

2025 Lee's Data Science Lab Summer URP

Final Meeting

Group1: 강재서, 신아영, 양인경, 유보현, 주수현



목차

1. 연구 주제 선정
2. 데이터 및 준비 과정
3. 방법론 선정
4. 분석 결과
5. 결론
6. End of Summer URP

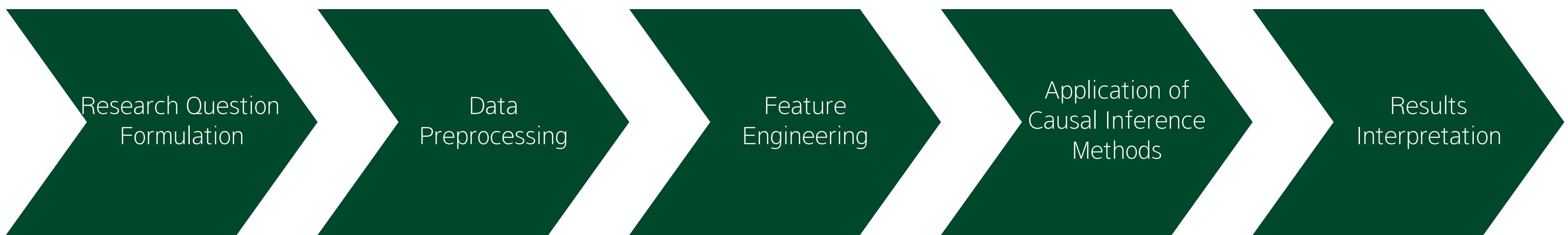
연구 주제 선정

Intro

연구 과정 목표

- 금융과 바이오를 연결하는 연구 경험
- 교수님의 연구 방법론 중 하나인 인과 추론 경험
- 이를 고려하여 주제 및 데이터 선정

연구 과정 Pipeline



4

Research Question

고비용 질병 진단 이후 의료비 과부담이 실제 금융파탄(신용하락/연체)으로 이어지는 취약 집단 탐색



고비용 질병 진단 이후 의료비 과부담이 가계 재정 상태에 미치는 영향 분석



경제활동인구의 만성질환 유무가 의료비 과부담에 미치는 영향 분석 및 인과 관계 탐색

데이터 및 전처리

Data & Preparation

분석 데이터 및 준비 과정

자료: 한국의료패널(KHP) 2020년 데이터 (N=3,605)

전처리 과정

1. 표본 선정: 경제활동인구 (20-61세, 소득 > 0)
2. 주요 변수 정의
 - 처치: 만성질환 보유 여부 (binary)
 - 결과: 의료비 과부담($\frac{\text{본인 부담 의료비}}{\text{개인 연간 총소득(근로소득+사업소득)}} > 0.1$) 여부
 - 공변량: 선행연구 기반으로 선택편향을 통제할 수 있는 15개 변수 설정
3. 결측치 처리

Data & Preparation

표본선정

- 만성질환 / 사업소득 / 의료비
→ ‘모름/무응답’으로 코딩된 값(-9)가 존재하는 표본 제외
- 소득의 합 0 이하인 표본 제외
- 태어난 연도가 1960년~2001년인 표본
- 공변량
→ ‘모름/무응답’으로 코딩된 값(-9 or 3)가 존재하는 표본 제외

```
df = df[  
    (df['CD'] != -9) & (df['CD1_1'] != -9) & (df['I_INC2'] != -9) & ((df['I_INC1'] + df['I_INC2']) > 0) &  
    (df['EROOP'] != -9) & (df['INOOP'] != -9) & (df['OUOOP_1'] != -9) & (df['OUOOP_2'] != -9) &  
    (df['BIRTH_Y'] >= 1960) & (df['BIRTH_Y'] <= 2001) & (df['S3'] != -9) & (df['D1'] != -9) & (df['I_PHI1_YN'] != 3)  
]
```

Data & Preparation

처치 & 결과 변수 정의

처치 변수 : 만성질환 유무 (이진형)

- CD: 주요 30가지 만성질환 유무
- CD1_1: 이외의 만성질환 유무
→ 둘 중 하나라도 있으면 1로 간주

```
df['CD_total'] = ((df['CD'] == 1) | (df['CD1_1'] == 1)).astype(int)
```

결과 변수 : 소득 대비 의료비 지출 (= 의료비 과부담, 이진형)

WHO 기준에 따라 소득 대비 의료비 지출이 0.1 이상이면 1로 간주

- 소득 : 개인의 근로소득과 사업소득의 합 $df['INCOME'] = df['I_INC1'] + df['I_INC2']$
- 의료비 : 본인 부담 의료비 항목의 합 $df['MED_EXP'] = ((df['EROOP'] + df['INOOP'] + df['OUOOP_1'] + df['OUOOP_2']) / 10000)$

```
df['overburden_yn'] = (df['MED_EXP'] / df['INCOME']) >= 0.1).astype(int)
```

Data & Preparation

파생변수 정의

파생변수

- AGE → 한국 나이로 변환
- 체질량 지수 BMI 생성

```
df['AGE'] = 2020 - df['BIRTH_Y']+1  
df['BMI'] = df['WT'] / ((df['HT'] / 100) ** 2)
```

공변량 이진 변환 수행

```
# SEX: 남성(1) -> 1, 여성(2) -> 0  
df['SEX'] = df['SEX'].replace({1: 1, 2: 0})  
  
# DISA_YN : 장애 있음(1) -> 1, 장애 없음(2) -> 0  
df['DISA_YN'] = df['DISA_YN'].replace({1: 1, 2: 0})  
  
# HS2_YN : 우울감 있음(1) -> 1, 나머지 -> 0  
df['HS2_YN'] = df['HS2_YN'].apply(lambda x: 1 if x == 1 else 0)  
  
# P1 : 운동을 규칙적으로 함(1) -> 1, 나머지 -> 0  
df['P1'] = df['P1'].apply(lambda x: 1 if x == 1 else 0)  
  
# MARR: 배우자 있음(1) -> 1, 없음(2) -> 0  
df['MARR'] = df['MARR'].apply(lambda x: 1 if x in [1, 2] else 0)  
  
# S3: 2이하 -> 1, 나머지 -> 1  
df['S3'] = df['S3'].apply(lambda x: 1 if x in [1, 2] else 0)
```

```
# EDU: 3이하 -> 0, 4,5 -> 1  
df['EDU'] = df['EDU'].apply(lambda x: 1 if x in [4, 5] else 0)  
  
# HEALTH_INS: 원한 인코딩  
df = pd.get_dummies(df, columns=['HEALTH_INS'], prefix='HEALTH_INS', drop_first=True, dtype=int)  
#첫 번째 카테고리를 제외하여 다중공선성을 방지합니다.  
  
# I_PHI1_YN : 민간의료보험 가입했음 -> 1, 나머지 -> 0  
df['I_PHI1_YN'] = df['I_PHI1_YN'].apply(lambda x: 1 if x == 1 else 0)  
  
# EC01: 4이하 -> 1, 5,6 -> 1  
df['EC01'] = df['EC01'].apply(lambda x: 1 if x in [1, 2, 3, 4] else 0)  
  
#순서형변수  
# HS1 (스트레스 인지정도): 4(없다) -> 0, 나머지(1,2,3) -> 1  
df['HS1'] = df['HS1'].apply(lambda x: 0 if x == 4 else 1)  
  
# D1 (음주 빈도): 1-4(거의 안 함) -> 0, 나머지(5,6,7,8) -> 1  
df['D1'] = df['D1'].apply(lambda x: 0 if x in [1, 2, 3, 4] else 1)
```

Data & Preparation

결측치 처리

소득 및 의료비 변수

→ 소득, 의료비 등 6개의 변수의 결측치를 0으로 대체

핵심 분석 변수

→ 만성질환 여부, 우울감 등 분석에 필요한 변수에 결측치가 있는 표본은 분석 대상에서 제외

```
[ ] #결측치
df['I_INC1'] = df['I_INC1'].fillna(0)
df['I_INC2'] = df['I_INC2'].fillna(0)
df['ERODP'] = df['ERODP'].fillna(0)
df['INODP'] = df['INODP'].fillna(0)
df['OUODP_1'] = df['OUODP_1'].fillna(0)
df['OUODP_2'] = df['OUODP_2'].fillna(0)
df.dropna(subset=['CD', 'CD1_1', 'HS2_YN', 'HS1', 'HT', 'WT'], inplace=True)
```

방법론 선정

Methodology

분석 방법론 선정 과정

선택 배경

- Confounding 문제
 - 만성질환 보유 여부와 의료비 지출은 나이, 소득, 건강 상태 등 다양한 요인에 동시에 영향을 받음
→ 단순 비교로는 편향 발생
- RCT 불가능:
 - 만성질환 여부를 무작위로 배정할 수 없음 → 관찰자료 기반의 준실험적 방법 필요

방법론 선정

IPTW: 공변량 분포를 균형화하여 RCT 환경을 근사 → 평균 처치 효과(ATE) 추정

PSM: 보조적 방법으로 사용하여 추정 결과의 강건성 검증

CRE: 평균 효과를 넘어 하위집단별 이질적 효과(HTE) 탐색, 정책적 타겟팅 근거 제공

Methodology

분석 방법론 선정 과정

- 1단계: 선행 연구 검토
- 2단계: 최종 방법론 선정
 - ATE 추정: IPTW, PSM
 - HTE 추정: CRE
- 3단계: 이론 학습 및 코드 구현
 - 각 방법론의 핵심 이론 및 가정 학습
 - 별도 튜토리얼 코드를 직접 생성하며 방법론의 자동 원리 숙지 후 본 연구에 적용

Methodology

Key Assumptions of Causal Inference

Conditional Independence Assumption, CIA	<ul style="list-style-type: none">처치 여부 T는 공변량 X를 통제했을 때 잠재적 결과 $Y(1), Y(0)$와 독립적즉, 모든 교란변수가 관측되어 모형에 포함되어야 함충족되지 않으면 추정된 처치 효과는 편향(bias)이 발생
Common Support / Overlap	<ul style="list-style-type: none">모든 공변량 조합에 대해 처치와 비처치 집단이 공존해야 함성향점수 $P(T=1 X)$가 0과 1에 치우치지 않고, 두 집단의 분포가 겹쳐야 함공통 지지 영역 밖의 표본은 제거하거나 제한하여 비교 가능성 확보
SUTVA (Stable Unit Treatment Value Assumption)	<ul style="list-style-type: none">각 개인의 잠재적 결과는 자신의 처치 여부에만 의존다른 개인의 처치 여부가 자신의 결과에 영향을 주면 안 됨(no interference)

방법론 결정

IPTW

방법론 결정

Inverse Probability Treatment Weighting (IPTW)

Theory

- 성향점수 기반 가중치로 공변량 분포 균형화
- 관찰연구를 RCT 환경에 근사 → 평균 효과(ATE) 추정

Process

- 성향점수(PS) 추정 → 역확률 가중치 부여 → 균형성 검증 → ATE 추정

방법론 결정

Inverse Probability Treatment Weighting (IPTW)

[분석 배경]

단순 회귀분석 → 처치군·대조군 간 기저 특성 차이로 인해 선택 편향 발생
무작위배정 임상시험은 현실적으로 불가능

[IPTW의 핵심 아이디어]

- 로지스틱 회귀로 성향점수 추정
- PS 기반 역학률 가중치 부여 → 처치군과 대조군의 공변량 분포 균형화
결과적으로, 관찰연구를 준실험적 환경으로 근사

방법론 결정

Inverse Probability Treatment Weighting (IPTW)

Logistic Regression 모형을 적합하여 개체별 성향 점수 추정

[성향점수]

$$PS_i = P(T_i = 1 \mid X_i)$$

[선정 근거]

교란변수들은 선행연구 및 도메인 지식을 기반으로 선택
치료 할당(T)과 결과(Y) 모두에 영향을 줄 수 있는 변수 포함

방법론 결정

Inverse Probability Treatment Weighting (IPTW)

처치군: $w_i = 1/PS_i$

대조군: $w_i = 1/(1 - PS_i)$

- PS 기반으로 각 개체에 처치 받을 확률의 역수를 가중치로 부여
- 집단 간 공변량 분포를 균형화하여 선택 편향 통제

[가중치 분포 확인]

대부분 표본은 안정적인 범위(1~5)에 집중

일부 큰 값 존재했으나 관리 가능한 수준

공변량 균형성이 확보되어 추가적인 극단값 보정은 적용하지 않음

방법론 결정

Inverse Probability Treatment Weighting (IPTW)

[Balance Diagnostics]

지표: 표준화 평균 차이(SMD)

- 기준: $|SMD| < 0.1$

IPTW 적용 전: 일부 변수에서 큰 불균형 존재

IPTW 적용 후: 모든 공변량이 기준 내로 수렴 \rightarrow 균형 확보

시각적 진단: 러브플롯(Love plot), 성향점수 분포 겹침도

\rightarrow IPTW를 통해 교란 요인 보정 및 인과적 해석 기반 마련

\rightarrow 단, **No unmeasured confounding** 가정 하에서 타당

방법론 결정

PSM

방법론 결정

Propensity Score Matching (PSM)

Theory

- 성향점수(처치 확률)를 바탕으로 유사한 개체를 짹지어 매칭
- 처치군과 대조군의 공변량 분포를 유사하게 만들어 무작위 배정 환경 근사
- 선택 편향(selection bias)을 최소화하여 인과적 효과 추정

Process

- 성향점수(PS) 추정 → 매칭 수행(최근접 이웃법 1:1 매칭) → 매칭 균형성 검증
→ 매칭된 표본 기반으로 ATE(평균 처치 효과) 추정

방법론 결정

Propensity Score Matching (PSM)

- 방법: 로지스틱 회귀를 통해 각 개인의 성향점수(PS) 추정
- 매칭 방법: 최근접 이웃 매칭(Nearest Neighbor Matching, 1:1)
- 성향점수 정의: $PS_i = P(T_i = 1 | X_i)$

방법론 결정

CRE

방법론 결정

Causal Rule Ensemble (CRE)

Theory

- 규칙(rule) 기반 학습으로 하위집단과 비선형성 탐지
- 명시적 규칙으로 결과 해석 가능
- 개별 효과(ITE) 개념을 기반으로 집단별 평균 효과(CATE) 추정

Process

- 후보 규칙 생성 → 필터링 → CATE 추정 → 해석

방법론 결정

Causal Rule Ensemble (CRE)

[Rule Generation]

- Decision Tree 기반으로 다양한 규칙 후보 생성
- ntrees, max_depth, node_size 등 파라미터로 다양성 확보
- 작은 하위집단까지 포착 가능
- 부트스트랩 반복(B)과 subsample 적용 → 규칙 안정성 평가

→ ntrees = 10000, max_depth = 10, node_size = 5

→ B = 200, subsample = 0.5

방법론 결정

Causal Rule Ensemble (CRE)

[Filtering]

- 후보 규칙 중 신뢰할 수 있는 규칙만 선별
- 통계적 유의성, 조건 만족 빈도, 이질성 등 선별 기준 적용
- 과적합 방지 및 신뢰성 확보

→ cutoff = 0.7, max_rules = 10000

→ 적은 표본 수를 고려하여 stability_selection = “no” 설정

방법론 결정

Causal Rule Ensemble (CRE)

[CATE Estimation]

- 선택된 규칙을 이진 변수로 변환
- 회귀 기반으로 하위집단별 평균 효과(CATE) 추정
- 최종 유의미한 규칙만 선택

→ 과적합 방지를 위해 변수 선택 + 정규화를 동시에 수행하는 `learner_ps = SL.glmnet` 설정
→ 결과 예측 정확도를 높이기 위해 `learner_y = SL.xgboost` 설정
→ ITE 추정 방식으로 cf (Causal Forest) 설정 : 하위집단 효과 탐색에 적합

방법론 결정

Causal Rule Ensemble (CRE)

[Interpretation]

- 최종 규칙으로 하위집단별 효과 해석
- 하위집단별 CATE 비교 및 시각화 가능
- 각 규칙이 의미하는 인과적 영향 설명

→ `summary(cre_results)`와 `plot(cre_results)`를 통해 결과 확인

분석 코드

분석 코드

IPTW

분석 코드

IPTW

성향점수 PS 추정

```
ps_model <- glm(CD_total ~ SEX + MARR + EDU + DISA_YN + HEALTH_INS + I_PHI1_YN + EC01 + HS2_YN +  
                  HS1 + P1 + S3 + D1 + AGE + BMI + INCOME, data = df, family = "binomial")  
df$ps <- predict(ps_model, type = "response")
```

역학률가중치 IPTW 계산

```
df$weight <- ifelse(df$CD_total == 1, 1 / df$ps, 1 / (1 - df$ps))
```

ATE 계산

```
ate_model <- svyglm(overburden_yn ~ CD_total, design = iptw_design)  
summary(ate_model)
```

분석 코드

PSM

분석 코드

PSM

성향 점수 매칭

```
mod_match <- matchit(match_formula, method = "nearest", caliper = 0.1, data = ecls_trimmed)
```

매칭 균형 평가

```
ggplot(combined_df, aes(x = pr_score, fill = CD_total)) +  
  geom_density(aes(weight = weights), alpha = 0.7) +  
  facet_wrap(~ status, ncol = 1) +  
  labs(  
    title = "Propensity Score Distribution Before and After Matching",  
    x = "Propensity Score",  
    y = "Density",  
    fill = "Group"  
) +  
  theme_bw()
```

최종 효과 분석

```
dta_m <- match.data(mod_match)  
t_test_result <- with(dta_m, t.test(overburden_numeric ~ CD_total))
```

분석 코드

CRE

분석 코드

CRE

변수 설정 (y : 결과, z : 치료, x : 공변량)

```
y <- df$overburden_yn  
z <- df$CD_total  
x <- df[, c("INCOME", "SEX", "MARR", "EDU", "DISA_YN", "HEALTH_INS_2.0", "HEALTH_INS_3.0", "HEALTH_INS_4.0",  
          "HEALTH_INS_5.0", "HEALTH_INS_6.0", "HEALTH_INS_8.0", "I_PHI1_YN", "EC01", "HS2_YN", "HS1", "P1",  
          "S3", "D1", "AGE", "BMI")]
```

파라미터 설정

CRE 실행

```
# CRE 실행  
set.seed(238)  
cre_results <- cre(y, z, x, method_params, hyper_params)
```

```
# Result  
# 결과 요약  
summary(cre_results)
```

분석 결과

TableOne

특성	만성질환 보유 여부			p-value ²
	Overall N = 3,605 ¹	만성질환 없음 N = 2,391 ¹	만성질환 보유 N = 1,214 ¹	
연령 (세)	45.8 ± 10.3	43.1 ± 10.3	51.0 ± 8.2	<0.001
체질량지수 (kg/m ²)	23.9 ± 3.5	23.6 ± 3.4	24.6 ± 3.5	<0.001
성별				0.9
여성	1,596 (44%)	1,061 (44%)	535 (44%)	
남성	2,009 (56%)	1,330 (56%)	679 (56%)	
혼인상태				<0.001
미혼	1,029 (29%)	739 (31%)	290 (24%)	
기혼	2,576 (71%)	1,652 (69%)	924 (76%)	
교육수준				<0.001
고졸 미만	458 (13%)	235 (9.8%)	223 (18%)	
고졸 이상	3,147 (87%)	2,156 (90%)	991 (82%)	
의료보장 형태				<0.001
직장가입자(본인)	2,195 (61%)	1,500 (63%)	695 (57%)	
건보: 직장피부양자	475 (13%)	310 (13%)	165 (14%)	
건보: 지역세대주	669 (19%)	417 (17%)	252 (21%)	

건보: 지역세대주	211 (5.9%)	142 (5.9%)	69 (5.7%)
의료급여 1,2종 세대주	41 (1.1%)	16 (0.7%)	25 (2.1%)
의료급여 세대원	13 (0.4%)	6 (0.3%)	7 (0.6%)
미가입	1 (<0.1%)	0 (0%)	1 (<0.1%)

경제활동 상태				0.3
비취업	71 (2.0%)	52 (2.2%)	19 (1.6%)	
취업	3,534 (98%)	2,339 (98%)	1,195 (98%)	

민간의료보험 가입				0.5
미가입	331 (9.2%)	226 (9.5%)	105 (8.6%)	
가입	3,274 (91%)	2,165 (91%)	1,109 (91%)	

장애 유무				<0.001
아니오	3,528 (98%)	2,366 (99%)	1,162 (96%)	
예	77 (2.1%)	25 (1.0%)	52 (4.3%)	

스트레스 인지				0.2
아니오	465 (13%)	320 (13%)	145 (12%)	
예	3,140 (87%)	2,071 (87%)	1,069 (88%)	

현재 흡연 여부				0.4
비흡연	2,664 (74%)	1,756 (73%)	908 (75%)	
현재 흡연	941 (26%)	635 (27%)	306 (25%)	

음주 빈도				0.007
월 2회 미만	1,637 (45%)	1,047 (44%)	590 (49%)	
월 2회 이상	1,968 (55%)	1,344 (56%)	624 (51%)	

규칙적 운동				0.2
아니오	1,958 (54%)	1,317 (55%)	641 (53%)	
예	1,647 (46%)	1,074 (45%)	573 (47%)	

우울감 경험				0.004
아니오	3,367 (93%)	2,254 (94%)	1,113 (92%)	
예	238 (6.6%)	137 (5.7%)	101 (8.3%)	

의료비 과부담 여부				<0.001
개인 소득	3,398.7 ± 2,634.6	3,431.0 ± 2,480.0	3,335.3 ± 2,915.3	0.009

¹ Mean ± SD; n (%)

² Wilcoxon rank sum test; Pearson's Chi-squared test

- 선택편향 존재 → PSM, IPTW

분석 결과

IPTW

IPTW 가중치 분포

가중치 분포 - 대부분 1-5 구간에 집중, 극단값 일부 존재하지만 최대값은 28.5로 관리 가능한 수준

[가중치 변동성]

처치군(CD=1): 변동계수 = 0.86

대조군(CD=0): 변동계수 = 0.41

→ 두 집단 모두 $CV < 1$ 로, 가중치 분포가 과도하게 불안정하지 않음을 의미

[유효 표본 크기(ESS)]

유효 표본 크기	치료 군	대조군
원자료 표본 수	1214명	2391명
IPTW 적용 후	699명	1361명

→ 표본 수는 감소하였으나, 여전히 statistical power 확보에 충분

공변량 균형성 검증

[SMD 기반 결과]

IPTW 적용 전 - 큰 불균형 존재 (예: BMI ≈ 0.30 , AGE ≈ 0.85 수준)

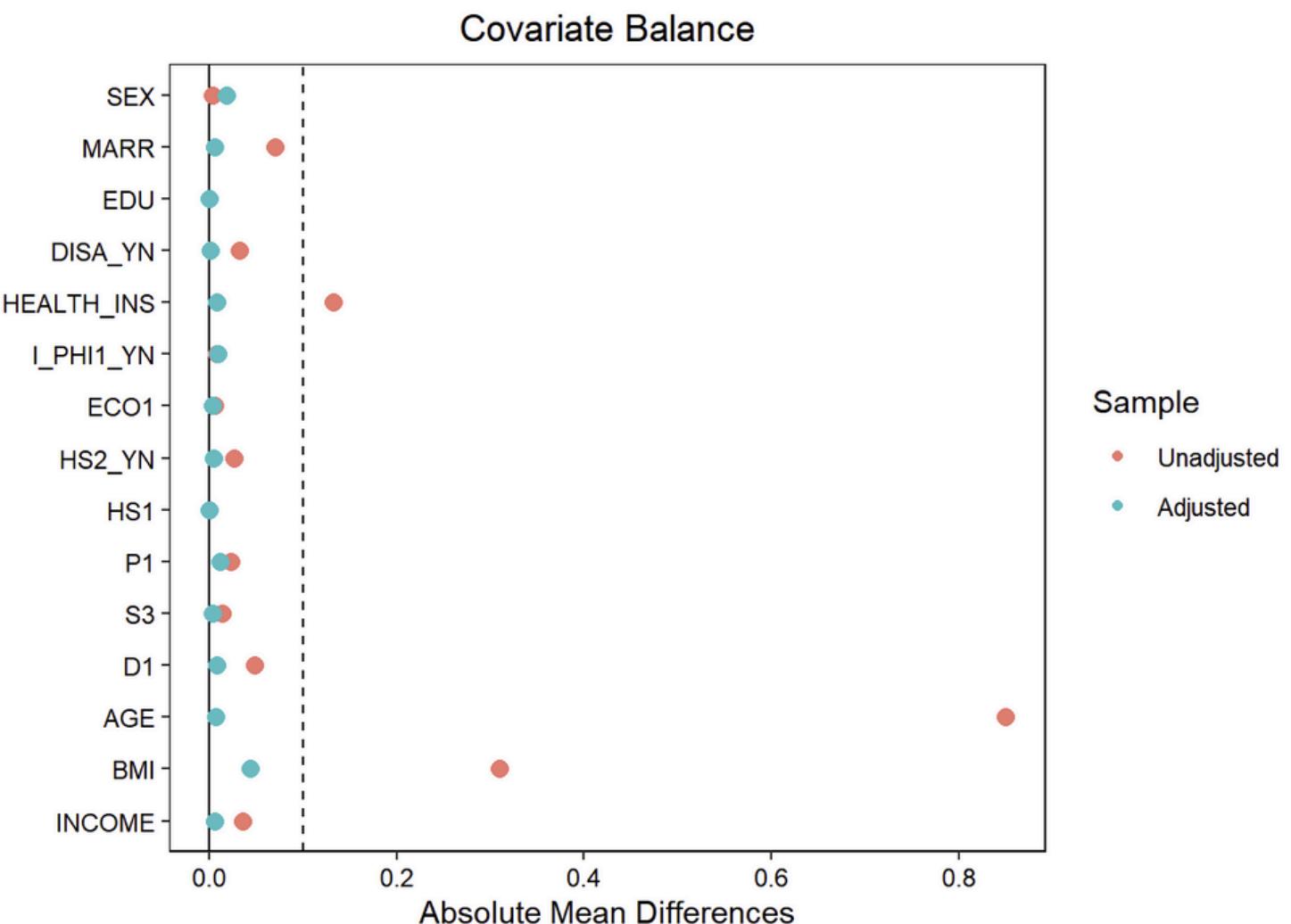
IPTW 적용 후 - 모든 변수의 $|SMD| < 0.1$ 로 감소, 가장 큰 값도 -0.03 수준 \rightarrow 균형 확보

[Love Plot]

적색(Unadjusted) 점은 일부 공변량에서 임계선(0.1)을 넘어섰으나,

청색(Adjusted) 점은 모두 임계선 내부로 수렴

\rightarrow 가중치 적용 효과 시각적으로 확인

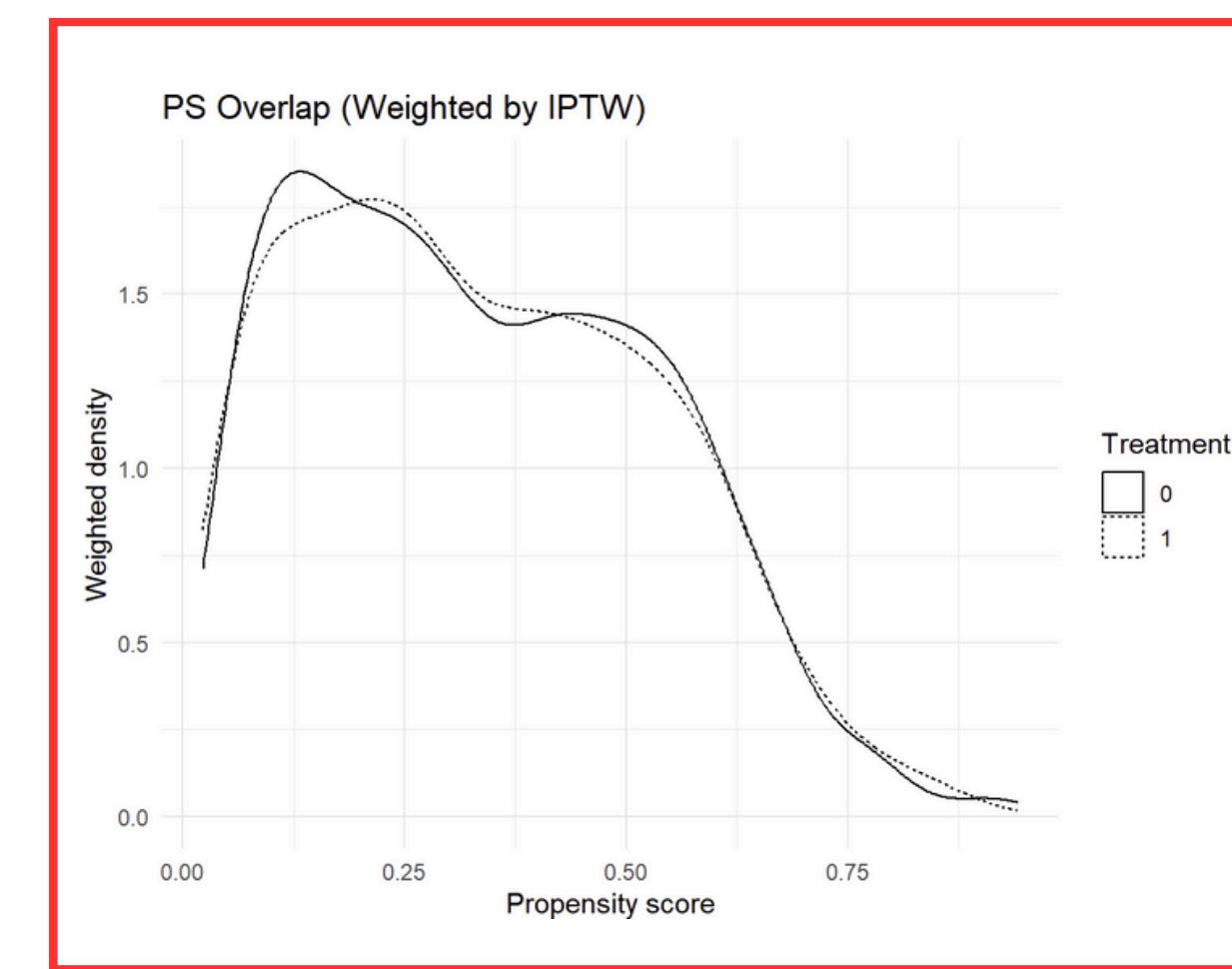
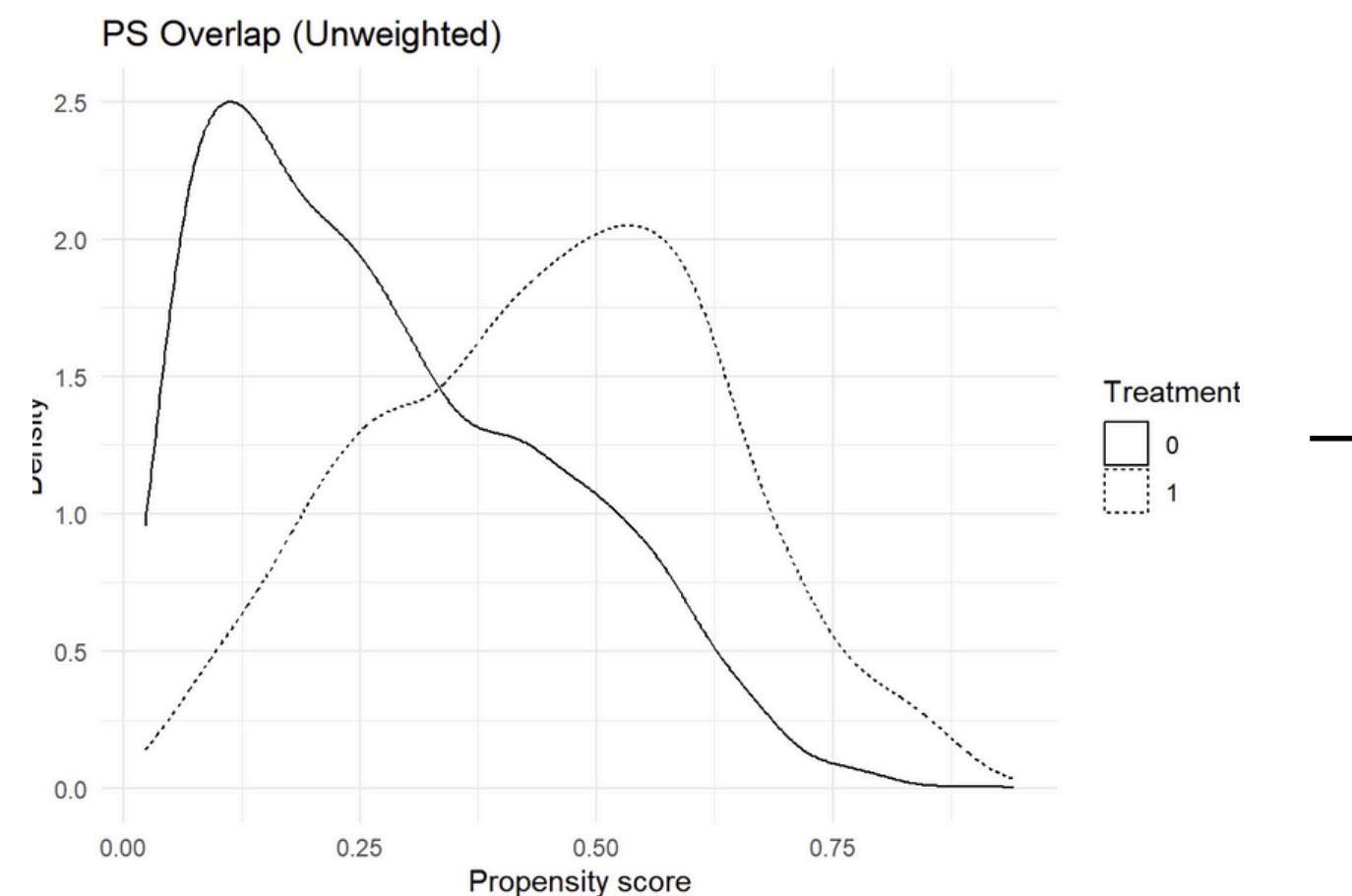


성향점수 분포 겹침도

목적: Positivity 가정 충족 여부 확인

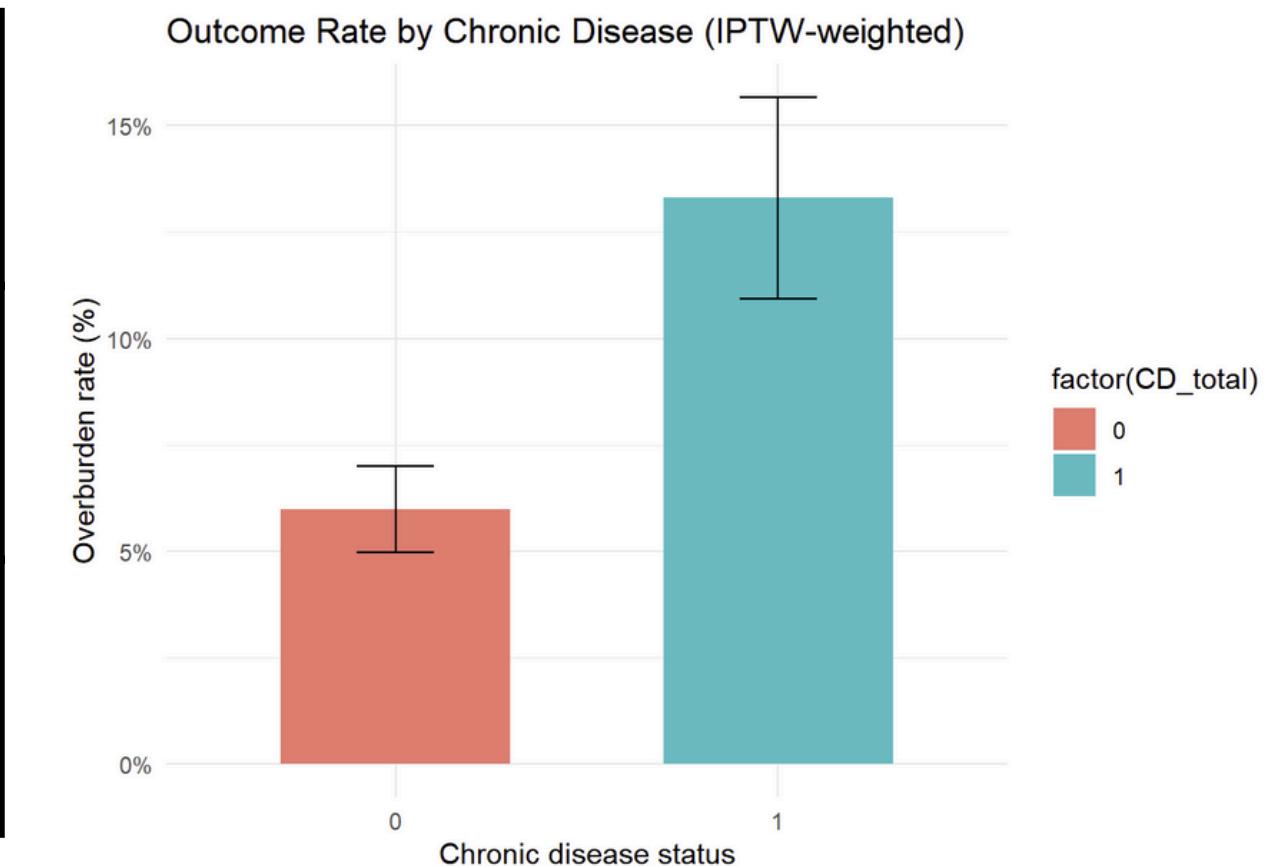
IPTW 전: 두 집단 분포 차이 뚜렷

IPTW 후: 분포가 겹쳐짐 → 공통 지지 영역 확보



ATE 추정 결과

CD_Total	Overburden_yn	Confidence interval
0	0.0597	0.0496 - 0.0698
1	0.1330	0.1094 - 0.1566



만성질환이 없는 가구에서는 약 6% 정도가 의료비 과부담을 경험하는 것으로 추정.
만성질환을 보유한 가구에서는 약 13.3%가 의료비 과부담을 경험하는 것으로 추정.
→ 만성질환 보유 여부가 과부담 발생 확률을 유의하게 증가시킴.

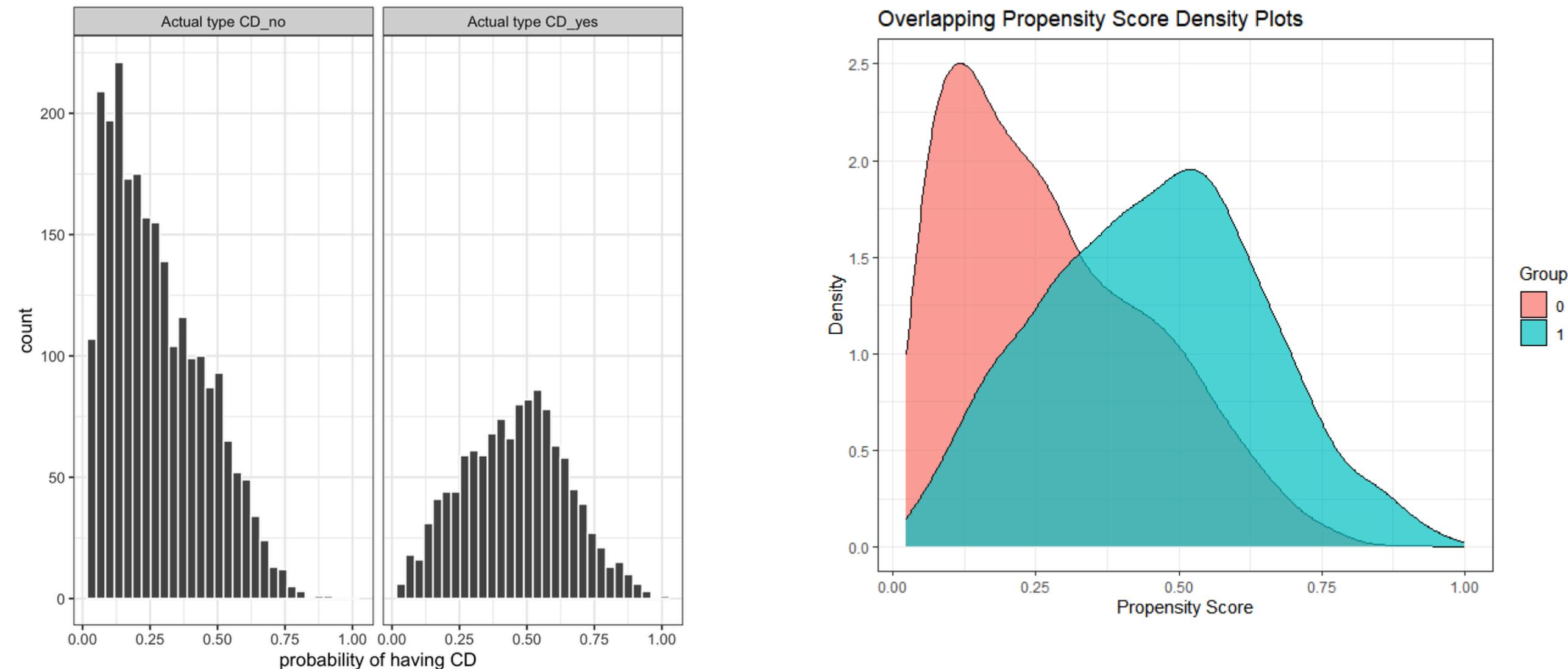
분석 결과

PSM

성향 점수 추정

- 성향 점수(PS) 추정: “어떤 사람이 만성질환에 걸릴 확률”
- 이 확률은 로지스틱 회귀(Logit) 모델로 추정
- 입력 변수: 사전 공변량들

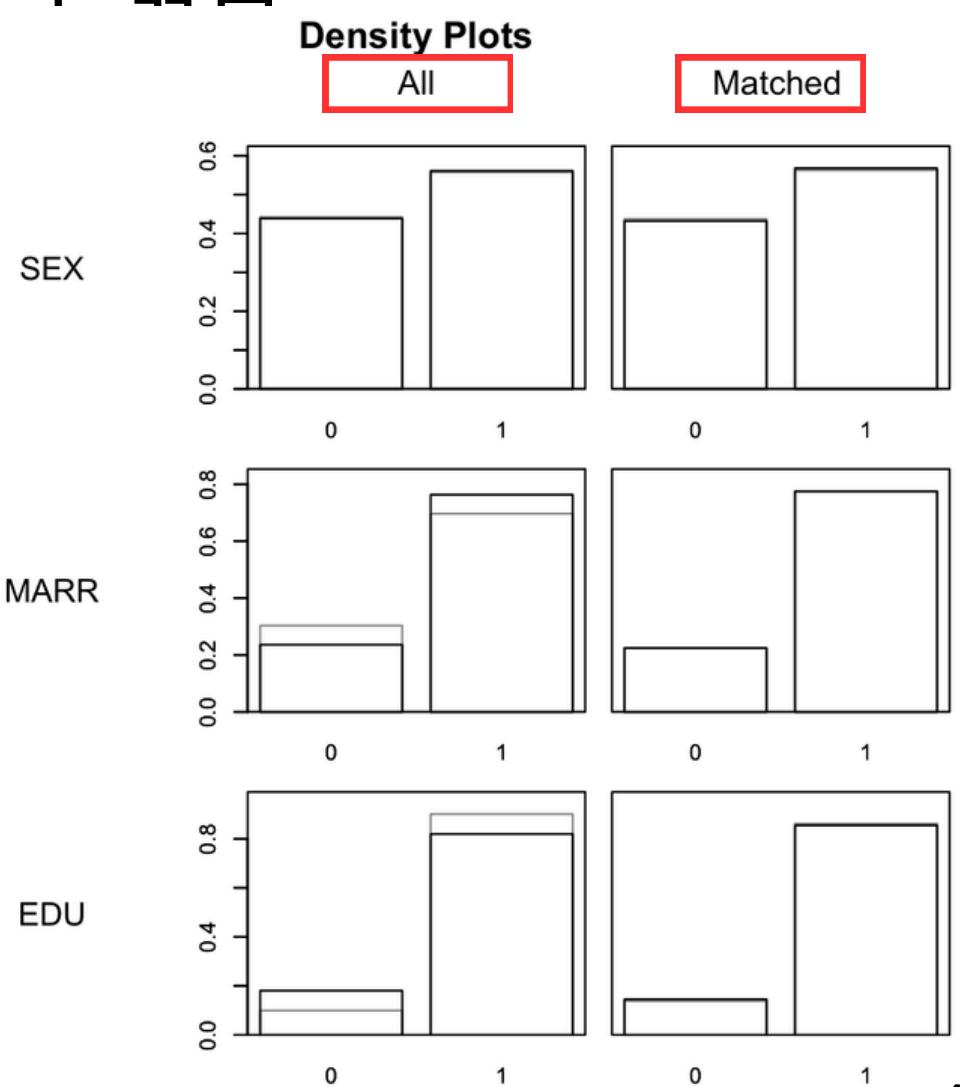
성향 점수 추정: 공통 지원 영역



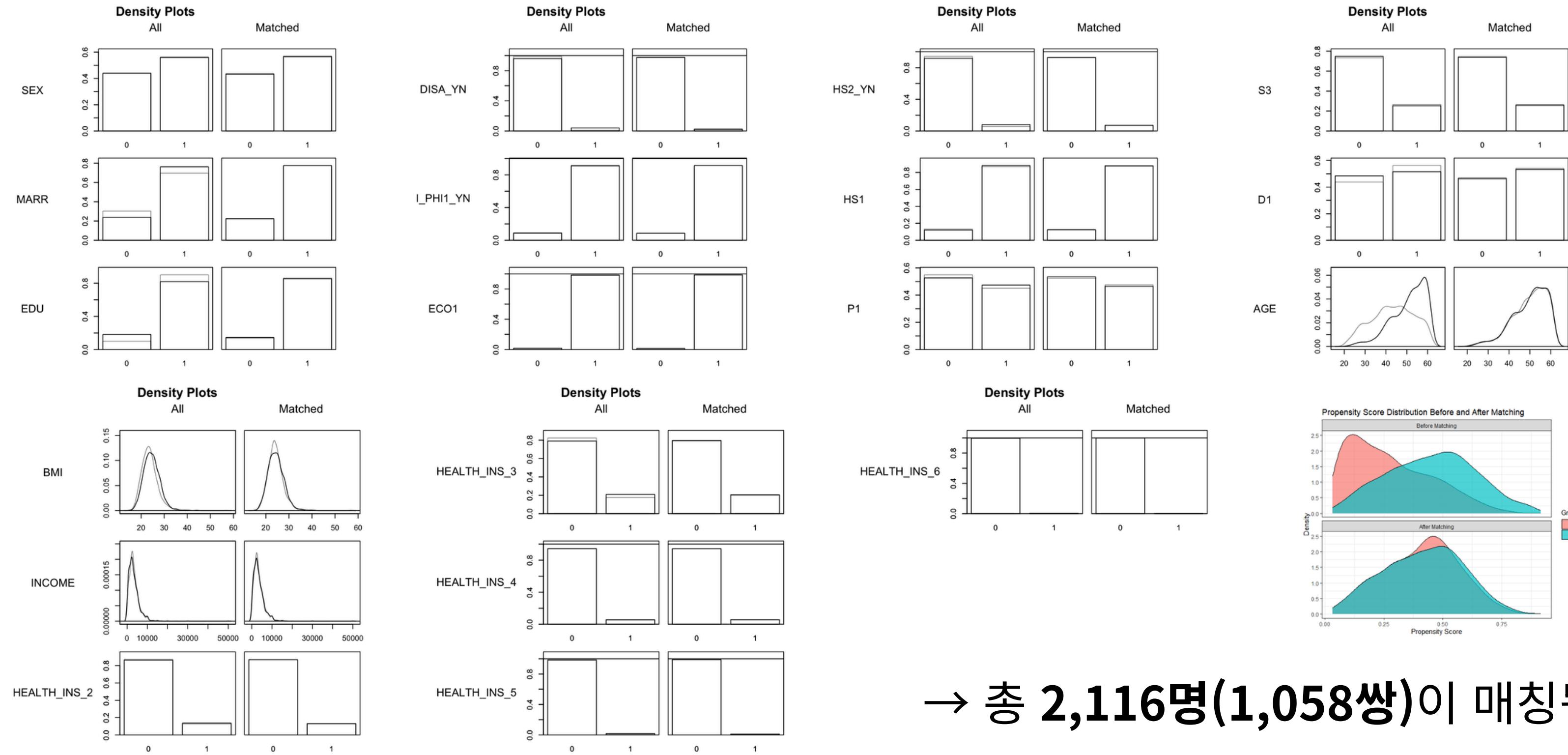
- 두 그룹의 점수가 겹치는 영역(공통 지원 영역)만 분석에 활용
- 약 0.0333~0.9153 구간에서 두 그룹이 모두 관측되어 이 구간에서 비교 진행

매칭 알고리즘(MatchIt) 실행

- MatchIt 패키지를 사용하여 실행
- 매칭 방법: 최근접 이웃 매칭(Nearest Neighbor Matching)
- 만성질환 유 · 무 사람 중, 성향 점수가 가장 비슷한 사람을 한 쌍으로 묶음
- All(매칭 전): 원래 데이터에서 두 집단(만성질환 유 · 무)의 분포
- Matched(매칭 후): 매칭된 표본에서 두 집단의 분포

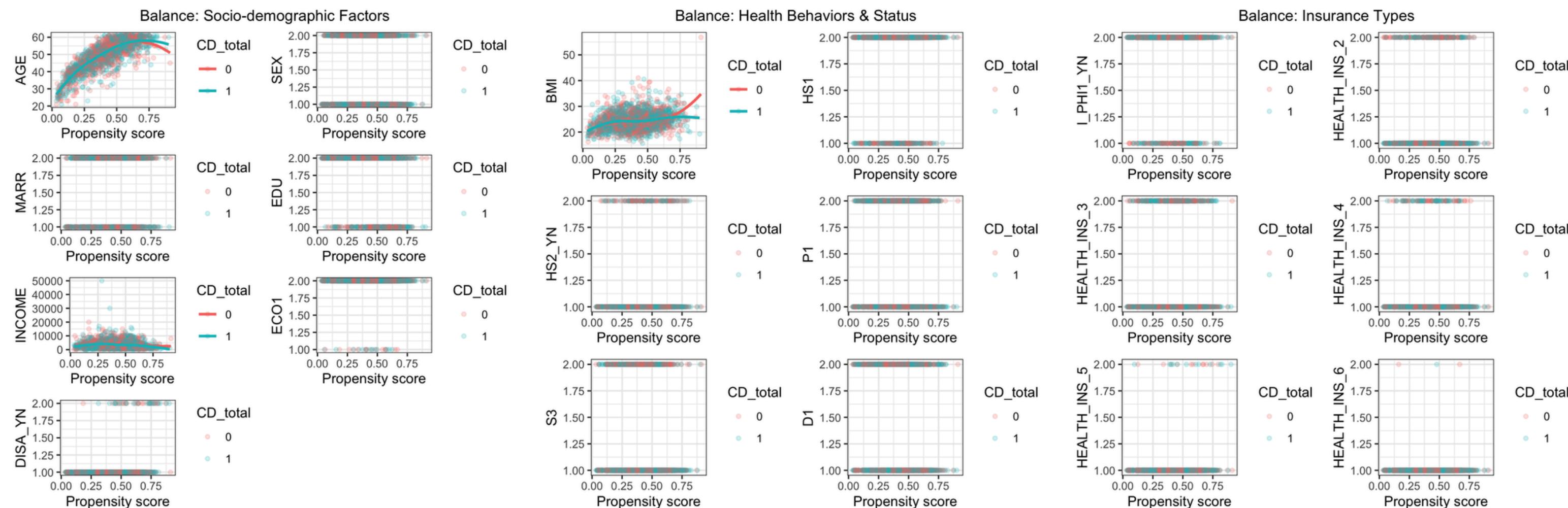


매칭 알고리즘(MatchIt) 실행



매칭된 샘플의 시각적 공변량 균형 검사

- **시각적 공변량 균형 검사:** 두 그룹이 얼마나 잘 겹쳐있는지 보는 것
- 시각화 결과, 모든 변수에서 두 선이 거의 완벽하게 겹쳐짐
→ 성향 점수(PS)의 모든 구간에 걸쳐 두 집단의 특성이 동일해졌음을 나타냄



평균처치효과(ATE) 추정

t-검정 → 매칭된 데이터에서 두 그룹의 결과 변수 평균 차이 계산

Welch Two Sample t-test

```
data: overburden_numeric by CD_total
t = -6.2473, df = 1860.9, p-value = 5.161e-10
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
-0.1055616 -0.0551189
sample estimates:
mean in group 0 mean in group 1
0.05860113 0.13894140
```

만성질환이 없는 가구에서는 약 5% 정도가 의료비 과부담을 경험하는 것으로 추정.
만성질환을 보유한 가구에서는 약 13.8%가 의료비 과부담을 경험하는 것으로 추정.
→ 만성질환 보유 여부가 과부담 발생 확률을 유의하게 증가시킴.

분석 결과

CRE

Summary 기반 CRE 분석 결과

[필터링 후 선택된 규칙]

후보 규칙 10,000개 생성 후, 다음 기준으로 필터링

- Irrelevant(무관한 규칙): 596개 제거
- Extreme(극단값 규칙): 396개 제거
- Correlated(상관 높은 규칙): 209개 제거

최종 LASSO 선택: 31개

→ 통계적으로 유의한 규칙 9개가 나타남

CAUSAL RULE ENSAMBLE - Summary

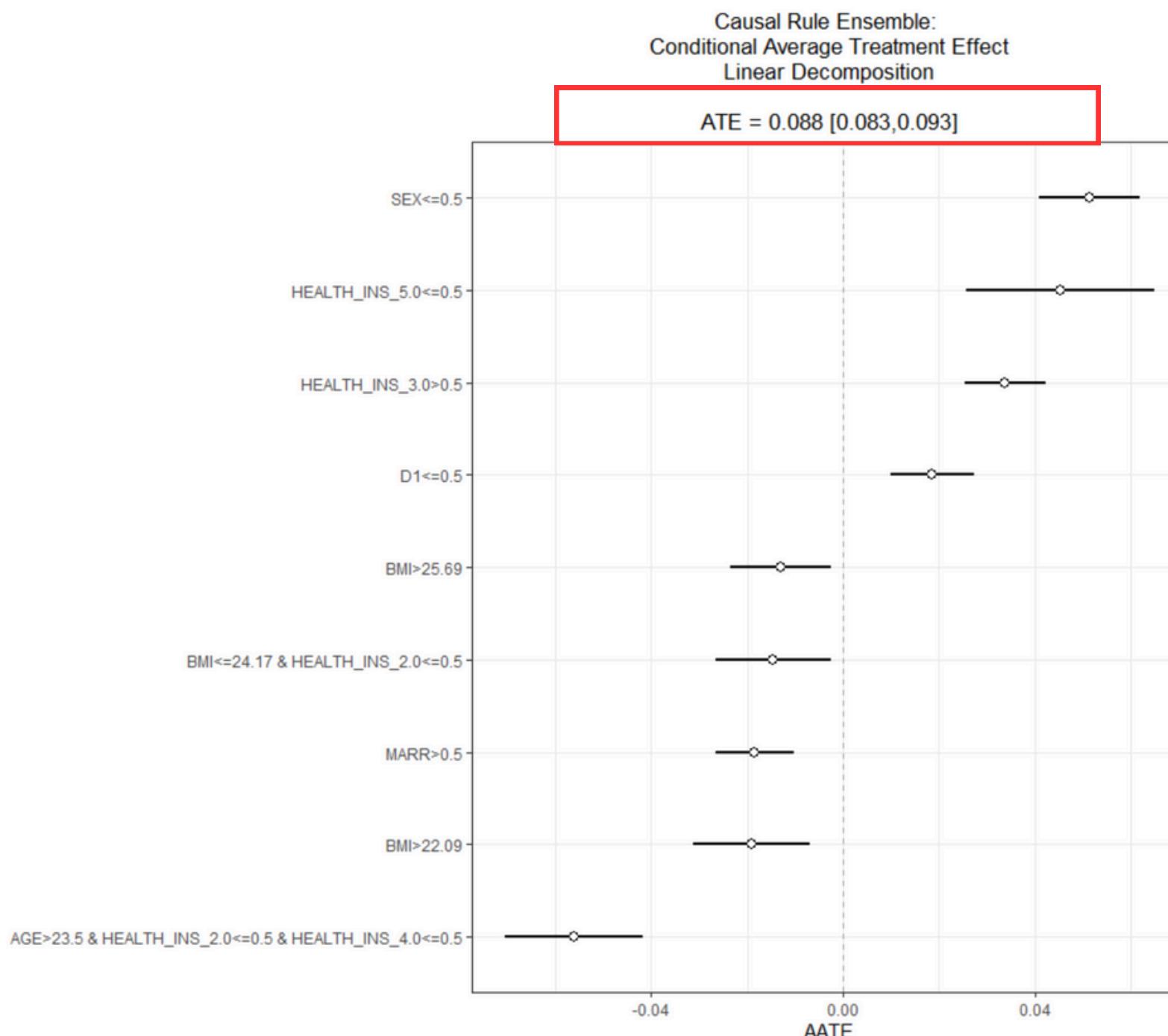
Model parameters

- Pseudo-Outcome estimation
 - Estimator : cf
 - Outcome : NA
 - Propensity Score: SL.glmnet
- Rules Generation
 - Intervention Variables: SEX MARR EDU DISA_YN I_PHI1_YN HS2_YN HS1_P1_S3
 - Number of Trees : 10000
 - Node Size : 5
 - Max Rules : 10000
 - Max Depth : 5
- Filtering
 - Threshold Decay (Irrelevant): 0.01
 - Threshold (Extreme) : 0.01
 - Threshold (Correlated) : 1
 - Threshold (p-Value) :
- No Stability Selection (only LASSO)

Rules

- Initial : 10000
- Filter (irrelevant) : 596
- Filter (extreme) : 396
- Filter (correlated) : 209
- Select (LASSO) : 31
- Select (significant) : 9

전체 평균처치효과(ATE) 분석 결과



ATE = 0.088

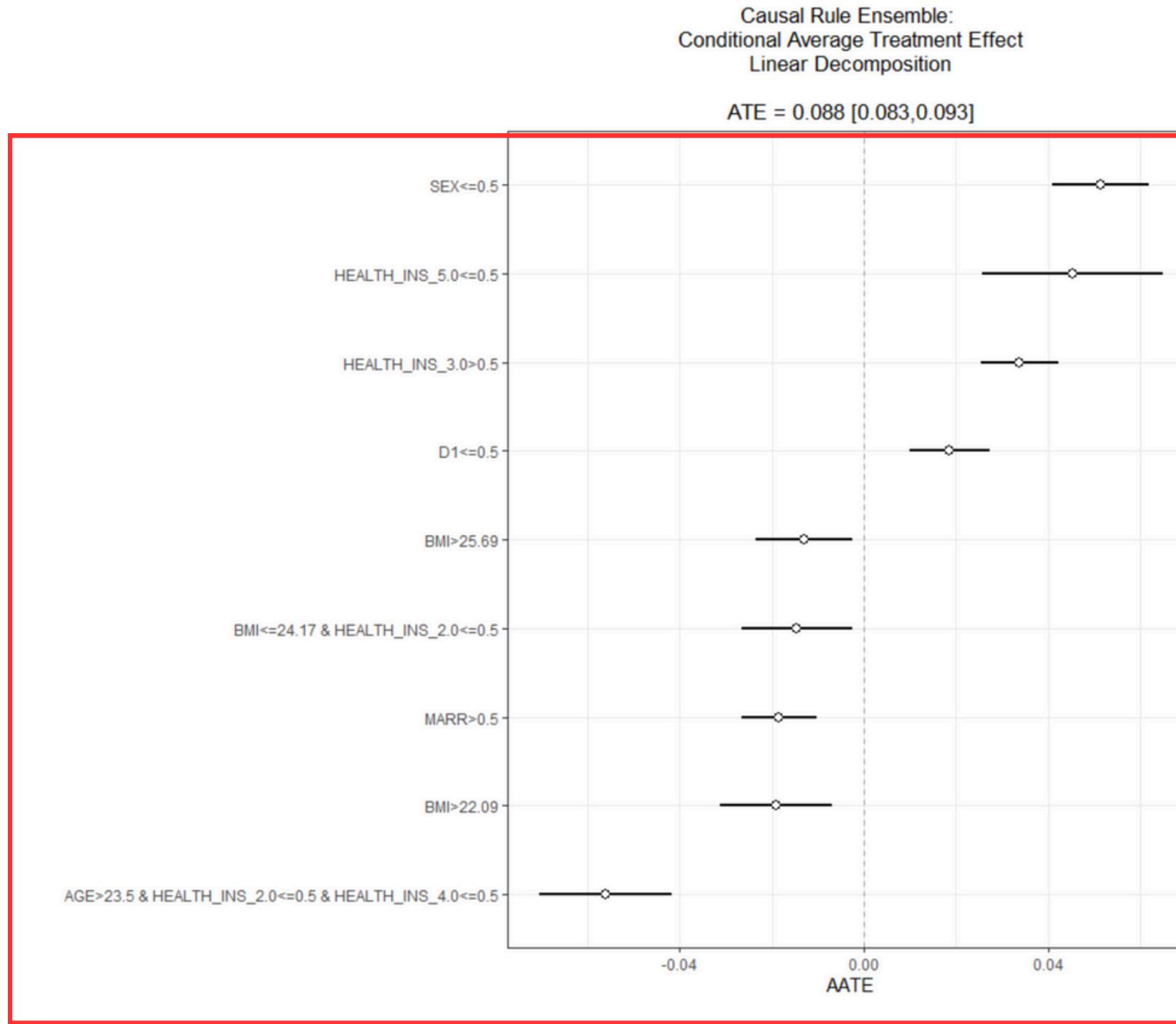
95% 신뢰구간 = [0.083, 0.093]

“만성질환을 보유한 사람은 보유하지 않은 사람에 비해 의료비 과부담을 경험할 확률이 평균적으로 8.8%p 더 높음”

→ IPTW, PSM 결과와 유사하게 추정

→ 만성질환 보유 시 의료비 과부담 발생 확률이 유의하게 증가함을 확인

주요 하위집단별 처치효과(CATE) 분석 결과



결과 변수: 이진형

→ ‘해당 조건을 만족하는 집단에서 만성질환이 의료비 과부담 발생 확률에 미치는 평균적 영향’이라고 해석해야 함

통계적으로 유의한 규칙 9개 도출 ($P\text{-value} < 0.05$)

- 단일 규칙: 7개
- 복합 규칙: 2개

Results			
- CATE Linear Decomposition:			
Rule	Estimate	CI_lower	CI_upper
(ATE)	0.08788490		
SEX<=0.5	0.05134887		
HEALTH_INS_3.0>0.5	0.03383062		
BMI>22.09	-0.01908939		
D1<=0.5	0.01852154		
BMI>25.69	-0.01303544		
BMI<=24.17 & HEALTH_INS_2.0<=0.5	-0.01454500		
MARR>0.5	-0.01847668		
HEALTH_INS_5.0<=0.5	0.04522546		
AGE>23.5 & HEALTH_INS_2.0<=0.5 & HEALTH_INS_4.0<=0.5	-0.05627070		
		P_Value	
1	0.083073232	0.092696559	0.000000e+00
2	0.040901827	0.061795913	0.000000e+00
3	0.025420084	0.042241155	7.771561e-15
4	-0.031156832	-0.007021948	1.982356e-03
5	0.009825336	0.027217741	3.228936e-05
6	-0.023594906	-0.002475981	1.570413e-02
7	-0.026604886	-0.002485114	1.826209e-02
8	-0.026601436	-0.010351916	9.174171e-06
9	0.025580568	0.064870360	7.124405e-06
10	-0.070650534	-0.041890865	3.863576e-14

주요 하위집단별 처치효과 분석 결과 해석

하위집단 조건	CATE
여성 (SEX ≤ 0.5)	+0.051
직장 건강보험 지역가입자 (HEALTH_INS_3 > 0.5)	+0.034
과체중 (BMI > 22.09)	-0.019
음주하지 않는 집단 (D1 ≤ 0.5)	+0.019
고도비만 (BMI > 25.69)	-0.013
정상체중 & 비피부양자 (BMI ≤ 24.17 & HEALTH_INS_2 ≤ 0.5)	-0.015
기혼 (MARR > 0.5)	-0.018
지역보험 가입자 아님 (HEALTH_INS_5 ≤ 0.5)	+0.045
상대적으로 낮은 연령 + 비피부양자 + 지역보험 세대원 아님 (AGE > 23.5 & HEALTH_INS_2 ≤ 0.5 & HEALTH_INS_4 ≤ 0.5)	-0.056

[경제적 취약성]

- 여성, 음주하지 않는 집단, 지역보험 가입자
→ 부담 증가

[경제적/사회적 안정성]

- 기혼, 비피부양자, 젊은 독립적 집단
→ 부담 감소

[BMI 관련 효과]

- 과체중·고도비만 집단은 단순 선형 관계가 아님
을 예측 가능
→ 일부 구간에서 부담 감소

결론

결론

[주요 결과]

- 평균적으로 만성질환 보유 시 과부담 발생 확률 약 8%p 증가
- CRE 분석에서 총 9개의 유의미한 규칙 확인
→ 집단별 효과 차이가 존재함을 확인 가능

결론

[의의]

- 만성질환 보유 여부가 의료비 과부담 발생 확률을 유의하게 증가시킴을 확인
- CATE 분석을 통해 집단별 이질적 효과 규명 → 평균 효과를 넘어선 세부적 이해 가능

[한계점]

- 표본 수 제한으로 일부 하위집단의 추정 안정성 낮을 수 있음
- 자기보고식 자료 활용으로 측정오차 가능성 존재
- 관찰자료 기반 분석 → 인과적 해석에 잔여 교란 가능성

End of Summer URP

Summer URP 소감

강재서

- Summer URP reflection
- Other achievements during the summer
 - 2025 문화 디지털혁신 및 데이터 활용 공모전
 - 2025 대국민 지하수 빅데이터 공모전
 - 수리통계학 스터디
 - TEPS 영어 공부
 - 체력 단련
- Future career plans (or the next semester)
 - 빅데이터분석기사 자격증 공부
 - 내년 전기 목표로 대학원 진학

Summer URP 소감

신아영

- Summer URP reflection
- Other achievements during the summer
 - 제 3회 BDA 채용 연계 공모전 참가
 - 수리통계학 스터디 진행
 - 파이썬 기초 공부
- Future career plans (or the next semester)
 - 대학원 진학 준비
 - 다양한 공모전 출전 계획
 - 영어 공부 (OPIC 등)
 - 빅데이터분석기사 공부

Summer URP 소감

양인경

- Summer URP reflection
- Other achievements during the summer
 - ADSP 자격증 시험
 - 정보처리기사 필기 시험
 - 계절학기 수강
 - 졸업 프로젝트를 위한 연구 주제 고민 및 연구실 컨택
 - 영어 공부
- Future career plans (or the next semester)
 - 컴퓨터공학과 졸업을 위한 연구실 참여 및 졸업 프로젝트 시작
 - 수학과 졸업 시험 준비
 - 공인 영어 시험

Summer URP 소감

유보현

- Summer URP reflection
- Other achievements during the summer
 - ADSP 자격증 시험
 - FBA Quant 학회
- Future career plans (or the next semester)
 - GitHub에 프로젝트를 문서화 / 기술 블로그 월 1편 작성
 - 학부생 연구·캡스톤·학회 활동을 통해 연구 설계와 발표 경험 확장
 - 전공 교과목 복습 및 심화학습으로 기초 이론·수리적 기반 강화

Summer URP 소감

주수현

- Summer URP reflection
- Other achievements during the summer
 - ADSP 자격증 시험
 - 영어 공부
 - 투운사 스터디
 - 경제 공부
- Future career plans (or the next semester)
 - 빅데이터분석기사 자격증 시험 공부
 - 투자자산운용사 시험 공부
 - 대학원 진학 준비
 - 통계 졸업시험 준비