ISSN: 3107-4707
DOI: doi.org/10.51219/JMMS/E-G-Stephen/24

URF PUBLISHERS
connect with research world

Journal of Medicine and Medical Studies
https://urfpublishers.com/journal/medicine-medical-studies

Vol: 2 & Iss: 2

Mini Review

Inside the "Black Box of Predictive Models" 6 Things the Cardiologist Should Know about AI and Precision Medicine

Stephen G Ellis

Citation: Ellis SG, Inside the "Black Box of Predictive Models" 6 Things the Cardiologist Should Know about AI and Precision Medicine. *J M Med Stu 2025;* 2(2): 126-127. DOI: doi.org/10.51219/JMMS/E-G-Stephen/24

Received: 02 April, 2025; **Accepted:** 09 April, 2025; **Published:** 11 April, 2025

***Corresponding author:** Stephen G Ellis, Email: elliss@ccf.org

As large amounts of patient data become increasingly available, apps for predicting clinical outcomes will proliferate. Imagine a scenario where you have an app on your device that seems to answer the clinical question you have, perhaps whether you should place the patient before you on statins or if their disease is more advanced, offer PCI or CABG? You plug in the key variables for your patient and come up with a predicted result. You have heard, correctly, that new AI-based outcome modelling is usually better than that with traditional statistical modelling1. Should you trust the results?

What should you ask?

1. Are a lot of patients like mine in the dataset from which the model was derived?

It is well known that patients with substantial comorbidities, those not at equipoise, as well as women and minorities are under-represented in U.S.-based randomized trials (RCT) that often drive clinical guidelines. This is not the case in broad-based registries such as the EPIC Cosmos, the US National Inpatient Registry or ACC/NCDR PCI Registry, but these registries have their own deficiencies (see below). To get data and models on Caucasians who were excluded from RCT, subset analysis from comprehensive datasets such as Swedeheart can be valuable. However, if your patient is African-American, Hispanic, Asian or from low socio-economic status, you'll need to look elsewhere, as all of these factors influence (particularly long-term) outcomes. Even models from large datasets such as the Pooled Cohort Equation (PCE) models perform less well in minority groups[2]. Finally, for AI-based modelling, such as that by XG Boost or Random Survival Forests, which iteratively model overlapping subsets of patients until their loss function (the difference between their prediction and reality) is stably minimized and outperform traditional model with complex datasets, truly large numbers of patients like yours are needed.

2. Are the data unbiased?

One would certainly like to think large broad-based databases should be unbiased, but what about predictive models that come from industry or use industry-funded data? We all know that data can be cherry-picked and conclusions "spun." Risk of bias is common even in RCT[3]. It is known that industry-funded studies that are "positive" are more likely to be published than those that are neutral or negative[4]. This can skew even well intentioned "neutral databases". As an example, and not to be judgmental, but industry-based models such as the QRISK-3 model have been criticized for overestimating patient risk of MI and CVA.

3. Is the nature of the dataset appropriate to the question you seek to answer?

There are many issues here. Many of the really big U.S. datasets are based only on ICD-10 codes, medications prescribed and lab values. Administrative datasets often lack important details, often have incomplete data, are subject to miscoding/misclassification and have incomplete patient follow up. They also lack information on patient quality of life and their desired outcomes. They may be reasonable, in concept, to access the relationships between baseline characteristics and later MACE such as used in the PCE, but they lack the nuance to accurately predict procedural outcomes that may be found in the NCDR PCI, STS and TVT Registries. On the other hand, the NCDR PCI dataset, even if supplemented by survival data, may also be poorly suited to predict long term general outcomes such as mortality. For instance, we have developed models to predict long-term mortality after PCI using data from ACC/NCDR supplemented by natural language processing (NLP)- extracted data from our patients' EPIC EMR. Five of the top 10 predictive correlates were not in the ACC/NCDR dataset (e.g. serum albumin, diuretic/dose and depression [all more important than LVEF]). Additionally, coding cause of death is notoriously unreliable, so focusing on cardiac death is highly problematic. Digging deeper, even in 2024, NLP-based data extraction from most databases is typically accurate only 75-90% of the time (free text is especially challenging)[5]. That said, large datasets often have results far more generalizable than those from single center or more limited datasets.

4. Has field evolved since the dataset was constructed?

The PCE and MESA cohorts are good examples of models developed from large datasets to inform decisions about preventive treatments such as aspirin and statins. As calcium scoring became available, it became clear that inclusion of these data improved the model's predictive capabilities[6]. Since non-calcified plaque is less stable that calcified plaque, it stands to reason that quantization of non-calcified coronary plaque, soon to be readily available, will provide even more information. Genetic data is also becoming increasingly available. Beyond this, the widespread use of GLP-1 inhibitors will likely reduce risk. There is no good solution to the problem of predicting long-term outcomes when background treatment and available tests are evolving quickly, but physicians need to be aware of the latest data.

5. Is the model good?

Currently, we use a number of models that really aren't that good. For example, the commonly used CHADSVASC2 score has validation c-statistics ranging from 0.59-0.67 and the DAPT score from 0.49-0.71. One might wonder why thought leaders and our societies haven't stressed these limitations. Perhaps it's because these models are better than a "gut choice." The thoughtful cardiologist should at least know the basics of how to critique a model. Models (AI generated or not) should be evaluated on how well does the tool discriminate risk [low, medium and high for example; typically measured by the c-statistic [0.50 no discrimination to 1.00 complete separation on the basis of risk; good: 0.7-0.8, very good 0.8-0.9]) or the statistically better F1 score [good 0.7-0.8, very good >0.8]), calibration (does the predicted risk match the actual risk? (The tool isn't helpful if it discriminates amongst patient's risks but under or overestimates it) assessed by the calibration plot, Brier score or Hosmer-Lemeshow test and generalizability (results of the model in datasets other than the one it which it was developed.) Beyond this and at a more nuanced level, they need to minimize confounder effect and avoid overfitting (AI based models have, on average, three times more predictors than non-AI models, so they are particular risk of this problem)[7] Guidelines for quality modelling with AI have been published recently[8].

6. Is there a good reason to think that the results at my practice/hospital should be different?

Predicting outcomes of treatments that involve physician skill (devices) are fundamentally more challenging than those that don't (drugs). Although many of our current procedures are largely standardized (PCI, TAVR), data exist that should make us question how well global results apply to your patient. For instance, CABG-related mortality in the SYNTAX trial varied widely by hospital. As another example, if we know that intra-coronary imaging (ICI) improves PCI outcomes and yet it is only used in a small minority of cases nationwide, should you trust national outcome data if your center uses ICI in 90% of its cases?

There are always trade-offs in medicine. RCTs eliminate the treatment biases that contaminate observational trials, but their results apply only to the typical patient in the study. Patients want and will increasingly expect treatment recommendations tailored specifically to their situation (think genetic risk factor-based cancer therapy). Large dataset, AI-based predictive apps have the potential to meet this need, but they should not be followed without question.

Should you have to know the basics of what we just reviewed? Perhaps not. It would be better if our cardiac societies would do the model evaluation for us, but from the review above it seems that they can't be fully relieved upon. In the end, your patients depend on you to know what's best.

7. References

1. Liu W, Laranjo L, Klimis H, et al. Machine-learning versus traditional approaches for atherosclerotic cardiovascular risk prognostication in primary prevention cohorts: a systematic review and meta-analysis. European Heart Journal-Quality of Care and Clinical Outcomes, 2023;9(4): 310-322.
2. Zinzuwadia AN, Mineeva O, Li C, et al. Tailoring Risk Prediction Models to Local Populations. JAMA cardiology, 2024;9(11): 1018-1028.
3. Baasan O, Freihat O, Nagy DU, Lohner S. Methodological Quality and Risk of Bias Assessment of Cardiovascular Disease Research: Analysis of Randomized Controlled Trials Published in 2017. Frontiers in Cardiovascular Medicine, 2022;9(Mar): 830070.
4. DeVito NJ, Goldacre B. Catalogue of bias: publication bias. BMJ Evidence-Based Medicine, 2019;24(2): 53-54.
5. Wornow M, Xu Y, Thapa R, et al. The shaky foundations of large language models and foundation models for electronic health records. npj Digital Medicine, 2023;6(July): 135.
6. Pletcher MJ, Sibley CT, Pignone M, et al. Interpretation of the coronary artery calcium score in combination with conventional cardiovascular risk factors: the Multi-Ethnic Study of Atherosclerosis (MESA). Circulation, 2013;128(10): 1076-1084.
7. Damen JA, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ, 2016;353(May): 2416.
8. Cohen JF, Bossuyt PM. TRIPOD+ AI: an updated reporting guideline for clinical prediction models. BMJ, 2024;385(Apr): 824.