Xu et al. Cancer Cell International (2024) 24:427
https://doi.org/10.1186/s12935-024-03601-0

Cancer Cell International

COMMENT Open Access

Multiple machine learning algorithms, validation of external clinical cohort and assessments of model gain effects will better serve cancer research on bioinformatic models

Fangshi Xu[1], Zongyu Li[1], Hao Guan[1] and Jiancang Ma[1*]

Abstract
Bioinformatics models greatly contribute to individualized assessments of cancer patients. However, considerable research neglected some critical technological points, including comparisons of multiple modeling algorithms, evaluating gain effects of constructed model, comprehensive bioinformatics analyses and validation of clinical cohort. These issues are worthy of emphasizing, which will better serve future cancer research.

Keywords Machine learning algorithm, Gain effect, Bioinformatics, Prognostic model, Clinical validation

Introduction

With the continuous emergence of high-throughput sequencing data, more scholars have developed predicting models for assessing the clinical status and therapeutic outcome of tumor patients using multiple bioinformatic approaches. Reasonable application of genomic data will drive high-quality survival outcome analytics, finally contributing personalized treatment [1]. Notably, several key points merit emphasizing to better serve the clinical application of data mining, including comparisons of multiple modeling algorithms, evaluating gain effects of constructed model, clinical multi-omics analysis and validation of clinical cohort.

*Correspondence:
Jiancang Ma
majiancangxjtu@163.com
[1]Department of Vascular Surgery, The Second Affiliated Hospital of Xi'an Jiaotong University, No. 157, West Five Road, Xi'an 710004, Shaanxi Province, China

The dilemma of model overfitting

Model overfitting is a commonly encountered conundrum in clinical modeling. Although utilizing more parameters for modeling is usually better to simulate the actual clinical situation, more parameters will also bring about cumbersome models and interference within the parameters, which are not conducive to clinical practice. Regularization is a pivotal tool for mitigating overfitting due to its excellent abilities to well generalize to invisible data [2]. The basic principle of regularization is to limit the model learning process by adding another parameter to the loss function we are trying to minimize. Typically, the regularization is performed through three types of regression, namely Lasso, Ridge, elastic net, with their respective advantages and limitations.

Algorithmically, Ridge regression is the addition of sum-of-squares regularization to the loss function. Lasso regression is the regularization method that adds absolute sums to the loss function. Elastic net regression is a

fusion of the above two regularization methods. Commonly, for the same value of lambda, the root mean square error (RMSE) of elastic net is between that of Ridge and Lasso algorithms, which is better than Lasso regression and worse than ridge regression.

Although it may not be the optimal option in term of the RMSE value, the model fitted by lasso regression offers significant interpretability, which is completely different from some 'black box' algorithms, such as random forest and support vector machine (SVM) [3]. Moreover, when no multicollinearity exists between independent variables, Ridge regression may diminish the predictive performance of the constructed model [4]. As for elastic net, it encourages group effects when variables are highly correlated, rather than zeroing some of them as in Lasso regression [5]. Thus, elastic net is more advantageous when multiple features are correlated with one another. At this point, Lasso tends to randomly select one feature, whereas elastic net prefers to select both. Given that different clinical features used for clinical modeling are often independent of each other, Lasso regression may be a safer modeling strategy.

Comparison of multiple modeling algorithms

Lasso regression is mainly employed to solve the encountered dilemma of traditional linear regression when dealing with high-dimensional data. In a high-dimensional space, traditional least squares regression (LSR) commonly struggles with difficult variable selection and model overfitting. However, lasso regression can force the coefficient values of some inconsequential variables to decay to zero through constructing a penalty function termed λ [6]. This effectively controls the complexity of model, which is very suitable for the modeling environment of clinical indicators. Thus, lasso regression is the most commonly used modeling algorithm in clinical research. Recently, Tang J et al. also utilized this method to construct a disulfidptosis-related (DR) model with outstanding prognostic analytical performance [7].

It should be noted that lasso regression is not the sole access to modeling. Due to the continuous updating of algorithms, such as elastic net, random survival forest and ridge regression dimensionality reduction algorithms, lasso is not necessarily optimal solution. For instance, Liu Z et al. compared the C-index values of multiple prognostic models established via 101 different machine learning algorithms in colorectal cancer [8]. Among them, the combination of lasso regression and step cox regression possessed the highest C-index of 0.696, followed by support vector machine-recursive feature elimination (SVM-RFE) of 0.659 [8]. Clearly, Liu Z et al. exhibited a more precise strategy of clinical modeling, that comparison of various algorithms.

Another team also employed a similar modeling strategy [9]. The authors compared the differences in C-index among 76 modeling methods. Due to the highest C-index of the combination of Lasso and stepwise Cox (0.777), this model was considered the optimal one [9]. Similarly, Zhang L et al.. constructed a programmed cell death (PCD)-related model to enhance the accuracy for predicting prognosis and immunotherapy efficacy through the comparison of 101 combinations of 10 machine learning algorithms in lung adenocarcinoma [10]. Clearly, selecting the optimal modeling approach by comparing the C index of multiple modeling algorithms is a reasonable and effective strategy.

Assessments of gain effects

To date, AJCC-stage and TNM-system are the most commonly used prognostic assessments for HCC patients. Although novel evaluation models have sprung up, they substituting for AJCC or TNM seems to be not realistic. One feasible solution is to determine whether novel gene signatures or risk scores can enhance the predictive performance of TNM or AJCC. In the case of m7G risk score, Ren B's team used decision curve analysis (DCA) to assess the net benefit of decision-making when introducing this novel score into AJCC-stage system [11]. Their results revealed that novel m7G risk score greatly elevated the predictive accuracy (AUC=0.787 vs. 0.891) and clinical decision-making benefit of traditional prognostic model. Hence, m7G risk score can be regarded as a valuable and pivotal complement to AJCC-based prognostic assessment in adrenocortical carcinoma (ACC).

Moreover, some in silico algorithm tools will broaden research horizon. For instance, molecular docking technology can contribute to design, synthesis and bio-evaluation of potential chemical targets, thereby driving its pharmaceutical research and development [12]. Take the DR model as an example, it consisted of 19 critical disulfidptosis regulators, providing multiple novel therapeutic targets against HCC [7]. To go a step further, using structure-based virtual screening (SBVS), we can screen the binding structures of these molecules from large-scale chemical compound libraries and identify potential hits by evaluating their binding affinity [13]. Further, scaffold hopping technology can be employed to optimize the physicochemical and pharmacokinetic (PK) properties of the above ligands, thereby obtaining novel chemical agents with bioactivities.

Five key steps of clinical cohort validation

Although the prognostic models constructed using bioinformatics approaches widen the boundaries of cancer evaluation system, this prediction means is extremely required the validation from external real cohorts. One primary reason for this is that modeling based on

bioinformatic algorithm is more like simplicial mathematical operation rather than tools with clinical significance. Take DR signature as an instance, despite tightly associations of this model with prognosis and immune microenvironment of HCC, it remains elusive that the relationships between different levels of risk scores and the activity of disulfidptosis, a novel pattern of programmed cell death (PCD) [7]. Are HCC patients with high DR scores accompanied by more active disulfidptosis process? Thus, it is essential to determine biological implications of constructed models. Notably, clinical validation is the indispensable access for novel bioinformatics models to move from theory to clinical practice. Regretfully, several studies published in 'Cancer cell international' all failed to address this point, such as Fatty acids synthesis and metabolism (FASM) gene signature [14] and oxidative stress-related model [15].

According to the related guideline from the British Medical Journal (BMJ), clinical verification process involves five critical steps [16]. First, obtaining a suitable clinical dataset. The data collected from prospective study is of high quality and suitable for external validation, but is more time-consuming and expensive. The data from retrospective studies is easily accessible, but additional attention to data quality is required. Especially, researchers should determine whether the content of external cohort meet the core purpose of study. For instance, the inclusion and exclusion criteria for study subjects should match the target population and the environment of model operating. Second, prediction based on models. In this step, the model is applied to the external cohort to calculate the predicted values through programming. Third, quantifying the predictive performance of models. This step includes the assessments of the overall fit, calibration, and discrimination ability in the external cohort. For instance, the consistency of observed event probability with model-estimated event probability is assessed through a calibration plot. Fourth, quantifying clinical utility. If the predictive model is to be used to guide medical decision-making, the overall benefit of the model to the participant and healthcare outcomes, i.e., the clinical utility, should also be assessed. This process is commonly accomplished through the decision clinical analysis (DCA). Fifth, clear and transparent report. The Transparent Reporting of Individual Prognostic or Diagnostic Multivariate Models Statement (TRIPOD) from BMJ can help us with this step [16].

It is worth noting that external validation also faces some difficult issues, such as batch effects. Differences in non-biological factors such as experimental design, sample handling, data collection and processing may lead to an alteration of gene expression between different cohorts. Such differences, termed as batch effects, may mask or obscure biologically true variation. Although data standardization (Z-score normalization etc.), batch correction algorithms (Minimum covariance determinant etc.), and multivariate analysis (PCA etc.) can be used to reduce the adverse impact of batch effects, effective approaches of reduction is reliant not only on the reasonable algorithms, but also on the rigorous design of clinical research [17]. Undoubtedly, this is a challenging task.

Conclusions

With the rapid development of bioinformatics technology, novel predictive models are emerging. However, most of them have great potential for algorithmic improvements, and the lack of external cohort validation also limits their application in clinical practice. Herein, we proposed that the comparison of multiple machine learning algorithms, assessments of model gain effects, and validation of external clinical cohort would contribute to addressing this dilemma.

Acknowledgements
All authors would like to thank Second Affiliated Hospital of Xi'an Jiaotong University for its support.

Author contributions
JCM conceived and designed the study. FSX, ZYL and HG wrote the manuscript. All authors have read and approved the manuscript.

Funding
This study was supported by Natural Science Foundation of Shaanxi Province (2024JC-YBQN-0905).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and informed consent
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 24 June 2024 / Accepted: 5 December 2024
Published online: 23 December 2024

References

1. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell. 2018;173(2):400–e1611.
2. Zhang WF, Dai DQ, Yan H. Framelet kernels with applications to support vector regression and regularization networks. IEEE transactions on systems, man, and cybernetics Part B. Cybernetics: Publication IEEE Syst Man Cybernetics Soc. 2010;40(4):1128–44.
3. Zhao Y, Long Q. Multiple imputation in the presence of high-dimensional data. Stat Methods Med Res. 2016;25(5):2021–35.
4. Yang R, He F, He M, Yang J, Huang X. Decentralized Kernel Ridge Regression Based on Data-Dependent Random Feature. IEEE transactions on neural networks and learning systems. 2024;Pp.

5. Paolino JP. Rasch Model Parameter Estimation via the Elastic Net. J Appl Meas. 2015;16(4):353–64.
6. Ternès N, Rotolo F, Michiels S. Empirical extensions of the lasso penalty to reduce the false discovery rate in high-dimensional Cox regression models. Stat Med. 2016;35(15):2561–73.
7. Tang J, Peng X, Xiao D, Liu S, Tao Y, Shu L. Disulfidptosis-related signature predicts prognosis and characterizes the immune microenvironment in hepatocellular carcinoma. Cancer Cell Int. 2024;24(1):19.
8. Liu Z, Liu L, Weng S, Guo C, Dang Q, Xu H, et al. Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer. Nat Commun. 2022;13(1):816.
9. Liu Z, Guo C, Dang Q, Wang L, Liu L, Weng S, et al. Integrative analysis from multi-center studies identities a consensus machine learning-derived lncRNA signature for stage II/III colorectal cancer. EBioMedicine. 2022;75:103750.
10. Zhang L, Cui Y, Zhou G, Zhang Z, Zhang P. Leveraging mitochondrial-programmed cell death dynamics to enhance prognostic accuracy and immunotherapy efficacy in lung adenocarcinoma. J Immunother Cancer. 2024;12(10).
11. Xu F, Cai D, Liu S, He K, Chen J, Qu L, et al. N7-methylguanosine regulatory genes well represented by METTL1 define vastly different prognostic, immune and therapy landscapes in adrenocortical carcinoma. Am J cancer Res. 2023;13(2):538–68.
12. Curcio A, Rocca R, Alcaro S, Artese A. The histone deacetylase family: structural features and application of combined computational methods. Pharmaceuticals (Basel Switzerland). 2024;17(5).
13. Carlsson J, Luttens A. Structure-based virtual screening of vast chemical space as a starting point for drug discovery. Curr Opin Struct Biol. 2024;87:102829.
14. Zhengdong A, Xiaoying X, Shuhui F, Rui L, Zehui T, Guanbin S, et al. Identification of fatty acids synthesis and metabolism-related gene signature and prediction of prognostic model in hepatocellular carcinoma. Cancer Cell Int. 2024;24(1):130.
15. Wang K, Xiao Y, Zheng R, Cheng Y. Immune cell infiltration and drug response in glioblastoma multiforme: insights from oxidative stress-related genes. Cancer Cell Int. 2024;24(1):123.
16. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. BMJ (Clinical Res ed). 2024;386:e078276.
17. Yu Y, Mai Y, Zheng Y, Shi L. Assessing and mitigating batch effects in large-scale omics studies. Genome Biol. 2024;25(1):254.

Publisher's note