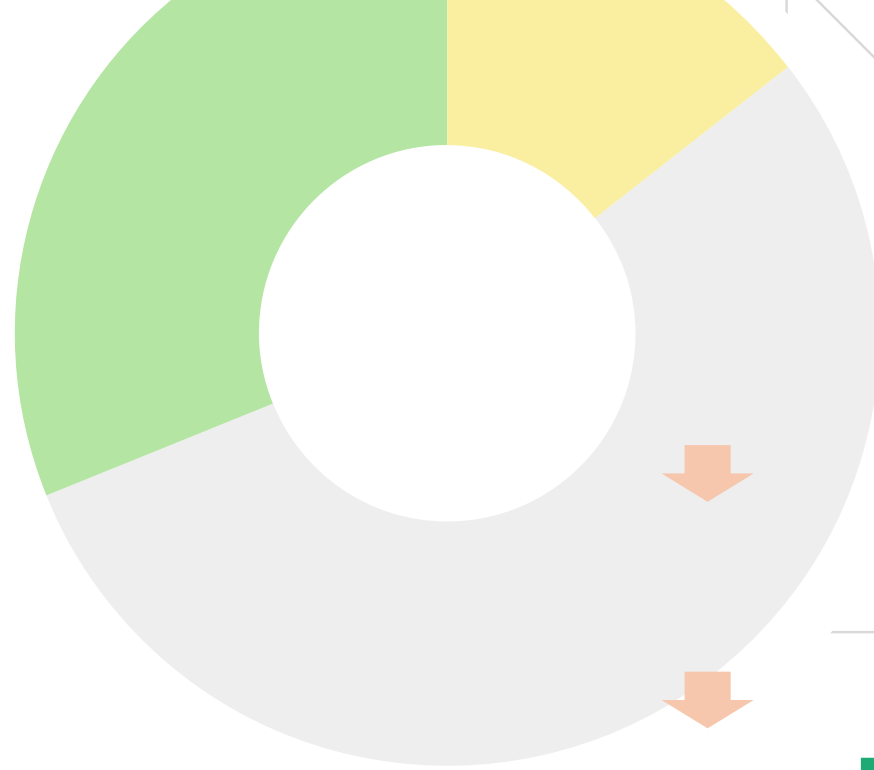


감염병 대응체계 개선을 위한 AI 기반 사망원인 예측모델 개발

이석민, 전고설, 최지원

01

Introduction



현행 업무 한계

01 수작업 사인분류의 한계

- 담당자 수작업 대조/원사인 분류
→ 오류 가능성 및 **공표시차 지연**
→ **보건분야 정책** 대응 한계

02 국제표준 35년 만의 대규모 개편

- 국제질병사인분류(ICD) 전면 개편
(ICD-10 → ICD-11 전환)
→ **사망원인 자동 코드화 제한**

03 사망원인분류 업무 급증

- 고령화로 인한 사망자 증가(연간 36만 명)
→ **업무량↑, 사인분류 복잡도↑**

연구 목적



사인분류 정확도 향상 + 행정효율 증가

- AI 모델이 사인분류 결과를 먼저 제안
→ 급증하는 업무량 대응
→ 약물 성분 등 사인분류 오류 감소



정책수요 대응 서비스 개발 기반 마련

- **감염병 등 잠정 원사인 예측 서비스**
→ 최종 공표 이전에 특정 사인 동향을 AI가 예측하여 서비스 할 수 있는 기반 구축
- **자살동기별 사망자 통계 서비스**
→ 자살동기를 AI가 분석하여 관련 정책을 지원할 수 있는 통계 개발 기반 마련

기존 사인분류사의 최종 판단 구조 및 WHO의 분류지침은 유지하되,
AI가 예측 결과를 먼저 제시하여 업무 효율을 높이고, **부가 서비스를 개발**하는 방식으로 설계

해외 사례 : 일본 인구동태통계 서비스

후생노동성, 단계별 인구동태 통계 제공 체계

STAGE 1

매월 2개월 후

인구동태통계 속보

조사표 작성 건수대상 집계

STAGE 2

매월 5개월 후 · 다음해 6월

인구동태통계 월보

잠정 인구동태 수치 제공
성/연령/지역/사망원인별 사망자 수
감염병 사망자 대상 별도 통계표

STAGE 3

다음해 9월

인구동태통계 연보

확정 인구동태 수치 보정 후 제공

REFERENCE

ひと、くらし、みらいのために



일본 후생노동성

Ministry of Health, Labour and Welfare

2025년 11월 인구동태 월보

(2026년 4월 발표)

주요 통계표

- 1 표 1 인구 통계 개요 - 전년도와의 비교 -
- 2 표 2 월별 인구 조사
- 3 표 3 인구 통계 (확진자 수, 부도부현별(특별 구역 - 지정 도시 재등록)
- 4 표 4 사망자 수, 사망률(인구 10만 명당 사망률), 사망 원인(사망 원인의 단순 분류) - 전년도와의 비교 -
- 5 표 5 영아 사망자 수, 영아 사망률(10만 명), 사망 원인(영아 사망의 단순 분류) - 전년도와의 비교 -
- 6 표 6 사망 원인별 사망자 수(선택적 사망 원인 분류), 성별, 연령(5세 연령대)
- 7 표 7 감염병으로 인한 사망자 수 및 사망 원인(감염병 분류) - 전년도와의 비교 -
- 8 저장 표 1 출생 수, 어머니 나이(각 연령), 출생 순서, 성별
- 9 보관소 표 2 출생 수, 부도부현별(특별 구역 - 지정 도시 재등록) 및 출생 순서
- 10 예금 표 3 일본 어머니 출생 수, 현(특별 구역 - 지정 도시 재등록), 아버지의 국적 및 성별
- 11 보관소 표 4 일본 아버지 출생 수, 부도부현(특별 구역 - 지정 도시 재등록), 어머니의 국적 및 성별
- 12 아카이브 표 5 사망자 수, 부도부현(특별 구역 - 지정 도시 재등록), 연령(5세 그룹), 성별
- 13 아카이브 표 6 사망자 수, 사망 원인별 단순 분류, 성별, 연령(5학년, 초등학교 및 중학생)
- 14 저장 표 7 사망자 수, 부도부현별(특별 구역 - 지정 시 재등록), 사망 원인(사망 원인 간단 분류), 성별 및 연령(5세 그룹)
- 15 저장 표 8 감염병으로 인한 사망자 수, 부도부현(특별 구역 - 지정 시 재등록) 및 사망 원인(감염병 분류)
- 16 저장 표 9 결혼 건수(레이와 7호에 보고), 결혼 당시 부부 연령(5년 등급), 첫 결혼과 재혼의 합산
- 17 저장 표 10 이혼 건수(레이와 7년 별거 보고), 별거 당시 부부 연령(5세), 동거 기간별
- 18 인구, 월별, 성별 인구 (일본 인구)

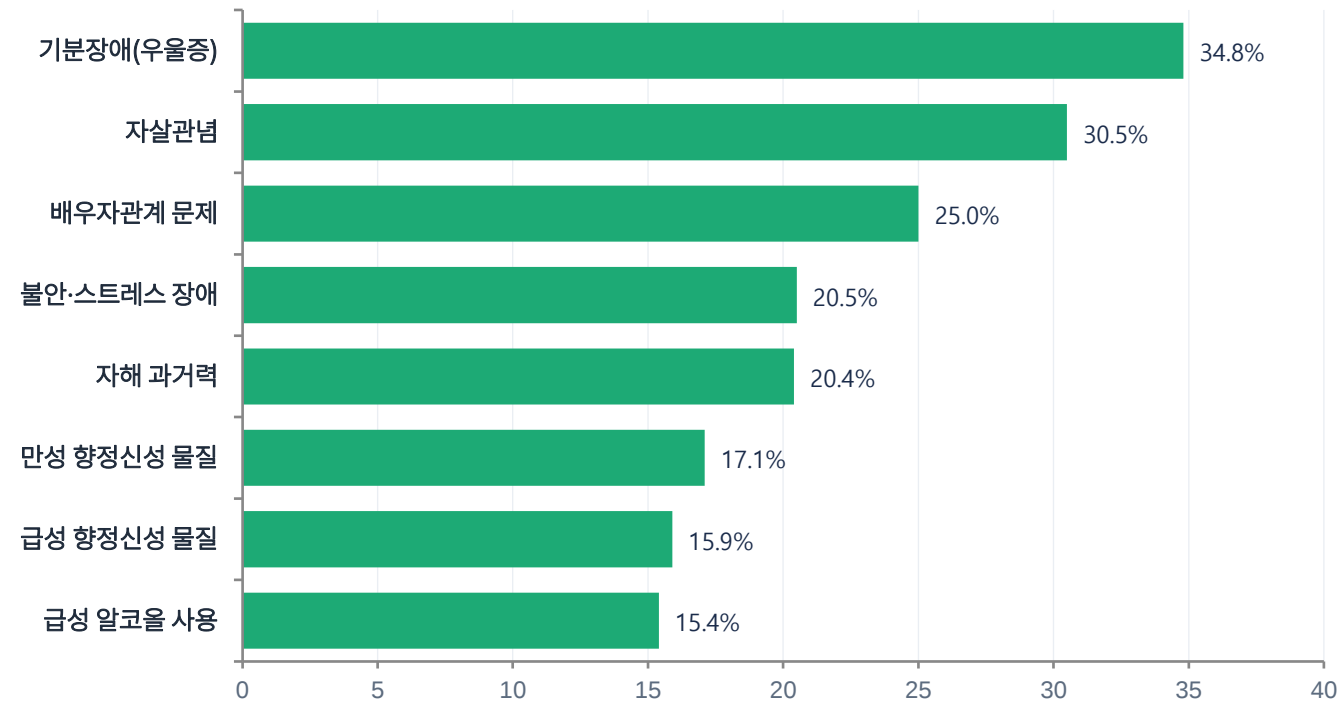
해외 사례 : 호주 자살원인별 사망분석

호주 통계청(ABS), Risk factors for intentional self-harm deaths (Suicide) in Australia, 2024

- 국립검시정보시스템(NCIS, National Coronial Information System)에 등록된 경찰·병리·독성·검시관 보고서 활용 작성
- 전체 자살자 3,307명 중 79.7% 자살 위험요인(동기) 확인, 1인당 평균 4개 요인

상위 위험요인 (2024, 전체)

단위: % (전체 자살 사망자 대비)



연령대별 핵심 위험요인

- 전 연령 : 기분 장애(우울증 등) 전 연령대 가장 흔한 위험 요인
- 85세+ : 질병 및 장애로 인한 활동 제한 큰 위험 요인
- 25~64세 : 고용, 부부 관계, 법적 문제와 관련된 어려움
- 25~44세 : 급성 알코올 중독을 위험 상대적으로 높음
- 45~64세 : 만성 알코올 남용 위험 상대적으로 높음

성별 위험요인

- 약물 남용은 여성보다 남성에게서 더 흔한 위험 요인 (25~44세 남성 만성/급성 약물 중독 위험 상대적으로 높음.)
- 배우자 관계 문제로 인한 위험 남성>여성
- 기분 장애와 불안 장애는 남성<여성

국내 사례 : 국내 감염병 사망 감시체계 및 경찰 자살원인 분류 한계

질병청 연구용역(국외 사례 고찰을 통한 국내 표본감시 감염병 사망신고 체계 개선방안 마련, 경찰청 통계연보

국내 감염병 사망 감시체계 한계

- 감염병 유병과 함께 사망지표는 보건학적 필수 지표
 - 감염병 확진자 수는 검사건수 영향, 중증도 고려 한계
 - 감염병 확진자 사망, 감염병으로 인한 사망 구분
 - 메르스, COVID19 때에도 초과사망 등 사망 모니터링 중요
 - 사망원인통계는 다음해 9월에 공표되어 정책시차 존재
- 법정 감염병 신고체계의 완전성 확보를 위한 노력 필요
 - 1~3급 감염병(전수), 4급 감염병(표본)
 - 환자 전원, 사후 검출, 원내 사망 등의 사유로 신고 누락 가능
 - 병원 의무기록 및 건강보험 청구 자료 공유 제한
- 사망 / 법정감염병 신고 / 건강보험 청구 자료의 분절

경찰 자살원인 분류 한계

- **(단일분류 원칙)** 개별 담당자가 사건에 대해 1개 원인만 코딩
 - 자살원인은 복합적으로 발생 가능
 - 호주의 경우, 개인별 평균 4개의 요인 존재
 - 예: 실직으로 인한 우울증 → 직업문제+정신문제+경제문제
- **(불완전성)** 분류체계의 불완전성으로 기타 비중 높음.
 - 학업스트레스, 군복무 등의 젊은 연령층 분류체계가 모호
 - 가정문제, 학대 및 폭력, 사별문제, 남녀문제는 상호 중첩
 - 다른 분류기준이 같은 레벨로 혼재
 - 가정문제, 남녀문제 → 관계 / 사별문제, 학대 및 폭력 → 생애사건
 - 경제생활문제, 직업직장문제 → 상황 / 육체적질병문제, 정신적 문제 → 건강상태
- **(변별력 문제)** Top-3 원인이 전체 자살의 80.2%

AI 적용모델

ML(기계학습) : 수많은 사례를 학습해 패턴을 찾는 방식
LLM(대형 언어 모델) : 문장 맥락을 이해해 답을 추론하는 방식



(모델1) 사망진단서 기재 사망원인 코드화

ML

사망진단서에 기재된 한글 사인 명(예: 폐렴)을
KCD 질병코드로 자동 변환하는 모델

입력

한글 사인명

출력

KCD 코드

행정효율 향상



(모델2) 행정자료 반영 잠정 원사인 예측

ML

사망진단서+건강보험 진료 기록을 결합,
AI가 잠정 원사인을 예측하는 모델

입력

사망진단서
+ 건강보험 기록

출력

원사인
KCD코드

맞춤형 정책지원



(모델3) 약물 중독 성분 분류

LLM

약물중독 사망자 약물 성분 추출
KCD분류 및 사망 우선순위 판단 모델

입력

서술형
행정자료

출력

약물
KCD 코드

행정효율 향상



(모델4) 자살 위험요인(원인) 분류

LLM

사망자 행정자료(서술문)에서 자살 동기 추출
자동 분류 → 자살원인 사망통계 서비스 모델

입력

서술형
행정자료

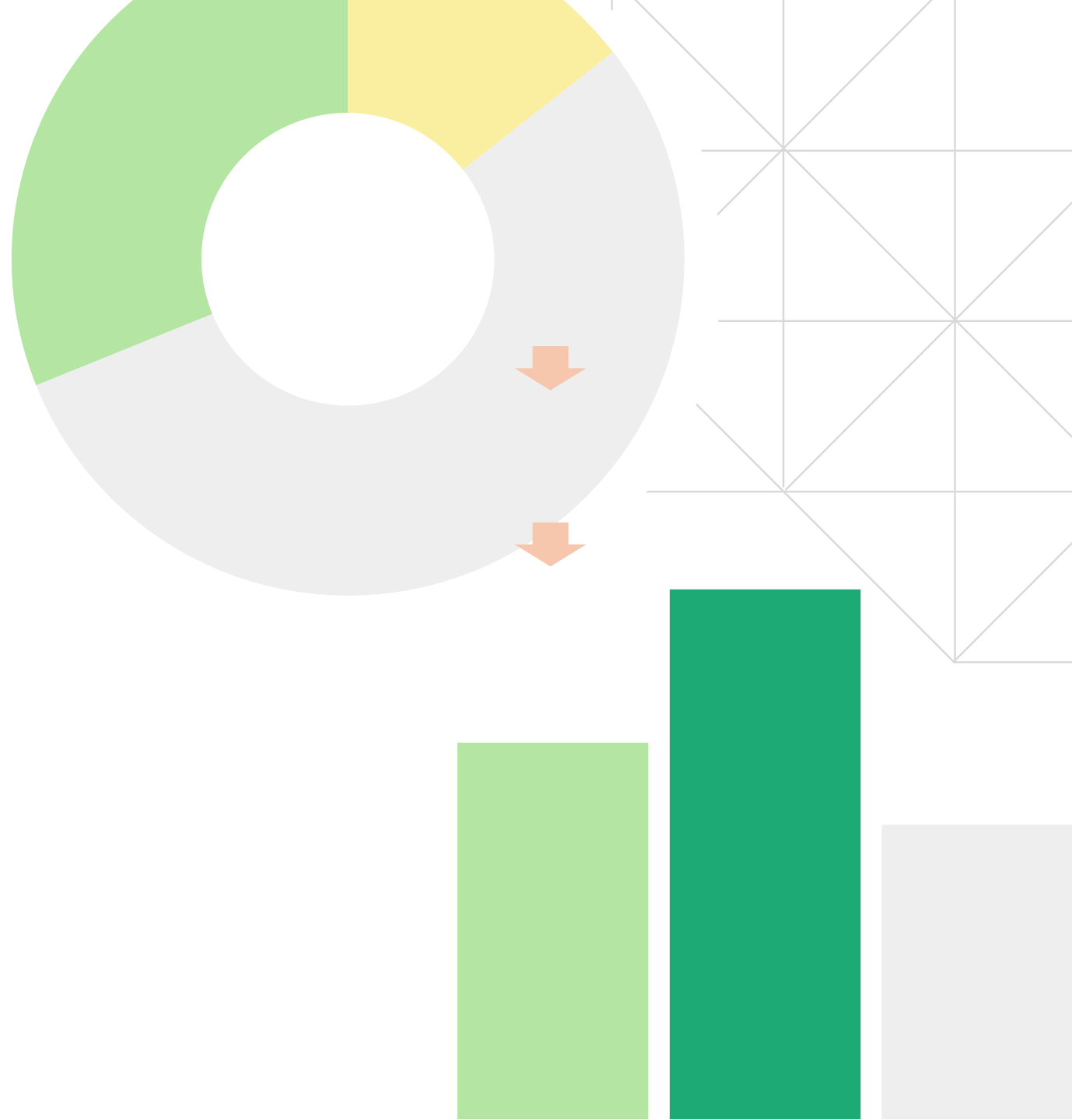
출력

자살원인
코드

맞춤형 정책지원

02

Method



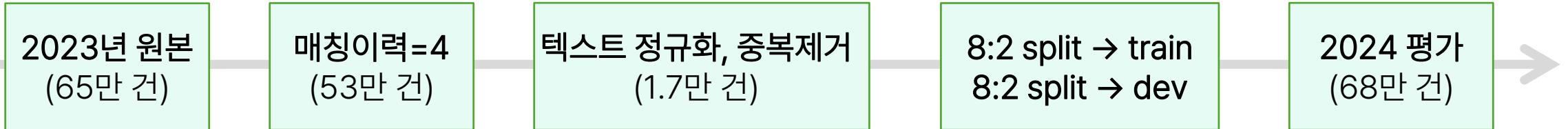
AI 모델설계 (모델1) 사망진단서 기재 사인 코드화(설계중)

2024년 사망진단서 기재 사인 68만 건 중 **폐렴** 5.8만 건(8.4%), **패혈증** 2.7만 건(4.0%), **미상** 1.7만 건(2.5%)

빈도 수가 1건 인 사망원인(희귀질환 or 자세한 표현 or 오타) 7.3만 건 → 압도적 다수 경향 존재 + 희귀 자료의 다수 분포

기존 한글사전의 정확도 하락 + 사인선정시스템 내 자동코딩 결과 오류 → 정답 없는 학습자료 & 정답 없는 검증자료

ICD-11은 한글사전 X, 기존 자동코딩 자료 X, 학습자료 X → ICD-11 원사인자동선정 불가



)	(가)	직접사인	폐색전증 의증, 기관지 폐색 의증, 뇌경색
	(나)	(가)의 원인	흡인성 폐렴, 뇌경색
	(다)	(나)의 원인	뇌경색
	(라)	(다)의 원인	당뇨, 심방세동, 고혈압
	(가)부터 (라)까지와 관계없는 그 밖의 신체상황		
	수술의사인		수술

- (언어모델) KM-BERT, KcELECTRA 비교 활용
- 롤업(Roll-up) 임베딩 적용

AI 모델설계

(모델2) 행정자료 반영 잠정 원사인 예측



인코딩

(언어모델) **KcELECTRA** vs KM-BERT

(임베딩) **계층형 임베딩*** vs 코드 단일 입력

* 대분류(I)+ 중분류(I21)+ 세부코드(I219)를 각각 별도 벡터로 분해 후 결합

모델조합

GCN 구조로 KCD 계층 관계 반영

모델구조	Accuracy	Macro F1	Weight F1
CNN	0.6623	0.0768	0.6350
Semantic	0.6300	0.0823	0.6146
GIN	0.6343	0.0711	0.6057
GATv2	0.6450	0.0781	0.6257
GCN	0.6547	0.0784	0.6247
HealthGAT	0.6465	0.0808	0.6279

Accuracy 65.5% · Top-3 포함률 82.7% · 3자리 일치율 76.8% · 104항목 일치율 83.2%

AI 모델설계 (모델2) 행정자료 반영 잠정 원사인 예측

사망진단서 TEXT

사망진단서 사인(직접사인, 가~다의 원인) **5개 항목**

(가)	직접사인	심폐정지
(나)	(가)의 원인	폐렴, 폐기능부전
(다)	(나)의 원인	
(라)	(다)의 원인	
(가)부터 (라)까지와 관계없는 그 밖의 신체상황		

잠정원사인 CODE

WHO지침에 따른
사망진단서 기반 원사인
1개 항목

잠정원사인
: J18.9(상세불명 폐렴)

건강보험 수진기록

사망직전 주상병, 진료비, 진료일수 등
20개 항목

자료구분	주상병	부상병	...
사망직전	A15.01	J69.0	...
사망직직전	J20.9	J45.9	...

A15.01

(배양 유무에 관계없이 가래 현미경 검사로 확인된 공동이 없거나 상세불명의 폐결핵)

AI 모델설계 (모델2) 행정자료 반영 잠정 원사인 예측



GCN 선정 사유 : 데이터 구조, 확장성, 통계 활용성 고려

① KCD 코드 체계와의 정합성

- KCD(ICD)의 계층 트리 구조 → 코드 간 위계 관계가 명확
- 평면 분류 방식(CNN)보다 계층·인접 관계를 그래프로 직접 학습

A 대분류
감염성 질환

A15 소분류
호흡기 결핵

A153 세분류
폐결핵

② 희소 코드 보완 가능성

- 사망원인 분포는 일부 코드에 집중된 long-tail 특성 — 희소 코드 다수 존재
- HIV·알츠하이머 등 희소 코드에서 GCN이 CNN 대비 F1 +10%p 이상 개선 확인
- 학습 데이터 부족 코드일수록 계층 정보로 보완 — GCN의 구조적 강점

<F1 score>

HIV : 45.1%(CNN) → 59.5%(GCN)

알츠하이머 : 51.6%(CNN) → 62.0%(GCN)

③ ICD-11 전환 대응력

- 향후 ICD-10 → ICD-11 전환 시 코드 체계 대규모 개편 예정
- CNN: 라벨 공간 변경 시 전면 재학습 필요 → 유지보수 부담 발생
- GCN: 그래프 구조 갱신만으로 신규 코드 적응 가능 → 장기 운영 유리

A15-19 중분류
결핵

A15(확진)

A16(임상진단)

1B1X(마이코박테리아)

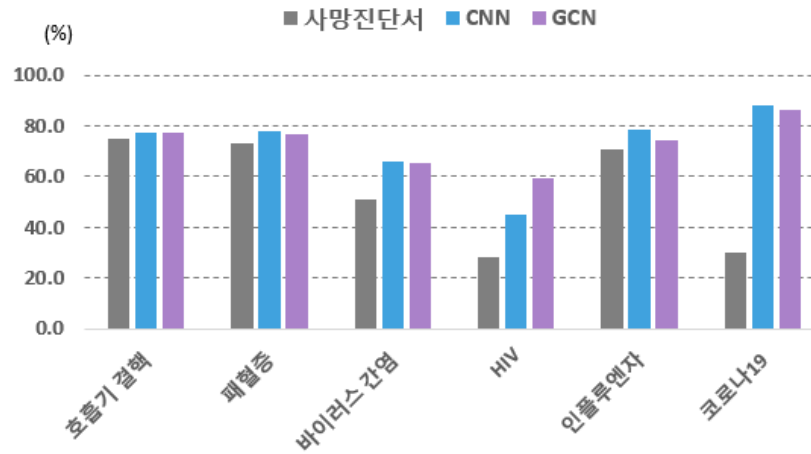
1B10(폐결핵)

AI 모델설계

(모델2) 행정자료 반영 잠정 원사인 예측

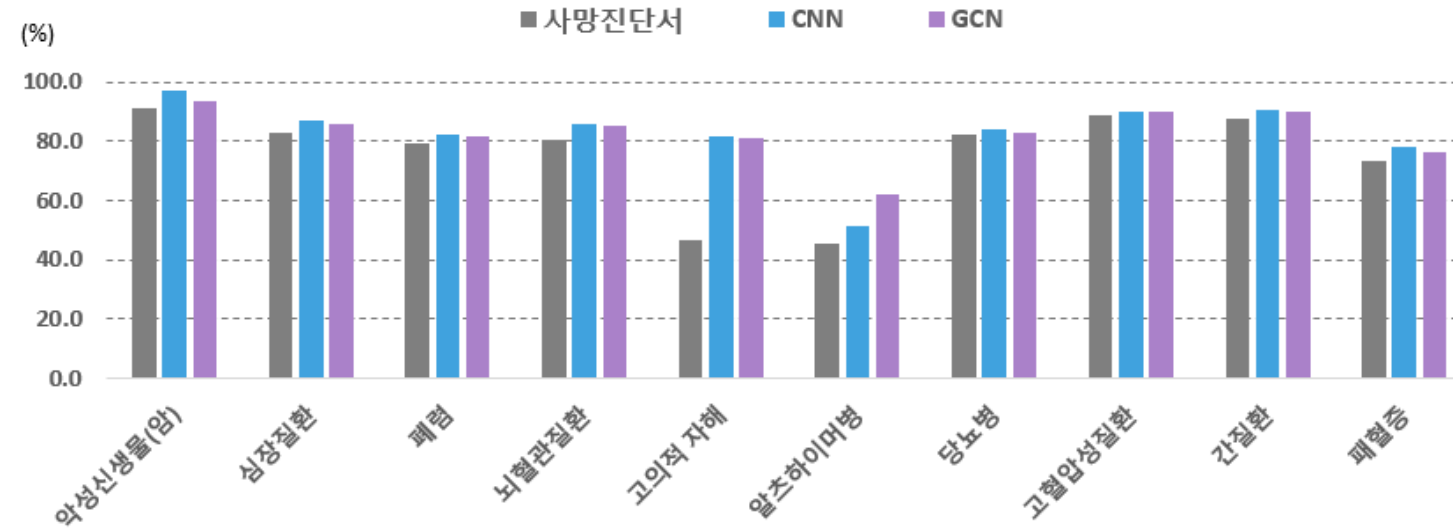
F1 Score

감염병



HIV, 알츠하이머와 같은 사망자 수가 작고,
학습데이터가 부족한 경우
GCN방식의 결과가 상대적으로 좋게 나타남.

10대 사인



AI 모델설계 (모델2) 행정자료 반영 잠정 원사인 예측 - 최적화 탐색

STRATEGY 1

In/Out 변수 조정

- ① Full code(1.8만 개)
- ② 104항목 제표기준
- ③ 16개 사인
- ④ 감염병 각각 예측

STRATEGY 2

AI 예측확률별 선별반영

정답과 오답의 분포에 따라
임계확률 식별 및 선별 반영

STRATEGY 3

특수조건 반영

감염병, 폐렴, 패혈증 등의
병태 간 조건을 특수조건으로 추가 조정

기준	일치율(%)	
	사망진단서 (기존 방식)	AI 예측
최하단 코드 (1.8만개)	49.9	65.5*
제표항목 (104개)	72.9	83.6

* Top-3 포함률은 82.7%



USE CASE 1

사인분류사 업무 참고용

입력 / 출력

1.8만개 full 코드

예측 예시

- 1위 J18.9 (26.3%)
- 2위 D47.1 (8.19%)
- 3위 A15.11 (5.73%)



USE CASE 2

잠정 사망원인통계 서비스

입력 / 출력

104개 제표코드

활용 예시

잠정 사망원인통계 공표
감염병 사망 감시체계 개발

AI 모델설계

(모델2) 행정자료 반영 잠정 원사인 예측 - 최적화 탐색

STRATEGY 1

In/Out 변수 조정

- ① Full code(1.8만 개)
- ② 104항목 제표기준
- ③ 16개 사인
- ④ 감염병 각각 예측

STRATEGY 2

AI 예측확률별 선별반영

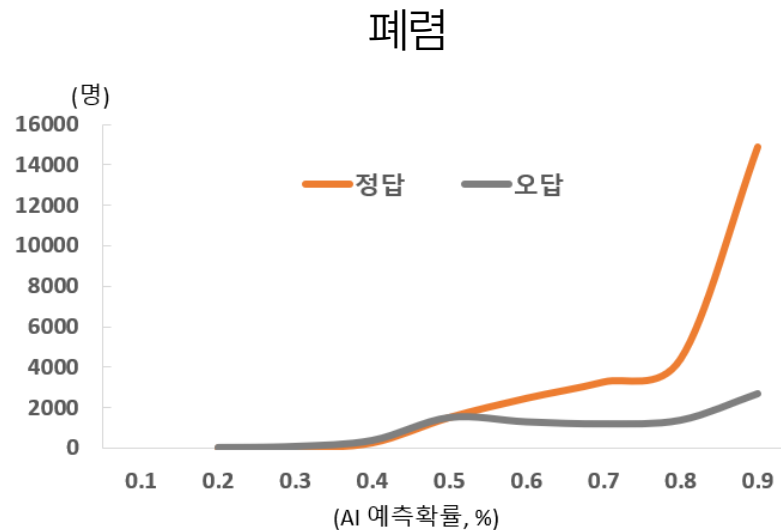
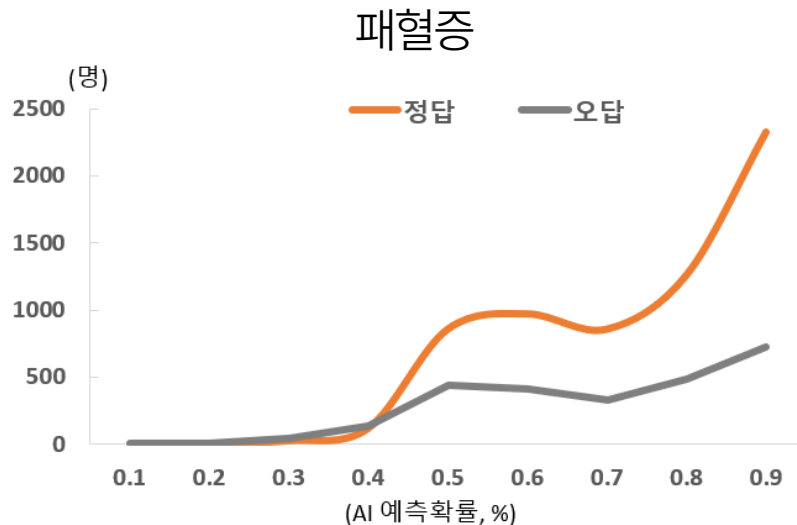
정답과 오답의 분포에 따라
임계확률 식별 및 선별 반영

STRATEGY 3

특수조건 반영

감염병, 폐렴, 패혈증 등의
병태 간 조건을 특수조건으로 추가 조정

<AI 예측확률 별 사망자 일치 정도, 2024년>



<병태 특수조건(예시)>

- 폐렴, 패혈증 등은 사망의 증상으로 고려 사망자의 사망진단서에 기본적으로 기재되는 사인임.
- WHO, ICD-10 vol2 지침서에 따르면 일부 감염성 질환은 다른 질환에 우선

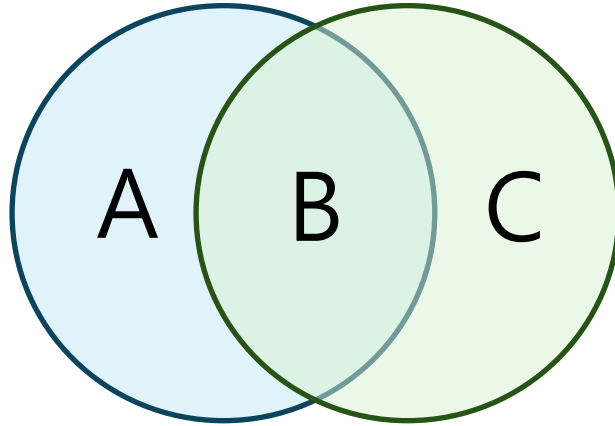
AI 모델설계 (모델2) 행정자료 반영 잠정 원사인 예측

- 2023년 선행연구에서는 의료 특수성을 고려하여 KM-Bert 언어모델 활용하였으나,
- 본 연구에는 사망진단서 기재 사인의 특성을 고려하여 KcELECTRA를 활용함.
- KM-BERT가 질병명 간 의미적 유사성, 의료적 인과 흐름, 전문용어 표현 해석에 강점
- 사망진단서의 정식 의료용어는 인코딩 난이도가 상대적으로 쉬움. 문제는 자연어기반의 증상 또는 오타

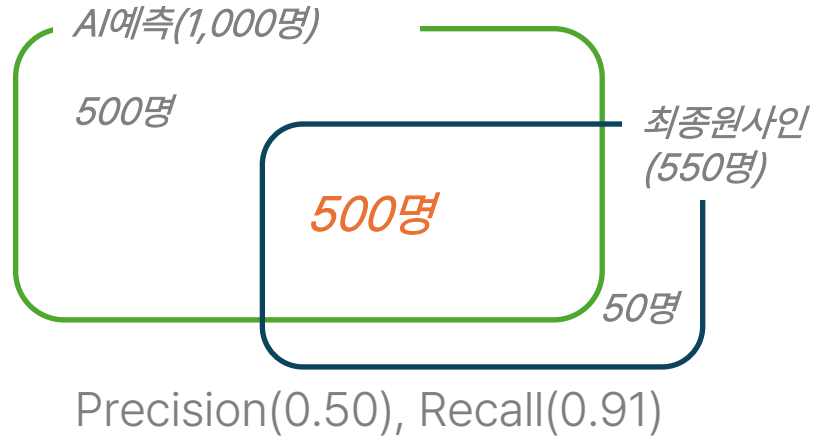
비교항목	KM-BERT	KcELECTRA
모델 설명	한국어 의료 텍스트 이해를 위해 의학 교과서, 논문 등으로 사전 학습된 의료 도메인 특화 BERT 모델	한국어 뉴스 댓글, 대댓글 같은 user-generated noisy text에 특화된 ELECTRA 기반 사전학습 언어모델
학습 데이터	의학 교과서 518개 파일, 건강정보 뉴스 약 72,844건, 의학 연구논문 15,698건(규모: 약 600만 문장, 1억 1,600만 단어)	네이버 뉴스 댓글, 대댓글 등 (규모: 약 17.3GB, 1억 8천만 문장 이상 규모)
핵심 강점	- 전문 의학 용어의 정확한 이해 - 의료 문맥 해석 정확도 우수	- 비정형/구어체(단어나열축약표현) 문장의 유연한 처리 - 학습 효율 높고 fine-tuning 안정적

AI 모델검증

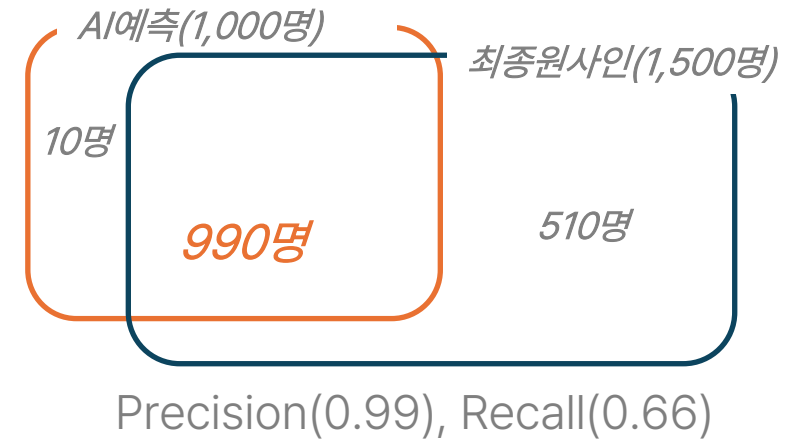
AI 예측(A+B) 참값(B+C)



<패혈증 사망자 수(예시)>



<호흡기 결핵 사망자 수(예시)>

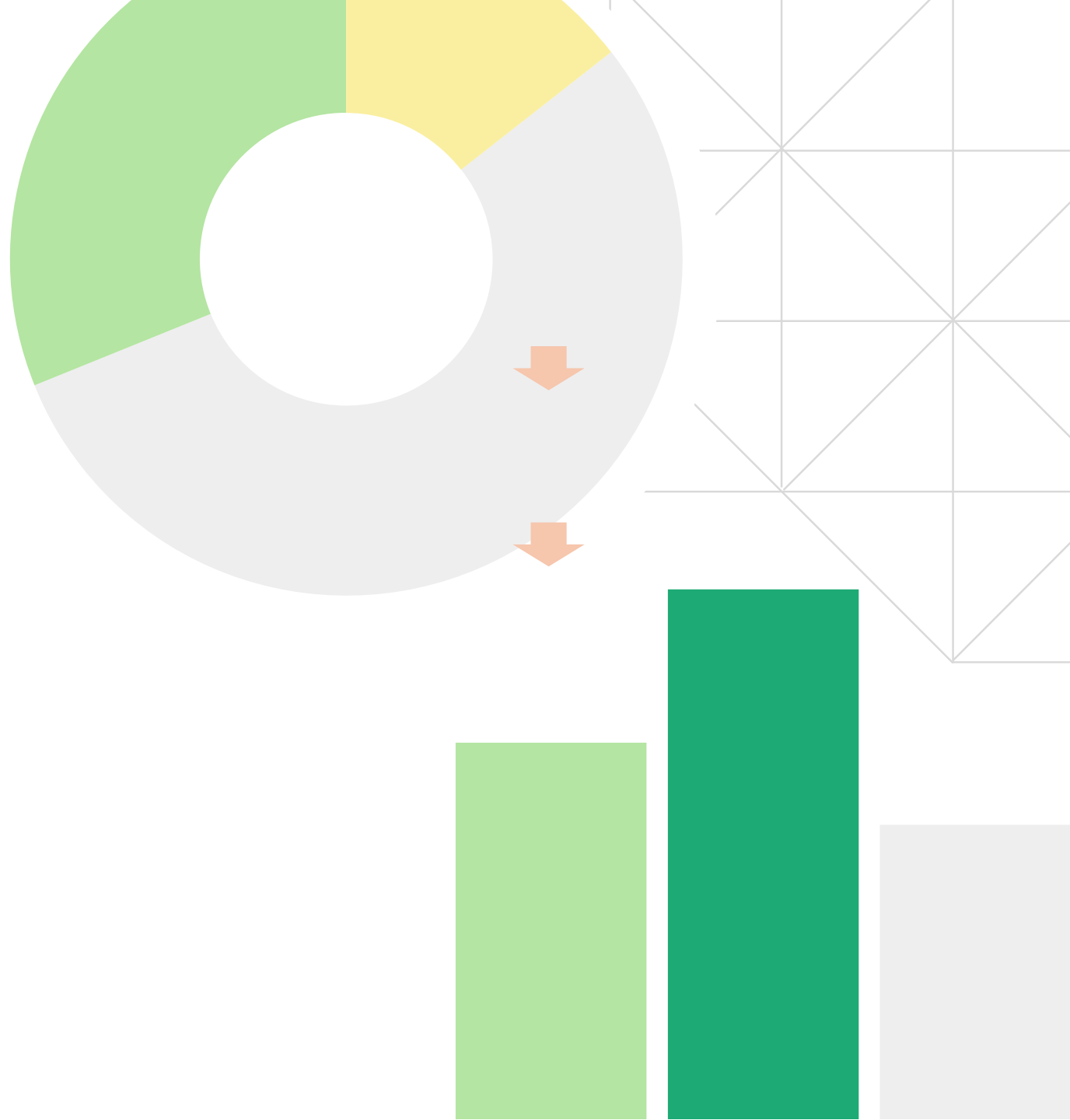


검증지표	내용
정밀도(Precision)	AI모델이 정답으로 예측한 값 중 정답 비율, $\frac{B}{(A+B)}$
재현율(Recall)	실제 참값 중 AI모델이 정답으로 맞춘 비율, $\frac{B}{(B+C)}$
F1 Score	정밀도와 재현율의 조화평균, $\frac{2 \times (Precision \times Recall)}{(Precision + Recall)}$
사망자 수 집계값	실제 잠정 사망자 수 서비스 지표, $\frac{(A+B)}{(B+C)}$

- 정밀도 및 재현율이 일정 수준 이상이면서 비슷한 수준이 적절함.
 - (정밀도大+재현율小) AI 예측한 것은 정확하지만 예측량이 적어 활용도 ↓ (대표적 사인: HIV)
 - (정밀도小+재현율大) AI가 과다예측하여 정답은 많지만 정확도가 ↓ (대표적 사인: 폐렴)
- 주 검증지표 : F1 score, 사망자 수 비교
- 보조 검토 : 정밀도와 재현율 수준

03

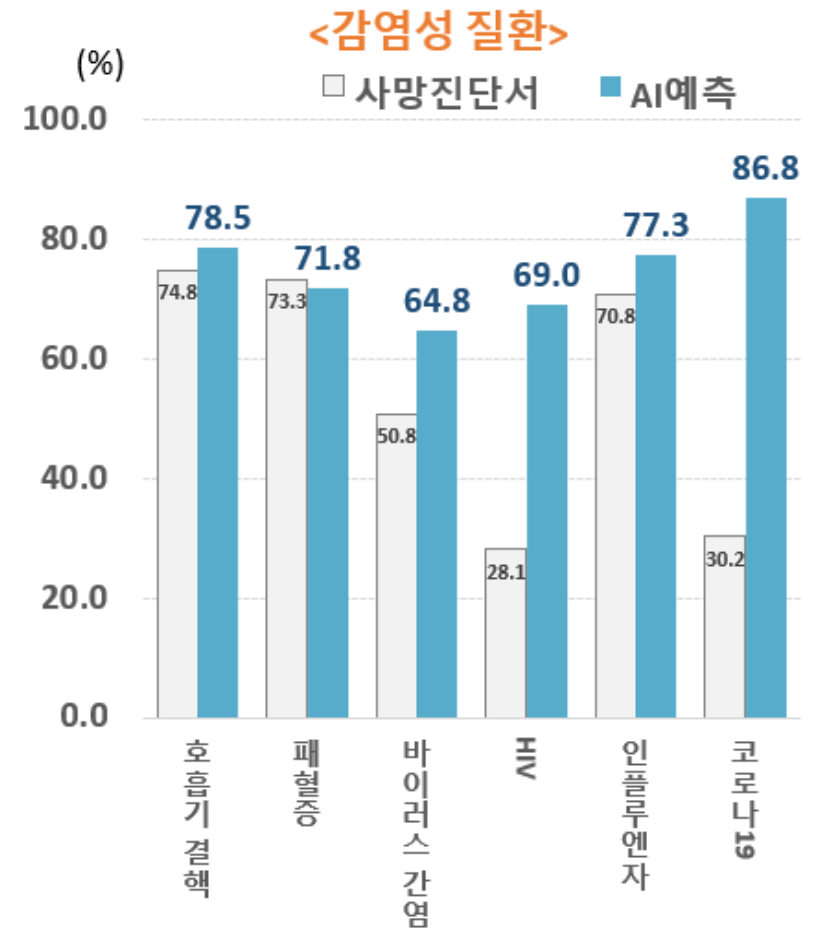
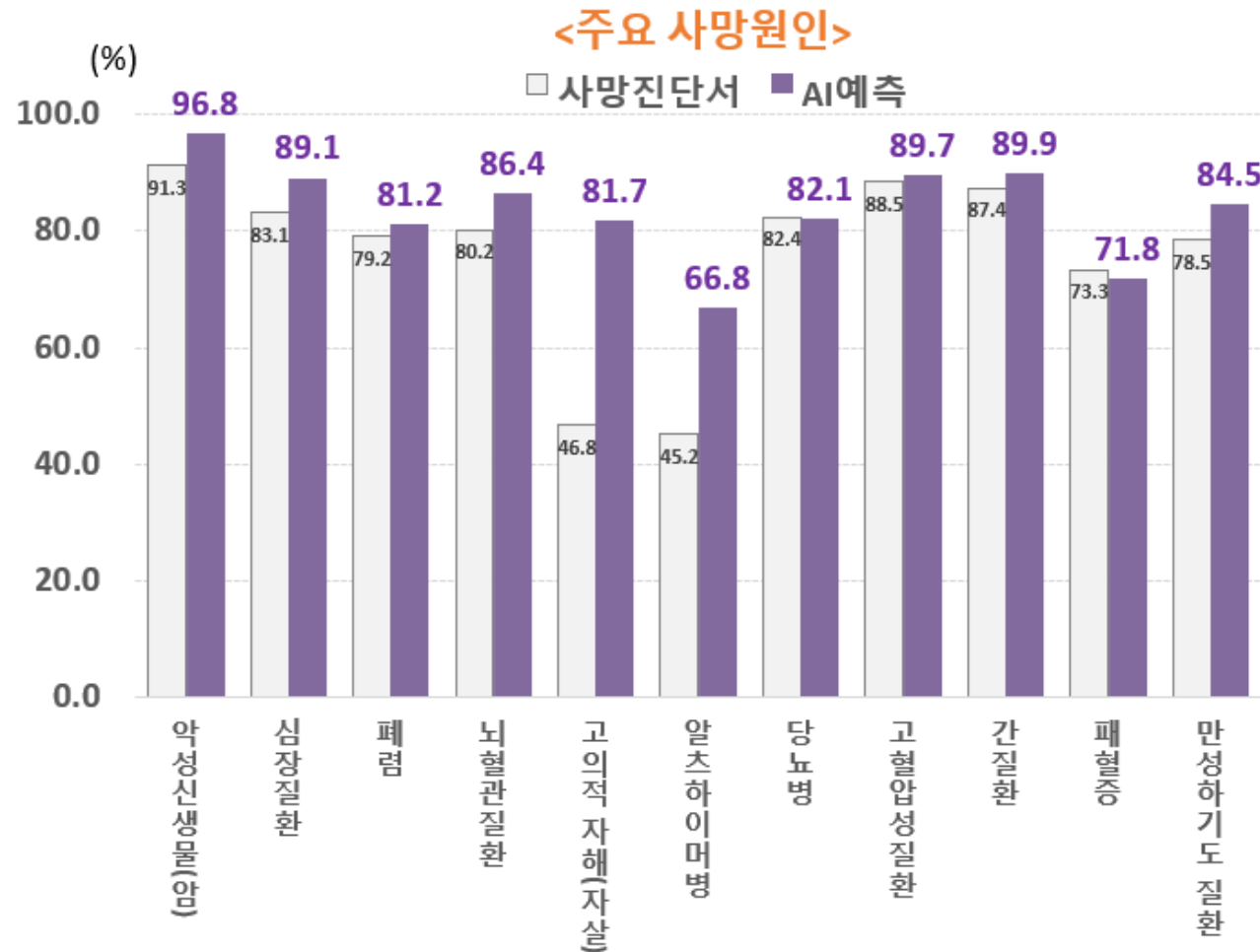
Result



AI 모델시산

(모델2) 행정자료 반영 잠정 원사인 예측

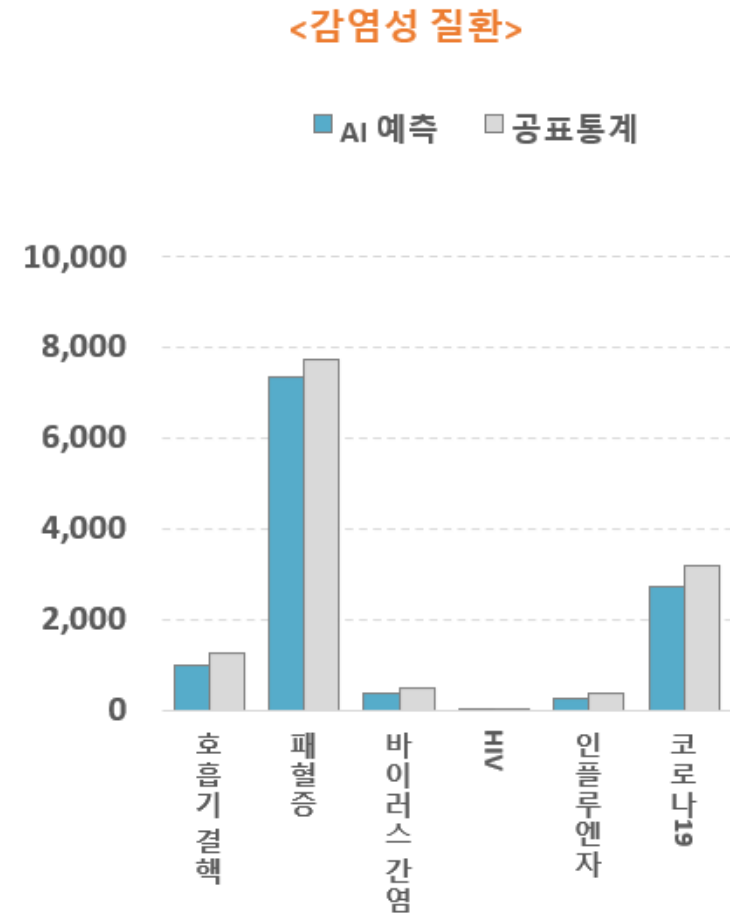
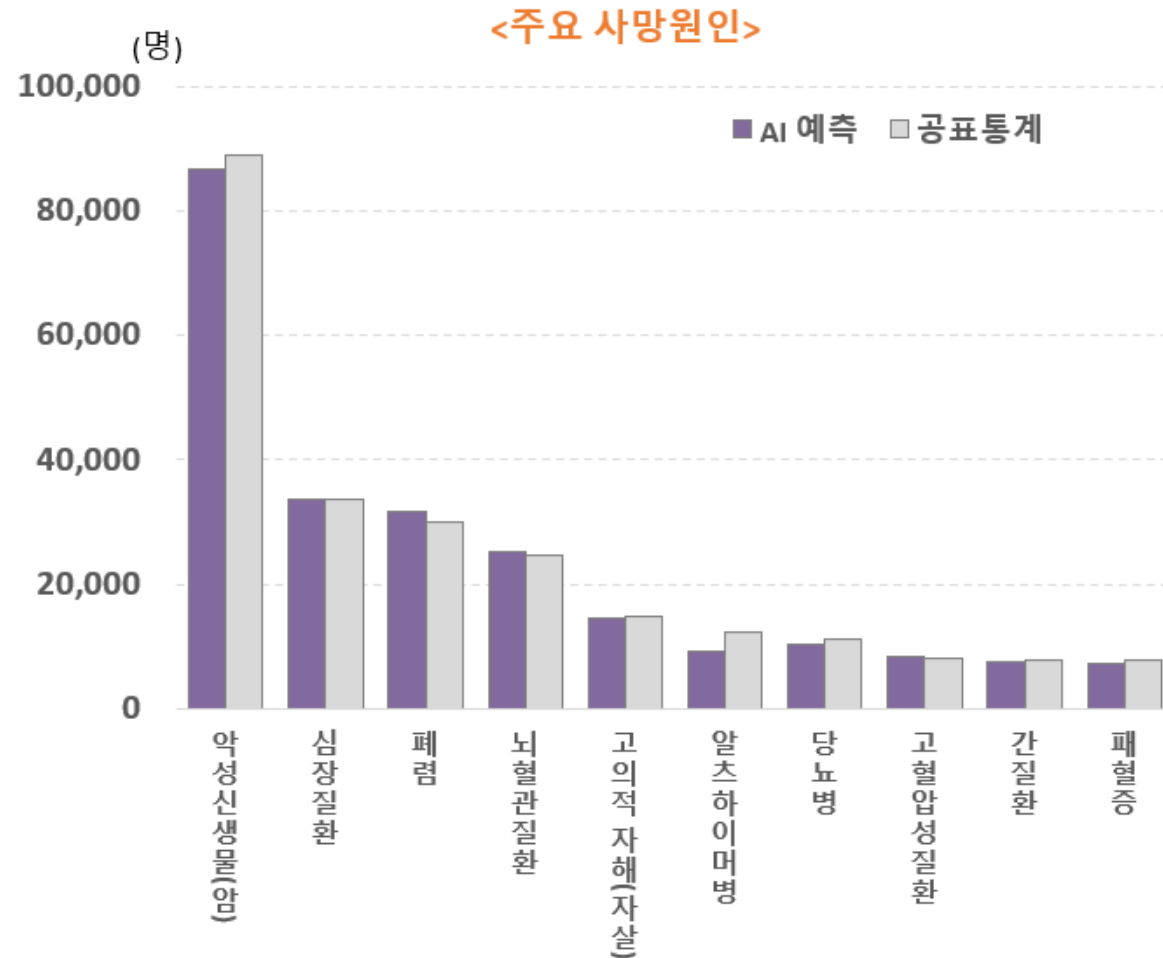
<사망진단서 및 AI 예측 원사인 F1 정확도, 2024년>



AI 모델시산

(모델2) 행정자료 반영 잠정 원사인 예측

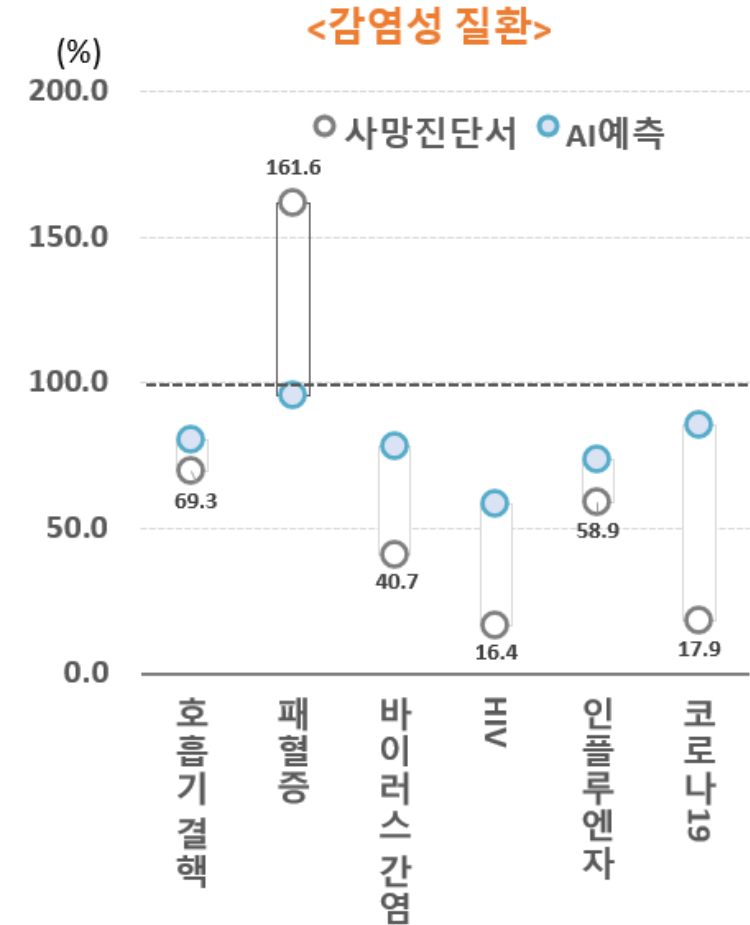
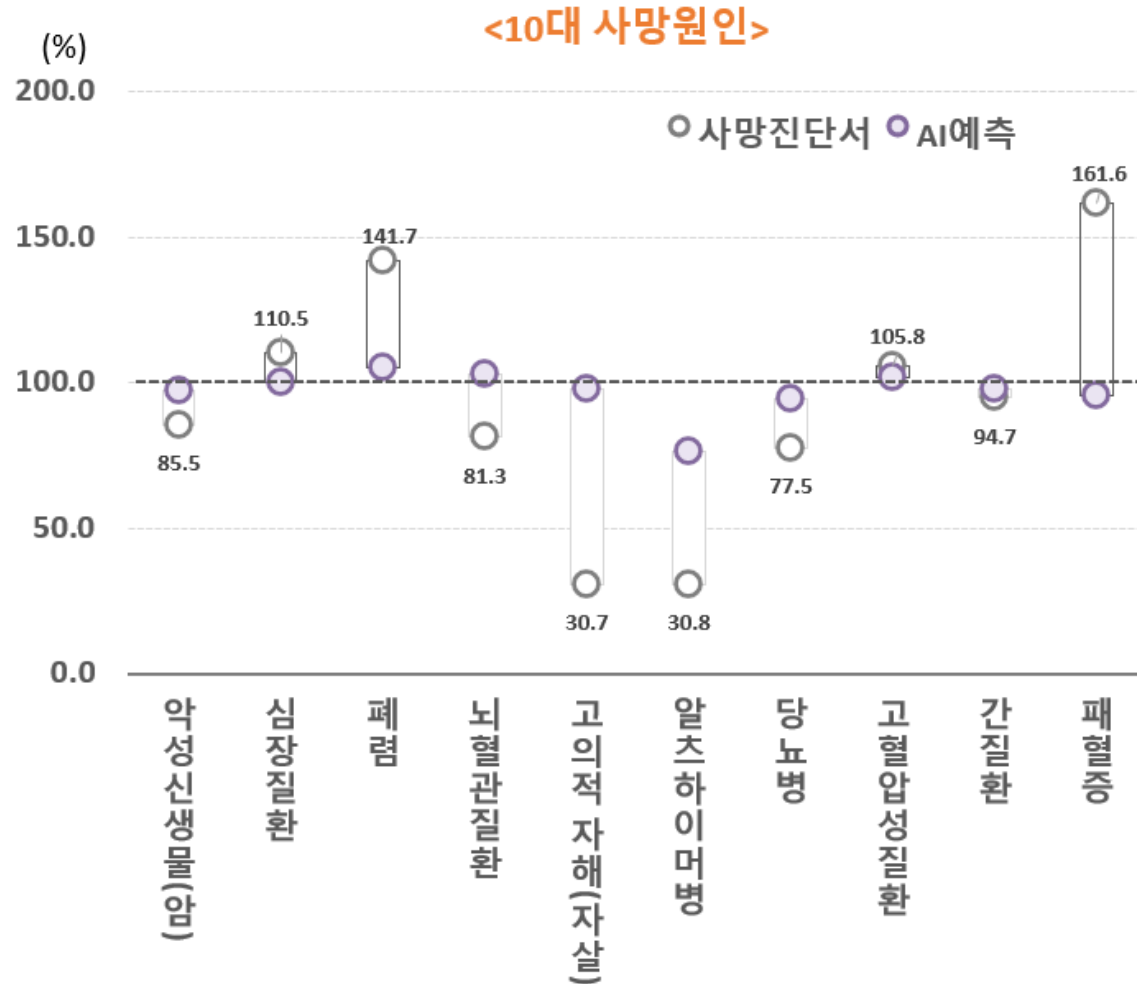
<AI 예측 사망자 수 비교, 2024년>



AI 모델시산

(모델2) 행정자료 반영 잠정 원사인 예측

<공표된 사망자 수 대비 일치비율, 2024년>



AI 모델시산

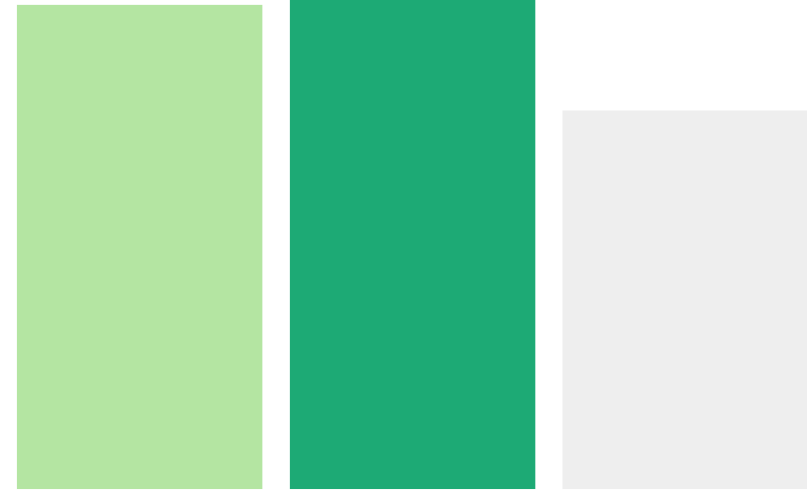
(모델2) 행정자료 반영 잠정 원사인 예측

<감염성 질환 사망률 예측 월별 사망자 수 - 특수조건 적용 이전>



04

Discussion



후속연구 추진방안

- ICD-11 도입일정 및 시스템 개발 사업 등을 고려하여 2025년 사망자 DB 구축 시 바로 후속 연구 실시



학습자료 추가(2개년)를 통한 성능 개선 및 안정성 확보

- AI모델의 안정성, 일관성 확인을 위하여 추가 데이터셋 확보 및 학습을 통한 검증 및 시스템 구현
- 건강보험 수진기록 데이터 전처리를 통한 성능개선(불필요 항목 제거, 산정특례 추가)

'26년 연구범위	2023년 사망자 ^{학습} → 2024년 사망자 ^{검증}
	2024년 사망자 ^{학습} → 2025년 사망자 ^{검증}
후속연구 범위	2023년+2024년 사망자 ^{학습} → 2025년 사망자 ^{검증}
	2023년+2025년 사망자 ^{학습} → 2024년 사망자 ^{검증}



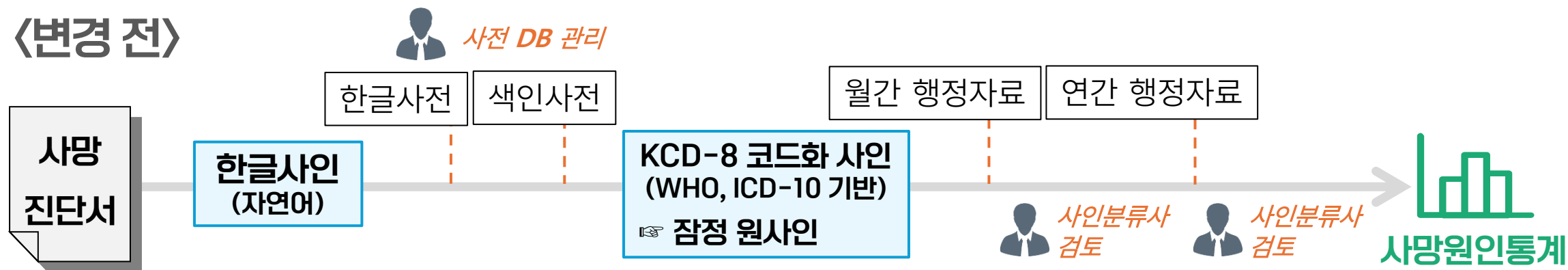
감염병 특수조건 추가를 통한 정확도 개선

- 정확도 향상을 위한 감염병 특수조건 및 파인튜닝 등 개선
- 사망진단서 키워드 포함 및 제외
- AI 예측확률별 선별 반영
- 감염병 인과관계에 따른 조건 추가

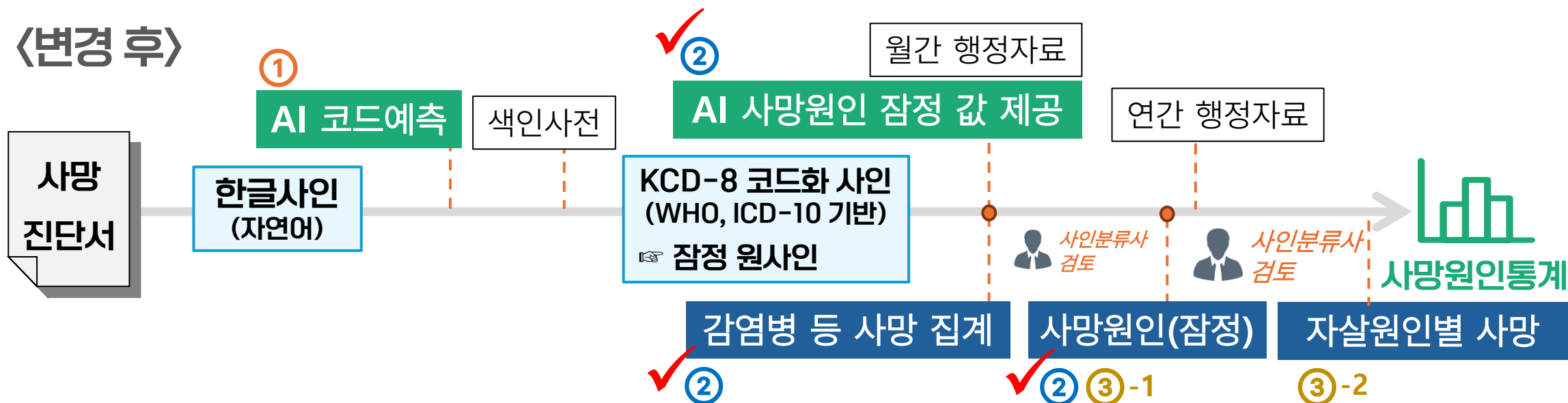


원사인 분류 프로세스 : 서비스 변경 전후 업무흐름도

〈변경 전〉



〈변경 후〉





감사합니다.