연구보고서

직업병 인과추론 가이드라인 및 통계분석법 개발 (2)
- 복합노출의 건강 영향평가 국문 가이드라인 개발 -

예신희, 이경은, 윤민주, 박동준, 마성원, 이영신, 이우주

산업재해예방
안전보건공단
산업안전보건연구원

요 약 문

● **연 구 기 간** 2022년 03월 ~ 2022년 11월
● **핵 심 단 어** causal inference, multiple exposure, g-formula, bayesian kernel machine regression
● **연구과제명** 직업병 인과추론 가이드라인 및 통계분석법 개발 (2)
- 복합노출의 건강 영향 평가 국문 가이드라인 개발 -

1. 연구 배경

- 작업환경에서 노출될 수 있는 여러 종류의 유해물질에 의한 근로자의 건강영향을 평가할 때, 건강근로자 편향 (healthy worker bias)과 같이 복합노출의 건강영향을 평가할 때 고려해야 할 특성들을 고려하지 않고 표준적인 회귀모형 (standard regression)만을 이용하여 분석한다면 노출 변수의 인과효과 추정치에 편향(bias)이 발생한다는 것이 잘 알려져 있음.
- g-formula는 건강근로자 효과에 의한 편향을 효과적으로 통제하기 위하여 치료-교란 요인 되먹임 (treatment-confounder feedback)을 반영하여 인과효과를 추론하는 방법임. 또한, g-formula는 두 가지 이상의 노출 (multiple exposures)을 동시에 고려할 수 있음. 하지만 역학 연구자들은 g-formula에 대한 개념과 기술적인 세부내용에 대한 이해 부족으로 g-formula의 사용에 어려움을 겪고 있으며, 현재 국내 산업보건 영역에서 많이 사용되고 있지 않음.
- 복합물질의 건강 영향 평가 방법은 (ⅰ) 복합노출 원인 물질 각각과 근로자의 건강 지표 사이의 관계가 비선형이며 비가산 관계일 수 있고, (ⅱ)

원인 물질들 사이의 교호작용이 건강 지표에 작용할 수 있으며, (iii) 원인 물질 사이의 연관성이 매우 클 수 있다는 점을 고려해야 함. 하버드 대학교의 Bobb JF, Valeri L 교수는 기존 kernel machine regression (KMR) 방법에 베이지안 접근법(bayesian approach)을 적용한 bayesian kernel machine regression(BKMR)을 개발함으로써 이러한 문제들을 해결하고자 하였음.

- 복합물질의 건강 영향 평가 방법 중 하나인 BKMR은 국외에서 널리 사용되고 있으며, 국내에서도 적용된 사례가 일부 있지만, 현재 국내 산업보건 영역에서 널리 사용되고 있지는 않음. 따라서, 복합물질의 건강 영향 평가 방법인 g-formula 또는 BKMR에 대한 장단점을 검토하고 복합물질과 관련된 국내 산업보건 자료에 각 방법을 적재적소에 쉽게 따라하여 적용할 수 있도록 가이드라인의 제공이 필요함. 특히 2021년 연구과제 (예신희 등, 2021)에서 작성된 g-formula를 활용한 인과추론에 초점을 맞춘 국문 가이드라인에 인과추론에서 사용하는 기본 용어를 정리한 내용을 추가하여, 산업보건 역학 연구자들이 g-formula 방법을 쉽게 이해할 수 있도록 돕고자 함.
- 이러한 국문 가이드라인을 통해 국내 산업보건 역학 연구자들이 두 방법을 근로자 종적 자료에 적용하고 역학연구에서 인과 관계를 도출할 수 있도록 하여 국내 산업보건 역학조사 및 역학연구 결과의 신뢰성을 높일 수 있도록 함. 나아가, 작업장 내 여러 복합물질에 대한 위험을 평가하고, 위험성이 높은 물질에 대해서는 제한 기준을 마련할 수 있는 과학적 근거를 제공하여, 작업장 내에서 발생하는 직업성 질환의 발생률과 사망률을 감소시켜 근로자의 건강을 보호하고자 함.

2. 주요 연구 결과

1) 근로자 종적 자료에 대한 가설 및 가설의 인과 그래프 작성

- ‘납과 카드뮴에 대한 장기적인 복합노출이 빈혈 발생의 위험을 높인다.’라는 가설과 ‘납과 크실렌에 대한 장기적인 복합노출이 빈혈 발생의 위험을 높인다.’라는 가설을 가설과 관련된 변수들 (빈혈 여부, 혈중 납 농도, 혈중 카드뮴 농도, 요중 메틸마뇨산 농도, 나이, 성별, 사업장 규모, 특수건강진단 수행 연도, 음주 유무, 흡연력 유무, 체질량 지수, 검진결과에 대한 의사의 판정 결과 그리고 그로 인한 사후관리조치 결과)을 활용하여 인과 그래프로 표현하였음.
- 실무지침 내 노출 기준치에 대해 납의 경우, 직업병 요관찰자에 해당하는 혈중 납 농도 기준치는 30 μg/dL이며, 직업병 유소견자에 해당하는 혈중 납 농도 기준치는 40 μg/dL이다. 카드뮴의 경우, 5 μg/L이며, 크실렌의 경우 요중 메틸마뇨산 농도 기준치는 1.5 g/g creatinine임.
- 따라서 납과 카드뮴 그리고 납과 크실렌에 대한 복합노출에 따른 빈혈 발생률을 산출하기 위해 혈중 납 농도 5, 10, 15, 20, 25, 30, 35, 40 (단위: μg/dL), 혈중 카드뮴 농도 1, 2, 3, 4, 5 (단위: μg/L) 그리고 요중 메틸마뇨산 농도 0.25, 0.50, 0.75, 1.00, 1.25, 1.50 (단위: mg/L)를 개입 노출량 (hypothetical intervention)으로 설정하여 각 조합에 따른 특수건강검진 대상 근로자의 빈혈 발생률을 산출하고, 등고선 그림 (contour plot)을 통해 시각적으로 그 추이를 살펴보고자 함.

2) 특수건강진단 자료 분석

- 여러 혈중 납 농도와 혈중 카드뮴 농도의 조합에 따른 빈혈의 누적 발생률을, 일반인구 집단의 노출 수준에 해당하는 혈중 납 농도 (1.6 μg/dL)와 혈중 카드뮴 농도 (0.9187 μg/L)일 때의 빈혈에 대한 누적 발생률로

나누어 구한 위험 비를 기술하였음. 혈중 카드뮴 농도뿐만 아니라 요중 메틸마뇨산 농도에 대해서도 분석하여 여러 혈중 납 농도와 요중 메틸마뇨산 농도의 조합에 따른 빈혈의 누적 발생률을, 일반인구 집단에 해당하는 혈중 납 농도 (1.6 μg/dL)와 요중 메틸마뇨산 농도 (0.234 mg/L) 일 때의 빈혈에 대한 누적 발생률로 나누어 위험 비를 기술하였음.

- 혈중 카드뮴 농도가 고정되어있는 경우, 혈중 납 농도가 증가함에 따라 빈혈의 누적 발생률의 위험 비가 증가하는 것을 확인할 수 있었으며, 마찬가지로 혈중 납 농도가 고정되어있을 때, 혈중 카드뮴 농도가 증가함에 따라 빈혈의 발생률의 위험 비가 증가하는 것을 확인할 수 있음.
- 요중 메틸마뇨산 농도가 고정되어있는 경우, 혈중 납 농도가 증가함에 따라 빈혈의 누적 발생률의 위험 비가 증가하는 것을 확인할 수 있었음. 하지만 혈중 납 농도가 고정되어있을 경우, 요중 메틸마뇨산이 증가함에 따라 빈혈의 누적 발생률의 위험 비가 소폭 감소하는 것을 확인하였음
- 혈중 납 농도와 혈중 카드뮴 농도에 따른 빈혈의 누적 발생률의 위험 비 그리고 혈중 납 농도와 요중 메틸마뇨산 농도에 따른 빈혈의 누적 발생률을 시각적으로 확인하기 위한 등고선 그림을 그렸으며, 이 등고선 그림을 통해 혈중 납 농도와 혈중 카드뮴 농도 그리고 혈중 납 농도와 요중 메틸마뇨산 농도 각 조합에 따른 위험 비를 직관적으로 확인할 수 있음.
- 자료의 현재 노출 패턴에 따른 비모수적 생존 곡선을 비슷하게 예측하는 모형을 선택하여 납과 카드뮴의 복합 노출에 대한 건강 영향을 평가하기 위해 신뢰구간을 추정하였으며, 납과 카드뮴에 대해 저농도(예; 납: 5 μg/dL, 카드뮴: 1 μg/L)에서 일부 유의하지 않은 결과가 나타났지만 고농도(예; 납: 20 μg/dL, 카드뮴: 3 μg/L)에서는 유의한 결과가 나타났음. 또한, 혈중 납 농도와 요중 메틸마뇨산 농도에 대해서는 혈중 납 농도가 20 μg/dL 인 경우부터 유의한 결과를 나타내었음.

3) g-formula와 BKMR에 대한 국문 가이드라인 작성

- **g-formula의 배경:** 근로자 종적 자료와 같은 반복측정 자료에서 인과 그래프의 시간에 따라 변하는 치료-교란 요인 되먹임 구조(treatment-confounder feedback)를 반영하여 결과 변수에 대한 노출 전략의 인과효과의 크기를 추론할 수 있도록 하버드 보건대학원의 Robins JM가 개발한 방법임.
- **g-formula의 이론:** 치료-교란 요인 되먹임이 존재하는 인과 그래프에서 g-formula가 올바른 추정치를 제공하기 위해 필요한 가정들을 소개함. 또한, 예제를 사용하여 이러한 자료에서 전통적인 통계 분석 방법을 적용하였을 때 나타나는 문제점을 소개하고 설명함으로써 g-formula의 필요성을 기술하였음.
- **g-formula에 대한 R 패키지 소개:** Robins JM의 g-formula를 통계 분석 프로그래밍 언어 R로 구현한 R 패키지인 'gfoRmula'를 소개하고, R 패키지에 내장되어있는 함수와 그 사용법에 대해 예제를 통해서 설명하였음. 또한, 2021년 연구과제 (예신희 등, 2021)에서 작성한 내용을 산업보건 역학 연구자가 보다 용이하게 이해할 수 있도록 내용 및 표현을 수정하고, g-formula를 사용하여 복합물질을 다룰 수 있도록 패키지에 대한 설명을 추가함.
- **BKMR의 배경:** 복합물질의 건강 영향 평가를 할 때, 평가 방법은 복합물질에 대한 노출과 건강영향이 복잡한 비선형 또는 비가산적 관계를 가질 수 있고, 건강영향과 복합물질 사이의 교호작용이 허락되어야 하며, 복합물질의 구성 성분 사이의 높은 상관성을 모형이 반영하여야 함. 이러한 3가지 고려해야할 점을 반영하여 복합물질에 대한 건강 영향 평가를 하기 위해 Bobb JF는 bayesian kernel machine regression (BKMR)을 개발하였음.
- **BKMR의 이론:** BKMR은 kernel machine regression (KMR)에 기초한 방법으로 여러 유해물질이 존재하는 상황에서 유해물질의 비선형성,

비가법적 관계 그리고 유해물질 사이의 높은 상관성을 각각 kernel 행렬을 이용한 혼합 모형 그리고 베이지안 관점의 변수 선택법을 사용하여 모형에 반영함.

- **BKMR에 대한 R 패키지 소개:** Bobb JF의 BKMR를 통계 분석 프로그래밍 언어 R로 구현한 R 패키지 'bkmr'를 소개하고, 패키지에 내장되어있는 함수와 그 사용법을 예제를 통해 설명하였음.

4) 2021년 작성된『직업병 인과추론 가이드라인: g-formula 국문 가이드라인』보완

- 산업보건 역학 연구자가 2021년 작성된『직업병 인과추론 가이드라인: g-formula 국문 가이드라인』을 읽고 g-formula에 대한 이해를 보다 용이하게 할 수 있도록, 2021년 작성된 가이드라인에 대한 내, 외부 직업환경의학 전문의의 자문 의견을 받아 산업보건 역학 연구자가 g-formula를 이해할 때 어려움을 호소하는 의견을 수렴하고, 그러한 점을 해소할 수 있도록 용어 부록 등을 보고서에 반영하였음.

5) g-formula와 BKMR에 대한 검토 및 장단점 평가

- **g-formula와 BKMR의 공통점:**
  - 근로자가 복합물질에 노출되었을 때, g-formula와 BKMR 모두 사용이 가능함.
- **g-formula의 장점:**
  - 치료-교란 요인 되먹임과 경쟁사건 (competing event)의 존재를 반영할 수 있음.
  - 산업보건 역학연구에서 발생할 수 있는 건강근로자 편향을 효과적으로 통제할 수 있음.
  - Marginal causal effect에 해당하는 위험도의 차이, 위험도의 비, 위

험도에 대한 오즈 비를 모두 제공하기 때문에 인과효과를 다양하게 해석할 수 있음.

- marginal structural model, g-estimation 등 다른 인과추론 방법론을 통해 일관된 결과가 나오는지 확인함으로써 분석 결과의 신뢰성을 일부 확인할 수 있음.

▪ g-formula의 단점:

- g-formula는 모수적 모형 (parametric model)을 사용하기 때문에 비모수적 모형 (nonparametric model)을 사용하는 BKMR과 같이 다양한 고차원 항 또는 교호작용 항을 모형에 반영하는 것에는 한계가 있음.
- 인과효과를 시각적으로 표현하기 위해서는 별도의 시각화 과정이 요구됨.
- 유해물질 사이의 교호작용의 효과를 다양한 노출 수준에서 확인하기 어려움(모든 시점에서 일정한 노출수준 (static exposure level)을 가진 복합노출의 경우에는 교호작용을 확인할 수 있음).

▪ BKMR의 장점:

- 유해물질의 수가 많아도 복합노출에 의한 건강영향 평가가 가능함.
- 유해물질 사이의 교호작용 효과를 시각적으로 확인할 수 있음.
- 여러 유해물질과 결과 변수 사이의 관계를 비모수적 함수를 통해 기술하기 때문에 다양한 고차원 항 또는 교호작용 항을 고려하여 모형화할 수 있음.

▪ BKMR의 단점:

- 치료-교란 요인 되먹임과 같은 건강근로자 편향을 반영하기 어려움.
- 작은 크기의 자료에서도 분석 속도가 매우 느림.
- 선행연구에 따르면, 자료의 형태가 복잡해지면 복합노출의 건강영향을 평가하는 여러 통계 방법론들의 분석결과 값들이 일관되지 않음.

6) 후속 연구계획 수립

- 연구진 회의를 통해 2023년에 진행할 과제의 목표인 g-formula와 BKMR의 통계분석법 개선안을 다음과 같이 논의하였음.
- G-formula의 통계분석법 개선안은 다음과 같음.
  - 용량-반응 곡선과 교호작용을 표현하는 시각화 코드 개발
  - 분석 결과의 안정성 평가 방법 개발
- BKMR의 통계분석법 개선안은 다음과 같음.
  - 분석 시간 단축 방법 개발
  - 로지스틱 회귀 모델로 확장 (현재는 프로빗 모델만 가능함)
  - 반복측정 자료에서 기울기에 랜덤 효과를 적용 (현재는 절편에만 랜덤 효과를 적용할 수 있음)
  - 노출 변수의 수와 튜닝 파라미터의 올바른 활용을 위한 분석 방법 가이드라인 작성

3. 연구 활용방안

- 건강근로자 편향을 통제할 수 있는 인과추론 통계 방법에 대한 이해를 높여 특수건강진단 자료와 같은 근로자 종적 자료 통계 분석의 질을 한 층 개선 및 발전시킬 수 있음.
- 국내 다양한 산업보건 역학연구에서 국문 가이드라인을 참고하여 작업장에서 발생하는 여러 복합노출에 대한 위험을 평가하고, 위험성이 높은 물질에 대해서는 제한 기준을 마련할 수 있는 과학적 근거를 제공함.
- 다양한 연구 결과를 통해 위험한 복합물질의 노출량을 제한하여 근로자가 안전한 작업장에서 근무할 수 있도록 작업환경을 개선할 수 있게 하고, 직업성 질환의 발생률 및 사망률을 감소시켜 근로자의 건강을 보호하고자 함.

- 본 국문 가이드라인을 통하여 산업보건 연구를 진행하는 연구자들이 복합노출에 대한 건강 영향 평가 방법인 g-formula와 BKMR를 자료에 적용하는데 필요한 시간을 단축시키며, 올바르게 사용할 수 있도록 하여 근로자의 사망 또는 건강 지표에 대한 복합노출의 효과를 올바르게 추정, 산출할 수 있도록 함.
- 복합노출을 다루는 g-formula와 BKMR에 대한 개념과 기술적인 세부 내용에 대해 산업보건 역학 연구자들의 이해도를 높여 국내 산업보건 역학 연구의 발전을 기대하며, 세계 수준의 학술논문 발표를 통해 국내 산업보건 분야의 학술 발전에 기여함.

4. 연락처

- 연구책임자: 산업안전보건연구원 중부권역학조사팀 팀장 예신희
  - ☎ 032) 510. 0754
  - E-mail shinheeye@kosha.or.kr

목 차

목 차

표 목차

그림목차

그림목차

Ⅰ. 서 론

Ⅰ. 서 론

1. 연구 목적 및 필요성

- 작업장 내 산재한 복합물질에 대해 근로자에게 노출된 양을 시간에 따라 변하는 노출 변수로 본다면 근로자들의 고용상태는 과거 노출 변수에 의해 영향을 받으면서 현재 노출 변수와 건강 상태에 영향을 주는 시간에 따라 바뀌는 교란 요인의 역할을 하게 되며, 이러한 구조를 인과 그래프가 가지고 있을 때, 치료-교란 요인 되먹임 (treatment-confounder feedback)이 존재한다고 함. 따라서 이러한 치료-교란 요인 되먹임을 고려하지 않고 표준적인 회귀모형 (standard regression)만을 이용하여 분석한다면 노출 변수의 인과효과 추정치에 편향 (bias)이 발생한다는 것이 잘 알려져 있음.
- 하버드대학교의 Robins JM 교수는 시간에 따라 변하는 노출 변수가 있는 복잡한 관찰 연구 자료로부터 이러한 치료-교란 요인 되먹임의 존재를 반영하여 인과효과를 추론하는 방법인 g-method를 개발하였음. 이는 건강근로자 생존 편향 (healthy worker survivor bias)과 같은 종적 산업보건 역학연구에서 발생할 수 있는 선택 편향(selection bias)을 효과적으로 통제할 수 있는 방법임. 하지만 역학 연구자들은 g-method에 대한 개념과 기술적인 세부내용에 대한 이해 부족으로 g-method 사용에 어려움을 겪고 있음. 국내 산업보건 역학연구에서도 g-method 사용의 필요성은 남정모 등 (2002) 및 이경무 등 (2011)이 언급한 바 있으나, 현재 국내 산업보건 영역에서 많이 사용되고 있지 않음.
- Robins JM 교수가 개발한 g-method에는 g-formula, marginal structural model (MSM) 그리고 g-estimation이 있음. g-formula (g-computation이라 불리기도 함)와 비교하여 MSM의 경우, inverse

probability weighting의 계산과정에서 positivity 가정이 위반되거나 계산된 weight의 극단 값으로 인한 추정치의 불확실성이 클 수 있으며, g-estimation의 경우 모형 지정의 유연성 (flexibility of model specification)을 가지고 있지만, 개념 자체가 복잡하여 연구자들에게 가장 친숙하지 않은 방법이므로 g-method 중 g-formula를 산업보건 역학연구에서 먼저 고려해야 할 필요성이 있다고 생각됨. g-formula는 복합물질 노출 자료를 다룰 수 있는 인과추론 방법론으로 두 가지 이상의 다중 노출 (multiple exposure)을 동시에 고려할 수 있음.

- 또한, 복합물질 노출 연구에서 사용하는 방법에는 Elasticnet regression, Partial least square, Weighted quantile sum regression (WQS) 등이 있음. 근로자의 건강상태에 대한 복합물질의 효과를 추정하기 위해 WQS 방법에 g-formula를 적용한 사례가 있음. 이는 g-formula가 역학연구에서 의미 있게 활용될 수 있는 가능성을 구체적으로 보여주었음.
- 복합물질의 건강 영향 평가 방법은 (ⅰ) 복합노출 원인 물질 각각과 근로자의 건강 지표 사이의 관계가 비선형이며 비가산 관계일 수 있고, (ⅱ) 원인 물질들 사이의 상호작용이 건강 지표에 작용할 수 있으며, (ⅲ) 원인 물질 사이의 연관성이 매우 클 수 있다는 점을 고려해야 함. 하버드 대학교의 Bobb JF, Valeri L 교수는 기존 kernel machine regression (KMR) 방법을 도입하여 이러한 문제들을 해결하고자 하였음(Bobb JF et al., 2015).
- 복합물질을 다루는 BKMR은 국외에서 사용되고 있으며, 국내에서도 적용된 사례가 일부 있지만, 현재 국내 산업보건 영역에서 널리 사용되고 있지는 않음. 따라서 복합물질의 건강 영향 평가 방법인 g-formula 또는 BKMR를 국내 산업보건 자료에 쉽게 따라하여 적용할 수 있도록 가이드라인의 제공이 필요함. 특히 2021년 연구과제 (예신희 등, 2021)에서 작성한 g-formula 가이드라인에 대하여 외부 직업환경의학 전문의

의 자문 의견을 받아 산업보건 역학 연구자가 g-formula를 이해할 때 어려움을 느끼는 내용에 대해 의견을 수렴하고, 그러한 점을 해소할 수 있도록 가이드라인을 수정하고 보완하여, 연구자들이 g-formula 방법을 쉽게 이해할 수 있도록 돕고자 함.

- 또한, 복합노출에 대한 건강 영향 평가 방법인 g-formula와 BKMR에 대한 장단점을 검토하고자 함. g-formula의 경우, 다중 노출이 있는 경우 각 노출 변수에 대한 모형을 적합할 수 있으며, 다중 노출 사이의 연관성을 허락할 수 있음.
- 이러한 국문 가이드라인을 통해 국내 산업보건 역학 연구자들이 두 방법을 근로자 종적 자료에 적용하여 역학연구에서 인과 관계를 도출할 수 있도록 함. 또한, 국내 산업보건 역학조사 및 역학연구 결과의 신뢰성을 높일 수 있도록 함. 나아가, 작업장 내 여러 복합물질에 대한 위험을 평가하고, 위험성이 높은 물질에 대해서는 제한 기준을 마련하여 작업장 내에서 발생하는 질병에 대한 누적 발생률 또는 사망률 감소시켜 국가사업의 사회적 기여도의 상승에 이바지하고자 함.

2. 관련 선행 연구에 대한 분석

- Taubman SL 등 (2009)의 연구는 Nurses' Health Study 자료에 등록된 78,746명의 2년마다 실시하는 간호사 건강검진 기록을 사용하여 g-formula를 통해 여러 개입에 대한 관상동맥질환 (coronary heart disease)의 1982년부터 2002년까지 follow-up된 사망률을 계산하였음. 관찰연구 자료로부터 직접 구한 20년 follow-up된 누적 발생률은 3.50%였으며, 5가지 활동 (금연, 매일 적어도 30분씩 운동, 매일 5g 이상의 알코올 섭취, diet score의 상위 40% 이내로 유지, BMI 지수 25 이하로 유지)을 모두 실행하였을 때에는 누적 발생률이 1.89% (신뢰구간: 1.46에서 2.41까지)으로 낮아지는 것을 확인하였음. 이 연구에서는

한 가지 개입뿐만 아니라 2가지 이상의 개입을 동시에 시행하여 개입의 조합에 대한 효과를 측정하였음.

- Keil AP 등 (2017)은 건강 근로자 생존 편향을 보정하기 위해 근로자의 고용상태 (employment status)를 인과적 방향성 비순환 그래프에 포함한 후, g-formula를 이용하여 8,014명의 백인 남성에 대하여 구리 용광로에서 공기 중 비소 흡입량에 따라 모든 질병, 심장병, 폐암으로 인한 초과 사망률이 나이에 따라 어떻게 변화하는지 연구하였음. 이때 노출량에 대한 개입 (intervention)은 노출을 시키지 않는 개입, 개입하지 않음, 심한 노출을 시키는 개입 총 3가지로 분류하여 적용하였으며, 각 개입을 시행하였을 때 발생하는 초과 사망률을 측정하였음.
- Valeri L 등 (2017)은 중금속 복합물질에 대한 임신 중 노출이 출생 후 20-40개월 영유아의 신경 발달 결과에 영향을 미치는지 파악하기 위해 어머니-아이 825쌍을 대상으로 BKMR을 적용하여 연구를 진행하였음. 이때, 영유아의 신경 발달 정도를 측정하기 위해 인지 발달 점수 (cognitive development score)와 언어 개발 종합 점수 (language development composite score)를 사용하였으며, 중금속 복합물질로는 비소, 마그네슘, 납을 고려하였음.
- 이슬비 등 (2019)은 어머니-아이 302쌍을 대상으로 중금속에 해당하는 납, 수은, 카드뮴과 대기오염물질에 해당하는 NO2, PM10, PM2.5 그리고 비스페놀 A, 프탈레이트 대사체 MEHHP, MEOHP, MnBP 3종을 포함하여 총 10가지의 환경유해물질에 대한 복합노출이 출생 후 6개월이 지난 영유아의 아토피 피부염 발생에 미치는 영향을 확인하고자 하였음. BKMR을 사용하여, 임신 말기에서 복합물질에 대한 누적 노출 양이 증가할 때 영유아의 아토피 피부염 발생의 위험이 증가한다고 보고하였음.
- Neophytou AM 등 (2019)은 광부 연구 코호트에서 디젤 배기가스와 호흡성 광산 분진의 노출 수준을 제한하는 가상의 시나리오 하에서 허혈성 심장질환 사망률의 반사실적 결과 (counterfactual outcome) 위험

을 평가하였음. 대기오염에 대한 일반적인 인구집단 대상 연구에서 미세먼지(특히, 디젤 배기가스 배출물질)가 심혈관질환의 잠재적 위험 요소임을 시사하지만, 정량적 노출 측정을 사용한 직업 코호트에서의 직접적인 근거는 제한적임. 이 연구는 디젤 장비가 각 시설에 도입된 후 8개의 비금속, 비석탄 광산에 고용된 10,778명의 남성 광부 데이터를 분석하였고, 1948년부터 1997년까지 추적하였으며, 이 중 297명이 허혈성 심장질환으로 인한 사망하였음. 연구자들은 호흡성 원소 탄소 (디젤 배기 물질에 대한 대체물)와 호흡성 분진에 대해 개별적으로 그리고 공동으로 다양한 제한 기준을 가진 가상 시나리오 하에서 위험을 평가하기 위해 g-formula를 적용하였음. 원소 탄소와 호흡성 분진에 대한 노출이 제거되는 가상 시나리오 하에서, 80세에서 관찰된 위험과 누적 허혈성 심장질환 위험을 비교한 risk ratio는 0.79였음. Risk difference는 -3.0%였음. 비금속 광부 코호트 자료를 기반으로 하는 이 연구 결과는 디젤 배기 물질 및 호흡성 분진에 대한 노출을 제거하기 위한 개입이 허혈성 심장질환 사망 위험을 감소시킬 것이라는 가설과 일치하였음.

- 예신희 등 (2020)은 특수건강진단 자료와 한국 국민건강영양조사 자료, 미국 국민건강영양조사 자료를 BKMR 방법으로 분석하여, 납과 카드뮴의 복합노출이 간기능검사 (AST, ALT, GGT) 결과 수치에 미치는 영향을 평가하였음. 그 결과 모든 모형과 세 가지 종류의 자료 (특수건강진단, KNHANES (한국 국민건강영양조사), NHANES (미국 국민건강영양조사))에서 납과 카드뮴 복합노출에 의한 건강영향 분석결과가 일관되게 나온 것은 GGT (감마글루타밀전이효소) 검사 결과였음. 혈중 납과 혈중 카드뮴은 각각 GGT와 연관성을 보였으며, 복합 노출도 GGT 증가와 연관성이 있었음. 또한 납과 카드뮴은 물질의 노출 수준이 높아질수록 나머지 물질이 GGT를 더 크게 증가시키는 교호작용이 모든 분석에서 일관되게 확인되었고, 혈중 납과 혈중 카드뮴 농도 값이 함께 높아질수록 GGT가 높아지는 복합 노출의 건강영향도 일관되게 관찰되었음.

3. 연구 목표

- g-formula와 BKMR을 근로자 종적 자료에 적용하여 작업장 내 납, 카드뮴, 크실렌 등에 대한 복합물질이 근로자의 빈혈 발생에 얼마나 영향을 미치는지 분석하고, 평가하고자 함.
- 본 연구를 통해 국내 산업보건 역학조사 및 역학연구에서 활용할 수 있는 g-formula와 BKMR에 대한 국문 가이드라인을 제공함으로써 두 가지 이상의 시간에 따라 변하는 노출 변수를 포함하는 복합노출을 가지는 근로자 종적 자료의 가설을 분석하는데 적합한 통계 분석 방법론을 보급하고자 함.
- 2021년에 작성된 『직업병 인과추론 가이드라인: g-formula 국문 가이드라인』에 대하여, 외부 직업환경의학 전문의의 자문 의견을 받아 산업보건 역학 연구자가 g-formula를 이해할 때 어려움을 호소하는 의견을 수렴하고, 그러한 점을 해소할 수 있도록 가이드라인을 보완함.
- 복합물질에 대한 노출의 건강 영향 평가 방법인 g-formula와 BKMR을 검토하고 두 방법의 장단점을 평가하여 국내 산업보건 역학 연구자가 근로자의 종적 자료의 특성에 알맞게 방법을 선택하여 쓸 수 있게 하고자 함.
- 근로자 종적 자료를 분석하면서 나타나는 g-formula와 BKMR의 한계점 등을 파악하고, 이를 개선하기 위한 3차 연도 연구계획을 수립하고자 함.

Ⅱ. 연구 방법

Ⅱ. 연구 방법

1. 연구 내용 및 범위

1) 직업병 인과추론 가이드라인 및 통계분석법 개발 (2) (자체)

(1) 복합노출에 대한 가설 선정 및 인과 그래프 초안 작성.

(2) 분석용 특수건강진단 자료 가공.

(3) 복합노출에 대한 인과추론 통계방법 활용 국문 가이드라인 초안 작성.

(4) 2021년에 작성한 인과추론 통계방법 활용 국문 가이드라인 (예신희 등, 2021) 수정.

2) 산업보건 역학연구 수행을 위한 인과추론 통계방법 검토 및 적용에 대한 연구 (2) (위탁)

(1) 복합노출에 대한 가설 및 인과 그래프 검토.

(2) g-formula와 BKMR로 근로자 종적 자료 (예: 특수건강진단자료) 분석

(3) 복합노출의 건강 영향 평가 국문 가이드라인 검토 및 수정.

(4) 2021년 작성 인과추론 통계 방법 활용 국문 가이드라인 재검토 및 수정.

(5) g-formula와 BKMR의 장단점을 검토하여, 근로자 종적 자료 (예: 특수건강진단 자료) 특성에 맞는 복합노출 건강 영향평가 방법을 제안하고 통계 방법론의 개선점을 파악.

(6) 3차 연도 연구계획 수립.

2. 연구 방법

1) 직업병 인과추론 가이드라인 및 통계분석법 개발 (2) (자체)

(1) 복합노출에 의한 건강영향을 평가하기 위한 가설을 연구진 회의를 통해 인과 그래프로 표현함.

- 복합노출에 의한 건강영향을 평가하기 위한 가설을 다음과 같이 인과 그래프로 작성함.
- 근로자 종적 자료에서 해당 가설을 검증하기 위해 성별, 연령 등 근로자의 기저변수를 포함하여 관심 있는 유해물질과 주요 건강 지표 사이의 인과적 방향성 비순환 그래프 (directed acyclic graph; DAG)를 그리고, 변수 사이의 관계들에 대한 문헌을 검토하여 상정한 인과적 방향성 비순환 그래프 및 가설의 타당성을 확보함.
- 특히 교란 요인 (confounder)에 의한 편향, 선택 편향, 측정 편향 (measurement bias)과 같이 산업 보건 역학연구에서 발생 가능한 편향 문제를 산업보건 역학 연구자들이 직관적으로 이해할 수 있도록 인과 그래프의 이론을 통해 시각화하여 표현함.

(2) 특수건강진단 자료를 개인식별정보를 제외한 연구 분석용 자료로 가공하여 부분위탁 용역기관에 연구 분석용 자료 제공함.

(3) g-formula와 BKMR 활용에 대하여 산업보건 역학연구자의 접근이 용이하도록 통계 분석 코드 사용 방법을 국문 가이드라인으로 구체적으로 초안을 작성함.

- 활용자료원 1: Bobb JF, Valeri L, Claus Henn B, Christiani DC, Wright RO, Mazumdar M, Godleski JJ, Coull BA. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics. 2015 Jul;16(3):493-508.

- 활용자료원 2: Bobb JF, Claus Henn B, Valeri L, Coull BA. Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. Environ Health. 2018 Aug 20;17(1):67.
- 활용자료원 3: https://jenfb.github.io/bkmr/overview.html
- 활용자료원 4: https://jenfb.github.io/bkmr/ProbitEx.html
- 활용자료원 5: McGrath S, Lin V, Zhang Z, Petito LC, Logan RW, Hernán MA, Young JG. gfoRmula: An R Package for Estimating the Effects of Sustained Treatment Strategies via the Parametric g-formula. Patterns (NY). 2020 Jun 12;1(3):100008.

(4) 2021년에 작성한 인과추론 통계방법 활용 국문 가이드라인 수정

- 활용자료원: Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC, 2020.
- 인과추론에서 활용되는 기본 용어에 대한 설명을 추가하였음.
- Time-dependent Cox model에서 발생할 수 있는 편향을 인과 그래프로 표현하였음.
- Time-dependent Cox model 분석 결과와 g-formula 분석 결과를 비교하여 g-formula의 장점을 구체적으로 설명하였음.

2) 산업보건 역학연구 수행을 위한 인과추론 통계방법 검토 및 적용에 대한 연구 (2) (위탁)

(1) 근로자 종적 자료에 대한 가설 및 가설의 인과 그래프 검토

- 작업환경에서 근로자들이 노출된 복합물질 (납, 카드뮴, 크실렌) 이 빈혈 발생에 미치는 영향을 알아보기 위한 가설과 가설의 인과 그래프를

연구진 회의를 통해 검토함.

- 가설과 인과 그래프를 검토한 후, g-formula, BKMR에 사용할 산업보건 종적 자료의 전처리를 진행함 (특수건강진단 자료는 산업안전보건연구원에서 분석용으로 제공하는 자료를 기본으로 함).

(2) g-formula와 BKMR을 사용한 산업보건 종적 자료 분석

- 시간에 따라 변화하는 노출 변수와 시간에 따라 변하는 교란 변수의 구조를 g-formula에 반영하여 산업보건 종적 자료를 분석함.
- g-formula에서 가정하고 있는 식별조건 (identifiability)을 명확히 하고 통계적 인과추론이 이루어지는 과정을 국내 산업 보건 역학 연구자가 쉽게 이해할 수 있도록 하였음. 결과 변수, 노출 변수 그리고 교란요인에 대한 모형 설정에서 인과성 추론에 필요한 가정조건에 큰 위반사항이 없는지 확인하면서 분석을 진행함.
- 복합물질의 건강영향 평가를 위해 널리 사용되는 BKMR의 경우, 분석 시 시간이 오래 걸릴 뿐 아니라 산업 보건 역학 연구에서 발생하는 건강근로자 생존 편향을 반영할 수 없어 분석에 사용하지 않음.

(3) g-formula와 BKMR에 대한 국문 가이드라인 검토 및 수정

- 산업안전보건연구원에서 제공하는 국문 가이드라인 초안을 기준으로 검토 및 수정 과정을 거쳐, 복합물질의 건강 영향 평가 방법인 g-formula와 BKMR의 개념과 작동하는 알고리즘에 대한 기술적인 내용을 산업보건 역학 연구자들이 쉽게 이해할 수 있도록 상세하게 정리함.
- g-formula와 BKMR를 자료에 적합할 수 있도록 관련 프로그램의 사용법을 구체적으로 작성하여 산업보건 역학 연구자들이 쉽게 활용할 수 있도록 함.

- 2022년 가이드라인에 대해 외부 직업환경의학 전문의의 자문 의견을 받아, 산업보건 역학 연구자가 가이드라인을 읽고 복합노출의 건강 영향평가 통계 방법론을 이해할 때 어려움을 호소하는 의견을 수렴하고, 사용자 친화적으로 가이드라인을 수정함.

(4) 2021년 작성 인과추론 통계 방법 활용 국문 가이드라인 재검토 및 수정

- 2021년 가이드라인에 대해 외부 직업환경의학 전문의의 자문 의견을 받아, 산업보건 역학 연구자가 가이드라인을 읽고 g-formula를 이해할 때 어려움을 호소하는 의견을 수렴하고, 사용자 친화적으로 가이드라인을 수정함.
- 산업안전보건연구원에서 수정한 2021년 작성 가이드라인을 재검토하고 수정함.

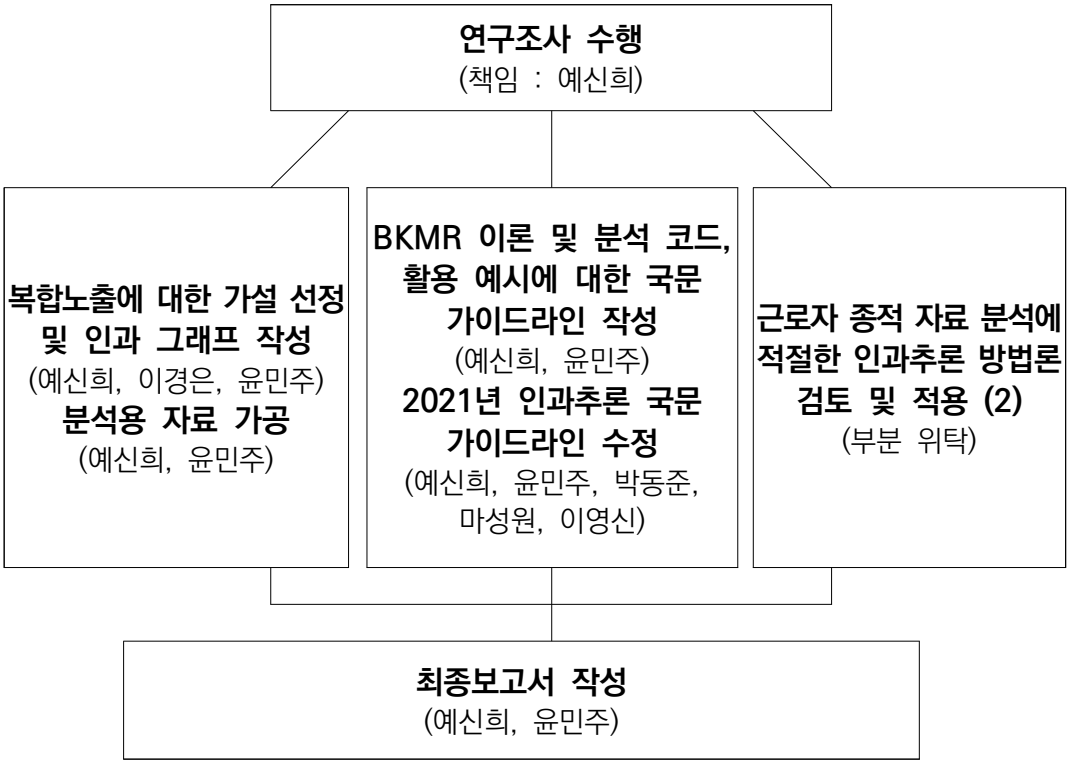
(5) g-formula 및 BKMR의 장단점 검토

- g-formula 및 BKMR을 적용한 다양한 문헌들을 검토하여 각 방법이 갖는 장/단점을 검토하여 요약하여 국내 산업보건 역학 연구자들이 적재적소에 방법을 사용할 수 있도록 돕고자 함.

(6) 후속연구 계획 수립

- 연구진 회의를 통해 결정함.
- 과제 진행 상황 공유, 관련 방법론 및 자료에 대한 논의를 위해 월 1회 이상 대면 또는 비대면 회의를 진행함. 산업보건 관련 전문 지식을 가지고 있는 과제 담당자와 통계적 인과추론 방법에 대한 연구 및 논문 출간 및 통계적 기법 개발 경험이 풍부한 부분위탁과제 연구진이 회의에 참여하였음.

3. 연구 추진 체계



4. 연구 윤리

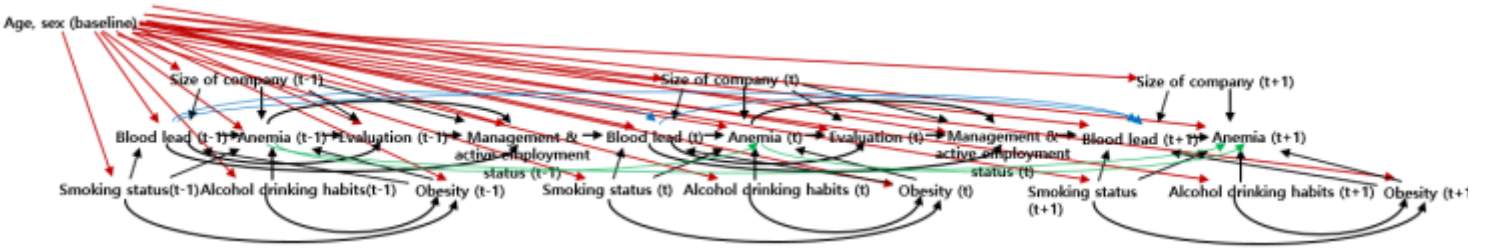
- 본 조사를 위하여 2022년 산업안전보건연구원 기관생명윤리위원회의 심의(institutional review board, IRB)를 통과하였음(승인번호: OSHRI-202206-HR-019).
- 본 조사를 위하여 2022년 서울대학교 기관생명윤리위원회의 심의 (institutional review board, IRB)를 통과하였음(승인번호: IRB No. 2208/002-005).

Ⅲ. 연구 결과

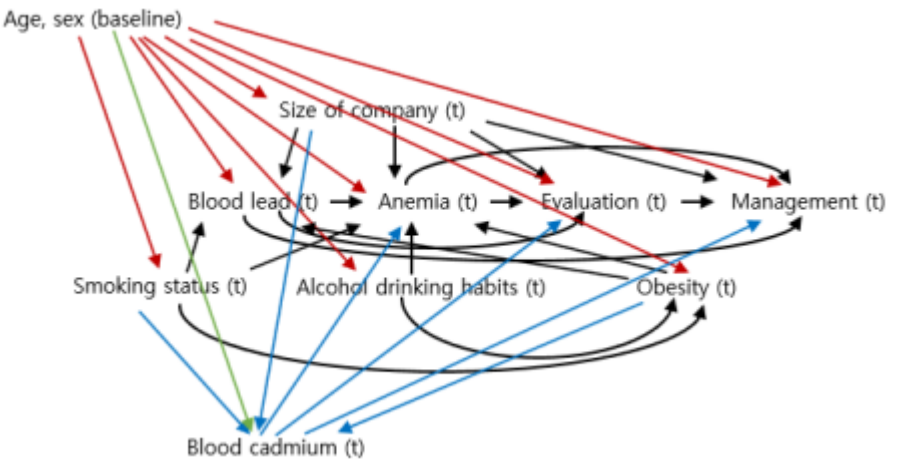
Ⅲ. 연구 결과

1. 근로자 종적 자료에 대한 가설 및 가설의 인과 그래프

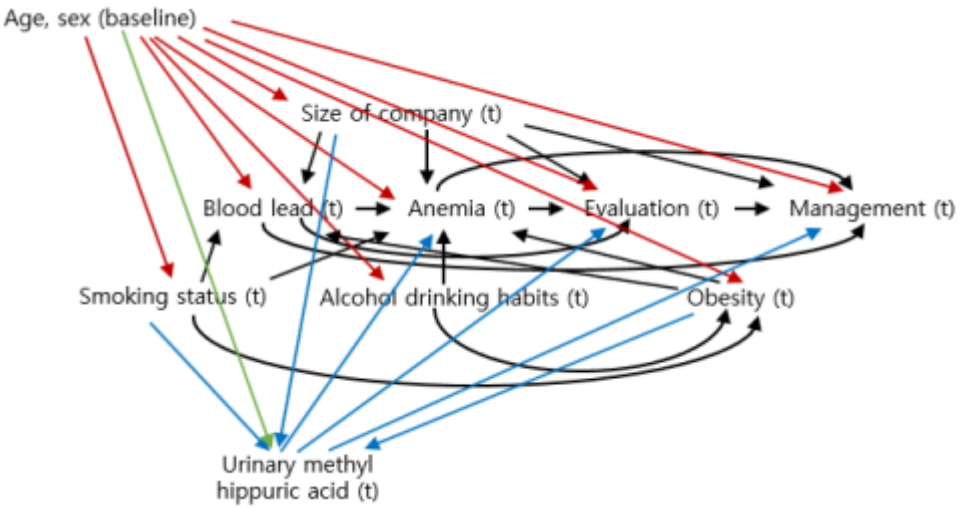
- 연구 가설 '납과 카드뮴에 대한 장기적인 복합노출이 빈혈 발생의 위험을 높인다.'와 '납과 크실렌에 대한 장기적인 복합노출이 빈혈 발생의 위험을 높인다.'를 근로자의 건강 관련 정보 (빈혈 여부, 혈중 납 농도, 혈중 카드뮴 농도, 요중 메틸마뇨산 농도, 나이, 성별, 사업장 규모, 음주 여부, 흡연 상태, 비만도, 건강진단 결과에 대한 의사의 판정 결과 및 그로 인한 사후관리조치 결과)를 활용하여 인과 그래프로 표현하였음.
- 실무지침 내 노출 기준치에 대해 납의 경우, 직업병 요관찰자에 해당하는 혈중 납 농도 기준치는 30 μg/dL이며, 직업병 유소견자에 해당하는 혈중 납 농도 기준치는 40 μg/dL이다. 카드뮴의 경우, 5 μg/L이며, 크실렌의 경우 요중 메틸마뇨산 농도 1.5 mg/L임.
- 따라서 납과 카드뮴 그리고 납과 크실렌에 대한 복합노출에 따른 빈혈 발생률을 산출하기 위해 혈중 납 농도 5, 10, 15, 20, 25, 30, 35, 40 (단위: μg/dL), 혈중 카드뮴 농도 1, 2, 3, 4, 5 (단위:μg/L) 그리고 요중 메틸마뇨산 농도 0.25, 0.50, 0.75, 1.00, 1.25, 1.50 (단위: mg/L)을 가상적인 노출 수준 (hypothetical exposure level)으로 설정하여 각 조합마다 특수건강검진 대상 근로자의 빈혈의 누적 발생률을 산출하고 등고선 그림을 통해 시각적으로 그 추이를 살펴보고자 함. 그림 Ⅲ-1은 납과 카드뮴 그리고 납과 크실렌에 관한 가설을 각각 인과 그래프로 표현하였을 때, 두 인과 그래프에 공통적으로 포함되는 시간에 따라 변하는 납에 대한 노출과 빈혈의 발생 사이의 관계를 인과 그래프임. 그림 Ⅲ-2는 납과 카드뮴에 대한 노출과 빈혈의 발생 사이의 관계를 세부적으로 그린 인과 그래프이고, 그림 Ⅲ-3은 납과 크실렌에 대한 노출과 빈혈의 발생 사이의 관계를 세부적으로 그린 인과 그래프임.



[그림 Ⅲ-1] 두 가설에 공통적으로 표현되는 시간에 따라 변하는 납에 대한 노출과 빈혈의 발생 사이의 관계를 표현한 인과 그래프



[그림 Ⅲ-2] 납과 카드뮴에 대한 노출과 빈혈의 발생 사이의 관계를 표현한 인과 그래프



[그림 Ⅲ-3] 납과 크실렌에 대한 노출과 빈혈의 발생 사이의 관계를 표현한 인과 그래프

2. g-formula를 사용한 산업보건 종적 자료 분석

1) 특수건강진단 자료의 특성

- 산업안전보건연구원의 직업건강연구실에서 수집하고 있는 근로자의 특수건강진단자료 중 2013년도부터 2019년까지 총 7년 동안 연 1회 이상 혈중 납 농도를 측정한 근로자의 수는 189,549명임. 이 자료는 반복 측정된 종적 자료로, 아이디 (개인식별변수), 검진 연도 (검진 순서), 성별, 나이, 사업장 규모, 흡연 상태, 음주 여부, 비만도, 혈중 납 농도, 혈중 카드뮴 농도, 요중 메틸마뇨산 농도, 혈중 헤모글로빈 수치, 사후관리 조치 결과, 판정결과에 대한 정보를 포함하고 있음.
- 특수건강검진자료 중 혈중 납 농도와 혈중 카드뮴 농도가 동시에 측정된 근로자는 총 23,336명이었으며, 혈중 납 농도와 요중 메틸마뇨산 농도(크실렌의 biomarker)가 동시에 측정된 근로자는 총 57,489명이었음. 빈혈 여부는 남성의 경우, 혈중 헤모글로빈 수치가 13g/dL 보다 작은 경우, 여성의 경우 12g/dL 보다 작은 경우, 빈혈이 있다고 정의하였음. 흡연 상태는 과거 흡연자 또는 현재 흡연자인 경우 1, 흡연 경험이 전혀 없으면 0이라고 정의하였고, 음주 여부는 특수건강진단을 받을 당시 음주를 하였으면 1, 음주를 하지 않았다면 0으로 정의하였음. 비만도는 근로자의 몸무게를 근로자의 키(m)의 제곱으로 나눈 체질량 지수를 사용하였음. 혈중 납 농도, 혈중 카드뮴 농도 그리고 요중 메틸마뇨산 농도는 근로자의 특수건강진단 결과로 확인할 수 있는 값으로 각각 μg/dL, μg/L 그리고 mg/L의 단위로 표현됨. 판정결과는 특수건강진단 결과에 따른 의사의 판단결과를 기록한 내용이며, 총 6가지의 범주(D1; D2 또는 DN; C1; C2 또는 CN; U 또는 R; A)를 가짐. 사후관리조치 결과는 판정결과에 따라 결정되는 사후관리조치 내용이며, 총 3가지의 범주(작업 장소 변경 및 타 업무로 전환조치 등 노출이 중단되는 경우; 보호구 착용 등 노출 수준이 낮아지는 경우; 사후관리가 필요 없는 경우)를 가

짐. 나이는 근로자의 나이를 의미하며, 현재 자료에는 18세 이상의 근로자에 대한 자료만 포함되어있음. 자료에서 나타나는 성별은 남성과 여성 두 종류의 성만 있으며, 사업장 규모는 사업장에서 근무하는 총 근로자의 수를 나타냄. 혈중 납 농도와 혈중 카드뮴 농도가 빈혈의 발생에 미치는 효과 그리고 혈중 납 농도와 요중 메틸마뇨산 농도가 빈혈의 발생에 미치는 효과를 알아보는 것이 연구 가설이기 때문에, 빈혈 여부가 결과 변수, 혈중 납 농도, 혈중 카드뮴 농도 그리고 요중 메틸마뇨산 농도가 노출 변수 그리고 그 외 나머지 변수는 교란 요인에 해당함.

2) g-formula를 사용한 특수건강진단 자료의 분석 방법

- 특수건강진단 자료를 사용하여 두 가지 연구 가설 “장기간에 걸쳐 납과 카드뮴에 노출될 가능성이 있는 작업장에서 일한 근로자들을 대상으로, 7년 동안 특정 농도로 일정하게 혈중 납 농도와 혈중 카드뮴 농도가 고정되었을 때, 빈혈의 발생률에 얼마나 영향을 미치는가?” 그리고 “장기간에 걸쳐 납과 크실렌 (xylene)에 노출될 가능성이 있는 작업장에서 일한 근로자들을 대상으로, 7년 동안 특정 농도로 일정하게 혈중 납 농도와 요중 메틸마뇨산 농도가 고정되었을 때, 빈혈의 누적 발생률이 얼마나 될 것인가? 그리고 일반인구 집단에 비교하여 그 누적 발생률은 얼마나 높게 나타날 것인가?”에 대하여 알아보고자 함. 두 연구 가설은 각각 두 종류의 유해물질(혈중 납 농도, 혈중 카드뮴 농도 / 혈중 납 농도, 요중 메틸마뇨산 농도)의 노출 수준의 조합에 따른 빈혈의 누적 발생률의 추이를 확인하고자 하며, 그에 따라 특수건강진단을 받은 근로자 집단에 각 유해물질의 특정 노출 수준에 대한 개입 (intervention)을 지정하고, 이를 등고선 그림을 통해 그 추이를 확인하고자 함. 각 연구 가설에서 확인하고자 하는 효과를 두 유해물질에 대한 근로자의 혈중/요중 농도가 특정 농도로 고정되어 있을 때의 빈혈의 누적 발생률과 일반인구 집

단의 노출 수준에 해당하는 혈중/요중 농도로 고정되었을 때 나타나는 빈혈의 누적 발생률의 비로 정의하였음. 예를 들어, 유해물질의 노출 수준에 대한 개입으로 “총 7년 동안 모든 근로자의 혈중 납 농도가 30 μg/dL, 혈중 카드뮴 농도가 5 μg/L로 유지되었을 때”를 지정할 경우, 이러한 노출 수준에 해당하는 개입을 받은 근로자들의 누적 발생률을, 대조군에 해당하는 일반인구 집단의 혈중 납 농도 (1.6 μg/dL[1])와 혈중 카드뮴 농도 (0.9187 μg/L[2])일 때의 빈혈의 발생률과 비교하여 유해물질의 노출 수준에 따른 빈혈의 누적 발생률을 산출하였음. 요중 메틸마뇨산 농도의 경우, 일반인구 집단에서의 요중 농도에 해당하는 0.234 mg/L[3]을 사용하였음.

- g-formula를 적용하여 특수건강검진자료를 분석하기 위해서는 결과 변수, 노출 변수 그리고 교란 요인에 대한 모형이 필요하며, 모형에 포함하는 요인은 인과 그래프를 통하여 결정되었음. 혈중 납 농도와 혈중 카드뮴 농도의 조합 그리고 혈중 납 농도와 요중 메틸마뇨산 농도에 관한 연구 가설을 분석하기 위한 모형에 포함된 변수는 각각 표 Ⅲ-1과 Ⅲ-2에 기술되어 있음.

1) 환경부 국립환경과학원의 국민환경보건기초조사 DB에서 2017년 자료 내 성인의 혈중 납 기하평균 1.6 μg/dL을 참고하였음.
2) 보건복지부 질병관리청의 국민건강영양조사 자료에서 2016년과 2017년에 측정한 자료 내 만 19세 이상 성인의 혈중 카드뮴 기하 평균 0.9187 μg/L를 참고하였음.
3) 국가통계포털 KOSIS에서 2011년부터 2014년까지 수록된 자료 내 성인의 요중 메틸마뇨산 농도의 기하평균 0.234 mg/L를 참고하였음.

〈표 Ⅲ-1〉 특수건강검진자료에서 혈중 납 농도와 혈중 카드뮴 농도에 대한 연구 가설을 분석하기 위해 설정한 결과 변수, 노출 변수 그리고 교란 요인에 대한 모형과 모형에 포함된 변수

모형의 종류	분석을 위한 모형에 포함된 변수	
결과 변수에 대한 모형	시점 t에서의 빈혈 여부	혈중 납 농도 (시점 t), 혈중 카드뮴 농도 (시점 t), 흡연 상태 (시점 t), 음주 여부 (시점 t), 비만도 (시점 t), 혈중 납 농도 (시점 t-1), 혈중 카드뮴 농도 (시점 t-1), 사후관리조치 결과 (시점 t-1), 혈중 납 농도 (시점 t-2), 혈중 카드뮴 농도 (시점 t-2) 그리고 나이, 성별, 사업장 규모
노출 변수에 대한 모형	시점 t에서의 혈중 납 농도	음주 여부 (시점 t), 흡연 상태 (시점 t), 비만도 (시점 t), 혈중 납 농도 (시점 t-1), 사후관리 조치 결과 (시점 t-1), 혈중 납 농도 (시점 t-2), 혈중 납 농도 (시점 t-3) 그리고 나이, 성별, 사업장 규모, 검진 순서
	시점 t에서의 혈중 카드뮴 농도	음주 여부 (시점 t), 흡연 상태 (시점 t), 비만도 (시점 t), 혈중 카드뮴 농도 (시점 t-1), 사후관리조치 결과 (시점 t-1), 혈중 카드뮴 농도 (시점 t-2), 혈중 카드뮴 농도 (시점 t-3) 그리고 나이, 성별, 사업장 규모, 검진순서
교란 요인에 대한 모형	시점 t에서의 음주 여부	음주 여부 (시점 t-1) 그리고 나이, 성별, 검진순서
	시점 t에서의 흡연 상태	흡연 상태 (시점 t-1), 흡연 상태 (시점 t-1) 그리고 나이, 성별, 검진순서
	시점 t에서의 비만도	음주 여부 (시점 t), 흡연 상태 (시점 t), 비만도(시점 t-1) 그리고 나이, 성별, 검진순서
	시점 t에서의 판정결과	혈중 납 농도 (시점 t), 혈중 카드뮴 농도 (시점 t) 그리고 나이, 성별, 사업장 규모
	시점 t에서의 사후관리조치	판정결과 (시점 t), 혈중 납 농도 (시점 t), 혈중 카드뮴 농도 (시점 t) 그리고 나이, 성별, 사업장 규모

〈표 Ⅲ-2〉 특수건강검진자료에서 혈중 납 농도와 요중 메틸마뇨산 농도에 대한 연구 가설을 분석하기 위해 설정한 결과 변수, 노출 변수 그리고 교란 요인에 대한 모형과 모형에 포함된 변수

모형의 종류	분석을 위한 모형에 포함된 변수	
결과 변수에 대한 모형	시점 t에서의 빈혈 여부	혈중 납 농도 (시점 t), 요중 메틸마뇨산 농도 (시점 t), 흡연 상태 (시점 t), 음주 여부 (시점 t), 비만도 (시점 t), 혈중 납 농도 (시점 t-1), 요중 메틸마뇨산 농도 (시점 t-1), 사후관리조치 결과 (시점 t-1), 혈중 납 농도 (시점 t-2) 그리고 나이, 성별, 사업장 규모, 검진연도
노출 변수에 대한 모형	시점 t에서의 혈중 납 농도	음주 여부 (시점 t), 흡연 상태 (시점 t), 비만도 (시점 t), 혈중 납 농도 (시점 t-1), 사후관리조치 결과 (시점 t-1), 혈중 납 농도 (시점 t-2), 혈중 납 농도 (시점 t-3), 그리고 나이, 성별, 사업장 규모, 검진 순서
	시점 t에서의 요중 메틸마뇨산 농도	음주 여부 (시점 t), 흡연 상태 (시점 t), 비만도 (시점 t), 요중 메틸마뇨산 농도 (시점 t-1), 사후관리조치 결과 (시점 t-1) 그리고 나이, 성별, 사업장 규모, 검진순서
교란 요인에 대한 모형	시점 t에서의 음주 여부	음주 여부 (시점 t-1) 그리고 나이, 성별, 검진순서
	시점 t에서의 흡연 여부	흡연 상태 (시점 t-1), 흡연 상태 (시점 t-2) 그리고 나이, 성별, 검진순서
	시점 t에서의 비만도	음주 여부 (시점 t), 흡연 여부 (시점 t), 비만도(시점 t-1) 그리고 나이, 성별, 검진순서
	시점 t에서의 판정결과	혈중 납 농도 (시점 t), 요중 메틸마뇨산 농도 (시점 t) 그리고 나이, 성별, 사업장 규모
	시점 t에서의 사후관리 조치	판정결과 (시점 t), 혈중 납 농도 (시점 t), 요중 메틸마뇨산 농도 (시점 t) 그리고 나이, 성별, 사업장 규모, 검진 순서

3) g-formula를 사용한 특수건강진단 자료의 분석 결과

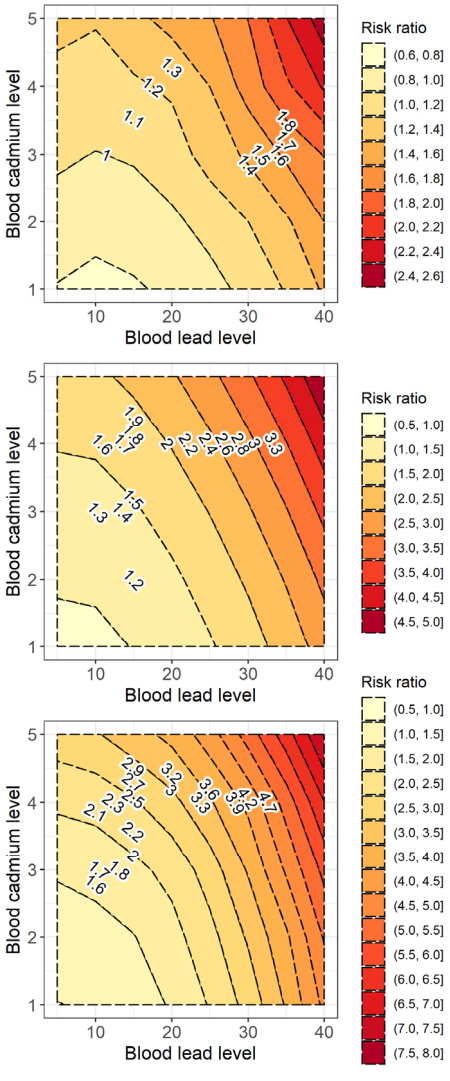
- 표 Ⅲ-3은 혈중 납 농도와 혈중 카드뮴 농도의 조합에 따른 빈혈의 누적 발생률을 일반인구집단에서 측정되는 혈중 농도인 혈중 납 농도 (1.6 μg/dL)와 혈중 카드뮴 농도 (0.9187 μg/L)일 때의 빈혈의 누적 발생률로 나누어 구한 위험 비를 기술한 표이며, 95% 신뢰구간은 bootstrap 방법을 사용하여 산출하였음. 표 Ⅲ-3에서 혈중 카드뮴 농도가 고정되어 있을 때, 혈중 납 농도가 증가함에 따라 위험 비가 증가하는 것을 확인할 수 있었으며, 마찬가지로 혈중 납 농도가 고정되어있는 경우, 혈중 카드뮴 농도가 증가함에 따라 위험 비가 증가하는 것을 확인할 수 있음. 저 농도에서의 혈중 납 농도와 혈중 카드뮴의 조합의 일부에서 빈혈에 대한 위험 비의 95% 신뢰구간들 중 일부가 1을 포함하고 있지만, 고 농도를 포함하여 그 외의 혈중 납 농도와 혈중 카드뮴 농도의 조합에서 95% 신뢰구간의 왼쪽 경계 값이 모두 1보다 크며, g-formula가 유의한 결과를 제공하고 있음을 알 수 있음.
- 더불어, 표 Ⅲ-4은 혈중 납 농도와 요중 메틸마뇨산 농도의 조합에 따른 빈혈의 누적 발생률을 일반인구집단에서 측정되는 혈중 납 농도와 요중 메틸마뇨산 농도 (0.234 mg/L)일 때의 빈혈의 누적 발생률로 나누어 산출한 위험 비를 기술한 표임. 신뢰구간은 표 Ⅲ-3과 마찬가지로 bootstrap 방법을 이용하여 계산하였음. 표 Ⅲ-4를 보면 요중 메틸마뇨산 농도가 고정되어있는 경우, 혈중 납 농도가 증가함에 따라 위험 비가 증가하는 것을 확인할 수 있음. 하지만 혈중 납 농도가 고정되어있는 경우, 요중 메틸마뇨산 농도가 증가함에 따라 위험 비가 소폭 감소하는 것을 확인할 수 있음. 또한, 낮은 혈중 납 농도에서 납과 카드뮴에 대한 노출로 인한 빈혈의 누적 발생률의 95% 신뢰구간이 1을 포함한 것과 같이 혈중 납 농도와 요중 메틸마뇨산 농도의 조합에서의 빈혈의 누적 발생률 또한 낮은 혈중 납 농도에서 95% 신뢰구간이 1을 포함하지만, 혈중 납 농도가 높아지는 경우 g-formula가 95% 신뢰구간의 왼쪽 경

계 값으로 1보다 큰 값을 제공함으로써 유의한 결과를 제공한다는 것을 알 수 있음.

- 그림 Ⅲ-4는 혈중 납 농도와 혈중 카드뮴 농도의 조합에 따른 빈혈의 누적 발생률의 위험 비를 시각적으로 확인하기 위한 등고선 그림으로, 표 Ⅲ-3에서 확인한 것과 같이 혈중 납 농도와 혈중 카드뮴 농도가 모두 증가함에 따라 빈혈의 발생률의 위험 비 또한 증가하는 것을 확인할 수 있음. 그림 Ⅲ-5는 혈중 납 농도와 요중 메틸마뇨산 농도의 조합에 따른 빈혈의 누적 발생률의 위험 비를 시각적으로 표현하기 위한 등고선 그림으로, 표 Ⅲ-4에서 확인한 것과 같이 혈중 납 농도가 증가함에 따라 위험 비가 증가하는 것을 시각적으로 확인할 수 있음. 위의 결과를 위해 사용된 g-formula가 특수건강검진자료를 적절히 설명하고 있는지 확인하기 위해 자료에서 나타나는 근로자들의 혈중 납 농도와 혈중 카드뮴 농도 또는 요중 메틸마뇨산 농도(자연 경과; natural course)에서 적합된 g-formula를 사용하여 산출한 빈혈에 대한 누적 발생률을 자료에서 나타나는 유해물질의 노출 수준에서의 빈혈에 대한 누적 발생률과 비교하였으며, 그림 Ⅲ-6은 그 결과를 제공함. 그림 Ⅲ-6에서 자연 경과일 때의 g-formula 적합 결과를 ’parametric g-formula estimates’라 표현하였고, 자료로부터 직접 산출한 빈혈의 누적 발생률을 ‘nonparametric estimates’라 표현하였음. 혈중 납 농도와 혈중 카드뮴 농도 그리고 혈중 납 농도와 요중 메틸 마뇨산 농도 각각의 자연 경과에서 산출한 빈혈의 누적 발생률에 대한 95% 신뢰구간이 자료로부터 직접 산출한 빈혈에 대한 누적 발생률을 포함하므로 g-formula가 자료를 올바르게 적합한다는 것을 알 수 있음.

〈표 Ⅲ-3〉 혈중 납 농도와 혈중 카드뮴 농도에 따른 빈혈의 발생률에 대한 위험비를 기술한 표. 혈중 납 농도에 대하여 5 μg/dL 단위 별로 굵은 글씨 및 회색 칸으로 표의 셀을 표시하였음.

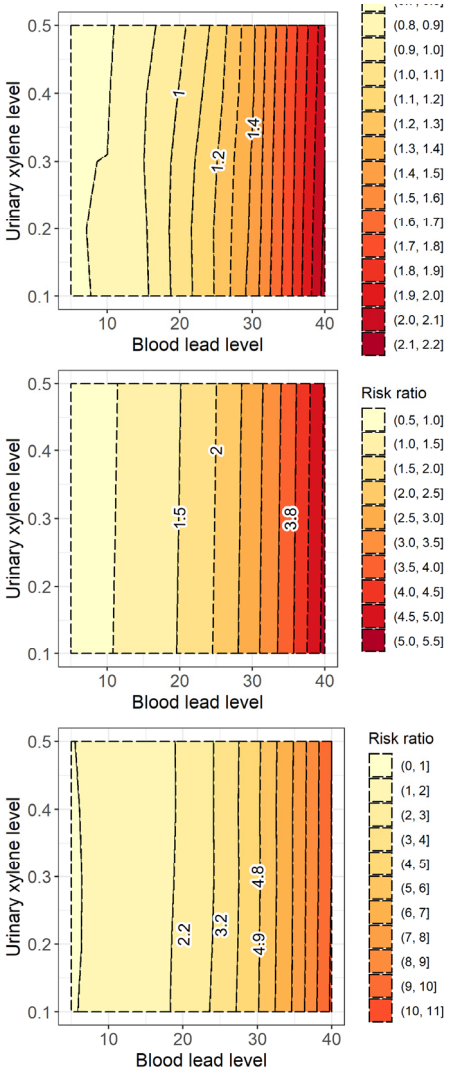
혈중 납 농도 (μg/dL)	혈중 카드뮴 농도 (μg/L)	위험 비(Risk ratio)	위험 비의 95% 신뢰구간	
			왼쪽 경계 값	오른쪽 경계 값
1.6	0.9187	Reference(1.0000)	-	-
5	**1**	**0.8811**	**0.7884**	**0.9854**
5	2	1.0453	0.9113	1.2160
5	3	1.2619	1.0384	1.5628
5	4	1.5321	1.1450	2.0900
5	5	1.8608	1.2495	2.7630
10	**1**	**0.8999**	**0.7428**	**1.0732**
10	2	1.0676	0.8621	1.2967
10	3	1.2892	0.9912	1.6799
10	4	1.5658	1.1543	2.1672
10	5	1.9010	1.2091	2.9435
15	**1**	**1.0138**	**0.7777**	**1.2560**
15	2	1.1991	0.8912	1.4774
15	3	1.4436	1.0230	1.8725
15	4	1.7463	1.1672	2.4428
15	5	2.1106	1.3490	3.2974
20	**1**	**1.1971**	**0.8372**	**1.5493**
20	2	1.4096	0.9637	1.8312
20	3	1.6879	1.1031	2.1498
20	4	2.0291	1.2316	2.8534
20	5	2.4356	1.4685	3.6415
25	**1**	**1.4504**	**0.9333**	**2.0399**
25	2	1.6979	1.0665	2.3277
25	3	2.0187	1.2843	2.7492
25	4	2.4080	1.4015	3.4354
25	5	2.8660	1.5829	4.2732
30	**1**	**1.7814**	**1.0544**	**2.6648**
30	2	2.0705	1.1997	3.0672
30	3	2.4417	1.4351	3.5540
30	4	2.8864	1.6837	4.1293
30	5	3.4062	1.8051	5.2302
35	**1**	**2.2006**	**1.2159**	**3.5577**
35	2	2.5371	1.3723	3.9055
35	3	2.9656	1.6347	4.5729
35	4	3.4811	1.9632	5.2125
35	5	4.0892	2.1902	6.2649
40	**1**	**2.7229**	**1.4304**	**4.6238**
40	2	3.1172	1.5998	5.0553
40	3	3.6244	1.8106	5.6505
40	4	4.2449	2.2264	6.4824
40	5	4.9756	2.5131	7.5492



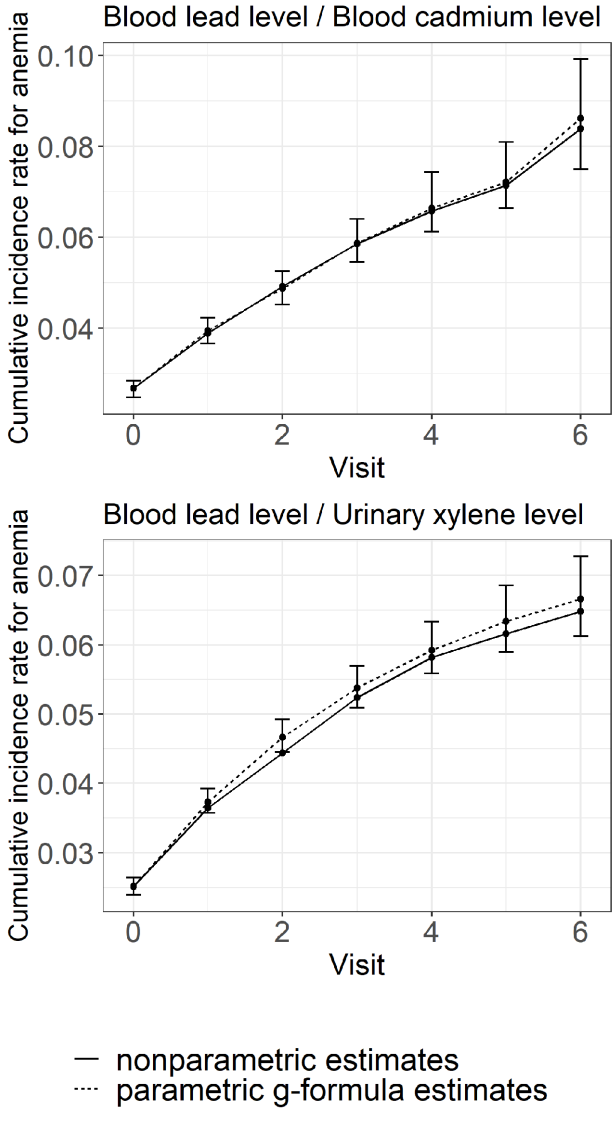
[그림 Ⅲ-4] 혈중 납 농도 (blood lead level)와 혈중 카드뮴 농도 (blood cadmium level)에 따른 빈혈의 발생률에 대한 위험 비를 표현한 등고선 그림. 상단의 그림은 빈혈의 누적 발생률 추정치에 대한 95% 신뢰구간의 하한을, 가운데 그림은 빈혈의 누적 발생률에 대한 추정치를 그리고 하단의 그림은 빈혈의 누적 발생률 추정치에 대한 95% 신뢰구간의 상한을 그린 등고선 그림임

〈표 Ⅲ-4〉 혈중 납 농도와 요중 메틸마뇨산 농도에 따른 빈혈의 발생률에 대한 위험비를 기술한 표. 혈중 납 농도에 대하여 5 μg/dL 단위 별로 굵은 글씨 및 회색 칸으로 표의 셀을 표시하였음.

혈중 납 농도 (μg/dL)	요중 메틸마뇨산 농도 (μg/dL)	위험 비(Risk ratio)	위험 비의 95% 신뢰구간	
			왼쪽 경계 값	오른쪽 경계 값
1.6	0.2340	Reference(1.0000)	-	-
5	**1**	**0.8879**	**0.7946**	**0.9727**
5	2	0.8820	0.7971	0.9542
5	3	0.8762	0.7980	0.9541
5	4	0.8705	0.7845	0.9640
5	5	0.8648	0.7670	0.9851
10	**1**	**0.9663**	**0.8045**	**1.1346**
10	2	0.9599	0.8040	1.1218
10	3	0.9535	0.8008	1.1178
10	4	0.9472	0.7930	1.1257
10	5	0.9410	0.7854	1.1343
15	**1**	**1.1812**	**0.8772**	**1.5509**
15	2	1.1732	0.8858	1.5277
15	3	1.1652	0.8986	1.5050
15	4	1.1574	0.8908	1.5003
15	5	1.1496	0.8623	1.5401
20	**1**	**1.5309**	**1.0410**	**2.2387**
20	2	1.5203	1.0482	2.2068
20	3	1.5098	1.0336	2.1592
20	4	1.4995	1.0075	2.1217
20	5	1.4892	0.9742	2.1180
25	**1**	**2.0506**	**1.2097**	**3.3032**
25	2	2.0364	1.2120	3.2104
25	3	2.0223	1.1951	3.1731
25	4	2.0083	1.1575	3.1884
25	5	1.9944	1.1277	3.1883
30	**1**	**2.7911**	**1.4447**	**4.9066**
30	2	2.7721	1.4219	4.8753
30	3	2.7532	1.4105	4.8242
30	4	2.7345	1.3888	4.8444
30	5	2.7159	1.3757	4.7998
35	**1**	**3.8116**	**1.7614**	**7.1765**
35	2	3.7867	1.7548	7.1367
35	3	3.7619	1.7455	7.1163
35	4	3.7373	1.7358	7.1083
35	5	3.7129	1.7181	7.0711
40	**1**	**5.1743**	**2.1391**	**10.2182**
40	2	5.1425	2.1214	10.1321
40	3	5.1095	2.1020	10.0472
40	4	5.0792	2.0814	10.0226
40	5	5.0483	2.0612	10.0168



[그림 Ⅲ-5] 혈중 납 농도 (blood lead level)와 요중 메틸마뇨산 농도 (urinary xylene level)에 따른 빈혈의 발생률에 대한 위험 비를 표현한 등고선 그림. 상단의 그림은 빈혈의 누적 발생률 추정치에 대한 95% 신뢰구간의 하한을, 가운데 그림은 빈혈의 누적 발생률에 대한 추정치를 그리고 하단의 그림은 빈혈의 누적 발생률 추정치에 대한 95% 신뢰구간의 상한을 그린 등고선 그림임



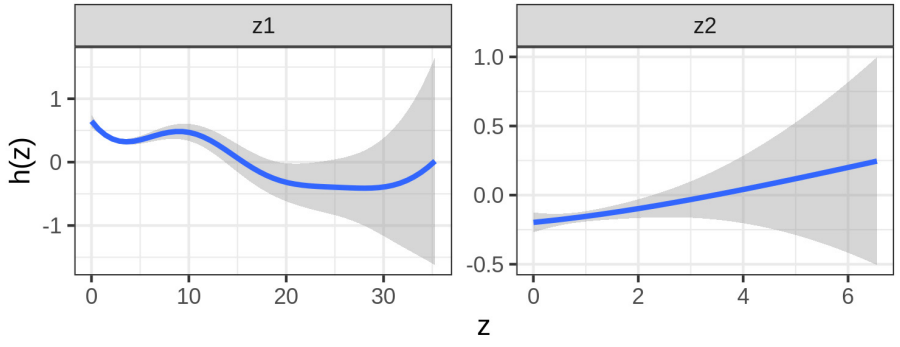
[그림 Ⅲ-6] 혈중 납 농도 (blood lead level)와 혈중 카드뮴 농도 (blood cadmium level)에 대한 빈혈의 누적 발생률을 산출하기 위해 적합한 g-formula의 자연 경과 (natural course)에서의 적합 결과 (위). 혈중 납 농도 (blood lead level)와 요중 메틸마뇨산 농도 (urinary xylene level)에 대한 빈혈의 누적 발생률을 산출하기 위해 적합한 g-formula의 자연 경과 (natural course)에서의 적합 결과 (아래)

4) BKMR을 사용한 특수건강진단 자료의 분석 방법

- 연구가설을 분석하기 위해 BKMR을 적합하였다. 다만, g-formula의 분석에 사용된 약 20,000명의 근로자 자료를 모두 사용하게 되면 BKMR의 결과를 확인하기까지 많은 시간이 소요되기 때문에 약 20,000명의 근로자 중 임의로 2,000명을 무작위 추출하여 현재 분석에 사용하였음.
- 반복측정 자료인 특수건강검진자료의 특징을 반영하기 위해 BKMR을 적합할 때, 근로자의 아이디를 입력하였다. 보정변수로는 나이와 성별, 음주 유무, 흡연 상태, 체질량지수, 사업장 규모를 사용하였음.

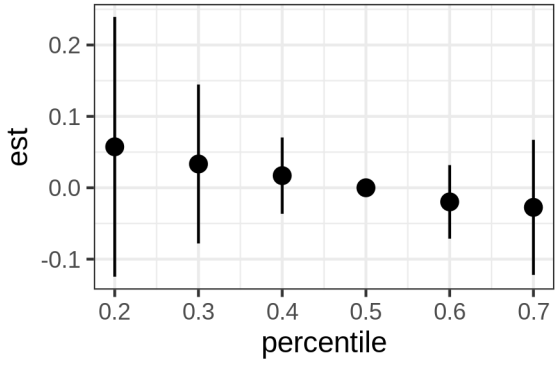
5) BKMR을 사용한 특수건강진단 자료의 분석 결과

- 그림 Ⅲ-7은 BKMR을 적합한 후, 혈중 납 농도(z1) 그리고 혈중 카드뮴 농도(z2) 각각에 대하여 h(·)의 추정치를 그린 그래프임. 혈중 납 농도의 경우, 혈중 납 농도가 증가할수록 h(·)의 값이 작아지는, 즉 혈중 납 농도가 증가할수록 빈혈의 발생률이 작아지는 경향이 나타났으며, 사실상 혈중 납 농도가 15 μg/dL 이상인 경우에는 h(·)의 추정치에 대한 95% 신뢰구간이 0을 포함하여 BKMR은 유의하지 않은 결과를 제공하였음. 혈중 카드뮴 농도의 경우, 혈중 카드뮴 농도가 증가할수록 빈혈의 발생률이 증가하는 추세를 보였음. 하지만 혈중 카드뮴 농도의 경우 또한, 혈중 카드뮴 농도가 약 3.3 μg/L 이상에서 h(·)의 추정치에 대한 95% 신뢰구간이 0을 포함하여 BKMR이 유의하지 않은 결과를 제공함. **혈중 납 농도와 혈중 카드뮴 농도에 대해 고 농도에서 유의한 결과를 제공한 g-formula와 달리 BKMR은 사실상 반대의 결과를 제공하였는데, 이는 BKMR이 g-formula와 달리 건강근로자 생존 편향을 고려하지 못하였기 때문에 발생하는 차이로 이해할 수 있음.**



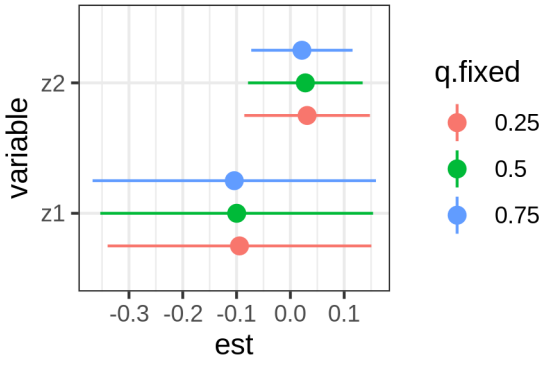
[그림 Ⅲ-7] 혈중 납 농도와 혈중 카드뮴 농도에 따른 h(·)의 추정치에 대한 그래프

- 그림 Ⅲ-8에서 근로자의 혈중 납 농도와 혈중 카드뮴 농도를 모두 특정 노출 수준 (percentile)인 경우에서의 h(·)의 추정치의 추세를 살펴볼 수 있음. 노출 수준이 중앙값일 때와 비교하여 노출 수준이 증가할수록 h(·)의 추정치가 작아지고 있다는 것을 확인할 수 있고, 이는 중금속에 대해 혈중 농도가 증가할수록 빈혈이 발생할 확률이 감소한다는 의미이며, g-formula와는 상반된 결과를 제공함.



[그림 Ⅲ-8] 혈중 납 농도와 혈중 카드뮴 농도를 각 percentile에 고정시킨 경우의 h(·)의 추정치에 대한 그래프

▪ 그림 Ⅲ-9는 빈혈의 발생과 관련하여 혈중 납 농도와 혈중 카드뮴 농도의 교호작용을 살펴볼 수 있는 그래프이며, 그림 29의 세로축에서 z1을 보면 혈중 카드뮴 농도가 25, 50, 75 percentile로 고정되어있을 때, 혈중 납 농도에 따른 h(·)의 추정치와 95% 신뢰구간을 보여줌. 서로 다른 혈중 카드뮴 농도에서 혈중 납 농도에 따른 h(·)의 추정치들의 차이가 크지 않고, 반대로 서로 다른 혈중 납 농도에서 혈중 카드뮴 농도에 따른 h(·)의 추정치들의 차이가 크지 않으므로 혈중 납 농도와 혈중 카드뮴 농도 사이의 교호작용이 유의하지 않을 수 있음을 시각적으로 확인할 수 있음.



[그림 Ⅲ-9] 혈중 납 농도와 혈중 카드뮴 농도 사이의 교호작용을 나타낸 그래프

▪ 이를 혈중 카드뮴 농도가 각각 75, 25 percentile로 고정되었을 때, 혈중 납 농도가 75 percentile에서 25 percentile로 감소하였을 때의 h(·)의 추정치의 변화에 대한 차이를 살펴봄으로써 (혈중 카드뮴 농도의 경우, 반대로 혈중 납 농도가 고정되어있을 때, 혈중 카드뮴 농도에 대한 추정치의 변화에 대한 차이) 교호작용이 존재하는지 수치적으로 확인할 수 있음. 표 Ⅲ-5를 보면 교호작용의 추정치는 -0.0097, 표준 오차는 0.0297로 95% 신뢰구간을 계산하게 되면 신뢰구간이 0을 포함하는 것을 확인할 수 있음.

〈표 Ⅲ-5〉 혈중 납 농도와 혈중 카드뮴 농도에 대한 교호작용의 추정치 및 표준 오차

변수	추정치	표준 오차
혈중 납 농도	-0.0097	0.0297
혈중 카드뮴 농도	-0.0097	0.0297

3. g-formula 및 BKMR에 대한 국문 가이드라인 검토 및 수정 의견 제시

- 산업안전보건연구원에서 작성한 인과추론 통계 방법 g-formula와 BKMR에 대한 국문 가이드라인 초안을 부분위탁 과제 연구진이 검토한 후 수정하였고, 주요 수정내용은 다음과 같음.

〈표 Ⅲ-6〉 BKMR에 대한 국문 가이드라인 수정 전과 후의 대조표.

수정 전	수정 후
근로자들은 유해물질 노출과 관련된 요인들로 인해 고용상태가 변화될 수 있고, 유해물질 노출과 교란요인이 시간에 따라 변화할 수 있다. 또한, 근로자들은 여러 가지 업무와 관련된 유해요인에 복합적으로 반복적으로 노출된다. 이러한 근로자를 대상으로 수집한 종적자료의 특성을 고려하지 않고, 노출과 질병 간의 관련성을 분석할 시에 다양한 바이어스가 발생할 수 있다. 예를 들어, 시간에 따라 변화하는 노출과 선행 노출에 영향을 받으며 시간에 따라 변화하는 혼란요인의 특성을 고려하지 않고 건강영향을 평가할 때 문제가 발생할 수 있는데, 그림 1과 같이 '판정결과' 변수는 '첫 번째 노출 측정'과 '두 번째 노출 측정' 매개변수임과 동시에 '두 번째 노출 측정' 변수와 '건강 결과' 변수의 혼란요인이다. 그림 1과 같은 경우 '판정 결과'를 보정해도 바이어스가 발생하고, 보정하지 않아도 바이어스가 발생하므로, 이러한 경우에는 전통적인 회귀 분석으로 바이어스를 통제할 수 없고, g methods와 같은 새로운 통계 분석방법으로 접근해야한다.	근로자들은 유해물질 노출과 그와 관련된 요인들로 인해 고용상태가 변화될 수 있고, 유해물질에 대한 노출과 교란 요인은 시간에 따라 변화할 수 있다. 또한, 근로자들은 여러 가지 업무와 관련된 유해요인에 복합적으로 그리고 반복적으로 노출된다. 근로자를 대상으로 수집한 종적 자료의 특성을 고려하지 않고, 노출과 질병 간의 관련성을 분석하게 되면 다양한 편향(bias)이 발생할 수 있다. 예를 들어, 시간에 따라 변화하는 노출과 선행 노출에 영향을 받으며 시간에 따라 변화하는 교란 요인의 특성을 고려하지 않는 경우가 대표적인 경우이다. 그림 1과 같이 '판정 결과' 변수는 '첫 번째 노출 측정'과 '두 번째 노출 측정' 매개변수임과 동시에 '두 번째 노출 측정' 변수와 '건강 결과' 변수의 교란 요인이기도 하다. '판정 결과' 변수를 보정하게 되면 첫 번째 노출 측정에서 건강 결과로의 경로 중 '판정 결과' 변수를 통해 지나가는 경로가 모두 차단되므로 건강 결과에 대한 첫 번째 노출의 인과효과는 총 효과가 아닌 직접적인 인과효과만 추정된다. 하지만 '판정 결과' 변수를 보정하지 않으면 건강 결과에 대한 두 번째 노출의 인과효과를 추정함에 있어 교란 비뚤림(confounding bias)이 발생한다. 이러한 특성이 있는 자료에서는 전통적인 회귀분석으로 이러한 비뚤림을 통제할 수 없으며, g-methods와 같은 새로운 통계 분석 방법의 접근이 필요하다.
G methods의 경우, 그림 2와 같이 근로자 종적연구에서 발생하는 '건강근로자 생존효과'로 인한 선택 바이어스를 제어하는 데도 활용할 수 있다.	또한, g-methods의 경우, 그림 2와 같이 근로자 종적 연구에서 주로 발생하는 '건강근로자 생존 효과(healthy worker survivor bias)'로 인한 선택 바이어스(selection bias)를 제어하는 데도 활용할 수 있다.
이 외에도, 아래 그림 3과 같이 복합유해물질 노출을 고려하지	위에서 설명한 두 가지 비뚤림 이 외에도, 아래 그림 3과 같이

수정 전	수정 후
않고, 단일 물질 1과 건강 결과 간의 관련성을 분석할 경우, 실제로는 물질 2가 건강 결과에 영향을 미치는 원인임에도 불구하고, 통계적으로 물질 1과 건강 결과 간의 유의한 관련성이 관찰될 수 있다.	복합유해물질에 대한 노출, 즉 근로자가 2가지 이상의 유해물질에 다중 노출되었을 때, 다중 노출을 고려하지 않고, 단일 물질 1과 건강 결과 간만의 관련성을 분석할 경우, 실제로는 물질 2가 건강 결과에 영향을 미치는 원인임에도 불구하고, 통계적으로 물질 1과 건강 결과 간의 유의한 관련성이 관찰될 수 있다.
하지만, 복합노출의 건강영향 평가에 인과추론 통계 접근법을 함께 적용한 연구는 거의 이루어지지 않고 있으며, 관련된 통계적 이론도 부족한 부분이 있다. 따라서, 국내 산업보건 영역에서는 '인과추론 통계 접근법'과 '복합노출의 건강영향 평가 방법' 각각에 대해서도 근로자 종적자료 분석에 적절하게 사용되지 않고 있으므로, 국내 산업보건 영역에서 이러한 통계적 접근법들을 활용하는 방법에 대하여 구체적인 국문 가이드라인이 필요하다. 2021년 연구과제(예신희 등. 2021)에서는 단일노출로 인한 건강영향을 인과추론하는 국문가이드라인을 작성하였고, 2022년 본 과제에서는 2개 이상의 노출로 인한 건강영향을 인과추론 하는데 활용할 수 있는 통계방법론에 대한 국문가이드라인을 작성하고자 한다.	하지만, 국내 산업보건 영역에서는 '인과추론 통계 방법'과 '복합노출의 건강영향 평가 방법' 각각 근로자 종적 자료 분석에 적절하게 사용되고 있지 않으며, 따라서 국내 산업보건 역학연구에서 이러한 통계적 방법론들이 다양하게 활용될 수 있도록 각 방법론에 대한 구체적인 국문 가이드라인이 필요하다. 2021년 연구과제(예신희 등. 2021)에서는 단일노출로 인한 건강영향을 추론하는 국문 가이드라인을 작성하였고, 2022년 본 과제에서는 2개 이상의 유해물질에 대한 노출이 근로자의 건강에 영향을 미치는지 그 영향을 추론하는데 활용될 수 있는 통계적 방법론에 대한 국문 가이드라인을 작성하고자 한다.
	Keil AP 등(2017)은 건강 근로자 생존 비뚤림을 보정하기 위해 근로자의 고용상태(employment status)를 인과적 방향성 비순환 그래프에 포함한 후, g-formula를 이용하여 8,014명의 백인 남성에 대하여 구리 용광로에서 공기 중 비소 흡입량에 따라 모든 질병, 심장병, 폐암으로 인한 초과 사망률이 나이에 따라 어떻게 변화하는지 연구하였다. 이때 노출량에 대한 개입(intervention)은 노출을 시키지 않는 개입, 개입하지 않음, 심한 노출을 시키는 개입 총 3가지로 분류하여 적용하였으며, 각 개입을 시행하였을 때 발생하는 초과 사망률을 측정하였다. Valeri L 등(2017)은 중금속 복합물질에 대한 임신 중 노출이 출생 후 20-40개월 영유아의 신경 발달 결과에 영향을 미치는지

수정 전	수정 후
	파악하기 위해 어머니-아이 825쌍을 대상으로 BKMR을 적용하여 연구를 진행하였다. 이때, 영유아의 신경 발달 정도를 측정하기 위해 인지 발달 점수(cognitive development score)와 언어 개발 종합 점수(language development composite score)를 사용하였으며, 중금속 복합물질로는 비소, 마그네슘, 납을 고려하였다. 이슬비 등(2019)은 어머니-아이 302쌍을 대상으로 중금속에 해당하는 납, 수은, 카드뮴과 대기오염물질에 해당하는 NO2, PM10, PM2.5 그리고 비스페놀 A, 프탈레이트 대사체 MEHHP, MEOHP, MnBP 3종을 포함하여 총 10가지의 환경유해물질에 대한 복합노출이 출생 후 6개월이 지난 영유아의 아토피 피부염 발생에 미치는 영향을 확인하고자 하였다. BKMR을 사용하여, 임신 말기에서 복합물질에 대한 누적 노출 양이 증가할 때 영유아의 아토피 피부염 발생의 위험이 증가한다고 보고하였다.
근로자를 포함한 대다수의 인구 집단은 여러 유해요인에 동시에 노출될 수 있으므로, 최근 몇 년간 복합유해물질 노출에 의한 건강영향을 통계적인 접근을 통해 정량적으로 추정하는 연구가 주목 받고 있다. 이러한 복합유해물질에 대한 건강영향을 추정할 때에는 몇 가지 고려해야 할 점들이 있다. 첫 번째로, 복합유해물질 노출과 건강영향이 복잡한 비선형 또는 비가법적 관계(non-additive relationship)를 가질 수 있다는 것이다. 두 번째로, 여러 노출의 유연한 노출-반응 함수와 여러 노출들 간에 상호작용이 허용되어야 하는데, 이러한 경우 관측치들(observations)의 수에 비해 모수들(parameters)의 수가 더 많아 불완전한 추정치를 산출하게 되는 고차원적 문제(high-dimensional problem)가 발생할 수 있다. 세 번째로는, 통계방법이 높은 상관성을 갖고 있는 노출들(multiple highly correlated exposures)로 구성된 혼합물의 복잡한 구조를 설명할 수 있어야 한다.	근로자를 포함한 대다수의 인구집단은 많은 경우 여러 유해요인에 동시에 노출되기 때문에, 최근 몇 년간 통계적인 접근을 통해 복합유해물질 노출에 의한 건강영향을 정량적으로 추정하려는 연구가 주목받고 있다. 이러한 복합유해물질의 노출로 인한 건강영향을 추정할 때에는 몇 가지 고려해야 할 점들이 있다. 첫 번째로, 복합유해물질 노출과 건강영향이 복잡한 비선형 또는 비가법적 관계(non-additive relationship)를 가질 수 있다는 것이다. 두 번째로, 결과 변수와 여러 유해물질의 노출 사이의 교호작용이 허용되어야 하는데, 이러한 경우, 모형에서 추정해야 하는 모수(parameters)의 수가 관측치(observations)의 수보다 더 많아지게 되어 즉, 고차원적 문제(high dimensional problem)가 발생하여 추정치 산출에 있어 불안정해질 수 있다. 세 번째로는, 사용된 통계 방법이 높은 상관성을 가지고 있는 노출들(multiple highly correlated exposures)로 구성된 혼합물의 복잡한 구조를 설명할

수정 전	수정 후
	수 있어야 한다.
혼합물 연구에 대한 기존의 접근방식은 위와 같은 복합유해물질에 대한 건강영향을 추정할 때 고려해야 할 점들 중 일부를 해결할 수 있지만, 뚜렷한 단점이 존재한다. 예를 들어, 군집방법(clustering method)은 연속변수 형태인 노출농도를 범주화하는 과정에서 정보의 누락을 야기할 수 있다. 통계적 학습 알고리즘(random forest 등)은 혼합물질의 변수선택에 활용될 수는 있지만 노출과 반응의 연관성 정도와 방향은 설명하기가 어렵다. 회귀모형(예: lasso 등) 내 변수선택방법의 경우, 일반적으로 혼합물의 상대적으로 단순한 모수 모형(relatively simple parametric model of mixture components)에 기반 둔다는 한계점이 있다. 계층적 모델 수식(hierarchical model formulation)은 개별 효과 추정치(individual effect estimates)를 그룹 평균(group means)으로 축소(shrinking)하여 상관관계가 높은 노출들 간의 특성을 고려하지만, 이러한 방법은 일반적으로 각 노출과 건강결과 간의 선형(linear) 연관성과 가법적(additive) 연관성을 가정하는 경우에 가능하다. Bobb JF 등(2015)은 Bayesian kernel machine regression(BKMR)을 복합물질의 건강영향을 추정하기 위한 새로운 방법으로 제안하였다. 이 방법을 수행하기 위해서는 건강결과를 smooth function인 'h'로 모델링하고, 교란변수들을 보정한 노출변수들의 kernel function을 사용한다. 건강결과는 여러 가지 노출물질들 중 일부 하위집단(subset)에만 의존적일 수 있기 때문에, BKMR은 혼합물질 중 건강영향에 연관성이 있는 일부 성분을 파악하기 위해 변수선택을 수행한다. 또한, 혼합물 성분의 다중공선성(collinearity)을 다루기 위해 혼합물의 구조에 대한 사전지식(prior knowledge)을 통합할 수 있는 계층적 변수선택 확장을 BKMR에 도입하였다.	혼합물 연구에 대한 기존의 접근 방식은 위와 같은 복합유해물질에 대한 건강영향을 추정할 때 고려해야 할 점들 중 일부를 해결할 수 있지만, 뚜렷한 단점이 존재한다. 예를 들어, 군집 방법(clustering method)은 연속변수 형태인 노출농도를 범주화하는 과정에서 정보의 누락을 야기할 수 있다. 통계적 학습 알고리즘(random forest 등)은 혼합물질의 변수 선택에 활용될 수는 있지만 노출과 반응의 연관성 정도와 방향은 설명하기가 어렵다. 회귀모형 내 변수선택방법(예를 들어 LASSO 등)은 여러 유해물질 사이의 높은 상관성을 모형에 반영하여 유해물질 중 일부를 선택하지만 일반적으로 결과 변수와 혼합물 사이의 관계를 상대적으로 단순한 모형(relatively simple parametric model of mixture components)으로 설계한다는 한계점이 있다. 계층적 모델 수식(hierarchical model formulation)은 개별 효과 추정치(individual effect estimates)를 그룹 평균(group means)으로 축소(shrinking)하여 노출 변수 사이의 높은 상관관계를 설명하지만, 이러한 방법은 일반적으로 각 노출과 건강 결과 간의 선형(linear)적, 가법적(additive) 연관성을 가정해야 한다. Bobb JF 등(2015)은 앞서 언급한 3가지 건강영향을 평가할 때 고려해야 할 점을 반영하기 위해 복합물질의 건강영향을 추정하기 위한 새로운 방법으로 Bayesian kernel machine regression(BKMR)을 제안하였다. 건강 결과는 여러 유해물질 중 일부에만 의존할 수 있기 때문에, BKMR은 혼합물질 중 건강영향에 연관성이 있는 일부 성분을 파악하기 위해 변수 선택을 수행한다. 또한, 혼합물 성분 사이의 높은 상관성을 고려하기 위해 혼합물의 구조에 대한 사전 지식(prior knowledge)을 통합하여 모형에 반영할 수 있도록 계층적 변수 선택법(hierarchical variable selection)을 BKMR에 도입하

수정 전	수정 후
	였다.
각 subject i = 1,2, ... n에서, Y_i를 건강결과, $z_i = (z_{i1},....z_{iM})^T$를 M개의 노출변수들(공기오염 등), x_i를 잠재적 교란변수들의 집합, $\epsilon_i \sim N(0, \sigma^2)$ (독립항등분포; i.i.d.) 이라고 두었을 때, 다음과 같은 수식으로 나타낼 수 있다. $$Y_i = h(z_i) + x_i^T\beta + \epsilon_i$$ 환경적 혼합물의 측면에서 보았을 때, h(·)는 일반적으로 혼합물 성분 간의 비선형성(non-linear) 및/또는 상호작용(interaction)을 통합하는 고차원적 노출-반응 함수들을 특징화시킨다. 이러한 상황에서, h(·)를 나타내거나 고차원적 모수 공간을 갖는 결과모델에 맞추기 위해 basis function set를 구체화하는 것은 어려울 수 있으므로, 본 논문에서는 kernel machine 표현을 사용하였다.	i번째 근로자 (i = 1,2, ... n)에 대하여 Y_i을 결과 변수, 노출 벡터 $z_i = (z_{i1},....z_{iM})^T$을 M개의 유해물질들로 이루어진 벡터 (납, 카드뮴, 미세먼지 등), x_i을 여러 교란 요인들로 이루어진 벡터, ε_i은 오차항이라 할 때, 결과 변수 Y_i을 다음과 같은 수식을 통해 표현할 수 있다. $$Y_i = h(z_i) + x_i^T\beta + \epsilon_i$$ 여기서 함수 h(·)는 일반적으로 혼합물 성분과 결과 변수 사이의 비선형성 (non-linear) 또는 혼합물 성분 간의 교호작용 (interaction)을 포함하는 고차원적 노출-반응 함수 (exposure-response function)를 나타내고, 오차항 ε_i는 평균이 0이며, 분산이 σ^2인 정규분포를 따른다. 하지만 고차원 노출-반응 함수 h(·)를 구체적으로 표현하는 것이 쉽지 않기 때문에 본 보고서에서는 커널 함수 (kernel function)를 기반으로 하는 kernel machine regression (KMR)을 사용하여 함수 h(·)를 표현하고자 하였다.
Kernel function인 K(z,z')에서 z는 한 개체의 환경적 노출 혹은 혼합물질의 벡터값을 의미하고(exposure profile), $(z_1,....z_M)^T$ 로 표현할 수 있다. z'는 $(z'_1,....z'_M)^T$ 두 번째 개체의 exposure profile을 의미한다. Gaussian kernel K(z,z')는 radial basis function(RBF)로도 불리며, 수식은 다음과 같다.	**2) BKMR의 개요** BKMR은 KMR 방법에 베이지안 변수 선택 접근법을 적용한 방법이다. BKMR에서 제공하는 변수 선택법은 구성 요소별 변수 선택법(component-wise variable selection)과 계층적 변수 선택법(hierarchical variable selection) 두 가지가 있다. 구성 요소별 변수 선택법을 설명하기 전에 위에서 설명한

수정 전	수정 후
$$K(z, z') = \exp\left(-\sum_{m=1}^{M} r_m (z_m - z_m{}')^2\right)$$ Bobb JF 등(2015)은 Gaussian kernel에 초점을 맞춰 살펴보았는데, 즉 KMR은 Gaussian process regression이라고도 볼 수 있으며 이는 머신러닝에서도 자주 등장하는 방법이다. KMR의 주요 아이디어는 특정 결과변수와 다수의 노출변수들 간의 연관성을 유연하게 모델링하는 데에 있다. Gaussian kernel의 경우에서도 h의 함수적 형태의 넓은 범위를 유연하게 잡아낼 수 있다. Gaussian kernel 하에서 z와 z'는 두 명의 다른 개인에서의 벡터 예측값을 의미하고, $r_m \geq 0$ 은 z_m 의 함수로 h의 부드러운 형태(smoothness)를 조정하는 튜닝 모수(parameter)를 의미한다. 직관적으로 보았을 때, 이러한 kernel function은 서로에 대한 비슷한 exposure profile을 갖고 있는 두 개인의 건강영향 추정을 축소한다. R 프로그램에서의 'bkmr' 은 위 모델에 대한 베이지안 추론을 수행하는 기능이 포함되어 있으며, kmbayes() 옵션을 이용해 모델을 적합(fitting)하고 다양한 방식으로 모델의 결과를 요약하고 시각적으로 나타낼 수 있는 기능을 수행할 수 있다. 아래에 기술된 2) Component-wise variable selection과 'bkmr' R 패키지의 기본 예제에서는 변수 선택방법 두 가지 중 하나인 Component-wise variable selection과 함께 'bkmr' R 패키지의 기본 예제를 설명한다. 3)Hierarchical variable selection에서는 두 가지 중 하나인 Hierarchical variable selection을 R로 구현한 예제를 설명한다. 4) 이분형 결과변수로의 확장에서는 이분형 결과변수를 포함한 자료를 'bkmr' R 패키지로 분석할 때 고려할 점을 설명한다. 5) 반복측정 자료로의 확장에서는 반복해서 측정된 자료를 'bkmr' R 패키지로 분석할 때 고려	kernel 행렬을 다음과 같이 다시 표현할 수 있다. $$K(z, z';r) = \exp\left(-\sum_{m=1}^{M} r_m (z_m - z_m{}')^2\right)$$ 이때 $r = (r_1, ..., r_M)^T$ 이고, r_m은 m번째 유해물질이 중요한 정도를 의미하며, 0과 1 사이의 값을 가진다. Component-wise variable selection은 유해물질 사이의 상관성을 고려하지 않고, 유해물질 각각을 독립된 하나의 물질로 여기는 변수 선택법이다. 또한, 구성 요소별 변수 선택법은 베이지안 방법에서 다중 회귀분석에서 변수를 선택할 때 사용하는 사전 분포(prior distribution)인 "slab-and-spike" 사전 분포를 사용하여 수식으로 표현할 수 있다. "slab"이란, 회귀 계수 값의 사전 분포(prior distribution)를 의미하고 "spike"는 회귀 계수의 값이 0이 될 확률이라고 볼 수 있다. "slab-and-spike" 사전 분포는 결과 변수와 여러 유해물질 사이의 사전 지식(prior knowledge)을 활용할 수 있다는 점에서 장점이 있다. $$r_m|\delta_m \sim \delta_m f_1(r_m) + (1-\delta_m)P_0, \quad m = 1, ..., M,$$ $$\delta_m \sim Bernoulli(\pi)$$ 위 수식에서 $f_1(\cdot)$는 r_m의 확률밀도함수이며, P_0는 r_m이 0의 값을 가질 확률밀도를 나타낸다. δ_m는 결과 변수에 m번째 유해물질이 영향을 미치는지 나타내는 지표로, 1이면 결과 변수에 m번째 유해물질이 영향을 주는 것으로, 0이면 영향을 주지 않는 것으로 해석할 수 있으며, π는 m번째 유해물질이 결과 변수에 영향을 주

수정 전	수정 후
할 점을 설명한다. 6) 분석 속도를 높이는 방법는 'bkmr' R 패키지의 단점인 느린 분석속도를 향상시키는 knot 옵션에 대하여 설명한다. 2) Component-wise variable selection과 'bkmr' R 패키지의 기본 예제 베이지안 기법 측면에서의 변수선택을 수행할 때, augmented Gaussian kernel function을 정의하여야 하고, 아래와 같이 표현할 수 있다. $$K(z, z';r) = \exp\left(-\sum_{m=1}^{M} r_m (z_m - z_m')^2\right)$$ 이때 $r = (r_1, \ldots., r_M)^T$ 이고 $K_{z,r}$ 을 (i, j)요소가 $K(z_i, z_j; r)$과 동일한 $n \times n$ 행렬이라고 정의한다. 이때 우리는 다중회귀문제에서의 베이지안 변수 선택 방법과 동일한 "slab-and spike" 가정을 하고 수식으로 표현하면 아래와 같다. "slab"이란, 회귀 계수값에서의 이전 분포(prior distribution)을 의미하고 "spike"는 모델에서 0의 값을 갖는 특정 계수의 확률이라고 볼 수 있다. 이러한 "slab-and-spike" 가정은 모델의 이전 지식(prior knowledge)을 활용할 수 있다는 점에서 베이지안 변수선택 기법으로서의 장점을 갖고 있다. $$r_m|\delta_m \sim \delta_m f_1(r_m) + (1-\delta_m) P_0, \quad m = 1, \ldots, M,$$ $$\delta_m \sim Bernoulli(\pi)$$	는 요인으로 뽑힐 확률을 의미한다. 지표 δ_m의 사후 평균은 m번째 유해물질이 혼합물에서 상대적으로 중요한 요소인지 나타내는 사후 포함 확률(posterior inclusion probability; PIP)로 해석할 수 있다. 혼합물의 구성성분 사이의 상관성이 높을 경우, 위에서 언급한 구성 요소별 변수 선택법은 상관성 있는 구성성분들을 구분하는 데 어려움이 생기기 때문에 실질적으로 적용하기 힘들 수 있다. 따라서 구성성분 간의 상관성을, 즉 혼합물의 구조에 대한 정보를 모형에 통합할 수 있는 계층적 변수선택법을 추가로 고려할 수 있다. 계층적 변수 선택법은 상관성이 높은 구성성분들을 하나의 그룹으로 묶어 혼합물을 구성하는 성분들을 여러 개의 그룹으로 분할한 후, 각 그룹 별로 중요한 변수를 선택하는 변수 선택법이다. 혼합물 $z_1, \ldots., z_M$이 총 G개의 그룹$(S_1, S_2, \ldots, S_G)$으로 분할된다고 가정해보자. 이때 그룹 간 상관성은 낮으면서 그룹 내 유해물질 간의 상관성은 높도록 분할할 수 있다. 예를 들어, 대기오염원에 대한 정보가 일반적으로 알려져 있으며, 이러한 정보는 오염물질을 여러 그룹으로 분할할 때 사용될 수 있다. 구성 요소별 변수 선택법과 비교하여 계층적 변수 선택법은 혼합물의 구성성분들을 여러 그룹으로 분할하는 과정이 추가된 변수 선택법이다. 따라서 구성 요소별 변수 선택법에서 결과 변수에 영향을 주는 유해물질을 파악하는 지표인 δ_m에 대응되는 지표로 결과 변수에 영향을 주는 그룹을 결정하는 지표 ω_g 와 g번째 그룹이 결정되었을 때, 그룹 S_g에서 중요한 유해물질을 선택하는 지표인 δ_{S_g}가 구성 요소별 변수 선택법과 비교하여 변수 선택과정에 추가된다.

수정 전	수정 후
위 수식에서 $f_1(\bullet)$는 r_m의 확률밀도함수 형태이고, P_0는 0 지점에서의 밀도를 나타낸다. δ_m는 지표의 사후평균으로, 성분 m이 혼합물의 중요한 요소 또는 성분 m의 사후 포함 확률의 형태로 해석될 수 있다.	$$\delta_{S_g}\|\omega_g \sim Mltinomial(\omega_g, \pi_{S_g}), \quad g=1,\ldots G,$$ $$\omega_g \sim Bernoulli(\pi)$$ 이때, π는 g번째 그룹이 결과 변수에 영향을 주는 범주로 뽑힐 확률을 의미하며, π_{S_g}는 그룹 S_g 안에 있는 구성 성분들이 결과 변수에 영향을 미치는지 결정하는 확률을 나타내는 벡터이다. 비록 이러한 방법을 사용할 때, 동일한 그룹의 두 구성 요소가 결과 변수에 대하여 독립적이거나 상호작용이 없다고 가정해야 하지만, 그룹 내 높은 상관관계가 있는 경우에서는 이러한 효과를 보다 일반적인 모델에서 식별하기는 힘들다.
다음은 R 프로그램으로 표현한 예제를 들어 component-wise variable selection을 설명하면 다음과 같다.	R 프로그램의 'bkmr' 패키지에는 BKMR의 수행에 필요한 다양한 함수들이 포함되어 있으며, 특히 kmbayes() 함수는 모형 적합(fitting)에 있어 가장 중요한 함수이다. 그 외 제공되는 함수들을 통해 다양한 방식으로 모형의 결과를 요약하고 시각적으로 나타낼 수 있는 기능을 수행할 수 있다. 다음은 R 프로그램의 'bkmr' 패키지를 설치하고 불러오는 코드이다.
임의의 데이터셋을 생성하여 R 패키지인 'bkmr'을 활용해보도록 한다.	임의의 데이터셋을 생성하여 R 패키지 'bkmr'에서 제공하는 함수들에 대하여 설명하고자 한다. SimData(n, M)은 모의 실험 자료를 생성하는 함수이며, 이때 n은 자료의 수, M은 유해물질의 수를 의미한다. 결과의 재현성을 위해 자료 생성 시 111 seed 번호를 사용하였다. 'y'는 결과의 벡터값을, 'Z'는 노출 변수의 행렬, 'X'는 공변량(covariate)을 포함하는 행렬을 나타낸다.
데이터 생성에 사용된 실제 노출-반응 함수를 살펴본다. res의 결과는 아래의 그래프와 같다.	모의 실험자료 생성에 사용된 참(true) 노출-반응 함수를 살펴보고자 한다. 유해물질의 수가 4개이고, 결과 변수와 유해물질의 사이의 그림을 그리기 위해서는 5차원이 필요하기 때문에 본 장에서

수정 전	수정 후
	는 첫 번째 유해물질과 두 번째 유해물질에 대해 노출된 양에 따라 결과 변수가 어떻게 변하는지 3차원 그림으로 표현하였으며, 그 결과는 아래의 그래프(res)와 같다. 아래의 그림을 통해 첫 번째 유해물질과 두 번째 유해물질이 증가하면 결과 변수가 증가하는 것을 확인할 수 있다.
그 다음으로, BKMR 모델을 적합하기 위하여 앞서 설명한 kmbayes() 함수를 활용하도록 한다. 이 함수는 Markov chain Monte Carlo(MCMC) 알고리즘에 기반한 것으로, 'iter' 은 MCMC sampler의 반복횟수를, 'y'는 결과의 벡터값을, 'Z'는 노출변수의 행렬, 'X'는 공변량 행렬을 나타낸다.	그 다음으로, BKMR 모델을 적합하기 위하여 앞서 언급한 kmbayes() 함수를 사용하고자 한다. BKMR은 모수를 추정할 때, Markov chain Monte Carlo(MCMC) 알고리즘을 사용하며, 알고리즘을 얼마나 반복시킬지 그 반복횟수가 필요하다. MCMC 알고리즘을 실행시킬 때 필요한 반복횟수는 'iter' 인수(argument)을 통해 결정할 수 있다.
얼마나 다양한 모수 값들이 sampler가 작동하면서 변하는지를 보기 위하여 시각적으로 나타내기 위한 작업을 진행한다.	MCMC를 통해 모수 값들이 안정적으로 수렴하는지 trace plot을 통해 시각적으로 확인하고자 한다. 어떤 모수에 대하여 trace plot을 그릴지 par 인수를 통해 지정할 수 있다.
	comp 인수를 통해 r_m의 수렴 여부를 시각적으로 확인할 수 있으며, 아래의 코드는 r_1의 수렴성을 확인하기 위한 plot을 그리는 코드이다.
각 노출그룹(z_{im})의 사후 포함확률(Posterior Inclusion Probability; PIP)은 다음과 같이 추정할 수 있다.	각 유해물질마다 추정된 사후 포함확률(PIP)은 다음과 같이 확인할 수 있다. 추정된 사후 포함확률을 크기순으로 열거하면 z1, z2, z4, z3임을 확인할 수 있으며, z1과 z2가 결과 변수 y에 중요하게 기여하고 있으며, z3의 기여가 가장 낮음을 확인할 수 있다.
또한, PIP는 다음과 같이 그림으로 표현할 수 있다.	또한, 사후 포함확률은 다음의 코드를 통해 시각적으로 표현할 수 있다.

수정 전	수정 후
이제 BKMR 분석결과를 요약하는데 사용할수 있는 'bkmr' 패키지에 포함된 다양한 기능을 살펴본다. 이 기능은 노출-반응의 연관성을 시각화하고, 그 외에 복합노출의 의한 건강영향을 평가할 수 있는 다양한 통계 결과를 시각적으로 표현한다. 연구자들은 BKMR 분석결과 중 $h(\bullet)$를 다음과 같이 시각화할 수 있는데, 우리는 고차원으로 표현하기 어렵기 때문에, 대신 하나 또는 두개의 노출과 결과의 관계에 초점을 맞추고 나머지 노출을 특정 값으로 고정한 상태로 시각적으로 표현한다. 다음은 나머지 노출변수는 특정 percentile로 고정시켰을 때, 하나의 특정 노출변수와 결과변수 간의 관계를 보여준다. 예를 들어, z2, z3, z4는 50 percentile로 농도를 고정시켰을 때, z1과 결과변수 간의 노출-반응 관계를 보여준다. 여기서 관심있는 분석결과는 각 zm과 결과 간의 일변량 관계이며, 여기서 다른 모든 노출은 특정 percentile로 고정된다. 이는 PredictorResponseUnivar 함수를 사용하여 수행할 수 있다. 다른 노출의 percentile을 지정하는 인수(argument)는 q.fixed로 지정하며, 기본 값은 50 percentile에 해당하는 'q.fixed = 0.5'이다.	지금까지 BKMR을 적합하고, 안정적으로 모수들이 추정되었는지 확인할 때 필요한 함수에 대하여 설명하였다. 하지만 지금부터는 적합된 BKMR의 분석결과를 요약할 때, 사용하는 'bkmr' 패키지에 포함된 다양한 함수들을 살펴보고자 한다. 이 함수들은 유해물질과 결과 변수 사이의 연관성을 시각화하고, 그 외에 복합물질 노출에 대한 건강영향을 평가할 수 있는 다양한 통계 분석결과를 시각적으로 표현한다. 연구자들은 BKMR 분석결과를 통해 고차원적 노출-반응 함수 h(·)를 시각화할 수 있다. 하지만 3차원 이상의 고차원을 그림을 통해 표현하기 어려우므로, 대안으로 관심 있는 변수를 하나 또는 두 개 결정하여 해당 유해물질과 결과 사이의 관계를 시각화한다. 이때 관심 있는 변수를 제외한 나머지 노출을 특정 값으로 고정한 상태로 관심 있는 변수들과 결과 사이의 관계를 시각적으로 표현한다. 함수를 시각적으로 표현할 때, 관심 있는 변수가 한 개인 경우 PredictorResponseUnivar 함수를 사용하고, 두 개인 경우에는 PredictorResponseBivarLevels 함수를 사용한다. 다음의 코드는 z1, z2, z3, z4 각각 하나의 변수에 대하여 노출-반응 함수를 시각적으로 표현하기 위해 사용한 코드이다. 관심 있는 변수(예; z1)를 제외한 나머지 유해물질(예; z2, z3, z4)의 값은 특정 percentile로 고정시킨 후 관심 있는 변수의 값을 변화해 나가며, 하나의 유해물질과 결과 사이의 노출-반응 관계를 시각화한다. 다른 유해물질의 percentile을 지정하는 인수는 q.fixed이며, 기본 값은 50 percentile로 설정되어 있다. 만약 70 percentile로 설정을 원하는 경우에는 'q.fixed = 0.7'를 입력하면 된다.
위의 예를 바탕으로 다른 모든 노출변수가 특정 percentile로 고정되어있을 때, 두 개의 노출변수에 대한 이변량 노출-반응 함수를 시각화할 수 있다. 이 접근은 두 번째 노출변수를 여러	위의 예에서는 관심 있는 유해물질이 한 개인 경우를 설명하였다. 관심 있는 유해물질이 2개인 경우, 그 유해물질들을 제외한 나머지 모든 유해물질이 특정 percentile로 고정되어있을 때, 두 유

수정 전	수정 후
percentile로 고정하였을 때, 첫 번째 노출변수와 결과변수 간의 노출반응 함수를 시각화 한다; 이 두개의 노출변수 외 나머지 노출변수는 특정 값으로 고정한다. 이는 PredictorResponseBivarLevels 함수를 사용하여 수행할 수 있다. 여기서 두 번째 노출변수의 percentile은 인수(argument) qs로 여러 개의 값을 지정한다.	해물질에 대한 이변량 고차원적 노출-반응 함수 h(·)를 시각화하고자 한다. 이때 시각화하는 방식은 두 번째 유해물질을 여러 percentile로 고정한 후, 첫 번째 유해물질과 결과 간의 노출-반응 함수를 시각화한다. 관심 있는 유해물질이 1개인 경우와 마찬가지로 두 유해물질을 제외한 나머지 유해물질은 특정 percentile 값으로 고정한다. 이는 PredictorResponseBivarLevels 함수를 사용하여 수행할 수 있다. 여기서 qs 인수에 여러 percentile 값을 지정함으로써 두 번째 유해물질에 대해 고정할 percentile을 지정할 수 있다.
추정된 노출-반응 함수 h를 시각적으로 표현하는 것 외에도, 모든 노출변수가 50 percentile일 때와 비교하여 모든 노출변수가 특정 percentile에 있을 때의 h 값을 비교하여 노출변수 전체의 효과를 계산할 수 있다. 이는 OverallRiskSummaries 함수와 인수(argument) qs를 사용하여 percentile의 시퀀스를 지정하고 인수(argument) q.fixed를 사용하여 고정 percentile(기본 값은 50 percentile)을 지정할 수 있다.	지금까지 관심 있는 유해물질이 1개 또는 2개의 유해물질 노출량에 따라 결과가 어떻게 변하는지 그 형태를 시각화하여 확인하였다. 하지만 bkmr 패키지에서는 각 유해물질 별 건강영향에 미치는 효과의 형태뿐만 아니라 모든 유해물질에 대한 노출량이 변화하였을 때, 건강영향에 미치는 효과의 값을 요약할 수 있으며, 그 형태 또한 확인할 수 있다. 모든 유해물질의 노출량이 특정 percentile에 있을 때의 결과 값을 모든 유해물질에 대한 노출량이 50 percentile일 때와 비교하여 유해물질 노출량 전체의 효과를 계산할 수 있다. 전체 효과는 OverallRiskSummaries 함수와 qs 인수를 사용하여 비교하고 싶은 percentile을 지정하고 q.fixed 인수를 사용하여 비교 기준인 percentile(기본 값은 50 percentile)을 지정하여 요약할 수 있다.
또한, 노출변수 전체의 효과를 시각적으로도 표현할 수 있는데, 이때 'ggplot' 패키지를 사용한다.	또한, 유해물질 노출량 전체에 대한 효과를 다음의 코드를 사용하여 시각적으로도 표현할 수 있다.
연구자들이 관심 있을 수 있는 h에 대한 또 다른 정보는 결과변수에 대한 각 노출변수의 기여도이다. 예를 들어, h 내에서 특정 노출변수가 75 percentile일 때와 그 노출변수가 25 percentile	연구자들이 관심 있을 수 있는 또 다른 정보는 결과 변수에 대한 각 노출 변수의 기여도이다. 이때 그 기여도는 특정 하나의 노출변수에 대하여 두 percentile에서의 위험을 비교하여 요약되며, 그

수정 전	수정 후
일 때의 위험을 비교하고 싶을 때 다음과 같이 시각화할 수 있다; 이 때 나머지 노출변수는 모두 특정 percentile로 고정된다. 이러한 시각화 정보의 의미는 단일 노출변수의 건강위험이라고 할 수 있고, SingVarRiskSummaries 함수를 사용하여 계산할 수 있다. 위험을 비교할 percentile을 지정할 때에는 인수(argument) qs.diff를 사용하고, 나머지 노출변수 값의 시퀀스를 고정할 때에는 인수(argument) q.fixed를 사용한다.	외 나머지 노출 변수는 모두 특정 percentile로 고정된다. 이러한 기여도를 단일 노출 변수 효과(single variable effect)라 할 수 있고, SingVarRiskSummaries 함수를 사용하여 계산할 수 있다. 위험을 비교할 percentile은 qs.diff 인수를 통해 값을 지정할 수 있으며, 나머지 고정할 노출 변수의 값은 q.fixed 인수를 통해 값을 지정할 수 있다.
위에서 계산된 예를 다음과 같이 시각화하면, 다음과 같다. 아래 그림은 노출변수 z1이 75 percentile에서 25 percentile로 바뀌었을 때, 위험의 변화를 보여주고 있다. 또한, 여기서 나머지 노출변수 z2, z3, z4는 모두 25 percentile로 고정되었을 때(빨강), 모두 50 percentile 고정되었을 때(초록), 모두 75 percentile(파랑)로 고정되었을 때 결과를 보여주고 있다.	위에서 계산된 결과를 아래의 코드를 사용하여 시각화할 수 있다. 아래 그림은 각 노출 변수마다 노출량이 75 percentile에서 25 percentile로 바뀌었을 때, 위험의 차이를 보여주고 있다. 여기서 나머지 노출 변수가 모두 25 percentile로 고정되었을 때(빨강), 모두 50 percentile 고정되었을 때(초록), 모두 75 percentile(파랑)로 고정되었을 때 결과를 보여주고 있다.
위의 예제를 보면, 노출변수 z3과 z4는 위험에 기여하지 않으며, z1과 z2는 노출수준이 높을수록 h 함수의 값을 더 높이는 것을 알 수 있다. 또한, 위의 그림을 보면 z1의 경우, 나머지 노출변수의 값이 25 percentile에서 75 percentile로 증가할 때 z1 노출에 의한 위험이 증가함을 보여주고 있다. z2에서도 이와 유사한 패턴을 보이고 있다. 즉, 위의 그림은 z1과 z2의 상호작용의 가능성을 나타낸다. 이러한 상호작용의 가능성을 좀 더 명확하게 하기 위하여, 상호작용 매개변수를 계산할 수 있다. 예를 들어, z1 외에 다른 모든 노출변수들이 75 percentile로 고정되었을 때 z1의 위험을, z1 외에 다른 모든 노출변수들이 25 percentile로 고정되었을 때 z1의 위험과 비교할 수 있다. 이는 이전 그림에서 파란색 원으로 표시된 추정치에서 빨간색 원으로 표시된 추정치를 빼는 것에 해당한다. 이는 SingVarIntSummaries 함수를 사용하여 수행할 수 있다.	위의 예제를 보면, 노출 변수 z3과 z4에 대해서 해당 변수의 노출량이 75 percentile과 25 percentile로 고정되었을 때의 차이가 0에 가까우므로 결과 변수에 대한 기여도가 크지 않다고 할 수 있다. 하지만 노출 변수 z1 z2에 대해서는 다른 노출 변수에 대한 노출량이 높을수록 결과 변수에 대해 노출 변수 z1과 z2의 노출량이 75 percentile과 25 percentile로 고정되었을 때의 차이가 점점 커지는 것을 확인할 수 있다. 즉, 위의 그림은 노출 변수 z1과 z2의 교호작용의 가능성을 나타낸다. 이러한 교호작용의 가능성을 명확하게 확인하기 위하여, 교호작용 모수에 대한 추정치를 각 노출 변수마다 계산할 수 있다. 예를 들어, z1 외 다른 모든 노출 변수들이 75 percentile로 고정되었을 때 z1의 위험을, z1 외에 다른 모든 노출 변수들이 25 percentile로 고정되었을 때 z1의 위험과 비교하여 교호작용의 크기를 확인할 수 있다. 이는 이전 그림에서 파란색 원으로 표시된

수정 전	수정 후
	추정치에서 빨간색 원으로 표시된 추정치를 빼는 것에 해당하며, SingVarIntSummaries 함수를 사용하여 수행할 수 있다.
3) Hierarchical variable selection 혼합물질이 매우 상관성이 높은 경우, 위에서 언급한 수식들은 데이터가 상관성 있는 요소들을 구분하는 데에 어려움이 있기 때문에, 실질적으로 적용하기 힘들 수도 있다. 따라서 이러한 경우, 혼합물의 구조에 대한 정보를 모델에 통합하는 계층적 변수선택법을 활용할 수 있다. 혼합물질 $z_1, \ldots . z_M$ 이 $S_g\,(g=1,\ldots . G)$ 그룹으로 분할된다고 가정해보면, 이 때 그룹 간 상관성은 낮은 반면에 그룹 내 상관성은 높을 수 있다. 예를 들어 대기오염원에 대한 정보가 일반적으로 알려져 있고, 이는 오염물질을 그룹화 하는 데 사용할 수 있다. 이러한 지표 변수들이 위에서 언급한 “slab-and-spike” 가정 하에서의 변수라고 보았을 때, 분포의 수식은 다음과 같다. $$\delta_{S_g}|\omega_g \sim Mltinomial(\omega_g, \pi_{S_g}),\quad g=1,\ldots G,$$ $$\omega_g \sim Bernoulli(\pi)$$ 이때 $\delta_{S_g} = (\delta_m)_{z_m} \epsilon S_g$이고, 이는 지표변수들의 벡터라고 볼 수 있다. π_{S_g} 는 그룹 S_g안에 있는 혼합성분 Z_m에 대한 사전확률에 대응하는 벡터값이다. 이러한 방법은 그룹의 최대 단일요소(상관성이 높은 구성요소)가 한번에 모델에 들어갈 수 있다. 비록 이러한 방법을 사용할 시, 동일한 그룹의 두 구성요소가 건강결과에 대하여 독립적이거나 상호작용이 없다고 가정해야 하지만, 높은 그룹 내 상관관계가 있는 경우에서는 이러한 효과를 보다 일반적인 모델	2) BKMR의 개요 부분으로 올려서 설명함.

수정 전	수정 후
에서 식별하기는 힘들다. 계층적 변수선택법에 대한 예제를 들면 아래와 같고, 이때 두 개의 노출변수의 상관성이 매우 높다고 가정해본다.	
	지금까지 구성 요소별 변수 선택법에 기초하여 BKMR 분석결과를 요약하고 시각화하는 방법에 대해 설명하였다. 계층적 변수 선택법 또한 비슷한 방식으로 BKMR 결과를 요약하고 시각화할 수 있다. 계층적 변수 선택법의 사용을 설명하기 위해 4개의 유해물질 중 2개의 유해물질이 높은 상관성을 갖는다고 가정하고, 모의실험 자료를 생성하였다. 생성된 모의실험 자료로부터 유해물질 사이의 상관계수(correlation coefficient)을 구해보면 유해물질 z1과 z3 사이의 상관계수가 다른 유해물질 사이의 상관계수와 비교하여 높게 생성되었다는 것을 확인할 수 있다.
노출 z1, z3이 다른 노출변수에 비하여 상관성이 매우 높음을 알 수 있다. 'Component-wise variable selection'의 대안으로서의 'hierarchical variable selection'은 상관성 높은 노출변수들을 각각 겹치지 않는 그룹으로 분류하여 진행한다. 앞서 설명하였던 'Component-wise variable selection' 결과와 비교하기 위해 R 코드를 정리하면 다음과 같다.	'구성 요소별 변수 선택법'의 대안으로서의 '계층적 변수 선택법'은 앞서 설명한 것과 같이 상관성 높은 유해물질들을 그룹 간 상관성은 낮지만 그룹 내 상관성은 높도록 분류를 진행한다. 앞서 재현하였던 '구성 요소별 변수 선택법' 결과와 '계층적 변수 선택법' 결과를 비교하기 위한 코드는 다음과 같다. 계층적 변수 선택법을 적용할 때, 구성 요소별 변수 선택법과의 차이점은 group 인수가 추가된다는 것이다. group 인수를 통해 각 유해물질이 어떤 그룹으로 분류할지 혼합물의 구조에 대한 사전 지식을 반영할 수 있다.
두 방법에 대한 Posterior Inclusion Probabilities(PIPs)를 비교하면 다음과 같다.	두 방법에 대한 사후 포함확률을 비교하면 다음과 같다.
'fitkm_corr'의 'Component-wise variable selection'의 경우, PIPs의 결과에서 z2가 모델에 포함되어야 하는 반면에 z1이	'fitkm_corr'는 구성 요소별 변수 선택법을 적용하여 BKMR을 적합한 결과에서 사후 포함확률에 대한 요약 결과를 보면 유해물질

수정 전	수정 후
들어가야 하는 증거는 강하지 않음을 알 수 있다. 반면, ‘fitkm_hier’의 ‘Hierarchical variable selection’의 경우, Group1의 group-specific PIP를 추정하며, Group1의 오염물질들 중 하나는 모델에 들어가야 함을 알 수 있고 z1과 z3 중에 condPIP의 값이 z1에서 더 높기 때문에 z1이 결과변수와 더 상관성이 높은 것을 확인하였다. 이러한 결과를 바탕으로, ‘Component-wise variable selection’에서는 z2만을 선택하여 잘못된 결과를 도출할 수도 있는 반면 ‘Hierarchical variable selection’ 에서는 z1과 z2를 결과의 예측변수로 올바르게 선택될 수 있음을 알 수 있다.	z2가 높은 사후 포함확률을 가지고 있기 때문에 z2가 모형에 포함되어야 하는 반면 z1이 들어가야 하는 증거는 강하지 않음을 알 수 있다. 계층적 변수 선택법을 적용한 BKMR 분석 결과인 ‘fitkm_hier’에서 그룹에 대한 사후 포함확률(group-specific PIP)에 대한 요약 결과를 보면 그룹 1과 2에 대해 높은 사후 포함확률이 추정되었음을 확인할 수 있다. 그러므로 그룹 1과 2 각각에 속한 유해물질 중 1개의 유해물질이 결과 변수에 영향을 준다는 것을 확인할 수 있다. 이때, 그룹 1에 속한 유해물질 z1과 z3에 대해서 유해물질 z1과 z3의 사후 포함 확률(condPIP) 중 유해물질 z1에서 사후 포함확률이 더 크기 때문에 유해물질 z1이 결과 변수와 더 상관성이 높은 것으로 확인하였다. 그룹 2의 경우, 속한 유해물질이 z2 1개이기 때문에 유해물질 z2에 대한 사후 포함확률(condPIP)가 1로 나타났다. 이러한 결과를 바탕으로, ‘구성 요소별 변수 선택법’은 유해물질 z2만을 선택하여 잘못된 결과를 도출할 수도 있지만 ‘계층적 변수 선택법’ 에서는 z1과 z2를 결과에 대한 유해물질로 올바르게 선택될 수 있음을 알 수 있다.
4) 이분형 결과변수로의 확장 이분형 결과변수를 사용하여 BKMR 모델을 fitting 하기 위하여 앞서 연속형 결과변수에 적용한 바와 같이 kmbayes() 함수를 활용하도록 한다. 이 함수는 Markov chain Monte Carlo(MCMC) 알고리즘에 기반한 것으로, ‘iter’ 은 MCMC sampler의 반복횟수를, ‘y’는 결과의 벡터값을, ‘Z’는 노출변수의 행렬, ‘X’는 공변량 행렬을 나타낸다. 다만, 이분형 결과변수의 경우 다음과 같이 family를 “binomial”로 지정해주는 코드를 추가로 사용하면 된다. 이후 과정은 위에서 설명한 과정과 동일하다.	BKMR을 연속형 변수뿐만 아니라 이분형 결과 변수(binary outcome)로도 확장하여 사용할 수 있다. 이분형 결과 변수에 대해서 앞서 설명한 연속형 결과 변수와 마찬가지로 kmbayes() 함수를 사용하여 BKMR을 적합한다. 연속형 결과 변수에서 사용한 모든 인수들을 이분형 결과 변수에 BKMR을 적용할 때, 동일하게 사용할 수 있다. 다만 이분형 결과 변수의 경우 로지스틱 회귀분석(logistic regression)을 glm 함수를 통하여 적합할 때 사용하는 family 인수에 “binomial”로 지정해주는 코드를 BKMR에서도 반드시 추가로 사용해야 한다. 이후 BKMR 분석결과를 요약하고 시각화하는 과정은 위에서 설명한 연속형 결과 변수와 같은 방식으로 진행할 수 있다.

수정 전	수정 후
5) 반복측정 자료로의 확장 한 개인의 정보를 반복하여 측정한 자료를 사용하여 BKMR 모델을 fitting 하기 위하여 앞서 적용한 바와 같이 kmbayes() 함수를 활용하도록 한다. 이 함수는 Markov chain Monte Carlo(MCMC) 알고리즘에 기반한 것으로, 'iter' 은 MCMC sampler의 반복횟수를, 'y'는 결과의 벡터값을, 'Z'는 노출변수의 행렬, 'X'는 공변량 행렬을 나타낸다. 다만, 개인의 id에 해당하는 정보를 subject로 지정해주시고, kmbayes() 함수 내에서 id를 subject로 지정해주시는 코드를 추가로 사용하면 된다. 이후 과정은 위에서 설명한 과정과 동일하다.	BKMR은 횡단 자료(cross-sectional data)뿐 아니라 반복측정 자료(repeated measurements) 분석을 위해서 사용될 수 있다. kmbayes() 함수를 사용하여 개인의 정보를 반복하여 측정한 자료를 BKMR으로 적합이 가능하다. 이러한 자료에 대해서도 연속형 결과 변수에서 다루었던 여러 인수, 요약 및 시각화하는 함수를 같은 방식으로 사용이 가능하다. 다만, 반복측정된 자료가 같은 환자로부터 측정되었다는 것을, 즉 개인의 id에 해당하는 정보를 id 인수로 지정한 후, kmbayes() 함수를 사용하면 된다. 이후 과정은 위에서 설명한 연속형 결과 변수와 동일하다.
6) 분석 속도를 높이는 방법 BKMR은 Markov Chain Monte Carlo(MCMC) 과정에서 $n_1 \times n_2$ 행렬의 역행렬을 여러 번 계산해야 하고, 이로 인해 BKMR의 fitting 과정에서 많은 시간이 소요된다; 여기서 n_1은 노출 1의 관측치 개수, n_2는 노출 2의 관측치 개수인데, BKMR로 분석하기 위해서는 연구대상자 별로 노출 1과 노출 2의 측정치가 함께 측정되어야 하므로 n_1과 n_2의 관측치 수는 같다. 따라서, 분석 속도를 줄이기 위해여, Bobb JF 등(2018)은 공간데이터 분석에 사용되는 Gaussian predictive process(Benerjee et al., 2008)를 사용하여, 분석속도를 향상시키는 knot 옵션을 소개하였다. 이 knot 옵션은 실제로 관측된 n 개의 수(예: 100) 보다 더 적은 수인 m 개의 수(예:10)로 knot를 설정하여, $n_1 \times n_2$ 행렬(예: 100×100)을 계산하는 것이 아니라 $m_1 \times m_2$ 행렬(예: 10×10)로 계산하게 하여, 분석속도를 향상시킨다. 쉽게 말하자면 $n_1 \times n_2$ 행렬(예: 100×100)에 해당하는 지점들 중 일부인 $m_1 \times m_2$ 행렬(예: 10×10)의 개수에 해당하는 일부 지점들만 선정하여 분석하	BKMR은 MCMC 과정에서 역행렬을 여러 번 계산해야 하므로 BKMR을 적합하는 과정에서 많은 시간이 소요된다. 이에 따라 소요되는 분석 시간을 줄이기 위하여, Bobb JF 등(2018)은 공간데이터 분석에 사용되는 Gaussian predictive process(Benerjee et al., 2008)를 사용하여, 분석 속도를 개선시키기 위한 knot 옵션을 소개하였다. 이 knot 옵션은 실제로 관측치의 수(예: 100)보다 더 적은 수인 m개의 수(예:10)로 knot를 설정하여, 즉 관측치에 해당하는 지점 중 일부 지점들만 선정하여 분석하는 것을 말한다. knot 인수는 다음의 코드를 통해 적용할 수 있으며, 아래 예시에서는 knot의 개수를 10으로 설정하였다. 다만, 주의할 점을 반복측정된 자료에 대해서는 현재 'bkmr' R 패키지에서 knot 옵션을 제공하지 않는다는 점이다.

수정 전	수정 후
는 것을 말한다. 또한, knot의 개수가 m개라면 노출변수 공간을 n차원에서 m차원으로 축소시키는 것으로도 설명할 수 있다. knot는 다음과 같이 'bkmr' R 패키지에서 적용할 수 있다. 아래 예시에서는 knot의 개수를 10으로 설정하였다.	
	KMR을 설명하기 앞서 KMR에 사용된 커널을 이해하기 위한 몇 가지 개념에 대하여 간략하게 소개하고자 한다. 먼저 노출 벡터 $z_i = (z_{i1}, z_{i2}, \ldots, z_{iM})^T$은 근로자 i에게 노출된 유해물질의 양(exposure profile)을 나타내며, 같은 방식으로 노출 벡터 z_j은 근로자 j에게 노출된 유해물질의 양을 의미한다고 하자. 이때, 길이가 M인 노출 벡터를 길이가 M보다 큰 벡터로 변환이 가능한데, 예를 들어, 유해물질의 개수가 2인 노출 벡터를 생각해보자. 이러한 경우, 근로자 1에게 노출된 유해물질의 양을 나타내는 노출 벡터 $z_1 = (z_{11}, z_{12})^T$라 하면, 이 노출 벡터 z_1을 길이가 3인 새로운 노출 벡터 $\phi(z_1) = (z_{11}^2, z_{12}^2, \sqrt{2}\, z_{11} z_{12})^T$로 변환하여 표현이 가능하다. 이렇게 변환된 새로운 노출 벡터를 공변량으로 생각하여 $h(z_1) = \sum_{j=1}^{3} w_j \phi_j(z_1)$ 으로 결과 변수 Y_1을 모형화할 때, 새로운 노출 벡터를 사용할 수 있다. 하지만 커널 함수를 사용하여 $\sum_{j=1}^{3} w_j \phi_j(z_1) = \sum_{i=1}^{n} K(z_i, z_1)\alpha_i$로 표현할 수 있으며 (이때 n은 표본의 수임), 커널 함수 $K(\cdot, \cdot)$로는 linear 커널, polynomial 커널, Gaussian 커널 등이 사용된다. 위의 예에서 근로자 i의 노출 벡터 z_i을 새로운 노출 벡터 $\phi(z_i)$로 변환하는 함수 $\phi(\cdot)$에 대응되는 커널 함수가 바로 polynomial 커널 함수이다. 이 커널 함

수정 전	수정 후
	수를 이용하여 n x n 커널 행렬을 구성할 수 있으며, (i,j)-원소는 $K(z_i, z_j) = (z_{i1}z_{j1} + z_{i2}z_{j2})^2$으로 산출이 가능하다. KMR의 주요한 아이디어는 결과 변수와 여러 유해물질 변수들 사이의 연관성을 유연하게 모형화하는 것이다. 위에서 언급한 것과 같이 함수 $\phi(\cdot)$을 이용하여 새로운 노출 벡터를 만들 수 있으며, 함수에 따라 이 새로운 노출 벡터는 이차항, 교호작용 항을 포함한다. 이러한 특징을 이용하여 결과 변수와 여러 유해물질 변수들 사이의 관계를 유연하게 모형화하는 것이 가능하다.
이때 그룹 간 상관성은 낮으면서 그룹 내 유해물질 간의 상관성은 높도록 분할할 수 있다. 예를 들어, 대기오염원에 대한 정보가 일반적으로 알려져 있으며, 이러한 정보는 오염물질을 여러 그룹으로 분할할 때 사용될 수 있다. 구성 요소별 변수 선택법과 비교하여 계층적 변수 선택법은 혼합물의 구성성분들을 여러 그룹으로 분할하는 과정이 추가된 변수 선택법이다. 따라서 구성 요소별 변수 선택법에서 결과 변수에 영향을 주는 유해물질을 파악하는 지표인 δ_m에 대응되는 지표로 결과 변수에 영향을 주는 그룹을 결정하는 지표 ω_g 와 g번째 그룹이 결정되었을 때, 그룹 S_g에서 중요한 유해물질을 선택하는 지표인 δ_{S_g}가 구성 요소별 변수 선택법과 비교하여 변수 선택과정에 추가된다.	
	“# bkmr 패키지의 설치” 등 코드의 설명을 위해 주석을 추가하였음.
	“(1) 패키지의 설치 및 모의연습 자료의 생성(Installation of package and generation of a simulated dataset)” 등 독자가 쉽게 이해할 수 있도록 BKMR의 적용, 즉 코드 설명 부분을 4개로 나누어 각각 소제목을 붙임.

수정 전	수정 후
	kmbayes() 함수를 사용할 때, 먼저 주의해야할 점은 사용되는 자료 y, Z, X에 결측치가 존재하면 안 된다는 것이다. 자료에 결측치가 있는 경우, 함수가 결측치가 존재하는 변수에 대하여 에러 메시지를 띄우며, 실행되지 않는다.
	그림을 통해 유해물질 z1, z2, z3 그리고 z4에 대하여 각 유해물질이 미치는 건강 영향을 확인할 수 있다. 유해물질 z1, z2는 노출 수준이 증가할수록 결과 변수의 값을 높이는 방향으로 작용한다. 또한, 그림을 보면 유해물질 z1의 경우, 높은 노출 수준에서는 오히려 결과 변수의 값을 낮추었다. 유해물질 z2의 경우, 유해물질 z1과 달리 결과 변수와의 관계가 선형적인 것을 확인할 수 있었던 반면, 유해물질 z3, z4는 결과 변수에 유의한 결과를 주지 못 하는 것으로 보인다.
	그림 11을 통해 유해물질 z1, z2, z3 그리고 z4에 대하여 각 유해물질이 미치는 건강 영향을 확인할 수 있다. 유해물질 z1, z2는 노출 수준이 증가할수록 결과 변수의 값을 높이는 방향으로 작용한다. 또한, 그림을 보면 유해물질 z1의 경우, 높은 노출 수준에서는 오히려 결과 변수의 값을 낮추었다. 유해물질 z2의 경우, 유해물질 z1과 달리 결과 변수와의 관계가 선형적인 것을 확인할 수 있었던 반면, 유해물질 z3, z4는 결과 변수에 유의한 결과를 주지 못 하는 것으로 보인다.

〈표 Ⅲ-7〉 g-formula에 대한 국문 가이드라인 수정 전과 후의 대조표

수정 전	수정 후
	g-formula에 대하여 2021년 연구과제(예신희 등. 2021)에서는 단일 노출로 인한 건강영향을 추론하는 국문 가이드라인을 작성하였다. 하지만 2022년 본 과제에서는 2개 이상의 유해물질에 대한 노출이 근로자의 건강에 영향을 미치는지 그 영향을 추론하는데 활용될 수 있는 통계적 방법론에 대한 국문 가이드라인을 작성하는 것뿐만 아니라 산업보건 역학 연구자가 2021년 연구과제에서 작성한 국문 가이드라인의 g-formula에 대한 이해를 보다 쉽게 이해할 수 있도록 하고자 한다. 본 장에 작성된 내용은 Grath 등(2020), 예신희 등(2021)을 기반으로 하여 작성되었으며, 복합물질에 대해 노출되었을 때 여러 노출 변수를 처리하는 방법이 중점적으로 추가되었다. 또한, R 패키지 gfoRmula에 대하여 업데이트된 내용 또한 추가로 서술하고자 한다.
	위의 예제에서 노출 변수로서 변수 E 1개만을 고려하였다. 하지만 근로자가 복합물질에 노출되는 경우, 노출 변수는 2개 이상이 될 수 있으며, 이러한 복합물질의 구조를 g-formula에 반영하여야 한다. 2개 이상의 노출 변수에 대하여 g-formula에 반영하는 예시를 설명하기 위해 복합물질은 2가지 유해물질 E1, E2로 이루어져 있으며, 근로자가 각 유해물질에 대해 노출되었는지 그 여부를 모형화한다고 하자. 이때, 각 유해물질에 대한 노출 여부는 서로 독립적으로 발생한다고 가정하자. 또한, 각 유해물질에 대한 노출 여부는 변수 L1, L2, L3, race, sex 그리고 자신의 이전 시점의 노출량에 영향을 받는다고 가정하자. 즉, 근로자는 두 유해물질에 대한 노출 여부에 대해 서로 독립적으로 발생하기 때문에 t 시점 또는 그 이전 시점의 유해물질 E1에 대한 노출 여부가 t 시점의 유해물질 E2에

수정 전	수정 후
	대한 노출 여부에 영향을 미치지 않으며, 역 또한 마찬가지일 것이다. 그러므로 앞서 설명한 단일 노출에 대한 코드를 응용하여 t 시점의 유해물질 E1과 E2에 대한 노출 여부에 대하여 모형화를 하면 아래와 같다.

	복합물질에 대한 노출과 같이 2개 이상의 유해물질에 동시에 노출되는 경우 또는 2가지 이상의 치료를 같이 처치하는 경우 또한, 패키지를 통해 잠재적 결과를 추정할 수 있다. 앞선 예시에서 노출변수에 해당하는 E 외에 추가적인 노출 변수 H가 있고, E에 대한 노출이 H에 대한 노출에 영향을 준다고 생각해보자. 이러한 경우, 요인 발생의 시간적 흐름이 L1, L2, L3, E 그리고 K이므로, covnames 인수에 c("L1", "L2", "L3", "E", "H")를 입력하면 된다. 코드는 다음과 같이 작성할 수 있다.
	`gformula(…,`
	`covnames = c('L1', 'L2', 'L3', 'E', 'H'),`
	`basecovs = c('race','sex')`
	`)`

	위의 예제에서는 노출 변수로서 변수 E 1개만을 고려하였다. 앞서 (4)에서의 예제와 같이 노출 변수 E뿐만 아니라 변수 G 또한 노출 변수로 고려될 수 있다. 각 노출 변수에 대해 앞서 기술한 노출변수 E가 노출 변수 G에 영향을 준다는 것 외에 각 노출 변수는 시간에 따라 변하는 교란 요인 L1, L2, L3와 시간에 따라 변하지 않는 교란 요인 race, sex 그리고 자신의 이전 시점의 노출 변수에 영향을 받는다고 생각해보자. 이러한 상황에서 앞서 설명한 단일 노출에 대한 코드를 응용하여 t 시점의 노출 변수 E와 G에 대해 모형화를 하면 아래와 같다.

수정 전	수정 후
	gformula(…, covnames = c('L1', 'L2', 'L3', 'E', 'G'), histories = c(lagged), histvars = list(c('L1', 'L2', 'L3', 'E', 'G')), covtypes = c('binary','normal','categorical','binary','binary'), covparams =list(covmodels=c(L1~race+sex+lag1_L1, L2~L1+race+sex+lag1_L2, L3~L1+L2+race+sex+lag1_L3, E~L1+L2+L3+race+sex+lag1_E1,G~L1+L2+L3+race+sex+E+lag1_G), covlink = c(NA, NA, NA, NA, NA)))

〈표 Ⅲ-8〉 인과추론 용어의 정리에 대한 수정 전과 후의 대조표

수정 전	수정 후
철수는 심장이식을 기다리는 환자이다. 1월 1일에 철수는 새로운 심장을 이식받았고, 5일 후에 사망하였다. 철수가 1월 1일에 심장이식을 받지 않았더라면, 5일 후에 살아있을 것이라는 사실을 우리가 어떻게든 알 수 있다고 상상해보자. 철수가 심장이식을 받았을 때 (5일 후 사망)와 심장이식을 받지 않았을 때 (5일 후 생존)의 결과를 모두 알고 있는 대부분의 사람들은 철수가 심장이식으로 인해 사망하였다는 것에 동의할 것이다. 즉, 심장이식은 철수의 5일 후 생존에 인과적 영향을 미쳤다.	개인에서의 인과효과를 간단한 예시를 통해 설명하고자 한다. 심장 이식을 기다리는 철수라는 환자가 있다고 상상해보자. 1월 1일에 철수는 새로운 심장을 이식받았고, 그로부터 5일 후 사망하였다. 이 사례에서 철수는 심장 이식으로 인해 사망한 것일까? 이러한 질문에 답변하기 위해서는 철수가 1월 1일에 심장 이식을 받지 않은 상태에서 철수의 5일 후의 상태를 알아야 한다. 예를 들어, 철수가 심장 이식을 받지 않았더라면, 5일 후에 살아있을 것이라는 사실을 우리가 알고 있다면 사람들은 철수가 심장 이식으로 인해 사망하였다는 것에 동의할 것이다.
$Y^{a=1}$를 치료 값이 $a=1$일 경우 관찰되는 결과변수라고 하고, $Y^{a=0}$를 치료 값이 $a=0$일 경우 관찰되는 결과변수라고 하자. 철수는 치료를 하였을 때 사망하였고, 치료를 하지 않았을 경우에 살았기 때문에 $Y^{a=1}=1$과 $Y^{a=0}=0$로 표현한다. 반면, 영희는 치료를 하였을 때도 살았고, 치료를 하지 않았을 때도 살았기 때문에 $Y^{a=1}=0$과 $Y^{a=0}=0$로 표현한다. 이제 위의 정보들을 바탕으로 '개인에 대한 인과효과'를 공식적으로 정의할 수 있다. 만약 $Y^{a=1}\neq Y^{a=0}$일 경우, 치료 A는 개인의 결과 Y에 인과효과를 가진다. 따라서, 철수의 경우 치료 유무에 따른 건강결과가 $Y^{a=1}=1\neq 0=Y^{a=0}$이므로, 치료는 철수의 건강결과에 인과효과를 가진다. 반면, 영희의 경우 치료 유무에 따른 건강결과가 $Y^{a=1}=0=Y^{a=0}$이므로, 치료는 영희의 건강결과에 인과효과를 가지지 않는다. 이 때, 변수 $Y^{a=1}$와 $Y^{a=0}$는 잠재적 결과들 (potential outcomes) 또는 반사실적 결과들 (counterfactual outcomes)라고 부른다. 일부 연구자들은 받은	$Y^{a=1}$을 치료 값이 1인 경우 관찰되는 결과라고 하고, $Y^{a=0}$을 치료 값이 0인 경우 관찰되는 결과라 정의할 수 있다. 이때, 변수 $Y^{a=1}$와 $Y^{a=0}$는 잠재적 결과(potential outcomes) 또는 반사실적 결과(counterfactual outcome)라고 부른다. 일부 연구자들은 받은 치료에 따라 이 두 가지 결과 중 하나가 잠재적으로 관찰될 수 있음을 강조하기 위해 "잠재적 결과(potential outcome)"이라는 용어를 선호한다. 다른 연구자들은 이러한 결과가 실제로 발생하지 않을 수 있는 상황 (즉, 사실과 반대되는 상황)을 나타내는 것을 강조하기 위해 "반사실적 결과(counterfactual outcome)"라는 용어를 선호한다. 현재 예제에서 철수를 인덱스 i로, 영희를 인덱스 j로 표현하면, 철수는 심장을 이식받았을 때(a=1) 사망하였고, 이식받지 않은 경우(a=0) 생존하였기 때문에 $Y_i^{a=1}=1$와 $Y_i^{a=0}=0$로 표현할 수 있다. 반면, 영희는 심장 이식과 무관하게 생존하였기 때문에 $Y_j^{a=1}=0$와 $Y_j^{a=0}=0$로 표현할 수 있다.

수정 전	수정 후
치료에 따라 이 두 가지 결과 중 하나가 잠재적으로 관찰될 수 있음을 강조하기 위해 “잠재적 결과들 (potential outcomes)”이라는 용어를 선호한다. 다른 연구자들은 이러한 결과가 실제로 발생하지 않을 수 있는 상황 (즉, 사실과 반대되는 상황)을 나타내는 것을 강조하기 위해 “반사실적 결과들 (counterfactual outcomes)”라는 용어를 선호한다.	$Y^{a=1} \neq Y^{a=0}$이면 개인에 대해 인과효과를 가진다고 할 수 있고, 치료 변수 A는 개인의 결과 변수 Y에 대해 인과효과를 가진다고 할 수 있다. 철수의 경우 심장 이식에 따른 생존 여부가 달라지므로 ($Y_i^{a=1} = 1 \neq 0 = Y_i^{a=0}$), 심장 이식 여부가 철수의 생존 여부에 인과적으로 영향을 미쳤다고 할 수 있는 반면에 영희의 경우 심장 이식 유무에 따라 생존 여부가 달라지지 않았으므로 ($Y_j^{a=1} = 0 = Y_j^{a=0}$), 심장 이식 여부는 영희의 생존 여부에 인과적으로 영향을 미치지 않았다.
하지만, 각 개인에서 이러한 결과들 중 하나만 실제로 관찰할 수 있는데, 이 관찰된 결과는 개인이 실제로 경험한 치료 값의 결과를 의미한다. 즉, 그 외 모든 반사실적 결과들은 관찰할 수 없고, 이렇게 확인할 수 없는 나머지 반사실적 결과들로 인해 개인의 인과효과는 확인할 수 없다. 따라서, 개인의 인과효과는 관찰된 데이터의 함수로 표현할 수 없다.	하지만, 현실에서는 각 개인에 대해 잠재적 결과들 중 하나만 실제로 관찰할 수 있는데, 이 관찰된 결과는 개인이 실제로 경험한 치료 값의 결과를 의미한다. 즉, 그 외 모든 잠재적 결과들은 관찰할 수 없고, 이렇게 확인할 수 없는 나머지 잠재적 결과들로 인해 개인의 인과효과는 자료로부터 직접 계산이 불가능하다. 따라서, 개인의 인과효과는 관찰된 데이터의 함수로 표현할 수 없다.
그러나 일반적으로 개인의 인과관계를 식별하는 것은 불가능하기 때문에, 이제 우리는 합쳐진 인과효과인 ‘집단의 평균 인과효과’에 주목한다. 이를 정의하려면 우리는 세 가지 정보가 필요하다: i. 관심 결과, ii. 비교할 치료들 a=1과 a=0, iii. 잘 정의된 집단 내 개인들의 비교될 결과들 $Y^{a=1}$와 $Y^{a=0}$. 우리의 관심 집단으로 철수의 대가족을 생각해보자. 표 1.1은 철수의 가족 20명 모두에 대해 치료를 한 경우 a=1과 치료를 하지 않은 경우 a=0 모두에서 반사실적 결과들을 보여준다. 예를 들어, 마지막 열은 치료를 받았을 경우에 각 개인에서 관찰된 결과들 $Y^{a=1}$이다. 그리고 이 열을 살펴보면, 철수를 포함한 가족 20명이 심장이식을 받은 경우 가족 중 절반 (20명 중 10명)이 사망하였다. 즉, 모든 인구집단의 개인들이 치료를 받은 경우 a=1, 결과 값 1이	(iii)의 값 중 일부를 자료로부터 확인할 수 없어 개인에 대한 인과효과를 직접 계산하는 것은 불가능하므로 현실적인 대안으로서 집단에 대한 인과효과인 ‘집단의 평균 인과효과’에 주목하고자 한다. 우리의 관심 집단으로 20명으로 구성된 철수의 대가족을 생각해보자. 표 1.1은 철수의 가족 20명 모두에 대해 심장 이식을 받은 경우(a=1)와 심장 이식을 받지 않은 경우(a=0)에 대한 잠재적 결과들을 보여준다. 예를 들어, 3번째 열은 각 개인이 심장 이식을 받았을 경우 관찰되는 결과들 $Y^{a=1}$이다. 그리고 이 열을 살펴보면, 철수를 포함한 가족 20명 모두 심장 이식을 받았더라면 가족 중 절반 (20명 중 10명)이 사망했을 것이다. 즉, 인구집단의 모든 개인들이 심장 이식을 받은 경우(a=1), 사망하는 비율은

수정 전	수정 후
의미하는 사망이 발생할 개인들의 비율은 $\Pr[Y^{a=1}=1]=\frac{10}{20}=0.5$이다. 마찬가지로, 표 1.1의 두 번째 열에서 우리는 철수를 포함한 가족 20명이 심장이식을 받지 않은 경우 가족 중 절반 (20명 중 10명)이 사망했을 것이라고 결론을 내릴 수 있다. 즉, 인구집단 내 모든 개인이 a=0일 때, 결과 값 1이 의미하는 사망이 발생할 개인들의 비율은 $\Pr[Y^{a=0}=1]=\frac{10}{20}=0.5$이다. 우리는 치료에 따른 반사실적 위험도를 사망자 수 (10)를 센 뒤 전체 개인의 수 (20)으로 나누어 0.5로 계산하였다. 이는 인구집단 내 모든 개인에 대한 반사실적 결과의 평균을 계산하는 것과 같다.	$\Pr[Y^{a=1}=1]=\frac{10}{20}=0.5$이다. 마찬가지로, 표 1.1의 두 번째 열로부터 가족 20명 모두 심장 이식을 받지 않았다면 가족 중 절반 (20명 중 10명)이 사망했을 것이라고 결론을 내릴 수 있다. 즉, 인구집단 내 모든 개인이 심장 이식을 받지 않았을 때, 사망하는 개인들의 비율은 $\Pr[Y^{a=0}=1]=\frac{10}{20}=0.5$이다. 우리는 심장 이식 여부에 따른 잠재적 결과(사망)가 발생한 사건의 수(10)를 센 뒤 전체 개인의 수 (20)로 나누어 0.5로 계산하였다. 이는 인구집단 내 모든 개인에 대한 잠재적 결과가 주어져있을 때, 잠재적 결과에 대한 기댓값을 계산하는 것과 같다.
그 정의는 다음과 같다: 건강 결과 Y에 대한 치료 A의 평균 인과효과는 연구자가 관심을 가지는 인구집단에서 $E[Y^{a=1}=1]\neq E[Y^{a=0}=1]$로 표현한다. 이 정의에 따르면, 인구집단 '철수의 대가족'에서 치료 A '심장이식'은 결과 Y '사망'에 평균 인과효과를 가지지 않는다. 왜냐하면, 치료를 받은 경우 죽음에 대한 위험도 $\Pr[Y^{a=1}=1]$와 치료를 받지 않은 경우 죽음에 대한 위험도 $\Pr[Y^{a=0}=1]$는 모두 0.5로 같기 때문이다. 즉 이 결과에 따르면 모든 개인이 심장이식을 받는지 받지 않는지 여부는 중요하지 않다. 왜냐하면, 두 경우 모두 절반은 사망하기 때문이다. 여기에서와 같이 인구집단의 평균 인과효과가 0일 때, 우리는 평균 인과효과가 없다는 귀무가설이 참이라고 말한다. 위험도가 평균과 같고, 문자 E가 일반적으로 인구집단의 평균을 나타내는데 사용되기 때문에, 우리는 '인구집단에서 귀무가설이 참이 아닌 평균 인과효과의 정의'를 다음과 같이 다시 쓸 수 있다: $E[Y^{a=1}]\neq E[Y^{a=0}]$. 이 정의는 이분형 변수와 이분형이 아닌	인구집단의 평균 인과효과에 대한 정의는 다음과 같다: 결과 변수 Y에 대한 치료 변수 A의 평균 인과효과는 잠재적 결과의 기댓값들 $E[Y^{a=1}]$, $E[Y^{a=0}]$의 대조로 정의할 수 있다. 여기서 E는 기댓값을 정의하기 위해 사용되었으며, 위의 예제에서 Y가 이분형 변수이기 때문에 $E[Y^{a=1}]=\Pr[Y^{a=1}=1]$, $E[Y^{a=0}]=\Pr[Y^{a=0}=1]$로 표현할 수 있다. 이 정의에 따라 관심 인구집단 '철수의 대가족'에서 '심장 이식'(치료 변수 A)은 '사망 여부'(결과 변수 Y)에 평균 인과효과를 가지지 않는다. 왜냐하면, 심장 이식을 받은 경우 사망하는 확률 $\Pr[Y^{a=1}=1]$과 심장 이식을 받지 않은 경우 사망하는 확률 $\Pr[Y^{a=0}=1]$이 모두 0.5로 같아 두 확률 값의 차이 또는 비가 0 또는 1이기 때문이다. 평균 인과효과가 없다는 것이 개인에 대한 인과 효과가 없다는 것을 암시하지는 않는다. 개인에 대한 인과 효과에서 설명하였듯이 철수에게 사망 여부에 대한 심장 이식의 개인에 대한 효과가 있음을 설명하였다. 추가적으로 표 1.1에서 철수 외에도 11명의 사람

수정 전	수정 후
변수 모두에 적용된다. ※ 인구집단에서의 평균 인과효과: $E[Y^{a=1}] \neq E[Y^{a=0}]$ 평균 인과효과가 없다는 것이 개별 효과가 없다는 것을 의미하지는 않는다. 표 1.1은 치료가 인구의 12명 (철수 포함)에게 개별 인과효과가 있음을 보여준다.	에 대하여 잠재적 결과들의 값 $Y^{a=1}$와 $Y^{a=0}$이 서로 다른 것을 확인할 수 있기 때문이다.
더 일반적으로, 조건부 확률 $\Pr[Y=1\|A=a]$는 치료 수준 a를 받은 경우 관심 인구집단에서 결과 Y가 발생한 개인의 비율로 정의할 수 있다. 치료를 받은 경우 결과가 발생한 개인의 비율 $\Pr[Y=1\|A=1]$과 치료를 받지 않은 경우 결과가 발생한 개인의 비율 $\Pr[Y=1\|A=0]$이 같은 경우, 우리는 A와 Y가 독립이라고 얘기한다. A와 Y가 독립이란 말은 A가 Y와 연관되어 있지 않다는 의미이고, 또는 A가 Y를 예측할 수 없다고도 말할 수 있다. 독립성 (independence)은 $Y \amalg A$ 또는 $A \amalg Y$라고 표현할 수 있고, 'Y와 A는 독립이다'라고 읽을 수 있다.	같은 방식으로 심장 이식을 받지 않은 사람 7명에 대한 사망률을 구해보면 사망한 사람은 3명이므로 사망률 $\Pr[Y=1 \mid A=0]$ 은 3/7이다. 따라서 $\Pr[Y=1 \mid A=1]$와 $\Pr[Y=1 \mid A=0]$의 차이를 살펴보면 7/13-3/7≠0이기 때문에 심장 이식 여부 A와 사망 여부 Y가 서로 의존적이라고 할 수 있다. 이때, 우리는 치료 변수 A와 결과 변수 Y가 서로 연관성이 있다고 표현한다.
인과성은 전체가 흰색인 마름모와 전체가 회색인 마름모 사이의 대비라고 정의할 수 있다. 인과성은 전체가 흰색인 마름모 (모든 개인이 치료를 받음)와 전체가 회색인 마름모 (모든 개인이 치료를 받지 않음) 사이의 대비라고 정의할 수 있다. 반면, 연관성은 전체 마름모의 일부인 흰색 부분 (치료 받음)과 나머지 일부인 회색 부분 (치료를 받지 않음) 사이의 대비라고 볼 수 있다.	연관성에 대하여 그림으로 표현하면 그림 21과 같이 표현할 수 있다. 연관성의 정의에서 사용되는 확률은 모두 조건부 확률로 관심 있는 인구 집단 내에서 치료 수준이 1인 집단과 0인 집단에서 결과가 발생할 확률이기 때문에 그림 21에서 표현되고 있는 것과 같이 관심 있는 인구집단에 해당하는 마름모가 흰색 영역(치료 수준이 1인 경우)과 회색 영역(치료 수준이 0인 경우)으로 나뉘는 것을 확인할 수 있다. 하지만 인과성은 위에서 설명하였듯이 인구집단 전체가 치료 수준이 1인 경우와 0인 경우의 결과를 비교하기 때문에 마름모 전체가 흰색인 경우와 마름모 전체가 회색인 경우의 대비라고 정의할 수 있다.

수정 전	수정 후
$\Pr[Y=1 \mid A=a]$는 조건부 확률이며, 치료 값 a (즉, A=a)를 실제로 받은 인구집단 내 부분집단에서 Y가 발생할 위험을 의미한다. 반면, 위험 $\Pr[Y^a=1]$는 비조건부 (unconditional 또는 marginal이라고 표현함) 확률이고, 인구집단 전체에서 Y^a의 위험을 의미한다. 따라서, 연관성은 개인에서 실제로 치료를 받은 값 (A=1 또는 A=0)에 의해 결정된 인구집단 내 부분집단들 간의 위험도의 차이로 정의한다. 반면, 인과성은 다른 두 가지 치료 값 하 (a=1 또는 a=0)에서 같은 인구집단의 서로 다른 위험으로 정의한다.	앞서 계산했던 $\Pr[Y=1 \mid A=a]$는 치료 수준 a를 실제로 받은 인구집단 내 부분집단에서 결과 Y가 발생할 조건부 확률을 의미하고, $\Pr[Y^a=1]$는 인구집단 전체가 치료 수준 a를 받았을 경우, 결과 Y가 발생할 비조건부 (unconditional 또는 marginal이라고 표현함) 확률이다. 따라서, 연관성을 개인에서 실제로 치료를 받은 값 (A=1 또는 A=0)에 의해 결정된 인구집단 내 서로 다른 부분집단들 사이의 결과가 발생할 확률의 대조 (예; 차이 $\Pr[Y=1 \mid A=1]-\Pr[Y=1 \mid A=0]$)로 정의하는 반면, 인과성은 인구집단이 서로 다른 두 가지 치료 수준 (a=1 또는 a=0)을 각각 받았을 경우, 결과가 발생할 확률의 대조로 정의된다.
결과적으로 결측치가 있음에도 불구하고, 무작위 시험에서는 효과 측정을 일관되게 추정하거나 계산할 수 있다. 아래에서 좀 더 살펴보자. 그림 1.1에서 마름모로 표시된 인구가 거의 무한대이고, 이러한 인구 각 개인에 대해 동전을 던져서 치료 유무를 결정했다고 가정해보자. 동전의 뒷면이 나오면 흰색 그룹에, 앞면이 나오면 회색 그룹에 개인을 할당했다.	이러한 결측 자료에도 불구하고, 평균 인과효과의 계산을 가능하게 해주는 연구 디자인인 무작위 시험을 소개하고자 한다. 무작위 시험 또한 표 2.1과 같이 잠재적 결과들 중 일부에 대해 결측 자료를 가지는 데이터를 생성한다. 그러나 무작위 시험에서 의미하는 치료의 '무작위' 배정은 이러한 결측치가 우연에 의해 발생했음을 보장한다.
즉, 이상적으로 무작위 할당이 된 시험에서는, a=1과 a=0 모두에 대하여, 과거에 치료를 받은 집단이 잠재적 치료 값인 a를 받게 될 경우의 위험도 $\Pr[Y=1 \mid A=1]$는, 과거에 치료를 받지 않은 집단이 잠재적 치료 값인 a를 받게 될 경우의 위험도 $\Pr[Y=1 \mid A=0]$와 같다.	즉, 이상적으로 무작위 할당이 된 시험에서는, A=1과 A=0 모두에 대하여, 치료를 받은 집단이 잠재적 치료 값인 a를 받게 될 경우의 사망률 $\Pr[Y^a=1 \mid A=1]$은, 치료를 받지 않은 집단이 잠재적 치료 값인 a를 받게 될 경우의 사망률 $\Pr[Y^a=1 \mid A=0]$와 같다는 것을 의미한다.
교환가능성 $Y^a \amalg A$는 반사실적 결과와 관찰된 치료가 독립적이라는 것으로 정의할 수 있다. 다시 말해서, 현재 치료를 받은 집단과 현재 치료를 받지 않은 집단이 같은 치료(a=0 이거나 a=1 이거나 상관없이)를 받게 된다면, 이 두 집단이 같은 사망 위험도를 나	교환가능성 $Y^a \amalg A$은 잠재적 결과와 관찰된 치료가 독립적이라는 것으로 의미하며, $Y \amalg A$은 관찰된 결과와 관찰된 치료가 서로 독립적이라는 것을 의미한다. 잠재적 결과가 관찰된 치료에 의해 정해지는 것은 아니지만 관찰된 결과는 가능한 잠재적 결과 중 관

수정 전	수정 후
타낸다는 것을 말한다. 하지만, 반사실적 결과와 관찰된 치료가 독립적이라는 $Y^a \amalg A$는 관찰된 결과와 관찰된 치료가 독립적이라는 $Y \amalg A$를 의미하는 것은 아니다. 예를 들어, 교환가능성을 만족하는 무작위 시험에서, 치료가 결과에 미치는 인과효과를 확인할 수 있고, 이 경우에 $Y \amalg A$는 만족하지 않는다. 왜냐하면 치료는 관찰된 결과와 연관되어 있기 때문이다.	찰된 치료에 따라 정해지기 때문에 관찰된 치료와 연관되어 있어 $Y^a \amalg A$가 $Y \amalg A$을 암시하지는 않는다.
즉, 수학적으로, a=0 일 때, 교환가능성 $Y^a \amalg A$는 만족되지 않는다고 우리는 증명하였다. a=1 일 때도 같은 방법으로 교환가능성이 만족되지 않는다고 증명할 수 있다. 따라서, 이 문단에서의 질문에 대한 답은 '아니오'이다. 하지만, 실제 세계에서 우리는 표 1.1과 같은 반사실적 데이터가 아니라 표 2.1과 같은 관찰된 데이터만 확인할 수 있다. 따라서 실제로는 치료받은 사람들을 치료를 하지 않는다면 관찰할 수 있는 위험도 $\Pr[Y^{a=0}=1 \mid A=1]=7/13$과 같은 반사실적 위험도를 계산하기에는, 표 2.1은 가지고 있는 정보가 부족하다.	즉, 수학적으로, 치료 수준 a=0일 때, 교환가능성 $Y^a \amalg A$은 만족되지 않는다는 것을 보일 수 있다. a=1일 때도 같은 방법으로 교환가능성이 만족되지 않는다는 것을 보일 수 있다. 따라서 이 문단에서의 물음에 대한 답은 '아니오'에 해당한다. 물음에 답하기 위해 우리는 표 1.11.1의 잠재적 결과에 대한 자료를 확보할 수 있다고 가정하였다. 하지만, 우리는 표 1.1과 같이 잠재적 결과를 포함하는 자료는 확보할 수 없고, 표 2.1과 같이 관찰된 자료만 수집할 수 있다. 따라서 실제로는 치료를 받은 사람들에 대하여 '치료를 받지 않았더라면'에 해당하는 사망률 $\Pr[Y^{a=0}=1 \mid A=1]$은 표 2.1에 있는 정보만으로는 계산이 불가하다.
연구 설계 1에서 우리는 인구의 65%를 무작위로 선택하였고, 여기서 선택된 각 개인에게 새로운 심장을 이식하였다. 이 내용은 20명 중 13명이 치료를 받은 이유를 설명한다. 연구 설계 2에서 우리는 모든 개인을 예후인자 (개인이 위독한 경우 L=1, 그렇지 않은 경우 L=0)에 따른 상태로 분류했다. 그런 다음 우리는 위독한 상태에 있는 개인의 75%와 그렇지 않은 상태에 있는 개인의 50%를 무작위로 선택하고 선택된 각 개인에게 새 심장을 이식했다.	연구 설계 1에서 우리는 관심 있는 인구집단의 구성원에게 65%의 확률로 심장 이식 여부를 무작위로 배정하고, 선택된 각 개인에게 새로운 심장을 이식하였다. 이러한 내용이 20명 중 13명이 심장 이식을 받은 이유를 설명한다. 연구 설계 2에서는 구성원을 예후 인자 L에 따라 분류하였고, 위독한 상태에 있는 개인에게는 75%의 확률로, 그렇지 않은 상태에 있는 개인에게는 50%의 확률로 심장 이식 여부를 무작위로 배정하고 선택된 각 개인에게 새 심장을 이식했다.

수정 전	수정 후
두 연구 설계 모두 무작위 시험이다. 다만, 연구 설계 1은 1) 무작위화 (randomization) 섹션에서 설명한 무작위 시험의 유형이다. 이 연구 설계에서는 모든 개인에게 치료를 할당할 때 한 개의 동전만 사용한다 (예: 뒷면이면 치료 함, 앞면이면 치료하지 않음): 뒷면이 나올 확률이 0.65인 동전을 사용하여, 개인의 65%가 치료를 받게 되었다. 연구 설계 2에서는 모든 개인을 위해 두 개의 동전을 사용한다. 위독한 상태의 개인에게는 0.75 확률로 뒷면이 나올 동전을 사용하였고, 위독하지 않은 상태의 개인에게는 0.50 확률로 뒷면이 나올 다른 동전을 사용하였다	두 연구 설계 모두 무작위 시험이며, 다만, 연구 설계 1은 구성원의 위독한 상태와는 무관하게 심장 이식 여부를 무작위로 배정하였으며(이전 절에서 설명한 무작위 시험의 유형임), 연구 설계 2은 구성원의 위독한 상태에 따라 심장 이식의 배정에 관한 확률을 달리한 무작위 시험이다. 연구 설계 1에서는 모든 개인에게 치료를 할당할 때 65%의 확률로 심장 이식을 시행하는 한 개의 동전만 사용하지만, 연구 설계 2에서는 총 두 개의 동전을 사용하는데, 하나는 위독한 상태의 개인에게는 75%의 확률로 심장 이식을 시행하는 동전과 다른 하나는 위독하지 않은 상태의 개인에게는 50%의 확률로 심장 이식을 시행하는 동전이다.
따라서 표 2.2의 데이터는 비조건부 무작위 시험에서 도출된 것이 아니다. 왜냐하면, 치료를 받은 사람 중에서는 69%가 위독한 상태였던 반면 치료를 받지 않은 사람 중에서는 43%가 위독한 상태였기 때문이다. 이 불균형은 실제 세계에서 치료를 받았던 사람들이 치료를 받지 않은 상태로 남아 있었다면, 실제 세계에서 치료를 받지 않았던 사람들의 사망 위험보다 더 높았을 것이라는 것을 나타낸다. 즉, 치료 A는 치료를 받지 않은 경우 사망할 반사실적 위험도를 예측할 수 있고, 교환가능성 $Y^a \amalg A$은 성립하지 않는다. 하지만, 우리 연구는 무작위 시험이었기 때문에, 이 연구는 L에 대한 조건부를 둔 상태에서 무작위화를 시행한 무작위 시험이라고 결론내릴 수 있다.	하지만 조건부 무작위 시험은 노출 군과 비노출 군 사이의 비조건부 무작위 시험의 교환가능성을 만족시키지 않는데, 그 이유는 조건부 무작위 시험의 설계에 따라 노출 군과 비노출 군 각각에서 나쁜 예후를 가진 개인의 비율이 다를 수 있기 때문이다. 다시 말해, 비조건부 무작위 시험의 교환 가능성이 성립한다면 치료를 받는 군과 치료를 받지 않는 군이 서로 교환가능하기 때문에 나쁜 예후를 가지는 사람들의 비율 또한 동일해야 한다. 하지만 현재 예제를 보면, 치료를 받은 사람 중에서 69%가 위독한 상태 ($\Pr(L=1 \mid A=1)=9/13$)이고, 치료를 받지 않은 사람 중에서는 43%가 위독한 상태 ($\Pr(L=1 \mid A=0)=3/7$)로 같지 않기 때문에 비조건부 무작위 시험의 교환가능성이 성립하지 않는다고 할 수 있다. 또한, 이러한 나쁜 예후를 가지는 사람의 비율의 차이(불균형(imbalance)이라 부름)는 실제 치료를 받은 사람들이 치료를 받지 않았더라면, 실제 치료를 받지 않은 사람들의 사망보다 더 높았을 것이라는 점을 암시한다.
조건부 무작위 시험은 단순히 두 개의 개별적인 비조건부 무작위 시험의 조합이다. 하나는 위독한 상태 (L=1)에 있는 개인의 하위집	조건부 무작위 시험은 연구 설계 2와 같이 단순한 두 개의 개별적인 비조건부 무작위 시험을 평행하게 결합한 것이다. 하나는 위

수정 전	수정 후
단에서 수행하고, 다른 하나는 위독하지 않은 상태 (L=0)에 있는 개인의 하위집합에서 수행한다. 먼저 위독한 상태의 개인이 모여 있는 하위집합에서 수행하는 무작위 시험을 고려해보자. 이 하위집단에서 노출된 군과 노출되지 않은 군은 교환가능성을 만족한다. 즉, 치료 할당 당시에 모두 위독한 상태였기 때문에, 공식적으로 각 치료 값 a에 따른 반사실적 사망 위험도는 치료를 받은 사람과 치료받지 않은 사람이 동일하다. 이 내용은 다음과 같은 식으로 표현할 수 있다.	독한 환자 (L=1)로 구성된 부분집단에서 수행하고, 다른 하나는 위독하지 않은 환자 (L=0)로 구성된 부분집단에서 수행한다. 먼저 위독한 환자로 구성된 부분집단에서 수행되는 무작위 시험을 고려해보자. 이 부분집단에서 무작위 시험이 시행되기 때문에 치료를 받는 군과 치료를 받지 않는 군은 교환가능성을 만족한다.
이와 비슷하게, 위독하지 않은 상태에 있던 개인으로 이루어진 하위집단에서, 무작위화는 치료를 받는 군과 치료를 받지 않는 군이 상호교환성을 만족하게 만든다. 즉, $Y^a \amalg A \mid L = 0$이다. 모든 값 l에 대하여 $Y^a \amalg A \mid L = l$이 만족할 때, 우리는 단순하게 $Y^a \amalg A \mid L$이라고 표현할 수 있다.	이와 같은 방식으로 위독하지 않은 환자로 구성된 부분집단에서 시행된 무작위 시험으로 인하여 치료를 받는 군과 치료를 받지 않는 군에 대하여 교환가능성이 성립한다.
표준화 (standardization)와 역확률가중치 (inverse probability weighting)는 조건부 무작위 시험에 대한 데이터를 사용하여 전체 모집단의 평균 인과효과를 계산할 수 있는 방법이다.	조건부 교환가능성이 만족하는 경우, 평균 인과효과를 추론하기 위해 널리 사용되는 방법으로 표준화 (standardization)와 역 확률가중치 (inverse probability weighting)이 있으며, 소개하고자 한다.
다만, 여기서 기억할 것은 이 8명이 위독하지 않은 상태를 가진 80억 명의 연구대상자를 대표한다는 것이다. 이 집단에서, 치료를 한 경우 사망 위험도는 $\Pr[Y=1 \mid L=0, A=1] = \frac{1}{4}$이고, 치료를 하지 않은 경우 사망 위험도는 $\Pr[Y=1 \mid L=0, A=0] = \frac{1}{4}$이다. 왜냐하면, L=0인 집단 내 연구대상자에게 치료를 무작위로 배정하였기 때문에, $Y^a \amalg A \mid L = 0$으로 표현할 수 있고, 이때 관찰된 위험도는 반사실적 위험도와 동일하다. 즉, L=0인 집단에서, 치	다만, 표본 추출로 인한 변동성을 논의에서 제외하기 위해 위독하지 않은 환자 8명이 위독하지 않은 환자 80억 명을 대표한다고 생각하자. 위독한 상태의 부분집단 또한 마찬가지이다. 우리의 목표가 인과 위험 비 $\Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1]$을 계산하는 것이다. 인과 위험 비의 분자는 모든 20명의 연구대상자가 치료를 받았을 때, 발생했을 잠재적 결과에 대한 확률 $\Pr[Y^{a=1}=1]$이다. 이 확률은 조건부 확률에 정의에 의하여 다음과 같이 표현할 수 있다.

수정 전	수정 후
료한 사람들의 위험도는 모든 사람들을 치료했다고 가정했을 때의 위험도($\Pr[Y=1\|L=0,A=1]=\Pr[Y^{a=1}=1\|L=0]$)와 같고, 치료하지 않은 사람들의 위험도는 모든 사람들을 치료하지 않았다고 가정했을 때의 위험도($\Pr[Y=1\|L=0,A=0]=\Pr[Y^{a=0}=1\|L=0]$)와 같다. 이와 동일한 접근으로, 우리는 위독한 상태에 있는 12명의 연구대상자들을 포함한 집단에서도 관찰된 위험도가 반사실적 위험도와 동일하다는 결론을 내릴 수 있다. 즉, $\Pr[Y=1\|L=1,A=1]=\Pr[Y^{a=1}=1\|L=1]=\frac{2}{3}$이고, $\Pr[Y=1\|L=1,A=0]=\Pr[Y^{a=0}=1\|L=1]=\frac{2}{3}$. 이제 우리의 목표가 인과 위험 비 $\Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1]$를 계산하는 것이라고 가정해보자. 인과 위험 비의 분자는 만약 모든 20명의 연구대상자가 치료를 받았다면 이라고 가정할 경우의 위험도이다. 이전 문단의 내용에 따라, 우리는 L=0인 8명의 연구대상자 모두가 치료를 받았다면 위험도가 1/4이고, L=1인 12명의 연구대상자 모두가 치료를 받았다면 위험도가 2/3라는 것을 알고 있다. 따라서 모든 사람이 치료를 받았다고 가정하였을 때의 위험도 $\Pr[Y^{a=1}=1]$는 0.5와 같다. 이와 동일하게 추론해보면, 모든 사람이 치료를 받지 않았다고 가정하였을 때의 위험도 $\Pr[Y^{a=0}=1]$는 0.5로 계산할 수 있다. 결론적으로, 인과 위험 비는 0.5/0.5=1이다. 더 공식적으로는, 비조건부 반사실적 위험도(marginal counterfactural risk) $\Pr[Y^{a}=1]$는 계층 내 위험도의 가중 평균이다. 이때 가중치는 전체 인구집단에서 L=0인 연구대상자 수의 비율, L=1인 연구대상자 수의 비율과 같다. 즉,	$\Pr[Y^{a=1}=1]=\Pr[Y^{a=1}=1\mid L=1]\Pr[L=1]+\Pr[Y^{a=1}=$ 여기서 $\Pr[Y^{a=1}=1\mid L=1]$은 조건부 교환가능성에 의해 $\Pr[Y=1\mid A=1,\ L=1]$ 와 같으며, $\Pr[Y^{a=1}=1\mid L=0]=\Pr[Y=1\mid A=1,\ L=0]$이 성립한다. 따라서 확률 $\Pr[Y^{a=1}=1]$을 계산하기 위해서는 $\Pr[L=1]$ ($\Pr[L=0]$은 $1-\Pr[L=1]$을 통해 계산이 가능함), $\Pr[Y=1\mid A=1,\ L=1]$ 그리고 $\Pr[Y=1\mid A=1,\ L=0]$을 계산해야한다. 표 2.2로부터 위독한 상태에서 심장 이식을 받은 환자의 사망률 $\Pr[Y=1\mid A=1,\ L=1]$을 구할 수 있으며, 그 값은 $\frac{2}{3}$이고, 같은 방법으로 위독하지 않은 상태에서 심장 이식을 받은 환자의 사망률 $\Pr[Y=1\mid A=1,\ L=0]$ 또한 계산이 가능하며, 그 값은 $\frac{1}{4}$이다. 마지막으로 위독한 환자의 비율 $\Pr[L=1]$을 계산하면 $\frac{3}{5}$의 값을 얻을 수 있다. 이 결과들로부터 $$\Pr[Y^{a=1}=1]=\frac{2}{3}\cdot\frac{3}{5}+\frac{1}{4}\cdot\frac{2}{5}=\frac{1}{2}$$ 임을 알 수 있다. 마찬가지 방법으로 인과 위험비의 분모를 계산하면 확률 $\Pr[Y^{a=0}=1]$ 또한 0.5임을 알 수 있다. 그러므로 인과 위험 비는 1이다. 예제에서 수행한 과정을 일반화하여 설명하면, 잠재적 결과에 대한 위험 $\Pr[Y^{a}=1]$은 계층 내 위험의 가중 평균이다. 이때 가중치는 전체 인구집단에서 예후 인자 L이 0인 환자 수의 비율, 예후 인자 L이 1인 환자 수의 비율과 같다. 즉, 앞서 기술하였던 것을

수정 전	수정 후
$\Pr[Y^a=1]=\Pr[Y^a=1\|L=0]\Pr[L=0]+\Pr[Y^a=1\|L=1]$이다. 더 간략하게 표현하자면, $\Pr[Y^a=1]=\Sigma_l\Pr[Y^a=1\|L=l]\Pr[L=l]$으로 표현할 수 있고, 여기서 Σ_l은 인구집단에서 발생할 수 있는 모든 l 값에 대하여 합한 것을 의미한다. 조건부 교환가능성에 의하여, 위의 식에서 반사실적 위험도 $\Pr[Y^a=1\|L=l]$를 관찰된 위험도 $\Pr[Y=1\|L=l,A=a]$로 대체할 수 있다. 즉, $\Pr[Y^a=1]=\Sigma_l\Pr[Y=1\|L=l,A=a]\Pr[L=l]$이다. 이 식의 왼쪽 식은 관찰되지 않은 반사실적 위험도인 반면 오른쪽 식은 L, A 및 Y에 대한 데이터를 사용하여 계산할 수 있는 관찰된 양만 포함한다. 반사실적 양은 관찰된 데이터의 분포(즉, 확률)의 함수로 표현할 수 있고, 이러한 경우, 반사실적 양은 식별 가능하다고 한다. 하지만 반사실적 양을 관찰된 데이터의 분포(즉, 확률)의 함수로 표현할 수 없을 경우, 반사실적 양은 식별 불가능하다고 한다.	일반화하여 기술하면, 모든 a에 대하여 $\Pr[Y^a=1]=\Pr[Y^a=1\|L=0]\Pr[L=0]+\Pr[Y^a=1\|L=1]$이다. 보다 간략하게 표현하자면, $\Pr[Y^a=1]=\sum_l \Pr[Y^a=1 \mid L=l]\cdot\Pr[L=l]$으로 표현할 수 있고, 여기서 Σ_l은 관심 있는 인구집단에서 발생할 수 있는 모든 예후인자 L의 값 l에 대하여 이어 나오는 항을 합산하는 것을 의미한다. 조건부 교환가능성에 의하여, 위의 식에서 조건부 잠재적 위험 $\Pr[Y^a=1\|L=l]$을 조건부 관찰된 위험 $\Pr[Y=1 \mid A=a,\ L=l]$로 대체할 수 있다. 즉, $\Pr[Y^a=1]=\sum_l \Pr[Y=1 \mid L=l,A=a]\cdot\Pr[L=l]$이 다 . 이 식의 왼쪽의 확률은 관찰할 수 없는 값을 포함하는 잠재적 위험인 반면 오른쪽 양은 자료에서 확인 가능한 예후 인자 L, 치료 여부 A 및 결과 Y를 사용하여 계산 가능한 확률만 포함한다. 이와 같이 조건부 교환가능성 조건 아래에서 잠재적 결과에 관한 양은 관찰된 자료의 분포(예, 확률)의 함수로 표현할 수 있으며, 이러한 경우, 잠재적 결과에 관한 양을 식별 가능 (identifiable)하다고 한다. 반대로, 잠재적 결과에 관한 양을 관찰된 데이터의 분포 (즉, 확률)의 함수로 표현할 수 없는 경우, 잠재적 결과에 관한 양이 식별 불가능하다고 한다.
이전 섹션에서는 표준화를 통해 조건부 무작위 시험에서 인과 위험 비를 계산하였고, 이 섹션에서는 역 확률 가중치를 통해 인과 위험 비를 계산하고자 한다.	이전 절에서는 조건부 무작위 시험에서 표준화를 통해 인과 위험 비를 계산하는 과정에 대하여 설명하였다. 이번 절에서는 역 확률 가중치 방법을 통해 인과 위험 비를 계산하는 방법에 대하여 설명하고자 한다.

수정 전	수정 후
트리의 가장 왼쪽 원에는 다음과 같이 첫 번째 가지가 포함되어 있다: 여기서 8명의 연구대상자는 위험하지 않은 상태 (L = 0)이고 12개의 연구대상자는 위험한 상태 (L = 1)이다. 괄호 안의 숫자는 다음과 같이 L=0 또는 L=1인 조건에 있을 확률이다: Pr[L = 0]=8/20 = 0.4 또는 Pr[L = 1]=12/20=0.6. 예를 들어 L=0인 조건 하에 해당하는 8명 중, 4명은 치료를 받지 않았고 (A = 0) 4명은 치료를 받았다 (A = 1). L=0일 때 치료받지 않을 조건부 확률은 괄호 안에 표시된 것처럼 Pr[A=0|L=0]=4/8=0.5이다. Pr[A=1|L=0]에 해당하는 치료받을 조건부 확률도 0.5이다. 가장 오른쪽의 맨 위쪽 원은 L=0과 A=0인 연구대상자 중 3명이 생존 (Y=0)하고 1명이 사망 (Y=1)한 것을 나타낸다. 즉, Pr[Y=0|L=0, A=0]=3/4이고, Pr[Y=1|L=0, A=0]=1/4이다. 트리의 다른 가지들도 유사하게 해석할 수 있다. 이제 이 트리를 사용하여 인과 위험 비를 계산해보자. 인과 위험 비의 분모인 Pr[Ya=0=1]은 인구집단의 모든 사람이 치료를 받지 않았을 경우 사망할 확률의 역수이다. 이 확률을 계산해 보면 다음과 같다.	나무의 가장 왼쪽 원 안에는 두 개의 가지가 포함되어 있다. 두 개의 가지는 위독하지 않은 환자와 위독한 환자를 구분하기 위해 사용된다. 그러므로 20명의 환자 중 위독하지 않은 환자 8명은 두 개의 가지 중 윗 방향으로 향하는 가지를 따라가며, 위독한 환자 12명은 아래 방향으로 향하는 가지를 따라간다. 가지의 위에 있는 괄호 안의 숫자는 예후 인자 L이 0 또는 1일 확률이다 ($\Pr[L=1]=\frac{3}{5}$, $\Pr[L=0]=\frac{2}{5}$). 같은 방식으로 위독하지 않은 환자를 다시 한 번 심장 이식을 받은 환자와 받지 못 한 환자로 분류할 수 있다. 위독하지 않은 환자 (L=0) 8명 중 4명은 심장 이식을 받았고 (A=1), 4명은 심장 이식을 받지 않았으므로 (A = 0), 예후 인자 L이 0일 때, 심장 이식을 받지 않을 조건부 확률은 괄호 안에 표시된 것처럼 $\Pr[A=1 \mid L=0]=\frac{4}{8}=\frac{1}{2}$이다. 심장 이식 여부에 대하여 분류한 다음 사망 여부를 통해 환자들을 한 번 더 분류할 수 있다. 가장 오른쪽의 맨 위쪽 원으로부터 위독하지 않은 환자 중 심장 이식을 받지 않은 환자에 대해서 3명의 환자가 생존하고 (Y=0), 1명의 환자가 사망한 것을 확인할 수 있다. 즉, $\Pr[Y=1 \mid A=0, L=0]=\frac{1}{4}$이고, 그러므로 $\Pr[Y=0 \mid A=0, L=0]=\frac{3}{4}$ 임을 알 수 있다. 이 가지들뿐만 아니라 나무의 다른 가지들도 비슷하게 해석할 수 있다. 이제 이 나무를 사용하여 인과 위험 비를 계산해보자. 표준화 방법을 통해 인과 위험 비를 계산할 때에는 분자를 먼저 계산하였지만 역 확률 가중치 방법에서는 분모를 먼저 계산해보고자 한다.

수정 전	수정 후
즉, 연구대상자 인구집단인 8+12=20명 모두가 치료를 받지 않았다면 2+8=10명이 사망했을 것이다. 따라서, 인과 위험 비의 분모인 Pr[Ya=0=1]은 10/20=0.5이다.	즉, 관심 있는 인구집단 20명 모두가 치료를 받지 않았다면 위독하지 않은 환자 군에서 2명, 위독한 환자 군에서 8명으로 총 10명이 사망했을 것이다. 따라서, 인과 위험 비의 분모인 $\Pr[Y^{a=0}=1]$은 $\frac{10}{20}=\frac{1}{2}$이다.
인과 위험 비의 분자인 Pr[Ya=1=1]는 인구집단의 모든 사람이 치료를 받았을 경우 사망할 반사실적 위험이다. 이전 단락에서와 같이 추론해보면, L=1이 주어졌을 때 교환가능성이 만족되며, 이 때 이 위험도는 10/20=0.5로 계산된다. 그림 2.2의 두 번째 트리는 인구가 모두 치료를 받았을 때를 보여준다. 이 단락과 이전 단락의 결과를 결합하면 인과 위험 비 Pr[Ya=1=1]/Pr[Ya=0=1]은 0.5/0.5=1과 같다. 이 방법이 어떻게 작동하는지 살펴보자. 그림 2.2의 두 나무는 모집단의 모든 연구대상자가 치료를 받지 않았거나 또는 모든 연구대상자가 치료를 받았다면 어떤 일이 일어났을 지를 시뮬레이션 한 것이다. 조건부 교환 가능성이 만족할 경우 이러한 시뮬레이션은 정확하게 수행될 수 있다. 이 두 시뮬레이션을 통합하여 모든 연구대상자가 치료를 받았을 경우의 가상 인구와 모든 연구대상자가 치료를 받지 않았을 경우의 가상 인구를 생성할 수 있습니다. 이 두 가상 인구는 원래 인구집단의 두 배이며, 이 가상 인구를 의사 인구 (pseudo-population)라고 합니다. 그림 2.3은 전체 의사 모집단을 보여줍니다. 원래의 모집단에서 조건부 교환가능성 $Y^a \amalg A \mid L$ 하에서, L이 A와 독립이기 때문에 의사 모집단에서 치료받지 않은 인구집단과 치료받은 인구집단이 (무조건적으로) 교환가능하다. 즉, 의사 모집단의 연관성 위험 비는 의사 모집단과 원래 모집단 모두에서의 인과 위험 비와 동일하다.	그림 2.2의 두 나무는 관심 있는 인구집단의 모든 환자가 심장 이식을 받지 않았거나 (왼쪽 그림) 또는 받았다면 (오른쪽 그림) 생기는 결과를 표현했던 그림이다. 조건부 교환가능성 아래에서 이러한 가상적인 결과를 상상해볼 수 있다. 이러한 결과를 통합하여 모든 환자가 심장 이식을 받았을 경우의 가상적인 인구집단과 모든 환자가 심장 이식을 받지 않았을 경우의 가상 인구집단을 생각해볼 수 있다. 이 가상 인구는 그림 2.2의 두 가상 인구를 통합한 인구집단이며, 그 결과 각 인구집단의 표본 수의 두 배에 해당한다. 또한, 이러한 가상 인구를 가상의 인구집단 (pseudo-population)라고 한다. 그림 24은 이 통합된 인구집단을 보여줍니다. 원래의 모집단에서의 조건부 교환가능성 $Y^a \amalg A \mid L$ 아래에서 가상의 인구집단은 예후 인자 L이 심장 이식 여부 A와 독립이 되기 때문에 가상의 인구집단에서 심장 이식을 받지 않은 부분집단과 심장 이식을 받은 부분집단이 (비조건적으로) 교환가능하다. 좀 더 구체적으로 이 방법이 어떻게 작동하는지 설명하고자 한다. 그림 2.1의 모집단에서 위독하지 않은 환자 (L=0) 중 심장 이식을 받지 않은 4명의 환자를 살펴보자. 이 환자들은 그림 24의 가상의 인구집단에서 심장 이식을 받지 않은 8명의 환자를 생성하는데 사용된다. 위독하지 않은 환자가 심장 이식을 받을 확률 $\Pr[A=1 \mid L=0]=0.5$의 역수인 2의 가중치를 기존의 4명에 곱하여 8명의 환자가 되었다. 같은 방식으로 그림 2.1에서 위독한

수정 전	수정 후
	환자 9명은 확률 $\Pr[A=1 \mid L=1]=\frac{3}{4}$의 역수에 해당하는 $\frac{4}{3}$의 가중치를 받아서 12명의 연구대상자가 되었다.
표준화와 역 확률 가중치 둘 다 변수 L이 치료받을 확률을 결정하는 데 사용되지 않았다면 관찰되었을 것을 시뮬레이션하기 때문에 우리는 종종 이러한 방법이 L에 대해 보정되었다고 말한다. 우리는 때때로 이러한 방법이 L을 통제한다고도 말한다. 이렇게 L을 통제하는 가장 좋은 방법이 무작위 시험이지만, 무작위 시험은 종종 비윤리적이거나 비현실적이거나 시기적절하지 않을 수 있다. 따라서, 관찰연구를 수행해야 할 수 있고, 다음 파트에서 관찰연구에서 표준화와 역 확률 가중치를 사용하여 L을 통제하는 방법에 대해 설명하고자 한다.	표준화와 역 확률 가중치 방법 모두 예후 인자 L이 방법 내에서 사용되기 때문에 우리는 종종 이러한 방법이 예후 인자 L을 보정한다고 말한다. 또는 우리는 때때로 이러한 방법이 L을 통제한다고 표현하기도 한다. 이렇게 예후 인자 L을 통제하는 가장 좋은 방법이 무작위 시험이지만, 예를 들어, 흡연 또는 중금속에 대한 노출로 인해 발생하는 건강 영향에 대해 평가하기 위한 무작위 시험을 시행하는 것과 같이, 관심 있는 주제 중 일부에 대한 무작위 시험은 비윤리적이거나 또는 비현실적이며, 시기적절하지 않을 수 있다. 따라서 그러한 주제들의 경우 무작위 시험을 시행하기 어려워 관찰연구를 수행해야만 할 수 있다.
이상적인 무작위 시험에서는 평균 인과효과 (average causal effect)를 정의하고 정량화할 수 있는데, 그 이유는 이러한 무작위 시험이 '상호교환성 (exchangeability)'이라는 성질을 만족하기 때문이다. '상호교환성'은 치료군 (treated group)과 비치료군 (untreated group)의 결과 (outcome)에 대한 값이 독립이라는 의미이다. 예를 들어 비조건부적인 (marginally) 심장이식과 사망에 대한 무작위 시험이 있다고 가정하자. 이식을 받은 사람들이 만약 실제로 이식을 받지 않았다면, 이식을 받지 않은 사람들과 동일한 사망위험이 있을 것으로 예상할 수 있다. 결과적으로 무작위 시험으로부터 도출된 위험 비는 인과 위험 비와 동일하다고 생각할 수 있다. 이러한 무작위 시험에 반해 관찰연구의 경우에는 위의 가정이 위배될 가능성이 더 높다. 이는 관찰연구가 무작위로 치료군을 배정할 수 없기 때문이다. 즉, 관찰연구에서 치료와 결과의 연관	그럴 수 있었던 이유는 비조건부 또는 조건부 무작위 시험이 '(비조건부 또는 조건부) 교환가능성 ((unconditional or conditional) exchangeability)'이라는 성질을 쉽게 만족하는 연구 디자인(study design)이기 때문이다. 교환가능성에 대한 정의를 상기시키기 위해, '(비조건부) 교환가능성'은 잠재적 결과는 치료 여부에 대해 독립이라는 의미이다. 예를 들어, 심장 이식의 효과를 알기 위해 수행한 무작위 시험이 있다고 가정하자((앞선 예제와는 예임). 이 무작위 시험에서 윤리적으로 문제가 생길 수 있으며, 비현실적이지만 어떤 환자에 대한 심장 이식 여부를 동전을 던져 앞면이 나오면 심장 이식 수술을 수행한다고 상상하여보자. 이 무작위 시험은 환자의 상태와 무관하게 심장 이식 여부를 환자에게 배정하기 때문에 (비조건부) 교환가능성이 성립하고, 이로부터 심장 이식을 받은 사람들이 만약 심장 이식을 받지 않았다면, 이식을 받

수정 전	수정 후
성은 결과에 대한 치료의 인과효과로 설명하는 것이 적절하지 않을 수 있다. 그러나, 무작위 시험이 인과관계 추론에 대한 본질적인 이점을 가진다는 것을 알하면서도 때때로 우리는 인과관계의 질문에 대한 답을 찾기 위해 관찰연구를 수행한다. 앞서 설명한 관찰연구에서의 인과추론에 대한 제한점을 해결하기 위한 기본적인 가정은 관찰연구가 조건부 무작위 시험 (conditionally randomized experiment)이라고 보는 것이다.	지 않은 사람들과 동일한 사망률이었을 것으로 예상할 수 있다. 결과적으로 무작위 시험으로부터 도출된 위험 비는 인과 위험 비와 동일하다고 생각할 수 있다. 하지만 이 무작위 시험과 달리 관찰연구의 경우에는 위의 가정이 성립하지 않을 가능성이 더 높다. 관찰연구에서는 환자에게 무작위로 심장 이식 여부를 배정할 수 없기 때문이다. 교환가능성이 성립되지 않을 수 있기 때문에 관찰연구에서 치료와 결과의 연관성을 결과에 대한 치료의 인과효과로 여기는 것이 적절하지 않을 수 있다. 그러나, 무작위 시험이 인과추론에 대한 본질적인 이점을 가진다는 것을 알고 있지만 우리는 앞서 언급한 흡연 또는 중금속의 노출에 관한 예시와 같이 인과관계의 질문에 대한 답을 찾기 위해 때때로 관찰 연구를 수행한다. 앞서 설명한 관찰연구에서의 인과추론에 대한 제한점을 해결하기 위해 우리는 기본적인 가정으로 관찰연구가 조건부 무작위 시험 (conditionally randomized experiment)이라고 볼 것이다.
비조건부 (marginally) 무작위 시험에서는 치료군이 실제로 치료를 받지 않았다면 비치료군과 동일한 평균 결과를 가질 것이라는 '상호교환성'의 성격을 갖고 있다. 즉, 이러한 교환가능성의 성질은 무작위적인 시험에서 결과에 대한 독립변수들이 치료군과 비치료군 간에 동일한 분포를 갖는 것을 의미한다. 예를 들어 설명해 보자. 아래의 표와 같이 치료군과 비치료군이 결과변수에 대한 분포가 균등하지 않은 경우 비조건부 (marginally) 무작위 시험으로 해결할 수 없고 대신 조건부 (conditionally) 무작위 시험으로 해결할 수 있다. 아래의 표를 보면 변수 L의 수준 하에서는 치료군과 비치료군 간에는 조건부 교환 (conditionally exchangeable)이 가능한 것을 알 수 있다. 즉, $Y^a \amalg A|L$의 형태로 볼 수 있다. 다시 관찰연구 측면으로 돌아가 보자. 치료여부가 조사자에 의한 무작위 배정이 아닌 경우, 치료를 받게 하는 원인들이 일부의 결과 예측치들	비조건부 (또는 조건부) 무작위 시험에서는 (특정 변수에 대해 값이 주어지면) 치료를 받은 군이 치료를 받지 않았더라면 치료를 받지 않은 군과 동일한 평균 잠재적 결과를 가질 것이라는 '교환가능성'이 성립한다는 특징을 가지고 있다. 즉, 조건부 교환가능성 $Y^a \amalg A|L$이 성립하는 자료라 볼 수 있다. 다시 관찰연구 측면으로 돌아가 보자. 치료 여부가 조사자에 의해 무작위로 배정되지 않는다면, 치료를 받게 하도록 영향을 주는 요인들 중 일부가 빠져있을 수 있으며, 이러한 요인들이 잠재적 결과와 연관성이 있을 수 있다. 즉, 치료를 받게 하도록 영향을 주는 요인이 주어졌을 때, 잠재적 결과들의 분포가 치료를 받은 군과 치료를 받지 않은 군이 서로 다를 수 있다. 표 3.1로 다시 예시를 들어보면, 연구의 형태가 관찰연구인 경우, 의사는 치료가 제일 필요한 사람들에게 심장 이식을 하려는 경향이 있을 수 있다. 만약 심

수정 전	수정 후
과 연관성이 있을 수 있다. 즉, 조건부 무작위 시험처럼 결과 예측치들의 분포가 치료군과 비치료군에서 일반적으로 다를 수 있다. 위에서 언급한 표의 내용을 다시 예시로 들어보면, 연구의 형태가 관찰연구일 경우 의사가 치료가 제일 필요한 사람들에게 심장이식을 하려는 경향이 존재할 수 있다. 만약 치료군과 비치료군 사이에 다르게 분포하는 유일한 결과 예측변수가 L이라면, 관찰연구와 조건부 무작위 시험은 논리적으로 봤을 때 동일하다고 볼 수 있다. 이러한 조건 하에서, 표준화 (standardization) 혹은 역 확률 가중치 (IP weighting)가 인과적 효과를 확인하기 위하여 활용될 수 있다. 그러나 만약 조건부 교환가능성 (conditional exchangeability)이 만족되지 않은 경우는 측정되지 않은 독립변수들이 존재할 때인데, 결국 분석을 수행할 때 충분한 데이터를 확보하여 연구의 가정이 조건부 교환성에 근접할 수 있도록 살펴봐야 할 필요가 있다.	장 이식을 받은 군과 심장 이식을 받지 않은 군의 분포를 다르게 하는 유일한 요인이 변수 L이라면, 관찰 자료와 조건부 무작위 시험은 논리적으로 봤을 때 동일하다고 볼 수 있다. 이러한 교환가능성 아래에서 표준화 (standardization) 혹은 역 확률 가중치 (IPW)가 평균 인과효과를 산출하기 위해 사용될 수 있다. 그러나 조건부 교환가능성 (conditional exchangeability)이 만족되지 않은 경우는 변수 L 외에 측정되지 않은 요인들이 존재하는 경우인데, 이 경우 조건부 교환가능성에 근접할 수 있도록 분석을 수행할 때 충분한 데이터를 확보가능한지 살펴봐야할 필요가 있다.
이는 조사자들이 몇몇 개인들에게는 치료 수준 A=1을, 다른 사람들에게는 치료 수준 A=0을 배정하였다고 볼 수 있는데, 즉 연구자들은 일부의 연구대상자가 각 치료 그룹에 속할 수 있도록 치료를 배정해야 한다. 즉, 연구대상자들이 각각의 치료 수준에 배정될 양의 확률 (positive probability)이 존재함을 알고 있어야 하며,	이 연구에서 연구자는 모든 환자를 심장 이식 수술을 받는 군 (A=1)과 심장 이식 수술을 받지 않는 군 (A=0) 중 하나의 군으로 배정한 후, 각 군에서의 사망률의 차이로 심장 이식의 효과를 추정할 것이다. 각 군에서의 사망률의 차이를 구하기 위해서는 각 군에 속하는 환자가 적어도 1명 이상이어야 하며, 즉, 각 환자가 각 군에 배정될 확률이 0보다 커야한다.
또한, 양의 조건은 상호교환성에 필요한 변수 L에 대해서만 필요하다. 예를 들어, 표 3.1의 조건부 무작위 실험에서 “연구대상자의 파란 눈 유무” 변수는 치료를 받은 사람과 치료를 받지 않은 사람 사이의 교환가능성을 달성하는 데 필요하지 않기 때문에 파란 눈을 가진 개인에서 치료를 받을 확률이 0보다 큰지 여부를 묻지 않는다. 즉, L만 보정하여도 표준화된 위험도와 역 확률 가중치를 적용한 위험도는 반사실적 위험과 동일하며, “연구대상자의 파란 눈 유	또한, 양의 조건은 조건부 교환가능성의 성립에 필요한 변수 L에 대해서만 필요하다. 예를 들어, 표 3.1의 조건부 무작위 실험에서 “연구대상자의 파란 눈 유무” 변수가 있다면 이 변수는 치료를 받은 환자와 치료를 받지 않은 환자 사이의 교환가능성을 달성하는 데 필요하지 않기 때문에 파란 눈을 가진 환자에게서 치료를 받을 확률이 0보다 큰지 여부를 묻지 않는다. 즉, 조건부 교환가능성의 성립에 요구되는 변수 L만 보정하면 표준화된 위험도와 역 확률 가

수정 전	수정 후
무"와 같이 보정할 필요가 없는 변수에는 양의 조건이 적용되지 않는다. 관찰연구는 양의 조건도 상호교환성도 만족하지 않는다. 예를 들어, 아래의 그림과 같이 만약 의사가 예후가 좋지 않은 상태 (L=1)인 상태일 때는 무조건 심장이식 (A=1)을 한다면, Pr[A=0|L=1]=0이 되고 이로 인해 양의 조건은 만족할 수 없다. 그러나, 양의 조건이 상호교환성과 다른 점은 양의 조건은 때때로 경험적으로 검증될 수도 있다는 것이다. 예를 들어, 표 3.1이 관찰연구라고 가정할 경우, 우리는 L의 모든 수준 (즉, L= 0 및 L=1)에서 모든 수준의 치료 (즉, A=0 및 A=1)에 해당되는 연구대상자들이 있기 때문에 L에 대해 양의 조건이 성립한다고 결론지을 수 있다. 또한 이러한 양의 조건이 만족해야 표준화 (standardization)와 역 확률 가중치 (IP weighting)를 활용할 수 있는데, 만약 양 (positivity)의 조건이 위배될 때 두 방법이 사용되기 어려운 이유는 아래의 그림을 통해 이해할 수 있다. 만약 L=1일 때 치료를 받지 않은 사람들이 존재한다면 Pr[A=0|L=1]=0이고, 해당 데이터는 L=1일 때 치료받지 않은 연구대상자들이 존재하지 않기 때문에, 모든 치료군들이 치료를 받지 않았더라면 어떠한 일이 일어날지를 시뮬레이션 할 수 있는 정보가 없다.	중치 방법을 사용한 얻은 위험은 잠재적 위험과 동일하기 때문에 "연구대상자의 파란 눈 유무"와 같이 보정할 필요가 없는 변수에는 양의 조건을 적용할 필요가 없다. 관찰 연구에서는 양의 조건이 만족하지 않을 수 있다. 예를 들어, 그림 25과 같이 만약 의사가 위독한 환자 (L=1)에게 무조건 심장이식을 수행한다면 (A=1), $P[A=0 \mid L=1]=0$이 되고 이로 인해 양의 조건은 성립할 수 없다. 그러나, 양의 조건이 교환가능성과 다른 점은 양의 조건은 자료로부터 검증할 수 있다는 것이다. 예를 들어, 표 3.1이 관찰연구로부터 수집한 자료라 하면 우리는 변수 L의 모든 수준 l (0 또는 1)에서 심장 이식을 받은 환자가 있는지 없는지 셈하여 양의 조건이 성립하는지 검증할 수 있다. 이러한 양의 조건이 성립해야 표준화 (standardization)와 역 확률 가중치 (IPW)를 활용할 수 있으며, 만약 양 (positivity)의 조건이 성립하지 않을 때, 두 방법이 사용되기 어려운 이유를 아래의 그림을 통해 이해할 수 있다. 만약 변수 L이 1 일 때, 치료를 받지 않은 사람이 없다면 $\Pr[A=0 \mid L=1]=0$이고, 표준화 방법에서는 변수 L이 1의 값을 가지며, 치료를 받지 않은 환자와 비교할 대상이 없어 표준화 방법이 적용되기 어려우며, 역 확률 가중치 방법의 경우 확률이 0이므로 가중치를 계산할 수 없어 적용되기 어렵다.
일관성을 만족하기 위해서는 첫 번째로, 반사실적 결과 (counterfactual outcome)를 정의해야 한다. "일관성 (consistency)"은 치료를 받은 모든 연구대상자들에 대해 관찰된 결과가 한 연구대상자가 치료를 받아 얻은 결과와 동일하고, 반대로 치료를 받지 않은 결과에 대해서도 동일하다는 것을 의미한다. 즉, 수식으로 표현하면 모든 연구대상자 A=a 에 대해 $Y^a = Y$로 나타낼 수 있다. 일관성의 주요 구성요소는 다음과 같다.	일관성을 만족하기 위해서는 잠재적 결과를 먼저 구체적으로 정의해야 한다. "일관성 (consistency)"는 치료를 받은 환자들에게서 관찰한 결과가 해당 환자가 치료를 받았을 때 얻게 될 결과와 동일하고, 반대로 치료를 받지 않은 환자들에게서 관찰한 결과가 해당 환자가 치료를 받게 되지 않았을 때 얻게 될 결과와 동일하다는 것을 의미한다. 즉, 수식으로 표현하면 모든 연구대상자에 대하여 치료 수준 a을 받았을 때, $Y^a = Y$ 로 나타낼 수 있으며, 또는

수정 전	수정 후
(1) 반사실적 결과 Y^a 의 정확한 정의는 세분화한 “a”의 분류를 통해 나타낼 수 있음 (2) 관찰된 결과에 대한 반사실적 결과의 연결성이 존재함 **(1) 일관성 (consistency): 반사실적 결과의 정의** 이 중, (1)번 내용을 중심으로 일관성의 조건을 살펴보겠다. 이러한 일관성의 조건은 서로 다른 버전의 치료가 각각 다른 인과효과를 갖는 경우에 문제가 발생할 수 있는데, 예를 들어 전통적인 수술기법을 사용한 연구에서의 심장이식에 관한 평균효과는 새로운 수술기법을 사용한 연구결과와 다를 수 있다. 따라서, 심장이식 A가 사망률 Y에 미치는 영향을 보고자 할 때 관심 있는 치료의 버전 “a”에 대한 부분을 구체화시킬 필요가 있다. 관찰연구에서는 연구자들이 이러한 “a”를 가능한 명확하게 지정해야 한다. 만약 비만에 의한 사망 연구에서 연구대상자들이 비만이었던 기간, 가장 최근 시점에서 비만 유무, 비만의 강도 등이 정의될 수 있다면 여러 측면에서의 개입 (intervention)들을 구체화할 수 있다. 예를 들어 한 연구대상자가 허리와 관상동맥의 지방조직을 증가시키기는 유전적 변형이 있거나, 높은 칼로리 섭취에 비해 적극적인 신체활동을 하지 않는 경우, 또는 장내 미생물군이 변하였거나, 수술을 하였거나 등의 여러 개입들을 고려할 수 있다. 문제는 이러한 선택지 각각이 어떻게든 비만도를 같은 수준으로 만든다고 하더라도 사망률에 다른 영향을 미칠 수 있다는 것이다.	$$Y = A \cdot Y^{a=1} + (1-A) \cdot Y^{a=0}$$ 와 같이 표현되기도 한다. 일관성의 주요 구성요소는 다음과 같다. (1) 서로 다른 치료 방법에 대한 잠재적 결과 Y^a을 가능한 구체적으로 기술해야 함. (2) 관찰된 결과와 잠재적 결과의 연결성이 존재함. **(1) 일관성 (consistency): 반사실적 결과의 정의** 앞서 언급한 일관성에 관한 두 가지 주요 구성 요소 중 (1)번 내용을 중심으로 일관성의 조건에 대해 살펴보고자 한다. 이러한 일관성의 조건은 어떤 치료에 대해 여러 방법이 있는데도 불구하고 이를 구분하지 않고 치료에 대한 효과를 추론하고자 할 때, 문제가 발생할 수 있다. 예를 들어 전통적인 수술 기법에 대한 평균 인과효과를 추론하기 위한 연구에서 전통적인 수술 기법의 심장 이식에 대한 평균 인과효과는 새로운 수술 기법을 사용한 연구결과와 다를 수 있다. 따라서 심장 이식 여부가 환자의 심근경색 여부에 미치는 영향을 볼 때, 관심 있는 치료 방법에 대해 구체화하여 기술할 필요가 있다. 특히 관찰연구에서 연구자들은 여러 치료 방법 중 관심 있는 치료 방법을 가능한 명확하게 지정해야 한다. 만약 비만의 사망에 대한 효과를 추정하기 위한 연구에서 비만을 비만이었던 기간, 가장 최근 시점에서 비만 유무, 비만의 강도 등을 통해 다양하게 정의할 수 있기 때문에 단순히 ‘비만’이라고 정의하는 것이 아니라 연구 가설에 부합하는 구체적인 비만의 정의를 기술하여 평균 인과효과의 크기를 산출할 수 있다.
보자 (Y=0). 이 때 비만인 경우 a=1일 때, 철수의 반사실적 결과 (counterfactual outcome) $Y^{a=1}$ 는 무엇일까? 우리는 그가	생각해보자 (Y=0). 이 때 비만인 경우의 철수의 잠재적 결과 $Y^{a=1}$은 어떻게 정의되어야 할까? 그가 비만 유전자로 인해 비만

수정 전	수정 후
비만 유전자로 인해 비만이 발생하였을 때 (A=1) 사망했지만, 운동 부족과 고열량 섭취로 인해 비만이 발생하였을 때는 (A=1) 죽지 않았을 것이라고 말했다. 즉, a=1일 때, 반사실적 결과 $Y^{a=1}$는 잘 정의되지 않았다.	이 발생 (A=1)했을 때, 심근경색이 발생하였지만, 운동 부족과 고열량 섭취로 인해 비만이 발생 (A=1)했을 때의 경우, 심근경색이 발생하지 않았다. 두 경우 모두 철수에게 비만이 발생하였지만, 비만이 생기게 된 경로에 따라 심근경색이 발생하기도 하고, 발생하지 않기도 하였다. 이와 같이 치료 또는 노출 변수를 분명히 정의하지 않으면 잠재적 결과 $Y^{a=1}$ 또한 명확히 정의되지 않는다.
이전에 등장한 내용을 고려하여, 연구자는 비만이 50세까지 사망률에 미치는 영향이라는 모호한 인과 질문을 보다 정확한 인과 질문으로 바꾸기로 결정했다. 연구자는 이제 다음과 같은 개입 (a=1)에 관심이 있다. 18세에서 40세까지 모든 연구대상자들을 대상으로 18세 당시의 체중보다 더 많은 체중이 나가지 않도록 하는 엄격한 식단을 따라야 한다. 특히, 각 연구대상자들은 18세 생일 전날부터 매일 체중을 쟀고, 18세 당시의 체중보다 클 때 기준치에 해당하는 체중 아래로 떨어질 때까지 (보통 1~3 일 이내) 칼로리 공급원과 미량 영양소의 일반적인 조합을 변경하지 않은 상태에서 열량 섭취는 제한한다. 따라서 1~2 kg의 오차를 무시하면 40세가 될 때까지 어떤 연구대상자도 기준 체중보다 더 많이 나가지 않을 것이다. 칼로리 제한이 없는 기간 동안의 운동이나 식단에 관한 지침이나 제한은 없다. 비교하고자 하는 개입 (a = 0)은 “개입하지 않음”이다. 전문가들이 이러한 개입 값 a=1 및 a=0이 충분히 잘 정의되어 있으므로 반사실적 결과 $Y^{a=1}$ 및 $Y^{a=0}$에 모호함이 남아 있지 않다는데 동의한다고 가정하자. 이제 A=a인 연구대상자에 대한 일관성의 조건 $Y^{a}=Y$로 넘어갈 수 있다. 다른 측면을 생각해보기 위해, 연구자의 엄격한 개입 a=1을 받지 않았음에도 불구하고 18세에서 40세 사이에 거의 일정한 체중을 유지한 성호를 생각해 보자. 성호는 좋은 유전자를 가지고 있고,	앞선 예에서 연구자는 비만이 50세까지 사망률에 미치는 영향이라는 모호한 연구 가설을 보다 구체적인 연구 가설로 수정하기로 하였다. 연구자는 이제 다음과 같은 개입 (a=1)에 관심이 있다. 18세에서 연구를 시작하여 40세까지 모든 연구대상자들은 18세 당시의 체중보다 더 많은 체중이 나가지 않도록 하는 엄격한 식단을 따라야 한다. 특히, 각 연구대상자들은 18세 생일 전날부터 매일 체중을 쟀으며, 18세 당시의 체중보다 커지면 기준치에 해당하는 체중 아래로 떨어질 때까지 (보통 1~3 일 이내) 칼로리 공급원과 미량 영양소의 일반적인 조합을 변경하지 않은 상태에서 열량 섭취를 제한한다. 따라서 1~2 kg의 오차를 무시하면 40세가 될 때까지 어떤 연구대상자도 18세에 측정한 기준 체중보다 더 많이 나가지 않을 것이다. 칼로리 제한이 없는 기간 동안의 운동이나 식단에 관한 지침이나 제한은 없다. 비교하고자 하는 개입 (a = 0)은 “개입하지 않음”이다. 이때, 전문가들이 이러한 개입은 충분히 잘 정의되어 있으며, 잠재적 결과 $Y^{a=1}$ 및 $Y^{a=0}$에 대해서도 모호함이 남아 있지 않다는 것에 동의한다고 가정하자. 이제 이러한 개입 a (a는 0 또는 1)인 연구대상자에 대한 일관성의 조건 $Y^{a}=Y$을 기술할 수 있다. 모든 연구대상자는 두 잠재적 결과 $Y^{a=1}$과 $Y^{a=0}$에 대한 정보를 모두 가지고 있지만 실제 처치된 치료는 하나이기 때문에 두 잠재적 결과를 모두 관측할 수 없으며, 이 중 하나만 관찰된다. 이

수정 전	수정 후
평소 활발한 신체활동을 하여 기본 체중을 유지했다. 따라서 성호의 관찰된 치료 값은 A=1이 아니므로 관찰된 결과 Y가 연구자의 가상 개입 a=1을 받았다면 경험했을 반사실적 결과 $Y^{a=1}$과 반드시 같지는 않다. 반사실적 결과 $Y^{a=1}$과 관찰된 결과 Y 사이의 연결을 유지하기 위해 분석에서 치료 버전 a=1을 받는 연구대상자만 치료된 연구대상자 A=1로 간주하고, 치료되지 않은 연구대상자도 이와 유사하게 간주한다. 이것이 의미하는 바는, 관찰 데이터를 사용하여 인과효과 Pr[Ya=1=1] − Pr[Ya=0=1]을 정량화하려면, 일부 연구대상자가 a=1 및 a=0과 일치하는 치료 값을 받은 데이터가 필요하다는 것이다. 즉 (비조건부) 양의 조건이 필요하다. 즉 개입을 잘 정의하였더라도 그 개입이 관찰된 데이터와 연결될 수 없는 경우 (즉, Ya=Y가 적어도 일부 연구대상자에 대해 성립한다고 합리적으로 가정할 수 없는 경우) 무용지물이 될 수 있다.	것을 수식으로 표현하면 다음과 같이 표현할 수 있다. $$Y = A \cdot Y^{a=1} + (1-A) \cdot Y^{a=0}$$ 이 식은 연구대상자가 치료를 받으면 (A=1), 잠재적 결과 $Y^{a=1}$가 관측되고, 치료를 받지 않으면 (A=0) 잠재적 결과 $Y^{A=0}$가 관측된다는 내용을 명확히 표현하고 있다. 그러나 예를 들어 비만을 획득하게 된 경로에 대한 자료가 종종 충분하지 않은 경우가 있다. 예를 들어 40세의 체중에 대한 자료를 수집하였지만, 개인의 체중, 운동습관 및 식이요법에 대한 평생 이력에 대한 자료는 수집하지 못 한 “비만에 관한 연구”가 있을 수 있다. 이 문제에서 벗어나는 한 가지 방법은 모든 치료 버전의 효과가 동일하다고 가정하는 것이다. 예를 들어, 뇌졸중에 대한 고혈압 대 정상 혈압의 인과관계에 관심이 있는 경우, 경험적 증거에 따르면 다양한 약리학적 기전을 통해 혈압을 낮추면 유사한 결과가 나타난다. 이 경우 잠재적인 결과와 관찰된 결과를 연결하기 위해 치료 “혈압을 낮추는 경로”에 대한 내용이 불필요하다고 주장할 수 있다. 그러나 다른 경우에는 이 가정에 대한 합리성이 의심될 수 있다.
표준화에 대해 알아보자. 심장이식 연구가 조건부 무작위 시험이라고 할 때 심장이식을 받은 군 (treatment group, A=1)은 무작위 할당을 통해 위험하지 않은 상태 (L=0)의 8명의 개인에게 50%의 심장이식 확률을 배정하였고, 12명의 개인이 위험한 상태 (L=1)일 때의 심장이식 확률 75%를 가정한다. 먼저, 8명의 위험하지 않은 상태인 사람들을 살펴보았을 때, 치료군 (A=1)에서의 사망위험은 Pr[Y=1|L=0, A=1] = 1/4 이고, 비치료군 (A=0)에서의 사망위험은 Pr[Y=1|L=0, A=0] =1/4 이다. 이때 treatment는 무	우리는 2장 3절에서 표준화 방법을 통하여 평균 인과효과를 계산하는 방법에 대하여 설명하였다. 2장 3절의 예제의 경우, 조건부 교환가능성 조건을 위해 요구되는 변수가 예후 인자 L 하나였다. 하지만 조건부 교환가능성의 성립을 위해 요구되는 변수들의 수가 많거나 결과 변수에 대한 정보의 부족으로 결과 변수에 대한 분포를 가정해야 하는 경우에는 모델 기반 (model-based)의 모수적 (parametric) 표준화 방법이 사용될 수 있다. 2장 3절에서 설명한 표준화 방법은 결과 변수의 분포에 관해 어떠한 가정도 하지 않았

수정 전	수정 후
작위 시험이기 때문에, 사망(Y)과 A|L 간에는 독립성이 만족되며, 즉 위험하지 않은 상태(L=0)인 그룹에서 치료받는 사람의 위험은 모든 사람이 치료를 받았을 때의 위험과 동일하다. 수식으로 표현하면 Pr[Y=1|L=0, A=1] = Pr[$Y^{a=1}$= 1 | L=0] 과 같다. 반대로 12명의 위험한 상태(L=1)인 개인에서 살펴보았을 때, Pr[Y=1|L=1, A=1] = Pr[$Y^{a=1}$ = 1 |L=1] = 2/3, Pr[Y=1|L=1, A=0] = Pr[$Y^{a=0}$ = 1 |L=1] = 2/3 이다. 우리의 관심사가 인과 위험 비 (causal risk ratio)이라고 할 때 Pr[$Y^{a=1}$ =1]/Pr[$Y^{a=0}$=1]로 표현할 수 있다. 여기서의 분자는 20명의 모든 사람들이 치료를 받았을 때의 risk인데, 이는 위험하지 않은 그룹(L=0)과 위험한 그룹(L=1)의 각각의 위험도에 대한 weighted average 값을 의미하며 값으로 나타내면 1/4 * 0.4 + 2/3 * 0.6 = 0.5 이다. 분모의 경우에도 마찬가지로 0.5 이므로 인과적 위험 비는 0.5/0.5 = 1 이다. 즉, 이를 수식적으로 표현하면 다음과 같다. $\Pr[Y^a=1] = \Pr[Y^a=1\|L=0]\Pr[L=0] + \Pr[Y^a=1\|L=$ $\Pr[Y^a=1] = \sum_l \Pr[Y^a=1\|L=l]\Pr[L=l]$ 표준화된 위험도 (standardized risk)는 '조건부 상호교환성 (conditional exchangeability)'의 특성으로 인해 반사실적 위험도 (counterfactual risk)와 동일한 개념으로 볼 수 있기 때문에, 인과 위험 비	기 때문에 비모수적 (non-parametric) 방법이라 불린다. 모수적 표준화는 앞서 설명한 비모수적 방법과 다르게, 많은 수의 공변량으로 생길 수 있는 고차원 적 문제와 연속형 치료 변수에 대한 문제를 해결하기 위해 활용될 수 있는 모델 기반의 방법이다. 이러한 모수적 표준화 방법을 사용하여 평균 인과효과를 추론하기 위해서는 치료 변수와 교란 요인을 조건으로 하는 결과 변수에 대한 조건부 기댓값을 추정해야한다. 조건부 기댓값을 추정할 때, 결과 변수의 분포에 따라 선형 회귀모형(linear regression model), 로지스틱 회귀모형(logistic regression model)을 포함하는 일반화 선형 모형(generalized linear model) 등의 모형을 사용할 수 있다. 2장 3절에서 기술했던 것과 같이 조건부 교환가능성 아래에서 잠재적 결과 Y^a의 기댓값 $E[Y^a]$은 치료를 받은 군과 치료를 받지 않은 군에 대한 가중 평균 (weighted mean) $$E[Y^a] = \sum_l E[Y \mid A=a, L=l] \cdot \Pr[L=l]$$

수정 전	수정 후
$\frac{\Pr[Y^{a=1}=1]}{\Pr[Y^{a=0}=1]}$ 는 $\frac{\sum_l \Pr[Y=1\|L=l, A=1]\Pr[L=l]}{\sum_l \Pr[Y=1\|L=l, A=0]\Pr[L=l]}$ 의 표준화된 형태로 계산할 수 있다. 다만 보정해야 하는 변수들의 수가 많거나 데이터의 한계로 분포를 가정해야 하는 경우에는 모델기반 (model- based)의 모수적 표준화 (parametric standardization) 방식을 활용할 수 있다. 모수적 표준화 (parametric standardization)는 앞서 설명한 비모수적 (non-parametric) 방법과 다르게, 많은 수의 공변량과 이분형이 아닌 (non-dichotomous) 치료에 대한 고차원적 문제를 해결하기 위해 활용될 수 있는 모델기반 (model-based)의 분석 방법이다. 이러한 모수적 추정치(parametric estimates)를 얻기 위하여, 치료군과 공변량으로 들어가는 교란변수들을 활용한 선형 회귀모형을 적합할 수 있다. 치료을 한 군과 치료를 하지 않은 군 간의 표준화된 평균 (standardized mean) $\sum_l E[Y\|A=a, C=0, L=l] * \Pr[L=l]$ 의 추정은 E[Y|A=a, C=0, L=l]의 가중평균값과 같다. 즉, 표준화된 평균은 E[E[Y|A=a, C=0, L]]와 같이 double-expectation 형태로 표현될 수 있기 때문에, 표준화된 평균에서의 Pr[L=l]을 구하지 않아도 추정값을 계산할 수 있다. 모수적 표준화를 수행하는 방법은 크게 데이터의 확장 (expension of dataset), 결과변수 모델링 (outcome modeling), 예측 (prediction), 그리고 평균화 (averaging)를 통한 표준화 네 가지 단계로 진행된다. 예를 들어	

수정 전	수정 후
설명하면 다음과 같다. 만약 조사대상자가 20명이 있다고 가정하자. 20명의 데이터 값을 세 번 확장하였을 때, 첫 번째 데이터셋을 원 데이터로 두고 나머지 두 번째와 세 번째 데이터셋을 각각 치료군과 비치료군의 데이터로 나눈다. 이를 첫 번째 단계인 데이터의 확장이라고 볼 수 있다. 그 다음으로 3개로 분류된 데이터셋을 이용하여 치료변수와 교란변수가 주어졌을 때의 평균 결과 (mean outcome)를 구하기 위한 회귀모형을 적합한다. 이 때 두 번째와 세 번째 데이터셋에서는 결과 값이 결측이기 때문에, 원 데이터가 들어있는 첫 번째 데이터셋만이 모델의 모수추정에 기여할 수 있다. 이 단계가 결과변수 모델링이다. 세 번째 단계로는 첫 번째 데이터셋으로부터의 모수추정치를 이용하여 두 번째와 세 번째 데이터셋의 모든 결과 값들을 예측 (predict)한다. 예측된 결과 값들은 각각 두 번째 데이터셋에서는 교란변수와 치료받지 않은 값의 조합으로 이루어진 평균 추정치이고, 세 번째 데이터셋에서는 교란변수와 치료된 값의 조합으로 이루어진 추정치라고 볼 수 있다. 마지막 단계로, 예측된 결과 값들을 평균해서 평균 결과를 추정하도록 한다.	

4. 2021년 작성된『직업병 인과추론 가이드라인: g-formula 국문 가이드라인』에 대한 자문 의견 반영

의견 1) "이러한 상황이 벌어졌을 때, 잠재적 결과를 추측하는 과정을 진행하고 있기 때문에 반드시 배정받는다는 의미인 것인지요?"

수정 사항: 부록 2의 Ⅰ.인과추론 용어의 정리에서 1장 인과효과의 정리에서 잠재적 결과의 의미에 대해 자세히 기술하였음.

의견 2) "g-formula의 이론 소개 부분에서 왜 전통적인 방식으로 분석한 결과가 편향된 결과를 가지게 되는지 그 이유에 대하여 한 번 더 짚어 주시면 좋겠습니다."

수정 사항: 부록 2의 Ⅰ.인과추론 용어의 정리에서 6장 전통적인 회귀분석과 g-formula의 차이에서 전통적인 방식으로 분석한 결과가 편향을 가지는 결과인지 인과 그래프를 통하여 기술하였음.

의견 3) "109페이지에서 건강근로자 생존 편향에 '취약'한 상황이라는 것이 어떤 의미인지요? 건강근로자 생존 편향이 잘 발생할 수 있다는 의미인 것인지요?"

수정 사항: 예를 들어, 라돈에 노출될 수 있는 작업장에서 근무하는 근로자 집단이 있다고 생각해보자. 이 작업장에서 건강한 근로자일수록 더 오래 근무하게 될 것이고, 오래 근무하게 됨에 따라 라돈에 대한 누적 노출량이 많아질 것임. 이러한 상황에서 라돈에 대한 누적 노출량과 근로자의 건강 지표 사이의 관계를 분석하게 되면 라돈에 대한 누적 노출량이 증가할수록 건강 지표가 개선되는 것으로 나오게 됨. 그 이유는 건강하지 않은 근로자는 중간에 직업을 그만두게 되어 자료에는 건강한 근로자만이 남게 되

기 때문임. 이러한 예시처럼 건강한 근로자가 자료에 많이 남게 되는 상황이 건강근로자 생존 편향에 취약한 상황임.

의견 4) "자주 사용되는 용어 및 개념에 대한 해설 또는 설명을 따로 분리해서 추가해주면 어떨까 싶습니다. 대부분 메뉴얼 내 줄글로 잘 설명이 되어있긴 하나, 내용에 관한 전반적인 이해력이 부족한 상태에서 매뉴얼을 읽어 내려가다 보니 중간중간 주요 용어에 대한 개념이 헷갈리는 경우가 있습니다. 따라서 소 부록처럼 용어 해설 형태의 파트가 따로 하나 있으면 필요할 때마다 해당 파트를 참고하며 읽어 내려갈 수 있을 것 같습니다."

수정 사항: 용어의 개념에 대한 어려움을 해소하고자 부록 2의 Ⅰ. 인과추론 용어의 정리를 작성하였음.

의견 5) "KMR에 대하여 통계 초급자도 이해할 수 있는 수준으로 정리되면 좋겠습니다."

수정 사항: 부록 1-Ⅱ. BKMR 이론과 적용-2. BKMR 이론-1) KMR의 개요 에 KMR에 대한 설명을 자세히 기술하였음.

의견 6) "BKMR은 정규성 가정이 필요 없는 비모수적 방법으로 분류되나요? 분석 진행 시 노출 변수와 결과 변수의 분포가 어떤 형태이건 관계없는지 알려주면 좋겠습니다."

수정 사항: 로지스틱 회귀분석(logistic regression)을 포함한 일반화 선형모형(generalized linear model; GLM)에서와 같이 BKMR에서 또한 결과 변수의 분포에 따라 정규성 가정의 필요 여부가 결정됨. 예를 들어, 결과 변수가 정규 분포를 따를 것이라 생각하면 정규성 가정을 하게 되며, 그렇지 않은 경우, 정규성을 가

정하지 않아도 됨. 하지만 BKMR은 일반화 선형 모형에서 결과 변수의 기댓값을 $X\beta$ 와 같이 모형화하는 것과 달리 결과 변수에 대해 커널을 사용함으로써 비모수적 모형을 사용함. 다만, 분석 진행 시 결과 변수의 분포와 무관하게 BKMR의 적용이 원리적으로는 가능하나 현재 R 패키지에는 정규 분포를 따르는 연속형 결과 변수와 이분형 결과 변수에 대해서만 분석이 가능함.

의견 7) "모형에 들어가는 노출 변수들의 scale이 상이한 경우에 effect size가 다르게 나타날 수 있다는 이야기가 있던데, 이 부분은 고려하지 않아도 되나요?(예를 들면, 혈중 카드뮴 농도와 소득을 동시에 변수로 보고 싶은데, 소득수준을 원 단위로 해서 1 ~ 10,000,000으로 넣는 경우와 표준화해서 -1부터 1 사이의 값으로 변환하여 넣는 경우에 해당합니다)."

수정 사항: 선형 회귀분석 모형에서 보정되는 변수의 단위의 크기 차이로 인하여 생기는 해석상의 어려움은 BKMR에서도 나타남. 즉, 선형 회귀분석에서 보정하는 변수의 단위를 모두 표준화(standardized)한 후 분석을 진행하게 되면 표준화하기 전과 비교하여 회귀계수가 달라지며, 변수를 표준화함으로써 변수의 해석이 어려워짐. 그러므로 BKMR에서도 노출 변수를 표준화하여 사용하는 경우, 표준화하기 전과 비교하여 effect size가 달라질 수 있음.

의견 8) "그림에 대한 해석을 간략히 적어주면 좋을 것 같습니다. 해당 내용을 통해 독자가 매뉴얼대로 적용해본 후 추출된 결과를 제대로 해석한 것인지에 대해 점검해볼 수 있을 것 같습니다."

수정 사항: 그림에 대한 해석을 부록1-Ⅱ-3. BKMR의 적용 부분에 추가 작성하였음.

의견 9) "시점 문제에 관해 예를 들어 1번 환자가 0,1,3개월에 방문하였고, 이것을 재정의 했을 때, 0,1,2시점의 측정으로 수정되고, 2번 환자의 0,4개월 방문이 0,1시점으로 수정된다면 1번 환자와 2번 환자의 '1' 이라는 시점은 물리적 시간으로는 다른 의미인데 이러한 데이터의 구조(?)가 분석에 미치는 영향은 없는 것인지요. 아니면 이러한 부분 또한 분석 단계에서 고려가 되어 분석이 이루어지는 것인지요."

수정 사항: 현재 개발된 g-formula는 연구대상자가 규칙적으로 방문하는 것(regular visit)을 전제로 하고 있음. 하지만 특수건강진단 자료 내 근로자들은 여러 가지 이유로 매년 받지 않는 사람이 존재함. 그러므로 특수건강검진자료를 통해 확인되는 근로자의 측정시점은 규칙적이지 않으며 비규칙적 방문(irregular visit)은 현재 개발되어 있는 g-formula로 반영하기 어려움. 비규칙적 방문을 고려한 g-formula 방법의 개발이 필요함.

의견 10) "g-formula 파트도 BKMR처럼 분석 결과에 따른 결과물 예시를 보여주고, 이에 관한 해석을 언급해주면 이해를 하는데 좀 더 편리할 것 같습니다."

수정 사항: g-formula에 대한 구체적인 예시는 예신희 등 (2021)에서 코드에 대한 설명과 함께 다루었음.

5. g-formula 및 BKMR에 대한 검토 및 장단점 평가

1) g-formula와 BKMR의 장점과 단점을 비교하는 목적

- g-formula는 산업보건 역학연구에서 사용되는 인과추론 통계 방법론 중의 하나로 2개 이상의 복합노출에 의한 건강 영향을 인과적으로 추론할 수 있는 방법임.
- BKMR은 최근 환경 역학연구에서 많이 사용되고 있는 복합노출의 건강영향을 평가하는 통계 방법론 중 하나로, 여러 노출 변수에 대한 건강영향을 평가할 수 있고 추정치의 다양한 시각화가 가능하며, 모델이 유연하다는 장점을 가지고 있음.
- 이 두 통계 방법론은 반복 측정된 자료에서 복합노출에 의한 건강 영향 평가가 가능하다는 공통점을 가지고 있는 반면, 서로 다른 장점과 단점을 가지고 있음. 따라서, 이 두 통계 방법론의 장점과 단점을 비교하여, 반복 측정된 산업보건 역학 자료에서 복합노출로 인한 건강 영향을 평가하기에 적절한 통계 방법론에 대해 논의해보고자 함.

2) g-formula의 장점과 단점

- g-formula 장점은 g-formula는 반복 측정된 자료를 분석할 때, 치료-교란 요인 되먹임의 존재를 반영할 수 있고, 건강근로자 생존 편향과 같은 산업보건 역학연구에서 발생할 수 있는 선택 편향을 효과적으로 통제할 수 있다는 것에 있음. g-formula는 계산시간이 오래 걸리는 단점이 있는 것으로 알려져 있지만, BKMR에 비해서는 분석 속도가 월등히 빠름. 또한, 연속형 노출 변수 외에 범주형 노출 변수 또한 분석이 가능함. 나아가, marginal causal effect에 해당하는 지표로 위험 비 (risk ratio), 위험 차이 (risk difference) 그리고 오즈 비 (odds ratio)를 구할 수 있음. g-formula로 계산한 marginal causal effect의 추정치는 marginal structural model과 g-estimation 등 다른 인과추론 방법론을 통해 일관된 값이 나오는지 확인하여 분석 결과의 신뢰성을 일부 평가할 수 있음.
- g-formula 단점은 g-formula는 복합노출에 의한 건강 영향을 추정할

수 있으나, 반복측정된 자료에서 교호작용을 평가하는 접근법에 대해서는 아직 잘 알려져 있지 않음. 분석한 결과를 BKMR과 같이 다양하게 시각화하기 위해서는, 현재 많이 사용되는 R 패키지에 더하여 추가적인 작업이 필요함. 선행 산업보건 역학연구에서 사용한 g-formula는 모수적 모형 (parametric model)을 사용하였으며, BKMR보다 덜 유연한 모형임.

3) BKMR의 장점과 단점

- BKMR의 장점은 반복측정된 자료를 분석할 수 있고, 많은 수의 노출 변수에 대해 분석이 가능하다. BKMR은 g-formula와 달리 시각적으로 교호작용을 평가할 수 있음. 또한, BKMR은 교호작용 외에도, 사후포함확률 (posterior inclusion probability; PIP), 단일 노출-반응 함수, 복합노출-반응 함수 등을 시각적으로 보여줌. BKMR은 모형 내에서 복합노출의 다양한 고차원 항 또는 교호작용 항을 반영하기 위해 커널 행렬 (kernel matrix)을 이용한 혼합 모형 (mixed model)을 사용하기 때문에 모수적 모형을 사용한 g-formula보다 모형의 유연성을 확보할 수 있음. 이를 통해 많은 노출 변수와 결과 변수 사이의 관계를 유연하게 모형화할 수 있음.
- BKMR의 단점은 반복측정 자료를 분석할 때 발생하는 치료-교란 요인 되먹임이나 선택 편향을 다룰 수 없음. 또한, 결과를 시각적으로 보여주어 직관적이지만, 노출량의 사분위 수를 개입하는 양 (intervention)으로 설정하기 때문에 해석이 쉽지 않음. 선행연구에 따르면, 노출 변수가 많아지고 데이터가 복잡해질수록, 다른 복합노출 통계 방법으로 분석한 결과와 비교했을 때 그 결과들이 일관되지 않다는 보고가 있으며, 따라서 노출 변수가 많고 자료의 형태가 복잡할 경우, 분석 결과의 신뢰성이 낮을 수 있음.

6. 후속 연구계획 수립

1) 2023년 연구 내용 및 방법

▪ 2022년 7월 22일에 연구진 회의를 통해 다음과 같이 후속 연구계획안을 수립하였음.

〈표 Ⅲ-9〉 복합노출의 건강 영향평가 통계분석법 개선안

항목	BKMR	g-formula
인과추론이 가능한 활용 데이터	• 반복측정 자료의 경우에서 외생 노출 변수에 대해 활용 가능함.	• 반복측정 자료에서 활용 가능함. • 내생 노출 변수에 대해서도 활용 가능
개선안	• 분석 시간 단축 • 이분혈 결과 변수에 대하여 현재는 probit model만 가능하기 때문에 logistic regression model로 확장이 필요함. • 반복측정 자료에서 intercept뿐만 아니라 slope에도 random effect를 적용할 수 있도록 하는 것이 필요함.	• 시각화 코드화 (dose-response curve, interaction등) • 분석 결과의 안정성 평가
주의할 점 (올바른 분석 방법 가이드)	• 노출 자료의 수 • tuning parameter의 올바른 활용; 다만, 집중적인 수치 연구(numerical study)가 필요함.	

2) 2023년 연구의 예상되는 기대효과 및 활용방안

▪ 복합노출과 건강 영향 간의 인과효과를 평가할 수 있는 통계 분석법을 개선하여, 국내 산업보건 집단 역학조사 및 역학연구에서 복합노출에 대한 인과추론 통계 방법의 적용을 원활하게 함.

- 이를 통해 근로자 종적 자료를 활용한 역학조사와 역학연구에서 연관성이 아닌 인과효과를 도출할 수 있도록 하여, 국내 산업보건 조사 및 연구 결과의 신뢰성을 높이고자 함.

Ⅳ. 고찰

Ⅳ. 고찰

1. 주요 결과

- 복합노출의 건강영향을 평가하는 것은 과거부터 산업보건 역학연구에서 많은 관심을 받던 주제이나, 통계 방법론의 부재 및 건강 근로자 생존 효가로 인한 연구 결과의 편향으로 인해 접근하기 힘든 부분이었음. 하지만 최근 복합노출의 건강영향을 평가할 수 있는 통계방법론들이 개발되고, 역학연구에서 이러한 통계방법론들의 활용이 증가되고 있는 추세임. 따라서 본 연구에서는 국내 산업보건 역학자들이 복합노출의 건강영향을 추론하는데 g-formula와 BKMR을 활용할 수 있도록 국문가이드라인을 작성하였음. 특히 g-formula는 시간에 따라 변화하는 노출수준과 정보적 중도절단(informative censoring)으로 인해 발생하는 건강 근로자 생존 효과를 제어하면서 반복측정된 자료에서 복합노출의 건강영향을 평가할 수 있는 방법임을 설명하였음.
- 7년 간 반복측정된 특수건강진단 자료를 g-formula로 분석하여 납과 카드뮴의 복합노출이 빈혈의 위험도를 높이고, 납과 카드뮴 간에 교호작용이 있는 것을 확인하였음. 다만 BKMR로 분석하였을 때는 이와 다른 결과를 확인하였는데, 이러한 결과는 BKMR이 노출-교란요인 되먹임 관계나 경쟁 사건(competing event)으로 인한 편향을 제어할 수 없는 연구결과이기 때문인 것으로 판단됨.
- g-formula와 BKMR의 장점과 단점을 평가하여, 단점을 해결하는 차년도 연구계획을 작성하였음.

2. 본 연구의 강점

- 복합노출의 건강영향 평가와 건강 근로자 생존 효과로 인해 발생하는 편향을 제어하는 통계 방법론은 국내 산업보건 역학연구 분야에서도 과거부터 많은 관심을 가지고 있었던 분야이나, 개념적으로 이해하기 힘들고 실제로 통계 분석을 수행하는데 기술적 장벽이 있어 활용하기 힘든 분야였음. 본 연구에서는 산업보건 역학자와 복합노출과 인과추론 역학연구 분야에 경험이 풍부한 통계학자가 협업하여, 복합노출의 건강영향을 인과적으로 평가하는데 적용할 수 있는 통계방법론을 국문으로 자세히 설명하는 가이드라인을 작성하였음. 또한, 직업환경의학 전문가 3인의 검토를 받아 이해하기 어려운 부분을 수정하고 보완하였음.
- 납과 카드뮴이 빈혈에 미치는 영향을 g-formula로 분석하였으며, 이렇게 많은 연구대상자 수를 가지고 납과 카드뮴이 빈혈에 미치는 복합노출의 건강 영향과 교호작용을 확인한 연구는 전 세계적으로 처음인 것으로 확인됨.

3. 본 연구의 제한점 및 제언

- 특수건강진단 자료에서 반복측정된 만성 노출과 건강 지표 간의 연관성을 분석할 때에는 건강 근로자 생존 효과로 인한 편향을 제어하기 위하여 g-formula를 적용하는 것이 더 적절한 방법이나, 현재 g-formula R package는 반복측정 되는 시간 간격이 일정하지 않은 경우에 대한 분석은 제공하고 있지 않기 때문에 이러한 제한점을 고려한 분석이 필요함. 또한, 데이터 기반의 모델 안정성 평가와 경쟁사건(competing event) 등을 고려한 민감도 분석이 필요함.
- 복합노출의 건강영향 평가 및 인과추론 통계분석법의 활용과 저변 확대를 위한 연수강좌 등이 필요함.

참고문헌

남정모, 김진흠, 강대룡, 안연순, 이후연, 이대희. Intermediate 변수의 영향을 통제하는 통계적 방법론에 대한 연구-건강근로자효과를 통제하기 위한 새로운 접근. Korean Journal of Epidemiology. 2002;24(1):7-16.

이경무, 전재범, 박동욱, 이원진. 건강근로자효과의 최소화 방안과 보정 방법. 한국환경보건학회지. 2011;37(5):342-347.

이슬비. 임신 중 복합 환경유해물질 노출이 6 개월 영유아 아토피 피부염 발생에 미치는 영향. 2019.

예신희, 이상길, 이지혜, 이경은, 성정민, 김민수. 저농도 복합유해물질 노출과 혈액검사 이상 관련성 탐색 연구. 산업안전보건연구원. 2020.

예신희, 이경은, 성정민, 박동준, 이우주. 직업병 인과추론 가이드라인 및 통계분석법 개발(1): g methods 국문 가이드라인 개발. 산업안전보건연구원. 2021.

Bobb JF. “Introduction to Bayesian kernel machine regression and the bkmr R package”, GitHub, 2017년 3월 24일 작성. 2022년 7월 10일 접속. URL https://jenfb.github.io/bkmr/overview.html#estimated_posterior_inclusion_probabilities.

Bobb JF, Valeri L, Claus Henn B, Christiani DC, Wright RO, Mazumdar M, Godleski JJ, Coull BA. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics. 2015 Jul;16(3):493-508.

Bobb JF, Claus Henn B, Valeri L, Coull BA. Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. Environ Health. 2018 Aug 20;17(1):67.

Carrico C, Gennings C, Wheeler DC, Factor-Litvak P. Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. J Agric Biol Environ Stat. 2015;20(1):100-20.

Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004 Sep;15(5):615-25.

Hernán MA. Selection bias without colliders. American Journal of Epidemiology 2017; 185 (11): 1048-1050.

Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC, 2020.

Joffe MM. Structural nested models, G-estimation, and the healthy worker effect: the promise (mostly unrealized) and the pitfalls. Epidemiology. 2012;23(2):220-222.

Keil AP, Richardson DB. Reassessing the Link between Airborne Arsenic Exposure among Anaconda Copper Smelter Workers and Multiple Causes of Death Using the Parametric g-Formula. Environ Health Perspect. 2017 Apr;125(4):608-614.

Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics. 2007

Dec;63(4):1079-88. doi: 10.1111/j.1541-0420.2007.00799.x. PMID: 18078480; PMCID: PMC2665800.

Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. Int J Epidemiol. 2017;46(2):756-762.

Neophytou AM, Costello S, Picciotto S, Brown DM, Attfield MD, Blair A, Lubin JH, Stewart PA, Vermeulen R, Silverman DT, Eisen EA. Diesel Exhaust, Respirable Dust, and Ischemic Heart Disease: An Application of the Parametric g-formula. Epidemiology. 2019 Mar;30(2):177-185.

Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods: application to control of the healthy worker survivor effect (erratum appear in Math Modelling 1987;14:917–921). Mathematical Modelling 1987; 7: 1393-1512.

Robins JM. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Sechrest L, Freeman H, Mulley A, eds. Health Services Research Methodology: A Focus on AIDS. Rockville, MD: National Center for Health Services Research, U.S. Public Health Service, 1989; 113-159.

Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11:550–560.

Taubman SL, Robins JM, Mittleman MA, Hernán MA.Intervening on

risk factors for coronary heart disease: an application of the parametric g-formula. International Journal of Epidemiology 2009; 38(6): 1599- 1611.

Valeri L, Mazumdar MM, Bobb JF, Claus Henn B, Rodrigues E, Sharif OI, Kile ML, Quamruzzaman Q, Afroz S, Golam M, Amarasiriwardena C. The joint effect of prenatal exposure to metal mixtures on neurodevelopmental outcomes at 20-40 months of age: evidence from rural Bangladesh. Environmental health perspectives. 2017 Jun 26;125(6):067015.

Abstract

Development of Korean guideline and statistical analysis method for causal inference of occupational disease (2) : Development of Korean guidelines for health effect assessment of multiple exposures

Objectives : *Bobb JF developed the bayesian kernel machine regression (BKMR) to figure out how much health outcomes are affected by exposure to environmental mixtures, The BKMR automatically reflects complex interaction terms between environmental mixtures and their higher-order nonlinear terms.*

On the other hand, it is well-known that standard regression methods often fail to take into account treatment-confounder feedbacks and healthy worker survival bias. To remove those biases, Robin JM proposed the g-formula, which is commonly used to correct the biases. This method can also be used for analyzing multiple exposure situations.

Despite the advantages of the BKMR and g-formula, occupational epidemiologists have difficulty in applying the methods to environmental mixtures-related studies since they are not familiar with the methods. Therefore, we help epidemiologists by offering the guideline, which describes their concepts and how to implement the methods with the statistical software R.

Method : *The BKMR explains the impact of environmental mixtures on the health outcomes by taking complex structures between environmental mixtures into account. Combining several parametric regression models based on causal graph with expert knowledge, g-formula handles the treatment-confounder feedback and healthy worker survival bias to estimate the impact of environmental mixtures on the health outcomes.*

Results : We write the Korean guideline about the concepts and use of the BKMR and g-formula. Further, we applied the g-formula using health examination data for workers in order to estimate a health impact of exposure to lead and cadmium (additionally, exposure to lead and xylene) on the health outcome in terms of risk ratio scale. The estimated causal effects of time-varying exposures, including blood lead and cadmium level, on the anemia are significant had all workers been exposed to lead and cadmium at least 15 μg/dL and 3 μg/L, respectively. Lastly, based on reviews for the methods, we suggest a future research plan to overcome their limitations.

Key words : *causal graph, multiple causes, g-formula, bayesian kernel machine regression*

부록

부록 1: 직업병 인과추론 가이드라인: 복합노출의 건강 영향 평가

부록 2: 직업병 인과추론 가이드라인: g-formula 국문 가이드라인 수정 내용

부록 1

직업병 인과추론 가이드라인: 복합노출의 건강영향 평가

부록 1 I

복합노출의 건강영향을 평가하는 통계방법의 배경

Ⅰ. 복합노출의 건강영향을 평가하는 통계 방법의 배경

1. 복합노출의 건강영향을 평가하는 통계방법의 필요성

근로자들은 시간의 흐름에 따라 다양한 업무와 유해요인에 노출될 수 있다. 다양한 유해요인에 복합적으로 그리고 반복적으로 노출될 수 있고, 유해요인과 건강영향 사이 교란 요인 또한 시간에 따라 변화할 수 있다. 더욱이 근로자들은 유해물질 노출에 따른 건강영향으로 인해 고용상태가 변화될 수 있다. 근로자를 대상으로 수집한 종적 자료의 특성을 고려하지 않고, 노출과 질병 간의 관련성을 분석하게 되면 다양한 편향 (bias)이 발생할 수 있다. 예를 들어, 시간에 따라 변화하는 노출과 선행 노출에 영향을 받으며 시간에 따라 변화하는 교란 요인의 특성을 고려하지 않는 경우가 대표적인 경우이다. 그림 1과 같이 '판정 결과' 변수는 '첫 번째 노출 측정'과 '두 번째 노출 측정' 매개변수임과 동시에 '두 번째 노출 측정' 변수와 '건강 결과' 변수의 교란 요인이기도 하다. '판정 결과' 변수를 보정하게 되면 첫 번째 노출 측정에서 건강 결과로의 경로 중 '판정 결과' 변수를 통해 지나가는 경로가 모두 차단되므로 건강 결과에 대한 첫 번째 노출의 인과효과는 총 효과가 아닌 직접적인 인과효과만 추정된다. 하지만 '판정 결과' 변수를 보정하지 않으면 건강 결과에 대한 두 번째 노출의 인과효과를 추정함에 있어 교란 편향 (confounding bias)이 발생한다. 이러한 특성이 있는 자료에서는 전통적인 회귀분석으로 이러한 편향을 통제할 수 없으며, g-methods와 같은 새로운 통계 분석 방법의 접근이 필요하다.

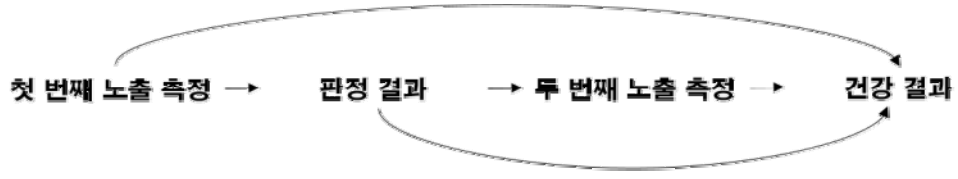


그림 1. 시간에 따라 변화하는 노출과 교란 요인, 건강 결과을 표현한 인과 그래프.

또한, g-methods의 경우, 그림 2와 같이 근로자 종적 연구에서 주로 발생하는 '건강근로자 생존 편향 (healthy worker survivor bias)'로 인한 선택 편향 (selection bias)을 제어하는 데도 활용할 수 있다.

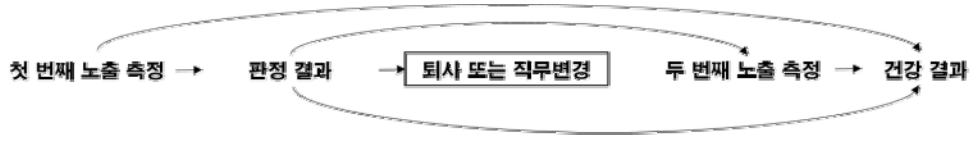


그림 2. 근로자 종적 연구에서 발생하는 '건강근로자 생존 효과'로 인한 선택 편향을 표현한 인과 그래프.

위에서 설명한 두 가지 편향 이 외에도, 아래 그림 3과 같이 복합 유해물질에 대한 노출, 즉 근로자가 2가지 이상의 유해물질에 다중 노출되었을 때, 다중 노출을 고려하지 않고, 단일 물질 1과 건강 결과 간만의 관련성을 분석할 경우, 실제로는 물질 2가 건강 결과에 영향을 미치는 원인임에도 불구하고, 통계적으로 물질 1과 건강 결과 간의 유의한 관련성이 관찰될 수 있다.

그림 3. 복합노출로 인한 건강영향을 고려하지 않을 시 발생할 수 있는 편향을 표현한 인과 그래프.

최근 이러한 특성을 고려한 인과추론 통계 방법인 g-methods와 복합노출 통계 방법 중 하나인 bayesian kernel machine regression (BKMR)의 활용이 증가하고 있는 추세다(그림 4, 5).

하지만, 국내 산업보건 영역에서는 '인과추론 통계 방법'과 '복합노출의 건강영향 평가 방법' 각각 근로자 종적 자료 분석에 적절하게 사용되고 있지 않으며, 따라서 국내 산업보건 역학연구에서 이러한 통계적 방법론들이 다양하게 활용될 수 있도록 각 방법론에 대한 구체적인 국문 가이드라인이 필요하다. 2021년 연구과제(예신희 등. 2021)에서는 단일노출로 인한 건강영향을 추론하는 국문 가이드라인을 작성하였고, 2022년 본 과제에서는 2개 이상의 유해물질에 대한 노출이 근로자의 건강에 영향을 미치는지 그 영향을 추론하는데 활용될 수 있는 통계적 방법론에 대한 국문 가이드라인을 작성하고자 한다.

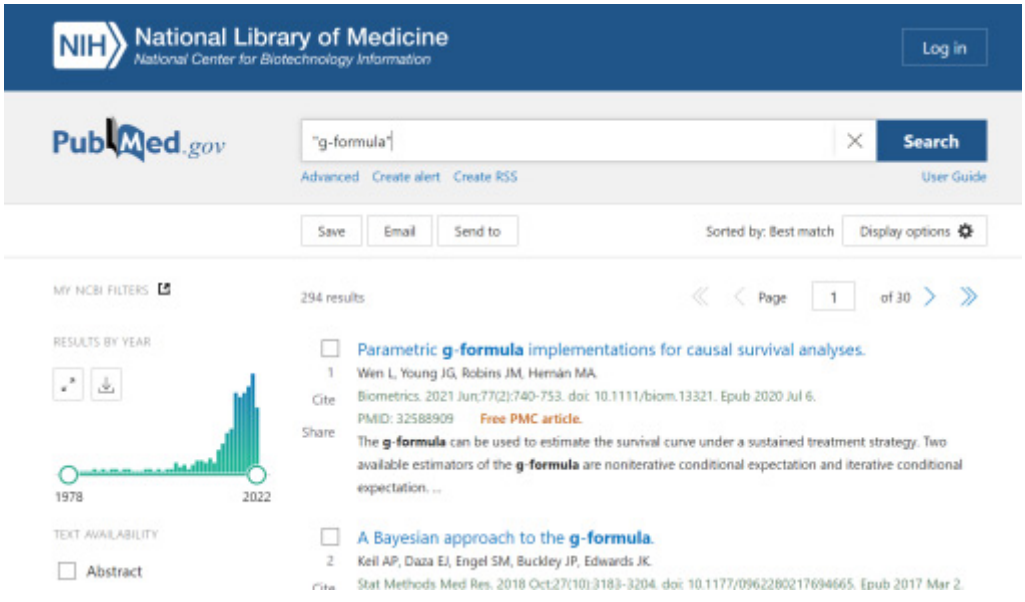


그림 4. G methods 중 하나인 'g-formula'의 Pubmed 검색 결과.

그림 5. 복합노출 건강영향 평가방법 중 하나인 'BKMR'의 Pubmed 검색 결과.

2. 근로자 자료에 g-formula와 BKMR을 적용한 복합노출의 건강영향을 평가한 산업보건 분야의 선행연구

Taubman SL 등 (2009)의 연구는 Nurses' Health Study 자료에 등록된 78,746명의 2년마다 실시하는 간호사 건강검진 기록을 사용하여 g-formula를 통해 여러 개입에 대한 관상동맥질환 (coronary heart disease)의 1982년부터 2002년까지 follow-up된 사망률을 계산하였다. 관찰연구 자료로부터 직접 구한 20년 follow-up된 누적 발생률은 3.50%였으며, 5가지 활동 (금연, 매일 적어도 30분씩 운동, 매일 5g 이상의 알코올 섭취, diet score의 상위 40% 이내로 유지, BMI 지수 25 이하로 유지)을 모두 실행하였을 때에는 누적 발생률이 1.89% (신뢰구간: 1.46에서 2.41까지)으로 낮아지는 것을 확인하였다. 이 연구에서는 한 가지 개입뿐만 아니라 2가지 이상의 개입을 동시에 시행하여 개입의 조합에 대한 효과를 측정하였다.

Keil AP 등 (2017)은 건강 근로자 생존 편향을 보정하기 위해 근로자의 고용상태 (employment status)를 인과적 방향성 비순환 그래프에 포함한 후, g-formula를 이용하여 8,014명의 백인 남성에 대하여 구리 용광로에서 공기 중 비소 흡입량에 따라 모든 질병, 심장병, 폐암으로 인한 초과 사망률이 나이에 따라 어떻게 변화하는지 연구하였다. 이때 노출량에 대한 개입 (intervention)은 노출을 시키지 않는 개입, 개입하지 않음, 심한 노출을 시키는 개입 총 3가지로 분류하여 적용하였으며, 각 개입을 시행하였을 때 발생하는 초과 사망률을 측정하였다.

Valeri L 등 (2017)은 중금속 복합물질에 대한 임신 중 노출이 출생 후 20-40개월 영유아의 신경 발달 결과에 영향을 미치는지 파악하기 위해 어머니-아이 825쌍을 대상으로 BKMR을 적용하여 연구를 진행하였다. 이때, 영유아의 신경 발달 정도를 측정하기 위해 인지 발달 점수 (cognitive development score)와 언어 개발 종합 점수 (language development

composite score)를 사용하였으며, 중금속 복합물질로는 비소, 마그네슘, 납을 고려하였다.

이슬비 등(2019)은 어머니-아이 302쌍을 대상으로 중금속에 해당하는 납, 수은, 카드뮴과 대기오염물질에 해당하는 NO2, PM10, PM2.5 그리고 비스페놀 A, 프탈레이트 대사체 MEHHP, MEOHP, MnBP 3종을 포함하여 총 10가지의 환경유해물질에 대한 복합노출이 출생 후 6개월이 지난 영유아의 아토피 피부염 발생에 미치는 영향을 확인하고자 하였다. BKMR을 사용하여, 임신 말기에서 복합물질에 대한 누적 노출 양이 증가할 때 영유아의 아토피 피부염 발생의 위험이 증가한다고 보고하였다.

Neophytou AM 등 (2019)은 광부 연구 코호트에서 디젤 배기가스와 호흡성 광산 분진의 노출 수준을 제한하는 가상의 시나리오 하에서 허혈성 심장질환 사망률의 반사실적 결과 (counterfactual outcome) 위험을 평가하였다. 대기오염에 대한 일반적인 인구집단 대상 연구에서 미세먼지 (특히, 디젤 배기가스 배출물질)가 심혈관질환의 잠재적 위험 요소임을 시사하지만, 정량적 노출 측정을 사용한 직업 코호트에서의 직접적인 근거는 제한적이다. 이 연구는 디젤 장비가 각 시설에 도입된 후 8개의 비금속, 비석탄 광산에 고용된 10,778명의 남성 광부 데이터를 분석하였고, 1948년부터 1997년까지 추적하였으며, 이 중 297명이 허혈성 심장질환으로 인한 사망하였다. 연구자들은 호흡성 원소 탄소 (디젤 배기 물질에 대한 대체물)와 호흡성 분진에 대해 개별적으로 그리고 공동으로 다양한 제한 기준을 가진 가상 시나리오 하에서 위험을 평가하기 위해 g-formula를 적용하였다. 원소 탄소와 호흡성 분진에 대한 노출이 제거되는 가상 시나리오 하에서, 80세에서 관찰된 위험과 누적 허혈성 심장질환 위험을 비교한 risk ratio는 0.79였다. Risk difference는 -3.0%였다. 비금속 광부 코호트 자료를 기반으로 하는 이 연구 결과는 디젤 배기 물질 및 호흡성 분진에 대한 노출을 제거하기 위한 개입이 허혈성 심장질환 사망 위험을 감소시킬 것이라는 가설과 일치하였다.

예신희 등 (2020)은 특수건강진단 자료와 한국 국민건강영양조사 자료, 미

국 국민건강영양조사 자료를 BKMR 방법으로 분석하여, 납과 카드뮴의 복합노출이 간기능검사 (AST, ALT, GGT) 결과 수치에 미치는 영향을 평가하였다. 그 결과 모든 모형과 세 가지 종류의 자료 (특수건강진단, KNHANES (한국 국민건강영양조사), NHANES (미국 국민건강영양조사))에서 납과 카드뮴 복합노출에 의한 건강영향 분석결과가 일관되게 나온 것은 GGT (감마글루타밀전이효소) 검사 결과였다. 혈중 납과 혈중 카드뮴은 각각 GGT와 연관성을 보였으며, 복합 노출도 GGT 증가와 연관성이 있었다. 또한, 납과 카드뮴은 물질의 노출 수준이 높아질수록 나머지 물질이 GGT를 더 크게 증가시키는 교호작용이 모든 분석에서 일관되게 확인되었고, 혈중 납과 혈중 카드뮴 농도 값이 함께 높아질수록 GGT가 높아지는 복합노출의 건강영향도 일관되게 관찰되었다.

부록 1 II

BKMR 이론과 적용

Ⅱ. BKMR 이론과 적용

BKMR 이론 소개에 대한 모든 내용은 Bobb JF 등 (2015), Bobb JF 등 (2018)의 논문과 Bobb JF이 2017년에 GitHub에 작성한 'bkmr' R 패키지 소개를 바탕으로 하여 작성되었고, 일부 내용은 예신희 등 (2020)이 작성한 산업안전보건연구원의 보고서를 참고하였다.

1. BKMR의 개발 배경

근로자를 포함한 대다수의 인구집단은 많은 경우 여러 유해요인에 동시에 노출되기 때문에, 최근 몇 년간 통계적인 접근을 통해 복합유해물질 노출에 의한 건강영향을 정량적으로 추정하려는 연구가 주목받고 있다. 이러한 복합유해물질의 노출로 인한 건강영향을 추정할 때에는 몇 가지 고려해야 할 점들이 있다. 첫 번째로, 복합유해물질 노출과 건강영향이 복잡한 비선형 또는 비가법적 관계 (non-additive relationship)를 가질 수 있다는 것이다. 두 번째로, 결과 변수와 여러 유해물질의 노출 사이의 교호작용이 허용되어야 하는데, 이러한 경우, 모형에서 추정해야 하는 모수 (parameters)의 수가 관측치 (observations)의 수보다 더 많아지게 되어 즉, 고차원적 문제 (high dimensional problem)가 발생하여 추정치 산출에 있어 불안정해질 수 있다. 세 번째로는, 사용된 통계 방법이 높은 상관성을 가지고 있는 노출들 (multiple highly correlated exposures)로 구성된 혼합물의 복잡한 구조를 설명할 수 있어야 한다.

혼합물 연구에 대한 기존의 접근 방식은 위와 같은 복합유해물질에 대한 건강영향을 추정할 때 고려해야 할 점들 중 일부를 해결할 수 있지만, 뚜렷한 단점이 존재한다. 예를 들어, 군집 방법 (clustering method)은 연속변수 형

태인 노출농도를 범주화하는 과정에서 정보의 누락을 야기할 수 있다. 통계적 학습 알고리즘 (random forest 등)은 혼합물질의 변수 선택에 활용될 수는 있지만 노출과 반응의 연관성 정도와 방향은 설명하기가 어렵다. 회귀모형 내 변수선택방법 (예를 들어 LASSO 등)은 여러 유해물질 사이의 높은 상관성을 모형에 반영하여 유해물질 중 일부를 선택하지만 일반적으로 결과 변수와 혼합물 사이의 관계를 상대적으로 단순한 모형 (relatively simple parametric model of mixture components)으로 설계한다는 한계점이 있다. 계층적 모형 설계 (hierarchical model formulation)은 개별 효과 추정치 (individual effect estimates)를 그룹 평균(group means)으로 축소 (shrinking)하여 노출 변수 사이의 높은 상관관계를 설명하지만, 이러한 방법은 일반적으로 각 노출과 건강 결과 간의 선형적 (linear), 가법적 (additive) 연관성을 가정해야 한다. Bobb JF 등 (2015)은 앞서 언급한 3가지 건강영향을 평가할 때 고려해야 할 점을 반영하기 위해 복합물질의 건강영향을 추정하기 위한 새로운 방법으로 Bayesian kernel machine regression (BKMR)을 제안하였다. 건강 결과는 여러 유해물질 중 일부에만 의존할 수 있기 때문에, BKMR은 혼합물질 중 건강영향에 연관성이 있는 일부 성분을 파악하기 위해 변수 선택을 수행한다. 또한, 혼합물 성분 사이의 높은 상관성을 고려하기 위해 혼합물의 구조에 대한 사전 지식 (prior knowledge)을 통합하여 모형에 반영할 수 있도록 계층적 변수 선택법 (hierarchical variable selection)을 BKMR에 도입하였다.

Bobb JF 등 (2015)은 BKMR의 개발을 통해 다음 두 가지 부분에 기여하였다. 첫째, Bobb JF 등 (2015)은 처음으로 kernel machine regression(KMR)을 복합노출에 의한 건강영향을 평가하는데 적용하였다. KMR을 사용한 이전 연구들은 통계적 검정 (testing)과 변수 선택 (variable selection), 위험 예측 (risk prediction)에 초점을 맞추었는데, Bobb JF 등 (2015)이 개발한 BKMR은 노출-반응 함수를 추정하는 것이 주요 목표이다. 둘째, 혼합물의 구조를 설명하고, 높은 상관성을 가진 노출들을 체계적으로

처리할 수 있는, BKMR 내 계층적 변수 선택법을 개발하였다.

다음 장인 BKMR의 이론과 적용에서는 KMR과 BKMR에 대한 개요와 R 프로그램에서의 BKMR의 적용에 대해서 설명하고자 한다.

2. BKMR 이론

i번째 근로자 (i = 1,2, ... n)에 대하여 Y_i을 결과 변수, 노출 벡터 $z_i = (z_{i1},....z_{iM})^T$을 M개의 유해물질들로 이루어진 벡터(납, 카드뮴, 미세먼지 등), x_i을 여러 교란 요인들로 이루어진 벡터, ε_i은 오차항이라 할 때, 결과 변수 Y_i을 다음과 같은 수식을 통해 표현할 수 있다.

$$Y_i = h(z_i) + x_i^T\beta + \epsilon_i$$

여기서 함수 h(·)는 일반적으로 혼합물 성분과 결과 변수 사이의 비선형성 (non-linear) 또는 혼합물 성분 간의 교호작용 (interaction)을 포함하는 고차원적 노출-반응 함수 (exposure-response function)를 나타내고, 오차항 ε_i는 평균이 0이며, 분산이 σ^2인 정규분포를 따른다. 하지만 고차원 노출-반응 함수 h(·)를 구체적으로 표현하는 것이 쉽지 않기 때문에 본 보고서에서는 커널 함수 (kernel function)를 기반으로 하는 kernel machine regression (KMR)을 사용하여 함수 h(·)를 표현하고자 하였다.

1) KMR의 개요

KMR을 설명하기 앞서 KMR에 사용된 커널을 이해하기 위한 몇 가지 개념에 대하여 간략하게 소개하고자 한다. 먼저 노출 벡터 $z_i = (z_{i1}, z_{i2}, ..., z_{iM})^T$은 근로자 i에게 노출된 유해물질의 양 (exposure profile)을 나타내며, 같은 방

식으로 노출 벡터 z_j은 근로자 j에게 노출된 유해물질의 양을 의미한다고 하자. 이때, 길이가 M인 노출 벡터를 길이가 M보다 큰 벡터로 변환이 가능한데, 예를 들어, 유해물질의 개수가 2인 노출 벡터를 생각해보자. 이러한 경우, 근로자 1에게 노출된 유해물질의 양을 나타내는 노출 벡터 $z_1 = (z_{11}, z_{12})^T$라 하면, 이 노출 벡터 z_1을 길이가 3인 새로운 노출 벡터 $\phi(z_1) = (z_{11}^2, z_{12}^2, \sqrt{2}\, z_{11} z_{12})^T$로 변환하여 표현이 가능하다. 이렇게 변환된 새로운 노출 벡터를 공변량으로 생각하여 $h(z_1) = \sum_{j=1}^{3} w_j \phi_j(z_1)$ 으로 결과 변수 Y_1을 모형화할 때, 새로운 노출 벡터를 사용할 수 있다. 하지만 커널 함수를 사용하여 $\sum_{j=1}^{3} w_j \phi_j(z_1) = \sum_{i=1}^{n} K(z_i, z_1)\alpha_i$로 표현할 수 있으며 (이때 n은 표본의 수임), 커널 함수 $K(\cdot, \cdot)$로는 linear 커널, polynomial 커널, Gaussian 커널 등이 사용된다. 위의 예에서 근로자 i의 노출 벡터 z_i을 새로운 노출 벡터 $\phi(z_i)$로 변환하는 함수 $\phi(\cdot)$에 대응되는 커널 함수가 바로 polynomial 커널 함수이다. 이 커널 함수를 이용하여 n x n 커널 행렬을 구성할 수 있으며, (i,j)-원소는 $K(z_i, z_j) = (z_{i1} z_{j1} + z_{i2} z_{j2})^2$으로 산출이 가능하다.

KMR의 주요한 아이디어는 결과 변수와 여러 유해물질 변수들 사이의 연관성을 유연하게 모형화하는 것이다. 위에서 언급한 것과 같이 함수 $\phi(\cdot)$을 이용하여 새로운 노출 벡터를 만들 수 있으며, 함수에 따라 이 새로운 노출 벡터는 이차항, 교호작용 항을 포함한다. 이러한 특징을 이용하여 결과 변수와 여러 유해물질 변수들 사이의 관계를 유연하게 모형화하는 것이 가능하다. Liu 등 (2007)은 KMR에서 $h(z_i)$ 항을 선형 혼합 모형(linear mixed model)으로 표현할 수 있다는 것을 보였으며, $h = (h(z_1), h(z_2), ..., h(z_n))$은 평균이 0이고 분산이 τK 인 다변량 정규분포를 따른다. 이때, 커널 행렬의 (i,j)-원소 $K(z_i, z_j)$은 두 근로자 i, j 사이의 복합물질에 대한 노출 유사성을 의미한다. 커널 행렬을 구성하기 위해 주로 사용되는 커널 함수는 Gaussian 커

널이며 (radial basis function (RBF)으로도 불림), 수식을 통해 $K(z_i, z_j)$을 다음과 같이 표현할 수 있다.

$$K(z_i, z_j) = \exp\left(-\sum_{m=1}^{M}\frac{1}{\rho}(z_{im} - z_{jm})^2\right)$$

여기서 ρ은 Gaussian 커널 함수의 튜닝 모수 (tuning parameter)이다. Bobb JF 등(2015)은 많은 커널 함수 중 다양한 고차원 항과 교호작용 항을 포함할 수 있는 함수 $\phi(\cdot)$에 대응되는 Gaussian 커널 함수에 초점을 맞춰 살펴보았다. Gaussian 커널 함수를 사용함으로써 KMR을 Gaussian process regression이라고도 볼 수 있으며 이 방법은 머신러닝 (machine learning)에서 자주 등장하는 방법이다.

2) BKMR의 개요

BKMR은 KMR 방법에 베이지안 변수 선택 접근법을 적용한 방법이다. BKMR에서 제공하는 변수 선택법은 구성 요소별 변수 선택법 (component-wise variable selection)과 계층적 변수 선택법 (hierarchical variable selection) 두 가지가 있다.

구성 요소별 변수 선택법을 설명하기 전에 위에서 설명한 커널 행렬을 다음과 같이 다시 표현할 수 있다.

$$K(z_i, z_j; r) = \exp\left(-\sum_{m=1}^{M} r_m (z_{im} - z_{jm})^2\right)$$

이때 $r = (r_1, \ldots., r_M)^T$ 이고, r_m은 m번째 유해물질이 중요한 정도를 의미하며, 0과 1 사이의 값을 가진다. 위에서 언급한 커널 행렬과 비교하면 $r_m = 1$

($m = 1, 2, \cdots, M$)으로 생각할 수 있다. 두 가지 변수 선택 방법 중 구성 요소별 변수 선택법은 유해물질 사이의 상관성을 고려하지 않고, 유해물질 각각을 독립된 하나의 물질로 여기는 변수 선택법이다.

또한, 구성 요소별 변수 선택법은 베이지안 방법에서 다중 회귀 분석에서 변수를 선택할 때 사용하는 사전 분포 (prior distribution)인 "slab-and-spike" 사전 분포를 사용하여 수식으로 표현할 수 있다. "slab"이란, 회귀 계수 값의 사전 분포 (prior distribution)를 의미하고 "spike"는 회귀 계수의 값이 0이 될 확률이라고 볼 수 있다. "slab-and-spike" 사전 분포는 결과 변수와 여러 유해물질 사이의 사전 지식 (prior knowledge)을 활용할 수 있다는 점에서 장점이 있다.

$$
\begin{aligned}
r_m|\delta_m &\sim \delta_m f_1(r_m) + (1-\delta_m)P_0, \quad m = 1, \ldots, M, \\
\delta_m &\sim Bernoulli(\pi)
\end{aligned}
$$

위 수식에서 $f_1(\cdot)$는 r_m의 확률밀도함수이며, P_0는 r_m이 0의 값을 가질 확률밀도를 나타낸다. δ_m는 결과 변수에 m번째 유해물질이 영향을 미치는지 그 여부를 나타내는 지표로, 1이면 결과 변수에 m번째 유해물질이 영향을 주는 것으로, 0이면 영향을 주지 않는 것으로 해석할 수 있다. δ_m은 0과 1의 값을 갖는 확률 변수로 생각할 수 있으므로, 베르누이 분포를 따른다고 할 수 있으며, 이때 $\pi = P(\delta_m = 1)$는 m번째 유해물질이 결과 변수에 영향을 주는 요인으로 뽑힐 확률을 의미한다. 지표 δ_m의 사후 평균은 m번째 유해물질이 혼합물에서 상대적으로 중요한 요소인지 나타내는 사후 포함 확률 (posterior inclusion probability; PIP)로 해석할 수 있다.

혼합물의 구성성분 사이의 상관성이 높을 경우, 위에서 언급한 구성 요소별 변수 선택법은 상관성 있는 구성성분들을 구분하는 데 어려움이 생기기 때문에 실질적으로 적용하기 힘들 수 있다. 따라서 구성성분 간의 상관성을, 즉 복합물질의 구조에 대한 정보를 모형에 반영할 수 있는 계층적 변수선택법을

추가로 고려할 수 있다. 계층적 변수 선택법은 상관성이 높은 구성성분들을 하나의 그룹으로 묶음으로써 혼합물을 구성하는 성분들을 여러 개의 군으로 분할한 후, 각 군 별로 중요한 유해물질을 선택하는 변수 선택법이다. 혼합물 $z_1, ..., z_M$을 총 G개의 군 $(S_1, S_2, ..., S_g, ..., S_G)$으로 분할하였다고 가정해보자. 또한, 군 간 상관성은 낮으면서 그룹 내 유해물질 간의 상관성은 높도록 분할하였다고 가정해보자. 따라서 구성 요소별 변수 선택법에서 결과 변수에 영향을 주는 물질인지 파악하는 지표인 δ_m에 대응되는 지표로 g번째 군이 결과 변수에 영향을 주는지 결정하는 지표 ω_g 와 g번째 군이 선택되었을 때, g번째 군 S_g에서 중요한 유해물질을 선택하는 지표인 δ_{S_g}가 구성 요소별 변수 선택법에서의 δ_m을 대신하여 변수 선택과정에 추가된다.

$$\delta_{S_g} \mid \omega_g \sim Multinomial(\omega_g, \pi_{S_g}), \quad g = 1, \cdots G,$$
$$\omega_g \sim Bernoulli(\pi)$$

이때, π는 g번째 군이 결과 변수에 영향을 주는 군으로 선택될 확률을 의미하며, π_{S_g}는 g번째 군 S_g 안에 있는 구성 성분들이 결과 변수에 영향을 미치는지 결정하는 확률을 나타내는 벡터이다. 예를 들어, g번째 군이 선택되었고 $(\omega_g = 1)$, 이 군이 3개의 유해물질 A, 유해물질 B 그리고 유해물질 C로 구성되어 있다고 하자. 이러한 경우, 3개의 유해물질 중 중요한 하나의 유해물질을 선택해야하므로 다항분포 (multinomial distribution) $\delta_{S_g} \mid \omega_g = 1 \sim Multinomial(\omega_g = 1, \pi_{S_g})$을 사용하여 중요한 유해물질 하나를 선택할 수 있다. 이때 유해물질 A, B, C가 선택될 확률을 p_A, p_B 그리고 p_C $(p_A + p_B + p_C = 1)$라 하면 g번째 군 S_g에서 유해물질이 선택될 확률 벡터 π_{S_g}을 (p_A, p_B, p_C)와 같이 표현할 수 있다. 비록 계층적 변수선택법을 사용할 때, 동일한 그룹의 두 구성 요소가 결과 변수에 대하여 독립적인 또는 상호작용 효과를 가지지 않는다고 가정해야 하지만, 군 내 높은 상관관계가 있는 경우에서는 이러한 효과를 보다 일반적인 모델에서 식별하기는 힘들다.

3. BKMR의 적용

1) 패키지의 설치 및 모의연습 자료의 생성(Installation of package and generation of a simulated dataset)

BKMR을 적합하기 위해서는 통계 프로그램 R (Version 4.1.2)의 'bkmr' 패키지 (Version 0.2.2)의 설치가 되어있어야 하며, 패키지에는 BKMR의 수행에 필요한 다양한 함수들이 포함되어 있다. 특히 kmbayes() 함수는 모형적합(fitting)에 있어 가장 중요한 함수이다. 그 외 제공되는 함수들을 통해 다양한 방식으로 모형의 결과를 요약하고 시각적으로 나타낼 수 있는 기능을 수행할 수 있다. 다음은 R 프로그램의 'bkmr' 패키지를 설치하고 불러오는 코드이다.

```
# bkmr 패키지의 설치
install.packages("bkmr")

# bkmr 패키지 불러오기
library(bkmr)
```

임의의 데이터셋을 생성하여 R 패키지 'bkmr'에서 제공하는 함수들에 대하여 설명하고자 한다. SimData(n, M)은 모의실험 자료를 생성하는 함수이며, 이때 n은 자료의 수, M은 유해물질의 수를 의미한다. 결과의 재현성을 위해 자료 생성 시 111 seed 번호를 사용하였다. 'y'는 결과 변수의 값을, 'Z'는 노출 변수의 행렬, 'X'는 공변량 (covariate)을 포함하는 행렬을 나타낸다.

```
# 재현성의 확보를 위한 코드
set.seed(111)

# 연습 자료의 생성
dat <- SimData(n = 50, M = 4)
y <- dat$y
Z <- dat$Z
X <- dat$X
```

모의 실험자료 생성에 사용된 참(true) 노출-반응 함수를 살펴보고자 한다. 유해물질의 수가 4개이고, 결과 변수와 유해물질의 사이의 그림을 그리기 위해서는 5차원이 필요하기 때문에 본 장에서는 첫 번째 유해물질과 두 번째 유해물질에 대해 노출된 양에 따라 결과 변수가 어떻게 변하는지 3차원 그림으로 표현하였으며, 그 결과는 아래의 그래프 res와 같다. 아래의 그림을 통해 첫 번째 유해물질과 두 번째 유해물질이 증가하면 결과 변수가 증가하는 것을 확인할 수 있다.

```
# 연습 자료의 시각화
z1 <- seq(min(dat$Z[, 1]), max(dat$Z[, 1]), length = 20)
z2 <- seq(min(dat$Z[, 2]), max(dat$Z[, 2]), length = 20)
hgrid.true <- outer(z1, z2, function(x,y) apply(cbind(x,y), 1, dat$HFun))
res <- persp(z1, z2, hgrid.true, theta = 30, phi = 20, expand = 0.5, col =
"lightblue", xlab = "", ylab = "", zlab = "")
```

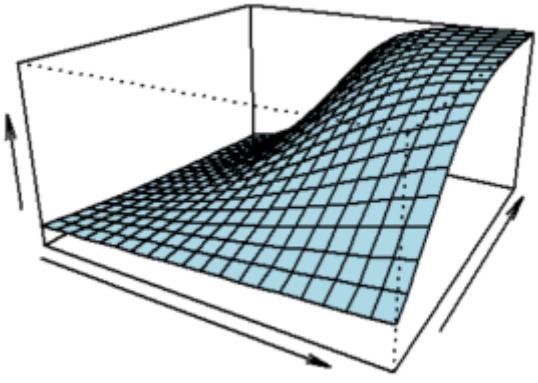


그림 6. 모의연습 자료를 시각적으로 표현한 그림

2) 모형의 적합 및 진단(Model fitting and diagnosis)

그 다음으로, BKMR 모델을 적합하기 위하여 앞서 언급한 kmbayes() 함수를 사용한다. BKMR은 모수를 추정할 때, 마코프 체인 몬테 카를로(Markov chain Monte Carlo; MCMC) 알고리즘을 사용하며, 알고리즘을 얼마나 반복시킬지 그 반복횟수가 필요하다. MCMC 알고리즘을 실행시킬 때 필요한 반복횟수는 'iter' 인수(argument)을 통해 결정할 수 있다. kmbayes() 함수를 사용할 때, 먼저 주의해야할 점은 사용되는 자료 y, Z, X에 결측치가 존재하면 안 된다는 것이다. 자료에 결측치가 있는 경우, 함수가 결측치가 존재하는 변수에 대하여 에러 메시지를 띄우며, 실행되지 않는다.

```
# BKMR 모형의 적합
set.seed(111)
fitkm <- kmbayes(y = y, Z = Z, X = X, iter = 10000, verbose = FALSE, varsel
= TRUE)
```

MCMC를 통해 모수 값들이 안정적으로 수렴하는지 trace plot을 통해 시각적으로 확인하고자 한다. 어떤 모수에 대하여 trace plot을 그릴지 par 인수를 통해 지정할 수 있다.

```
# 모수 β에 대한 MCMC 결과 확인
TracePlot(fit = fitkm, par = "beta")
```

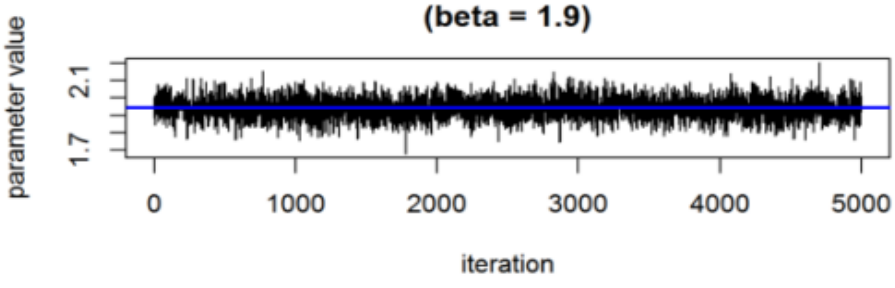


그림 7. 모수 β의 수렴성을 확인하기 위한 그래프

```
# 모수 σ²에 대한 MCMC 결과 확인
TracePlot(fit = fitkm, par = "sigsq.eps")
```

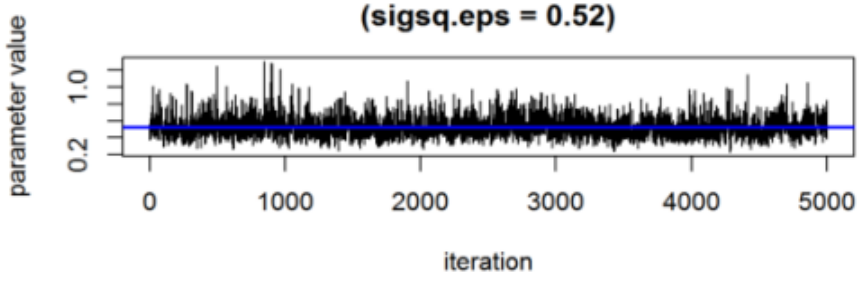


그림 8. 모수 σ^2의 수렴성을 확인하기 위한 그래프

comp 인수를 통해 각 유해물질이 건강 결과에 영향을 주는지 그 여부에 대한 확률의 수렴성을 시각적으로 확인할 수 있으며, 아래의 코드는 유해물질 z1의 수렴성을 확인하기 위한 plot을 그리는 코드이다.

```
# 유해물질 z1이 건강 결과에 영향을 주는지 그 여부에 대한 확률에 대한 MCMC 결과 확인
TracePlot(fit = fitkm, par = "r", comp = 1)
```

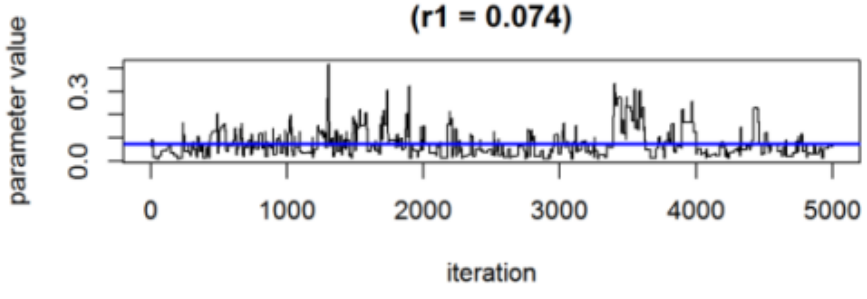


그림 9. 유해물질 z1이 결과 변수에 영향을 주는지 그 여부를 결정하는 모수 r_1의 수렴성을 확인하기 위한 그래프

3) 모형 적합 결과의 해석(Interpretation of BKMR)

(1) 연속형 변수에서의 BKMR의 결과 해석(Interpretation of BKMR for a continuous outcome)

각 유해물질마다 추정된 사후 포함확률 (PIP)은 다음과 같이 확인할 수 있다. 추정된 사후 포함확률을 크기순으로 열거하면 z1, z2, z4, z3임을 확인할 수 있으며, z1과 z2가 결과 변수 y에 중요하게 기여하고 있으며, z3의 기여가 가장 낮음을 확인할 수 있다.

```
# 각 유해물질의 기여도 산출 및 출력
ExtractPIPs(fitkm)
##    variable      PIP
## 1         z1 1.0000
## 2         z2 1.0000
## 3         z3 0.1122
## 4         z4 0.3034
```

또한, 사후 포함확률은 다음의 코드를 통해 시각적으로 표현할 수 있다(그림 10).

```
# 각 유해물질에 대한 기여도를 시각화하기 위한 패키지의 설치
install.packages("ggplot2")

# 패키지 불러오기
library(ggplot2)

# 각 유해물질의 기여도 산출
pips <- ExtractPIPs(fitkm)

# 각 유해물질에 대하여 번호 지정
pips$exposure <- as.factor(1:nrow(pips))

# 각 유해물질의 기여도의 시각화
ggplot(pips, aes(exposure, PIP)) +
  geom_point() +
  ylab("PIP") +
  ylim(0, 1)+theme_bw()
```

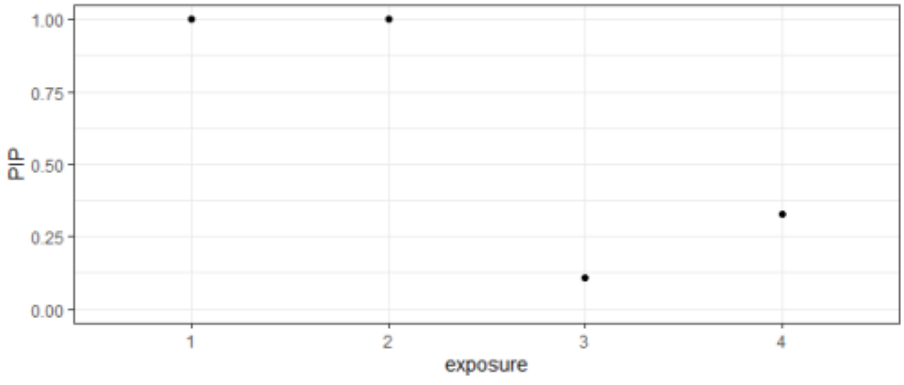


그림 10. 각 유해물질별 사후확률을 나타낸 그래프

지금까지 BKMR을 적합하고, 안정적으로 모수들이 추정되었는지 확인할 때 필요한 함수에 대하여 설명하였다. 하지만 지금부터는 적합된 BKMR의 분석결과를 요약할 때, 사용하는 'bkmr' 패키지에 포함된 다양한 함수들을 살펴보고자 한다. 이 함수들은 유해물질과 결과 변수 사이의 연관성을 시각화하고, 그 외에 복합물질 노출에 대한 건강영향을 평가할 수 있는 다양한 통계 분석결과를 시각적으로 표현한다.

연구자들은 BKMR 분석결과를 통해 고차원적 노출-반응 함수 h(·) 항을 시각화할 수 있다. 하지만 3차원 이상의 고차원을 2차원 그래프를 통해 표현하기 어렵기 때문에 대안으로 관심 있는 유해물질을 하나 또는 두 개 결정하여 유해물질의 노출 수준과 결과 사이의 관계를 시각화한다. 이때 관심 있는 유해물질을 제외한 나머지 유해물질에 대한 노출 수준을 특정 값으로 고정한 상태로 관심 있는 변수들과 결과 사이의 관계를 시각적으로 표현한다. 함수를 시각적으로 표현할 때, 관심 있는 유해물질이 하나인 경우 PredictorResponseUnivar 함수를 사용하고, 둘인 경우에는 PredictorResponseBivarLevels 함수를 사용한다. 다음의 코드는 z1, z2, z3, z4 각각 하나의 변수에 대하여 노출-반응 함수를 시각적으로 표현하기 위해 사용한 코드이다. 관심 있는 유해물질 (예; z1)를 제외한 나머지 유해물

질(예; z2, z3, z4)의 값은 특정 percentile로 고정시킨 후 관심 있는 유해물질의 값을 변화해 나가며, 이 유해물질과 결과 사이의 노출-반응 관계를 그래프를 통하여 보여준다. 다른 유해물질의 percentile을 지정하는 인수는 q.fixed이며, 기본 값은 50 percentile로 설정되어 있다. 만약 70 percentile로 설정을 원하는 경우에는 'q.fixed = 0.7'을 입력하면 된다.

```
# 하나의 유해물질 외 나머지 유해물질이 각 유해물질의 중앙값으로 고정되어있을 때,
하나의 유해물질의 노출 수준에 따른 h(·) 항 값 계산
pred.resp.univar <- PredictorResponseUnivar(fit = fitkm)

# 그 하나의 유해물질의 노출 수준에 따른 h(·) 항 값의 그래프
ggplot(pred.resp.univar, aes(z, est, ymin = est - 1.96*se,
  ymax = est + 1.96*se)) +
  geom_smooth(stat = "identity") +
  facet_wrap(~ variable,scales="free") +
  ylab("h(z)")
```

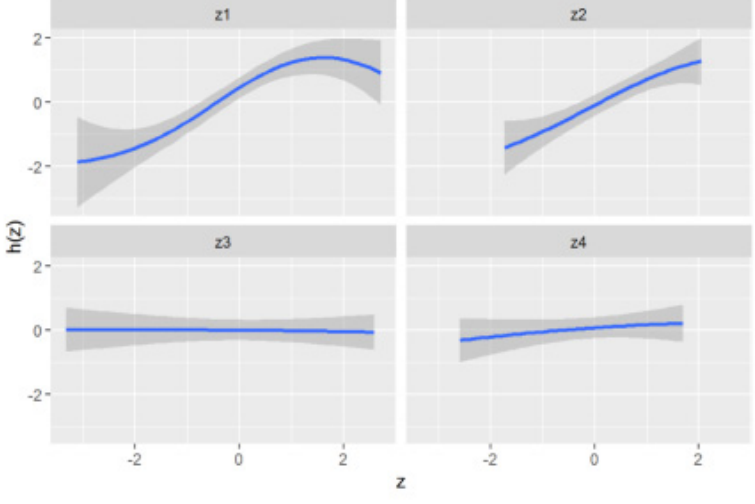


그림 11. 유해물질 z1, z2, z3 그리고 z4 각각에 대해 노출 수준에 따른 h(·)의 추정치에 대한 그래프

그림 11을 통해 유해물질 z1, z2, z3 그리고 z4에 대하여 각 유해물질이 미치는 건강 영향을 확인할 수 있다. 유해물질 z1, z2는 노출 수준이 증가할수록 결과 변수의 값을 높이는 방향으로 작용한다. 또한, 그림을 보면 유해물질 z1의 경우, 높은 노출 수준에서는 오히려 결과 변수의 값을 낮추었다. 유해물질 z2의 경우, 유해물질 z1과 달리 결과 변수와의 관계가 선형적인 것을 확인할 수 있었던 반면, 유해물질 z3, z4는 결과 변수에 유의한 결과를 주지 못 하는 것으로 보인다. 지금까지 위의 예에서 관심 있는 유해물질이 한 개인 경우를 설명하였다. 관심 있는 유해물질이 2개인 경우, 그 유해물질들을 제외한 나머지 모든 유해물질이 특정 percentile로 고정되어있을 때, 두 유해물질에 대한 이변량 고차원적 노출-반응 함수 h(·) 항을 시각화하고자 한다. 이때 시각화하는 방식은 두 번째 유해물질을 여러 percentile로 고정한 후, 첫 번째 유해물질과 결과 간의 노출-반응 함수를 시각화한다. 관심 있는 유해물질이 1개인 경우와 마찬가지로 두 유해물질을 제외한 나머지 유해물질은 특정 percentile 값으로 고정한다. 이는 PredictorResponseBivarLevels 함수를 사용하여 수행할 수 있다. 여기서 qs 인수에 여러 percentile 값을 지정함으로써 두 번째 유해물질에 대해 고정할 percentile을 지정할 수 있다.

```
# 두 개의 유해물질 외 나머지 유해물질의 노출 수준을 각 유해물질의 중앙값으로
고정되어 있을 때, 두 유해물질의 노출 수준에 따른 h(·) 항 값 계산
pred.resp.bivar <- PredictorResponseBivar(fit = fitkm,
  min.plot.dist = 1)

# 두 유해물질에 대하여 한 유해물질의 노출 수준을 qs로 설정한 후, 남은 유해물질의
노출 수준에 따른 값을 추출
pred.resp.bivar.levels <- PredictorResponseBivarLevels(
  pred.resp.df = pred.resp.bivar, Z = Z, qs = c(0.1, 0.5, 0.9))

# 두 유해물질의 노출 수준에 따른 h(·)의 추정치의 그래프
ggplot(pred.resp.bivar.levels, aes(z1, est)) +
  geom_smooth(aes(col = quantile), stat = "identity") +
  facet_grid(variable2 ~ variable1) +
  ggtitle("h(expos1 | quantiles of expos2)") +
  xlab("expos1")
```

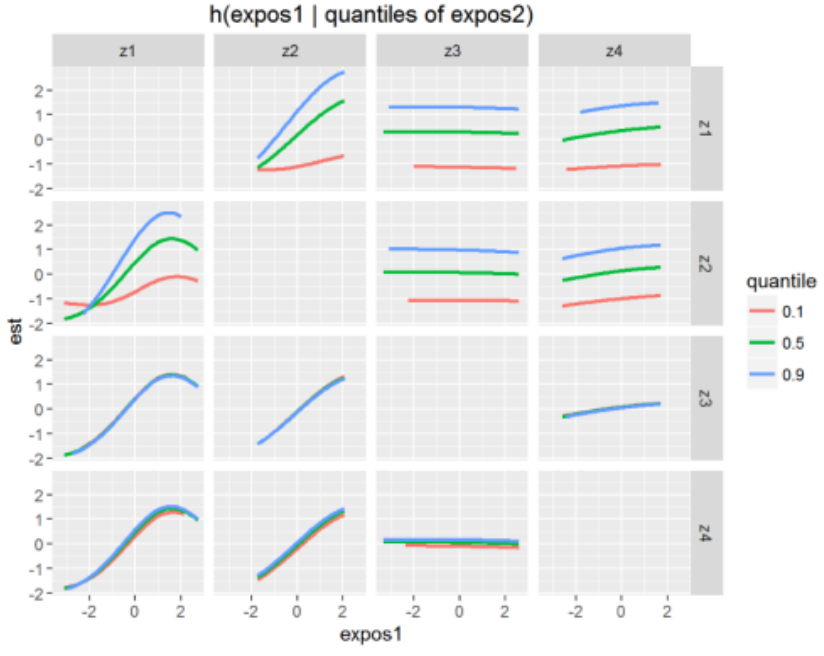


그림 12. 두 유해물질의 노출 수준에 따른 h(·)의 추정치에 대한 그래프

그림 12는 그림 11에서 설명하였던 것과 같이 하나의 유해물질과 결과 변수 사이의 관계를 설명하는 것과 같은 방식으로 설명이 가능하다. 현재 그림 12는 bkmr을 적합할 때, 4개의 유해물질을 사용하였기 때문에 4 x 4 행렬의 각 원소를 그림으로 채운 것과 같이 나타난다. 그림 12에서 대각선은 자기 자신이 특정 노출 수준으로 고정되어있을 때, 자기 자신에 대한 노출 수준이 되면 건강 결과가 어떻게 변화하는지 나타낼 수 없기 때문에 빈 그림으로 채워져있다. 그림 12에서 1행의 2열을 보면 유해물질 z2의 노출 수준이 PredictorResponseBivarLevels 함수에 입력된 qs의 값 각각으로 고정되어 있을 때, 유해물질 z1의 노출 수준에 따른 h(·) 항의 값을 보여준다. 자세히 살펴보면, 유해물질 z2가 10, 50, 90 percentile로 고정되어있을 때, 유해물질 z1이 결과 변수에 미치는 영향에 대한 그래프가 평행하지 않는 것으로 보아 유해물질 z1, z2 사이의 교호작용을 의심해볼 수 있다.

지금까지 관심 있는 유해물질이 1개 또는 2개인 경우에, 노출 수준에 따라 결과 변수가 어떻게 변하는지 그 형태를 시각화하여 확인하였다. 하지만 bkmr 패키지에서는 각 유해물질 별 건강 영향에 미치는 효과의 형태뿐만 아니라 모든 유해물질에 대한 노출량이 변화하였을 때, 건강영향에 미치는 효과의 값을 요약할 수 있으며, 그 형태 또한 확인할 수 있다. 모든 유해물질의 노출량이 특정 percentile에 있을 때의 결과 값을 모든 유해물질에 대한 노출량이 중앙값일 때와 비교하여 유해물질 노출량 전체의 효과를 계산할 수 있다. 전체 효과는 OverallRiskSummaries 함수와 qs 인수를 사용하여 비교하고 싶은 percentile을 지정하고 q.fixed 인수를 사용하여 비교 기준인 percentile (기본 값은 50 percentile)을 지정하여 요약할 수 있다.

```
# 모든 유해물질이 특정 노출 수준 (qs)일 때의 결과 값 산출 및 출력
risks.overall <- OverallRiskSummaries(fit = fitkm, y = y, Z = Z,   X = X, qs
= seq(0.25, 0.75, by = 0.05), q.fixed = 0.5,
 method = "exact")

risks.overall
##      quantile        est        sd
## 1        0.25 -1.24022859 0.20999691
## 2        0.30 -1.07409824 0.17907496
## 3        0.35 -0.57098339 0.10056121
## 4        0.40 -0.37647134 0.06862383
## 5        0.45 -0.06985162 0.02041499
## 6        0.50  0.00000000 0.00000000
## 7        0.55  0.28945731 0.05089159
## 8        0.60  0.51300959 0.09587608
## 9        0.65  0.67740749 0.13012125
## 10       0.70  0.88153349 0.16111132
## 11       0.75  1.10039068 0.19844371
```

또한, 유해물질 노출량 전체에 대한 효과를 다음의 코드를 사용하여 시각적으로도 표현할 수 있다(그림 13).

```
# 모든 유해물질이 qs에 해당하는 노출 수준으로 고정되었을 때의 결과 값에 대한 그래프
ggplot(risks.overall, aes(quantile, est,
ymin = est - 1.96*sd, ymax = est + 1.96*sd)) +
geom_pointrange()
```

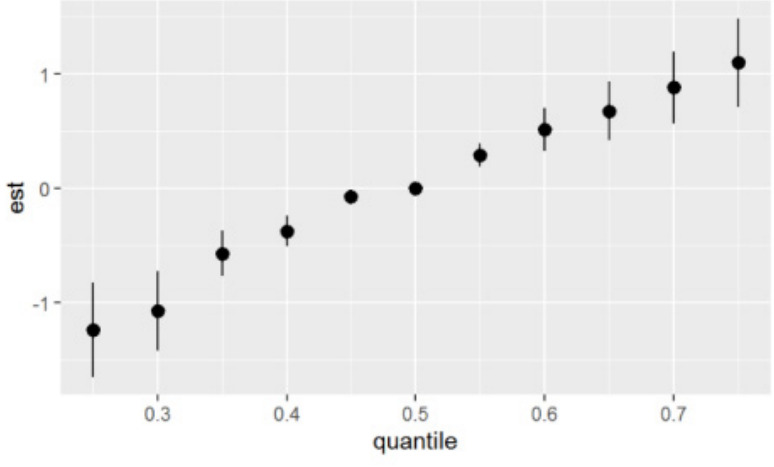


그림 13. 모든 유해물질의 노출 수준이 특정 노출 수준으로 고정되었을 때의 h(·)의 추정치에 대한 그래프.

연구자들이 관심 있을 수 있는 또 다른 정보는 결과 변수에 대한 각 노출 변수의 기여도이다. 이때 그 기여도는 특정 하나의 노출 변수에 대하여 두 percentile에서의 위험을 비교하여 요약되며, 그 외 나머지 노출 변수는 모두 특정 percentile로 고정된다. 이러한 기여도를 단일 노출 변수 효과 (single variable effect)라 할 수 있고, SingVarRiskSummaries 함수를 사용하여 계산할 수 있다. 위험을 비교할 percentile은 qs.diff 인수를 통해 값을 지정할 수 있으며, 나머지 고정할 노출 변수의 값은 q.fixed 인수를 통해 값을 지정할 수 있다.

```
# 각 유해물질에 대해 0.75 percentile과 0.25 percentile의 값의 차이 계산 및 출력
risks.singvar <- SingVarRiskSummaries(fit = fitkm, y = y, Z = Z, X = X, qs.diff
= c(0.25, 0.75), q.fixed = c(0.25, 0.50, 0.75), method = "exact")

risks.singvar
## # A tibble: 12 × 4
##    q.fixed variable          est          sd
##     <fctr>   <fctr>        <dbl>       <dbl>
## 1    0.25       z1  0.9900863616 0.27305455
## 2    0.25       z2  0.6538145079 0.25585205
## 3    0.25       z3  0.0013583345 0.07331792
## 4    0.25       z4 -0.0028078663 0.11460949
## 5     0.5       z1  1.3820953638 0.26906484
## 6     0.5       z2  1.1379994710 0.23862341
## 7     0.5       z3 -0.0008100762 0.05338260
## 8     0.5       z4  0.0365624832 0.10957626
## 9    0.75       z1  1.6852732951 0.32844982
## 10   0.75       z2  1.3105576463 0.27502934
## 11   0.75       z3 -0.0047191015 0.06051743
## 12   0.75       z4  0.0481485003 0.12991681
```

위에서 계산된 결과를 아래의 코드를 사용하여 시각화할 수 있다. 아래 그림은 각 유해물질마다 노출 수준이 75 percentile에서 25 percentile로 바뀌었을 때, 위험의 차이를 보여주고 있다. 여기서 나머지 유해물질의 노출 수준이 모두 25 percentile로 고정되었을 때 (빨강), 모두 50 percentile 고정되었을 때 (초록), 모두 75 percentile (파랑)로 고정되었을 때 결과를 보여주고 있다.

```
# 두 percentile에서의 결과 값의 차이의 시각화
ggplot(risks.singvar,aes(variable, est, ymin = est - 1.96*sd,
  ymax = est + 1.96*sd, col = q.fixed)) +
  geom_pointrange(position = position_dodge(width = 0.75)) +
  coord_flip()
```

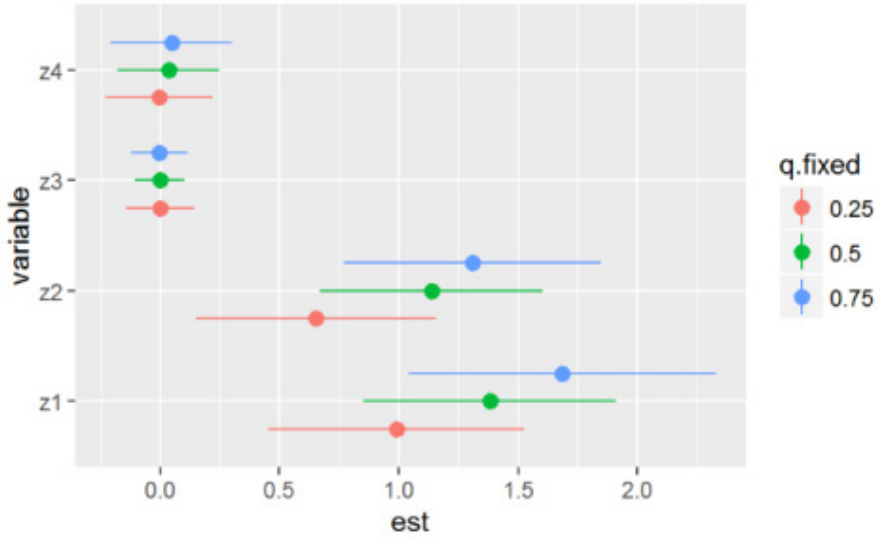


그림 14. 하나의 유해물질 외 나머지 유해물질의 노출 수준이 25, 50, 75 percentile로 고정되어있을 때, 그 유해물질의 노출 수준에 따른 h(·)의 추정치 및 95% 신뢰구간에 대한 그래프

위의 예제에서 보면, 유해물질 z3, z4에 대해서 해당 유해물질의 노출 수준이 75 percentile과 25 percentile로 고정되었을 때의 차이가 0에 가까우므로 결과 변수에 대한 기여도가 크지 않다고 할 수 있다. 하지만 유해물질 z1, z2에 대해서는 다른 유해물질에 대한 노출 수준이 높을수록 결과 변수에 대해 유해물질 z1, z2의 노출 수준이 75 percentile과 25 percentile로 고정되었을 때의 차이가 점점 커지는 것을 확인할 수 있다. 즉, 그림 14은 유해물질 z1, z2의 교호작용의 가능성을 나타낸다.

이러한 교호작용의 가능성을 명확하게 확인하기 위하여, 교호작용 모수에 대한 추정치를 각 유해물질마다 계산할 수 있다. 예를 들어, 유해물질 z1 외 다른 유해물질의 노출 수준이 75 percentile로 고정되었을 때, 유해물질 z1의 노출 수준이 75 percentile와 25 percentile에서의 위험의 차이를, 유해물질 z1 외에 다른 유해물질들이 25 percentile로 고정되었을 때, 유해물질

z1의 노출 수준이 75 percentile와 25 percentile에서의 위험의 차이를 비교하여 교호작용의 크기를 확인할 수 있다. 이는 이전 그림에서 파란색 원으로 표시된 추정치에서 빨간색 원으로 표시된 추정치를 빼는 것에 해당하며, SingVarIntSummaries 함수를 사용하여 수행할 수 있다.

```
# 각 유해물질마다 교호작용의 크기를 산출 및 출력
risks.int <- SingVarIntSummaries(fit = fitkm, y = y, Z = Z,
  X = X, qs.diff = c(0.25, 0.75), qs.fixed = c(0.25, 0.75),
  method = "exact")

risks.int
## # A tibble: 4 × 3
##   variable        est         sd
##   <fctr>        <dbl>      <dbl>
## 1      z1  0.695186934 0.31647915
## 2      z2  0.656743138 0.31785138
## 3      z3 -0.006077436 0.09010269
## 4      z4  0.050956367 0.16267033
```

위의 분석결과를 보면, 유해물질 z2, z3, z4가 75 percentile로 고정될 때와 비교하여 유해물질 z2, z3, z4가 25 percentile로 고정될 때, 유해물질 z1이 25 percentile에서 75 percentile로 증가할 때 건강 위험이 0.7만큼 증가한다는 것을 알 수 있다.

지금까지 구성 요소별 변수 선택법에 기초하여 BKMR 분석결과를 요약하고 시각화하는 방법에 대해 설명하였다. 계층적 변수 선택법 또한 비슷한 방식으로 BKMR 결과를 요약하고 시각화할 수 있다. 계층적 변수 선택법의 사용을 설명하기 위해 4개의 유해물질 중 2개의 유해물질이 높은 상관성을 갖는다고 가정하고, 모의실험 자료를 생성하였다. 생성된 모의실험 자료로부터 유해물질 사이의 상관계수 (correlation coefficient)를 구해보면 유해물질 z1, z3 사이의 상관계수가 다른 유해물질 사이의 상관계수와 비교하여 높게 생성되었다는 것을 확인할 수 있다.

```
# 계층적 변수선택법의 설명을 위한 모의연습 자료 생성
set.seed(111)
d2 <- SimData(n = 100, M = 4, Zgen = "corr", sigsq.true = 2.2)
round(cor(d2$Z), 2)
##         z1    z2    z3    z4
## z1   1.00  0.25  0.96 -0.06
## z2   0.25  1.00  0.42 -0.06
## z3   0.96  0.42  1.00 -0.04
## z4 -0.06 -0.06 -0.04  1.00
```

'구성 요소별 변수 선택법'의 대안으로서의 '계층적 변수 선택법'은 앞서 설명한 것과 같이 상관성 높은 유해물질들을 그룹 간 상관성은 낮지만 그룹 내 상관성은 높도록 분류를 진행한다. 앞서 재현하였던 '구성 요소별 변수 선택법' 결과와 '계층적 변수 선택법' 결과를 비교하기 위한 코드는 다음과 같다. 계층적 변수 선택법을 적용할 때, 구성 요소별 변수 선택법과의 차이점은 groups 인수가 추가된다는 것이다. groups 인수를 통해 각 유해물질이 어떤 그룹으로 분류할지 혼합물의 구조에 대한 사전 지식을 반영할 수 있다.

```
set.seed(111)
# 구성요소별 변수선택법의 사용
fitkm_corr <- kmbayes(y = d2$y, Z = d2$Z, X = d2$X,
  iter = 10000, varsel = TRUE, verbose = FALSE)

# 계층적 변수선택법의 사용
fitkm_hier <- kmbayes(y = d2$y, Z = d2$Z, X = d2$X,
  iter = 10000, varsel = TRUE, groups = c(1, 2, 1, 3),
  verbose = FALSE)
```

두 방법에 대한 사후 포함확률을 비교하면 다음과 같다.

```
# 구성요소별 변수선택법을 사영한 경우에서의 각 유해물질별 기여도
ExtractPIPs(fitkm_corr)
##    variable      PIP
## 1         z1 0.8076
## 2         z2 1.0000
## 3         z3 0.6702
## 4         z4 0.4252

# 계층적 변수선택법을 사용한 경우에서의 각 유해물질별 기여도
ExtractPIPs(fitkm_hier)
##    variable group groupPIP    condPIP
## 1         z1     1   0.9976 0.6046512
## 2         z2     2   1.0000 1.0000000
## 3         z3     1   0.9976 0.3953488
## 4         z4     3   0.3896 1.0000000
```

'fitkm_corr'는 구성 요소별 변수 선택법을 적용하여 BKMR을 적합한 결과에서 사후 포함확률에 대한 요약 결과를 보면 유해물질 z2가 높은 사후 포함확률을 가지고 있기 때문에 유해물질 z2가 모형에 포함되어야 하는 반면 유해물질 z1이 들어가야 하는 증거는 강하지 않음을 알 수 있다. 계층적 변수선택법을 적용한 BKMR 분석 결과인 'fitkm_hier'에서 그룹에 대한 사후 포함확률 (group-specific PIP)에 대한 요약 결과를 보면 그룹 1과 2에 대해 높은 사후 포함확률이 추정되었음을 확인할 수 있다. 그러므로 그룹 1과 2 각각에 속한 유해물질 중 1개의 유해물질이 결과 변수에 영향을 준다는 것을 확인할 수 있다. 이때, 그룹 1에 속한 유해물질 z1과 z3에 대해 유해물질 z1, z3의 사후 포함 확률 (condPIP) 중 유해물질 z1에서 사후 포함확률이 더 크기 때문에 유해물질 z1이 결과 변수와 더 상관성이 높은 것으로 확인하였다. 그룹 2의 경우, 속한 유해물질이 z2 1개이기 때문에 유해물질 z2에 대한 사후포함확률 (condPIP)가 1로 나타났다. 이러한 결과를 바탕으로, '구성 요소

별 변수 선택법'은 유해물질 z2만을 선택하여 잘못된 결과를 도출할 수도 있지만 '계층적 변수 선택법' 에서는 유해물질 z1, z2를 결과에 대한 유해물질로 올바르게 선택될 수 있음을 알 수 있다.

2) 이분형 자료에서 BKMR의 적합(Interpretation of BKMR for a binary outcome)

BKMR을 연속형 변수뿐만 아니라 이분형 결과 변수 (binary outcome)로도 확장하여 사용할 수 있다. 이분형 결과 변수에 대해서 앞서 설명한 연속형 결과 변수와 마찬가지로 kmbayes() 함수를 사용하여 BKMR을 적합한다. 연속형 결과 변수에서 사용한 모든 인수들을 이분형 결과 변수에 BKMR을 적용할 때, 동일하게 사용할 수 있다. 다만 이분형 결과 변수의 경우, kmbayes() 함수에서 family 인수에 "binomial"를 반드시 추가로 입력해야 한다. 이후 BKMR 분석 결과를 요약하고 시각화하는 과정은 위에서 설명한 연속형 결과 변수와 같은 방식으로 진행할 수 있다.

```
# 이분형 자료에서 BKMR의 적합
set.seed(222)
fitkm <- kmbayes(y = y, Z = Z, X = X, iter = 10000,
  verbose = FALSE, varsel = TRUE, family = "binomial")
```

3) 반복 측정 자료에서의 BKMR의 결과 해석(Interpretation of BKMR on repeated measurements)

BKMR은 횡단 자료 (cross-sectional data)뿐 아니라 반복측정 자료 (repeated measurements) 분석을 위해서 사용될 수 있다. kmbayes() 함수를 사용하여 개인의 정보를 반복하여 측정한 자료를 BKMR으로 적합이 가능하다. 이러한 자료에 대해서도 연속형 결과 변수에서 다루었던 여러 인수, 요

약 및 시각화하는 함수를 같은 방식으로 사용이 가능하다. 다만, 반복측정된 자료가 같은 환자로부터 측정되었다는 것을, 즉 개인의 id에 해당하는 정보를 id 인수로 지정한 후, kmbayes() 함수를 사용하면 된다. 이후 과정은 위에서 설명한 연속형 결과 변수와 동일하다.

```
# 반복측정된 대상의 지정
subject<- a$ID

# 반복측정자료에서의 BKMR의 적합
set.seed(333)
fitkm <- kmbayes(y = y, Z = Z, X = X, iter = 10000,
   verbose = FALSE, varsel = TRUE, id = subject)
```

4) BKMR에 필요한 계산 부담의 절감(Reduction for computational burden of BKMR)

BKMR은 MCMC 과정에서 역행렬을 여러 번 계산해야 하므로 BKMR을 적합하는 과정에서 많은 시간이 소요된다. 이에 따라 소요되는 분석 시간을 줄이기 위하여, Bobb JF 등 (2018)은 공간데이터 분석에 사용되는 Gaussian predictive process (Benerjee et al., 2008)를 사용하여, 분석 속도를 개선시키기 위한 knot 옵션을 소개하였다. 이 knot 옵션은 실제로 관측치의 수 (예: 100)보다 더 적은 수인 m개의 수 (예:10)로 knot를 설정하여, 즉 관측치에 해당하는 지점 중 일부 지점들만 선정하여 분석하는 것을 말한다. knot 인수는 다음의 코드를 통해 적용할 수 있으며, 아래 예시에서는 knot의 개수를 10으로 설정하였다. 다만, 주의할 점을 반복측정된 자료에 대해서는 현재 'bkmr' R 패키지에서 knot 옵션을 제공하지 않는다는 점이다.

```
# 계산 속도 개선을 위한 knot의 지정
knots<-fields::cover.design(nd=10)$design

# knot을 이용한 BKMR의 분석 속도 개선
set.seed(444)
fitkm <- kmbayes(y = y, Z = Z, X = X, iter = 10000, verbose = FALSE, varsel
= TRUE, knots=knots)
```

부록 1 III

g-formula의 이론과 적용

Ⅲ. g-formula 이론과 적용

g-formula에 대하여 2021년 연구과제 (예신희 등, 2021)에서는 단일 노출로 인한 건강 영향을 추론하는 국문 가이드라인을 작성하였다. 하지만 2022년 본 과제에서는 2개 이상의 유해물질에 대한 노출이 근로자의 건강에 영향을 미치는지 그 영향을 추론하는데 활용될 수 있는 통계적 방법론에 대한 국문 가이드라인을 작성하는 것뿐만 아니라 산업보건 역학 연구자가 2021년 연구과제에서 작성한 국문 가이드라인의 g-formula에 대한 이해를 보다 쉽게 이해할 수 있도록 하고자 한다. 본 장에 작성된 내용은 Grath 등 (2020), 예신희 등 (2021)을 기반으로 하여 작성되었으며, 복합물질에 대한 여러 노출변수를 처리하는 방법이 중점적으로 추가되었다. 또한, 여러 노출 변수가 있는 경우, R 패키지 gfoRmula(Version 1.0.0)를 사용하는 방법을 기술하였다.

1. g-formula의 개발 배경

Robins JM의 g-formula를 사용하면 연구자가 관심 있는 결과 변수에 대한 시간에 따라 변하는 치료 (treatment), 처치 또는 개입 (intervention) 또는 유해물질의 노출이 미치는 효과의 크기를 추정할 수 있다.

반복측정자료 중 치료-교란 요인 되먹임을 인과 그래프의 구조로 포함하는 자료에서 인과효과를 구하기 위해 전통적인 방법 (예: 선형 회귀분석, 로지스틱 회귀분석, 콕스비례위험 모형)을 사용하면 인과 그래프의 구조로 인하여 인과효과에 대한 편향이 발생한다. 반면, g-formula는 전통적인 방법에서 편향을 발생시키는 치료-교란 요인 되먹임 구조를 모형에 반영하여 편향이 제거된 인과효과 추정치를 제공한다. 치료-교란 요인 되먹임에 대한 설명은 예신희 등 (2021)의 2장 인과 그래프, 단일 노출에서의 g-formula의 이론은

예신희 등 (2021)에서 설명하였으므로, 본 장에서는 복합물질의 여러 노출 변수에 대한 설명에 집중하여 기술하고자 한다.

2. g-formula의 이론

g-formula는 단일 노출뿐만 아니라 복합노출 또는 다중 노출에 대해서도 적합이 가능하도록 만들어진 방법이기 때문에 단일 노출에서 작성된 g-formula에 대한 이론이 복합물질에 대해서도 유지된다. 다만 단일 노출이 아닌 복합물질에 대한 다중 노출에 관해서는 여러 유해물질에 동시에 노출되기 때문에 적용되는 가정이 단일 노출보다 복잡하게 표현되며, 이에 대해 본 장에서 집중적으로 다루고자 한다.

예신희 등 (2021)에서 단일 노출에서의 g-formula의 적합에 요구되는 3가지 가정 (consistency, sequential exchangeability, positivity)을 설명하였다. 복합물질 노출에 의한 건강 영향을 평가하기 위해서는 단일 노출에서와 마찬가지로 위의 3가지 가정이 동일하게 요구된다. 예신희 등 (2021)의 5장에서 g-formula에 대해 설명하기 위해 CD4 수 (Y)에 대한 HIV 치료 (E)의 효과를 구하는 예제를 사용하였다. 이 단일 노출에 대한 예제에서 HIV 치료는 두 시점 E0 (시점 0), E1 (시점 1)에서 이루어졌으며, 치료를 받은 경우, 1, 치료를 받지 않은 경우, 0의 값을 갖는다. 교란 변수 (L)로서 치료 전 HIV 바이러스 부하량이 높아졌는지 여부를 고려하였으며, 바이러스 부하량이 200 copies/ml보다 크면 1, 작으면 0의 값을 부여하였다. 모든 연구대상자는 연구 시작 시점 이전에는 모두 200 copies/ml 이상이라 가정하였다 (L0=1).

이 예제에서 g-formula의 적합에 요구되는 3가지 가정을 열거하면 다음과 같다. 첫째로, ‘consistency’ 가정은 치료 그룹에서의 관찰된 결과는 치료 그룹에서의 잠재적 결과와 동일하고, 비치료 그룹에서의 관찰된 결과는 비치료 그룹에서의 잠재적 결과와 동일하다는 것을 의미한다. 수학적으로 다음과 같이 표현할 수 있다.

$$Y = E_0E_1Y^{E_0=1,E_1=1} + (1-E_0)E_1Y^{E_0=0,E_1=1} + E_0(1-E_1)Y^{E_0=1,E_1=0} + (1-E_0)(1-E_1)Y^{E_0=0,E_1=0}$$

두 번째로 'sequential exchangeability' 가정은 시점 0까지의 치료 이력과 시점 1까지의 교란 요인의 이력이 주어지면 관찰 연구의 자료가 무작위 임상시험의 자료와 비슷함을 의미한다. 즉, 매 시점마다 환자의 HIV 치료 여부가 이전 치료 변수의 이력 (E_0), 교란 요인의 이력 (L_1) 등의 정보로부터 산출된 확률에 따라 무작위로 배정된다는 것을 의미한다. sequential exchangeability를 수학적으로 표현하면 0과 1의 값을 갖는 e_0, l_1, e_1에 대하여 다음이 성립한다 (L_0에 대해서는 모든 연구대상자가 1의 값을 가지므로 아래의 식에서 모두 생략하였음).

$$Y^{E_0=e_0,E_1=e_1} \amalg E_0,\ \ Y^{E_0=e_0,\,E_1=e_1} \amalg E_1 \,|\, E_0=e_0,\ L_1=l_1$$

이 가정은 'no unmeasured confounder assumption'으로 다르게 불리기도 한다. 위의 식이 성립한다는 것은 현재 관찰 연구 자료가 가정의 성립을 위해 필요한 변수를 모두 포함하고 있다는 것이기 때문에 연구 자료에 있는 변수 외에 측정하지 못 한 변수는 없다고 표현하기도 한다.

셋째로, 'positivity' 가정으로 이전 치료 변수의 이력과 교란 요인의 이력으로 구성된 부분 집단에 대해 환자가 치료에 배정될 확률이 0과 1 사이에 있다는 것을 의미한다. 즉, 예제에서 시점 0에서 치료를 받았고 (E_0=1), 시점 1에서 바이러스 부하량이 200 copies/ml보다 작은 (L_1=0) 환자에 대해 HIV 치료를 받은 환자와 받지 않은 환자가 모두 자료에 존재해야 함을 의미한다. 수학적으로 표현하면 다음과 같다.

$$0 < P(E_0=1) < 1,\ 0 < P(E_1=1 \mid E_0=e_0,\ L_1=l_1) < 1$$

지금까지 단일 노출에서 설명하였던 g-formula에 대한 3가지 가정에 대하여 설명하였다. 3가지 가정의 의미는 복합물질에 대한 자료에서도 동일하지만, 수학적으로 표현이 달라지며, 그 점에 대하여 설명하고자 한다. 위의 예제에서는 CD4 수를 낮추기 위해 HIV 치료 한 가지만 고려하였다면, 여러 치료변수를 고려하기 위해 운동 여부 (H)도 같이 고려하여 3가지 가정들이 어떻게 다르게 표현되는지 기술하고자 한다.

첫 번째로, 'consistency' 가정은 처치 변수의 수와 무관하게 단일 노출과 동일한 의미를 갖지만 수학적으로는 다음과 같이 다소 복잡하게 기술된다.

$$Y = E_0E_1H_0H_1\,Y^{(E_0,E_1)=(1,1),(H_0,H_1)=(1,1)} + (1-E_0)E_1H_0H_1\,Y^{(E_0,E_1)=(0,1),(H_0,H_1)=(1,1)} + \cdots + (1-E_0)(1-E_1)(1-H_0)(1-H_1)\,Y^{(E_0,E_1)=(0,0),(H_0,H_1)=(0,0)}$$

처치 변수가 두 개 이상이기 때문에 단일 노출과 같이 잠재적 결과 변수를 표현할 때, 단일 노출 변수에 대해서만 기술하는 것이 아닌 벡터의 형태로 다중 노출 변수를 표현해야 하기 때문이다.

두 번째로, 확장된 예제에서 운동 여부 또한 처치 변수로 고려하고 있기 때문에 운동 여부 또한 HIV 치료 여부와 마찬가지로 각 시점마다 무작위로 배정해야 하므로 운동 여부에 대한 'sequential exchangeability' 가정이 단일 노출과 비교하여 추가되어야 한다. 가정을 추가로 기술하기 위해서는 두 처치변수 사이의 관계가 고려돼야 한다. 예를 들어, 현재 확장된 예제에서 시점 t (t는 0 또는 1의 값)에서의 HIV 치료 여부에 따라 시점 t에서의 운동 여부가 결정된다고 가정하면 g-formula를 적용하기 위해 다음에 해당하는 'sequential exchangeability' 가정들이 요구된다.

$$Y^{(E_0,\,E_1)=(e_0,\,e_1),\,(H_0,\,H_1)=(h_0,\,h_1)} \amalg E_0,$$

$$Y^{(E_0,\,E_1)=(e_0,\,e_1),\,(H_0,\,H_1)=(h_0,\,h_1)} \amalg E_1 \,|\, E_0 = e_0,\, H_0 = h_0,\, L_1 = l_1,$$

$$Y^{(E_0,\,E_1)=(e_0,\,e_1),\,(H_0,\,H_1)=(h_0,\,h_1)} \amalg H_0 \,|\, E_0 = e_0,$$
$$Y^{(E_0,\,E_1)=(e_0,\,e_1),\,(H_0,\,H_1)=(h_0,\,h_1)} \amalg H_1 \,|\, E_0 = e_0,\, H_0 = h_0,\, L_1 = l_1,\, E_1 = e_1$$

여기서 h0 그리고 h1은 0 또는 1의 값을 갖는 상수이다. 앞선 두 조건은 단일 노출에서 기술하였던 가정과 비슷하게 HIV 치료가 시점마다 무작위로 배정되어야 한다는 가정으로 이전과의 차이점은, 특히 두 번째 가정에 대하여, 가정의 조건으로 운동 여부가 포함되어있다는 점이다. 이러한 조건을 통해 g-formula는 처치 변수 사이의 관계를 고려하여 인과효과를 추정한다. 세 번째와 네 번째 가정은 운동 여부가 HIV 치료의 이력, 교란 요인의 이력, 운동 여부의 이력이 주어지면 산출되는 확률에 따라 무작위로 배정되어야 함을 의미한다.

셋째, 'positivity' 가정과 관련하여 운동 여부에 배정될 확률에 대한 조건이 다음과 같이 기술되어야 한다.

$$0 < P(E_0 = 1) < 1,\ 0 < P(E_1 = 1 \,|\, E_0 = e_0,\, H_0 = h_0,\, L_1 = l_1) < 1,$$
$$0 < P(H_0 = 1 \,|\, E_0 = e_0) < 1,$$
$$0 < P(H_1 = 1 \,|\, E_0 = e_0,\, H_0 = h_0,\, L_1 = l_1,\, E_1 = e_1) < 1$$

처음 두 개의 가정은 단일 노출에서 가정한 조건과 비슷한 가정이며, 세 번째와 네 번째 가정은 운동 여부에 대한 처치 변수를 고려함으로써 추가된 가정이다. 위의 가정에서 3, 4번째 조건의 의미에 대해 확률 $P(H_1 = 1 \,|\, E_0 = 1,\, H_0 = 0,\, L_1 = 0,\, E_1 = 1)$을 예를 들어 설명하면, 현재 예제에서 시점 0에서 HIV 치료를 받고, 운동은 하지 않았으며, 바이러스 부하량이 200 copies/ml이고, 시점 1에서 치료를 받은 환자 가운데 운동한 환자가 있어야 함을 의미한다.

단일 노출뿐만 아니라 복합물질에 대한 노출에 대해서도 consistency 가정과 sequential exchangeability 가정은 관측자료로부터 검증 불가능한 가정이며, positivity 가정은 자료로부터 가정을 만족하는지 검증이 가능하다.

3. g-formula의 적용

역학 연구자들은 장기간의 노출와 같은 시간에 따라 변하는 개입의 인과효과를 추정하는 것에 관심이 있는데, 관찰 자료를 가지고 이러한 효과를 추정할 때에는 교란 요인의 보정이 반드시 필요하다. 그러나 기존의 전통적인 방법으로 치료-교란 요인 되먹임이 있는 상황에서 교란 요인을 적절하게 보정하기 어렵다. 반면에 Robins JM의 g-formula는 치료-교란 요인 되먹임이 존재하더라도 교란 요인을 적절하게 보정할 수 있다. g-formula에서 다룰 수 있는 개입 또는 치료 전략 (treatment regime)은 자연적 치료 (natural value of treatment)에 대한 개입뿐만 아니라 정적 (static)이거나 동적 (dynamic) 또는 결정적 (deterministic)이거나 무작위 (random)인 개입과 시간에 따라 변하는 이분형 (binary) 또는 연속형 (continuous) 개입의 인과효과를 추정하기 위해 사용할 수 있다.

본 가이드라인에서는 g-formula가 R에서 구현된 gfoRmula R 패키지에 대한 간략한 소개를 하고 예제를 통하여 사용법에 대하여 설명하고자 한다. 특히 gfoRmula R 패키지에서 결과 변수로 추적관찰 (follow-up)이 끝나는 시점에서 측정된 이분형, 연속형 결과 (end of follow-up outcome) 그리고 생존 여부 (survival)에 대해서 다룰 수 있다.

1) gfoRmula R 패키지란?

연구자들은 종적 자료를 사용하여 인구집단에서 시간의 흐름에 따라 변하는 가상의 개입(예: 치료, 유해물질에 대한 노출, 정책)이 시행되었을 때 예상되는 잠재적 결과를 추론하고자 한다. 하지만 예신희 등 (2021)에서 살펴본 바와 같이 전통적인 방법은 치료-교란 요인 되먹임이 있을 때, 해당 교란 요인을 적절하게 보정할 수 없는 한계를 가지고 있다. 이러한 상황에서 Robins JM의 g-formula (g-computation 또는 plug-in g-formula라 불리기도

함)는 5장에서 소개한 3가지 가정 (sequential exchangeability, consistency, positivity)이 만족되면 연구자가 원하는 시간의 흐름에 따라 결정되는 개입 또는 전략의 인과효과를 추정할 수 있다.

gfoRmula R 패키지는 연구자들이 Robins JM의 g-formula를 사용할 수 있도록 통계학 분야에서 많이 사용되는 R이라는 프로그래밍 언어를 통하여 구현된 함수를 포함하고 있는 패키지이며, 2019년 Lin VL와 McGrath S 등이 gfoRmula R 패키지의 사용법을 설명하는 논문을 발표하였다(Lin VL 등 (2019)) . gfoRmula R 패키지는 R CRAN에서 확인 및 다운 받을 수 있으며, 패키지가 제공하는 다양한 함수에 대해서 더 알고 싶은 연구자는 개발자가 제공하는 reference manual을 참고하기 바란다 (https://cran.r-project.org/web/packages/gfoRmula/index.html).

사용자가 각 시점에서의 어떤 치료전략 (user-specified treatment intervention)을 설정하면 g-formula는 해당 시점에서 설정된 치료전략과 교란 요인의 값들을 고려하여 설정된 치료전략이 시행되었을 때 예상되는 잠재적 결과의 추정치를 제공한다. g-formula는 로지스틱 회귀분석과 같은 일반화 선형 모형 등의 모수적 모형을 통하여 아래의 (1), (2) 그리고 (3)을 추정하고 해당 추정치를 사용하여 몬테카를로 시뮬레이션 (Monte Carlo simulation)을 통하여 잠재적 결과의 값를 추정한다.

(1) 관찰된 치료와 교란 요인의 이력을 보정한 결과의 평균 (결과가 생존 여부일 경우 위험 (hazard))

(2) 관찰된 치료와 교란 요인의 이력을 보정한 각 시간 k마다 관찰되는 교란 요인의 평균

(3) 시간에 따라 변하는 사용자가 지정한 치료전략을 정의하는 개입의 평균.

gfoRmula R 패키지는 일반적인 개입 또는 치료전략을 다룰 수 있으며, 개입 또는 치료전략은 정적이거나 동적이거나 더 나아가 결정적이거나 무작위적일 수 있다. 또한, 정적, 동적, 결정적 또는 무작위적 치료전략뿐만 아니라

"기준"에 따라 다른 치료를 시행하는 치료전략에 대한 효과를 추정할 수 있다. 예를 들면, 연구자가 운동에 대해서 개입을 하려고 할 때, 하루 평균 운동을 30분 이하로 하는 사람들에 대해서만 운동시킨다고 하면 이때 운동 여부를 결정시키는 30분이 "기준"에 해당하며, 이러한 치료전략을 치료의 자연적 값 (natural value of treatment)이라고 한다. gfoRmula R 패키지가 다룰 수 있는 노출 변수의 형태, 결과 변수의 형태는 아래와 같다.

노출 변수의 형태

(1) 이분형, 연속형 그리고 다 수준 (multi-level) 형태의 시간에 따라 변하는 치료 또는 노출.

(2) 2개 이상의 유해물질이 동시에 노출되는 경우의 결합 개입 (joint intervention).

결과 변수의 형태

(1) 생존 여부

(2) 추적관찰이 끝난 시점에서의 이분형 또는 연속형 결과

g-formula를 구현하는 gfoRmula R 패키지는, GFORMULA SAS 매크로의 기능을 많이 가지고 있으며, stata의 gformula 명령어와도 일부 중복된 기능을 가지고 있다. 그러나 gfoRmula R 패키지는 무료이고 오픈소스 플랫폼이므로 R로만 데이터 분석을 수행할 수 있는(혹은 선호하는) 많은 연구자의 접근을 용이하게 한다.

본 가이드라인에서는 (4) 2개 이상의 치료가 동시에 진행되는 경우의 결합 개입에 집중하여 gfoRmula R 패키지의 핵심 기능 및 사용법에 대하여 설명하고자 한다.

2) gfoRmula R 패키지 사용법

gfoRmula R 패키지에서 가장 주요한 함수는 Robins JM의 g-formula를 구현하는 gformula 함수다. gformula 함수와 그 인수는 아래와 같으며, 각 인수에 대해서 설명하고자 한다. 아래의 상자는 설명하기 위한 예시 코드이며, 예시 코드에 대한 설명을 Lin VL 등 (2019)를 기반으로 하여 작성하였다. 이후 등장하는 상자들은 아래의 상자에서 해당 장에서 설명하고자 하는 내용에 대한 인수를 포함한다.

```
gformula(obs_data = obs_data,
id = 'ID',
time_name = 't0',
outcome_type = 'survival',
outcome_name = 'Y',
compevent_name = 'D',
covnames = c('L1', 'L2, 'L3', 'E'),
basecovs = c('race', 'sex'),
histories = c(lagged),
histvars = list(c('L1', 'L2', 'L3')),
covtypes = c('binary', 'normal', 'categorical', 'binary'),
covparams = list(covmodels = c(L1 ~ race+sex+lag1_L1,
                               L2 ~ L1+race+sex+lag1_L2,
                               L3 ~ L1+L2+race+sex+lag1_L3,
                               E ~ L1+L2+L3+race+sex+lag1_E),
                 covlink = c('logit', 'identity', 'logit', 'logit')),
ymodel = Y ~ E+L1+L2+L3+lag1_L1+lag1_L2+race+sex,
compevent_model = D ~ E+L1+L2+L3+lag1_L1+race+sex,
intvars = list(c('E'), c('E')),
interventions = list(list(c(static, rep(0, 10))),
                     list(c(static, rep(1, 10)))),
int_descript = c('Never treat', 'Always treat'),
ref_int = 0,
nsimul = 10000,
ci_method = "percentile",
nsamples = 1000,
seed = 1234,
parallel = TRUE,
ncores = ncores, ...)
```

(1) 입력자료의 구조

gformula 함수에 자료를 입력하기 위해서는 자료의 자료형이 data.table 형태이어야 한다. 다만 data.frame의 형태이더라도 함수는 오류 없이 실행되지만, 주의 메시지 (warning message)가 발생한다. 결과 유형과 관계없이, 입력자료에는 각 연구대상자에 대해 시점 k마다 하나의 기록이 있어야 하며, 시점을 의미하는 k는 0 (baseline 시점)에서 시작하고 1씩 증가해야 한다. 예를 들어, 1번 환자에 대한 기록이 시점 0, 1, 3에서 측정된 자료는 시점 값이 1씩 증가하지 않기 때문에 입력자료로 사용하기 어렵다. 이러한 시점 변수를 입력자료로 사용하기 위해서는 기존 시점에 해당하는 변수를 측정 시점에 관계 없이 방문순서로 변형하여 자료에 재입력해야 한다. 1번 환자는 기존 시점 변수에서 시점 0, 1, 3에 측정하였으므로 이를 방문순서로 재정의하면 1번 환자는 0, 1, 2 시점에 측정한 것으로 수정할 수 있다. 다른 예시로 2번 환자가 기존 시점 변수에서 시점 0, 4에 측정하였다면 2번 환자의 방문순서는 0, 1 시점에 측정한 것으로 수정할 수 있다. 자료의 각 열은 시간에 따라 변하는 교란 요인 (time-varying covariate)과 시간에 의존하지 않는 교란 요인 (baseline covariate 또는 time-invariant covariate)으로 구성된다. 정리하면 입력자료에는 행 (row)을 단위로 하여 환자에 대한 관측치가 반복측정된 횟수만큼 기록되어있으며, 열 (column)은 시간에 따라 변하는 또는 변하지 않는 교란 요인, 노출 변수 그리고 결과 변수에 대한 정보가 기록되어있다.

(2) 결과 변수 유형을 지정하는 법

Outcome_type은 연구 가설에서 주요한 관심 결과 변수의 유형을 지정하기 위해 사용되며, 사용 가능한 결과 변수로는 생존 여부 (survival)와 추적관찰 종료 시점에서 관찰된 연속형 (continuous_eof) 또는 이분형 (binary_eof) 결과 변수이다:

1) 'survival'은 생존 또는 사망 여부에 해당할 때 사용해야 함(예: 각 시점

에서 모든 원인으로 인한 사망 여부)

2) 'continuous_eof'는 추적관찰이 종료된 시점에서의 관찰된 연속형 결과 (예: 추적관찰 종료 시점 K+1에서의 혈압)를, 'binary_eof'는 추적관찰이 종료된 시점에서 관찰된 이분형 결과 (예: 추적관찰 종료 시점 K+1에서의 혈압이 특정 기준보다 높은지에 대한 지표)를 분석하기 위해 사용한다.

▪ 생존 여부 (survival)

연구 가설에서 주요한 관심 결과 변수의 유형이 생존 여부인 경우, 자료에 포함된 결과 변수는 사망한 시점에서는 1, 그 외의 시점에서는 0의 값을 갖는다. 즉, 시점 k에서 환자 또는 연구대상자가 사망하였으면 Yk = 1, 사망하지 않았으면 Yk = 0으로 표현한다. 또한, 생존 분석 (survival analysis)에서 나타날 수 있는 중도절단 사건과 경쟁 사건을 변수 C, D를 사용하여 표현하고자 한다. 알파벳 우측 하단에 있는 값은 요인이 발생한 시점을 의미하며, 중도절단 사건 Ck는 시점 k에서 환자 또는 연구대상자가 추적관찰을 종료한 경우 1, 그렇지 않은 경우, 0의 값, 경쟁 사건 Dk는 시점 k에서 환자 또는 연구대상자가 연구 대상에 해당하는 질병이 아닌 다른 질병 또는 상태로 인하여 추적관찰이 종료된 경우 1, 그렇지 않으면 0의 값을 갖는다. 생존 여부의 경우, gformula 함수는 사용자가 지정한 치료전략에서 추적관찰 기간동안 질병이 발생할 위험, 즉 해당 시간까지 질병이 발생할 확률의 추정치를 제공한다.

경쟁 사건을 중도절단으로 취급하기로 선택한 경우 입력자료에는 경쟁 사건에 대한 변수를 따로 지정할 필요가 없다. 하지만 경쟁 사건을 중도절단으로 처리하지 않는 경우 연구자는 경쟁사건 D를 표현하는 열을 생성하여 자료에 포함해야 한다. 어떤 연구대상자에 대하여 시점 k를 기준으로 다음 시점 k+1에서 경쟁 사건이 일어난다면 (D_{k+1}=1) 다음 시점 k+1에서의 결과 Y_{k+1}은 경쟁 사건으로 인하여 관측할 수 없기 때문에 NA로 입력되어있어야 한다.

▪ 추적관찰 종료 시점에서 관찰된 결과 (continuous_eof, binary_eof)

추적관찰이 종료된 시점에서 관찰된 연속형/이분형 결과의 경우, g-formula 함수는 추적관찰이 종료된 시점에서의 결과 변수의 평균값의 추정치를 제공한다. 또한, 이 경우, 반복측정된 결과 변수의 값 중에서 추적관찰이 종료된 시점에서 관찰된 결과만 알고리즘에 사용된다. 만약 추적관찰 중 연구대상자가 사망한다면, 중도절단 사건으로 취급한다. K+1보다 작은 시점 k에서 중도절단된 연구대상자의 경우, Y를 제외한 공변량에는 시점 k까지의 추적관찰 결과가 포함되어있으며, Y는 시점 k에서의 값만 입력되며, 그 이전 시점의 Y는 NA로 입력하면 된다. 아래의 표 1은 연구대상자의 최대 반복측정이 4회인 가상의 자료를 보여준다.

표 1. 추적관찰 종료 시점에서 관찰된 결과가 자료에서 표현되는 형태를 기술한 표

ID	t0	Y
1	0	NA
1	1	NA
1	2	NA
1	3	52
2	0	NA
2	1	NA
2	2	NA
2	3	76

위의 자료로부터 1번과 2번 연구대상자는 0 시점부터 3 시점까지 4회 모두 관찰되었으며, 추적관찰 종료 시점인 3 시점에만 결과 변수 Y에 값이 있고, 그 이전 시점에 대해서는 NA 값이 입력되어있다.

(3) 입력자료, 연구대상자 식별 아이디, 시간 지표, 결과 및 경쟁사건 지정 (obs_data, id, time_name, outcome_name, compevent_name)

입력자료는 obs_data, 연구대상자의 식별 아이디는 id, 연구대상자의 반복측정 시간은 time_name, 분석하고자 하는 결과 변수는 outcome_name, 경쟁 사건에 해당하는 변수는 compevent_name을 통해 입력할 수 있다.

```
gformula(…,
  obs_data = obs_data,
  id = 'ID',
  time_name = 't0',
  outcome_type = 'survival',
  outcome_name = 'Y',
  compevent_name = 'D'
)
```

위의 예제 구문을 통해 입력자료는 obs_data이며, 연구대상자를 식별할 때 사용되는 변수는 ID, 자료에서 연구대상자에 대하여 반복측정시간을 나타내는 변수는 t0, 분석하고자 하는 결과 변수는 Y, 결과 변수의 관찰 여부에 영향을 주는 경쟁 사건은 D임을 알 수 있으며, obs_data 자료는 ID, t0, Y, D 4개의 변수를 모두 포함하고 있어야 한다. 위의 예시 구문에서 소개하지 않은 인수로 time_points가 있다. 이 인수는 연구의 최대 반복측정 횟수를 의미한다. 하지만 gformula 함수 내에서 최대 반복측정 횟수를 계산하여 자동으로 입력하는 것이 기본값으로 설정되어있어 일반적인 경우, 별도의 설정이 필요하지 않다.

(4) 시간에 따라 변하는 교란 요인과 기준선 교란 요인 지정 (covnames, basecovs)

시간에 따라 변하는 내생 교란 요인과 노출 변수는 covnames으로 정의한다. covnames으로 요인의 이름을 기술할 때 주의해야 할 사항은 요인이 발생하는 시간적 흐름을 반드시 고려하여 입력해야 한다는 점이다. 시간에 따라 변하지 않는 교란 요인 또는 시간에 따라 변하는 외생 요인은 basecovs으로 정의한다. 이때 외생 요인 (exogenous variable)이란 연구대상자의 건강 상태와 무관하게 그 값이 결정되는 요인을 의미하며, 그 예시로는 연구대상자의 나이, 연구대상자가 거주하는 시군구의 미세먼지 농도 등이 있다. 반면 내생 요인 (endogenous variable)이란 외생 요인에 해당하지 않는 요인을 의미하며, 연구대상자의 건강 상태에 따라 그 값이 결정되는 요인 등을 포함한다. 흡연 상태, 혈중 중금속 농도 등을 예시로 볼 수 있다. 시간에 따라 변하는 교란 요인과 기준선 교란 요인 지정에 관한 예시 구문은 아래와 같다.

```
gformula(…,
 covnames = c('L1', 'L2', 'L3', 'E'),
 basecovs = c('race','sex')
)
```

위의 예시 구문에서 covnames = c('L1', 'L2', 'L3', 'E')의 의미는 시간에 따라 변하는 내생 교란 요인 및 노출 변수가 L1, L2, L3, E 임을 나타내며 또한, 요인 발생의 시간적 흐름이 L1, L2, L3, E 순으로 일어나는 것을 내포한다.

복합물질에 대한 노출과 같이 2개 이상의 유해물질에 동시에 노출되는 경우 또는 2가지 이상의 치료를 같이 처치하는 경우 또한, 패키지를 통해 잠재적 결과를 추정할 수 있다. 앞선 예시에서 노출 변수에 해당하는 E 외에 추가적인 노출 변수 H가 있고, E에 대한 노출이 H에 대한 노출에 영향을 준다고

생각해보자. 이러한 경우, 요인 발생의 시간적 흐름이 L1, L2, L3, E 그리고 K이므로, covnames 인수에 c("L1", "L2", "L3", "E", "H")를 입력하면 된다. 코드는 다음과 같이 작성할 수 있다.

```
gformula(…,
 covnames = c('L1', 'L2', 'L3', 'E', 'H'),
 basecovs = c('race','sex')
)
```

(5) 교란 요인 및 노출 변수에 대한 이력 생성 (histories, histvars)

histvars 및 histories는 R의 모델 구문 내에서 정의할 수 없지만, 노출 변수 및 교란 요인의 이력을 표현하기 위한 함수를 지정하기 위해 사용한다. 예를 들어, 입력자료에서 시간에 따라 변하는 어떠한 교란 요인 L4를 고려해보자. 이때, 연구자가 교란 요인 L4에 대해서 k 시점 이후의 이 요인의 분포가 시점 k까지의 이 요인의 누적 평균에 의존한다고 생각한다고 해보자. 교란 요인 L4의 이력을 표현하는 함수는 시점 k가 0보다 큰 경우 $\frac{1}{k}\sum_{t=0}^{k} L_{4,t}$, 그렇지 않으면 L4,0으로 표현된다. 연구자가 고려하는 이력에 대한 함수를 histvars 및 histories를 사용하여 g-formula 함수에 반영할 수 있으며, histvars 및 histories를 사용하면 자료 obs_data에 각 시점 k에서 교란 요인 L4의 이력을 표현하는 시간에 따라 변하는 요인이 생성된다. gfoRmula R 패키지에서 이력을 표현하기 위해 제공하는 기본적인 함수는 다음과 같다.

- lagged: 모든 i=1…,r (여기서 r은 원하는 지연 수)에 대하여 추적관찰 시점 k에서 Lj의 i 번째 지연을 포함하는 lagi_Lj라는 변수를 자료에 추가한다. lagi_Lj는 k<i 인 행에서 0으로 설정된다.

- cumavg: 추적관찰 시점 k>0까지 Lj의 누적 평균을 포함하는 cumavg_Lj라는 변수를 자료에 추가한다. k=0에서는 Lj로 설정된다.

- lagavg: 추적관찰 시점 k와 i=1,…,r에 대한 Lj 누적 평균의 i 번째 지연을 포함하는 lag_cumavgi_Lj라는 변수를 자료에 추가한다. lag_cumavgi_Lj는 k<i 인 행에서 0으로 설정된다.

이 g-formula 함수에서 histories의 q 번째 요소에 나열된 이력 함수를 histvars의 q 번째 요소에 입력한 모든 변수에 적용한다 (q = 1,2,3, ...). 그러므로 histories 벡터의 길이와 histvars 리스트의 길이가 일치해야 한다.

```
gformula(…,
 histories = c(lagged),
 histvars = list(c('L1', 'L2', 'L3')
)
```

위의 예시 구문은 L1, L2, L3에 대해서 적어도 1 시점 이상 지연된 변수(lagged)를 사용한다는 것을 의미한다. 조건부 분포/위험/평균의 추정에 사용하기 위해, 입력자료 obs_data에 새 변수로 교란 요인에 대해 지연된 변수, 누적 평균 그리고 지연된 누적 평균 함수를 추가하는 예제는 아래와 같다.

```
gformula(…,
 histories = c(lagged, cumavg, lagavg),
 histvars = list(c('L1', 'L2', 'L3'), c('L1', 'L2'), c('L1', 'L3')
)
```

위의 예시에서 보면 histories에 lagged, cumavg, lagavg가 입력되어 있으며 길이가 3이고, histvars에는 list의 형태로 c(‘L1’, ‘L2’, ‘L3’), c(‘L1’, ‘L2’), c(‘L1’, ‘L3’)으로 입력되어 있으며, 길이가 3으로 histories의 길이와 같음을 확인할 수 있다. 이 예제를 통하여 교란 요인 L1, L2, L3에 대해서 지연된 변수를 생성하고, 교란 요인 L1, L2에 대해서 누적 평균 변수를 생성하고, 교란 요인 L1, L3에 대해서는 지연 누적 평균 변수를 생성하는 것을 의미한다.

(6) 교란 요인 및 노출 변수에 대한 분포 지정 (covtypes, covmodels, covlink)

covtypes는 교란 요인 및 노출 변수의 이력이 주어졌을 때 시간에 따라 변하는 교란 요인 및 노출 변수의 평균을 모형화하기 위해 그리고 몬테카를로 시뮬레이션을 통하여 관측치를 생성할 때 필요한 분포를 지정하기 위해 사용되며, covnames과 하나씩 대응되기 때문에 covnames와 길이가 같아야 한다. gfoRmula R 패키지는 covtypes에 대해 미리 구현된 여러 옵션(‘binary’, ‘normal’, ‘categorical’, ‘bounded normal’, ‘zero-inflated normal’, ‘truncated normal’, ‘absorbing’, ‘categorical time’)을 제공하며, covtypes에 해당하는 분포를 모형화할 때 필요한 요소 (covmodels, covlink)는 covparams에 list의 형태로 입력해야 한다.

```
gformula(…,
  covnames = c(‘L1’, ‘L2, ‘L3’),
  histories = c(lagged),
  histvars = list(c('L1', 'L2', 'L3’)),
  covtypes = c('binary', 'normal', 'categorical', ‘binary’),
  covparams = list(covmodels=c(L1~race+sex+lag1_L1,
                               L2~L1+race+sex+lag1_L2,
                               L3~L1+L2+race+sex+lag1_L3,
                               E~L1+L2+L3+race+sex+lag1_E),
                   covlink = c(‘logit’, ‘identity’, ‘logit’, ‘logit’))
)
```

위의 예시 구문에서 교란 요인과 노출 변수 L1, L2, L3, E 각각을 모형화하기 위한 분포로 교란 요인 L1은 이분형 변수이므로 이항 분포 (binomial distribution), 교란 요인 L2는 정규분포를 따르는 변수이므로 정규 분포 (normal distribution), 교란 요인 L3은 범주형 변수이므로 다항 분포 (multinomial distribution) 그리고 노출 변수 E는 노출 여부에 해당하는 이분형 변수이므로 이항 분포를 가정하였다. 이항 분포와 정규 분포의 경우 glm 함수를, 다항 분포의 경우 'nnet' 패키지의 multinom 함수를 통하여 모형을 적합하기 때문에 모형 적합에 기본적으로 필요한 요소는 formula와 link 함수이다. 각각의 함수에 입력할 formula와 link 함수를 벡터의 형태로 각각 covmodels 과 covlink 에 입력한다. covmodels를 통해 모형을 기술할 때, 모형화하고자 하는 시간에 따라 변하는 내생 교란 요인 또는 노출 변수는 ~의 왼쪽에 위치하며, 오른쪽에는 시간에 따라 변하는 내생 교란 요인 또는 노출 변수에 영향을 주는 요인들이 위치한다. 예를 들어, L1 ~ race + sex + lag1_L1은 ~의 왼쪽에 위치하는 시간에 따라 변하는 내생 교란 요인 L1을 모형화하고자 하는 것이며, 내생 교란 요인 L1은 race, sex 그리고 이전 시점의 교란 요인 L1에 의하여 영향을 받는다는 것을 나타낸다. 모형을 기술할 때, 특히 주의해야 하는 점은 요인들의 발생 시점에 대한 고려가 모형에 반영되어 인과 그래프가 순환되지 않도록 해야 한다는 점이다. 예를 들어, 연구자가 음주 여부가 흡연 여부에 의해 영향을 주고, 흡연 여부가 음주 여부에 영향을 준다고 가정하였다고 하자. 이 경우, 두 요인 간 발생 시점에 대한 고려가 반영되지 않아 음주와 흡연에 대한 모형으로 각각 '흡연 여부 ~ 음주 여부', '음주 여부 ~ 흡연 여부'와 같이 기술될 수 있다. 이 경우 서로 같은 시점에서 흡연 여부가 음주 여부에 영향을 주며 음주 여부가 흡연 여부에 영향을 주기 때문에 논리적으로 모순이 된다. 이러한 문제를 피하기 위해 각 요인의 발생 시점을 고려하여 모형을 기술해야 하며, t 시점에서의 흡연은 t 시점에서의 음주 여부 그리고 t 시점에서의 음주 여부는 t-1 시점에서의 흡연 여부와 같이 기술할 수 있다. 각 분포를 모형화하기 위해 사용하는 함수 (위의

예시에서 glm, multinom에 해당함)에서 기본값으로 생각되는 인수들은 따로 작성하지 않아도 된다. 위의 예시 구문에서는 link 함수를 작성하였지만 glm 함수에서 기본값으로 사용하는 link 함수를 gformula 함수 내에서도 사용한다면 covlink에 NA를 입력하면 된다.

위의 예제에서는 노출 변수로서 변수 E 1개만을 고려하였다. 앞서 (4)에서의 예제와 같이 노출 변수 E뿐만 아니라 변수 G 또한 노출 변수로 고려될 수 있다. 각 노출 변수에 대해 앞서 기술한 노출 변수 E가 노출 변수 G에 영향을 준다는 것 외에 각 노출 변수는 시간에 따라 변하는 교란 요인 L1, L2, L3와 시간에 따라 변하지 않는 교란 요인 race, sex 그리고 자신의 이전 시점의 노출 변수에 영향을 받는다고 생각해보자. 이러한 상황에서 앞서 설명한 단일 노출에 대한 코드를 응용하여 t 시점의 노출 변수 E와 G에 대해 모형화를 하면 아래와 같다.

```
gformula(…,
  covnames = c('L1', 'L2, 'L3', 'E', 'G'),
  histories = c(lagged),
  histvars = list(c('L1', 'L2', 'L3', 'E', 'G')),
  covtypes = c('binary','normal','categorical','binary','binary'),
  covparams =list(covmodels=c(L1~race+sex+lag1_L1,
                              L2~L1+race+sex+lag1_L2,
                              L3~L1+L2+race+sex+lag1_L3,
                              E~L1+L2+L3+race+sex+lag1_E1,
                              G~L1+L2+L3+race+sex+E+lag1_G),
                  covlink = c(NA, NA, NA, NA, NA))
)
```

(7) 결과 및 경쟁 사건에 대한 모델 지정 (ymodel, compevent_model)

ymodel은 결과 변수에 대한 모형을 지정할 때 사용한다. gfoRmula R 패키지에서 gformula 함수는 이분형, 생존 여부 또는 경쟁 사건인 결과 변수의 평균에 대한 모형으로 R의 stats 패키지의 glm 함수를 사용하고 있기 때문에

glm에서 formula를 입력하는 방식과 동일하게 gformula의 ymodel을 입력하면 된다. 또한, ymodel 입력 시 covmodels을 입력할 때와 마찬가지로 결과 변수를 ~의 왼쪽에, 결과 변수에 영향을 주는 요인을 ~의 오른쪽에 입력하면 된다. gfoRmula R 패키지에서 gformula 함수는 결과 변수에 대하여 각 시점에서의 모형을 적합하는 방식이 아닌 모든 시점의 자료를 통합하여 모형을 적합하는 방식인 pooled regression 접근 방식을 사용하고 있다. 다만 경쟁 사건에 대한 모형은 결과 변수의 유형이 survival인 경우에만 적용될 수 있다.

gfoRmula R 패키지는 추적관찰이 종료된 시점의 이분형 또는 연속형인 결과 변수 또한, R의 glm 함수를 사용한 선형 또는 로지스틱 회귀분석을 적합하므로 위와 동일하게 ymodel을 입력하면 된다.

```
gformula(…,
  ymodel = Y ~ E+L1+L2+L3+lag1_L1+lag1_L2+race+sex,
  compevent_model = D ~ E+L1+L2+L3+lag1_L1+race+sex
)
```

(8) 개입에 대한 지정 (intvars, interventions)

"개입 전략 (intervention rule)"이란 한 가지 이상의 치료에 대한 개입 전략으로 정의한다. 그러나 "개입 (intervention)", "개입 전략 (intervention strategy)" 또는 "전략 (strategy)"이라는 용어는 두 가지 이상의 치료에 대한 개입 전략을 가리키는 "공동 개입"과 혼용하여 사용되기도 한다.

intvars, interventions, int_times는 연구자가 비교하고자 하는 치료 개입을 지정할 때 함께 사용된다. intvars은 연구자가 확인하고자 하는 개입 규칙이 list 형태로 입력된다. 즉, 연구자가 확인하고자 하는 개입 전략이 두 가지이면 intvars은 두 가지 개입 전략을 가지는 list로 구성된다. 또한, 한 가지

변수에만 개입하는 경우 단일 개입 전략 (single intervention rule)이 되며, 두 가지 이상의 변수에 개입하는 경우 공동 개입 전략 (joint intervention rule)이 된다.

interventions는 개입하고자 하는 변수의 특정 값과 개입 전략의 유형을 입력한 list를 구성요소로 하는 list이다. 좀 더 구체적으로 표현하면 interventions은 list를 구성요소로 하는 list의 형태이기 때문에 개입을 지정하는 목록을 “외부 목록 (outer lists)”이라고 하고, 이 외부 목록의 구성요소를 “내부 목록 (inner lists)”이라고 한다. 외부 목록의 길이는 연구자가 확인하고자 하는 개입 전략의 수와 일치하여야 한다. 각 내부 목록은 개입 전략의 형태를 지정한 list이며, 이 list는 개입하고자 하는 변수의 특정 값과 개입 전략의 유형으로 구성된다. 개입하고자 하는 변수의 특정 값으로는 한 가지 변수에만 개입하는 경우 해당 변수에 대한 단일 벡터만 입력하면 되고, 2개 이상의 노출 변수에 대한 개입일 경우 각 노출 변수에 대한 벡터를 입력해야한다.

int_descript는 연구자가 설정한 개입을 명명하기 위해 사용된다. 위의 설명에 대한 예제 구문은 아래와 같다.

```
gformula(…,
            time_points = 10,
            intvars = list(c('E'), c('E')),
            interventions = list(list(c(static, rep(0, 10))),
                                 list(c(static, rep(1, 10)))),
            int_descript = c(‘Unexposed’, ‘Exposed’)
)
```

위의 예시 구문은 노출 변수 E에 대하여 연구자가 두 가지 개입 전략을 시행하였을 때의 잠재적 결과를 추정하고자 하는 코드이다. 코드를 설명하면 연구자는 두 가지 개입 전략에 대하여 모두 0 시점부터 9 시점까지 개입을 시행하고자 하며, 개입하려는 변수는 모두 노출 변수 E이다. 또한, 개입 전략의

유형은 모두 static이고, 두 가지 개입 전략 중 하나의 개입 전략은 0 시점에서 9 시점까지 노출 변수 E의 값이 모두 0인 전략, 즉 10개의 시점에서 모두 유해물질에 노출되지 않은 경우를, 나머지 하나의 전략은 0 시점에서 9 시점까지 노출 변수 E의 값이 모두 1인 전략, 즉 10개의 시점에서 모두 유해물질에 노출된 경우를 의미한다. 여기서 개입 전략의 유형 static은 매 시점에서의 노출 변수 E의 값이 교란 요인의 이력에 영향받지 않고, 특정 값으로 고정되는 개입 전략을 의미한다. 10개의 시점에서 모두 노출되지 않은 경우와 모두 노출된 경우를 전략으로 생각하고 있으므로 int_descript를 통해 각 전략에 대한 이름으로 'Unexposed'와 'Exposed'로 명명하였음을 알 수 있다.

```
gformula(…,
            time_points = 10,
            intvars = list(c('E', 'G'), c('E', 'G')),
            interventions = list(list(c(static, rep(0, 10)),
                                      c(static, rep(0, 10))),
                                 list(c(static, rep(1, 10)),
                                      c(static, rep(1, 10))))
)
```

위의 예시 구문은 첫 번째 구문을 (4) 시간에 따라 변하는 교란 요인과 기준선 교란 요인 지정, (6) 교란 요인 및 노출 변수에 대한 분포 지정 각각의 마지막 문단에서 다루었던 2개 이상의 노출 변수로 확장한 것이다. 이 구문을 통해 연구자가 노출 변수 E, G에 대하여 0 시점부터 9 시점까지 총 10개의 시점에 대하여 유해물질 E와 G에 모두 노출되지 않은 경우와 모두 노출된 경우에 대한 잠재적 결과를 추정하려는 것임을 알 수 있다.

```
gformula(…,
        intvars = list(c('L2'), c('L2')),
        interventions = list(list(c(threshold, 2, Inf)),
                             list(c(threshold, 3, Inf)))
)
```

- natural: natural 함수는 현재 자료의 노출 패턴을 의미하는 자연 경과 (natural course)를 지정한다. 자연 경과는 함수에서 항상 제공하는 값이기 때문에 분석 시 따로 지정할 필요가 없다.

위의 예시 구문은 static과는 다른 개입 전략의 유형인 threshold을 지정한 코드이다. threshold는 앞서 3-1) gfoRmula R 패키지란?에서 언급한 치료의 자연적 값 (natural value of treatment)에 해당하는 치료전략을 의미한다. 구문을 통해 앞선 예시와는 달리 노출 변수 E와 G를 처치 변수로 지정하지 않고, 교란 요인 L2를 처치 변수로 지정한 것을 알 수 있으며, 이때 치료전략으로써 L1의 값이 2보다 큰 경우 개입을 시행하였을 경우와 3보다 큰 경우 개입을 시행하였을 경우에 대한 잠재적 결과를 추정하려는 것을 알 수 있다. 치료전략으로 static, threshold 외에 dynamic과 natural이 있으며 설명은 아래와 같다.

- dynamic: dynamic 인수는 노출 변수, 교란 요인의 이력에 따라 달라지는 치료 전략을 지정하기 위해 사용된다.
- natural: natural 인수는 현재 자료의 노출 패턴을 의미하는 자연 경과 (natural course)를 지정한다. 자연 경과는 함수에서 항상 제공하는 값이기 때문에 분석 시 따로 지정할 필요가 없다.

"reference"로 사용되는 개입은 차이, 위험 비 또는 오즈 비 등을 계산할 때 기준값으로 사용되는 개입이다. 기본적으로 ref_int는 0이며, 0은 자연 경과를 의미한다. 여기서 자연 경과란 별 다른 개입을 시행하지 않는 치료전략, 즉 현재 자료에서 나타나는 노출 패턴을 의미한다. 자연 경과가 아닌 다른 치

료전략을 reference로 설정하고 싶은 경우에는 ref_int 인수를 통해 해당 치료전략으로 설정하면 된다.

필요한 경우 연구자는 int_times를 사용하여 개입이 적용되는 시점을 지정할 수 있다. intvars와 마찬가지로 int_times는 list를 구성요소로 하는 list의 형태를 갖는다. int_times의 구성요소는 개입을 적용하고자 하는 시점을 지정한다. 개입이 적용되지 않은 시점에서는 실제 자료에서 시행된 노출 변수의 값이 사용된다. int_times을 지정하지 않는 경우, 기본적으로 모든 시점에 개입이 시행되는 것으로 간주하여 잠재적 결과를 추정한다. 아래 예제 구문은 시간에 따라 변하는 교란 요인 L2에 대하여 두 개의 개입을 비교하는데 시점 0과 시점 1에는 개입이 적용되지 않아 실제 자료에서 관측된 치료의 값이 할당되고, 시점 2 이후에는 threshold 개입을 적용하여 교란 요인 L2가 2보다 큰 경우에 나타나는 잠재적 결과와 3보다 큰 경우에 나타나는 잠재적 결과를 추정하게 된다.

```
gformula(…,
            time_points=10,
            intvars = list(c('L2'), c('L2')),
            interventions = list(list(c(threshold, 2, Inf)),
                                 list(c(threshold, 3, Inf))),
            int_times = list(list(2:9), list(2:9))
)
```

(9) 신뢰구간의 측정 및 몬테카를로 시뮬레이션 (ci_method, nsamples, nsimul)

gformula 함수는 몬테카를로 시뮬레이션을 통하여 결과 변수의 평균 또는 위험의 추정치를 산출하며 이때 몬테카를로 시뮬레이션을 몇 번 반복할지 nsimul을 통하여 설정할 수 있다. nsimul의 값은 자료 안에 있는 연구대상자의 수가 10,000명보다 작은 경우에는 10,000으로, 연구대상자가 10,000

명보다 큰 경우 연구대상자의 수로 기본값이 설정되어있다.

gformula 함수는 추정치에 대한 신뢰구간을 제공하기 위해 붓스트랩(bootstrap)을 이용한다. nsamples를 통해서 현재 연구자료에서 몇 번의 붓스트랩을 할지 결정하며, 붓스트랩의 결과를 사용하여 신뢰구간을 어떻게 산출할지 ci_method로 결정할 수 있다. 또한, 붓스트랩를 이용하여 신뢰구간의 추정치를 구하기 때문에 같은 자료로 gformula 함수를 다시 시행하였을 때 랜덤 넘버로 인하여 다른 결과가 나올 수도 있는데, seed를 사용하여 랜덤 넘버를 고정하고 결과의 재현성을 확보할 수 있다.

```
gformula(…,
            nsimul = 10000,
            ci_method = 'percentile',
            nsamples = 1000
)
```

(10) 계산 속도 (parallel, ncores)

gformula 함수는 계산량이 많이 요구되는 몬테카를로 시뮬레이션과 붓스트랩을 모두 사용하기 때문에 결과를 제공하기까지 오랜 시간이 소요된다. 하지만 이러한 문제는 병렬계산을 통하여 다소 해결할 수 있으며, 병렬계산을 사용할 수 있도록 gformula 함수는 parallel과 ncores를 제공한다. 병렬계산을 하고자 할 때, parallel의 값을 TRUE로 설정하고, 이때 사용할 CPU core의 개수를 ncores에 입력하면 된다. 다만 ncores를 입력하는 경우에는 gformula 함수 앞 뒤에 아래의 예시와 같은 준비 코드가 필요하다. 첫 번째 줄에서 -1를 하는 이유는 데스크탑 또는 노트북의 CPU가 gformula 함수 외에 다른 프로그램을 처리할 수 있도록 여유분을 설정하려는 목적으로 입력한 임의의 숫자이며, 연구자의 데스크탑이 보유하고 있는 CPU를 모두 gformula 함수를 처리하는데 사용하려는 연구자는 -1가 아닌 0을 또는 여유분을 더 남기려는 연구자는 -3, -4 등의 숫자를 사용해도 된다.

```
ncores <- parallel::detectCores() - 1

gformula(…,
         parallel = TRUE,
         ncores = ncores
)
```

(11) 출력 결과 (outputs)

gformula 함수는 다음과 같은 주요 결과들을 보여준다.

(1) 자연 경과 시의 비모수적 위험 추정치, 자연 경과를 포함하여 연구자가 설정한 개입 전략에서의 g-formula 위험 추정치와 추정치들 사이의 차이, 위험 비 그리고 위험 오즈비를 계산한 결과를 표로 제공한다.

(2) 알고리즘의 1단계에 적합했던 모델의 계수 (coefficient), 표준 오차 (standard error), 모델의 평균 제곱근 오차(Root Mean Square Error; RMSE) 값 및 분산-공분산 행렬 (variance-covariance matrix)을 제공한다.

(3) (model_fits를 TRUE로 지정한 경우) 알고리즘의 1단계에서 적합된 모델(예: glm objects)을 확인할 수 있도록 자료를 제공한다.

(4) (sim_data_b를 TRUE로 지정한 경우) 각 지정된 개입 하에서 알고리즘의 2 단계에서 시뮬레이션 된 자료를 제공한다.

print, summary, plot, coef 및 vcov 함수를 이용하여 위에서 언급한 gfoRmula 함수의 결과물을 출력하는데 사용할 수 있다. plot 함수는 자연 경과가 적용되었을 때 비모수적 방법과 g-formula의 방법으로 각 시점에서 산출한 위험 추정치의 비교를 직관적인 그림을 통하여 보여줌으로써 g-formula에 사용된 각 교란 요인에 대한 모형이 잘 적합되었는지 확인이 가능하도록 한다.

부록 1

Ⅳ

BKMR과 g-formula의 장점과 단점

Ⅳ. BKMR과 g-formula의 장점과 단점

1. g-formula와 BKMR의 장점과 단점을 비교하는 목적

g-formula는 산업보건 역학연구에서 사용되는 인과추론 통계 방법론 중의 하나로 2개 이상의 복합노출에 의한 건강 영향을 인과적으로 추론할 수 있는 방법이다. BKMR은 최근 환경 역학연구에서 많이 사용되고 있는 복합노출의 건강 영향을 평가하는 통계 방법론 중 하나로, 여러 노출 변수에 대한 건강 영향을 평가할 수 있고 추정치의 다양한 시각화가 가능하며, 모델이 유연하다는 장점을 가지고 있다. 이 두 통계 방법론은 반복측정된 자료에서 복합노출에 의한 건강 영향 평가가 가능하다는 공통점을 가지고 있는 반면, 서로 다른 장점과 단점을 가지고 있다. 따라서, 이 두 통계 방법론의 장점과 단점을 비교하여, 반복 측정된 산업보건 역학 자료에서 복합노출로 인한 건강 영향을 평가하기에 적절한 통계 방법론에 대해 논의해보고자 한다.

2. g-formula의 장점과 단점

g-formula 장점은 다음과 같다. 먼저, g-formula는 반복측정된 자료를 분석할 때, 치료-교란 요인 되먹임의 존재를 반영할 수 있고, 건강근로자 생존 편향과 같이 산업보건 역학연구에서 발생할 수 있는 선택 편향을 효과적으로 통제할 수 있다. g-formula는 계산시간이 오래 걸리는 단점이 있는 것으로 알려져 있는데, BKMR에 비해서는 분석 속도가 월등히 빠르다. 또한, 연속형 노출 변수 외에 범주형 노출 변수에 대해서도 분석이 가능하다. 또한, marginal causal effect에 해당하는 위험 차이 (risk difference), 위험 비 (risk ratio) 그리고 오즈 비 (odds ratio)를 구할 수 있다. g-formula로 계산한 marginal causal effect의 추정치는 marginal structural model과

g-estimation 등 다른 인과추론 방법론을 통해 일관된 값이 나오는지 확인하여 분석 결과의 신뢰성을 일부 평가할 수 있다.

g-formula 단점은 다음과 같다. g-formula는 복합노출에 의한 건강 영향을 추정할 수 있으나, 반복측정된 자료에서 교호작용을 평가하는 접근법에 대해서는 아직 잘 알려져 있지 않다. 분석한 결과를 BKMR과 같이 다양하게 시각화하기 위해서는, 현재 많이 사용되는 R 패키지에 더하여 추가적인 작업이 필요하다. 선행 산업보건 역학연구에서 사용한 g-formula는 모수적 모형(parametric model)을 사용하였으며, BKMR보다 덜 유연한 모형이다.

3. BKMR의 장점과 단점

BKMR의 장점은 다음과 같다. BKMR은 반복측정된 자료를 분석할 수 있고, 많은 수의 노출 변수에 대해서 분석이 가능하다. BKMR은 g-formula와 달리 교호작용을 시각적으로 나타낼 수 있다. 또한, BKMR은 교호작용 외에도, 사후포함확률 (posterior inclusion probability; PIP), 단일 노출-반응 함수, 복합노출-반응 함수 등을 시각적으로 보여준다. BKMR은 모형 내에서 복합노출의 다양한 고차원 항 또는 교호작용 항을 반영하기 위해 kernel 행렬을 이용한 혼합 모형(mixed model)을 사용하기 때문에 모수적 모형을 사용한 g-formula보다 모형의 유연성을 확보할 수 있다. 이를 통해 많은 노출 변수와 결과 변수 사이의 관계를 유연하게 모형화할 수 있다.

BKMR의 단점은 다음과 같다. BKMR은 반복측정 자료를 분석할 때 발생하는 치료-교란 요인 되먹임이나 선택 편향을 효과적으로 통제하기 어렵다. 또한, 결과를 시각적으로 보여주어 직관적이지만, 노출량의 사분위 수를 개입하는 양 (intervention)으로 설정하기 때문에 해석이 쉽지 않다. 선행연구에 따르면, 노출 변수가 많아지고 데이터가 복잡해질수록, 다른 복합노출 통계 방법으로 분석한 결과와 비교했을 때 그 결과들이 일관되지 않다는 보고가 있으며, 따라서 노출 변수가 많고 자료의 형태가 복잡할 경우, 분석 결과의 신

뢰성이 낮을 수 있다.

4. g-formula와 BKMR의 장단점 정리

표 2는 g-formula와 BKMR의 장단점을 정리한 표이다.

표 2. g-formula와 BKMR의 장단점을 정리한 표

	g-formula	BKMR
공통점	• 복합물질에 근로자가 노출되었을 때, 두 방법 모두 건강 영향 평가 방법으로 사용이 가능함.	
장점	• 치료-교란 요인 되먹임과 경쟁 사건의 존재를 모형에 반영할 수 있음. • 건강근로자 편향을 통제할 수 있음. • Marginal causal effect에 해당하는 위험의 차이, 위험 비 그리고 위험 오즈 비를 모두 제공하기 때문에 인과효과를 다양하게 해석할 수 있음.	• 유해물질의 수가 많아도 복합노출에 의한 건강 영향 평가가 가능함. • 유해물질 사이의 교호작용의 효과를 시각적으로 표현해주는 함수가 개발되어 있어 별도의 시각화 과정이 요구되지 않음. • 커널을 이용하여 복합물질과 건강 결과 사이의 관계를 비모수적 함수를 통해 나타내기 때문에 다양한 고차원 항 또는 교호작용 항을 고려하여 모형화할 수 있음.
단점	• 모수적 모형을 사용하기 때문에 비모수적 모형을 사용하는 BKMR과 같이 다양한 고차원 항 또는 교호작용 항을 모형에 반영하는 것이 한계가 있음. • 인과 효과를 시각적으로 표현하기 위해서는 별도의 시각화 과정이 요구됨. • 유해물질 사이의 교호작용의 효과를 다양한 노출 수준에서 확인하기 어려움(다만, 모든 시점에 일정한 노출 수준을 가지는 복합노출의 경우에는 교호작용의 효과를 확인할 수 있음).	• 치료-교란 요인 되먹임으로 발생 가능한 건강근로자 편향을 통제하기 어려움. • 작은 크기의 자료에서도 분석 속도가 매우 느림. • 자료의 형태가 복잡해지면 복합노출의 건강영향을 평가하는 여러 통계 방법론들의 분석 결과 값들이 일관되지 않음.

부록 1 V

특수건강진단 자료를 활용한 복합노출 분석

V. 특수건강진단 자료를 활용한 복합노출 분석

1. 근로자 종적 자료에 대한 가설 및 가설의 인과 그래프

‘납과 카드뮴에 대한 장기적인 복합노출이 빈혈 발생의 위험을 높인다.’라는 가설과 ‘납과 크실렌에 대한 장기적인 복합노출이 빈혈 발생의 위험을 높인다.’라는 가설을 가설과 관련된 변수들 (빈혈 여부, 혈중 납 농도, 혈중 카드뮴 농도, 요중 메틸마뇨산 농도, 나이, 성별, 사업장 규모, 음주 여부, 흡연 상태, 비만도, 건강진단 결과에 대한 의사의 판정 결과 그리고 그로 인한 사후관리조치 결과)을 활용하여 인과 그래프로 표현하였다.

실무지침 내 노출 기준치에 대해 납의 경우, 직업병 요관찰자에 해당하는 혈중 납 농도 기준치는 30 μg/dL이며, 직업병 유소견자에 해당하는 혈중 납 농도 기준치는 40 μg/dL이다. 카드뮴의 경우, 5 μg/L이며, 크실렌의 경우 요중 메틸마뇨산 농도 1.5 mg/L이다.

따라서 납과 카드뮴 그리고 납과 크실렌에 대한 복합노출에 따른 빈혈 발생률을 산출하기 위해 혈중 납 농도 5, 10, 15, 20, 25, 30, 35, 40(단위: μg/dL), 혈중 카드뮴 농도 1, 2, 3, 4, 5 (단위:μg/L) 그리고 요중 메틸마뇨산 농도 0.25, 0.50, 0.75, 1.00, 1.25, 1.50 (단위: mg/L)을 개입 노출량 (hypothetical intervention)으로 설정하여 각 조합마다 특수건강검진 대상 근로자의 빈혈의 누적 발생률을 산출하고 등고선 그림을 통해 시각적으로 그 추이를 살펴보고자 한다. 그림 15는 납과 카드뮴 그리고 납과 크실렌에 관한 가설에서 공통적으로 포함되는 시간에 따라 변하는 납 노출과 빈혈 발생에 대한 인과 그래프이다. 그림 16는 납과 카드뮴의 복합노출에 대하여 세부적으로 그린 인과 그래프이고, 그림 17은 납과 크실렌의 복합노출에 대하여 세부적으로 그린 인과 그래프이다.

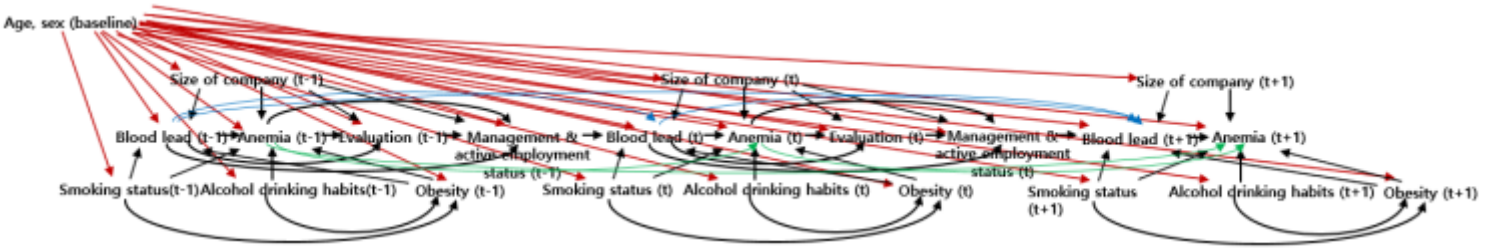


그림 15. 시간에 따라 변하는 납과 빈혈에 관한 가설을 표현한 인과 그래프

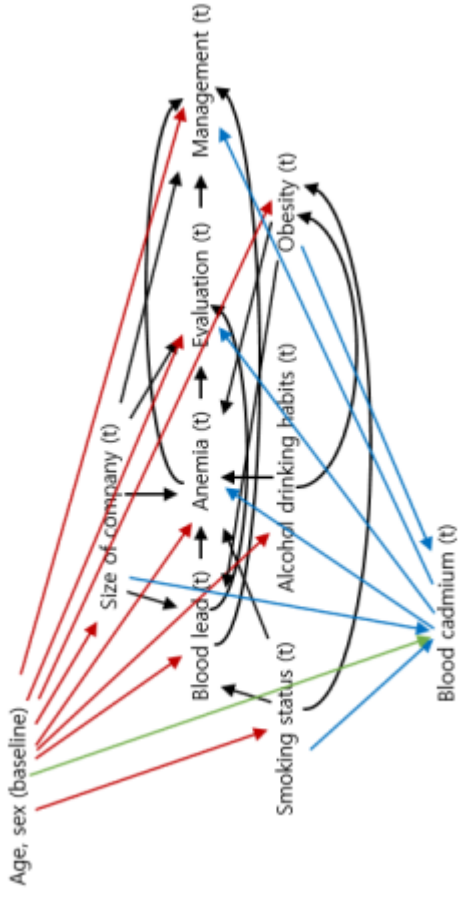


그림 16. 남과 카드뮴의 복합노출이 빈혈에 미치는 영향에 관한 가설을 표현한 인과 그래프

그림 17. 납과 크실렌의 복합노출이 빈혈에 미치는 영향에 관한 가설을 표현한 인과 그래프

2. g-formula를 사용한 산업보건 종적 자료 분석

1) 특수건강진단 자료의 특성

산업안전보건연구원의 직업건강연구실에서 수집하고 있는 근로자의 특수건강진단자료 중 2013년도부터 2019년까지 총 7년 동안 연 1회 이상 혈중 납 농도를 측정한 근로자의 수는 189,549명임. 이 자료는 반복 측정된 종적 자료로, 아이디 (개인식별변수), 검진 연도 (검진 순서), 성별, 나이, 사업장 규모, 흡연 상태, 음주 여부, 비만도, 혈중 납 농도, 혈중 카드뮴 농도, 요중 메틸마뇨산 농도, 혈중 헤모글로빈 수치, 사후관리조치 결과, 판정결과에 대한 정보를 포함하고 있다.

특수건강검진자료 중 혈중 납 농도와 혈중 카드뮴 농도가 동시에 측정된 근로자는 총 23,336명이었으며, 혈중 납 농도와 요중 메틸마뇨산 농도 (크실렌의 biomarker)가 동시에 측정된 근로자는 총 57,489명이다. 빈혈 여부는 남성의 경우, 혈중 헤모글로빈 수치가 13g/dL 보다 작은 경우, 여성의 경우 12g/dL 보다 작은 경우, 빈혈이 있다고 정의하였다. 흡연 상태는 과거 흡연자 또는 현재 흡연자인 경우 1, 흡연 경험이 전혀 없으면 0이라고 정의하였고, 음주 여부는 특수건강진단을 받을 당시 음주를 하였으면 1, 음주를 하지 않았다면 0으로 정의하였다. 비만도는 근로자의 몸무게를 근로자의 키(m)의 제곱으로 나눈 체질량 지수를 사용하였다. 혈중 납 농도, 혈중 카드뮴 농도 그리고 요중 메틸마뇨산 농도는 근로자의 특수건강진단 결과로 확인할 수 있는 값으로 각각 μg/dL, μg/L 그리고 mg/L의 단위로 표현된다. 판정결과는 특수건강진단 결과에 따른 의사의 판단결과를 기록한 내용이며, 총 6가지의 범주(D1; D2 또는 DN; C1; C2 또는 CN; U 또는 R; A)를 가진다. 사후관리조치 결과는 판정결과에 따라 결정되는 사후관리조치 내용이며, 총 3가지의 범주(작업 장소 변경 및 타 업무로 전환조치 등 노출이 중단되는 경우; 보호구 착용 등 노출수준이 낮아지는 경우; 사후관리가 필요 없는 경우)를 가진다.

나이는 근로자의 나이를 의미하며, 현재 자료에는 18세 이상의 근로자에 대한 자료만 포함되어있다. 자료에서 나타나는 성별은 남성과 여성 두 종류의 성만 있으며, 사업장 규모는 근로자가 근무하는 사업장에서 근무하는 총 근로자의 수를 나타낸다. 혈중 납 농도와 혈중 카드뮴 농도가 빈혈의 발생에 미치는 효과 그리고 혈중 납 농도와 요중 메틸마뇨산 농도가 빈혈의 발생에 미치는 효과를 알아보는 것이 연구 가설이기 때문에, 빈혈 여부가 결과 변수, 혈중 납 농도, 혈중 카드뮴 농도 그리고 요중 메틸마뇨산 농도가 노출 변수 그리고 그 외 나머지 변수는 교란 요인에 해당한다.

2) g-formula를 사용한 특수건강진단 자료의 분석 방법

특수건강진단 자료를 사용하여 두 가지 연구 가설 “장기간에 걸쳐 납과 카드뮴에 노출될 가능성이 있는 작업장에서 일한 근로자들을 대상으로, 7년 동안 특정 농도로 일정하게 혈중 납 농도와 혈중 카드뮴 농도가 고정되었을 때, 빈혈의 발생률에 얼마나 영향을 미치는가?” 그리고 “장기간에 걸쳐 납과 크실렌 (xylene)에 노출될 가능성이 있는 작업장에서 일한 근로자들을 대상으로, 7년 동안 특정 농도로 일정하게 혈중 납 농도와 요중 메틸마뇨산 농도가 고정되었을 때, 빈혈의 누적 발생률이 얼마나 될 것인가? 그리고 일반인구 집단에 비교하여 그 누적 발생률은 얼마나 높게 나타날 것인가?”에 대하여 알아보고자 한다. 두 연구 가설은 각각 두 종류의 유해물질(혈중 납 농도, 혈중 카드뮴 농도 / 혈중 납 농도, 요중 메틸마뇨산 농도)의 노출 수준의 조합에 따른 빈혈의 누적 발생률의 추이를 확인하고자 하며, 그에 따라 특수건강진단을 받은 근로자 집단에 각 유해물질의 특정 노출 수준에 대한 개입 (intervention)을 지정하고, 이를 등고선 그림을 통해 그 추이를 확인하고자 한다. 각 연구 가설에서 확인하고자 하는 효과를 두 유해물질에 대한 근로자의 혈중/요중 농도가 특정 농도로 고정되어 있을 때의 빈혈의 누적 발생률과 일반인구 집단의 노출 수준에 해당하는 혈중/요중 농도로 고정되었을 때 나타나는 빈혈의

누적 발생률의 비로 정의하였다. 예를 들어, 유해물질의 노출 수준에 대한 개입으로 “총 7년 동안 모든 근로자의 혈중 납 농도가 30 μg/dL, 혈중 카드뮴 농도가 5 μg/L로 유지되었을 때”를 지정할 경우, 이러한 노출 수준에 해당하는 개입을 받은 근로자들의 누적 발생률을, 대조군에 해당하는 일반인구 집단의 혈중 납 농도 (1.6 μg/dL[4])와 혈중 카드뮴 농도 (0.9187 μg/L[5])일 때의 빈혈의 발생률과 비교하여 유해물질의 노출 수준에 따른 빈혈의 누적 발생률을 산출하였다. 요중 메틸마뇨산 농도의 경우, 일반인구 집단에서의 요중 농도에 해당하는 0.234 mg/L[6]을 사용하였다.

g-formula를 적용하여 특수건강검진자료를 분석하기 위해서는 결과 변수, 노출 변수 그리고 교란 요인에 대한 모형이 필요하며, 모형에 포함하는 요인은 인과 그래프를 통하여 결정되었다. 혈중 납 농도와 혈중 카드뮴 농도의 조합 그리고 혈중 납 농도와 요중 메틸마뇨산 농도에 관한 연구 가설을 분석하기 위한 모형에 포함된 변수는 각각 표 3과 4에 기술되어 있다.

4) 환경부 국립환경과학원의 국민환경보건기초조사 DB에서 2017년 자료 내 성인의 혈중 납 기하평균 1.6 μg/dL을 참고하였음.
5) 보건복지부 질병관리청의 국민건강영양조사 자료에서 2016년과 2017년에 측정한 자료 내 만 19세 이상 성인의 혈중 카드뮴 기하 평균 0.9187 μg/L를 참고하였음.
6) 국가통계포털 KOSIS에서 2011년부터 2014년까지 수록된 자료 내 성인의 요중 메틸마뇨산 농도의 기하평균 0.234 mg/L를 참고하였음.

표 3. 특수건강검진자료에서 혈중 납 농도와 혈중 카드뮴 농도에 대한 연구 가설을 분석하기 위해 설정한 결과 변수, 노출 변수 그리고 교란 요인에 대한 모형과 모형에 포함된 변수

모형의 종류	분석을 위한 모형에 포함된 변수	
결과 변수에 대한 모형	시점 t에서의 빈혈 여부	혈중 납 농도 (시점 t), 혈중 카드뮴 농도 (시점 t), 흡연 상태 (시점 t), 음주 여부 (시점 t), 비만도 (시점 t), 혈중 납 농도 (시점 t-1), 혈중 카드뮴 농도 (시점 t-1), 사후관리조치 결과 (시점 t-1), 혈중 납 농도 (시점 t-2), 혈중 카드뮴 농도 (시점 t-2) 그리고 나이, 성별, 사업장 규모
노출 변수에 대한 모형	시점 t에서의 혈중 납 농도	음주 여부 (시점 t), 흡연 상태 (시점 t), 비만도 (시점 t), 혈중 납 농도 (시점 t-1), 사후관리 조치 결과 (시점 t-1), 혈중 납 농도 (시점 t-2), 혈중 납 농도 (시점 t-3) 그리고 나이, 성별, 사업장 규모, 검진 순서
	시점 t에서의 혈중 카드뮴 농도	음주 여부 (시점 t), 흡연 상태 (시점 t), 비만도 (시점 t), 혈중 카드뮴 농도 (시점 t-1), 사후관리조치 결과 (시점 t-1), 혈중 카드뮴 농도 (시점 t-2), 혈중 카드뮴 농도 (시점 t-3) 그리고 나이, 성별, 사업장 규모, 검진순서
교란 요인에 대한 모형	시점 t에서의 음주 여부	음주 여부 (시점 t-1) 그리고 나이, 성별, 검진순서
	시점 t에서의 흡연 상태	흡연 상태 (시점 t-1), 흡연 상태 (시점 t-1) 그리고 나이, 성별, 검진순서
	시점 t에서의 비만도	음주 여부 (시점 t), 흡연 상태 (시점 t), 비만도(시점 t-1) 그리고 나이, 성별, 검진순서
	시점 t에서의 판정결과	혈중 납 농도 (시점 t), 혈중 카드뮴 농도 (시점 t) 그리고 나이, 성별, 사업장 규모
	시점 t에서의 사후관리조치	판정결과 (시점 t), 혈중 납 농도 (시점 t), 혈중 카드뮴 농도 (시점 t) 그리고 나이, 성별, 사업장 규모

표 4. 특수건강검진자료에서 혈중 납 농도와 요중 메틸마뇨산 농도에 대한 연구 가설을 분석하기 위해 설정한 결과 변수, 노출 변수 그리고 교란 요인에 대한 모형과 모형에 포함된 변수

모형의 종류	분석을 위한 모형에 포함된 변수	
결과 변수에 대한 모형	시점 t에서의 빈혈 여부	혈중 납 농도 (시점 t), 요중 메틸마뇨산 농도 (시점 t), 흡연 상태 (시점 t), 음주 여부 (시점 t), 비만도 (시점 t), 혈중 납 농도 (시점 t-1), 요중 메틸마뇨산 농도 (시점 t-1), 사후관리조치 결과 (시점 t-1), 혈중 납 농도 (시점 t-2) 그리고 나이, 성별, 사업장 규모, 검진연도
노출 변수에 대한 모형	시점 t에서의 혈중 납 농도	음주 여부 (시점 t), 흡연 상태 (시점 t), 비만도 (시점 t), 혈중 납 농도 (시점 t-1), 사후관리조치 결과 (시점 t-1), 혈중 납 농도 (시점 t-2), 혈중 납 농도 (시점 t-3), 그리고 나이, 성별, 사업장 규모, 검진 순서
	시점 t에서의 요중 메틸마뇨산 농도	음주 여부 (시점 t), 흡연 상태 (시점 t), 비만도 (시점 t), 요중 메틸마뇨산 농도 (시점 t-1), 사후관리조치 결과 (시점 t-1) 그리고 나이, 성별, 사업장 규모, 검진순서
교란 요인에 대한 모형	시점 t에서의 음주 여부	음주 여부 (시점 t-1) 그리고 나이, 성별, 검진순서
	시점 t에서의 흡연 여부	흡연 상태 (시점 t-1), 흡연 상태 (시점 t-2) 그리고 나이, 성별, 검진순서
	시점 t에서의 비만도	음주 여부 (시점 t), 흡연 여부 (시점 t), 비만도(시점 t-1) 그리고 나이, 성별, 검진순서
	시점 t에서의 판정결과	혈중 납 농도 (시점 t), 요중 메틸마뇨산 농도 (시점 t) 그리고 나이, 성별, 사업장 규모
	시점 t에서의 사후관리조치	판정결과 (시점 t), 혈중 납 농도 (시점 t), 요중 메틸마뇨산 농도 (시점 t) 그리고 나이, 성별, 사업장 규모, 검진 순서

3) g-formula를 사용한 특수건강진단 자료의 분석 결과

표 5는 혈중 납 농도와 혈중 카드뮴 농도의 조합에 따른 빈혈의 누적 발생률을 일반인구집단에서 측정되는 혈중 농도인 혈중 납 농도 (1.6 μg/dL)와 혈중 카드뮴 농도 (0.9187 μg/L)일 때의 빈혈의 누적 발생률로 나누어 구한 위험 비를 기술한 표이며, 95% 신뢰구간은 bootstrap 방법을 사용하여 산출하였다. 표 5에서 혈중 카드뮴 농도가 고정되어있을 때, 혈중 납 농도가 증가함에 따라 위험 비가 증가하는 것을 확인할 수 있었으며, 마찬가지로 혈중 납 농도가 고정되어있는 경우, 혈중 카드뮴 농도가 증가함에 따라 위험 비가 증가하는 것을 확인할 수 있다. 저 농도에서의 혈중 납 농도와 혈중 카드뮴의 조합의 일부에서 빈혈에 대한 위험 비의 95% 신뢰구간들 중 일부가 1을 포함하고 있지만, 고 농도를 포함하여 그 외의 혈중 납 농도와 혈중 카드뮴 농도의 조합에서 95% 신뢰구간의 왼쪽 경계 값이 모두 1보다 크며, g-formula가 유의한 결과를 제공하고 있음을 알 수 있다

더불어, 표 6은 혈중 납 농도와 요중 메틸마뇨산 농도의 조합에 따른 빈혈의 누적 발생률을 일반인구집단에서 측정되는 혈중 납 농도와 요중 메틸마뇨산 농도 (0.234 mg/L)일 때의 빈혈의 누적 발생률로 나누어 산출한 위험 비를 기술한 표이다. 신뢰구간은 표 5와 마찬가지로 bootstrap 방법을 이용하여 계산하였다. 표 6을 보면 요중 메틸마뇨산 농도가 고정되어있는 경우, 혈중 납 농도가 증가함에 따라 위험 비가 증가하는 것을 확인할 수 있다. 하지만 혈중 납 농도가 고정되어있는 경우, 요중 메틸마뇨산 농도가 증가함에 따라 위험 비가 소폭 감소하는 것을 확인할 수 있다. 또한, 혈중 납 농도와 혈중 카드뮴 농도에서의 결과가 비슷하게 혈중 납 농도와 요중 메틸마뇨산 농도의 조합에서의 결과에서도 혈중 납 농도가 옅을 때의 경우, 95% 신뢰구간이 1을 포함하지만, 혈중 납 농도가 높아지는 경우 g-formula가 95% 신뢰구간의 왼쪽 경계 값으로 1보다 큰 값을 제공함으로써 유의한 결과를 제공한다는 것을 알 수 있다.

그림 18은 혈중 납 농도와 혈중 카드뮴 농도의 조합에 따른 빈혈의 누적 발생률의 위험 비를 시각적으로 확인하기 위한 등고선 그림으로, 표 5에서 확인한 것과 같이 혈중 납 농도와 혈중 카드뮴 농도가 모두 증가함에 따라 빈혈의 발생률의 위험 비 또한 증가하는 것을 확인할 수 있다. 그림 19는 혈중 납 농도와 요중 메틸마뇨산 농도의 조합에 따른 빈혈의 누적 발생률의 위험 비를 시각적으로 표현하기 위한 등고선 그림으로, 표 6에서 확인한 것과 같이 혈중 납 농도가 증가함에 따라 위험 비가 증가하는 것을 시각적으로 확인할 수 있다. 위의 결과를 위해 사용된 g-formula가 특수건강검진자료를 적절히 설명하고 있는지 확인하기 위해 자료에서 나타나는 근로자들의 혈중 납 농도와 혈중 카드뮴 농도 또는 요중 메틸마뇨산 농도 (자연 경과; natural course)에서 적합된 g-formula를 사용하여 산출한 빈혈에 대한 누적 발생률을 자료에서 나타나는 유해물질의 노출 수준에서의 빈혈에 대한 누적 발생률과 비교하였으며, 그림 19는 그 결과를 제공한다. 그림 20에서 자연 경과일 때의 g-formula 적합 결과를 'parametric g-formula estimates'라 표현하였고, 자료로부터 직접 산출한 빈혈의 누적 발생률을 'nonparametric estimates'라 표현하였다. 혈중 납 농도와 혈중 카드뮴 농도 그리고 혈중 납 농도와 요중 메틸 마뇨산 농도 각각의 자연 경과에서 산출한 빈혈의 누적 발생률에 대한 95% 신뢰구간이 자료로부터 직접 산출한 빈혈에 대한 누적 발생률을 포함하므로 g-formula가 자료를 올바르게 적합한다는 것을 알 수 있다.

표 5. 혈중 납 농도와 혈중 카드뮴 농도에 따른 빈혈의 발생률에 대한 위험 비를 기술한 표. 혈중 납 농도에 대하여 5 μg/dL 단위 별로 굵은 글씨 및 회색 칸으로 표의 셀을 표시하였다.

혈중 납 농도 (μg/dL)	혈중 카드뮴 농도 (μg/L)	위험 비(Risk ratio)	위험 비의 95% 신뢰구간	
			왼쪽 경계 값	오른쪽 경계 값
1.6	0.9187	Reference (1.0000)	-	-
5	**1**	**0.8811**	**0.7884**	**0.9854**
5	2	1.0453	0.9113	1.2160
5	3	1.2619	1.0384	1.5628
5	4	1.5321	1.1450	2.0900
5	5	1.8608	1.2495	2.7630
10	**1**	**0.8999**	**0.7428**	**1.0732**
10	2	1.0676	0.8621	1.2967
10	3	1.2892	0.9912	1.6799
10	4	1.5658	1.1543	2.1672
10	5	1.9010	1.2091	2.9435
15	**1**	**1.0138**	**0.7777**	**1.2560**
15	2	1.1991	0.8912	1.4774
15	3	1.4436	1.0230	1.8725
15	4	1.7463	1.1672	2.4428
15	5	2.1106	1.3490	3.2974
20	**1**	**1.1971**	**0.8372**	**1.5493**
20	2	1.4096	0.9637	1.8312
20	3	1.6879	1.1031	2.1498
20	4	2.0291	1.2316	2.8534
20	5	2.4356	1.4685	3.6415
25	1	1.4504	0.9333	2.0399
25	2	1.6979	1.0665	2.3277
25	3	2.0187	1.2843	2.7492
25	4	2.4080	1.4015	3.4354
25	5	2.8660	1.5829	4.2732
30	**1**	**1.7814**	**1.0544**	**2.6648**
30	2	2.0705	1.1997	3.0672
30	3	2.4417	1.4351	3.5540
30	4	2.8864	1.6837	4.1293
30	5	3.4062	1.8051	5.2302
35	**1**	**2.2006**	**1.2159**	**3.5577**
35	2	2.5371	1.3723	3.9055
35	3	2.9656	1.6347	4.5729
35	4	3.4811	1.9632	5.2125
35	5	4.0892	2.1902	6.2649
40	**1**	**2.7229**	**1.4304**	**4.6238**
40	2	3.1172	1.5998	5.0553
40	3	3.6244	1.8106	5.6505
40	4	4.2449	2.2264	6.4824
40	5	4.9756	2.5131	7.5492

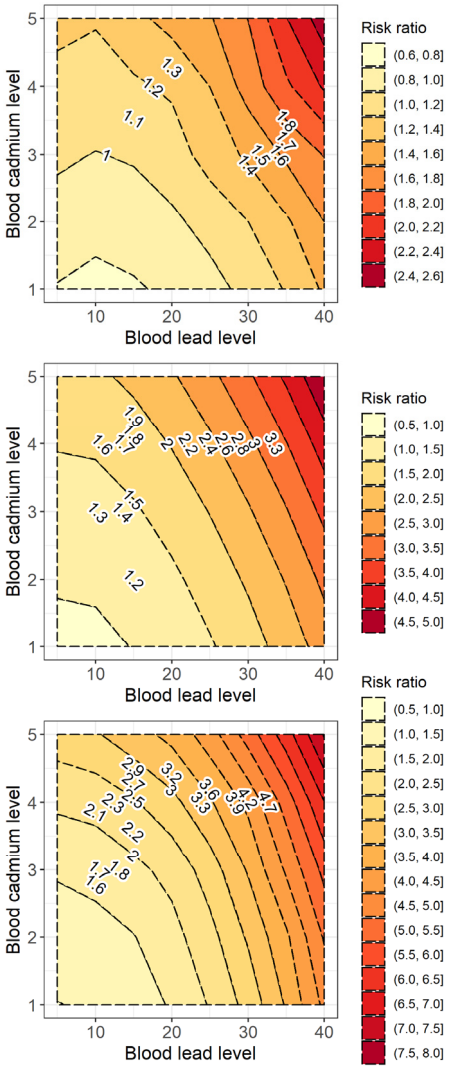


그림 18. 혈중 납 농도 (blood lead level)와 혈중 카드뮴 농도 (blood cadmium level)에 따른 빈혈의 발생률에 대한 위험도 비를 표현한 등고선 그림. 상단의 그림은 빈혈의 누적 발생률 추정치에 대한 95% 신뢰구간의 하한을, 가운데 그림은 빈혈의 누적 발생률에 대한 추정치를 그리고 하단의 그림은 빈혈의 누적 발생률 추정치에 대한 95% 신뢰구간의 상한을 그린 등고선 그림이다.

표 6. 혈중 납 농도와 요중 메틸마뇨산 농도에 따른 빈혈의 발생률에 대한 위험 비를 기술한 표. 혈중 납 농도에 대하여 5 μg/dL 단위 별로 굵은 글씨 및 회색 칸으로 표의 셀을 표시하였다.

혈중 납 농도 (μg/dL)	요중 메틸마뇨산 농도 (μg/dL)	위험 비(Risk ratio)	위험 비의 95% 신뢰구간	
			왼쪽 경계 값	오른쪽 경계 값
1.6	0.2340	Reference(1.0000)	-	-
5	**1**	**0.8879**	**0.7946**	**0.9727**
5	2	0.8820	0.7971	0.9542
5	3	0.8762	0.7980	0.9541
5	4	0.8705	0.7845	0.9640
5	5	0.8648	0.7670	0.9851
10	**1**	**0.9663**	**0.8045**	**1.1346**
10	2	0.9599	0.8040	1.1218
10	3	0.9535	0.8008	1.1178
10	4	0.9472	0.7930	1.1257
10	5	0.9410	0.7854	1.1343
15	**1**	**1.1812**	**0.8772**	**1.5509**
15	2	1.1732	0.8858	1.5277
15	3	1.1652	0.8986	1.5050
15	4	1.1574	0.8908	1.5003
15	5	1.1496	0.8623	1.5401
20	**1**	**1.5309**	**1.0410**	**2.2387**
20	2	1.5203	1.0482	2.2068
20	3	1.5098	1.0336	2.1592
20	4	1.4995	1.0075	2.1217
20	5	1.4892	0.9742	2.1180
25	**1**	**2.0506**	**1.2097**	**3.3032**
25	2	2.0364	1.2120	3.2104
25	3	2.0223	1.1951	3.1731
25	4	2.0083	1.1575	3.1884
25	5	1.9944	1.1277	3.1883
30	**1**	**2.7911**	**1.4447**	**4.9066**
30	2	2.7721	1.4219	4.8753
30	3	2.7532	1.4105	4.8242
30	4	2.7345	1.3888	4.8444
30	5	2.7159	1.3757	4.7998
35	**1**	**3.8116**	**1.7614**	**7.1765**
35	2	3.7867	1.7548	7.1367
35	3	3.7619	1.7455	7.1163
35	4	3.7373	1.7358	7.1083
35	5	3.7129	1.7181	7.0711
40	**1**	**5.1743**	**2.1391**	**10.2182**
40	2	5.1425	2.1214	10.1321
40	3	5.1095	2.1020	10.0472
40	4	5.0792	2.0814	10.0226
40	5	5.0483	2.0612	10.0168

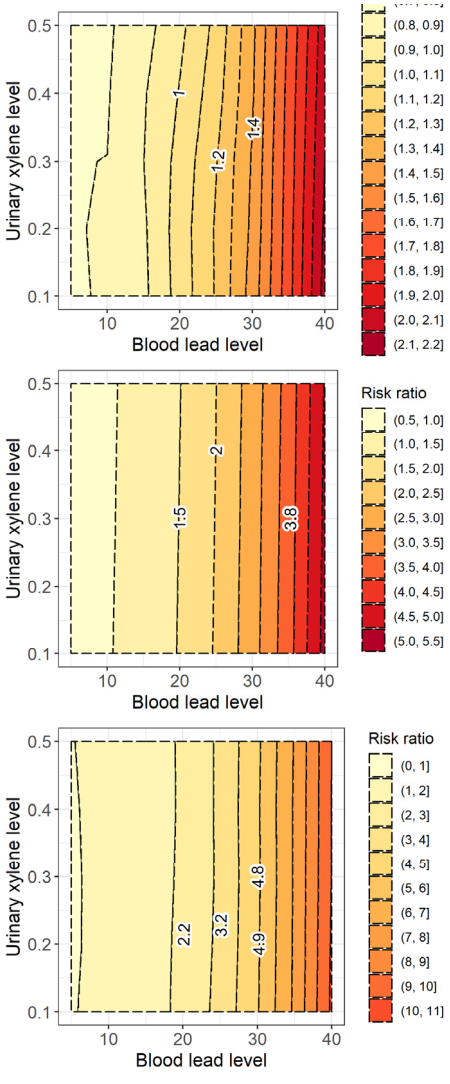


그림 19. 혈중 납 농도 (blood lead level)와 요중 메틸마뇨산 농도 (urinary xylene level)에 따른 빈혈의 발생률에 대한 위험도 비를 표현한 등고선 그림. 상단의 그림은 빈혈의 누적 발생률 추정치에 대한 95% 신뢰구간의 하한을, 가운데 그림은 빈혈의 누적 발생률에 대한 추정치를 그리고 하단의 그림은 빈혈의 누적 발생률 추정치에 대한 95% 신뢰구간의 상한을 그린 등고선 그림이다.

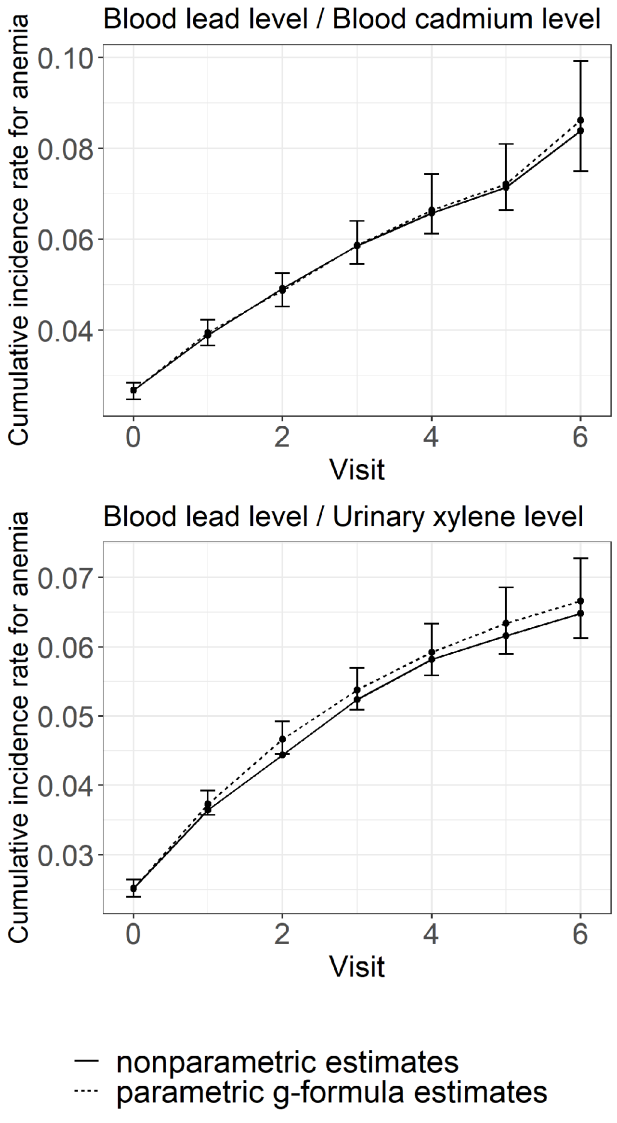


그림 20. 혈중 납 농도 (blood lead level)와 혈중 카드뮴 농도 (blood cadmium level)에 대한 빈혈의 누적 발생률을 산출하기 위해 적합한 g-formula의 자연 경과 (natural course)에서의 적합 결과 (위). 혈중 납 농도 (blood lead level)와 요중 메틸마뇨산 농도 (urinary xylene level)에 대한 빈혈의 누적 발생률을 산출하기 위해 적합한 g-formula의 자연 경과 (natural course)에서의 적합 결과 (아래).

3. BKMR을 사용한 산업보건 종적 자료 분석

1) BKMR을 사용한 특수건강진단 자료의 분석 방법

연구가설을 분석하기 위해 BKMR을 적합하였다. 다만, g-formula의 분석에 사용된 약 20,000명의 근로자 자료를 모두 사용하게 되면 BKMR의 결과를 확인하기까지 많은 시간이 소요되기 때문에 약 20,000명의 근로자 중 임의로 2,000명을 무작위 추출하여 현재 분석에 사용하였다. 또한, 반복측정 자료인 특수건강검진자료의 특징을 반영하기 위해 BKMR을 적합할 때, 근로자의 아이디를 입력하였다. 보정변수로는 나이, 성별, 음주 유무, 흡연 상태, 체질량지수, 사업장 규모르 사용하였다.

2) BKMR을 사용한 특수건강진단 자료의 분석 결과

그림 21은 BKMR을 적합한 후, 혈중 납 농도(z1) 그리고 혈중 카드뮴 농도(z2) 각각에 대하여 h(·)의 추정치를 그린 그래프이다. 혈중 납 농도의 경우, 혈중 납 농도가 증가할수록 h(·)의 값이 작아지는, 즉 혈중 납 농도가 증가할수록 빈혈의 발생률이 작아지는 경향이 나타났으며, 사실상 혈중 납 농도가 15 μg/dL 이상인 경우에는 h(·)의 추정치에 대한 95% 신뢰구간이 0을 포함하여 BKMR은 유의하지 않은 결과를 제공한다. 혈중 카드뮴 농도의 경우, 혈중 카드뮴 농도가 증가할수록 빈혈의 발생률이 증가하는 추세를 보였다. 하지만 혈중 카드뮴 농도의 경우 또한, 혈중 카드뮴 농도가 약 3.3 μg/L 이상에서 h(·)의 추정치에 대한 95% 신뢰구간이 0을 포함하여 BKMR이 유의하지 않은 결과를 제공한다. **혈중 납 농도와 혈중 카드뮴 농도에 대해 고 농도에서 유의한 결과를 제공한 g-formula와 달리 BKMR은 사실상 반대의 결과를 제공하였는데, 이는 BKMR이 g-formula와 달리 건강근로자 생존 편향을 고려하지 못하였기 때문에 발생하는 차이로 이해할 수 있다.**

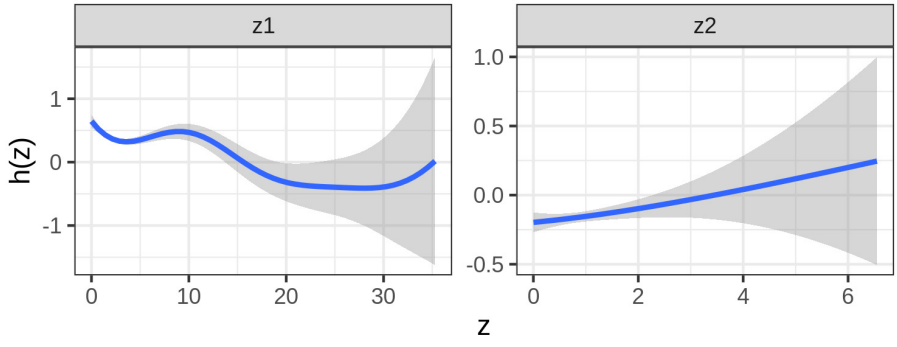


그림 21. 혈중 납 농도와 혈중 카드뮴 농도에 따른 h(·)의 추정치에 대한 그래프

그림 22에서 근로자의 혈중 납 농도와 혈중 카드뮴 농도를 모두 특정 노출 수준 (percentile)인 경우에서의 h(·)의 추정치의 추세를 살펴볼 수 있다. 노출 수준이 중앙값일 때와 비교하여 노출 수준이 증가할수록 h(·)의 추정치가 작아지고 있다는 것을 확인할 수 있고, 이는 중금속에 대해 혈중 농도가 증가할수록 빈혈이 발생할 확률이 감소한다는 의미이며, g-formula와는 상반된 결과를 제공한다.

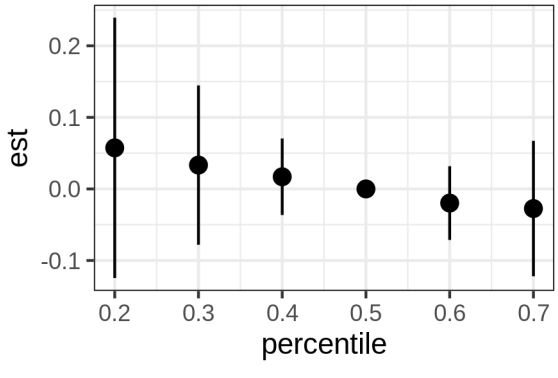


그림 22. 혈중 납 농도와 혈중 카드뮴 농도를 각 percentile에 고정시킨 경우의 h(·)의 추정치에 대한 그래프

그림 23은 빈혈의 발생과 관련하여 혈중 납 농도와 혈중 카드뮴 농도의 교호작용을 살펴볼 수 있는 그래프이며, 그림 29의 세로축에서 z1을 보면 혈중 카드뮴 농도가 25, 50, 75 percentile로 고정되어있을 때, 혈중 납 농도에 따른 h(·)의 추정치와 95% 신뢰구간을 보여준다. 서로 다른 혈중 카드뮴 농도에서 혈중 납 농도에 따른 h(·)의 추정치들의 차이가 크지 않고, 반대로 서로 다른 혈중 납 농도에서 혈중 카드뮴 농도에 따른 h(·)의 추정치들의 차이가 크지 않으므로 혈중 납 농도와 혈중 카드뮴 농도 사이의 교호작용이 유의하지 않을 수 있음을 시각적으로 확인할 수 있다.

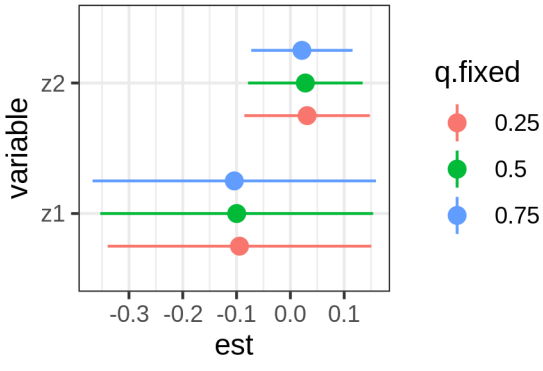


그림 23. 혈중 납 농도와 혈중 카드뮴 농도 사이의 교호작용을 나타낸 그래프

이를 혈중 카드뮴 농도가 각각 75, 25 percentile로 고정되었을 때, 혈중 납 농도가 75 percentile에서 25 percentile로 감소하였을 때의 h(·)의 추정치의 변화에 대한 차이를 살펴봄으로써 (혈중 카드뮴 농도의 경우, 반대로 혈중 납 농도가 고정되어있을 때, 혈중 카드뮴 농도에 대한 추정치의 변화에 대한 차이) 교호작용이 존재하는지 수치적으로 확인할 수 있다. 표 16을 보면 교호작용의 추정치는 -0.0097, 표준 오차는 0.0297로 95% 신뢰구간을 계산하게 되면 신뢰구간이 0을 포함하는 것을 확인할 수 있다.

표 16. 혈중 납 농도와 혈중 카드뮴 농도에 대한 교호작용의 추정치 및 표준 오차

변수	추정치	표준 오차
혈중 납 농도	-0.0097	0.0297
혈중 카드뮴 농도	-0.0097	0.0297

부록 1의 참고문헌

이슬비. 임신 중 복합 환경유해물질 노출이 6 개월 영유아 아토피 피부염 발생에 미치는 영향. 2019.

예신희, 이상길, 이지혜, 이경은, 성정민, 김민수. 저농도 복합유해물질 노출과 혈액검사 이상 관련성 탐색 연구. 산업안전보건연구원. 2020.

예신희, 이경은, 성정민, 박동준, 이우주. 직업병 인과추론 가이드라인 및 통계분석법 개발(1): g methods 국문 가이드라인 개발. 산업안전보건연구원. 2021.

Bobb JF. “Introduction to Bayesian kernel machine regression and the bkmr R package”, GitHub, 2017년 3월 24일 작성. 2022년 7월 10일 접속. URL https://jenfb.github.io/bkmr/overview.html#estimated_posterior_inclusion_probabilities.

Bobb JF, Valeri L, Claus Henn B, Christiani DC, Wright RO, Mazumdar M, Godleski JJ, Coull BA. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics. 2015 Jul;16(3):493-508.

Bobb JF, Claus Henn B, Valeri L, Coull BA. Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. Environ Health. 2018 Aug 20;17(1):67.

Keil AP, Richardson DB. Reassessing the Link between Airborne

Arsenic Exposure among Anaconda Copper Smelter Workers and Multiple Causes of Death Using the Parametric g-Formula. Environ Health Perspect. 2017 Apr;125(4):608-614. doi: 10.1289/EHP438. Epub 2016 Aug 19. PMID: 27539918; PMCID: PMC5381993.

Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics. 2007 Dec;63(4):1079-88. doi: 10.1111/j.1541-0420.2007.00799.x. PMID: 18078480; PMCID: PMC2665800.

Neophytou AM, Costello S, Picciotto S, Brown DM, Attfield MD, Blair A, Lubin JH, Stewart PA, Vermeulen R, Silverman DT, Eisen EA. Diesel Exhaust, Respirable Dust, and Ischemic Heart Disease: An Application of the Parametric g-formula. Epidemiology. 2019 Mar;30(2):177-185.

Taubman SL, Robins JM, Mittleman MA, Hernán MA. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. Int J Epidemiol. 2009 Dec;38(6):1599-611.

Valeri L, Mazumdar MM, Bobb JF, Claus Henn B, Rodrigues E, Sharif OI, Kile ML, Quamruzzaman Q, Afroz S, Golam M, Amarasiriwardena C. The joint effect of prenatal exposure to metal mixtures on neurodevelopmental outcomes at 20–40 months of age: evidence from rural Bangladesh. Environmental health perspectives. 2017 Jun 26;125(6):067015.

부록 2

직업병 인과추론 가이드라인: g-formula 국문 가이드라인 수정 내용

부록 2 I

인과추론 용어의 정리

Ⅰ. 인과추론 용어의 정리

1. 인과효과의 정의

1) 개인에 대한 인과효과 (individual causal effects)

개인에서의 인과효과를 간단한 예시를 통해 설명하고자 한다. 심장 이식을 기다리는 철수라는 환자가 있다고 상상해보자. 1월 1일에 철수는 새로운 심장을 이식받았고, 그로부터 5일 후 사망하였다. 이 사례에서 철수는 심장 이식으로 인해 사망한 것일까?

이러한 질문에 답변하기 위해서는 철수가 1월 1일에 심장 이식을 받지 않은 상태에서 철수의 5일 후의 상태를 알아야 한다. 예를 들어, 철수가 심장 이식을 받지 않았더라면, 5일 후에 살아있을 것이라는 사실을 우리가 알고 있다면 사람들은 철수가 심장 이식으로 인해 사망하였다는 것에 동의할 것이다. 즉, 심장 이식은 철수의 5일 후 생존에 인과적 영향을 미쳤다.

또 다른 환자인 영희의 사례를 살펴보자. 영희는 1월 1일에 심장 이식을 받았고, 5일 후에 살아남았다. 영희가 1월 1일에 심장 이식을 받지 않았더라면, 철수와는 달리 5일이 지난 후에도 여전히 살아있을 것이라는 사실을 우리가 알 수 있다고 해보자. 즉, 영희는 심장 이식을 받아도 5일 후에 살아남았고, 심장 이식을 받지 않아도 5일 후에 살아남았다. 따라서 심장 이식은 영희의 5일 후 생존에 인과적 영향을 미치지 않았다.

위의 두 사례는 인간이 인과적 효과에 대해 추론하는 방식을 보여준다. 우리는 행동 A를 했을 때의 결과와 행동 A를 하지 않았을 때의 결과를 (보통 생각으로만) 비교하고, 두 결과가 서로 다르면 행동 A는 결과에 인과적 영향 또는 예방적 영향을 미친다고 말한다. 반면 두 결과가 서로 같은 경우 행동 A는

결과에 인과적 영향을 미치지 않는다고 말한다. 역학자, 통계학자, 경제학자 및 기타 사회 과학자들은 이러한 행동 A를 개입, 노출 또는 치료라고 한다.

인과관계에 대한 직관을 수학적 분석 및 통계적 분석에 적용할 수 있도록 하기 위한 몇 가지 표기법을 소개한다. $Y^{a=1}$을 치료 값이 1인 경우 관찰되는 결과라고 하고, $Y^{a=0}$을 치료 값이 0인 경우 관찰되는 결과라 정의할 수 있다. 이때, **변수 $Y^{a=1}$와 $Y^{a=0}$는 잠재적 결과(potential outcomes) 또는 반사실적 결과(counterfactual outcome)라고 부른다. 일부 연구자들은 받은 치료에 따라 이 두 가지 결과 중 하나가 잠재적으로 관찰될 수 있음을 강조하기 위해 "잠재적 결과(potential outcome)"이라는 용어를 선호한다. 다른 연구자들은 이러한 결과가 실제로 발생하지 않을 수 있는 상황 (즉, 사실과 반대되는 상황)을 나타내는 것을 강조하기 위해 "반사실적 결과(counterfactual outcome)"라는 용어를 선호한다.**

현재 예제에서 철수를 인덱스 i로, 영희를 인덱스 j로 표현하면, 철수는 심장을 이식받았을 때(a=1) 사망하였고, 이식받지 않은 경우(a=0) 생존하였기 때문에 $Y_i^{a=1}=1$와 $Y_i^{a=0}=0$로 표현할 수 있다. 반면, 영희는 심장 이식과 무관하게 생존하였기 때문에 $Y_j^{a=1}=0$와 $Y_j^{a=0}=0$로 표현할 수 있다.

이제 위의 정보들을 바탕으로 '개인에 대한 인과효과'를 공식적으로 정의할 수 있다. $Y^{a=1} \neq Y^{a=0}$이면 개인에 대해 인과효과를 가진다고 할 수 있고, 치료 변수 A는 개인의 결과 변수 Y에 대해 인과효과를 가진다고 할 수 있다. 철수의 경우 심장 이식에 따른 생존 여부가 달라지므로 ($Y_i^{a=1}=1 \neq 0=Y_i^{a=0}$), 심장 이식 여부가 철수의 생존 여부에 인과적으로 영향을 미쳤다고 할 수 있는 반면에 영희의 경우 심장 이식 유무에 따라 생존 여부가 달라지지 않았으므로 ($Y_j^{a=1}=0=Y_j^{a=0}$), 심장 이식 여부는 영희의 생존 여부에 인과적으로 영향을 미치지 않았다.

앞서 말한 바와 같이 개인에 대한 인과효과는 잠재적 결과 값들의 대조(contrast)로 정의할 수 있다. 하지만, 현실에서는 각 개인에 대해 잠재적 결

과들 중 하나만 실제로 관찰할 수 있는데, 이 관찰된 결과는 개인이 실제로 경험한 치료 값의 결과를 의미한다. 즉, 그 외 모든 잠재적 결과들은 관찰할 수 없고, 이렇게 확인할 수 없는 나머지 잠재적 결과들로 인해 개인의 인과효과는 자료로부터 직접 계산이 불가능하다. 따라서, 개인의 인과효과는 관찰된 데이터의 함수로 표현할 수 없다.

2) 평균 인과효과

앞선 절에서 개인에 대한 인과효과를 정의하기 위해서는 다음의 세 가지 정보가 필요했다: (i) 관심 결과, (ii) 비교할 치료들(a=1과 a=0) 그리고 (iii) 비교될 치료에 대한 잠재적 결과들($Y^{a=1}$와 $Y^{a=0}$). (iii)의 값 중 일부를 자료로부터 확인할 수 없어 개인에 대한 인과효과를 직접 계산하는 것은 불가능하므로 현실적인 대안으로서 집단에 대한 인과효과인 '집단의 평균 인과효과'에 주목하고자 한다.

우리의 관심 집단으로 20명으로 구성된 철수의 대가족을 생각해보자. 표 7은 철수의 가족 20명 모두에 대해 심장 이식을 받은 경우(a=1)와 심장 이식을 받지 않은 경우(a=0)에 대한 잠재적 결과들을 보여준다. 예를 들어, 3번째 열은 각 개인이 심장 이식을 받았을 경우 관찰되는 결과들 $Y^{a=1}$이다. 그리고 이 열을 살펴보면, 철수를 포함한 가족 20명 모두 심장 이식을 받았더라면 가족 중 절반 (20명 중 10명)이 사망했을 것이다. 즉, 인구집단의 모든 개인들이 심장 이식을 받은 경우(a=1), 사망하는 비율은 $\Pr[Y^{a=1}=1]=\frac{10}{20}=0.5$이다. 마찬가지로, 표 7의 두 번째 열로부터 가족 20명 모두 심장 이식을 받지 않았다면 가족 중 절반 (20명 중 10명)이 사망했을 것이라고 결론을 내릴 수 있다. 즉, 인구집단 내 모든 개인이 심장 이식을 받지 않았을 때, 사망하는 개인들의 비율은 $\Pr[Y^{a=0}=1]=\frac{10}{20}=0.5$이다. 우리는 심장 이식 여부에 따른 잠재적 결과(사망)가 발생한 사건의 수(10)를 센 뒤 전체 개인의 수 (20)로 나누

어 0.5로 계산하였다. 이는 인구집단 내 모든 개인에 대한 잠재적 결과가 주어져있을 때, 잠재적 결과에 대한 기댓값을 계산하는 것과 같다.

표 7. 인과효과 설명을 위한 예시 자료

이름	$Y^{a=0}$	$Y^{a=1}$
서연	0	1
민준	1	0
유진	0	0
동현	0	0
지혜	0	0
지훈	1	0
영희	0	0
철수	0	1
지영	1	1
정훈	1	0
미영	0	1
성호	1	1
미경	1	1
영수	0	1
영숙	0	1
지훈	0	1
순자	1	1
현우	1	0
영자	1	0
성민	1	0

이제 우리는 인구집단의 평균 인과효과에 대한 공식적인 정의를 제공할 준비가 되었다.

인구집단의 평균 인과효과에 대한 정의는 다음과 같다: 결과 변수 Y에 대

한 치료 변수 A의 평균 인과효과는 잠재적 결과의 기댓값들 $E[Y^{a=1}]$, $E[Y^{a=0}]$의 대조로 정의할 수 있다. 여기서 E는 기댓값을 정의하기 위해 사용되었으며, 위의 예제에서 Y가 이분형 변수이기 때문에 $E[Y^{a=1}] = \Pr[Y^{a=1}=1]$, $E[Y^{a=0}] = \Pr[Y^{a=0}=1]$로 표현할 수 있다. 이 정의에 따라 관심 인구집단 '철수의 대가족'에서 '심장 이식'(치료 변수 A)은 '사망 여부'(결과 변수 Y)에 평균 인과효과를 가지지 않는다. 왜냐하면, 심장 이식을 받은 경우 사망하는 확률 $\Pr[Y^{a=1}=1]$과 심장 이식을 받지 않은 경우 사망하는 확률 $\Pr[Y^{a=0}=1]$이 모두 0.5로 같아 두 확률 값의 차이 또는 비가 0 또는 1이기 때문이다.

평균 인과효과가 없다는 것이 개인에 대한 인과 효과가 없다는 것을 암시하지는 않는다. 개인에 대한 인과 효과에서 설명하였듯이 철수에게 사망 여부에 대한 심장 이식의 개인에 대한 효과가 있음을 설명하였다. 추가적으로 표 7에서 철수 외에도 11명의 사람에 대하여 잠재적 결과들의 값 $Y^{a=1}$와 $Y^{a=0}$이 서로 다른 것을 확인할 수 있기 때문이다. 이 12명 중, 철수를 포함한 6명은 치료를 받아서 사망하였고 ($Y^{a=1} - Y^{a=0} = 1$), 나머지 6명은 치료를 받아서 살 수 있었다 ($Y^{a=1} - Y^{a=0} = -1$). 하지만, 인구집단의 모든 개인에게 인과효과가 없는 경우를 표현하는 용어가 따로 존재하는데, 인구집단 내 모든 개인에서 $Y^{a=1} = Y^{a=0}$일 때, sharp causal 귀무가설 (null hypothesis)이 참이라고 한다. Sharp causal 귀무가설은 모든 개인에게서 인과 효과가 없음을 의미하기 때문에 평균 인과효과가 없다는 귀무가설을 내포한다.

이번 장에서는 잠재적 결과를 사용하여 개인에 대한 인과효과와 집단의 평균 인과효과를 산출하고 그 값을 해석하는 것에 대하여 설명하였다. 다음 장에서는 데이터로부터 집단의 평균 인과효과는 추론할 수 있다는 것을 설명하려고 한다. 또한, 앞으로 우리는 '집단의 평균 인과효과'를 단순히 '인과효과 (causal effect)'라고 얘기하고, 평균 인과효과가 없다는 귀무가설을 '인과 귀무가설 (causal null hypothesis)'이라고 하고자 한다.

3) 인과성과 연관성 (causation and association)

사실 실제 연구에서 볼 수 있는 데이터는 표 7과 다르다. 예를 들어, 우리는 실제로 철수가 심장 이식을 받았을 때의 결과 $Y^{a=1}$와 철수가 심장 이식을 받지 않았을 때의 결과 $Y^{a=0}$을 모두 관찰할 수 없다. 현실에는 철수가 심장 이식을 받거나 또는 받지 않으므로 둘 중 하나에 의한 결과만 관찰할 수 있다. 이와 같이 심장 이식을 받았는 지 여부와 같은 치료 변수를 A, 그에 따라 관찰된 결과를 Y 라고 한다. 따라서 우리는 예시로부터 각 개인이 심장 이식을 받았는지 여부 A와 그에 따라 관찰된 결과 (사망 여부) Y만 확인할 수 있다(표 8).

우리는 표 8에 있는 데이터로, 심장 이식 여부 A에 따라 인구집단 내 사망한 개인의 비율을 계산할 수 있다. 예를 들어, 표 8에서, 심장 이식을 받은 사람 (A=1) 13명 중 7명이 사망하였다 (Y=1). 따라서, 심장 이식을 받은 사람에 대한 사망률 $\Pr[Y=1 \mid A=1]$[7] 은 7/13이다. 같은 방식으로 심장 이식을 받지 않은 사람 7명에 대한 사망률을 구해보면 사망한 사람은 3명이므로 사망률 $\Pr[Y=1 \mid A=0]$ 은 3/7이다. 따라서 $\Pr[Y=1 \mid A=1]$와 $\Pr[Y=1 \mid A=0]$의 차이[8]를 살펴보면 $7/13-3/7 \neq 0$이기 때문에 심장 이식 여부 A와 사망 여부 Y가 서로 의존적이라고 할 수 있다[9]. 이때, 우리는 치료 변수 A와 결과 변수 Y가 서로 연관성이 있다고 표현한다.

---

7) 조건부 확률 $\Pr[Y=1|A=a]$의 의미는 인구집단 중 치료 수준 a를 받은 부분집단에서 결과 Y가 발생한 확률, 즉 치료 수준 a를 받은 사람 중 결과 Y가 발생한 비율로 계산될 수 있다.

8) 현재 예제에서는 확률의 차이를 보았지만, 경우에 따라서 차이뿐만 아니라 비 또는 오즈 비를 사용할 수 있으며, 그 경우에는 0이 아닌 1이 연관성 유무를 확인하는 기준이 된다.

9) 치료 변수 A와 결과 변수 Y가 서로 독립이면 조건부 확률의 정의에 따라 $\Pr[Y=1|A=0]=\Pr[Y=1|A=1]$가 성립한다. 이 명제의 대우 명제에 의하여 '$\Pr[Y=1|A=0] \neq \Pr[Y=1|A=1]$이면 치료 변수 A와 결과 변수 Y는 서로 독립이 아니다. 즉, 치료 변수 A와 결과 변수 Y가 서로 의존적이다(dependent).'가 성립하게 된다.

표 8. 관찰된 치료 A와 결과 Y를 설명하기 위한 예시 자료

이름	A	Y
서연	0	0
민준	0	1
유진	0	0
동현	0	0
지혜	1	0
지훈	1	0
영희	1	0
철수	1	1
지영	0	1
정훈	0	1
미영	0	0
성호	1	1
미경	1	1
영수	1	1
영숙	1	1
지훈	1	1
순자	1	1
현우	1	0
영자	1	0
성민	1	0

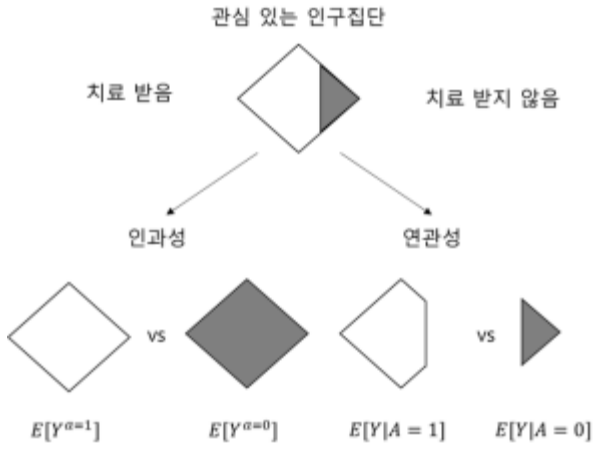


그림 21. 인과성과 연관성을 표현한 그림.

연관성에 대하여 그림으로 표현하면 그림 21과 같이 표현할 수 있다. 연관성의 정의에서 사용되는 확률은 모두 조건부 확률로 관심 있는 인구 집단 내에서 치료 수준이 1인 집단과 0인 집단에서 결과가 발생할 확률이기 때문에 그림 21에서 표현되고 있는 것과 같이 관심 있는 인구집단에 해당하는 마름모가 흰색 영역(치료 수준이 1인 경우)과 회색 영역(치료 수준이 0인 경우)으로 나뉘는 것을 확인할 수 있다. 하지만 인과성은 위에서 설명하였듯이 인구집단 전체가 치료 수준이 1인 경우와 0인 경우의 결과를 비교하기 때문에 마름모 전체가 흰색인 경우와 마름모 전체가 회색인 경우의 대비라고 정의할 수 있다. 즉, 인과성에 대한 추론은 “모든 사람이 치료를 받았다면 위험도가 어느 정도인가?” 및 “모든 사람이 치료를 받지 않았다면 위험도가 어느 정도인가?”와 같은 잠재적 세계에서의 질문과 관련이 있다. 반면 연관성에 대한 추론은 “치료를 받는 사람들의 위험은 어느 정도인가?” 및 “치료를 받지 않은 사람들의 위험은 어느 정도인가?”와 같은 실제 세계의 질문과 관련이 있다.

우리는 지금까지 설명한 표기법을 사용하여 인과성과 연관성을 구분하여 표현할 수 있다. 앞서 계산했던 $\Pr[Y=1 \mid A=a]$는 치료 수준 a를 실제로 받은 인구집단 내 부분집단에서 결과 Y가 발생할 조건부 확률을 의미하고,

$\Pr[Y^a=1]$는 인구집단 전체가 치료 수준 a를 받았을 경우, 결과 Y가 발생할 비조건부 (unconditional 또는 marginal이라고 표현함) 확률이다. 따라서, 연관성을 개인에서 실제로 치료를 받은 값 (A=1 또는 A=0)에 의해 결정된 인구집단 내 서로 다른 부분집단들 사이의 결과가 발생할 확률의 대조 (예; 차이 $\Pr[Y=1 \mid A=1]-\Pr[Y=1 \mid A=0]$)로 정의하는 반면, 인과성은 인구집단이 서로 다른 두 가지 치료 수준 (a=1 또는 a=0)을 각각 받았을 경우, 결과가 발생할 확률의 대조로 정의된다.

이렇게 인과성과 연관성은 근본적으로 다르며, "연관성은 인과성이 아니다"라는 잘 알려진 격언이 이러한 둘 사이의 차이를 설명한다. 철수의 대가족에서 치료를 받은 부분집단의 사망 위험 (7/13) 은 치료를 받지 않은 부분집단의 사망 위험 (3/7) 보다 더 컸다. 하지만, 철수의 대가족 구성원이 모두 치료를 받았을 때와 모두 치료를 받지 않았을 때의 사망 위험은 10/20으로 같았다. 이 예제와 같이 인과성과 연관성에 대한 정의의 차이로 인해 확률의 대조를 계산했을 때, 그 결과가 다를 수 있다. 그리고 이러한 불일치는 교란 (confounding)에 의해 발생할 수 있다.

위의 예제와 같이 평균 인과효과를 산출하기 위해서는 표 7의 가상 데이터와 같은 데이터가 필요하지만, 우리가 수집할 수 있는 자료는 표 8과 같은 실제로 관찰할 수 있는 자료뿐이다. 문제는 '평균 인과효과를 추정하기 위해 관찰 자료를 사용할 때, 어떤 조건이 필요한가'이다. 다음 장에서는 한 가지 답을 제시하는데, 이것은 무작위 시험을 수행하는 것이다.

2. 무작위 시험 (randomized experiments)

1) 무작위화 (randomization)

실제 세계에서는 철수의 잠재적 결과들인 심장 이식을 받았을 때의 결과 $Y^{a=1}$와 심장 이식을 받지 않았을 때의 결과 $Y^{a=0}$을 동시에 알 수 없다. 다

만, 우리는 실제로 받은 치료 A 아래에서 관찰되는 잠재적 결과 Y만 알 수 있다. 즉, 실제로 받은 치료 값의 결과에 해당하는, 각 개인에서 나타날 수 있는 두 개의 잠재적 결과들 중 오직 하나만 알 수 있고, 나머지 다른 하나의 잠재적 결과는 데이터에서 결측 자료가 되며, 이를 표 9에서 보여주고 있다. 이전 장에서 말했듯이, 치료의 효과를 측정하기 위해서는 두 개의 잠재적 결과 값이 모두 필요하기 때문에 이러한 결측 자료를 가진 데이터는 평균 인과효과의 계산을 어렵게 한다.

이러한 결측 자료에도 불구하고, 평균 인과효과의 계산을 가능하게 해주는 연구 디자인인 무작위 시험을 소개하고자 한다. 무작위 시험 또한 표 2.1과 같이 잠재적 결과들 중 일부에 대해 결측 자료를 가지는 데이터를 생성한다. 그러나 무작위 시험에서 의미하는 치료의 '무작위' 배정은 이러한 결측치가 우연에 의해 발생했음을 보장한다. 결과적으로 결측치가 있음에도 불구하고, 무작위 시험에서는 효과 측정을 일관되게 추정하거나 계산할 수 있다. 아래에서 좀 더 살펴보자.

그림 21에서 마름모로 표시된 관심 있는 인구집단의 표본 수가 거의 무한대이고, 이러한 인구집단 각 개인에 대해 동전을 던져서 치료 약의 사용 여부를 결정했다고 가정해보자(이어 나오는 예시는 앞선 심장 이식 예제와는 다른 예제이다). 동전의 뒷면이 나오면 흰색 그룹에, 앞면이 나오면 회색 그룹에 개인을 배정했다. 하지만, 그림 21을 보면 흰색 영역의 크기가 회색 영역보다 더 크므로, 동전의 앞면이 나올 확률이 50% 미만이라는 의미이며, 이러한 사실을 통해 공정한 동전 던지기가 아니라고 볼 수 있다. 다음으로, 우리는 연구 조교에게 치료 약 (A=1)를 흰색 그룹의 개인에게, 위약 (A=0)을 회색 그룹의 개인에게 투여하도록 요청하였다. 5일 후, 연구가 끝날 때 우리는 각 그룹에서의 사망률을 계산하였으며, 치료를 받은 군에서의 사망률 $\Pr[Y=1 \mid A=1]$은 0.3, 치료를 받지 않은 군에서의 사망률 $\Pr[Y=1 \mid A=0]$은 0.6으로 나타났다. 이때, 연관성의 위험도 차이(associational risk difference)는 0.3-0.6=-0.3, 연관성의 위험도 비 (associational risk

ratio)는 0.3/0.6=0.5로 계산되었다. 여기서 우리는 인과추론에 대한 몇 가지 핵심개념의 소개를 위해 이상적인 무작위 시험을 가정하고자 한다. 즉, 추적관찰이 중단된 연구대상자가 없고, 연구기간 동안 할당받은 치료를 완전히 준수하고, 연구 기간동안 단일한 버전의 치료 (single version of treatment)를 하고, 이중 맹검 할당 (double blind assignment)을 하였고 가정하였다.

표 9. 연관성 측정을 설명하는 예시 자료

이름	A	Y	$Y^{a=0}$	$Y^{a=1}$
서연	0	0	0	?
민준	0	1	1	?
유진	0	0	0	?
동현	0	0	0	?
지혜	1	0	?	0
지훈	1	0	?	0
영희	1	0	?	0
철수	1	1	?	1
지영	0	1	1	?
정훈	0	1	1	?
미영	0	0	0	?
성호	1	1	?	1
미경	1	1	?	1
영수	1	1	?	1
영숙	1	1	?	1
지훈	1	1	?	1
순자	1	1	?	1
현우	1	0	?	0
영자	1	0	?	0
성민	1	0	?	0

이제는 연구 조교가 우리 지시를 잘못 이해하고 흰색 그룹이 아닌 회색 그룹에게 치료 약을 투여했다면, 어떤 일이 일어났을지 상상해보자. 연구가 끝난 후에 연구 조교가 오해한 내용에 대하여 알게 되었다고 가정해보자. 이러한 치료 상태의 역전은 연구 결과에 어떠한 영향을 줄까? 결론적으로, 연구 결과는 전혀 영향을 받지 않는다. 우리는 여전히 치료를 받은 군 (회색 그룹)의 사망률 $\Pr[Y=1|A=1]$은 0.3이고, 치료를 받지 않은 군 (흰색 그룹)의 사망률 $\Pr[Y=1|A=0]$은 0.6이다. 이 무작위 시험에서 연관성을 측정하더라도 연구 결과는 바뀌지 않는다. 왜냐하면, 연구대상자 개인들은 흰색 그룹과 회색 그룹에 무작위로 할당되었기 때문에, 치료 받은 사람들의 사망 비율인 $\Pr[Y=1|A=1]$은 흰색 그룹이 치료를 받았든, 회색 그룹이 치료를 받았든 동일할 것으로 예상되기 때문이다. 두 군의 구성원이 무작위로 할당되면, 치료를 받은 특정 그룹은 $\Pr[Y=1|A=1]$의 값과 연관성이 없다. 물론 $\Pr[Y=1|A=0]$의 경우에도 마찬가지다. 이와 같이 이상적인 무작위 시험에서 무작위 할당된 두 그룹은 교환 가능 (exchangeable)하다고 말할 수 있다.

교환가능성 (exchangeability)은 흰색 그룹의 연구 대상자들이 회색 그룹에 제공된 치료를 받았다면, 흰색 그룹의 사망률이 회색 그룹의 사망률과 동일했을 것임을 의미한다. 즉, 이상적으로 무작위 할당이 된 시험에서는, A=1과 A=0 모두에 대하여, 치료를 받은 집단이 잠재적 치료 값인 a를 받게 될 경우의 사망률 $\Pr[Y^a=1 \mid A=1]$은, 치료를 받지 않은 집단이 잠재적 치료 값인 a를 받게 될 경우의 사망률 $\Pr[Y^a=1 \mid A=0]$와 같다는 것을 의미한다. 이러한 교환가능성을 만족하는 이상적인 조건 하에서, 인구집단 내 치료 상태에 의해 정의된 모든 부분집단의 사망률이 동일하다는 결과는, 이 부분집단의 사망률이 전체 모집단에서 치료 유무 하에 측정되는 사망률과 동일하다는 것을 의미한다: $\Pr[Y^a=1|A=1]=\Pr[Y^a=1|A=0]=\Pr[Y^a=1]$. 왜냐하면, 치료 값 a 하에서 잠재적 결과에 대한 확률은 치료를 받은 군 (A=1)과 치료를 받지 않은 군 (A=0)에서 같기 때문이다. 우리는 이러한 내용을 치료 여부 A가 잠재적 결과 Y^a를 예측하지 못한다고 얘기할 수 있으며, 교환가능성은 잠재적 결과와 실제 치료 여부가 독립임을 의미하며, 따라서 모든 값 a에 대하여

$Y^a \amalg A$ 라고 표현할 수 있다. 이때, 무작위화는 교환가능성을 가능하게 만들기 때문에 매우 중요한 의미를 가지고 있다. 치료를 받는 군과 치료를 받지 않는 군이 교환 가능할 경우, 우리는 치료가 외생적 (exogenous)이라고 말하고, 외생 (exogeneity)은 일반적으로 교환가능성의 동의어로 사용된다.

즉, 이상적인 무작위 시험은, 교환가능성을 만족시키고, 전체 인구집단에서 치료에 따른 잠재적 사망률을 계산할 수 있게 해준다. 따라서, 이상적인 무작위 시험 아래에서 인구집단 전체 구성원이 치료를 받은 경우의 사망률 $\Pr[Y^{a=1}=1]$은 치료를 받은 부분집단에서의 사망률 $\Pr[Y=1|A=1]=0.3$ 과 같다. 치료를 받지 않은 경우에도 마찬가지이다. 인구집단 전체 구성원이 치료를 받지 않은 경우의 사망률 $\Pr[Y^{a=0}=1]$은 치료를 받은 부분집단에서의 사망률 $\Pr[Y=1|A=0]=0.6$과 같다. 따라서 인과 위험 비 (causal risk ratio)는 0.5이고 인과 위험 차이 (causal risk difference)는 -0.3이다. **이상적인 무작위 시험에서 연관성은 인과성이다.**

다음으로 넘어가기 전에 $Y^a \amalg A$와 $Y \amalg A$은 다른 의미를 가진다는 것을 알아두어야 한다. 교환가능성 $Y^a \amalg A$은 잠재적 결과와 관찰된 치료가 독립적이라는 것으로 의미하며, $Y \amalg A$은 관찰된 결과와 관찰된 치료가 서로 독립적이라는 것을 의미한다. 잠재적 결과가 관찰된 치료에 의해 정해지는 것은 아니지만 관찰된 결과는 가능한 잠재적 결과 중 관찰된 치료에 따라 정해지기 때문에 관찰된 치료와 연관되어 있어 $Y^a \amalg A$가 $Y \amalg A$을 암시하지는 않는다.

이제 표 9의 심장 이식 자료는 교환가능성을 만족하는지 확인하여 보자. 이 물음에 답하기 위해서는 a=0일 때와 a=1일 때, $Y^a \amalg A$가 성립하는지 성립하지 않는지 확인해야 한다. 먼저 a=0 때를 살펴보자. 표 7의 잠재적 결과에 대한 자료를 확보할 수 있다고 가정해보자. 우리는 실제 세계에서는 치료를 받은 13명의 환자 모두 치료받지 않은 경우의 사망률 $\Pr[Y^{a=0}=1|A=1]=7/13$을 계산할 수 있다. 또한, 실제 세계에서 치료를 받지 않은 7명의 환자 모두 치료를 받지 않은 경우의 사망률 $\Pr[Y^{a=0}=1|A=0]=3/7$을 계산할 수 있다. 여기서 실제 세계에서 치료를 받지 않았던 사람보다 치료를 받았던 사람들이

치료를 받지 않은 경우의 사망률이 더 높기 때문에(즉, 7/13>3/7) 우리는 치료를 받은 사람이 치료를 받지 않은 사람보다 예후가 더 나쁘다는 결론을 내릴 수 있으며, 따라서 표 9에서 치료를 받은 사람과 치료를 받지 않은 사람 간의 교환가능성은 만족하지 않는다고 답변할 수 있다. 즉, 수학적으로, 치료 수준 a=0일 때, 교환가능성 $Y^a \amalg A$은 만족되지 않는다는 것을 보일 수 있다. a=1일 때도 같은 방법으로 교환가능성이 만족되지 않는다는 것을 보일 수 있다. 따라서 이 문단에서의 물음에 대한 답은 '아니오'에 해당한다.

물음에 답하기 위해 우리는 표 7의 잠재적 결과에 대한 자료를 확보할 수 있다고 가정하였다. 하지만, 우리는 표 7과 같이 잠재적 결과를 포함하는 자료는 확보할 수 없고, 표 9과 같이 관찰된 자료만 수집할 수 있다. 따라서 실제로는 치료를 받은 사람들에 대하여 '치료를 받지 않았더라면'에 해당하는 사망률 $\Pr[Y^{a=0}=1 \mid A=1]$은 표 9에 있는 정보만으로는 계산이 불가하다. 그러므로 우리는 연구에서 일반적으로 교환가능성이 성립하는지 자료를 통해 검증할 수 없다.

우리가 표 7과 같은 자료를 수집할 수 있어서, 결론적으로 심장 이식 연구에서 교환가능성이 성립하지 않는다고 가정해 보자. 그렇다면 우리 연구가 무작위 시험이 아니라는 결론을 내릴 수 있을까? 대답은 '아니요'다. 그 이유는 다음 두 가지 이유 때문이다. 첫째, 독자가 이미 생각하고 있는 것처럼, 20명을 대상으로 한 연구는 명확한 결론을 내리기에는 대상자 수가 너무 적다. 샘플링 가변성 (sampling variability)으로 인해 무작위 변동 (random fluctuation)이 발생할 수 있기 때문이다. 따라서, 우리는 여기서 설명하는 인구집단의 각 개인이 10억 명을 대표한다고 가정하자. 둘째, 무한한 연구 대상자에서 교환가능성이 유지되지 않더라도 연구가 무작위 시험일 가능성이 여전히 있다. 그러나 이 절에서 설명하는 무작위 시험 유형과 달리 연구자가 치료를 무작위로 할당하기 위해 두 개 이상의 코인을 사용하는 무작위 시험일 경우에 해당한다. 다음 절에서는 하나 이상의 동전을 사용한 무작위 시험에 대해 설명하고자 한다.

2) 조건부 무작위화 (conditional randomization)

표 10는 심장 이식 여부를 무작위로 배정한 무작위 시험의 데이터를 보여주고 있다. 심장 이식 여부 A (이식을 받은 경우 1, 그렇지 않은 경우 0) 및 사망 여부 Y (개인이 사망한 경우 1, 그렇지 않은 경우 0)에 대한 자료 외에도 표 10에는 환자에게 치료가 할당되기 전에 측정한 예후 인자 L (위독한 경우 1, 그렇지 않은 경우 0)에 대한 자료도 포함되어 있다. 여기서 두 가지 상호 배타적인 연구 설계를 살피면서 표 10의 자료가 이 두 연구 설계 중 어떤 설계에서 발생할 수 있는지 논의하여 보자.

연구 설계 1에서 우리는 관심 있는 인구집단의 구성원에게 65%의 확률로 심장 이식 여부를 무작위로 배정하고, 선택된 각 개인에게 새로운 심장을 이식하였다. 이러한 내용이 20명 중 13명이 심장 이식을 받은 이유를 설명한다. 연구 설계 2에서는 구성원을 예후 인자 L에 따라 분류하였고, 위독한 상태에 있는 개인에게는 75%의 확률로, 그렇지 않은 상태에 있는 개인에게는 50%의 확률로 심장 이식 여부를 무작위로 배정하고 선택된 각 개인에게 새 심장을 이식했다. 따라서 위독한 상태에 있는 12명의 개인 중 9명과 위독하지 않은 상태에 있는 8명의 개인 중 4명이 치료를 받은 이유를 설명한다.

두 연구 설계 모두 무작위 시험이며, 다만, 연구 설계 1은 구성원의 위독한 상태와는 무관하게 심장 이식 여부를 무작위로 배정하였으며(이전 절에서 설명한 무작위 시험의 유형임), 연구 설계 2은 구성원의 위독한 상태에 따라 심장 이식의 배정에 관한 확률을 달리한 무작위 시험이다. 연구 설계 1에서는 모든 개인에게 치료를 할당할 때 65%의 확률로 심장 이식을 시행하는 한 개의 동전만 사용하지만, 연구 설계 2에서는 총 두 개의 동전을 사용하는데, 하나는 위독한 상태의 개인에게는 75%의 확률로 심장 이식을 시행하는 동전과 다른 하나는 위독하지 않은 상태의 개인에게는 50%의 확률로 심장 이식을 시행하는 동전이다. 연구 설계 1과 달리 연구 설계 2는 환자의 예후 인자 L의 값에 의존하는 (조건부) 여러 개의 동전을 사용하기 때문에 조건부 무작위 시험

(conditionally randomized experiments)이라고 하며, 연구 설계 1은 모든 개인에게 공통적인 (비조건부) 하나의 동전을 사용하기 때문에 비조건부 무작위화 시험 (marginally randomized experiments)이라고 한다.

이전 절에서 논의한 바와 같이, 비조건부 무작위 시험은 노출 군과 비노출 군에 대해 교환가능성이 성립할 것으로 생각된다. 즉,

모든 a에 대하여 $\Pr[Y^a = 1|A = 1] = \Pr[Y^a = 1|A = 0]$ or $Y^a \amalg A$

하지만 조건부 무작위 시험은 노출 군과 비노출 군 사이의 비조건부 무작위 시험의 교환가능성을 만족시키지 않는데, 그 이유는 조건부 무작위 시험의 설계에 따라 노출 군과 비노출 군 각각에서 나쁜 예후를 가진 개인의 비율이 다를 수 있기 때문이다. 다시 말해, 비조건부 무작위 시험의 교환 가능성이 성립한다면 치료를 받는 군과 치료를 받지 않는 군이 서로 교환가능하기 때문에 나쁜 예후를 가지는 사람들의 비율 또한 동일해야 한다. 하지만 현재 예제를 보면, 치료를 받은 사람 중에서 69%가 위독한 상태 ($\Pr(L = 1 \mid A = 1) = 9/13$)이고, 치료를 받지 않은 사람 중에서는 43%가 위독한 상태 ($\Pr(L = 1 \mid A = 0) = 3/7$)로 같지 않기 때문에 비조건부 무작위 시험의 교환가능성이 성립하지 않는다고 할 수 있다. 또한, 이러한 나쁜 예후를 가지는 사람의 비율의 차이(불균형(imbalance)라 부름)는 실제 치료를 받은 사람들이 치료를 받지 않았더라면, 실제 치료를 받지 않은 사람들의 사망보다 더 높았을 것이라는 점을 암시한다. 즉, 치료 여부 A는 치료를 받지 않는 경우의 잠재적 결과 (사망 여부)를 예측할 수 있기 때문에 교환가능성 $Y^a \amalg A$은 성립하지 않는다. 하지만, 우리 연구는 무작위 시험이었기 때문에, 이 연구는 예후 인자 L을 조건으로 하여 치료 여부를 무작위로 배정한 무작위 시험이라고 결론을 내릴 수 있다.

표 10. 조건부 무작위화를 설명하기 위한 예시 자료

이름	L	A	Y
서연	0	0	0
민준	0	0	1
유진	0	0	0
동현	0	0	0
지혜	0	1	0
지훈	0	1	0
영희	0	1	0
철수	0	1	1
지영	1	0	1
정훈	1	0	1
미영	1	0	0
성호	1	1	1
미경	1	1	1
영수	1	1	1
영숙	1	1	1
지훈	1	1	1
순자	1	1	1
현우	1	1	0
영자	1	1	0
성민	1	1	0

조건부 무작위 시험은 연구 설계 2와 같이 단순한 두 개의 개별적인 비조건부 무작위 시험을 평행하게 결합한 것이다. 하나는 위독한 환자 (L=1)로 구성된 부분집단에서 수행하고, 다른 하나는 위독하지 않은 환자 (L=0)로 구성된 부분집단에서 수행한다. 먼저 위독한 환자로 구성된 부분집단에서 수행되는 무작위 시험을 고려해보자. 이 부분집단에서 무작위 시험이 시행되기 때문에 치료를 받는 군과 치료를 받지 않는 군은 교환가능성을 만족한다.

모든 a에 대해

$$\Pr[Y^a = 1 \mid A = 1, L = 1] = \Pr[Y^a = 1 \mid A = 0, L = 1] \text{ or } Y^a \amalg A \mid L = 1$$

여기서 $Y^a \amalg A \mid L = 1$은 L=1이 주어졌을 때 Y^a와 A가 독립이라는 의미이다. 이와 같은 방식으로 위독하지 않은 환자로 구성된 부분집단에서 시행된 무작위 시험으로 인하여 치료를 받는 군과 치료를 받지 않는 군에 대하여 교환가능성이 성립한다. 즉, $Y^a \amalg A | L = 0$이다. 예후 인자 L의 모든 값 l에 대하여(현재 예제에서는 0과 1) $Y^a \amalg A | L = l$가 성립하면, 우리는 간단히 $Y^a \amalg A | L$이라고 표현하며, 조건부 교환가능성(conditional exchangeability)이라 부른다. 따라서 조건부 교환가능성 $Y^a \amalg A | L$이 비조건부 교환가능성 $Y^a \amalg A$을 보장하지 않지만, 조건부 교환가능성은 예후 인자 L의 수준 내에서 교환가능성을 보장한다. 요약해서 말하자면, 무작위화는 비조건부 교환가능성(연구 설계 1)과 조건부 교환가능성(연구 설계 2) 중 하나를 성립하도록 만들어낼 수 있다.

우리는 비조건부 교환가능성이 만족하는 경우, 결과에 대한 노출의 효과를 다음과 같이 측정할 수 있다. 비조건부 무작위 시험에서 인과 위험 비(causal risk ratio)는 인구집단의 구성원이 모두 치료 수준 1일 때의 잠재적 결과가 발생할 확률 $\Pr[Y^{a=1} = 1]$ 을 모두 치료 수준 0일 때의 잠재적 결과가 발생할 확률 $\Pr[Y^{a=0} = 1]$으로 나눈 값($\Pr[Y^{a=1} = 1] / \Pr[Y^{a=0} = 1]$) 으로 연관 위험 비 (associational risk ratio) $\Pr[Y = 1 | A = 1] / \Pr[Y = 1 | A = 0]$와 같다. 왜냐하면, 비조건부 교환가능성은 치료 수준이 a 일 때 계산되는 잠재적 결과가 발생할 확률 $\Pr[Y^a = 1]$가 실제로 치료를 받은 사람들에서 관찰된 위험도 $\Pr[Y = 1 | A = a]$와 같다는 것을 보장하기 때문이다. 따라서, 표 10의 데이터가 비조건부 무작위 시험에서 수집된 경우 인과 위험 비는 표 10에서 $\Pr[Y = 1 \mid A = 1] = 7/13$, $\Pr[Y = 1 \mid A = 0] = 3/7$ 그리고 $\Pr[Y = 1 \mid A = 1] / \Pr[Y = 1 \mid A = 0] = \frac{7/13}{3/7} = 1.26$으로 쉽게 계산할 수 있다.

이제 조건부 무작위 시험에서 인과 위험 비를 계산해보자. 먼저, 앞서 설명하였듯이 조건부 무작위 시험은 관심 있는 인구집단 내에서 서로 다른 부분집단에서 각각 수행한 두 개 (또는 그 이상)의 개별 비조건부 무작위 시험의 조합이라는 것을 기억해야 한다. 따라서 두 가지 옵션이 있다.

첫째, 각 계층에서 평균 인과효과 (average causal effect)를 계산할 수 있다. 조건부 교환가능성에 의해 각 부분집단 내에서의 연관성은 인관성과 같기 때문에 위독한 사람들로만 구성된 계층 내 인과 위험 비 $\Pr[Y^{a=1}=1|L=1]/\Pr[Y^{a=0}=1|L=1]$은 위독한 사람들로만 구성된 부분집단(L=1)내에서의 연관 위험 비 $\Pr[Y=1|L=1,A=1]/\Pr[Y=1|L=1,A=0]$과 같다. 그리고 위독하지 않은 사람으로 구성된 부분집단에 대해서도 같은 방식으로 설명할 수 있다. 우리는 이러한 방법을 계층화(stratification)를 통해 계층 별 인과효과를 계산한다고 말한다. 위독한 사람으로 구성된 부분집단 내 인과 위험 비는 위독하지 않은 사람들로 구성된 부분집단 내 인과 위험 비와 다를 수 있다. 이러한 경우, 우리는 치료의 효과가 L에 의해 조정되거나 (the effect of treatment is modified by L) L에 의해 효과 조정이 있다고 (there is effect modification by L) 말한다.

둘째, 이전 단락에서 설명한 계층화 방법 외에 지금까지 산출해오던 관심 있는 인구집단에 대한 평균 인과효과 (예; $\Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1]$)을 계산할 수 있다. 연구자의 주요 관심은 계층 별 평균 인과효과일 수도 있고, 인구집단의 평균 인과효과일 수도 있다. 조건부 교환가능성이 만족하는 경우, 평균 인과효과를 추론하기 위해 널리 사용되는 방법으로 표준화 (standardization)와 역 확률 가중치 (inverse probability weighting)이 있으며, 소개하고자 한다.

3) 표준화 (standardization)

우리가 예시로 설명하고 있는 심장 이식 자료 (표 10)는 조건부 무작위 시

험이다. 그 이유는 위독하지 않은 상태 (L=0)의 환자 8명에게 50%의 확률로 심장 이식 여부 (A=1)를 무작위로 배정하였고, 위독한 상태 (L=1)의 환자 12명에게 75%의 확률로 심장 이식 여부 (A=1)를 무작위로 배정하였기 때문이다. 먼저 위독하지 않은 상태를 가진 8명의 환자에 대하여 살펴보자. 다만, 표본 추출로 인한 변동성을 논의에서 제외하기 위해 위독하지 않은 환자 8명이 위독하지 않은 환자 80억 명을 대표한다고 생각하자. 위독한 상태의 부분집단 또한 마찬가지이다. 우리의 목표가 인과 위험 비 $\Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1]$을 계산하는 것이다. 인과 위험 비의 분자는 모든 20명의 연구대상자가 치료를 받았을 때, 발생했을 잠재적 결과에 대한 확률 $\Pr[Y^{a=1}=1]$이다. 이 확률은 조건부 확률에 정의에 의하여 다음과 같이 표현할 수 있다.

$$\Pr[Y^{a=1}=1]=\Pr[Y^{a=1}=1 \mid L=1]\Pr[L=1]+\Pr[Y^{a=1}=1 \mid L=0]\Pr[L=0]$$

여기서 $\Pr[Y^{a=1}=1 \mid L=1]$은 조건부 교환가능성에 의해 $\Pr[Y=1 \mid A=1,\ L=1]$ 와 같으며, $\Pr[Y^{a=1}=1 \mid L=0]=\Pr[Y=1 \mid A=1,\ L=0]$이 성립한다. 따라서 확률 $\Pr[Y^{a=1}=1]$을 계산하기 위해서는 $\Pr[L=1]$($\Pr[L=0]$은 $1-\Pr[L=1]$을 통해 계산이 가능함), $\Pr[Y=1 \mid A=1,\ L=1]$ 그리고 $\Pr[Y=1 \mid A=1,\ L=0]$을 계산해야한다. 표 2.2로부터 위독한 상태에서 심장 이식을 받은 환자의 사망률 $\Pr[Y=1 \mid A=1,\ L=1]$을 구할 수 있으며, 그 값은 $\frac{2}{3}$이고, 같은 방법으로 위독하지 않은 상태에서 심장 이식을 받은 환자의 사망률 $\Pr[Y=1 \mid A=1,\ L=0]$ 또한 계산이 가능하며, 그 값은 $\frac{1}{4}$이다. 마지막으로 위독한 환자의 비율 $\Pr[L=1]$을 계산하면 $\frac{3}{5}$의 값을 얻을 수 있다. 이 결과들로부터

$$\Pr[Y^{a=1}=1]=\frac{2}{3}\cdot\frac{3}{5}+\frac{1}{4}\cdot\frac{2}{5}=\frac{1}{2}$$

임을 알 수 있다. 마찬가지 방법으로 인과 위험비의 분모를 계산하면 확률 $\Pr[Y^{a=0}=1]$ 또한 0.5임을 알 수 있다. 그러므로 인과 위험 비는 1이다.

예제에서 수행한 과정을 일반화하여 설명하면, 잠재적 결과에 대한 위험 $\Pr[Y^a=1]$은 계층 내 위험의 가중 평균이다. 이때 가중치는 전체 인구집단에서 예후 인자 L이 0인 환자 수의 비율, 예후 인자 L이 1인 환자 수의 비율과 같다. 즉, 앞서 기술하였던 것을 일반화하여 기술하면, 모든 a에 대하여 $\Pr[Y^a=1]=\Pr[Y^a=1|L=0]\Pr[L=0]+\Pr[Y^a=1|L=1]\Pr[L=1]$이다. 보다 간략하게 표현하자면, $\Pr[Y^a=1]=\sum_l \Pr[Y^a=1 \mid L=l]\cdot\Pr[L=l]$으로 표현할 수 있고, 여기서 Σ_l은 관심 있는 인구집단에서 발생할 수 있는 모든 예후인자 L의 값 l에 대하여 이어 나오는 항을 합산하는 것을 의미한다. 조건부 교환가능성에 의하여, 위의 식에서 조건부 잠재적 위험 $\Pr[Y^a=1|L=l]$을 조건부 관찰된 위험 $\Pr[Y=1 \mid A=a,\ L=l]$로 대체할 수 있다. 즉, $\Pr[Y^a=1]=\sum_l \Pr[Y=1 \mid L=l, A=a]\cdot\Pr[L=l]$이다. 이 식의 왼쪽의 확률은 관찰할 수 없는 값을 포함하는 잠재적 위험인 반면 오른쪽 양은 자료에서 확인 가능한 예후 인자 L, 치료 여부 A 및 결과 Y를 사용하여 계산 가능한 확률만 포함한다. 이와 같이 조건부 교환가능성 조건 아래에서 잠재적 결과에 관한 양은 관찰된 자료의 분포(예, 확률)의 함수로 표현할 수 있으며, 이러한 경우, 잠재적 결과에 관한 양을 식별 가능 (identifiable)하다고 한다. 반대로, 잠재적 결과에 관한 양을 관찰된 데이터의 분포 (즉, 확률)의 함수로 표현할 수 없는 경우, 잠재적 결과에 관한 양이 식별 불가능하다고 한다.

지금까지 설명한 방법을 역학 분야에서 표준화(standardization)라고 한다. 예를 들어, 인과 위험 비의 분자 $\Sigma_l\Pr[Y=1|L=l, A=1]\Pr[L=l]$은 관심 있는 인구집단을 표준으로 사용하여 치료를 받은 집단에 대해 표준화된 위험

도다. 조건부 교환가능성이 만족되는 경우, 이 표준화된 위험을 관심 있는 인구집단의 모든 개인이 치료를 받았더라면 관찰되었을 (잠재적) 위험으로 해석할 수 있다.

치료를 받은 부분집단과 치료를 받지 않은 부분집단에 대해 표준화된 위험은 각각 모두 치료를 받은 경우의 잠재적 위험과 모두 치료를 받지 않은 경우의 잠재적 위험과 같다. 따라서 인과 위험 비 $\frac{\Pr[Y^{a=1}=1]}{\Pr[Y^{a=0}=1]}$은 표준화를 통해 $\frac{\sum_l \Pr[Y=1|L=l, A=1]\Pr[L=l]}{\sum_l \Pr[Y=1|L=l, A=0]\Pr[L=l]}$으로 계산할 수 있다.

4) 역 확률 가중치 (inverse probability weighting)

이전 절에서는 조건부 무작위 시험에서 표준화를 통해 인과 위험 비를 계산하는 과정에 대하여 설명하였다. 이번 절에서는 역 확률 가중치 방법을 통해 인과 위험 비를 계산하는 방법에 대하여 설명하고자 한다. 표 10에 포함되어있는 20명의 환자를 그림 22과 같이 왼쪽에서 오른쪽으로 진행하는 나무를 통해 분류할 수 있다.

나무의 가장 왼쪽 원 안에는 두 개의 가지가 포함되어 있다. 두 개의 가지는 위독하지 않은 환자와 위독한 환자를 구분하기 위해 사용된다. 그러므로 20명의 환자 중 위독하지 않은 환자 8명은 두 개의 가지 중 윗 방향으로 향하는 가지를 따라가며, 위독한 환자 12명은 아래 방향으로 향하는 가지를 따라간다. 가지의 위에 있는 괄호 안의 숫자는 예후 인자 L이 0 또는 1일 확률이다 ($\Pr[L=1]=\frac{3}{5}$, $\Pr[L=0]=\frac{2}{5}$). 같은 방식으로 위독하지 않은 환자를 다시 한 번 심장 이식을 받은 환자와 받지 못 한 환자로 분류할 수 있다. 위독하지 않은 환자 (L=0) 8명 중 4명은 심장 이식을 받았고 (A=1), 4명은 심장 이식을 받지 않았으므로 (A = 0), 예후 인자 L이 0일 때, 심장 이식을 받지

않을 조건부 확률은 괄호 안에 표시된 것처럼 $\Pr[A=1 \mid L=0]=\frac{4}{8}=\frac{1}{2}$이다. 심장 이식 여부에 대하여 분류한 다음 사망 여부를 통해 환자들을 한 번 더 분류할 수 있다. 가장 오른쪽의 맨 위쪽 원으로부터 위독하지 않은 환자 중 심장 이식을 받지 않은 환자에 대해서 3명의 환자가 생존하고 (Y=0), 1명의 환자가 사망한 것을 확인할 수 있다. 즉, $\Pr[Y=1 \mid A=0, L=0]=\frac{1}{4}$이고, 그러므로 $\Pr[Y=0 \mid A=0, L=0]=\frac{3}{4}$ 임을 알 수 있다. 이 가지들뿐만 아니라 나무의 다른 가지들도 비슷하게 해석할 수 있다. 이제 이 나무를 사용하여 인과 위험 비를 계산해보자.

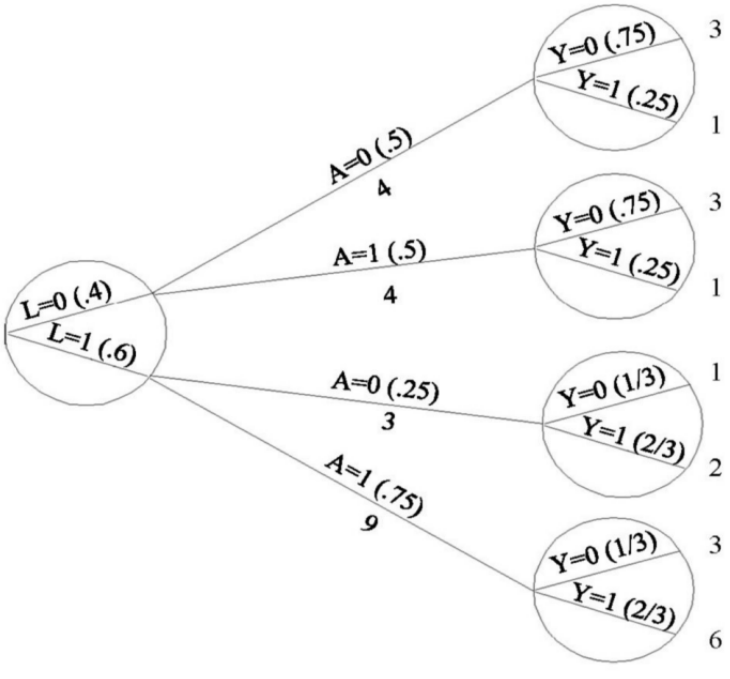


그림 22. 환자의 상태에 따른 심장 이식 여부와 사망 여부를 나타낸 나무 그림

표준화 방법을 통해 인과 위험 비를 계산할 때에는 분자를 먼저 계산하였지만 역 확률 가중치 방법에서는 분모를 먼저 계산해보고자 한다. 그림 22에서 위독하지 않은 환자 (L=0) 8명 중 4명은 심장 이식을 받지 않았고, 이 중 1명은 사망했다. 그럼 위독하지 않은 환자 8명 모두가 심장 이식을 받지 않았더라면 얼마나 많은 사망이 발생했을까? 4명이 아닌 8명이 모두 심장 이식을 받지 않은 채로 남아 있었다면 1명의 사망자가 아닌 2명의 사망자가 관찰되었을 것이기 때문에 2명의 사망이 관찰되었을 것이다. 즉, 심장 이식을 받지 않은 환자의 수가 2배가 되면, 사망자도 2배가 된다. 그림 2.1에서 위독한 환자 12명 중 3명이 치료를 받지 않았고 이 중 2명이 사망했다. 위독하지 않은 환자의 경우와 마찬가지로, 위독한 환자 12명 모두가 치료를 받지 않았다면 얼마나 많은 사망이 발생했을까? 2명이 아닌 8명의 사망이 관찰되었을 것이다. 즉, 관심 있는 인구집단 20명 모두가 치료를 받지 않았다면 위독하지 않은 환자 군에서 2명, 위독한 환자 군에서 8명으로 총 10명이 사망했을 것이다. 따라서, 인과 위험 비의 분모인 $\Pr[Y^{a=0}=1]$은 $\frac{10}{20}=\frac{1}{2}$이다. 그림 2.2의 첫 번째 나무는 관심 있는 인구집단 내 모든 환자들이 심장 이식을 받지 않은 채 남아 있는 인구집단을 보여주고 있다. 물론, 이러한 계산은 위독하지 않은 환자 (L=0) 중 심장 이식을 받은 환자가 만약 심장이식을 받지 않았더라면, 실제로 심장 이식을 받지 않은 사람들과 동일한 사망률을 가졌을 것이라는 조건부 교환가능성 $Y^{a=0}\amalg A \mid L=0$ 에 의존한다.

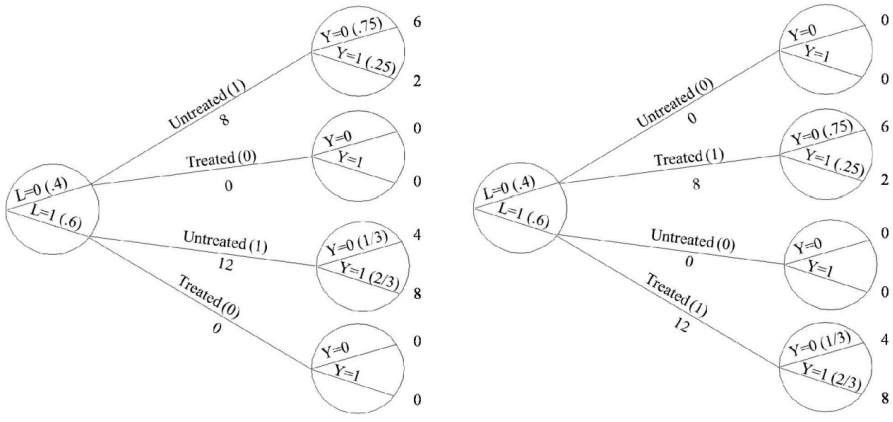


그림 23. 환자의 상태와 무관하게 모든 환자가 심장 이식을 받지 않았을 때의 사망 여부 (왼쪽)와 심장 이식을 받았을 때의 사망 여부 (오른쪽)를 나타낸 나무 그림

인과 위험 비의 분모와 마찬가지로 분자인 잠재적 사망률 $\Pr[Y^{a=1}=1]$은 관심 있는 인구집단의 모든 환자가 심장 이식을 받았을 경우의 잠재적 위험이다. 조건부 교환가능성 아래에서 분모와 같은 방법으로 계산하면 모든 환자가 심장 이식을 받았을 때의 사망률 $\Pr[Y^{a=1}=1]$은 10/20=0.5로 계산된다. 그림 23의 두 번째 나무는 모든 환자가 모두 심장 이식을 받았을 경우를 보여준다. 이 결과로부터 인과 위험 비 $\Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1]=\frac{0.5}{0.5}=1$ 임을 알 수 있다.

그림 23의 두 나무는 관심 있는 인구집단의 모든 환자가 심장 이식을 받지 않았거나 (왼쪽 그림) 또는 받았다면 (오른쪽 그림) 생기는 결과를 표현했던 그림이다. 조건부 교환가능성 아래에서 이러한 가상적인 결과를 상상해볼 수 있다. 이러한 결과를 통합하여 모든 환자가 심장 이식을 받았을 경우의 가상적인 인구집단과 모든 환자가 심장 이식을 받지 않았을 경우의 가상 인구집단을 생각해볼 수 있다. 이 가상 인구는 그림 2.2의 두 가상 인구를 통합한 인구집단이며, 그 결과 각 인구집단의 표본 수의 두 배에 해당한다. 또한, 이러

한 가상 인구를 가상의 인구집단 (pseudo-population)라고 한다. 그림 24은 통합된 인구집단을 보여줍니다.

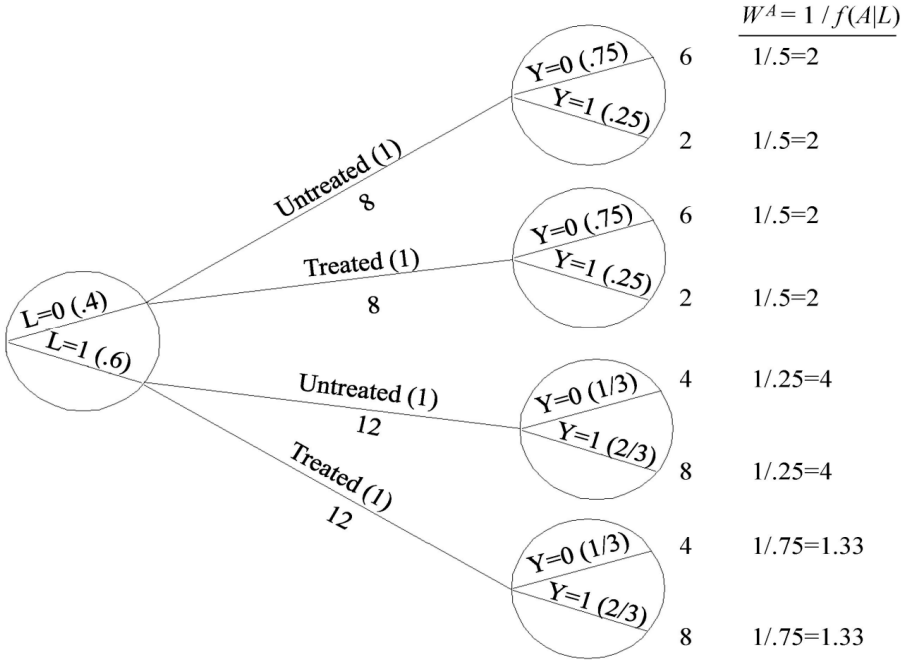


그림 24. 그림 23의 두 나무 그림을 하나의 그림으로 합쳐놓은 나무 그림

원래의 모집단에서의 조건부 교환가능성 $Y^a \amalg A|L$ 아래에서 가상의 인구집단은 예후 인자 L이 심장 이식 여부 A와 독립이 되기 때문에 가상의 인구집단에서 심장 이식을 받지 않은 부분집단과 심장 이식을 받은 부분집단이 (비조건적으로) 교환가능하다. 즉, 가상의 인구집단에서 연관성 위험 비는 가상의 인구집단과 원래 모집단 모두에서의 인과 위험 비와 동일하다.

좀 더 구체적으로 이 방법이 어떻게 작동하는지 설명하고자 한다. 그림 22의 모집단에서 위독하지 않은 환자 (L=0) 중 심장 이식을 받지 않은 4명의 환자를 살펴보자. 이 환자들은 그림 24의 가상의 인구집단에서 심장 이식을 받

지 않은 8명의 환자를 생성하는데 사용된다. 위독하지 않은 환자가 심장 이식을 받을 확률 $\Pr[A=1 \mid L=0]=0.5$의 역수인 2의 가중치를 기존의 4명에 곱하여 8명의 환자가 되었다. 같은 방식으로 그림 2.1에서 위독한 환자 9명은 확률 $\Pr[A=1 \mid L=1]=\frac{3}{4}$의 역수에 해당하는 $\frac{4}{3}$의 가중치를 받아서 12명의 연구대상자가 되었다. 즉, 가상의 인구집단은 예후 인자 L이 주어졌을 때 심장 이식을 받거나 또는 받지 않을 조건부 확률의 역수를 원래 인구집단의 각 환자에 대한 가중치로 부여하여 생성된다. 이러한 역 확률 가중치는 그림 24의 가장 우측 열에서 확인할 수 있다. 이렇게 가상의 인구집단을 생성한 후, 평균 인과효과를 계산하는 방법을 역 확률 가중치 (inverse probability weighting, IPW)라고 한다.

역 확률 가중치는 이전 절에서 표준화 (인과 위험 비가 1로 계산되었음)와 동일한 결과를 산출했다. 이것은 우연이 아니다. 표준화와 역 확률 가중치가 서로 동일한 결과를 제공한다는 사실은 널리 알려져 있으며, 수학적으로 증명된 사실이다.

표준화와 역 확률 가중치 방법 모두 예후 인자 L이 방법 내에서 사용되기 때문에 우리는 종종 이러한 방법이 예후 인자 L을 보정한다고 말한다. 또는 우리는 때때로 이러한 방법이 L을 통제한다고 표현하기도 한다.

이렇게 예후 인자 L을 통제하는 가장 좋은 방법이 무작위 시험이지만, 예를 들어, 흡연 또는 중금속에 대한 노출로 인해 발생하는 건강 영향에 대해 평가하기 위한 무작위 시험을 시행하는 것과 같이, 관심 있는 주제 중 일부에 대한 무작위 시험은 비윤리적이거나 또는 비현실적이며, 시기적절하지 않을 수 있다. 따라서 그러한 주제들의 경우 무작위 시험을 시행하기 어려워 관찰연구를 수행해야만 할 수 있다. 다음 파트에서 관찰연구에서 표준화와 역 확률 가중치를 사용하여 예후 인자 L을 통제하는 방법에 대해 설명하고자 한다.

3. 관찰연구 (observational studies)

지금까지 우리는 이상적인 무작위 시험에서 평균 인과효과 (average causal effect)를 정의하고 정량화하는 방법들에 대하여 설명하였다. 그럴 수 있었던 이유는 비조건부 또는 조건부 무작위 시험이 '(비조건부 또는 조건부) 교환가능성 ((unconditional or conditional) exchangeability)'이라는 성질을 쉽게 만족하는 연구 디자인(study design)이기 때문이다. 교환가능성에 대한 정의를 상기시키기 위해, '(비조건부) 교환가능성'은 잠재적 결과는 치료 여부에 대해 독립이라는 의미이다. 예를 들어, 심장 이식의 효과를 알기 위해 수행한 무작위 시험이 있다고 가정하자((앞선 예제와는 예임). 이 무작위 시험에서 윤리적으로 문제가 생길 수 있으며, 비현실적이지만 어떤 환자에 대한 심장 이식 여부를 동전을 던져 앞면이 나오면 심장 이식 수술을 수행한다고 상상하여보자. 이 무작위 시험은 환자의 상태와 무관하게 심장 이식 여부를 환자에게 배정하기 때문에 (비조건부) 교환가능성이 성립하고, 이로부터 심장 이식을 받은 사람들이 만약 심장 이식을 받지 않았다면, 이식을 받지 않은 사람들과 동일한 사망률이었을 것으로 예상할 수 있다. 결과적으로 무작위 시험으로부터 도출된 위험 비는 인과 위험 비와 동일하다고 생각할 수 있다. 하지만 이 무작위 시험과 달리 관찰연구의 경우에는 위의 가정이 성립하지 않을 가능성이 더 높다. 관찰연구에서는 환자에게 무작위로 심장 이식 여부를 배정할 수 없기 때문이다. 교환가능성이 성립되지 않을 수 있기 때문에 관찰연구에서 치료와 결과의 연관성을 결과에 대한 치료의 인과효과로 여기는 것이 적절하지 않을 수 있다. 그러나, 무작위 시험이 인과추론에 대한 본질적인 이점을 가진다는 것을 알고 있지만 우리는 앞서 언급한 흡연 또는 중금속의 노출에 관한 예시와 같이 인과관계의 질문에 대한 답을 찾기 위해 때때로 관찰 연구를 수행한다. 앞서 설명한 관찰연구에서의 인과추론에 대한 제한점을 해결하기 위해 우리는 기본적인 가정으로 관찰연구가 조건부 무작위 시험 (conditionally randomized experiment)이라고 볼 것이다.

다음 절에서는 관찰 연구에서 인과추론을 수행하기 위해 요구되는 세 가지 조건들을 살펴보고자 한다.

1) 교환가능성 (exchangeability)

비조건부 (또는 조건부) 무작위 시험에서는 (특정 변수에 대해 값이 주어지면) 치료를 받은 군이 치료를 받지 않았더라면 치료를 받지 않은 군과 동일한 평균 잠재적 결과를 가질 것이라는 '교환가능성'이 성립한다는 특징을 가지고 있다. 표 11은 변수 L의 수준(0과 1)에서 치료를 받은 군과 치료를 받지 않은 군에 대해 조건부 교환가능성 (conditional exchangeability)이 성립하는 자료를 제공한다. 즉, 조건부 교환가능성 $Y^a \amalg A|L$이 성립하는 자료라 볼 수 있다. 다시 관찰연구 측면으로 돌아가 보자. 치료 여부가 조사자에 의해 무작위로 배정되지 않는다면, 치료를 받게 하도록 영향을 주는 요인들 중 일부가 빠져있을 수 있으며, 이러한 요인들이 잠재적 결과와 연관성이 있을 수 있다. 즉, 치료를 받게 하도록 영향을 주는 요인이 주어졌을 때, 잠재적 결과들의 분포가 치료를 받은 군과 치료를 받지 않은 군이 서로 다를 수 있다. 표 11로 다시 예시를 들어보면, 연구의 형태가 관찰연구인 경우, 의사는 치료가 제일 필요한 사람들에게 심장 이식을 하려는 경향이 있을 수 있다. 만약 심장 이식을 받은 군과 심장 이식을 받지 않은 군의 분포를 다르게 하는 유일한 요인이 변수 L이라면, 관찰 자료와 조건부 무작위 시험은 논리적으로 봤을 때 동일하다고 볼 수 있다. 이러한 교환가능성 아래에서 표준화 (standardization) 혹은 역 확률 가중치 (IPW)가 평균 인과효과를 산출하기 위해 사용될 수 있다. 그러나 조건부 교환가능성 (conditional exchangeability)이 만족되지 않은 경우는 변수 L 외에 측정되지 않은 요인들이 존재하는 경우인데, 이 경우 조건부 교환가능성에 근접할 수 있도록 분석을 수행할 때 충분한 데이터를 확보 가능한지 살펴봐야할 필요가 있다.

표 11. 변수 L의 수준 하에서 상호교환성을 설명하기 위한 예시 자료

이름	L	A	Y
서연	0	0	0
민준	0	0	1
유진	0	0	0
동현	0	0	0
지혜	0	1	0
지훈	0	1	0
영희	0	1	0
철수	0	1	1
지영	1	0	1
정훈	1	0	1
미영	1	0	0
성호	1	1	1
미경	1	1	1
영수	1	1	1
영숙	1	1	1
지훈	1	1	1
순자	1	1	1
현우	1	1	0
영자	1	1	0
성민	1	1	0

2) 양 (positivity)의 조건

몇몇 연구자들이 심장 이식 A로 인한 5년 후 사망 여부 Y에 대한 평균 인과효과를 연구한다고 가정해보자. 이 연구에서 연구자는 모든 환자를 심장 이식 수술을 받는 군 (A=1)과 심장 이식 수술을 받지 않는 군 (A=0) 중 하나의 군으로 배정한 후, 각 군에서의 사망률의 차이로 심장 이식의 효과를 추정할

것이다. 각 군에서의 사망률의 차이를 구하기 위해서는 각 군에 속하는 환자가 적어도 1명 이상이어야 하며, 즉, 각 환자가 각 군에 배정될 확률이 0보다 커야한다. 이를 "양 (positivity)"의 조건이라고 한다.

우리는 앞서 (비조건부 또는 조건부) 무작위 시험에 대해 집중하기 위해 양의 조건에 대한 설명을 강조하지 않았다. 비조건부 무작위 시험에서 확률 $\Pr[A=1]$ 및 $\Pr[A=0]$ $(1-\Pr[A=1])$은 연구 설계상 모두 양의 확률을 가진다. 조건부 무작위 시험에서도 마찬가지로 모든 l에 대한 조건부 확률 $\Pr[A=1 \mid L=l]$은 연구 설계상 양의 확률을 가진다. 예를 들어, 표 3.1의 데이터가 조건부 무작위 시험에서 나온 것이라면 심장 이식에 대한 조건부 할당 확률은 위독한 상태에 있는 사람들의 경우 $\Pr[A=1 \mid L=1]=0.75$이고, $\Pr[A=1 \mid L=0]=0.5$이다. 따라서 예후 인자 L이 주어졌을 때, 심장 이식을 받을 확률이 양의 값을 가진다.

또한, 양의 조건은 조건부 교환가능성의 성립에 필요한 변수 L에 대해서만 필요하다. 예를 들어, 표 11의 조건부 무작위 실험에서 "연구대상자의 파란 눈 유무" 변수가 있다면 이 변수는 치료를 받은 환자와 치료를 받지 않은 환자 사이의 교환가능성을 달성하는 데 필요하지 않기 때문에 파란 눈을 가진 환자에게서 치료를 받을 확률이 0보다 큰지 여부를 묻지 않는다. 즉, 조건부 교환가능성의 성립에 요구되는 변수 L만 보정하면 표준화된 위험도와 역 확률 가중치 방법을 사용한 얻은 위험은 잠재적 위험과 동일하기 때문에 "연구대상자의 파란 눈 유무"와 같이 보정할 필요가 없는 변수에는 양의 조건을 적용할 필요가 없다.

관찰 연구에서는 양의 조건이 만족하지 않을 수 있다. 예를 들어, 그림 25과 같이 만약 의사가 위독한 환자 (L=1)에게 무조건 심장 이식을 수행한다면 (A=1), $P[A=0 \mid L=1]=0$이 되고 이로 인해 양의 조건은 성립할 수 없다. 그러나, 양의 조건이 교환가능성과 다른 점은 양의 조건은 자료로부터 검증할 수 있다는 것이다. 예를 들어, 표 3.1이 관찰연구로부터 수집한 자료라 하면 우리는 변수 L의 모든 수준 l (0 또는 1)에서 심장 이식을 받은 환자가 있는

지 없는지 셈하여 양의 조건이 성립하는지 검증할 수 있다. 이러한 양의 조건이 성립해야 표준화 (standardization)와 역 확률 가중치 (IPW)를 활용할 수 있으며, 만약 양 (positivity)의 조건이 성립하지 않을 때, 두 방법이 사용되기 어려운 이유를 아래의 그림을 통해 이해할 수 있다. 만약 변수 L이 1 일 때, 치료를 받지 않은 사람이 없다면 $\Pr[A=0 \mid L=1]=0$이고, 표준화 방법에서는 변수 L이 1의 값을 가지며, 치료를 받지 않은 환자와 비교할 대상이 없어 표준화 방법이 적용되기 어려우며, 역 확률 가중치 방법의 경우 확률이 0이므로 가중치를 계산할 수 없어 적용되기 어렵다.

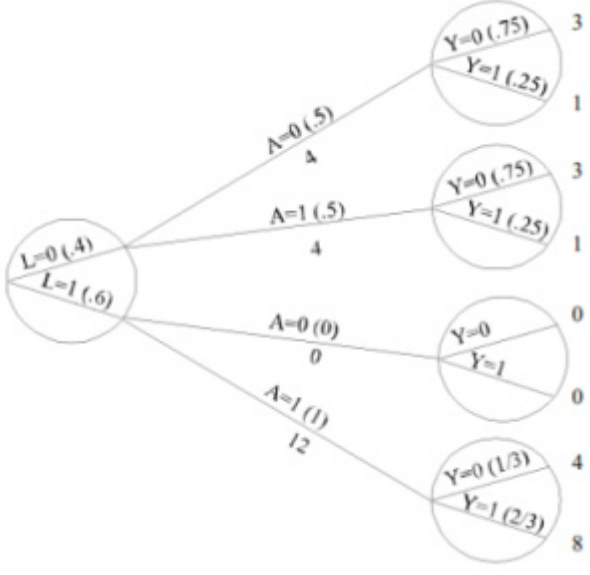


그림 25. Positivity를 설명하기 위한 예시 나무 그림

3) 일관성 (consistency)의 조건

일관성을 만족하기 위해서는 잠재적 결과를 먼저 구체적으로 정의해야 한다. “일관성 (consistency)”는 치료를 받은 환자들에게서 관찰한 결과가 해당 환자가 치료를 받았을 때 얻게 될 결과와 동일하고, 반대로 치료를 받지 않은

환자들에게서 관찰한 결과가 해당 환자가 치료를 받게 되지 않았을 때 얻게 될 결과와 동일하다는 것을 의미한다. 즉, 수식으로 표현하면 모든 연구대상자에 대하여 치료 수준 a을 받았을 때, $Y^a = Y$ 로 나타낼 수 있으며, 또는

$$Y = A \cdot Y^{a=1} + (1-A) \cdot Y^{a=0}$$

와 같이 표현되기도 한다. 일관성의 주요 구성요소는 다음과 같다.

(1) 서로 다른 치료 방법에 대한 잠재적 결과 Y^a을 가능한 구체적으로 기술해야 함.
(2) 관찰된 결과와 잠재적 결과의 연결성이 존재함.

(1) 일관성 (consistency): 반사실적 결과의 정의

앞서 언급한 일관성에 관한 두 가지 주요 구성 요소 중 (1)번 내용을 중심으로 일관성의 조건에 대해 살펴보고자 한다. 이러한 일관성의 조건은 어떤 치료에 대해 여러 방법이 있는데도 불구하고 이를 구분하지 않고 치료에 대한 효과를 추론하고자 할 때, 문제가 발생할 수 있다. 예를 들어 전통적인 수술 기법에 대한 평균 인과효과를 추론하기 위한 연구에서 전통적인 수술 기법의 심장 이식에 대한 평균 인과효과는 새로운 수술 기법을 사용한 연구결과와 다를 수 있다. 따라서 심장 이식 여부가 환자의 심근경색 여부에 미치는 영향을 볼 때, 관심 있는 치료 방법에 대해 구체화하여 기술할 필요가 있다. 특히 관찰연구에서 연구자들은 여러 치료 방법 중 관심 있는 치료 방법을 가능한 명확하게 지정해야 한다. 만약 비만의 사망에 대한 효과를 추정하기 위한 연구에서 비만을 비만이었던 기간, 가장 최근 시점에서 비만 유무, 비만의 강도 등을 통해 다양하게 정의할 수 있기 때문에 단순히 '비만'이라고 정의하는 것이 아니라 연구 가설에 부합하는 구체적인 비만의 정의를 기술하여 평균 인과효과의 크기를 산출할 수 있다.

앞서 등장하였던 철수를 예로 들어보자. 철수가 40세의 나이에 비만으로 진단받았고 (A=1), 50세의 나이에 심각한 심근경색 (Y=1)이 발생하였다고 가정해보자. 이때 철수가 허리와 관상동맥에서 지방조직이 많이 생기는 비만 유전자를 갖고 있었기 때문에 적당한 운동, 건강한 식단, 그리고 유익한 장내 미생물을 가지고 있었음에도 불구하고 심근경색이 발생하였다고 생각해보자. 또한, 사실과 다르게, 철수가 비만 유전자는 갖고 있지 않았지만, 운동 부족과 고열량 섭취 등으로 인해 비만이 되었고 (A=1), 50 세의 나이에 심근경색이 발생하지 않았다고 생각해보자 (Y=0). 이 때 비만인 경우의 철수의 잠재적 결과 $Y^{a=1}$은 어떻게 정의되어야 할까? 그가 비만 유전자로 인해 비만이 발생 (A=1)했을 때, 심근경색이 발생하였지만, 운동 부족과 고열량 섭취로 인해 비만이 발생 (A=1)했을 때의 경우, 심근경색이 발생하지 않았다. 두 경우 모두 철수에게 비만이 발생하였지만, 비만이 생기게 된 경로에 따라 심근경색이 발생하기도 하고, 발생하지 않기도 하였다. 이와 같이 치료 또는 노출 변수를 분명히 정의하지 않으면 잠재적 결과 $Y^{a=1}$ 또한 명확히 정의되지 않는다.

잠재적 결과 $Y^{a=1}$뿐만 아니라 철수가 비만이 아니었을 경우에 해당하는 잠재적 결과 $Y^{a=0}$도 잘 정의되지 않는다. 철수가 비만이 아닌 경우 또한 마찬가지로 어떻게 비만이 되지 않았느냐에 따라 50세까지 죽거나 죽지 않았을 수 있기 때문이다. 잘못 정의된 잠재적 결과는 모호한 연구 가설로부터 도출된다. 따라서 연구자가 비만 A=1이 사망률에 미치는 영향에 관심이 있다면, 연구자는 잠재적 결과 $Y^{a=1}$와 $Y^{a=0}$을 보다 구체적으로 연구 가설을 기술해야한다.

연구자가 충분히 잘 정의된 연구 가설을 기술하는 데에 집중할수록 원래의 연구 가설에서 더 멀어질 수 있으나, 변수의 모호성이 남아 있지 않도록 인과관계에 대한 질문을 정제하는 것이 인과추론에서의 기본 구성요소이다. 즉, 합의에 도달할 때까지 치료법의 세부사항을 수정하여 인과 관계에 대한 질문을 명확하게 할 필요가 있으며, 질문의 구체성이 짙어질수록 모호한 질문으로 인한 모호한 결과를 피할 가능성이 높아진다.

(2) 일관성 (consistency): 관찰된 결과와 잠재적 결과의 연결성

일관성의 정의를 만족하기 위한 두 번째 고려사항으로, 관찰된 결과와 잠재적 결과를 연결해야 한다. 앞선 예에서 연구자는 비만이 50세까지 사망률에 미치는 영향이라는 모호한 연구 가설을 보다 구체적인 연구 가설로 수정하기로 하였다. 연구자는 이제 다음과 같은 개입 (a=1)에 관심이 있다. 18세에서 연구를 시작하여 40세까지 모든 연구대상자들은 18세 당시의 체중보다 더 많은 체중이 나가지 않도록 하는 엄격한 식단을 따라야 한다. 특히, 각 연구대상자들은 18세 생일 전날부터 매일 체중을 쟀으며, 18세 당시의 체중보다 커지면 기준치에 해당하는 체중 아래로 떨어질 때까지 (보통 1~3 일 이내) 칼로리 공급원과 미량 영양소의 일반적인 조합을 변경하지 않은 상태에서 열량 섭취를 제한한다. 따라서 1~2 kg의 오차를 무시하면 40세가 될 때까지 어떤 연구대상자도 18세에 측정한 기준 체중보다 더 많이 나가지 않을 것이다. 칼로리 제한이 없는 기간 동안의 운동이나 식단에 관한 지침이나 제한은 없다. 비교하고자 하는 개입 (a = 0)은 "개입하지 않음"이다. 이때, 전문가들이 이러한 개입은 충분히 잘 정의되어 있으며, 잠재적 결과 $Y^{a=1}$ 및 $Y^{a=0}$에 대해서도 모호함이 남아 있지 않다는 것에 동의한다고 가정하자. 이제 이러한 개입 a (a는 0 또는 1)인 연구대상자에 대한 일관성의 조건 $Y^a = Y$을 기술할 수 있다.

모든 연구대상자는 두 잠재적 결과 $Y^{a=1}$과 $Y^{a=0}$에 대한 정보를 모두 가지고 있지만 실제 처치된 치료는 하나이기 때문에 두 잠재적 결과를 모두 관측할 수 없으며, 이 중 하나만 관찰된다. 이것을 수식으로 표현하면 다음과 같이 표현할 수 있다.

$$Y = A \cdot Y^{a=1} + (1 - A) \cdot Y^{a=0}$$

이 식은 연구대상자가 치료를 받으면 (A=1), 잠재적 결과 $Y^{a=1}$가 관측되고, 치료를 받지 않으면 (A=0) 잠재적 결과 $Y^{A=0}$가 관측된다는 내용을 명확히 표현하고 있다.

그러나 예를 들어 비만을 획득하게 된 경로에 대한 자료가 종종 충분하지 않은 경우가 있다. 예를 들어 40세의 체중에 대한 자료를 수집하였지만, 개인의 체중, 운동습관 및 식이요법에 대한 평생 이력에 대한 자료는 수집하지 못한 "비만에 관한 연구"가 있을 수 있다.

이 문제에서 벗어나는 한 가지 방법은 모든 치료 버전의 효과가 동일하다고 가정하는 것이다. 예를 들어, 뇌졸중에 대한 고혈압 대 정상 혈압의 인과관계에 관심이 있는 경우, 경험적 증거에 따르면 다양한 약리학적 기전을 통해 혈압을 낮추면 유사한 결과가 나타난다. 이 경우 잠재적인 결과와 관찰된 결과를 연결하기 위해 치료 "혈압을 낮추는 경로"에 대한 내용이 불필요하다고 주장할 수 있다. 그러나 다른 경우에는 이 가정에 대한 합리성이 의심될 수 있다.

물론 전문가들이 서로 다른 치료 버전이 유사한 인과적 영향을 미친다는 것에 동의한다면 자료에 있는 치료 버전의 구체화는 불필요할 것이다. 그러나 전문가들의 의견도 오류가 있을 수 있기 때문에, 우리가 할 수 있는 최선은 이러한 논의와 우리의 가정을 가능한 한 투명하게 만들어 다른 연구자들이 우리 주장에 구체적으로 이의를 제기할 수 있도록 하는 것이다.

4. 표준화와 모수적 g-formula (standardization and parametric g-formula)

우리는 2장 3절에서 표준화 방법을 통하여 평균 인과효과를 계산하는 방법에 대하여 설명하였다. 2장 3절의 예제의 경우, 조건부 교환가능성 조건을 위해 요구되는 변수가 예후 인자 L 하나였다. 하지만 조건부 교환가능성의 성립을 위해 요구되는 변수들의 수가 많거나 결과 변수에 대한 정보의 부족으로 결과 변수에 대한 분포를 가정해야 하는 경우에는 모델 기반 (model-based)의 모수적 (parametric) 표준화 방법이 사용될 수 있다. 2장 3절에서 설명한

표준화 방법은 결과 변수의 분포에 관해 어떠한 가정도 하지 않았기 때문에 비모수적 (non-parametric) 방법이라 불린다.

모수적 표준화는 앞서 설명한 비모수적 방법과 다르게, 많은 수의 공변량으로 생길 수 있는 고차원 적 문제와 연속형 치료 변수에 대한 문제를 해결하기 위해 활용될 수 있는 모델 기반의 방법이다. 이러한 모수적 표준화 방법을 사용하여 평균 인과효과를 추론하기 위해서는 치료 변수와 교란 요인을 조건으로 하는 결과 변수에 대한 조건부 기댓값을 추정해야한다. 조건부 기댓값을 추정할 때, 결과 변수의 분포에 따라 선형 회귀모형(linear regression model), 로지스틱 회귀모형(logistic regression model)을 포함하는 일반화 선형 모형(generalized linear model) 등의 모형을 사용할 수 있다. 2장 3절에서 기술했던 것과 같이 조건부 교환가능성 아래에서 잠재적 결과 Y^a의 기댓값 $E[Y^a]$은 치료를 받은 군과 치료를 받지 않은 군에 대한 가중 평균 (weighted mean)

$$E[Y^a] = \sum_l E[Y \mid A=a, L=l] \cdot \Pr[L=l]$$

으로 표현될 수 있다. 2장 3절에서는 변수 L에 해당하는 변수가 예후 인자 1개뿐이었지만, 모수적 표준화 방법이 사용되는 경우는 변수 L에 해당하는 변수는 2개 이상인 경우이다. 모수적 표준화를 수행하는 방법은 크게 데이터의 확장 (expansion of dataset), 결과 변수에 대한 모형 기술 (specification of outcome model), 예측 (prediction), 그리고 평균화 (averaging)로 네 가지 단계로 진행된다. 구체적인 예시를 통하여 설명하면 다음과 같다. 먼저 20명의 환자에 대한 정보를 포함한 자료가 있다고 가정하자. 20명의 환자에 대한 자료와 같은 자료를 추가로 2개를 복사할 것이다. 다만 복사한 자료에서 결과 변수는 제거한다. 이를 첫 번째 단계인 데이터의 확장이라고 볼 수 있다. 총 3개의 자료에서 첫 번째 원래의 자료는 그대로 두고 추가적으로 복사한 두 번째 자료에서 모든 환자에 대한 치료 변수의 값을 치료받은 상태로 변

경한다. 나아가, 세 번째 자료에서는 모든 환자에 대한 치료 변수의 값을 치료 받지 않은 상태로 변경한다. 그 다음으로 첫 번째 원래의 자료를 가지고 치료 변수와 교란 요인을 공변량으로 보정하여 결과 변수에 대한 모형을 구축한다. 이때 앞서 언급한 선형 회귀모형 또는 로지스틱 회귀모형 등의 모형이 사용된다. 이 단계가 결과 변수에 대한 모형 기술에 해당하는 단계이다. 이때 두 번째와 세 번째 자료에서 결과 값이 존재하지 않기 때문에 앞서 적합한 결과 변수에 대한 모형을 토대로 두 번째, 세 번째 자료를 사용하여 잠재적 결과를 예측한다. 예측된 잠재적 결과 값들은 두 번째 자료에서는 교란 요인과 치료받지 않은 값의 조합으로 이루어진 추정치이고, 세 번째 자료에서는 교란 요인과 치료받은 값의 조합으로 이루어진 추정치라 볼 수 있다. 마지막 단계로, 이와 같이 두 번째, 세 번째 각 자료에서 예측된 잠재적 결과 값들의 평균을 계산한 후, 평균들의 비를 통해 인과적 위험 비를 평균 인과효과로서 추정할 수 있다. 첫 번째, 두 번째 그리고 세 번째 자료는 각각 표 12, 13 그리고 14에 해당한다.

이렇듯 주어진 고차원적 문제를 해결함에 있어 모형을 이용한 모수적인 방법을 통해 표준화를 적용할 수 있다. 시간에 따라 변하지 않는 자료에서 g-formula를 적용하는 방법을 표준화라 하며, g-formula는 1986년에 Robins JM가 시간에 따라 변하는 치료 변수 또는 교란 요인에 있는 자료에서 평균 인과효과를 추론하기 위해 처음 소개한 인과추론 방법이다. g-formula는 g-computation이라 불리기도 한다.

표 12. 첫 번째 자료: 원 자료

이름	L	A	Y
서연	0	0	0
민준	0	0	1
유진	0	0	0
동현	0	0	0
지혜	0	1	0
지훈	0	1	0
영희	0	1	0
철수	0	1	1
지영	1	0	1
정훈	1	0	1
미영	1	0	0
성호	1	1	1
미경	1	1	1
영수	1	1	1
영숙	1	1	1
지훈	1	1	1
순자	1	1	1
현우	1	1	0
영자	1	1	0
성민	1	1	0

표 13. 두 번째 자료: 모두 치료받은 경우

이름	L	A	Y
서연	0	1	·
민준	0	1	·
유진	0	1	·
동현	0	1	·
지혜	0	1	·
지훈	0	1	·
영희	0	1	·
철수	0	1	·
지영	1	1	·
정훈	1	1	·
미영	1	1	·
성호	1	1	·
미경	1	1	·
영수	1	1	·
영숙	1	1	·
지훈	1	1	·
순자	1	1	·
현우	1	1	·
영자	1	1	·
성민	1	1	·

표 14. 세 번째 자료: 모두 치료받지 않은 경우

이름	L	A	Y
서연	0	0	·
민준	0	0	·
유진	0	0	·
동현	0	0	·
지혜	0	0	·
지훈	0	0	·
영희	0	0	·
철수	0	0	·
지영	1	0	·
정훈	1	0	·
미영	1	0	·
성호	1	0	·
미경	1	0	·
영수	1	0	·
영숙	1	0	·
지훈	1	0	·
순자	1	0	·
현우	1	0	·
영자	1	0	·
성민	1	0	·

5. 치료-교란 요인 피드백 (treatment-confounder feedback)

치료-교란 요인 되먹임 (treatment-confounder feedback)이란 다음의 조건을 만족하는 시간에 따라 변하는 교란 요인 L을 의미한다.

(1) 요인 L이 시간에 따라 변하는 노출 변수 (또는 치료 변수; time-varying exposure or treatment)와 영향을 주고 받는다.

(2) 인과 그래프에서 요인 L이 t 시점의 노출 변수와 결과 변수 (outcome variable)의 교란 변수이면서 또한 충돌 변수 (collider)도 될 수 있다.

조건 (1)은 시점 t-1, t, t+1에 대하여 t-1 시점의 변수 L이 t 시점의 노출 변수의 원인이 되며, 또한 t 시점의 노출 변수가 t+1 시점의 변수 L의 원인이 되어 영향을 서로 주고받는 구조가 인과 그래프에 표현되는 것을 의미한다. 조건 (2)는 인과 그래프에서 t 시점의 변수 L이 t 시점의 노출 변수와 결과 변수의 교란 변수 또는 충돌 변수 역할을 하게 되어 결과 변수에 대한 t 시점의 노출 변수의 인과효과를 산출하기 위해 교란 변수들을 고려할 때, 변수 L을 보정하게 되면 충돌 변수를 보정하게 되어 인과효과에 편향이 발생하고, 변수 L을 보정하지 않게 되면 교란 편향이 발생하는 것을 의미한다. 따라서 이러한 치료-교란 변수 되먹임의 두 번째 조건으로 인하여 치료-교란 변수 되먹임이 있는 인과 그래프에서 인과효과를 구하기 위해 전통적인 회귀분석 방법을 사용할 때는 주의가 요구된다.

6. 전통적인 회귀분석과 g-formula의 차이

1) Time-dependent Cox Model

(1) Time-dependent Cox Model과 인과 그래프

전통적인 회귀분석에 해당하는 time-dependent Cox model를 통해 현재 구하고자 하는 causal estimand는 '총 7개의 관측 시점에서 연구에 참여하는 모든 근로자가 특정 혈중 납 농도를 유지하고, 중도절단 (censoring) 되지 않은 경우에서의 빈혈의 발생률'이다. 하지만 time-dependent Cox model에서 시점 t에서의 노출 변수 E_t의 회귀 계수와 시점 t-1에서의 노출 변수 E_{t-1}의 회귀 계수는 위에서 언급한 causal estimand에 대응되지 않는다. 예를 들어, time-dependent Cox model을 통해 시간에 따라 변하는 노출 변수의 효과를 확인한다고 하면 아래와 같은 인과 그래프를 생각할 수 있다. 아래의 그림에서 Yt는 시점 t에서의 결과 변수를 의미한다. 또한, 네모 상자는 time-dependent Cox model에서 보정하는 공변량을 의미한다. 여기서, 시점 t-1에서의 결과 변수 Y_{t-1}는 모형에서 직접적으로 보정되지는 않지만 time-dependent Cox model에서 모형화하는 위험(hazard)의 정의로부터 자동적으로 보정되는 것으로 생각할 수 있다.

Time-dependent Cox model에서 두 노출 변수 E_t, E_{t-1}를 보정하게 되면, 시점 t에서의 노출 변수 E_t와 시점 t에서의 결과 변수 Y_{t-1}로 인하여 시점 t-1에서의 노출 변수 E_{t-1}에서 결과 변수 Y_t로 가는 길의 일부 (E_{t-1}->E_t->Yt 및 E_{t-1}->Y_{t-1}->Y_t)가 차단되는 것을 그림 26에서 확인할 수 있다. 두 길의 차단으로 인해 시점 t-1에서의 노출 변수 E_{t-1}의 회귀 계수에 대한 편향이 발생하기 때문에 time-dependent Cox model에서 노출 변수 E_{t-1}의 회귀계수를 인과적 의미를 가지는 값으로 해석하기에는 어려움이 따른다.

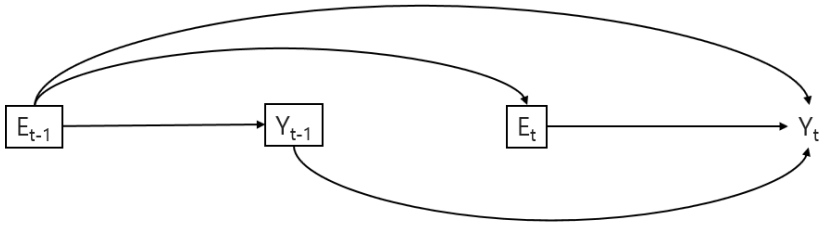


그림 26. time-dependent Cox model을 적용할 때, 보정되는 변수를 표현한 인과 그래프.

(2) 중도절단 (Censoring)

건강이 악화되어 빠르게 중도절단되는 근로자를 고려하여 분석하지 않으면 건강근로자 생존 편향이 나타난다는 것이 잘 알려져 있으며, 이러한 중도절단은 전형적인 informative censoring의 한 형태이다. 하지만 time-dependent Cox model의 경우는 non-informative censoring을 가정하고 있기 때문에 time-dependent Cox model에서 얻어진 노출 변수에 대한 회귀 계수는 편향이 포함된 값일 가능성이 높다. 지금까지 언급한 (1)과 (2)의 이유로 현재 특수건강검진자료 문제에서는 time-dependent Cox model보다는 g-formula가 더 적절한 방법이라 할 수 있다.

(3) Time-dependent Cox Model을 적합한 결과

표 15에는 time-dependent Cox model의 결과가 담겨있다. 혈중 납 농도(시점 t)의 효과를 구하기 위해 음주 여부(시점 t), 흡연 여부(시점 t), 비만도(시점 t) 그리고 혈중 납 농도(시점 t-1), 음주 여부(시점 t-1), 흡연 여부(시점 t-1), 비만도(시점 t-1)을 공변량으로 보정하였고, 그 결과 혈중 납 농도(시점 t)의 효과는 유의하지 않게 나왔다 (P-값 = 0.7161). 또한, 혈중 납 농도(시점 t-1)에 대해서도 혈중 납 농도(시점 t)와 동일하게 효과가 유의하지 않게 나왔다 (P-값: 0.4154).

표 15. Time-dependent Cox model을 적합한 결과. 각 변수에 대한 회귀계수, 표준 오차, 95% 신뢰구간 및 P-값.

역할	요인	회귀계수	표준 오차	회귀계수에 대하여 exponential을 취한 값	95% 신뢰구간의 왼쪽 경계 값	95% 신뢰구간의 오른쪽 경계 값	P-값
시간에 따라 변하는 노출 변수	혈중 납 농도 (시점 t)	0.0033	0.0092	1.0034	0.9854	1.0216	0.7161
	혈중 납 농도 (시점 t-1)	0.0072	0.0088	1.0072	0.9900	1.0247	0.4154
시간에 따라 변하는 내생 교란 요인	음주 여부 (시점 t)	-0.2021	0.0822	0.8171	0.6954	0.9599	0.0140
	흡연 여부 (시점 t)	-0.0515	0.1081	0.9498	0.7684	1.1739	0.6334
	비만도 (시점 t)	-0.0176	0.0210	0.9826	0.9430	1.0238	0.4026
	음주 여부 (시점 t-1)	0.2118	0.1111	1.2360	0.9941	1.5367	0.0566
	흡연 여부 (시점 t-1)	-0.1254	0.1089	0.8821	0.7125	1.0920	0.2494
	비만도 (시점 t-1)	-0.0621	0.0213	0.9398	0.9013	0.9799	0.0036
시간에 따라 변하는 외생 교란 요인	나이	-0.0932	0.1458	0.911	0.6845	1.2125	0.5229
	사업장 규모 (2인 사업장)	-0.2630	0.1445	0.7688	0.5791	1.0205	0.0688
	사업장 규모 (3인 사업장)	-0.1906	0.1499	0.8265	0.6161	1.1088	0.2038
	사업장 규모 (4인 사업장)	-0.2113	0.1515	0.8096	0.6016	1.0894	0.1631
	사업장 규모 (5인 사업장)	1.7372	0.0614	5.6816	5.0372	6.4085	< 0.0001
시간에 따라 변하지 않는 교란 요인	성별 (여성)	0.0347	0.0022	1.0353	1.0309	1.0397	< 0.0001

부록 2의 참고문헌

Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC, 2020.

연구진

연 구 기 관: 산업안전보건연구원

연구책임자: 예신희(팀장, 중부권역학조사팀)
연 구 원: 이경은(선임연구위원, 중부권역학조사팀)
연 구 원: 윤민주(과장, 중부권역학조사팀)
연 구 원: 박동준(전공의, 역학조사부)
연 구 원: 마성원(전공의, 역학조사부)
연 구 원: 이영신(전공의, 역학조사부)

부분위탁

연구책임자: 이우주(부교수, 서울대학교 보건대학원)
연 구 원: 조성일(부교수, 인하대학교 통계학과)
연 구 원: 이동환(부교수, 이화여자대학교 통계학과)
연 구 원: 김양우(전임의, 한양대학교 구리병원 직업환경의학과)
연구보조원: 심현만(박사과정 대학원생, 서울대학교 보건대학원)

연구기간

2022. 01. 07. ~ 2022. 11. 30.

본 연구보고서의 내용은 연구책임자의 개인적 견해이며,
우리 연구원의 공식견해와 다를 수도 있음을 알려드립니다.

산업안전보건연구원장

직업병 인과추론 가이드라인 및 통계분석법 개발 (2)
- 복합노출의 건강 영향평가 국문 가이드라인 개발-
(2022-산업안전보건연구원-738)

발 행 일 : 2022년 11월 30일
발 행 인 : 산업안전보건연구원 원장 김은아
연구책임자 : 중부권역학조사팀 팀장 예신희
발 행 처 : 안전보건공단 산업안전보건연구원
주 소 : (44429) 울산광역시 중구 종가로 400
전 화 : 032-510-0754
팩 스 : 032-510-0759
Homepage : http://oshri.kosha.or.kr
I S B N : 979-11-92782-04-1
공공안심글꼴 : 무료글꼴, 한국출판인회의, Kopub바탕체/돋움체