Supplementary material

Development and Internal Validation of a Machine Learning-Based Colorectal Cancer Risk Prediction Model

Deborah Jael Herrera †, [1], Daiane Maria Seibert †, [2], Karen Feyen [2], Marlon van Loo [3], Wessel van de Veerdonk §, [1,3], Guido Van Hal §, *, [1]

† **Shared co-first authorship**
§ **Shared co-last authorship**
* Corresponding author: guido.vanhal@uantwerpen.be

Affiliations:
[1] Family Medicine and Population Health Department (FAMPOP), Faculty of Medicine and Health Sciences, University of Antwerp, 2610 Antwerp, Belgium; deborahjael.herrera@uantwerpen.be (D.J.H.); guido.vanhal@uantwerpen.be (G.V.H.)
[2] Centre of Expertise – Design and Technology, Campus De Nayer, Thomas More University of Applied Sciences, 2860 Sint-Katelijne-Waver, Belgium; daiane.seibert@thomasmore.be (D.M.S.), karen.feyen@gmail.com (K.F.)
[3] Centre of Expertise - Care and Well-being, Campus Zandpoortvest, Thomas More University of Applied Sciences, 2800 Mechelen, Belgium; wessel.vandeveerdonk@thomasmore.be (W.v.d.V.), marlon.vanloo@thomasmore.be (M.v.L.)

Table S1. Key terms and definitions to machine learning in clinical research

Term (Abbreviation)	**Definition**
Machine Learning (ML)	A mathematical model that can improve its performance on a task by exposure to data.
Light Gradient Boosting Machine (LightGBM)	LightGBM is an advanced machine learning technique that improves prediction accuracy by combining many smaller, simpler models called decision trees. Each decision tree makes predictions, and when combined, they create a more powerful model. LightGBM uses a leaf-wise (best-first) growth strategy, splitting the leaf with the maximum loss reduction at each step. To prevent LightGBM from making overly specific predictions that may not work well on new data (known as overfitting), it uses a method called 'gradient-based one-side sampling' (GOSS). GOSS focuses on the most important data points that contribute the most to prediction errors, allowing LightGBM to learn from the data more efficiently. In machine learning, gradients are values that measure how much a function's output changes with respect to changes in its input. For LightGBM, gradients represent the residual errors between the current predictions and the actual target values. They help optimize the model by indicating the direction and magnitude of adjustments needed to reduce prediction errors.
Decision trees	They are essentially flow charts that can be used to predict outcomes based on branching logic. Given input feature values, the model undertakes a series of binary decisions to reach the proper outcome. Over the course of training, the model learns to navigate each branch point with increasing accuracy [50]. These models are best used when a high degree of interpretability is sought, but their performance may suffer relative to more complex modelling strategies
Training, test, and validation sets	*Training set:* Dataset from which the model learns the optimal parameters to accomplish the task.

	Test set: Dataset on which the performance of a trained, parameterized model is evaluated. *Validation set:* Dataset that is used to evaluate the model's performance during training. Differs from a test set in that it is used during training to establish hyperparameters of the model.
Supervised learning	A subset of ML in which the outcomes to be learned by the model ("labels") are provided in the training set. For example, teaching a model to estimate patient's risk of CRC for prognostic purposes would require training the model on a training set containing labelled patients with and without CRC prior to validating that model on a new set of unlabelled patients with and without CRC.
SHapley Additive exPlanations | SHAP	Method used to explain the output of machine learning models. It leverages concepts from cooperative game theory, specifically Shapley values, to assign an importance value to each feature for a given prediction. SHAP values represent the contribution of each feature to the prediction by considering all possible combinations of features. SHAP guarantees three key properties: local accuracy, missingness, and consistency, making it a reliable and interpretable method for understanding complex models. (https://arxiv.org/pdf/1705.07874)

Note: Definitions of terms are adapted and modified from Weissler et.al. 2021 [1] to fit the context of CRC risk prediction model.

References:

1. Weissler, E Hope et al. "The role of machine learning in clinical research: transforming the future of evidence generation." Trials vol. 22,1 537. 16 Aug. 2021, doi:10.1186/s13063-021-05489-x