<산업 안전 보건 연구원 - Meeting 자료>

- 6월 21일 Version

- 작성자 : 이은경

<What TO DO>

: EDA 진행 통해 얻은 데이터 이용해 Demographic Information(생년월일, 성별, 사망 일자 + 사망 여부) 정의, "entry" 변수 정의, Outcome(폐암 / 백혈병 발병 여부), Cancer History 변수 정의

<Share & Result>

1) Demographic Information 가져오기

1)-① 생년월일 + 성별 + 최초 고용보험 등록일 정보 가져오기

: EDA 진행한 데이터를 "INDI_ID", "ECNY_DT" 기준으로 정렬

: "first.INDI_ID"만 가져와 최초 고용보험 등록 일자 확인(distinct INDI_ID = 29,189,766)

↳ 생년월일과 성별은 time-invariant covariate 이므로 모든 반복 측정치에 대해 값이 동일

1)-② 사망 일자 + 사망 여부 정의

: 모든 반복 측정치에 대해 사망 일자(DTH_DATE1 ~ DTH_DATE3) 변수값이 모두 동일하므로 각 객체 당 첫 번째 반복 측정치의 사망 일자 변수 이용해 해당 변수 값이 존재하면 사망으로 간주

1)-③ "Entry" 변수 정의

: 1)-①을 통해 각 객체 별 최초 고용보험 등록일(ECNY_DT)을 얻었으므로 "ECNY_DT" 변수의 맨 앞 4자리만 가져와 해당 값 이용해 "Entry" 변수 정의

(변수 정의 방식)

최초 고용보험 등록 연도	Entry 변수 값
1995 ~ 1999년	0 (Baseline)
2000 ~ 2004년	1
2005 ~ 2009년	2
2010 ~ 2014년	3
2015 ~ 2018년	4

∴"mine.Demographic_Info" table

(생성한 Data example)

성별	INDI_ID	BYEAR	DTH_DATE	Entry	Death
남	1000000757001	1940	19990521	0	1
남	1000000757002	1940	19980106	0	1
남	1000000757003	1940		2	0
남	1000000757004	1940		0	0
남	1000000757005	1940		0	0
남	1000000757006	1940		0	0
남	1000000757007	1940		0	0
남	1000000757008	1940		2	0
남	1000000757009	1940	20160703	0	1

2) Outcome(폐암 / 백혈병 발병 여부) + Cancer History 정의

: 해당 변수는 모두 fdx1-fdx6, icd10_1 ~ icd10_6 변수 이용해 정의

↳ 정의하고자 한 변수는 "lung_cancer"(폐암 발병 여부), "lung_cancer_date"(폐암 진단 일자), "leukemia"(백혈병 발병 여부), "leukemia_date"(백혈병 진단 일자), "Cancer_History"(암 이력 여부)

2)-① 폐암 / 백혈병 발병 여부 + 진단일자 변수 정의

(코드 전개 : 폐암 예시)

① "icd10_1"이 "C34%"이면, "lung_cancer"=1 대입 + "lung_cancer_date"는 "fdx1"로 대입

② "icd10_1"은 "C34%"가 아닌데, "icd10_2"가 "C34%"이면, "lung_cancer"=1 + "lung_cancer_date"는 "fdx2" 값 대입

... 해당 과정을 "icd10_6"까진 진행한 후, 마지막으로 "lung_cancer"값이 결측인 관측치는 폐암 이력이 존재하지 않는 것을 의미하므로 "lung_cancer" = 0 대입 -- "백혈병"에 대해서도 동일한 방식으로 진행

```
/* lung cancer */
IF substr(icd10_1,1,3) = "C34" then do;
       lung_cancer=1;
       lung_cancer_date=fdx1;
end;
ELSE IF substr(icd10_2,1,3) = "C34" then do;
       lung_cancer=1;
       lung_cancer_date=fdx2;
end;
ELSE IF substr(icd10_3,1,3) = "C34" then do;
       lung_cancer=1;
       lung_cancer_date=fdx3;
end;
ELSE IF substr(icd10_4,1,3) = "C34" then do;
       lung_cancer=1;
       lung_cancer_date=fdx4;
end;
ELSE IF substr(icd10_5,1,3) = "C34" then do;
       lung_cancer=1;
       lung_cancer_date=fdx5;
end;
ELSE IF substr(icd10_6,1,3) = "C34" then do;
       lung_cancer=1;
       lung_cancer_date=fdx6;
end;
ELSE lung_cancer=0;
```

2)-②. 암 이력 변수 정의

: 해당 변수는 "fdx1"(첫 번째 암 발병 시 진단일자)가 결측이 아니면서 "lung_cancer"(폐암 발병 여부) = 0 이고 "leukemia"(백혈병 발병 여부) = 0인 관측치 대상으로 "Cancer_History"(암 이력 존재 여부) = 1 부여 / 나머지 관측치에 대해서는 모두 0 값 부여

∴ 2) 과정 통해 “mine.Cancer” tbl 얻음

(“mine.Cancer” tbl example)

INDI_ID	lung_cancer	lung_cancer_date	leukemia	leukemia_date	Cancer_History
1000000757026	1	20120725	0		0
1000000757027	1	20030112	0		0
1000000757028	0		0		0
1000000757029	0		0		0
1000000757030	0		0		0
1000000757031	0		0		1
1000000757032	1	20081122	0		0
1000000757033	0		0		1
1000000757034	0		0		0
1000000757036	0		0		1

---

3) 1), 2) 통해 생성한 두 table 합침

(example)

성별	INDI_ID	BYEAR	DTH_DATE	Entry	Death	lung_cancer	lung_cancer_date	leukemia	leukemia_date	Cancer_History
남	1000000757001	1940	19990521	0	1	0		0		1
남	1000000757002	1940	19980106	0	1	0		0		0
남	1000000757003	1940		2	0	0		0		0
남	1000000757004	1940		0	0	0		0		0
남	1000000757005	1940		0	0	0		0		0
남	1000000757006	1940		0	0	0		0		0
남	1000000757007	1940		0	0	0		0		0
남	1000000757008	1940		2	0	0		0		1
남	1000000757009	1940	20160703	0	1	0		0		0
남	1000000757010	1940	20160713	2	1	0		0		0
남	1000000757011	1940	20011125	1	1	0		0		0

---

4) Study population에 포함되는 객체 대상으로, Outcome 발생 비율 + 사망 사건 발생 비율 파악

Outcome	빈도	백분율
폐암 발병	90,523	90,523 / 29189766 (0.31%)
백혈병 발병	17,350	17,350 / 29189766 (0.06%)
사망	808,940	808,940 / 29189766 (2.77%)

↳ 사건 발생 수가 많을 것이라 예상했던 “폐암”도 사건 발생 수가 그다지 많지는 않다.