Artificial intelligence model validation before its application in clinical diagnosis assistance

Gustavo Jesus Vazquez-Zapien, Monica Maribel Mata-Miranda, Francisco Garibay-Gonzalez, Miguel Sanchez-Brito

ORCID number: Gustavo Jesus Vazquez-Zapien 0000-0002-6657-0722; Monica Maribel Mata-Miranda 0000-0002-1811-4511; Francisco Garibay-Gonzalez 0000-0002-1071-5208; Miguel Sanchez-Brito 0000-0003-4371-0486.

Author contributions: Vazquez-Zapien GJ and Sanchez-Brito M designed the research; all authors contributed to the writing and revision of the letter.

Conflict-of-interest statement: The authors declare having no competing interests.

Supported by SEDENA Budgetary Program, No. A022-2021.

Country/Territory of origin: Mexico

Specialty type: Gastroenterology and hepatology

Provenance and peer review: Invited article; Externally peer reviewed.

Peer-review model: Single blind

Peer-review report's scientific quality classification
Grade A (Excellent): 0
Grade B (Very good): 0
Grade C (Good): C
Grade D (Fair): 0
Grade E (Poor): E

Gustavo Jesus Vazquez-Zapien, Embryology Lab, Escuela Militar de Medicina, Ciudad de Mexico 11200, CDMX, Mexico

Monica Maribel Mata-Miranda, Cell & Tissue Biology Lab, Escuela Militar de Medicina, Ciudad de Mexico 11200, CDMX, Mexico

Francisco Garibay-Gonzalez, Department of Research, Escuela Militar de Medicina, Ciudad de Mexico 11200, CDMX, Mexico

Miguel Sanchez-Brito, Instituto Tecnológico de Zacatepec, Industrial Engineering, TecNM, Zacatepec 62780, Morelos, Mexico

Miguel Sanchez-Brito, Instituto Tecnológico de Aguascalientes, Computational Sciences, TecNM, Aguascalientes 20256, Mexico

Corresponding author: Miguel Sanchez-Brito, MD, PhD, Research Assistant Professor, Instituto Tecnológico de Zacatepec, Industrial Engineering, TecNM, Plan de Ayala, Zacatepec 62780, Morelos, Mexico. miguel.sb@zacatepec.tecnm.mx

Abstract

The process of selecting an artificial intelligence (AI) model to assist clinical diagnosis of a particular pathology and its validation tests is relevant since the values of accuracy, sensitivity and specificity may not reflect the behavior of the method in a real environment. Here, we provide helpful considerations to increase the success of using an AI model in clinical practice.

Key Words: Artificial intelligence; Diagnostic assistance; Validation tests; Leave-one-out cross-validation; K-fold validation; Hold-out validation

©The Author(s) 2022. Published by Baishideng Publishing Group Inc. All rights reserved.

Core tip: The validation tests and the process to adopt a particular artificial intelligence (AI) model are relevant. The percentages of accuracy, sensitivity and specificity obtained through validation techniques are strong indicators of whether the AI model is suitable for implementation in clinical practice or whether it will be necessary to
After studying the interesting article “Non-occlusive mesenteric ischemia: Diagnostic challenges and perspectives in the era of artificial intelligence” by Bourcier et al.[1], who analyzed the current state of artificial intelligence (AI) in assisting clinical diagnosis and its possible application in diagnosing nonocclusive mesenteric ischemia, we are in full agreement with the AI techniques that the authors mention. However, a greater emphasis on the evaluation process for AI models could yield better results; when a rigorous testing stage is lacking, these models show poor performance upon transfer from the laboratory to real practice.

It is essential to mention that AI models using machine learning techniques, such as decision tree, support vector machine, artificial neural networks, naïve Bayesian classifier, Bayesian network K-nearest neighbor, and random forest, are predictive methods that are indispensable to performance of the three stages of training, validation and testing[3].

In this sense, the scarcity of validation tests provokes a reduction in the percentages of accuracy (Ac), sensitivity (Se) and specificity (Sp) of the AI models at the time of transferring them to a real environment[4].

These validation tests consist of segmenting the total of the samples available in different proportions to force the AI model to look for a robust solution (a representative pattern) due to the variance in the data. However, how to define the proportions in which the database will be segmented is a subject under development. Therefore, cross-validation strategies such as leave-one-out cross-validation (LOOCV) or k-fold cross-validation have been used more frequently than techniques such as hold-out validation because they obtain better Ac, Se and Sp in laboratory tests[4-6]; moreover, they consider a larger population in the training process compared to hold-out.

Following the LOOCV guidelines, a sample is left out of the database and the AI model is trained with the rest; once the training is finished, the separated spectrum is evaluated with the trained model. This process is repeated until all the collected spectra are evaluated, and the percentages of Ac, Se and Sp are calculated based on the number of correct and incorrect evaluations that the models have carried out. By involving most of the data in the training process, the result obtained by LOOCV usually reflects an overtrained model, making the generalization process of future samples complex by reducing their Ac, Se and Sp in a real scenario.

The k-fold model is similar to LOOCV, except that the database is divided into k groups with approximate numbers of samples instead of separating a single sample. Thus, one group is left out, and the rest is used for training; the process ends when all groups have been evaluated. The conflict with this strategy lies in defining the number of k groups created, since there is currently no formal methodology to calculate them. However, the most common values are k = 5 and 10; as such, the base data are segmented into five or 10 groups. In contrast, hold-out divides the populations that make up the database into percentages of 80–20 (one of the most used); that is, 80% of the samples from each population that make up the database are used for training. As this process is subjected to a more significant variance, the evaluation process usually shows lower percentages of Ac, Se and Sp compared to LOOCV and k-fold[4]. The above does not mean that the AI model is inadequate; instead, it indicates that the number of samples collected is insufficient to detect a sufficiently robust pattern. Thus, if the Ac, Se and Sp percentages are not reliable enough, acquiring more samples is a good option before using this AI methodology in clinical practice.

Although no studies have been carried out in this regard, an excellent strategy for evaluating whether an AI model is ready to be tested in a real environment is to...
analyze several techniques, first using LOOCV and selecting the techniques with the best results to study their performance. Subsequently, k-fold evaluates the performance of the previously selected models thanks to the LOOCV strategy. As a result of the study of the models using the k-fold strategy, the model with the best performance should be selected. Finally, the best AI technique can be studied using the hold-out strategy; upon separating a considerable number of samples from each population according to the database (20%), the training/learning process of the AI models is subject to a more significant variance in the data of each population. In this way, they focus on particular features of the same group and not on characteristics of the samples that make up a particular database (overfitting), as could occur in the case of considering the LOOCV strategy only[4,5]. If the accuracy metrics of the model evaluated with hold-out are similar to those obtained when it was evaluated using LOOCV, it is possible to expect that the AI model will perform well in a real environment.

The use of AI methods in clinical diagnosis is new, and there are many subjects to investigate in this field; however, it is fascinating how the use of these technologies has reached medical science and how the new generations of researchers venture to use and combine the different sciences (physics, chemistry, mathematics, engineering, computer science, biology, among others) to generate new knowledge. We hope that the recommendations made here will help explore this AI field in the biological and medical sciences.

REFERENCES
1 Bourcier S, Klug J, Nguyen LS. Non-occlusive mesenteric ischemia: Diagnostic challenges and perspectives in the era of artificial intelligence. World J Gastroenterol 2021; 27: 4088-4103 [PMID: 34326613 DOI: 10.3748/wjg.v27.i26.4088]
2 Sakr S, Elshawi R, Ahmed AM, Qureshi WT, Brawner CA, Keteyian SJ, Blaha MJ, Al-Mallah MH. Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercise testing (FIT) project. BMC Med Inform Decis Mak 2017; 17: 174 [PMID: 29258510 DOI: 10.1186/s12911-017-0586-6]
3 Maleki F, Muthukrishnan N, Ovens K, Reinhold C, Forghani R. Machine Learning Algorithm Validation From Essentials to Advanced Applications and Implications for Regulatory Certification and Deployment. Neuroimaging Clin N Am 2020; 30: 433-445 [PMID: 33038994 DOI: 10.1016/j.nic.2020.08.004]
4 Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. Concepts, Tools, and Techniques to Build Intelligent Systems. 2nd Ed. O’Reilly Media, Inc, 2019: 851
5 Rafało M. Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis. ICT Express 2021 [DOI: 10.1016/j.ictex.2021.05.001]
6 Wainer J, Cawley G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. Expert Syst Appl 2021; 182: 115222 [DOI: 10.1016/j.eswa.2021.115222]
