JDCR
Journal of Digestive Cancer Research
https://doi.org/10.52927/jdcr.2023.11.3.157
pISSN: 2950-9394 eISSN: 2950-9505
Journal of Digestive Cancer Research 2023;11(3):157-164
REVIEW ARTICLE

예측모형의 구축과 검증: 소화기암연구 사례를 중심으로

권용한[1], 한경화[2]

[1]연세대학교 일반대학원 의학전산통계학협동과정, [2]연세대학교 의과대학 영상의학교실 방사선의과학연구소 의료영상데이터사이언스센터

Development and Validation of a Prediction Model: Application to Digestive Cancer Research

Yonghan Kwon[1], Kyunghwa Han[2]

[1]Department of Biostatistics and Computing, Yonsei University Graduate School, [2]Department of Radiology, Research Institute of Radiological Science and Center for Clinical Imaging Data Science, Yonsei University College of Medicine, Seoul, Korea

Received November 1, 2023
Revised December 4, 2023
Accepted December 4, 2023

Corresponding author:
Kyunghwa Han
E-mail: khhan@yuhs.ac
https://orcid.org/0000-0002-5687-7237

Prediction is a significant topic in clinical research. The development and validation of a prediction model has been increasingly published in clinical research. In this review, we investigated analytical methods and validation schemes for a clinical prediction model used in digestive cancer research. Deep learning and logistic regression, with split-sample validation as an internal or external validation, were the most commonly used methods. Furthermore, we briefly introduced and summarized the advantages and disadvantages of each method. Finally, we discussed several points to consider when conducting prediction model studies.

Key Words: Machine learning; Clinical decision rules; Precision medicine

INTRODUCTION

임상연구에서 예측모형에 대한 수요와 연구는 계속 증가하고 있다[1,2]. 예를 들어 소화기암연구에서는 유경성 대장용종에서 내시경 협대역 영상에서의 여러 인자들을 이용하여 T1 대장암 진단을 하기 위한 모형이나[3] 조기 위암에서 병리학적 병기 정보를 활용하여 수술적 절제 후 위 이외의 재발 위험을 예측하는 모형 등이 있다[4]. 이처럼 환자의 기초적인 정보나 영상, 병리학적 소견들을 종합하여 환자의 질병 상태를 진단하거나 예후를 예측하는 것이 예측모형의 목적이고, 이를 위해 전통적으로 통계학적 방법 중 다변수 회귀분석(multivariable regression model)을 이용해왔다[5]. 최근에는 단일 기관뿐 아니라 다기관 자료를 활용하는 경우가 늘어나고 의료인공지능에 대한 연구가 활발하게 이루어지면서 전통적인 통계 방법을 포함하여 랜덤 포레스트, 딥러닝 등 머신러닝에 기반한 모형들이 등장하였고, 모형 구축 방법이 복잡해짐에 따라 예측모형 성능에 대한 다각도에서의 검증이 요구되고 있다.

본 논문에서는 소화기암연구 분야에서의 예측모형 개발과 검증의 형태를 조사하여 정리하고, 각 방법론에 대한 설명과 장단점을 살펴보고자 한다.

MAIN SUBJECTS

소화기암연구 분야의 예측모형 연구 사례

소화기 연구의 예측모형 구축에 사용된 통계학적 모형과 검증 방법을 조사하기 위해 PubMed에서 2019년 1월 1일부터 2023년 6월 30일까지의 논문을 조사하였다.

조사 대상 학술지로는 소화기 분야 관련성과 impact factor를 종합적으로 고려하여 Gastroenterology, Gut,

Endoscopy, Gastrointestinal Endoscopy, Alimentary Pharmacology & Therapeutics, Gastric Cancer, Journal of Gastroenterology, 총 7개의 저널을 선정하여 살펴보았다. 구체적인 PubMed 검색 쿼리는 Supplementary Material에 명시했다.

검색 결과 총 462편의 논문을 선정하였고 초록을 기준으로 예측모형 구축에 사용한 모형 방식과 검증 유형을 분류했다. 두 개 이상의 예측모형 방식과 검증 유형을 사용한 경우 각각 따로 세어서 중복된 경우도 있다.

모형분류 결과 예측모형 방식과 검증 유형을 각각 Tables 1, 2에 정리했다. 소화기 연구에서는 딥러닝(deep learning)과 로지스틱 회귀(logistic regression)가 대표적인 예측모형으로 활용되며, 모형 검증에는 외부 검증(external validation)과 분할 샘플(split sample) 방법이 주로 사용되는 것으로 나타났다. 초록에 구체적인 예측모형 방식을 기술하지 않은 논문은 총 260편(56.3%)이었고, 검증 방식을 파악할 수 없었던 논문은 총 123편(26.6%)이었다.

각 방법에 대한 간략히 소개는 다음과 같고, Table 3에 각 모형의 장단점에 대하여 정리했다.

Table 1. Statistical and Machine Learning Methods Used for Developing a Prediction Model in Gastrointestinal Research

Method	Frequency
Per method	217 (100.0)
Deep learning	82 (37.8)
Logistic regression	62 (28.6)
Cox proportional-hazards regression	36 (16.6)
Random forest	15 (6.9)
Gradient boosting	5 (2.3)
LASSO	4 (1.8)
Support vector machine	3 (1.4)
Others	10 (4.6)
Per article	462 (100.0)
Classified	202 (43.7)
Not classified	260 (56.3)

Values are presented as number (%).
LASSO, Least Absolute Shrinkage and Selection Operator.

Table 2. Validation Methods for a Prediction Model Used in Gastrointestinal Research

Method		Frequency
Per validation method		366 (100.0)
External validation	External validation	227 (62.0)
Internal validation	Split sample	85 (23.2)
	Cross validation	30 (8.2)
	Not specified	12 (3.3)
	Bootstrapping	11 (3.0)
	Repeated random validation	1 (0.3)
Per article		462 (100.0)
Classified		333 (72.1)
Not classified		129 (27.9)
Internal & external both used		38

Values are presented as number (%) or number only.

예측모형 구축 방법

딥러닝(deep learning)

딥러닝은 대용량 데이터에서 패턴을 인식하여, 이를 바탕으로 예측, 분류, 의사 결정 등의 작업을 수행하는 인공지능의 한 분야로, 매우 복잡한 비선형 관계를 모형화할 수 있다는 강점이 있다. 이러한 딥러닝 모형은 여러 개의 은닉층을 포함한 신경망으로 구성되어 있으며, 각 층은

Table 3. Comparison of Machine Learning Methods: Advantages and Disadvantages

Models	Advantages	Disadvantages
Deep learning	• Modeling complex nonlinear relation	• Expensive cost and time • Overfitting • Difficult to understand
Logistic regression	• Easy to understand	• Multicollinearity
Cox proportional-hazards regression	• Time-to-event data • Time-dependent covariate	• Proportional hazard assumption
Random forest	• Prevent overfitting	• Difficult to understand
Gradient boosting	• Parallel construction • Prevent overfitting • Handling missing value	• Memory consumption • Difficult to understand
LASSO	• Variable selection	• Randomly select variable(s) when there are highly correlated variables
Support vector machine	• Effective in heterogeneous high dimensional data • Rare event classification	• Difficult to tune parameters and select kernel

LASSO, Least Absolute Shrinkage and Selection Operator.

https://doi.org/10.52927/jdcr.2023.11.3.157

입력 데이터의 다양한 특징을 추출하고 학습한다. 학습 과정에서는 데이터의 특징점을 찾아내는 것이 중요하며, 이를 위해 초기에는 무작위로 설정된 파라미터를 업데이트 해 나가며 모형의 성능을 개선한다. 이러한 과정을 통해 모형은 데이터의 복잡한 특징들을 자동으로 학습하게 되며, 사전에 정의된 특징이 아닌, 데이터로부터 직접 특징을 학습하므로, 전문가의 지식이 없어도 높은 성능의 모형을 구축할 수 있다는 장점이 있다[6]. 하지만 이러한 딥러닝 모형을 훈련하는 데에는 상당한 자원과 시간이 필요하다. 이 과정에서 딥러닝은 학습 데이터의 노이즈를 학습하면서 과적합이 발생할 수 있으며, 이는 모형의 일반화 능력을 저하시켜 새로운 데이터에 대한 성능을 떨어뜨릴 수 있다. 또한, 딥러닝 모형의 복잡한 구조는 해석을 어렵게 만든다[7]. 딥러닝을 활용한 연구의 예로, Gong 등[8]은 딥러닝을 활용해 내시경(endoscopy) 백색광 영상에서 위 신생물(gastric neoplasms)을 실시간으로 자동 감지하고 분류하는 시스템을 개발하였다.

로지스틱 회귀(logistic regression)

로지스틱 회귀는 이분형 분류 문제(예: 질병 유무)를 해결하기 위한 통계 모형이다. 로지스틱 회귀는 종속변수의 로짓(logit), 즉 관심 있는 범주에 해당할(예: 질병 발생) 오즈비(odds ratio)에 자연로그를 취한 값이 독립변수들의 선형 결합과 선형 관계를 맺는다고 가정한다. 이러한 관계를 기반으로 독립변수의 선형 결합을 로지스틱 함수(logistic function)를 통해 0과 1 사이의 확률값으로 해석할 수 있다. 로지스틱 회귀는 독립변수들의 영향력을 해석하는 데 유용하지만, 변수 간의 강한 상관관계, 즉 다중공선성이 존재하는 경우나 관심 있는 범주에 해당하는 대상자 수에 비해 예측인자가 많은 경우, 회귀 계수의 표준 오차의 과대 추정을 초래하여 편향된 결과를 낳게 한다[9]. 로지스틱 회귀를 활용한 연구의 예로, Geng 등[10]은 식도 내시경 점막하 박리술(esophageal endoscopic submucosal dissection)을 받은 환자를 대상으로 로지스틱 회귀를 활용하여 인구통계학적, 임상학적 변수와 협착증(stenosis) 발생 간의 관계를 파악하였다.

콕스 비례 위험 회귀(Cox proportional-hazards regression)

콕스 비례 위험 회귀는 생존분석에서 주로 사용되는 모형으로, 환자의 예후(생존이나 재발)와 같은 관심 사건이 발생했는지 뿐만 아니라 관심 사건이 발생하기까지의 시간을 함께 고려할 수 있어 예후 예측에 널리 쓰인다[11]. 추정된 회귀 계수는 공변량과 위험함수의 관계를 나타내며, 이는 모형의 해석에 있어 중요한 역할을 한다[12]. 콕스 비례 위험 회귀의 핵심 가정은 '비례 위험 가정'으로, 주어진 공변량에 따른 위험률 비(hazard ratio) 가 시간에 따라 일정하다는 것이다. 이 가정이 타당하다면 콕스 비례 위험 회귀를 통해 개개인의 위험률을 예측할 수 있으며, 특정 변수가 위험률에 미치는 영향을 평가할 수 있다. 그러나 현실에서는 비례 위험 가정이 종종 성립하지 않는 경우가 많다. 이런 경우에 변수를 여러 계층으로 나누는 계층화된(stratified) 콕스 비례 위험 회귀 모형을 사용하거나, 시간 가변 공변량(time-varying covariate)을 고려하는 것도 유용하다. 콕스 비례 위험 회귀를 활용한 연구의 예로, Bae 등[4]은 후향적 코호트 디자인에서 조기 위암으로 근치적 절제술(curative resection)을 받은 환자를 대상으로 위외 재발(extragastric recurrence) 무진행 생존 예측을 위해 콕스 비례 위험 모형을 활용해 예측인자를 찾아내는 연구를 하였다.

랜덤 포레스트(random forests)

랜덤 포레스트는 의사 결정 나무를 활용한 앙상블 학습 방법으로, 다수의 나무 모형을 독립적으로 훈련해 그들의 예측을 통합함으로써 성능을 향상한다. 앙상블 학습은 여러 모형의 결과를 결합해 단일 모형보다 우수한 성능을 달성하려는 접근 방식이다. 랜덤 포레스트에서, 각 나무는 데이터의 부트스트랩 샘플을 이용해 훈련되며, 각 노드에서는 무작위로 선택된 변수의 부분 집합만을 고려하여 샘플을 분할한다. 이렇게 하여, 각 나무는 조금씩 다른 변수를 가지게 되어 모형의 다양성이 증가하고, 이에 따라 과적합이 방지된다. 랜덤 포레스트는 여러 나무의 예측을 종합함으로써 과적합의 문제를 완화하고, 일반화 능력을 키운다. 더욱이, 병렬 계산의 이점으로 인해 랜덤 포레스트는 큰 데이터 세트에서도 효율적으로 훈련될 수 있다[13,14]. 랜덤 포레스트의 가장 큰 단점은 해석의 어려움이다. 단일 의사 결정 나무의 단순함과 달리, 랜덤 포레스트는 여러 나무 모형을 결합함으로써 복잡해진다[15]. 랜덤 포레스트를 활용한 연구의 예로, Liwinski 등[16]은 분변 마이크로바이옴(fecal microbiome)을 분석하여 자가면역성 간염(autoimmune hepatitis)과 원발성 담즙성 담관염(primary biliary cholangitis)을 랜덤 포레스트를 활용해 효과적으로 분류하는 연구를 수행하였다.

그래디언트 부스팅(gradient boosting)

그래디언트 부스팅은 주로 결정 나무와 같은 약한 학습기를 사용하여, 모형의 전반적인 성능을 점진적으로 향상시키는 앙상블 방법이다. 이 방법은 각 학습 단계에서 발생한 예측 오차를 줄이기 위해 새로운 모형을 순차적으로 추가함으로써, 전체 모형의 오차를 조절하고 예측 성능을 점차 개선한다[17]. 그러나, 높은 메모리 사용, 평가 속도 지연, 순차적 학습으로 인한 병렬 처리의 어려움, 그리고 복잡한 모형 구조로 인해 해석이 어렵다는 단점이 있다. 이를 일부 개선하기 위해 eXtreme Gradient Boosting (XGBoost) [18], lightGBM [19] 등이 제안되었다. 그중 가장 대표적으로 많이 쓰이는 XGBoost는 병렬 처리, 정규화를 통한 과적합 방지 그리고 결측치 처리 기능을 제공하여 빠른 속도와 높은 예측 성능 및 편의성을 제공한다. 하지만 여전히 모형 해석의 어려움은 남아있다[20]. XGBoost를 활용한 연구 중 하나로, Kwon 등[21]은 위암 수술 후 제2형 당뇨병 예측을 위한 점수를 만들고 비교하는 연구를 진행했다.

라쏘(Least Absolute Shrinkage and Selection Operator)

라쏘(LASSO, Least Absolute Shrinkage and Selection Operator)는 회귀분석에서 변수 선택과 정규화(regularization)를 동시에 수행하는 방법으로, 이는 특히 모형에 넣고자 하는 예측인자 변수의 개수가 샘플의 수보다 많은 고차원 자료에서 유용하다. 라쏘는 계수의 절댓값에 비례하여 페널티를 부여함으로써 모형의 복잡도를 조절하고 일반화 성능을 개선한다. 람다 파라미터로 조절되는 페널티의 강도에 따라, 람다가 클수록 모형이 단순화되며 일부 회귀 계수가 정확히 0이 되어 변수 선택이 가능하게 된다. 이 특징은 모형이 더 해석하기 쉬워지게 하며, 불필요한 변수를 제거하여 예측 성능을 향상시킨다[22,23]. 그러나 라쏘는 변수 수가 샘플 수를 초과할 때 제한된 수의 변수만을 선택할 수 있는 한계를 가지고 있으며, 높은 상관관계를 가진 변수 그룹 내에서는 선택이 임의적일 수 있어 문제가 된다[24]. 라쏘를 활용한 연구의 예로, Ali 등[25]은 내시경소매위성형술(endoscopic sleeve gastroplasty) 후 재중재술(reintervention)의 가능성에 영향을 미치는 변수 선정에 라쏘를 사용하였고, 후보 변수 26개 중 11개를 선정하여 예측모형 구축에 활용하였다.

서포트 벡터 머신(support vector machine)

서포트 벡터 머신(support vector machine)은 데이터를 잘 구분하는 결정 경계(또는 초평면)를 찾는 모형이다. 서포트 벡터 머신의 목적은 서로 다른 범주의 데이터 사이에 위치한 결정 경계의 마진을 최대화하는 것이다. 여기서 마진은 결정 경계와 이 경계에 가장 가까운 샘플, 즉 서포트 벡터 사이의 거리를 의미한다. 서포트 벡터 머신의 일반화 성능은 마진을 최대화함으로써 향상될 수 있다. 특히, 데이터가 선형적으로 구분되지 않을 때 커널 트릭을 사용하여 데이터를 더 높은 차원으로 매핑함으로써 데이터를 선형적으로 구분할 수 있게 만든다. 커널 트릭은 원래의 변수 공간에서 선형적으로 구분이 어려운 경우에도, 높은 차원에서는 구분이 가능하게끔 데이터를 변환한다[26,27]. 서포트 벡터 머신은 고차원의 이질적 데이터와 종속변수의 관심 범주 발생 빈도가 적은 데이터 세트를 효율적으로 처리하는 장점을 가지고 있다. 하지만 복잡한 파라미터 튜닝과 적절한 커널 선택 및 결합이 어렵다는 단점이 있다[28]. 서포트 벡터 머신을 활용한 연구의 예로, Yu 등[29]이 수행한 세포 외 소포 긴 RNA (extracellular vesicle long RNA) 프로파일링을 기반으로 췌관 선암(pancreatic ductal adenocarcinoma) 검출을 위한 진단 방법 개발 연구가 있다.

검증 방식

내부 검증(internal validation)

예측모형에서의 내부 검증은 모형의 학습 과정에 사용한 데이터 세트를 통하여 모형의 성능을 평가하는 것이다. 내부 검증은 사용하기 간단하고 드는 비용이 적다는 장점이 있다. 하지만 모형이 학습한 데이터를 통해 모형의 성능을 평가하는 것이기 때문에 새로운 데이터에 대한 일반화 능력을 평가하기는 어렵다. 대표적인 방법으로는 분할 샘플(split sample), 교차 검증(cross validation), 부트스트랩핑(bootstrapping) 그리고 반복적인 무작위 하위 샘플링 검증(repeated random sub-sampling) 방법이 있다[30].

분할 샘플(split sample)

분할 샘플 검증은 예측모형에서 내부적으로 모형의 성능을 평가하는 방법으로, 데이터 세트를 구축용 세트와 검증 세트로 구분한다. 이 방법에서 모형은 구축용 세트

https://doi.org/10.52927/jdcr.2023.11.3.157

를 통해 학습하며, 학습에 사용되지 않은 나머지 부분은 모형의 성능을 측정하기 위한 검증 세트로 사용된다. 분할 샘플 검증의 주요 목적은 구축용 세트로 학습된 모형이 구축 과정에서 보지 않은 데이터에 대한 예측이 어느 정도 가능한지 확인하는 것이다. 이 과정은 모형이 구축용 데이터에 과적합되어 일반적인 패턴이 아닌 특정 데이터의 노이즈를 학습했는지 평가하는 데 중요한 역할을 한다. 학습과 검증에 할당되는 데이터의 비는 다양할 수 있지만, 일반적으로 구축용 세트는 전체의 70–80%, 검증 세트는 20–30%로 설정된다. 어떤 경우에는 구축용 세트를 다시 훈련용과 구축 중 검증용으로 다시 분리하여 전체 데이터를 세 부분으로 나누기도 하는데, 이때 구축 중 검증용 자료는 모형의 파라미터를 최종 평가하기 이전에 미세조정하는 데 사용되는 일종의 중간 평가 세트로 활용될 수 있으며, 딥러닝 모형의 훈련 과정에서는 조기 종료(early stopping) 규칙을 적용해 과적합을 방지할 수 있다[31-33].

교차 검증(cross validation)

교차 검증은 모형의 일반화 능력을 정밀하게 평가하기 위한 유용한 내부 검증 방법이다. 이 방식은 전체 데이터 세트를 여러 개의 부분 집합으로 분리하고, 각 부분 집합을 테스트 세트로 활용하며, 나머지는 모형 학습에 이용한다. 이 과정을 모든 부분 집합에 대해 반복한다. k-겹 교차 검증이 교차 검증의 대표적인 형태로, 데이터를 같은 크기의 k개의 폴드로 나눈다. 각 폴드는 차례대로 테스트 세트로 취급되며, 나머지 k–1개의 폴드는 훈련 세트로 정의된다. 이러한 절차를 k번 반복한 후, 도출된 성능 지표의 평균을 모형의 최종 성능으로 간주한다. 교차 검증의 주된 이점은 모든 샘플이 훈련과 테스트에서 적어도 한 번씩 사용된다는 점이며, 이에 따라 성능 평가가 특정 데이터 세트에 의존하지 않고, 일반적으로 더 안정적이며, 덜 편향된다. 이 방법은 데이터의 임의 분할로 인한 변동성을 완화하므로, 모형의 일반화 능력에 대한 더 신뢰성 있는 추정을 가능하게 한다. 그러나, 모형을 여러 번 학습시켜야 하므로, 계산 비용이 상대적으로 높을 수 있다. 교차 검증은 모형 선택, 하이퍼파라미터 조정 및 일반화 성능 평가 등 다양한 모형화 작업에서 핵심적인 역할을 수행한다. 때로는 데이터의 특성이나 분포에 따라 변형된 교차 검증 방법, 예를 들어 계층화 교차 검증 등이 적용되기도 한다[34].

부트스트랩핑(bootstrapping)

부트스트랩핑은 예측모형의 성능을 평가하는 데 있어 널리 사용되는 재표본추출 방법이다. 이 방법은 동일한 샘플을 여러 번 선택할 수 있게 함으로써 다양한 샘플 세트 조합으로 생성한 모형을 활용한다. 이 과정을 통해, 모형 성능의 변동성을 더욱 정밀하게 파악할 수 있게 되며, 모형의 예측 불확실성에 대한 더 깊은 이해를 얻을 수 있다. 부트스트랩핑을 검증 단계에서 활용하면, 모형이 미래 데이터에 얼마나 잘 일반화될 수 있는지에 대한 거의 편향되지 않은 평가를 할 수 있어, 과적합의 위험을 줄이고 모형의 신뢰도를 높일 수 있다[35,36].

반복적인 무작위 하위 샘플링 검증(repeated random sub-sampling)

반복적 무작위 검증 방법은 데이터 세트를 여러 번, 무작위로 훈련 데이터와 검증 데이터로 분할하는 방식이다. 각 분할에서 모형은 훈련 데이터에 기반하여 학습하며, 검증 데이터를 사용해 예측 정확도를 평가한다. 모든 분할의 결과를 종합하여 평균을 내, 최종적인 결과를 도출한다. 이 방법의 특징은 훈련 및 검증 데이터의 분할 비율이 분할 수 즉, 반복 횟수에 영향을 받지 않는다는 점이며, 이는 K-겹 교차 검증과는 대조된다. 추가적으로, 이 방식에서는 일부 샘플이 검증 세트에 한 번도 포함되지 않을 수 있거나, 반대로 여러 번 평가에 사용될 수 있다. 또한, 무작위 샘플링의 변동성 때문에, 분석을 다른 무작위 분할에 대해 반복할 때마다 결과가 달라질 수 있다[37].

외부 검증

예측모형에서 외부 검증은 모형의 학습이 완료된 후 독립적인 데이터 세트를 사용하여 모형의 성능을 평가하는 방법이다. 여기서, 독립적인 데이터란 시간이나 지리적 위치로 분리될 수 있는 어떤 특정 데이터 세트를 의미한다. 이러한 외부 검증은 모형의 성능에 대한 일반화와 새로운 데이터에의 적용 가능성에 대한 보다 믿을 수 있는 평가를 제공한다[30,36].

예측모형 연구에서 고려할 점

앞서 설명한 바와 같이 예측모형의 구축과 검증에 다양한 방법들이 적용되고 있어 연구 계획 단계에서의 적절한 선택이 중요하다.

예측모형의 구축에 사용할 모형 유형은 예측하고자 하

는 종속변수의 형태(분류/생존 등)에 따라 1차적으로 선택할 수 있고, 이후 예측인자들과 종속변수의 관계, 대상자 수와 변수의 개수의 비에 따라 모형을 탐색하여 선정하게 된다. 최근 의료인공지능 연구가 활발하게 이루어지면서 머신러닝이나 딥러닝을 활용한 예측모형 연구가 증가하였으나 여전히 전통적인 통계학적 방법인 회귀분석에 기반한 로지스틱 회귀분석과 콕스 비례 위험 회귀분석도 많이 쓰이는 것으로 나타났다.

적절한 성능을 가지는 모형 구축도 중요하지만 실제 임상에서 유용하게 사용되려면 적절한 모형 적용 형태(presentation form)도 논문에서 정보를 얻을 수 있어야 한다. 예를 들어 딥러닝이나 랜덤 포레스트 등의 머신러닝을 사용한 예측모형은 실제 해당 프로그램(소프트웨어)을 웹사이트 등을 통해 공개하여 접근 가능해야 임상 현장에서 쓸 수 있는 모형이 된다.

모형 구축에 필요한 대상자 수와 변수의 개수는 한 모의실험 연구를 통해 events per variable이 10 이상이면 모형이 안정적으로 추정된다는 점이 알려져 있어 이를 참조하는 편이다[5]. 예를 들어 관심 있는 종속변수의 특정 군에 해당하는 대상자가 50명이라면 50/10 = 5개의 변수를 예측인자로 넣는 것이 적절하다고 판단하는 것이다. 최근에는 모형에 들어갈 변수들의 분포를 설정하여 목표하는 성능에 도달하는 데에 필요한 대상자 수를 실험적으로 구하는 방법도 제안되었다[38].

대상자 수 선정 후 자료 수집 단계에 들어서면 검증 방식을 검토해야 한다. 본 연구를 통해 조사한 결과 다기관 외부 검증 방식이 많았고 그 다음으로 내부 검증 방식 중 자료 분할 형태가 많았는데, 아무래도 현실적으로 다기관 자료 수집이 어려운 연구 분야의 경우 단일 기관 자료를 활용하여 예측모형 연구를 수행할 수밖에 없고, 이 때 단순한 자료 분할보다는 교차 검증이나 부트스트래핑과 같은 일반화 정도를 높일 수 있는 방식에 대한 검토도 필요하다고 할 수 있겠다. 외부 검증이 필요한 경우 타기관 자료나 공개 자료 등 외부 자료를 수집할 필요가 있는데, 외부 검증 방식의 경우 최근에는 기관별로 분류한 자료를 이용한 구축과 검증보다는 각 기관을 폴드로 고려해서 모든 기관의 자료를 구축과 검증에 한 번씩 사용하도록 하는 leave-one-hospital-out 방법을 사용하거나, 더욱 다양한 구성의 대상자 집단 자료를 구축하여 정교한 검증을 해야 한다는 주장들도 제기되고 있다[39].

CONCLUSION

본 논문에서는 소화기암연구 분야에서의 예측모형 연구에 대해 구축 방법과 검증 방식 중심으로 리뷰하고 각 방법론에 대한 소개를 하였다. 예측모형 연구를 준비하는 독자들에게 예측모형에 대한 보고 가이드라인인 TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement [5] 및 설명 문서[40]를 참고할 것을 권장하며, 출간 준비 중으로 알려져 있는 머신러닝 등 인공지능을 활용한 예측모형에 대한 가이드라인인 TRIPOD-AI statement [41]도 추후 유용하게 활용할 것으로 기대한다.

FUNDING

이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2021R1I1A1A01059893).

연구비 제공자는 연구 설계, 데이터 수집 및 분석, 출판 준비 및 결정에 아무런 역할을 하지 않았다.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

AUTHOR'S CONTRIBUTIONS

Conceptualization: Kyunghwa Han. Data acquisition: Yonghan Kwon. Formal analysis: Yonghan Kwon. Funding: Kyunghwa Han. Supervision: Kyunghwa Han. Writing—original draft: Yonghan Kwon, Kyunghwa Han. Writing—review & editing: Yonghan Kwon, Kyunghwa Han.

ORCID

Yonghan Kwon, https://orcid.org/0000-0001-7951-1142
Kyunghwa Han, https://orcid.org/0000-0002-5687-7237

SUPPLEMENTARY MATERIALS

Supplementary data is available at https://doi.org/10.52927/jdcr.2023.11.3.157.

REFERENCES

1. Yang C, Kors JA, Ioannou S, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. J Am Med Inform Assoc 2022;29:983-989. https://doi.org/10.1093/jamia/ocac002
2. van den Boorn HG, Engelhardt EG, van Kleef J, et al. Prediction models for patients with esophageal or gastric cancer: a systematic review and meta-analysis. PLoS One 2018;13:e0192310. https://doi.org/10.1371/journal.pone.0192310
3. Backes Y, Schwartz MP, Ter Borg F, et al. Multicentre prospective evaluation of real-time optical diagnosis of T1 colorectal cancer in large non-pedunculated colorectal polyps using narrow band imaging (the OPTICAL study). Gut 2019;68:271-279. https://doi.org/10.1136/gutjnl-2017-314723
4. Bae JS, Chang W, Kim SH, et al. Development of a predictive model for extragastric recurrence after curative resection for early gastric cancer. Gastric Cancer 2022;25:255-264. https://doi.org/10.1007/s10120-021-01217-1
5. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement.. Ann Intern Med 2015;162:55-63. https://doi.org/10.7326/m14-0697
6. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-444. https://doi.org/10.1038/nature14539
7. Shrestha A, Mahmood A. Review of deep learning algorithms and architectures. IEEE Access 2019;7:53040-53065. https://doi.org/10.1109/ACCESS.2019.2912200
8. Gong EJ, Bang CS, Lee JJ, et al. Deep learning-based clinical decision support system for gastric neoplasms in real-time endoscopy: development and validation study. Endoscopy 2023;55:701-708. https://doi.org/10.1055/a-2031-0691
9. Agresti A. Categorical data analysis. Hoboken: John Wiley & Sons, 2012.
10. Geng ZH, Zhu Y, Li QL, et al. Muscular injury as an independent risk factor for esophageal stenosis after endoscopic submucosal dissection of esophageal squamous cell cancer. Gastrointest Endosc 2023;98:534-542.e7. https://doi.org/10.1016/j.gie.2023.05.046
11. Cox DR. Regression models and life-tables. J Royal Stat Soc Ser B 1972;34:187-220.
12. George B, Seals S, Aban I. Survival analysis and regression models. J Nucl Cardiol 2014;21:686-694. https://doi.org/10.1007/s12350-014-9908-2
13. Breiman L. Random forests. Mach Learn 2001;45:5-32. https://doi.org/10.1023/A:1010933404324
14. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer, 2009. p.587-604.
15. Ziegler A, König IR. Mining data with random forests: current options for real-world applications. WIREs 2014;4:55-63. https://doi.org/10.1002/widm.1114
16. Liwinski T, Casar C, Ruehlemann MC, et al. A disease-specific decline of the relative abundance of Bifidobacterium in patients with autoimmune hepatitis. Aliment Pharmacol Ther 2020;51:1417-1428. https://doi.org/10.1111/apt.15754
17. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist 2001;95:1189-1232. https://doi.org/10.1214/aos/1013203451
18. Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting. R package version 0.4-2. 2015. https://xgboost.readthedocs.io/en/stable/R-package/xgboostPresentation.html (accessed Oct 1, 2023).
19. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree [abstract]. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4-9; Long Beach, USA. p.3149-3157.
20. Sagi O, Rokach L. Approximating XGBoost with an interpretable decision tree. Inf Sci 2021;572:522-542. https://doi.org/10.1016/j.ins.2021.05.055

21. Kwon Y, Kwon JW, Ha J, et al. Remission of type 2 diabetes after gastrectomy for gastric cancer: diabetes prediction score. Gastric Cancer 2022;25:265-274. https://doi.org/10.1007/s10120-021-01216-2
22. Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Stat Soc Ser B 1996;58:267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
23. Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: the lasso and generalizations. Boca Raton: CRC Press, 2015.
24. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Stat Soc Ser B 2005;67:301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
25. Ali H, Patel P, Malik TF, et al. Endoscopic sleeve gastroplasty reintervention score using supervised machine learning. Gastrointest Endosc 2023;98:747-754.e5. https://doi.org/10.1016/j.gie.2023.05.059
26. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Appl 1998;13:18-28. https://doi.org/10.1109/5254.708428
27. Chen PH, Lin CJ, Schölkopf B. A tutorial on ν-support vector machines. Appl Stoch Models Bus Ind 2005;21:111-136. https://doi.org/10.1002/asmb.537
28. Salcedo-Sanz S, Rojo-Álvarez JL, Martínez-Ramón M, Camps-Valls G. Support vector machines in engineering: an overview. WIREs 2014;4:234-267. https://doi.org/10.1002/widm.1125
29. Yu S, Li Y, Liao Z, et al. Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma. Gut 2020;69:540-550. https://doi.org/10.1136/gutjnl-2019-318860
30. Bleeker SE, Moll HA, Steyerberg EW, et al. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol 2003;56:826-832. https://doi.org/10.1016/s0895-4356(03)00207-5
31. Xu Y, Goodacre R. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. J Anal Test 2018;2:249-262. https://doi.org/10.1007/s41664-018-0068-2
32. Gholamy A, Kreinovich V, Kosheleva O. Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation. 2018 Feb. Report No.: UTEP-CS-18-09.
33. Prechelt L. Early stopping-but when? In: Orr GB, Müller KR, eds. Neural networks: tricks of the trade. Berlin, Heidelberg: Springer, 2002:55-69.
34. Berrar D. Cross-validation. Encycl Bioinform Comput Biol 2019;1:542-545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X
35. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman and Hall, 1994.
36. Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Cham: Springer International Publishing, 2015.
37. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer, 2013.
38. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020;368:m441. https://doi.org/10.1136/bmj.m441
39. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Med 2023;21:70. https://doi.org/10.1186/s12916-023-02779-w
40. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1-W73. https://doi.org/10.7326/M14-0698
41. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;393:1577-1579. https://doi.org/10.1016/S0140-6736(19)30037-6