einstein
Official Publication of the Instituto Israelita de Ensino e Pesquisa Albert Einstein

e-ISSN: 2317-6385

LETTER TO THE EDITOR

Precision over quantity: identifying trustworthy genomic risk stratification models

Valbert Oliveira Costa Filho[1], Pedro Robson Costa Passos[1]

[1] Center of Research and Drug Development, Universidade Federal do Ceará, Fortaleza, CE, Brazil.

DOI: 10.31744/einstein_journal/2025CE1726

Dear Editor,

The Brazilian Society of Medical Genetics and Genomics recently issued a statement against using polygenic risk scores for breast cancer risk stratification in Brazilian women.(1) In fact, an ongoing surge could be observed in genomic signatures and prognostic models, with over 200 LASSO-based TCGA signatures appearing in *Scientific Reports* last year, based on a focused keyword-based search, several of questionable quality. Discerning reliable models from unreliable ones is crucial for clinicians in order to avoid poorly validated tools.

The first step is to assess how the model was built and whether it was properly validated. Currently, we distinguish two common validation types. The first involves splitting the patient dataset into proportions such as 80/20 or 70/30 for model training and testing. Unfortunately, this approach is less effective, as machine learning after a single random training-test split might yield unreliable results.(2) The second and more robust method involves validation using an entirely external dataset from independent studies. This approach is more reliable, as internal data splitting often fails to truly validate a model and cannot replace independent replication across different researchers, populations, and methodologies.(3)

The next step is the assessment of the area under the receiving operating curve (AUC), representing the integral of all points along the curve and capturing both sensitivity and specificity to provide a comprehensive measure of the overall performance of the model. In general, a model with an AUC <0.6 could be considered unreliable and acceptable models should display an AUC of at least 0.7.(4) Subsequently, evaluating the biological plausibility of the genes included in the model is essential. For instance, models in which a tumor suppressor expression predicts worse survival or an oncogene expression predicts better prognosis, could lack reproducibility. Models and signatures should (or should aim to) retain biological relevance and plausibility. Therefore, assessing whether the applied genes have been validated in the disease through methods beyond transcriptomics (*e.g.*, real-time PCR or immunohistochemistry) is also necessary.

Finally, most models and signatures are developed using publicly available datasets, such as those from TCGA or the Gene Expression Omnibus, and are typically produced and evaluated retrospectively. To ensure reproducibility and enable the integration of these models into clinical practice, model validation in prospective studies would be essential. This step is crucial, as prospective validation provides an important first assessment of the real-world machine learning model performances.(5)

These steps could potentially prove particularly valuable for clinicians, helping them avoid poorly designed studies, as this emerging field is underrepresented

How to cite this article:
Costa Filho VO, Passos PR. Precision over quantity: identifying trustworthy genomic risk stratification models [letter]. einstein (São Paulo). 2025;23:eCE1726.

Associate Editor:
Kenneth Gollob
Hospital Israelita Albert Einstein, São Paulo, SP, Brazil
ORCID: https://orcid.org/0000-0003-4184-3867

Received on:
Mar 8, 2025

Accepted on:
Mar 27, 2025

in the undergraduate curricula and its literature could be hard to navigate. Figure 1 presents a flowchart, summarizing the algorithm for identifying robust models.

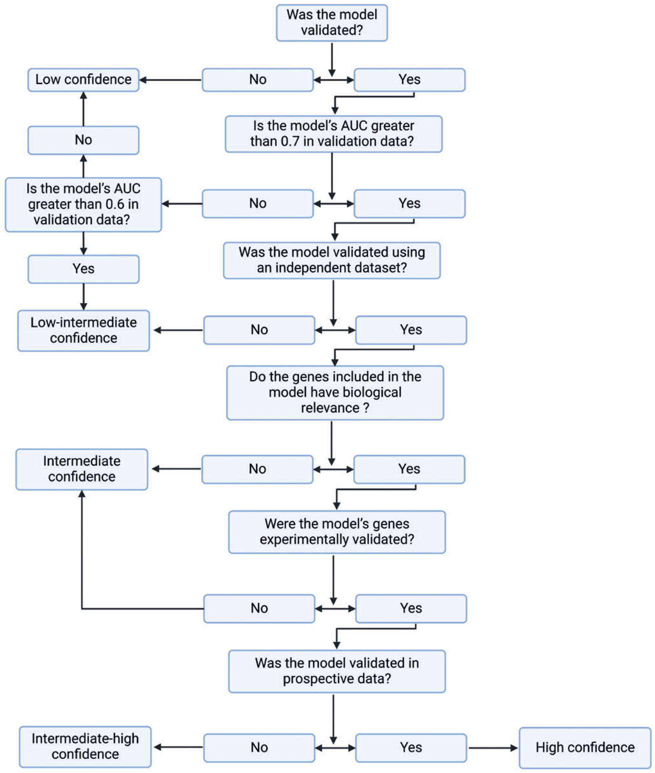
Was the model validated?
Low confidence
No
Yes
Is the model's AUC greater than 0.7 in validation data?
No
Is the model's AUC greater than 0.6 in validation data?
No
Yes
Yes
Was the model validated using an independent dataset?
Low-intermediate confidence
No
Yes
Do the genes included in the model have biological relevance ?
Intermediate confidence
No
Yes
Were the model's genes experimentally validated?
No
Yes
Was the model validated in prospective data?
Intermediate-high confidence
No
Yes
High confidence

Figure 1. Algorithm for identifying robust models

AUTHORS' CONTRIBUTION

Valbert Oliveira Costa Filho: conceptualization, data curation, formal analysis, investigation, validation, visualization, writing - original draft, and writing - review & editing. Pedro Robson Costa Passos: data curation, writing - original draft, and writing - review & editing.

AUTHORS' INFORMATION

Costa Filho VO: http://orcid.org/0009-0003-2864-0966
Passos PR: http://orcid.org/0009-0007-9049-0996

REFERENCES

1. Sociedade Brasileira de Genética Médica (SBGM). Posicionamento da SBGM sobre o uso de escores de risco poligênico para estratificação de risco para câncer de mama em mulheres brasileiras. Porto Alegre: SBGM; 2025 [citado 2025 Mar 8]. Disponível em: https://www.sbgm.org.br/detalhe.aspx?id=6379&area=4
2. An C, Park YW, Ahn SS, Han K, Kim H, Lee SK. Radiomics machine learning study with a small sample size: Single random training-test set split may lead to unreliable results. PLoS One. 2021;16(8):e0256152.
3. Källberg H, Alfredsson L, Feychting M, Ahlbom A. Don't split your data. Eur J Epidemiol. 2010;25(5):283–4.
4. Polo TC, Miot HA. Aplicações da curva ROC em estudos clínicos e experimentais. J Vasc Bras. 2020;19:e20200186.
5. Brajer N, Cozzi B, Gao M, Nichols M, Revoir M, Balu S, et al. Prospective and external evaluation of a machine learning model to predict in-hospital mortality of adults at time of admission. JAMA Netw Open. 2020;3(2):e1920733.