난소암 백금계 항암제 반응 예측 모형 개발 보고서

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

Introduction

2019년 12월 국립암센터에서 발표한 국가암등록사업 연례 보고서(2017년 암등록통계)에 따르면, 난소암은 2017년 여성 전체 암 중 8번째로 빈번히 발생하는 암으로, 1999년부터 2017년까지 꾸준히 증가하는 추세이다. 낮은 등급의 난소암의 경우 수술만으로도 높은 생존율을 보이지만, 높은 등급의 난소암의 경우에는 항암치료를 병행하여 생존율을 높이는 치료를 진행하고 있다(1). 따라서 높은 등급의 난소암 환자의 항암치료 효율성을 높이기 위해, 여기서는 환자의 임상 자료를 기반으로 백금 계열의 항암제 반응을 예측하기 위한 마커를 찾고, 그 마커들을 통해 반응성을 예측하기 위한 여러 통계 모델들을 제시하고 비교 할 것이다.

Method

서울대학교병원, 서울아산병원, 그리고 신촌세브란스병원의 난소암 환자 임상자료 각각 568례, 246례, 188례를 서울대학교 생물정보통계 연구실을 통해 전달 받았다. 1002례의 임상자료에서 결측값이 다수 존재하여**(그림 1)**, MICE package를 이용한 결측치 예측하는 방법과 결측치를 가진 임상 자료를 제거하는 방법(complete case analysis)을 진행하였다**(그림 2)**. MICE package는 데이터의 특성에 따라 predictive mean matching(pmm), logistic regression model(logreg), Bayesian polytomous regression model(polyreg)의 3가지 방법 중 적절한 것을 사용하여 결측값을 대체한다. 이 방법들은 R의 mice package mice 함수의 default method로 pmm 방법은 이산형 자료, polyreg 방법은 범주 개수가 2보다 큰 범주형자료에도 사용될 수 있어 이들을 결합하였을 때 이산형, 범주형, 3 이상의 다범주형 변수에 대하여 plausible한 값으로 대체 할 수 있다. 이렇게 결측치가 적절히 대체된 변수는 그것이 결측치가 전혀 없이 complete한 상태와 비교하였을 때 상대적으로 분산이 작아지므로 5개의 데이터 셋을 생성하고 그것을 따로따로 분석하는 과정을 거쳐 이를 통해 변동성을 추가해 감소한 분산을 어느 정도 상쇄하게 된다.

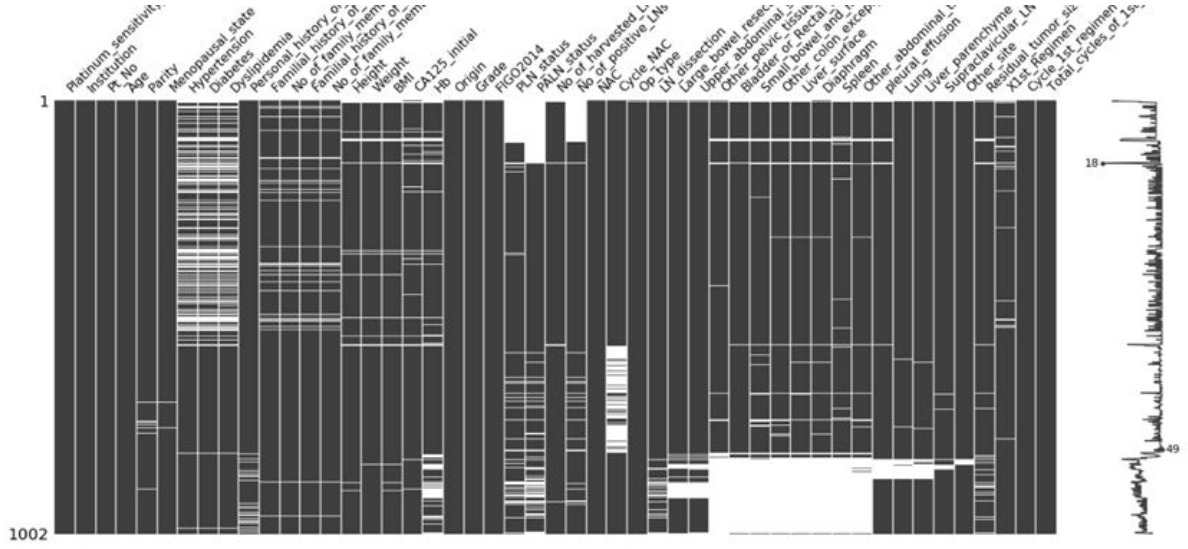

그림 1. 결측치 시각화. Missingno plot을 사용하여 1002례의 샘플의 결측값을 나타내었다.

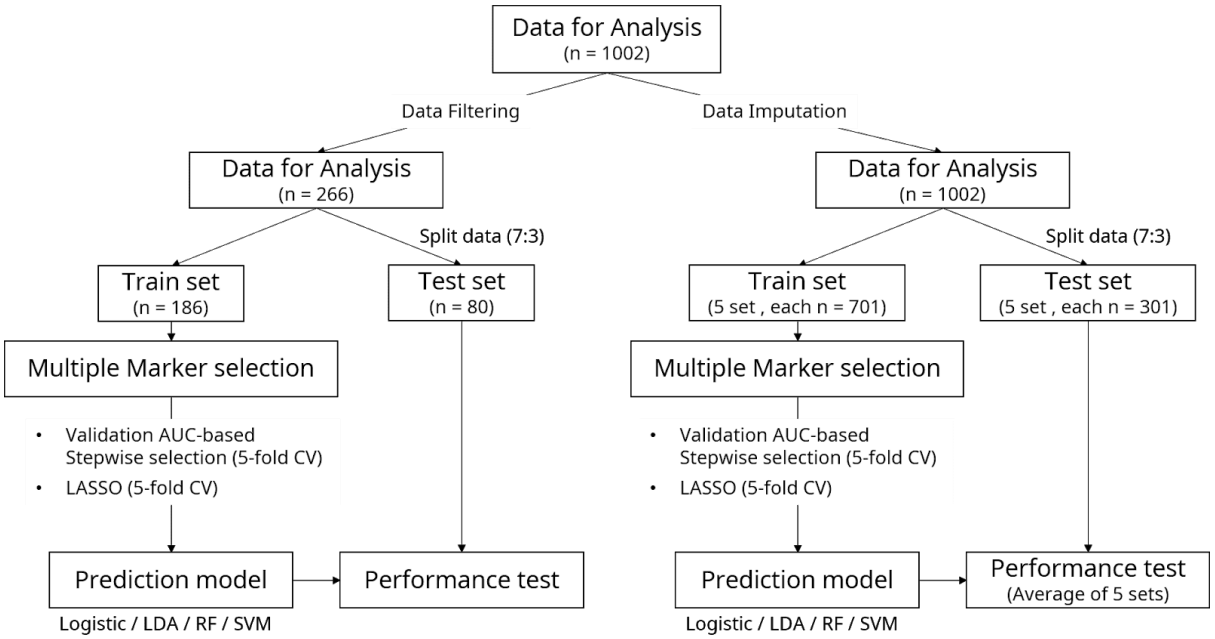

그림 2. 항암제 반응 예측 모형 개발 절차.

그림 2의 Workflow는 기존의 데이터셋에 들어있는 결측치들을 어떻게 처리할 것이냐에 따라 2가지 갈래로 나뉜다. 한 방법은 NA 값을 포함하는 모든 열을 제거하는 Complete cases analysis이며, 다른 방법은 Complete cases analysis의 경우 기존 데이터의 약 30%(1002에서 266으로 축소) 밖에 사용하지 못하게 되는 점을 고려, Multiple imputation을 이용하여 기존 데이터와 같은 차원을 갖는 5개의 multiply imputed 데이터셋을 만든 뒤 그것들을 독립적으로 생각하여 따로따로 분석하였다. missing value가 처리된 데이터셋들은 공통적으로 Stepwise selection, Lasso 방법을 적용하여 차원 축소 과정을 거친다. 이를 통해 모든 종속변수를 그대로 사용하였을 때 발생하는 모델의 과한 복잡성과 해석의 어려움을 줄일 수 있다. 그 결과 50개에 가까운 종속변수 중 prediction model에 실제로 사용되는 마커들은 10개 내외로

줄어든다. 차원축소, 즉 변수 선택은 AUC 기반, 5-fold cross validation을 이용한 stepwise selection과 5-fold cross validation을 이용한 LASSO, 2가지 기법을 모두 적용하였다.

Multiply imputed 데이터셋의 경우에는 적절한 변수 선택하는 것이 challenging할 수 있다. 그 이유는 서로 다른 5개의 데이터셋에 차원 축소 기법을 적용하면 5가지 서로 다른 변수 선택 결과가 도출되고, 이 5개의 선택된 변수 집합 중 어떤 것을 최종 선택해야하는지를 결정하는 것이 까다롭기 때문이다. 이에 대해 stef van buuren은 5개의 데이터셋에 변수 공통된 변수 선택 기법을 적용한 뒤 전체에 걸쳐 50% 이상의 비율로 등장하는 변수들만 고려하라고 권장한다(Majority method)(2). 구체적으로 설명하면 5개의 데이터셋을 각각 MI_1, …, MI_5라고 두고 각각 변수 선택 결과 얻은 final formula가 f_1, .. , f_5라고 했을 때 f_1,.. f_5에 모두 포함되는 변수(100%의 비율로 등장), 이들 중 4개에만 포함되는 변수(80%의 비율), 3개에만 포함되는 변수(60%의 비율)들을 사용하는 것을 고려한다. 5번 모두 등장한 변수는 모두 최종 formula에 포함하고, 4번 또는 3번만 포함되는 변수는 좀 더 확실한 statistical justification을 위해 Wald test를 거쳐 변수를 최종 formula에 포함해도 되는지 여부를 결정한다. 이렇게 Majority method을 이용하여 구성한 최종 formula는 5개의 데이터셋 전체에 공통적으로 사용되어 바로 다음 단계에 이어질 prediction model 적합 과정의 분석 대상이 된다.

Variable selection으로 선택된 마커들로 Logistic Model, LDA Model, RandomForest, Support Vector Machine(이하 SVM) 을 이용한 prediction model을 만들었다**.** Prediction model은 tuning hyperparameter의 유무에 따라 Logistic과 LDA, 그리고 Random forest와 SVM으로 나뉜다. Logistic과 LDA에서는 R의 Mass package를 이용하여 confusion matrix와 ROC curve를 그리고 이로부터 train, test accuracy와 train, test auc를 구할 수 있다.

Tuning hyperparameter가 존재하는 Random forest와 SVM 분석에서는 위와 같은 성능 측도 계산 이전에, best parameter 값을 찾아내야한다. 본 연구에서는 R의 caret package를 이용하여 10-fold cross validation을 거쳐, training AUC가 가장 큰 모델을 제공하는 parameter의 조합을 선택하는 방식을 택하였다. 그리고 본 연구에서는 caret에서 성능 측정의 default metric은 training accuracy이지만, training accuracy보다는 AUC가 모델 성능의 척도로서 의미있다고 판단하여 metric을 AUC로 바꾸어 진행하였다.

모델 성능 평가 척도와 더불어 tune grid의 설정 역시 tuning parameter를 선정하는데에 있어서 굉장히 중요하다. grid의 범위가 너무 좁다면 best parameter를 발견하지 못할 가능성이 상당히 증가하므로 적절한 grid의 범위를 고려하여 사용하는 것이 중요하다. caret에서 training data 학습을 수행하는 train() 함수에서는 tune grid를 설정하는 방법은 크게 2가지로 나뉜다. 사용자 검색 그리드(custom search grid)를 이용하면 사용자가 직접 원하는 범위를 설정해 그 범위 내에서 best parameter을 찾게 된다. 예를 들어, Random forest의 tuning parameter인 mtry(각각의 tree node마다 고려할 후보 변수의 수)를 결정하기 위해 tune grid를 1부터 10까지의 자연수로만 제한하여 살펴볼 수 있다. 본 연구에서는 Randomforest 분석에서 사용자 검색 그리드를 사용, mtry grid를 1부터 fomula에 속한 마커

개수의 절반까지로 설정하였다. 이것에 대한 justification은 mtry가 너무 크면 decision tree의 tree의 분기를 해석하는 것이 어려워진다는 점, 그리고 practical하게 mtry = 총 변수 개수의 ½ 또는 약 ⅓ 을 취한다는 것이다. SVM의 경우 tuning parameter은 C와 sigma로 2개인데, sigma는 scale에 관련된 것이고 C는 misclassification cost로서 이 값이 크다면 해당 모델은 주어진 training data를 mis classify 하지 않는데에 굉장한 비중을 두게 된다(그리고 이것은 future data를 대비한 wiggle room을 남겨두지 않는다는 의미이기도 하다). Best model을 제공하는 sigma, C 값의 조합을 찾는 것은 2차원 모수공간에서 단 하나의 최적 target point에 최대한 근접한 point를 찾는 문제와 같다. 이를 위해서 랜덤 검색 그리드(random selection of tuning parameter combinations) 방법을 사용하였는데 이 방법은 SVM과 같이 tuning parameter가 2개 이상인 경우 사용되며, 랜덤하게 searching할 point의 개수(tuneLength)를 지정하여 최적의 target point에 근접한 값을 찾도록 하였다(tuneLength가 길수록 computation은 늘지만 최적의 target point에 근접할 가능성은 커진다). 이렇게 사용자 검색 그리드, 랜덤 검색 그리드를 각각 Random forest, SVM에 적용하여 tuning parameter를 찾은 뒤에는 4가지 모델 성능 척도(train accuracy, train AUC, test accuracy, test AUC)를 계산하였다.

Complete case analysis에서는 위에 제시된 방법을 통해 모델 성능 척도를 unique하게 계산할 수 있지만, Multiple Imputation analysis에서는 5개의 서로 다른 데이터셋에 대하여 서로 다른 성능 척도 값이 나오므로 결과값들을 pool하여야 한다. 5개의 데이터셋에 variable selection with Majority method를 이용해 얻은 최종 formula를 공통으로 사용, 총 4개의 prediction model에 대한 총 4가지 모델 성능 척도 값을 구한 뒤 각각 얻어지는 5개의 값에 대한 평균을 최종 성능 척도로 판단하였다. 예를 들면 MI_1, .. ,MI_5 데이터에 각각 LDA를 적용하였을 때 얻어지는 test AUC를 test_AUC_1,.. test_AUC_5라고 하면 이 5개의 AUC 값을 평균한 것이 multiple imputation method에서의 LDA prediction model에 대한 test AUC라고 보는 것이다(각 데이터셋 안에서 train, test split을 하였고 각 데이터셋에서 얻어지는 test data를 통해 그 데이터셋의 training data를 통해 얻어진 모델을 평가, 5개의 데이터셋으로부터 얻어지는 총 10개의 train test data간의 독립성을 보장하였음을 밝힌다). 평균값을 취한 것에 대한 justification은 5개의 데이터셋의 중요도를 동일하게 보아야하며 따라서 weight를 동일하게 주어 평가해야한다는 점이다.

Result

임상자료에서 사용된 변수는 총 46개 종속변수와 1개의 독립변수로 이는 아래와 같이 구성되어 있다. 종속변수는 카테고리 변수 36개 (Parity, Menopausal state, Hypertension, Diabetes, Dyslipidemia, Personal history of breast cancer, Total cycles of 1st regimen, Familial history of breast/gynecologic cancer, Origin, Grade, FIGO2014, PLN/PALN status, NAC, Cycle NAC, Cycle 1st regimen, Op_type, LN dissection, Large bowel resection, pleural effusion, Upper abdominal surgery, Other pelvic tissue except Ut tube and LN, Bladder or Rectal mucosa, Small bowel and mesentery, Other colon except rectosigmoid, Liver surface,

Diaphragm, Spleen, Residual tumor size 1st debulking, Other site, 1st Regimen, Other abdominal tissue outside pelvis, Lung, Liver parenchyme, Supraclavicular LN)와 연속형 변수 10개(Age, No of family member with breast/gynecologic cancer upto 2nd degree, Height, Weight, BMI, CA125, Hb, No of harvested/positive LNs) 이다.

먼저 결측치가 제거된 Complete cases analysis의 경우, Stepwise selection을 통해 선택된 마커들은 총 13개로 아래 표 2에 서술되어 있다. 이 마커들로 네가지 모델에 적합 한 경우, SVM모델과 Random Forest모델에서 각각 0.80, 0.79의 test accuracy를 보이며 좋은 성능을 보였으나, SVM모델은 낮은 specificity로 인해 test AUC가 random forest모델보다 0.1가량 낮은 값인 0.62를 보였다. 따라서 sensitivity와 specificity를 모두 고려한 Random forest 모델이 가장 좋은 성능을 보였다고 할 수 있다.

또한 Lasso로 변수를 추려내었을 경우, Stepwise selection의 추려낸 변수보다 더 적은 7개의 변수를 추려낸 한편 PLN_status와 Liver_surface가 새롭게 추가 되었다. 이 변수들로 모델을 적합 한 경우 Logistic regression model에서 test auc 0.74로 가장 좋은 성능을 보였다. Stepwise selection을 통한 Random Forest와 Lasso를 통한 Logistic regression model의 test AUC 차이는 0.02로 비슷한 성능을 보였지만, 더 적은 변수로 비슷한 성능을 보인 Lasso-Logistic regression model이 Complete cases analysis에 더 적합한 모델이라 생각된다.

NA preprocessing	**Complete cases(remove all rows containing NA)**							
variable selection method	AUC-based stepwise selection				lasso			
selected variables (formula)	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				7 variables Hypertension + CA125_initial + PLN_status + No_of_harvested_LNs + NAC + Liver_surface + Residual_tumor_size_1st_debulking			
prediction model	**Logistic**	**LDA**	**RF (mtry=2) tuneGrid (1:13)**	**SVM (sigma=0.098, C=2) tuneLength=50**	**Logistic**	**LDA**	**RF (mtry=2) tuneGrid (1:7)**	**SVM (sigma=0.015, C=0.2) tuneLength=50**
training accuracy	0.81	0.81	0.85	0.84	0.82	0.82	0.88	0.80
training AUC	0.85	0.84	0.71	0.67	0.80	0.80	0.71	0.69
test accuracy	0.75	0.75	0.79	0.80	0.79	0.75	0.80	0.75
test AUC	0.66	0.68	0.72	0.62	**0.74**	0.71	0.71	0.66

표2. Complete cases analysis의 Prediction model 성능 비교.

기존 데이터와 같은 차원을 가진 5개의 Multiply imputed 데이터셋에서 Stepwise selection을 통해 공통으로 추려낸 변수는 총 9개로 아래 표에 명시되어있다**(표 3)**. 아래 변수에 추가적으로 No_of_positive_LNs와 Familial_history_of_breast_cancer가 각각 4개와 3개의 데이터셋에서 포함되었지만, Wald test의 p-value 값이 각각 0.176과 0.475로 최종 모델에 불필요하다 판단되어 제외하였다. Complete cases analysis와 마찬가지로 Logistic regression, LDA, Random forest, SVM을 이용한 prediction model 적합에서 SVM prediction model이 test AUC 0.70으로 가장 높은 값을 나타내었다.

Lasso를 통해 공통으로 추려낸 변수는 총 10개이며, Stepwise selection에서의 No_of_harvested_LNs가 제외되고, FIGO2014와 Total_cycles_of_1st_regimen이 추가되었다. FIGO_2014는 종양의 범위와 림프절 전이 및 원격 전이 등을 근거로 암의 병기를 나타내는 변수이다(3). 암의 병기는 예후에 직접적인 연관이 있지만, Prediction model에서

Stepwise selection된 모델과 큰 성능차이를 보이지 않았고, 더 나아가 FIGO stage를 포함하지 않는 stepwise selection된 SVM prediction model이 가장 좋은 성능을 보여주었다.

NA preprocessing	Multiply imputed data(pooled result)							
variable selection method	AUC-based stepwise selection				Lasso			
selected variables (formula)	9 variables CA125_initial + Liver_parenchyme + Menopausal_state + NAC + No_of_harvested_LNs + Op_type + PLN_status + Residual_tumor_size_1st_debulking + Small_bowel_and_mesentery				10 variables CA125_initial + FIGO2014 + Liver_parenchyme + Menopausal_state + NAC + Op_type + PLN_status + Residual_tumor_size_1st_debulking + Small_bowel_and_mesentery + Total_cycles_of_1st_regimen			
prediction model	Logistic	LDA	RF	SVM	Logistic	LDA	RF	SVM
training accuracy	0.79	0.79	0.77	0.80	0.78	0.78	0.78	0.81
training AUC	0.77	0.76	0.70	0.67	0.77	0.76	0.69	0.63
test accuracy	0.78	0.78	0.80	0.79	0.78	0.77	0.80	0.78
test AUC	0.67	0.67	0.63	**0.70**	0.65	0.65	0.63	0.57

표3. Multiply imputed data analysis의 Prediction model 성능 비교(성능 척도의 평균값).

Discussion and Limitation

앞서 언급된바 총 1002례의 데이터 중, complete cases는 단지 266례로 전체 데이터의 약 30%에 불과하였다. 비록 모델 성능은 괜찮았지만 기존 데이터의 상당 부분을 소실한 점은 여전히 큰 한계점으로 작용한다. 또한 Multiply imputed data 에 사용된 변수 중 NAC, FIGO2014, 그리고 Total_cycles_of_1st_regimen을 제외한 모든 변수들은 각 selection method 별 평균 6.5%에서 8.5%정도로 분포해있다. 따라서 본 연구에서 적합한 Prediction model들의 성능은 추가적인 검증이 필요할 것이라 생각된다. 또한 본 연구에서는 비슷한 성능이라면 변수가 적게 선택된 것이 더 좋은 모델이라는 판단을 하였지만, 개수와 별개로 실제 어떤 변수가 선택되었고 그것이 실제적으로 해석이 되는지에 대한 여부도 중요하다. test AUC가 높고 변수 개수가 적어서 가장 좋다고 판단한 모델이 실제 이 분야 관련 전문가가 보았을 때 전혀 의미 없는 변수를 포함한다면, 그 모델은 결코 제일 좋다고 할 수 없을 것이다. 따라서 선택된 변수를 가지고 이것을 채택하여야 하는지에 관해 관련 전문가들과 논의를 하는 과정이 필요하며, 본 연구에 이러한 과정이 포함되었다면 best model 선정 결과가 통계적뿐만 아니라 실무적으로도 justified될 수 있을 것이다.

Reference

1. Suh, D. H., Chang, S. J., Song, T., Lee, S., Kang, W. D., Lee, S. J., ... & Kim, H. S. (2018). Practice guidelines for management of ovarian cancer in Korea: a Korean Society of Gynecologic Oncology Consensus Statement. *Journal of gynecologic oncology*, *29*(4).
2. “Flexible imputation of missing data.”, accessed December 10, https://stefvanbuuren.name/fimd/sec-stepwise.html.
3. Paik, E. S., Lee, Y. Y., Lee, E. J., Choi, C. H., Kim, T. J., Lee, J. W., ... & Kim, B. G. (2015). Survival analysis of revised 2013 FIGO staging classification of epithelial ovarian cancer and comparison with previous FIGO staging classification. *Obstetrics & gynecology science*, *58*(2), 124-134.