

2021학년도 1학기 의학통계학 실습강의

경상국립대학교 의과대학 예방의학교실
김미지

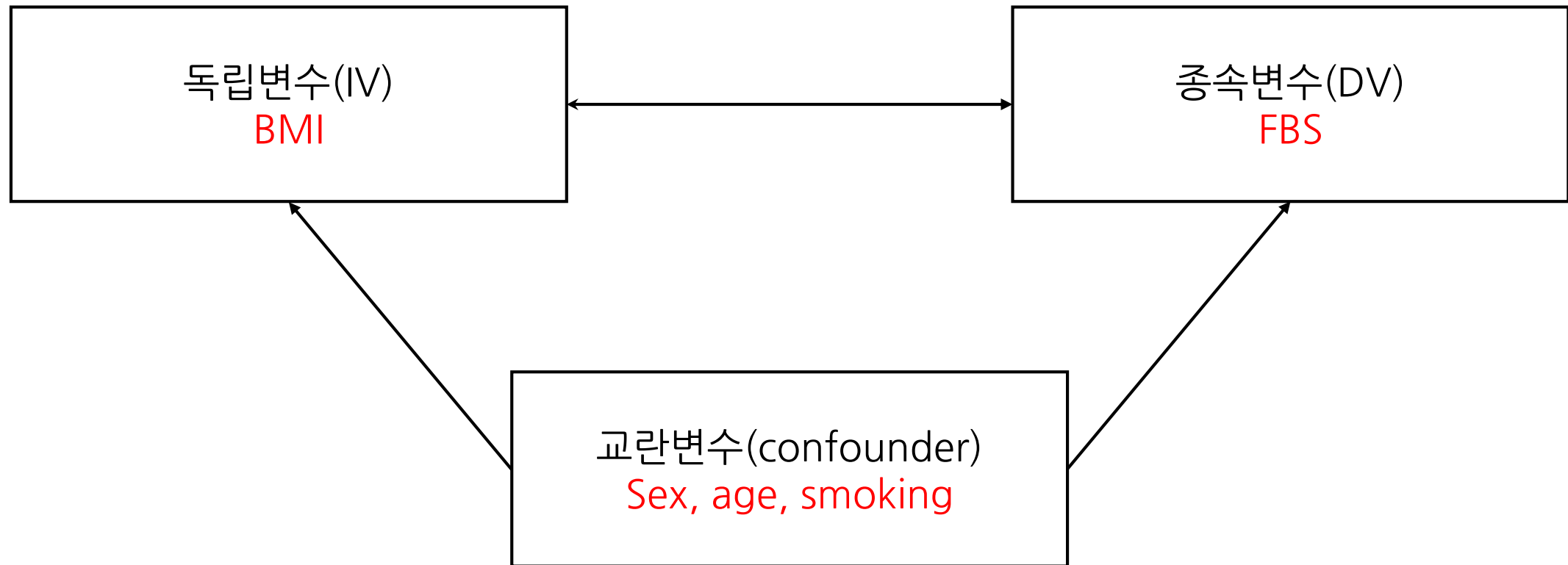
주차	일시	강의내용	강의교수	비고
7	4. 14. (수) 09:00-11:50	8. 데이터 수집 및 정리	김미지	비대면 강의(실시간)
8	4. 21. (수) 09:00-11:50	중간고사	김미지	대면 시험
9	4. 28. (수) 09:00-11:50	9. 두 집단 간 차이 비교: t 검정, 카이제곱 검정, 비모수 검정	김미지	비대면 강의(실시간)
10	5. 05. (수) 09:00-11:50	10. 세 집단 간 차이 비교: ANOVA, 사후분석, 비모수 검정	김미지	비대면 강의(동영상)
11	5. 12. (수) 09:00-11:50	11. 단순 연관성 확인: 상관분석	김미지	비대면 강의(동영상)
12	5. 19. (수) 09:00-11:50	12. 인과적 연관성 확인: 회귀분석	김미지	비대면 강의(동영상)
13	5. 26. (수) 09:00-11:50	13. 빅데이터 분석 및 생물정보학 (bioinformatics) I	유원기	비대면 강의(실시간/동영상) 과제 제출
14	6. 02. (수) 09:00-11:50	14. 빅데이터 분석 및 생물정보학 (bioinformatics) II	유원기	비대면 강의(실시간/동영상)
15	6. 09. (수) 09:00-11:50	기말고사	김미지	대면 시험(예정)



8. 데이터 수집 및 정리

2021-04-14 (수)

연구가설: 우리나라 30세 이상 성인에서 체질량
지수(BMI)의 증가와 공복혈당(FBS)의 증가는 연
관성이 있다.



데이터 수집(E)

sex	age
1. 남자	00세
2. 여자	1~79. 1~79세
	80. 80세 이상

id	year	sex	age	smk1	smk2	ht	wt	glu
A901020501	2018	1	20	2	3	162.4	56	103
A901021301	2018	1	38	3	8	167.7	76.4	127
A901021302	2018	1	35	3	8	157.7	53.2	134
A901022601	2018	1	33	3	8	151	53.5	99
A901023301	2018	1	30	3	8	146.1	57.6	111
A901023304	2018	1	31	3	8	157.2	53.2	51
A901023305	2018	1	34	3	8	180	85.6	163
A901023306	2018	1	32	3	8	168.8	54.2	83
A901023307	2018	1	33	8	8	171.6	54.8	
A901023308	2018	1	31	8	8	155.2	38	89
A901023309	2018	1	36	3	8	172.6	69	86
A901023310	2018	1	30	3	8	153.9	54.7	104
A901023905	2018	1	32	8	8	87.5	11.3	
A901024801	2018	1	57	2	1	177.1	94.1	99

코드북 작성(new variables)

변수명	변수설명	내용
smk	현재흡연 여부	1. 과거흡연, 비흡연
		2. 현재흡연(평생 담배 5갑(100개비) 이상 피웠고 현재 담배를 피우는 경우)
bmi	체질량지수	체중(kg)/신장(m) ²
bmi_c	체질량지수 범주화	1. 저체중($0 < \text{bmi} < 18.5$)
		2. 정상($18.5 \leq \text{bmi} < 25$)
		3. 비만($\text{bmi} \geq 25$)
smk	현재흡연 여부	1. 과거흡연, 비흡연



9. 두 집단 간 차이 비교: t 검정, 카이제곱 검정, 비모수 검정

2021-04-28 (수)

e.g., A 고혈압 약을 먹은 군과 B 고혈압 약을 먹은 군의 혈압(mmHg) 차이

Table 1. Demographic and clinical characteristics of study participants

Characteristics	Total (n=)
Sex	
Male	
Female	
Age (year)	
Smoking	
No	
Yes	
BMI (kg/m ²)	
FBS (mg/dl)	

BMI, body mass index; FBS, fasting blood sugar.
Data are presented as mean (SD) or n (%).

기술통계학

- 범주형 변수
 - 빈도(%)
- 연속형 변수
 - 평균(표준편차), 중앙값(IQR)
 - 평균을 사용할 것인지, 중앙값을 사용할 것인지 결정하기 위해서는 반드시 정규성 검사를 해보는 것이 바람직함

정규성 검정

- 연속형 변수를 다루는 통계법을 사용할 때 변수의 분포가 정규분포 (normal distribution, Gaussian distribution)을 이룬다는 가정 하에서 사용할 수 있음
- 많은 연구자가 이런 정규분포를 하는지 여부를 검정하지 않고 통계법을 사용하는 실수를 범함
- 정규분포를 하는지 보는 검사법이 정규성 검정(normality test)
 - Shapiro-Wilk test, Kolmogorov-Smirnov test 등
- 정규성 검정의 가설
 - 귀무가설 H_0 : 자료는 정규분포를 따른다.
 - 대립가설 H_1 : 자료는 정규분포를 따르지 않는다.
 - $p \geq 0.05$ 귀무가설 채택 \rightarrow 정규성 O

$n \geq 30$, 중심극한정리



Table 1. Demographic and clinical characteristics of study participants

Characteristics	Total (n=5520)
Sex	
Male	2279 (43.7)
Female	2941 (56.3)
Age (year)	55.1 ± 14.04
Smoking	
No	4289 (82.2)
Yes	931 (17.8)
BMI (kg/m ²)	24.1 ± 3.40
FBS (mg/dl)	102.9 ± 23.87

BMI, body mass index; FBS, fasting blood sugar.
Data are presented as mean (SD) or n (%).

(2군 비교) Demographic and clinical characteristics of study participants by sex

Characteristics	Total (n=)	Male (n=)	Female (n=)	<i>p</i> value
Age (year)	55.1 ± 14.04			
Smoking				
No	4289 (82.2)			
Yes	931 (17.8)			
BMI (kg/m ²)	24.1 ± 3.40			
FBS (mg/dl)	102.9 ± 23.87			

BMI, body mass index; FBS, fasting blood sugar.

Data are presented as mean (SD) or n (%).

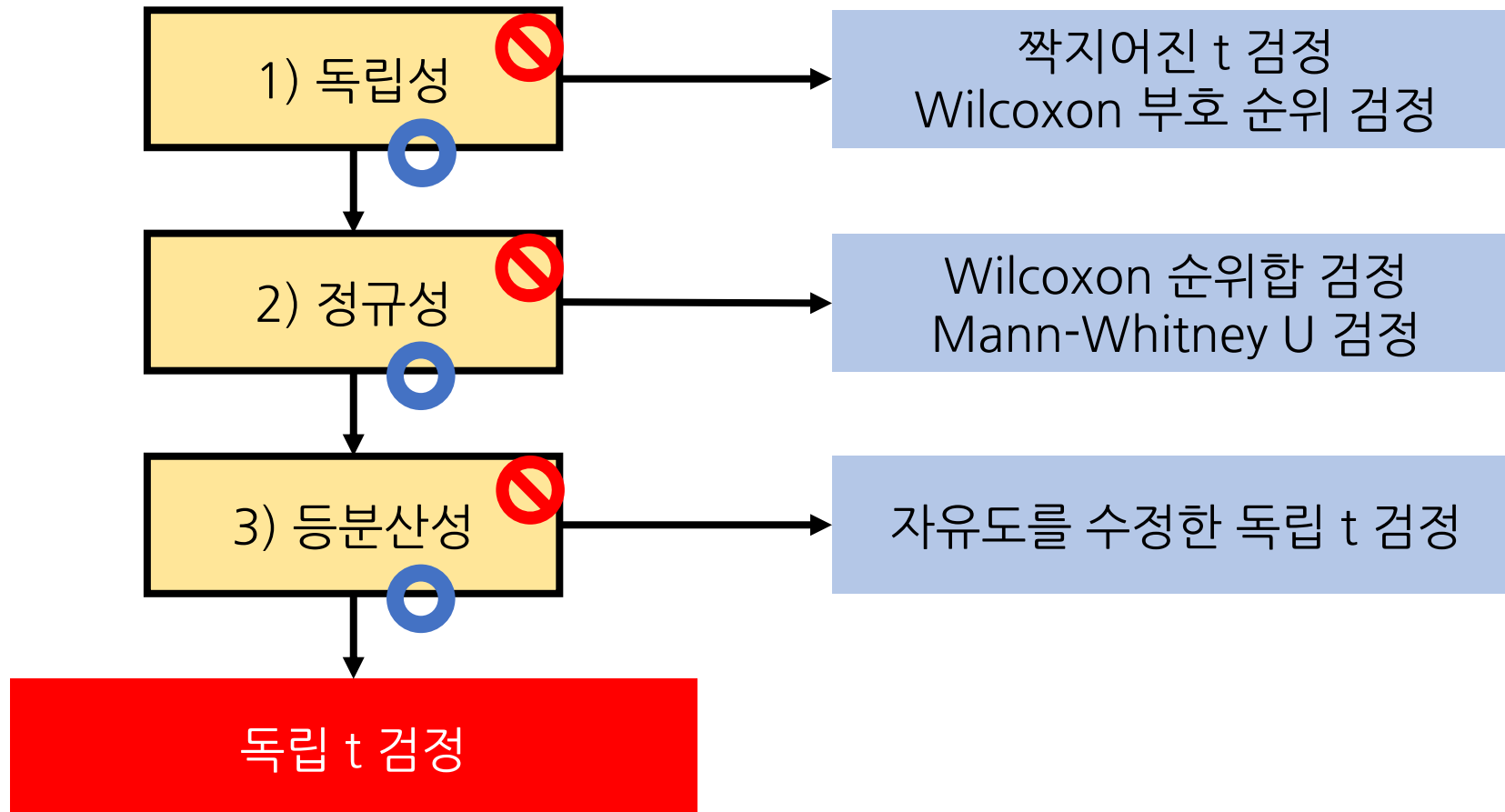
군 간 차이 검정을 위한 통계분석법

		독립적 결과변수				
		연속형		범주형		시간관찰
		정규분포	비정규분포	순서형	명목형	이분형*시간
명목형 설명변수	두 군	독립 t-검정	Wilcoxon 순위합 검정 Mann-Whitney U 검정	카이제곱 검정 Fisher의 정확성검정	Log-rank 검정	
	세 군 이상	일원분산분석	Kruskal-Wallis 검정	카이제곱 검정		
		의존적 결과변수				
		연속형		범주형		
		정규분포	비정규분포	순서형	명목형	
명목형 설명변수	두 군	짝지어진 t-검정	Wilcoxon 부호 순위 검 정	McNemar 검정		
	두 군 이상	반복측정 분산분석	Friedman 검정	Cochran Q 검정		

독립 t 검정

- Independent t-test, student's t-test, independent sample t-test
- 연구 집단을 2개의 군으로 나누어서 관심 있는 결과의 값에 차이가 있는 지 보는 것
- 독립변수: 명목형 변수
 - 1군/2군, 투약군/위약군, 수술군/약물치료군, A 수술군/B 수술군, 남녀, 한국인/미국인
- 종속변수: 연속형 변수
 - 혈압, 키, 체중

독립 t 검정을 위한 가정



정규성 검정

- 정규분포를 하는지 보는 검사법이 정규성 검정(normality test)
 - Shapiro-Wilk test, Kolmogorov-Smirnov test 등
- 정규성 검정의 가설
 - 귀무가설 H_0 : 자료는 정규분포를 따른다.
 - 대립가설 H_1 : 자료는 정규분포를 따르지 않는다.
 - $p \geq 0.05$ 귀무가설 채택 \rightarrow 정규성 O

등분산성 검정

- 분산이 일정한 지 보는 검사법이 등분산성 검정(homogeneity of variance test)
 - Levene test 등
- 등분산성 검정의 가설
 - 귀무가설 H_0 : 두 군의 분산은 같다.
 - 대립가설 H_1 : 두 군의 분산은 같지 않다.
 - $p \geq 0.05$ 귀무가설 채택 \rightarrow 등분산성 O



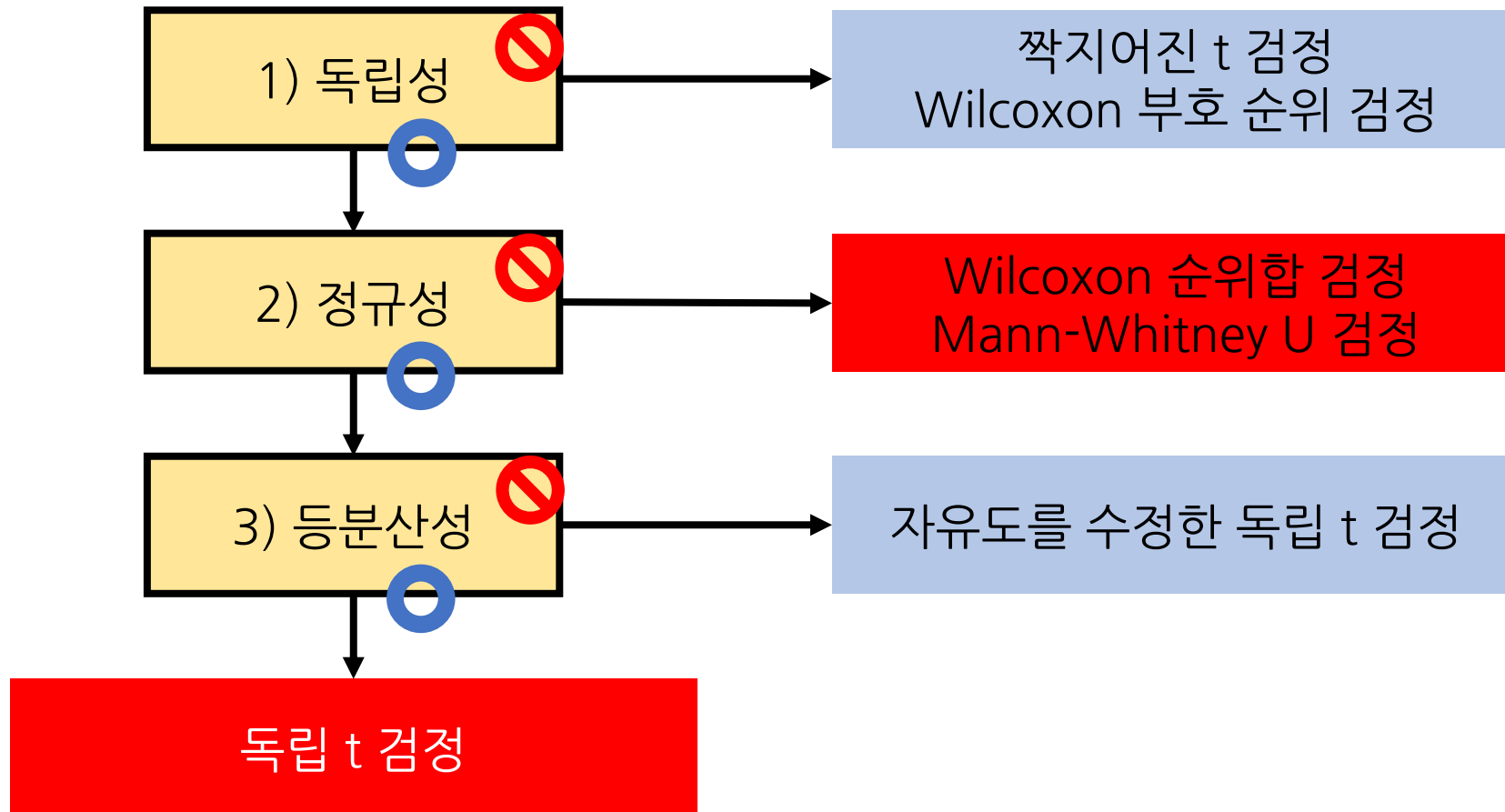
(2군 비교) Demographic and clinical characteristics of study participants by sex

Characteristics	Total (n=5520)	Male (n=2279)	Female (n=2941)	<i>p</i> value
Age (year)	55.1 ± 14.04	54.9 ± 14.10	55.3 ± 13.99	0.383
Smoking				
No	4289 (82.2)			
Yes	931 (17.8)			
BMI (kg/m ²)	24.1 ± 3.40	24.6 ± 3.17	23.7 ± 3.53	<0.001
FBS (mg/dl)	102.9 ± 23.87	106.9 ± 27.51	99.8 ± 20.07	<0.001

BMI, body mass index; FBS, fasting blood sugar.

Data are presented as mean (SD) or n (%).

독립 t 검정을 위한 가정



비모수 검정

- 독립 t 검정의 비모수적 검정: Wilcoxon 순위합 검정(Wilcoxon rank sum test), Mann-Whitney U 검정
- 종속변수가 정규분포를 이루지 않거나 서열변수(순서형)일 때



군 간 차이 검정을 위한 통계분석법

		독립적 결과변수				
		연속형		범주형		시간관찰
		정규분포	비정규분포	순서형	명목형	이분형*시간
명목형 설명변수	두 군	독립 t-검정	Wilcoxon 순위합 검정 Mann-Whitney U 검정	카이제곱 검정 Fisher의 정확성검정	Log-rank 검정	
	세 군 이상	일원분산분석	Kruskal-Wallis 검정	카이제곱 검정		
		의존적 결과변수				
		연속형		범주형		
		정규분포	비정규분포	순서형	명목형	
명목형 설명변수	두 군	짝지어진 t-검정	Wilcoxon 부호 순위 검 정	McNemar 검정		
	두 군 이상	반복측정 분산분석	Friedman 검정	Cochran Q 검정		

카이제곱 검정

- Chi-squared test, X^2 test
 - 연구 집단을 n개의 군으로 나누어서 관심 있는 결과의 빈도수에 차이가 있는 지 보는 것
 - 독립변수: 명목형 변수
 - 종속변수: 명목형 변수
- | 2*2 | 남성 | 여성 |
|-----|-------|-------|
| 비흡연 | 1502명 | 2787명 |
| 흡연 | 777명 | 751명 |
- 피셔의 정확성검정(Fisher's exact test): 샘플의 숫자가 작을 때 사용(기대빈도가 5 이하인 셀이 20% 이상일 때 사용)



(2군 비교) Demographic and clinical characteristics of study participants by sex

Characteristics	Total (n=5520)	Male (n=2279)	Female (n=2941)	<i>p</i> value
Age (year)	55.1 ± 14.04	54.9 ± 14.10	55.3 ± 13.99	0.383
Smoking				<0.001
No	4289 (82.2)	1502 (65.9)	2787 (94.8)	
Yes	931 (17.8)	777 (34.1)	154 (5.2)	
BMI (kg/m ²)	24.1 ± 3.40	24.6 ± 3.17	23.7 ± 3.53	<0.001
FBS (mg/dl)	102.9 ± 23.87	106.9 ± 27.51	99.8 ± 20.07	<0.001

BMI, body mass index; FBS, fasting blood sugar.

Data are presented as mean (SD) or n (%).