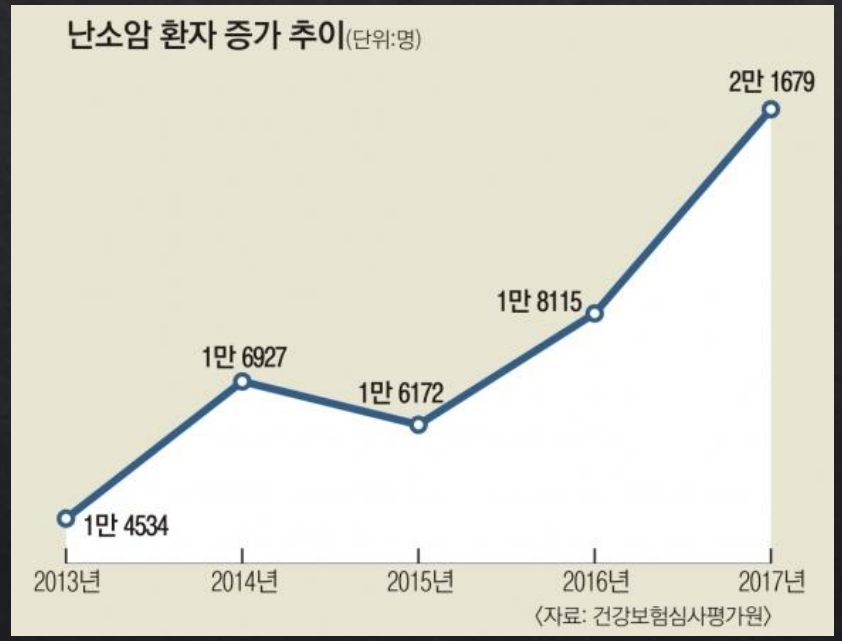
난소암의 백금계항암제 반응 예측 모형 개발

문정섭(2020-21324) 김민준 (2013-13430) 백대현(2014-10451)

난소암 백금계 항암제 반응

- “3대 부인암”으로 분류되는 난소암
- 최근 년 **10%**씩 증가 추세
- 그 중 고등급 장액성 난소암(HGSOC) 환자들의 백금계 항암제 치료 후 효과(TFI)를 연구

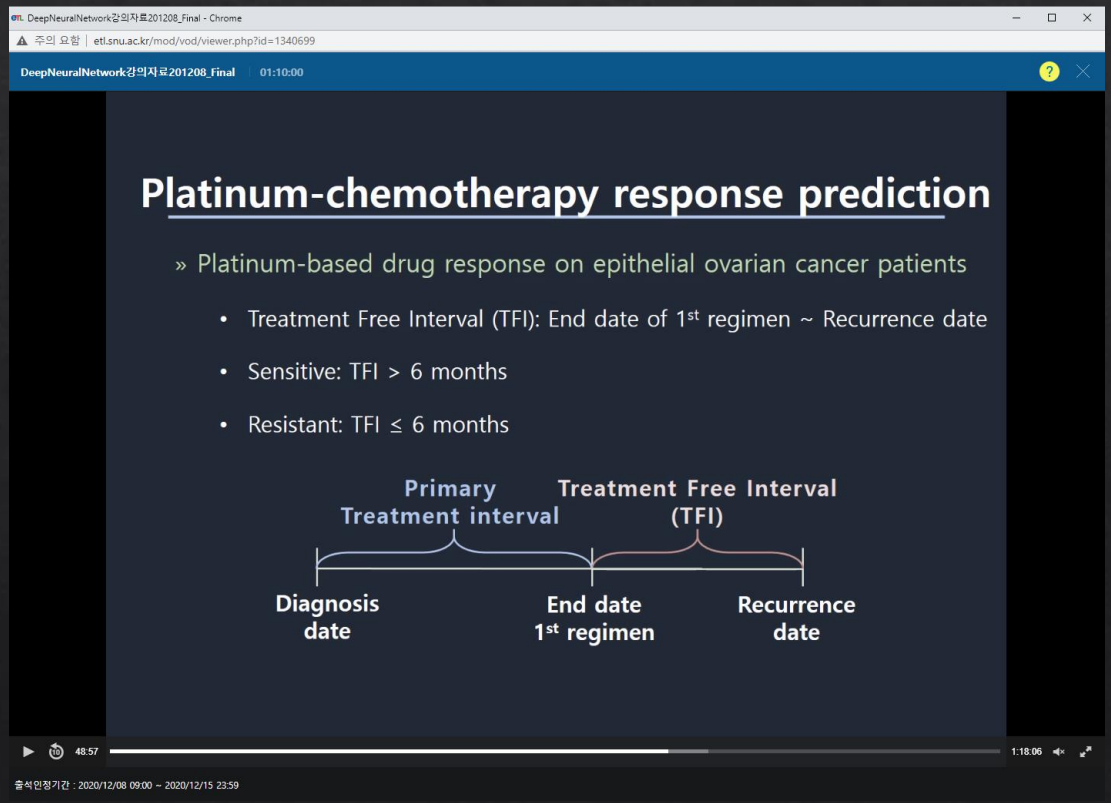


주제가 어디서 들어본 것 같은데...

- 공교롭게도 이번 주 화요일 수업에서
  정확히 같은 주제의 예시가 등장!

- 다만 이미 작업의 막바지여서 cross-check정도
  도움만 받았음

- 두 연구의 차이를 주목 (⇔ 하늘색으로 표기)



Data

- 3개의 기관에서 받은 총 n=1002건의 HGSOC 임상 데이터 (⇔2개기관 n=638)
- 총 46개의 변수 (⇔65개)

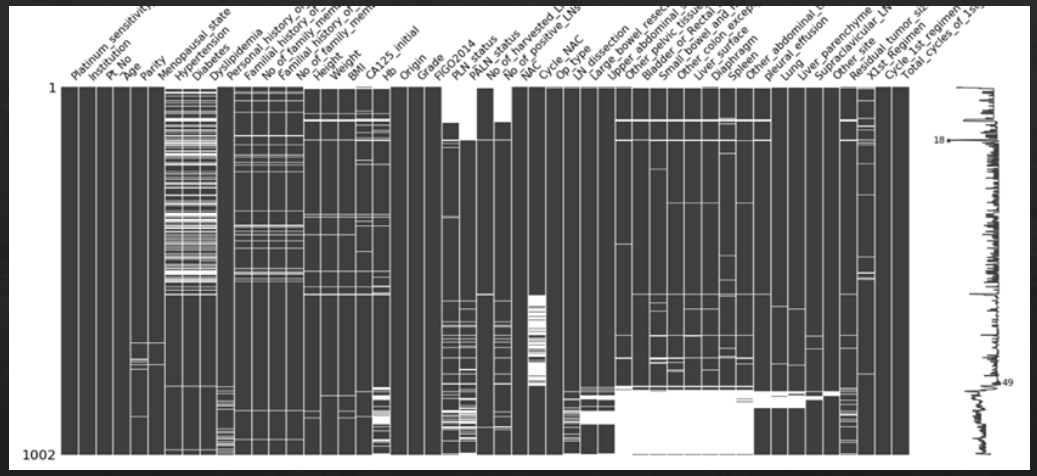
Institution	Resistant	Sensitive	Total
SNUH	114	454	568
AMC	62	184	246
Severance	47	141	188
Total	223	779	1002

chisq test: p=0.1636 [no resistance rate difference per institution]

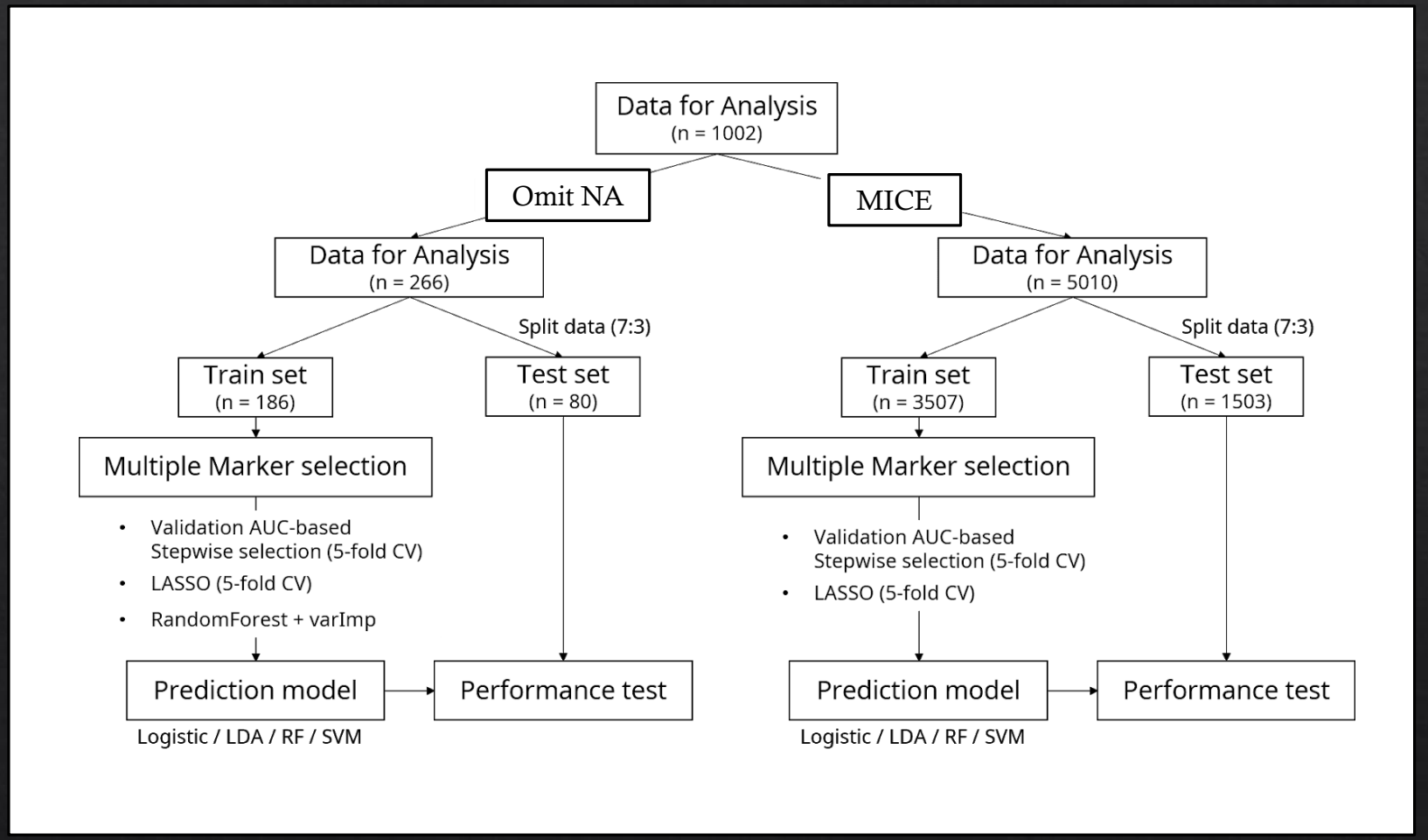
Features	Means ± σ
Age	55.84 ± 10.42
Height	156.12 ± 5.70
Weight	57.47 ± 8.61
BMI	23.58 ± 3.45
Hemoglobin	12.26 ± 1.31

주요 착안점

- 데이터에 Missing Data가 많은 데 어떻게 처리할 것인가?
  - OmitNA vs MICE
- Feature를 어떻게 선택할 것인가?
  - Lasso vs Stepwise(AUC based)
- Prediction 모델을 어떻게 만들 것인가?
  - Logistic vs LDA vs RandomForest vs SVM
  - (⇔ Deep Learning Model)
- 총 2x2x4 = 16 건의 모델을 만든 후 비교



Workflow


Data for Analysis
(n = 1002)
Omit NA
MICE
Data for Analysis
(n = 266)
Data for Analysis
(n = 5010)
Split data (7:3)
Split data (7:3)
Train set
(n = 186)
Test set
(n = 80)
Train set
(n = 3507)
Test set
(n = 1503)
Multiple Marker selection
Multiple Marker selection
• Validation AUC-based Stepwise selection (5-fold CV)
• LASSO (5-fold CV)
• RandomForest + varImp
• Validation AUC-based Stepwise selection (5-fold CV)
• LASSO (5-fold CV)
Prediction model
Performance test
Prediction model
Performance test
Logistic / LDA / RF / SVM
Logistic / LDA / RF / SVM

Variables Selected

Data	Selection Method	Selected Variables	Selected Variables
NA omitted	Stepwise(AUC)	13 variables Hypertension + Diabetes + Dyslipidemia + No_of_family_member_with_breast_cancer_upto_2nd_degree + Familial_history_of_gynecologic_cancer + CA125_initial + No_of_harvested_LNs + NAC + Upper_abdominal_surgery + Bladder_or_Rectal_mucosa + pleural_effusion + Lung + Residual_tumor_size_1st_debulking	3 variables log_CA125initial, Response_1st_regimen, Pleural_effusion
NA omitted	Lasso	7 variables Hypertension + CA125_initial + PLN_status + No_of_harvested_LNs + NAC + Liver_surface + Residual_tumor_size_1st_debulking	9 variables Age, SBP, DBP, No_of_harvested_LNs,No_of_positive_LNs, log_CA125initial, count_Segmentedneutrophil, count_Lymphocyte, Response_1st_regimen1
MICE	Stepwise(AUC)	45 variables (All except Diabetes)	Stepwise와 상동
MICE	Lasso	41 variables (All except Familial_history_of_breast_cancer, BMI, Large_bowel_resection, Liver_surface, Supraclavicular_LN)	Lasso와 상동

Results (NA omitted)

Feature Selection	Prediciton Model	Training Acc	Training AUC	Test Acc	Test AUC	Test AUC
Lasso	Logistic	0.81	0.85	0.75	0.66	0.7215
Lasso	LDA	0.81	0.84	0.75	0.68	-
Lasso	RF(1)	0.85	0.71	0.79	**0.72**	**0.7250**
Lasso	SVM(1)	0.84	0.67	0.82	0.62	0.6718
Stepwise(AUC)	Logistic	0.82	0.80	0.79	**0.74**	**0.7184**
Stepwise(AUC)	LDA	0.82	0.80	0.75	0.71	-
Stepwise(AUC)	RF(2)	0.88	0.71	0.80	0.71	0.6806
Stepwise(AUC)	SVM(2)	0.82	0.69	0.75	0.66	0.6225

RF(1): (mtry=2) tuneGrid (1:13)
RF(2): (mtry=2) tuneGrid (1:7)
SVM(1): (sigma=0.098, C=2) tuneLength=50
SVM(2): (sigma=0.015, C=0.2) tuneLength=50

10 fold CV를 통해 AUC 기준으로 선정한 tuning parameter

Results (MICE Imputed)

Feature Selection	Prediciton Model	Training Acc	Training AUC	Test Acc	Test AUC	Test AUC
Lasso	Logistic	0.80	0.78	0.78	0.75	0.7215
Lasso	LDA	0.80	0.78	0.78	0.74	-
Lasso	RF(1)	1	1	0.99	1	0.7250
Lasso	SVM(1)	0.93	0.91	0.89	0.92	0.6718
Stepwise(AUC)	Logistic	0.80	0.77	0.78	0.75	0.7184
Stepwise(AUC)	LDA	0.80	0.77	0.78	0.74	-
Stepwise(AUC)	RF(2)	1	1	0.99	1	0.6806
Stepwise(AUC)	SVM(2)	0.93	0.91	0.90	0.91	0.6225

RF(1): (mtry=11) tuneGrid (1:20)
RF(2): (mtry=13) tuneGrid (1:20)

SVM(1): (sigma=0.011, C=1) tuneLength=3
SVM(2): (sigma=0.012, C=1) tuneLength=3

결과 해석

- MICE로 Imputation을 하는 경우 Test acc, auc가 너무 좋게 나왔다
  - Train / Test 분리가 잘 안 된 것으로 추정
  - Variable selection과정까지는 문제가 없음
  - **최종 보고서에는 보완할 예정**
- 선행연구와 똑같이 Lasso방식에서는 RF가, Stepwise에서는 Logistic이 AUC가 좋았다
  - 데이터가 겹쳐서 그런 것인지 아니면 feature selection과 model의 관련이 있는지?
- NA를 omit하여 n=266임에도 불구하고 선행연구와 모델의 성과가 비슷하거나 좋았다.
  - Test 데이터가 같지 않아 절대적인 비교는 불가능하지만

Discussion

- MICE Imputation을 하는 경우 제대로 Test/Train 분리를 어떻게 해야하는가?
- Selected Variable이 너무 많을 경우 모델 해석을 어떻게 해야 하는가?
- 난소암에서 유전력과 breast cancer여부는 필드에서 known factor인데 선행연구과 저희 연구 둘 다 variable selection 단계에서 잡아내지 못한 이유는?