DSRI
Data
Research
Brief

Vol. 2026-3
ISSN 2765-1002(Online)

데이터 리서치 브리프



◈ 국가통계 개념의 체계적 관리를 위한 GSIM 활용 방안 ····· 1
◈ 사망원인통계 원사인 분류 AI 적용 연구 ····· 7
◈ 2025년 기준 경제총조사 조사표(도·소매업) 개선 연구 ····· 12

OPEN 공공누리 출처표시 변경금지 공공저작물 자유이용허락

국가데이터처
국가데이터연구원

* 표지 이미지 생성형 AI 활용

DSRI
Data
Research
Brief

vol. 2026-3

ISSN 2765-1002(Online)



국가통계 개념의 체계적 관리를 위한 GSIM 활용 방안

김능경
경제동향통계심의관
물가동향과

배미경
국가데이터연구원
데이터방법연구실

- **(연구 목적)** 통계 간 개념의 공통점과 차이를 식별·비교할 수 있는 일관된 통계 정보 제공을 위해 UNECE 승인 통계정보 참조모델인 **GSIM(Generic Statistical Information Model) 2.0**과 캐나다·노르웨이의 적용 사례를 분석하여 **국가통계 개념정보 관리체계 방안**을 제시하고자 한다.
- **(핵심 결과)** GSIM은 **설계와 구현을 분리**함으로써 재사용성과 일관성을 확보한다. 캐나다는 **표준 정책 기반**의 중앙 관리, 노르웨이는 **기능별 시스템** 연계 방식이며, 양국 모두 개념변수에 판정기준, 도출방법 등 구체적인 사항을 포함하고 있다.
- **(제안 사항)** **Unit Type**(단위유형), **Variable**(변수), **Code List**(코드목록)를 주요 관리대상으로 설정하고 표준 변수와 통계 프로그램별 변수로 **변수 관리를 이원화**하되 통계 프로그램별 정보 구축 후 공통성이 높은 정보부터 표준화하는 단계적 접근을 제안한다.
- **(향후 활용)** 향후 다양한 통계로 **실증 범위를 확대**하여 실무 적용을 위한 최적안을 마련하고 표준의 제정·개정·승인 등 **표준 관리 절차를 확립**함으로써 통계 간 비교 가능성과 재사용성을 획기적으로 향상시키는 기반이 되기를 기대한다.

연구배경 및 목적

❍ **(연구 배경)** 같은 명칭의 개념이라도 조사대상, 측정 단위·방법, 범주 구성에 따라 의미가 다르고, 유사한 개념이 서로 다른 명칭으로 제공되기도 함(예: 노동시장 규모 관련 통계표의 '취업자', '종사자', '일자리')

❍ **(문제 인식)** 통계설명자료서비스의 용어 조회는 상세 작성 기준 없이 기관별로 작성되어 통계 간 비교에 한계가 있으며, 통계용어사전은 단위유형 등 정보의 구조화와 관련 통계와의 연계가 부족함

❍ **(연구 필요성)** 이용자의 통계 선별·해석, 중복 생산 방지와 통계 간 연계는 물론 생성형 AI의 기계 가독성 측면에서도 체계적인 통계 개념 정보가 필요함

❍ **(연구 목표)** GSIM과 이를 도입한 해외 사례를 분석하여 국가통계 전반에 적용할 수 있는 개념 정의·기술 체계의 공통 기준을 제시하고자 함

☞ **(연구 질문) 국가통계 개념정보를 어떤 대상과 구조로, 어떤 절차로 관리해야 통계 간 비교가능성과 재사용성을 높일 수 있는가?**

데이터 및 방법

❍ **(국제표준 검토)** GSIM 2.0에 대한 다양한 문서(클래스 정의, 그룹 다이어그램, User Guide, Linking GSBPM and GSIM, DDI-CDI Specification, ISO/IEC 11179-1 등

❍ **(해외 사례)** 캐나다 통계청과 노르웨이 통계청의 표준관리 정책, 메타데이터 조회 시스템을 통한 개념정보 관리체계와 정보 클래스별 관리 현황 비교

❍ **(국내 시범 적용)** 통계정책관리시스템, 통계설명자료서비스, 국가통계포털 자료로 현황을 진단하고, 노동 분야를 중심으로 GSIM 정보 클래스를 시범 적용

주요 결과

❍ **GSIM의 구조: 5개 그룹과 변수 연쇄(Variable Cascade)**

- GSIM은 통계 생산 전 과정을 포괄하며, 역할에 따라 정보 클래스를 5개 그룹으로 구분하여 기초 그룹은 기술적 정보, 비즈니스 그룹은 업무 프로세스, 교환 그룹은 수집·배포, 구조 그룹은 데이터·메타데이터 구조화, 개념 그룹은 데이터 의미·표현에 대한 정보를 담고 있음

<표1> GSIM 그룹별 정보 클래스

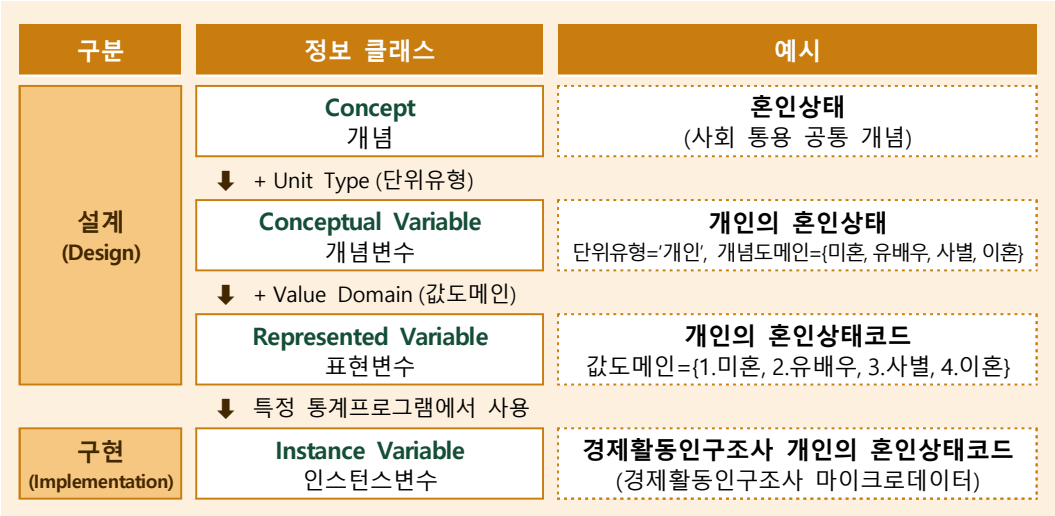
그룹	역할	주요 정보 클래스	개수*
기초(Base)	식별, 버전 관리 등	주체, 역할, 이력, 식별가능 아티팩트 등	12개
비즈니스(Business)	통계 프로그램 설계, 수행	통계 프로그램, 업무 프로세스 등	32개
교환(Exchange)	데이터 수집 및 배포	조사표, 산출물, 정보 이용자 등	21개
구조(Structures)	데이터 구성 방법	데이터세트 구조, 구성요소 등	23개
개념(Concepts)	데이터의 의미 및 표현	개념, 변수, 단위유형, 분류체계 등	39개

* Sentinel Value Domain 등 User Guide에서 설명하지 않는 일부 정보 클래스를 모두 포함한 개수임
출처: UNECE, GSIM Version v2.0 (December 2023), https://unece.github.io/GSIM-2.0

- 개념 그룹의 Concept는 ❶ 측정할 특성(개념변수), ❷ 대상이 되는 집합(단위유형·유니버스·모집단), ❸ 외연적 정의를 위한 범주(통계분류·코드목록)의 세 가지 역할을 수행함

- 추상적 개념이 개념변수(개념+단위유형), 표현변수(개념변수+값도메인)를 거쳐 개별 인스턴스변수로 구체화되는 변수 연쇄(Variable Cascade) 구조를 통해 설계와 구현을 분리함으로써 공통 정보의 재사용성과 조사별 차별성을 동시에 확보함

[그림1] 변수 연쇄(Variable Cascade) 구조와 적용 예시



출처: UNECE(2024), Generic Statistical Information Model (GSIM): Version 2.0 User Guide

주요 결과 (계속)

❍ 해외 사례: 중앙 표준 관리(캐나다)와 기능별 시스템(노르웨이)

- 두 기관 모두 GSIM을 활용하여 개념, 통계단위, 변수, 분류 및 코드목록을 체계적으로 관리하나, 캐나다는 표준의 제정과 재사용에, 노르웨이는 기능별 시스템 구현 및 연계에 중점을 둠

- (표준 의무 수준) 캐나다는 표준을 '부처 표준', '권고 표준', '프로그램별 표준'으로 구분하여 표준화의 강제력과 유연성을 함께 확보함

- (코드목록 제공 방식) 노르웨이는 전체 코드 구조와 버전·대응표·변형을 구조적으로 제공하지만 범주 정의는 없고, 캐나다는 범주 정의와 용법을 제공하지만 전체 구조 조회에는 제약이 있어 상호 보완적 고려가 필요함

- (GSIM 현지화) 두 기관 모두 변수의 개념도메인(Conceptual Domain) 정보는 제공하지 않으며, 캐나다는 하나의 표준 변수에 여러 분류를 연계하는 등 GSIM을 실제 통계작성 환경에 맞게 변형하여 적용하고 있음

<표2> 캐나다·노르웨이 통계 개념정보 관리체계 비교

구분	캐나다 통계청	노르웨이 통계청
관계 체계	표준정책 및 중앙 메타데이터 저장소 중심	기능별 메타데이터 시스템 구조
시스템	IMDB(Integrated Metadata Base)	Datadok, Vardok, Klass 등
거버넌스	표준부서, 방법·표준 위원회, 프로그램 담당자의 역할 분담 명확	변수 정보는 통계 프로그램 담당자가 입력
Unit Type	Statistical Unit 33개 - 맥락이 부여된 Universe까지 포괄	Unit Type 31개 - GSIM의 Unit Type 개념을 충실히 적용
Variable	표준 변수 162개 - 표준 변수와 통계 프로그램별 변수 별도 관리 - Classification의 범주는 불포함하기도 함	Conceptual Variable 1,193개 - 표준 여부 미구분, 범주 여부 무관 - 통계표 조회 시스템과 연계
Node Set (통계분류/ 코드목록)	Classification 226개 - 통계분류와 코드목록 미구분 - 범주 정의와 사용법 등을 제공	Statistical Classification 151개, Codelist 596개 - 통계분류와 코드목록 엄격히 구분 - 코드구조·버전·변경이력·대응표·변형 등 제공

❍ 해외 사례: 변수 정의에 구체적 내용까지 포함

- (취업자) GSIM상 개념변수는 추상성을 유지할 것으로 예상했으나, 양국 모두 '취업자' 변수에 수입 목적 근로, 무급가족종사, 질병·휴가 등에 따른 일시적 결근 등 판정기준을 구체적으로 명시함

* 캐나다는 정의(Definition)와 도출방법(Derivation)을 분리하여 세부 기준을 기술하고, 노르웨이는 '주 1시간 이상 소득창출활동', '3개월 이내 일시해고' 같은 수치 기준을 정의문에 직접 포함함

- (혼인상태) 캐나다는 사실혼 포함 여부에 따라 변수를 둘로 나누어 각각 표준으로 승인한 반면, 노르웨이는 법률혼 기준의 단일 변수에 2009년 동성혼 법제화 이전의 관계 유형 범주를 그대로 유지함

* 같은 상위 개념이라도 변수와 범주 구성은 이용 목적과 제도적 맥락에 따라 달라질 수 있음

주요 결과 (계속)

❍ 국내 시범 적용: 노동 분야

- SIM 정보 클래스를 국내 노동 분야에 시범 적용한 결과는 <표3>와 같음
- 개념은 추상적으로 정의되나, 개념변수는 단위유형을 특정하고 개념도메인 1개를 참조하여 구체화되며, GSIM 최상위 정보 클래스인 Identifiable Artefact의 Description 속성을 활용하여 조작적 정의에 해당하는 구체적인 설명을 추가하였음
- 표현변수는 개념도메인에서 표현방식이 추가된 값도메인을 참조하며, 유니버스는 정보 클래스의 자기참조 관계를 이용하여 조사 및 변수 수준으로 정의할 수 있음

<표3> 노동 분야 GSIM 정보 클래스 시범 적용 결과 (일부)

정보 클래스		사례1	사례2	사례3
주제		노동		
개념	명칭	Employment		
	정의	임금 또는 수익을 위해 재화를 생산하거나 서비스를 제공하는 활동에 참여		
개념변수	단위유형	개인	사업체	
	명칭	개인의 취업여부	사업체의 종사자 수	
	개념도메인	{취업, 미취업}	인원수	
	설명	수입목적1시간이상근로 또는 18시간이상 무급가족일 또는 일시휴직	소속외근로자(파견, 용역), 3개월 이상 장기 결근자 제외	자영업자, 무급가족종사자, 소속외근로자(파견, 용역) 제외
표현변수	값도메인	{1 취업, 2 미취업}	{0 이상의 정수}	{0 이상의 정수}
	측정단위	-	명	명
유니버스		만 15세이상 인구	모든 사업체	종사자 1인 이상을 고용하고 있는 모든 사업체
통계프로그램		경제활동인구조사, 지역별고용조사	전국사업체조사	사업체노동력조사

❍ 국내 활용 방안: 정보 클래스별 관리

- 관리 실익과 재사용성을 고려하여 Unit Type, Code List, Variable을 주요 표준 관리 대상으로 선정하고, Universe는 작성 지침 중심으로 관리할 것을 제안함

* Conceptual Domain은 개념변수당 1개만 참조해야 하는 제약관계로 인해 재사용성이 제한되므로 표현변수의 Value Domain으로 관리 갈음

<표4> 주요 정보 클래스별 국내 현황 및 관리 방안 제안

정보 클래스	국내 현황	관리 방안 제안
Unit Type	조사 수준에서만 관리, '기타' 36%	변수 수준으로 확장, 표준값 집합 마련
Universe	'작성 대상 범위'로 조사 수준에서 기술	표준값 대신 작성 원칙·지침 제공
Statistical Classification	통계법에 따라 체계적으로 제정·관리	현행 방식 유지
Code List	별도 관리체계 부재, 같은 범주도 통계별 명칭 상이	표준 관리대상으로 설정, 범주 정의·위계·버전·이력·대응표·변형 관리
Variable	작성 틀 부재로 통계 프로그램별 편차	표준 변수와 통계 프로그램별 변수 이원화

주요 결과 (계속)

- (Code List) '혼인상태' 범주와 같이 인구주택총조사 '배우자있음', 경제활동인구조사 '유배우', 생활시간조사 '배우자있음(동거포함)'처럼 동일 개념이면서 범주 명칭이 다른 경우, 표준 코드목록을 정의하고 각 통계가 표준과의 대응표를 작성하면 비교가능성 제고

- (Variable) 표준 변수는 공통 개념과 정의를, 통계 프로그램별 변수는 작성 목적과 맥락에 따른 구체적 정의와 용법을 이원화하여 관리하고, 코드목록, 관련 통계표 등은 링크를 통해 재사용성 및 정보 클래스 간 연계성 제고

<표5> 표준 변수와 통계 프로그램 변수에 대한 작성 정보(안)

구분	표준 변수	통계 프로그램 변수
정보 클래스	• 주제영역 | 개념 | 개념 정의 • 변수명 | 단위 유형 | 변수 설명 • 코드목록(Link)	• 주제영역 • 변수명 | 단위 유형 | 변수 설명 • 코드목록 | 표준 코드목록 대응표 | 측정단위
추가 정보	• 관련 통계 프로그램 변수(Links) • 관련 통계표(Links)	• 통계 프로그램명 • 관련 통계표(Links)

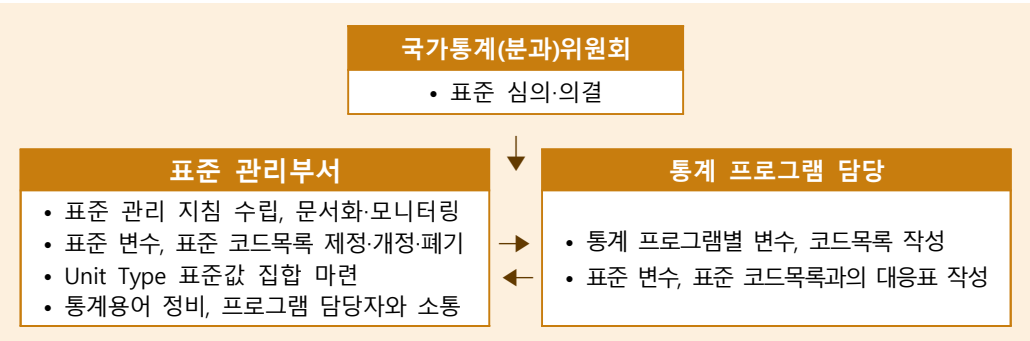
❍ 국내 활용 방안: 개념정보 관리 거버넌스

- (현황) 국내에는 국가통계 개념을 정의하는 틀이 일원화되어 있지 않고, 통계분류 외에는 개념을 공식적으로 표준화하고 관리·조정하는 절차가 마련되어 있지 않음

- (표준 관리 지침) 캐나다의 '표준에 관한 정책'을 벤치마킹하여 표준 관리 범위, 명칭, 의무 수준을 규정하고, 제정·개정·승인·폐기 절차와 관리 주체를 명확화 필요

- (역할 구분) 표준의 심의·의결은 국가통계위원회를 활용하고, 표준 관리부서는 공통 개념과 표준값을, 통계 프로그램 담당자는 개별 통계의 개념정보를 관리

[그림2] 국가통계 개념정보 관리 거버넌스(안)



❍ 실무 적용 고려사항

- (작성 범위) 조사표 문항 및 마이크로데이터 수준에서 나타나는 개념을 관리하는 것이 이상적이나, 관리 가능성을 고려하여 공표 통계표에 사용되는 개념과 변수를 우선 포함하되, 통계 프로그램 담당자에게 일정한 자율을 부여하는 방안 고려 가능

  * 교육정도 산출에 쓰이는 학력, 수학구분 코드처럼 공표 통계표에 드러나지 않지만 여러 통계에서 활용되는 개념이 있을 수 있음

- (개념 체계) 개념을 체계적으로 이해하기 위한 보조수단으로, Concept System 정보 클래스 활용하거나 현행 30개 주제 영역을 세분화하는 방안 고려 가능

- (개념의 역할 연계) '취업'은 개념변수, Universe, 범주로 모두 쓰일 수 있으므로, 개념 저장소 설계 시 개념의 역할 간 범용 활용성을 고려해야 함

시사점

❍ (선별 적용) GSIM 정보 클래스를 그대로 도입하기보다, 각 정보 클래스의 관리 실익과 국내 통계환경을 고려하여 선별·현지화하여 적용하는 것이 바람직함

❍ (벤치마킹) 캐나다의 표준 정책, 의무 수준 차등화와 범주 정의 제공, 노르웨이의 코드목록 구조 조회와 버전·대응표 관리를 상호 보완적으로 참조할 필요가 있음

❍ (단계적 접근) 통계 개념 표준 관리체계의 연착륙을 위해 통계 프로그램별 변수와 코드목록을 먼저 구조화하여 확보한 뒤, 공통성과 활용성이 높은 정보부터 단계적으로 표준을 지정·확대하는 접근 방식을 제안함

* 국가데이터처 작성 통계 → 중앙부처 작성 통계 → 지역통계 순으로 단계적 접근 고려 가능

❍ (기대 효과) 개별 통계에 분산된 개념 정보를 공통 구조로 통합 관리하고 표준 개념과 연계함으로써 통계 간 비교가능성과 재사용성을 획기적으로 향상하고, 나아가 통계·데이터 연계·융합을 촉진하는 핵심 기반이 되기를 기대함

한계 및 향후 연구

❍ (적용 범위) 본 연구는 캐나다·노르웨이 등 일부 국가의 사례를 검토하고 노동 분야를 중심으로 시범 적용했다는 한계가 있으므로, 다양한 통계 분야로 실증 범위를 넓히고 통계작성기관과의 협의를 통해 실효성을 검증할 필요

❍ (매뉴얼) GSIM 정보 클래스에 대응하는 국문 통계용어를 정립하고, 통계작성기관 담당자가 이해할 수 있도록 정보 클래스별 상세 작성 기준을 명확히 해야 함

❍ (시스템 연계) 변수·코드목록 정보를 국가통계포털과 연계하여, 하나의 개념정보를 여러 시스템에서 일관되게 조회·활용할 수 있는 구조 마련 필요

참고 문헌

○ 국가데이터처(2026),「2026년 경제활동인구조사 지침서 외부용(2026년)」
○ 국가데이터처, 통계설명자료서비스, https://www.k-stat.go.kr/metasvc
○ 국가데이터처, 통계정책시스템, https://www.narastat.kr/pms/index.do
○ DDI Alliance(2025), DDI-Cross Domain Integration Specification Overview
○ ISO/IEC(2023), ISO/IEC 11179-1:2023, Information technology — Metadata registries (MDR) — Part 1: Framework, Fourth edition
○ Statistics Canada, Classifications, variables and statistical units, https://www.statcan.gc.ca/en/concepts/search
○ Statistics Canada, Policy on standards (revised July 14, 2004), https://www.statcan.gc.ca/en/about/transparency/policy/standards
○ Statistics Canada, Surveys and statistical programs, https://www23.statcan.gc.ca/imdb-bmdi/pub/indexa-eng.htm
○ Statistics Norway, About SSBs variable system (Published: 3 December 2021), https://www.ssb.no/en/metadata/om-ssbs-variabelsystem
○ Statistics Norway, Metadata, https://www.ssb.no/en/metadata
○ UNECE(2024), Generic Statistical Information Model (GSIM): Version 2.0 User Guide
○ UNECE, GSIM Version v2.0 (December 2023), https://unece.github.io/GSIM-2.0
○ UNECE Statistics, GSIM 2.0 Webinar 2025, YouTube (@UNECEStats), 2025.9.16., https://www.youtube.com/playlist?list=PLe6uBaBVebZDQXgqM-G__clN-MAVhXQ08
○ UNECE Statswiki, Generic Statistical Information Model, https://statswiki.unece.org/spaces/gsim/pages/59703371/Generic+Statistical+Information+Model

⇨ **(연구보고서 원본)** 국가데이터연구원홈페이지 > 연구간행물 > 연구보고서 > **국가통계 개념의 체계적 관리를 위한 GSIM 활용 방안**



사망원인통계 원사인 분류 AI 적용 연구

이석민
전고설
최지원

국가데이터연구원
데이터과학연구팀

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

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

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

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

연구배경 및 목적

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

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

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

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

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

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

데이터 및 방법

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

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

[그림1] 사망원인통계 예측분석 시스템 개념도



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

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

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

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

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

<표1> AI 적용 모델 개요

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

주요 결과

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

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

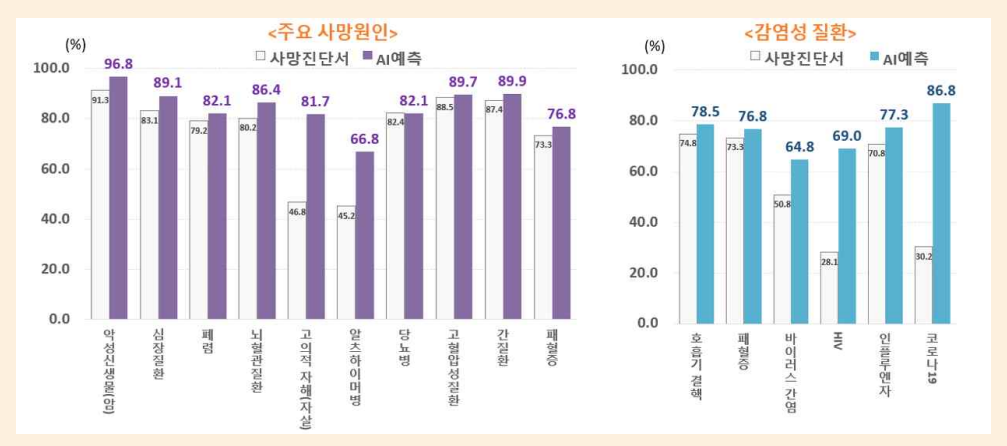
❍ **(모델2) 행정자료를 반영하여 잠정 원사인을 예측한 결과 기존 사망진단서 기반 결과보다 높은 성능을 보여, 공표시점 단축과 정확도 제고가능성을 모두 검증**

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

<표2> 잠정 원사인 예측 정확도 비교

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

[그림2] 사망진단서 및 AI 예측 원사인 정확도(2024년)



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

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

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

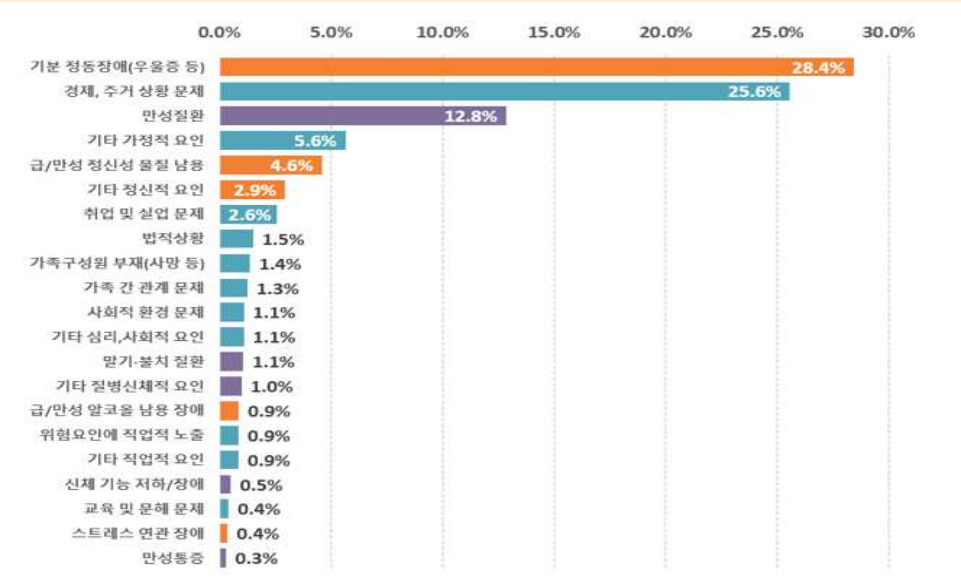
- 1인당 평균 1.16건의 자살 위험요인이 검출되었으며, 주로 기분 정동장애(28.4%)와 경제적 요인(25.6%)에 집중되어 있었음
- 그리고 기존 통계에서는 포착하기 어려웠던 세부 위험요인 및 위험요인 조합을 추가로 식별함

주요 결과 (계속)

<표3> 2024년 자살 위험요인 대분류별 분포

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

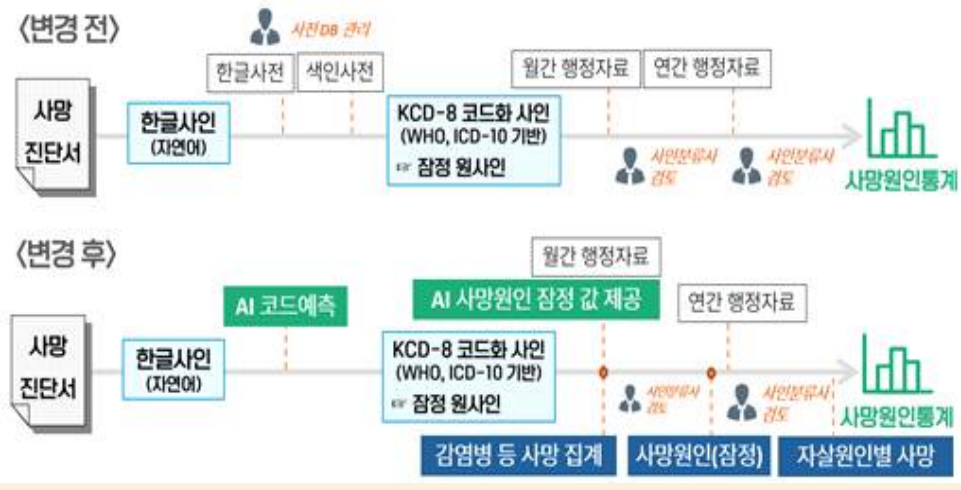
[그림3] 모델 예측결과 자살 위험요인 구성비(2024년)



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

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

[그림4] 원사인분류 프로세스 : 서비스 변경 전·후 업무흐름도



시사점

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

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

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

한계 및 향후 연구

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

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

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

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

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

참고 문헌

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

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



2025년 기준 경제총조사 조사표(도·소매업) 개선 연구

- 인지면접과 조사원 FGI를 통한 조사표 사전평가 -

남상민
국가데이터연구원
데이터과학연구팀

박주언
국가데이터연구원
데이터방법연구실

- **(연구목적)** 본 연구는 2025년 기준 경제총조사 도·소매업 조사표의 신규 및 변경 항목을 대상으로 **응답 가능성, 이해도 및 조사현장 적용성**에 대해 **개선 의견**을 **도출**하고자 하였다.
- **(연구방법)** 도·소매업 **사업체 응답자** 19명 대상 **인지면접(cognitive interviewing)**과 **시범 예행조사 조사원** 30명 대상 **초점집단면접(FGI)을 병행**하였다.
- **(주요결과)** 대부분의 문항은 응답이 가능했으나, 일부 문항(**AI 활용 여부, 건물 연면적, 인터넷 판매 중 모바일 비중** 등)에서 **개념 이해, 자료 확보, 응답 기준**에 대한 기준의 **검토**가 **필요**하였다.

  인공지능(AI) 활용 문항 단계화, 문항별 예시 및 지침 강화, 건물 연면적 조사 보조자료 제시, 모바일 비중 문항 재설계 및 사업실적 배치의 종합 검토

- **(활용)** 본조사 조사표 · 조사지침 · 조사원 교육을 개선하여 응답 부담과 측정오차를 줄이기 위한 근거자료로 활용할 수 있다.

연구배경 및 목적

❍ **(연구배경)** 국가데이터처의 2025년 기준 경제총조사는 2020년 대비하여 조사표가 10종에서 7종으로 변경되었으며, 일부 조사 항목에서 변경, 신규, 행정자료 대체 등의 변화가 있었음

* 신규항목: 외국인 종사자 수, 인공지능(AI) 활용 여부, 무인 매장 운영 여부 등

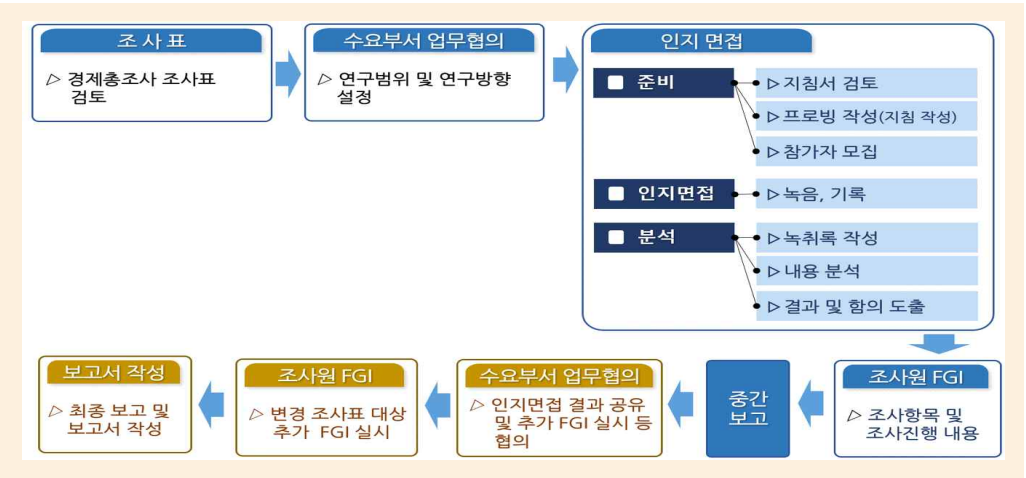
❍ **(검토관점)** 응답자가 질문을 어떻게 이해·회상·판단하고 답하는지 뿐 아니라, 조사원이 실제 현장에서 어떤 설명과 순서로 질문하는지에 대한 종합적인 검토

❍ **(연구목적)** 본 연구에서는 도·소매업 조사표를 대상으로 전체 문항을 검토하되, 신규 및 변경 문항과 사업실적 항목 배치를 중점적으로 살펴보고자 함

연구방법

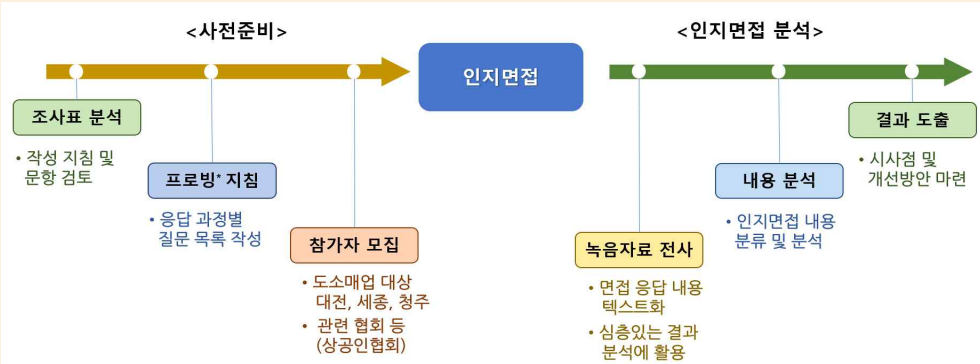
- **(대상)** 경제총조사 조사표 7종 중 도소매업 조사표
- **(범위)** 항목 배치와 신규 항목을 중심으로 조사표 전반에 대한 검토
- **(연구방법)** 응답자 대상 인지면접(cognitive interviewing)과 시범예행조사 조사원 대상 초점집단면접(FGI)을 통한 조사표 및 항목 평가

[그림1] 연구추진 방법 및 과정



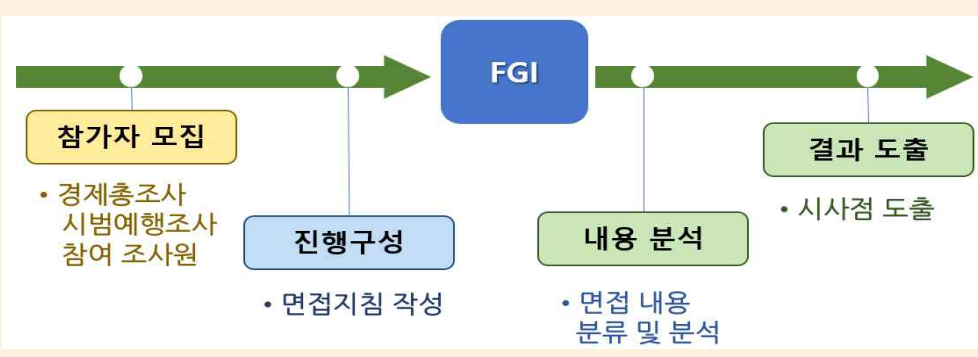
- **(인지면접)** 자기기입 후 일대일 프로빙(캐어묻기, Probing)을 통해 문항 이해, 판단 기준, 응답 근거와 오류 가능성을 확인하고, 면담 내용을 녹음·기록하여 분석함

[그림2] 인지면접 절차



- **(FGI)** 조사표 배부, 작성, 회수 과정, 질문 방식, 난이도 높은 항목, 신규문항, 사업실적 질문 순서와 개선 의견을 논의

[그림3] FGI 절차



주요 결과

❍ 보고서에서는 도소매업 조사표 항목별로 의견이 제시되어 있으나, 본 브리프에서는 신규문항(AI 활용 여부) 및 주요 변경 항목에 대해서만 제시하도록 함

❍ **(AI 활용 여부 항목 검토)** 본 연구를 추진하는 동안 조사표가 변경됨에 따라 FGI에서 변경된 조사표에 대해 추가로 의견을 검토함

[최초 안]

- **(인지면접 주요 의견)**
  - AI에 대한 개념을 이해하지 못함
  - 추가적인 예시가 필요
- **(조사원 FGI 주요 의견)**
  - 조사 문항에 대한 단계적 질의
  - 응답할 수 있는 응답자 또는 부서 확인
  - 조사원 교육시 명확한 해설 및 설명 필요
  - 활용 분야에 대한 단순화 필요
- **(시사점)**
  - "활용하지 않음"으로 응답시 하위 활용분야를 확인하지 않고 다음 문항으로 이동하여 누락 가능성 있었음
  - 인공지능을 활용하고 있는 분야에서 일반화할 수 있는 예시 제시 → 일반화에 어려움 있으나, 최대한 제시 필요
  - 생산공정, 시스템 등에 대해 정확히 답할 수 있는 담당자 또는 부서를 확인하도록 조사지침과 조사원 교육 보완

<그림4> AI 활용 여부 최초(안)

❽ 인공지능(AI) 활용 여부

[1] 활용하고 있음 → ❽-1 인공지능(AI)을 활용하고 있는 분야 [2] 활용하지 않음

- 인공지능(A.I.: Artificial Intelligence) : 텍스트 마이닝, 컴퓨터 비전, 음성 인식, 머신러닝, 딥러닝 등과 같은 기술을 사용하여 인간의 지각 능력, 학습 능력, 자연어 처리 능력 등을 컴퓨터가 실행할 수 있도록 프로그램으로 구현한 기술을 말함.
- '활용'은 **사업체에서 비용을 지불하는 공식적인 업무상 활용**을 말함(개인적으로 업무에 활용하는 경우는 포함하지 않음)

❽-1 인공지능(AI)을 활용하고 있는 분야 (중복 선택 가능)

(1) 분야		활용 여부	설 명
기획	① 시장 조사·예측	□	소비자 수요, 경쟁사 동향, 매출 흐름 등을 분석하여 제품 전략이나 판매 예측 등에 활용
	② 제품·서비스 개발	□	신제품 기획이나 서비스 고도화를 위해 소비자 의견, 데이터 분석을 통한 아이디어 도출 및 제품 테스트에 활용
생산	③ 생산	□	생산 공정 최적화, 생산 라인의 불량 감지 등 자동화를 통해 효율을 높이는 데 활용
	④ 품질·판매·물류 관리	□	품질 개선, 수요 기반 공급계획, 재고 예측, 배송 경로 최적화 등 판매 및 물류 효율화를 위해 활용
마케팅	⑤ 고객 지원	□	고객 문의 자동 응답(챗봇, 보이스봇 등), 불만 사항 분류 등 고객 상담 및 서비스 대응을 자동화하는 데 활용
	⑥ 홍보	□	홍보 콘텐츠 생성, 광고 효과 분석 등 고객 맞춤형 홍보 활동에 활용
조직 관리	⑦ 안전·보안 관리	□	작업장 사고 예방, 출입 통제, 보안 감시 영상 분석 등 직원 및 시설의 안전 확보와 데이터 보안에 활용
	⑧ 경영 지원	□	회계, 세무, 인사 등 내부 관리 업무의 자동화로 활용
	⑨ 직원 교육·훈련	□	직원 맞춤형 교육 콘텐츠 제공, 역량 평가 자동화 등 사내 교육에 활용
⑩ 기타(_____)		□	① ~ ⑨에 해당하지 않는 다른 분야

⬇

[변경안]

- 변경안은 조사원과 응답자 모두에게 최초안 보다 직관적으로 이해 되었으며, 대다수 조사원이 변경안을 선호함. 다만 분야별 예시를 지침서에 제시하고, 조사원 교육시 문항 의미화 사례를 상세히 설명할 필요는 남아있음

* AI 활용 여부 변경안에 대해 중간보고 이후 2개 시범예행조사 지역 조사원 20명을 대상으로 추가 검토

<그림5> AI 활용 여부 변경(안)

❾ 인공지능(AI) 활용

- **인공지능(A.I.: Artificial Intelligence)**은 인간의 지능으로 할 수 있는 사고, 학습, 문제 해결 등의 능력을 컴퓨터가 수행하도록 만드는 기술로 생성형 AI, 머신러닝(딥러닝) 등이 해당됨
  - 활용 예시 : ChatGPT·뤼튼(Wrtn) 등을 통한 정보 검색·문서 작성, AI 비서, 챗봇, 지능형 CCTV 등
- **활용**은 사업체에서 **'업무 목적'**으로 활용하는 모든 경우를 의미함
- **비용**은 인공지능(AI) 활용 관련 **서비스 유료 구입비, 본사 자체 개발·투자** 등 포함함

[1] 활용하고 있음 →

(1) 분야		활용 여부 (중복 선택 가능)
① 기획	상품 개발, 시장 조사·예측, 경쟁사 동향 파악, 아이디어 도출, 의사결정 지원 등	□
② 생산	생산·품질·판매·재고·물류 관리 및 업무 자동화, 고객 맞춤형 서비스 생성(교육·보건·금융서비스 등)	□
③ 마케팅	홍보 활동, 광고 콘텐츠 생성, 고객 상담 자동화 등	□
④ 조직관리	회계·세무·인사 업무 자동화, 안전·보안 관리 등	□
⑤ 기타	(_____)	□

(2) 활용 관련 비용 지출	① 비용 있음	② 비용 없음

[2] 활용하지 않음

❍ **(추가적인 지침 제공 필요)** 경제총조사는 사업체명, 대표자, 사업의 종류 등 출력 조사표를 제공하여 수정하지만, 신규 사업체의 경우는 모든 사항을 작성해야 하므로 조사표 내 최소한의 작성 지침 제공 필요

- **(사업체 대표자, 조직형태)** 공동 대표 2명을 모두 기재하는 오류가 있어 1명만 기재, 개인 사업체가 법인 대상 사업체구분란까지 응답하는 오류 등을 줄이도록 대상 기준 및 주의사항을 제시

주요 결과 (계속)

- **(사업의 종류)** 출력 조사표가 제공되는 사업체와는 달리 신규사업체 응답자는 투입, 과정, 산출을 작성하는데 어려움이 있어 이를 위해 조사표 하단 및 기입 요령을 제작할 때 산업별 대표적인 예시의 제시가 필요

- **(사업체 건물 연면적)** 응답자나 조사원이 가장 어려워하는 항목으로 인지면접에서도 불응이 많았고, 현장 조사에서도 조사원마다 조사 방식의 차이가 있었음. 이에 따라 조사 지침서에 현장 조사 불응 시 조사 방법, 활용 사이트 등을 제시하여 조사원이 원활하게 조사할 수 있도록 검토

❍ **(조사표 항목 디자인)**

- **(사업체명)** 사업체명에서 사업체명과 실제 사용하는 상호가 다름에도 사업체명만 작성하는 경우가 있었음. 이에 따라 사업체명과 간판명을 구분하여 작성하도록 항목 구분에 대한 검토가 필요

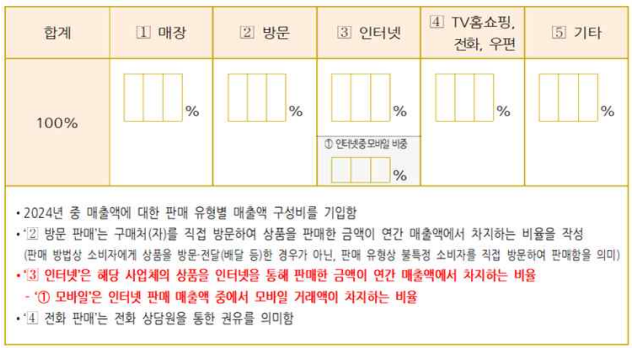
- **(상품 판매 유형별 매출액 구성비)** 인터넷 중 모바일 매출액 구성비가 인지면접에서 오류가 많았으며, 조사원 면접에서도 응답자가 회계장부 등 세부적인 자료를 가지고 있지 않아 정확한 응답이 어려웠다고 제시. 모바일 비중에 대한 항목의 활용 여부를 파악하여 항목 삭제를 검토하거나, 일부 항목의 디자인 변경을 통해 오류의 최소화가 필요

<그림6> 상품 판매 유형별 매출액 구성비 조사표 변경안

[기존]



[변경안]



<표1> 조사항목별 주요 검토 내용 요약

조사문항	주요내용	검토 의견
사업체명	• 간판명(실제 상호명) 누락	• 사업체명과 간판명을 분리
사업체 대표자	• (출력 조사표) 연령대 오류	• 행정자료 연계시 검토
조직형태	• (신규 사업체) 개인사업체임에도 사업체구분 작성	• 지침서에 기준 명시, 오류 사례 검토에 활용
사업의 종류	• (신규 사업체) 응답자가 투입, 과정, 산출 작성의 어려움	• 조사표 내 주요 산업 예시 제공 • 조사표 기입요령 제공
종사자 수 및 연간 급여액 /외국인 종사자	• (변경안) 종사자 수, 연간급여액 문항과 외국인 종사자 문항의 포함 및 제외 관계가 혼동되지 않도록 문항의 관계와 기준을 지침서에 제시	• 지침서에 외국인에 대한 종사자 수 및 연간 급여액 포함하여 조사 내용 제시
사업체 건물 (사업장) 연면적	• 건물 연면적은 19명 중 7명만 응답한 반면 매장 연면적은 대다수가 응답, 건물 연면적 불응 및 조사원 작성 어려움	• 건물 연면적을 활용할 수 있는 사이트(K-GEO 등)나 확인 방법 지침서 수록
상품 판매 유형별 매출액 구성비	• 인터넷 중 모바일 비중 응답 오류 가능성	• 모바일을 인터넷 하위 항목으로 시각적으로 배치하여 포함관계를 명확히 하고, 활용도가 낮다면 삭제도 검토

주요 결과 (계속)

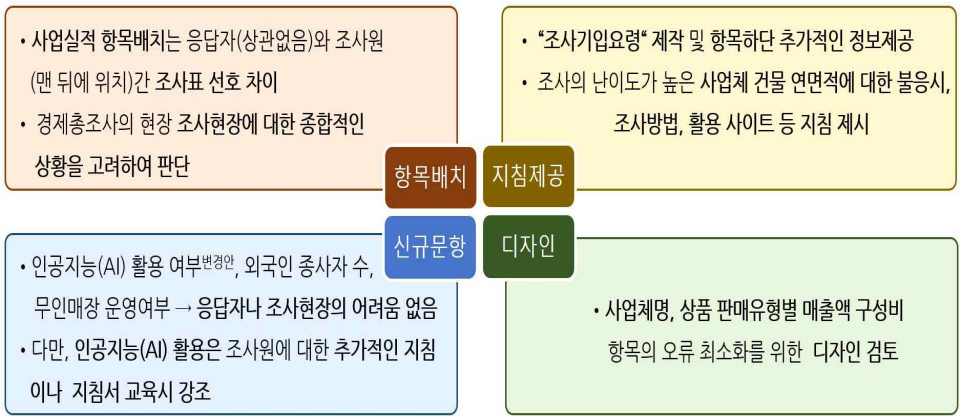
❍ **(항목 배치)** 조사표상 사업실적의 항목배치에 대한 검토

- 2015년 기준 조사부터 조사표 종류의 축소에 따라 산업별 공통으로 조사하는 항목은 앞부분에, 그 외 산업별 특성항목은 공통항목 뒤에 배치. 이에 따라 사업실적은 공통항목 마지막에 배치하여 조사함
- **(인지면접)** 사업실적의 조사표 위치가 중간 부분이든 맨 뒷부분이든 차이가 없다(10/19명)는 응답이 많았음
- **(조사원 FGI)** 다수의 조사원은 맨 뒤에 배치하는 것을 선호(25/30명) 하였음. 다수의 조사원은 응답자가 조사표를 직접적으로 기재하기보다 조사표 분실, 훼손을 방지하기 위해 현장에서 조사원이 면담을 통해 응답받기 때문
- **(시사점)** 응답 흐름 측면에서는 맨 뒤 배치 선호가 높지만, 마지막 배치는 누락·불응 가능성을 높일 수 있어 응답 편의와 누락 위험을 함께 고려해 결정할 필요

맺음말

❍ 본 연구에서는 응답자 인지면접과 조사원 FGI를 병행하여 도·소매업 조사표를 사전 평가하고, 응답 과정과 현장 조사 과정에서 발생할 수 있는 측정오차 요인을 점검함.

❍ 그 결과, 대부분의 문항은 응답이 가능했으나 AI 활용 여부와 같이 개념 설명이 필요한 문항, 건물 연면적과 같이 자료 확보가 어려운 문항, 인터넷 판매 중 모바일 비중과 같이 문항 구조상 혼동이 발생하는 항목에서는 추가적인 개선이 필요한 것으로 나타남



❍ 향후 AI 활용 문항의 단계적 질의와 조사원 교육 강화, 신규사업체 대상 작성지침 및 예시 보완, 건물연면적 확인을 위한 보조자료 제공, 모바일 비중 문항의 재설계 또는 삭제 검토가 필요하다. 사업실적 항목배치는 조사원의 현장 선호와 응답누락 가능성을 함께 고려하여 결정할 필요가 있다. 이러한 개선은 조사표의 이해 가능성과 현장 적용성을 높이고 응답 부담과 측정오차를 줄이는데 활용될 수 있음

참고 문헌

○ 백선미, 박선희. (2020). 2021 경제총조사 조사표 개선 연구: 도매 및 소매업. 통계개발원(현 국가데이터연구원)

○ 박선희, 박현주, 박주언. (2018). "조사표 평가 방법: 인지면접". 통계개발원(현 국가데이터연구원)

○ 박주언, 양경진. (2015). "2015년 기준 경제총조사 시범예행조사 인터넷조사표 평가". 통계개발원(현 국가데이터연구원)

○ 양경진, 박주언, 박현정. (2014). "인지적 기법을 활용한 「경제총조사」 조사표 평가: FGI 결과보고". 통계개발원(현 국가데이터연구원)

○ Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). Survey Methodology. John Wiley & Sons, Inc

⇨ **(연구보고서 원본)** 국가데이터연구원홈페이지 > 연구간행물 > 연구보고서 > **2025년 기준 경제총조사 조사표(도·소매업) 개선 연구**

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

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