<산업보건연구원 - Meeting 자료>

- 6월 27일 Version
- 작성자 : 이은경

<What TO DO>

: EDA 재진행한 Data Double-check

: 다시 EDA 진행한 Data 기반으로 "Entry" 변수, "사망 여부", "사망 일자", Outcome(폐암 / 백혈병 발병 여부 + 진단 일자), 개인 암 이력 변수 재정의

<Share & Result>

1) EDA 재진행한 Data double-check

1)-① 새롭게 정의한 "AGE"(입사 시 연령), "DURATION"(근속 연수) 변수 값이 결측 / 음수인 관측치 존재하는지 재확인

: 확인한 결과 해당 조건 만족하는 관측치 없음.

1)-② "DTH_DATE1" ~ "DTH_DATE3" 변수 / "fdx1" ~ "fdx6" 변수 값이 결측이 아닌 관측치 대상, 해당 변수를 날짜형으로 변환했을 때 결측인 관측치가 존재하는지 확인(결측은 달력 상 존재하지 않는 날짜임을 뜻함.)

: 확인한 결과 "fdx2" 변수 값이 이상하게 기입되어 있는 관측치가 "1개" 존재함. (나머지 변수는 모두 ok)

	fdx1	tcode1	mcode1	method1	icd10_1	seer_grp1	fdx2	tcode2	mcode2	method2	icd10_2	seer_grp2	fdx3	tcode3	mcode3
1	20100125	C187	81403	7	C187	2	20180073	C343	81403	6	C343	7			

↳ 해당 객체의 관측치를 모두 제외함. -- "dir.Preproc1_New" tbl

∴ "Preproc1_New" tbl 총 행 수는 89,408,537이고 distinct INDI_ID는 총 26,633,951명이다.

2) "Entry"(최초 고용보험 등록 연도) 변수, "사망 여부" + "사망 일자" 변수 정의

2)-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

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

2)-② 사망 일자 + 사망 여부 정의

: "cats(DTH_DATE1-DTH_DATE3)" 구문 통해 "사망 일자"(DTH_DATE) 변수 자체 정의한 후, 해당 변수 값이 결측이 아니면 사망으로 간주(Death 변수)

2)-③ "Entry" 변수 정의

: 2)-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_1999	Entry_2004	Entry_2009	Entry_2014	Entry_2018
1995 ~ 1999년	1	0	0	0	0
2000 ~ 2004년	0	1	0	0	0
2005 ~ 2009년	0	0	1	0	0
2010 ~ 2014년	0	0	0	1	0
2015 ~ 2018년	0	0	0	0	1

↳ 기존 방식에서 Dummy variable 형태로 바꿈.

(생성한 Data example)

	성별	INDI_ID	BYEAR	DTH_DATE	Entry_1999	Entry_2004	Entry_2009	Entry_2014	Entry_2018	Death
1	남	1000000757001	1940	19990521	1	0	0	0	0	1
2	남	1000000757002	1940	19980106	1	0	0	0	0	1
3	남	1000000757003	1940		0	1	0	0	0	0
4	남	1000000757004	1940		1	0	0	0	0	0
5	남	1000000757005	1940		1	0	0	0	0	0
6	남	1000000757006	1940		1	0	0	0	0	0
7	남	1000000757007	1940		1	0	0	0	0	0
8	남	1000000757008	1940		0	1	0	0	0	0
9	남	1000000757009	1940	20160703	1	0	0	0	0	1

3) 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), "Lung_Cancer_History", "Leukemia_Cancer_History"(암 이력 여부)

3)-① 폐암 / 백혈병 발병 여부 + 진단일자 변수 정의

: 이전에 사용했던 방식 그대로 이용 / 하지만, 폐암 icd-10 code를 "C34%"만 고려하지 않고 "C33%"도 추가로 고려

3)-②. 암 이력 변수 정의

: 각 Outcome 별로 암 이력 변수 다르게 정의

-- "fdx1" 변수가 결측이 아니면서 "lung_cancer" = 0이면 "Lung_cancer_History" = 1을 부여
마찬가지로, "fdx1" 변수는 결측이 아니지만 "leukemia"=0이면 "Leukemia_History" = 1을 부여

(생성한 Data Example)

	INDI_ID	lung_cancer	lung_cancer_date	leukemia	leukemia_date	Lung_Cancer_History	Leukemia_Cancer_History
1	1000000757001	0		0		1	1
2	1000000757002	0		0		0	0
3	1000000757003	0		0		0	0
4	1000000757004	0		0		0	0
5	1000000757005	0		0		0	0
6	1000000757006	0		0		0	0
7	1000000757007	0		0		0	0
8	1000000757008	0		0		1	1
9	1000000757009	0		0		0	0
10	1000000757010	0		0		0	0
11	1000000757011	0		0		0	0
12	1000000757012	0		0		0	0
13	1000000757013	0		0		0	0
14	1000000757014	0		0		0	0
15	1000000757015	0		0		0	0
16	1000000757016	0		0		0	0
17	1000000757018	0		0		0	0
18	1000000757019	0		0		1	1
19	1000000757020	0		0		0	0
20	1000000757021	0		0		0	0
21	1000000757022	0		0		0	0
22	1000000757023	0		0		1	1
23	1000000757024	0		0		0	0
24	1000000757025	0		0		0	0
25	1000000757026	1	20120725	0		0	1
26	1000000757027	1	20030112	0		0	1
27	1000000757028	0		0		0	0
28	1000000757029	0		0		0	0
29	1000000757030	0		0		0	0
30	1000000757031	0		0		1	1
31	1000000757032	1	20081122	0		0	1

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

Outcome	빈도	백분율
폐암 발병	85,999	85,999 / 26,633,951(0.32%)
백혈병 발병	15,851	15,851 / 26,633,951(0.06%)
사망	770,281	770,281 / 26,633,951(2.89%)