

산보연 Data check - 6월 9일 Version

작성자 : 이은경

1. CAL(입사 시기)의 각 범주에 대하여 입사 시기 5년 이후까지의 follow-up year(YEAR 변수)에 대응하는 추적 인년 합계(FY_BZ), 누적 근로자수(ENR_BZ)에 결측값이 있는지 확인.

: 예를 들어, CAL=0이면 입사 시기가 ~1995년까지이므로, YEAR가 1996년 ~ 2000년 동안 추적 인년 합계, 혹은 누적 근로자 수에 결측값이 있는지 확인

[Result]

CAL 범주	Follow-up year 범위	FY_BZ 결측값 개수	ENR_BZ 결측값 개수
0	1996년 ~ 2000년	116	0
1	2001년 ~ 2005년	3262	0
2	2006년 ~ 2010년	4959	0
3	2011년 ~ 2015년	5476	0
4	2016년 ~ 2020년	2465	0

→ “추적 인년 합계” 변수에는 결측값이 있는 것으로 보임 / “누적 근로자 수” 변수에는 결측값이 없음.

2. YEAR 변수 기준으로 백혈병 발생 건수(LEUKEMIA_BZ) 합한 후, 국립암센터에서 매년 보고되는 백혈병 발병률과 비슷한지 확인.

2-1) 국립암센터 백혈병 발병률

(참고 link : <https://ncc.re.kr/cancerStatsView.ncc?bbsnum=598&searchKey=total&searchValue=&pageNum=1>)

year	LEUKEMIA_BZ	CR
1 2000	200700000	4.2
2 2001	220900000	4.6
3 2002	231900000	4.8
4 2003	228300000	4.7
5 2004	237200000	4.9
6 2005	234600000	4.8
7 2006	244300000	5.0
8 2007	247700000	5.0
9 2008	260100000	5.3
10 2009	271700000	5.5
11 2010	276000000	5.5
12 2011	290000000	5.8
13 2012	287300000	5.7
14 2013	307200000	6.1
15 2014	311400000	6.1
16 2015	328600000	6.4
17 2016	343700000	6.7
18 2017	339700000	6.6
19 2018	352000000	6.9

LEUKEMIA_BZ : 각 YEAR에 발생한 총 백혈병 발생 건수

CR(조발생률) :
$$\text{조발생률} = \frac{\text{새롭게 발생한 암환자수}}{\text{연앙인구}} \times 100,000$$

2-2) 산보연 데이터

YEAR	sum_LEUKEMIA_BZ	sum_FY	rate_per_year_FY
<int>	<int>	<int>	<dbl>
2000	121	18857990	0.642
2001	269	22612823	1.19
2002	473	29265205	1.62
2003	680	41100139	1.65
2004	955	51892516	1.84
2005	1322	65058466	2.03
2006	1704	78544059	2.17
2007	2166	93595599	2.31
2008	2646	109762417	2.41
2009	3242	126932284	2.55
2010	3936	145737309	2.70
2011	4701	165984731	2.83
2012	5519	187676209	2.94
2013	6481	210808770	3.07
2014	7514	234846595	3.20
2015	8718	261033631	3.34
2016	10011	289216437	3.46
2017	11392	319175197	3.57
2018	12977	350391620	3.70

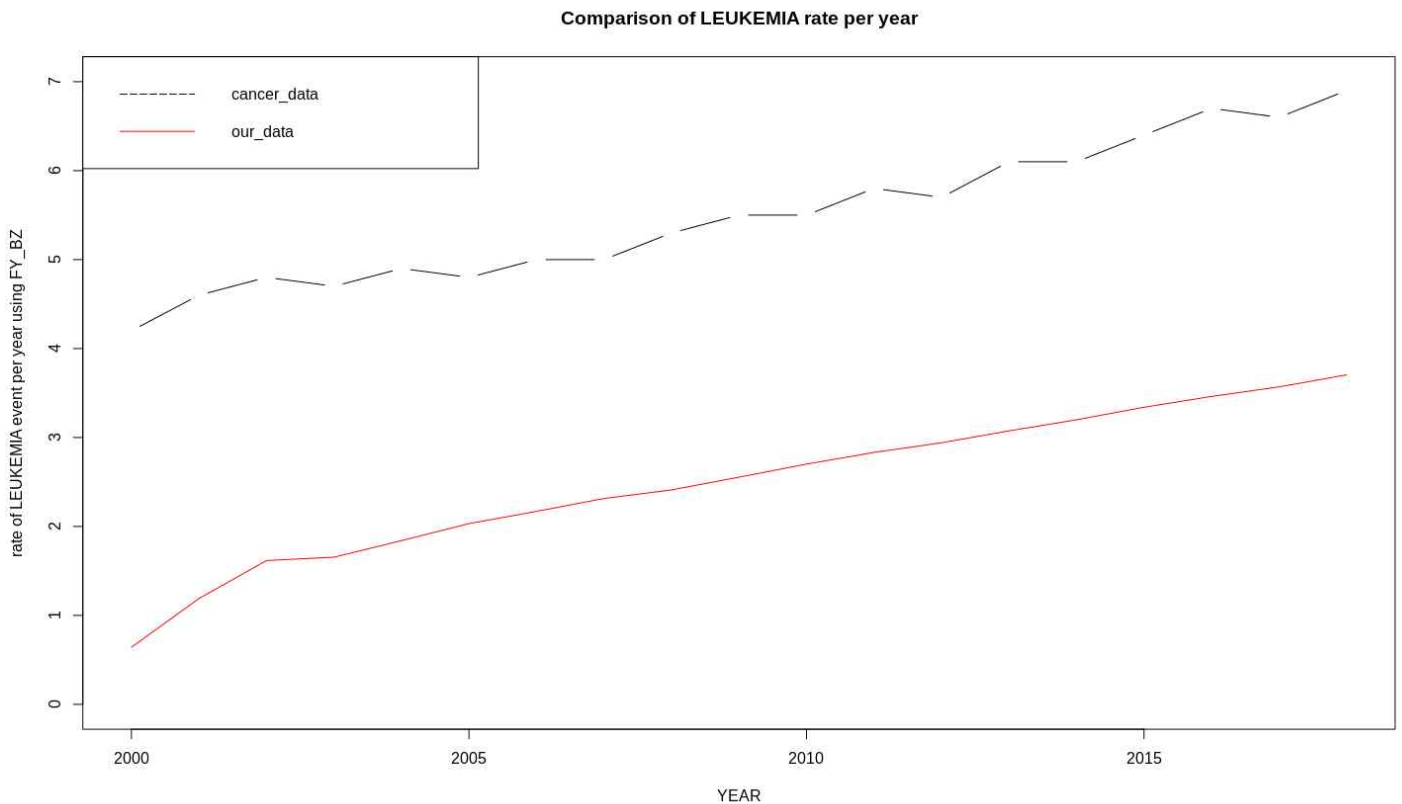
sum_LEUKEMIA_BZ : 각 YEAR에 발생한 총 백혈병 발생 건수

sum_FY_BZ : YEAR에 대응하는 총 추적인년 합계

rate_per_year_FY : 위의 CR과 대응되는 개념으로,
(sum_LEUKEMIA_BZ / sum_FY_BZ) * 100000 공식으로 계산됨.

[Graph]

: 빨간색 실선이 our data, 검은색 점선이 국립암센터에서 보고된 백혈병 추세



→ 국내 총 백혈병 발생 건수가 집계된 국립암센터 데이터값이 더 큰 것이 당연한 부분 / 이때, YEAR에 따른 추세는 비슷한 것으로 보임. (이로 보아, 백혈병 발생률 계산 때, 추적 인년 합계로 나누는 것이 타당해 보임.)

4. follow up year(YEAR)가 2018년 일 때, 사업장 중분류별 LEUKEMIA_BZ(백혈병 발생 건수), FY_BZ(추적인년합계) TOP 5는 어느 사업장인가.

1) 백혈병 발생 건수(LEUKEMIA_BZ)

UP2(중분류)	업종명	발생 건수
41	건설업	844
46	도매 및 소매업	803
49	운수업	666
26	제조업	648
75	사업시설관리, 사업지원 서비스업	551

2) 추적 인년 합계(FY_BZ)

UP2(중분류)	업종명	합계
46	도매 및 소매업	23213023
26	제조업	21837127
41	건설업	21328582
75	사업시설관리, 사업지원 서비스업	15367009
86	보건업 및 사회복지 서비스업	14999701

5. 사업장 중분류 category 기준으로, YEAR에 따른 백혈병 발생률의 변화량 그래프 확인

: 백혈병 발생률 계산하는 방법을 2가지로 나눔.

① 중분류별, YEAR로 grouping → 그 후, (백혈병 건수 합) / (추적 인년 합계) 계산 = 통합 누적 발생률

② 각, 사업장별로 rate = (백혈병 발생 건수) / (추적 인년 합계) 계산 → 중분류별, YEAR로 grouping

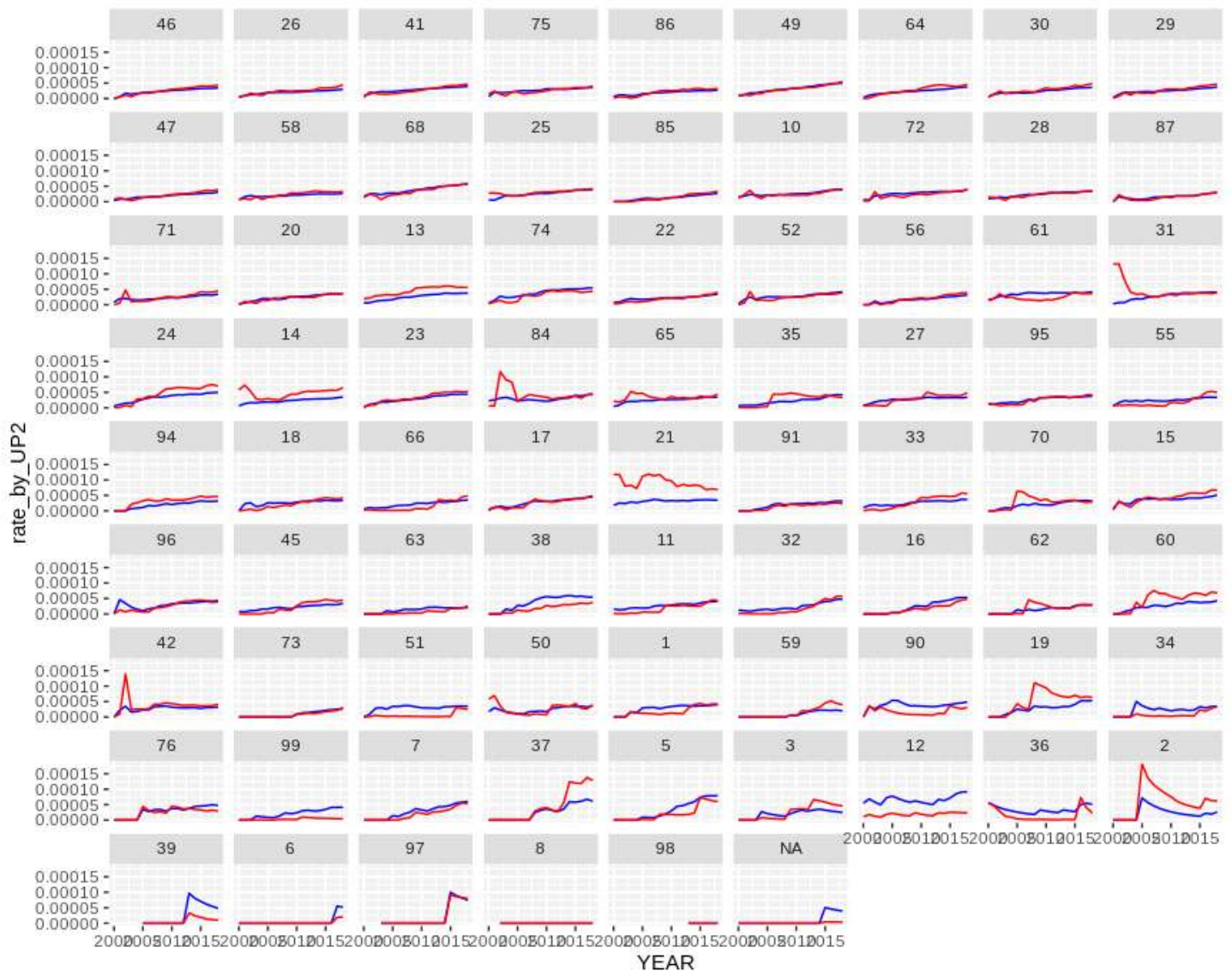
→ rate들의 평균 계산 = 평균 누적 발생률

: ①, ② 값을 사업 중분류 별로 겹쳐서 그림. (파란색이 ①값, 빨간색이 ②값 의미)

: 추적 인년 합계 기준으로 plot들을 오름차순 정렬 (plot 순서 = 추적 인년 합계 순서)

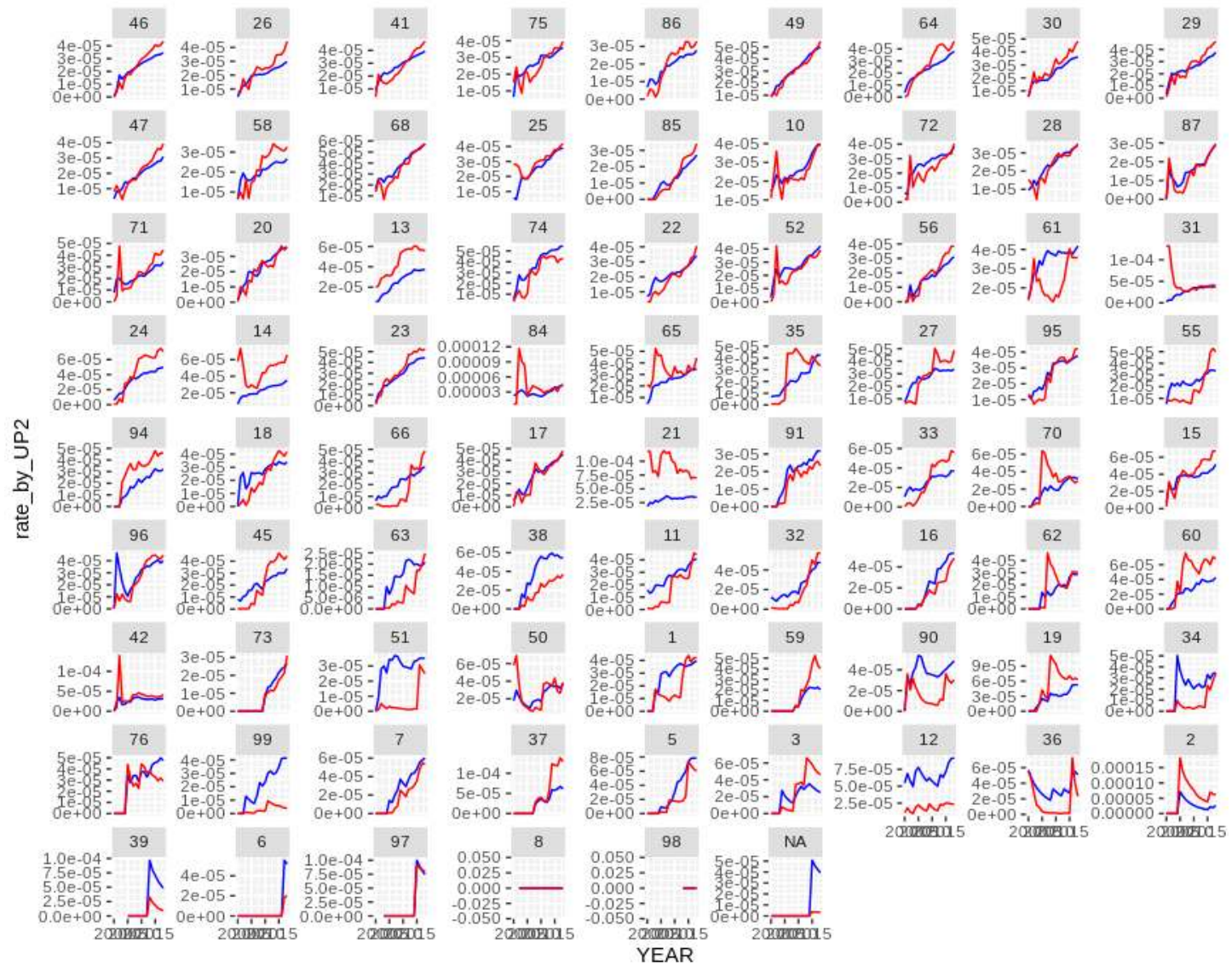
[Version1]

: y축의 scale이 모두 동일하게 적용한 경우



[Version 2]

: y축의 scale을 각 사업장에 따라 다르게 적용한 경우



[Follow up year(YEAR)가 2018년 일 때, 백혈병 발생 건수, 2가지 방법으로 계산한 백혈병 발생 비율 TOP10 파악]

1) 백혈병 발생 건수(LEUKEMIA_BZ)

UP2(중분류)	업종명	발생 건수
41	건설업	844
46	도매 및 소매업	803
49	운수업	666
26	제조업	648
75	사업시설관리, 사업지원 서비스업	551
68	부동산업 및 임대업	511
64	금융 및 보험업	478
30	제조업(자동차 부품 관련)	450
29	제조업(고무, 종이, 금속기계, 냉장장비, 펌프, 압축기)	449
86	보건업	415

2) 통합 누적 발생률 - 빨간색 추세선

UP2(중분류)	업종명	발생률
12	제조업(담배)	0.0000918
5	광업(석탄, 원유, 채굴업)	0.0000788
97	가구 내 고용 활동	0.0000744
37	하수, 폐기물 정화 업	0.0000611
7	광업(화학물 채취, 광물, 비금속 관련)	0.0000602
68	부동산, 임대업	0.0000572
38	건설 폐기물, 방사성 폐기물 정화 업	0.0000546
74	사업시설 관리, 사업지원 서비스업	0.0000544
16	제조업(나무 관련)	0.0000531
19	제조업(석탄, 운할유, 원유, 정제처리 관련)	0.0000527

3) 평균 누적 발생률 - 파란색 추세선

UP2(중분류)	업종명	발생률
37	하수, 폐기물 정화	0.000129
97	가구 내 고용 활동	0.0000798
24	제조업(제철, 철강)	0.0000704
21	제조업(의약품)	0.0000692
60	방송 통신업	0.0000679
15	제조업(가죽 관련 제품)	0.0000669
14	제조업 (옷 관련)	0.0000658
19	제조업(석탄, 원할유, 원유 정제처리)	0.0000632
2	임업	0.0000620
5	광업(석탄, 원유, 채굴업)	0.0000601