DSRI
Data
Research
Brief

vol. 2026-3

ISSN 2765-1002(Online)



사망원인통계 원사인 분류 AI 적용 연구

이석민
전고설
최지원

국가데이터연구원
데이터과학연구팀

● **(연구목적)** 사망원인통계 원사인 분류업무의 정확도와 행정 효율성을 높이고, 감염병 등 특정 사망원인의 잠정 원사인 예측과 자살 위험요인별 사망자 통계 등 보건정책 수요에 대응할 수 있는 AI 기반 서비스 개발 기반을 마련하고자 하였다.

● **(연구내용)** 사망원인통계 원사인 분류에 대한 AI 적용 가능성을 사인 코드화(모델1), 잠정 원사인 예측(모델2), 약물중독 코드 자동분류(모델3), 자살 위험요인 자동분류(모델4) 등 네 가지 모델 개발을 통해 실증적으로 검증하였다.

- 그리고 이를 통해 사인분류 업무체계의 개편방향과 사망원인통계 서비스의 확대 방안을 제시하였다.

● **(향후활용)** 잠정 사망원인통계를 신속하게 제공하고 자살 위험요인별 사망자 통계 서비스 기반을 마련함으로써, 감염병 감시체계 구축과 자살예방 정책 등 시의성 있는 보건정책 수요에 대응할 수 있다.

연구배경 및 목적

❍ **(통계작성환경 변화)** 사망원인통계는 「통계법」제17조 및 제18조에 따른 국가승인 통계로, 보건의료정책, 감염병 대응, 자살예방 등 다양한 정책분야의 기초자료로 활용됨

- 그러나 고령화에 따른 사망자 수 증가, 세계보건기구(WHO)의 제11차 국제질병사인분류(ICD-11) 도입, 감염병과 같은 특정 사인에 대한 시의성 있는 통계 수요 증가로 통계 작성 환경이 구조적 변화에 직면함

❍ **(업무환경 변화)** 사망진단서 사인 표현의 복잡·다양화로 원사인 분류 업무의 난이도와 담당 인력의 업무부담이 지속적으로 증가

❍ **(정책현장 요구)** 관계부처 합동 「10대 청소년 자살예방 범정부 추진대책」(2026.5) 수립, 국무조정실·질병관리청의 자살통계·감염병 감시체계 개선 수요 등 정책현장의 요구가 커지고 있음

❍ **(연구목적)** 원사인 분류 업무의 정확도와 행정 효율성을 높이고, 보건정책 수요에 대응할 수 있는 AI 기반 통계서비스의 기반을 마련하기 위해 수행됨

☞ **(연구질문) 사망진단서에 행정자료를 결합하여 AI 모델을 활용하면 사망원인 예측이 얼마나 빠르고 정확해지는가?**

데이터 및 방법

❍ **(연구개요)** ① 사망진단서 사인 정보의 자동 코드화, ② 행정자료를 반영한 잠정 원사인 예측, ③ 약물중독 사망의 약물 성분 추출·코드화, ④ 서술형 행정자료의 자살 위험요인 분류 등 네 가지 모델을 설정, 각각에 AI 적용 가능성을 단계적으로 검증하는 실증연구로 수행

❍ **(기초자료)** 사망진단서, 건강보험 수진기록(월간), 국립과학수사연구원 부검자료, 경찰청·해양경찰청 변사자료 등을 활용

[그림1] 사망원인통계 예측분석 시스템 개념도



❍ **(연구방법)** 학습형 분류모델과 LLM 기반 분류모델 두 방법으로 모델을 구축함

- **(모델1·모델2)** 학습형 분류모델로서 한국어 자연어 처리에 특화된 KcELECTRA 인코더를 기반으로 구축하였음

  · 모델1은 KCD코드에 대한 멀티라벨 분류구조를 적용하였으며, 모델2는 KCD 계층 구조를 반영한 그래프 합성곱 신경망(GCN)을 결합하여 잠정 원사인 코드와 건강보험 수진기록을 함께 학습하였음

- **(모델3·모델4)** 서술형 텍스트를 다루는 LLM 기반 분류모델로서, 시스템 프롬프트와 few-shot 예시를 결합하는 방식으로 설계함

  · 모델3은 부검자료 서술문에서 약물성분·우선순위를 추출하도록 LLM과 규칙엔진을 결합하였으며, 모델4는 근거추출과 종합판단을 분리한 2단계 파이프라인과 규칙엔진을 적용하여 분류체계 이탈과 라벨링 중복을 방지하였음

<표1> AI 적용 모델 개요

모델	목적	입력 데이터	핵심 방법
1	사망진단서 기재 사인 코드화	사망진단서 사인명 (한글 자연어)	KcELECTRA 기반 텍스트 분류
2	잠정 원사인 예측	사망진단서 + 건강보험 수진기록	그래프 합성곱신경망(GCN)
3	약물중독 성분 분류	국과수 부검자료 서술문	LLM + 규칙엔진
4	자살 위험요인 분류	경찰·해경·국과수 서술형 행정자료	LLM 2단계 파이프라인 (근거추출+종합판단)

주요 결과

❍ **(모델1) 사망진단서 사인 코드화를 통한 방법론적 기반은 마련하였고, 학습데이터의 품질에 따라 모델 성능에서 차이가 있음을 확인**

- 한글사전 매칭 영역에서는 높은 일치율(98.9%)을, 반면 색인사전 매칭 영역에서는 상대적으로 낮은 일치율(62.1%)을 보임
- 전반적인 AI 모델의 분류정확도는 90.9%에 달하여 실무적용 가능성을 확보
- 모델 오분류의 주요 원인은 학습데이터(사전 자동분류 코드)의 정확도와 편향으로 확인하였음(학습데이터의 정확도는 87.2%로 추정)

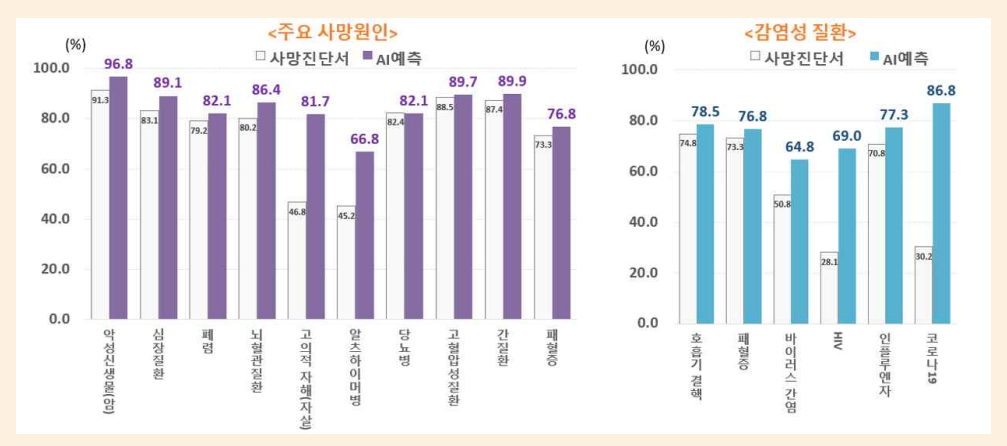
❍ **(모델2) 행정자료를 반영하여 잠정 원사인을 예측한 결과 기존 사망진단서 기반 결과보다 높은 성능을 보여, 공표시점 단축과 정확도 제고가능성을 모두 검증**

- 기존 사망진단서 기반의 시스템 75.8%에서 AI 예측결과 84.6%로 8.8%p 성능개선
- 질환별로 정확도가 상이하나, 악성신생물, 고혈압성질환, 간질환 등에서는 높은 성능을 보였고, 상대적으로 바이러스간염, HIV에서는 낮은 성능을 보임
- 월간 입수되는 행정자료 확보만으로 향상된 정확도와 시의성을 동시에 확보할 수 있는 기반을 마련

<표2> 잠정 원사인 예측 정확도 비교

구분	기존시스템	AI 모델	개선폭
예측 정확도(%)	75.8	84.6	+8.8%p

[그림2] 사망진단서 및 AI 예측 원사인 정확도(2024년)



❍ **(모델3) 약물중독 성분 자동분류 결과, 국립과학수사연구원 부검 연계사망자 약 7천명을 대상으로 약물검출 사망자 347명, 검출약물건수 529건을 도출**

- 다만 약물사전의 유사계열이 누락되거나 정규화 실패 등과 같은 오류가 반복 확인되어 약물사전 정비와 자동 재검증 절차 구축이 필요

❍ **(모델4) 자살 위험요인 자동분류를 통해 2024년 자살사망자 약 14,800명을 대상으로 약 17,000건의 위험요인을 분류하였음**

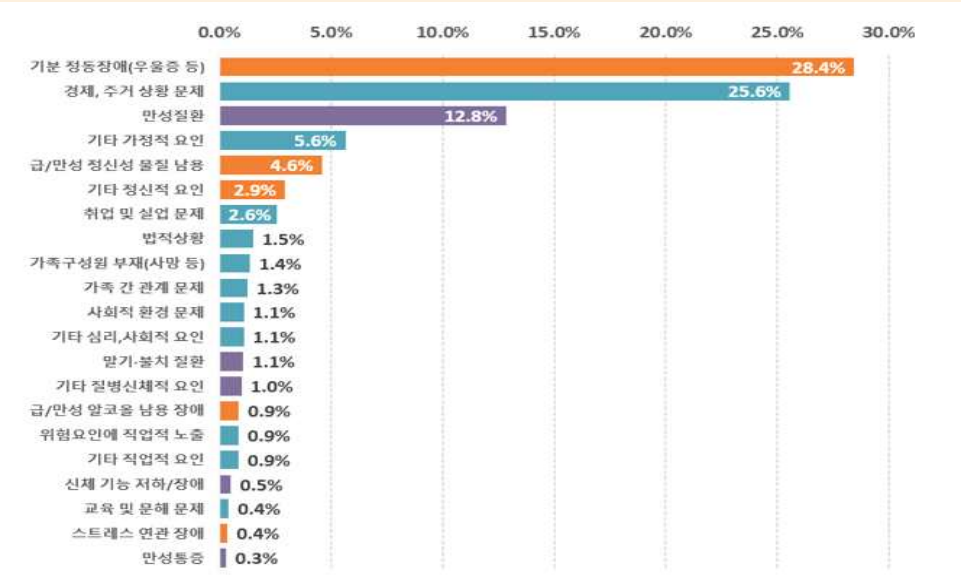
- 1인당 평균 1.16건의 자살 위험요인이 검출되었으며, 주로 기분 정동장애(28.4%)와 경제적 요인(25.6%)에 집중되어 있었음
- 그리고 기존 통계에서는 포착하기 어려웠던 세부 위험요인 및 위험요인 조합을 추가로 식별함

주요 결과 (계속)

<표3> 2024년 자살 위험요인 대분류별 분포

대분류	빈도(건)	비율(%)	주요 소분류
심리・사회적 요인	7,306	42.2	경제적 요인(25.6%), 가정적 요인(8.3%)
정신적 요인	6,440	37.2	기분 정동장애(28.4%), 물질남용(4.6%)
질병・신체적 요인	2,720	15.7	만성질환(12.8%)
미상	852	4.9	-

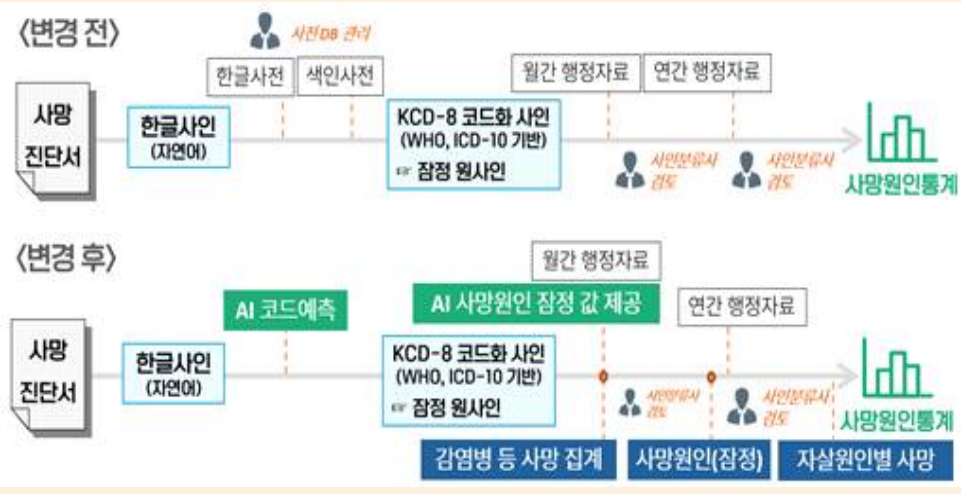
[그림3] 모델 예측결과 자살 위험요인 구성비(2024년)



❍ **(종합) AI는 원사인 분류를 자동 확정하는 도구가 아니라, 사인분류사의 검토를 지원하는 업무지원형 모델이면서, 예측값을 활용하여 통계를 추가 생산하는 서비스 확장형 모델임**

- AI가 한글 사인을 코드로 예측하여 사인매칭 과정을 대체하고, AI가 예측한 잠정 원사인 값을 제공함으로써 감염병 등 잠정 사망원인통계를 신속하게 산출

[그림4] 원사인분류 프로세스 : 서비스 변경 전·후 업무흐름도



시사점

❍ **(통계적 시의성 제고)** AI 기반 잠정통계 서비스의 도입은 사망원인통계를 사후적 기록 통계에서 정책 대응을 지원하는 시의성 있는 정보체계로 전환하는 계기가 될 수 있음
- 기존 연간 확정통계(다음 해 9월 공표)에 앞서 사망 후 5개월 시점 잠정 사망원인 통계를 제공하고, 익년 12월에 자살 위험요인 통계 서비스를 단계적으로 제공 가능

❍ **(업무체계 개편)** 한글사전·색인사전 기반의 매칭 방식을 AI 기반 코드화로 전환하되, 최종 판단 권한은 사인분류사에게 유지하는 인간-AI 협업형 업무체계로 개편 가능

❍ **(정책지원 기능 강화)** 감염병 사망 감시체계 구축과 다중 위험요인 기반 자살 위험요인 분석을 통해 데이터 기반 정책지원 기능을 강화하고, 보건정책의 대응력과 정밀성을 높이는 데 기여 가능

한계 및 향후 연구

❍ **(자료한계)** 본 연구는 자료 확보의 한계로 학습데이터와 평가데이터 각각 1개년치만 활용
- 학습데이터 추가에 따른 모델성능 재검증이 후속과제로 남아있음
- AI 기반 통계의 성능은 학습데이터 품질에 크게 좌우됨을 확인하였음

❍ **(학습데이터 품질검사 및 정제 효과 분석)** 수작업 검수로 모델 적합도를 재평가하고, 정제 전·후 성능변화를 정량 분석하여 모델성능 제고를 위한 전략 마련

❍ **(자살 위험요인 심층분석)** 2024년과 2025년 자료를 대상으로 다중 위험요인의 상호작용 분석 및 정책 활용에 적합한 서비스 형태를 모색하고 연말 기획보도 예정(12월)

❍ **(모델별 고도화)** 학습데이터 확충·정제, 행정자료 추가 연계, 실제 실무에 직접 적용, 환각 감소를 위한 프롬프트·후처리 규칙 개선

❍ **(ICD-11 적용을 위한 학습데이터 구축)** 생성형 AI를 활용한 학습자료 구축과 전문가 검수를 병행하여 신뢰성을 확보하고 ICD-11 도입준비

참고 문헌

○ Shin H.Y., et al.(2018), 'How to write a death certificate: from a statistical point of view', Journal of the Korean Medical Association/Taehan Uisa Hyophoe Chi 61.4
○ Duarte, F., Martins, B., Pinto, C. S., & Silva, M. J.(2018), 'Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text', Journal of Biomedical Informatics, 80, 64-77
○ Kim, C., You, S. C., Reps, J. M., Cheong, J. Y., & Park, R. W.(2021), 'Machine-learning model to predict the cause of death using a stacking ensemble method for observational data', Journal of the American Medical Informatics Association, 28(6), 1098-1107
○ Ward, P. J., et al.(2023), 'Deep Neural Networks for Fine-Grained Surveillance of Overdose Mortality', American Journal of Epidemiology, 192(2), 257-266
○ Goodman-Meza, D. et al.(2022), 'Development and Validation of Machine Models Using Natural Language Processing to Classify Substances Involved in Overdose Deaths', JAMA network open, 5(8), e2225593
○ Wang, S., et al.(2023), 'An NLP approach to identify SDoH-related circumstance and suicide crisis from death investigation narratives', Journal of the American Medical Informatics Association, 30(8), 1408–1417
○ 국가데이터처(2023), '인공지능 통계분류 자동화 확대 적용 연구', 통계개발원(현 국가데이터연구원)
○ 질병관리청(2025), '국외 사례 고찰을 통한 국내 표본감시 감염병 사망신고 체계 개선 방안 마련', 질병관리청

⇨ **(연구보고서 원본)** 국가데이터연구원홈페이지 > 연구간행물 > 연구보고서 > **사망원인통계 원사인 분류 AI 적용 연구**

출처를 밝히지 않고 국가데이터연구원 데이터 리서치 브리프의 내용을 무단전재 또는 복제하는 것을 금합니다.
본 리서치 브리프의 내용은 필자들의 개인적 의견이며, 국가데이터연구원의 공식적인 의견이 아님을 밝힙니다.

발 행 일 2026년 10월 30일
발 행 처 국가데이터처 국가데이터연구원
발 행 인 김 서 영
주 소 35220 대전광역시 서구 한밭대로 713 통계센터 6~8층
전화번호 042-366-7118
홈페이지 mods.go.kr/dsri