# ‹건강검진코호트DB Covariate Missing rate 정리›

- 2월 21일 Version
- 작성자 : 이은경

‹What TO DO›

: 다시 정의한 Study population을 바탕으로, Covariate 정의한 후, 결측 비율 확인 작업 진행

‹Share & Result›

1) "Covariates" Define part

1)-① "Demographic" 관련 Covariate : Missing value는 없는 것으로 확인.

1)-② "가족력"

| Family History | Missing obs / N |
|---|---|
| Heart Disease | 0 / 168,339 |
| Diabetes | 0 / 168,339 |
| Cancer | 0 / 168,339 |

1)-③ "Biomarker" Covariate

| Biomarker | Missing obs / N |
|---|---|
| TOT_CHOLE | 168 / 168,339 (0.099%) |
| BLDS | 2 / 168,339 (0.0011%) |
| OLIG_PROTE_CD | 641 / 168,339 (0.38%) |
| SGOT_AST | 30 / 168,339 (0.017%) |
| SGPT_ALT | 29 / 168,339 (0.017%) |
| GAMMA_GTP | 26 / 168,339 (0.015%) |

1)-④ "음주습관" Covariate

: Missing인 객체의 수는 4,868명 (2.89%)

1)-⑤ "운동습관"(빈도) Covariate

: Missing인 객체의 수는 4,007명 (2.38%)

1)-⑥ "운동지속시간"(METS Minutes) Covariate

: Missing인 객체의 수는 16,130명 (9.58%)

1)-⑦ "흡연상태", "흡연빈도" 관련 Covariate

: "흡연상태" Missing인 객체의 수는 361명이다.(0.214%)

: "현재" 흡연자 대상, 아래 공변량이 Missig인 객체의 수는 표와 같다.

| Covariate | Missing |
|---|---|
| 흡연 기간 | 364 |
| 하루 흡연량 | 364 |
| Pack year | 364 |

↳ “과거”에 흡연했던 객체들 대상, 아래 공변량이 Missing인 객체의 수는 표와 같다.

| Covariate | Missing |
|---|---|
| 흡연기간 | 363 |
| 하루 흡연량 | 13,291 |
| Pack year | 13,291 |

1)-⑧ “SODA” 관련 Covariate
: 07~08년도 생애구강검진DB & 14~19년도 일반 + 생애구강검진DB로 “SODA” 변수 정의한 후, “SODA” 변수의 요약 통계량 확인
↳ 위 과정에서 얻은 요약 통계량 치를 이용해 02~08년도 구강검진DB 변수 값 대체 방안 설립
-- 1(그렇다) -> 0.1538 (3) 과정에서 얻은 SODA 변수 값 중 3분위수 이상 값들의 평균치)
-- 2(아니다) -> 0 (3) 과정에서 얻은 SODA 변수 값 중 1분위수 이하 값들의 평균치)
-- 3(모르겠다) -> 0 (3) 과정에서 얻은 SODA 변수 값 중 1분위수 이상 3분위수 이하 값들의 평균치)

∴ “SODA” Covariate 가 결측인 객체는 총 168,339명 중 164,081명이며,
해당 공변량의 Missing rate는 약 97.47%이다

1)-⑨ “SNCK” 관련 Covariate
: 07~08년도 생애구강검진DB & 14~19년도 일반 + 생애구강검진DB로 “SODA” 변수 정의한 후, “SNCK” 변수의 요약 통계량 확인
↳ 위 과정에서 얻은 요약 통계량 치를 이용해 02~08년도 구강검진DB 변수 값 대체 방안 설립
-- 1(그렇다) -> 1.391 (3) 과정에서 얻은 SNCK 변수 값 중 3분위수 이상 값들의 평균치)
-- 2(아니다) -> 0 (3) 과정에서 얻은 SNCK 변수 값 중 1분위수 이하 값들의 평균치)
-- 3(모르겠다) -> 0.262 (3) 과정에서 얻은 SNCK 변수 값 중 1분위수 이상 3분위수 이하 값들의 평균치)

∴ “SNCK” Covariate 값이 결측인 객체는 총 171,084명 중 92,751명이며,
해당 공변량의 Missing rate는 약 55.09%이다.

1)-⑩ “과일, 채소 섭취 빈도” Covariate
: 07~08년도 생애구강검진DB에서 “과일, 채소 섭취 빈도” 변수가 정의되는 객체 수는 없음.
즉, 이 변수의 Missing rate는 100%이다.

1)-⑪ “개인과거병력” Covariate
[병력 변수 값이 1인 관측치 비율]

| 개인과거병력 | 객체 수 |
|---|---|
| Myocardical infraction | 0 / 168,339 |
| Heart failure | 0 / 168,339 |
| TIA or Stroke | 0 / 168,339 |
| Hypertension | 55,970 / 168,339 |
| Cancer | 0 / 168,339 |

<개인과거병력 변수 값이 1인 객체가 없는 이유 추측>
: “MI”, “Heart failure”, “Cancer” 과거 병력은 T20만을 이용해 정의하는 부분인데, 이미 Study population 정의할 때 Cohort entry date 이전에 해당 병력이 있는 사람은 제외하였기 때문이라 생각한다.
/ 반면, Hypertension은 제외조건에 포함하지 않았던 병명이며 “TIA or Stroke”는 T20이 아닌 검진DB 통해 파악된 부분이다.

1)-⑫ “Comorbidites” Covariate
[변수명]

| diagnosis | table 변수명 |
|---|---|
| **hypertension** | **Hypertension_result** |
| Asthma | Asthma_result |
| Chronic obstructive pulmonary disease (COPD) | COPD_result |
| Atrial fibrillation | Atrial_fibrillation_result |
| Thromboembolism | Thromboembolism_result |
| Chronic liver disease | CLD_result |
| Chronic Kidney disease | CKD_result |
| Coronary artery disease (동맥경화, 심장혈관질환) | CAD_result |
| Peripheral vascular disease | PVD_result |
| Dementia | Dementia_result |
| GI disorders | GI_disorders_result |
| Hyperlipidemia | Hyperlipidemia_result |
| Pneumonia | Pneumonia_result |
| Psychiatric disorders | Psychiatric_disorders_result |

[병력 변수 값이 1인 관측치 비율]

| 변수명 | 객체 수 |
|---|---|
| Asthma_result | 19,908 / 168,339 |
| COPD_result | 31,162 / 168,339 |
| Atrial_fibrillation_result | 779 / 168,339 |
| Thromboembolism_result | 1,129 / 168,339 |
| CLD_result | 25,480 / 168,339 |
| CKD_result | 343 / 168,339 |
| CAD_result | 10,803 / 168,339 |
| PVD_result | 11,238 / 168,339 |
| Dementia_result | 500 / 168,339 |
| GI_disorders | 100,952 / 168,339 |
| Hyperlipidemia_result | 23,394 / 168,339 |
| Pneumonia_result | 7,394 / 168,339 |
| Psychiatric disorders_result | 25,871 / 168,339 |
| Hypertension_result | 141,293 / 168,339 |

1)-⑬ “Medication” Covariate

[약물 이름 - table 변수명 pair]

| Medication | table 변수명 |
|---|---|
| insulin | insulin_result |
| 경구당뇨약 | oral_diabetes_result |
| Non-statin antihyperlipidemic | Non_statin_result |
| statin | statin_result |
| 혈압약 (renin angiotensin system (교감신경) pathway 관련) | Hypertension_renin_result |
| 혈압약 (RAS 제외) | Hypertension_except_RAS |
| CNS (central nerve system) | cns_result |
| 진통제(non-opioid analgesic) | non_opioid_result |
| Anticoagulants | Anticoagulants result |
| Antiplatelets including Aspirin | Antiplatelets_result |
| immunosuppressant | immunosuppressant result |

[병력 변수 값이 1인 관측치 비율]

| 변수명 | 객체 비율 |
|---|---|
| insulin_result | 2,370 / 168,339 |
| oral_diabetes_result | 21,145 / 168,339 |
| Non_statin_result | 4,303 / 168,339 |
| statin_result | 19,896 / 168,339 |
| Hypertension_renin_result | 24,937 / 168,339 |
| Hypertension_except_RAS | 60,095 / 168,339 |
| Cns_result | 95,112 / 168,339 |
| non_opioid_result | 135,229 / 168,339 |
| Anticoagulants_result | 398 / 168,339 |
| Antiplatelets_result | 16,957 / 168,339 |
| immunosuppressant_result | 662 / 168,339 |