 개발 10건 이상<br>• AI 모델 예측값 실험 검증 보고 10건(주제별 1건) | • 연구성과 발표 SCI(E) 급이상 논문 5건, 특허 5건 |

| 성과지표별 최소 요구성과 | 성과지표명 | 최소 요구성과 | 비고 |
|---|---|---|---|
| | 연합학습 플랫폼 운영 | 3건 | • 플랫폼의 구동여부 확인<br>• 플랫폼 구축 및 운영 건수는 누적치로 계산('26-'28년, 연도별 1건씩) |
| | 연합학습 플랫폼 운영 SOP | 2건 | • 외부전문가의 평가를 통해 객관성 확보 필수<br>• '24년 1건, '25년 1건 |
| | 플랫폼 인증 | 1건 | • 플랫폼의 TTA SW 인증 시험 통과 |
| | 플랫폼 개발 보고 | 1건 | • 플랫폼의 설계, 기술, 개발 내용을 확인할 수 있도록 소프트웨어 개발 보고서 작성 |
| | 플랫폼 검증 보고 | 1건 | • 플랫폼 작동·보안 문제 등에 대한 외부전문가의 감사보고서 제출 필수<br>• '26년 1건 |
| | 데이터 품질 및 생산 보고 | 10건 | • 최소 10개 기관 이상 참여(기관 별 1건의 품질 및 생산 보고) |
| | AI 모델 개발 | 10건 | • AI 모델 개발 및 성능에 대한 외부 전문가의 검증을 받은 보고서 제출 필수<br>• '26년 4건, '27년 3건, '28년 3건<br>• 활용사업 주제별 최소 1건(10개 주제 예상) |
| | AI 모델 예측값 실험 검증 보고 | 10건 | • 신약개발 주제별 모델당 1건 수행<br>• 모델 입력에 사용한 화합물의 실제 실험 결괏값과의 일치정도를 검증<br>• 실험값과의 비교 분석으로 모델 실용성 검증 |
| | 연합학습 프레임워크 개발 및 검증 보고 | 1건 | • 연합학습 프레임워크 개발 결과 및 구동 및 성능 보고 |
| | 연구성과<br>논문, 특허 | 논문 4건,<br>특허 4건 | • 연합학습 보안 강화, 알고리즘, 절차, 기여도 평가 별 논문과 특허 등록 |

## 5.5. 타당성 분석

### 5.5.1. 사업추진 시의성

❑ (수요 확인) 20대 혁신형 제약기업의 프로젝트 수요 확인

❍ (일시) 2022년 11월 16일

❍ (행사명) AI 신약개발 연구실장 자문 간담회

❍ (목적) 기획안 소개 및 과제 참여 잠재 제약기업 및 AI 신약개발기업 파악

❍ (참여기업) 21개 제약기업의 사업 참여 의향 확인

❍ (발표 내용) EU MELLODDY 사례 소개 및 K-MELLODDY의 비전, 목표 등

❍ (응답 결과) ADME/Tox 데이터는 다른 데이터에 비해 비교적 오랜 기간 축적되었기에 기획 중인 사업 수요가 있음을 확인

- 데이터 보유량의 편차가 있지만, 데이터 공급 가능 확인
- 제약기업의 특허권이 소멸된 화합물 경우 활용가능확인
- 특허권이 있는 경우, 데이터를 연합학습에 활용했을 때, 해당 모델에 대한 용도특허가 가능한지 법적 해석과 합의가 필요하며, 모델의 유료 서비스로 인한 수익 공유로 데이터의 권리를 보장
- '22년 11월 7개 제약기업 대상 자문회의 결과

표 34. 기업별 데이터 종류 및 보유 현황 추정치

| 순번 | 기관명 | 데이터 종류 | 보유 추정치 |
|---|---|---|---|
| 1 | D사 | (1) ADME/Tox Prediction Model<br>(2) pre-clinical/clinical Prediction Model | 약 50만개 |
| 2 | I사 | ADMET & Physicochemical property | ADME 7,000개,<br>Tox 3,000개,<br>DTI(Kinase) 5,000개 |
| 3 | Y사 | (1) 대사안정성, (2) CYP inhibition,<br>(3) 투과성 예측, (4) DTI 예측 | 각 항목별 500~2,000개,<br>(화합물 4만종 이상) |
| 4 | D사 | 물질에 대한 In vitro ADME 관련 예측 | 총 2만 5천여 개 |
| 5 | J사 | (1) 용해도 (Solubility) 예측<br>(2) 투과도 (Permeability) 예측<br>(3) DTI 예측 | 총 2,000개 이하<br>(3) 문항의 경우 수백개 |
| 6 | H사 | 물질에 대한 In vitro ADME 관련 예측 | 화합물 약 1만종 |
| 7 | H사 | | 화합물 1만 5천여종 |

### 5.5.2. 플랫폼 개발의 가능성

❑ 오픈소스에서 솔루션으로의 진화하였으나 한국은 늦음

❍ 아래 표에 제시한 연합학습 프레임워크는 오픈소스로 연합학습 기술재현하고 각자 활용을 위한 특징을 강화하여 개발되었음

❍ 약 2년전부터 오픈소스 프레임워크를 빗대어 솔루션화하여 비즈니스를 gk는 솔루션 업체가 속속 등장하고 있으나, 한국에서는 현재까지 구축한 사례가 없음

- 해외 사례: Owkin, Google, Nvidia, Apheris, Adap, Intel labs, Edge Delta, Bitfount, Rhino Health, lhasa limited 등
- 국내 사례: 카카오헬스케어 덱스콤, 구글 클라우드, 시그니처 헬스케어 3사 업무 협력 체결, 모바일 기반 혈당관리 솔루션 제공을 위해 연합학습 기술 활용(23.04.28)
- 다수의 오픈소스 프레임워크에서 기술 개발에 참고(레퍼런스 역할)

표 35. 연합학습 프레임워크 개발 현황

| 제공 단체 | 프레임워크명 | 주요 추진 내용 |
|---|---|---|
| Adap | Flower | • 연합학습을 위한 기본 프레임워크를 제공하며, 기계학습 프레임워크의 제약이 없이 모델 개발이 가능<br>• 안드로이드, IOS, Raspberry Pi, Nvidia Jetson까지 지원하여 Device부터 Silo 데이터까지 연합학습 할 수 있음 |
| NVIDIA | Clara | • 실시간, 보안 및 확장성 있는 솔루션을 만들기 위한 개발자, 데이터 과학자, 연구원을 위한 전체 스택 GPU 가속 라이브러리, SDK 및 참조 응용 프로그램 제공<br>• 유전자 염기서열 데이터 분석용 가속 플랫폼 Clara Genomics 제공 |
| OpenMind | Pysyft | • Pytorch, Tensorflow 기반 프레임워크에서 모델 개발이 가능하며, 개인정보보호를 위한 DP, MPC, HE 기능이 있음<br>• 현재 기관에서 추후 운영하지 않아 비활성화 상태인 것으로 확인됨 |
| Webank | Fate | • 연합학습 기본 아키텍처 제공, 다양한 연합 기계학습 알고리즘 제공, 멀티 클러스터 연합학습 구축을 위한 패키지 제공<br>• 모든 함수가 모듈화되어있어 딥러닝 알고리즘으로 연합학습 모델 개발 시 제약사항이 많음 |
| IBM | IBM Federated Learning | • 연합학습 기본 프레임워크를 제공, 기계학습 프레임워크의 제약 없이 모델개발이 가능함<br>• ●프레임워크에는 클라이언트, 서버, 통신 지정 기능 |
| USC 대학 | Fedml | • 기계학습, 자연어처리, 그래프, 컴퓨터 비전, IoT, 모바일에 대한 폭넓은 연합학습 라이브러리 제공<br>• 벤치마크 데이터 세트로 Federated EMNIST를 제공하고 있으며, 연합학습의 벤치마크를 처음으로 시도했다는 점이 특이사항 |

❍ 연합학습 기술의 국내 연구팀도 존재

– 카이스트 예종철 교수팀이 'COVID-19 흉부 X-Ray(CXR) 진단을 위한 연합 분석 비전 트랜스포머 FeSTA(Federated Split Task-Agnostic) 모델 연구'를 통해 기술의 안전성과 성능 향상을 실증한 바 있음(NeurIPS, 2021)

– 서울대학교 김중헌 교수팀이 'COVID-19, X-Ray, 콜레스테롤 분류 모델 학습을 위한 다중 사이트 분할학습 타당성 연구'를 통해 안전성을 실증한 바 있음(Nature, 2022)

❑ 연합학습 이외의 플랫폼은 서비스 차원의 개발

❍ 연합학습 기술(백엔드) 외 개발내용은 주로 팀 구성, 팀 참여, 해체, 클라우드 기반의 데이터 업로드, 관리, 기록 확인, 접근 제어, 연합학습 모니터링, 성능 확인, 커뮤니티, 위험 대응 등의 프론트엔드 개발 중심

❍ 국내 클라우드를 활용할 계획으로 국내 클라우드 운영과 AI·SW 개발을 동시에 수행하고 있는 기업이 적합할 것으로 판단

– 후보) 카카오 엔터프라이즈, KT 클라우드, 네이버 클라우드

## 5.5.3. 정부 지원 필요성

❑ 참여 기관의 부담 완화

❍ 인공지능 신약개발을 위한 협업이라는 새로운 협력 방식은 전 세계에서도 사례가 1개밖에 없을 정도로 선도적인 분야이므로 기업이 해당 분야에 투자하기에는 리스크가 크며, 경쟁적 관계에 있는 기업들의 협력이 필요하기에 개별 기관이 프로젝트를 맡아서 수행하기 어려움

❍ 제약사, 바이오, AI, IT 기업 등 다양한 이해관계자가 참여할 필요가 있는 사업 특성상 정부가 주도하여 안전한 신약개발 데이터의 협력을 통한 고성능 인공지능 기술 확보의 성공사례, 성공 가능성을 보여주면, 향후 기업이 적극적으로 따라올 것으로 예상함

❍ 데이터 공유 협력을 위해서는 참여 기관이 보유한 데이터의 가치 평가가 중요한데, 다른 주체보다 정부에서 데이터의 가치를 객관적으로 평가·산정해줄 수 있다고 봄

❑ 제5차 과학기술 기본계획의 민간 중심 혁신생태계 조성과 데이터 플랫폼 연계 및 맞춤형 데이터 확산의 디지털 전환 전략에 부합

❍ [전략2] 혁신 주체의 역량 제고 및 개방형 생태계 조성에서의 민간 주도 전략의 목표에 제시된 기업의 연구 역량과 기술 혁신을 중심으로 연구개발 지원 체계 고도화와 본과제의 기업 주도 참여 전략이 일치함

❍ [전략3] 과학기술기반 국가 현안 해결 및 미래 대응의 디지털 전환 전략의 목표에 제시된 공공·민간 데이터 통합, 데이터 유통 및 활용 활성화와 K-MELLODDY의 공공 민간 데이터 협력 학습 방법이 일치함

❍ 제5차 과학기술 기본계획(22.12.14.)



그림 78. 제5차 과학 기술 기본계획의 주요 방향(민간 참여 확대)

❑ 제약산업 육성 및 지원에 관한 특별법

❍ 2022.6.10. 발표된 제약산업 육성 및 지원에 관한 특별법의 제4조에 제약산업 발전 기반 조성 및 국제 경쟁력 강화를 촉진하기 위하여 5년마다 제약산업육성·지원종합계획을 수립하라는 조항이 있는데 종합계획에 포함되는 사항에 인공지능을 이용한 신약개발 지원계획이 있음[28)]

❍ 신약개발에 인공지능 기술 적용을 가능하게 하기 위한 제약사 간 협업의 새로운 방식을 제안하고 협력 체계를 구축하는 K-MELLODDY 프로젝트의 목적은 제약산업법의 육성·지원 계획과 부합함

❑ 제3차 제약산업 육성지원 종합 계획 비전 및 목표[29)]

❍ 2023.03.24. 보도된 바이오헬스 글로벌 중심국가 도약을 위한 제3차 제약바이오산업 육성·지원 5개년 종합계획발표자료의 글로벌 신약 창출을 위한 R&D 투자 확대의 AI·빅데이터 등 신약개발의 디지털 전환 촉진 언급

❍ 보건복지부와 과학기술정보통신부의 협업을 통한 공공 인공지능 신약개발 플랫폼의 고도화 및 후보물질 도출부터 임상시험 신청까지의 성과 창출을 위해 수행 예정 프로젝트 2가지 중 하나가 K-MELLODDY 사업

- 연합학습 모델을 기반으로 다기관에 분산된 보건의료 데이터 등 민감 정보를 효과적으로 활용하는 'K-MELLODDY' 사업을 통해 신약개발을 가속화

28) 제약산업 육성 및 지원에 관한 특별법(제약산업법), 2022.06.10. 시행,
29) 대한민국 정책브리핑, 제3차 제약바이오산업 육성·지원 5개년 종합계획 발표, 2023.03.24.

### 5.5.4. 벤치마크 사업과의 차별성

❑ 벤치마크 사업과의 차별성

❍ EU의 MELLODDY 프로젝트와 다양한 환경 요인의 차이 존재. 제안하는 프로젝트는 제약바이오기업 중심의 문제 해결형 프로젝트 전략을 수립

- 정부지원금으로 생산된 데이터는 정부에 기탁하는 것이 원칙이나, 특수한 경우는 참여자들과의 합의가 필요 (기업이 보유한 화합물 미공개, 정부지원금으로 데이터 생산을 통해 실험값이 만들어졌을 경우, 정부지원과 기업의 이익이 충돌)

표 36. 벤치마크 사업과의 차별성

| 비교 항목 | EU MELLODDY | K-MELLODDY |
|---|---|---|
| 목적 | 연합학습 기술에 대한 실증사업 (사업의 활용방안 부분이 부족) | 연합학습 기술에 대한 활용사업 (본사업을 플랫폼화하여 데이터 활용 생태계를 조성하고자 함) |
| 추진 주체 | Owkin(연합학습 솔루션)과 Kubermatic(클라우드 플랫폼)과 같은 인프라 ICT 기업을 중심으로 추진 | 다기관(학교, 공공기관, 연구소, 제약·바이오, AI, IT 기업 등)의 협력이 필요하고, 자국의 신약개발 경쟁력을 확보하기 위하여 산·학·정 협력 필요 |
| 데이터 현황 | 참여 제약사는 글로벌 빅파마로 오랜 기간에 걸친 신약개발 연구 경험과 다양한 신약개발 파이프라인을 보유해 공유가능 데이터가 많이 존재 | 글로벌 제약사보다 규모가 영세해 공유가능 데이터가 상대적으로 적지만, 약물 발견과정의 실험 결과인 ADME/Tox는 공유 활용이 가능한 수준이며, 부족 데이터는 제약사를 지원하여 생산 |
| ICT·인프라 환경 | NVIDIA, Owkin, Kubermatic의 연합학습, 인프라 전문기업이 참여해 원천기술을 확보한 상태로 사업수행 | 자체 조사에 따르면 국내의 경우, 대학, IT 및 클라우드 기업이 협력하여 연합학습 원천기술 연구개발이 필요 |
| 데이터 공개 범위 | 데이터 공개 없음, 기탁도 없음 (단, 기본 모델 학습에 활용한 전처리된 공개 화합물 데이터는 공개) *동 사업에서 해당 전처리 데이터를 사전 훈련된 모델 개발에 활용 가능 | 기업이 기보유한 데이터는 공개하지 않음 (단, 특허로 공개된 화합물인 경우, 과제비로 생산한 데이터의 경우 공개) 국내 연구자들에게만 공개 |
| 모델 공개 범위 | 모델의 구조와 사용법만 공개 학습된 모델 결과(가중치)는 비공개 (단, 참여기업은 연합학습된 모델 공유) | 모델 연합학습에 기여하는 참여자들에게 공개하는 것을 원칙으로 함(추후 참여자도 협의를 통해 승인받으면 가능) 모델의 구조와 사용법은 공개 |
| 플랫폼 공개 | 개발된 플랫폼 미공개(참여자들이 활용) (Owkin, substra 연합학습 솔루션만 공개) | 개방형 플랫폼으로 공개 (연합학습 팀 구성, 데이터, 전처리기, 모델, 연합학습 추적 등 기능 탑재 예상) |

### 5.5.5. 기존 플랫폼 사업과의 차별성

❑ 수행 중인 데이터 구축사업, AI 신약개발 플랫폼과의 비교 분석을 통해 “**신약개발 민간 데이터의 고립**을 **해제**하고, **민간 데이터의 지속 가능한 협력 체계 및 AI 신약개발 모델의 개발**”이 제안 사업의 차별점임

표 37. 수행 완료 및 수행 중 기존 플랫폼사업과의 차별성

| 구분 | 국가 바이오 빅데이터 구축 시범사업 | AI 신약개발 플랫폼 구축사업 (KAIDD) | AI 활용 혁신신약발굴사업 | K-MELLODDY 사업 |
|---|---|---|---|---|
| 사업 규모 | 총 345.51억<br>‘19~‘21(2, 2년) | 총 142.05억<br>‘19~‘21(2+1, 3년) | 총 88.6억<br>‘22~‘26(5년) | 475억<br>‘24~’28(5년) |
| 선정 규모 | 총 4개 과제 | 총 6개 과제 | 총 4개 과제 | 총 4개 과제(예상)<br>과기부(2개)<br>복지부(2개) |
| 주무 부처 | 복지부, 과기부, 산업부 | 과기부, 복지부 | 과기부 | 과기부, 복지부 |
| 전담 기관 | 연구재단 | 연구재단 | 연구재단 | - |
| 사업 목표 | 정부재정이 투입된 연구개발 사업과 국가 바이오 빅데이터 구축 시범사업과의 연계 | 글로벌 신약개발에 필요한 인공지능 플랫폼을 구축 소요 시간 및 비용 단축 | 공공플랫폼(KAIDD) 고도화 및 후속 성과 도출 | 분산된 민간 공공 데이터의 안전한 활용 및 협력이 가능한 한국형 연합학습 기반 AI 신약개발 플랫폼 구축, 사례 발굴 |
| 사업 내용 | 검체 확보 및 임상·유전체 데이터 연계, 100만명 연계 규모 확대를 고려한 업무 프로세스 검증, 고품질의 임상·유전체 데이터 및 활용·연구성과 등의 확보 | 후보물질 발굴, 약물재창출, 약물감시 분야에 특정질환 적용 AI 모델 개발, 공공플랫폼(KAIDD) 개방 및 운영 | AI 플랫폼을 활용하여 IND 신청가능 수준의 신약 후보물질 발굴 AI 모델 추가 개발을 통한 공공 플랫폼 운영 및 고도화, 서비스 활성화 | 민간-공공 데이터 연구 비밀의 안전을 보장한 협력 체계 및 AI 신약개발 모델 학습 및 개발이 지속 가능한 플랫폼 구축, 이를 활용하여 산재한 신약개발 데이터 연합학습으로 거대 AI 모델개발 |
| 사업 특징 | 정부재정이 투입된 **연구데이터를 통합**·연계하고, **직접적인 임상·유전체 데이터를 생성 확보**, **활용 체**계 **구축** | 신약개발에 활용 가능한 **AI 기반 도구(모델)를 개발**하는 것이 목표로 **도구 개발에 공개된 데이터 활용** | 신약개발에 활용 가능한 **AI 기반 도구(모델)**의 **분야**를 **확장**하여 **추가 개발**하고, **기존 플랫폼을 고도화** | 신약개발 **민간 데이터 고립을 연합학습** 으로 해제하고, **민간 데이터의 지속 가능한 협력 생태계와 AI 신약개발 도구 개발 활용** 가능 |

## 5.6. 활용 방안

❑ AI 신약개발을 위한 데이터 활용 중개 플랫폼

❍ K-MELLODDY 사업으로 구축한 연합학습 플랫폼은 사업 종료 후에도 산업계에서 다양한 용도로 활용 가능하며, 기업 간 서로의 미충족 영역을 충족시켜 줄 수 있는 창구로서 역할 기대

❍ (활용 및 확산) K-MELLODDY 사업으로 구축한 연합학습 플랫폼은 사업 종료 후에도 산업계에서 신약개발의 다양한 용도로 활용할 수 있음

❍ (수요와 공급의 연결) 연합학습 플랫폼은 모델 학습의 참여 기관(로컬 클라이언트) 구성을 여러 형태로 조정할 수 있기에 데이터의 수요처와 공급처를 연결할 수 있음

- AI 신약개발기업은 모델 학습을 위해 다양한 다량의 데이터가 필요하지만, 충분한 데이터를 확보하는 것이 어려움
- 반면, 공공기관이나 제약바이오기업은 데이터를 보유하고 있지만, 프로젝트 종료 등의 이유로 이를 활용하지 못하고 방치되는 경우가 있음



그림 79. 데이터 중개 플랫폼(안)

❑ 데이터의 안전한 AI 학습을 위한 플랫폼

❍ (데이터 중심 협업 도구) AI 신약개발 기업과 제약바이오기업이 협업을 진행할 시, 제약바이오기업의 보유 데이터를 활용하는 것이 프로젝트 성공률을 높일 수 있으나, 이를 제공하는 것은 부담으로 작용. 연합학습 플랫폼을 활용하면 안전하게 유출 위험 부담을 줄이고 데이터 협력 진행 가능

❍ (거대 지식 AI 확보) 연합학습 참여기관의 로컬 데이터 학습 결과인 학습 파라미터를 취합하는 방식 즉, 로컬 데이터의 지식 학습 결과 융합이라는 점에서 방대한 지식이 학습된 AI의 확보가 가능하고, 신약개발뿐만 아니라 다른 분야에 활용할 수 있음

❍ (사업 확장) 약물 탐색 이외에도, 표적 발굴이나 마커 발굴 등 신약개발의 다른 분야에서도 사업을 확장하여 활용할 수 있음

❑ 연합학습된 AI 기반 예측 모델을 활용한 유/무료 서비스 시행

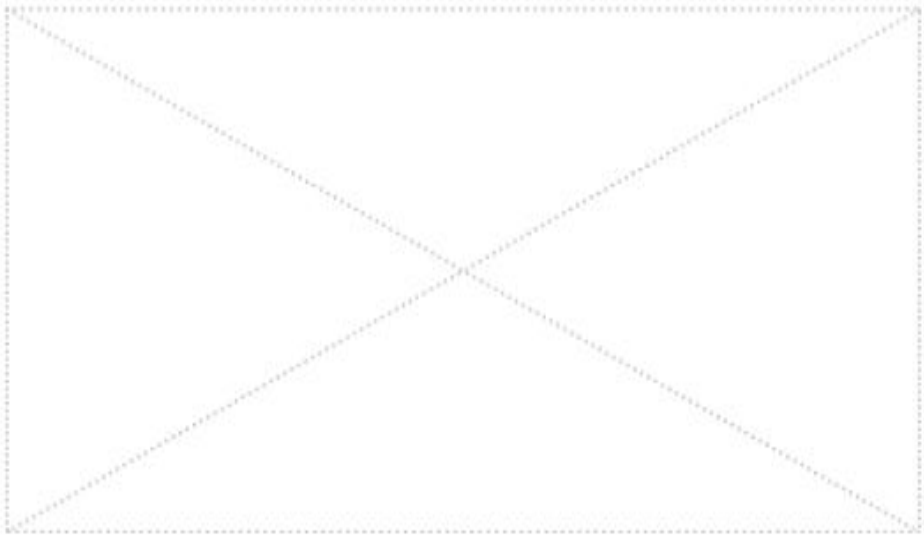
❍ 일본 “창약지원 인포매틱스 시스템 구축사업”에서는 ADME, 약동학 예측 모델의 예측 서비스를 유료화

– 기본판 DB(공공 DB)로 만들어진 시스템은 그대로 공개하고, 민감한 기업 데이터가 들어간 AI 예측 모델은 해당 기업에서만 활용함

– 동 사업에서도 개발한 연합학습 기반 AI 신약개발 시스템을 참여한 기업들의 신약개발에 활용하고, 사업 성과로 도출되는 예측 AI 응용 프로그램을 기본판 버전(무료), 유료화 버전, 기업 내부용 등 다양한 형태로 개발하여 국내 기업, 학계에도 공개하여 사용할 수 있도록 함

❍ 예상되는 서비스 형태로는 사용자들이 웹사이트의 시스템을 통해 예측 결과를 얻는 것 외에도, 사용자들이 학습데이터를 직접 업로드하여 자체적인 AI 모델을 개발할 수 있는 서비스를 제공함

❍ 이를 통해 데이터 확보가 어려워 자체 솔루션을 갖기 어려운 중소기업에게도 새로운 사업 기회 제공



(출처 : Fujitsu)

그림 80. 후지츠큐슈 시스템의 SCIQUICK 서비스 개요

❍ 일본 “창약 지원 인포매틱스 시스템 구축사업”(15.4 – 20.3)을 통해 개발된 AI 신약개발 플랫폼 “SCIQUICK”는 공공 DB와 7개 기업이 제공한 데이터로 구축됨

– 플랫폼의 예측 모델은 ADME, 심장 독성, 간 독성, 후지츠 자체 모델 보유

• (ADME) 막투과성, 용해도, 흡수율, P-gp, 뇌 homogenate 결합, 혈장 단백질 결합, 간 고유 클리어런스, 요중배설형, 요중미변화체배설률,

신장 클리어런스

- (심 독성) hERG 저해 모델
- (간 독성) 담즙 울체성 간 독성, 세포 장애성 간 독성, 간암, 약물 유발성 간 독성
- (후지츠 자체 모델) AMES, 피하 감작성, 발암성, CYP 저해
- 사업화 프로젝트는 후지츠가 실용화 프로젝트를 통해 서비스를 추가 개발하고 제공

❍ 상기한 기계학습 모델을 SCIQUICK에서 모델 작성 서비스를 제공하고 있음

– 이 서비스는 유료로 공개되고 있으며, 이용료는 연간 라이선스 80만엔, 영구 라이선스 200만 엔으로 책정되어 있음

## 5.7. 기대효과

❑ R&D 효율화

❍ 연합학습 기반 ADME/Tox 예측 모델 개발로 4,000억 원의 직접적 R&D 투자비 20% 절감, 인산화효소 활성 저해 예측 모델 개발 등 사업을 확대해 나간다면 국가 및 민간의 신약개발 R&D 비용을 1조원 이상 절감

| ADME/Tox(흡수, 분포, 대사, 배설, 독성)분석 비용은 화합물 1종 당 1억 원 내외이며, 1개의 파이프라인 당 10종의 화합물을 분석하고, 국내에 진행 중인 후보 물질 및 비임상 시험 단계 파이프라인이 407건임(국가 신약개발 사업단, 2022.06.29.)<br>• 추산 [1억 원(1종 분석비) X 10종(화합물) X 407개(파이프라인 수)=4,070억 원]<br>• 연합학습 기반 ADMET 예측모델은 이 비용을 1/2수준으로 낮추고 분석 기간(6개월)도 1개월 이내로 단축할 것으로 예상 |
|---|

❑ 공동연구 촉진

❍ 제약바이오산업의 디지털 전환과 AI 기술 도입을 촉진하여 저비용 고효율 AI 신약개발로 패러다임 전환 가속화

❑ 국내 제약산업 글로벌 경쟁력 확보

– 연합학습 기반 플랫폼 구축을 통한 국내 제약산업 협력 체계를 조성하여 분산된 제약산업의 투자와 방향을 응집함으로써 투자의 결집화, 신약개발 변화혁신 창출로 글로벌 제약기업과의 경쟁력 강화

❍ 학습된 모델의 부분 또는 전체 공개를 통한 스타트업도 활용할 수 있게 하여 AI 활용 신약개발 비용 절감 효과 확산

❑ 기술 패권주의에 대응

❍ 글로벌 빅테크 기업의 AI·IT 기술의 승자독식 구조의 심화, 한국의 기술 및 데이터 주권을 확보하기 위해서는 데이터를 안전하게 AI에 활용하는 연합학습 기술의 실용화와 실용화 주도로 글로벌 표준국 도달이 우선 목표