20230623_Date_EDA_Update

담당자	은경 이
상태	완료
마감일	@2023년 6월 23일
태그	My work

What TO DO)

: Outcome을 정의하고, 데이터를 다시 살펴보던 중 이상 case 발견하여 Data EDA 다시 진행

Share & Result)

1. Outcome (폐암 / 백혈병 발병 여부) + 사망 여부 / 사망 일자 정의하고 데이터 다시 살펴본 결과 이상 Case 존재

(Case1) 마지막 반복측정치의 상실일(OUT_DT)가 2018.12.31. 이후이면서 어떤 사건(Outcome 혹은 사망 사건)도 발생하지 않은 경우

연번	취득일	상실일	INDI_ID	성별	BYEAR	Entry	Cancer_History	Outcome_Category	Outcome_date
5546155	20180102	20190321	1000008126694	여	1962	4	0	last_follow	20181231

(Case2) 사망 사건이 발생한 사람 중, 사망일보다 고용보험 상실일(OUT_DT)가 더 미래시점인 경우

: 행정 상의 문제인 것으로 판단됨

1000000757235	19950701	19960913	Death	19960912
1000000757236	19950701	19971014	Death	19971013

(Case3) INDI_ID와 고용보험 취득일(ECNY_DT)이 동일한데 고용보험 상실(OUT_DT)은 다른 반복 측정치들이 존재하는 경우

1000000757060	20170101	20190101	last_follow	20181231
1000000757060	20170101	20200101	last_follow	20181231

→ (Case1) + (Case2) 경우를 만족하는 관측치는 총 5,048,631건, (Case3)을 만족하는 관측치는 총 1,871,046건임을 확인함.

```
/* 2023.06.23 Update - 상실일이 2018.12.31이후인 관측치 개수 확인
+ 사망일이 상실일보다 과거 시점인 관측치 확인 */
DATA check;
SET cohort2 (KEEP = INDI_ID NO ECNY_DT OUT_DT DTH_DATE1-DTH_DATE3);
DTH_DATE = cats(DTH_DATE1, DTH_DATE2, DTH_DATE3);
IF OUT_DT>'20181231' or (not missing(DTH_DATE) and DTH_DATE < OUT_DT);
RUN;
```

2. 해당 case를 모두 정제하기 위해 Raw data EDA 재진행

2)-1 고용보험 상실일(OUT_DT)가 2018.12.31이후인 경우, 해당 값을 모두 2018.12.31.로 변경

2)-2 사망일(DTH_DATE1-DTH_DATE3)이 고용보험 상실일(OUT_DT) 보다 과거 시점인 경우 상실일 값을 사망일로 대체

```
/* 상실일, 취득일의 입력 방식이 옳지 않은 데이터 delete + OUT_DT가 결측인 관측치 대체 */
/* 2023.06.23 Update - 상실일이 2018.12.31이후인 관측치는 모두 2018.12.31로 변경
+ 사망일이 상실일보다 과거 시점인 경우 상실일 값을 사망일로 변경 */
DATA cohort3;
SET cohort2;
IF length(ECNY_DT) ^= 8 or (not missing(OUT_DT) and length(OUT_DT) ^= 8) then delete;
IF (OUT_DT = '') or  (OUT_DT > '20181231') then OUT_DT = '20181231';
DTH_DATE = cats(DTH_DATE1, DTH_DATE2, DTH_DATE3);
IF (not missing(DTH_DATE) and DTH_DATE < OUT_DT) then OUT_DT = DTH_DATE;
DROP DTH_DATE;
RUN;
```

2)-2 (Case3)을 만족하는 관측치들의 모든 변수를 살펴보니, "사업장 정보"가 다름을 확인

: "투잡" 뛰는 사람 반영한 것으로 추측됨.

(Example)

168	2800768	남	20170101	20181231		N	75121	75120	75121	019	3	20686153330	2068615333	주식회사 피누스이앤씨	32
169	6915684	남	20170101	20181231	계약만료, 공사종료	N		75120	75121			91918247571	2068615333	(주)피누스이앤씨-강촌2단지한신	27

→ 해당 경우는 사업장 정보가 다르므로 두 관측치 모두 KEEP하는 것이 맞다고 판단함.

2)-3 DATA EDA 다시 진행한 후 얻은 “New_raw” DB의 총 Row 수는 100681137건이고, distinct INDI_ID 수는 29189326명임을 확인

3. 해당 Data 이용해 다시 Outcome 정의 재진행

: 방식은 이전에 사용한 방식과 동일

(Outcome 발생 비율)

Event	빈도	백분율
Lung cancer	90,523	90,523/29,189,326 (0.31%)
Leukemia	17,349	17,349/29,189,321 (0.06%)
Death	808,500	808,500/29,189,321 (2.77%)