

인공지능(AI) 기반 조사데이터 품질 제고 방안

- 한국의료패널 사례를 중심으로 -

2026.06.25 | 이기호

CONTENTS

Korea Institute for Health and Social Affairs

PART.1

개요

PART.2

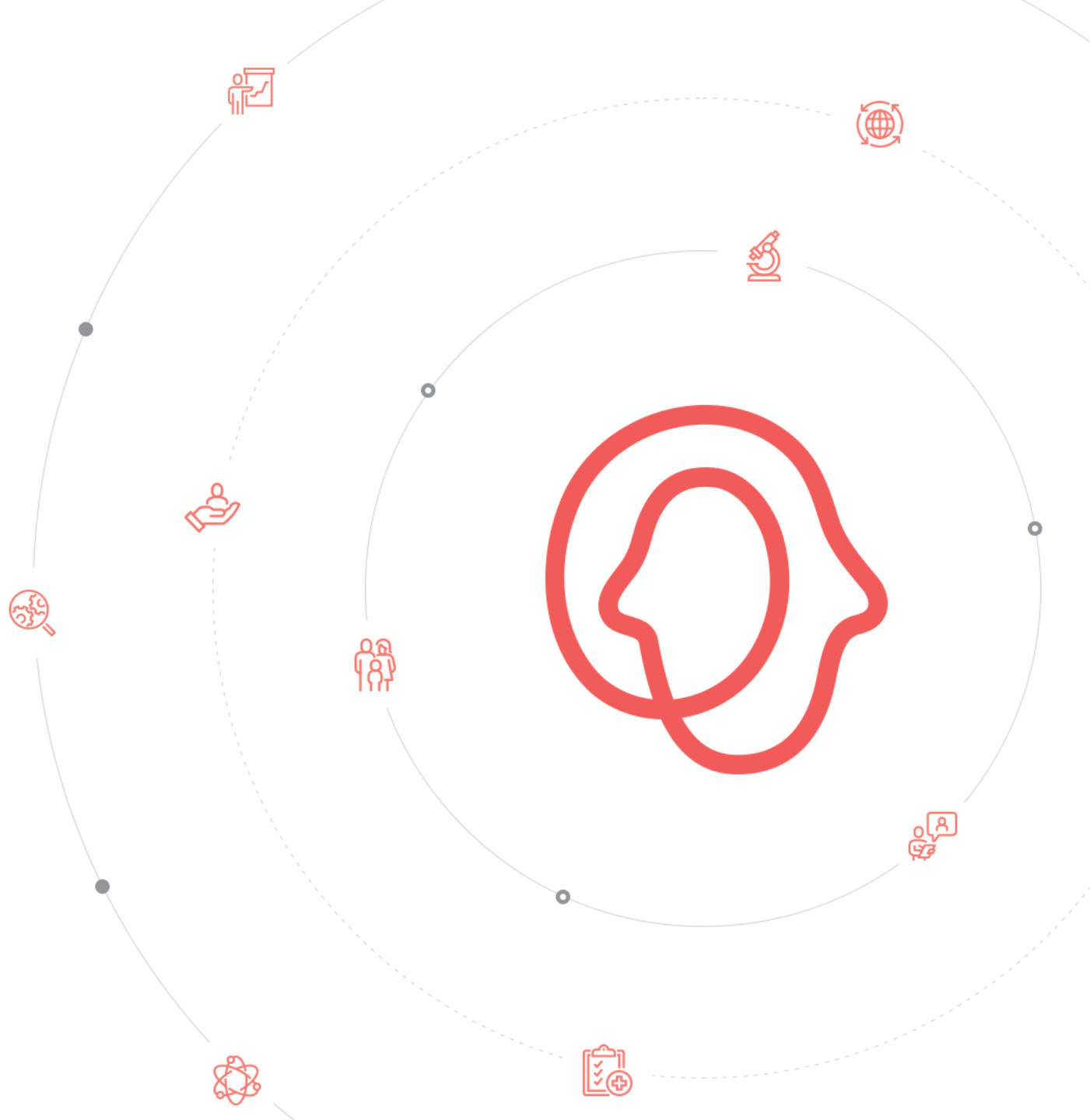
한국의료패널

PART.3

데이터 실증

PART.4

성과 및 향후 과제

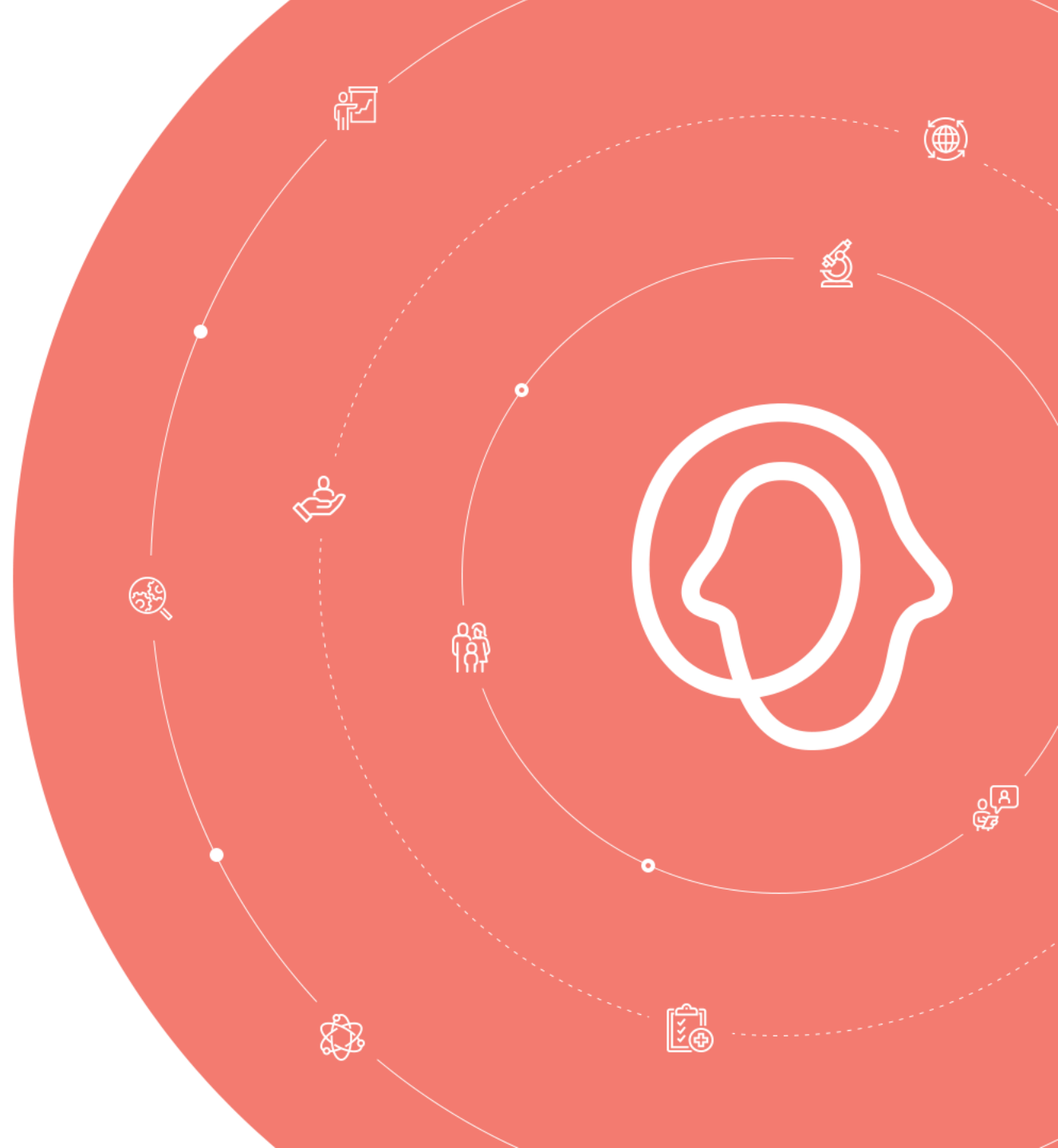


PART.1

개요

01 | 배경 및 필요성

02 | 인공지능 기반 통계작성 선행 연구 및 사례 검토



01. 개요

01

배경 및 필요성

- ✔ **우리원은 국가승인통계 3종을 비롯하여 각종 연구수행을 위해 매년 70~110건의 조사데이터 생산**

구분	2016년	2017년	2018년	2019년	2020년	2021년	2022년
조사데이터 생산(건)	79	80	76	98	116	98	74

- ✔ **조사데이터 오류는 연구결과의 신뢰성 저하**

- 조사데이터는 연구의 기반이 되는 데이터로서 수정되지 않은 데이터 오류로 인해 품질 문제가 생기면 해당 연구결과의 신뢰성 저하

- ✔ **조사데이터 담당부서 및 담당자의 부담 증가**

- 연구 종료 후 이관된 조사데이터는 외부 공개(마이크로데이터 배포)를 원칙으로 하고 있어, 외부 공개된 조사데이터의 오류는 국책연구기관의 이미지에 악영향을 미칠 수 있기에 조사데이터 담당부서 및 담당자의 부담 증가

01. 개요

02

연구 목적

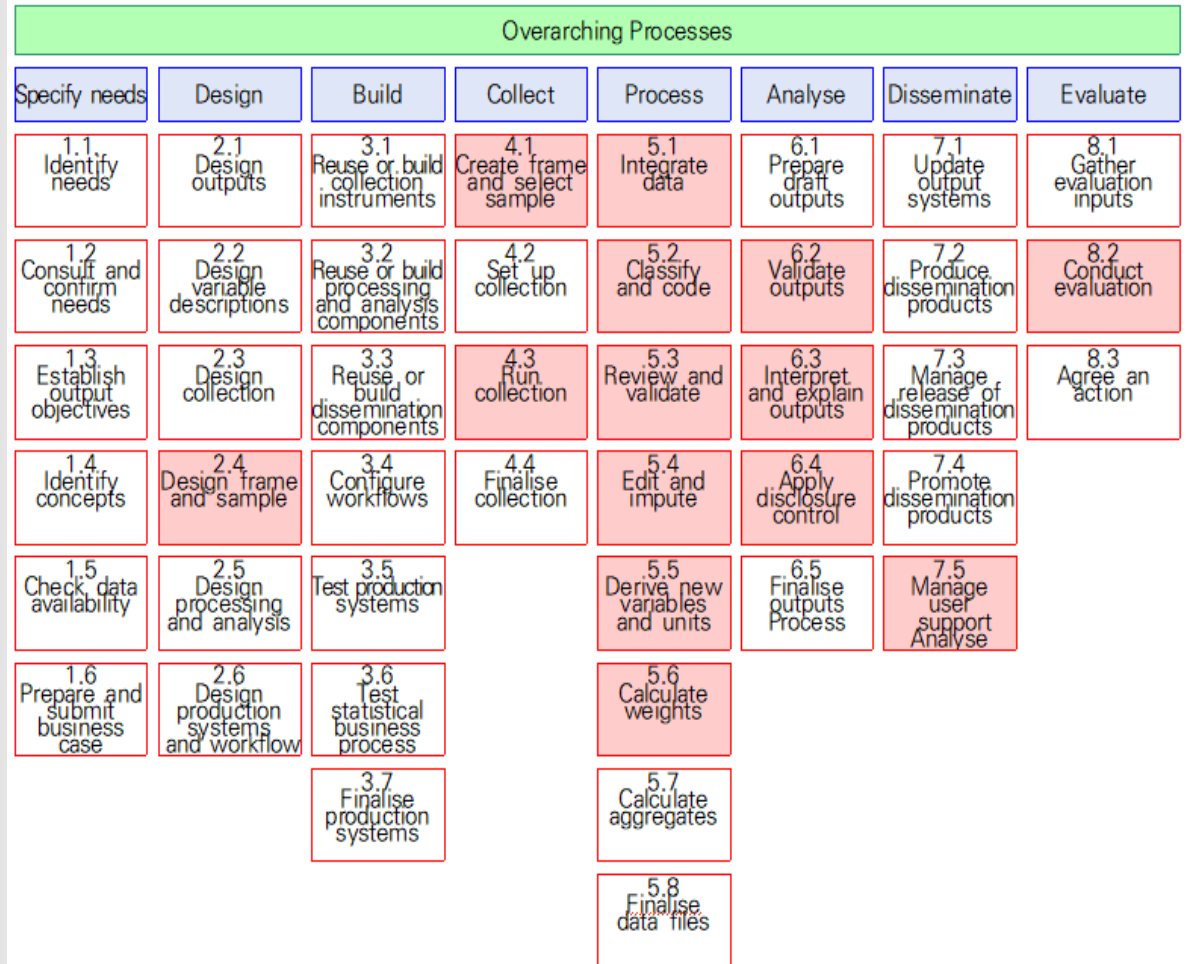
- ✔ **한국의료패널 조사데이터 자료 내검 업무를 중심으로 조사데이터 품질 제고 및 업무 효율화를 위한 인공지능(AI) 기술 활용 방안 모색**
- 국내외 선행연구 및 사례분석을 통해 통계생산 및 관리를 위한 인공지능(AI) 기술 활용 가능성을 검토
- 한국의료패널 데이터 실증을 통해 우리 연구원 조사데이터 품질 및 업무 효율성 제고 위한 인공지능(AI) 기술 활용 방안을 제시

02. 인공지능(AI) 기반 통계작성 선행연구 및 사례 검토

01

인공지능(AI) 적용가능 업무 검토(국내외 선행연구 고찰)

- (Beck 등., 2018) 독일연방통계청 국가 및 국제 통계생산기관의 머신러닝 활용 실태
- (Yung 등, 2018) 유럽경제위원회(UNECE) 머신러닝팀의 1차 데이터 머신러닝 활용 가능성 검토
- (UNECE, 2021) 머신러닝 공식 통계 활용 촉진을 위한 프로젝트 결과보고서
- (UNECE, 2023) 공식 통계 분야에서 대형 언어모델(LLM) 백서
- (이상현, 2021) AI를 활용한 국내외 통계편제 현황
- (김정민, 2021) 국가통계 AI도입 활성화 방안
- (이기재, 2023) 조사연구에서 AI 활용 방안
- (김정민, 2024) 통계분야의 AI 도입-국내외 사례 및 향후과제



02. 인공지능(AI) 기반 통계작성 선행연구 및 사례 검토

02

통계작성 사례분석

✓ 코딩 및 분류(Coding and Classification)

- (캐나다) 산업 및 직업 코딩, (미국 노동통계청) 업무상 상해 및 질병 코딩, (아이슬란드) 사회 통계 종사에서 직업 및 산업의 자동 코딩, (노르웨이) 표준 산업 코드 분류, (우리나라) 산업분류, 한류콘텐츠 분류 등 다수 국가에서 자동 코딩 및 분류에 활용

✓ 편집 및 대체(Edit and Imputation)

- (이탈리아) 개인 기본 등록부 ‘교육수준’의 대체, 공공행정 등록부 편집, (영국) 생활비 및 식품 조사 소득 데이터 편집, (벨기에) 머신러닝을이용한 에너지 밸런스 통계 추정 등 다수 국가에서 활용

✓ 이미지 분석(Imagery Analysis)

- 호주, 네덜란드, 멕시코, UNECE 등에서 위성사진 분석에 활용

02. 선행연구 및 통계작성사례 검토

03

기타: 대형 언어모델(LLM) 적용 사례분석

✓ 호주통계청(ABS)

- 통계 분류 정의 업데이트 시 대형 언어모델(LLM)을 활용하여 직업별 수행 작업 목록 초안을 생성 → 분석가의 업무 효율성을 향상

✓ 아일랜드 통계청(CSO Ireland)

- SAS코드를 R코드로 변환하는 도구 개발 → 코드 변환 작업 간소화 및 자동화

✓ 국제통화기금(IMF)

- StatGPT를 개발 → 자연어 기반 인터페이스를 통해 사용자가 IMF 통계 데이터에 쉽게 접근하고 활용할 수 있도록 지원

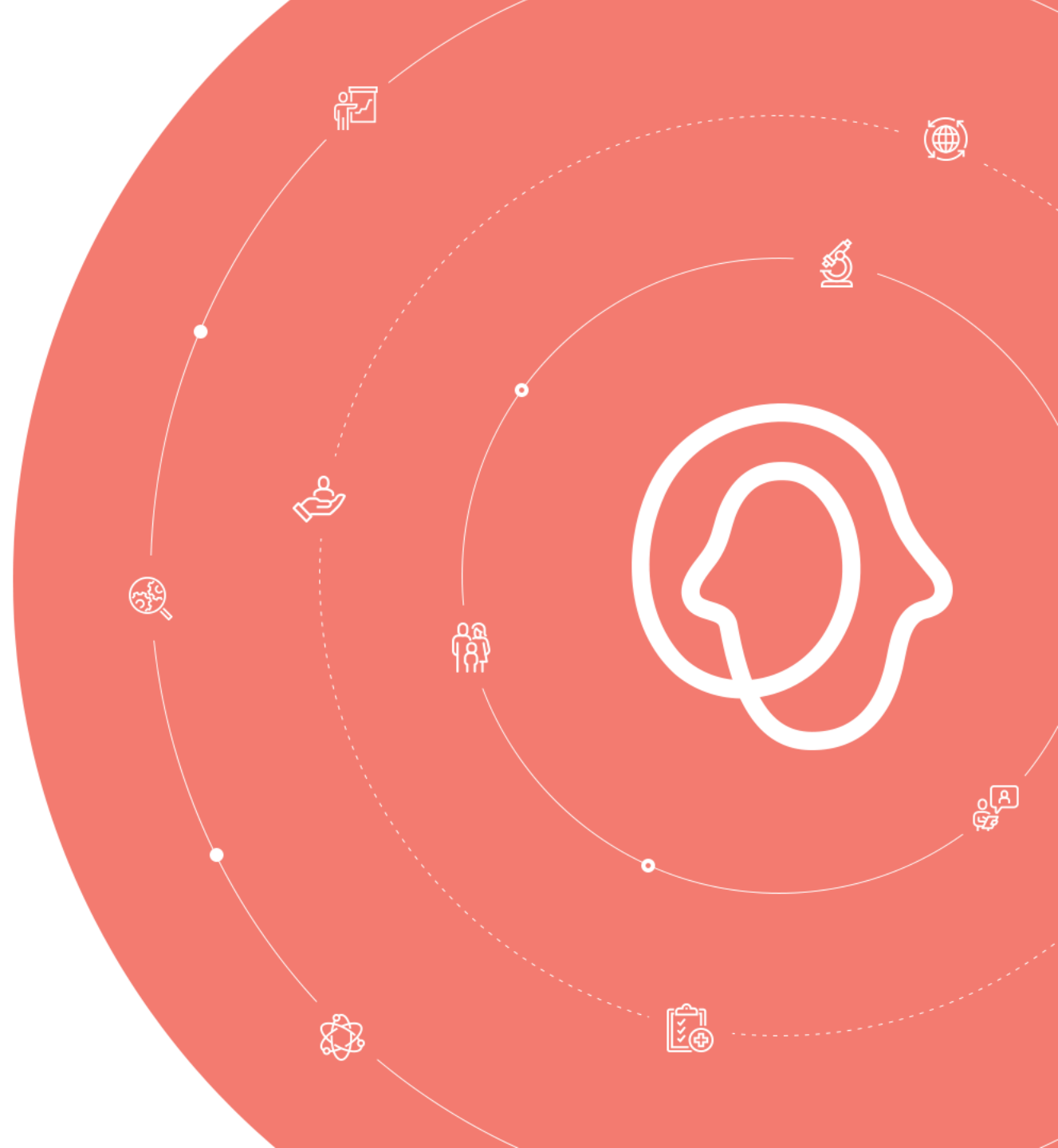
✓ 우리나라(통계청)

- 일상언어로 묻고 답하는 초거대 AI 통계 챗봇 서비스 개발 → 대형 언어모델(LLM)기반으로 자연어 질의로부터 의도와 맥락을 파악하여 통계정보를 검색하고 생성결과를 제공

PART.2

한국의료패널

- 01 | 조사 소개
- 02 | 조사내용 구성 및 조사 영역
- 03 | 조사데이터 수집 과정
- 04 | 조사데이터 내검
- 04 | 조사데이터 품질관리 이슈



01. 한국의료패널 개요

01

조사 소개

✓ 조사목적

- 의료이용 행태와 의료비 지출 규모에 관한 정보 뿐만 아니라 의료이용 및 의료비 지출에 영향을 미치는 요인들을 포괄적이고 심층적으로 분석할 수 있는 패널 데이터 구축

단기적목적

- 국민 개개인 및 가구 단위의 의료비 규모 산출
- 의료비 재원 분석
- 보건의료서비스 수요자의 이용행태 및 공급자 특성 분석
- 보건의료정책수립을 위한 보건복지관련 지표 생산
- 건강보험급여자료와의 연계성을 통한 의료비 데이터 생산의 완전성 구축

중장기목적

- 국민의료비 산출 및 변화 양상 추적
- 의료비 지출 양상과 패턴에 대한 지속적인 모니터링
- 주기적, 종단적 데이터를 분석하여 의료비의 흐름(인과관계) 분석
- 보건의료정책 수립 및 평가를 위한 동태적 보건복지관련 지표 생산

✓ 조사연혁

- (제1기 한국의료패널) 2008년 1차 조사를 시작으로 2019년 총 14차 조사까지 진행
- (제2기 한국의료패널) 2020년부터는 한국의료패널을 개편하여 2025년까지 6차 조사 진행

✓ 조사근거: 통계청 승인 일반통계(승인번호 제920012호)

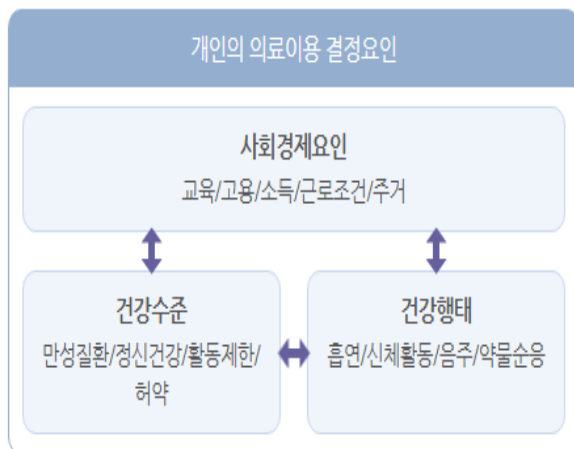
01. 한국의료패널 개요

02

조사내용 구성 및 조사영역

✓ 조사내용 구성

보건의료서비스 이용		
보건의료이용	반응성	접근성/형평성
- 응급/입원/외래 서비스 - 건강검진/예방접종 - 의약품/건강기능식품 이용 - 장기요양 서비스	- 서비스 및 보건의료제도 만족도 - 의사에 대한 신뢰도 - 공유된 의사결정 참여 경험 - 기타 환자경험(과잉진료 경험 등)	- 의료서비스 공급 - 경제적, 물리적 접근성 - 미충족 의료 - 가계부담 의료비



의료비 재원
- 건강보험, 장기요양보험 - 의료급여 - 민간의료보험 - 산재보험 - 자동차보험 - 본인부담 - 기타(민간 및 제도적 지원 등)

✓ 조사영역

제2기 한국의료패널 조사	
가구 조사	가구원 조사
▪ 가구 및 가구원 일반 사항 ▪ 경제활동상태 ▪ 소득(가구원, 가구) ▪ 부채 및 자산 ▪ 지출 ▪ 만성질환 이환 ▪ 일반의약품 및 의료관련 지출 ▪ 의료서비스 이용 - 응급(환자경험 포함) - 입원(환자경험 포함) - 외래 - 건강검진 ▪ 장기요양 및 기타 돌봄 서비스(출산 포함) ▪ 민간의료보험1	▪ 건강수준 - 정신건강 - 삶의 질 ▪ 건강생활 습관 ▪ 의료접근성 ▪ 상용치료원 ▪ 민간의료보험2

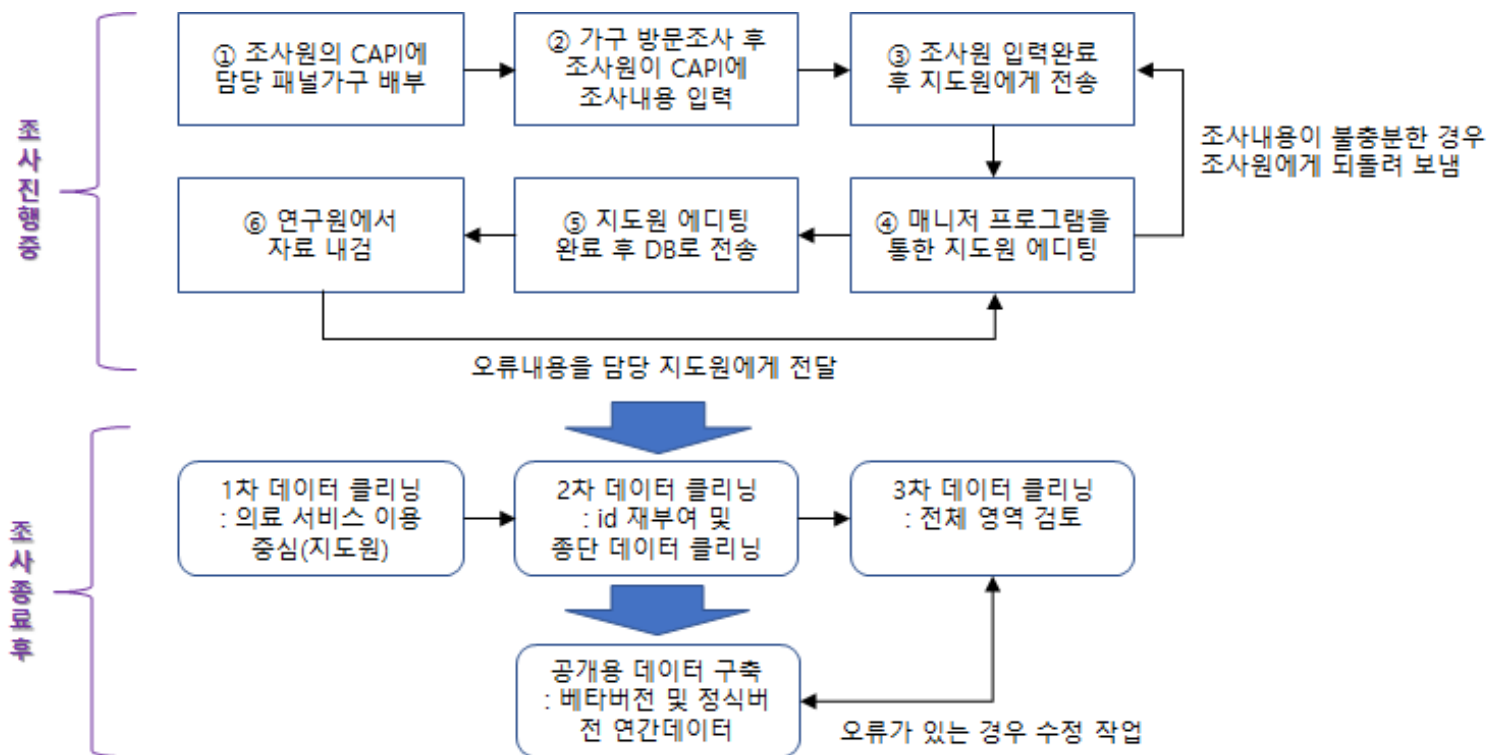
01. 한국의료패널 개요

03

조사데이터 수집 과정

✓ 데이터 내검

- (조사 진행 중) 조사원, 지도원, 데이터 클리닝 담당자(연구원)가 데이터 내검 실시
- (조사 완료 후) 데이터 클리닝 담당자(연구원)가 데이터 내검 실시



01. 한국의료패널 개요

04

조사데이터내검



데이터클리닝 담당자 내검(연속형 변수 60개, 다년도 검토 변수 3개)

- 데이터클리닝 담당자(연구원)는 기본적인 범위 내검이나 설문로직 내검 뿐 아니라 CAPI시스템에 구현하기 복잡하고 조사원 또는 지도원이 발견하기 어려운 연속형 변수의 이상치 탐색이나 다년도 검토가 필요한 데이터 내검 수행

구분	설명	비고	구분	설명	비고
소득1	- [근로소득] 이상치 재확인 * 고소득 근로소득자의 경우 “경제활동 상태”의 [주된 경제활동 상태]와 [직업분류] 재확인 * <u>육아휴직인</u> 경우 근로소득 0일 수 있음	- 이상치 탐지	소득2	- 작년(이전) 소득과 비교 재확인 * 별도로 에러체크는 하지 않음	- 다년도 검토
	- “가구원 일반사항2”의 [의료보장 형태]가 ‘⑤ 의료급여(1.2종)세대주’ 또는 ‘⑥ 의료급여 세대원’일 때 [근로소득] 재확인 연 근로소득이 2000만원을 초과하거나 50만원 미만인 경우	- 이상치 탐지	만성질환 이환	- [전체 질환별 만성질환 유무] 개인의 만성질환 이환 변화 비교 재확인	- 다년도 검토
	- [사업소득] 이상치 재확인 * 상위 1% 등 고소득 사업소득자의 경우 “경제활동 상태”의 [주된 경제활동 상태]와 [직업분류]를 재확인 * [사업소득]은 마이너스(-)가 있을 수 있으므로 <u>모름(-9)</u> 와 잘 구분하여야 함	- 이상치 탐지	건강생활 습관	- 지난 차수와 비교하여 [신장], [몸무게] 불일치 재확인 * 키 5cm 이상, 몸무게 15kg 이상 차이날 때 재확인 요청	- 다년도 검토
	- “가구원 일반사항2”의 [의료보장 형태]가 ‘⑤ 의료급여(1.2종)세대주’ 또는 ‘⑥ 의료급여 세대원’일 때 [사업소득] 재확인 * 연 사업소득이 1000만원 이상인 경우	- 이상치 탐지			

01. 한국의료패널 개요

05

조사데이터 품질관리 이슈(데이터 클리닝 담당자 인터뷰)



변수간 논리적 관계 오류 문제



완료 과제



향후 과제



개선 과제

- 유관 변수 간 논리적 불일치가 발생 → 로직 개발 및 CAPI시스템 구현(구현 완료)



연속형 변수의 이상치 탐지 문제

- ①데이터 입력시 빈번한 오류, ②복수개의 변수에 대한 여러 조건들 확인해야 하는 담당자의 업무부담 가중, ③통계패키지(SAS) 활용한 이상치 탐지 어려움 → 보다 직관적으로 이상치를 판단할 수 있는 시각화된 분석방법



의료서비스 이용 데이터(이용금액) 입력 오류 문제

- 영수증 등 의료서비스 이용 기록을 토대로 데이터 입력시 데이터 입력 오류 발생 → 이미지 분석 및 자동 입력 구현



질병코드 분류의 문제

- 의무기록사 등 전문인력이 텍스트로 작성한 의료서비스 이용 내용 분석하여 질병코드 분류 → 질병코드 자동 분류 방법 구현(다만, 제2기 패널에서는 분류하고 있지 않은 비공개 변수)



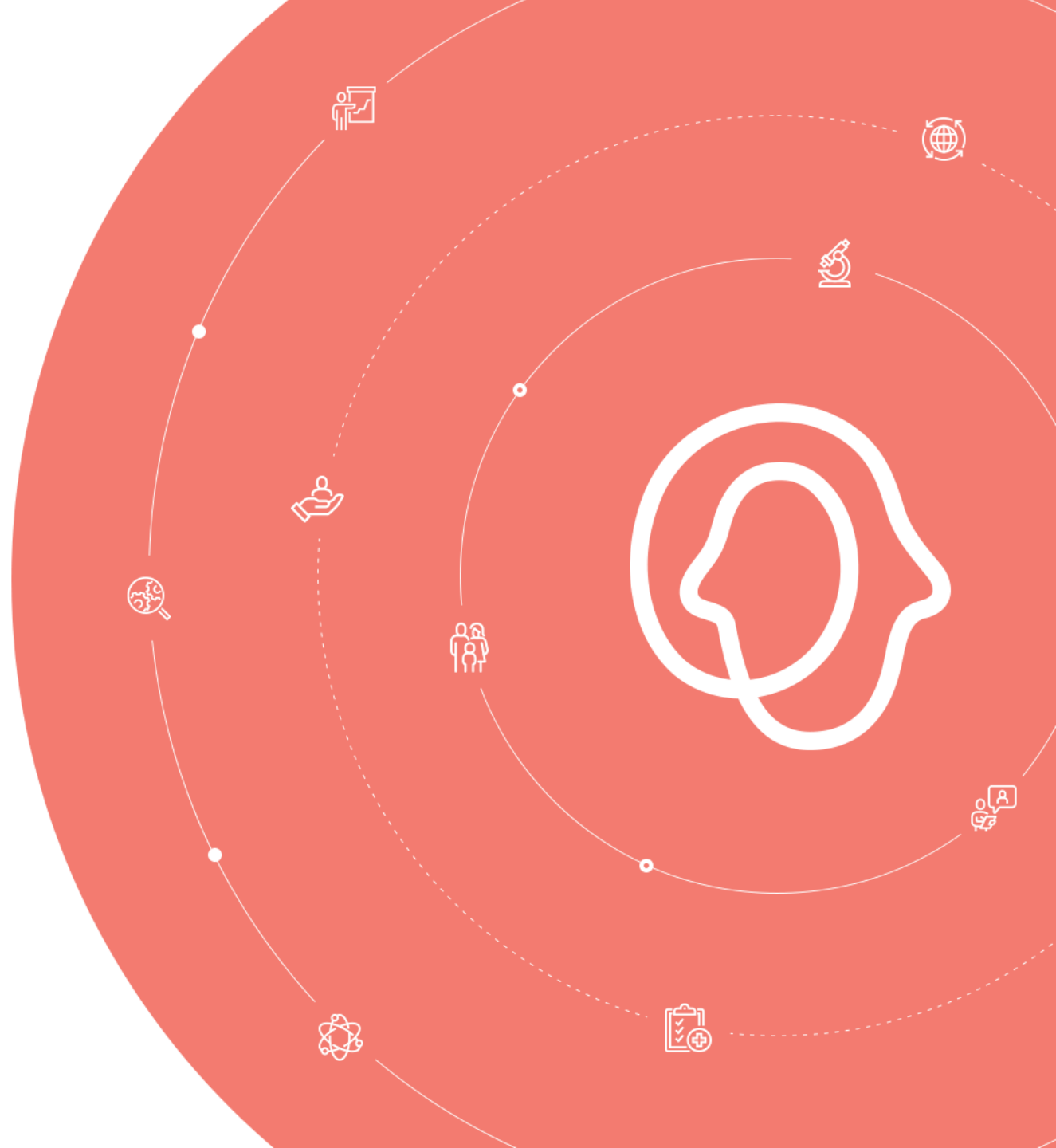
이상치 탐지를 위한 기존 점검 기준 개선 및 새로운 점검 기준 개발

- 기존 63개의 점검 기준 개선 및 새로운 점검 기준 개발 필요 → LLM을 활용한 이상치 탐지 검수 질의문 생성

PART.3

데이터 실증

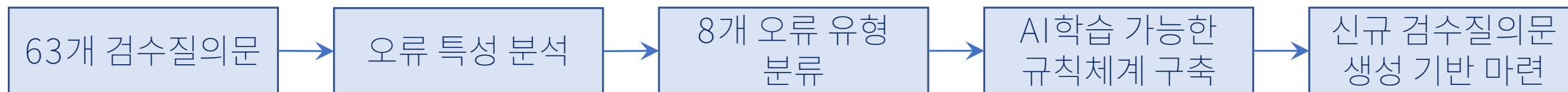
- 01 | 검수 질문 유형 분류
- 02 | 연속형 변수 이상치 탐지를 위한 시각화 방법
- 03 | LLM을 활용한 이상치 탐지 점검기준 개발
- 04 | 조사 영역 11~13
- 05 | 조사 영역 14



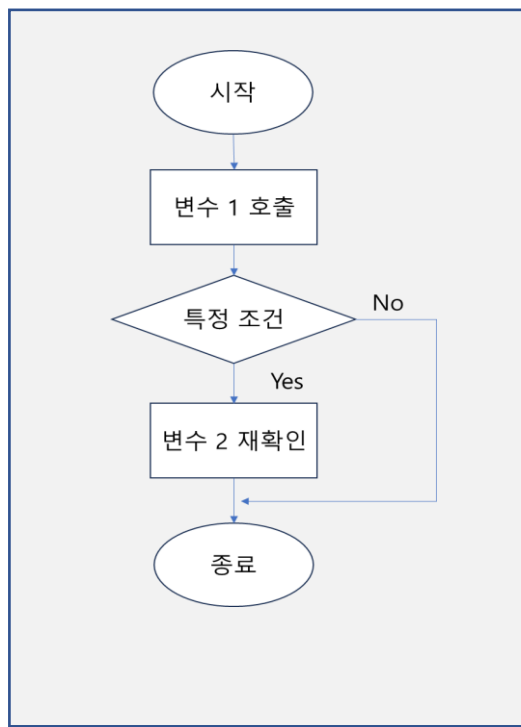
01. 검수 질문 유형 분류

✓ 63개의 검수 질의문을 SI가 학습 가능한 구조로 전환하기 위해 오류 유형 분류

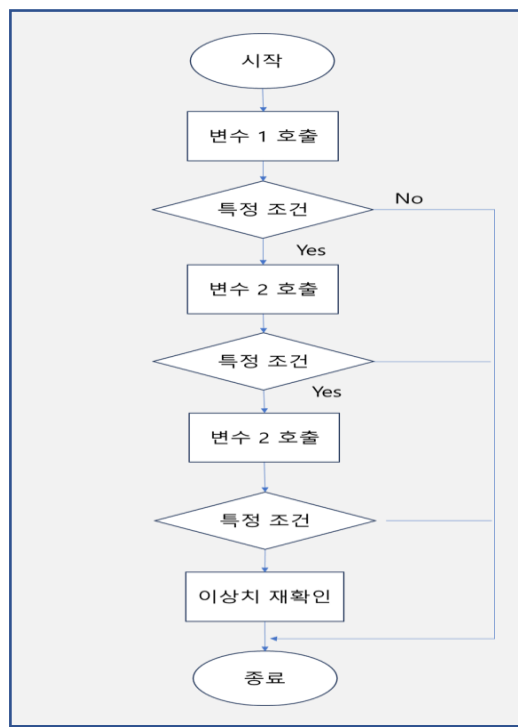
- ① 단순 조건 기반 오류, ② 다중 조건 기반 오류, ③ 논리적 관계 오류, ④ 단일 변수 이상치, ⑤ 변수 간 비교 오류, ⑥ 분포 기반 이상치, ⑦ 변수 조합 기반 이상치, ⑧ 복합 이상치 등 8개 유형으로 분류



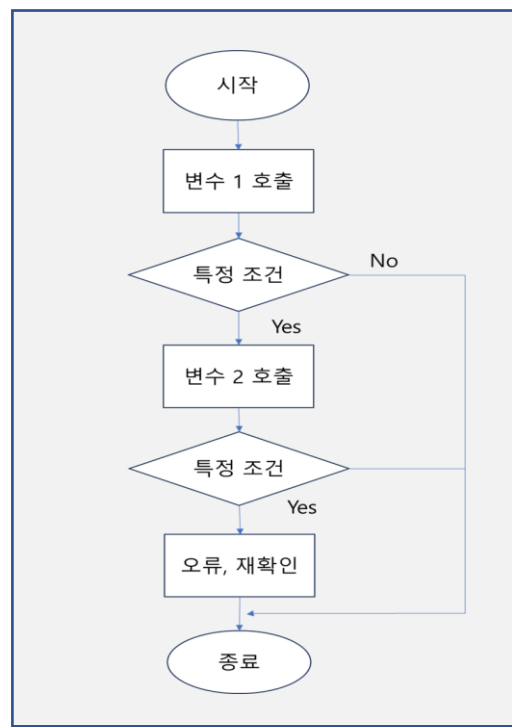
✓ 유형1



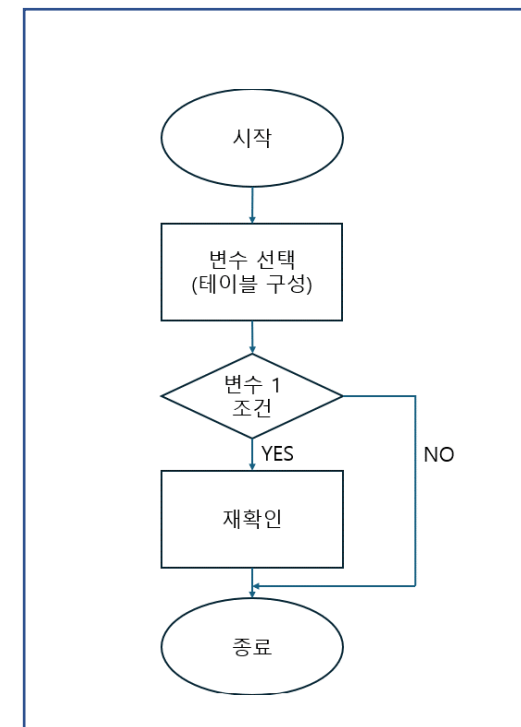
✓ 유형2



✓ 유형3



✓ 유형8



02. 연속형 변수 이상치 탐지를 위한 시각화 방법

✓ 연속형 변수 이상치 탐지를 위한 시각화 방법

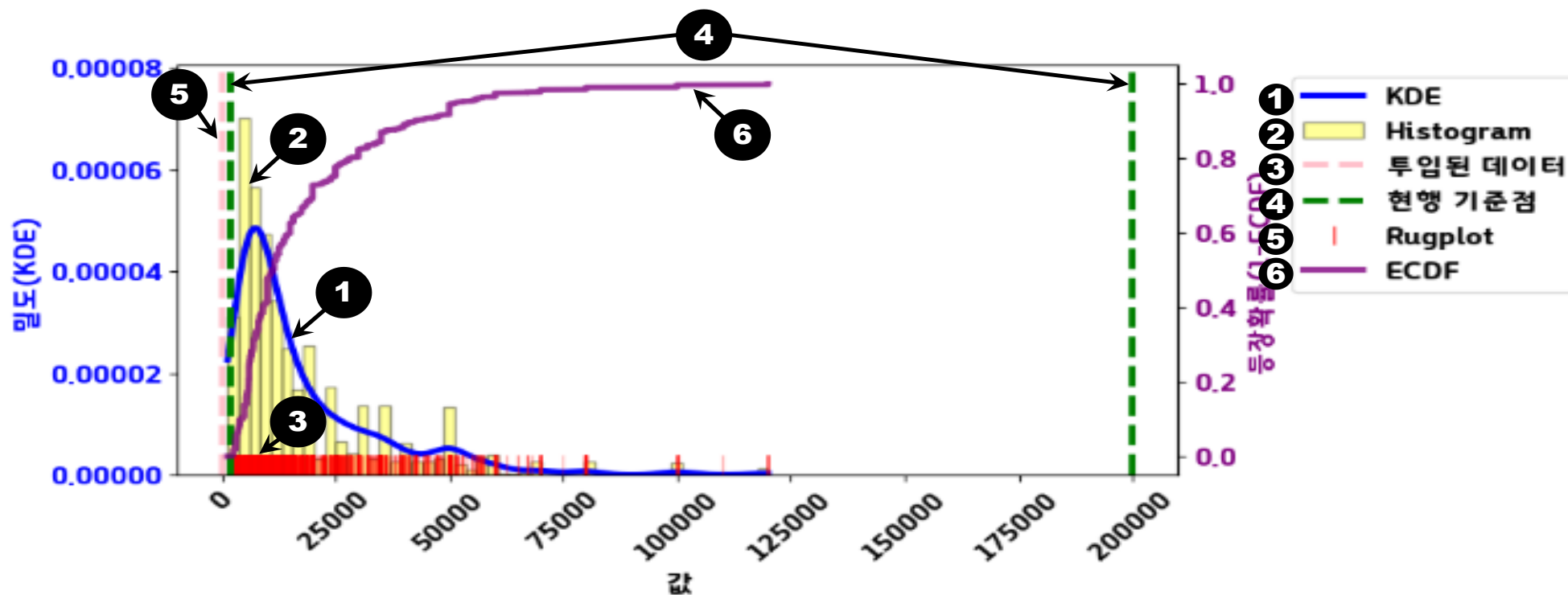
- 담당자 인터뷰를 통해 한국의료패널 전체 데이터에서 연속형 변수 이상치를 파악하기 위해 필요한 시각화 방법 선정

구분	목적	특징
커널확률밀도함수 (KDE)	데이터가 특정 구간에 얼마나 집중되어 있는지 파악	데이터 전체의 분포를 나타내기 위해 개별 데이터 포인트가 전체 분포에 기여하는 방식을 합산하여 전체 곡선을 나타냄
히스토그램	데이터의 전반적인 빈도 분포를 확인	데이터의 빈도를 직관적으로 표현
Rugplot	개별 데이터 포인트의 분포 확인	데이터가 특정 값에 얼마나 집중되어 있는지 직관적으로 파악 가능
현행 기준점	이상치 판별 기준값에 대해 분포 내 등장 위치	설정된 기준 값이 분포에서 얼마나 극단적인지 확인 가능
투입된 데이터	투입된 데이터에 대해 분포 내 등장 위치	데이터 포인트가 분포에서 어느 정도 위치해 있는지 확인 가능
경험적누적분포함수 (ECDF)	데이터의 누적 특성 및 특정 값을 기준으로 분위수 확인	y축이 0~1 사이에서 변화하며 누적 비율을 나타냄

02. 연속형 변수 이상치 탐지를 위한 시각화 방법

✓ (예시) 연속형 변수 이상치 탐지를 위한 시각화 방법 적용

- (예시: 유형1-문항4) [의료보장형태(HEALTH_INS)]에 해당하는 첫 번째 변수에 대해 '[의료보장형태]가 5 또는 6인 경우'로 필터링을 진행한 후 두 번째 변수인 [사업소득(I_INC2)]의 수치적 분포와 시각적 분포를 확인

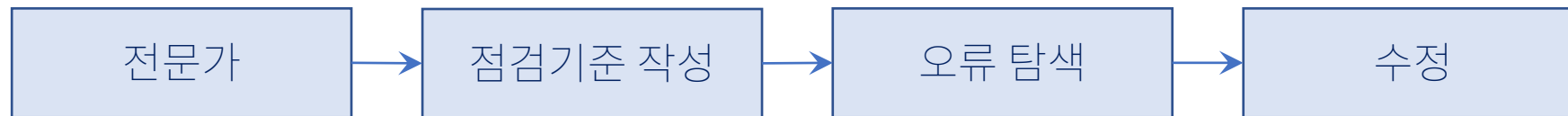


- ✓ (실무자 검토) 기존방식(SAS통계패키지)과 비교해 직관적으로 이상치를 확인 할 수 있다는 점에서 매우 유용

03. LLM을 활용한 이상치 탐지 점검기준 개발

✓ LLM을 활용한 이상치 탐지 점검기준 개발

- **(기존방식)** 전문가가 직접 조사데이터 이상치 탐지를 위한 점검기준을 개발하고 오류 탐색 후 수정과정을 반복
 - * 명확한 규칙이 존재하는 오류에는 효과적이지만 변수 간 복합적인 관계나 데이터 분포에 따라 새롭게 발생하는 이상치를 탐지하는 데에는 한계



- **(개선방식)** LLM을 활용하여 이상치 탐지부터 검수질의문 생성, SAS 코드 자동생성
 - * 기존의 규칙기반 품질관리체계를 확장하여, LLM과 RAG를 활용해 새로운 검수기준을 생성하고, 데이터 분포를 반영한 이상치 기준을 도출하며, 최종적으로 SAS코드 생성까지 연결하는 AI, 품질관리 체계 제안



03. LLM을 활용한 이상치 탐지 점검기준 개발

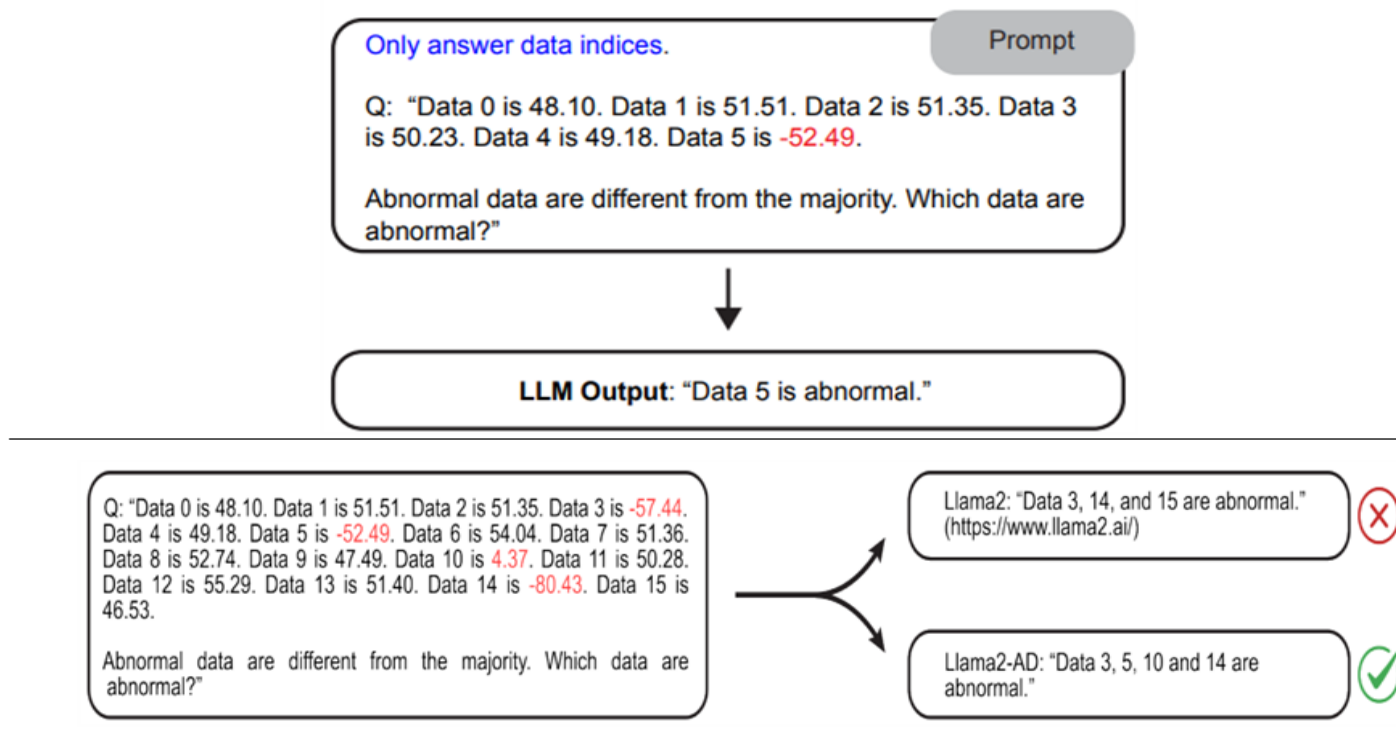
✓ 기존 방식과 개선 방식의 차별성

구분	기존 방식	개선 방식
검수 기준	전문가가 사전 정의	AI가 기존 기준을 학습하여 신규 기준 생성
오류 탐지	규칙 기반	규칙 + 데이터 분포 + 변수 간 관계 반영
코드 작성	담당자가 SAS 직접 작성	자연어 질의문 기반 SAS 자동 생성
확장성	기존 기준 중심	새로운 변수조합 및 이상치 조건 생성 가능

03. LLM을 활용한 이상치 탐지 점검기준 개발

✓ 연구개념

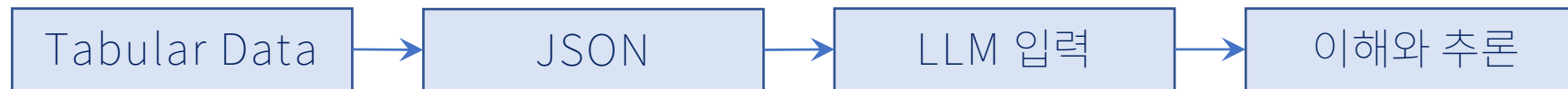
- LLM이 데이터 분포를 이해하고 다른 값들과 현저히 차이나는 데이터를 이상치로 탐지하는 최근 연구결과 반영



03. LLM을 활용한 이상치 탐지 점검기준 개발

✓ LLM 적용을 위한 데이터 직렬화

- LLM은 표(테이블) 구조보다 텍스트 기반 구조를 더 잘 이해하는 특성 → 조사데이터를 JSON형태로 직렬화



기존 데이터

ID	v1	v2	v11	v12
101	1	5.3	0.1	0.9
102	0	3.2	NA	NA
103	1	4.8	0.5	1.2
104	0	2.7	NA	NA

변환 데이터

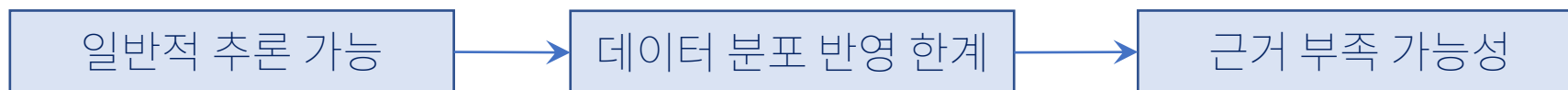
```
{  
  "id": 101,  
  "v1": 1,  
  "v2": 5.3,  
  "v11": 0.1,  
  "v12": 0.9},  
{
```

03. LLM을 활용한 이상치 탐지 점검기준 개발

✓ LLM 단독 적용의 한계와 RAG 도입 필요성

- LLM은 자연어 이해와 추론 능력이 우수하지만, 조사데이터 품질관리에서는 단순한 언어 이해만으로 불충분
- 이는 검수질의문은 변수의 의미, 조사영역, 기존 점검 기준, 실제 데이터 분포를 함께 고려해야 하기 때문임.
- RAG는 기존 검수질의문과 변수정보, 주변분포, 결합분포 정보를 검색하여 LLM이 보다 근거 있는 검수 기준을 생성할 수 있도록 지원

LLM 단독



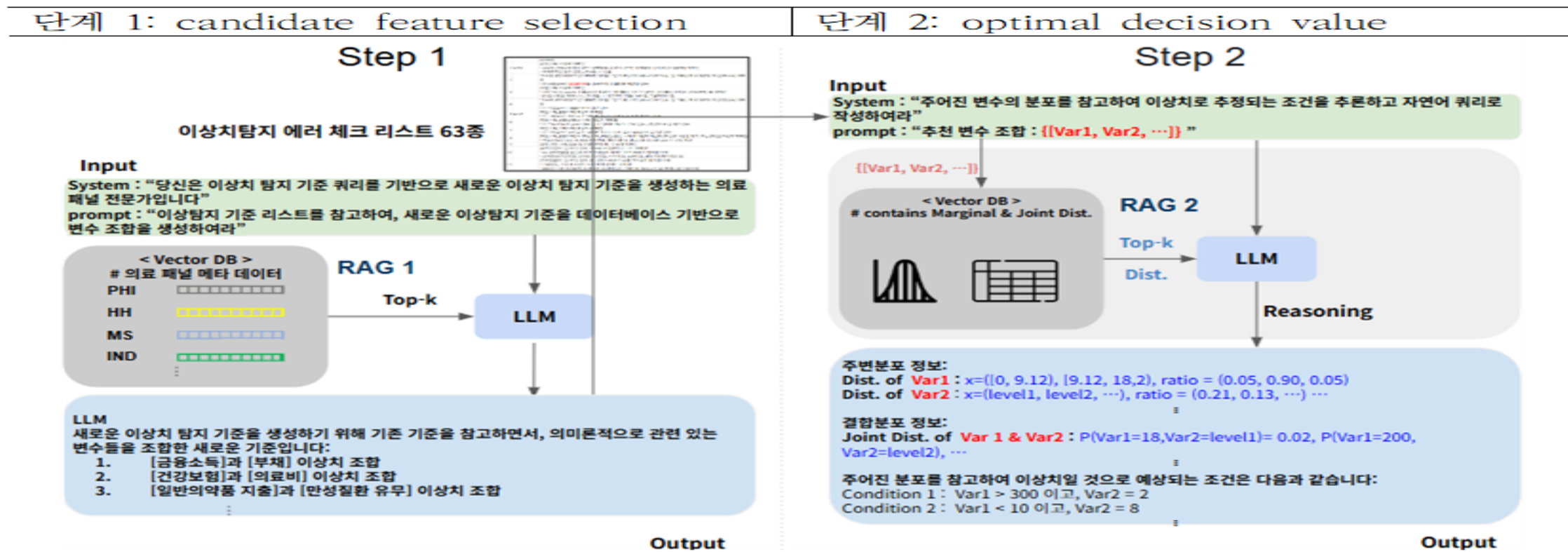
LLM + RAG



03. LLM을 활용한 이상치 탐지 점검기준 개발

✓ 분포인지 RAG 기반 검수 질의문 추천 알고리즘 개발

- (단계1) 기존 점검기준(63종)을 참고하여 이상치가 존재할 수 있는 새로운 변수조합 추천
- (단계2) 단계1에서 추천된 변수조합으로 부터 RAG_marginal, RAG_joint 내에서 유사성이 높은 documents를 찾고, 유사한 documents에서 서술된 내용을 바탕으로 분포 인지 기반의 cut-off-value를 생성



03. LLM을 활용한 이상치 탐지 점검기준 개발

✓ (예시) 단계1: RAG1 이상치 탐지기준 생성

Input	위 이상치 탐지 기준 리스트 60개를 참고하되, 리스트에는 없는 새로운 이상치 탐지 기준을 벡터 데이터베이스를 기반으로 한글변수명을 통해 2개의 의미론적인 변수 조합을 생성하여라. 다음은 제공된 리스트를 기반으로 새롭게 생성한 이상치 탐지 기준입니다. 각 기준은 서로 다른 변수를 의미 있게 조합하여 설정하였습니다: 1. [근로소득] & [사업소득] 조합: [근로소득]과 [사업소득]이 모두 0으로 보고되었으나, 개인이 고용 상태 또는 자영업자로 등록되어 있는 경우 오류 가능성이 있습니다. 2. [재산소득_금융소득] & [재산소득_부동산/동산 소득] 조합: 두 소득 유형이 모두 매우 높게 보고되었으나(예: 금융소득 > 1억, 부동산/동산 소득 > 2억), [자산]이 낮거나 0으로 보고된 경우 불일치가 있을 수 있습니다. 3. [공적연금] & [기초연금] 조합: 개인이 [공적연금]과 [기초연금]을 모두 수령하지만, 나이가 일반적인 수급 연령(예: 65세 미만)에 맞지 않는 경우 오류로 간주될 수 있습니다. 4. [의료비_수납 금액] & [의료비_비급여 금액] 조합: [의료비_수납 금액]이 매우 낮고 [의료비_비급여 금액]이 매우 높은 경우, 카테고리의 오해 또는 보고 오류일 수 있습니다.
Output	5. [교통비] & [의료비_수납 금액] 조합: 의료 방문을 위한 [교통비]가 [의료비_수납 금액]에 비해 매우 높은 경우, 교통비 보고에 오류가 있을 수 있습니다. 6. [보험료] & [보험금 수령액] 조합: [보험료]가 지속적으로 높게 보고되지만 [보험금 수령액]이 0이거나 매우 낮은 경우, 보험 청구 또는 보고에 문제가 있을 수 있습니다. 7. [신체활동별 비용] & [건강보조식품_복용] 조합: [신체활동별 비용]이 높지만 [건강보조식품_복용]이 0인 경우, 건강 관련 지출 습관에 불일치가 있을 수 있습니다. 8. [와병일수] & [결근결석일수] 조합: [와병일수]가 높지만 [결근결석일수]가 낮은 경우, 질병 보고와 실제 결근 사이에 불일치가 있을 수 있습니다. 9. [소득_소비지출] & [소득_비소비지출] 조합: [소득_소비지출]과 [소득_비소비지출]의 합이 보고된 소득을 크게 초과하는 경우, 소득 또는 지출 보고에 오류가 있을 수 있습니다. 10. [장기요양 및 기타 돌봄 서비스 비용] & [산후조리원 비용] 조합: 두 비용이 동시에 높게 보고된 경우, 돌봄 관련 비용의 중복 또는 분류 오류가 있을 수 있습니다. 이 기준들은 서로 다른 변수 간의 논리적 관계를 통해 잠재적인 불일치나 오류를 식별하는 데 목적이 있습니다.

03. LLM을 활용한 이상치 탐지 점검기준 개발

✓ (예시) 단계2: RAG2 이상치 범위 생성

Input 8번 변수 조합 기반으로 분포를 확인하고, 확인한 정보를 바탕으로 이상치 조건(cut-off value)을 새로 추론하여라.

주변분포 정보:

Dist. of [와병일수]:

$x = ([0, 10.34), [10.34, 26.28), [26.28, \infty))$, ratio = (0.68, 0.25, 0.07)[1]

Dist. of [결근결석일수]:

$x = ([0, 11.59), [11.59, 32.93), [32.93, \infty))$, ratio = (0.70, 0.25, 0.05)[2]

결합분포 정보:

Output Joint Dist. of [와병일수] & [결근결석일수]:

$P(\text{와병일수}=120, \text{결근결석일수}=5)=0.01$, $P(\text{와병일수}=60, \text{결근결석일수}=90)=0.01$, ...[3]

주어진 분포를 참고하여 이상치일 것으로 예상되는 조건은 다음과 같습니다:

Condition 1: [와병일수] > 3 * [결근결석일수]

Condition 2: [와병일수] > 26.28 (상위 7%) 이고, [결근결석일수] < 11.59 (하위 70%)

이러한 조건은 [와병일수]가 상대적으로 높지만 [결근결석일수]가 낮은 경우를 이상치로 간주하여, 질병 보고와 실제 결근 사이의 불일치를 탐지하는데 사용됩니다.

03. LLM을 활용한 이상치 탐지 점검기준 개발

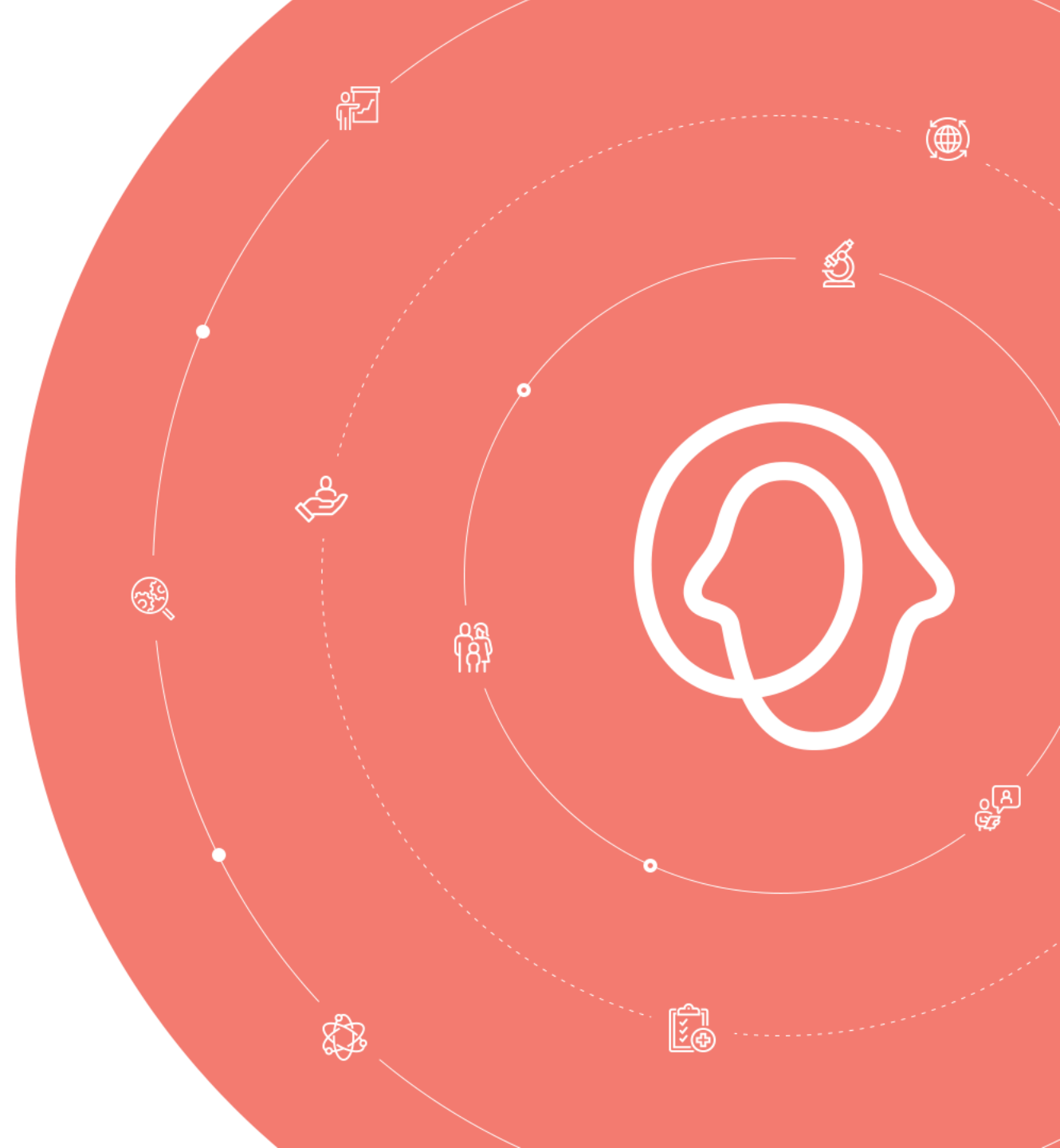
✓ (실무자 검토) 생성된 탐지기준의 실무 적용 가능성

- 신규 생성된 탐지기준 10개에 대해 실무 적용 가능성을 검토한 결과, 3개는 적용 가능, 2개는 수정 후 적용가능으로 평가
→ 즉, 총 5개 기준은 실무 활용 가능성이 확인
- 추후 더 자세한 정보를 학습시킨다면 실무에 적용 가능한 더 많은 질의문을 확보하여 내검 목록 고도화 기대

구분	생성된 탐지기준	활용
1	[근로소득]과 [사업소득]이 모두 0으로 보고되었으나, 개인이 고용 상태 또는 자영업자로 <u>등록되어 있는 경우</u> 오류 가능성이 있습니다.	△
2	[재산소득_금융소득], [재산소득_부동산/동산 소득] 두 소득 유형이 모두 매우 높게 보고되었으나(예: 금융소득 > 1억, 부동산/동산 소득 > 2억), [자산]이 낮거나 0으로 보고된 경우 불일치가 있을 수 있습니다.	○
3	개인이 [공적연금]과 [기초연금]을 모두 수령하지만, 나이가 일반적인 수급 연령(예: 65세 미만)에 맞지 않는 경우 오류로 <u>간주될</u> 수 있습니다.	○
4	[의료비_수납 금액]이 매우 낮고 [의료비_비급여 금액]이 매우 높은 경우, 카테고리의 오해 또는 보고 오류일 수 있습니다.	×
5	의료 방문을 위한 [교통비]가 [의료비_수납 금액]에 비해 매우 높은 경우, 교통비 보고에 오류가 있을 수 있습니다.	×
6	[보험료]가 지속적으로 높게 보고되지만 [보험금 수령액]이 0이거나 매우 낮은 경우, 보험 청구 또는 보고에 문제가 있을 수 있습니다.	×
7	[신체활동별 비용]이 높지만 [건강보조식품_복용]이 0인 경우, 건강 관련 지출 습관에 불일치가 있을 수 있습니다.	×
8	[와병일수]가 높지만 [<u>결근결석일수</u>]가 낮은 경우, 질병 보고와 실제 결근 사이에 불일치가 있을 수 있습니다.	△
9	[소득_소비지출]과 [소득_비소비지출]의 합이 보고된 소득을 크게 초과하는 경우, 소득 또는 지출 보고에 오류가 있을 수 있습니다.	○
10	[<u>장기요양</u> 및 기타 돌봄 서비스 비용], [산후조리원 비용] 두 비용이 동시에 높게 보고된 경우, 돌봄 관련 비용의 중복 또는 분류 오류가 있을 수 있습니다.	×

PART.4

성과 및 향후 과제



연구 성과

- ✓ 63개 검수질의문을 8개 오류유형으로 분류하여 AI가 학습 가능한 구조로 전환
- ✓ LLM을 활용하여 조사데이터 이상치 탐지 가능성 확인
- ✓ 분포인지RAG 알고리즘을 개발하여 기존 검수기준을 확장하고 신규 검수기준 생성
- ✓ 자연어 질의문을 SAS코드로 변환하는 Query-to-SAS 자동 생성 가능성 검증

성과	내용	의미
검수체계 구조화	• 63개 질의문 → 8개 오류유형	• AI 학습 기반 마련
LLM 적용	• 이상치 탐지 가능성 확인	• 규칙 기반 탐지 보완
RAG 개발	• 신규 검수질의문 생성	• 기존 검수기준 확장
Query-to-SAS	• 자연어 → SAS 코드 생성	• 업무 자동화 가능성 확인

기대효과

✓ 품질 측면

- 이상치 탐지 정확성 향상
- 신규 오류 패턴 발견 유용

✓ 운영 측면

- 반복 검수작업 시간 단축
- SAS코드 작성 부담 완화

✓ 인력 측면

- 전문가 의존도 완화
- 신규 담당자 업무 적응 지원

✓ 정책 측면

- 국가승인통계 품질관리고도화
- 연구데이터 신뢰성 제고

향후 과제

✓ 단기 과제: 한국의료패널 실증 적용 및 정확도 검증

- 한국의료패널 데이터에 대해 생성된 검수기준을 실제 내검 과정에 적용하고, 실무자 평가를 통해 정확도와 활용성을 검증

✓ 중기 과제: 우리원 조사데이터 전반으로 확대

- 한국의료패널 외에도 우리원에서 생산하는 다양한 조사데이터에 AI 기반 품질관리 기술을 확대 적용 필요

✓ 장기 과제: AI 기반 통계품질관리 플랫폼 구축

- 개별 조사 단위의 AI 적용을 넘어, 조사데이터 품질관리 전 과정을 지원하는 AI 기반 통계품질관리 플랫폼 구축 필요

감사합니다

