

2026년 연구보고서

사망원인통계 원사인 분류 시 적용 연구

2026. 9.



국가데이터처
국가데이터연구원

연구보고서 2026-02

사망원인통계 원사인 분류 AI 적용 연구

이석민 · 전고설 · 최지원 · 오교중



Ministry of Data and Statistics
Data and Statistics Research
Institute

발간사

국가데이터·통계는 사회와 경제의 변화를 정확히 이해하고 미래를 준비하기 위한 핵심 자산입니다. 최근 인공지능(AI)을 비롯한 디지털 기술의 발전과 데이터 활용 환경의 변화는 국가데이터·통계의 생산과 관리, 활용 방식에도 새로운 혁신을 요구하고 있습니다.

국가데이터연구원은 이러한 변화에 능동적으로 대응하여 국가데이터·통계의 정확성과 활용성을 높이고, 데이터에 기반한 정책 수립을 지원하기 위한 연구를 지속적으로 수행해 왔습니다.

2026년 연구보고서에는 국가데이터·통계의 체계적 관리와 인공지능(AI) 기반 통계 생산 혁신, 그리고 우리 사회와 경제의 주요 변화를 심층적으로 분석한 연구 성과를 담았습니다.

첫째, 국가데이터·통계를 체계적으로 관리하고 통계생산의 효율성을 높이기 위한 연구를 강화하였습니다. 국가데이터·통계에서 사용되는 개념과 정보를 일관성 있게 관리하기 위해 국제표준인 GSIM을 활용할 수 있는 방안을 연구하였으며, 또한, 사망 원인통계의 원사인 분류와 기업통계등록부 정비과정의 산업분류에 인공지능(AI) 기술 적용 가능성을 연구하여, 전문적인 판단과 많은 시간이 필요한 통계분류 업무의 정확성과 효율성을 높이기 위한 방안을 제시하였습니다.

둘째, 다양한 데이터를 연계·활용하여 청년층의 경제적 이동과 삶의 변화를 입체적으로 파악하고자 하였습니다. 통계등록부를 활용하여 청년층의 경제적 자립과 지역 이동의 특성을 이해할 수 있는 소득 변화와 이동성을 분석하였습니다. 이러한 연구는 청년정책과 지역정책을 보다 정교하게 설계하기 위한 기초자료로 활용될 수 있을 것입니다.

셋째, 국민의 삶과 지속가능한 발전을 종합적으로 측정하기 위한 연구를 수행하였습니다. 주요국의 삶의 질 지표와 지속가능발전목표(SDG) 지표체계를 비교·분석하여 각국의 지표 구성과 활용방식을 살펴보고, 우리나라 지표체계의 발전 방향과 정책적 활용도를 높이기 위한 시사점을 도출하였습니다. 이는 사회 발전을 경제적 성과뿐만 아니라 삶의 질과 지속가능성의 관점에서 폭넓게 진단하는 데 기여할 것으로 기대합니다.

이번 연구들은 국가데이터·통계의 생산과 관리에 새로운 기술과 국제표준을 적극적으로 접목하는 한편, 다양한 데이터를 활용하여 사회·경제 현상을 심층적으로 이해하려는 노력의 결과입니다.

국가데이터연구원은 앞으로도 최신 기술과 다양한 데이터를 적극 활용하여 국가데이터·통계의 전문성과 신뢰성을 높이고, 시대적 변화에 선제적으로 대응할 수 있는 연구를 이어가겠습니다.

본 연구보고서가 통계 생산자와 이용자 모두에게 실질적인 도움이 되고, 데이터에 기반한 정책 수립과 의사결정에 유용한 자료로 활용되기를 기대합니다.

많은 관심과 성원을 부탁드립니다.

2026년 9월

국가데이터연구원 연구기획실장



목 차

제1장 서론	1
제1절 연구 배경	1
제2절 연구 목적 및 필요성	2
제3절 연구 범위 및 수행 방법	3
 제2장 사망원인통계 작성 체계 및 환경 변화	4
제1절 사망원인통계 작성 절차와 원사인 분류체계	4
제2절 제11차 국제질병사인분류 도입 영향	9
제3절 사망원인통계 작성 현장의 과제와 개선 수요	10
 제3장 사망원인통계 AI 적용 선행사례 검토	12
제1절 국내 사망원인통계 관련 연구 동향	12
제2절 해외 사망원인통계 관련 연구 동향	17
제3절 선행연구의 시사점	21
 제4장 연구 데이터 및 분석 방법론	23
제1절 연구 기초자료 특성 및 현황	23
제2절 수행 방법론 및 성능평가 지표	33

제5장 사망원인통계 AI 적용 모형 분석 결과	51
제1절 (모델1) 사망진단서 사인 코드화 모형	51
제2절 (모델2) 행정자료 반영 잠정 원사인 예측 모형	65
제3절 (모델3) 약물중독 코드 자동분류 모형	101
제4절 (모델4) 자살 위험요인 자동분류 모형	108

제6장 AI 적용에 따른 업무 개선 및 통계서비스 확장 방안	122
제1절 사망원인통계 서비스의 확대와 정책적 함의	123
제2절 사인분류 업무체계 및 업무시스템의 구조적 전환	125
제3절 통계서비스 확장과 데이터 기반 정책지원 체계 구축	134

제7장 결론	137
--------------	-----

참고문헌	143
부 록	144
Abstract	151

요 약

본 연구는 사망원인통계 원사인 분류에 대한 AI 적용 가능성을 네 가지 모델 개발을 통해 실증적으로 검증하고, 이를 바탕으로 사인분류 업무체계의 개편방향과 사망원인통계 서비스의 확대방안을 제시하였다.

사망원인통계는 사망진단서에 기재된 사인정보를 기반으로 국민의 사망 수준과 사망 구조를 파악하는 통계로, 국제질병사인분류체계에 따라 작성된다. 그러나 고령화에 따른 사망자 수 증가, 세계보건기구(WHO)의 제11차 국제질병사인분류(ICD-11) 도입, 감염병 등 특정 사인에 대한 시의성 있는 통계 수요의 증가로 인해 사망원인통계의 작성 환경은 구조적 변화에 직면해 있다. 이에 본 연구는 원사인 분류 업무의 정확성과 행정 효율성을 높이고, 보건정책 수요에 대응할 수 있는 AI 기반 통계서비스의 토대를 마련하기 위해 수행되었다.

본 연구에서는 사망원인통계 작성 및 서비스 과정의 주요 업무에 AI를 적용하는 네 가지의 모델을 개발하였다. 모델1은 사망진단서의 사인 정보에 한국표준질병사인분류(KCD) 코드를 자동 부여하는 사인 코드화 모형이며, 모델2는 사망진단서와 건강보험 수진기록을 결합하여 잠정 원사인을 예측하는 모형이다, 모델3은 약물 중독 사망으로부터 약물 성분을 자동 추출 및 분류하는 모형이며, 모델4는 행정 자료의 서술형 정보로부터 자살 위험요인을 자동으로 분류하는 모형이다. 이 중 모델1과 모델3은 사인분류사의 검토 업무를 지원하는 실무적 성격을 띠고, 모델2와 모델4는 세부 정보를 구조화하는 서비스 확장의 성격을 가진다.

모델의 성능과 활용 가능성을 검증한 결과, 사망진단서 사인 코드화 모형(모델1)의 주요 오류는 모델 구조보다 학습에 활용된 데이터 품질에서 비롯되는 경우가 많았다. ICD-11 코드화의 경우에도 방법론적 기반은 마련되었으나, 정답 데이터가 충분히 축적되지 않아 현장에 바로 적용할 수준에는 이르지 못하였다.

잠정 원사인 예측 모형(모델2)는 104항목 분류 기준으로 84.6%의 정확도를 기록하여 기존 사망진단서로부터 도출한 잠정 원사인 정확도(75.8%)보다 8.8%p 높은 성능을 보였다.

약물중독 코드 자동분류 모형(모델3)은 국립과학수사연구원의 부검자료와 연계된 사망자 7,423명을 대상으로 약물 성분 추출 및 우선순위 판정 규칙을 적용하였다. 분석 결과 일부 유사계열 약물의 매핑 등에서 보완이 필요한 것으로 확인되었다.

자살 위험요인 분류모형(모델4)을 적용한 결과, 2024년 자살 사망자의 위험요인은 정신·행동장애(37.2%)와 경제적 요인(25.6%)에 집중된 것으로 나타났다. 그리고 기존 자료만으로는 포착하기 어려웠던 세부 요인을 추가로 식별함으로써 위험 요인 통계의 해상도를 높이는 효과를 보였다.

이를 종합하면 전반적인 사망원인통계 생산체계를 구조적으로 전환할 수 있는 가능성이 확인된다. 한글 사전과 색인 사전에 의존하던 기존의 매칭 방식을 AI 기반 코드화 방식으로 전환하되, 최종 판단 권한은 사인분류사에게 부여하는 형태로 업무 체계를 개편함으로써, 통계 품질을 떨어뜨리지 않으면서 업무 효율성을 높일 수 있다.

또한 기존의 연간 확정통계(사망 다음 해 9월 공표)에 잠정 사망원인통계와 자살 위험요인 통계를 단계적으로 추가하는 방향으로 통계 서비스를 확장할 수 있다. 아울러 감염병 사망 감시체계 구축과 다중 위험요인에 기반한 자살 위험요인 분석 등을 통해 데이터 기반 정책 기능을 강화하고, 보건정책의 대응력과 정밀성을 제고하는 데 기여할 것으로 기대된다.

주요용어: 사망원인통계, AI(인공지능), 원사인 분류, 자살 위험요인, 약물 사망, ICD-11

제 1 장

서 론

제1절 연구 배경

사망원인통계는 사망진단서에 기재된 사인 정보를 기반으로 국민의 사망 수준과 사망 구조를 체계적으로 파악하기 위한 국가승인통계로서, 「통계법」에 따른 지정통계이다. 이 통계는 보건의료 정책 수립, 감염병 대응, 산업재해 및 안전 정책, 사회적 위험 요인 분석 등 다양한 정책 분야에서 핵심 기초자료로 활용되고 있으며, 국제적으로는 세계보건기구(WHO)가 제시하는 국제질병사인분류(ICD)를 기준으로 작성된다.

사망원인통계 주요 기초자료는 의사가 발급하는 사망진단서와 유가족이 신고하는 사망신고서에 포함된 사망자 정보이다. 이러한 자료를 바탕으로 개인의 사망에 이르게 한 근원적인 요인을 원사인으로 선정한 후 이를 집계하여, 사망 규모와 사망원인의 구조적 특성을 분석한다. 특히 사망진단서에 기재된 다수의 사인 정보를 국제질병사인분류(ICD)에 따라 원사인으로 정확히 분류하는 과정은 통계의 신뢰성과 국제 비교 가능성을 좌우하는 핵심 절차라 할 수 있다.

그러나 최근 사망원인통계 작성 환경은 여러 측면에서 구조적 변화를 맞이하고 있다. 첫째, 사망진단서에 기재되는 사인 표현이 점차 복잡·다양화되면서 원사인 분류 업무의 난이도와 담당 인력의 업무 부담이 지속적으로 증가하고 있다. 특히 고령화에 따른 사망자 수 증가는 통계 작성 전반의 업무 부담을 가중시키는 주요 요인으로 작용하고 있다.

둘째, 세계보건기구(WHO)는 약 35년 만의 대규모 개편인 제11차 국제질병사인분류(ICD-11)의 도입을 추진하고 있으며, 이는 기존의 사전 기반·규칙 기반 분류체계에 근본적인 전환을 요구한다. 셋째, 감염병 등 특정 사망원인에 대해 보다 시의성 있는 통계 제공에 대한 정책적 수요가 증가하고 있으나, 현행 통계 생산 체계로는 잠정 통계의 정확성과 제공 시점에 한계가 존재한다.

이와 같은 환경 변화는 수작업 중심의 분류 방식만으로는 대응에 구조적 한계가 있음을 시사하며, 분류 정확도와 행정 효율성을 동시에 제고할 수 있는 지능형 업무 지원 체계의 도입 필요성을 부각시키고 있다.

제2절 연구 목적 및 필요성

본 연구는 사망원인통계 작성의 핵심 단계인 원사인 분류 업무의 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 마련하는 것을 목적으로 한다.

본 연구는 다음과 같은 측면에서 정책적·통계적 필요성을 가진다.

첫째, 기존 사망원인 분류체계는 한글사전과 색인사전에 대한 의존도가 높아 새로운 질병 개념이나 복합적 서술 표현에 대응하는 데 구조적 한계가 있다. 특히 한국표준질병사인분류(KCD)의 개정이나 신규 용어 추가 시마다 사전 전반에 대한 수정이 필요하며, 장기간 누적된 사전 규모는 시스템 유지·운영 부담을 가중시키고 있다.

더욱이 ICD-11과 같이 분류체계 자체가 대폭 변동되는 경우에는 기존 사전 기반 접근 방식만으로는 업무 수행에 상당한 어려움이 발생할 수 있다. AI 기반 분류 방식은 이러한 제약을 완화하고, ICD-11 도입에 필요한 데이터 축적과 분류체계 전환의 실질적인 기반을 제공할 수 있다.

둘째, 행정자료를 활용한 AI 기반 예측은 특정 사인에 대해 보다 이른 시점에서 유용한 통계 정보를 제공할 가능성을 보여줌으로써, 감염병 등 정책적 관심 사인에 대한 신속한 대응을 지원할 수 있다. 현재 사망원인통계는 사망 발생 연도의 다음 해 9월에 공표되고 있는데, 이는 사망신고 지연과 사망진단서 정보의 보정·검증을 위한 행정자료 연계 과정이 필요하기 때문이다.

그러나 보건 및 안전 정책에서는 보다 신속한 통계 제공이 요구되고 있다. 실제로 고의적 자해 사망(자살)의 경우 이러한 정책적 관심에 따라 국가 자살동향 감시시스템을 구축하고, 월별 잠정 자살 사망자 수를 공표하고 있다. 이러한 사례는 특정 사인에 대한 잠정 통계서비스 확대의 필요성을 시사한다.

셋째, 서술형 정보 비중이 높은 영역에서 AI를 활용한 정보 추출과 분류 지원은 통계 품질 개선과 함께 현장 업무 부담 경감과 정책수요 신규 지표 개발에 기여할 수 있다. 약물중독 사망의 경우, 정확한 성분 식별과 코드화는 분류 난이도가 높은 영역에 해당한다. AI 기반 정보 추출은 이러한 복잡 사례에서 오분류 가능성을 낮추고 시간 절감에 기여할 수 있다.

한편, 자살 위험요인과 같이 자료를 해석하는 사람의 주관이 개입할 여지가 큰 영역에서는 일관된 기준에 따른 분류와 복합적인 분석이 가능해짐에 따라 자살 위험요인별 사망자 통계와 같은 정책 수요 대응형 통계 개발의 기반을 마련할 수 있다.

제3절 연구 범위 및 수행 방법

본 연구는 사망원인통계 작성 전 과정을 전면적으로 자동화하거나 재설계하는 것을 목표로 하기보다는, AI 적용의 실현 가능성과 효과를 단계적으로 검증하기 위한 실증 연구로 수행된다. 연구 범위는 ① 사망진단서 사인 정보의 자동 코드화, ② 행정자료를 반영한 특정 사인의 잠정 원사인 예측, ③ 약물중독 사망에서의 약물 성분 추출 및 코드화와 우선순위 판단, ④ 서술형 행정자료에서의 자살 위험요인 분류 등 네 개 영역으로 구성된다.

본 연구에서는 먼저 제2장에서 사망원인통계의 작성 체계와 환경 변화를 살펴보고, 제3장에서 국내외 선행사례를 검토하였다. 이어 제4장에서는 앞에서 제시한 각 연구 영역의 세부 분석을 위해 기초자료 수집 및 전처리, AI 학습 및 예측, 결과 검증, 업무 적용 방안 도출의 단계를 중심으로 기술하였다. 구체적으로는 사망진단서 및 관련 행정자료를 대상으로 데이터 표준화와 학습·검증 데이터 분리를 수행한 후, 각 분석 모형별 AI 학습 및 예측 과정을 제시하였다.

제5장에서는 각 영역의 AI 예측 결과를 기존 원사인 선정 결과 및 최종 사망원인통계와 비교·검증하여 분류 정확도와 한계를 분석하였다. 마지막으로 제6장에서는 이러한 분석 결과를 토대로 사망원인통계 작성 과정에서 AI를 활용한 업무 개선 방안과 잠정 통계 서비스 확대 가능성을 종합적으로 검토하였다. 이를 바탕으로 사망원인통계의 과학화·자동화 수준을 제고하고, 향후 ICD-11 도입과 보건정책 수요 대응형 통계 서비스 개발을 위한 기반을 마련하고자 하였다.

제 2 장

사망원인통계 작성 체계 및 환경 변화

제1절 사망원인통계 작성 절차와 원사인 분류체계

사망원인통계는 「통계법」 제17조 및 제18조에 따라 지정된 국가승인통계(승인번호 제101054호)로서, 「가족관계의 등록 등에 관한 법률」과 「인구동향조사규칙」을 법적 근거로 한다. 전국 읍·면·동 및 시·구(재외국민은 재외공관)에 접수된 사망신고를 기반으로 작성되는 전수 통계이며, 영아사망·무연고 사망 등 신고 누락 가능성이 상대적으로 높은 사례에 대해서는 화장장 신고자료, 사망원인 보완조사, 무연고 사망 공고자료 등을 추가로 연계·반영함으로써 포괄성을 확보하고 있다.

사망원인통계의 핵심 개념은 원사인(Underlying Cause of Death)이다. 원사인이란 직접적으로 사망에 이르게 한 일련의 사건을 야기한 질병 또는 손상, 혹은 치명적 손상을 유발한 사고나 폭력의 상황을 의미한다. 사망진단서에는 직접사인과 그 선행 원인이 인과적 순서에 따라 기재되며, 이 가운데 근원적인 사망원인을 식별하여 통계 작성의 기본 단위로 활용하는 것이 원사인 선정의 목적이다.

사망원인 분류에는 세계보건기구(WHO)가 제정한 국제질병사인분류(ICD)와 이를 기반으로 국내 실정에 맞게 조정·고시한 한국표준질병사인분류(KCD)가 기준으로 활용된다. ICD는 질병과 사망원인을 국제적으로 표준화하여 비교 가능성을 확보하기 위해 제정되었으며, 복합적인 사망원인 중 보건의료적 개입이 가능한 근원 요인을 식별하기 위한 원사인 선정 방법을 포함하고 있다. 우리나라 사망원인통계는 WHO가 권고하는 원사인 개념과 선정 규칙(SP1~SP8, M1~M4) 및 특별지침을 단계적으로 적용하여 원사인을 결정하며, 통계 공표 시에는 WHO 일반사망요약분류표, 한국사인요약분류표 등 목적별 요약 분류체계를 활용하여 사망원인을 집계한다.

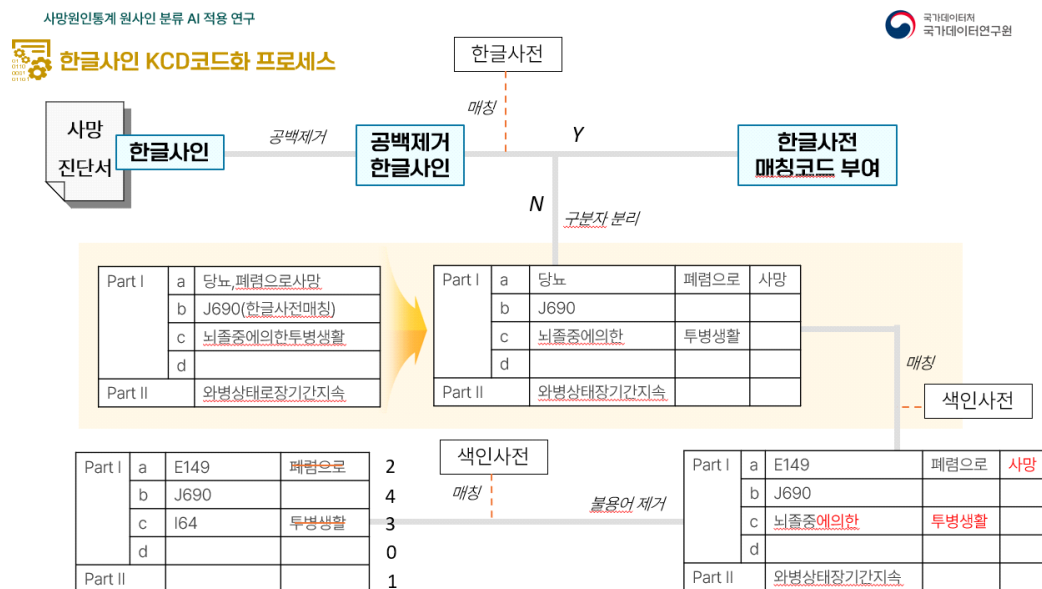
사망원인통계 작성 절차는 크게 자료 입수, 사인 코드화, 원사인 선정, 결과 검증의 단계로 구성된다. 자료 입수 단계에서는 지자체가 인구동향조사시스템을 통해 사망신고서와 사망진단서(시체검안서)를 입력하며, 해당 자료는 사망원인선정시스템으로 연계된다. 이후 사인 코드화부터 결과 검증까지 자동화된 시스템 처리와 사인분류사의 검토가 결합된 형태로 운영되며, 각 단계의 처리 결과가 다음 단계의 입력으로 연계되는 다단계 구조를 이룬다.

백을 제거한 '공백 제거 사망원인명'을 먼저 생성한 후 한글사전과 대조한다.

한글사전 매칭이 실패한 경우에는 쉽표·조사 등 구분자를 기준으로 사인 문구를 분리하여 색인어를 생성하고, 색인사전을 통해 매칭한다. 사망진단서 작성 지침에는 한 줄에 하나의 사망원인명을 기재하게 되어 있으나, 실제로는 여러 사망원인이 병렬로 기재되거나(예: 당뇨, 고혈압) 인과관계에 따라 연결어로 구성된 경우(예: 폐렴 등 호흡 부전에 의한 심폐기능 정지)가 빈번하다. 이에 따라 한 줄당 최대 10칸까지 색인어를 분리하여 각각 코드를 부여한다.

이러한 과정을 거쳐 사망진단서에 기재된 사망원인 5개 항목(직접사인, 가의 원인, 나의 원인, 다의 원인, 기타 신체상황)에 대한 KCD 코드가 매칭되며, 매칭 이력과 정확도가 별도 데이터로 관리된다. 두 사전은 용량이 크고 기준 변경에 대한 대응이 어려운 구조이다. 따라서 표준화된 질병 명칭에 대해서는 높은 정확도를 보이지만, 복합적·서술형 표현이나 새로운 질병 개념에 대해서는 분류의 한계를 가진다.

<그림 2-3> 사망원인 코드화 단계



잠정 원사인 선정 단계에서는 코드화된 사인 정보를 바탕으로 근원적 사망원인을 도출한다. 사망진단서에 기재된 사망원인이 1개인 경우에는 단독사인으로 해당 사인이 곧바로 원사인이 된다. 반면 여러 개의 사망원인이 기재된 경우에는 WHO가 규정하는 약 9만 개의 병태 간 인과관계와 약 5만 개의 특별 관계 지침을 근거로 원사인을 선정한다. 이 규정은 ICD-10 제2권 지침서에 수록되어 있으며, 시작점 선정규칙(SP1~SP8)과 수정준칙(M1~M4) 등 총 12개의 규칙을 단계적으로 적용하여 잠정 원사인을 선정한다.

선정규칙은 사망에 이르게 한 인과적 연쇄의 시작점을 식별하는 단계로, 이를 위해 시작점 선정규칙(SP1~SP8)이 단계적으로 적용된다. 이 규칙들은 사망진단서에 단일 병태만 기재된 단순 사례부터, 여러 병태가 복합적으로 보고된 복잡 사례까지 포괄하도록 설계되어 있다. 특히 사망진단서의 하단에 기재된 병태가 상단의 모든 병태를 설명할 수 있는지, 명백한 원인이 존재하는지, 또는 여러 개의 인과 연쇄가 동시에 존재하는지를 체계적으로 검토하도록 규정하고 있다. 이때 선정규칙에 사용하는 병태 간 인과관계는 단순한 의학적 개연성뿐 아니라, 역학적·공중보건적 중요성이 인과관계의 수용 여부를 판단하는 핵심 기준으로 작용한다.

시작점이 확인된 이후에는 특정 질환군이나 상황에 대해 적용되는 특별 수정지침(M1~M4)이 검토된다. 일부 경우에는 시작점 자체가 최종 원사인으로 확정되지만, 공중보건적 중요성이 더 큰 병태를 반영하기 위해 시작점과 다른 병태가 원사인으로 수정·선정되기도 한다. 이러한 수정 절차는 즉상경화증과 심장질환의 관계, 외부원인 손상, 중독, 모성사망, 의료행위 관련 사망 등에서 특히 중요하게 적용된다. 이는 원사인 선정 과정이 순수한 의학적 판단이 아니라 공중보건 통계 작성이라는 목적에 종속되어 있음을 분명히 한다.

행정자료 보완 단계에서는 시스템에서 자동 선정된 잠정 원사인에 대해 행정자료를 활용한 검토와 보완이 이루어진다. 이는 사망진단서의 기재 오류나 정보 부족으로 인해 발생할 수 있는 원사인 오분류를 최소화하기 위한 절차이다. 사망진단서를 기반으로 선정된 원사인 중 절반 정도는 사인분류사의 행정자료의 검토 및 확인 과정을 거쳐 최종 원사인으로 보완된다.

사망원인통계는 법적 신고의무에 기반한 전수자료로서 높은 포괄성을 갖지만, 기초 투입자료의 특성상 구조적인 한계를 내포하고 있다. 우선 사망신고 시 사망원인 파악을 위해 첨부되는 사망진단서(또는 시체검안서)의 제출률은 점진적으로 개선되고 있음에도 불구하고, 여전히 일정 수준의 미첨부 및 부실 기재 문제가 지속되고 있다. 특히 ‘노환’, ‘심정지’, ‘사인미상’, ‘의도불명 외인사’와 같이 병인적 인과관계가 명확하지 않은 사인은 원사인 선정에 직접적인 제약으로 작용하며, 이러한 증상·징후 중심 사망의 비중은 사망원인통계의 국제 비교 가능성과 분석 정밀도를 저하시킨다. 이러한 기초자료의 한계를 보완하기 위해 사망원인통계는 다양한 행정자료를 연계·활용하고 있다.

행정자료 활용의 주된 목적은 사망진단서에 기재된 사인이 불명확하거나 누락된 경우, 또는 사회적·문화적 요인으로 인해 사망원인이 왜곡될 가능성이 있는 사례(예: 자살의 사고사 분류)를 보정하기 위함이다. 특히, 사망원인 오분류 문제를 최소화하기 위해 건강보험 자료, 경찰 변사자료, 압등록자료 등 다수의 행정자료가 사망자 개인 단위로 연계되어 사망원인 결정 과정에 활용되고 있다. 실제로 행정자료에 의해 사망원인

이 보완된 사례의 비중은 시간이 지남에 따라 지속적으로 증가해 왔으며, 이는 행정자료 연계가 사망원인통계의 정확성 제고에 실질적으로 기여하고 있음을 시사한다.

결과 검증 단계에서는 원사인 선정 결과가 단계별 절차를 거쳐 체계적으로 점검·확정된다. 우선 기초 사망신고 자료는 입력 단계에서부터 항목 간 논리적 정합성에 대한 자동 내검을 거치며, 이후 사망원인선정시스템을 통해 사망진단서 기재 내용과 행정자료를 연계한 원사인(underlying cause of death) 선정 과정이 수행된다. 이 과정에서는 질병 간 인과관계, 연령·성별 특이 사인, 외인사 코드의 적합성 여부 등 WHO 국제기준에 따른 다차원적 내검 규칙이 적용되어, 자동 선정된 원사인과 수동 보정 결과 간의 일관성이 반복적으로 검토된다.

또한 사망원인통계는 사망진단서에 기재된 정보만으로 원사인 선정이 곤란한 사례를 고려하여, 건강보험 자료, 경찰 변사자료, 압등록자료 등 다수의 행정자료를 개인 단위로 연계·대조함으로써 결과 검증을 강화하고 있다. 행정자료는 사망원인의 누락·오분류 가능성을 점검하는 보조적 검증 수단으로 활용되며, 특히 불명확한 사인(R 코드), 외인사, 자살 관련 사망의 경우 행정자료와의 교차 확인을 통해 사망원인 결정의 타당성을 확보한다. 이러한 연계 검증은 단순 보완을 넘어, 사망원인 결과의 일관성과 재현 가능성을 제고하는 핵심 절차로 기능한다.

통계처리 및 분석 단계에서는 집계된 사망원인 결과에 대해 수준분석(level analysis)과 구조적 검토가 병행된다. 연령·성별·지역별 사망률 분포, 특정 사인의 연령 집중 여부, 국제적으로 보고되지 않는 비정상적 사인 패턴 등은 결과 검증의 주요 점검 항목으로 활용되며, 이상치 또는 비논리적 분포가 확인될 경우 원자료 및 원사인 선정 이력까지 역추적하여 재검토가 이루어진다. 이러한 과정은 사망원인통계 결과가 단순 집계 산출물이 아니라, 반복적 검증과 수정 과정을 거쳐 확정된 분석 결과임을 제도적으로 보장한다.

종합하면, 사망원인통계는 입력 내검, 행정자료 연계, 원사인 선정 및 검증, 수준분석으로 이어지는 다층적 절차를 통해 수행되며, 이는 사망원인 결과의 정확성뿐 아니라 시계열적 일관성과 국제 비교 가능성을 동시에 확보하기 위한 장치로 이해할 수 있다.

제2절 제11차 국제질병사인분류 도입 영향

WHO의 국제질병사인분류(ICD)는 1990년 ICD-10 채택 이후 약 30년 만인 2019년 5월 새로운 분류체계인 ICD-11이 채택되었으며, 2022년부터 공식 발효되었다. ICD-11은 분류체계의 범위·구조·운용 방식 전반에서 ICD-10과 근본적인 차이를 보이며, 이러한 변화는 사망원인통계 작성 체계 전반에 대한 점검과 대응을 요구하고 있다.

ICD-10에서는 'A00.0 - Z99.9' 형태의 영문자-숫자 조합을 사용하는 약 14,400개의 코드가 운용되었다. 반면 ICD-11에서는 코드체계가 '1A00.00 - ZZ9Z.ZZ' 범위로 확장되어 34,000개 이상의 코드가 제공되며, 최대 6단위까지 세분화된다. 코드 수의 증가는 질병 개념의 세분화를 반영한 것이지만, 동시에 사인 코드화 과정에서의 분류 복잡도를 높이는 요인으로 작용한다.

ICD-11에서는 분류의 구조 자체도 개편되었다. 대표적으로 ICD-10에서 제4편 형태 분류(Morphology)로 별도 관리되던 신생물의 조직학적 형태 유형이 ICD-11에서는 제2장 신생물 본문에 통합되었다. 이 외에도 장(chapter)의 구성과 포함 범위, 코드 간 위계 관계에서 광범위한 변경이 이루어져, 기존 분류체계에 익숙한 실무 환경에서는 상당한 전환 부담이 발생할 것으로 예상된다.

현행 사인 코드화 체계는 한글사전과 색인사전을 기반으로 한 규칙 기반 매칭 방식에 의존하고 있다. ICD-11로의 전환은 코드체계와 분류 구조가 동시에 변경되는 것을 의미하므로, 현재 운영 중인 한글사전과 색인사전 역시 전면적인 재구축이 불가피하다. 또한 약 14,400개에서 34,000개 이상으로 코드가 확대됨에 따라 사전 기반 매칭 방식만으로는 전환기의 분류 정확도를 안정적으로 유지하는 데 한계가 있을 것으로 예상된다. 원사인 선정 규칙 역시 ICD-11 체계에 맞추어 개정이 진행되고 있어, 약 9만 개의 병태 간 인과관계 테이블과 수정준칙 전반에 대한 갱신이 함께 이루어질 것으로 예상된다.

우리나라의 한국표준질병사인분류(KCD)는 2025년 7월 고시되어 2026년 1월부터 시행 중인 제9차 개정이 최신 버전이며, 여전히 ICD-10 체계를 기반으로 한다. ICD-11을 반영한 KCD 제10차 개정은 2027년 7월 고시를 목표로 개발이 진행되고 있다. 이와 같은 코드체계의 전환은 사전 구축, 정보시스템 개편, 업무 절차 개선, 담당 인력 교육 등 전반적인 대응을 필요로 하며, 전환기에도 분류 정확도와 시계열의 연속성을 유지할 수 있는 기술적 기반을 사전에 마련하는 것이 중요한 과제로 부각되고 있다.

이러한 변화는 기존 규칙 기반 분류체계만으로는 효과적인 대응에 한계가 있음을 보여주며, 새로운 분류체계에 유연하게 대응할 수 있는 AI 기반 분류 기술의 필요성을 더욱 부각시키고 있다.

제3절 사망원인통계 작성 현장의 과제와 개선 수요

ICD-11 도입에 따른 국제표준 개편은 기존 작성 체계에서 누적되어 온 실무적 한계와 맞물리며, 사망원인통계 작성 현장의 부담을 더욱 가중시키고 있다. 사망원인통계 작성 현장의 과제를 실증적으로 파악하기 위하여, 국가데이터처 인구동향과 소속 사인분류사 11명을 대상으로 2026년 2월 의견수렴을 실시하였다. 의견수렴은 사인분류사업부 부담이 큰 영역인 사인 코드화, 행정자료 연계 분석, 서술형 자료 처리를 중심으로 구성되었다.

사인 코드화 영역에서는 응답자 전원이 기존 사인선정시스템의 오분류 사례를 업무과정에서 자주 경험하며, 이로 인해 사인분류 업무에 영향을 미친다고 응답하였다. 한글사전 및 색인사전의 품질에 대해서는 사전이 장기간 갱신되지 않아 매칭이 원활하지 않으며 사전의 활용성이 크게 저하되었다는 의견이 제시되었다. 특히 한글사전의 일부 항목은 KCD 7차 개정 수준에서 이후 갱신되지 않았다는 의견이 제시되어, 사전의 현행화 지연이 분류사 업무 부담을 구조적으로 가중시키는 요인으로 작용하고 있음을 보여준다.

행정자료 연계 분석 영역에서는 AI가 사망진단서와 건강보험 자료 등 행정자료를 종합 분석하여 잠정 원사인을 제시하는 방식의 유용성 자체에 대해서는 부정적 응답이 확인되지 않았다. 다만 그 결과의 신뢰수준에 대해서는 응답 간 온도 차이가 존재하였다. 일부 응답자는 객관성을 높이는 보정 수단이자, 1차 분류 지원 도구로서의 활용 가능성을 높게 평가하였다. 반면, 잠정 수치 확인이 빨라지므로 대국민 서비스 측면에서는 긍정적이나 분류 정확도 측면에서 업무에 큰 도움이 되지 않을 수 있다는 의견, AI 예측 결과의 정확도에 따라 활용 가능성이 달라질 것이라는 의견도 제시되었다. 공통적으로 부각된 의견은 결과 제시 시점이 분류사가 원사인을 선정하기 이전일 것, 최종 결정은 분류사를 반드시 거쳐야 하므로 의사결정을 지원하는 참고자료로 활용되어야 한다는 의견이었다.

서술형 행정자료 처리 영역에서는 경찰·해경·국립과학수사연구원에서 제공하는 사건개요 및 분류 텍스트로부터 우선적으로 요약·표출되기를 희망하는 항목으로 손상부위 및 손상우선순위, 약물순위 및 약물명, 발생장소 및 장소코드, 사망구분이 다수 의견으로 나타났다. 약물중독 사망에서는 하나의 의약품에 대응되는 성분명이 존재하지만, 사건 자료에서는 상품명이나 오탈자가 포함된 형태로 기재되는 경우가 많아 사인분류사가 이를 개별적으로 검색·확인하는 부담이 발생하고 있다. 둘 이상의 약물이 기재된 경우에는 약물·약제 및 생물학적 물질에 의한 중독에 대한 특별 지침(M4 단계)에 따라 약물 우선순위를 판정해야 하나, 이 역시 수작업에 의존하고 있어 상당한 처리 시간이 소요된다.

사인분류사 의견수렴에서 확인된 실무 부담은 외부의 정책 수요 변화와 맞물리면서 그 시급성이 더욱 부각되고 있다. 특히 감염병을 비롯한 특정 사인에 대한 시의성 있는 통계 제공 요구가 확대되면서, 현행 통계 산출 체계의 시차 구조가 정책 대응의 제약 요인으로 지적되고 있다.

현행 사망원인통계는 전년도 사망 자료를 다음 해 9월에 확정·공표하는 연간 산출 구조를 따르고 있다. 이에 따라 사망 발생 시점과 통계 공표 시점 사이에는 약 9개월에서 21개월까지의 시차가 발생한다. 이러한 시차가 발생하는 이유는 사망신고 의무자가 유가족이고 신고 기한이 사망 후 1개월로 설정되어 있는 점, 앞서 1절에서 살펴본 것과 같이 사망통계 작성 절차의 구조적 특성에서 비롯된다. 이 시차는 정확성과 국제 비교 가능성을 확보하기 위한 제도적 산물이지만, 특히 감염병처럼 신속한 현황 파악과 정책 대응이 요구되는 영역에서는 제약 요인으로 작동한다.

질병관리청(2025)은 감염병 사망 감시체계의 구조적 한계를 분석하면서, 확진자와 사망자 자료가 분절적으로 관리되고 사망원인통계가 다음 해 9월에 공표되는 현행 체계에서는 감염병 유행 시 신속한 정책 대응에 한계가 있음을 지적하였다. 메르스와 COVID-19 대유행을 통해 초과사망을 포함한 사망 모니터링의 중요성이 확인되었으나, 확정 통계 중심의 생산 구조로 인해 시의성 확보에는 제약이 존재한다.

또한 법정 감염병 신고체계 자체의 완전성 확보를 위한 노력도 요구되는 상황이다. 1~3급 감염병은 전수 신고 대상이나 4급 감염병은 표본 감시 방식으로 운영되고 있고, 환자 전원, 사후 검출, 원내 사망 등의 사유로 사망 신고가 누락될 가능성이 존재한다. 병원 의무기록과 건강보험료 청구 자료의 공유 역시 제한적이어서, 감염병 확진 이후 사망에 이르는 경과를 자료 수준에서 통합적으로 추적하기 어려운 구조가 지속되고 있다. 이러한 정책 수요에 대응하기 위해서는 현재 확보 가능한 행정자료를 종합적으로 연계·분석하여 확정 통계 이전 단계에서 감염병 사망의 규모를 신속하게 추정할 수 있는 기술적 접근이 병행될 필요가 있다.

이상의 검토는 사망원인통계 작성 현장이 직면한 과제를 실무와 정책 두 측면에서 확인시킨다. 실무적으로는 사전 기반 코드화 체계의 한계와 ICD-11 전환에 따른 분류체계 개편 부담이 지속되고 있으며, 서술형·행정자료 처리에 대한 업무 부담도 증가하고 있다. 정책적으로는 감염병 등 특정 사인에 대한 신속한 통계 제공이 요구되고 있으나, 현행 통계 생산체계는 이러한 수요에 충분히 대응하기 어려운 구조이다. 따라서 기존의 규칙 기반 분류체계만으로는 변화하는 작성 환경에 효과적으로 대응하기 어려우며, 통계 작성 전반을 보조·고도화할 수 있는 새로운 접근 방식이 요구된다. 본 연구는 이러한 문제의식을 바탕으로 사망원인통계의 원사인 분류체계에 인공지능(AI) 도입을 중심으로 개선 가능성을 실증적으로 검토하고자 한다.

제 3 장

사망원인통계 AI 적용 선행사례 검토

제1절 국내 사망원인통계 관련 연구 동향

1. 연구 개요

국가데이터처(구 통계청)는 비정형 텍스트 입력을 기반으로 하는 통계 분류 업무의 효율성과 정확성을 제고하기 위해, 인공지능 기술을 활용한 통계분류 자동화 연구를 단계적으로 추진해 왔다. 그 일환으로 2023년 국가데이터연구원(구 통계개발원)에서 수행된 「인공지능 통계분류 자동화 확대 적용 연구」는 생활시간조사, 가계동향조사와 함께 사망원인통계를 주요 적용 대상으로 선정하여, 기존 자동화 시스템의 한계를 보완하고 AI 기반 분류 방법론의 적용 가능성을 실증적으로 검토하였다.

이 연구에서 사망원인통계 분야는 자연어 처리(NLP)와 기계학습 기반 모델을 활용하여 사망진단서에 기재된 사인 정보를 입력값으로 하고, 한국표준질병사인분류(KCD-8)에 따른 원사인을 예측하는 방식으로 분석하였다. 특화 언어모델로는 KM-BERT를 활용하였는데, KM-BERT는 한국어 의료 도메인에 특화된 BERT 기반 사전훈련 언어모델로서, 의학관련 지식(논문 및 기사 등)을 통해 약 6백만 개의 문장과 약 1억 1천6백만 개의 단어로 이루어진 한국어 말뭉치를 기반으로 의료 용어의 언어적 특성을 잘 반영할 수 있도록 학습된 것이 특징이다.

선행연구에서는 사망원인통계 작성 과정에서 생성되는 대규모 행정자료를 활용하여 AI 분류 실험을 진행하였다. 분석 대상은 2021~2022년 사망진단서 693,587건이며, 질병사인 분류 전반을 포괄하였다. 분류체계는 대분류 22개, 중분류 267개, 소분류 2,085개, 세분류 12,503개에 이르는 다층적인 구조를 그대로 적용하였다.

이는 사망원인통계가 다른 사회통계에 비해 분류 항목 수가 매우 많고, 희귀 사례가 다수 존재하는 구조적 특성을 반영한 것이다. 기초자료로 활용한 데이터 항목은 사망진단서의 사망원인 5개 항목, 사망진단서 상 원사인, 행정자료를 반영한 수정된 원사인을 활용하였다. 비교방법은 다음과 같이 3가지로 나누어 수행하였다.

① 행정자료 보완된 원사인 비교

사망진단서에 기재된 5개 사망원인 정보만으로 인공지능이 예측한 원사인과, 사망진단서 내용과 행정자료까지 보완한 사망원인통계의 최종 원사인을 비교하였다. 이 비교는 행정자료를 활용하지 않은 상태에서 최종 원사인을 얼마나 정확하게 예측할 수 있는지를 평가하기 위한 것이다.

② 행정자료 보완되지 않은 원사인 비교

WHO의 원사인 선정 규칙에 따라 도출된 사망원인통계의 잠정 원사인과 인공지능의 학습된 결과와의 일치성을 비교하였다. 즉, 행정자료 보정 이전 단계의 잠정 원사인에 대한 정확도를 측정한 모델이다. 다만 기본정보가 부족하여 외인사로 사망한 사람을 제외하고 분석하였다.

③ 사망진단서 기재된 단일 사망원인 비교

사망진단서에 기재되는 5개 사망원인 각각의 자연어를 KCD-8(ICD-10)으로 코드화하는 과정을 비교하였다. 다수 자연어에 대한 매핑테이블을 활용한 방식인 기존 방식과 인공지능으로 용어 매핑 테이블과 KCD-8(ICD-10)사전을 학습한 뒤 예측한 코드를 비교하였다.

2. AI 기반 사망원인 분류 실험 결과

가. 행정자료 보정된 최종 원사인과의 비교

인공지능의 원사인 예측 정도를 파악하기 위하여 사망진단서만을 학습하여 인공지능이 예측한 원사인과 사망진단서 및 다수 행정자료를 활용하여 공식통계에 공표한 원사인을 비교하였다. 2021년 사망신고된 317,356건을 학습데이터로 이용하여 2022년 신고된 306,898건의 데이터를 평가 비교하였다.

그 결과 300 steps(=약 50 epoch)에서 최적의 분류 모델이 학습되는 것을 확인할 수 있었다. 2022년 공표 데이터(306,898건)에 대하여 전수 평가 수행 결과, 62.65% 일치하는 것을 알 수 있었고, 약 37.35% 정도의 사건에 대해서 사고사/외인사로 인해 내검수정 발생한다는 것을 간접적으로 파악할 수 있었다.

<표3-1> 인공지능의 최종 원사인 예측 결과

Accuracy	0.6265
Weighed F1	0.5940 (Precision 0.6029 / Recall 0.6265)
Macro F1	0.1503 (Precision 0.1739 / Recall 0.1560)
Loss	1.8495

나. 행정자료 보정되지 않은 잠정 원사인과의 비교

다음은 행정자료가 보정되지 않은 잠정 원사인과 비교하였다. 잠정 원사인은 사망 진단서의 한글로 기재된 병태를 용어 매핑테이블로 자동 변환한 뒤, WHO에서 규정한 ICD-10 vol.2 지침서를 기반으로 자동 선정한다. 이러한 일련의 단계를 인공지능이 예측할 수 있는지 측정하는 것은 매핑테이블 관리로 인한 업무부담을 경감할 수 있는 실마리를 찾게 해준다.

자동 원사인과 최종 원사인이 같은 정제 데이터(총 310,034건)를 학습에 사용하였으며, 정제된 학습데이터를 바탕으로 자동 원사인이 만들어 내는 결과와 유사하게 분류하는 기초모델을 학습하였다. 학습은 270 steps(약 90 epoch)에서 최적화되었고, 평가결과는 95.35%의 높은 정확도가 나타났다. 다만, 분류 항목별 예측 결과를 상세 분석한 결과를 살펴보면 인공지능의 예측이 취약한 부분이 발견된다. 평가데이터에서 1,026개 코드 중 신뢰할 수 있는 수준(F1-score 0.9이상)의 코드는 291개(E272, I7110, C740 등)인 반면, 신뢰할 수 없는 수준(0점)인 코드 410개(L892, A1830, R578 등)로 나타났다.

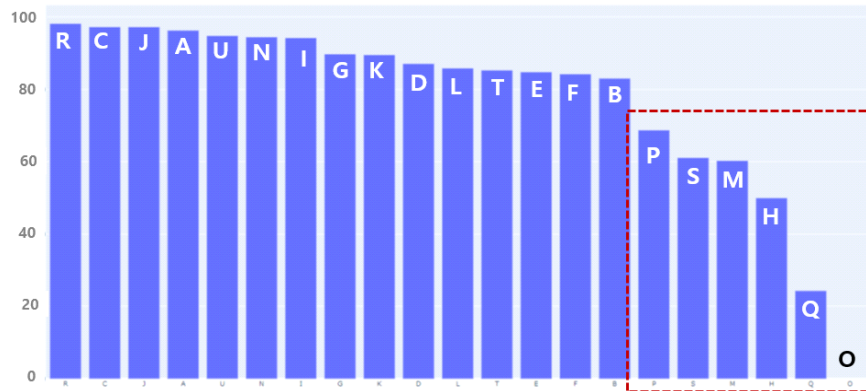
특히 P.출생전후기 병태, S.손상중독, M.근골격계질환, H.눈귀질환의 경우 신뢰할 수 없는 결과가 도출되었고, Q.선천기형과 O.임신, 출산 및 산후기 질환은 필수적으로 행정자료가 필요한 것으로 나타났다. 이를 통하여 세부 질병별로 반드시 내검이 필요한 영역이 존재함을 알 수 있었다.

오분류된 사망원인은 18,116개로 학습/검증/평가 데이터셋이 부족한 무뇌증, 콜레라 등 희귀질환이 대다수임을 알 수 있었으며, 데이터 수가 부족한 세분류 항목이 많아 분류 목적에 맞춘 코딩 사례 데이터 확보가 필요하다.

<표3-2> 인공지능의 잠정 원사인 예측 결과

Accuracy	0.9535
Weighed F1	0.9516 (Precision 0.9510 / Recall 0.9535)
Macro F1	0.4996 (Precision 0.5104 / Recall 0.5119)
Loss	0.3506

<그림3-1> ICD-10 대분류별 인공지능 잠정 원사인 정확도



다. 사망진단서 기재된 단일 사망원인 비교

사망진단서의 한글로 기재된 사망원인의 ICD-10 코드화 정확도를 인공지능과 용어 매핑 테이블을 비교하였다. 학습데이터는 ICD-10 코드리스트와 국가데이터처에서 사용 중인 용어 매핑 테이블을 활용하였다. 사망진단서에 기재된 단일 사망원인 자연어를 코드화한 결과를 비교한 결과, 미세조정 분류 모델은 1,550 steps(=약 50 epoch)에서 최적으로 학습이 됨을 알 수 있었다. 평가결과 개별 사망원인의 인공지능 정확도는 79.51%로 다소 낮은 수준이 나타났다. 단, Macro F1 점수가 0.925로 매우 높아 예측 결과가 나오는 항목에 대해서는 일치율이 높음을 알 수 있었다.

<표3-3> 인공지능의 개별 사망원인 ICD-10 코드 예측 결과

Accuracy	0.7951
Weighed F1	0.7741 (Precision 0.8512 / Recall 0.7951)
Macro F1	0.9250 (Precision 0.9346 / Recall 0.9589)

3. 인공지능 오분류 유형

대표적으로 인공지능이 사망원인통계 작성 과정에서 오류를 범할 수 있는 유형은 3가지이다. 첫째, 사망진단서에 흔히 나타나는 증상에 대하여 인공지능이 혼동을 발생할 수 있다. 예를 들어 패혈증의 경우 다양한 사망원인의 결과로 나타나는 증상으로 사망진단서에 흔히 기재되는 병태이다. 예를 들어 패혈증, 척추염, 폐렴이 사망진단서에 같이 기재된 경우 M46.59(기타 감염성 척추병증)으로 분류되어야 함에도 불구하고 빈번하게 학습된 M46.99(상세불명의 염증성 척추병증)으로 오분류된다.

둘째, 사망원인 선정 규칙에 대한 학습 미흡으로 질병의 인과관계와 관련한 오류가 발생한다. 예를 들어 당뇨병 신부전의 경우 E14.28(신장 합병증을 동반한 상세불명의 당뇨병)로 분류되어야 함에도 불구하고 E14.9(합병증을 동반하지 않은 상세불명의 당뇨병)으로 인공지능이 예측하였다.

셋째, 사망진단서 자연어에 대한 이해가 부족한 경우 오류가 발생한다. 예를 들어 사망진단서에 ‘동맥 관 개방’이라고 기재된 경우 ‘개방’에 대한 혼동으로 Q25.0(동맥관 개존)로 분류되어야 하는 것은 Q21.4(대동맥 폐동맥 중격 결손)로 분류하였다. 이와 같은 사례는 AI가 학습데이터의 빈도 분포에 크게 의존하는 특성으로 인해 제한적인 문맥 정보에서는 오분류가 발생할 수 있음을 보여준다.

이러한 오분류는 인공지능이 학습데이터의 빈도 분포에 의존하는 구조적 한계를 보여주며, 사인 선정 규칙과 질병 간 인과관계에 대한 추가적인 학습 기제가 필요함을 시사한다. 세 가지 실험의 주요 결과를 종합하면 다음과 같다.

<표3-4> 인공지능 사망원인통계 적용 실험 결과

실험	학습데이터	입력항목 (X)	출력항목 (Y)	평가데이터 (비교대상)	일치율	Weighed F1	Macro F1
1	사망진단서 사망원인코드, 원사인, 행정자료 보완 원사인	사망진단서 사망원인 코드	원사인 (ICD-10코드)	최종 원사인 (행정자료 보완)	0.6265	0.5940	0.1503
2	사망진단서 사망원인코드 (행정자료 보완 전후 원사인이 동일한 경우)	사망진단서 사망원인 코드	원사인 (ICD-10코드)	잠정 원사인 (사망진단서)	0.9535	0.9516	0.4996
3	용어매핑 테이블, ICD-10 코드리스트	사망진단서 사망원인 항목	개별사망 원인코드 (ICD-10)	기존방식 수행결과	0.7951	0.7741	0.9250

이상의 선행연구는 사망원인통계에 AI 기반 분류 방법론을 적용하여, 질병사망 영역에서 일정 수준 이상의 분류 가능성을 실증적으로 확인하였다는 점에서 의의가 있다. 다만 연구 범위가 사망진단서 입력 정보에 한정됨에 따라, 행정자료 연계를 통한 예측 범위 확장, 서술형 텍스트에서의 약물 성분·자살 위험요인 등 세분화된 정보 추출, ICD-11 전환 대응 등은 다루어지지 못하였다. 한편, 이러한 과제 영역에서는 국외에서 다양한 데이터 원천과 알고리즘을 활용한 접근이 축적되고 있어, 국내 적용 방향을 모색하기 위한 참고 기반을 제공한다.

제2절 국외 사망원인통계 관련 연구 동향

1. 사망진단서 기반 ICD-10 자동 코드화

Duarte et al.(2018)은 포르투갈 보건총국(DGS)의 사망정보시스템(SICO)에 등록된 사망진단서, 임상정보지(clinical bulletin), 부검 보고서의 자유텍스트로부터 ICD-10 원사인 코드를 자동으로 분류하는 심층신경망 모델을 제안하였다.

포르투갈에서는 의사가 SICO에 사망진단서를 전자 형식으로 제출하며, 사망진단서의 Part I(직접사인~원사인의 인과 사슬, 최대 4개 필드)과 Part II(기여 병태), 그리고 임상정보지와 부검 보고서의 자유텍스트를 전문 코더가 수작업으로 ICD-10 코드를 부여한다. 이 수작업 코딩은 비용이 크고 시간이 소요되어, 1년치 사망자료의 처리에 약 9개월이 소요되는 것으로 보고되었다.

모델의 구조는 워드 임베딩, 2단계 양방향 게이트 순환 유닛(bi-directional Gated Recurrent Units, GRU), 신경 어텐션(neural attention)을 결합한 계층적 심층신경망이다. 1단계(word-level)에서는 사망진단서의 개별 기재란 내 단어 시퀀스로부터 필드 표현을 생성하고, 2단계(field-level)에서는 사망진단서·임상정보지·부검 보고서의 필드 표현을 다시 결합하여 전체 입력의 표현을 산출한다. 각 단계에 어텐션 메커니즘을 적용하여 분류에 기여하는 단어와 필드에 가중치를 부여하였으며, 이를 통해 모델의 해석 가능성을 확보하였다. 출력 노드는 ICD-10 세분류(full-code) 분류를 위한 소프트맥스, 블록 수준 분류를 위한 소프트맥스, 기여 병태의 다중 코드 분류를 위한 시그모이드의 3개로 구성되며, ICD-10 코드 간 동시 출현(co-occurrence) 정보를 활용한 비음수 행렬 분해(Non-negative Matrix Factorization, NMF) 기반 가중치 초기화 기법이 적용되었다.

실험에는 2013~2015년 사망진단서 121,536건이 사용되었으며, 신생아·주산기 사망은 제외되었다. 학습 75%·평가 25%로 분할하여 분류 정확도를 측정한 결과, ICD-10 대분류(chapter, 19개 클래스) 수준 89.2%, 블록(611개 클래스) 수준 81.2%, 세분류(full-code, 1,418개 클래스) 수준 75.9%의 정확도를 달성하였다. 전체 사망의 56.6%를 차지하는 신생물(chapter II)과 순환계통 질환(chapter IX)에서는 F1-score가 각각 96.4%, 92.8%에 이르렀다. 선형 분류기(SVM) 기반 베이스라인과 비교한 결과, 세분류 수준에서 SVM의 정확도 67.4% 대비 약 8.5%p 이상의 성능 향상이 확인되었다.

나아가 2016년 사망자료 86,071건에 대한 외부 검증에서도 대분류 89.1%, 블록

80.6%, 세분류 75.9%로 시간 경과에 따른 일반화 성능이 유지되었다. 이 연구는 주간 단위 시계열 분석을 통해 허혈성심장질환, 뇌혈관질환, 인플루엔자·폐렴 등 특정 사인의 발생 추이를 자동 분류 결과만으로도 추적할 수 있음을 보여주었으며, 이를 기반으로 포르투갈 보건부의 근실시간 사망원인 감시 플랫폼에 실제 적용되었다.

다만 본 모델은 포르투갈어 사망진단서에 특화되어 있으며, 한국어 기반 사망진단서의 기재 방식 및 용어 체계와는 언어적 차이가 존재한다. 또한 세분류 수준(1,418개 코드)에서의 macro-averaged F1-score는 27.0%로, 빈도가 낮은 사인에 대한 분류 성능에는 한계가 있다.

2. 행정자료 기반 사망원인 예측

Kim et al.(2021)은 사망진단서가 아닌 행정자료만으로 사망원인을 예측할 수 있는 기계학습 모델을 개발하였다. 기존 임상 연구에서는 사망 여부는 확인할 수 있더라도 구체적인 사망원인 정보에 접근하기 어려운 경우가 많아, 질병군별 사망률 분석에 제약이 존재하였다. 이 연구는 건강보험료 청구 자료와 의료 이용 기록 등 관찰 데이터(observational data)를 활용하여 사망원인을 예측하는 스택킹 앙상블(stack ensemble) 모델을 제안하였다.

모델은 2단계 구조로 설계되었다. 1단계 기초 학습기(base learner)로 라소 로지스틱 회귀(Lasso Logistic Regression)와 그래디언트 부스팅 머신(Gradient Boosting Machine)을 적용하여, 생존 여부 및 한국의 8대 주요 사망원인(악성신생물, 뇌혈관질환, 허혈성심장질환, 폐렴, 만성하기도질환, 간질환, 당뇨병, 고혈압성질환)에 대한 개별 예측값을 산출한다. 2단계 메타 학습기(meta-learner)는 이들 기초 학습기의 예측 결과 18개를 입력 변수로 받아 최종 사망원인을 판정하며, 랜덤 포레스트, GBM, XGBoost를 비교하여 AUROC 기준으로 최적 알고리즘을 선택하였다.

모델 개발에는 한국 국민건강보험공단 표본코호트(NHIS-NSC) 174,747명의 자료가 활용되었으며, 아주대학교병원 전자의무기록(AUSOM) 729,065명의 자료로 외부 검증을 수행하였다. 모든 데이터베이스는 OMOP 공통데이터모델(CDM) 형식으로 변환하여 분석하였다. 개발 데이터 기준 60일 이내 사망원인 예측의 일반화 AUROC는 0.9511, 외부 검증 AUROC는 0.8887로 나타났다. 나아가 사망원인 정보가 없는 미국의 3개 청구자료 데이터베이스(MDCD, MDCR, Optum DOD)에 모델을 적용하여 사망원인을 추정하는 활용 사례도 제시하였다. 다만 예측 대상이 8대 주요 사망원인과 기타로 한정되어 있어 ICD-10 세분류 수준의 원사인 분류와는 과제의 성격이 다르며, 외부 검증에서 AUROC가 하락한 점은 데이터 원천 간 이질성에 따른 일반화 한계를 시사한다.

3. 서술형 사망기록 기반 약물 관련 정보 추출

사망원인통계에서 약물 관련 사망의 감시는 ICD-10 코드만으로는 구체적인 관여 약물이나 약물 조합 패턴을 파악하기 어렵다는 구조적 한계를 가진다. 이에 따라 사망기록의 서술형 텍스트에서 약물 정보를 직접 추출하려는 자연어 처리 기반 연구가 국외에서 수행되고 있다.

가. 개체명 인식 기반 약물 식별

Ward et al.(2023)은 약물 과다복용 사망건의 사망진단서 텍스트에서 관여 약물을 자동 식별하기 위한 심층학습 기반 개체명 인식(Named Entity Recognition, NER) 모델을 개발하였다. 기존의 룩업(Look-up) 테이블 기반 감시 방식은 신규 약물이나 오타자에 대응하기 어렵고 지속적 갱신이 필요하다는 한계가 있었다.

연구에서는 미국 켄터키주 약물 과다복용 사망 감시체계의 2014~2019년 사망진단서 8,146건(토큰 수 95,566개)을 대상으로 하였다. 모델은 양방향 장단기 기억 네트워크-조건부 랜덤 필드(BiLSTM-CRF) 구조를 채택하였으며, 문맥에 따라 동일 단어에 다른 벡터 표현을 부여하는 두 가지 모델을 비교하였다. 평가 결과 Flair 임베딩 기반 모델의 F1-score는 99.13%로, 기존 룩업(Look-up) 테이블에 등재되지 않은 약물명의 오타자나 신규 물질도 인식할 수 있는 수준의 성능을 보였다.

나. 다중 알고리즘 비교 기반 약물 분류

Goodman-Meza et al.(2022)은 검시관·법의학관 보고서의 서술형 텍스트에서 과다복용 사망에 관여한 약물을 자동 분류하는 자연어 처리·기계학습 모델을 개발·검증하였다. 미국 내 10개 관할지역에서 2020년 수집된 사망기록 35,433건을 대상으로 하였다.

이 연구의 특징은 단일 알고리즘이 아닌 다수의 방법론을 체계적으로 비교한 점에 있다. 특성 추출 방법으로 TF-IDF, GloVe 워드 임베딩, CUI 임베딩의 세 가지를 적용하고, 분류기로는 로지스틱 회귀, 나이브 베이즈, 랜덤 포레스트, XGBoost, K-최근접 이웃, 서포트 벡터 머신, 단층 신경망의 7종을 조합하여 각 약물별 이진 분류 성능을 평가하였다. 분류 대상은 오피오이드, 펜타닐, 알코올, 코카인, 메스암페타민, 헤로인, 처방 오피오이드, 벤조디아제핀 등이었으며, 벤조디아제핀을 제외한 대부분의 약물에서 거의 완벽에 가까운 분류 성능이 보고되었다. 이 연구는 향후 약물 관련 사망 감시체에 자연어 처리를 도입할 때 알고리즘 선택의 실증적 근거를 제공하였다.

4. 자살 위험요인 분류

Wang et al.(2023)은 미국의 질병통제예방센터(CDC)가 운영하는 국가 폭력사망 보고시스템(NVDRS)에 축적된 자살 사건 서술 텍스트로부터, 건강의 사회적 결정요인(SDoH)과 관련된 자살 정황(circumstance)과 위기(crisis)를 자동 분류하는 자연어 처리 모델을 개발하였다. 여기서 정황은 자살에 기여한 것으로 판단되는 선행 사건 전반을 의미하며, 위기는 자살 전 2주 이내에 발생한 급성 사건(acute event)으로 정의된다.

NVDRS에는 법의관·검시관(CME) 보고서와 수사관(LE) 보고서의 서술형 텍스트가 포함되어 있으며, 각 사건에 대해 자살에 기여한 정황 변수가 코딩되어 있다. 그러나 정황 변수의 20% 이상이 2013년 이후에야 추가되어 그 이전 데이터에는 코딩이 누락되어 있으며, 수작업 코딩의 비용과 한계도 존재한다. 연구에서는 2003~2019년의 자살 사건 267,804건을 대상으로, 기존 SDoH 온톨로지(SDoHO)를 자살 맥락에 맞게 적응시킨 Suicide-SDoHO를 마련하였다. 이 온톨로지는 57개 세분화 정황 변수를 계층적으로 구조화한 것으로, 정신건강 문제, 대인관계 문제, 경제적 어려움, 약물 남용, 법적 문제 등 16개 핵심 변수에 대한 분류 모델을 개발하였다.

분류 모델은 생의학 도메인에 특화된 사전훈련 언어모델인 BioBERT 기반의 트랜스포머 모델을 활용하였다. CME 보고서와 LE 보고서를 연결하여 입력으로 사용하며, 각 정황 변수의 존재 여부를 다중 레이블 분류 방식으로 판정한다. 정황과 위기는 동일한 변수라 하더라도 시간적 맥락의 판별이 요구되므로 별도의 모델로 학습하였다. 2013년 이후 데이터(133,254건)를 학습·평가에 활용하고, 2013년 이전 데이터(103,352건)에는 모델을 소급 적용하여 미코딩 위기 변수를 추정하였다. 평가 결과 정황 분류의 weighted AUC는 0.966(F1 0.837), 위기 분류의 weighted AUC는 0.942(F1 0.563)로, 급성 시간적 맥락을 텍스트에서 판별하는 위기 분류가 정황 분류에 비해 난도가 높은 것으로 확인되었다. 소급 적용을 통한 추세 분석에서는 경제적 안정성 관련 위기 비율이 2007~2009년 대침체 시기에 유의미하게 상승하는 패턴이 관측되어, 거시 경제 변동과 자살 위기 간의 시계열적 연관이 확인되었다.

다만 NVDRS의 서술 텍스트는 미국의 법의·수사 체계에 기반한 자료로서, 국내의 경찰 변사자료나 국립과학수사연구원 부검기록과는 자료 구조와 기재 방식에 차이가 존재한다. 또한 주(state)별 문서화 관행 차이에 따른 성능 편차 존재(최고 F1 0.906 버지니아 ~ 최저 F1 0.719 하와이)한다는 한계가 있었다.

제3절 선행연구의 시사점

국내의 선행연구는 사망원인통계에 AI 기반 분류 방법론을 적용하여, 기존 전문가 시스템 중심의 원사인 선정 체계를 데이터 기반 방식으로 보완할 수 있는 가능성을 실증적으로 제시하였다는 점에서 의의가 있다. 특히 국내 선행연구는 사망진단서 입력 정보만을 활용한 분류 실험을 통해 질병사망 영역에서 일정 수준 이상의 분류 가능성을 확인하였고, Duarte et al.(2018)의 심충신경망 모델은 포르투갈 국가 통계 시스템에 실제 적용되어 실시간 사망원인 감시에 활용되고 있다. 이러한 사례는 인공지능의 도입 단계와 주의해야 할 점을 제시하는 데 기여하였다.

첫째, 사망원인통계는 전체적인 사망자 규모를 파악하는 것도 중요하지만, 세부적인 사망원인의 추세와 희귀질환에 의한 사망자 관리도 중요하다. 사망원인통계는 약 18,000개에 달하는 세분화된 분류체계를 가지며, 국내 선행연구에서 가장 높은 일치율(95.35%)을 보인 잠정 원사인과의 비교 실험에서도, 평가 대상 1,026개 코드 중 F1-score가 0인 코드가 410개에 달하였다는 점, Duarte et al.의 세분류 수준 macro F1-score가 27.0%에 그쳤다는 점은 단순한 학습 기반 모델만으로는 희귀 질환이나 외인사 사례에 대한 분류 정확도를 충분히 확보하기 어렵다는 한계를 보여준다. 일부 희귀질환 또는 소수 사망자가 발생하는 사망원인의 경우에도 정확하게 분류되어야 하며, 이를 위해서는 코드화 모델의 세분류 성능을 보완하는 별도의 접근이 필요하다.

둘째, 사망진단서 정보만으로는 최종 원사인 예측에 구조적 한계가 존재한다. 국내 선행연구의 사망진단서만을 학습한 AI의 예측 결과를 행정자료까지 보완한 최종 원사인과 비교한 실험에서 일치율이 62.65%에 그쳤다는 점은, 사망진단서에 기재되지 않은 건강보험 진료 기록 등의 정보가 원사인 결정에 상당한 영향을 미침을 보여준다. Kim et al.(2021)은 행정자료만으로도 주요 사망원인 예측이 가능함을 실증하였으며, OMOP CDM을 통한 다기관 적용 가능성까지 제시하였다. 다만 예측 대상이 8대 사망원인으로 한정되어 있어, 세분류 수준의 원사인 분류를 위해서는 행정자료의 연계 범위와 활용 방식을 보다 정밀하게 설계할 필요가 있다.

셋째, 사망진단서나 행정자료의 정형 코드만으로는 파악하기 어려운 정보가 서술형 텍스트에 내포되어 있다. 약물 관련 사망의 경우 ICD-10 코드만으로는 구체적 관여 약물이나 약물 조합을 식별하기 어렵고, 자살 사망의 경우에는 사망원인 분류를 넘어 기여한 위험요인을 파악해야 정책적 활용 가치가 확보된다. Ward et al.(2023)과 Goodman-Meza et al.(2022)은 서술형 사망기록에서 약물 관련 정보를 추출하는 방법론적 기반을 제시하였으며, Wang et al.(2023)은 자살 사건 서술로부터 사회적 결정요인과 위기 변수를 BioBERT 기반으로 분류하여 정책 수요에 대응하는 분석까지 확장될

수 있음을 보여주었다.

다만 이들 국외 연구는 각국의 사망등록 체계와 법의·수사 기록 구조에 기반하고 있어, 국내의 사망진단서 기재 방식, 경찰·해경 변사자료의 서술 구조, 국립과학수사연구원 부검기록의 형식 등과는 차이가 존재한다. 또한 ICD-11 도입을 직접적으로 고려한 분류체계 검증이나 행정자료를 활용한 잠정 원사인 예측까지는 국내 선행연구의 범위에 포함되지 못하였다. 아울러 사망진단서에 기재된 사망원인 정보를 이용하여 잠정 원사인을 선정하는 과정은 WHO ICD에서 권고한 규칙을 적용해야 하기 때문에, 인공지능 학습에 의한 잠정 원사인은 국제표준에 부합하지 않아 통계 신뢰성에 문제가 발생할 수 있다는 과제도 남아 있다.

이상의 선행연구 결과는 사망원인통계에 대한 AI 적용이 기술적으로 가능함을 확인하는 기초 연구로서 중요한 의미를 가지며, AI 예측 결과에 대한 사후 검증 체계와 전문가 판단을 결합한 하이브리드 접근이 필수적이라는 점, 전면 자동화보다는 현행 사망원인통계 작성 체계 내 특정 단계의 보조적·점진적 AI 도입을 목표로 설정하는 것이 바람직하다는 점을 공통적으로 시사한다.

본 연구는 이러한 선행연구의 성과와 한계를 바탕으로, 사망진단서 사인 정보의 코드화 정확도 평가, ICD-11 전환 대응, 행정자료 연계를 활용한 잠정 원사인 예측, 약물 성분 추출 및 자살 위험요인 분류 등 보다 확장된 영역을 대상으로 AI 활용 방안을 검토한다는 점에서 기존 연구와 차별성을 가진다.

제 4 장

연구 데이터 및 분석 방법론

제1절 연구 기초자료 특성 및 현황

1. 사망진단서

가. 작성기준

사망진단서에 기재되는 사망원인은 「의료법 시행규칙」 제9조(진단서의 기재사항) 제1항 제2호에 따라 한국표준질병사인분류(KCD)를 준수하여 작성하여야 한다. 세계보건기구(WHO)가 제시한 사망원인 작성 지침과 한국표준질병사인분류에 따르면 사망원인 기재란은 크게 두 부분으로 구분되며, 국제적으로는 1부(Part 1)와 2부(Part 2)로 명명된다.

1부에서는 사망에 이르게 한 직접적인 인과관계를 순차적으로 기재한다. 즉, 직접사인을 첫 번째 칸에 기재하고, 그 원인이 되는 선행사인을 인과관계에 따라 아래로 이어서 작성하며, 근원 선행사인은 가장 아래 칸에 기재한다. 반면, 사망에 영향을 미쳤으나 1부에 기재된 사인과 직접적인 인과관계가 없는 중요한 병태는 2부에 기재한다.

사망원인 각 기재란의 작성 지침은 다음과 같다. 첫째, ‘심정지’나 ‘호흡정지’와 같이 사망에 수반되는 증상이나 징후는 기재하지 않는다. ‘심정지’와 ‘호흡정지’ 등은 사망원인으로서의 의미가 없으므로 사망진단서의 기재란을 이러한 표현으로 채우지 말아야 한다.

둘째, 약어 사용을 배제하고 진단명은 완전하고 명확하게 기재하며, 추측성 표현을 사용하지 않는다. 약어는 의료기관이나 전문분야에 따라 다르게 해석될 수 있어 오인의 소지가 있으므로 사용을 지양해야 하며, 명확한 진단명 기재는 보다 정확한 사망원인 규명에 기여한다.

셋째, 현상에 대한 기술을 피하고, 한 줄에 두 개 이상의 사인을 기재하지 않는다. 한 줄에 두 개 이상의 사인을 기재할 경우 사망 인과관계가 복수로 해석될 수 있어 정확한 사망원인 규명이 어려워지고, 통계 자료로서의 활용 가치도 저하된다.

따라서 한 줄에는 하나의 사인만 기재하고, 인과관계에 따라 다른 줄에 작성하거나 직접적인 관련이 없는 경우에는 ‘그 밖의 신체상황’란에 기재하여야 한다. 하지만 실제 의료현장에서는 한 줄에 여러 가지 병태명을 기재하는 경우도 많이 존재한다.

나. 작성현황

2024년에 접수된 사망진단서는 약 37.5만 건으로, 2024년 사망자 수(358,569명)와 과거 발생 사망의 지연 신고분을 포함한 수치이다. 사망원인 KCD 코드화 작업 단위인 사망진단서에 기재된 전체 사망원인 줄은 약 68.4만 줄로 사망진단서 1건당 평균 1.8개 줄의 사망원인이 기재되었다.

이 중 직접사인 1줄만 기재한 경우는 전체 사망신고의 51.7%를 차지하였다. 반면 나머지 절반가량의 사망진단서는 둘 이상의 사인이 기재되어 있어 원사인 선정 과정에서 인과관계 판단이 필요한 사례에 해당한다.

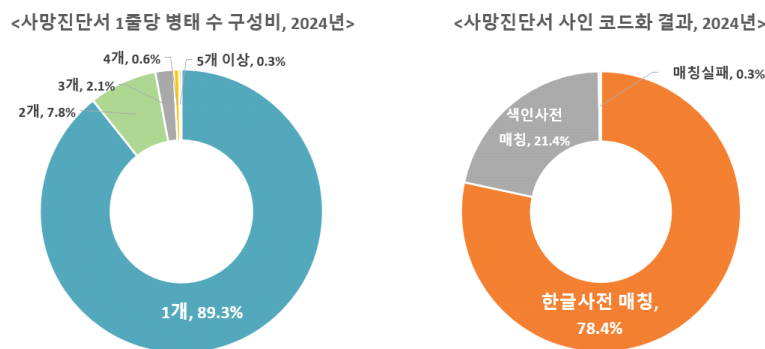
한 줄에 1개 사인이 기재된 경우는 전체 사망진단서 줄의 89.3%로 대부분을 차지하였으며, 2개 병태는 7.8%, 3개 병태는 2.1%, 4개 이상은 0.9%였다.

다. 코드화 현황 및 한계

사망진단서에 기재된 한글 자연어가 얼마나 정확하게 코드로 매칭되었는지를 확인하기 위해 한글사전과 색인사전 매칭 결과를 분석한 결과, 한글사전으로 완전하게 매칭된 건은 전체 68.4만 줄 중 53.7만 줄로 78.4% 수준이었다.

그리고 한글사전으로 매칭되지 않은 코드를 다시 색인사전으로 매칭되었는지를 확인한 결과, 14.6만 줄로 전체 중 21.4%였으며, 나머지 0.3%는 매칭이 실패하였다.

<그림4-1> 사망진단서 기재 사인 현황, 2024년 신고 기준



WHO지침에 따라 사망진단서 한 줄에 1개의 사인만 기재되어야 한다. 그러나 실제로는 한 줄에 여러 사망원인이 기재된 사례가 전체의 10.7%를 차지하며, 이는 한글사전 기반 매칭을 어렵게 하는 주요 요인으로 작용한다. 그 이유는 약 1만 4천 가지에 달하는 사망원인 조합을 사전에 모두 구축하는 것이 현실적으로 불가능하기 때문이다.

이러한 구조적 한계는 21.4%에 달하는 불완전한 KCD 코드화로 이어진다. 더 나아가, 현재 한글사전이 구축되어 있지 않은 ICD-11 체계에서는 코드화 자체가 어려운 실정이다. 기존 ICD-10 기준의 한글사전은 매년 누적·확충되어 온 결과물로, 이를 ICD-11 기준으로 새롭게 구축하는 것은 상당한 업무 부담으로 작용한다.

2. 잠정 원사인 정보

가. 개념 및 선정절차

잠정 원사인은 행정자료 보정 이전 단계에서 사망진단서에 기재된 정보만을 활용하여 선정한 개인별 사망의 근원 원인을 의미한다. 이는 사망진단서 외에 각종 행정자료를 연계·검토하여 보정한 최종 원사인과는 구별되는 개념이다. 사망진단서에 기재된 다수의 사망원인 중 잠정 원사인을 선정하는 방법은 앞서 제2장에서 기술한 바와 같이 세계보건기구(WHO)의 ICD-10 제2권 지침에 따른다.

행정자료 연계가 불가능한 다수의 외국에서는 이처럼 사망진단서 정보만을 기반으로 산출한 원사인을 그대로 사망원인통계로 공표한다. 반면 우리나라는 고령자 사망이 증가하고 있고 불명확하게 작성된 사망진단서 발급 사례가 다수 발견됨에 따라 사망진단서 정보만으로는 사망원인을 충분히 규명하기 어렵다는 한계가 존재한다. 이에 약 20종의 행정자료를 추가로 검토하여 최종 원사인을 도출하고 있다.

원사인 선정 과정을 이해하기 위해서는 WHO가 제시한 원사인 개념과 지침에 대한 이해가 필요하다. WHO는 국제적으로 사망원인통계 작성의 기준을 원사인으로 규정하고 있으며, 사망을 예방하는 관점에서 치료 또는 개입 시점을 식별할 수 있도록 원사인을 정의하고 있다. 이에 따른 공식적인 원사인의 정의는 다음과 같다.

- (a) 직접적으로 사망에 이르게 한 연속적 병적 사건들의 시작점이 된 질병 또는 손상
- (b) 치명적인 손상을 일으킨 사고 또는 폭력의 상황

아울러 WHO는 원사인 선정의 핵심 목적을 국제 비교 가능성 확보에 두고 있다. 이는 사망원인통계 작성 시 국가별로 상이한 기준이나 절차를 적용할 경우, 국가 간 비교가 불가능해지는 문제 때문이다. 따라서 원사인 선정은 인공지능이나 통계적 예측 결과에 따라 임의로 결정될 수 있는 영역이 아니라, WHO 지침에 근거하여 일관되게 수행되어야 하는 절차적 영역이라 할 수 있다.

나. 잠정, 최종 원사인 일치정도

WHO 지침에 따라 선정된 잠정 원사인은 사망진단서 정보만을 반영하므로, 행정자료를 활용해 보완한 최종 원사인에 비해 정확도가 상대적으로 낮다. 실제로 잠정 원사인과 최종 원사인을 비교한 결과, 전체 원사인 코드 기준으로는 35.9만 명 중 17.8만 명이 일치하여 일치율은 49.6% 수준에 그쳤다.

일치 수준을 분류체계별로 살펴보면, KCD 소분류 수준(3자리)에서 일치율은 67.3%, 사망원인통계의 제표기준인 104항목 분류 기준에서는 75.8%, KCD 대분류(1자리) 기준에서는 79.7%였다. 이는 분류수준이 상위체계로 갈수록 동일한 범주로 분류될 가능성이 높아지기 때문이다.

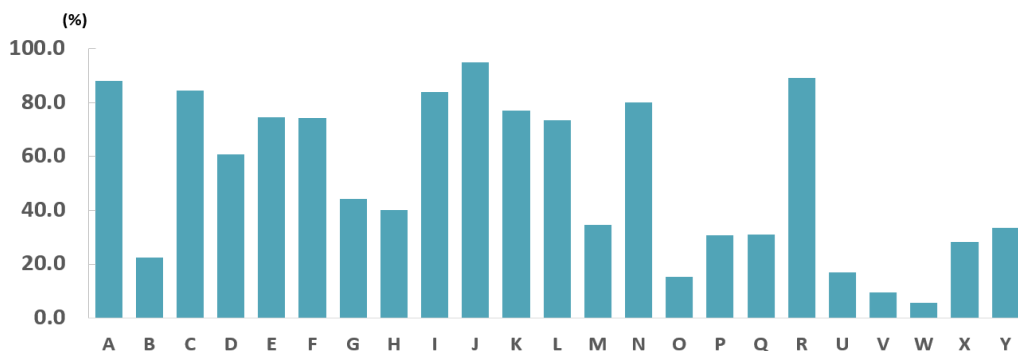
다만, 대분류 수준에서조차 약 20%의 사례는 일치하지 않았으며, 이러한 불일치는 질병군 또는 외인 분류 자체가 달라지는 것을 의미하므로, 사망원인통계 활용 측면에서 상당한 영향을 미칠 수 있다.

<표4-1> 잠정 원사인과 최종 원사인 일치 정도, 2024년

구분	전체	전체 일치	소분류 일치	대분류 일치	104항목 일치
사망자 수(명)	358,569	177,986	241,238	285,907	271,347
일치율(%)	100.0	49.6	67.3	79.7	75.8

잠정 원사인과 최종 원사인의 불일치 양상은 질환군에 따라 차이를 보였다. KCD 대분류별로 일치 정도를 살펴보면, J(호흡계통 질환)가 95.2%로 가장 높았고, 뒤를 이어 R(불명확한 병태) 89.3%, A(감염성 질환) 88.2%로 나타났다. 그러나 V~Y(외인사)의 경우 일치율이 상대적으로 낮았고, B(감염성 질환)와 O(임신, 출산, 산후기 질환)에서도 비교적 낮은 일치율을 보였다.

<그림4-2> 잠정 원사인과 최종 원사인의 KCD 대분류별 일치율, 2024년



다. 104항목별 집계값 비교

2024년 사망자 수 기준 상위 30개 질환의 104항목 분류 기준의 일치율을 살펴보면, 폐혈증(95.9%), 폐렴(95.7%), 고혈압성 질환(91.0%) 등 사망진단서에 흔히 기재되는 사망원인의 일치율은 높고, 추락사고(2.5%), 운수사고(10.8%), 코로나19(17.8%) 등 외인사 및 특정 감염병의 일치율은 낮았다. 질병 중에서는 알츠하이머병(29.5%)과 간 질환(52.9%)의 일치율이 낮게 나타났다.

다만 일치율만으로 통계서비스의 정확도를 판단하기는 어려우며, 실제 통계서비스에서는 집계값을 함께 비교해야 한다. 특정 사망원인의 잠정 원사인 집계 사망자 수가 많은 경우, 최종 원사인과 일치하는 사례도 증가하여 일치율이 높게 나타날 수 있으나, 실제 사망원인별 집계 사망자 수는 최종 원사인과 상당한 차이를 보일 수 있다. 이에 104항목 분류 기준의 사망자 수 집계값을 비교하였다.

<표4-2> 104항목 분류별 사망자 수 일치 정도(사망자 수 상위 30개 원인), 2024년

사망원인	전체 사망자 수 (명)	일치 사망자 수 (명)	일치율 (%)
달리 분류되지 않은 증상, 징후 (R00-R99)	41,787	37,212	89.1
폐렴 (J12-J18)	30,103	28,804	95.7
뇌혈관 질환 (I60-I69)	24,612	17,814	72.4
기관, 기관지 및 폐의 악성신생물 (C33-C34)	19,401	16,818	86.7
기타 심장 질환 (I26-I51)	19,277	17,083	88.6
고의적 자해(자살) (X60-X84)	14,872	4,550	30.6
허혈성 심장 질환 (I20-I25)	14,262	10,591	74.3
알츠하이머병 (G30)	12,223	3,610	29.5
나머지 악성신생물 (Re. C00-C97)	12,096	8,736	72.2
당뇨병 (E10-E14)	11,064	8,088	73.1
간 및 간내 담관의 악성신생물 (C22)	10,432	7,035	67.4
나머지 비뇨생식계통 질환 (N17-N98)	10,349	8,266	79.9
결장, 직장 및 항문의 악성신생물 (C18-C21)	9,683	7,257	74.9
나머지 호흡계통 질환 (Re. J00-J98)	8,917	6,250	70.1
나머지 신경계통 질환 (Re. G00-G98)	8,493	5,318	62.6
고혈압성 질환 (I10-I13)	8,229	7,492	91.0
췌장의 악성신생물 (C25)	8,190	7,052	86.1
간 질환 (K70-K76)	7,787	4,118	52.9
폐혈증 (A40-A41)	7,730	7,411	95.9
모든 기타 외인 (Re. V01-Y89, U12)	7,371	2,991	40.6
위의 악성신생물 (C16)	7,213	6,068	84.1
만성 하기도 질환 (J40-J47)	6,348	3,987	62.8

<표4-2> 104항목 분류별 사망자 수 일치 정도(사망자 수 상위 30개 원인), 2024년(계속)

사망원인	전체 사망자 수 (명)	일치 사망자 수 (명)	일치율 (%)
나머지 소화계통 질환 (Re. K00-K92)	6,320	4,776	75.6
운수사고 (V01-V99)	3,451	371	10.8
코로나19 (U07.1, U07.2, U10)	3,208	572	17.8
유방의 악성신생물 (C50)	2,970	2,495	84.0
나머지 정신 및 행동장애 (Re. F01-F99)	2,946	2,330	79.1
낙상(추락) (W00-W19)	2,856	71	2.5
전립선의 악성신생물 (C61)	2,839	1,926	67.8
비호지킨 림프종 (C82-C86)	2,343	1,490	63.6

104항목 분류별 사망자 수 집계값 비교 결과, 최종 원사인 기준 높은 일치율을 보였던 사망원인에서도 집계값 차이가 발생하는 사례가 확인되었다.

먼저 감염성 질환을 살펴보면, 호흡기 결핵과 기타 결핵의 잠정 원사인 사망자는 사망원인통계 최종 원사인 대비 각각 69.2%, 47.8%였다. 이에 비해 패혈증은 161.6%로 잠정 원사인 기준의 사망자 수가 최종 원사인보다 크게 집계되었다. 이는 패혈증이 다양한 질환의 진행 과정에서 발생하는 합병증으로, 사망진단서에 빈번하게 기재되기 때문으로 해석된다. 인체면역결핍바이러스병(HIV)과 코로나19는 각각 16.4%, 17.9%에 불과하여, 사망진단서 정보만으로는 해당 질환의 원사인을 정확하게 선정하는 데 한계가 있음을 보여준다.

악성신생물(암)로 잠정 원사인 사망자 수는 최종 원사인 대비 대부분 70% 이상으로 나타났다. 다만, 구강암(62.6%), 수막·뇌 및 기타 중추신경계통암(47.5%)·비호지킨림프종(66.3%)은 낮은 포괄성을 보인다. 이는 악성신생물의 질병 특성상 질병에 대한 위험도가 높아 사망진단서에 비교적 명확하게 기재되는 경우가 많기 때문으로 보인다.

다만 대부분의 암종에서 90% 이상의 포괄성이 나타나지 않는 것은 전이단계가 고려되지 않아 원발부위정보가 누락된 영향으로 보인다. 악성신생물의 원발/속발 정보는 연간 보정에 포함된 국립암센터의 암등록정보의 검토가 필수적으로 이루어져야 완전성 있는 통계가 작성될 수 있다.

내분비 질환 중 대표적인 질환인 당뇨병의 경우 최종 원사인 사망자 수 대비 잠정 원사인 사망자 수가 77.5% 수준으로 나타났다. 당뇨병은 대표적인 합병증으로 사망진단서 기재 빈도가 높은 편에 속함에도 잠정 원사인의 포괄성이 상대적으로 낮게 나타난다.

이에 최종 원사인이 당뇨병인 사망자의 잠정 원사인의 세부적인 분포를 확인한 결과, 잠정 원사인 오분류 건 중 31.3%는 비뇨생식계통 질환, 25.5%는 순환계통 질환, 21.4%는 불명확한 증상(R00-R99)으로 분류되었다. 비뇨생식계통 질환 중에는 신부전

(N17~N19), 순환계통 질환 중에는 심부전(I50)이 많았다. 이는 사망진단서에 당뇨병보다 합병증이나 증상이 우선 기재되는 경우가 많아 당뇨병이 원사인으로 과소 집계된 결과로 해석된다.

정신 및 행동장애는 다른 사망원인과 달리 잠정 원사인 사망자 수(7,208명)가 최종 원사인 사망자 수(3,872명)보다 더 많았다. 신경계통의 질환 중 대표적인 알츠하이머병은 잠정 원사인 사망자 수가 최종 원사인 대비 30.8%에 불과하였다. 이는 정신 및 행동장애로 인한 잠정 원사인 사망자 수가 과다추정된 것과 관련이 있는 것으로 보인다.

순환계통 질환을 살펴보면, 심장 질환의 유형에 따라 차이가 나타남을 알 수 있다. 협심증 등의 허혈성 심장질환은 잠정 원사인 사망자 수가 77.9%로 과소추정되었으나 심부전 등의 기타 심장질환은 잠정 원사인 사망자 수가 최종 원사인 사망자 수의 131.8% 수준으로 과다추정 되었다. 이는 심전도 검사 등으로 알 수 있는 허혈성 심장질환보다 심부전 등의 증상 및 징후가 사망진단서에 빈번히 기재되기 때문이다.

호흡계통의 질환 중 법정감염병인 인플루엔자는 잠정 원사인 사망자가 최종 원사인 사망자의 58.9% 수준으로 낮게 나타났다. 고령자의 사망진단서에 흔히 기재되는 폐렴은 141.7%로 과다추정 되었다. 만성 하기도 질환은 66.2% 수준으로 정확도가 높지 않았다.

소화계통의 질환 중 대표적인 질환인 간 질환은 58.4%로 잠정 원사인에 대한 정확도가 상대적으로 낮게 나타났다. 사망자 수가 소규모인 피부 및 피하조직의 질환, 근골격계통의 질환, 임신·출산 및 산후기 질환에서도 잠정 원사인과 최종 원사인의 차이를 발견할 수 있었다.

사고사 등의 질병이환 및 사망의 외인으로 사망한 잠정 원사인 사망자 수는 20,593명으로 최종 원사인 사망자 수(29,748명)의 69.2% 수준이었다. 다만, 운수사고(11.3%), 추락(4.0%), 익사(2.3%), 중독(7.8%), 화재(0.0%), 고의적 자해(30.7%) 등 세부요인의 정확도는 현저하게 낮았다.

이렇게 104항목 분류 기준으로 사망자 수를 파악하기 위하여 사망진단서 정보만을 기반으로 잠정 원사인을 선정하여 사망원인통계를 작성할 경우, 정확도 측면에서 구조적인 한계가 나타난다. 따라서 사망진단서 정보만으로 선정되는 잠정 원사인의 한계를 보완하고, 행정자료를 활용한 시의성 있는 원사인 예측 체계를 마련하기 위해서는 AI 기반 보조 분석 방법에 대한 검토가 필요하다.

3. 건강보험 수진기록(월간)

가. 자료 특성 및 연계현황

사망원인통계의 정확도 향상을 위하여 국가데이터처는 국민건강보험공단으로부터

건강보험 수진기록을 월간 및 연간 자료로 구분하여 입수하고 있다. 건강보험 자료의 연계 대상은 사망 신고된 사망자이며, 2024년 월간 자료 기준으로 전체 사망자 35.9만 명 중 35.6만 명이 연계되어, 연계율은 99.2%에 달한다.

건강보험 수진기록은 월 단위로 입수되어 시의성이 높고, 대부분의 사망자에 대해 자료 연계가 가능하다는 점에서 사망원인 보완 자료로써 활용 가치가 크다. 다만 감염병, 암, 외인사 등 일부 질환군에서는 진단의 세부 내용이나 사망과의 직접적인 인과관계를 파악하는 데 한계가 있다.

국민건강보험공단으로부터 입수되는 주요 항목은 수진일자, 주상병, 부상병, 진단종류 등으로, 사망 시점을 기준으로 사망 전 5년간의 진료 이력이 포함된다. 이로 인해 개인별 자료의 양이 방대하므로, 사망원인 선정 시스템에서는 사인분류사의 가독성과 자료 활용도를 고려하여 건강보험 자료를 사전에 정제·가공한다.

나. 자료 정제 및 가공

우선 주상병 또는 부상병 코드가 기재되어 있지 않은 자료 중 진단종류가 ‘조제(20, 21)’이거나 ‘치과외래(05)’에 해당하는 기록과 약국 자료는 분석 대상에서 제외한다. 이러한 진료기록은 사망과의 직접적인 연관성이 낮거나, 질병의 중증도에 비해 진료비 규모가 커 사망원인 분석 과정에서 혼동을 유발할 가능성이 있기 때문이다.

다수의 의료기관 진료기록 중에서는 사망 직전의 최근 진료기록과 진료비가 가장 높은 수진기록이 입원이나 수술 등 질병 부담이 큰 의료 이용을 반영할 가능성이 높아, 사망에 영향을 미친 질환과 관련될 개연성이 크다. 이에 따라 사망 직전 첫 번째 및 두 번째 진료기록에 별도 표시를 하고, 진료비가 가장 높은 기록 역시 구분하여 표시한다. 또한 동일 질병으로 장기간 통원 치료가 이루어질 수 있음을 고려하여, 진단일 수 합계와 진단비용 합계 항목을 추가로 산출한다.

이처럼 정제·가공된 건강보험 수진기록 중에서 원사인 AI 예측에는 사망 직전 첫 번째 및 두 번째 진료기록과 최고 진료비 기록에 포함된 주상병 및 부상병 코드가 활용되었다.

다. 단순 주상병 매칭의 한계

사망 직전 건강보험 수진기록의 주상병을 원사인으로 단순히 가정한 경우, 최종 원사인과 일치하는 사망자는 전체 사망자 35.9만 명 중 8.6만 명으로 일치율은 24.0% 수준이었다. 이는 사망진단서 정보만을 활용하여 선정한 잠정 원사인의 최종 원사인 대비 일치율(49.6%)과 비교할 때 크게 낮은 수준이다.

따라서 사망 직전 건강보험 수진기록의 주상병만을 이용하여 원사인을 단순히 추정하는 방식에는 구조적 한계가 있다. 건강보험 수진기록과 사망진단서 정보를 종합적으로 연계·분석하여 사망원인을 예측하는 접근이 필요하며, 인공지능 기반 예측모형이 잠정 원사인을 보완하는 데 중요한 역할을 할 것으로 기대된다.

4. 외인사 관련 정보

가. 분류체계 및 검시환경

외인사의 한국표준질병사인분류 정식 명칭은 ‘질병이환 및 사망의 외인(External causes of morbidity and mortality)’이다. 외인사는 질병 자체가 아니라 손상이나 질병을 유발한 외부적 요인에 의해 발생한 사망 및 질병이환의 원인을 의미한다. 의도성 여부에 따라 사고, 고의적 자해, 타살, 의도 미확인으로 나뉘며, 사망종류에 따라서는 크게 운수사고, 중독, 화재, 낙상, 익수 등으로 분류된다.

우리나라 사망에 대한 검시체계는 사망 시 의사가 작성하는 사망진단서를 기반으로 함에 따라 검시관(coroner) 또는 검시의(medical examiner) 제도를 운용하는 국가에 비해 외인사 관련 정보가 제한적이다. 특히, 사망 이후 경찰 수사 또는 부검을 거쳐 사망 원인이 밝혀지는 경우, 사망진단서만으로는 사망자의 정확한 사망원인을 파악하기 어렵다.

나. 행정자료 현황

이러한 한계를 보완하기 위하여 국가데이터처에서는 경찰 및 해양경찰의 수사기록, 국립과학수사연구원의 부검기록 등을 확인하여 사망원인을 분류하고 있다. 다만, 해당 자료는 사법행정 목적으로 기록된 자료이므로 일부 항목만 코드화 되어 있고, 중요한 정보는 자연어 기반의 서술문 형태의 기록으로 구성되어 있다. 따라서 이러한 행정자료를 해석·적용하기 위해서는 사인분류사의 수작업이 필요하다.

한 해 동안 입수되는 사망자의 부검자료는 약 7,400건, 경찰청 및 해양경찰청의 변사자료는 약 14,800건이다. 그 중 서술문 형태의 필드는 시스템 가독성을 위해 글자 수를 최대 1,300자로 제한하고 있는데, 1,300자를 초과하는 기록은 52건으로 확인되었다. 이는 A4용지 한 페이지 분량으로 연간 35만 명 이상의 사망 사례를 검토하는 업무 환경을 고려할 때, 이러한 장문의 기록을 확인하고 사인분류에 반영하는 과정은 상당한 업무 부담으로 작용한다.

다. 약물중독 서술자료 처리 부담

또한, 최근 약물중독의 사망이 증가하면서 약물의 성분을 파악하고, 약물 간의 우선 순위를 판단하는 업무부담이 커지고 있다. 실제 행정자료에서는 약물의 계열(예: 벤조디아제핀), 성분(예: 클로나제팜), 제품명(예: 리보트릴), 효능(예: 항불안제) 등이 혼용 기재되어 있어 한국표준질병사인분류(KCD)에 따른 정확한 코드를 부여하기 위해서는 다양한 정보를 검색하여야 한다. 또한 여러 종류의 약물이 사망에 기여하였을 때는 ICD-10 2권 지침서의 규정에 따라 28개 코드 종류에 대한 우선순위를 적용하여야 한다. 이러한 약물중독 사망에 대한 코드화와 원사인 선정은 상당한 전문성과 업무부담이 요구되는 과정이다.

라. 추가정보 추출가능성

이러한 외인사 기초자료의 특성을 고려할 때, 사인분류 업무에서 가장 큰 부담으로 작용하는 부분은 행정자료 내 서술형 정보를 해석하고 분류 결과에 반영하는 과정이다. 또한 서술형 정보에는 기존 사망원인 분류 항목뿐만 아니라 자살 위험요인 등 정책적 활용 가치가 높은 추가 정보를 포함하고 있어, 자연어 처리 기반 AI 기술을 활용할 경우 사망원인통계의 품질 향상과 신규 지표 개발 측면에서 활용 가능성이 있다.

제2절 수행 방법론 및 성능평가 지표

1. (모델1) 사망진단서 기재 사인 코드화

가. 모델 목적 및 설계

본 모델의 목적은 기존 한글사전 및 색인사전 기반의 사인 코드화 과정에서 발생하는 관리 부담을 완화하고, 향후 ICD-11 체계 적용을 위한 기술적 기반을 마련하는 데 있다. 이를 위해 기존 코드화 절차에서 한글사전 매칭 및 색인사전 매칭 과정을 인공지능 기반 예측모델로 대체하는 것을 핵심 목표로 설정하였다.

나. 학습 및 평가데이터 구성

ICD-10 기준 모델의 학습데이터는 KCD-8 마스터파일, 2023년 사망원인통계 분류 결과, 한글사전 및 색인사전 정보를 활용하여 구축하였다. 평가데이터는 2024년 사망원인통계에서 확정된 KCD 사망원인 코드를 사용하였다.

학습에 투입된 자료는 2023년 사망진단서 654,101건이다. 이 중 한글사전과 완전히 매칭된 자료(MTCHG_CODE=4)인 532,044건을 학습데이터로 사용하였으며, 나머지 한글사전과 완전 매칭되지 않아 색인사전으로 일부 매칭된 약 18.7%는 노이즈 발생 가능성을 고려하여 제외하였다.

학습데이터의 구성을 살펴보면, 단일라벨로 이루어진 데이터가 513,609건(96.54%)으로 대부분을 차지하며, 2개 이상의 라벨로 구성된 멀티라벨은 18,435건(3.46%)이었다. 멀티라벨 자료의 비중은 크지 않으나, 두 개 이상의 라벨이 존재하는 경우 사망원인 코드의 순서와 연결어까지 함께 고려해야 하므로 예측의 복잡성이 크게 증가한다.

또한 학습자료에는 출현 빈도가 1건에 불과한 희귀 라벨(long-tail)이 261개로 전체 라벨의 18.6%를 차지하였다. 이는 일부 희귀 사인명에 대한 예측 성능 확보가 중요한 과제를 시사한다. 입력텍스트의 길이는 중앙값(p50) 기준 약 3토큰으로 매우 짧은 편이며, 90백분위수는 8토큰, 99백분위수는 15토큰 수준이다. 이는 사망진단서에 기재되는 사인명이 일반적으로 수 단어 이내의 짧은 문구로 작성됨을 보여준다.

모델의 목표 출력값은 최대 5개의 KCD 코드(Code1~Code5)이며, 코드의 순서를 고려하는 순차적 멀티라벨(prediction) 예측 문제로 설계하였다. 평가데이터는 2024년 사망진단서 684,841건 전체 자료를 활용하였으며, 이 중 한글사전과 완전히 매칭된 자료는 536,687건이다.

<표4-3> 사망진단서 학습데이터 및 평가데이터 구성

구분	전체 레코드 수	한글사전 매칭 자료
학습데이터(2023년)	654,101건	532,044건
평가데이터(2024년)	684,841건	536,687건

한편, ICD-11 기준 모델은 학습데이터가 없는 상태로 별도의 학습자료 구축 단계가 필요하였다. 이에 ICD-11 MMS 2023을 기반으로 상·하위 계층을 정비하고, 코드명을 한글화한 ICD-11 한글코드 마스터파일과 ICD-10과 ICD-11의 1:1매칭 코드들을 이용한 자체 학습자료를 구축하여 학습데이터로 활용하였다.

<표4-4> 사망진단서 기재 사인 코드화 모델 설계

입력변수	출력변수	학습데이터	평가데이터
사망진단서 기재 사인 (한글/자연어)	ICD-10코드(KCD-8)	KCD-8 마스터파일, 2023년 사망원인통계, 한글사전	'24년 사망원인통계 결과 (KCD 사인코드)
	ICD-11코드	ICD-11 한글코드 마스터 파일(계층정비, MMS 2023), 자체 학습 자료	'24년 시스템 사인코드 ICD-11 일부 매핑 결과

다. 후보 모델 검토 및 선정

비정형 텍스트 처리에 적합한 KcELECTRA와 KoELECTRA를 주요 후보 모델로 선정하였으며, 함께 KLUE/RoBERTa를 포함하여 성능과 활용 가능성을 비교·검토하였다.

KcELECTRA는 온라인 댓글과 대댓글을 기반으로 사전학습된 한국어 언어모델로, 구어체, 신조어, 오타자 등 비정형 텍스트 처리에 강점이 있다. 따라서 사망진단서에 기재되는 짧고 비정형적인 한글 사인명 분류에 적합한 특성을 가진다. 모든 입력 토큰을 활용하는 학습 방식을 적용하여 짧은 텍스트에서도 효율적으로 의미를 학습할 수 있는 장점이 있다.

KoELECTRA는 뉴스와 위키 등 정제된 한국어 말뭉치를 기반으로 학습된 사전학습 언어모델로, 문어체 중심의 텍스트 처리에 강점을 가진다. KLUE/RoBERTa는 대규모 한국어 말뭉치로 학습된 모델로, 다양한 자연어 처리 과제에서 우수한 성능을 보이나 모델 규모가 커 학습 및 추론 비용이 상대적으로 높다. 본 연구의 대상인 사망진단서 사인명은 일반 문장보다 짧고 비정형적인 특성을 가지므로, 일반적인 문어체 중심의 사전학습 모델보다 비정형 텍스트에 특화된 모델이 적합한 것으로 판단하였다.

후보모델을 종합적으로 검토한 결과, 특화분야 성격, 학습 및 추론비용, 정확도 측면에서 KcELECTRA를 최종 모델로 적합하다는 결론을 내렸다. 그리고 선정된 모델의 예측 성능을 높이기 위하여 학습자료의 전처리와 품질 개선에 따른 성능비교를 지속적으로 수행하였다. 구체적으로 학습자료의 중복 제거, 사전 정보 정비, 라벨 오류 수정 등 데이터 품질을 개선하는 과정을 반복함으로써 모델을 고도화하였다.

라. 모델구조 및 학습전략

본 연구에서 최종 채택한 KcELECTRA가 사망진단서를 처리하는 과정은 크게 두 단계로 이루어진다. 먼저 사망진단서에 기재된 사망원인 텍스트의 의미를 파악하여 하나의 수치화된 벡터로 요약한다. 그리고 이 벡터를 바탕으로 별도의 분류층이 KCD 코드 각각에 대해 독립적으로 0~1 사이의 확률값을 계산한다.

질병분류체계의 계층적 구조를 반영하기 위하여 롤업(Roll-up) 임베딩 기법을 적용하였다. 이 기법은 KCD 코드를 대분류(1자리), 중분류(3자리), 전체 코드의 세 단계로 나누어 각각의 임베딩 벡터를 생성한 후 이를 결합하는 방식이다. 이를 통해 단순히 문자열이 비슷한지를 보는 것이 아니라 질병분류체계의 상하위 계층적 정보를 모델이 학습하도록 설계하였다.

이러한 멀티라벨 구조에서는 어떤 코드를 최종 예측값으로 선택할지를 결정하는 임계값(threshold)의 설정이 성능에 직접적인 영향을 미친다. 임계값이 지나치게 낮으면 불필요한 코드가 함께 선택되는 과잉 예측이 증가하고, 반대로 지나치게 높으면 실제 정답 코드가 선택되지 않는 누락 예측이 증가한다.

최적 임계값을 결정하기 위하여 0.05부터 0.90까지 단계적으로 변화시키는 탐색을 수행하였으며, 각 임계값 별 정확도(accuracy)를 기준으로 비교하였다. 그 결과 임계값 0.5에서 가장 우수한 성능을 보여 이를 최종 임계값으로 채택하였다.

모든 코드의 예측 확률이 임계값 미만인 경우에는 어떠한 라벨도 선택되지 않는 문제가 발생할 수 있다. 이를 방지하기 위하여 보조적인 선택규칙을 두었다. 임계값을 넘는 코드가 하나도 없을 경우 확률이 가장 높은 코드 1개를 대신 채택하도록 한 것이다. 다만 이때 채택한 코드의 확률이 매우 낮다면(예를 들어 0.1 이하) 이는 예측 결과의 신뢰도가 낮다는 의미이므로 사인분류사의 추가 검토가 필요하다.

마. 평가방법 및 기준

성능 평가는 사망진단서에 기재된 사인을 기준으로 예측된 KCD 코드와 사인선정 시스템에서 확정된 코드 간의 일치 여부를 중심으로 수행하였다.

구체적인 평가지표는 Ordered accuracy, Subset accuracy를 사용하였다. Ordered accuracy는 예측 라벨과 정답 라벨의 구성뿐 아니라 순서까지 모두 일치한 경우만 정답으로 인정하는 지표로, 가장 엄격한 평가 기준이다. Subset accuracy는 정답 라벨 집합과 예측 라벨 집합의 일치 여부를 평가하지만, Ordered accuracy와 달리 라벨의 순서는 고려하지 않는다.

정량적 성능 평가와 함께 추가 수작업 검토도 병행하였다. 전체 평가 대상 약 68만 건의 추론 결과 가운데 기존 결과와 불일치하는 13만 건 중 약 1%를 무작위로 추출하여 직접 검토하였다. 이 과정에서는 모델의 예측이 정답 라벨과 일치한 경우에도 정답 라벨 자체에 오류가 존재하는지(라벨 노이즈)를 함께 확인하였으며, 모델이 오분류한 사례에 대해서는 오분류 유형과 발생 원인을 분석하였다.

분석 결과는 향후 다시 학습자료 구축 단계에 반영하여 학습 라벨을 정제·보정한 후 재학습 및 재평가를 수행할 수 있도록 정리하였다. 이처럼 자료 구축-학습-평가의 각 단계에서 오류를 지속적으로 환류하고 보완하는 순환적 품질관리 절차를 적용함으로써 「AI 데이터 품질관리 가이드」의 자가점검 기준을 최대한 준수하고자 하였다.

2. (모델2) 행정자료 연계 잠정 원사인 예측

가. 모델 목적 및 설계

본 모델의 목적은 사인선정시스템에서 분류되는 잠정 원사인의 정확도를 제고하고, 특정 사망원인에 대한 잠정 통계 서비스 제공 가능성을 검토하는 데 있다. 현재 잠정 원사인은 사망진단서 기재정보만을 기반으로 산출되므로 최종 원사인과의 일치율이 약 49.6% 수준에 머물며, 건강보험 수진기록을 자동으로 반영하는 체계가 구축되어 있지 않다는 점에서 모델 개발의 필요성이 제기되었다.

모델의 출력은 활용 목적에 따라 두 가지 체계로 구분하였다. 첫째, 전체 코드(17,193개 KCD 사망원인 세부코드) 기준 예측 결과는 사인분류사의 업무 지원을 목적으로 하며, 정확도(Accuracy)를 주요 성능지표로 사용하였다. 둘째, 104항목 분류 기준 예측 결과는 잠정 사망원인통계 서비스 제공을 목적으로 하며, F1-score와 사망자 수 추정비(예측 사망자 수 대비 실제 사망자 수)를 주요 평가 지표로 설정하였다.

연구 초기에는 전체 코드체계만을 대상으로 모델을 설계하였다. 그러나 전체 코드 가운데 약 48%가 학습자료 1건만을 포함하는 극심한 long-tail 분포를 보여 희귀 코드에 대한 학습이 사실상 어려웠으며, 정책 활용 측면에서도 안정적인 통계 산출에 한계가 있음을 확인하였다. 이에 사망원인통계 공표 기준 가운데 하나인 104항목 분류를 병행하여 예측하는 체계를 추가로 도입하였다.

또한 본 모델은 사망진단서와 건강보험 수진기록이 모두 질병코드 형태의 구조화된 자료라는 점을 고려하여, 사망진단서의 텍스트 정보와 코드 기반 정보를 함께 활용하는 혼합형 모델 구조를 적용하였다.

나. 입력데이터 구조 및 특성

모델의 입력값은 WHO 로직에 따라 산출된 잠정 원사인과 월간 건강보험 자료에 포함된 주상병 및 부상병 등의 코드화된 변수이며, 출력값은 행정자료 연계를 통해 수정된 예측 원사인 KCD 코드이다.

학습데이터는 총 17개 입력 필드로 구성하였다. 세부적으로 잠정 원사인(1개), 사망진단서 사인 텍스트(직접사인, 가의 원인, 나의 원인, 다의 원인: 4개), 건강보험 KCD 코드(주상병 4개, 부상병 4개), 진단종류(4개)로 구성된다. 진단일수와 진료비는 원사인과 과의 인과적 관련성이 명확하지 않고 학습 과정에서 불필요한 변동성을 유발할 가능성이 있어 입력 변수에서 제외하였다. 정답 라벨은 사인분류사가 행정자료를 검토하여 최종 확정된 원사인 코드를 사용하였다.

<표4-5> 모델2 입력데이터 구조

구분	필드수	설명
잠정 원사인	1	WHO 국제표준 기반 사망진단서 잠정 원사인
사망진단서 사인	4	직접사인, 가/나/다의 원인 텍스트
건강보험 KCD	8	사망 직전 1-2번째 및 최고 금액 기록의 주상병·부상병
진단종류	4	입원·외래·조제 등 진료유형

입력자료는 다음과 같은 특징을 가진다. 첫째, 사망진단서 사인 텍스트 4개 필드는 결측 비율이 매우 높다. 2024년 기준 직접사인만 기재된 사례가 전체의 49.8%를 차지하여, 나의 원인 이하 필드는 대부분 결측으로 존재한다. 그러나 이러한 결측은 단순한 누락이 아니라 추가적인 인과관계가 존재하지 않는다는 정보를 포함하므로, 결측값을 제거하지 않고 희소성을 허용하는 구조로 모델을 설계하였다.

둘째, 잠정 원사인 코드는 최종 원사인과 약 49.6%의 일치율을 보여 단일 입력 변수 가운데 가장 중요한 예측 변수로 판단하였다. 이에 잠정 원사인 코드를 단순한 범주값으로 처리하지 않고 대분류(1자리), 중분류(3자리), 세부 코드의 3단계 계층형 임베딩 방식으로 표현하였다. 예를 들어 I21.9는 I, I21, I21.9로 분해하여 각각의 임베딩 벡터를 생성한 후 결합함으로써 질병분류체계의 계층적 의미를 학습하도록 하였다.

셋째, 건강보험 수진기록은 개별 코드보다 코드 간의 조합과 순서가 실제 원사인 판단에 중요한 정보를 제공한다. 이를 반영하기 위하여 사망 직전 진료기록의 주상병과 부상병을 하나의 입력 단위로 구성하는 쌍(pair) 임베딩 기법을 적용하였다.

<표4-6> 모델2의 학습·평가 데이터 구성

입력변수	출력변수	학습데이터	평가데이터
잠정 원사인 code, 사망진단서, 건강보험 자료	예측 원사인 code	'23년 사망진단서 잠정 원사인, 2023년 건강보험 자료(월간)	2024년 최종 원사인

다. 학습 및 평가데이터 구성

학습자료는 2023년 사망진단서를 기반으로 산출된 잠정 원사인 자료 353,383건을 활용하였다. 이를 학습과 검증자료로 8:2 비율로 분할하여 학습자료 280,343건과 검증자료 73,040건을 구성하였다.

모델 학습은 검증 성능이 더 이상 향상되지 않을 경우 조기 종료하였다. 각 epoch마다 검증 성능을 평가하여 가장 우수한 성능을 보인 시점의 모델을 최종평가 대상의 모델로 선정하였다. 평가자료는 2024년 사망원인통계 자료 358,569건을 사용하였으며, 정답 라벨은 최종 원사인 코드로 설정하였다.

<표4-7> 학습데이터 및 평가데이터 구성

연도	구분	건수
2023	학습	전체 353,383건(80%학습 및 20%검증에 활용)
2024	평가	전체 358,569건

라. 후보모델 검토 및 선정

모델2에서는 KcELECTRA 텍스트 인코더(사망진단서 사인 텍스트 처리용)와 코드 인코더(잠정 원사인, 건강보험 KCD 코드 처리)를 결합하는 2중 인코더 구조를 채택하였다. 코드 인코더 후보 모델은 CNN, FT-Transformer, GCN, HealthGAT의 네 가지로 구성하였으며, 각 모델의 성능을 비교 및 평가하였다.

<표4-8> 코드 인코더 모델별 특징

모델	특징
CNN	임베딩된 코드를 시퀀스로 구성하여 1D 합성곱으로 국소 패턴 추출, 코드 간 계층관계 미반영
FT-Transformer	범주형 변수 처리에 특화된 Transformer 구조. Self-Attention으로 코드 간 상호작용 반영
GCN	KCD 17,193개 코드를 노드로 하고 의학적으로 가까운 코드끼리 약 6만 개 엣지로 연결한 그래프 구조. 코드 간 계층·인접관계를 그래프 컨볼루션으로 학습
HealthGAT	KCD 계층별 Graph Attention Network. 미국 eICU(중환자실 EHR) 기반 사전학습. 이웃 코드 중요도를 가중치로 반영

동일 학습조건에서 4종 인코더의 성능을 비교하였으며, 분류체계는 전체코드 기준과 104항목 분류 기준으로 나누어 측정하였다. 전체코드 기준에서는 그래프계열(HealthGAT, GCN)이 Accuracy 0.64~0.65, Weighted-F1이 0.62~0.63수준으로 높은 성능을 보였다. 104항목 분류 기준에서는 그래프 계열이 Accuracy 0.83 수준으로 크게 향상되었으나 CNN은 이 체계에서도 Accuracy 0.2035에 그쳤다. Macro F1에서는 GCN이 FT-Transformer 및 HealthGAT에 비해 우위에 있었다.

<표4-9> 코드 인코더 후보별 성능 비교

모델	Accuracy	Macro F1	Weight F1
CNN	0.6623	0.0768	0.6350
FT-Transformer	-	-	-
GCN	0.6547	0.0784	0.6247
HealthGAT	0.6465	0.0808	0.6279

성능 수치만으로는 HealthGAT이 GCN에 비해 Weighted F1이 미세하게 높으나, 본 연구에서는 GCN을 최종 모델로 선정한 이유는 다음과 같다.

첫째, KCD 분류체계와의 도메인 정합성이 높다. HealthGAT는 미국 eICU(중환자실 전자의무기록, 2014~2015년)를 기반으로 환자, 진단, 약물, 시술 등의 관계를 보조과제(auxiliary task)를 통해 사전학습한 그래프 신경망 모델이다. 따라서 사전학습 가중치가 미국 임상환경과 ICD-9, 영문 의료텍스트에 특화되어 있어 한국의 KCD-8환경과 차이가 있다. 반면 GCN은 KCD-8 코드체계를 직접 그래프 구조로 구성하여 한국 분류 환경에 쉽게 적용할 수 있다.

둘째, 희귀 코드에 대한 분류성능(Macro F1)이 우수하다. GCN(0.691)의 Macro F1 점수가 HealthGAT(0.680)를 상회하여 희귀코드 분류 성능이 우위에 있음을 확인하였다. 그리고 희귀코드(HIV, 알츠하이머 등)에서 GCN이 CNN 대비 F1-Score가 10%p 이상 개선됨을 확인하였다.

셋째, ICD-11 전환 시 그래프 구조만 갱신하면 되므로 확장성이 높다. 향후 ICD-10에서 ICD-11로 분류체계 전환 시 CNN은 라벨공간 변경에 따른 전면 재학습이 필요하나, GCN은 그래프 구조의 노드와 엣지를 갱신하는 방식으로 대응할 수 있어 장기 운영 측면에서 유리하다.

마. 모델 구조

이상의 검토를 종합하여 GCN을 코드 인코더로서 최종 선정하였다. 최종 모델은 사망진단서의 사인 텍스트와 잠정 원사인, 건강보험 질병코드 정보를 함께 활용하도록 설계하였으며, 텍스트 인코더, 코드 인코더, 퓨전 레이어(정보 결합), 최종 예측의 네 단계로 구성된다.

텍스트 처리 단계에서는 사망진단서의 직접사인과 가·나·다의 원인에 기재된 사인명을 분석하여 의미있는 정보를 추출한다. 코드 처리 단계에서는 잠정 원사인과 건강보험 수진기록의 질병코드 간의 관계를 반영하여 사망원인 예측에 필요한 정보를 생성한다.

<그림4-3> GCN기반 모델2의 처리단계 구조도



정보 결합 단계(Fusion Layer)에서는 텍스트 코드와 코드 정보를 하나의 통합 표현으로 결합한다. 이후 분류기 헤드(Classification Head)는 이 통합 표현을 바탕으로 17,193개의 원사인 KCD 코드에 대한 예측 확률을 산출한다.

또한 세부 원사인 코드와 함께 대분류 수준의 사망원인을 동시에 예측하는 보조 분류기(Auxiliary head)를 추가하였다. 이는 세부 코드만을 직접 예측할 때 발생하는 학습의 어려움을 완화하기 위한 것으로, 대분류 정보를 함께 학습함으로써 희귀 코드의 분류 성능을 향상시키도록 설계하였다.

바. 데이터 불균형 보정을 위한 보조기법

모델2의 분류대상인 KCD 세부코드 17,193개는 학습데이터 분포에 있어 극단적인 Long-tail 구조를 가진다. 전체 코드 중 학습데이터에 1건만 등장하는 코드가 약 3,100개에 달하며, 이 코드들은 충분한 학습이 어렵다. 반면 암·폐렴과 같은 출현 빈도 높은 코드들은 전체 사망사례의 대부분을 차지하여 최빈코드와 희귀코드 간 학습건수 격차가 최대 30,000배에 이른다.

<표4-10> 보조기법 및 역할

기법	역할	수식	적용방법
Focal Loss	쉬운 사례 별점 감소, 희귀코드 집중학습	$FL(p, \gamma) = (1-p)^\gamma \times (-\log(p))$	$\gamma=1.5$ 적용
Label Smoothing	정답확신 억제, 일반화 성능 향상	$y'_k = (1 - \epsilon) \times y_k + \epsilon/K$	$\epsilon=0.1$ 적용
Combined Loss	Focal Loss + Label Smoothing + CE(대분류)	$1.0 \times FL(\gamma=1.5) + 0.1 \times LS(\epsilon=0.1) + 0.4 \times CE_{\text{대분류}}$	
Logit Adjustment	추론단계에서 빈도 높은 코드에 대한 감점	$f'_k = f_k - \tau \times \log(\pi_k)$	$\tau=0$
보조분류기 (Aux head)	대분류 동시학습으로 큰 방향 보조	-	전체코드에 적용

이와 같은 데이터 불균형 환경에서 표준 Cross-Entropy(CE) 손실함수만으로 모델을 학습한다면 빈도 높은 코드 중심으로 최적화가 진행되어 희귀코드 예측력이 떨어지는 현상이 발생할 수 있다. 이는 사망원인통계의 완전성과 활용성을 저하시킬 수 있으므로, 데이터 불균형을 완화하기 위한 보조기법을 함께 적용하였다.

본 연구에서는 손실함수를 Focal Loss, Label Smoothing, 보조 분류기(Auxiliary Head)를 조합하여 구성하였으며, 다만 대분류 예측에는 Cross-Entropy 손실함수를 사용하였다. 또한 추론 단계에서는 Logit Adjustment 기법을 적용하는 것을 검토하였다.

사. 평가방법 및 기준

모델2의 성능 평가는 행정자료를 반영하여 AI가 예측한 원사인과 사망원인통계 최종 원사인 간의 일치도를 중심으로 수행하였다. 분류체계 및 활용 목적에 따라 두 가지로 나누었다.

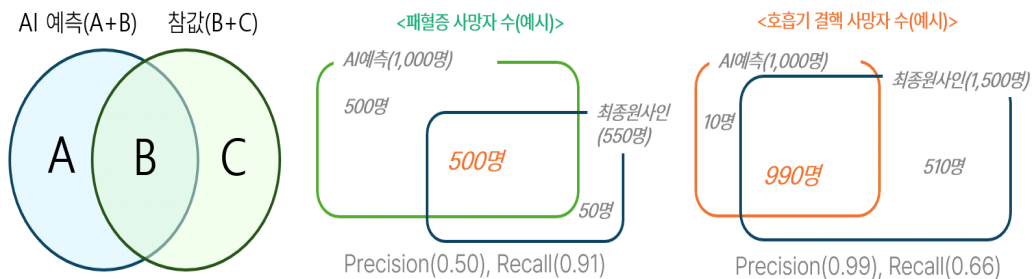
전체코드(KCD 세부코드, 17,193개) 기준은 사인분류사의 업무보조 목적에서, 104항목 분류 기준은 잠정 예측서비스 목적에서 평가한다. 특히 10대 사망원인과 시의성이 요구되는 감염성 질환을 대상으로 질환별 평가를 실시하였다.

평가지표는 정밀도(Precision), 재현율(Recall), F1-score, 사망자 수 추정비 등 4가지로 구성된다. 이 모델의 목표 수준은 정밀도 및 재현율이 일정 수준 이상이면서, 두 지표가 균형적으로 도출되도록 하여 사망자 수 예측 정확도를 확보하는 것을 목표 수준으로 잡았다. 검증지표의 개념은 다음과 같다.

<표4-11> 모델2 검증지표 개념

검증지표	내용	산식
정밀도(Precision)	AI 모델이 정답으로 예측한 값 중 실제 정답의 비율	$\frac{B}{(A+B)}$
재현율(Recall)	실제 참값 중 AI 모델이 정답으로 맞춘 비율	$\frac{B}{(B+C)}$
정확도(F1 Score)	정밀도와 재현율의 조화평균	$\frac{2 \times (Precision \times Recall)}{(Precision + Recall)}$
사망자 수 추정비	실제 잠정 사망자 수 대비 AI 모델이 예측한 사망자 수의 비(ratio)	$\frac{(A+B)}{(B+C)}$

<그림4-4> 모델2 검증지표 개념



정밀도가 높고 재현율이 낮은 경우는 AI가 예측한 것은 정확하지만, 예측량이 적어서 활용도가 낮은 경우이다. 반대로 정밀도가 낮고 재현율이 높은 경우는 AI가 과다 예측하여 예측값에 정답을 많이 포함하지만 정확도가 낮은 경우이다.

사망자 수 추정비는 이 두 지표의 균형상태를 확인하고 실제로 예측 사망자 수를 추정 및 서비스할 때 사용하게 되는 지표로서, 추정비가 1에 가까울수록 AI가 예측한 사망자 수가 실제 사망자 수에 근접함을 의미한다.

지표 계산 방법은 코드별 집계 방식에 Weighted와 Macro 방식으로 계산한다. Weighted는 코드별 F1-Score에 빈도 가중평균으로서 실제 운용성능 평가에 적합하다. Macro F1은 코드별 F1-Score의 단순평균으로 희귀코드에서 학습이 되지 못해 F1-Score가 0이 되는 경우 낮게 나타날 수 있다.

정확도 비교 수준은 KCD 코드 분류체계의 단위에 따라 세부코드 수준, 3자리 정확도(KCD 소분류 3자리 코드), 104항목 수준으로 구분하여 분석하였다.

<표4-12> 평가지표 계산방법 및 비교수준

구분	지표	정의	활용 목적
계산 방법	Weighted F1	코드별 F1-Score에 빈도를 가중	실제 운용성능 평가
	Macro F1	코드별 F1-Score의 단순 산술평균	희귀코드 성능 중점 평가
비교 수준	세부코드	KCD 세부코드 6,443개 기준	사인분류사 업무보조, 정확도 판단
	3자리코드	KCD 소분류(3자리) 코드 기준	질환별 세부 분석
	104항목	통계 공표단위 104항목 기준	감염성 질환 잠정치 예측(시의성 중요)

이상의 지표를 종합적으로 검토하되, 정책적 활용도가 높은 10대 사망원인과 시의성 있는 예측이 중요한 감염성 질환에 대해서는 질환별 분리평가를 실시하였다. 10대 사망원인은 사망원인통계의 핵심 공표 항목으로서 예측 정확도가 특히 중요하다. 암, 심장질환, 뇌혈관질환 등 상위 10개 사인에 대해 AI 예측 정확도를 중심으로 검토한다.

감염성질환(코로나19, 인플루엔자, 바이러스성 감염 등)은 계절성이 뚜렷하게 나타나거나, 특정 사건에 따라 사망자 수가 단기적으로 급증하는 특성이 있어 연간 집계 수준의 성능으로는 실용성을 충분히 검증하기 어렵다.

따라서 연간 집계 수준의 정확도뿐 아니라 월별 추이를 함께 분석하여 AI 예측값이 실제 사망자 수의 시계열 변동을 얼마나 정확하게 반영하는지를 평가하였다. 이를 통해 감염병 유행이나 예상치 못한 사망자 증가 상황에서 잠정 통계서비스의 활용 가능성을 검토하였다.

3. (모델3) 약물중독 코드 자동분류

가. 모델의 목적 및 설계

본 모델의 목적은 국립과학수사연구원의 5단 분류 서술형 자료로부터 약물중독 사망자의 약물 정보를 추출하고, 이를 KCD T코드(T40~T43, T45, T48, T50)와 약물 우선순위 체계로 자동 분류하는 데 있다. 약물 우선순위 체계는 국제표준질병사인분류 제2권 지침서의 「약물, 약제 및 생물학적 물질에 의한 중독에 대한 특별지침(M4)」을 준용하였다.

현행 사인분류 업무에서는 사인분류사가 서술형 행정자료를 직접 판독하여 약물명과 약물 우선순위를 확인하고 있어 상당한 업무 부담이 발생하고 있다. 이에 인구동향과 소속 사인분류사 11명을 대상으로 의견을 수렴한 결과, 서술형 자료에서 약물명과 우선순위 정보를 우선적으로 요약·표출하는 기능에 대한 수요가 높은 것으로 확인되었다. 이러한 현장 수요를 반영하여 약물중독 사망 사례의 약물 정보에 대해 AI를 활용하여 자동 추출·분류함으로써 분류 업무의 효율성과 일관성을 높이하고자 하였다.

<표4-13> 국과수 5단분류 자료 기재사례

기재 예시	의미
급성중독-치료약물-졸피뎀 등-자	단일 약물 케이스
급성중독-치료약물-에스시탈로프람, 트라마돌-자	복수약물(2종)
온도이상-고온-화재시-사-차량화재, 알코올 0.117%, 졸피뎀	약물중독 외 케이스이나 약물 등장
급성중독-약물 및 알코올-약물(7-아미노플루니트라제팜, 트리아졸람, 아미트리프틸린, 쿠에티아핀 및 트라조돈 등) 및 알코올 중독-사	복잡한 복수 약물조합

<표4-14> 약물코드 자동분류 모델설계

입력변수	출력변수	참조자료
국과수 부검자료 (5단분류 텍스트)	검출약물별 T코드, 약물 우선순위 R코드	약물분류표(약물사전)

모델의 입력값은 국립과학수사연구원의 5단 분류 서술형 자료이며, 출력값은 검출된 약물별 KCD T코드와 우선순위 코드이다. 입력자료는 '대분류-중분류-세분류-사망구분-추가설명'의 5단계 구조로 구성되며, 각 항목은 하이픈(-)으로 구분된다. 약물명은 주로 세분류 항목에 기재되지만 다른 항목에 포함되는 경우도 존재한다.

또한 실제 자료에는 단일 약물 사례뿐만 아니라 복수의 약물이 함께 기재된 사례, 약물과 무관한 정보가 혼재된 사례 등 다양한 형태가 존재하여 자동 분류 과정에서 추가적인 해석이 요구된다.

나. 약물 분류체계

복수의 약물이 검출된 경우에는 사망에 가장 직접적으로 기여한 약물을 특정하기 위하여 부록의 표와 같이 우선순위 분류의 16단계 기준을 적용하였다. 우선순위가 높은 약물일수록 사망 기여도가 높은 것으로 판단하며, 복수의 약물이 검출된 경우에는 가장 높은 우선순위를 가진 약물을 기준으로 사망원인통계에 집계한다.

다. 후보모델 검토 및 선정

국립과학수사연구원의 부검기록은 부검의에 따라 약물명이 영문, 한글 이체자, 대사체, 제품명 등 다양한 형태로 기재되며, 오타자와 부가적인 설명이 함께 포함되는 경우도 많다. 또한 약물 정보와 비약물 정보가 하나의 서술문에 혼재되어 있어 규칙 기반 방식만으로는 안정적인 분류가 어렵다.

한편, 본 연구에 활용한 자료는 정답 라벨이 사전에 구축되어 있지 않아 지도학습을 적용하기 어려웠다. 이에 별도의 학습 없이 프롬프트만으로 정보 추출이 가능한 대규모 언어모델(LLM)을 적용하는 것이 적합하다고 판단하였다.

다만 T코드 부여 및 우선순위 결정까지 모두 LLM에 의존하는 방식은 신뢰성 측면에서 한계가 있었다. 실제로 약 20억 개(2B) 규모의 소형 언어모델을 단독으로 적용한 결과 출력 형식을 준수하지 않거나 우선순위 규칙을 위반하는 사례가 확인되었다. 반면 대형 언어모델은 높은 성능을 기대할 수 있으나, GPU 자원 소모가 크고 추론 시간이 증가한다는 제약이 있었다.

이에 본 연구에서는 LLM과 규칙 기반 엔진을 결합한 하이브리드 구조를 채택하였다. 1단계에서는 LLM이 서술형 자료로부터 약물명을 추출하고 표기를 정규화하며, 2단계에서는 Python 기반 규칙 엔진이 사전에 구축한 약물사전을 조회하여 해당 약물의 T코드와 우선순위 코드를 자동 부여하도록 설계하였다. 이러한 구조는 LLM의 생성 결과를 규칙기반으로 사후 검증할 수 있다는 장점이 있다.

LLM은 개인정보 보호와 망분리 환경을 고려하여 내부서버에서 운영할 수 있는 공개 모델을 대상으로 선정하였다. 이에 공개모델인 Gemma 4 계열을 후보모델로 선정하고, 모델 규모에 따른 성능과 운영 효율성을 비교·평가하였다.

<표4-15> 모델3 관련 언어모델 검토

검토모델	파라미터	검토결과
14B 이상 대형	대형	GPU 자원 · 토큰비용 · 추론속도 측면에서 분리 LLM과 규칙엔진 분리로 LL의 역할이 추출 · 정규화만 담당하게 됨에 따라 4B급 대응이 가능해져 미채택
Gemma-4-e2b	약 2B	초소형 범용모델. 텍스트 주체(주어) 분석 취약, 시스템 프롬프트 규칙 위반(약물 미기재 시에도 임의 라벨 부여), 추론과정과 출력라벨 간 자기모순 확인
Gemma-4-e4b	약 4B	한국어 범용모델. 프롬프트와 후처리 안전장치 결합 시 구조적 제약 준수와 임상적 추론 정확도 달성 가능(최종 채택)
Gemma-4-31b	대형	향후 신약 추가 · 임상 가이드라인 보강 시 활용을 위한 대비 후보

검토 결과, Gemma-4-e2b 모델은 서술형 자료의 의미 해석과 규칙 준수 측면에서 한계가 확인되어 적용이 어려운 것으로 판단하였다. 반면 Gemma-4-e4b 모델은 규칙 기반 엔진과 결합할 경우 안정적인 성능을 보였으며, 운영 효율성 측면에서도 적합한 것으로 판단되어 최종 모델로 선정하였다. 또한 Gemma-4-31b 모델은 향후 신약추가 및 임상지침 변화 등에 대응하기 위한 확장가능성을 고려하여 후보 모델로 검토하였다.

라. 언어모델 입력자료 설계방법

언어모델의 성능을 안정적으로 확보하기 위하여 시스템 프롬프트와 few-shot 예시를 함께 설계하였다. 시스템 프롬프트에는 모델의 역할, 수행 규칙, 출력 형식 및 예외 처리 기준을 포함하였으며, few-shot 기법을 활용하여 대표적인 입력 · 출력 사례를 제시함으로써 모델이 일관된 형식과 판단 기준에 따라 응답하도록 하였다.

시스템 프롬프트에는 연구 과정에서 구축한 약물사전을 포함하였다. 약물사전에는 약물의 표준명, 소분류 코드 및 우선순위 정보 등이 포함되어 있으며, LLM은 서술형 자료에서 약물명을 추출하고 표준 한글명으로 정규화하는 역할만 수행하도록 설계하였다.

반면 질병코드(T코드) 부여, 우선순위 결정, 정렬 및 중복 제거는 규칙 엔진이 담당하도록 역할을 분리하였다. 이를 통해 LLM은 자연어 이해와 정보 추출에 집중하고, 규칙 기반 엔진은 코드 부여와 우선순위 결정의 일관성을 확보하도록 설계하였다.

마. 평가방법 및 기준

본 연구는 LLM과 규칙 기반 엔진을 결합한 하이브리드 구조를 적용하였으므로, 두 단계의 성능을 구분하여 평가하였다.

LLM 단계에서는 서술형 자료에서 약물명을 정확하게 추출하고 정규화하는 능력을 중심으로 평가하였으며, 환각(hallucination) 발생 여부와 환각 차단율을 주요 점검 항목으로 설정하였다.

규칙 기반 엔진 단계에서는 추출된 약물에 대해 T코드와 우선순위 분류가 정확하게 부여되었는지, 약물분류표 및 우선순위 규칙을 올바르게 적용하였는지를 평가하였다. 또한 출력 결과가 사전에 정의한 출력 형식을 일관되게 준수하는지 여부를 함께 확인하였다.

4. (모델4) 자살 위험요인 자동분류

가. 모델의 목적 및 설계

본 모델의 목적은 자살 사망자에 대해 결합된 행정자료로부터 자살 위험요인을 추출하고, 이를 분류체계에 맞추어 자동분류 하는 데 있다. 자살 위험요인 분류체계는 제 10차 국제질병사인분류(ICD-10)와 호주통계청(ABS, Australian Bureau of Statistics)의 자살 위험요인(Risk Factors for suicide) 분류체계를 준용하되, 우리나라 현실에 맞게 묶거나 분기하여 구성하였다.

현재 경찰청은 매년 공표하는 경찰통계연보에서 자살원인에 대한 통계표를 작성하고 있으며, 사건개요·사망수단 등 비정형 텍스트 자료를 바탕으로 자살동기를 분류한다. 그러나 현행 경찰청의 자살동기 분류체계는 다음과 같은 한계를 가진다.

첫째, 단일분류 원칙의 문제이다. 자살은 여러 요인이 복합적으로 작용하나 하나의 항목으로 코딩이 강제됨에 따라 작성자의 자의적 판단이 개입될 여지가 크다. 예컨대 ‘배우자의 외도로 인한 이혼’은 남녀문제와 가정문제에, ‘만성질환으로 인한 직장상실’은 육체적질병·직업·경제생활문제에 모두 해당될 수 있다.

둘째, 불완전성의 문제이다. 인간관계·상황·건강·사건·피해 등 5개 이상의 서로 다른 분류기준이 동일한 레벨에 혼재하고 있어, 이를 상호 배타적으로 집계할 경우 많은 정보가 유실될 수 있다.

셋째, 변별력의 문제이다. 법적문제·학업스트레스·군복무 관련 문제 등 사회적으로 중요한 항목이 분류체계에 부재하고, 단일분류 원칙까지 적용되어 ‘경제생활’, ‘육체적질병’, ‘정신적·정신과적 문제’ 등 3개 항목에 전체의 약 80%(2024년 기준)가 편중된

다. 또한 기타·미상 비율이 8.7%에 달하여 분류항목 부재로 인한 누락가능성이 있다.

이러한 한계를 보완하기 위하여 본 연구에서는 경찰청 사건개요(월간·연간), 사망수단(월간·연간), 해양경찰청 사건개요, 국과수 5단분류 및 사건개요를 이용하여 자살 위험요인 텍스트를 추출하고, 분류체계에 맞추어 자살 위험요인을 자동 추출하고 분류 체계에 따라 다중 분류하는 방식으로 모델을 설계하였다.

나. 입력데이터 구조 및 특성

입력데이터는 2024년 14,875명의 자살사망자로 구성되며, 사망자별로 경찰청, 해양경찰, 국과수 행정자료를 결합하여 구축하였다. 구성항목은 다음과 같다.

<표4-16> 모델4 입력데이터 구성항목

자료원	주요 필드	레코드 수	레코드비율
경찰청	사건개요_연간(POL_YEAR_COMMENT)	14,523	(97.6%)
	사건개요_월간(POL_MONTH_COMMENT)	14,497	(97.5%)
	사망수단_연간(POL_YEAR_COMMENT)	14,537	(97.7%)
	사망수단_월간(POL_MONTH_COMMENT)	14,546	(97.8%)
해양경찰청	사건개요(NMPA_COMMENT)	110	(0.7%)
국과수	5단분류(NFS_GUBUN)	1,343	(9.0%)
	사건개요(NFS_COMMENT)	1,345	(9.0%)

입력데이터는 두 가지 특성을 가진다. 첫째, 사망수단 등 코드성 자료와 사건개요·5단분류 등 비정형 텍스트 자료가 혼재한다. 사망수단은 수사관이 모든 정황을 바탕으로 사망동기를 판단한 자료이고, 경찰사건개요는 수십~수백자 수준의 서술형 자료로서 자살 위험요인을 추출하는데 중요한 정보를 제공한다.

둘째, 경찰청 자료는 연간자료와 월간자료를 모두 활용하였으며 2024년 전체 레코드 14,875건 중 약 14,500건이 채워져 있어 가장 많은 정보를 포함하고 있는 자료이다. 특히 경찰청에서 작성한 사망구분 및 사망수준과 같은 코드성 자료는 LLM이 사망위험요인을 분류하는데 결정적 단서를 제공하는 경우가 많다.

다. 자살 위험요인 분류체계

본 분류체계는 ICD-10과 호주 통계청의 자살 위험요인 통계 작성방법을 참고하였다. 호주통계청은 사망원인통계의 하위 통계로 자살 위험요인을 공표하며, 한 명의 사망자에게 복수의 위험요인을 부여하는 다중요인 코딩 방식을 적용한다. 또한 위험요인을 ①심리·사회적 위험요인, ②정신적 요인, ③질병·신체적 요인의 3종 대분류로 구분하여 ICD-10 코드와 매핑함으로써 국제표준 정합성과 통계 완전성을 확보하였다. 본 연구에서는 이를 준용하여 3개 대분류와 21개 소분류로 구성된 분류체계를 구축하였다.

<표4-17> 모델4 자살 위험요인 분류체계

대분류	KCD 범위	중분류
1. 심리·사회적 요인	Z55-Z65	교육적, 직업적, 경제적, 사회환경, 가정적, 법적상황, 기타 요인
2. 정신적 요인	F00-F99	기분정동장애, 스트레스 연관 장애, 급·만성 정신성물질 남용, 급·만성 알코올 남용, 기타
3. 질병·신체적 요인	A00-R99	만성질환, 말기·불치 질환, 만성통증, 신체기능 저하·장애, 기타

이 분류체계는 앞서 기술한 경찰청 분류체계의 세 가지 한계를 해소한다.

첫째, 여러 위험요인이 복합적으로 발생할 수 있음을 인정하여 하나의 자살사건에 대해 다중 코딩을 허용한다.

둘째, 위험요인을 3개 대분류로 체계화하고 ICD-10 코드와 연계할 수 있도록 설계하여 국제 비교 가능성과 통계의 일관성을 확보하였다.

셋째, 기타·미상 항목으로의 편중을 방지하기 위하여 대분류별로 빈도가 높은 중분류를 충분히 세부화하고, 직업·경제·가정 관련 요인을 사회·심리적 요인으로 분류하여 정신·행동장애 요인과 명확히 구분함으로써 변별력 문제를 해소하였다.

라. 언어모델 입력자료 설계

본 모델은 사전에 구축된 정답 라벨이 없는 자료를 대상으로 분류를 수행해야 하므로, 모델3과 동일하게 별도의 학습 과정 없이 프롬프트 기반으로 추론이 가능한 대규모 언어모델(LLM)을 활용하였다. 다만 서술형 자료에서 LLM이 위험요인을 직접 결정하는 방식은 추론 과정에서 오류가 발생할 가능성이 있으므로, (1단계) 근거 추출과 (2단계) 종합 판단을 분리하는 2단계 프롬프트 구조를 적용하였다.

1단계 근거 추출에서는 7개 입력 필드에 포함된 자살 위험요인 관련 근거 구절을 모두 추출하고, 각 구절에 대해 대분류-중분류-소분류체계의 21개 위험요인 라벨을 부여하도록 설계하였다.

2단계 종합 판단에서는 1단계에서 추출된 근거와 부여된 라벨 정보를 종합하여 최종 자살 위험요인을 결정하도록 하였으며, 출력 위험요인은 최대 2개로 제한하였다. 이때 필수 위험요인 1개와 부가 위험요인 0~1개를 출력하도록 설정하였다.

출력 결과의 신뢰성을 확보하기 위하여 다음과 같은 제약 조건을 적용하였다. 먼저 출력 라벨은 사전에 정의한 자살 위험요인 분류체계(21개 소분류) 내에서만 선택하도록 폐쇄형 라벨 집합을 적용하였다. 또한 동일 위험요인이 중복 출력되는 것을 방지하기 위하여 자동 중복 제거 규칙을 적용하였다. 아울러 최종 위험요인은 반드시 근거 구절에 부여된 후보 라벨 중에서만 선택하도록 제한하여 근거가 부족한 라벨 생성을 방지하였다. 마지막으로 발견자 및 신고자의 행동, 주변 상황 등 자살 위험요인 자체로 보기 어려운 정황 정보는 분류 대상에서 제외하도록 기준을 설정하였다.

마. 평가방법 및 기준

본 모델은 사전에 정의된 정답 라벨이 존재하지 않아 일반적인 지도학습 기반 평가 지표(accuracy, F1-score 등)를 적용하기 어렵다. 이에 따라 모델 분류 결과에 대한 품질 검토를 통해 분류 타당성을 검증하였다.

검증 대상은 2024년 자살사망자 14,875건이며, 평가 목적에 따라 ① 미상 검토, ② 희소 분류 검토, ③ 임의 추출 검토의 세 단계로 구분하여 수행하였다. 이를 통해 전체 분류 결과의 대표성을 확보하는 동시에 취약 영역에 대한 집중 검토를 병행하였다.

① 미상 검토는 ‘미상’으로 분류된 사례가 실제로 위험요인 관련 정보가 부족한건지, 관련 정보가 존재함에도 모델이 이를 적절히 추출하지 못한 사례인지를 확인하기 위한 것이다.

② 희소분류 검토는 상대적으로 발생빈도가 낮은 위험요인 항목을 대상으로 분류 결과의 타당성을 점검하기 위한 것으로 모델의 판단 일관성을 검토하였다.

③ 임의추출 검토는 특정 위험요인이나 분류 유형에 편중되지 않은 전체적인 분류 품질을 확인하기 위하여 전체 결과 중 1%를 무작위 추출하여 검토하였다.

<표4-18> 품질검사 단계 및 기준

단계	검사기준	검사건수
① 미상	미상으로 분류된 레코드의 추가분류 가능성 점검	852건
② 희소분류	빈도가 낮은 소분류 위주 점검	297건
③ 임의추출	기타 표본 중 약 1%를 임의추출하여 점검	132건

제 5 장

사망원인통계 AI 적용 모형 분석 결과

제1절 (모델1) 사망진단서 사인 코드화 모형

1. 학습데이터 구성에 따른 성능 비교

가. 예측모델 및 평가 개요

본 연구는 KcELECTRA 기반 인코더 위에 KCD 코드에 대한 멀티라벨 분류 구조를 기본으로 유지하면서, 학습데이터의 중복제거와 예측확률의 임계값 설정 등에 따른 결과비교 및 조정을 통해 성능을 개선하였다. 아울러 학습데이터의 구성 방식에 대한 성능변화를 추가로 검토하였다. 인코더 구조와 학습 하이퍼파라미터는 동일하게 유지하고, 학습자료의 구성만 달리하여 두 개의 후보 모델을 구축하였으며, 본 연구에서는 이를 각각 조건 A와 조건 B로 정의하였다.

<표5-1> 예측모델별 학습데이터 구성

예측모델	학습데이터 구성
조건 A	2023년 사망진단서 학습데이터 단독
조건 B	2023년 사망진단서 학습데이터 + 한글사전

조건 A는 2023년 사망진단서 자료에서 한글사전 매칭이 확정된 레코드만을 학습데이터로 사용한 기본 구성이다. 조건 B는 여기에 KCD 한글사전에서 중복을 제거한 자료를 추가로 결합한 구성이다. 한글사전의 결합을 통해 실제 사망진단서 자료에서 관측 빈도가 낮은 표준 사인 표기의 코드 매핑 정보를 모델이 학습하도록 하여, 희귀 질환에 대한 예측 안정성을 높이하고자 하였다. 또한 이러한 학습자료의 확장이 실제 예측 성능 향상으로 이어지는지를 두 모델의 비교를 통해 확인하였다.

성능 비교의 기준선(baseline)은 현재 사인선정시스템이 부여하는 사전분류 코드로 설정하였다. 사전분류 코드는 본 연구의 예측모델(모델1)이 대체하고자 하는 처리 단계의 산출물이므로, 이를 기준으로 AI 예측모델의 상대적 성능을 비교하는 것은 실무 적

용 가능성을 판단하는 중요한 근거가 된다. 본 절에서 제시하는 모든 성능 수치는 임
계값 0.5 기준의 결과이다.

나. 학습데이터 구성별 성능 비교

조건 A와 조건 B의 예측 결과를 2024년 사망진단서 평가데이터 684,841건에 대해
산출하고, 사전분류 코드를 정답으로 가정하여 전체 일치율을 비교하였다.

<표5-2> 학습데이터 구성별 사망원인 코드 일치율

예측모델	일치 건수	불일치 건수	일치율
조건 A	561,683	123,158	82.02%
조건 B	563,317	121,524	82.26%

두 모델의 전체 일치율은 각각 82.02%와 82.26%로 나타났으며, 한글사전을 결합한
조건 B가 조건 A보다 0.24%p 높은 일치율을 보였다. 한편, 위 82%대의 일치율은 사전
분류 코드를 기준선으로 삼은 상대 지표라는 점에 유의할 필요가 있다. 예측모델의 절
대적 정확도를 나타내는 값이 아니므로, 이 결과만으로 예측모델의 실질 성능을 단정
하기는 어렵다.

다. 매칭 영역별 일치율 차이

전체 일치율을 매칭코드별로 구분하여 살펴보면 두 영역 간 일치율 차이가 매우 크
게 나타난다. 두 매칭 영역 간 일치율 격차는 75%p에 달한다. 한글사전 매칭 영역
(MTCHG_CODE=4)에서는 두 모델 모두 98%대의 높은 일치율을 보인 반면, 색인사전
매칭 영역(MTCHG_CODE=2)에서는 23%대로 극히 낮았다.

<표5-3> 사망진단서 사인 매칭코드별 일치율

매칭 영역(건수)	조건 A		조건 B	
	일치 건수	일치율	일치 건수	일치율
색인사전(146,389)	34,173	23.34%	34,015	23.24%
한글사전(536,687)	527,510	98.29%	529,302	98.62%
전체(684,841)	561,683	82.02%	563,317	82.26%

한글사전 매칭 영역의 높은 일치율은 학습자료와 평가자료가 동일한 매칭 경로를 통해 확정된 자료라는 점에서 예상 가능한 결과이다. 학습 과정에서 모델이 접한 표기 패턴과 평가 시점의 표기 패턴이 동질적이므로, 사전분류 결과와 예측 결과가 대부분 수렴한다. 조건 B가 조건 A보다 다소 높은 결과를 보인 것은 KCD 한글사전 자료의 추가 결합이 표준 표기의 코드 매핑을 보완하는 효과를 일부 제공하였기 때문으로 해석된다.

반면, 색인사전 매칭 영역의 저조한 일치율은 두 요인이 복합적으로 작용한 결과로 볼 수 있다. 첫째, 학습자료 구성에 따른 영향이다. 본 연구의 학습자료는 한글사전 매칭이 확정된 레코드에 한정되어 있으므로, 색인사전 매칭 영역에 나타나는 비표준 표기·오타자·복합 사인 등의 텍스트 패턴이 학습 과정에서 직접적으로 노출되지 않았다. 따라서 이러한 유형의 텍스트에 대해서 모델이 사전분류와 동일한 코드를 예측하지 못할 가능성이 존재한다.

둘째, 기준선인 사전분류 코드 자체의 신뢰도 문제이다. 제2장에서 확인한 바와 같이, 사인분류사 의견수렴 결과 한글사전과 색인사전의 현행화 지연으로 매칭 원활성과 사전 활용성이 떨어진다는 의견이 있었다. 특히, 색인사전 연계 과정은 한글사전 직접 매칭에 실패한 텍스트에 대해 색인 규칙을 적용하는 것으로, 사전 품질 저하의 영향이 더 크게 나타날 수 있다.

다만, 두 요인의 상대적 기여도는 사전분류 결과와 예측 결과의 수치적 비교만으로는 구분하기 어렵다. 예측모델의 오분류 양상과 그 발생 원인을 개별 사례 수준에서 진단하는 작업이 선행되어야, 이후의 성능 개선 방향과 실무 적용 가능성을 종합적으로 판단할 수 있다.

2. 오분류 유형 및 발생 원인 분석

가. 오분류 대표 유형 및 사례

오분류 유형 분석은 조건 A의 예측 결과를 대상으로 수행하였다. 조건 B에서도 유사한 오분류 양상이 정성적으로 확인되었으므로, 이하에서는 조건 A 결과를 중심으로 정리하였다. 예측 결과에서 반복적으로 관찰되는 오분류 대표 유형과 사례는 다음과 같다.

<표5-4> 사망진단서 사인코드 예측 오분류 대표 유형

오류 유형	한글명	정분류	오분류	건수
문맥 동반 오류	상세불명의 폐렴	J189	C159-J189(식도암-폐렴)	3,543
토큰화 민감도	대사성 산증	E872	R688(전신증상 및 징후)	304
표기 변형 + 계통 오답	해당사항없음	/	J459-F03(천식-치매)	1,526

유형 1은 '상세불명의 폐렴' 표기 3,543건에서 관찰된 오분류 사례이다. 이 표기에 대해 모델은 C159(식도암)을 동반하여 예측하였고, 이 가운데 3,528건(99.6%)은 오히려 C159를 1순위로, 정답인 J189를 2순위로 강등시켰다. 반면 '상세불명의폐렴'(공백 제거) 또는 '상세불명 폐렴'('의' 생략) 등 미세한 표기 변형에서는 99~100%가 J189로 정상 예측되어 상이한 예측 결과를 보였다.

유형 2는 공백 유무에 따라 예측 결과가 정반대로 나타난 사례이다. 공백이 없는 '대사성산증' 304건은 100%가 R688(전신 증상 및 징후)로 오분류된 반면, 공백을 포함한 '대사성 산증' 572건은 99.8%가 정답인 E872(대사성 산증)로 예측되었다. 두 표기는 의미가 동일하지만, 공백 한 칸의 차이가 토큰열을 바꾸어 서로 다른 토큰 식별자로 처리되면서 예측 결과가 전혀 상반되게 나타난 것으로 확인되었다.

유형 3은 정답 코드가 공란(null)인 '해당사항없음' 계열 표기 1,529건에서 관찰된 계통 오답 수렴이다. 이 사례에서는 '해당 사항 없음'(684건), '해당사항없음'(342건), '해당사항 없음'(94건), '해당없음'(48건), '해당상항없음' 등 오타자·공백·마침표의 조합으로 발생한 20여 종의 표기 변형이 존재했다. 그럼에도 불구하고 예측 결과는 전체 1,529건 중 1,526건(99.8%)이 J459(천식)과 F03(치매)이라는 특정 코드 조합으로 반복 예측되었다. 이 때 예측 코드에 대한 확률도 평균 0.865로 높은 확신을 가진 환각을 확인할 수 있었다.

나. 오분류 유형의 공통 특성

세 유형의 오분류의 사례마다 양상은 서로 다르지만, 발생 구조를 분석하면 다음의 공통 요인이 확인된다.

첫째, 학습자료 필터링 정책의 부작용이다. 학습자료는 한글사전과 완전히 매칭된 라벨(MTCHG_CODE=4)에 한정되어 있으나, 고빈도 표준 문구의 정답 라벨이 색인사전 매칭 영역(MTCHG_CODE=2)에 존재하여 모델이 학습할 기회가 없는 것이다. '상세 불명의 폐렴'의 정답 표기 3,224건이, '대사성산증'의 정답 표기 338건이 모두 색인사전 매칭 영역에 해당하여 학습에서 배제되었다. 그 결과 학습에 활용된 표기의 수가 매우 제한적이어서 모델이 올바른 코드 매핑을 충분히 학습하기 어려운 환경이 형성되었다.

둘째, 정답 라벨 자체의 노이즈이다. 유형 1의 경우 학습에 남은 정확 표기 5건 중 2건이 '식도암, 폐렴'의 다중사인으로 C159가 붙어있었고, 유형 2의 경우 무공백 표기에 대해 학습에 남은 유일한 신호가 오라벨된 R688 2건이었다. 모델은 이러한 라벨의 노이즈를 충실히 학습하였고, 그 결과 동일한 유형의 오분류를 반복적으로 산출한 것으로 해석된다.

셋째, 토큰화의 공백 민감성이다. 텍스트 인코더는 공백을 기준으로 토큰 경계를 인식하여 공백 한 칸의 차이는 서로 다른 토큰 식별자와 임베딩을 산출한다. 유형 2의 경우 공백을 포함한 '대사성 산증'은 학습데이터에 없었지만, E872로 코딩된 '대사성 산증 추정/의증' 등과 토큰을 공유함으로써 일반화를 통해 정답 코드를 예측한 것으로 해석된다.

한편, 유형 3의 경우 예측 로직이 항상 1개 이상의 코드를 강제 출력하는 설계를 감안하여 '해당사항없음', '공란' 등 유사한 텍스트에 대해 '코드없음'을 뜻하는 '/(R99)'을 출력하도록 하는 후처리 규칙을 설정할 필요가 있다.

이상의 진단을 종합하면, 본 항에서 확인된 오분류는 모델 구조나 학습 알고리즘 자체의 한계라기보다는 학습자료의 구성, 전처리 정책 및 출력 정책의 영향이 크게 작용한 결과로 판단된다. 이에 따라 학습자료 필터링 정책 재검토, 학습·추론 단계의 공백 표준화 규칙 정비, 후처리 규칙 도입 등이 후속 연구의 검토 대상이 될 수 있다.

3. 표본검수를 통한 성능 재평가

가. 검수 배경 및 대상 선정

앞서 언급된 색인사전 매칭 영역(MTCHG_CODE=2)의 23%대 낮은 일치율은 기준선으로 사용한 사전분류 코드의 품질과도 관련되었을 가능성이 있으므로, 이에 대한 검토가 필요하였다. 이에 사전분류 결과와 모델 예측의 정답 여부를 재판정하기 위한 수작업 표본검수를 수행하였다.

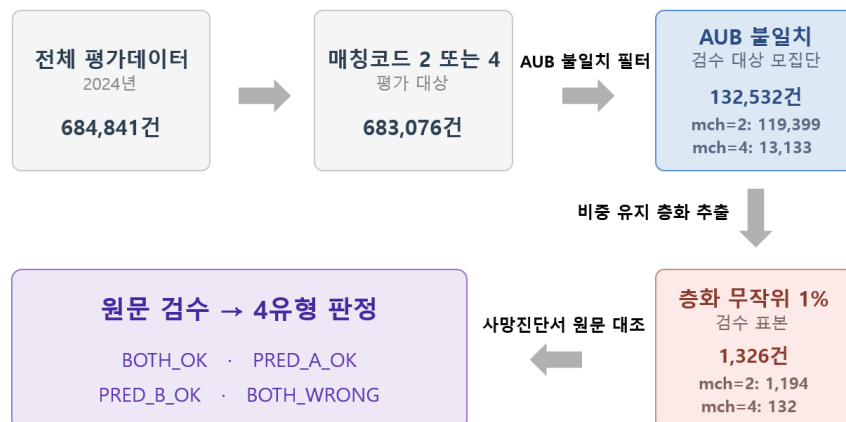
검수 대상은 매칭코드 2 또는 4에 해당하는 자료 중 조건 A 또는 조건 B의 예측 결과가 사전분류 코드와 불일치한 사례의 합집합으로 설정하였다. 두 모델 가운데 하나라도 사전분류와 어긋난 사례를 모두 포함함으로써, 두 모델의 오류를 동시에 비교·검증할 수 있도록 하였다.

<표5-5> 한글사전 매칭 여부별 사인 코드 불일치 대상 분포

매칭영역	건수	사망진단서 자동코딩 결과와 일치여부		
		A조건 불일치	B조건 불일치	A 또는 B 불일치
색인사전	146,389	112,216	112,374	119,399
한글사전	536,687	9,177	7,385	13,133
합계		121,393	119,759	132,532

색인사전 매칭 영역에서 전체 자료의 약 81.6%가 검수 대상에 포함된 반면, 한글사전 매칭 영역에서는 약 2.4%만이 검수 대상이 되었다. 두 영역의 예측 성능 차이가 검수 대상 규모에도 그대로 반영된 결과이다.

<그림5-1> 검수 대상 선정 및 표본 추출 흐름도



나. 검수 표본 설계

검수 대상 132,532건은 매칭 코드별 비중이 약 9:1로 편중되어 있으므로, 모집단의 구성 비율을 유지하기 위하여 층화 무작위 추출을 사용하였다. 검수 대상에서 1%의 표본을 추출하여 최종 표본은 총 1,326건이다.

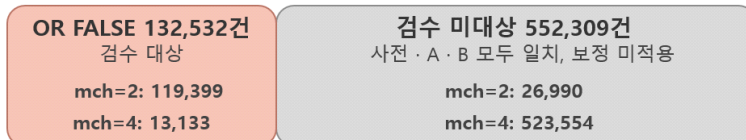
<표5-6> 수동 검수 표본의 구성(층화 무작위, 1%)

매칭 영역	모집단(비중)	표본(비중)
한글사전 미연계(색인사전)	119,399(90.1%)	1,194(90.0%)
한글사전 완전 연계	13,133(9.9%)	132(10.0%)
합계	132,532(100%)	1,326(100%)

검수는 각 표본의 사망진단서 텍스트(CODE_KOR)를 직접 확인하여, 조건 A와 조건 B 예측 결과 각각에 대해 독립적으로 판정하였다. 본 검수는 사인선정시스템의 사전분류 코드가 완전한 정답을 보장하지 않는다는 전제 위에서 이루어졌으며, 원문 텍스트에 기재된 실제 사인을 기준으로 사전분류 코드와 조건 A, 조건 B의 예측 결과에 대해 각각의 정답 여부를 재판정하였다.

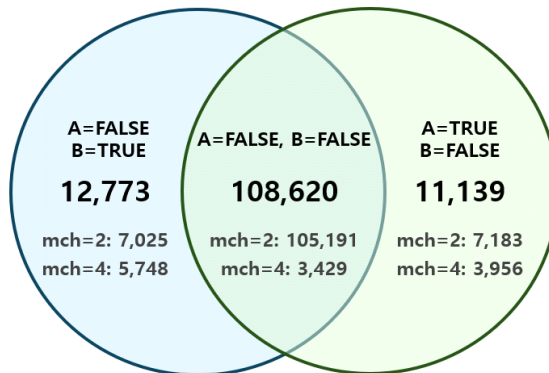
<그림5-2> 검수 대상의 세부 구성

[2024년 사망진단서 전체 684,841건]



↓ 검수대상(OR FALSE 영역) 132,532건의 세부 구성

조건 A ≠ 사전 A=FALSE, n=121,393 조건 B ≠ 사전 B=FALSE, n=119,759



* 검수 미대상 552,309 = mch=2(26,990) + mch=4(523,554) + 그 외 mch(1,765)

다. 검수 결과 집계

검수 결과는 조건 A, B의 예측 정오 조합에 따라 네 가지 판정 유형으로 분류하였고, 이 때 BOTH_WRONG으로 분류된 경우에는 조건 A, B 모두 정답으로 판단되지 않아 기존 사전분류 코드를 유지한 사례도 포함된다.

<표5-7> 검수 결과 판정 유형별 분포

판정 유형	의미	색인사전 연계	한글사전 연계	합계
BOTH_OK	A, B 모두 정답	388	14	402
PRED_A_OK	A만 정답	143	33	176
PRED_B_OK	B만 정답	250	56	306
BOTH_WRONG	A, B 모두 오답	413	29	442
합계		1,194	132	1,326

사전분류 코드와 조건 A, 조건 B의 정답판정 기준은 다음과 같다.

- ① 조건 A가 정답: 판정이 BOTH_OK 또는 PRED_A_OK인 경우
- ② 조건 B가 정답: 판정이 BOTH_OK 또는 PRED_B_OK인 경우
- ③ 사전분류가 정답: 판정이 BOTH_WRONG이 아니나, 정답인 AI 예측과 사전코드가 일치하는 경우, 또는 판정이 BOTH_WRONG인 경우

<표5-8> 조건별 정답 건수

매칭영역	사전 정답	조건 A 정답	조건 B 정답	검수 건수
색인사전	518 (43.4%)	531 (44.5%)	638 (53.4%)	1,194
한글사전	117 (88.6%)	47 (35.6%)	70 (53.0%)	132
전체	635 (47.9%)	578 (43.6%)	708 (53.4%)	1,326

따라서 본 검수에서 BOTH_WRONG 442건에는 실제 정답이 아니더라도 조건 A, B 가운데 어느 것도 정답으로 판단할 수 없어 사전분류 코드를 유지한 사례가 포함되어 있다. 한글 사전 매칭 영역에 대해선 기존 방식이 더 우수하긴 하나, 전체 검수 표본에서는 조건 B의 정답률이 53.4%로 가장 높게 나타났다.

라. 보정비율 산출 및 보정 후 성능 평가

검수 표본의 판정 결과를 조건 A 또는 조건 B가 불일치인 모집단 전체($A \cup B$)에 확대 적용하여 보정 비율을 산출하였다. 이는 사전분류 코드가 완전한 정답을 보장하지 않는다는 점을 고려하여, 사망원인에 대한 실제 사망원인 코드와 일치정도를 비교하기 위함이다. 이를 위하여 사전분류 및 $A \cdot B$ 모델에 대한 검수결과를 바탕으로 실제 정확도와 일치율을 추정하였다.

보정 비율과 보정 후 정확도 산출 절차는 다음과 같다. 보정 비율은 기존 판정이 오답(FALSE)이었으나 검수 결과 실제로는 정답(TRUE)이었던 비율($F \rightarrow T$)과 기존 판정이 정답(TRUE)이었으나 실제로는 오답(FALSE)이었던 비율($T \rightarrow F$)로 구분하여 산출하였다.

<표5-9> 보정 비율 산출결과

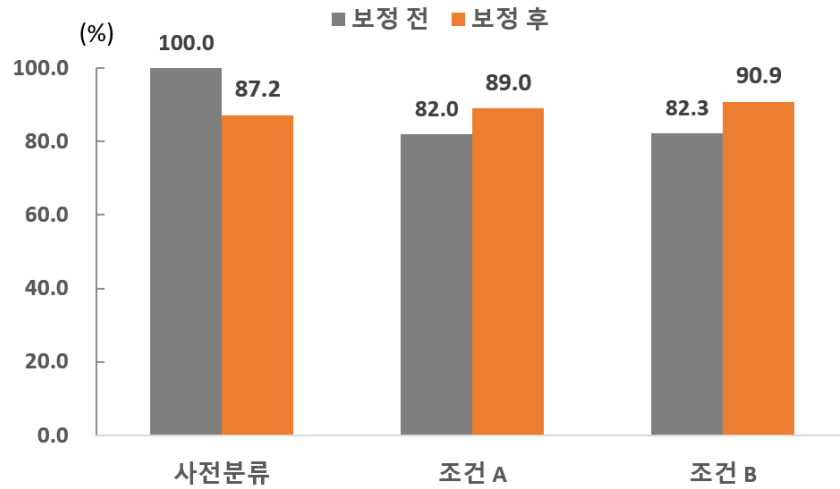
매칭영역	조건 A		조건 B	
	$F \rightarrow T$	$T \rightarrow F$	$F \rightarrow T$	$T \rightarrow F$
색인사전	43.2% (485/1,123)	35.2% (25/71)	51.3% (579/1,128)	10.6% (7/66)
한글사전	13.5% (12/89)	18.6% (8/43)	21.8% (17/78)	1.9% (1/54)
전체	41.0% (497/1,212)	28.9% (33/114)	49.4% (596/1,206)	6.7% (8/120)

여기서 두 보정 비율은 검수대상의 모집단인 $A \cup B$ 불일치 영역에만 적용되며, 조건 $A \cdot B$ 가 모두 사전분류와 일치했던 영역(552,309건)은 기존 판정을 유지한다. 이후 $A \cdot B$ 모델에 대한 매칭코드별 보정 비율을 이용하여 실제 정확한 사망원인 코드와의 일치율을 산정하였다. 별도로 사전분류는 매칭코드별 정확도를 검수하여 정확도를 산출하였다.

<표5-10> 보정 전·후 일치율 비교

매칭영역	사전분류 (보정 후)	조건 A		조건 B	
		보정 전	보정 후	보정 전	보정 후
색인사전	41.1%	23.3%	54.7%	23.2%	62.1%
한글사전	99.7%	98.3%	98.4%	98.6%	98.9%
전체	87.2%	82.0%	89.0%	82.3%	90.9%

<그림5-3> 보정 전·후 일치율 비교



보정 후 일치율은 다음 두 가지 방향의 결론을 제시한다.

첫째, 사전분류 코드를 정답으로 가정한 초기 평가에서는 예측모델의 실질 성능이 과소평가 되었음을 확인하였다. 매칭코드별로 살펴보면, 색인사전 매칭 영역에서 조건 B의 실질 일치율은 62.1%로 보정 전(23.2%) 대비 38.9%p 상승하였으며, 조건 A 역시 54.7%로 보정 전(23.3%)을 크게 상회하였다. 반면 한글사전 매칭 영역에서는 사전분류 (99.7%)가 여전히 조건 B(98.9%)와 조건 A(98.4%)보다 높으나, 그 격차는 1%p 이내로 좁혀졌다.

둘째, 전체 평가 대상에서도 조건 B의 보정 후 일치율은 90.9%로, 기존 사전분류와 조건 A보다 높게 나타났다. 특히 색인사전 매칭 영역에서 조건 B가 사전분류 대비 뚜렷한 우위를 보인 결과가 전체 결과에 반영되면서, AI 예측모델이 현행 사인선정시스템의 사전분류 기능을 대체할 수 있는 가능성을 보여주는 결과로 해석된다.

다만 본 검수는 1% 규모의 표본에 근거한 결과이며, BOTH_WRONG으로 분류된 442건에는 사전분류 유지 처리된 사례가 포함되어 있다는 한계가 있다. 향후 검수 규모를 확대하고, 오분류 사례에서 실제 정답을 확인하여 학습데이터를 추가로 정제한다면, 예측모델의 성능이 한층 더 향상될 것으로 기대된다.

4. ICD-11 기반 코드화 확장

4장 제2절에서 서술한 바와 같이, 본 연구는 향후 ICD-11 체계 적용에 대비하기 위하여 ICD-11 기반 코드화 모형을 별도로 구축하고 분석하였다. 다만 국내 사망원인 코드가 현재 ICD-10으로만 부여되어 있어 ICD-11 기준의 학습자료와 정답 라벨이 존재하지 않는다는 근본적 제약이 있다. 이에 자체 학습자료 구축 과정과 정성적 검수 결과를 중심으로 기술한다.

가. 학습자료 구축

ICD-11 기반 학습자료는 두 종류의 자료를 결합하여 구성하였다. 하나는 ICD-11 번역본에서 계층정보를 반영한 자료이며, 다른 하나는 2023년 실제 사망진단서 자료에서 ICD-10과 ICD-11이 1:1로 매칭되는 코드를 추출·병합한 학습 자료이다.

ICD-11 계층정보 반영 자료는 ICD-11 공식 한글 번역본이 완성되지 않은 상태에서 영어 코드집(ICD-11-MMS-en)을 기준으로 전처리 규칙을 통해 구축하였다. ICD-11 중간산출물에서 Kind=title 항목만을 추출하여 1~26장 범위로 한정하고, 코드의 점(dot) 개수와 코드 길이를 기준으로 6단계 Depth 체계(대/중/소/세/세세/세세세)를 부여하였다. 최종적으로 18,331개의 코드를 분석 대상으로 확정하였으며, 이 중 최하위코드(Leaf)는 15,115개로 판정되었다.

실제 사망진단서 기반 학습 자료는 2023년 사망진단서 654,101건 중 한글사전 매칭이 확정된 532,044건(MTCHG_CODE=4)을 대상으로 구축하였다. WHO에서 제공하는 '10To11 Map To Multiple Categories' 매핑표를 활용하여 Category 항목이면서 1:1 매칭이 성립하는 코드만을 추출하였다(15,630개 → 7,099개). 사망진단서 한 건에 복수의 사인이 기재된 경우 각 사인을 각각 별도의 열로 처리하되, 한 줄에 기재된 모든 사망원인이 ICD-11 코드로 변환된 사례만을 최종 학습자료로 채택하였다. 그 결과 532,044건 중 344,722건이 최종 학습자료로 산출되었다.

두 자료를 결합한 학습자료를 8:2 비율로 학습·검증 데이터로 분리하여 모델1과 동일한 구조를 적용하여 학습을 수행하였다. 인코더·손실함수·하이퍼파라미터 등 모델 구조는 ICD-10 기반 모형과 동일하며, 최종 분류 헤드의 코드 공간만 ICD-11 체계로 대체하였다.

나. 예측 수행 결과

학습된 모델을 2024년 사망진단서 684,841건에 대해 적용하여 ICD-11 코드 예측을 수행하였다. 예측 결과의 특성은 다음 두 가지 측면에서 분석하였다.

첫째, 예측 확신도의 저하가 두드러졌다. ICD-11의 경우 예측확률이 임계값 0.5 이상인 건의 비율이 76.6%로, 동일 임계값 기준의 ICD-10(98.2%) 대비 20%p 이상 낮았다. 이는 모델이 ICD-10에 비해 상대적으로 낮은 확신도로 예측을 산출하고 있음을 보여주며, 학습자료의 규모와 다양성이 충분히 확보되지 않았음을 반영하는 지표로 해석된다.

둘째, 코딩 개수 불일치가 상당한 수준으로 관찰되었다. 사망진단서 한 줄에서 산출되어야 할 코드 개수와 모델의 예측 결과가 일치하지 않는 사례가 684,841건 중 84,948건(12.4%)에 이르렀다. 이는 반드시 오류로 분류될 수밖에 없는 사례이며, 마찬가지로 모델의 예측 불안정성을 보여주는 지표로 해석된다.

다. 표본검수 결과 및 분석사례

평가데이터인 2024년 예측 결과의 경우 정답 라벨이 부재한 상태이므로 예측 결과를 검증하기 위하여, 전체 684,841건 중 0.1%인 685건을 표본으로 추출하여 직접 검수를 수행하였다. 검수 결과 407건이 정분류로 판정되어 정확도는 59.4%로 나타났다.

<표5-11> ICD-11 기반 사인코드 예측 오분류 대표 유형

사망진단서 사인	ICD-10	ICD-11	판정	오류성격
영양부전	E63.9	CB41.2Z(상세불명의 호흡부전)	오분류	무관 코드 예측
다발성 장기부전	R68.8	2C77.Z(상세불명의 자궁경부 악성신생물)	오분류	징후를 질환으로 과다예측
호흡 부전	J96.99	2B90.Z(상세불명의 결장 악성신생물)	오분류	무관 코드 예측
자궁암으로 인한 사망	C55	2D10	오분류	최하위(Leaf) 코드 미도달
뇌간기능부전	G93.88	KB07(신생아의 뇌 압박)	오분류	부적절 계통(연령) 오예측
직장구불결장접합부 악성신생물	D48.9	2B91.Z(직장구불결장접합부 악성신생물)	정분류	기존 ICD-10 라벨 오류 교정

검수 사례의 대부분은 학습자료의 부족으로 인해 징후를 질환으로 과다예측하거나 사망진단서의 사인과 무관한 코드를 예측하는 형태의 오분류로 나타났다. 다만 '직장구불결장접합부 악성신생물'의 경우 기존 ICD-10 코드 D48.9(행동양식 불명 신생물)가 정답이 아니었음에도 ICD-11 예측은 2B91.Z로 정확하게 산출되어, 기존 ICD-10 라벨의 오류에도 올바르게 예측한 사례로 확인되었다.

이상의 검수 결과는 두 가지 결론을 시사한다. 첫째, ICD-11 기반 모델의 성능 저하는 모델 구조의 문제가 아니라 학습자료의 절대적 부족에 기인한다. ICD-10 기반 모델에서 확보 가능했던 학습자료(532,044건)에 비해 ICD-11 기반 모델의 학습자료(344,722건)는 규모가 감소하였을 뿐 아니라, 1:1 매칭 조건으로 인해 특정 코드에 편중된 구성을 가진다.

둘째, 본 연구에서 적용한 방법은 정답 라벨이 없는 환경에서도 ICD-11 체계에 대한 예측모형을 구축할 수 있음을 확인하였으며, 일부 사례에서는 기존 ICD-10 라벨의 오류를 올바르게 예측한 사례도 확인되었다. 향후 KCD 제10차 개정(2027년 7월 고시 목표)에 맞추어 ICD-11 기반의 정답 라벨 자료가 축적되면, 본 연구에서 구축한 학습 방법을 바탕으로 학습자료를 확대함으로써 실무 적용이 가능한 수준으로 성능을 향상시킬 수 있을 것으로 기대된다.

5. 종합해석

가. 분석결과 요약

첫째, 표본검수를 통해 확인된 조건 B의 보정 후 일치율(90.9%)이 사인선정시스템의 사전분류(87.2%)보다 높았다. 초기 평가에서 82%대에 그쳤던 조건 B의 일치율은 사전 분류 코드를 정답으로 가정한 상대적 평가 결과였다. 그러나 사망진단서 원문을 기준으로 재판정한 결과, 조건 B의 실질 일치율이 사전분류보다 높은 것으로 확인되었다. 이는 AI 예측모델이 사인분류 실무의 초기 단계에서 사전분류 기능을 대체할 수 있는 수준에 이르렀음을 실증적으로 보여준다.

둘째, 매칭 경로별로 예측모델의 위상이 뚜렷이 갈린다. 색인사전 매칭 영역에서는 조건 B(62.1%)가 사전분류(41.1%)를 21%p 상회하여 예측모델이 상대적으로 높은 성능을 보인 반면, 한글사전 매칭 영역에서는 사전분류(99.7%)가 조건 B(98.9%)보다 여전히 높으나 격차가 1%p 이내로 좁혀졌다. 두 영역에서 상이한 방향의 결과가 도출된 것은, 예측모델의 실무 적용은 사전분류를 일괄적으로 대체하기보다 매칭 경로별 특성에 맞춘 단계적 도입으로 접근되어야 함을 시사한다.

셋째, 오분류의 발생 구조는 모델 자체가 아니라 학습자료의 구성과 처리 정책에 있음이 확인되었다. 조건 A의 오분류 사례에서 확인된 세 유형은 모두 모델 구조나 학습 알고리즘의 문제가 아니라 학습자료 필터링 정책의 부작용, 토큰화의 공백 민감성, 강제 출력 정책의 세 가지 요인에 기인하는 것으로 분석되었다. 이는 향후 성능 개선의 여지가 예측모델의 재설계보다는 학습자료와 전처리·후처리 규칙의 정비에 있음을 의미한다.

넷째, ICD-11 기반 코드화 확장은 적용 가능성을 탐색하는 수준의 결과를 확인하였다. 정답 라벨이 부재한 조건에서 자체적으로 학습자료를 구축하고 예측모형을 구축할 수 있음을 확인하였으나, 학습자료의 절대적 부족으로 실무 적용 수준의 정확도(59.4%)에는 아직 도달하지 못하였다.

나. 향후 과제

오분류 분석 결과를 반영한 학습자료 개선이 우선적으로 추진될 필요가 있다. 학습자료가 한글사전 완전 매칭영역(MTCHG_CODE=4)에 한정된 현행 방식은 고빈도 표준표기의 정답 라벨이 색인사전 매칭 영역에 존재하는 경우 학습에서 배제되는 부작용을 낳는다.

색인사전 매칭영역(MTCHG_CODE=2)의 자료를 학습에 포함하는 필터링 정책의 검토, 오류 패턴에 대응하는 정답 자료의 추가 반영, 학습·추론 단계에 동일한 공백표준화 규칙을 적용하는 방안, '해당없음', '공란'과 같은 입력에 대한 별도 후처리 규칙의 도입 등이 검토 대상이 된다. 아울러 표준 사인 표기와 코드명 등의 추가 결합을 통한 학습자료의 다각적 확장 역시 예측 안정성을 높이기 위한 방안으로 검토할 수 있다.

표본검수 규모의 확대와 학습자료의 지속적인 정제도 중요한 후속 과제이다. 본 절의 실증 결과는 조건 A 또는 조건 B에서 모델 예측값과 기존 사전기반 코드가 불일치한 약 13만 건의 1% 표본만 살펴보았다. 이에 검수 표본의 규모와 판정 처리 방식에서 비롯되는 한계가 존재한다. 이에 검수 규모를 확대하거나 검수 과정에서 실제 정답 코드를 함께 구축하여 학습데이터를 추가 정제한다면 예측모델의 성능이 한층 더 향상될 것으로 기대된다.

ICD-11 기반 모형의 후속 연구는 학습자료의 축적에 있다. ICD-11 기반의 정답 라벨 자료가 축적되면, ICD-10 대상으로 구축한 학습 방법을 바탕으로 학습자료를 확대함으로써 실무 적용이 가능한 수준으로 성능을 향상시킬 수 있을 것으로 기대된다.

제2절 (모델2) 행정자료 반영 잠정 원사인 예측 모형

1. 예측모델별 평가

가. 예측모델 개요

본 연구는 제4장에서 서술한 바와 같이 약 30차례의 학습데이터 구성 및 결과 검토를 반복함으로써 모델구조·손실함수·하이퍼파라미터를 고도화하였으며, 이를 바탕으로 3개의 예측모델 후보를 선정하였다. 세 모델은 동일한 기초자료(사망진단서 텍스트 항목, 잠정 원사인코드, 건강보험 수진기록 등)를 사용하되, 학습데이터와 정답라벨의 코드 단위를 다르게 구성하였다.

본 연구에서는 인코딩 모델과 학습·정답라벨 코드 단위를 따서 예측모델을 CNN_full, GCN_full, GCN_104로 명명하기로 하였다. 또한 예측모델의 성능을 비교하기 위한 기준으로 잠정 원사인(사망진단서 기반)의 예측 결과를 함께 검토하였다.

<표5-12> 예측모델별 인코더 구조 및 학습·정답라벨 코드 단위

모델	인코더 구조	학습 · 정답라벨 코드 단위
CNN_full	KcELECTRA + CNN	KCD 세부코드(full code)
GCN_full	KcELECTRA + GCN	
GCN_104		104항목 분류

CNN_full과 GCN_full은 원천 KCD 코드체계를 그대로 유지한 채 세부코드 단위로 학습·예측하는 반면, GCN_104는 104항목 분류로 사전 매핑한 후 해당 분류체계를 기준으로 학습 및 평가한다.

CNN과 GCN의 본질적인 차이는 KCD 코드 간의 위계관계를 모델 구조에 반영하는지 여부에 있다. CNN은 각 코드를 동등한 분류 대상으로 취급하는 반면, GCN은 KCD 코드의 계층 구조를 그래프로 표현하고, 그래프 상에서 정보를 전파함으로써 코드 간 인접관계와 상하위 관계를 함께 학습한다. 이 구조적 차이로 인해 GCN은 학습데이터가 적은 희소코드에 대해서도 관련 코드의 정보를 활용할 수 있어 예측성능 향상에 유리한 특성을 가진다.

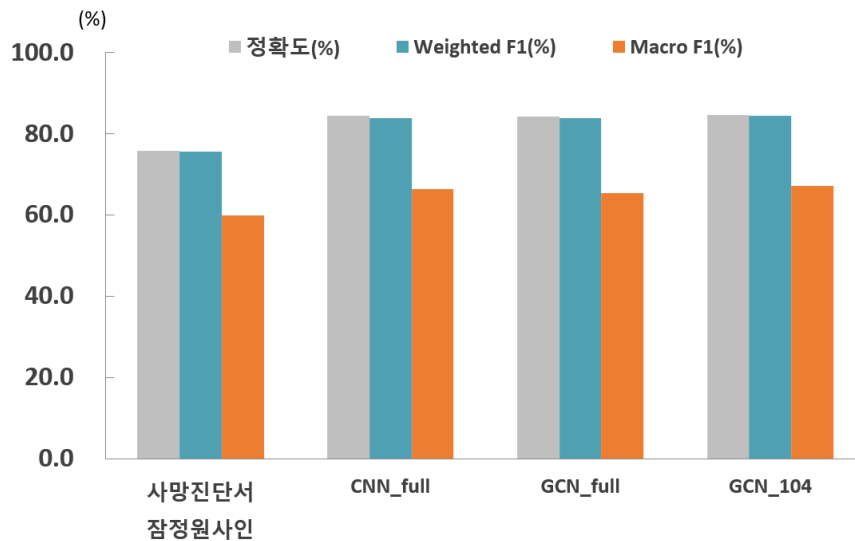
나. 잠정 원사인과 예측모델들 간 성능비교

예측모델의 성능을 평가함에 앞서 비교기준 단위는 104항목 분류로 통일하였다. 이는 GCN_104 모델이 104항목 분류로 학습 및 평가되었기 때문에 세부코드(full code) 기준으로는 GCN_104와 다른 두 모델(CNN_full, GCN_full)의 성능을 동일한 기준에서 비교하기 어렵기 때문이다. 평가지표는 정확도, Weighted F1, Macro F1 등 세 가지를 사용한다.

<표5-13> 잠정 원사인과 예측모델별 성능비교(정확도, Weighted F1, Macro F1)

모델	정확도(%)	Weighted F1(%)	Macro F1(%)
잠정 원사인	75.8	75.6	59.9
CNN_full	84.4	83.9	66.5
GCN_full	84.2	83.8	65.5
GCN_104	84.6	84.4	67.2

<그림5-4> 잠정 원사인과 예측모델별 성능비교(정확도, Weighted F1, Macro F1)



세 예측모델은 모든 평가항목에서 잠정 원사인보다 우수한 성능을 보였다. 정확도를 기준으로 예측모델의 성능은 84.2~84.6%이고 잠정 원사인의 정확도는 75.8%로 나타나 예측모델이 잠정 원사인보다 약 8~9%p 높은 정확도를 보인다. 이는 예측모델이 잠정 원사인뿐만 아니라, 사망진단서 텍스트와 건강보험 수진기록 등 추가 정보를 함께 활용하여 예측함으로써 실제 최종 원사인에 더 근접하게 예측하고 있음을 의미한다.

세 예측모델 간에는 GCN_104가 모든 평가지표에서 가장 우수한 성능을 보였다(정확도 84.6%, Weighted F1 84.4%, Macro F1 67.2%). 특히 Macro F1은 GCN_104가 CNN_full 대비 0.7%p, GCN_full 대비 1.7%p 높게 나타났다. Macro F1은 클래스의 코드 출현빈도와 관계없이 모든 클래스를 동일한 비중으로 평균하여 산출하는 지표이므로, 이러한 차이는 GCN_104가 다른 모델에 비해 희소코드에 대한 예측성능이 상대적으로 우수함을 의미한다.

다. 최종 예측모델 선정 및 근거

앞서 살펴본 바와 같이 세 예측모델의 정확도, Weighted F1, Macro F1의 차이는 0.4~1.7%p 수준으로 크지 않다. 따라서 정량적 지표만으로 모델 간 우열을 명확히 판단하기는 어렵다. 이에 본 연구에서는 기초자료의 계층형 성격을 가장 잘 반영하고, 희소코드에 대한 예측성능을 확보할 수 있는 모델 구조를 우선한다는 원칙에 따라 GCN 계열을 최종 예측모델로 채택하였다.

GCN 계열 모델 간 비교에서는 GCN_104가 균형성능을 나타내는 Macro F1과 전반적인 정확도 측면에서 GCN_full보다 근소하게 우수한 성능을 보였다. 그러나 본 연구의 목적은 104항목 분류 수준의 잠정 사망원인통계 공표가능성을 확보하는 데 그치지 않고, 현장에서 사인분류사의 사인분류 업무를 지원하는 것까지 포함한다. 이에 따라 시스템의 최종 구현형태는 다음과 같이 두 가지 활용목적을 전제로 설계되었다.

<표5-14> 잠정 원사인 예측시스템의 최종 구현형태

구현형태	활용
공표용(104항목 분류 기준)	잠정 사망원인통계 공표, 감염병 감시체계 등에 활용
실무 지원용(세부코드)	사인분류사가 실제 최종 원사인을 확정하는 과정에서 참고자료로 활용. 예측코드 Top1~3의 세부KCD 코드와 확률에 대해 시스템을 통해 함께 제시

실제로 사인분류사 11명을 대상으로 한 의견수렴 결과에서는 AI가 사망진단서, 잠정 원사인 및 건강보험 수진기록을 종합하여 수정원사인을 제시하는 방식에 대해 6명은 ‘많이 도움됨’, 4명은 ‘어느정도 도움됨’으로 응답하여 실무에서의 활용수요를 확인하였다. 다만 응답자들은 AI가 제시한 예측결과를 그대로 채택하기보다는 최종 검토와 보완과정이 필요하다는 의견을 공통적으로 제시하였다.

이처럼 시스템의 최종 산출물이 104항목 분류 수준의 요약정보뿐 아니라 세부코드 수준의 예측결과도 함께 제공할 필요가 있다. 이러한 점을 고려하면 세부코드 체계를 그대로 유지한 상황에서 학습과 예측을 수행하는 GCN_full이 실무지원 목적에 보다 적합하다. 이에 따라 본 연구에서는 GCN_full을 심층분석 및 시스템 구현의 기준모델로 채택하였다.

GCN_full을 기준모델로 선택함에 따른 성능저하는 제한적이다. GCN_full의 세부코드 예측결과를 104항목 분류 기준으로 변환하여 평가한 성능(정확도 84.2%, Weighted F1 83.8%, Macro F1 65.5%)은 104항목 분류로 직접 학습한 GCN_104(정확도 84.6%, Weighted F1 84.4%, Macro F1 67.2%)와 비교할 때, 정확도와 Weighted F1은 최대 1.7%p의 차이에 불과하다. 즉 GCN_full은 세부코드 단위의 실무 활용성을 확보하면서도 104항목 분류 기준의 통계적 예측성능 역시 GCN_104와 유사한 수준을 유지한다.

따라서 본 연구에서는 GCN계열 모델을 최종 예측모델로 선정하였다. 다만 활용 목적에 따라 결과를 구분하여 제시한다. 공표용 성능은 104항목 분류를 기준으로 학습한 GCN_104의 결과를 제시하고, 예측결과의 특성분석, 오분류패턴 분석 및 시스템 구현은 세부코드 수준의 예측이 가능한 GCN_full을 기준으로 수행하였다. 이를 통해 통계 공표의 정확성과 실무활용성을 모두 확보하고자 하였다.

2. 예측결과의 특성분석

가. 모델 예측코드의 출처

GCN_full 모델의 1순위 예측코드(Top1)가 기초자료의 어느 항목에 존재하는지를 확인하여, 모델이 실제로 어떠한 정보에 근거하여 원사인을 예측하는 경향이 있는지를 분석하였다. 비교 단위는 세부코드(full code) 기준이며, 출처는 예측코드와 각 기초자료의 코드값이 세부코드 수준에서 일치하는지를 기준으로 판정하였다. 비교 대상이 되는 기초자료는 다음의 네 가지이다.

1순위 예측코드(Top1)의 출처를 살펴보면, 잠정 원사인이 62.1%로 가장 높은 비중을 차지하였으며, 주상병(34.9%)이 그다음을 차지하였다. 반면 부상병(12.3%)과 사망진단서 텍스트(20.6%)는 상대적으로 낮은 비중을 보였다. 예측 범위를 Top1~Top3까지 확대하면 잠정 원사인의 비중은 79.7%로 증가하였으며, 주상병(47.1%)과 부상병(20.6%) 역시 증가하는 경향을 보였다. 반면 사망진단서 텍스트는 10.5%로 나타났다. 이는 모델이 Top2와 Top3 후보를 생성하는 경우에도 대부분 잠정 원사인, 주상병, 부상병 등 기초자료에 존재하는 코드를 후보로 제시하는 경향이 있음을 보여준다.

<표5-15> 모델에 활용한 기초자료 출처와 비교기준

기초자료 출처	비교기준
잠정 원사인	사망진단서 선정규칙에 따라 작성된 잠정 원사인코드(T_DTH_CODE)과 일치여부
주상병	사망 전후 4개 시점의 건강보험 주상병 코드 (JUSAN_1LA/2LA/MO/LAMO) 중 하나 이상과 일치여부
부상병	사망 전후 4개 시점의 건강보험 부상병 코드 (BUSAN_1LA/2LA/MO/LAMO) 중 하나 이상과 일치여부
사망진단서 텍스트	상기 항목 어디에도 없어서 사망진단서 텍스트(가·나·다·라 사인)에서 모델이 직접 추론했을 것으로 추측되는 경우

<표5-16> 모델 예측코드(Top1~Top3)의 출처별 빈도 및 비중

구분	N	잠정 원사인	주상병	부상병	사망진단서 텍스트
Top1	358,569	222,564 (62.1%)	125,011 (34.9%)	43,913 (12.3%)	73,926 (20.6%)
Top1~Top3		285,714 (79.7%)	169,000 (47.1%)	73,725 (20.6%)	37,755 (10.5%)

각 항목의 비중을 합산하면 100%를 초과하는데, 이는 하나의 예측코드가 잠정 원사인과 주상병 등 둘 이상의 기초자료에 동시에 존재하는 경우가 있기 때문이다. 다만 본 분석은 예측코드가 어느 기초자료의 코드값과 문자열 수준에서 일치하는지를 확인한 결과이며, 모델이 실제로 어떠한 정보를 근거로 해당 코드를 예측하였는지를 직접 규명한 것은 아니라는 점에 유의할 필요가 있다. GCN_full 모델은 사망진단서 텍스트를 처리하는 KcELECTRA 인코더와 코드 정보를 처리하는 인코더를 결합한 구조이며, 학습 과정에서 두 인코더의 출력이 최적화된다.

따라서 예측코드가 잠정 원사인, 주상병 또는 부상병과 같은 자료와 일치하더라도 실제 예측 과정에서는 사망진단서 텍스트 정보가 함께 영향을 미쳤을 가능성이 있다. 반대로 사망진단서 텍스트 출처로 분류된 경우 역시 텍스트 정보만으로 예측되었다고 단정하기는 어렵다. 오히려 범주형 정보가 부족하거나 충분한 단서를 제공하지 못하는 상황에서 텍스트 정보의 영향력이 상대적으로 커진 결과로 해석하는 것이 타당하다.

나. 정답라벨의 출처

앞에서는 모델의 예측코드(Top1)가 어느 기초자료의 코드와 일치하는지를 살펴보았다. 이번에는 학습·평가데이터의 정답라벨(T_DTH_WON_CODE)이 기초자료의 어느 항목에 존재하는지를 확인하여, 학습·평가데이터에서 정답라벨과 기초자료 간의 관계를 분석하고자 한다. 분석 대상은 GCN_full의 평가데이터이며, 출처 판정은 세부코드(full code) 기준에서 정답라벨과 각 기초자료의 코드 일치 여부를 기준으로 하였다.

전체 평가데이터에서 정답라벨은 잠정 원사인(49.6%)과 가장 많이 일치하였으며, 주상병(34.9%), 부상병(13.4%) 순으로 나타났다. 한편, 어떠한 범주형 기초자료와도 일치하지 않는 '출처없음'의 비중이 28.0%에 달한다는 점도 주목할 필요가 있다. 이는 정답라벨의 상당수는 잠정 원사인이나 주상병 등 기초자료에서 일정한 단서를 얻을 수 있지만, 약 1/3에 가까운 사례는 이러한 정보만으로는 설명되지 않음을 의미한다.

정분류와 오분류를 구분하여 살펴보면 출처 분포의 차이가 더욱 뚜렷하게 나타난다. 잠정 원사인과 일치하는 비중은 정분류에서 68.6%, 오분류에서 13.8%로 약 5배의 차이를 보였다. 반면 '출처없음'의 비중은 정분류에서 12.0%, 오분류에서 58.9%로 오분류에서 약 5배 높게 나타났다. 즉, 정답라벨이 잠정 원사인이나 주상병과 일치하는 사례일수록 모델이 정답을 예측할 가능성이 높았으며, 반대로 정답라벨이 범주형 기초자료 어디에도 존재하지 않는 사례일수록 오분류 될 가능성이 크게 증가하였다.

한편, 부상병은 잠정 원사인이나 주상병만큼 뚜렷한 차이를 보이지 않았다. 이는 부상병이 원사인을 직접적으로 나타내는 정보라기보다 원사인과 관련된 보조 정보의 성격을 가지므로, 원사인 예측에 대한 기여도가 상대적으로 낮기 때문으로 해석할 수 있다.

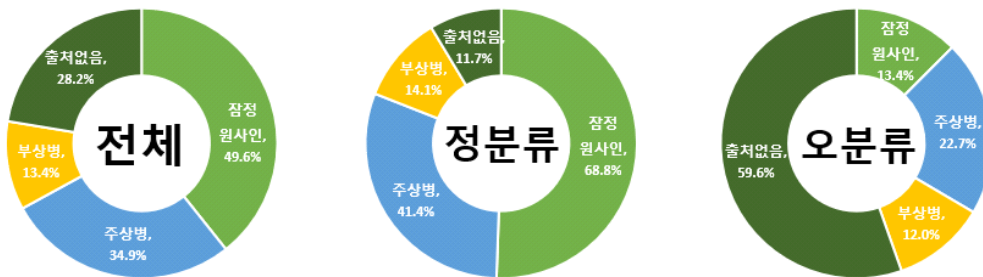
이러한 결과는 정답라벨이 범주형 기초자료, 특히 잠정 원사인이나 주상병과 일치하는 사례일수록 모델의 예측 정확도가 높다는 상관관계를 보여준다. 다만 이를 범주형 기초자료 자체의 효과로 단정하기는 어렵다. 정답라벨이 잠정 원사인이나 주상병과 일치하는 사례는 사인이 비교적 명확하여 WHO 선정규칙에 따른 잠정 원사인과 건강보험 진료기록이 최종 원사인과 잘 부합하는, 즉 상대적으로 판단이 용이한 사례일 가능성이 있다.

반대로 '출처없음'에 해당하는 사례는 사인분류사가 사망진단서 텍스트와 기타 정황을 종합적으로 검토하여 범주형 기초자료에는 존재하지 않는 최종 원사인을 결정한 판단 난도가 높은 사례일 가능성이 있다. 따라서 코드 일치 여부와 모델의 예측 정확도는 모두 사례 자체의 판단 난이도의 영향을 받았을 가능성이 있으며, 이에 대해서는 다음 절에서 기초자료의 난이도 분석을 통해 추가 검토한다.

<표5-17> 정답라벨의 기초자료 출처별 빈도 및 비중

구분 ¹⁾	N	빈도(건)				비중(%)			
		잠정 원사인	주상병	부상병	출처 없음	잠정 원사인	주상병	부상병	출처 없음
전체	358,569	178,811	125,282	48,019	100,448	49.6	34.9	13.4	28.2
정분류	235,970	161,931	97,208	33,209	28,206	68.6	41.2	14.1	12.0
오분류	122,599	16,880	28,074	14,810	72,242	13.8	22.9	12.0	58.9

<그림5-5> 정답라벨의 기초자료 출처별 비중



1) 정답라벨과 예측코드의 세부코드 단위 일치여부. 같은경우 정분류, 다른경우 오분류로 명시

3. 질환별 평가결과 분석

가. 10대 사망원인별 결과

앞서 GCN_full 모델을 기준으로 전체 예측 결과의 특성을 분석하였다. 여기서는 실제 통계 공표에 사용되는 104항목 분류를 기준으로 학습한 GCN_104 모델을 대상으로 10대 사망원인별 예측 성능을 살펴본다. 이를 통해 실제 통계 공표 기준에서 질환별 예측 성능과 신뢰도가 어떻게 나타나는지 확인하고, 나아가 실무 적용 가능성을 평가하고자 한다.

평가지표는 정밀도(Precision), 재현율(Recall), F1-score, 사망자 수 추정비의 네 가지를 사용하였다. 이 중 사망자 수 추정비는 AI 모델의 예측 건수를 실제 사망자 수로 나눈 값으로, 다음과 같이 산출하였다.

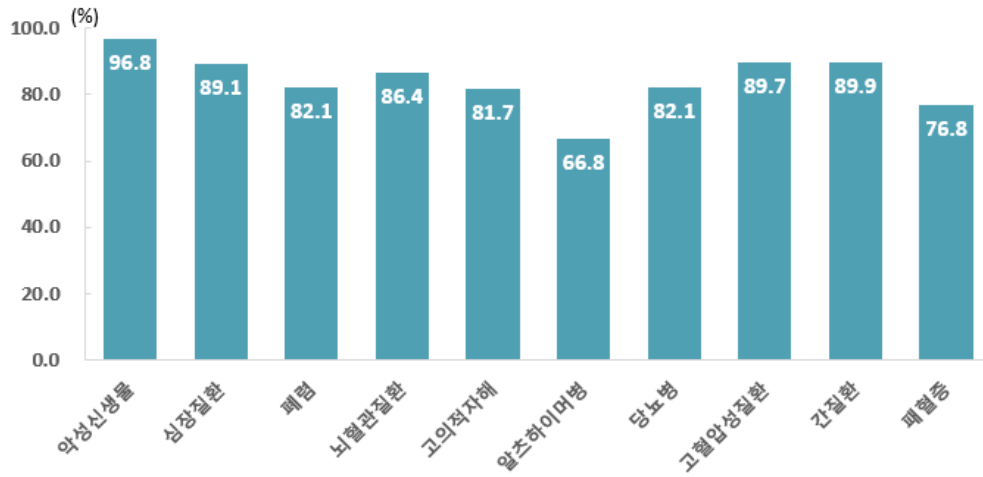
$$\cdot \text{사망자 수 추정비} = \frac{\text{AI예측 건수}}{\text{실제 사망자 수}} \times 100$$

사망자 수 추정비가 100보다 크면 AI 모델이 해당 질환의 사망자 수를 실제보다 많이 예측하는 과대추정(overestimation) 경향을, 100보다 작으면 실제보다 적게 예측하는 과소추정(underestimation) 경향을 의미한다.

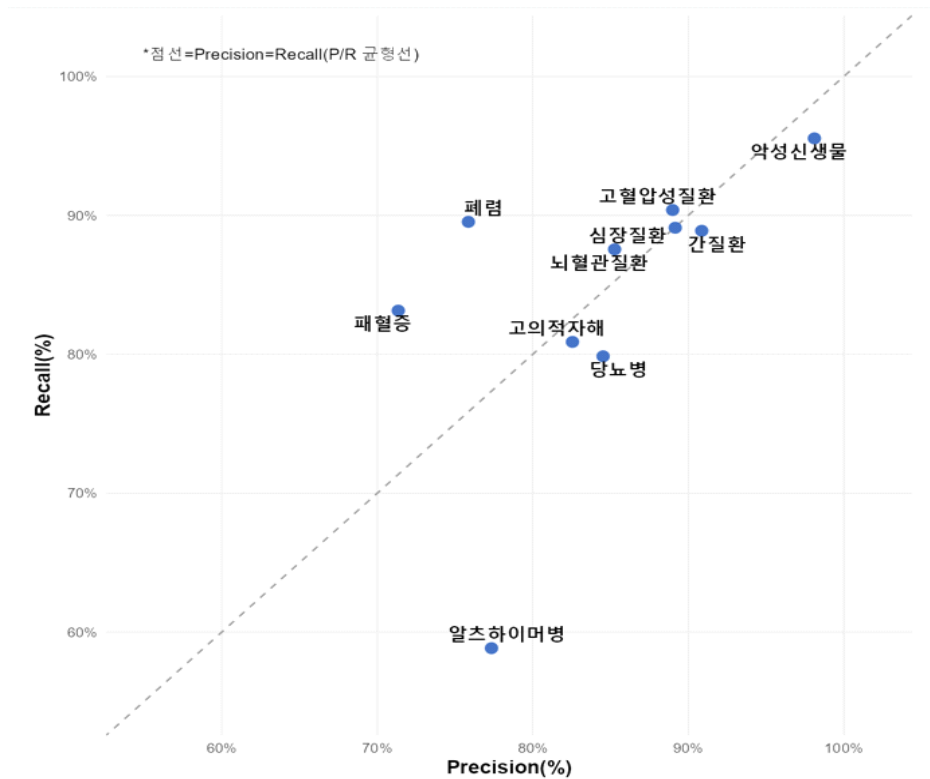
<표5-18> 10대 사망원인별 사망자 수, AI 예측 건수 및 예측정확도(104항목 기준)

질환	사망자 수	AI예측 건수	일치 건수	정밀도 (Precision)	재현율 (Recall)	F1 Score	사망자 수 추정비
악성신생물	88,933	86,606	84,970	98.1	95.5	96.8	97.4
심장질환	33,539	33,520	29,886	89.2	89.1	89.1	99.9
폐렴	30,103	35,530	26,952	75.9	89.5	82.1	118.0
뇌혈관질환	24,612	25,274	21,549	85.3	87.6	86.4	102.7
고의적자해	14,872	14,573	12,030	82.5	80.9	81.7	98.0
알츠하이머병	12,223	9,299	7,192	77.3	58.8	66.8	76.1
당뇨병	11,064	10,453	8,835	84.5	79.9	82.1	94.5
고혈압성질환	8,229	8,358	7,438	89.0	90.4	89.7	101.6
간질환	7,787	7,618	6,922	90.9	88.9	89.9	97.8
패혈증	7,730	9,008	6,427	71.3	83.1	76.8	116.5

<그림5-6> 10대 사망원인별 예측정확도(F1-Score)



<그림5-7> 10대 사망원인별 정밀도(Precision) 및 재현율(Recall)



10대 사망원인 중 사망자 수가 가장 많은 악성신생물은 정밀도(98.1%), 재현율(95.5%), F1-score(96.8%) 등 모든 평가지표에서 가장 우수한 성능을 보였으며, 사망자 수 추정비 역시 97.4%로 실제 사망자 수에 매우 근접하였다. 심장질환, 뇌혈관질환, 고혈압성질환, 간질환 또한 정밀도와 재현율이 85~91% 수준으로 높게 나타났으며, 사망자 수 추정비도 97~103% 범위에 분포하여 실제 사망자 수와의 차이가 크지 않았다. 따라서 이들 질환군은 현재 모델의 성능만으로도 비교적 높은 신뢰성을 확보하고 있어 실무 활용 가능성이 높은 것으로 판단된다.

반면 알츠하이머병은 정밀도(77.3%)에 비해 재현율(58.8%)이 크게 낮아 F1-score가 66.8%로 10대 사망원인 중 가장 낮게 나타났다. 재현율이 낮다는 것은 실제 알츠하이머병으로 확정된 사례 가운데 상당수를 모델이 다른 질환으로 예측하였음을 의미하며, 이러한 경향은 사망자 수 추정비(76.1%)에서도 확인된다. 즉, 모델이 알츠하이머병을 과소추정하는 경향을 보이며, 이를 통계 공표에 그대로 활용할 경우 실제 사망자 수보다 적게 집계될 가능성이 있음을 시사한다.

반대로 폐렴과 폐혈증은 상반된 양상을 보였다. 정밀도는 각각 75.9%와 71.3%로 10대 사망원인 중 가장 낮은 수준이었으나, 재현율은 각각 89.5%와 83.1%로 비교적 높게 나타났다. 또한 사망자 수 추정비는 각각 118.0%와 116.5%로 실제보다 약 16~18% 많은 사망자를 예측하는 과대추정 경향이 확인되었다.

이는 모델이 폐렴과 폐혈증을 다른 질환의 합병증이나 임종기 상태까지 원사인으로 판단하는 경향이 있을 가능성을 시사한다. 폐렴과 폐혈증은 다양한 기저질환의 진행 과정에서 흔히 동반되는 임상 상태이므로, 사망진단서 텍스트나 건강보험 진료기록에 관련 소견이 함께 기록된 경우 이를 원사인으로 우선 예측하였을 가능성이 있다. 따라서 이들 질환은 정밀도 향상을 위한 추가적인 보완이 필요한 질환군으로 판단된다.

종합하면 악성 신생물, 심장질환, 뇌혈관질환, 고혈압성질환, 간질환은 정밀도, 재현율 및 사망자 수 추정비가 모두 양호하여 실무 및 통계 공표에 활용할 수 있는 수준의 예측 성능을 보였다. 반면 알츠하이머병은 과소추정 경향, 폐렴과 폐혈증은 과대추정 경향이 뚜렷하게 나타났으며, 이들 질환은 사망자 수 추정비가 100%에서 유의하게 벗어나므로 통계 공표 시 별도의 보정이나 추가 검토 절차를 적용할 필요가 있는 질환군으로 판단된다.

나. 감염병 질환별 결과

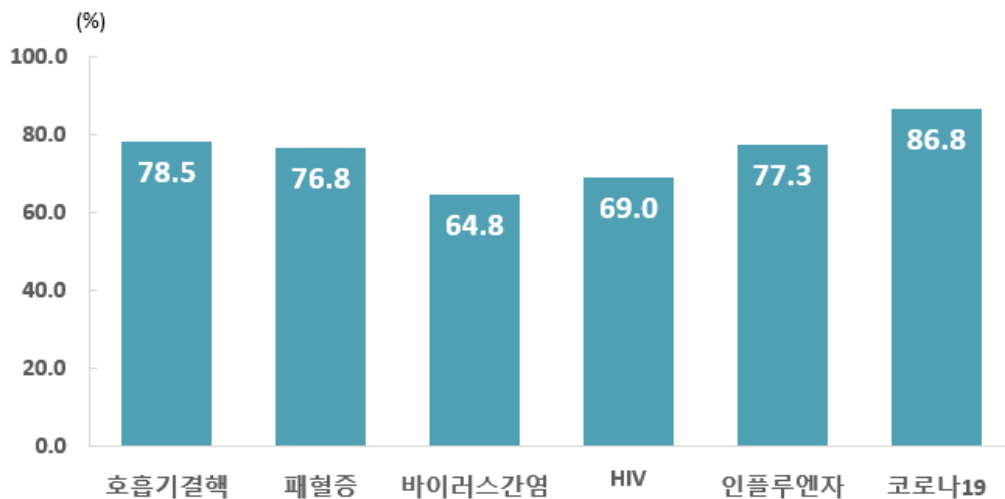
감염병 사망원인(호흡기결핵, 패혈증, 바이러스간염, HIV, 인플루엔자, 코로나19)의 예측 성능을 검토한다. 감염병은 절대적인 사망자 수가 상대적으로 적고 계절적 유행이나 감염병 확산 상황에 따라 월별 발생 규모의 변동성이 크다는 특성이 있다. 따라서 연간 성능뿐만 아니라 월별 시계열 추이를 함께 분석하여, 실시간 감염병 사망 감시체계 등에서의 활용 가능성을 평가하고자 한다.

분석 단위는 104항목 분류를 기준으로 하였으며, 평가지표는 정밀도(Precision), 재현율(Recall), F1-score, 사망자 수 추정비의 네 가지를 사용하였다.

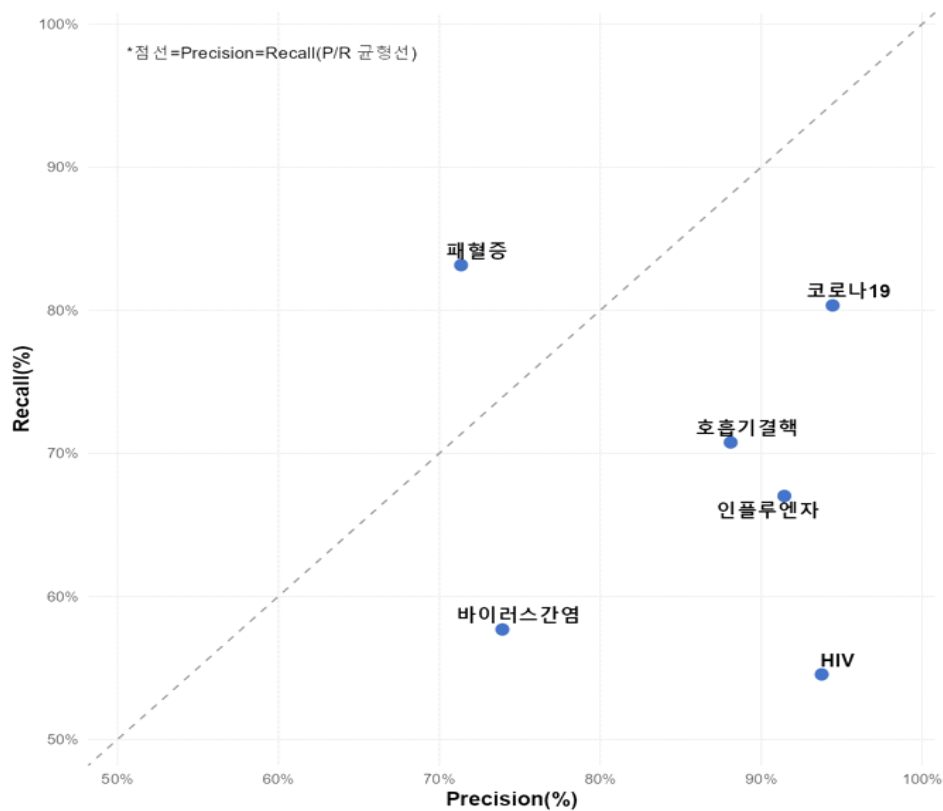
<표5-19> 감염병 질환별 사망자 수 및 예측정확도(104항목 기준)

감염병 질환	사망자 수	AI예측 건수	일치 건수	정밀도 (Precision)	재현율 (Recall)	F1 Score	사망자 수 추정비
호흡기결핵	1,255	1,008	888	88.1	70.8	78.5	80.3
패혈증	7,730	9,008	6,427	71.3	83.1	76.8	116.5
바이러스간염	501	391	289	73.9	57.7	64.8	78.0
HIV	55	32	30	93.8	54.5	69.0	58.2
인플루엔자	382	280	256	91.4	67.0	77.3	73.3
코로나19	3,208	2,729	2,577	94.4	80.3	86.8	85.1

<그림5-8> 감염병 질환별 예측정확도(F1-Score)



<그림5-9> 감염병 질환별 예측정확도(Precision, Recall)



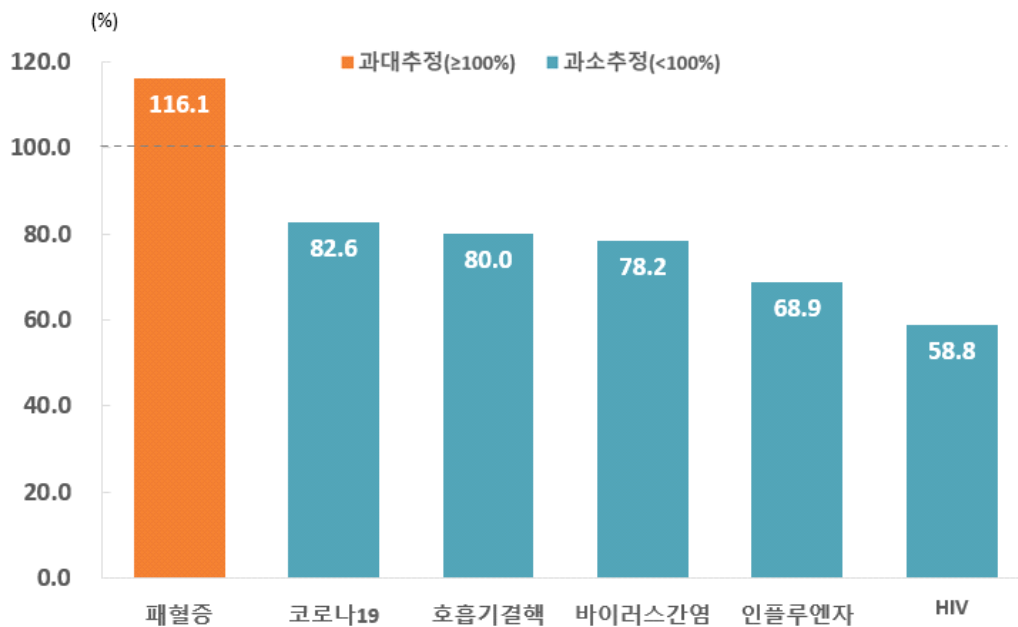
10대 사망원인과 비교하면 감염병 질환군의 전반적인 예측 성능은 다소 낮은 수준으로 나타났다. 이는 감염병 질환 대부분이 학습데이터에서 발생빈도가 낮은 희소코드에 해당하여, 모델이 해당 질환의 특성을 충분히 학습할 기회가 상대적으로 적었기 때문으로 판단된다.

다만 감염병 질환군 내에서도 질환별 성능 차이는 뚜렷하게 나타났다. 코로나19(F1-score 86.8%), 호흡기결핵(78.5%), 패혈증(76.8%)은 비교적 높은 예측 성능을 보인 반면, 바이러스간염(64.8%)과 HIV(69.0%)는 감염병 질환군 내에서도 상대적으로 낮은 성능을 나타냈다.

<표5-20> 감염병 질환별 월평균 사망자 수 추정비

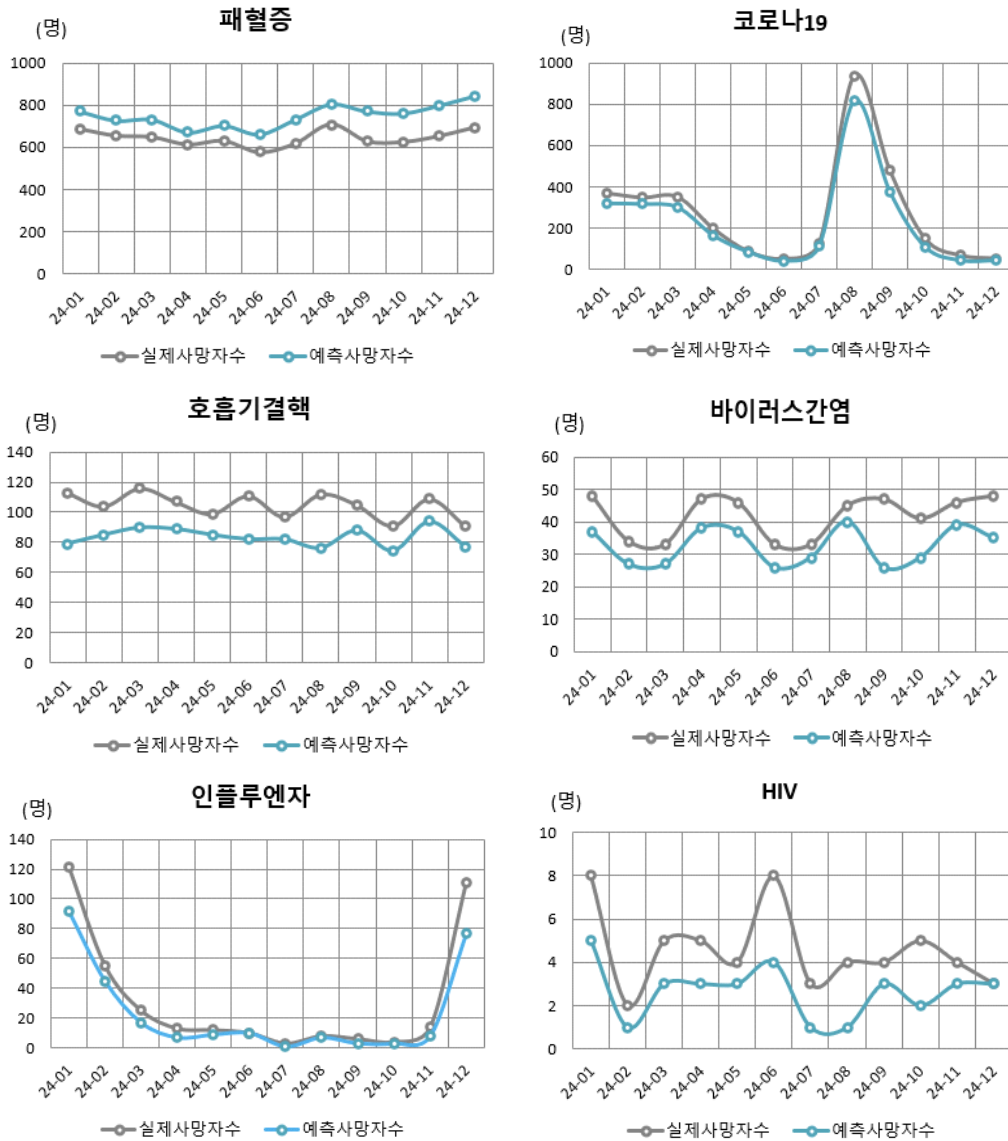
질환	월평균 사망자 수 추정비(%)	최소~최대 (%)	편차방향	RMSE ²⁾
패혈증	116.1	109.5~122.6	매월 일관된 과대추정	0.168
코로나19	82.6	62.7~96.5	일관된 과소추정, 변동 적은 편	0.194
호흡기결핵	80.0	67.9~86.2	일관된 과소추정, 변동 적은 편	0.208
바이러스간염	78.2	55.3~88.9	일관된 과소추정	0.234
인플루엔자	68.9	33.3~100.0	과소추정 경향, 변동 큼	0.356
HIV	58.8	25.0~100.0	과소추정 경향, 변동 매우 큼	0.458

<그림5-10> 감염병 질환별 월평균 사망자 수 추정비



2) $RMSE = \sqrt{(\sum(\text{월별 추정비}/100 - 1)^2 / 12)}$. 값이 작을수록 월별 추정비가 100%(실제값과 일치)에 가깝고 안정적임을 의미

<그림5-11> (감염병 질환별 × 월별) 실제 사망자 수 및 예측사망자 수



패혈증을 제외한 5개 감염병 질환(코로나19 · 호흡기결핵 · 바이러스간염 · 인플루엔자 · HIV)은 모두 월평균 추정비가 100% 미만으로서, 감염병 질환군 전반적으로 AI 모델이 실제 사망자 수보다 적게 예측하는 과소추정 경향이 공통적으로 나타났다. 이는 10대 사망원인 분석에서 확인된 알츠하이머병의 과소추정 양상과 유사한 패턴으로, 재현율이 낮은 희소질환일수록 실제 사례를 놓치는 방향으로 오차가 발생하는 경향이 있음을 시사한다.

반면 패혈증은 10대 사망원인 분석에서도 확인된 것처럼 사망자 수 추정비가 매월 109~123% 범위에서 일관적으로 과대추정되는 경향을 보였다. 그러나 RMSE(0.168)는 가장 작아 편차의 방향과 크기가 안정적으로 유지되는 특징을 보였다.

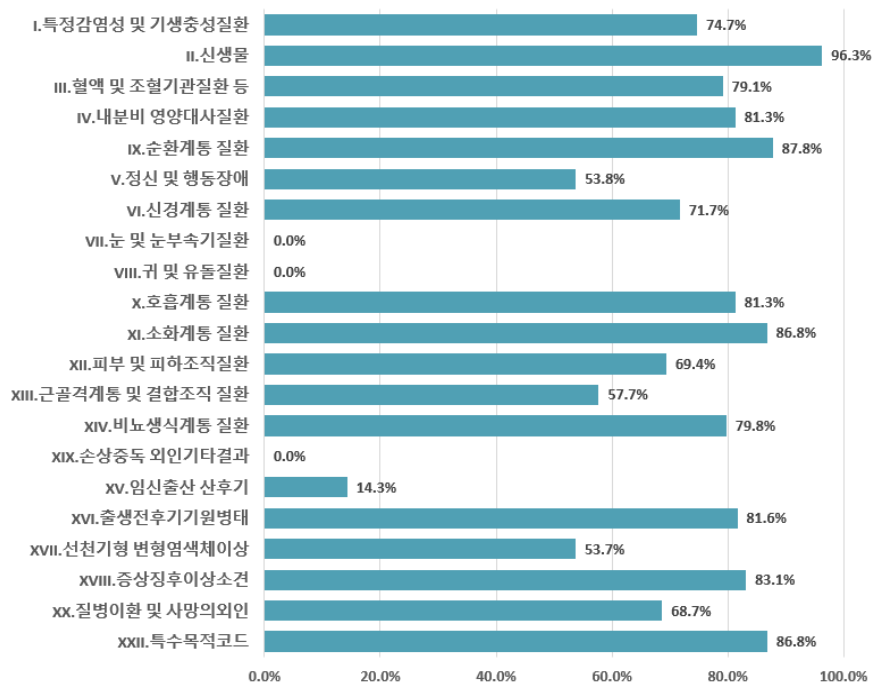
편차의 안정성을 나타내는 RMSE를 본다면, 패혈증(0.168), 코로나19(0.194), 호흡기 결핵(0.208)이 상대적으로 낮아 월별 추정비가 비교적 예측 가능한 범위에서 움직이는 반면, 인플루엔자(0.356)와 HIV(0.458)는 변동폭이 크다. 특히 HIV는 월별 실제 사망자 수 자체가 2~8건에 불과한 극소표본이므로, 한두 건의 오분류만으로도 사망자 수 추정비가 25%에서 100%까지 크게 변동할 수 있다는 점이 영향을 미친 것으로 보인다.

코로나19, 패혈증, 호흡기결핵은 연간 예측성능과 월별 안정성(RMSE) 모두 상대적으로 양호하여, 방향성(과대·과소추정)에 대한 보정을 전제로 감염병 사망 월별 감시 체계에 대한 참고지표로 활용할 가능성이 있다. 반면 HIV, 바이러스간염, 인플루엔자는 절대적인 발생 건수가 적고 월별 변동성이 커서, 현재 수준의 모델 성능만으로는 단독 지표로서 신뢰하기 어렵다.

다. 대분류별 결과

‘II.신생물’(96.3%), ‘IX.순환계통질환’(87.8%), ‘XI.소화계통질환’(86.8%) 등 사망자 수가 많은 대분류는 10대 사망원인 분석과 일관되게 높은 정확도를 보인다.

<그림5-12> 대분류별 예측정확도(F1-Score)



반면 ‘XV.임신출산 및 산후기’(사망자 수 26명, 재현율 7.7%), ‘XVII.선천기형, 변형 및 염색체 이상’(재현율 41.3%), ‘V.정신및행동장애’(F1 53.8%)는 표본이 적거나 코드가 세분화되어 있어 모델의 예측성능이 낮으며, ‘VII.눈및부속기질환’(3명), ‘VIII.귀및부속기질환’(3명)은 표본이 극소하거나 없어 지표 산출 자체가 무의미하다.

‘XVIII.상세불명이상소견’은 재현율(91.0%) 대비 정밀도(76.5%)가 낮고 추정비도 118.8%로서 과대예측 경향이 10대 사망원인의 폐렴·패혈증과 유사하게 나타난다.

<표5-21> 대분류별 사망자 수 및 예측정확도

대분류	사망자 수	AI 예측 건수	일치 건수	정밀도 (Precision)	재현율 (Recall)	F1 Score	사망자 수 추정비
I.특정감염성 및 기생충성질환	11,678	12,095	8,880	73.4	76.0	74.7	103.6
II.신생물	90,971	88,307	86,324	97.8	94.9	96.3	97.1
III.혈액 및 조혈기관질환 등	912	870	705	81.0	77.3	79.1	95.4
IV.내분비 영양대사질환	12,552	11,993	9,976	83.2	79.5	81.3	95.5
V.정신 및 행동장애	3,872	2,643	1,751	66.3	45.2	53.8	68.3
VI.신경계통질환	20,763	17,079	13,569	79.4	65.4	71.7	82.3
VII.눈 및 눈부속기질환	3	-	-	0.0	0.0	0.0	0.0
VIII.귀 및 유돌질환	3	-	-	0.0	0.0	0.0	0.0
IX.순환계통질환	68,723	69,075	60,477	87.6	88.0	87.8	100.5
X.호흡계통질환	45,816	50,626	39,200	77.4	85.6	81.3	110.5
XI.소화계통질환	14,659	14,242	12,539	88.0	85.5	86.8	97.2
XII.피부 및 피하조직질환	508	524	358	68.3	70.5	69.4	103.1
XIII.근골격계통 및 결합조직질환	1,795	1,292	891	69.0	49.6	57.7	72.0
XIV.비뇨생식계통질환	10,933	11,010	8,750	79.5	80.0	79.8	100.7
XV.임신출산 산후기	26	2	2	100.0	7.7	14.3	7.7
XVI.출생전후기지원병태	261	224	198	88.4	75.9	81.6	85.8
XVII.선천기형 변형염색체이상	351	189	145	76.7	41.3	53.7	53.8
XVIII.상세불명이상소견	41,787	49,660	38,007	76.5	91.0	83.1	118.8
XX.질병이환 및 사망의외인	29,748	25,876	19,094	73.8	64.2	68.7	87.0
XXII.특수목적코드	3,208	2,729	2,577	94.4	80.3	86.8	85.1

4. 심층분석

가. 오분류패턴 분석

오분류패턴 분석은 기초자료의 양적·질적 한계로 인해 예측정확도가 저하되는 현상을 파악하고, 질환별 오분류의 패턴을 확인하여 이후 모델의 개선방향을 탐색하기 위한 목적으로 수행한다. 아울러 모델의 예측코드를 그대로 신뢰하지 않고, 검증 후 환류를 통해 성능을 제고하려는 시도의 기초자료로 활용한다.

<표5-22> 정답질환별 예측오분류 목적지(GCN_full모델 성과기준)

정답질환	정답 라벨수	예측 일치수	오분류 건수	예측오분류 목적지(오분류 건수)		
				Top1	Top2	Top3
악성신생물	88,933	85,038	3,895	폐렴(981)	상세불명이상소견 (878)	패혈증(310)
심장질환	33,539	29,916	3,623	상세불명이상소견 (1,732)	폐렴(370)	고혈압성질환(241)
폐렴	30,103	27,702	2,401	기타호흡기질환 (462)	상세불명이상소견 (362)	뇌혈관질환(278)
뇌혈관질환	24,612	20,167	4,445	폐렴(1,427)	상세불명이상소견 (953)	패혈증(563)
고의적자해	14,872	12,078	2,794	상세불명이상소견 (2,033)	손상의 외부원인(교통사고 등)(139)	손상의 외부원인 (낙하물 등)(125)
알츠하이머병	12,223	6,260	5,963	정신및행동장애 (1,671)	상세불명이상소견 (1,552)	폐렴(1,193)
당뇨병	11,064	8,606	2,458	신장질환(819)	상세불명이상소견 (416)	폐렴(267)
고혈압성질환	8,229	7,483	746	심장질환(370)	상세불명이상소견 (80)	신장질환(66)
간질환	7,787	6,925	862	상세불명이상소견 (251)	폐렴(79)	악성신생물(간암) (78)
패혈증	7,730	6,939	791	뇌혈관질환(114)	신장질환(109)	상세불명이상소견 (98)
코로나19	3,208	2,625	583	폐렴(390)	기타호흡기질환 (36)	상세불명이상소견 (36)
호흡기결핵	1,255	875	380	폐렴(154)	상세불명이상소견 (64)	기타호흡기질환 (32)
바이러스간염	501	289	212	간질환(149)	상세불명이상소견 (10)	폐렴(9)
인플루엔자	382	238	144	폐렴(104)	상세불명이상소견 (9)	만성하기도질환(7)
HIV	55	25	30	악성신생물 (림프종 등)(5)	기타감염성질환 (4)	폐렴(4)

분석 대상인 정답라벨 범위는 10대 사망원인과 감염병 질환으로 한정하되, 오분류 목적지는 실제 예측된 104항목 분류로 구분한다. 이는 전체 질환 중 10대 사망원인과 감염병 질환은 일부이고, 예측결과를 검토한 결과 정답라벨이 해당 질환에 속하지만 오분류 목적지는 그 외의 질환을 향한 경우가 많았기 때문이다. 일치/오분류는 104항목 분류로 판정하였으나, 다수의 104항목 분류가 포함된 악성신생물 또는 심장질환의 경우 같은 질환(악성신생물 또는 심장질환)으로 분류했다면 정분류로 판정하였다.

10대 사망원인과 감염병 질환 각각에 대해 예측모델이 제시한 104항목 분류 코드 중 오분류이면서 절대 건수와 비중이 높고 정책적 함의가 도출될 수 있는 5가지 유형을 선별하여 세부코드(full code) 단위까지 심층분석하였다. 정답질환별 예측오분류 최다목적지를 종합하면, ‘상세불명이상소견(R코드)’와 ‘폐렴’ 등 두 코드가 자주 등장한다.

<표5-23> 오분류 유형과 주요 오분류 목적지의 빈도 및 비중

오분류 유형	정답질환	전체오분류 빈도	오분류 목적지 빈도	오분류 전체건수 대비 목적지 비중
① 원사인과 직접사인 혼동	뇌혈관질환	4,445	폐렴.패혈증 (1,990)	44.8%
② 병합분류코드 미학습	고혈압성질환	746	심장질환 (370)	49.6%
③ 만성 합병증	당뇨병	2,458	신장질환 (819)	33.3%
④ 사인불명 대체형	고의적자해	2,794	상세불명 이상소견(2,033)	72.8%

① 원사인과 직접사인 혼동 : 뇌혈관질환(정답) → 폐렴/패혈증(예측), 1,990건

뇌졸중(뇌혈관질환) 환자는 사망진단서의 직접사인 항목에 폐렴·패혈증이 흔히 함께 기재된다. 모델은 이 직접사인 표현에 반응하여 원사인에도 불구하고 상세불명의 폐렴 또는 패혈증 코드를 최상위로 예측하는 경향을 보인다. 뇌혈관질환의 오분류 4,445건에는 폐렴이 1,427건(32.1%), 패혈증이 563건(12.7%)으로, 두 종류를 합하면 전체 오분류의 44.8%를 차지한다.

<표5-24> 뇌혈관질환의 오분류 목적지별 빈도 및 비중

A41.9(상세불명의 패혈증)	J18.9(상세불명의 폐렴)	기타 폐렴
563 (오분류 목적지가 패혈증인 데이터 중 100.0%)	1,397 (오분류 목적지가 폐렴인 데이터 중 97.9%)	30 (오분류 목적지가 폐렴인 데이터 중 2.1%)

자세히 살펴보면, 뇌혈관질환이 폐렴으로 오분류된 1,427건 중 1,397건(97.9%)이 J18.9(상세불명 폐렴)로, 패혈증으로 오분류된 563건은 모두 A41.9(상세불명 패혈증)로 집중되었으며, 방향성도 뚜렷하다. 폐렴이나 패혈증이 정답인데 뇌혈관질환으로 예측된 사례는 392건으로 정방향의 1/5 수준에 그쳤다.

이는 텍스트상 '폐렴', '패혈증'이라는 표현 자체가 모델 예측에 강하게 작용했으며, 최종 원사인과 직접사인을 구분하는 판단규칙을 모델이 충분히 학습하지 못했음을 시사한다. 그러나 원사인과 직접사인 구분은 사인분류사도 임상 경과 전반을 종합하여 판단하는 영역이므로, 데이터 보완만으로는 해결하기 어려운 구조적 한계를 수반한다.

② 병합분류코드 미학습 : 고혈압성질환(정답) → 심장질환(예측), 370건

고혈압성질환 오분류 746건 중 370건(49.6%)이 심장질환으로 예측되었다. 반대방향(심장질환이 정답인데, 고혈압성질환으로 예측)은 241건으로 심장질환 전체 오분류 3,623건 중 6.7%에 그쳐 마찬가지로 방향성이 뚜렷하다.

오분류 사례를 정답의 세부코드 기준으로 살펴보면, I11(고혈압성 심장병)이 277건(74.9%)을 차지하였다. 그리고 해당유형 오분류 370건 중 200건(54.1%)이 I50.9(상세불명의 심부전)으로 오분류되었다.

I11(고혈압성 심장병), I13(고혈압성 심장 및 신장병)은 고혈압으로 인한 심부전을 이미 포함하는 병합분류 코드로 원칙상 고혈압을 원사인으로 판단해야 하나, 사망진단서 텍스트에 '심부전'이라는 표현이 함께 등장하면 모델이 최종 임상소견에 따라 심장질환 단독코드(I50.9 등)로 예측하는 경향을 보였다.

<표5-25> 고혈압성 심장병의 오분류 목적지별 빈도

정답 세부코드	오분류 건수	심장질환[Z68]	
		I50.9	기타 심장질환
I10.9(상세불명의 고혈압)	37	2	35
I11.0(고혈압성 심부전)	246	165	81
I11.9(고혈압성 심장병)	31	0	31
I13.2(심부전 및 신부전을 동반한 고혈압성 심장 및 신장병)	36	27	9
기타	20	6	14
합계	370	200	170

즉 텍스트의 임상소견에 반응하는 경향도 있지만, 보다 근본적으로는 병합분류 코드에 대한 원사인 선정규칙을 모델이 학습하지 못했기 때문으로 판단된다. I11(고혈압성 심장병), I13(고혈압성 심장 및 신장병) 여부와 심부전 또는 신부전의 동시 표현은 명시적인 규칙으로 판별할 수 있을 것으로 예상되므로, 향후 규칙기반 후처리방식을 도입한다면 우선 적용 후보로 판단된다.

③ 만성 합병증 : 당뇨병(정답) → 신장질환(예측), 819건

당뇨병의 104항목 분류 기준 오분류 2,458건 중 819건(33.3%)이 신장질환으로 예측되었다. 정답 세부코드별로는 E11.22(말기신장병을 동반한 2형 당뇨병)에서 432건, E11.78(기타 다발성 합병증을 동반한 2형 당뇨병)에서 300건으로, 신장질환으로 예측한 오분류건수 819건 중 두 코드가 732건(89.4%)을 차지한다.

<표5-26> 당뇨병의 오분류 목적지별 빈도

정답 세부코드	오분류 건수	N18.5 (만성, 말기 신장병)	N18.9 (상세불명의 만성신장병)	기타 신장질환
E11.22(말기신장병을 동반한 2형당뇨병)	432	365	38	29
E11.78(기타 다발성 합병증을 동반한 2형 당뇨병)	300	215	34	51
기타	87	54	4	29
합계	819	634	76	109

E11.22는 이미 ‘당뇨병성 신장합병증(말기신장병)’을 포함하는 결합분류 코드로, 원칙상 당뇨병을 원사인으로 판단해야 한다. 그러나 사망 직전 임상 경과가 말기신부전(N18.5)으로 진행된 경우, 모델은 이 최종 임상소견에 영향을 받아 원사인을 신장질환으로 예측하는 경향을 보인다. 이는 ‘②고혈압성질환→심장질환’에서 확인된 결합분류 코드에 대한 원사인 선정규칙을 미학습했던 메커니즘이 당뇨병에서도 반복되는 것으로서, 두 사례 모두 결합분류 코드를 식별하기 위한 규칙기반 로직 탐색을 통해 개선여지가 있다고 판단된다.

④ 사인불명 대체 : 고의적자해(정답) → 상세불명이상소견(예측, R코드), 2,033건

고의적자해(자살)의 104항목 분류 기준 오분류 2,794건 중 2,033건(72.8%)이 상세불명이상소견(R코드)으로 예측되어, 10대 사망원인 및 감염병 질환 16종 단일 목적지 비중이 가장 높다. 정답 전체(14,872건) 대비로도 13.7%에 달해, 자살로 확정된 사망 중

상당수에 대해 모델이 잠정 원사인 및 사망진단서 등 기초자료를 기반으로 원사인 불명 판정을 내리고 있음을 의미한다.

정답 쪽 세부코드는 X70(목매 · 교액 및 질식에 의한 자해) 294건, X67(기타 가스 및 휘발성 물질에 의한 자의의 중독 및 노출) 계열 500여 건, X80(높은 곳에서 뛰어내림에 의한 자해) 216건, X68.0(농약에 의한 자해, 주택) 98건 등 자살 수단별로 넓게 분산되어 있어, 자살 수단 전반에 걸쳐 모델이 구체적인 X코드 대신 ‘원인불명’으로 판정을 내리는 경향이 확인된다.

이는 특정 자살 수단 코드의 문제가 아니라, 정보 자체의 부재에서 비롯된 것으로 보인다. 정답라벨이 ‘고의적자해’이나 R99로 예측된 사례를 확인한 결과, 사망진단서 텍스트가 대부분 ‘미상’, ‘심정지’ 등 구체적 수단을 특정하지 않았고, 잠정 원사인도 대부분이 R99로 처리되어있었다. 즉 사망진단서 작성 시점에는 구체적인 자살 수단이 확정되지 않은 경우가 많으며, 모델은 주어진 정보의 한계 내에서 R99를 예측한 것으로 추측된다. 다만 건강보험 진료기록 중 일부 손상 관련 코드가 드물게 확인되기도 한다.

4가지 유형에 대해 조치방안 및 개선 가능성을 몇 가지로 나눌 수 있다. 유형①은 원사인과 직접사인의 구분 자체가 구조적으로 모호한 영역으로, 데이터의 보완만으로는 완전히 해소하기 어렵고, 사인분류사가 임상 경과 전반을 종합하여 판단할 영역이다.

유형② · ③은 결합분류 코드에서 원사인 선택규칙을 모델이 학습하지 못한 경우로, 두 사례 모두 비교적 원사인이 명확하고 판별조건을 규칙으로 정의할 여지가 있어 향후 규칙기반 후처리 프로세스의 우선 적용 후보로 판단된다.

유형④는 학습자료 자체가 부족하여 상세불명코드 외의 코드를 예측하는 것이 사실상 어렵다. 자살 수단별 학습자료 자체의 질적 보강이 근본 해법에 가깝다 할 수 있다.

나. 모델 예측코드의 정답라벨 포괄비율

앞선 오분류패턴 분석이 ‘예측이 틀렸을 때 예측이 어디로 향했는가’를 다루었다면, 이번에는 ‘모델의 예측코드가 정답을 얼마나 포괄하는지’를 살펴본다.

모델은 Top1(최상위 확률) 코드 외에도 Top2, Top3 후보를 함께 산출하므로, 정답라벨이 Top1뿐 아니라 Top1~2, Top1~3 범위까지 확장했을 때 얼마나 누적되어 포함되는지를 확인하면 모델의 예측코드를 실무적용 시 보조지표로서의 활용가능성을 검토할 수 있다. 실제 시스템에 구현된다면 사인분류사는 AI가 예측한 3개의 후보 코드를 고려하여 최종 코드를 선정하게 된다.

먼저 정답라벨 수(N) 대비 Top1의 적중률, Top1~2의 적중률, Top1~3의 적중률을

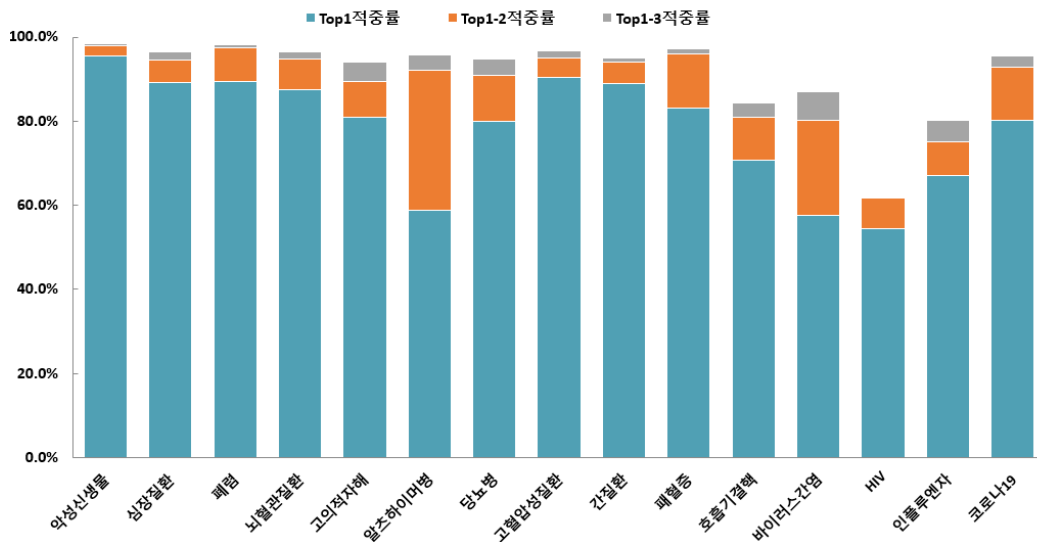
산출하였다. 적중률은 104항목 분류 기준으로 판단하였으며, 이는 정답라벨 수 중 예측 코드와 정답라벨이 일치한 건수를 의미하는 재현율(Recall)과 동일한 개념이다. 그리고 Top1과 Top1~3 간의 격차가 큰 순서로 정렬하였다. 이 증가폭이 클수록 모델의 확신이 Top1 하나에 집중되지 않고 Top2와 Top3으로 분산되어 있음을 의미하며, 반대로 증가폭이 작으면 Top1 예측 자체의 확신도가 이미 높음을 의미한다.

<표5-27> 질환별 Top1, Top1~2, Top1~3 적중률 및 증가폭

질환	Top1 적중률(A)	Top1~2 적중률(B)	Top1~3 적중률(C)	증가폭 (C-A)
악성신생물	95.5%	97.8%	98.6%	3.0%p
심장질환	89.1%	94.6%	96.5%	7.4%p
폐렴	89.5%	97.4%	98.3%	8.8%p
뇌혈관질환	87.6%	94.7%	96.6%	9.0%p
고의적자해	80.9%	89.4%	94.1%	13.2%p
알츠하이머병	58.8%	92.1%	95.7%	36.9%p
당뇨병	79.9%	90.9%	94.8%	14.9%p
고혈압성질환	90.4%	95.1%	96.8%	6.4%p
간질환	88.9%	94.1%	95.1%	6.2%p
패혈증 ³⁾	83.1%	96.1%	97.2%	14.0%p
만성하기도질환	80.0%	92.2%	95.1%	15.1%p
호흡기결핵	70.8%	80.9%	84.4%	13.6%p
바이러스간염	57.7%	80.2%	87.0%	29.3%p
HIV	54.5%	61.8%	61.8%	7.3%p
인플루엔자	67.0%	75.1%	80.1%	13.1%p
코로나19	80.3%	92.9%	95.5%	15.1%p

3) 패혈증은 10대 사망원인과 감염병 질환(6종) 모두에 포함됨

<그림5-13> 질환별 Top1, Top1~2, Top1~3 적중률(104항목 기준)



증가폭을 기준으로 보면 두 극단이 뚜렷하게 대비된다. 알츠하이머병은 Top1 적중률이 58.8%로 10대 사망원인 중 가장 낮지만, Top1~3까지 확장하면 95.7%까지 상승하여 36.9%p의 압도적인 상승폭을 보인다.

이는 모델이 알츠하이머병 여부 자체는 어느 정도 인지하고 있으나, Top1 하나로는 확신하지 못하고 후보 2와 후보 3으로 상당히 분산되어 있음을 의미한다. 앞서 확인한 것처럼 알츠하이머병은 104항목 분류 기준 오분류율이 41.2%로 가장 높았는데, 이번 결과는 그 오분류의 상당 부분이 완전히 다른 질환으로 예측된 것이 아니라 모델의 두 번째 또는 세 번째 후보 안에는 있었던 근소한 오분류임을 보여준다.

반면 악성신생물은 상반된 양상을 보인다. Top1 적중률이 95.5%로 이미 매우 높아, Top1~3 확장에 따른 증가폭은 3.0%p에 그쳤다. 모델이 Top1 예측에 강한 확신을 가지고 있으며, 그 확신이 대체로 정확함을 의미한다. 고혈압성질환(6.4%p)과 간질환(6.2%p) 역시 증가폭이 작아 Top1 예측의 확신도가 높은 질환군으로 볼 수 있다. 감염병 질환 중에서는 바이러스간염(29.3%p)이 상대적으로 큰 증가폭을 보였다.

Top1의 적중률만으로 모델 활용가능성을 판단하면 알츠하이머병이나 바이러스간염과 같은 질환은 과소평가될 수 있다. 그러나 Top1~3까지 범위를 확대하면 적중률이 90% 내외까지 향상된다는 점은 사인분류사 의사결정을 지원하는 보조지표로서 중요한 의미를 가진다. 따라서 이러한 질환군은 AI가 최종 코드를 자동으로 확정하는 방식보다, Top3 후보를 함께 제시하고 사인분류사가 최종 선택하는 방식이 더 적합할 수 있다. 반면 악성신생물과 폐렴처럼 Top1 예측의 확신도가 이미 높은 질환군은 자동화를 적용하는 방안을 검토할 수 있다.

다. 모델 예측확률 분포

AI 모델은 원사인 코드를 예측할 때 해당 예측에 대한 확신 정도(예측확률, softmax 변환값)를 함께 산출한다. 여기서는 예측확률 구간별로 실제 정확도가 어떻게 나타나는지를 확인하여, 예측확률이 모델의 신뢰도 지표로서 실질적으로 기능하는지, 즉 확률이 높을수록 실제로 더 정확한지를 검증한다. 이는 향후 예측확률 기준으로 AI 예측코드를 선별 반영하는 운영전략의 기초자료가 된다.

<표5-28> GCN_104모델의 예측확률구간별 실제정확도

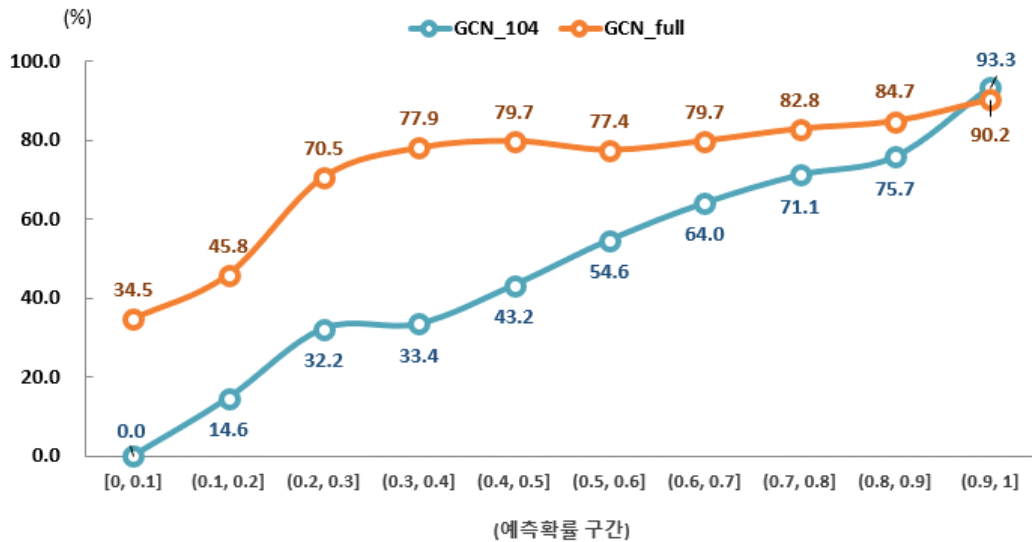
모델예측확률 구간 (Top1_prob)	데이터수(N)	예측정확건수 (N_correct)	전체대비비중 (N/sum(N))	실제정확도 (N_correct/N)
[0, 0.1]	1	-	0.0%	0.0%
(0.1, 0.2]	41	6	0.0%	14.6%
(0.2, 0.3]	382	123	0.1%	32.2%
(0.3, 0.4]	1,574	526	0.4%	33.4%
(0.4, 0.5]	6,096	2,634	1.7%	43.2%
(0.5, 0.6]	21,936	11,986	6.1%	54.6%
(0.6, 0.7]	23,102	14,778	6.4%	64.0%
(0.7, 0.8]	25,418	18,071	7.1%	71.1%
(0.8, 0.9]	33,739	25,537	9.4%	75.7%
(0.9, 1]	246,280	229,782	68.7%	93.3%
전체	358,569	303,443	100%	84.6%

<표5-29> GCN_full모델의 예측확률구간별 실제정확도

예측확률 구간 (Top1_prob)	데이터수(N)	예측정확건수 (N_correct)	전체대비비중 (N/sum(N))	실제정확도 (N_correct/N)
[0, 0.1]	304	105	0.1%	34.5%
(0.1, 0.2]	2,845	1,302	0.8%	45.8%
(0.2, 0.3]	10,078	7,102	2.8%	70.5%
(0.3, 0.4]	18,670	14,551	5.2%	77.9%
(0.4, 0.5]	27,350	21,796	7.6%	79.7%
(0.5, 0.6]	31,586	24,452	8.8%	77.4%
(0.6, 0.7]	32,200	25,674	9.0%	79.7%
(0.7, 0.8]	38,073	31,522	10.6%	82.8%
(0.8, 0.9]	48,450	41,045	13.5%	84.7%
(0.9, 1]	149,013	134,445	41.6%	90.2%
전체	358,569	301,994	100%	84.2%

모델예측확률(Top1_prob)을 0.1 단위의 10개 구간으로 나누고, 구간별 해당건수, 전체대비비중, 실제정확도(정답라벨과 예측코드가 일치한 비율, 104항목 분류 기준)를 GCN_104과 GCN_full 두 모델 각각에 대해 산출하였다.

<그림5-14> GCN_104, GCN_full 모델에 대한 예측확률 구간별 실제정확도



두 모델 모두 예측확률 구간이 높아질수록 실제정확도가 함께 높아지는 단조증가 패턴을 보인다. 이는 모델이 제시하는 예측확률이 모델의 신뢰도를 반영하는 유효한 지표로 기능하고 있음을 의미하며, 향후 확률 기반 선별반영 전략(임계확률 설정)의 전제조건이 충족됨을 뜻한다.

다만 두 모델의 확신 분포는 뚜렷한 차이를 보인다. GCN_104는 예측확률 0.9 초과 구간이 전체의 68.7%를 차지하며, 이 구간의 실제정확도는 93.3%에 달한다. 반면 GCN_full은 0.9 초과 구간의 비중이 41.6%로 GCN_104의 약 60% 수준에 그치고, 이 구간의 실제 정확도도 90.2%로 상대적으로 낮은 편이다.

6천여 종의 세부코드 단위로 학습한 GCN_full은 104항목 분류로 학습한 GCN_104보다 분류대상 자체가 훨씬 많아, 정답을 맞히더라도 전체 확률이 여러 후보 코드로 분산되면서 최상위 확률이 상대적으로 낮게 형성되는 구조적 특성을 가진 것으로 추측된다.

이 차이는 중간 확률구간에서 더 뚜렷하게 나타난다. GCN_104는 예측확률 구간과 실제정확도가 비교적 근접하게 대응한다. 예를 들면 (0.3,0.4] 구간의 정확도 33.4%는 해당 구간과 큰 차이가 없다. 반면 GCN_full은 상대적으로 낮은 확률구간에서도 실제정확도가 구간 범위보다 현저히 높게 나타난다. (0.3,0.4] 구간의 정확도가 77.9%에 달

하고 (0.6,0.7] 구간도 79.7%로 구간 상단에 근접한다. 이는 GCN_full이 스스로의 예측 확신을 다소 과소평가하는 경향이 있음을 시사한다.

그러나 두 모델 모두 예측확률과 실제정확도 간에 유의한 상관관계를 보여, 예측확률을 활용한 선별반영 전략은 두 모델 모두에 적용 가능하다. 다만 동일한 임계확률(예: 0.9)을 두 모델에 그대로 적용할 경우 실질적 의미가 달라진다. GCN_104는 0.9 임계값만으로 전체의 68.7%를 포괄하면서 93.3% 정확도로 자동 채택할 수 있는 반면, GCN_full은 같은 임계값에서 채택 가능한 비중이 41.6%로 줄고 정확도도 90.2%로 상대적으로 낮은 수준이다.

따라서 하이브리드 운영 시 임계확률은 모델별 특성을 고려하여 최적화할 필요가 있으며, GCN_full처럼 예측확률을 다소 과소평가하는 모델에는 상대적으로 완화된 임계확률을 적용할 수 있다. 반면 사인분류사의 최종 검토를 거치지 않거나 최소한의 검증만 수행하는 경우에는, 높은 예측확률에서 높은 정확도를 보장하는 GCN_104 모델이 더 적합한 모델로 판단된다.

마. 기초자료 난이도

앞서 모델의 예측결과 자체를 분석하였다면, 이번에는 관점을 바꾸어 기초자료 자체의 난이도가 모델 성능에 어떤 영향을 미치는가를 확인한다. 잠정 원사인과 최종 원사인의 계층적 거리, 정답라벨의 기초자료 출처, 학습데이터의 절대적 규모 및 평가데이터 대비 상대적 규모, 희소코드 여부 등 네 가지 관점에서 GCN_full 모델을 중심으로 분석하였다.

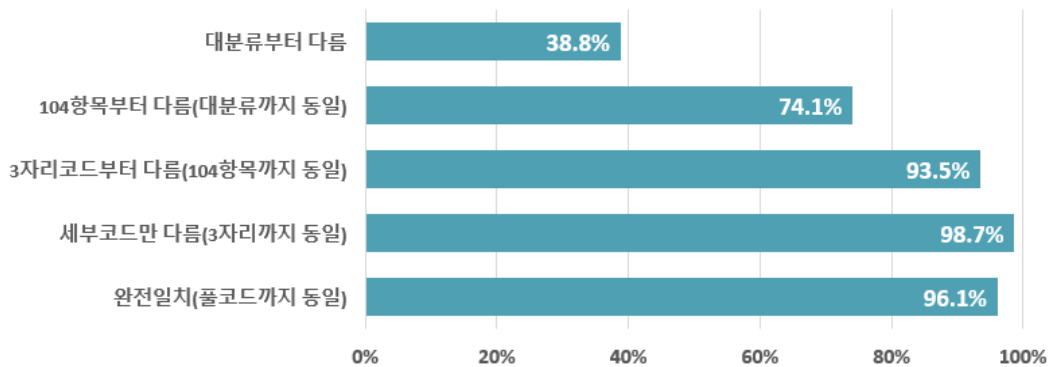
기초자료의 난이도는 본질적으로 학습데이터와 평가데이터 간의 균형문제 및 학습데이터 구성정보의 일관성과 관련되며, 이러한 특성이 악화될수록 모델 성능이 저하될 것이라는 가설에 기반한다. 이에 따라 특정 군의 정답라벨을 모델이 얼마나 놓치지 않고 포착하는 것을 평가하는 것이 적합하다고 판단하여, 성능 평가지표로 재현율을 사용하였다.

<표5-30> 잠정 원사인과 최종 원사인의 계층적 거리와 정확도

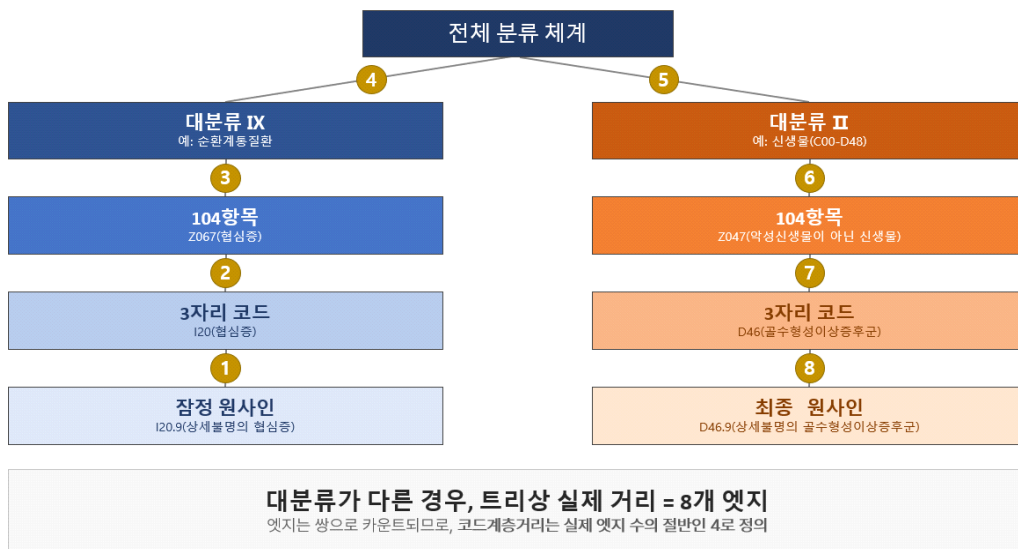
계층적 거리	판정기준	N	N_correct	정확도 (N_correct/N)
0	완전일치(세부코드까지 동일)	177,986	171,027	96.1%
1	세부코드만 다름(3자리까지 동일)	55,476	54,735	98.7%
2	3자리코드부터 다름(104항목까지 동일)	31,190	29,157	93.5%
3	104항목부터 다름(대분류까지 동일)	30,207	22,374	74.1%
4	대분류부터 다름	63,710	24,701	38.8%

먼저 잠정 원사인과 최종 원사인(정답라벨)이 코드 계층 상 얼마나 떨어져 있는지를 5단계 거리로 구분하고, 거리별 GCN_full의 정확도(104항목 분류 기준)를 확인하였다.

<그림5-15> 잠정 원사인과 최종 원사인의 계층적 거리와 정확도



<그림5-16> 잠정 원사인과 최종 원사인의 계층적 거리 도식



흥미로운 점은 정확도가 거리 0(완전일치, 96.0%)보다 거리 1(세부코드만 다름, 98.7%)에서 오히려 더 높다는 점이다. 이는 잠정 원사인이 세부코드까지는 다르지만 3자리코드까지는 일치하는 경우, 모델이 그 근접성에 대한 정보를 활용하여 더 정확하게 예측하고 있음을 추측할 수 있다. 반면 거리가 멀어질수록 정확도는 급격히 하락하여, 잠정 원사인이 정답라벨과 대분류 수준부터 다른 경우에는 정확도가 38.8%까지 떨어진다.

즉 잠정 원사인이 104항목 분류까지만 일치해도 모델정확도는 90% 이상을 유지하지만, 대분류 자체가 다르다면 모델이 예측에 어려움을 겪는다. 이는 잠정 원사인의 정답라벨로의 근접도가 결과적으로 모델 성능을 설명하는 유의미한 지표임을 보여준다.

다음으로 정답라벨이 잠정 원사인, 주상병, 부상병 중 어디에서 유래했는지에 따라 GCN_full의 정확도(104항목 분류기준)가 어떻게 달라지는지 확인하였다.

<표5-31> 정답라벨의 출처 및 정확도

정답라벨의 출처	N	N_correct	정확도 (N_correct/N)
1) 잠정 원사인	178,811	171,629	96.0%
2) 주상병	64,568	52,502	81.3%
3) 부상병	14,742	9,448	64.1%
4) 출처없음	100,448	68,415	68.1%

먼저 1) 잠정 원사인부터 4) 출처없음까지 배타적으로 집계한 값이라는 점에 유의해야 한다. 즉 1) 잠정 원사인이 정답라벨의 출처가 아닌 경우, 2) 주상병이 정답라벨의 출처인지 여부를 판단하는 방식으로, 단계적으로 분류하여 작성한 값이다.

정답라벨이 잠정 원사와 일치하는 경우 모델의 정확도가 96.0%로 가장 높고, 주상병(81.3%), 부상병(64.1%) 순으로 낮아진다. 다만 기초자료 어디에도 없는 ‘출처없음’ 케이스의 정확도는 68.1%로, 부상병(64.1%)보다 유사하지만 약간 높게 나타났다. 이는 범주형 자료에 자료가 없더라도 모델이 사망진단서 텍스트 자체에서 상당한 정보를 추출해 활용하고 있고, 세부코드는 아니더라도 104항목 분류 단위에서는 예측 가능성을 시사한다.

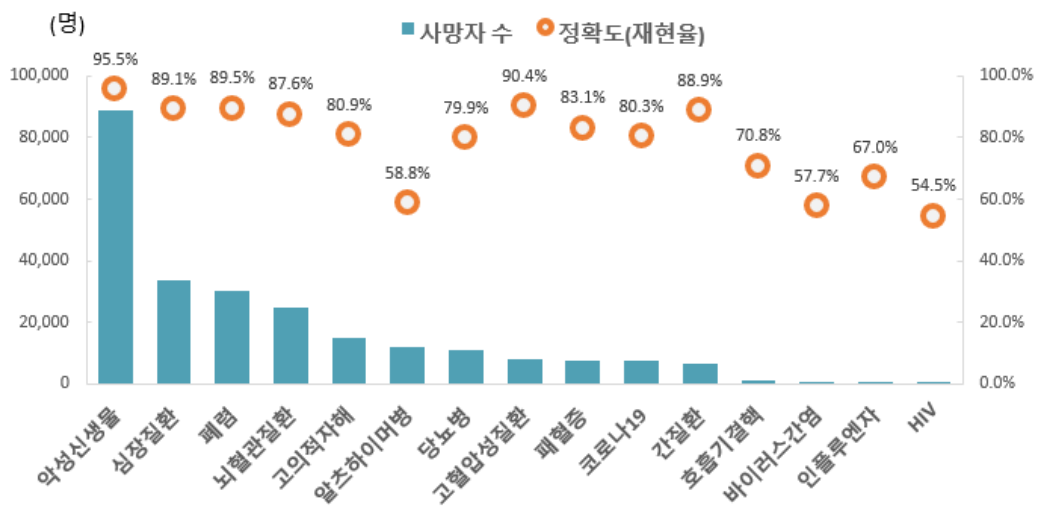
그 다음으로 질환별 규모와 GCN_104모델 간 정확도 관계를 확인하였다. 전반적으로 학습데이터 규모(N_2023)와 정확도는 유의하게 양(+)의 관계를 보인다. 1천 건 미만인 HIV, 인플루엔자, 바이러스간염 3개 질환은 모두 정확도 70%에 미달하는 반면, 1만 건 이상 질환의 대부분은 80~92%대의 높은 정확도를 보인다. 다만 알츠하이머병은 예외적이다. 정답라벨 수가 12,223건으로 1만 건 이상 구간에 속함에도 정확도는 58.8%로, 오히려 1천~1만 구간에 해당하는 질환들보다 낮다. 이는 학습데이터 절대 수에 따른 성능 개선 효과보다 알츠하이머병 고유의 난이도가 더 크게 작용하고 있음을 의미한다.

유사하게 2023년(학습) 대비 2024년(평가) 데이터 규모의 상대적 비율과 정확도의 관계를 확인하였다.

<표5-32> 사망자 수 규모구간별 예측정확도(재현율)

규모 구간	질환	N_2023	N_2024	예측정확도(GCN_104)
1천건 미만	HIV	45	55	54.5%
	인플루엔자	316	382	67.0%
	바이러스간염	459	501	57.7%
1천~1만	호흡기결핵	1,227	1,255	70.8%
	간질환	7,263	7,787	88.9%
	코로나19	7,442	3,208	80.3%
	폐혈증	7,809	7,730	83.1%
	고혈압성질환	7,988	8,229	90.4%
1만 이상	당뇨병	11,058	11,064	79.9%
	알츠하이머병	11,109	12,223	58.8%
	고의적자해	13,978	14,872	80.9%
	뇌혈관질환	24,194	24,612	87.6%
	폐렴	29,422	30,103	89.5%
	심장질환	33,147	33,539	89.1%
	악성신생물	85,271	88,933	95.5%

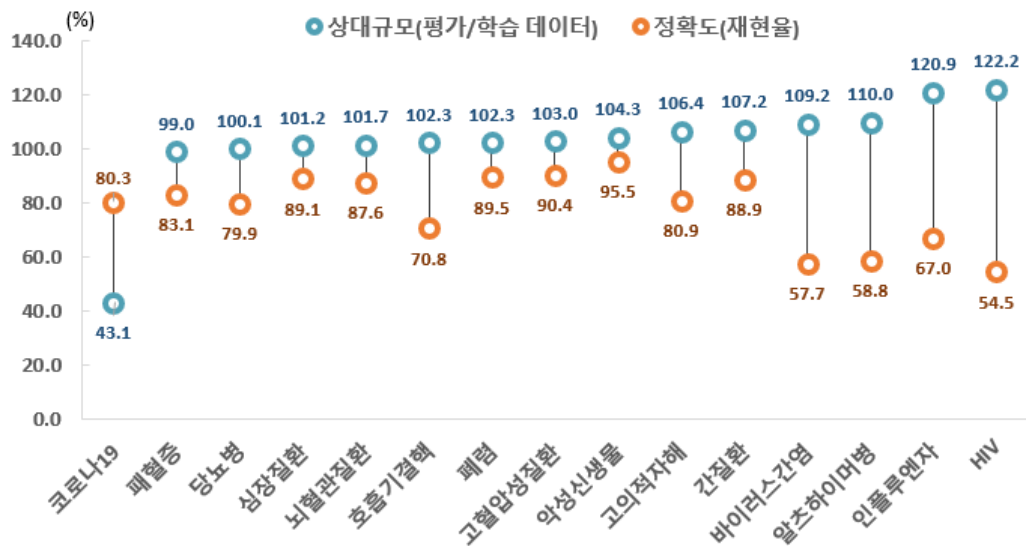
<그림5-17> 사망자 수 규모구간별 예측정확도(재현율)



<표5-33> 학습데이터(2023년) 대비 평가데이터(2024년)의 상대규모와 예측정확도(재현율)

구간(2024/2023 비)	+질환	상대규모(%)	N_2024	예측정확도(GCN_104)
50% 미만	코로나19	43.1%	3,208	80.3%
50~100%	패혈증	99.0%	7,730	83.1%
100~110%	당뇨병	100.1%	11,064	79.9%
	심장질환	101.2%	33,539	89.1%
	뇌혈관질환	101.7%	24,612	87.6%
	호흡기결핵	102.3%	1,255	70.8%
	폐렴	102.3%	30,103	89.5%
	고혈압성질환	103.0%	8,229	90.4%
	악성신생물	104.3%	88,933	95.5%
	고의적자해	106.4%	14,872	80.9%
	간질환	107.2%	7,787	88.9%
	바이러스간염	109.2%	501	57.7%
110%~	알츠하이머병	110.0%	12,223	58.8%
	인플루엔자	120.9%	382	67.0%
	HIV	122.2%	55	54.5%

<그림5-18> 질환별 학습데이터 대비 평가데이터의 상대적 규모와 예측정확도(재현율)



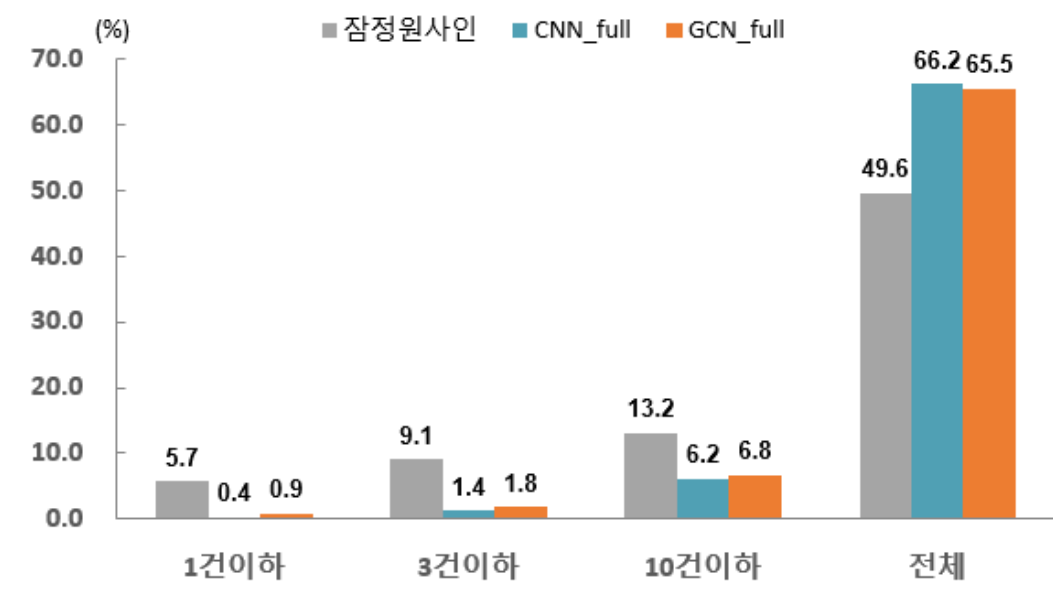
2023년 대비 2024년 데이터가 상대적으로 크게 늘어난 질환(HIV 122%, 인플루엔자 121%, 알츠하이머병 110%, 바이러스간염 109%)일수록 정확도가 낮게 나타나는 경향이 확인된다. 이는 학습(2023)과 평가(2024) 데이터 간 규모 불균형이 클수록, 다시 말해 학습데이터에서 학습하지 못한 새로운 패턴이 평가데이터에 많이 추가될수록 모델의 성능이 저하된다는 의미이다.

마지막으로 2024년 정답라벨 기준 코드별 발생빈도에 따라 '1건 이하', '3건 이하', '10건 이하', '전체' 등 4개 구간으로 나누어, 구간별 CNN_full, GCN_full 모델과 사망 진단서 기반 잠정 원사의 정확도를 비교하였다.

<표5-34> 희소코드 기준별 정확도(잠정 원사인, CNN_full, GCN_full)

희소코드 기준	해당코드 수	정답라벨 수	정확도		
			잠정 원사인	CNN_full	GCN_full
1건 이하	853	853	5.7%	0.4%	0.9%
3건 이하	1,479	2,352	9.1%	1.4%	1.8%
10건 이하	2,121	6,333	13.2%	6.2%	6.8%
전체	3,273	358,569	49.6%	66.2%	65.5%

<그림5-19> 희소코드 기준별 정확도(잠정 원사인, CNN_full, GCN_full)



희소코드일수록 세 방식 모두 정확도가 급격히 낮아지지만, 그 양상은 뚜렷이 다르다. 극희소 구간(1건 이하, 3건 이하, 10건 이하)에서는 사망진단서(잠정 원사인) 방식이 CNN_full, GCN_full보다 오히려 높은 정확도를 보인다. 예를 들어 1건 이하 구간에서 사망진단서는 5.7%인 반면, GCN_full은 0.9%, CNN_full은 0.4%에 불과하다.

이는 잠정 원사인은 매 건에 대해서 개별적으로 WHO 지침을 적용하여 판단하지만, AI는 학습에 근거해서 예측하기 때문에 사실상 학습이 불가능한 희소코드에서는 사망진단서보다 성능이 떨어지는 것으로 추정된다.

다만 전체 기준으로는 GCN_full(65.8%), CNN_full(66.2%)이 사망진단서(49.6%)를 크게 상회하여, 희소코드를 제외한 대부분의 사례에서는 AI 모델의 우위가 뚜렷하다.

앞서 실시한 네 가지 분석은 공통적으로 모델 성능이 순수한 예측 알고리즘의 문제만이 아니라 기초자료 자체의 난이도에 크게 좌우됨을 보여준다. 잠정 원사인과의 근접도, 정답라벨의 범주형 자료 존재 여부, 학습데이터의 절대·상대적 규모, 세부코드의 희소성이 모두 유의미한 영향을 미친다.

바. 예측코드의 보정(하이브리드 방식)

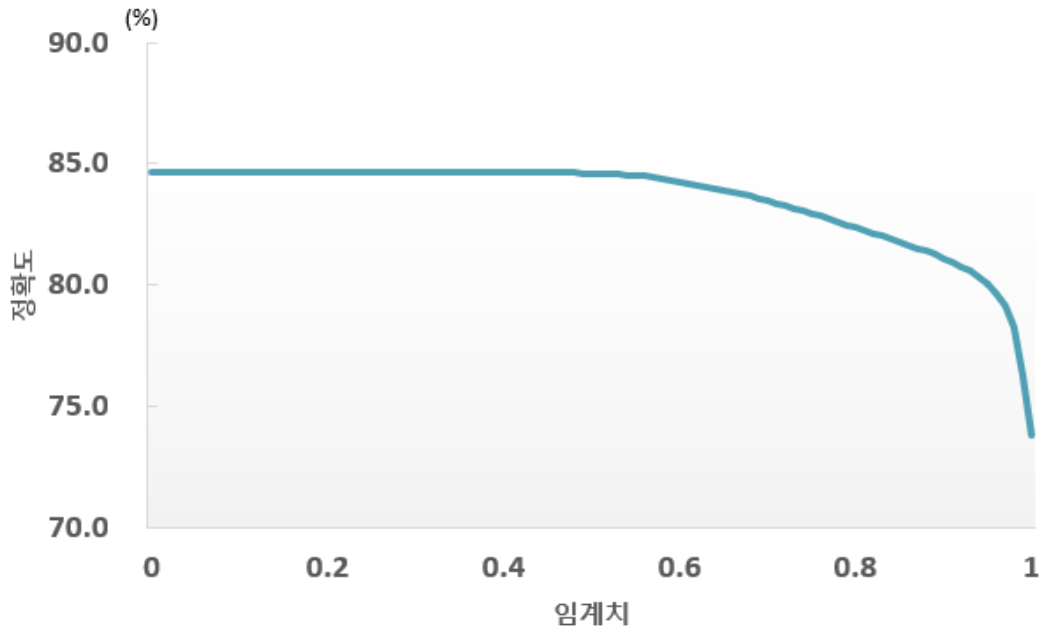
하이브리드 방식은 모델의 예측확률(Top1_prob)이 특정 임계치 이상이면 AI 예측코드를 채택하고, 임계치 미만이면 잠정 원사인(WHO 규칙 기반 산출값)을 대신 채택하는 보정 방식이다. 임계치를 0~1 구간에서 0.01 단위로 조정하면서 하이브리드 방식의 정확도가 최대화되는 최적임계치를 그리드서치(Grid Search)로 탐색하였다.

<표5-35> 예측확률(Top1_prob)의 최적 임계치 및 정확도 개선율

모델	최적 임계치	최적하이브리드 정확도	임계치_Top1 채택비율	Top1 단독정확도	개선율
GCN_104	0.4	84.6%	99.4%	84.6%	0.0%

하이브리드 방식으로 인한 개선폭은 미미하였다. 최적임계치 0.4에서 Top1 채택비율은 99.4%(즉 0.6%만 잠정 원사인으로 대체)로 정확도가 84.6% → 84.6%로 개선폭이 거의 없었고 대부분 케이스에서 예측코드를 그대로 채택하는 것이 가장 높은 정확도를 보였다. 이는 모델 자체의 정확도(84%수준)가 잠정 원사인을 이미 큰 폭으로 상회하므로, 일부 사례를 잠정 원사인으로 대체하더라도 전체 정확도 향상에는 거의 기여하지 못함을 의미한다.

<그림5-20> 예측확률(Top1_prob)의 임계치별 정확도



5. 종합해석

가. 잠정 원사인 대비 예측결과의 정확도 순위와 그 함의

1절에서 확인한 바와 같이 예측모델(정확도 84.2~84.6%)은 잠정 원사인(75.8%) 대비 8~9%p 높은 정확도를 보였다. 이 부분은 기초자료 난이도별 결과를 같이 보면 더 분명하다. 모델 Top1 예측코드의 62.07%가 잠정 원사인 코드와 일치한다는 점에서, 모델은 잠정 원사인과 완전히 별개의 추론을 하기보다는 이를 핵심정보로 활용하여 추론을 한다고 볼 수 있다. 여기에 주상병, 부상병 등 건강보험 기록과 사망진단서 텍스트를 추가로 종합함으로써, 잠정 원사인만으로 도달하기 어려운 최종 원사인에 근접한 판단을 내릴 수 있는 것이다.

실제로 잠정 원사인과 최종 원사인의 코드 계층 거리가 가까울수록(거리 0~1) 모델 정확도가 96~99%에 달하지만, 거리가 멀어질수록 정확도가 38.8%(거리 4)까지 급락함을 확인하였다. 즉 모델의 정확도 순위는 잠정 원사인을 기반으로 추가 정보를 종합함으로써 도달한 결과이며, 잠정 원사인 자체가 크게 빗나간 사례에서는 모델도 함께 어려움을 겪는다는 한계를 동시에 가진다.

다만 이 우위에 예외는 있다. 희소코드 분석에서는 오히려 잠정 원사인이 CNN_full, GCN_full 모델보다 높은 정확도를 보였고(1건 이하 구간에서 잠정 원사인은 5.7%, CNN_full은 0.4%, GCN_full은 0.9%), 질환 단위에서도 폐혈증·폐렴처럼 원자료에서 사인이 비교적 명확한 질환은 잠정 원사인만으로 평가한 정확도가 모델과 큰 차이가 나지 않는 수준을 보였다. 요컨대 예측모델의 정확도 우위는 대체로 유효하나, 전면적으로는 그렇지 않다고 볼 수 있으며, 이 경계가 어디에 있는지를 앞서 규명하였다.

나. 데이터 품질·특성(질환, 정답라벨 출처)에 따른 평가결과의 차이

종합하면, 모델 성능은 균질하지 않고 정답라벨의 출처와 질환 특성에 따라 뚜렷한 편차를 보인다. 먼저 정답라벨의 기초자료 출처 여부가 정확도에 미치는 영향은 매우 크다. 정답라벨이 잠정 원사인과 일치하는 경우에는 정분류 비중이 68.6%에 달하지만, 범주형 자료 어디에도 없는 ‘출처없음’의 경우에는 정분류 비중이 12.0%에 그친다. 이는 정답라벨이 범주형 정보로 뒷받침되는 사례일수록 사인이 상대적으로 명확하고 판단이 용이한 사례일 가능성이 높기 때문으로 해석되며, 실무적으로는 ‘출처없음’ 사례의 레코드에 대한 별도의 품질관리 및 검증이 필요함을 시사한다.

질환별로는 두 가지 이질적 패턴이 확인되었다. 첫째, 알츠하이머병은 재현율이 58.8%로 가장 낮고 사망자 수 추정비도 76.1%로 과소추정 경향이 뚜렷하다. 둘째, 폐렴·폐혈증은 정밀도가 낮고(각 75.9%, 71.3%), 사망자 수 추정비가 116~118%로 과대추정 경향을 보인다.

이 두 패턴은 오분류패턴 분석에서 그 메커니즘이 구체적으로 드러났다. 알츠하이머병 오분류의 상당수(41.2%)가 정신및행동장애(F), 상세불명이상소견(R), 폐렴 등으로 ‘임종기의 임상소견에 이끌려 대체’되는 경향이 강한 반면, 뇌혈관질환 등 다른 질환의 경우에는 오분류 상당수(44.8%)가 폐렴·폐혈증으로 수렴하는 모습을 보였다.

아울러 오분류 패턴에서 확인된 4가지 유형은 각기 다른 개선 방향을 요구한다는 점에서 의미가 크다. 고혈압성질환과 심장질환, 당뇨병과 신장질환 두 사례(유형②·③)는 서로 다른 질환에서 동일한 메커니즘(병합 분류코드의 원사인 직접사인 선별규칙 미학습 경향)이 반복되고 있는데, 원인이 한 질환에서만 나타나는 것이 아닌 구조적인 패턴임을 보여준다.

이는 기초자료의 난이도와도 관련이 있다. 결합 코드는 정답라벨과 잠정 원사인·건강보험기록 간 거리가 가까운(거리 0~1) 사례에 해당함에도 오분류가 발생한다는 점에서, 단순히 데이터가 부족해서가 아니라 모델이 분류 규칙을 학습하지 못한 데서 비롯된 오류임을 뒷받침한다.

다. 데이터 난이도(전체)와 예측성능 간 관계

네 가지 난이도 지표(잠정 원사인과의 계층적 거리, 정답라벨 출처, 학습데이터의 절대·상대적 규모, 희소코드 여부)는 개별적으로도 유의미하지만, 종합하면 모델 성능은 예측 알고리즘 자체도 중요하지만, 기초자료의 난이도에 의해 상당 부분이 설명된다고 볼 수 있다.

물론 이 결론에는 중요한 예외가 하나 있다. 알츠하이머병은 정답라벨 수가 12,223건(1만 건 이상 구간)으로 데이터 절대량 자체는 충분함에도 정확도가 58.8%에 그쳐, '1천~1만 구간'의 다른 질환들보다도 낮다.

이는 데이터가 많다고 해서 반드시 난이도가 낮아지는 것은 아니라는 사실을 보여주는 사례이며, 알츠하이머병 고유의 난이도, 즉 나향에서 확인한 '임종기 임상소견으로의 대체' 경향이 데이터 양과는 별개의 요인으로 작용하고 있음을 뒷받침한다.

다만 Top1~3 포함비율에서 보았듯이 알츠하이머병의 Top1~3 누적적중률이 95.7%(Top1 대비 36.9%p 개선)까지 오른다는 점은, 모델이 알츠하이머병에 대해 전혀 알지 못하는 것은 아니며 예측후보 안에는 정답을 포함하고 있으나 하나의 후보로 확신을 좁히지 못하는 상태라고 볼 수 있겠다. 즉 알츠하이머병의 높은 난도는 정보 부재가 아닌, 확신이 분산되었기 때문에 기인한다.

반대로 코드 당 데이터 수가 1건 이하인 희소코드 구간에서는 잠정 원사인(5.7%)이 GCN_full(0.9%), CNN_full(0.4%)보다 오히려 높은 정확도를 보인다는 결과는, AI 모델의 근본적 한계인 '학습데이터가 부족할수록 예측 성능이 저하된다'는 특성을 명확히 드러낸다.

반면 잠정 원사인은 매 건마다 WHO 규칙을 독립적으로 적용하였으므로 학습데이터 양과 무관하게 각 레코드별 데이터 품질에 따라 정확도가 결정되는 경향을 보인다. 이는 모델의 전반적 우위가 희소코드 영역에서는 역전된다는 것을 보여주는 근거이며, 난도가 매우 높은 예외 영역에서는 규칙기반 방식이 여전히 유효한 대안이 될 수 있다는 점과 연결된다.

라. 실무적용을 위한 전략

지금까지의 분석결과를 종합하면, 시스템의 실무 적용은 전 질환·전 사례에 하나의 기준을 일괄 적용하기보다 다음과 같이 질환별·상황별로 차별화된 전략을 취하는 것이 합리적이다.

① 신뢰 가능 질환군의 자동화 확대 : 악성신생물·고혈압성질환·간질환 등은 정밀도·재현율·사망자 수 추정비가 고르게 양호하고, Top1 확신도도 이미 높아

Top1~3 확장에 따른 개선폭이 3~6%p에 그친다. 이들 질환군은 예측확률 임계치를 비교적 높게 설정하더라도 채택비율을 충분히 확보할 수 있어, 시스템을 통한 자동화 수준을 우선적으로 적용할 수 있는 후보질환군이다.

② 후보를 제시하는 등의 보조적 활용이 필요한 질환군 : 알츠하이머병·인플루엔자·HIV처럼 Top1 확신도는 낮지만, Top1~3 누적적중률이 크게 개선되는 질환군은, AI가 단일 코드를 자동 확정하기보다 Top3 후보를 사인분류사에게 제시하고 최종 선택을 맡기는 방식이 더 적합하다. 실제 사인분류사들을 대상으로 한 의견수렴에서도 'AI 제시값을 그대로 채택하기보다 최종 검토·보완이 필요하다'는 의견이 공통적으로 확인된 바 있다.

③ 규칙 기반 후처리 적용대상 질환군(또는 조합) : 오분류 패턴 유형② (고혈압성 질환을 심장질환으로 예측), 유형③ (당뇨병을 신장질환으로 예측)은 결합 코드 여부와 특정 임상소견(심부전 등)의 동시출현이라는 명시적 조건으로 판별이 가능하다. 이 두 사례는 서로 다른 질환이지만 동일한 메커니즘이 반복된다는 점에서 같은 규칙기반 로직으로 개선할 수 있다고 보인다.

④ 감염병 질환의 선별적 월별 감시체계 적용 : 감염병 질환 중 코로나19·폐렴 중·호흡기결핵은 연간 성과와 월별 추정비 안정성(RMSE) 모두 상대적으로 양호하여, 과대 또는 과소추정 방향성에 대한 보정을 전제로 감염병 사망 월별 감시체계의 참조 지표로 활용할 여지가 있다. 다만 HIV·바이러스간염·인플루엔자는 절대 발생 건수가 적고 월별 변동성이 커서 현재 수준에서는 단독 예측지표로 신뢰하기는 어렵다.

⑤ 근본적 데이터 보강이 필요한 영역 구별 : 오분류패턴 유형① (원사인과 직접사인 혼동)과 유형④ (상세불명이상소견 대체)는 규칙기반 접근이나 확률 임계치 조정만으로는 해소되지 않는다. 유형①은 사인분류사도 임상 경과 전반을 종합해야 원사인과 직접사인을 구분하는 판단이 가능한 영역이며, 유형④(고의적자해)는 사망진단서 작성 시점에 구체적 자살수단이 확정되지 않은 경우가 대부분이라는 점, 즉 정보 자체가 없다는 것이 확인되었다. 이 두 영역은 모델 개선보다 학습자료 자체의 질적 보강이 근본적인 해결책에 가깝다. 특히 유형④의 경우 자살수단에 대한 정보가 후속 검시절차를 통해 확정되는 시점에 해당 정보를 반영할 수 있는 데이터 연계체계를 구축하는 방법 등이 가능할 것이다.

제3절 (모델3) 약물중독 코드 자동분류 모형

1. 분석개요

가. 분석대상 및 목적

분석 대상 모집단은 2024년 사망원인통계 중 국과수 부검자료가 연계된 사망자 7,423명이다. 기초자료는 일련번호와 국과수 5단분류로만 구성되며, 정답라벨이 없는 상태에서 분석을 진행하였다. 이에 따라 앞선 장들과 달리 정량적 정확도 지표를 산출하기보다는, 분류표 상 약물 강제 추출 규칙 및 약물 우선순위 체계 준수 여부를 점검하고 이를 통해 모델의 오분류 내역을 점검하는 데 중점을 두었다.

나. 추론결과 구성

다음 표는 모델 3의 약물중독 성분 추론 결과를 필드별로 정리한 예시이다. K_RESULT, drugs_json 등 5개의 필드로 구성되어 있다.

<표5-36> 약물중독 성분 추론 결과 필드구성

필드명	주요내용	추론예시
K_RESULT	국과수 5단분류	급성중독-치료약물-급성 복합약물중독(에스시탈로프람, 데스벤라팍신 등)-불-
drugs_json	추출약물종류, 정식약물종류, 질병코드, 소분류명칭, 우선순위를 json형태로 추론	[[{'추출약물종류': '에스시탈로프람', '정식약물종류': '에스시탈로프람', '질병코드': 'T43.2', '소분류명칭': '기타 항우울제(SSRI 등)', '우선순위': 'R10'}, {'추출약물종류': '데스벤라팍신', '정식약물종류': '벤라팍신', '질병코드': 'T43.2', '소분류명칭': '기타 항우울제(SSRI 등)', '우선순위': 'R10'}]]
top_정식 약물 종류	drug_json에서 정식약물종류 추출	에스시탈로프람
reason	추론이유 설명	데스벤라팍신(파생체) → 벤라팍신으로 매핑. 에스시탈로프람은 직접 매칭.
추론결과 (성공여부)	추론결과 정리	TRUE

K_RESULT는 국과수 5단분류 원본으로, 급성중독 여부에서부터 세부 약물 성분에 이르는 계층적 분류 정보를 담고 있다. drugs_json은 모델이 K_RESULT 텍스트에서 약물성분을 추출하여 구조화한 핵심 필드로, 원문에 기재된 표현인 추출 약물종류, 이를 표준 약물명으로 매핑한 정식 약물종류, 해당 질병코드(T코드), 소분류 명칭, 그리고 132종 약물 허용 목록상의 우선순위를 약물별로 함께 제공한다. 한 사례에서 여러 약물 성분이 동시에 검출될 수 있으므로 drugs_json은 리스트 형태로 구성된다.

top_정식 약물종류는 이 중 우선순위 체계에 따라 최종 채택된 대표 약물명만을 별도로 추출한 필드이다. reason은 원문 표현(K_RESULT)을 정식 약물종류로 매핑한 근거를 설명하는 필드로, 예시에서는 데스벤라팍신이 벤라팍신의 파생체이므로 벤라팍신으로 매핑되었고, 에스시탈로프람은 원문과 정식 명칭이 동일하여 직접 매칭되었음을 보여준다.

마지막으로 추론결과(성공 여부)는 이러한 매핑 및 우선순위 판정이 규칙에 부합하게 이루어졌는지를 TRUE/FALSE로 표시한 필드로, 이후 분석에서 규칙 준수 여부를 집계하는 기준으로 활용되었다.

2. 분석결과

가. 약물 검출 빈도 분석

2024년 부검자료 연계 사망자 7,423명 중 약물이 검출된 사망자는 347명이며, 나머지는 약물 미검출로 분류되었다. 동일 사망자에서 2개 이상의 약물이 동시에 검출되는 경우가 있어, 검출된 약물의 총 건수(529건)는 약물 검출 사망자 수(347명)보다 많다. 사망자 1인당 평균 약 1.52건의 약물이 검출되었다.

약물분류표에 작성된 132개 약물 중 실제로 1건 이상 검출된 약물은 48종으로, 본 연구에서 사전에 구성한 약물분류표 대비 검출 실적이 있는 약물의 비율은 약 36.4% 수준이다. 검출 빈도 구간별 분포를 보면, 검출 빈도가 1건에 그친 약물이 12종으로 가장 많고, 20건 이상 검출된 약물은 8종이다.

<표5-37> 약물 검출 빈도별 약물종류 수

검출 빈도	약물종류 수	검출 빈도	약물종류 수
1건	12종	10건 이상 20건 미만	6종
2건 이상 5건 미만	11종	20건 이상	8종
5건 이상 10건 미만	11종	합계	48종

최다 검출 약물은 졸피뎀으로, 전체 529건 중 116건(21.9%)을 차지하여 단일 약물로는 압도적 1위를 기록하였다. 이는 졸피뎀이 처방 빈도가 높은 수면유도제라는 점과, 접근성이 높은 약물이라는 특성이 반영된 결과로 볼 수 있다. 원자료 기준 상위 검출 약물은 졸피뎀(116건), 쿠에티아핀(57건), 에스시탈로프람(37건), 메트암페타민(29건), 플루니트라제팜(28건) 순으로 나타났다.

<표5-38> 약물별 검출 빈도 및 비율 Top-10

검출 순위	약물종류	약물특성	검출 빈도	비율(%)
1	졸피뎀	항정신성의약품 지정 수면유도제. 처방모집단이 넓고 상대적으로 덜 위험하다는 인식 있음.	116	21.9
2	쿠에티아핀	비정형 항정신병제. 수면 목적 처방빈도가 높은 편.	57	10.8
3	에스시탈로프람	SSRI계열 항우울제로서 우울증 환자에게 가장 많이 처방되고 있음.	37	7.0
4	메트암페타민	항정신성의약품 지정 불법각성제(필로폰)	29	5.5
5	플루니트라제팜	벤조디아제핀계 수면제로서 남용위험성이 커 마약류로 관리됨.	28	5.3
6	미르타자핀	항우울제. 진정작용이 강하여 불면을 동반한 우울증에 처방	26	4.9
7	디펜히드라민	1세대 항히스타민제. 수면 목적 오남용 가능	25	4.7
8	알프라졸람	벤조디아제핀계 항불안제	20	3.8
9	트라마돌	마약성(오피오이드계) 진통제	17	3.2
10	벤라팍신	SNRI계열 항우울제	16	3.0
기타	-		158	29.9

상위 10종의 구성을 보면, 졸피뎀, 플루니트라제팜(수면제), 쿠에티아핀, 알프라졸람(항정신병제·항불안제), 에스시탈로프람, 미르타자핀, 벤라팍신, 디펜히드라민(항히스타민제), 트라마돌(진통제) 등 대부분 정신건강의학과 또는 통증클리닉에서 합법적으로 처방되는 항정신성의약품이 차지하고 있으며, 불법 약물은 메트암페타민이 유일하다. 이는 약물중독 사망이 불법 약물 남용보다는 처방 또는 시판 약물의 과량 복용·오남용과 밀접하게 연관되어 있을 수 있음을 시사하며 검출된 약물 대부분이 우울증, 불안증, 불면증 등 정신과적 치료력과 연관되어 있을 가능성이 있음을 보여준다.

나. 약물 우선순위 규칙 준수 여부 및 적용현황

약물이 검출된 347명 중 1명에 2개 이상의 약물이 동시에 기재된 경우, 약물 우선순위 체계에 따라 우선순위 약물 1종을 원사인 대표 약물로 선정한다. 여기서는 우선순위 약물 선정 결과가 우선순위 체계와 정합 여부와 우선순위 적용 현황을 면밀히 살펴본다.

우선순위 규칙 준수 여부는 별도의 프로그램을 통해 347건 전수를 대상으로 두 단계에 걸쳐서 검증하였다. 먼저 drugs_json에 기재된 개별약물(정식약물종류)마다 부여된 우선순위 코드가 약물분류표의 우선순위 코드와 일치하는지 대조하였고, 그 결과 예외 없이 일치함을 확인하였다. 즉 모델이 추론한 약물별 우선순위 코드는 약물분류표와 정합한다.

다음으로는 사망자별로 검출된 약물 중 우선순위 코드값이 가장 높은(최우선 순위) 약물을 식별한 뒤, 이 약물이 실제로 대표약물(추론결과 필드에서는 Top_정식약물종류)로 선정되었는지를 대조하였다. 그 결과 최우선 순위 약물과 추론결과의 Top_정식약물종류가 불일치한 사례는 0건으로, 우선순위 규칙이 전 사례에 걸쳐 정확하게 적용되었음을 확인하였다.

이처럼 규칙 자체의 정합성이 검증된 결과를 바탕으로 약물 우선순위 적용 현황을 우선순위 그룹별로 집계하면 다음과 같다.

<표5-39> 우선순위 그룹별 우선적용 비율

우선 순위	해당 소분류(예시)	검출건수	우선적용 건수	우선적용 비율
R01d~e	마약성 진통제(옥시코돈 · 트라마돌 · 펜타닐 등)	40	36	90%
R02	마취제(프로포폴 · 케타민)	16	15	94%
R03	삼환계 항우울제(아미트립틸린 등)	13	13	100%
R06	항정신병제(쿠에티아핀 · 올란자핀 등)	82	68	83%
R07	기타 진정제(졸피뎀 등)	125	100	80%
R09	정신흥분제(메트암페타민, 디펜히드라민 등)	58	48	83%
R10	SSRI 등 항우울제(에스시탈로프람 등)	112	43	38%
R11	벤조디아제핀 계열	80	21	26%
R12	기타(리도카인 등, 표본 5건)	3	3	100%

우선적용 비율은 해당 약물이 다른 약물과 동반 검출되었을 때, 최종 대표약물로 선정된 비율을 의미한다. 이 비율은 해당 약물이 실제 검출상황에서 다른 약물과 동반 검출되는 빈도와 그 상대적 우선순위를 반영하는 지표로 해석할 수 있다.

R01~R09(마약성 진통제·마취제·삼환계 항우울제·항정신병제·정신흥분제·항히스타민제 등 상대적 고순위 그룹)는 80~100%의 높은 우선적용비율을 보인 반면, R10~R11(SSRI 계열 항우울제·벤조디아제핀 계열 등 상대적 저순위 그룹)은 26~38%로 뚜렷이 낮아지는 패턴이 확인되었다. 이는 해당 약물들이 다른 우선순위의 약물과 동시에 검출되는 사례가 빈번함을 의미한다. 다만 R12는 표본이 3건(리도카인 3건)에 불과해 우선적용비율(100%)의 대표성이 낮음을 유의해야 한다.

한편, 트리아졸람, 미다졸람, MDMA(엑스터시), 가바펜틴, 에스조피클론, 클로바잠, 퍼페나진 등 7종은 검출 건수 자체는 1~6건으로 적고, 검출될 때마다 매번 더 높은 순위의 약물과 동반 검출되어 대표약물로 선정된 사례가 한 번도 없었다.

다. 오분류 패턴 분석

오분류의 발생 원인에 따라 ① 분류표 미반영, ② 발음 유사 오분류, ③ 유사약물 미검출 처리의 세 가지 유형으로 구분된다. 이 가운데 분류표 미반영 유형, 특히 디펜히드라민 및 독시라민 관련 오분류의 비중이 가장 크며, 이는 단일 원인(사전 미등재)에서 비롯된 구조적 문제라는 점에서 우선 개선 대상으로 판단되었다.

<표5-40> 약물 추론 오분류 유형 및 오류 내용

오분류 유형	오분류 내용	해당 약물 표기명
분류표 미반영	기초자료(K_RESULT)에 약물명이 명시되어 있음에도 허용 약물사전에 해당 성분이 없어 강제 제외됨	디펜히드라민(17건), 독시라민(6건)
발음유사 오분류	원문의 축약이나 오타자 표기를 사전 내 유사발음 약물로 잘못 매핑	‘엠이엠에이’→‘메트암페타민’(실제는 MDMA(엑스터시)로 추정) ‘트리돌’을 ‘티오리다진’으로 오분류(실제는 트라마돌로 추정)
유사약물 미검출 처리	사전에 등재된 표기와 원문 표기가 불일치하여 미검출	암페타민, 대마, 아미트리프틸린, 알프랄줄람, 설프랄린, 졸리덴, 엠이엠에이, 트리돌

세 유형의 성격은 서로 다르다. 분류표 미반영은 모델의 추론 오분류가 아니라 모델 설계 초기에 제시한 약물분류표의 포괄 범위 한계에서 비롯된 것으로서, 아무리 원문에 명시되어 있어도 시스템 설계 상 추출되지 않는다. 이는 약물분류표를 보완함으로써 해소가 가능하다. 반면 발음유사 오분류는 사전매칭 단계에서 축약어나 구어체 표기를 유사발음의 등재 약물로 추정한 경우이며, 유사약물 미검출 처리는 반대로 표기 차이에 지나치게 엄격하게 대응하여 실제 약물정보를 놓친 경우이다.

분류표 미반영 문제에 대해서는 분류표 상 T45.0(항알레르기제 및 항히스타민제에 의한 중독)에 해당 약물(디펜히드라민, 독시라민)을 추가하는 방식으로 수정 조치하였다. 이후 모델 추론 결과 다시 검토한 결과, 기존 미반영 건이 모두 정상검출로 전환되어 해당 문제는 해소되었다.

발음 유사 오분류나 유사약물 미검출 처리는 재검증 결과, 잘못된 라벨이 나타나지는 않았으나 아예 약물 미검출로 처리된 사례가 포착되었다(엠이엠에이, 트리돌). 또한, 재추론 과정에서는 아미트리프틸린, 알프랄줄람과 같은 약물 표기에 대해 약물 미검출 사례가 추가로 발견되었다. 즉 모델이 틀린 값을 내지 않는다는 지침은 처리하였으나 여전히 제대로 된 값을 낸다는 보장을 하지 못한다는 점에서 해당 오류가 해결된 것으로 보기는 어렵다.

3. 종합해석

가. 분석결과 요약

첫째, 약물중독 사망은 불법 약물보다 처방, 시판 약물의 과량 복용, 오남용과 밀접하게 연관되어 있다. 검출 빈도 상위 10종 중 메트암페타민을 제외한 9종이 수면제·항정신병제·항우울제·진통제 등 정신건강의학과 또는 통증클리닉에서 합법적으로 처방되거나 약국에서 구입 가능한 약물이었다.

둘째, 약물 우선순위 규칙은 설계된 대로 정확하게 작동하고 있음이 확인되었다. 347건 전수 검증에서 우선순위 코드 불일치가 0건이었고, 2개 이상의 약물이 발견된 경우에는 우선순위에 따라 대표약물을 추출하였음을 확인했다.

셋째, 사전에 작성하여 모델에 입력한 약물사전의 포괄 범위 한계는 오분류에 있어 중요한 요인이다. 디펜히드라민 사례처럼 사전에 없는 성분은 원문에 명시되어 있어도 전혀 추출되지 않는 반면, ‘엠이엠에이’, ‘트리돌’ 같은 축약·오타자 표기는 유사 발음의 등재 약물로 잘못 추정되기도 한다. 즉 허용 약물사전 기반 강제 매칭 원칙은 환각 차단에는 효과적이지만, 사전의 포괄 범위가 완전하지 않음으로 구조적인 오류가 발생할 수밖에 없다.

넷째, 프롬프트 개정만으로는 의도한 오류를 모두 차단할 수 없음이 실증적으로 확인되었다. 1차 분석 결과에 대한 오분류 검사 과정에서 기존에는 정상적으로 처리되던 대마→대마초, 아미트리프틸린→아미트립틸린, 알프라졸람→알프라졸람, 설트랄린→세르트랄린, 졸리템→졸피템 등 5건의 오타자 정규화가 오히려 사라졌다. 이는 특정 오류를 개별적으로 수정하는 방식만으로는 전체 품질을 담보할 수 없으며, 검사 전후의 비교 체계가 갖춰져야 함을 보여준다.

나. 개선방안

약물사전의 계열 단위 일괄 정비가 필요하다. 현재까지의 약물사전 보완은 디펜히드라민, 독시라민 사례처럼 오류가 발견될 때마다 건별로 이루어지고 있다. 이후에는 개별 약물명을 하나씩 검토하기보다, 유사계열(예: 항히스타민제 전체 등)을 단위로 일괄 검토·등재하는 방식으로 전환하여 유사한 누락이 반복되지 않도록 할 필요가 있다.

약물분류표나 프롬프트를 개정할 때마다 기존 검사에서 확인된 오류사례를 자동으로 재검증하는 절차를 모델에 구조화할 필요가 있다. 예를 들면 매 검사마다 오분류건에 대한 few-shot 예시 및 과거 오타자 정규화 성공사례를 재검증하는 절차를 추가하는 것이다. 매번 수작업으로 과거 포착한 오분류를 다시 검증해야 하는 번거로움과 비효율을 방지할 수 있다.

개정 이력의 명시적 관리가 필요하다. 두 번에 걸친 검증과정에서 상이한 약물별 검출건수 및 오분류 내역을 추적하는데 상당히 많은 시간이 소요되었다. 향후에는 분류표 버전, 개정일자, 변경 내용(추가/삭제된 약물, 매핑 규칙 등)을 명시적으로 로그화하여, 결과 수치가 달라질 때 즉시 원인을 추적할 수 있는 체계를 갖출 필요가 있다.

제4절 (모델4) 자살 위험요인 자동분류 모형

1. 분류결과

가. 분류개요

2024년 자살사망자 14,875명을 대상으로 자살 위험요인을 추론하였다. 자살 위험요인 분류체계는 심리·사회적 요인, 정신적 요인, 질병·신체적 요인의 3개 대분류 아래 총 21개의 소분류로 구성되며, 각 소분류는 KCD 코드와 매핑된다(심리·사회적 요인 Z55~Z65, 정신적 요인 F00~F99, 질병·신체적 요인 A00~R99).

분류는 근거추출과 종합판단을 분리한 2단계 파이프라인으로 수행하였다. 1단계에서는 경찰(연간·월간) 사망구분 및 사건개요, 해양경찰청 사건개요, 국립과학수사연구원 5단분류 및 사건개요로 구성된 변사자료 7개 필드를 검토하여 위험요인의 근거가 되는 구절을 모두 추출하고, 각 구절에 대분류·중분류·소분류 라벨을 부여하였다.

2단계에서는 1단계에서 추출된 근거들을 종합하여 근거가 더 많거나 명확한 라벨을 반드시 1개 선정하고, 그 선정된 1개의 위험요인과 함께 작용한 다른 근거가 명확한 경우에 한하여 2순위 위험요인을 추가로 선정하도록 하였다. 이렇게 근거수집과 종합판단의 단계를 분리함으로써 단일 추론단계에서 발생할 수 있는 오류를 줄이고자 하였다.

아울러 추론한 위험요인이 21개 분류체계를 벗어나지 않고 복합위험요인을 추론하더라도 동일한 소분류에 속하는 위험요인을 중복하여 선택하는 것을 방지하기 위해 규칙엔진을 적용하였다. 만약 기초자료 7개 필드 어디에서도 위험요인 근거가 발견되지 않는 경우에는 종합판단을 생략하고 '미상'으로 분류하였다.

이러한 절차를 거쳐 자살사망자 14,875명에 대해 총 17,318건의 자살 위험요인을 추론하였다(1인당 평균 1.16건). 추론 결과에 대해 분류체계 이탈이나 중복 부여 사례가 없는지를 재검증한 결과, 규칙엔진이 의도한 대로 정상 작동하였음을 확인하였다.

나. 위험요인별 빈도

2024년 자살사망자로부터 추론한 자살 위험요인 17,318건 중 대분류 기준으로는 심리·사회적 요인이 42.2%(7,306건)로 가장 큰 비중을 차지하였고, 정신적 요인 37.2%(6,440건), 질병·신체적 요인 15.7%(2,720건)가 뒤를 이었다. 근거를 찾지 못해 미상으로 분류된 사례도 4.9%(852건) 존재하였다.

<표5-41> 자살 위험요인 분류별(대분류, 중분류, 소분류) 빈도 및 비율

대분류	중분류	소분류	빈도4)	비율
심리· 사회적요인	교육적 요인	교육 및 문해에 관련된 문제	72	0.4%
	직업적 요인	취업 및 실업에 관련된 문제	443	2.6%
		위험요인에 직업적 노출	149	0.9%
		기타 및 미상	149	0.9%
	경제적 요인	경제, 주거 상황에 관련된 문제	4,428	25.6%
	사회적 환경 요인	사회적 환경에 관련된 문제	188	1.1%
	가정적 요인	가족 간 관계 문제(Z63.0-Z63.1)	217	1.3%
		가족구성원 부재 (사망, 이혼 등)(Z63.3-Z63.5)	235	1.4%
기타 및 미상		978	5.6%	
법적상황	법적상황(Z65.0-Z65.4)	259	1.5%	
기타 심리· 사회적 요인	기타 심리· 사회적 요인	188	1.1%	
정신적 요인	정신, 행동장애 요인	기분 정동장애(우울증 등)	4,923	28.4%
		스트레스 연관 장애	63	0.4%
		급/만성 정신성 물질 남용	795	4.6%
		급/만성 알코올 남용 장애	150	0.9%
		기타 정신적 요인	509	2.9%
질병· 신체적 요인	질병· 신체적 요인	만성질환	2,224	12.8%
		말기· 불치 질환	185	1.1%
		만성통증	52	0.3%
		신체 기능 저하 / 장애	89	0.5%
		기타 질병· 신체적 요인	170	1.0%
미상	-	-	852	4.9%5)
합계			17,318	100.0%

중분류·소분류 단위로 살펴보면 대분류별로 특정 요인에 집중되는 양상이 뚜렷하다. 정신적 요인(6,440건) 중에서는 기분정동장애(우울증 등)가 4,923건으로 76.4%를 차지해 압도적인 비중을 보였으며, 급·만성 정신성 물질남용(795건, 12.3%)이 뒤를 이었다. 질병·신체적 요인(2,720건) 역시 만성질환이 2,224건(81.8%)으로 대부분을 차지하였고, 말기 불치 질환(185건, 6.8%)이 뒤를 이었다.

4) 자살 위험요인 수를 기준으로 집계하였으므로 자살사망자 수와 다름

5) 미상건수 852건을 자살 위험요인 수 17,318명으로 나눈 값임. 본 연구에서 미상비율 = 미상건수(852명) / 자살사망자 수(14,875명) = 5.7%와 혼동하지 않도록 유의

반면 심리·사회적 요인(7,306건)은 상대적으로 세부 요인이 고르게 분산된 편이다. 단일 소분류로는 경제적 요인이 4,428건(60.6%)으로 전체 위험요인 중에서도 가장 높은 비중(전체 대비 25.6%)을 차지하였으며, 가정적 요인이 1,430건(19.6%)으로, 그다음으로 높았다. 가정적 요인 내에서는 세부 사유가 특정되지 않은 '기타 및 미상'이 978건으로 가정적 요인의 68.4%, 심리·사회적 요인 전체의 13.4%를 차지해, 가족관계 문제의 구체적인 성격을 파악하기 어려운 사례가 상당수 포함되어 있음을 시사한다.

이러한 분포는 자살의 위험요인이 정신 또는 질병 영역에서는 특정 소분류(우울증, 만성질환)로 뚜렷하게 수렴하는 반면, 심리·사회적 영역에서는 경제·가정·직업·법적상황 등 여러 요인이 복합적으로 작용함을 보여준다. 이는 자살예방 정책 설계 시 정신건강 관리와 같은 단일 개입만으로는 충분하지 않으며, 특히 심리·사회적 요인에 대해서는 경제적 지원과 가정적 요인을 진단하는 등 다양한 접근이 필요한 것으로 해석할 수 있다.

다. 라벨 분포

사망자 수 기준으로는 단일라벨이 83.6%(12,432명), 복합라벨이 16.4%(2,443명)를 차지하였다. 단일라벨 12,432명 중에서는 위험요인 근거를 찾지 못해 '미상'으로 분류된 852명을 제외한 11,580명은 분류체계 내 하나의 소분류로 할당되었다.

복합라벨 2,443건은 두 위험요인이 같은 대분류 내 서로 다른 소분류로 구성된 경우와 서로 다른 대분류에 걸쳐 구성된 경우로 나누어 볼 수 있다. 서로 다른 대분류 간 조합이 1,502건(61.5%)으로 다수를 차지하였으며, 그중에서도 심리·사회적 요인과 정신적 요인의 결합이 915건(전체 복합라벨의 37.5%)으로 가장 많이 나타났다. 이는 경제적·가정적 어려움과 같은 심리·사회적 스트레스가 우울증 등 정신적 문제와 함께 작용하는 경우가 흔함을 보여준다.

정신적 요인과 질병·신체적 요인의 결합도 458건(18.7%)으로 두 번째로 높아, 만성질환·말기질환 등 신체적 문제가 정신적 위기와 동반되는 사례가 상당수임을 시사한다. 심리·사회적 요인과 질병·신체적 요인의 결합은 129건(5.3%)으로, 상대적으로 드물었다.

같은 대분류 내 서로 다른 소분류가 복합적으로 작용한 경우는 941건(38.5%)이었다. 이 중에는 기분정동장애와 물질남용의 동반, 심리·사회적 요인 내부 조합, 질병·신체적 요인 내부 조합 등이 포함된다.

<표5-42> 자살 위험요인의 라벨분류별 빈도 및 비율

라벨분류	빈도(건)	비율(%)
단일라벨	12,432	83.6
복합라벨	2,443	16.4
계	14,875	100.0

<표5-43> 자살 위험요인 대분류간 조합건수

위험요인1	위험요인2			
	심리·사회적 요인	정신적 요인	질병·신체적 요인	없음(단일라벨)
심리·사회적 요인	385	915	129	5,492
정신적 요인		411	458	4,245
질병·신체적 요인			145	1,843
미상				852

요약하면 복합라벨의 다수는 서로 다른 영역(심리·사회적↔정신적, 정신적↔질병·신체적)이 함께 작용한 사례이며, 이는 자살위험이 단일 원인이 아니라 여러 영역의 취약성이 중첩되어 나타나는 경우가 많다는 점을 의미한다. 특히 심리·사회적 요인과 정신적 요인의 결합 빈도가 가장 높다는 점은, 정신건강 지원 정책 설계 시 경제·가정 등 심리·사회적 스트레스 요인에 대한 고려가 필수적이라는 점을 시사한다.

라. 10대 청소년 연령층의 위험요인 분석

10대 청소년(11세~20세) 연령층의 위험요인을 분석한 결과, 대분류 단계에서는 정신적 요인(55.8%), 심리·사회적 요인(36.6%), 질병·신체적 요인(2.1%) 순으로 나타났다. 정신적 요인 중에서는 기분 정동장애(우울증 등)가 47.7%로 대부분을 차지하였으며, 심리·사회적 요인에서는 가정적 요인이 15.9%, 교육적 요인이 12.1%로 높은 비중을 보였다. 위험요인 1만 단일 집계한 자살사망자 수 기준에서도 대분류 및 중분류 순위와 비중은 유사하게 나타났다. 미상 비율은 자살 위험요인 기준 5.5%, 자살사망자 수 기준 6.1%였다.

<표5-44> 10대 청소년(11세~20세)의 위험요인별 구성비

대분류	중분류	소분류	비중 ⁶⁾ (%)	
			자살위험 요인기준	자살사망자 기준
심리· 사회적 요인	교육적 요인	교육 및 문해에 관련된 문제	12.1	11.1
	직업적 요인	취업 및 실업에 관련된 문제	0.8	0.9
		위험요인에 직업적 노출	0.2	0.2
		기타 및 미상	0.4	0.2
	경제적 요인	경제, 주거 상황에 관련된 문제	1.9	1.9
	사회적 환경 요인	사회적 환경에 관련된 문제	1.3	0.5
	가정적 요인	가족 간 관계 문제	1.7	1.4
가족구성원 부재(사망, 이혼 등)		0.4	0.2	
		기타 및 미상	13.8	13.9
	법적상황	법적상황	1.9	2.1
	기타 심리·사회적 요인	기타 심리·사회적 요인	2.1	1.9
정신적 요인	정신, 행동장애 요인	기분 정동장애(우울증 등)	47.7	52.7
		스트레스 연관 장애	1.1	0.2
		급/만성 정신성 물질 남용	3.2	1.7
		급/만성 알코올 남용 장애	0.2	0.2
		기타 정신적 요인	3.6	3.3
질병· 신체적 요인	질병·신체적 요인	만성질환	1.3	0.5
		기타 질병·신체적 요인	0.8	0.7
미상	미상	미상	5.5	6.1
합계			100.0	100.0

2. 추가분석

가. 근거분류 및 근거필드 인용현황

전체 근거 인용 29,144건 중 범주형 텍스트 필드(경찰 사망구분, 국과수 5단분류) 인용이 81.2%(23,675건), 서술형 텍스트 기반 필드(사건개요 등 서술자료) 인용이 18.8%(5,469건)로, 범주형 텍스트 기반 필드가 근거 판단의 주된 기준으로 활용되었다.

6) 자살 위험요인 기준: 자살사망자 1명에서 위험요인이 복수(2개)로 검출된 경우 각각을 모두 집계
자살자 수 기준: 자살사망자 1명당 위험요인1만 집계. 자살사망자 수와 일치

필드별로는 '경찰_연간_사망구분'의 인용 비중이 70.6%(20,572건)로 압도적으로 높았으며, 서술형 필드 중에서는 '경찰_연간_사건개요'가 17.6%(5,128건)로 가장 많이 인용되었다. 해석에 유의해야 하는 점은 동일 성격의 연간-월간 자료 간 인용빈도 격차이다.

사망구분 코드는 연간자료(20,572건)가 월간자료(2,680건)의 7.7배이고, 사건개요 서술자료는 연간자료(5,128건)가 월간자료(96건)의 53배에 달한다. 이는 월간자료의 정보량 자체가 제한적이라는 의미보다는 연간자료와 월간자료가 상당부분 중복되지만, 모델이 연간자료부터 순차적으로 참조하면서 월간자료에서는 연간자료에서 이미 입수한 내용은 건너뛰는 내부적 메커니즘이 작동했을 것으로 추측된다.

한편 해경_사건개요의 인용은 16건(0.1%)에 불과했는데, 이는 해경_사건개요의 빈도가 매우 낮고 자살장소의 특성에 따른 현상으로 추정된다.

<표5-45> 근거분류기준 및 근거필드 인용 현황

근거분류기준	근거필드	빈도(건)	비율(%)
범주형 텍스트 기반	경찰_연간_사망구분	20,572	70.6
	경찰_월간_사망구분	2,680	9.2
	국과수_5단분류	423	1.5
	합계	23,675	81.2
서술형 텍스트 기반	경찰_연간_사건개요	5,128	17.6
	경찰_월간_사건개요	96	0.3
	국과수_사건개요	229	0.8
	해경_사건개요	16	0.1
	합계	5,469	18.8
합계		29,144	100

이러한 인용 패턴은 모델이 서술형 텍스트보다 구조화된 범주형 텍스트에 크게 의존하고 있음을 보여준다. 범주형 텍스트 필드는 담당기관이 이미 어느 정도의 해석이 포함되어 있는 단순화 및 집약한 정보이기 때문이다. 그러나 서술형 텍스트 기반 필드는 사건의 맥락을 보다 구체적으로 담고 있으므로, 범주형 텍스트 기반 필드에 대한 의존도가 지나치게 높을 경우 위험요인의 세부적 맥락(예: 복합적 사유)을 놓칠 가능성도 함께 고려할 필요가 있다.

<표5-46> 자살 위험요인 분류별 근거필드 인용현황

대분류	중분류	소분류	인용비율(%)		
			범주형 기반	서술형 기반	혼합형
심리· 사회적 요인	가정적 요인	가족 간 관계 문제	59.4	33.6	6.9
		가족구성원 부재(사망, 이혼 등)	61.3	37.4	1.3
		기타 및 미상	77.3	14.1	8.6
	경제적 요인	경제, 주거 상황에 관련된 문제	79.4	4.1	16.5
	교육적 요인	교육 및 문해에 관련된 문제	62.5	31.9	5.6
	법적상황	법적상황	75.3	14.7	10.0
	사회적 환경 요인	사회적 환경에 관련된 문제	1.1	98.9	0.0
	직업적 요인	기타 및 미상	66.4	26.2	7.4
		위험요인에 직업적 노출	64.4	27.5	8.1
		취업 및 실업에 관련된 문제	55.8	38.4	5.9
	기타 심리·사회적 요인	기타 심리·사회적 요인	26.6	71.3	2.1
정신적 요인	정신, 행동장애 요인	급/만성 알코올 남용 장애	25.3	70.0	4.7
		급/만성 정신성 물질 남용	46.4	41.6	11.9
		기분 정동장애(우울증 등)	72.4	8.1	19.5
		기타 정신적 요인	76.8	10.8	12.4
		스트레스 연관 장애	4.8	95.2	0.0
질병· 신체적 요인	질병·신체적 요인	기타 질병·신체적 요인	50.6	45.3	4.1
		만성질환	66.9	14.3	18.8
		만성통증	1.9	98.1	0.0
		말기·불치 질환	1.6	84.3	14.1
		신체 기능 저하 / 장애	1.1	96.6	2.2

나. 경찰청 자살동기 체계와 교차비교

모델4의 21개 소분류체계와 경찰청 자살동기 분류(10개 대분류) 간 대응관계를 경찰청 분류 항목 기준으로 살펴보면, 각 항목이 모델4의 특정 소분류 하나에 뚜렷하게 수렴하는 경향이 확인된다. 경찰청의 '경제생활 문제'로 분류된 사건의 85.4%가 모델4의 경제적 요인으로, '가정 문제'의 89.0%가 가정적요인 문제로, '정신적정신과적문제'의 72.8%가 기분정동장애(우울증 등)로, '육체적 질병문제'의 73.7%가 질병·신체적요인 중 만성질환으로 대응되어, 모두 단일 소분류가 지배적인 대응관계를 보였다. '직장 또는 업무상의 문제' 역시 90.7%가 직업적 요인으로 매핑되었다.

이러한 대응관계는 10대 청소년(11세~20세)만 살펴봐도 비슷하게 나타났다. ‘경제생활문제’와 ‘직장 또는 업무상의 문제’는 각각 ‘경제적 요인’과 ‘직업적 요인’으로 100% 매칭되었고, ‘정신적·정신과적 문제’의 ‘기분 정동장애(우울증 등)’ 대응비율도 81.3%로서 전체기준(72.8%)보다 높았다. ‘육체적 질병 문제’ 역시 만성질환, 말기불치질환 등 질병·신체적 요인에 대부분 대응되었다.

<표5-47> 자살 위험요인 분류결과와 경찰청 자살동기 분류체계 교차비교(전체연령)

(단위: %)

대분류	중분류	소분류	1.가정 문제	2.경제생활 문제	3.육체적 질병 문제	4.정신적 정신과적 문제	5.직장 또는 업무상의 문제	6.남녀 문제	7.사별 문제	8.학대 또는 폭력 문제	9.기타, 10.미상	합계
심리,사회적요인	교육적 요인	교육 및 문화에 관련된 문제	0.3	1.9	0.3	0.2	49.1	0.7		16.7	3.2	0.4
		취업 및 실업에 관련된 문제		0.3		0.2	21.4				0.6	2.6
	직업적 요인	위험요인에 직업적 노출		0.2	0.1	0.2	20.3				0.4	0.9
		기타 및 미상									1.0	0.9
	경제적 요인	경제, 주거 상황에 관련된 문제	0.8	85.4	0.4	1.6	1.5	1.3	0.6		2.4	25.6
	사회적 환경 요인	사회적 환경에 관련된 문제	0.4	1.3	0.6	1.1				16.7	2.1	1.1
		가족 간 관계 문제	18.2	0.4	0.1	0.6	0.2				1.0	1.3
	가정적 요인	가족구성원 부재(사망, 이혼 등)	0.4	0.2	0.4	1.0			83.1		0.3	1.4
		기타 및 미상	70.4	1.0	0.2	0.8	1.1	91.3		66.7	3.3	5.6
	법적상황	법적상황		0.1		0.1	0.2				14.1	1.5
정신적 요인	기타 심리,사회적 요인	기타 심리,사회적 요인	0.3	1.3	0.2	0.4	0.6	0.7			4.8	1.1
		기분 정동장애(우울증 등)	2.1	3.3	4.8	72.8	3.6	4.3	9.3		4.3	28.4
	정신, 행동장애 요인	스트레스 연관 장애		0.1	0.3	0.7	0.2		0.6		0.2	0.4
		급/만성 정신성 물질 남용	4.4	2.6	3.3	6.1	0.9	1.3	3.5		8.7	4.6
		급/만성 알코올 남용 장애	0.1	0.2	0.4	1.7		0.3	1.2		1.1	0.9
		기타 정신적 요인	0.5	0.3	0.5	7.3	0.2				1.3	2.9
		만성질환	1.5	0.9	73.7	2.8	0.8		1.7		2.3	12.8
		말기·불치 질환	0.1		6.4	0.2					0.2	1.1
	질병, 신체적 요인	만성질환		0.0	1.7	0.1					0.0	0.3
		신체 기능 저하 / 장애		0.1	2.3	0.3					0.2	0.5
	미상	기타 질병신체적 요인	0.1	0.1	4.2	0.5					1.1	1.0
			0.4	0.1	0.3	0.1					47.3	4.9
합계			100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

<표5-48> 자살 위험요인 분류결과와 경찰청 자살동기 분류체계 교차비교(11~20세)

(단위: %)

대분류	중분류	소분류	1.가정 문제	2.경제생활 문제	3.육체적 질병 문제	4.정신적 정신과적 문제	5.직장 또는 업무상의 문제	6.남녀 문제	7.사별 문제	8.학대 또는 폭력 문제	9.기타, 10.미상	합계
심리,사회적요인	교육적 요인	교육 및 문화에 관련된 문제				3.4					40.3	12.1
		취업 및 실업에 관련된 문제					75.0				0.8	0.8
	직업적 요인	위험요인에 직업적 노출					25.0				0.0	0.2
		기타 및 미상				0.4					0.8	0.4
	경제적 요인	경제, 주거 상황에 관련된 문제		100.0		0.4					0.0	1.9
	사회적 환경 요인	사회적 환경에 관련된 문제				1.5				33.3	0.8	1.3
		가족 간 관계 문제	14.3			0.4					0.8	1.7
	가정적 요인	가족구성원 부재(사망, 이혼 등)				0.4			100.0		0.0	0.4
		기타 및 미상	73.8			0.4		87.0		66.7	9.2	13.8
	법적상황	법적상황									7.6	1.9
정신적 요인	기타 심리,사회적 요인	기타 심리,사회적 요인	2.4			0.4					6.7	2.1
		기분 정동장애(우울증 등)			20.0	81.3		8.7			4.2	47.7
	정신, 행동장애 요인	스트레스 연관 장애				1.5					0.8	1.1
		급/만성 정신성 물질 남용	4.8			3.0		4.3			3.4	3.2
		급/만성 알코올 남용 장애				0.4					0.0	0.2
		기타 정신적 요인	4.8			5.2					0.8	3.6
		만성질환			60.0	1.1					0.0	1.3
		기타 질병신체적 요인			20.0	0.4					1.7	0.8
	질병, 신체적 요인	질병, 신체적 요인										
	미상	기타 질병신체적 요인									21.8	5.5
합계			100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

한편, 경찰청 분류상 ‘기타’ 및 ‘미상’ 항목은 새로운 자살 위험요인 분류체계에서 오히려 다양한 위험요인으로 세분되어 나타났다. 전체 기준으로는 ‘기타’ 및 ‘미상’의 47.3%가 여전히 ‘미상’으로 분류되었으나, 나머지는 ‘법적상황’(14.1%), ‘급/만성 정신성

물질 남용'(8.7%), '기분 정동장애(우울증 등)'(4.3%) 등으로 흩어져 분류되었다.

반면 10대 청소년 기준으로는 '기타' 및 '미상' 중 '미상'으로 재분류된 비중이 21.8%에 그쳤고, 오히려 '교육 및 문화에 관련된 문제'(40.3%)로 분류되는 비중이 가장 컸으며, 이 밖에 '가정적 요인'의 '기타 및 미상'(9.2%), '법적상황'(7.6%) 등으로 다양하게 분류되었다. 이는 경찰청의 단일분류체계에서는 포착하지 못했던 세부 위험요인들이 실제로는 다수 존재했으나, 그동안 '기타' 및 '미상'이라는 범주 안에 묻혀 유실되었을 가능성을 추측할 수 있다.

3. 품질검사 결과

가. 품질검사 개요

모델4의 결과에 대한 품질검사는 총 2차에 걸쳐 이루어졌다. 먼저 1차 분석은 경찰 연간자료, 해양경찰청 자료, 국립과학수사연구원 자료를 이용하여 추론한 결과를 바탕으로 분석하였다. 그리고 2차 분석에서는 기초자료에 경찰 월간자료(사망구분·사건개요)를 추가하였고, 1차 분석에서 발견된 오분류 패턴을 환류하여 반영하였다.

<표5-49> 품질검사 기초자료 필드 구성

필드명	내용
SEQ	일련번호
T_DTH_WON_CODE	최종 원사인
POL_YEAR_MEAN	경찰 연간 사망구분
POL_YEAR_COMMENT	경찰 연간 사건개요
POL_MONTH_MEAN	경찰 월간 사망구분
POL_MONTH_COMMENT	경찰 월간 사건개요
NMPA_COMMENT	해경 사건개요
NFS_GUBUN	국과수 5단분류
NFS_COMMENT	국과수 사건개요

품질검사는 데이터를 성격이 다른 세 가지 위험군으로 나누고, 각각에 대해 ①미상 분류 건, ②희소분류, ③임의추출 표본의 3단계로 실시하였다. ①미상 분류 건은 모델이 근거를 찾지 못해 '미상'으로 처리한 사례 중 추가분류 가능성을 점검하기 위한 것이고, ②희소분류는 학습·판단 근거가 절대적으로 적어 오분류 위험이 상대적으로 큰 특정 소분류를 선별하여 점검한 것이며, ③임의추출은 위 두 범주에 해당하지 않는 일반 표본 중 약 1%를 무작위로 추출하여 전반적인 분류의 품질을 검증하기 위한 것이다.

즉 ①·②는 오류 발생 가능성이 상대적으로 높은 취약 지점을 점검하는 단계이고, ③은 그 외 영역의 품질을 대표성 있게 확인하는 절차로서 서로 보완적인 관계에 있다.

<표 5-50> 품질검사 단계별 검사대상

구분	1차 분석	2차 분석
전체 분석대상자 수	14,872명	14,875명
전체 분석대상 위험요인	17,381건	17,318건
미상 검사대상	1,526건	852건(5.7% ⁷⁾)
희소분류 검사대상	297건 (5개 소분류)	276건 (4개 소분류)
임의추출 검사대상	153건	132건

1차 분석결과에 대해 오류분석 결과를 반영하여 2차 분석까지 진행한 결과, 미상 검사대상은 1,526건에서 852건으로 44.2% 감소하였다. 희소분류 검사대상도 5개 소분류(297건)에서 4개 소분류(276건)로 줄어, 오분류 품질검사 및 환류가 모델의 성능을 일부 개선한 것으로 판단된다.

1차 분석 품질검사에서는 미상 사례 중 사망구분·사망수단 필드가 채워진 경우에만 추가분류가 가능했던 반면, 해당 필드가 비어있고 사건개요 등 서술자료만 있는 경우에는 대부분 정황 정보만 존재하여 분류가 어려웠음을 확인하였다. 2차 분석에서는 이러한 취약점을 보완하고 경찰 월간자료를 추가하였으며 이것이 모델 성능 향상의 주요 요인으로 작용했을 가능성이 있다.

나. 미상 검사 단계

1차 분석결과에 대한 피드백과 경찰 월간자료 추가 반영으로 미상 모수 자체가 1,526건에서 852건으로 줄었으며, 그 중 사망구분 필드가 채워져 있는 26건은 전부 분류가 가능하였다. 나머지 826건 중 서술자료(사건개요 등)가 있는 638건에 대해 검사한 결과, 해당 데이터는 정황 판단에 그쳐 분류가 불가능했고, 서술자료조차 없는 188건은 해석 근거가 전무하였다. 최종적으로 852건 중 26건(3.1%)만 추가 분류 가능한 것으로 확인되었다.

7) 위험요인 미상인 건수는 852건, 자살사망자 수 14,875명 대비 5.7%수준

<표 5-51> 미상 검사 단계 검사결과

구분	자살 위험요인 추출가능성	미상 건수	검사결과			
			추출 가능	추출불가		
				합계	정황만 있음	자료 없음
10_사망구분 있음	자살 위험요인 추출가능	26	26	-	-	-
20_사망구분 비어있음		826	-	826	-	-
21_사건개요 및 기타 필드가 채워져있음	사건개요 및 기타필드에서 해석여지 있음	638	-	638	638	-
22_사건개요 및 기타필드가 비어있음	해석할 수 있는 자료가 없음	188	-	188	-	188
전체		852	26	826	638	188

이 결과는 역설적으로 1차 분석의 품질개선이 실질적으로 이루어졌음을 뒷받침한다. 1차 분석에서 미상으로 남았던 사례 중 상당수(549건, 1차 미상의 36.0%)는 사망구분 필드가 채워져 있었음에도 미상으로 처리되었었다. 이 점을 보완하고 2차 분석에서 경찰 월간자료를 반영한 결과 2차 분석 당시에는 미상(852건) 중에는 사망구분이 채워진 채로 남은 사례가 26건으로 줄어들었다. 따라서 범주형 필드를 통해 손쉽게 회수할 수 있었던 미상 사례는 1차 분석결과 후 환류 과정에서 대부분 처리되었고, 2차 분석에서는 애초에 범주형 정보가 존재하지 않는 근본적으로 정보량이 거의 없는 데이터들만으로 압축된 것이다.

서술자료 활용 측면에서도 유사한 패턴이 나타난다. 1차 분석에서는 사건개요 등 서술자료가 있는 767건 중 7건이 추가로 해석 가능했던 반면, 2차 분석에서는 서술자료가 있는 638건 전부가 정황 판단에 그쳐 단 1건도 추가로 분류하지 못하였다.

이는 1차 검사에서 확인되었던 ‘사건개요에서 드물게 해석 가능한 사례’가 모델 개선을 통해 반영되었고, 2차 분석 시에는 포착되지 않았음을 의미한다. 그리고 2차 분석의 미상 852건은 더이상의 규칙이나 프롬프트 개선으로는 축소하기 어려운, 원천 자료의 부재에서 기인하는 구조적 미상으로 볼 수 있다.

다. 회소분류 검사 단계

회소분류 검사대상 소분류 중 두 곳에서 특징적인 오분류 패턴이 확인되었다. 첫째, 심리·사회적 요인의 교육적 요인에서는 학생신분이라는 정황 하나만으로 교육적 문제로 분류하는 경향이 나타났다. 그러나 학생 신분 자체는 자살 위험요인이자보다 인적 속성에 가까우므로, 다른 위험요인(예: 학업 스트레스, 따돌림 등)과 결부되어 언급된 경우에 한해 교육적 문제로 인정하고, 단순히 재학 중이라는 사실만으로는 교육적 문제로 분류하지 않도록 판단 기준을 보완할 필요가 있다.

둘째, 정신적 요인의 정신·행동장애 요인에서는 정황만으로 판단하기 어려운 사례가 다수 확인되었다. 특히 신체장애나 기능 저하가 있다는 사실만으로 스트레스 연관 장애로 연결 짓는 오분류가 나타났는데, 신체적 장애·기능 저하는 그 자체로는 질병·신체적 요인에 해당하는 정보일 뿐 정신적 스트레스의 근거가 되기 위해서는 다른 정황과 결합되어야 하는 것이다. 두 위험요인이 함께 언급되는 경우와 신체적 정황만 있고 정신적 반응에 대한 별도 서술이 없는 경우를 구분하는 판별기준 보완이 필요하다.

<표 5-52> 회소분류 검사 단계에서 포착된 오분류 특징

대분류	중분류	오분류 특징
심리·사회적 요인	교육적 요인	학생신분 정황만으로 교육적 문제로 오인
정신적 요인	정신·행동장애 요인	정황만으로 판단하기 어려운 사례 특히 신체장애, 기능저하가 있다는 이유만으로 스트레스 요인으로 잘못 연결

라. 임의추출 검사 단계

미상 단계와 회소분류 단계 대상을 제외한 나머지 데이터에서 약 1%를 임의 추출하여 품질을 검증하였다. 검사 결과 표본 132건 중 오분류 또는 개선 여지가 있는 사례는 29건(22.0%)으로 확인되었다.

오분류 유형은 다양하게 발견되었다. 그 중 구체적 병명이 확인되지 않는데도 ‘만성 질환’으로 분류한 경우가 가장 많았다. 사실 이는 ‘기타 질병·신체적 요인(3-1-9)’으로 분류하는 것이 타당하며, 소분류 판단에 필요한 최소 정보(구체적 병명)가 없을 때는 상위 포괄 범주가 아니라 ‘기타’ 범주로 처리해야 한다는 원칙을 명확히 적용할 필요가 있다.

자살 수단을 위험요인으로 오인한 경우도 빈번히 포착되었다. 농약 음독을 위험요인으로 오인한 사례가 일부 확인되었는데, 음독은 자살의 수단이지 위험요인 자체가 아니다. 수단과 위험요인을 혼동하는 이 오류 유형은 앞서 오분류 패턴 분석(유형④ 사인불명 대체형)에서도 확인된 바 있다.

그 외 근거 없는 정황만으로 위험요인을 단정하는 경우가 있었다. ‘무직’이라는 정황만으로 직업적 요인을 부여한 사례가 소수 확인되었는데, 무직 상태 자체가 아니라 그로 인한 경제적·심리적 고통이 원인으로 명시되어야 분류가 타당하다. 유사하게 단순 음주 정황만으로 알코올 중독으로 오분류한 사례도 있어, 정황과 위험요인을 곧바로 등치시키는 패턴이 여러 소분류에서 반복적으로 나타남을 확인하였다.

임의추출 단계에서 확인된 오류들은 회소분류 단계와 달리 특정 소분류에 국한되지 않고 전반에 걸쳐 나타난다는 점에서, ①구체적인 요인이 확인되지 않을 경우 상위 분류체계로 임의분류하는 성향의 방지장치 마련, ②수단과 위험요인의 구분, ③정황 정보만으로 위험요인을 단정하지 않는다는 원칙 등 세 가지를 모델에 가이드라인으로 규정할 필요가 있다.

<표5-53> 임의추출 검사 단계에서 포착된 오분류 유형

유형	세부 사유
만성질환 임의분류	구체적 병명 없이 만성질환으로 분류 → 3-1-9가 타당
자살 수단을 위험요인으로 오인(농약)	농약은 위험요인이 아니고 수단
정황을 위험요인으로 오인(무직)	무직은 정황이고 위험요인이 드러나지 않음
자살 수단을 위험요인으로 오인(질식사)	질식사는 위험요인이 아님
정황을 위험요인으로 오인(음주)	음주는 단순 정황
단순 오분류	연인·친구 사망은 가족구성원 부재 아님

4. 종합해석

1~3의 결과를 종합하여 모델4의 현황과 한계를 정리하면, 먼저 위험요인 구조적 특징이 확인된다. 2024년 자살 사망자의 위험요인은 정신·행동장애(37.2%)와 경제적 요인(25.6%)이 전체의 62.8%를 차지하였다. 10대(11세~20세)에서는 기분 정동장애(우울증 등)가 47.7%로 대부분을, 그리고 가정적 요인이 15.9%, 교육적 요인이 12.1%로서 다른 패턴을 보인 바 있다.

근거 판단에 있어 모델의 특징이 확연히 나타난다. 전체 근거의 81.2%가 범주형 텍스트 기반 필드(사망구분 등)에서 도출되어, 서술형 텍스트(사건개요)의 기여도는 제한적이다. 다만 사회적 환경 요인·스트레스 연관장애·말기질환 등 일부 소분류는 텍스트 기반 근거 비중이 높아, 범주형 필드만으로는 포착되지 않는 세부 요인 식별에 서술자료가 필수적임을 보여준다.

경찰청 분류체계와의 관계에서 위험요인의 식별도가 향상될 가능성을 확인하였다. 모델4는 경찰청의 포괄적인 '기타·미상' 분류에 흡수되던 교육적 요인·법적상황 등 세부 요인을 추가로 식별하고 있어, 기존 경찰청 통계보다 위험요인의 해상도를 높이는 효과가 있는 것으로 판단된다.

품질개선의 효과도 확인되었다. 경찰 월간자료를 반영하고, 1차 분석결과에 대한 검사를 수행한 결과 미상 건수는 44.2% 감소하였으나, 잔여 미상 852건 중 분류가능 비율은 오히려 1차(36.4%) 대비 2차(3.1%)에서 크게 낮아졌다. 이는 모델 성능이 악화된 것이 아니라, 상대적으로 판단이 쉬운 사례(사망구분·사망수단이 명확한 경우)가 먼저 해소되었고, 남은 미상 건은 서술자료로도 판단이 어렵고 정보가 부족한 사례로 압축되었기 때문으로 해석된다. 즉 잔여 '미상' 건수의 추가 감축은 수사기록 자체의 정보보장이 필요한 영역임을 시사한다.

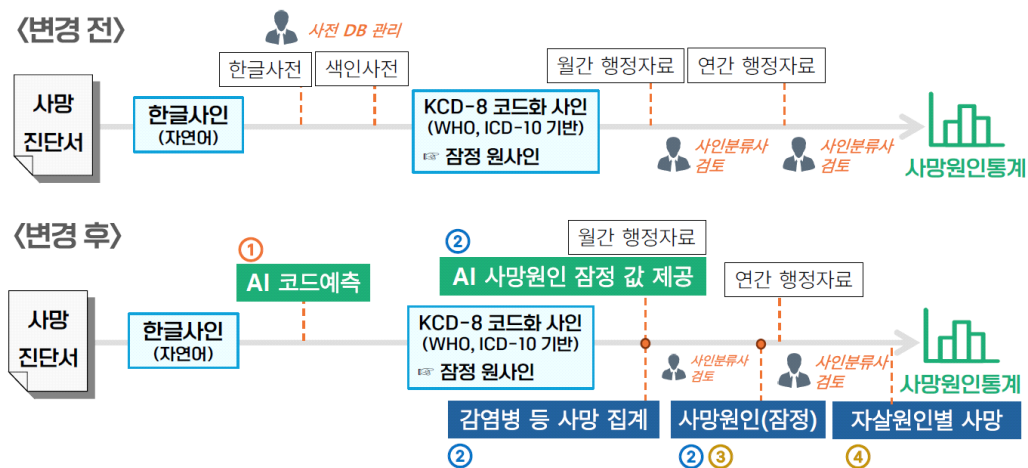
반복되는 오분류 패턴에 대한 모델 개선이 필요하다. 임의추출·희소분류 검사 결과 1차·2차 공통적으로 ① 자살 수단(농약, 질식 등)을 위험요인으로 오인, ② 정황 정보만으로 특정 요인을 단정, ③ 복합 요인이 있음에도 주된 요인만 반영하는 경우가 확인되었다. 이는 모델 프롬프트와 후처리 로직에서 공통적으로 보완이 필요한 지점으로 판단된다.

제 6 장

AI 적용에 따른 업무 개선 및 통계서비스 확장 방안

본 장에서는 제4장에서 제시한 4종 AI 모델의 개발 결과와 제5장의 분석 성과를 바탕으로, 사망원인통계 작성업무의 개선 방향과 통계서비스의 확장 방안을 제시한다. 본 연구의 AI 모델은 사망원인분류 업무를 자동으로 확정하는 도구가 아니라, 원사인 분류 결과를 사인분류사에게 우선 제안하여 검토 업무를 지원하는 업무지원형 모델이면서 AI 예측값을 이용하여 정책적으로 필요한 서비스를 추가 개발하는 모델이다. 이에 대하여 현재의 사망원인통계 작성 과정 및 시스템 구조의 전반적인 개선방향은 <그림 6-1>과 같다. 기존의 담당자 및 사인분류사의 수작업 검토 업무를 AI 적용을 통해 간소화 및 대체하여 행정 효율성을 증대시키고, 그 과정에서 만들어진 결과를 정책적 서비스에 활용하는 방안이다.

<그림6-1> 사망원인통계 작성 과정 개선방안



이러한 관점에서 본 장은 ① 사망원인통계 서비스 구조의 개선 방향, ② 사인분류 업무체계 및 사인선정시스템의 단계별 개선안, ③ 감염병 사망 감시 및 자살 위험요인 분석 등 정책수요 대응 서비스의 확장 방안을 차례로 다룬다.

제1절 사망원인통계 서비스의 확대와 정책적 함의

사망원인통계는 보건정책 수립의 핵심 기초자료임에도, 현행 생산체계는 확정통계 중심으로 운영됨에 따라 시의성 측면에서 구조적 한계를 지닌다. 특히 사망진단서 기반 원사인 분류 이후 사인분류사의 검토를 거쳐 최종 확정되는 과정은 개별 사망자의 원사인 정확도 확보에 기여하는 반면, 보건정책 대응의 측면에서는 상당한 시간 지연을 초래한다. 특히 감염병과 같이 즉각적인 정책 대응이 요구되는 영역에서는 정보 활용의 적시성을 제약한다. 이러한 문제는 고령화에 따른 사망자 수 증가와 사인 복잡도 상승으로 인해 더욱 심화되고 있다.

이러한 한계를 보완하기 위한 방안으로 인공지능 기반의 잠정 원사인 예측 체계를 도입하고, 이를 활용한 잠정 사망원인통계 서비스의 구현을 위한 업무절차 및 사인선정시스템 개선을 제안한다. 본 접근은 사인분류사의 최종 판단 구조와 WHO 원사인선정 기준을 유지하면서, AI가 사전적으로 원사인 후보를 제시하는 보조적 구조를 채택함으로써 정확성과 효율성을 동시에 확보하는 것을 목표로 한다.

본 연구에서 개발한 잠정 원사인 예측 모델(모델 2)은 사망진단서 텍스트와 건강보험 진료기록을 결합하여 원사인을 예측하는 구조로 설계되었다. 한국어 자연어 처리에 특화된 언어모델(KcELECTRA)에 KCD 계층 구조를 반영한 그래프 합성곱 신경망(GCN) 기반 분류 구조를 결합함으로써, 텍스트의 의미적 정보와 KCD 코드 간 계층 패턴을 동시에 반영하였다. 그 결과 제5장 제2절에서 제시한 바와 같이 정책 활용이 가능한 수준의 예측 성능을 확보하였다.

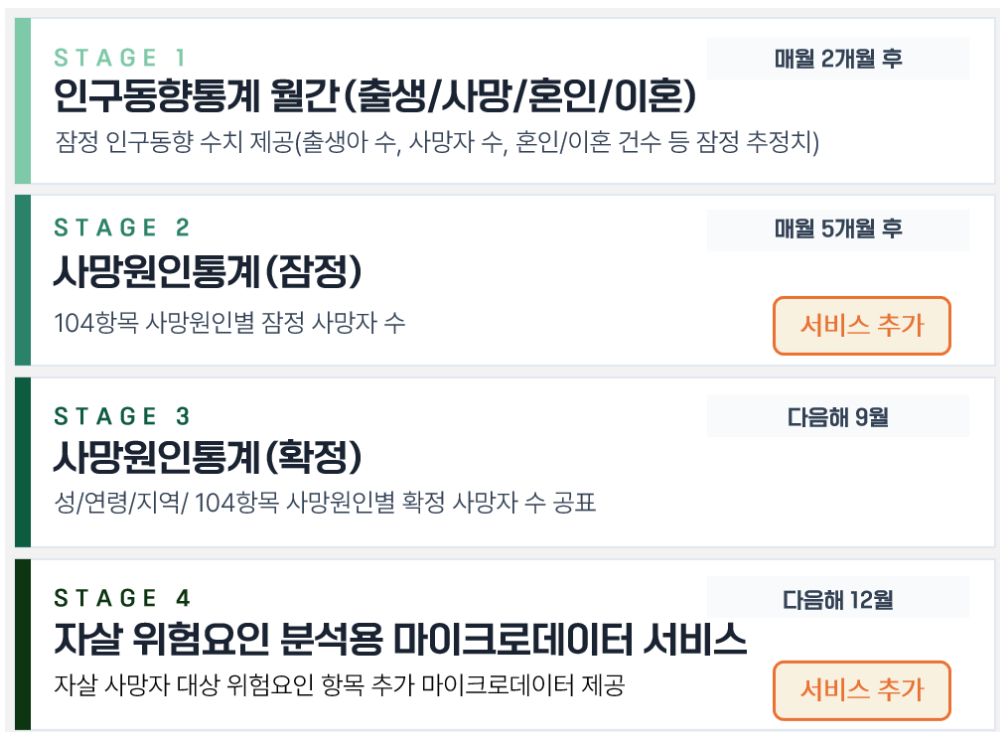
또한 자살 위험요인 분류 모델(모델4)은 경찰청, 해양경찰청, 국립과학수사연구원의 행정자료를 이용하여 자살 위험요인을 자동 분류하는 구조로 설계하였다. 특히, 해당 모델에서는 행정자료의 특성상 자유형 텍스트 형태에서 필요한 정보를 추출하는 과정이 필요하다. 이를 위해 거대언어모델을 이용하여 여러 가지 제약조건과 명령문을 통해 유의미한 정보를 추출할 수 있는 환경을 개발하였다.

이러한 기술적 기반 위에서 잠정 사망원인통계 서비스는 다음과 같은 방향으로 구현될 수 있다. 첫째, 월 단위 주기의 잠정통계 생산을 통해 기존 확정통계 대비 공표 시차를 큰 폭으로 단축할 수 있다. 둘째, 감염병 사망자에 대한 적시성 있는 감시체계를 구축함으로써 특정 질병군의 급증 또는 계절적 변동을 조기에 탐지할 수 있다. 셋째, 확정통계 공표 이후에는 자살 위험요인 자동분류 결과를 통계와 연계함으로써 정책 수요에 부합하는 심층 통계의 생산이 가능해진다.

결과적으로 AI 기반 잠정통계 서비스의 도입은 사망원인통계를 사후적 기록 통계에서 정책 대응을 지원하는 동적 정보체계로 전환하는 계기가 될 것으로 판단된다. 이는 통계의 생산 주기를 단축하는 차원을 넘어, 통계의 기능을 '설명' 중심에서 '예측 및 대응' 중심으로 확장하는 구조적 변화로 이해될 수 있다.

AI 적용에 따른 사망 및 사망원인통계의 서비스 구조는 <그림6-2>과 같이 기존 공표 체계에 ① 사망원인통계 잠정통계와 ② 자살 위험요인 분석용 데이터 제공의 2가지 서비스를 추가하여 구현될 수 있다. 세부적으로는 사망 발생으로부터 매월 2개월 후에는 사망자 수에 대한 잠정 수치를 공표하고, 5개월 후에는 사망원인에 대한 잠정 추정치를 공표한다. 이후 다음 해 9월에는 사인분류사의 검토를 거친 최종 사망원인통계를 공표하고, 12월에는 자살 위험요인을 추가한 마이크로데이터를 서비스한다. 그 결과 사망원인통계의 서비스 단계는 ① 잠정 사망자 수(2개월), ② 잠정 사망원인(5개월), ③ 확정 사망원인(다음 해 9월), ④ 자살 위험요인 연계 데이터(다음 해 12월)의 4단계로 구성되어 정책 및 민간 연구 수요에 맞는 서비스 제공 체계가 구축된다.

<그림6-2> 사망원인통계 서비스 구조(안)



제2절 사인분류 업무체계 및 업무시스템의 구조적 전환

현행 사망원인 분류 시스템은 한글 사인명에 대한 한글사전과 색인사전 매칭을 기반으로 KCD 코드가 자동 부여되고, WHO 지침에 따라 잠정 원사인이 도출된 후 사인분류사가 이를 수작업으로 검토·확정하는 방식으로 운영된다. 이러한 구조는 국제 기준에 부합하는 일관성을 유지한다는 장점이 있으나, 대량 자료 처리에 상당한 인적 자원을 요구하며, 반복적인 수작업으로 인해 오류 발생 가능성과 공표 지연 문제가 발생할 수 있다.

AI 도입 이후의 사인분류 업무는 기존 구조와 다른 방식으로 재편된다. 핵심적인 변화는 사망진단서 입력 단계에서 AI가 원사인 후보를 확률 기반으로 사전 제시하는 '선행 제안 구조'의 도입이다. 이에 따라 사인분류사의 역할은 직접 분류 수행자에서 AI 제안 결과를 검토하고 최종 판단을 내리는 검토자로 전환된다. 이는 인간과 AI의 협업 체계를 기반으로 정확성과 효율성 간의 균형을 확보하는 데 목적이 있다.

업무 흐름 측면에서, 기존에는 '한글사전 기반 코드 매칭 → 시스템 자동분류 → 사인분류사 검토'의 순서로 진행되었으나, AI 도입 이후에는 'AI 코드화 → 사전 매칭 보완 → 시스템 자동분류 → AI 원사인 예측 → 사인분류사 검토'의 순서로 프로세스가 재구성된다. 이 과정에서 AI는 단순한 코드 연계 및 분석을 넘어, 다수의 후보 코드와 확률 정보를 함께 제공함으로써 분류 판단의 근거를 제시하는 역할을 수행한다.

이와 같은 시스템 전환을 통해 업무 효율성 측면의 개선 효과를 기대할 수 있다. 우선 정확한 원사인을 찾기 위한 반복적인 코드 분석 작업이 일정 수준 자동화됨에 따라 사인분류사의 업무 부담이 완화되고, 대량의 사망자료를 더욱 신속하게 처리할 수 있다. 또한 약물중독 사망과 같이 복합적 판단이 요구되는 영역에서는 AI 기반 약물 성분 추출 및 우선순위 판단 결과를 검토에 활용함으로써 분류 오류 가능성을 낮추는데 기여할 수 있다. 더 나아가 WHO 기준 및 ICD 분류체계 적용의 일관성을 유지하면서도, ICD-11 전환과 같은 환경 변화에 유연하게 대응할 수 있는 기반이 마련된다.

한편, AI 기반 사인분류 시스템은 단순한 코드 분류 기능을 넘어 확장할 수 있는 구조를 가진다. 대표적으로 행정자료의 서술형 텍스트를 활용한 자살 위험요인 자동분류체계는 기존 사망원인통계가 제공하지 못했던 위험요인 구조에 대한 심층 정보를 제공할 수 있다. 또한 약물중독 사망자의 경우 약물명 추출과 KCD 코드 매핑을 자동화함으로써 다중 원인 구조를 체계적으로 분석할 수 있는 기반을 제공한다.

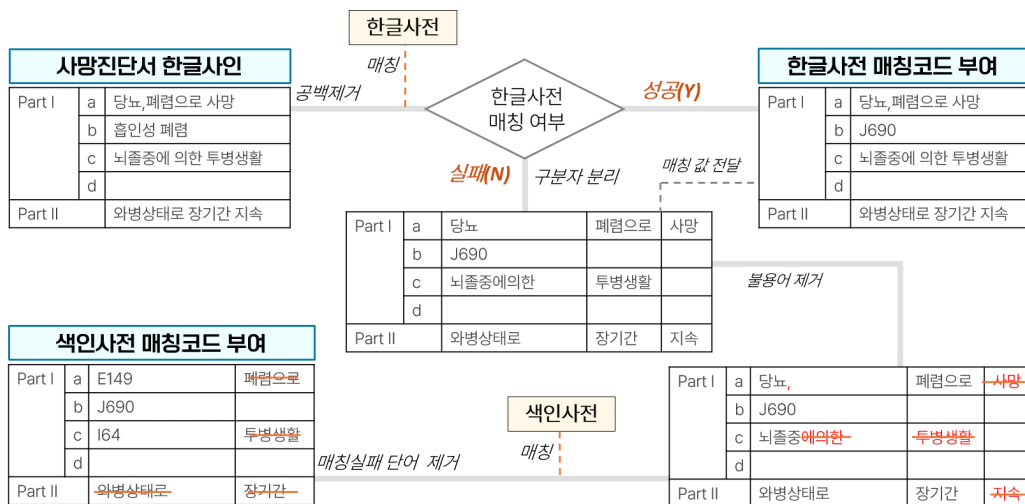
이상의 전환은 사인분류 업무를 단순 분류 작업에서 분석·해석 중심의 고부가가치 업무로 전환시키며, 통계 생산체계 전반의 효율성과 품질을 동시에 향상시키는 요인으

로 작용한다. 본 절에서는 이러한 전환을 구현하기 위한 사망원인선정시스템의 단계별 개선방향을 다음과 같이 제시한다. 각 절차는 본 연구 제4장과 제5장에서 제시한 4종 AI 모델과 연계된다.

1. 사망진단서 사망원인 코드화

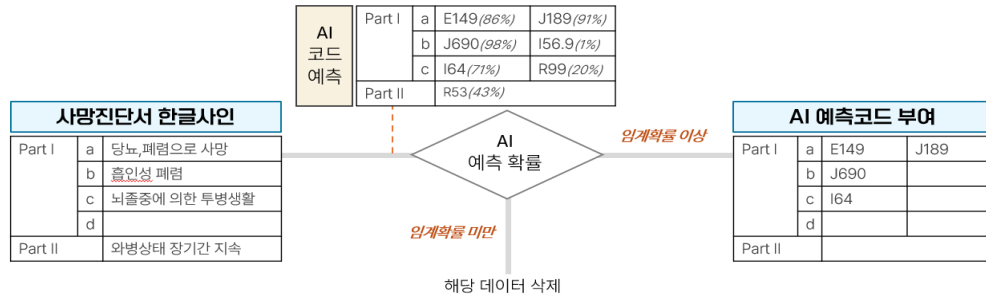
사망원인통계 작성 과정에서 시스템이 가장 먼저 수행하는 단계는 사망진단서에 기재된 사망원인을 한국표준질병사인분류(KCD)로 부호화(coding)하는 과정이다. 현재, 이 과정은 한글사전과 색인사전의 매칭을 통해 이루어지고 있으나, 사전 데이터 관리 부담이 크다. 특히, 한글사전의 경우 현재 9.6만 레코드로 구성되어 있어 기준 변경 또는 신규 용어 추가 시 활용에 한계가 있다. 또한, 자주 나오는 단어에 대한 매칭은 정확하게 이루어지나, 희귀질환 등의 저빈도 사망원인의 경우 코드화가 실패하여 원사인 선정의 정확도를 저하시키는 요인으로 작용한다.

<그림6-3> 현재 시스템 사망진단서 사망원인 코드화 절차



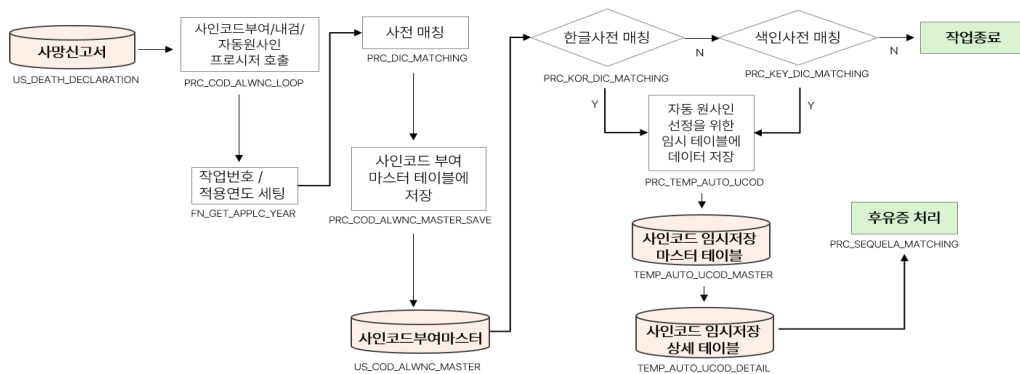
이러한 사망진단서의 사망원인 코드화 단계를 AI 기계학습 기반 예측 방식으로 전환할 경우 업무환경 개선에 기여할 수 있다. 기존 사전 매칭 방식은 기준 변경과 환경 변화에 따라 시간이 지날수록 정확도가 저하되는 반면, AI 기반 예측 방식은 학습데이터의 누적에 따라 정확도가 개선될 여지가 있다.

<그림6-4> 사망진단서 사망원인 코드화 절차 개선(안)



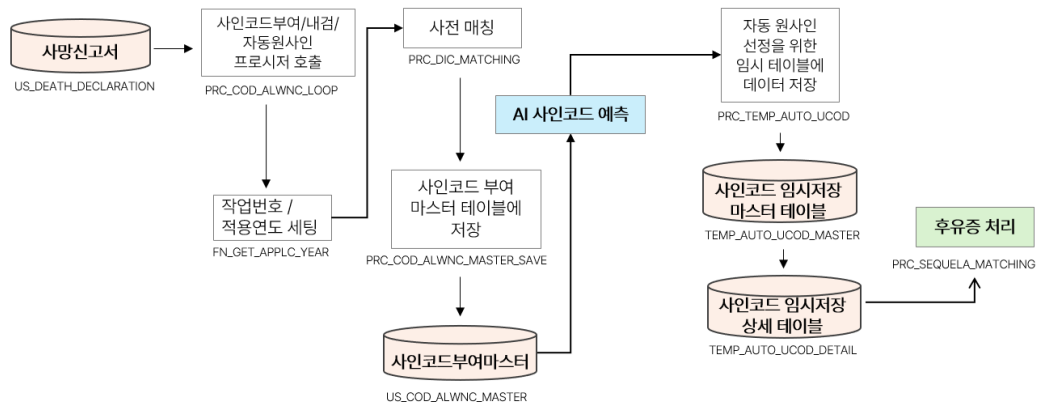
사망진단서 코드화 과정은 실제 통계생산 업무에서 자동 원사인 선정, 행정자료 연계, 사망신고서 내검 등의 과정과 함께 매월 지방자치단체의 사망신고서 제출이 마무리되는 시기에 수행된다. 각 단계는 사인선정시스템의 프로시저 모듈별로 구성되어 있으며, 한 단계가 완료되면 다음 절차가 호출되는 방식으로 작동한다. 사인코드 부여 프로시저는 한글사전 및 색인사전 매칭, 후유증 처리, 합병증 처리 등 3단계로 구성되어 있고, 종료 이후 내검 프로시저로 연결된다. 사망진단서 사망원인 코드화(모델1)에 대한 단위 프로시저는 다음과 같다.

<그림6-5> 현재 시스템 사인코드 부여 프로시저 호출도



AI 기계학습 기반 사망원인 코드화 예측이 도입되면, 단기적으로는 기존 '한글사인 연계' 과정 및 '색인사전 연계' 과정을 대체한다. 이는 5장의 분석결과에서 색인사전을 통한 자동코딩 결과보다 AI 예측 결과가 더 정확했기 때문이다. 사망진단서의 내용을 사인코드 부여할 수 있는 테이블 형태로 전처리한 이후, 각각의 사망원인에 대해 AI가 최적 코드값과 예측확률을 생성한다. 이때 본 연구 결과에 따라 예측확률 50% 이상의 자료만 반영하고, 신뢰도가 낮은 데이터는 반영하지 않는다. 이에 따른 사인선정시스템의 업무 프로세스를 살펴보면 <그림6-5>, <그림6-6>과 같다.

<그림6-6> AI 예측기능 개선 시스템 사인코드 부여 프로시저 호출도(장기)



AI가 매월 사망원인 코드를 예측하는 기능과 함께, 사망원인 코드 수정 기능과 AI 학습자료 추가 기능도 동시에 구현되어야 한다. AI의 코드 분류가 개선되더라도 오류 예측 결과는 일정 수준 존재하며, 그 결과가 정제되지 않은 채 향후 학습자료에 반영될 경우 추가적인 오류를 야기할 수 있다. 따라서 사인분류사가 코드화 결과를 수정할 수 있는 기능의 개발이 필요하다.

또한 AI 학습자료 추가는 향후 모델의 안정성과 정확도 향상을 위해 필수적인 요소로 최소 3년간의 학습자료가 학습되도록 누적되도록 구축되어야 한다. 시점 측면에서는 매년 사망원인통계 확정 이후 1년 간의 사망진단서 코드화 결과에 대한 정기 학습과 긴급히 추가되는 사망원인 코드에 대한 수시 학습이 함께 이루어져야 한다. 그리고 실제 자동코딩 결과에 대한 표본검수 등 품질관리 절차가 지속적으로 이루어져야 한다.

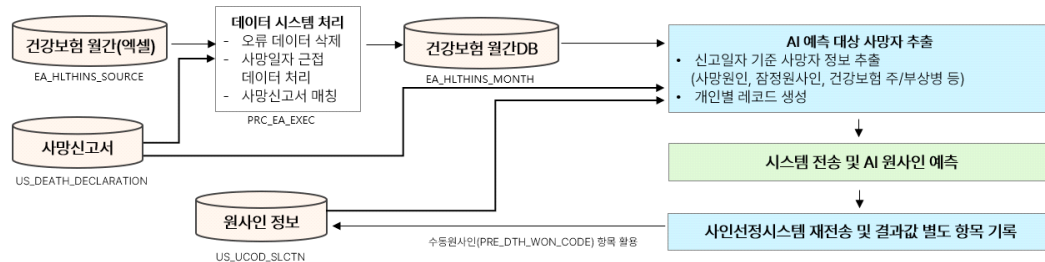
2. 행정자료 반영 원사인 AI 예측

앞서 진행된 사망진단서 사망원인 코드 부여 단계 이후, 사망신고서 내검 → 사인 코드 내검 → 자동원사인 부여 → 자동원사인 내검 → 결과 코드 및 로그 저장 등의 과정을 거치게 된다. 이 단계에서는 사망진단서만으로 시스템에서 WHO의 가이드라인에 따라 선정한 원사인 정보만 존재한다. 이후 단계에서는 사망자의 다수 행정자료를 반영하여 원사인의 정확도와 활용성을 높이는 과정이 이어진다.

본 연구에서 활용한 건강보험 수진기록(월간 자료)은 사망신고 시기와 건강보험 기록의 완성도를 고려하여 사망 발생 월로부터 2개월이 경과한 시점에 요청되며, 그 다음 달에 입수된다. 입수된 건강보험 수진기록은 행정자료 업로드 과정을 거쳐 데이터 정제와 사망신고서와의 연계가 이루어진다. 전처리가 완료된 자료는 가장 최근 자료와 두 번째 최근 자료의 순서로 주상병과 부상병이 정리된다.

월간 건강보험 수진기록 업로드 절차가 완료되면, 사망진단서 사망원인, 사망진단서 자동 원사인, 건강보험 수진기록을 연계하여 'AI통계분류시스템'으로 전송한다. 이때 건강보험 수진기록에서는 가장 최근 자료, 두 번째 최근 자료, 최고 진료비 기록, 최근 최고 진료비 기록에 해당하는 주상병, 부상병, 진단일 합계, 진단액 합계를 횡으로 연계하여 사망자 1인당 단일 레코드로 구성하는 전처리 작업이 필요하다. 또한, 전송하는 데이터는 건강보험 월간 수진기록의 연계여부와 관계없이 월별 사망 신고자료 전체를 작업대상으로 설정하여야 한다. 해당 업무처리 절차는 다음과 같다.

<그림6-7> AI 원사인 예측 업무 처리과정(안)



'행정자료 반영 원사인 AI 예측(모델2)' 과정은 한 달에 한 번씩 수행하게 되며, 행정자료 업로드 프로시저 직후에 실행한다. 원사인 예측은 업무 처리 복잡도와 그래픽 처리장치(GPU) 사양을 고려할 때, 사인선정시스템이 아닌 별도의 'AI통계분류시스템'으로 데이터를 이동하여 처리하는 것이 적절하다.

전송된 사망자별 데이터에 대해서는 제4장 및 제5장의 모델 2를 활용하여 AI 원사인 예측을 수행한다. 이후 사망자별 연계 식별자(key)와 AI 예측 원사인 값만 사인선 정시스템으로 반환한다. 반환된 'AI 예측 원사인' 값은 사망신고서 테이블의 'AI 원사인' 항목에 입력되어 사인분류사의 검토화면에 표출되도록 구성한다. 이때 사인분류사의 업무 효율성 제고를 위해 예측 1~3순위까지 표출되도록 설계한다.

<그림6-8> AI 예측 원사인 시스템 표출 화면(안)

한편, 감염병 등 사망원인 잠정통계 서비스를 위해서는 내부 테이블 데이터에 사인분류사 제공용 전체 코드가 아닌 104개 집계항목 기준의 AI 예측 잠정 원사인 코드가 추가되어야 한다. 테이블에 104개 집계항목 정보가 포함되면 담당자가 직접 DB를 조회하여 잠정 통계수치를 작성할 수 있게 된다. 다만, 감염병의 경우에는 AI 예측 이후 정확도를 개선하기 위하여 다음에 추가 조건을 선택적으로 검토하여 실행할 수 있다. 감염병 사망에 대한 추가조건은 국제질병사인분류 제2권 지침서를 근거로 병태 간의 관계와 국내 사망진단서 작성 실태를 고려하였다.

<표 6-1> 특정 감염성 질환 추가 추출 조건(예시)

병태명	특수조건
호흡기 결핵	① 사망진단서 사망원인 중 '결핵' 포함 + '비결핵' 또는 '후유증' 제외 ② AI예측 값 중 Top-3 중 호흡기 결핵이 포함 ③ 잠정 원사인과 AI 예측 원사인이 악성신생물, 손상/중독이 아님
바이러스 간염	① 사망진단서 사망원인 중 '바이러스 간염' 포함 ② 건강보험 사망직전 주상병이 '바이러스 간염' ③ 잠정 원사인과 AI 예측 원사인이 손상/중독이 아닌 사망자
HIV	① 사망진단서 사망원인 중 '면역결핍' 또는 'HIV' 포함 ② 건강보험 사망직전 주상병이 'HIV' ③ 잠정 원사인과 AI 예측 원사인이 손상/중독이 아닌 사망자

추가적인 보완 방법은 전체 코드 예측값과 104항목 예측값 두 개 데이터를 비교하여 우선순위대로 선별·반영하는 방법도 적용할 수 있다. 이에 대해서는 과소·과대 예측되는 정도를 기준으로 감염병에 대해서는 HIV, 호흡기결핵, 바이러스 간염, 인플루엔자, 코로나19 순으로 반영하고, 10대 질환 중에는 알츠하이머병, 악성신생물, 고의적 자해, 당뇨병, 간질환, 고혈압성질환, 뇌혈관질환, 심장질환, 폐렴, 패혈증 순으로 반영하여 과소·과대 집계 차이를 감소시킬 수 있다.

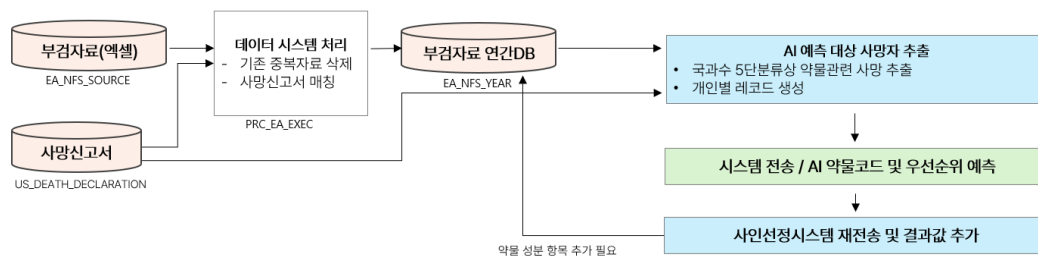
현재 1개년 자료의 학습을 통해 얻은 정확도 및 검증지표 결과는 향후 학습자료의 확충을 통해 더욱 좋은 결과가 도출될 것으로 기대된다. 또한 AI 예측 모델의 기초자료 중 하나인 사망진단서 사망원인 및 잠정 원사인의 경우, 사전매칭 방식에서 모델1의 사망진단서 사망원인 AI 코드화 도입을 통하여 품질이 향상된다면 모델2의 예측값의 품질도 함께 개선될 수 있다.

이번 연구의 단계에서는 건강보험 수진기록 중 주상병, 부상병, 진단종류 항목을 사망직전 시기별로 반영하였는데, 중장기적으로 품질제고를 위하여 건강보험 산정특례 코드를 활용할 수 있다. 산정특례 코드는 희귀·난치성 질환, 결핵, 외상 등 특정 병태에 대하여 좀 더 정확한 정보를 얻을 수 있는 항목이다. 이를 위해서는 추가적인 데이터의 검토 및 분석 등의 후속연구가 필요하다.

3. 약물중독 성분 분류

약물성분의 AI 자동분류의 기초자료는 국립과학수사연구원의 부검자료이다. 부검자료의 경우 사망 기준연도의 다음 해 3월에 입수되며, 연간 사망자료에 대해 약물성분 및 사망수단에 대한 분류가 포함되어 있다. 부검자료 입수 이후 부검자료에 대한 업로드 절차를 진행하게 되는데, 이때 행정자료의 시스템 업로드 후 약물중독 성분 분류를 위한 대상자 추출을 시작한다. 대상자 추출은 해당연도 DB에 업로드된 부검자료만을 대상으로 추출하여 약물과 관련한 사망만 분리하고, 개인별 데이터를 구축하여 '사인선정시스템'에서 'AI통계분류시스템'으로 부검자료를 전송한다. 이때 전송되는 항목은 연계 식별자와 사망에 유효한 약물성분이 기재된 부검 5단 분류 항목이다.

<그림6-9> AI 약물 성분 및 우선순위 예측 업무 처리과정(안)



'AI통계분류시스템'에서는 내부 거대언어모델(LLM)을 활용하여 본 연구 제4장 및 제5장에서 제시한 방법으로 약물 성분 KCD 코드와 우선순위를 생성한다. 이후 사망자별 연계 식별자와 AI 예측 약물 성분 코드 및 우선순위를 사인선정시스템으로 반환한다. 반환된 AI 예측값은 국립과학수사연구원 부검 행정자료 테이블에 항목으로 추가되며, 사인선정시스템 행정자료 검토화면에 표출되어 사인분류사의 검토 업무에 참고자료로 활용된다. 이때 사인선정시스템의 행정자료 검토화면은 약물성분을 표출할 수 있는 공간의 재편성이 필요하다.

이 모델은 학습데이터를 이용한 기계학습 방식이 아닌 거대언어모델에서 시스템 프롬프트와 few-shot 방식을 사용하였기 때문에 향후 추가적인 약물성분이 식별되거나 AI 분류 오류가 나타나는 경우에는 시스템 프롬프트 추가만으로 모델 수정이 가능하다. 따라서 사망원인통계 담당자 및 사인분류사의 지속적인 품질관리가 이루어져야 한다.

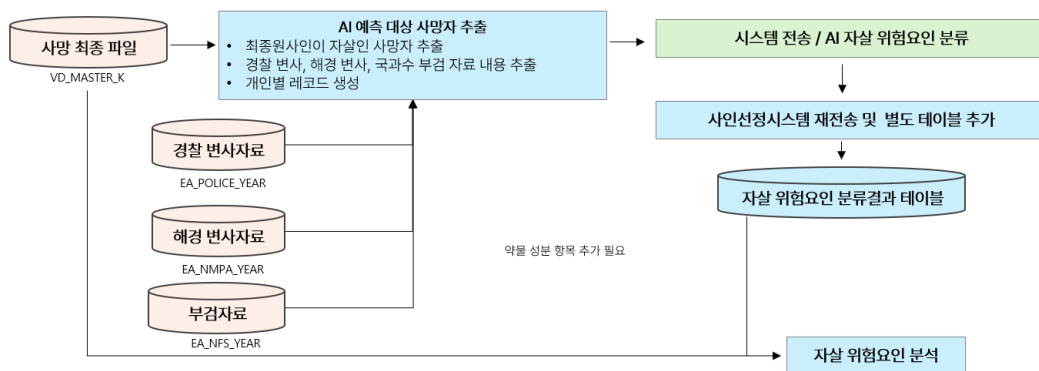
4. 자살 위험요인 분류

자살 사망자에 대한 동향 파악과 관련해서는 현재 매월 잠정 자살 사망자 수가 국가통계포털(KOSIS)을 통해 공표되고 있다. 이와 별도로 본 연구에서 추진한 자살 위험요인별 분류는 신속한 동향 파악이 아니라 자살 사망자의 다각적 분석을 목적으로 하며, 따라서 신속성보다는 정확도와 정보 활용도가 중요하다.

현재 우리나라의 공식적인 자살 사망통계는 다음 해 9월 사망원인통계에 포함되어 공표된다. 따라서 자살 위험요인에 대한 AI 분류는 자살 사망자가 확정되는 다음 해 9월 이후 진행한다. 기초자료는 사망신고서, 경찰청 변사자료, 해양경찰청 변사자료, 국립과학수사연구원 부검자료이다. 이 4종의 자료를 개인별 1개 행으로 데이터를 구성한 후 ‘AI통계분류시스템’으로 전송한다.

‘AI통계분류시스템’에서는 로컬 환경에서 운영되는 거대언어모델(LLM)을 활용하여 자살 위험요인을 개인 단위로 분석하고, 결과를 사인선정시스템으로 반환한다. 이후 반환된 자살 위험요인과 사망원인통계 최종 결과(마스터 파일)의 항목을 연계하여 자살 사망에 대한 특성을 분석하고, 마이크로데이터를 생성·서비스한다.

<그림6-10> AI 자살 위험요인 분류 업무 처리과정(안)



본 연구에서 제안한 자살 위험요인 분류체계는 직업적 요인, 경제적 요인, 가정적 요인, 정신적 요인, 질병·신체적 요인 등의 체계로 구성하였다. 이는 기존 사망원인통계의 분류인 KCD-8(ICD-10)을 기반으로 호주 통계청 분류를 준용하였다. 현 단계에서는 정책적 활용성을 고려하여, 개인별로 2개까지의 복합요인이 추출가능하도록 설계되어 있으나 향후 정책목적에 따라 최대 복합요인 수 및 분류체계의 수정이 가능하다.

제3절 통계서비스 확장과 데이터 기반 정책지원 체계 구축

본 절에서는 제1절과 제2절의 업무체계 개편을 토대로, AI 기반 사망원인통계가 보건정책 수요 대응에 어떻게 기여할 수 있는지를 구체화한다. AI 기반 사망원인 분류체계의 도입은 단순한 업무 효율화에 그치지 않고, 통계서비스의 본질적 확장을 가능하게 한다. 기존 사망원인통계가 연간 단위의 집계 중심 서비스에 머물렀다면, AI 도입 이후에는 시계열적·구조적 분석이 결합된 다차원 통계서비스로 발전할 수 있다.

1. 감염병 사망 감시체계

국내 감염병 사망 감시체계는 시의성, 완결성, 그리고 기관 간 연계 측면에서 구조적 한계를 가진다. 특히 감염병 발생 정보와 사망 정보가 분절적으로 관리됨에 따라, 감염병의 실제 중증도 변화나 치명률을 신속하게 파악하기 어려운 문제가 지속되고 있다.

첫째, 표본감시 중심의 구조로 인해 전체 사망 규모를 대표하기 어렵다. 제4급 감염병의 경우 일부 의료기관에 의존한 자료 수집이 이루어지고 있어 사망 규모 및 변동을 실시간으로 파악하기 어렵고, 실제 사망 양상은 사후적으로만 확인되는 경향이 있다.

둘째, 사망 신고체계가 발생 신고 중심으로 설계되어 있어 사망 정보 누락 가능성이 존재한다. 특히 의료기관 간 전원 과정이나 감염 이후 일정 기간 경과 후 발생한 사망은 체계적으로 포착되지 못하는 한계가 있다.

셋째, 질병관리청의 감염병 감시체계와 국가데이터처의 사망원인통계 간 연계가 미흡하여 통합적 분석이 어렵다. 이로 인해 감염병 발생과 사망을 결합한 정책지표 생산이 제한되고, 통계 활용의 시의성도 저하된다.

넷째, 전자적 신고 및 실시간 모니터링 체계가 미흡하여 정책 대응에 필요한 시의성 있는 정보 제공이 어려운 구조이다. 해외 주요국이 전자사망신고 시스템을 기반으로 단기 시차 내 사망지표를 제공하는 것과 대비되는 부분이다.

이러한 구조적 한계를 극복하기 위해서는, 사망정보를 신속하게 수집·분석하고 정책에 활용할 수 있는 실시간 기반의 통합 감시체계 구축이 필요하다. 본 연구에서 제안한 AI 기반 사망원인 예측 모델은 이러한 전환을 가능하게 하는 기술적 수단으로 기능한다.

특히 감염병 사망 감시체계에 AI를 적용할 경우 다음과 같은 기능적 개선이 기대된다. 월 단위 또는 그 이하 주기의 잠정 사망통계를 생성하여 기존 통계 대비 공표 시

차를 획기적으로 단축할 수 있다. 또한 감염병별 사망 추이를 시계열적으로 모니터링을 통해 유행 확산이나 치명률 변화에 대한 조기 경보 기능을 수행할 수 있다. 실제로 AI 예측 기반 감염병 사망자 추이는 기존 통계와 유사한 흐름을 보이며, 단기 변동을 효과적으로 반영하는 것으로 나타났다.

결과적으로 AI 기반 감염병 사망 감시체계는 기존의 '사후적 통계 생산 구조'를 '적기 감시 및 정책 대응 체계'로 전환시키는 핵심 인프라로 기능한다. 이는 단순한 업무 자동화를 넘어, 감염병 대응 정책의 시의성과 정확성을 동시에 제고하는 데이터 기반 의사결정 체계 구축에 기여할 것으로 판단된다.

특히, 사망에 대한 분류체계를 살펴볼 때, 감염성 질환 등의 병태가 포함되는 회피가능 사망 중 치료가능 사망의 경우 적기에 치료를 받거나 의료적 개입으로 예방이 가능하며, 예방가능 사망의 경우 위험요인 제거 및 건강 정책으로 사망 감소 효과를 거둘 수 있다. 감염병 사망 감시체계에서 주요 질환 또는 회피가능 사망에 대한 모니터링 시스템으로 확장된다면 보건정책에 큰 도움이 될 것으로 기대한다.

2. 자살 위험요인별 사망 분석

기존의 사망원인통계에서 자살 사망에 대한 분석은 성·연령·지역·자살수단에 한정되어 있어, 자살예방 정책을 모니터링하고 수립하는 데 한계가 있었다. 경찰청에서 변사자에 대한 자살원인별 자살자 수를 공표하고 있으나, 분류체계가 표준화되어 있지 않고 자살의 특수성이 충분히 반영되어 있지 않다. 특히, 자살 위험요인에 대해서는 복합적인 요인이 작용하는 것과 분류의 중첩성이 한계점이다.

이와 관련하여 국제적으로 주목할 만한 사례는 호주 통계청(Australian Bureau of Statistics, ABS)의 자살 통계 생산 방식이다. 호주 통계청은 단순히 사망원인을 분류하는 데 그치지 않고, 검시자료(coronial data)와 사망진단서를 활용하여 자살과 관련된 심리 사회적·신체적·정신적 위험요인을 체계적으로 코드화하여 통계로 생산하고 있다.

호주 통계청의 자살 위험요인별 통계는 단일 원인 분석이 아닌 다중 요인 구조를 전제로 한다. 실제 통계에 따르면 자살 사망자의 약 79.7%는 하나 이상의 위험요소가 보고되며, 평균적으로 약 4개의 위험요소가 동시에 존재하는 것으로 분석하였다. 또한 자살동기와 같은 위험요소는 연령·성별·지역 등에 따라 상이하게 나타나기 때문에 사망원인을 단일 사건으로 보는 것이 아니라, 사회경제적·심리적 요인이 결합된 결과로 이해하는 다층적 분석 체계라는 점에서 중요한 의의를 가진다.

본 연구에서 제시한 AI 기반 자살 위험요인 분류 모델은 이러한 호주 통계청 사례와 높은 정합성을 가진다. 본 연구는 행정자료의 서술형 텍스트를 활용하여 자살 위험요인을 직업·경제·가정·정신·신체 질병 등의 영역으로 자동 분류하는 모델을 제안하고 있으며, 이는 기존 통계에서 포착되지 않았던 위험요인 구조를 계량적으로 분석할 수 있는 기반을 제공한다.

특히 본 연구에서는 AI를 통해 자동 추출·분류할 수 있다는 점에서 다음과 같은 확장이 가능하다. 첫째, 텍스트 기반 정보 활용의 확대이다. 자살 관련 행정자료에는 정형화되지 않은 서술 정보가 다수 포함되어 있으나, 기존 통계 생산 과정에서는 충분히 활용되지 못하였다. AI는 이러한 비정형 데이터를 구조화하여 통계로 전환할 수 있다. 둘째, 정책 목표 집단에 대한 정밀도의 향상이다. 위험요인별 자살 통계는 특정 집단(예: 경제적 취약계층, 정신건강 위험군)에 대한 정책 개입을 보다 정밀하게 설계할 수 있도록 지원한다.

이상의 두 영역(감염병 사망 감시 및 자살 위험요인 분석)은 본 연구가 지향하는 '보건정책 수요 대응을 위한 AI 기반 서비스 개발 기반 마련'의 구체적 구현 방향이다.

제 7 장

결 론

본 연구는 사망원인통계 작성의 핵심 단계인 원사인 분류 업무의 정확도와 행정 효율성을 제고하고, 보건정책 수요에 부응하는 통계서비스를 확장하기 위해 인공지능(AI) 기반 분석모형을 개발·평가한 연구이다. 본 장에서는 앞 장들에서 다룬 연구 배경과 방법론, 4종 AI 모델의 분석 결과, 그리고 이를 반영한 업무체계 개편 방안을 종합적으로 정리하고, 정책적 함의와 후속 연구과제를 제시한다.

제1절 연구 요약

사망원인통계는 「통계법」에 따른 지정통계로서 보건의료 정책, 감염병 대응, 산업 재해 및 안전 정책, 사회적 위험요인 분석 등 다양한 정책 분야의 핵심 기초자료이다. 그러나 최근 사망진단서 사인 표현의 복잡·다양화, 고령화에 따른 사망자 수 증가, 세계보건기구(WHO)의 제11차 국제질병사인분류(ICD-11) 도입, 감염병 등 특정 사망원인에 대한 시의성 있는 통계 수요 증가 등으로 사망원인통계 작성 환경은 구조적 변화를 맞이하고 있다. 이러한 변화는 수작업과 사전매칭 중심의 현행 분류 방식만으로는 대응에 한계가 있음을 시사하며, 분류 정확도와 행정 효율성을 동시에 높일 수 있는 업무지원 체계의 도입 필요성을 보여준다.

본 연구에서는 이러한 문제의식에 따라 사망원인통계 작성 전 과정에 AI를 적용하는 4종의 모델을 개발하였다. 각 모델은 업무 단계와 정책 수요를 고려하여 설계되었으며, 사망진단서 사인 코드화(모델1), 행정자료를 반영한 잠정 원사인 예측(모델2), 약물중독 코드 자동분류(모델3), 자살 위험요인 자동분류(모델4)로 구성된다. 모델1과 모델3은 사인분류사의 검토 업무를 지원하는 업무 지원형 모델이며, 모델2와 모델4는 기존 통계에서 포착하지 못했던 세부 정보를 구조화하여 정책 활용도가 높은 심층 통계 생산의 기반을 마련하는 서비스 확장형 모델이다.

제2절 주요 분석 결과

1. 사망진단서 사인 코드화(모델1)

모델1은 KcELECTRA 인코더에 KCD 코드에 대한 멀티라벨 분류 구조를 결합한 모델로, 사망진단서의 자연어 사인명을 KCD 세부코드로 자동 부여한다. 학습데이터 구성만 달리한 조건 A(사망진단서 학습데이터 단독)와 조건 B(사망진단서 학습데이터와 한글사전 결합)를 비교한 결과, 한글사전 매칭 영역에서는 두 조건 모두 사전분류 결과와 높은 수준의 일치도를 보였으나, 한글사전이 매칭되지 않은 색인사전 매칭 영역에서는 상대적으로 낮은 일치율이 나타났다.

이에 대한 표본검수 결과, 색인사전 매칭 영역에서 관찰된 불일치의 상당 부분은 학습데이터(사전 자동분류 코드) 자체의 신뢰도 한계와 학습자료 구성상의 편향에서 비롯된 것으로 확인되었다. 즉, 오분류의 주요 원인은 모델 알고리즘 자체보다 학습데이터 및 원시데이터 오류와 토큰화 과정의 공백 민감성 등에 있는 것으로 분석되었다. 이는 향후 모델1의 성능 개선의 핵심이 예측 모델의 재설계보다 학습자료의 품질개선과 전·후처리 규칙의 정비에 있음을 시사한다. 아울러 ICD-11 자동코딩에 대해서도 방법론적 기반은 확보하였으나, 학습자료가 절대적으로 부족하여 실무 적용이 가능한 수준의 정확도에는 아직 이르지 못하였다.

2. 행정자료 반영 잠정 원사인 예측(모델2)

모델2는 사망진단서 텍스트와 건강보험 수진기록을 결합하여 잠정 원사인을 예측하는 모델로, 한국어 자연어 처리에 특화된 KcELECTRA에 KCD 계층구조를 반영한 그래프 합성곱 신경망(GCN)을 결합한 구조를 적용하였다. 세 개의 후보 모델(CNN_full, GCN_full, GCN_104)을 비교한 결과, 104항목 분류 기준 정확도는 GCN_104가 84.6%, CNN_full이 84.4%, GCN_full이 84.2%로 나타났으며, 세 모델 모두 기존 잠정 원사인 정확도(75.8%)보다 8~9%p 높은 성능을 보였다.

이는 사망진단서 텍스트만을 활용하던 기존 방식보다 사망진단서 자동 코드화 결과와 건강보험 수진기록 등 다양한 정보를 결합한 예측이 실제 최종 원사인에 더욱 근접한 결과를 산출함을 보여준다. 최종 예측모델로는 KCD 코드의 계층구조를 학습에 반영할 수 있고 희소 코드에 대한 예측 성능이 상대적으로 우수한 GCN 계열 모델을 채택하였다.

3. 약물중독 코드 자동분류(모델3)

모델3은 국립과학수사연구원 부검자료의 서술형 텍스트로부터 약물 성분 코드와 우선순위를 추출하는 모델로, 거대언어모델(LLM)에 「약물, 약제 및 생물학적 물질에 의한 중독에 대한 특별지침(M4)」을 준용한 규칙을 결합하여 구현하였다. 2024년 부검자료 연계 사망자 7,423명을 대상으로 분석한 결과, 약물 강제 추출 규칙과 우선순위 적용이 전반적으로 지침에 부합하는 것으로 확인되었다. 다만 약물사전의 유사계열 누락과 파생제 매핑 오류가 반복적으로 관찰되어, 유사계열 단위의 약물사전 정비, 오분류 사례에 대한 자동 재검증 절차의 구축, 개정 이력의 체계적 관리가 필요한 것으로 나타났다.

4. 자살 위험요인 자동분류(모델4)

모델4는 경찰청·해양경찰청·국립과학수사연구원의 변사자료로부터 자살 위험요인을 자동 분류하는 모델로, 심리·사회적 요인, 정신적 요인, 질병·신체적 요인의 3개 대분류와 21개 소분류로 구성된 분류체계를 호주 통계청(ABS) 사례를 준용하여 설계하였다. 또한 근거 추출과 종합판단을 분리한 2단계 파이프라인과 규칙을 적용하여 분류체계 이탈과 중복 부여를 방지하였다.

2024년 자살사망자 14,875명을 대상으로 적용한 결과, 총 17,318건의 위험요인이 분류되었으며(1인당 평균 1.2건), 대분류 기준으로는 심리·사회적 요인 42.2%, 정신적 요인 37.2%, 질병·신체적 요인 15.7%, 미상 4.9%의 분포를 보였다. 특히 경찰청의 포괄적인 '기타·미상' 범주에 포함되던 교육적 요인, 법적 상황 등의 세부 위험요인을 추가로 식별함으로써 기존 통계보다 자살 위험요인에 대한 세분화 분석이 가능하게 되었다. 다만 자살 수단을 위험요인으로 오인하거나 정황 정보만으로 위험요인을 단정하는 사례가 반복적으로 확인되어, 프롬프트와 후처리 로직의 추가적인 보완이 필요한 것으로 나타났다.

제3절 정책적 합의 및 통계서비스 확장 방안

본 연구에서 개발한 4종의 AI 모델은 개별 모델의 성능 향상을 넘어 사망원인통계 생산체계와 통계서비스 전반의 구조적 전환 가능성을 다음과 같이 뒷받침한다.

첫째, 사망원인통계 서비스가 확대된다. AI 도입 이후에는 기존의 연간 확정통계(다음 해 9월)에 더하여 사망 발생 후 2개월 시점의 잠정 사망자 수, 5개월 시점의 잠정 사망원인통계, 그리고 다음 해 12월의 자살 위험요인 연계 데이터를 추가로 제공함으로써 4단계 서비스 체계가 구축 가능하다. 이는 사망원인통계를 사후적 기록 통계에서 정책 대응을 지원하는 시의성 있는 정보체계로 발전시키는 계기가 될 것으로 기대된다.

둘째, 사인분류 업무체계가 개편된다. 모델1의 AI기반 사망진단서 자동 코드화는 기존의 한글사전 및 색인사전 매칭을 대체하고, 모델2의 잠정 원사인 예측은 사인분류사의 최종 판단을 지원하는 도구로 활용될 수 있다. 이러한 업무지원형 체계는 WHO 원사인 선정 기준과 사인분류사의 최종 판단 권한을 유지하면서도 행정 효율성을 높인다는 점에서 통계의 정확성과 신뢰성을 유지하는 업무체계 개편 방안으로 활용될 수 있다.

셋째, 데이터 기반 정책지원 기능이 강화된다. 감염병 사망 감시체계에서는 AI 기반 잠정통계 생산을 통해 유행 확산과 치명률 변화에 대한 조기 파악이 가능해질 것으로 기대된다. 또한 회피가능 사망에 대한 모니터링을 확대할 경우 예방정책 수립의 실효성을 높일 수 있을 것으로 예상된다. 자살 위험요인 분석 역시 기존의 성·연령·지역·자살 수단 중심 분석에서 벗어나 다중 위험요인의 구조적 분석이 가능해짐에 따라 자살예방 정책의 정밀성을 높이는 데 기여할 것으로 기대된다.

제4절 연구의 한계와 향후 과제

본 연구는 사망원인통계 원사인 분류에 AI를 체계적으로 적용한 연구로서 방법론적 기반과 실증적 성과를 확보하고자 하였다. 그러나 실무 적용을 위해서는 학습데이터의 품질 제고와 모델 고도화라는 두 축의 후속 연구가 지속적으로 이루어질 필요가 있다.

본 연구에서는 기초자료 확보의 한계로 인하여 1개년 학습데이터만 사용함에 따라 학습데이터 추가에 따른 지표 검증이 후속과제로 이루어져야 한다. 또한, AI 기반 통계의 성능은 학습데이터의 품질에 크게 좌우되므로, 보다 정교하고 심층적이며 정밀한 통계를 구현하기 위해서는 학습데이터 정제와 그에 따른 효과분석이 수행되어야 한다.

이러한 단계적인 학습데이터 구축·정제와 지속적인 모델 개선이 병행될 때 실제 서비스 시행이 가능해진다. 이러한 관점에서 후속과제를 단기과제와 장기과제로 구분하여 다음과 같이 제시한다.

1. 단기 후속과제

첫째, 학습데이터 품질검사와 정제 효과에 대한 분석이 필요하다. 수작업 검수를 통해 모델 적합도를 재추정하고, 이를 바탕으로 오류 데이터를 정제하여 학습데이터의 품질을 높일 필요가 있다. 이후 학습데이터 개선에 따른 성능 변화를 정량적으로 분석함으로써 향후 자원 투입의 우선순위를 설정하기 위한 근거를 마련한다.

둘째, 자살 위험요인에 대한 심층 분석이 필요하다. 2024년 및 2025년 사망원인통계를 대상으로 자살 위험요인의 분포와 구조를 분석하고, 다중 위험요인의 상호작용에 대한 통계적 분석과 이에 적합한 서비스 형태를 모색한다. 이는 모델4의 실용성을 검증하는 동시에 정책 수요에 부합하는 통계서비스의 구체적인 활용 방안을 마련하는 데 기여할 것으로 기대된다.

2. 장기 후속과제

첫째, AI 적용 시스템 구현을 위한 학습데이터 확충과 모델 개선이 이루어져야 한다. 구체적으로 모델1은 학습데이터의 규모확충과 데이터 정제 등을 실시하고, 모델2는 질병별 특수 조건을 반영하는 한편 건강보험 산정특례 자료 등 추가 행정자료를 연계

하여 예측 정확도를 높일 필요가 있다. 모델3은 실제 통계 생산 과정에 적용하여 개선 사항을 도출하고, 모델4는 환각(hallucination) 현상을 줄이기 위한 프롬프트와 후처리 방식을 지속적으로 개선할 필요가 있다.

둘째, ICD-11 적용을 위한 학습데이터 구축 방안을 마련해야 한다. 본 연구에서는 ICD-11 기반 모델의 방법론적 기반은 확보하였으나, 정답 라벨의 부족으로 실무 적용이 가능한 수준의 정확도에는 이르지 못하였다. 이에 생성형 AI 등을 활용한 ICD-11 학습데이터 구축 및 개선방안을 마련하고, 전문가 검수 결과와 생성형 AI 검수 결과를 비교·분석하여 데이터의 신뢰성을 확보할 필요가 있다. 아울러 학습데이터 개선 효과를 체계적으로 분석하여 ICD-11 도입 시점에 부합하는 준비 로드맵을 수립하여야 한다.

제5절 맺음말

본 연구는 사망원인통계라는 국가통계에 AI를 적용하는 과정에서, AI가 사인분류사의 전문성을 대체하는 것이 아니라 이를 보완·지원하는 도구로 기능할 때 통계의 정확성과 서비스 확장이 동시에 확보될 수 있음을 확인하였다. 4종 AI 모델은 사망진단서 코드화의 정확도 제고, 잠정 사망원인통계의 생산, 약물중독 및 자살 위험요인의 자동 분류에 이르는 전 과정을 포괄하며, 이는 사망원인통계 서비스의 시의성·세밀성·정밀성을 동시에 제고하는 기반이 될 수 있다.

궁극적으로 감염병 사망 감시체계 마련, 자살예방 정책의 정밀화, 회피가능 사망에 대한 정책 개입 등 국민의 생명과 직결되는 정책 영역에서 통계가 실질적 대응 수단으로 기능하기 위해서는, 본 연구에서 제시한 방법론적 기반 위에 지속적인 학습데이터 정제와 모델 고도화가 축적되어야 한다. 본 연구가 그러한 여정의 첫 단계로서, 향후 사망원인통계 뿐만 아니라 AI의 데이터 활용 분야에 기여할 수 있기를 기대한다.

참고문헌

- World Health Organization (WHO). International statistical classification of diseases and related health problems: 10th revision.
- World Health Organization (WHO). International statistical classification of diseases and related health problems: 11th revision.
- 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(2025), 제9차 한국표준질병사인분류
- 국가데이터처(2023), '인공지능 통계분류 자동화 확대 적용 연구', 통계개발원.
- 질병관리청(2025), '국외 사례 고찰을 통한 국내 표본감시 감염병 사망신고 체계 개선 방안 마련', 질병관리청

부 록

1. 약물의 우선순위 분류표(16단계)
2. 자살 위험요인 분류표(3개 대분류, 9개 중분류, 21개 소분류)
3. 용어정의집

1. 약물의 우선순위 분류표(16단계)

우선 순위	코드	소분류명칭	개별약물명
R01a	T40.1	헤로인 (Heroin)	디아세틸모르핀
R01b	T40.3	메타돈 (Methadone)	디피파논, 이소메타돈, 메타돈, 페나독손
R01c	T40.0	아편(Opium)	아편
R01d	T40.2	기타 천연 아편유사제	코데인, 디히드로코데인, 에틸모르핀, 하이드로코돈, 하이드로 모르핀, 모르핀, 옥시코돈, 옥시모르핀, 테바인
R01e	T40.4	기타 합성 아편유사제	알펜타닐, 부프레노르핀, 부토르파놀, 데즈오신, 에토르핀, 펜타닐, 날부 핀, 펜타조신, 페티딘, 레미펜타닐, 서펜타닐, 타펜타돌, 틸리딘, 트라마돌
R01f	T40.6	기타 및 상세 불명의 마약	디페녹실레이트
R02	T41.0	흡입 마취제	데스플루레인, 엔플루레인, 할로테인, 이소플루레인, 세보플루레인, 티오펜탈, 에토미데이트, 메토헥시탈, 에스케타민, 케타민, 프로포폴
R03	T43.0	삼환계 항우울제(TCA)	아미트립틸린, 이미프라민, 마프로틸린, 노르트립틸린
R04	T42.3	바비류레이트 계열	아모바르비탈, 펜토바르비탈, 페노바르비탈, 세코바르비탈
R06	T43.3	페노티아진 항정신병제	클로르프로마진, 퍼페나진, 티오리다진, 트리플루오페라진, 드로페 리돌, 플루펜틱솔, 할로페리돌, 주클로펜틱솔, 아리피프라졸, 클로자 핀, 올란자핀, 쿠에티아핀, 리스페리돈, 솔피리드, 지프라시돈
R07	T42.0	히단토인 유도체	에토토인, 페니토인, 카르바마제핀, 옥스카르바제핀, 클로랄히드레이트, 에스 조피클론, 가바펜틴, 라모트리진, 프레가발린, 토피라메이트, 발프로산, 졸피 뎀, 조피클론, 아만타딘, 브로모크립틴, 레보도파, 로피나롤, 셀레길린
R08	T40.5	코카인 (Cocaine)	코카인
R09	T43.6	정신흥분제	메트암페타민, 메틸페니데이트, MDMA(엑스터시), 디펜히드라 민
R10	T43.1	MAOI	이소카복사지드, 모클로베미드, 페넨진, 부프로피온, 돌록세틴, 에스 시탈로프람, 플루옥세틴, 미르타자핀, 파록세틴, 세르트랄린, 벤라팍 신
R11	T42.4	벤조디아제핀 계열	알프라졸람, 브로마제팜, 클로르디아제폭시드, 클로바잠, 클로나제팜, 클 로티아제팜, 디아제팜, 에틸로플라제페이트, 에티졸람, 플루니트라제팜, 플루라제팜, 로라제팜, 메사졸람, 미다졸람, 쿠아제팜, 레미마졸람, 트리아 졸람
R12	T40.7	대마 (Cannabis)	칸나비디올, 대마초, 해시시, 마리화나, THC, 리세르산 디에틸아미드 (LSD), DMT, 메스칼린, 펜시클리딘(PCP), 프실로시빈, 합성 칸나비노이 드, 벤조카인, 부피바카인, 리도카인, 프로카인, 로피바카인, 테트라카인

2. 자살 위험요인 분류표(3개 대분류, 9개 중분류, 21개 소분류)

대분류	중분류	소분류	해당내용	기록표현 예시
1. 심리·사회적 요인 (Z55-Z65)	1. 교육적 요인 (Z55)	1. 교육 및 문해에 관련된 문제	실패한 시험, 학교 성취미달, 학교불화, 온라인 언어폭력, 사진영상유포, 신상공개, 집단따돌림 등	'학급, 단톡방에서 지속적 욕설 및 따돌림을 당함', '시험성적 비관', '입시실패후 재수 중에 가족의 압박'
	2. 직업적 요인 (Z56-Z57)	1. 취업 및 실업에 관련된 문제	실직, 소득상실, 해고, 계약종료, 사업실패, 경영악화	'수입이 끊김', '직장을 잃은 후 생활고', '폐업 이후 어려움'
		2. 위험요인에 직업적 노출	직장 환경(소음, 온도 등), 직장 관계(동료, 상사 등)	'직장 내 왕따', '지속적인 폭언', '직무 스트레스 호소'
		9. 기타 및 미상	-	-
	3. 경제적 요인 (Z59)	1. 경제, 주거 상황에 관련된 문제	노숙, 극도의 빈곤 저소득 등 주택, 음식, 빈곤문제	'채무로 고민', '생계곤란', '경제적 사정', '보증채무로 재산압류', '온라인도박에 따른 채무'
	4. 사회적 환경 요인(Z60)	1. 사회적 환경에 관련된 문제	무연고 독거, 문화적응 어려움, 사회적소외, 차별 및 박해, 은둔형 외톨이, 복지 사각지대, 다문화 차별 등	'이웃과 왕래없음', '사망추정일로부터 3주후 발견', '무연고자로 등록', '가족과 연락두절', '다문화가정, 언어장벽 으로 지역사회고립'
	5. 가정적 요인 (Z63)	1. 가족 간 관계 문제	-	'도박 문제로 가족 갈등'
		2. 가족구성원 부재 (사망, 이혼 등)	사별, 이혼 등의 부재에 대한 내용	'배우자 사망 후 우울', '이혼 이후 심리적 불안'
		9. 기타 및 미상	-	-
	6. 법적상황 (Z65.0-4)	1. 법적상황	민·형사소송, 양육소송	'소송 중 스트레스'
	7. 기타 심리· 사회적 요인 (Z58,Z61,Z62)	1. 기타 심리·사회적 요인	물리적환경, 소아기 부정사건, 양육 문제, 재난피해경험 등	'○○○사고 생존자로서 가까운사람의 죽음목격'

대분류	중분류	소분류	해당내용	기록표현 예시
2. 정신적 요인 (F00-F99)	1. 정신행동장애 요인 (F00-F99)	1. 기분 정동장애 (우울증 등)	조증, 양극성정동장애, 우울증 등	'우울증으로 정신과 치료 중', '항우울제 복용'
		2. 스트레스 연관 장애	공황장애, 불안장애, 강박장애, 극심한 불안, 공황발작 등	'공황장애 치료 이력', '극심한 스트레스 상태'
		3. 급/만성 정신성 물질 남용	졸피뎀, 항정신성 의약품, 대마(카바노이드), 수면제 등	'수면제를 과다 복용', '차량 없이 약을 구해 복용'
		4. 급/만성 알코올 남용 장애	음주	'지속적인 음주', '과도한 음주'
		9. 기타 정신적 요인	병적도박	'수년간 온라인 불법도박', '스포츠 도박앱 다수 설치확인'
3. 질병· 신체적 요인 (A00-R99)	1. 질병·신체적 요인	1. 만성질환	암, 심장질환, 당뇨, 만성 폐질환, 신장질환 등	'지병을 앓아옴', '오랜 투병 생활'
		2. 말기·불치 질환	회복 가능성 거의 없음, 예후 불량 등	'항암 말기', '말기 암 판정'
		3. 만성통증	허리 통증, 신경통, 관절통 등	극심한 통증을 호소', '진통제 복용에도 통증 지속'
		4. 신체기능저하/장애	마비, 시각·청각 장애, 거동 불편 등	'거동이 불편함', '장애로 인해 일상생활 곤란'
		9. 기타 질병·신체적 요인		

3. 용어정의집

○ 핵심개념

용어	정의
사망원인통계	「통계법」에 근거한 국가승인통계로서, 사망신고와 사망진단서를 기반으로 국민의 사망 수준과 사망원인 구조를 파악하기 위해 작성되는 전수 통계
원사인 (Underlying Cause of Death)	직접적으로 사망에 이르게 한 일련의 사건을 야기한 질병 또는 손상, 혹은 치명적 손상을 유발한 사고나 폭력의 상황을 의미하며, 사망원인통계의 기본 집계 단위
잠정 원사인	사망진단서에 기재된 정보만을 활용하여 WHO의 ICD-10 제2권 지침에 따라 도출한 개인별 사망의 근원 요인이며, 행정자료 보완 이전 단계의 원사인
최종 원사인	잠정 원사인에 건강보험 수진기록·경찰 변사자료·부검자료 등 약 20종의 행정자료를 연계·검토하여 사인분류사가 확정한 원사인
ICD (국제질병사인분류)	세계보건기구(WHO)가 질병과 사망원인을 국제적으로 표준화하여 비교 가능성을 확보하기 위해 제정한 분류체계로, 원사인 선정 규칙을 포함함
KCD (한국표준질병사인분류)	ICD를 기반으로 국내 실정에 맞게 조정·고시한 국내 표준 분류체계, 2026년 1월부터 KCD 제9차 개정안(ICD-10 기반)이 시행 중이며 ICD-11이 반영된 제10차 개정안은 2027년 7월 고시를 목표로 개발 중
한글사전	KCD를 기반으로 사망진단서에 자주 기재된 사망원인 표현과 해당 코드를 축적한 매칭 테이블로, 사인 코드화 단계에서 가장 먼저 대조되는 사전
색인사전	한글사전 매칭이 실패한 경우 심표·조사 등 구분자로 분리한 색인어를 대조하기 위해 사용하는 보조 매칭 사전
자살 위험요인	자살 사망에 기여한 심리·사회적 요인, 정신·행동장애 요인, 질병·신체적 요인 등을 포괄하는 다중 요인 개념으로, 본 연구에서는 호주통계청 분류체계와 ICD-10을 준용하여 3개 대분류·21개 소분류체계로 구축함, 경찰청 원자료의 '자살동기'와는 구분됨
사인분류사	사인선정시스템에서 자동 선정된 잠정 원사인에 대해 행정자료를 연계·검토하여 최종 원사인을 확정하는 전문 인력
외인사 (外因死)	KCD 정식 명칭은 '질병이환 및 사망의 외인(External causes of morbidity and mortality)'으로, 병적 요인이 아닌 외부 요인에 의한 사망을 의미하며 의도성에 따라 사고·고의적 자해·타살·의도미확인으로 구분됨

용어	정의
R코드	KCD 대분류 R에 해당하는 '달리 분류되지 않은 증상, 징후 및 임상·검사이상 소견'을 지칭하며, 원사인이 불명확한 사망의 분류에 사용됨
104항목	사망원인통계 제표 기준으로 활용되는 사망원인 집계 항목 체계로, 본 연구 모델2의 잠정치 예측 서비스 평가 기준으로 사용됨
건강보험 수진기록	국민건강보험공단이 관리하는 진료 이력 자료로, 국가데이터처는 이를 입수하여 원사인 보완에 활용하며 사망자 연계율은 약 99.2%

○ AI 및 모델 관련

용어	정의
자연어 처리 (NLP, Natural Language Processing)	사람이 사용하는 언어를 컴퓨터가 이해·분석·생성하도록 하는 기술 영역으로, 본 연구에서는 사망진단서의 자연어 사인명 처리와 서술형 행정자료의 정보 추출에 활용됨
사전학습 언어모델	방대한 텍스트 데이터로 사전에 언어 표현을 학습해 둔 뒤, 개별 과제에 맞춰 추가 학습(파인튜닝)하여 활용하는 언어모델
KcELECTRA	온라인 뉴스 댓글 데이터로 학습된 한국어 사전학습 언어모델로, 구어체·오탈자·짧은 텍스트에 강점을 지녀 사망진단서 사인명처럼 짧고 비정형적인 텍스트의 분류에 적합함. RTD(Replaced Token Detection) 방식으로 학습되어 짧은 입력에서도 학습 효율이 높음
GCN (Graph Convolutional Network, 그래프 합성곱 신경망)	데이터의 노드와 엣지로 구성된 그래프 구조로부터 이웃 정보를 전파·집약하여 표현을 학습하는 신경망. 본 연구 모델2에서는 KCD 6,443개 코드를 노드로, 코드 간 의학적 근접성을 엣지로 구성하여 계층·인접관계를 반영하는 코드 인코더로 활용함
LLM (Large Language Model, 거대언어모델)	대규모 파라미터와 방대한 텍스트로 학습되어 별도 학습 없이도 프롬프트만으로 다양한 언어 과제를 수행할 수 있는 언어모델. 본 연구 모델3·4에서는 서술형 자료로부터의 정보 추출에 활용됨
시스템 프롬프트	LLM에 부여하는 업무지침에 해당하는 입력으로, 역할·규칙·출력형식 등을 명시하여 모델의 응답을 제약함
환각 (hallucination)	LLM이 사실이 아닌 내용이나 근거가 없는 답을 생성하는 현상으로, 본 연구에서는 폐쇄 라벨집합·규칙엔진 결합·근거 추출 강제 등의 장치로 이를 차단함
멀티라벨 분류	하나의 입력에 여러 개의 정답 라벨이 부여될 수 있는 분류 방식으로, 본 연구 모델1에서 한 줄에 여러 사망원인이 기재된 경우의 코드 부여에 활용됨
임계값 (threshold)	멀티라벨 분류에서 각 라벨을 최종 예측으로 채택할지를 결정하는 확률 기준값

용어	정의
Long-Tail (희귀라벨 편중 구조)	학습데이터에서 극소수 사례만 존재하는 라벨이 다수를 차지하는 분포 구조로, 본 연구 모델2에서는 KCD 세부코드 6,443개 중 약 48%가 학습데이터에 1건만 등장하여 이러한 구조에 대한 보조 학습기법이 요구됨
계층형 임베딩 (롤업 임베딩)	KCD 코드를 대분류(1자리) · 중분류(3자리) · 세부코드의 세 단계로 분해하여 각각 벡터로 표현한 뒤 결합하는 방식으로, 분류체계의 계층 구조를 모델 학습에 반영함
손상/중독 코드 (약물중독)	KCD의 T40~T50 계열 코드로, 본 연구 모델3에서는 T40 · T41 · T42 · T43 · T45 · T48 · T50을 약물중독 사망 분류 대상으로 삼음
우선순위 코드(16단계)	「약물, 약제 및 생물학적 물질에 의한 중독에 대한 특별 지침(M4)」을 준용하여 복수 약물 검출 시 사망 기여도가 높은 순으로 부여하는 우선순위 체계로, 통계표 작성 시 가장 높은 순위 약물을 기준으로 집계함

○ 성능 평가 지표

용어	정의
일치율 (Accuracy)	전체 예측 건수 중 정답과 일치한 건수의 비율
정밀도 (Precision)	AI가 정답으로 예측한 값 중 실제 정답의 비율
재현율 (Recall)	실제 정답 중 AI가 정답으로 맞춘 비율
F1-score	정밀도와 재현율의 조화평균. 코드별 F1-score에 빈도를 가중한 평균은 Weighted F1, 단순 산술평균은 Macro F1, Weighted F1은 실제 운용성능 평가에, Macro F1은 희귀코드 성능 평가에 활용됨
사망자 수 추정비	실제 사망자 수 대비 AI가 예측한 사망자 수의 비율로 1에 가까울수록 예측 규모가 실제와 근접함
ordered accuracy	멀티라벨 분류에서 정답 라벨 집합과 예측 라벨 집합이 구성과 순서 모두 일치한 비율로, 본 연구 모델1의 평가지표 중 가장 엄격한 기준
subset accuracy	멀티라벨 분류에서 정답과 예측의 집합 자체가 일치하는지를 판단하되, 순서는 고려하지 않는 지표

○ 시스템, 인프라

용어	정의
AI통계분류 시스템	사인선정시스템과 별도로 운영되는 AI 예측 전용 시스템으로, GPU 사양과 처리 복잡도를 고려하여 데이터를 이관받아 AI 예측을 수행하고 결과를 사인선정시스템으로 반환하는 구조
사망원인 선정시스템	사망신고서 · 사망진단서 자료를 입수하여 사인 코드화, 잠정 원사인 선정, 행정자료 연계, 사인분류사 검토 등의 업무 전 과정이 이루어지는 통합 업무시스템

Abstract

A Study on AI-based Classification of Underlying Causes of Death for Causes of Death Statistics

Seokmin Lee, Goseol Jeon, Jiwon Choi, Kyojoong Oh

This research empirically verified the applicability of AI to the classification of underlying causes of death when compiling Causes of Death Statistics, by developing four models. Based on these findings, the study proposed directions for upgrading business processes of the classification of causes of death, and strategies for expanding the dissemination of Causes of Death Statistics.

Causes of Death Statistics, which are designed to identify the death levels and structures based on information from death certificates, are compiled in accordance with the International Classification of Diseases (ICD). However, the environment for producing these statistics faces structural changes, due to factors such as an increasing number of deaths caused by population aging, the 11th revision of the ICD (ICD-11) adopted by the World Health Organization and a rising demand for the timely releases of statistics on specific causes of death including infectious diseases. Therefore, this research was carried out to enhance the accuracy and efficiency of the classification of underlying causes of death and to lay the foundation for AI-driven statistical services to meet health policy demands.

Four AI models were developed for application throughout the production and dissemination of Causes of Death Statistics. Model 1 automatically assigns Korean Standard Classification of Diseases (KCD) codes to information on causes of death in death certificates. Model 2 predicts the preliminary underlying causes of death by integrating death certificate data with health insurance medical records. Model 3 automatically extracts and classifies drug substances causing deaths from drug poisoning. Model 4 automatically classifies suicide risk factors based on text-type information in administrative data. Models 1 and 3 are designed to support the business of the classification staff members in practice. On the other hand, Models 2 and 4 facilitate statistical service expansion by structuring detailed information.

The verification results of the models' performance and applicability revealed that the primary errors of Model 1 (Death certificate coding model) generally

resulted from the quality of training data rather than the model's architecture itself. The methodological foundation for ICD-11 coding has been established, but labeled data has not been sufficiently accumulated for practical implementation.

The accuracy of Model 2 (Prediction model for preliminary underlying causes of death) recorded 84.6% based on 104 item classification, which was higher by 8.8%p compared to the accuracy of preliminary underlying causes of death derived from existing death certificate coding (75.8%).

Model 3 (Automated classification model for drug poisoning code) applied rules for extracting drug substances and determining priority to 7,423 deaths linked to autopsy records from the National Forensic Service (NFS). The analysis results showed a need for a supplementation task, particularly, in the mapping of some drugs within similar classes.

The application results of Model 4 (Classification model for suicide risk factors) revealed that the risk factors identified among suicide deaths in 2024 were concentrated in mental and behavioral disorders (37.2%) and economic factors (25.6%). In addition, Model 4 enhanced the quality of risk factor statistics by identifying detailed factors that were difficult to capture using existing data alone.

In summary, this research demonstrated the potential for a structural transformation of the production system of Causes of Death Statistics. By shifting from the traditional matching method relying on Korean and index dictionaries to AI-based coding system, business efficiency can be strengthened without downgrading statistical quality while assigning final decision-making authority to the classification staff members.

Furthermore, statistical services can be expanded by progressively introducing preliminary cause-of-death statistics and suicide risk factor statistics alongside the existing annual final statistics, which are currently released in September of the year following death. Besides, it is expected that these AI-enabled models will strengthen data-driven policy implementations, through measures including the establishment of a monitoring system for infectious disease-related deaths and the analysis of suicide risk factors based on multiple risk factors, and contribute to enhancing the responsiveness and precision of health policies.

Key words: Causes of Death Statistics, Artificial Intelligence, classification of underlying causes of death, suicide risk factor, drug poisoning deaths, ICD-11

연구진

- 이석민(국가데이터처 국가데이터연구원 데이터과학연구팀 사무관)
- 전고설(국가데이터처 국가데이터연구원 데이터과학연구팀 주무관)
- 최지원(국가데이터처 국가데이터연구원 데이터과학연구팀 주무관)
- 오교중(아일리스프런티어(주) 이사)

* 연구진의 소속 및 직급은 연구과제 완료 시 기준임을 알려드립니다.

연구보고서 2026-02

사망원인통계 원사인 분류 시 적용 연구

인 쇄	2026년 9월
발 행	2026년 9월
발 행 인	백지선
발 행 처	국가데이터처 국가데이터연구원 35220 대전광역시 서구 한밭대로 713 Tel.(042)366-7100 Fax.(042)366-7123
홈페이지	https://mods.go.kr/dsri/
ISSN(Online)	2733-4120



국가데이터처
국가데이터연구원