〈건강검진코호트DB Data Cleaning Plan〉

작성자: 이은경

1. Study Population 정의

[회의록]

- Study population정의
 - <u>검진DB에 BMI 한 건 이상 & WAIST 한 건 이상 &</u> 공복혈당 기록 한 건 이상 & 명세서 한 건 이상 & 원외처방전 한 건 이상 있는 자 중.
 - 포함 기준
 - 각 개체에 대하여 Onset date (당뇨 발병일, "A"): 셋 중에 가장 이른 시점을 계산
 - 검진DB: 최초로 (공복혈당 >= 126 OR 당뇨과거력 YES)인 시점
 - 당뇨발병시점: 연도의 중간값(6월 30일)
 - 명세서DB(T20): 주상병, 부상병, 제2부상병을 포함한 모든 상병들 중에 다음의 당뇨 코드 (E11~, E12~, E13~,E14~)가 기록된 시점
 - 당뇨발병시점: 요양개시일자
 - 원외처방전DB(T60): 당뇨 경구치료제 (A10B~)가 기록된 시점
 - 당뇨발병시점: 요양개시일자 (가정: 당뇨병 진단/처방받으러 온 사람들은 대부분 외래이고, 입원 중 진단되었다 하여도, 입원기간이 길지 않을 것이다.)
 - 제외 기준: 제1형 당뇨코드(E10~) 와 임신중 당뇨병(O24~) 명세서 소지자를 제외
 - n수를 확인해가며

Note1) 검진DB에도 정확한 검진일자가 존재하는 것을 확인: "HME_DT" 변수

Note2) 60T에는 "RECU_FR_DT" 변수가 존재 : 이는 최초내원일자, 조제투여일자를 의미.

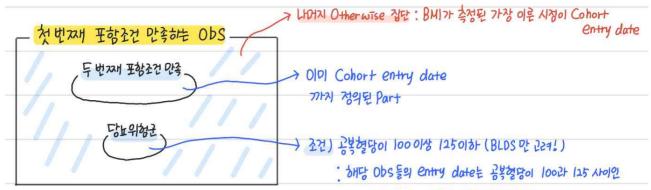
Note3) 사망여부는 사망일자 변수값 존재 유무로 판단 (가능)

- * 60T의 Risk Factor(가정)
- : 당뇨병 약물 처방 받은 사람은 대부분 외래일 것이며 입원해도 입원 기간이 길지 않을 것이다.
- * 자격변수, 즉 의료급여자격관련 변수보다는 소득 변수가 분석에 더 도움이 될 것이다.
- --> 소득분위변수(CTRB PT TYPE CD)가 자격DB에 존재. / 의료급여수급권자는 "0분위"

1-1) 모집단 포함 기준

	첫 번째 조건	검진DB에 BMI, WAIST, 공복 혈당 기록이 적어도 한 건 이상 존재해야 함.	
		+ 진료기록, 처방 기록 적어도 한 건 이상 존재해야 함.	
	두 번째 조건	검진DB	공복혈당수치가 126 이상 OR 당뇨병 과거력 존
전제조건		& 생애전환검진DB	재
		진료DB - 20T, 40T	상병 변수에 type 2 당뇨병에 해당하는 ICD-10
			코드가 존재("E11~", "E12~", "E13~", "E14~")
		진료DB - 60T	처방된 약물 중 당뇨병 경구 복용 치료제 처방 이
			력(ATC Code - "A10B"~)이 존재

: 첫 번째 조건을 만족하는 개체들 중, 두 번째 조건 중 하나만 만족해도 모집단에 포함.



시정기준 BMI가 기록된 가장 가까운 이래

- : Disease onset date는 조건을 만족하는 시점 중 "가장 이른" 시점.
- : Cohort entry date 파악 필요.
- : Cohort entry date와 Disease onset date 구분해서 파악

Disease Onset Date	Cohort entry date
<u>(후보)</u>	Disease onset date에서 가장 가까운 미래
- 검진DB 기준으로 공복혈당이 126 이상인 첫 시점	<u>(조건)</u>
- 진료DB 기준으로 상병 코드에 당뇨병 코드가 기	- Cohort entry date 이전에 진료, 처방 기록 적어
록된 첫 시점	도 한 건 이상 존재해야 함.
- 처방전 table 기준으로 처방 약물이 당뇨 경구 치	- Cohort entry date 이후로 BMI가 한 번 이상
료제인 첫 시점	추적되어있어야 함.
-> 세 후보들 중 <u>가장 이른 시점</u>	- Cohort entry date는 2016년이 최대 연도

1-2) 모집단 제외기준

: 1-1 포함조건을 만족하는 개체 중, cohort entry date 이전 진료기록에서 type I 당뇨("E10~") 혹은 임신 중 당뇨("O24~")를 진단받은 개체는 제외

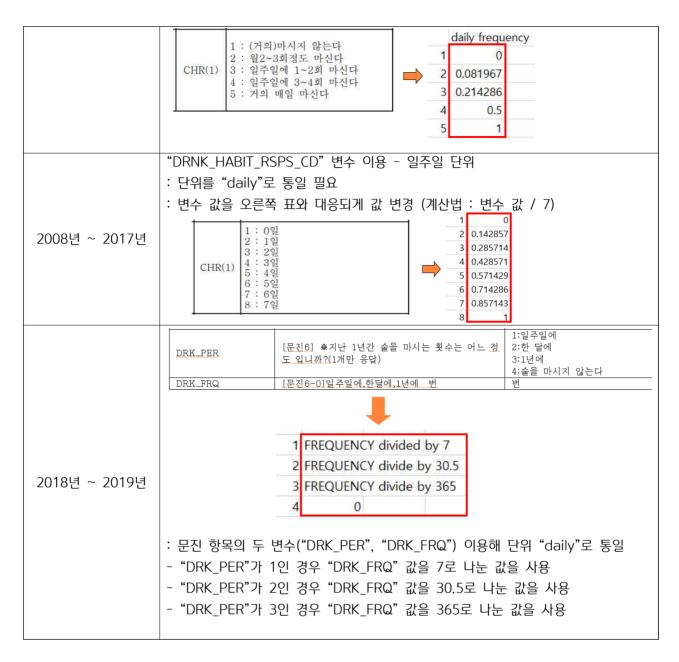
: Baseline cancer(Maligancy : ICD-10 code가 "C00" ~ "C97") 이력이 있는 사람 제외(이때, Baseline cancer 항목에서 피부암(단, 멜라노마는 제외조건에 포함되지 않음), 갑상선암은 제외) -- 고려해야 하는 ICD-10 code는 <u>"C00" ~ "C43", "C45" ~ "C75", "C77" ~ "C97"</u>

: Myocardial infraction, Heart failure, TIA or Stroke 이력이 있는 사람 제외 이때, <u>Baseline은 cohort entry date 이전 & 그 당시 timing</u>을 의미함.

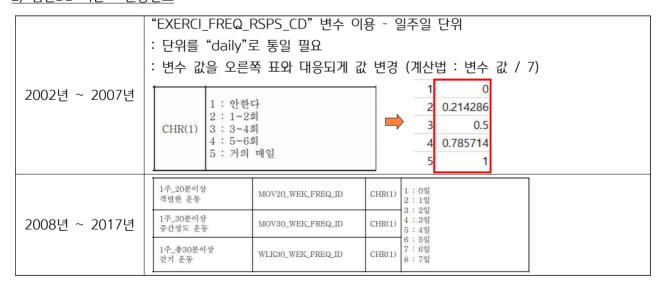
2. Covariate 정의 — Master Document 확인

1) 검진DB 기반 - 음주습관

	"DRNK_HABIT_RSPS_CD"(음주습관) 변수 이용
2002년 ~ 2007년	: 문진 항목 값을 오른쪽 표와 대응되게 값 변경 (2 -> 2.5/30.5 3 -> 1.5/7 4 -> 3.5/7 5->1)
	: 단위를 "daily"로 모두 통일



2) 검진DB 기반 - 운동빈도



		네 가지 변수 모두 이용 변수 값 + 중간정도 운동 변수 값 + 7	건기 운동 변수 값) / 3} / 7
	PA_VD_FRQ	[문진7-1]평소1주일간숨이많이차게만드는고강도 체활동을며칠하십니까?주당 일 *고강도 신체활동의 예)달리기,에어로빅,빠른속 로 자전거 타기,건설 현장 노동,계단으로 물건 나 기 등	도 일
2018년 ~ 2019년	PA_MD_FRQ	[문진8-1]평소1주일간,숨이약간차게만드는중강도신 체활동을며철하십니까?주당 일 *7번 응답에 관련 된 신체활동은 제외하고 답해주십시오. * 중강도 신체활동의 예)빠르게 걷기, 복식태니스,보통 속도 로 자전거 타기, 가벼운 물건 나르기, 청소 등	일
	PA_MUSL_FRQ	[문진9]최근1주일동안팔굽혀펴기,웟몸일으키기,아 령,역기,철봉동근력운동을한날은며칠입니까?주당 일	일
	_	는 운동 관련 세 가지 변수 모두 이용 변수 값 + 중간정도 운동 변수 값 + 급	 근력 운동 변수 값) / 3} / 7

3) 운동지속 기간 -> 추가로 생성하는 변수

: 운동시간을 나타내는 단위로 "mets"가 있음

2002년 ~ 2007년	문진 항목에 세분화 되어있지 않던 시기이므로 2009년 ~ 2017년 기록에서 가 장 과거의 기록 가져와 "운동 지속 기간" 변수 생성
2008년 ~ 2017년	2)에서 제시한 운동 관련 변수 중 "격렬한 운동", "중간정도 운동" 변수만 이용 (걷기 운동 관련 변수는 2018년 ~ 2019년에 측정되지 않으므로) - 운동량에 따른 Met 기준 Moderate intensity excercise 3 mets Vigorous acitivity 6 mets met minutes mets*minutes Weekly met minutes calculation example every day vigorous exercise more than 20min 6 mets*20min*2 240 every day wigorous exercise more than 30min 3 mets*30min*3 270 every day walk more than 30min 3 mets*30min*4 360 total met minutes per week : 위의 Met 기준 이용하여 "격렬한 운동", "중간 정도 운동" 변수 값 이용해 일 주일 단위 총 total met minute 계산
2018년 ~ 2019년	2)에서 제시한 운동 관련 변수 중 "격렬한 운동", "중간정도 운동" 변수만 이용 (근육 운동 관련 변수는 2009년 ~ 2017년에 측정되지 않으므로) - 운동량에 따른 Met 기준 Moderate intensity excercise 3 mets Vigorous acitivity 6 mets met minutes mets*minutes

PA_VD_FRQ	[문진7-1]평소1주일간숨이많이차게만드는고강도신 체활동을며원하십니까?주당 일 *고강도 신체활동의 예)달리기,에어로빅,빠른속도 로 자전거 타기,건설 현장 노동,계단으로 물건 나르 기 등	ପୂ
PA_VD_HRS	[문진7-2-1]평소하루에숨이많이차게만드는고강도 신체활동을몇시간하십니까?하루에 시간	시간
PA_VD_MINS	[문진7-2-2]평소하루에숨이많이차게만드는고강도 신체활동을몇시간하십니까?하루에 분	분
PA_MD_FRQ	[문진8-1]평소1주일간,숨이약간차계만드는증강도신 체활동을며원하십니까?주당 일 *7번 응답에 관련 된 신체활동은 제외하고 답해주십시오. * 증강도 신체활동의 예)빠르게 걷기, 복식태니스.보통 속도 로 자전거 타기, 가벼운 물건 나르기, 청소 등	인
PA_MD_HRS	[문진8-2-1]평소하루에숨이약간차게만드는중강도 신체활동을몇시간하십니까?하루에 시간	시간
PA_MD_MINS	[문진8-2-2]평소하루에숨이약간차게만드는중장도 신체활동을몇시간하십니까?하루에 분	분
PA_MUSL_FRQ	[문진9]최근1주일동안팔굽혀퍼기,윗몸일으키기,아 령,역기,철봉등근력운동을한날은며칠입니까?주당 일	일

: 위의 Met 기준 이용하여 "격렬한 운동", "중간 정도 운동" 변수 값 이용해 빈도 파악

→ 각 꼬리 문항 ("~HRS", "_MINS") 이용해 minutes 파악(이때, "~MINS" 변수 값이 20 이상 이거나 "~HRS"이 결측이 아닌 경우 "격렬한 운동"은 minutes을 "30분"으로, "중간 정도의 운동"은 "20분"으로 시간 파악, (이유는 사람들이 작성할 때 "시간", "분" 고려하지 않고 작성할 수 있어 시간 과대 추정 가능성 있음 & 2008년 ~ 2017년 자료와 기준 통일)
→ 일주일 단위 총 total met minute 계산

4) 검진DB 기반 - 흡연습관

i) 2002년 ~ 2008년

변수명	데이터 값	Cleaned Data에서의 값	
흡연상태	1 2 3		변함 없음.
흡연지속기간(year가 단위)	1 2 3 4 5	2.5 7 14.5 24.5 30	: 각 범주가 의미하는 연도 범위의 중간값으로 통일
	1	0.25(5/20)	(전제조건) : 흡연상태가
하루 흡연량(1pack = 20개비	2	0.73((10+19)/2)/2	"current smoker"(3)인 사람만 해당 값 계산
가정)	3	1.48((20+39)/20)/2	/ 단위를 pack으로 맞추기 위해
	4	2(40/20)	20으로 나눔
Pack years (갑년) : 흡연의 누적 damage 계산	새로 생성하는 변수	하루 흡연량 × 흡연지속기간	(전제조건) : 흡연상태가 "current smoker"(3)인 사람만 해당 값 계산

ii) 2008년 ~ 2017년

흡연상태	SMK_STAT_TYPE_RSPS_CD	CHR(1)	1 : 피우지 않는다. 2 : 과거에 피웠으나 지금은 끊었다 3 : 현재도 피운다
(과거)흡연기간	PAST_SMK_TERM_RSPS_CD	NUM(3)	_년
(과거)하루흡연량	PAST_DSQTY_RSPS_CD	NUM(3)	_개피
(현재)흡연기간	CUR_SMK_TERM_RSPS_CD	NUM(3)	_년
(현재)하루흡연량	CUR_DSQTY_RSPS_CD	NUM(3)	_개피

[변수값 조정해야 하는 변수]

- For current smoker

(현재) 흡연기간	"years" 기준 — 변동 필요없음
(현재) 하루 흡연량	"#of cigarettes / 20"으로 변환
Pack years(새로 추가)	(현재) 하루 흡연량 × (현재) 흡연기간

- For past smoker

(과거) 흡연기간	"years" 기준 — 변동 필요없음
(과거) 하루 흡연량	"#of cigarettes / 20"으로 변환
Pack years(새로 추가)	(과거) 하루 흡연량 × (과거) 흡연기간

iii) 2018년 ~ 2019년

SMK_STAT_TYPE_R SPS_CD	흡연상태	1:피우지않는다. 2:과거에피웠으나지금은끊었다 3:현재도피운다
PAST_SMK_TERM_ RSPS_CD	(과거)흡연기간	
PAST_DSQTY_RSPS _CD	(과거) <u>하루흡연</u> 량	
CUR_SMK_TERM_R SPS_CD	(현재)흡연기간	
CUR_DSQTY_RSPS_ CD	(현재) <u>하루흡연</u> 량	

[변수값 조정해야 하는 변수]

- For current smoker

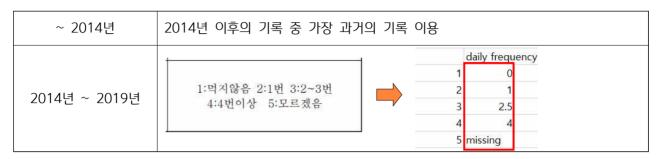
(현재) 흡연기간	"years" 기준 — 변동 필요없음
(현재) 하루 흡연량	"#of cigarettes / 20"으로 변환
Pack years(새로 추가)	(현재) 하루 흡연량 × (현재) 흡연기간

- For past smoker

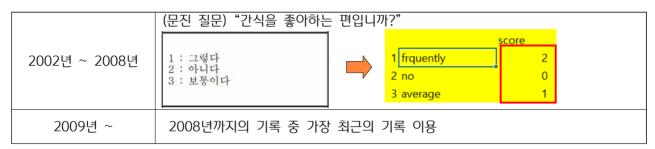
(과거) 흡연기간	"years" 기준 — 변동 필요없음

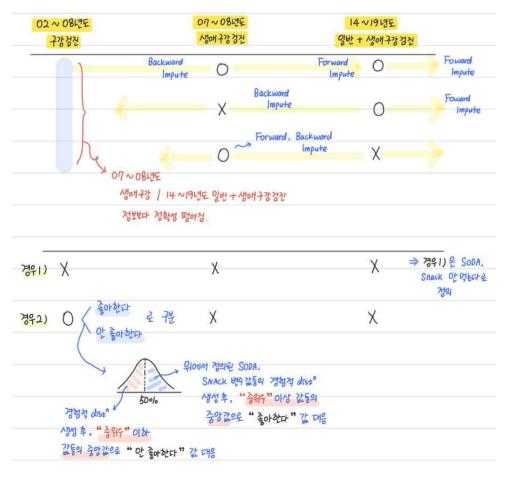
(과거) 하루 흡연량	"#of cigarettes / 20"으로 변환
Pack years(새로 추가)	(과거) 하루 흡연량 × (과거) 흡연기간

5) 구강검진DB 기반 - Soda 섭취 습관



6) 구강 검진DB 기반 -간식 섭취 습관





7) 검진DB 기반 - 가족력

- : 한 번이라도 가족력이 있다면, always positive로 간주(모든 time에 가족력이 존재하는 것으로 판단)
 - 가족력은 유전적인 의미를 가지고 가기 때문.

Heart disease	검진DB의 문진항목 이용		
Diabetes mellitus	검진DB의 문진항목 이용		
Cancer	2002년 ~ 2007년 검진DB 문진 항목에 가족력이 있다고 기록되어 있다면,		
	always positive로 간주		

8) 개인 과거 병력

: 개인 과거병력은 검진DB, 진료DB 모두에서 확인할 수 있다.

	Hypertension (고혈압)	Stroke or TIA (뇌졸중)	Myocardical infraction (심근경색 - 혈관 관련)	Heart failure (심장마비 - 근육 관련)	Cancer (암)
검진DB	사용	사용	사용하지 않음	사용하지 않음	사용하지 않음
진료DB	사용	사용	사용	사용	사용

9) Comorbidities

- * Myocardical infraction(심근경색)은 입원 때만 잡히는 병명, Angina(협심증)는 외래에서도 기록되는 병명
- * Baseline covariate 중 "검진" 관련 항목은 Cohort entry date 이전까지 모든 검진 연도마다 기록 follow up 하여 table 따로 생성해 놓기

[수정된 Proposal 기반 Comorbidities]

ii. Comorbidities:

- Capturing both physical examination and ICD-10 code: hypertension
- ICD-10 based: Asthma, Chronic obstructive pulmonary disease (COPD), Atrial fibrillation, Thromboembolism, Chronic liver disease, Chronic Kidney disease, Coronary artery disease, Peripheral vascular disease, Dementia, GI disorders, Hyperlipidemia, Pneumonia, Psychiatric disorders.

10) Medications

: code-book Update Version follow up 하기

[수정된 Proposal 기반 Medications]

iii. Medications: +

- 당뇨약제들↓
 - a. Insulin-
 - b. Insulin 제외 경구약제: Metformin, meglitinides, SGLT2 inhibitors, thiazolidinediones, sulfonylureas, GLP-1 receptor agonists, DPP-4 inhibitors, and α-glucosidase inhibitors.
- 혈압약들
 - a. renin angiotensin system (교감신경 pathway) 관련: Angiotensin converting enzyme inhibitors, Angiotensin receptor II blockers
 - b. RAS 를 제외한 <u>혈압약제들</u>: β-blockers, Calcium channel blockers, Diuretics, vasodilators, anti-adrenergics
- Central nerve system: Antidementia, Antiarrhythmics, Anxiolytics, Antipsychotics, Antidepressants,
- 진통제(non-opioid analgesic): NSAIDs, acetaminophen₄
- Anticoagulants and antiplatelets: Anticoagulants. Antiplatelets including Aspirin.
- None of the above: Non-statin antihyperlipidemic, statin.

3. Outcome 정의

: Outcome에 따른 population 제외 조건 동일!

Primary Outcome	all cause death	Baseline cancer(피부암, 갑상선암 제외 - 단, 피	
Secondary Outcome	Cancer	부암에서 멜라노마는 제외조건에 포함 시키지 (음), Baseline MI, Baseline heart failure,	
	Myocardical infraction		
	Heart failure		
	TIA or Stroke	Baseline TIA or Stroke 있는 개체 제외	

- * 제외기준 적용함에 따라 남은 N수 파악 필요
- * Cancer는 제거했다가 다시 늘어난 Outcome 항목

4. Exposure 정의

- : cohort entry date에서 가장 가까운 시점의 BMI 값
- * Exposure의 후보는 여러 가지가 있고, 각 Exposure 마다 분석 방법은 달라진다.
 - 1) 당뇨병 진단 이후 시점 중 가장 과거 시점에 측정한 WAIST 값
 - : 이 경우에서 각 개체들의 Cohort entry date는 BMI / WAIST가 측정된 시점이다.
 - : Cohort entry date가 2016년 12월 31일 이후인 개체들은 제외.
 - 2) BMI, WAIST trajectory