

헬스케어 Text-to-SQL 자연어 질의 시나리오 정리

본 문서는 건강검진·자격/보험료 샘플 데이터를 기반으로, Text-to-SQL 시스템에서 활용 가능한 자연어 질의 시나리오를 **난이도 단계별**로 정리한 것입니다.

현재 구축된 샘플 시스템에서는 아래 질의들이 자연어 → SQL 생성 → DB 조회 흐름 자체는 대부분 성공적으로 동작하고 있으며, 구체적인 결과값의 통계적/도메인적 검증 및 정교화는 OCI 프로젝트를 통해 진행 예정입니다.

1. 단순 집계/분포 조회 시나리오

목적

- 단일 테이블 또는 단순 집계 위주의 기본적인 데이터 조회
- 값의 분포와 평균 수준을 직관적으로 확인하는 단계

예시 질의

1. “남녀별로 음주습관 분포를 보여줘.”
2. “보험료 10분위별로 평균 BMI를 알려줘.”
3. “흡연 상태별 평균 BMI를 보여줘.”
4. “성별에 따라 비흡연자 비율이 어떻게 다른지 알려줘.”
5. “최근 검진 기준으로, 연령대별(20대·30대·40대 등) 평균 BMI를 보여줘.”

2. 자격 + 건강검진을 함께 보는 시나리오

목적

- 자격/보험료 정보와 건강검진 정보를 함께 활용
- 소득, 성별, 지역 등과 생활습관·BMI 간의 기본적인 패턴을 확인하는 단계

예시 질의

6. “남성과 여성의 보험료 10분위별 음주습관 분포를 알려줘.”
7. “소득분위(보험료 10분위)가 높을수록 비흡연자 비율이 어떻게 변하는지 보여줘.”
8. “성별과 보험료 10분위별로 평균 BMI를 비교해서 표로 보여줘.”
9. “시군구(지역)별 평균 BMI와 비만(BMI 25 이상) 비율을 같이 보여줘.”
10. “보험료 10분위별로 ‘거의 술을 안 마시는 사람’과 ‘자주 마시는 사람’ 비율을 비교해줘.”

3. 조건이 조금 더 복잡한 시나리오

목적

- 연령, 흡연/음주 여부, BMI 기준 등 조건을 추가해
특정 부분집합(subset)에 대한 집계/분포를 조회하는 단계

예시 질문

- “30세 이상 성인만 대상으로, 성별·보험료 10분위별 평균 BMI를 보여줘.”
- “현재 흡연자만 대상으로, 보험료 10분위별 음주습관 분포를 알려줘.”
- “BMI 25 이상인 사람들만 골라서, 성별·보험료 10분위별 인원수를 집계해줘.”
- “최근 3년 검진 데이터를 합쳐서, 성별·연령대별 비만(BMI 25 이상) 비율 변화를 연도별로 보여줘.”
- “음주를 거의 하지 않는 사람들만 대상으로, 성별·보험료 10분위별 평균 BMI를 비교해줘.”

4. 복합 질문(조회 + 설명용) 시나리오

목적

- 여러 조건과 축을 동시에 사용하는 복합 질문
- 단순 수치 조회를 넘어, 그룹 간 비교·요약을 한 번에 요청하는 형태

예시 질문

- “보험료 10분위가 낮은 그룹과 높은 그룹(1분위 vs 10분위)의 평균 BMI와 자주 음주하는 사람 비율을 성별로 나눠서 비교해줘.”
- “현재 흡연자이면서 자주 술을 마시는 사람들만 대상으로,
성별과 보험료 10분위별 인원수와 평균 BMI를 함께 보여줘.”
- “전체 검진 대상자를 기준으로,
① 성별, ② 연령대, ③ 보험료 10분위, ④ 음주습관을 축으로 해서
각각 비만(BMI 25 이상) 비율이 어떻게 달라지는지 요약해서 보여줘.”

5. 샘플 시스템 구현 현황 및 한계

1) 현재 구현 상태

- 위 18개 자연어 질문 시나리오는 현재 구축된 샘플 **Text-to-SQL 시스템**에서
 - 자연어 입력 → SQL 자동 생성 → DB 실행 → 결과 반환
까지의 동작 자체는 대부분 성공한 사례를 기준으로 정리한 것입니다.
- 즉, **엔드 투 엔드 플로우(질문 → 조회 결과 확인)**가 가능한 수준까지는 구현 완료된 상태입니다.

2) 결과값 검증 및 정교화 수준

- 다만, 각 질의에 대한 결과값이
 - 도메인 전문가 관점에서 “완전히 정확하다”고 검증된 단계는 아닙니다.
 - 컬럼 매핑, 조건 정의, 그룹핑 방식 등은 향후에 추가 검토와 튜닝이 필요합니다.
- 또한, 프롬프트/모델 특성상
 - 1~2번 시도에서 실패하고, 3번째 시도에서 성공하는 경우도 간헐적으로 발생합니다.
 - 예:
 - 처음 생성된 SQL에서 컬럼명이 일부 잘못되거나
 - JOIN 조건이 누락되는 경우
 - 질문을 살짝 바꾸거나 다시 입력하면 정상 동작하는 케이스 등
- 현재 단계에서는
 - “샘플 시나리오 기반으로 시스템이 어느 정도까지 동작하는지”를 확인하는 수준이며,
 - **안정적인 실서비스 수준의 정확도와 재현성 확보, 정교화는 OCI 프로젝트 진행 과정에서 고도화 예정입니다.**