Part I 2차자료 연관 분석 — 근로환경조사로 하는 단면연구

AI 산업보건 석사논문

2주차· 3주차

이 Part 에서 답할 질문

- 이미 있는 자료로 하는 연관 분석은 무엇을 묻고 무엇을 답하는가
- 연구 질문의 변수는 어떻게 독립변수·종속변수·보정변수로 나뉘고, 표 1과 표 2는 무엇을 보여 주는가
- 어디서 판단이 갈리는가 — 그리고 그 판단을 AI 가 대신하면 무슨 일이 생기는가

0. 문제로 시작합니다

7차 근로환경조사(KWCS, 2023)에는 취업자 50,195명의 응답이 439개 변수로 들어 있습니다. 근로시간, 교대근무, 소음, 자세, 직무스트레스, 그리고 요통·수면장애·불안우울 같은 건강 문제까지. 자료는 이미 있습니다.

한 대학원생이 묻습니다. "주 52시간을 넘게 일하는 사람이 요통이 더 많은지 보고 싶습니다. 자료를 열어서 돌리면 되지요?" 돌리면 숫자는 나옵니다. 그러나 그 숫자가 **무엇을 뜻하는지**는 열기 전에 정한 것들 — 질문, 변수, 기준군, 보정변수, 제외 기준 — 이 정합니다. 이 Part 는 그것을 정하는 법과, 표 1·표 2 를 만들어 논문 한 편의 뼈대까지 가는 길입니다.

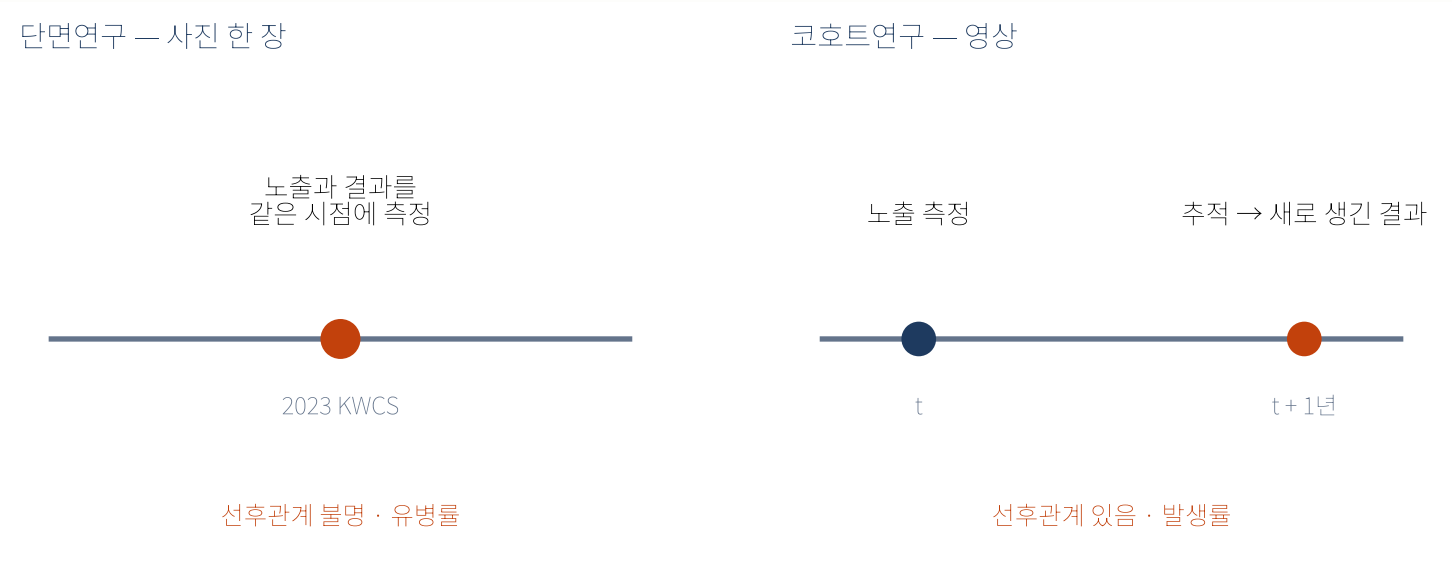
1. 무엇인가 — 1.1 역학 연구와 연관 분석

역학은 **인구 집단에서 질병의 분포와 결정요인**을 연구합니다. 어떤 사람들이, 왜, 어떤 병에 더 잘 걸리는가. 연관 분석은 그 중 "노출 A 가 있는 사람에게 결과 B 가 더 많은가"를 숫자(교차비, 상대위험도)로 답하는 것입니다.

1. 무엇인가 — 1.2 2차자료

남이 다른 목적으로 모아 둔 자료입니다. KWCS 는 근로환경 실태 파악을 위해 산업안전보건연구원이 3년마다 가구 방문 면접으로 모읍니다. 특수건강진단·작업환경측정·산재 자료도 2차자료입니다. 장점은 크고 싸고 빠르다는 것, 단점은 **내가 원하는 변수가 내가 원하는 방식으로 측정되어 있지 않다**는 것입니다. 자료가 설계를 정하게 두면 질문이 사후에 붙습니다.

1. 무엇인가 — 1.3 단면연구와 코호트연구



1. 무엇인가 — 1.3 단면연구와 코호트연구

	단면연구	코호트연구
측정	노출과 결과를 **같은 시점**에	노출 측정 뒤 **추적**해 새로 생긴 결과
지표	유병률, 교차비(OR)	발생률, 상대위험도(RR)·위험비(HR)
인과	선후관계 불명 — "관련이 있다"	선후관계 있음
비용	낮음 (KWCS 한 번)	높음
산업보건 예	KWCS 분석	특수건진 연계 코호트

KWCS 분석은 단면연구입니다. 논문에는 "장시간 노동이 요통을 **유발**한다"가 아니라 "장시간 노동은 요통과 **관련이 있었다**"고 씁니다.

1. 무엇인가 — 1.4 KWCS 의 특성과 한계

항목	내용
표본	50,195명, 15세 이상 취업자, 다단계 층화 집락추출
변수	439개 — 인구학·고용·근로시간·작업환경·심리사회·건강 결과
측정	**자기보고** — 요통은 진단이 아니라 "있다"는 응답
한계	단면 · 자기보고 · 건강근로자효과(심한 사람은 이미 퇴직) · 잔여 교란

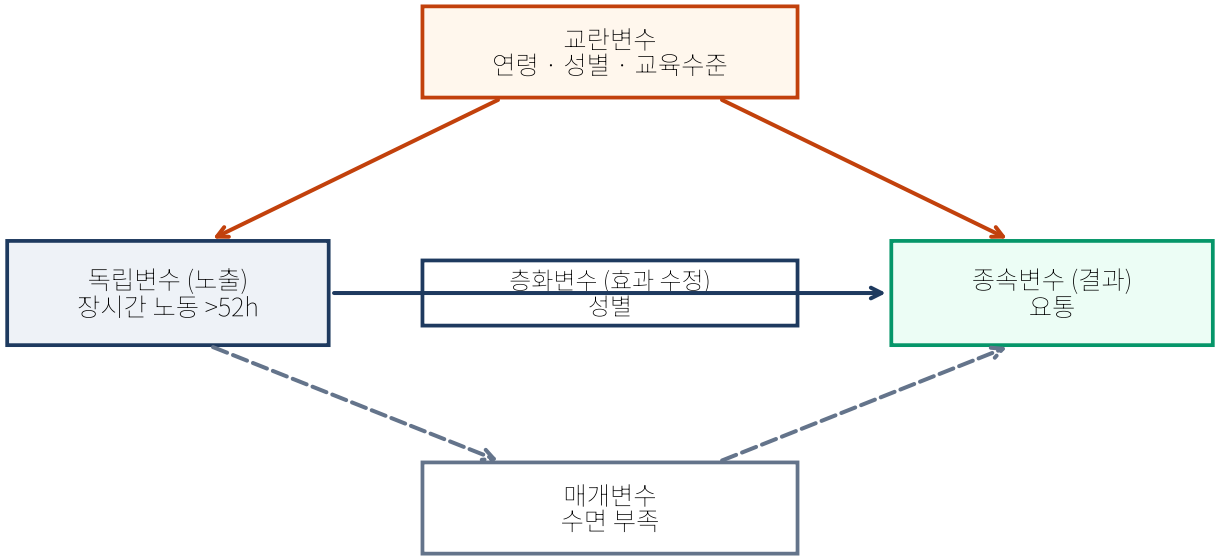
2. 기본 구조 — 2.1 연구 질문 — PECO

	뜻	예
P	대상	7차 KWCS 응답 취업자
E	노출	주 52시간 초과 근무
C	비교	정형 근로 35–40시간
O	결과	요통

"한국 취업자에서, 주 52시간 초과 근무는 정형 근로 대비 요통과 관련이 있는가." 추상 개념을 측정 변수로 바꾸는 것이 **조작화**입니다 — 장시간 노동 → `wtime_week > 52`, 요통 → `heal_prob3 == 1`.

2. 기본 구조 — 2.2 변수의 네 종류

교란은 보정하고, 매개는 보정하지 않는다 (보정하면 직접효과만 남아 총효과 과소)



층화변수는 보정이 아니라 나누어 본다 — 남성 OR 2.1 / 여성 OR 1.3 이면 효과 수정

2. 기본 구조 — 2.2 변수의 네 종류

종류	무엇	예	분석에서
독립변수 (노출)	관심 원인	근로시간 4단계 (<35 / 35–40 / 40–52 / >52)	기준군을 정한다 — 정형 근로
종속변수 (결과)	건강 결과	요통 있음/없음 (이분형)	로지스틱 회귀
보정변수 (교란)	노출과 결과 양쪽에 영향	연령·성별·교육수준 (+고용형태·산업·직종)	모형에 넣는다
층화변수	효과가 달라지는 집단	성별	나누어 본다 (효과 수정)

매개변수(장시간 노동 → 수면 부족 → 요통의 "수면 부족")는 보정하지 않습니다. 보정하면 직접효과만 남아 총효과가 작아집니다.

2. 기본 구조 — 2.3 산출물

	내용
그림 1	대상자 흐름 — 전체 50,195 → 결측 제외 → 연령 제외 → 최종 n
표 1	대상자 특성 — 독립변수 수준별 열, 보정변수 행, 결과 유병률
표 2	로지스틱 회귀 — Model 1(조) · 2(연령·성별) · 3(전체 보정) 의 OR (95% CI)
표 3·그림 2	민감도 분석 · 층화 · 용량–반응 (선택)
보고 지침	STROBE

3. 하는 법 단계별

단계	하는 일	규칙
① 질문	PECO · 가설(H_0/H_1) · 조작화	자료를 열기 전에
② 변수 찾기	변수 카탈로그(atoms)·주제별 가이드(molecules)에서 권장 변수와 코딩 확인	변수 속성을 추측하지 않는다 — 명세를 읽는다
③ 대상자 확정	포함·제외 기준, 단계별 인원 기록	이 기록이 그림 1
④ 결측	결측률 확인. <5% 완전제거, 5–20% 제거+민감도, >20% 변수 재검토	**비해당과 결측을 구분**
⑤ 재코딩	근로시간 4단계, 결과 0/1, 기준군 지정	기준군은 "정상" 또는 "가장 건전한" 집단
⑥ 기술통계	범주형 빈도·χ^2, 연속형 평균·t, 용량–반응 추세	

3. 하는 법 단계별

단계	하는 일	규칙
⑦ 표 1	독립변수 수준별 특성 (`tabf`)	가로합 100% 인지 세로합인지 확인
⑧ 표 2	로지스틱 Model 1→2→3 (`glm` , `oddsTabf`)	보정변수는 선행연구·DAG 로
⑨ 심화	민감도(결과 정의 변경·대상 제한·보정 추가) · 교호작용·층화 · 용량–반응	계획한 것만
⑩ 논문	그림 1 → 방법 → 결과 → 고찰(주요 결과·선행연구·기전·한계)	사실 목록에서 문장화

3. 하는 법 단계별 — 3.1 로지스틱 회귀와 OR

결과가 이분형이면 로지스틱 회귀를 씁니다. $\ln \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$, 교차비 $OR = e^{\beta_1}$. OR 1.82 (95% CI 1.55–2.14) 는 "기준군 대비 요통의 오즈가 1.82배, 신뢰구간이 1 을 포함하지 않으므로 유의". 결과가 흔하면(유병률 >10%) OR 은 RR 보다 큽니다 — "위험이 1.8배"라고 쓰지 않습니다.

3. 하는 법 단계별 — 3.2 다단계 보정

Model 1 조 → Model 2 연령·성별 → Model 3 전체 보정. 보정 후에도 OR > 1 이고 유의하면 "연령·성별·교육수준을 보정한 후에도 관련이 있었다"고 씁니다. 보정할수록 OR 이 조금 줄어드는 것이 보통이고, 크게 바뀌면 그 변수가 강한 교란입니다.

3. 하는 법 단계별 — 3.3 판단이 갈리는 지점

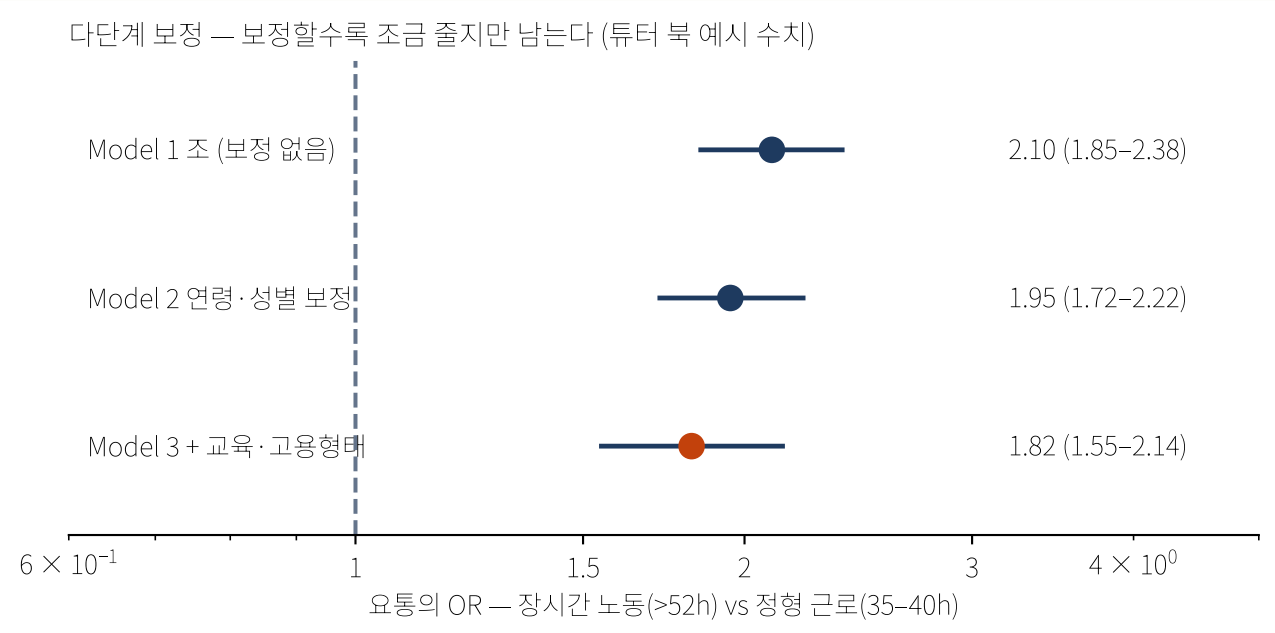
같은 자료에서 n 이 5만이 되기도 3천이 되기도 합니다. 교대근무 변수의 "해당 없음" 47,000건을 **결측**으로 버리는가, **비해당(주간근무)** 으로 두는가에 따라서입니다. 정답이 하나가 아닙니다. 그러나 **누가 그 판단을 했고 왜 했는지**는 방법 절에 적혀야 하고, 저자가 설명할 수 있어야 합니다.

4. 시범 — 끝까지 한 번

장시간 노동과 요통 (튜터 북의 예시 수치입니다 — 실제 값은 여러분이 돌려서 얻습니다).

단계	결과
그림 1	50,195 → 결측 제외 1,463 (근로시간 823·요통 640) → 미성년 523 제외 → **48,209**
표 1	근로시간 4단계 열 × 연령·성별·교육·고용형태 행. 장시간군이 남성·저학력·고연령
표 2	아래 그림

4. 시범 — 끝까지 한 번


다단계 보정 — 보정할수록 조금 줄지만 남는다 (튜터 북 예시 수치)
Model 1 조 (보정 없음)
2.10 (1.85–2.38)
Model 2 연령 · 성별 보정
1.95 (1.72–2.22)
Model 3 + 교육 · 고용형태
1.82 (1.55–2.14)
6 × 10−1
1
1.5
2
3
4 × 10^0
요통의 OR — 장시간 노동(>52h) vs 정형 근로(35–40h)

4. 시범 — 끝까지 한 번

방법 문장 — "요통은 '지난 12개월 요통 문제 있음' 응답으로 정의하였다. 근로시간은 정형 근로(35–40시간)를 기준으로 4단계로 나누었다. 연령·성별·교육수준·고용형태를 보정한 로지스틱 회귀로 교차비를 추정하였다. 주요 변수 결측 1,463명과 15–17세 523명을 제외하였다." 결과 문장 — "연령·성별·교육수준·고용형태를 보정한 후에도 장시간 노동 근로자는 정형 근로 근로자에 비해 요통의 교차비가 1.82배 높았다 (95% CI 1.55–2.14)."

5. 주의점과 흔한 오류

오류	무엇이 잘못인가	대신
"자료가 있으니 이 설계를"	질문이 사후에 붙음	PECO 먼저, "이 자료가 없어도 이 설계였을까"
"장시간 노동이 요통을 유발"	단면 — 선후관계 불명	"관련이 있었다"
결측과 비해당을 같이 버림	n 5만 → 3천	분포를 보고 판단, 방법 절에 기록
변수 속성 추측	결측 코드 8·9·999 를 값으로	카탈로그 명세를 읽는다
기준군을 안 정함	정렬순 첫 범주가 기준 → 해석 반전	기준군 명시 (정형 근로)
보정변수를 많이	매개(수면 부족)까지 → 총효과 소실	선행연구·DAG 로 교란만

5. 주의점과 흔한 오류

오류	무엇이 잘못인가	대신
표 1 의 % 방향 혼동	가로합 100 vs 세로합 100	어느 쪽인지 표에 명시
OR 을 RR 처럼	유병률 높으면 과대	"교차비" 로 쓰고, 필요하면 PR
유의한 결과를 찾아 변수 교체	사후 가설화	계획을 먼저 커밋
건강근로자효과 무시	심한 사람은 퇴직 → 과소	한계에 기술, 근속 제한 민감도
하위군을 결과 보고 정함	다중검정	미리 정한 층화만

5. 주의점과 흔한 오류 — AI 가 대신 판단하면

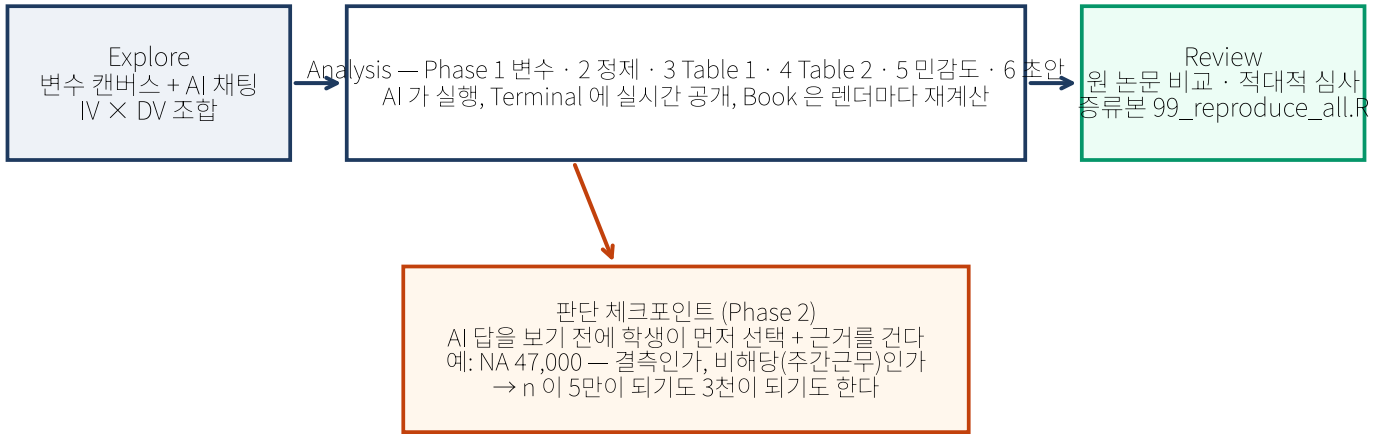
hitlpaper 실측: 학생 총 입력 102자("Phase 3 진행해줘")로 논문이 완성된 사례에서, 표본 47,220 vs 2,800 을 가르는 판단을 AI 가 혼자 했습니다. 그 저자는 자기 방법 절을 설명하지 못합니다. **논문은 나왔지만 저자는 남지 않은 것**입니다.

6. AI 와 함께 — hitlpaper 의 구조

> 자율성 수준 L0(쓰지 않음) · L1(보조 — 사람이 전부 검토) · L2(초안 — 전면 재검토) · L3(위임 — 표본 검수) · L4(자동 — 테스트 통과 시 채택). 정의는 Part 0 §2.2, 약어는 `00_공통/용어집.md`.

이 Part 의 AI 협업은 프롬프트가 아니라 **플랫폼의 구조**로 강제됩니다.

6. AI 와 함께 — hitlpaper 의 구조



AI 는 실행하고 사람은 판단한다 — 판단·검증은 전부 사람만 누르는 버튼이고 전부 로그로 남는다

6. AI 와 함께 — hitlpaper 의 구조

과업	수준	AI 가 하는 것	사람이 하는 것 · 틀리면 어떻게 알아채나
변수 탐색 (Explore)	L2	카탈로그에서 IV×DV 후보 추천	카탈로그 명세로 확인. 카탈로그에 없는 변수는 쓰지 않는다
Phase 1 변수 선정	L2	PECO → 변수·코딩 제안	기준군·절단점 결정
Phase 2 정제 — ✋ 판단 체크포인트	L1	분포를 보여 주고 **자기 선호를 밝히지 않은 채** 질문하고 멈춤	**AI 답을 보기 전에** 선택 + 근거(15~300자)를 건다. 제출 뒤 수정 불가. AI 가 판단을 공개하고 다르면 근거를 인용해 설명. **그 뒤 무엇을 할지는 학생이 정한다**
Phase 3–4 표 1·표 2	L3	`tabf` · `glm` · `oddsTabf` 실행	Terminal 에 raw output 실시간 공개 · Book 은 렌더마다 **재계산** — 채팅 숫자와 렌더 숫자가 다르면 그것이 환각 검출
Phase 5 민감도·층화	L3	계획한 분석 실행	계획에 없던 것은 "탐색적"
Phase 6 초안	L2	서론·고찰 초안. 인용은 placeholder 만	인용 실존은 사람의 bib 이 소유 (litwiki 검증완료 항목만)

6. AI 와 함께 — hitlpaper 의 구조

과업	수준	AI 가 하는 것	사람이 하는 것 · 틀리면 어떻게 알아채나
Review	L3	원 논문 비교 · **적대적 심사**(결함 찾기)	지적의 수용/기각 판정은 사람 — 판정률이 기록된다
증류본	L4	`99_reproduce_all.R` — 원자료에서 표까지 한 파일	새 R 세션에서 `stopifnot` 통과 = 제3자 재현
AI 사용 로그	—	자동 기록	15주 기술문은 기억이 아니라 이 로그에서

코드는 책(Book)에 **보이게** 들어갑니다 — `tabf(stratas=…, catVars=…)` 호출이 청크에 그대로. 숫자가 진짜인지(재계산)와 어떻게 구했는지(보이는 코드)는 다른 요구입니다.

7. 실습 — hitlpaper 로 논문 한 편의 뼈대

회차	하는 일	주는 것 (발판)
2주	Explore 에서 IV×DV 정하기 → PECO·가설·조작화를 **계획서로 커밋** → Phase 1–2 (✋ 체크포인트 통과, 근거 기록) → 그림 1	계획서 양식 · 튜터 북 · 변수 가이드 (molecules)
3주	Phase 3–4 (표 1·표 2, Model 1→3) → Phase 5 (계획한 민감도·층화 1개) → Book 렌더 → 적대적 심사 1회 → 증류본 재현	스크립트 골격은 플랫폼이 준다
제출	P1 결과 보고 I — 계획 커밋 · 체크포인트 근거와 AI 판정 · 그림 1 · 표 1·2 · 심사 판정 · 방법 사실 목록 + 결과 문장 · AI 사용 로그	루브릭 `30_실습.md`

8. 되돌아보기

- 오늘 잡은 것 — 체크포인트에서 내 예측과 AI 판정이 달랐는가. 달랐다면 누가 옳았고 어떻게 알았는가
- 내 방법 절의 문장 중 내가 결정하지 않은 것이 있는가
- 다음 Part 로 가져갈 것: **"AI 답을 보기 전에 내 답을 먼저 건다"** — 메타분석 추출에서는 손 추출을 AI 추출본보다 먼저 합니다

참고문헌

- von Elm et al. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. doi:10.1136/bmj.39386.490150.94
- Greenland et al. (1999). Causal diagrams for epidemiologic research. doi:10.1097/00001648-199901000-00008
- Schisterman et al. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. doi:10.1097/ede.0b013e3181a819a1
- Zhang et al. (1998). What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. doi:10.1001/jama.280.19.1690
- Li et al. (1999). The healthy worker effect in occupational epidemiology. doi:10.1093/occmed/49.4.225

(실존 확인: `reference/refs_verified.csv`)

가져갈 다섯 줄

1. 2차자료 연관 분석은 **이미 있는 자료**로 "노출이 있으면 결과가 더 많은가"를 묻는다. 질문이 자료보다 먼저다
2. KWCS 는 **단면·자기보고**다. "관련이 있었다"까지만 쓴다
3. 변수는 **독립·종속·보정·층화**로 나뉜다. 교란은 보정하고 매개는 보정하지 않으며 기준군을 명시한다
4. 산출물은 **그림 1 · 표 1 · 표 2(Model 1→3)** 다. 판단이 갈리는 지점(결측 vs 비해당)은 저자가 정하고 기록한다
5. hitlpaper 는 **AI 답을 보기 전에 내 답을 걸게** 만든다. 논문이 아니라 판단력이 남아야 한다