

Risk Stratification Strategies: From Logistic Regression to Artificial Intelligence

Thomas F. Imperiale, MD (1, 3-6) and Patrick O. Monahan, PhD (2, 5, 6)

From the Departments of Medicine (1) and Biostatistics (2), Indiana University School of Medicine; Health Services Research and Development, Richard L. Roudebush VA Medical Center (3); Regenstrief Institute, Inc.(4); Indiana University Melvin and Bren Simon Cancer Center (5), and the Indiana University School of Public Health (6); Indianapolis, IN

CONTACT INFORMATION:

Thomas F. Imperiale, MD

Regenstrief Institute, Inc.

1101 West 10th Street

Indianapolis, IN 46202

TEL: 317.274.9046

FAX: 317.274.9304

Email: timperia@iu.edu

Patrick O. Monahan, PhD

Health Information and Translational Sciences

410 West 10th Street Suite 3000

Indianapolis, IN, 46202

TEL: 317.278.8086

FAX: 317.274.2678

pmonahan@iu.edu

Corresponding author: Thomas F. Imperiale, MD

Disclosure statement: The Authors have nothing to disclose.

Word count (synopsis): 107

Word count (manuscript without references): 7620

Tables: 0

Figures: 4

Key Words: risk stratification, colorectal cancer screening, risk prediction models, risk prediction, cancer prevention, and multivariate methods.

KEY POINTS

This is the author's manuscript of the article published in final edited form as:

Imperiale, T. F., & Monahan, P. O. (2020). Risk Stratification Strategies for Colorectal Cancer Screening: From Logistic Regression to Artificial Intelligence. *Gastrointestinal Endoscopy Clinics of North America*, 30(3), 423–440. <https://doi.org/10.1016/j.giec.2020.02.004>

- Risk stratification in colorectal cancer (CRC) screening involves using factors associated with CRC or advanced neoplasia to estimate individual patient risk. It requires a decision or question, a risk and its timeframe, risk factors, and a system or method for integrating them.
- Among the methods for integrating risk factors (collectively referred to as artificial intelligence), two distinguishing features are: 1) whether the system uses clinical pre-processing or instead a ground up approach in which the machine explores unfiltered categories of candidate predictor variables, and; 2) the flexibility-interpretability tradeoff inherent in the method.
- Several risk prediction models / risk stratification schemes predict current risk of advanced neoplasia, while far fewer predict future risk of CRC. The models vary widely in their performance metrics, which include calibration, discrimination, and risk separation.
- Risk stratification in CRC screening has the potential to improve the uptake, efficiency, and effectiveness of screening. This area requires further research on model validation and impact assessment.

SYNOPSIS

Risk stratification is a system or process by which clinically-meaningful separation of risk is achieved in a group of otherwise similar persons. While parametric logistic regression dominates risk prediction, use of nonparametric methods such as classification and regression trees, artificial neural networks, and other machine-learning methods are increasing. Collectively, these learning methods are referred to as “artificial intelligence” (AI). The persuasive nature of AI requires knowledge of study validity, an understanding of model metrics, and determination of whether and to what extent the model can and should be applied to the patient or population under consideration. Further investigation is needed, especially in model validation and impact assessment.

INTRODUCTION

Risk stratification is an important tool in both clinical medicine and population health. In the current era of precision / personalized medicine, risk stratification may be used to tailor preventive, diagnostic, and therapeutic strategies for individual patients or for subgroups of a population. For more than a half century, parametric regression modeling (mainly logistic regression) has been the dominant statistical technique used for creation of risk models and subsequent risk stratification tools. However, within the last decade, more automated (i.e., “ground-up”) approaches, and nonparametric methods such as neural networks, have established a presence in risk prediction; these newer methods promise to have a larger role going forward. To most, these are considered “black box” methods; however, they have an understandable and underlying theoretical and statistical foundation. Although these methods have flexibility to fit nuances in relationships, making their results more difficult to interpret, ad-hoc techniques are emerging to help improve interpretability.

The aim of this article is to define and discuss the current and emerging methods of risk prediction and risk stratification, and to do so within the clinical and population health context of colorectal cancer (CRC) screening. This article will discuss the what, how, and why of risk stratification as it applies to CRC screening, along with consideration of barriers to using risk prediction models for CRC screening. We also describe an agenda for research.

RISK STRATIFICATION

The need to understand, estimate, and incorporate risks into decision making is omnipresent in both clinical medicine and public health. Risk is used in diagnosis (“What is the likelihood of this patient having disease X?”); prognosis (“What is this patient’s chance of surviving 5 years with condition X?”); response to treatment (“How likely is this patient to respond to Drug X?”), and; disease prevention (“Is screening this patient for X required at this time?”). We may begin the process of risk estimation in each of these settings with a general sense of the risk in the population under consideration or in the particular clinical setting. We then use risk factors (demographics, signs, symptoms, and other test results) to revise the risk estimate. The estimate moves up or down until it crosses a threshold for further testing, treating, or no further consideration of a particular diagnosis.

Risk stratification is an extension of the process of risk estimation, consisting of a system or process by which clinically-meaningful separation of risk is achieved in a group of otherwise similar persons. It is a tool used in both clinical medicine and in population health to estimate risk for a particular outcome – for either an individual person or a group of individuals – and to separate individuals into different levels of risk - generally low, intermediate, and high risk. The separation is usually achieved based on specific demographic, lifestyle, clinical, and more recently, genetic, features of each member of the group. Risk stratification enables providers and policy makers to identify an appropriate level of health care and health services for the distinct subgroups (“strata”). It has the potential to improve the quality, appropriateness, and efficiency of health care provided.

One of the first and best-known examples of risk stratification is the Framingham Risk Score, created to estimate a person’s 10-year risk of developing coronary heart disease (CHD).¹ It was created from a large cohort study of disease-free residents of Framingham, MA who were followed for 20 years, during which time some residents experienced clinical events of CHD. Six factors were independently related to CHD risk in both gender-stratified analyses: age, blood pressure, total cholesterol level, high-density lipoprotein level, diabetes, and cigarette smoking.² Gender-specific risk scores were subsequently developed, for which “plugged-in” patient-specific values provide an estimate of the 10-year risk of a CHD event.³

In the areas of gastroenterology and hepatology, risk prediction models and risk scores are available for a myriad of diagnoses. Best known are the prognostic models for acute upper gastrointestinal bleeding (the Blatchford and Rockall scores^{4,5}, among several); acute pancreatitis (APACHE-II and BISAP scores^{6,7}); and progression of cirrhosis (Child-Turcotte-Pugh, MELD^{8,9}), to name a few. In the domain of disease prevention, risk scores are less prevalent due to the challenges of risk prediction for conditions such as cancer, which are uncommon if not rare at the population level. Nevertheless, there are risk prediction models for breast cancer that are used in decision-making about when and how to screen women for breast cancer.¹⁰ Below, we discuss the ones created for risk prediction of CRC and its precursor lesion, the advanced adenoma. We first want to consider the “why” of risk prediction models for CRC.

Why Risk Stratification?

There are many reasons – both individual patient-wise and societal – to use risk stratification in general. An overarching reason is that more and more people expect it. They hear the terms “personalized medicine,” “precision medicine,” and “providing the right care to the right patient at the right time”, and want this for their own health care. Risk stratification may be used to tailor whether, when, and how to screen, test, or treat for a specific condition. Tailoring involves adjusting the intensity of management based on an individual’s risk for a particular diagnosis. For patients, the adjustment is intended to optimize the balance between benefits and harms; for society, it is intended to improve efficiency – the use of current health care resources while minimizing waste and cost of unnecessary, low-yield, or low-value tests and treatments.

These principles are readily applied to CRC screening. From the individual patient perspective, there are at least a few reasons why CRC screening should be tailored more aggressively toward those at high risk. First, CRC is common among cancers (the 2nd most common cancer and cancer death in men and women combined). Second, screening is effective in reducing CRC incidence and mortality, and does so cost-effectively. Third, screening is underutilized; current adherence rates are 60-65%.¹¹ In the same way, there are reasons to use risk stratification to tailor screening away from low-risk persons. Screening has risks. For example, false-positive test results can lead to unnecessary procedures with their more significant risks. These risks are of greater concern and relatively greater magnitude when the potential benefit is small or none. Further, when it comes to prevention of a relatively rare condition such as CRC, the great majority of persons do not benefit from screening.

From a societal perspective, using risk stratification to tailor CRC screening towards high-risk persons should be done to minimize lost productivity and downstream economic effects. Further, care of the cancer patient is expensive, particularly for advance stage disease, with cost of some of the newer “personalized” agents exceeding \$100,000 per year.¹² Similarly, reducing screening intensity for low-risk persons has the favorable effects of reducing the total cost of screening (i.e., CRC screening may be cost-effective, but costs in the tens of billions of dollars annually)¹³ and making resources available for other patients and other societal needs.

What is Required for Risk Stratification?

Any use of risk stratification requires four elements. The first is the clinical or public health decision or question under consideration – for example, “What is this patient’s prognosis given condition X?” or “How should this patient (or the population) be screened for CRC?” The second element is the condition or outcome. For CRC, the most relevant clinical outcomes are CRC itself and advanced neoplasia ([AN], the combination of CRC and advanced precancerous polyps). Since advanced precancerous polyps nearly always lend themselves to definitive treatment at the time of colonoscopy, they are an attractive outcome. Further, because of their higher prevalence (vs. CRC), AN has driven the great majority of studies on risk factors for CRC. The third element is identifying the risk of interest. What is the best overall estimate for risk? Part of considering this element is to identify factors affecting the risk under consideration, as well as its timeframe: is it a current risk or future risk? Regardless of whether a future or current risk is chosen, there needs to be factors associated with that risk. Under current implementation of CRC screening, only age and family history are considered in decision-making about whether, when, and how to screen. Depending on the specific guideline and family history, age is dichotomized at 50, 45, 40, or 10 years earlier than the youngest first-degree relative was at the time of CRC diagnosis. For average-risk persons, CRC risk nearly doubles for each decade between ages 50 and 80,¹⁴ yet this information is not used in decision-making. We know much more about the factors related to risks of CRC and AN, yet we do not use them currently. These factors include cigarette smoking history, diet, physical activity, BMI, metabolic syndrome, diabetes, regular use of NSAIDs and aspirin, to name a few.¹⁵ One reason these factors are usually not considered is the lack of a way to readily integrate or combine these factors to estimate risk. This shortcoming leads us to the final element required for risk estimation / stratification: a statistical method that allows simultaneous consideration of several risk factors to measure individual patient risk and place that risk into a stratum (or provide an estimated probability of risk). Let’s now consider some of the current as well as up-and-coming methods for risk estimation and risk stratification. All of these methods are considered to perform both “machine learning” and “statistical learning” in that they use computers to generate algorithms using parametric or nonparametric methods that have a statistical foundation.¹⁶

SELECTED MODELING TECHNIQUES FOR RISK PREDICTION

Overview

“Artificial intelligence” can refer to either simple rules or to “identifying patterns in the data” using specific machine learning and statistical learning methods (Figure 1). Two important features distinguish artificial learning methods. The first is whether the analyst applies clinical judgment to pre-process the data or instead applies a purely “ground-up” approach in which the machine explores the unfiltered or uncombined categories of the original predictors (called “features” in machine learning). The clinical pre-processing approach is often used in the biostatistical community, and involves understanding the modeling process in context of the underlying science. This process includes making decisions about how to categorize continuous or multi-category predictors into more clinically meaningful categories, such as pack-years for smoking or intervals for BMI. In the ground up approach, the analyst purposely attempts to minimize clinical pre-processing to reduce analyst time (and thereby increase scalability) and to increase the potential to uncover potential relationships that are not preconceived. Nevertheless, the ground-up approach often involves data pre-processing decisions such as how to handle missing data. The second important feature is the flexibility-interpretability tradeoff of the specific method used. The salient distinction between methods is their flexibility (and consequently their interpretability), not their terminology or their origins (e.g., machine vs statistical learning). For example, regarding origins, logistic regression is a traditional statistical method but is included in machine learning text books. Neural network analysis is commonly used in the machine learning community, but has a statistical foundation, is included in statistical learning text books, and is similar to the statistical methods of project pursuit regression¹⁶ and polynomial logistic regression.¹⁷

The difference among these methods is the extent of their flexibility¹⁶ for classifying or predicting a categorical response variable such as colorectal cancer or AN. These methods are called “supervised” learning in the computer science community because training is supervised by known values of the dependent variable. In unsupervised learning, the analysis of a set of variables is not connected to an outcome variable. Examples of unsupervised learning include variable-based factor analysis, or person-based cluster analysis, performed to understand whether risk factors can be represented by a fewer number of factors or whether the data indicate there are groups of persons with similar combinations of

risk factors, respectively. We will not discuss unsupervised methods further because they are less relevant for risk stratification. Supervised parametric models such as logistic regression provide: (1) a test and an estimate of the magnitude of association (i.e., odds ratio [OR]) between each predictor and the response (also known as the outcome or dependent variable); (2) easily interpretable results because a single regression coefficient describes the relationship between the feature and the response, and; (3) an estimate of the probability of the response variable at the subject level. In logistic regression, the shape of the relationship between a continuous predictor (independent) variable and the response (dependent) variable is not as flexible because it assumes linearity between continuous features and the log-odds of the response variable. Models with this restriction usually perform very well, and are preferred for their parsimony, if the true population relationship is reasonably linear.

If the population relationship between the predictor and outcome variables is markedly nonlinear, more accurate learning can be achieved by using semi-parametric and nonparametric methods that attain greater flexibility. These nonparametric and semi-parametric methods include classification tree analysis, support vector, random forest with bagging and boosting, neural networks, and semi-parametric logistic regression. In addition, “ensemble” methods incorporate results from more than one learning method or from more than one resampled data set within a learning method. Increased flexibility in these methods is generally attained through one or both of two general avenues: by incorporating non-linear relationships and/or interactions. Interactions allow the relationship between one feature and the response to depend on one or more other features, resulting in identifying subgroups. These subgroups are defined by combinations of features and can be illustrated by branches of trees, such as in the classification tree methods. Classification tree methods also conveniently search for the data-driven thresholds for categorizing continuous features when forming these subgroups. Nonlinear relationships are incorporated in methods such as neural networks and generalized additive models.

Generalized additive models, such as nonparametric and semi-parametric logistic regression, have origins in the statistical community. They incorporate nonlinear relationships between continuous predictors and the response variable by using scatterplot smoothers, while retaining a high degree of interpretability through the ability to provide probability estimates and tests of association.¹⁶

Although some flexible methods do not routinely provide probability estimates or tests of association and therefore have lower interpretability, their process for classification does have a probabilistic foundation. For example, the artificial neural network method was inspired by brain neural networks but can be demystified by realizing that it essentially involves creating new variables (i.e., hidden layers), derived from weighted combinations of features, which are then used to predict the response through nonlinear (e.g., sigmoid [or S-shaped]) functions. Furthermore, these “activation” functions are not deterministic or binary as mistakenly thought by some practitioners; they assume values ranging from 0 to 1, which are then categorized into a 0 vs 1 response classification. The neural network “cost function” is similar to the sum of squared residuals between observed and fitted values. Partial derivatives from calculus are used to iteratively update and estimate weights that minimize the cost function. Although calculus is used in neural networks, no learning method is inherently deterministic. Even if probabilities are not explicitly represented in formulas, all learning about complex phenomena, by definition, is affected by statistical sampling error. Sampling error is captured in the validation process, explained below, making validation one of the most important aspects of evaluating model performance.

It is necessary to understand the metrics by which all models are evaluated. The two most common measures for model evaluation are calibration and discrimination, both of which are statistical measures. Calibration conveys how well the model “fits” the data, and is measured by comparing the difference between observed and expected values for each observation in sample. A small difference indicates a good-fitting model. Discrimination refers to how well a model can discriminate between an observation with the outcome and one without it. In logistic regression, a popular discrimination measure is the area under the receiver operating characteristic curve (AUROC; estimated with the c-statistic), which plots sensitivity versus (1-specificity) (Figure 2). The diagonal bisecting line in Figure 2 represents no discrimination (as good as tossing a coin). The AUROC for the model in Figure 2 is 0.83, representing very good discrimination. Calibration and discrimination are “statistical” aspects of model performance, discrimination being the more important of the two. There is a third measure that is more “clinical”; it is the risk separation between strata or categories. This measure is more of a judgment about whether the difference(s) in estimated risk between or among categories is meaningful enough to affect decision-making about management. For example, if the respective risks for AN in the low-risk, intermediate-risk,

and high-risk groups are 1%, 3%, and 6%, it's questionable whether these differences are "important" enough to affect which test is used for screening – a non-invasive test or lower endoscopy. On the other hand, if the respective risks for AN are 2%, 7% and 25%, most persons in the high-risk group might best consider colonoscopy, whereas those in the low-risk group could safely and efficiently choose non-invasive screening.

For any model, the proof of how "good" it is requires testing it, a process known as validation. If allowed by larger samples sizes, the validation and test process can be separated into a validation data set used for evaluating particular "tuning" parameters and model building steps, and a test data set reserved for the final test of competing models. Some degree of model validation should be expected by potential adopters of the model. Indeed, the lack of literature on robust validation of models may be one reason limiting their more widespread use in clinical practice and public health settings. The rigor of testing determines the robustness of the model. Internal methods rely on a single data set and involve submitting it to resampling (e.g., bootstrap) or multi-split-sample cross-validation methods. A more robust test of a model is a prospective evaluation by independent investigators in another setting from the one(s) in which the model was developed.

Among the internal methods, an estimate of the out-of-training-sample prediction error can be obtained either indirectly and approximately through statistical indices such as the Akaike Information Criterion (AIC) or Mallow's C, or directly through cross-validation.¹⁶ The most commonly used internal cross-validation method is the "split-sample" technique, where a portion of the dataset (typically 50-80%) is used for model development, and the remaining portion is used to test the model's performance. However, an alternative, the *k*-fold cross-validation procedure, commonly performed with 5 or 10 folds, has several advantages.¹⁶ First, its results have less bias than those of the split-sample procedure and lower variability (i.e., better precision) than those of the leave-one-out procedure.¹⁶ Second, unlike the split-sample procedure, the *k*-fold procedure provides an estimate of the mean and standard error of the cross-validation prediction error across the (e.g., 5 or 10) hold-out validation samples.¹⁶ Third, non-statisticians tend to find split-sample and *k*-fold cross-validation more intuitive and easier to understand than the "leave-one-out" or bootstrap resampling procedures.

Validation is crucial because more flexible methods, by definition, usually fit data and classify responses in training (or derivation) samples better than less flexible methods. More complex models are those that have more parameters, more non-parametric “smoothing” (and therefore more non-linearity), and more interaction terms. In cross-validation, more flexible and complex methods may fit a validation or test set poorly compared to less flexible methods. For example, age as a continuous variable will perfectly classify the outcome (e.g., presence or absence of AN) in a training sample of 100 persons if 99 polynomial terms for age are entered into logistic regression. However, this high-degree-polynomial model would perform poorly on cross-validation metrics (and probably worse than a simple linear model) because the 98 nuanced bends in the curves between age and the logit of the outcome will be very different in one sample versus another sample.

Logistic Regression

Logistic regression is most frequently used in risk stratification. What is being modeled is the logarithm of the odds of the outcome (or dependent) variable. The underlying assumptions are that the outcome variable has a binomial distribution, and that the relationship between the log odds of the outcome and the predictors is linear.

Logistic regression is the appropriate method when the dependent variable is binary or dichotomous (e.g., presence or absence of disease); it explains the relationship between this single dependent dichotomous variable and one or more independent variables. It allows development of a predictive equation based on the probability of the outcome (such as present/absent, dead/alive). As with all modeling methods, the goal of logistic regression is to maximize the number of correct predictions of outcome for individual observations in the sample dataset using the most parsimonious model, meaning the simplest model that satisfies the underlying assumptions with the fewest variables and greatest explanatory power. The output of logistic regression is an equation that follows the format:

$$\text{Log}_e (p/1-p) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n,$$

.....where p is the probably of a binary event (e.g., death, or presence / absence of AN), x_1 , x_2 , and x_n are the independent (or predictor) variables, and $\beta_{1 \rightarrow n}$ are the regression coefficients associated with

reference group and the $x_{1 \rightarrow n}$ variables. The reference group, β_0 , is derived from those individuals who have the reference level for all of the independent variables.

Even though the outcome for each observation is known to be present or absent, the estimated probability for any single observation in the dataset is calculated by plugging in the specific values for the predictor variables (for example, age, sex, a first-degree relative with CRC), multiplying by its respective coefficient, summing the products (which will yield a log odds value between -1 and +1), and exponentiating that value, which provides the odds of the outcome for an individual. The odds may be converted to a probability using the formula: $p(\text{probability}) = \text{odds} / (\text{odds} + 1)$, and provides the probability (or risk) of the outcome for that individual based on parameter estimates from the entire sample.

As an alternative to providing an odds or probability, the regression coefficients may be used to create a risk score for each individual. Each independent variable's contribution to the risk score depends on the value of its coefficient.¹⁸ The coefficients can be rescaled, using a clinically meaningful metric such as the risk associated with a 5-year age interval, to produce a risk score that has a manageable point range (e.g., 0 to 10 points instead of 0 to 158 points).¹⁸ For a risk prediction from our group,¹⁹ we developed a risk score by re-scaling the log odds coefficients. Consistent with the Framingham approach, we prefer to derive risk scores using log odds coefficients instead of odds ratios because the former are symmetric around their null value of 0, whereas the latter are asymmetric around their null value of 1.¹⁸

Nearly all of the "CRC" risk prediction models use current risk of AN as the outcome (present/absent); only a few predict a future risk of CRC.^{20,21} Peng and colleagues recently reviewed risk scores for predicting AN, finding 22 studies of 17 original risk scores that include a median number of 5 independent risk factors. The models most commonly include age, sex, BMI, cigarette smoking, and a first degree relative with CRC. The AUROCs ranged from 0.62 to 0.77, representing fair-to-good discrimination.²²

There are fewer risk prediction models for estimating future risk of CRC because CRC is a much less common outcome than is AN, making model development more challenging. The National Cancer Institute's Risk Assessment Tool for Colorectal Cancer (<http://www.cancer.gov/colorectalcancerrisk/>) surveys the user for demographic features, personal and family medical history, diet and lifestyle

information, and OTC medication use. The risk assessment tool estimates 5-year and lifetime risks of CRC, and compares them with the average risk for the same age, gender, and race/ethnicity. In addition to future CRC risk, the model also estimates current risk for AN,^{21,23} demonstrating its versatility in predicting two outcomes in two different timeframes.

There are other risk prediction models that use logistic regression to predict future CRC risk. Using cohort data from the Physicians' Health Study, Driver and colleagues developed a model predicting 20-year risk of CRC among male physicians, identifying age by decade, smoking history, alcohol use, and BMI as independent risk factors.²⁴ Using the odds ratios from these variables, the investigators created a risk score that ranged from 0 to 10, collapsing the scores to form three risk groups: low-risk (scores of 0-3); intermediate risk (scores of 4-6), and high-risk (scores of 7-10), in which the respective 20-year risks of CRC were 1%, 3%, and 6%. The model demonstrated good calibration (goodness-of-fit P-value of 0.91) and fair discrimination (c-statistic by bootstrap validation of 0.69). However, the clinical importance of the risk separation is questionable. Further, the generalizability of the model and risk score – beyond U.S. male physicians – is uncertain. A recent model was created to determine the starting age for CRC screening. Jeon and colleagues²⁵ used two large case-control studies, the GRECC Consortium and the Colorectal Transdisciplinary study with nearly 10,000 CRC cases and just over 10,000 controls, to estimate whether a person's 10-year CRC risk exceeded a threshold of 1%, which was based on the average 10-year risk for 50-year old men and women combined. Four models were created: Model I used family history alone; Model II used family history and an "E" score, the latter measuring risk based on 19 environmental (i.e., phenotypic) factors; Model III used family history and a "G" score, which measured genetic risk based 63 CRC-associated SNPs; and Model IV used family history, "E" score and "G" score combined. Model metrics indicated that adding either "E" or "G" score to family history provided fair discrimination, with little improvement shown by Model IV, the combined model. Interestingly, the "E" and "G" components appeared to result in the same degree of risk prediction.

Classification and Regression Trees (CART)

Introduced in 1984, CART was first used in the clinical setting to identify patients at high risk of myocardial infarction within the first 24 hours of hospitalization. CART is a method of risk stratification that

uses decision tree algorithms for classification or regression predictive modeling. The tree is constructed through a method known as binary recursive partitioning, an iterative process that splits the data into branches with the goals of: 1) identifying which factors are most important in a model predicting the dependent (or “target”) outcome, and; 2) dividing the study population or dataset into smaller and more homogeneous subgroups. Classification trees are used when the target variable is categorical (race/ethnicity, gender, marital status, eye color, or findings on colonoscopy such as no neoplasia, non-advanced neoplasia, and AN) or binary (CRC or AN present / absent), while regression trees are used when the outcome or target variable is continuous (age, height, or BMI, for example). The product is an algorithm in the form of a multi-level tree that explains the relationship among variables and categorizes subsequent data into specific classes or subgroups.

A generic illustration of CART output is shown in Figure 3. The tree begins with the entire study population represented by a single group. This single group is known as a parent “node” because it gives rise to two child nodes based on the independent variable that provides the greatest difference with respect to the target variable. CART allows for examination of all candidate predictor variables at each level of the algorithm/tree to identify the one that results in the cleanest split or separation based on the algorithm learned by the machine. Each parent node splits into only two child nodes. Each of the remaining independent variables is examined to determine which one now results in the largest difference (or “best split”) based on pre-determined outcome variable, with each of the child nodes potentially becoming parent nodes for the two child nodes that result from the second independent variable. With CART, the question asked at each step of the algorithm is based on the answer to the previous question. This procedure continues for each branch / node of the tree until criteria for a pre-determined stopping rule are reached. A parameter known as the “complexity parameter” (CP) determines the number of splits in a tree by defining the minimum benefit that must be gained at each level (or split) to make that split worthwhile. The CP eliminates splits that add little or no value to the tree and, in doing so, provides a stopping rule for the tree. At this point, no further splitting occurs and each of the last nodes becomes a terminal node, representing mutually exclusive and exhaustive subgroups of the study population.

Advantages of CART include: 1) it is a non-parametric method of model building, which means that it requires minimal underlying assumptions; 2) it categorizes independent variables, and thus, can handle skewness without requiring transformation of variables to reduce overly influential observations; 3) it is more of an automatic “ground-up” learning method that does not require much analyst input (e.g. cut points for continuous independent variable are discovered from the data) and; 4) it is easy for users to interpret.

For CRC risk prediction, there is at least one CART-based study worth mentioning. Shi and colleagues used CART on regular health examination data for early detection of CRC, based on analysis of more than 7,000 CRC cases and more than 140,000 controls without known CRC.²⁶ The investigators validated the derived CART model on independent datasets, and compared its effectiveness in CRC detection with FIT-based screening. The CART model included four variables that were dichotomized: albumin (≥ 44 g/L vs. < 44 g/L), % lymphocytes ($\geq 16\%$ vs. $< 16\%$), age (≥ 70 years vs < 70 years), and hematocrit ($\geq 36\%$ vs $< 36\%$). At a specificity of 99%, model sensitivity for CRC was 62.2%, with these test characteristics remaining fairly constant within subgroups of the test set having different CRC prevalence, aging rates, gender ratios, and distribution of cancer stages and locations. CART-base screening had a positive predictive value of 1.6%, which was higher than FIT (0.3%). This model and setting illustrate the potential of CART for risk stratification for CRC screening. CART-generating programs are available in both well-known commercial statistical computing packages such as SPSS, SAS, and STATA, and in open-source programs, most notably R.^{27,28}

Neural Networks and Deep Learning

The basic structure of a neural network is shown in Figure 4, with each circle in the diagram representing a node, which some consider analogous to a neuron. A neural network is organized into three layers: an input layer, which consists of the independent (predictor) variables; an output layer, which generates a response to the input in the form of a dependent (predicted) variable; and a hidden layer, which connects the inner and outer layers based on intermediate values that are generated by the network. Hidden layers can be demystified by thinking of them as latent variables that are derived from weighted combinations of input values.²⁹ A neural network with many hidden layers is referred to as a deep neural network. Hidden

nodes enable the modeling of complex, non-linear relationships between predictor variables and outcome. Each input node is connected to each hidden node by connection weights, which represent the neural network counterpart of β coefficients in a regression model. These weights, like weights from other models, contain information about the strength of the relationship between input, hidden layers, and outcome.

Most neural networks “learn” through the process of backpropagation, short for “backward propagation of errors.” The name is derived from how the error in a neural network system becomes minimized.

Backpropagation uses the error associated with an incorrect guess to adjust its parameters towards the endpoint of minimizing error. In a multi-level (or deep) neural network, this process occurs sequentially at each level, in which the relationship between the network’s error and parameters of its last layer are optimized, then the process repeats between the last layer and second to last layer, and so forth.

A couple of examples of neural networks within the domain of CRC screening deserve mention, one involving classification of (diminutive) colorectal polyps using computer-aided diagnosis, and the other involving risk prediction for CRC using self-reportable personal health data. Both examples use artificial neural networks to accomplish these goals. Chen and colleagues used images of diminutive neoplastic polyps and hyperplastic polyps to train a deep neural network, with histology used as the reference standard.³⁰ The network was then tested on a separate set of images, and the results were compared to 2 expert and 4 novice endoscopists’ assessments. The network identified neoplastic or hyperplastic polyps with 96.3% sensitivity and 78.1% specificity, and negative predictive value of 91.5%, the latter measure exceeding the PIVI threshold of 90%. The neural network outperformed the endoscopists, among whom both intra- and inter-observer agreement was low. In the second example, Nartowt and colleagues used data from the National Health Interview Survey to estimate CRC risk, constructing a neural network trained on 12 to 14 categories of personal health data from 1269 cases of CRC and over 500,000 controls.³¹ Input variables included age, sex, race/ethnicity, BMI, smoking frequency, first-degree relatives with CRC, exercise frequency, and several comorbid conditions, while the output was a number between 0 and 1, and was treated as a risk score. When tested on an independent sample of 140 CRC cases and more than 58,000 controls, the neural network had a negative predictive value of 0.999. When

compared with the USPSTF guidelines to stratify subjects into low-, intermediate-, and high-risk categories, the network outperformed the guidelines in subjects misclassified as low risk (from 35% to 5%) and misclassified as high-risk (from 53% to 6%). These results exemplify the clinical and public health utility of neural networks for stratifying CRC risk, with the potential to improve the efficiency, effectiveness, and cost-effectiveness of screening.

Artificial Intelligence (AI)

We have already discussed some of the tools of AI. Until relatively recently, AI conjured thoughts of robots, IBM's Watson supercomputer, smart speakers, facial recognition, and others. AI's most active area within Gastroenterology and specifically relevant to CRC screening has been for both polyp detection during colonoscopy³² and real-time histological determination.^{33,34} This technology is approved for use in Europe and is under study in the U.S.

In health care and other sectors, AI refers to the collection of concepts and technologies that facilitate the simulation of human intelligence processes (such as problem solving, learning, and pattern recognition) by machines, particularly computer systems. The technologies in AI's toolbox include natural language processing, artificial neural networks, data mining, and modeling methods that fit linear and non-linear relationships. The utility of these techniques has been amplified by access to large databases and enhanced computational ability. Deep neural networks has emerged as a popular method underlying AI, as it is well-suited to discovering complex nonlinear relationships; however, because many parameters are estimated, it is well suited for large samples sizes and is often outperformed by parametric logistic regression in cross-validation analysis when the underlying population data contains very few nonlinear relationships.

While the potential of AI in medicine appears limitless, its greatest utility to date has been in the area of image analysis – discriminating between basal cell carcinoma and benign lesions, identifying retinal findings of diabetic retinopathy, and colorectal polyp recognition³² and real-time histologic diagnosis³³, among many others. Image analysis and other data sets that provide many observations are ideal for nonparametric methods that involve estimating many coefficients (such as neural networks) because

greater numbers of observations are needed to reliably estimate models with large numbers of coefficients.

A MEDLINE search intersecting the terms “colorectal cancer screening,” “risk prediction,” and “artificial intelligence” yielded 12 references, 3 of which are directly relevant to this discussion. The three studies report on the same model, which was used to identify patients who may have had CRC based on specific factors identified from examination of routinely collected clinical data from large numbers of patients with and without CRC. Kinar and colleagues used a method known as random forest modeling³⁵, which combines CART with ensemble learning by aggregating results from multiple decision trees obtained from bootstrap-derived repeat random re-samples of the original data set.¹⁶ They used a dataset of more than 600,000 Israeli patients, 3,135 of whom had a CRC diagnosis, deriving the model on 80% of the sample and validating it on both the remaining 20% and an entirely separate dataset from the UK that included 5,061 CRC cases and 25,613 controls. This final model was then tested completely independently on an independent dataset for external validation. The criteria by which model performance was evaluated included: 1) AUROC for overall performance; 2) determination of the odds of CRC above versus below a threshold corresponding to a specificity of 99.5%, and; 3) determination of the proportion of controls correctly classified as such (i.e., specificity) at the model threshold corresponding to a sensitivity of 50%.

The model identified age and variables from a CBC (e.g., hemoglobin, MCV, RDW, platelet count, others) obtained 3-6 months before the CRC diagnosis as the most discriminating features. The model’s AUROC was 0.82, while the odds ratio for CRC was 26 at the threshold required for a false positive rate of 0.5% (or specificity of 99.5%). At a CRC sensitivity of 50%, model specificity was 88%. Based on these performance metrics, the authors concluded that the model may help to detect CRC earlier in clinical practice by supplementing detection through screening. While this review is not the forum for critical evaluation of the article, a couple of limitations are worth discussing. One is that the stage spectrum of CRCs in the Israeli and UK cohorts is not described. The model would be of greater clinical utility if most or all of the cancers detectable were of a curable stage (0, I, II). Further, the 3-6 month interval between CBC and CRC diagnosis is short and may have discrimination only because of such close proximity to the

cancer diagnosis. The short interval might not allow the opportunity for intervention that favorably affects prognosis compared to the opportunity provided by a longer interval.

In the second article identified, which addressed the time interval issue, Birks and colleagues tested the model using the UK's Clinical Practice Research Datalink³⁶, from which patients 40 to 89 years with full CBC data were risk stratified by the model and followed for a CRC diagnosis. In this retrospective cohort study, the investigators identified CBCs at least 18 months prior to the CRC diagnosis for cases and prior to the end date of follow-up for controls. The CBCs obtained during the 18-24 month interval were considered as the primary analysis, with secondary analyses consisting of intervals of 3-6 months, 6-12 months, 12-18 months and 24-36 months prior to the CRC diagnosis. For the primary analysis, which included 5,141 CRC cases and 2,220,108 controls, AUROC was 0.776 (95% CI, 0.771-0.781). At a specificity of 99.5% (i.e., a false positive rate of 0.5%), the positive predictive value was 8.8%. As expected, model performance improved to an AUROC of 0.84 for the 3-6 month interval, validating the model developed by Kinar and colleagues. Most of the model discrimination was due to age, as evidenced by an age-matched case-control analysis for which AUROC was 0.583 (CI, 0.574-0.591), indicating poor discrimination of the residual factors. The investigators concluded that this model, with an 18-24 month lead-time, offered an additional way to identify patients with CRC. The CRC stage distribution was not described in this study.

The third article is another evaluation of the same model utilizing data from Kaiser Permanente's Northwest Region.³⁷ Hornbrook and colleagues identified a random sample of 900 CRC cases and 18 controls per case (age range 40-89 years for both groups) selected from the same year of the case's diagnosis and matched on the general population's distribution of 10-year age groups and length of enrollment. Intervals of 0-180 days and 181-360 days between CBC date and CRC diagnosis were considered, with respective ORs, at 99% specificity, of 34.7 and 20.4 and an overall AUROC of 0.80. Individuals in the highest 1% of scores had a 20-fold greater CRC risk within the subsequent 12-18 months. The AUROC for age alone was 0.73, supporting the dominance of age in the model. Among the 605 (67%) cases for which CRC stage was available, ORs at a specificity of 99% ranged from 12.1 for *in situ* cancers to 40.4 for stage IV. ORs were higher for CRCs in the proximal colon, although a more

advanced stage distribution may account for this finding. These U.S. findings are consistent with the other two studies, and suggest that routinely collected electronic health data may be used to supplement screening in the early identification of CRC.

Cumulatively, these articles illustrate the data mining and computational capacities and potential of current technology for identifying patterns in large, complex datasets and for deriving and testing multivariable models to identify how these patterns provide diagnostic discrimination. Given the persuasive nature of the computational capacity of AI methods, it behooves the reader / user to determine whether the underlying study methods and results are reliable and valid, what the results mean both quantitatively and clinically, and whether and to what extent the model can be applied to patients in other settings. Liu and colleagues recently published a user's guide for understanding articles that use machine learning.³⁸ It provides step-by-step guidance for evaluating this body of literature with a critical eye.³⁹

BARRIERS TO USE OF RISK PREDICTION MODELS

Despite the proliferation of risk prediction models in medicine in general, and the several models available for prediction of current risk for AN and future CRC risk, the great majority of models are not used in clinical practice. Several factors negatively impact their use, foremost among which are: 1) absence of robust validation, and; 2) lack of an impact analysis. Many models have had no "external" (i.e., independent) validation. A hierarchy of increasing stringent validation includes temporal, geographic, and domain validation.⁴⁰ Temporal validation tests a prediction model at a different time, but is usually done by the same investigators at the same institution and with the same population. Geographic validation varies time and place, and is carried out in a population defined by the setting and inclusion and exclusion criteria of the development study. The most rigorous test of a model, domain validation, varies time, place, setting, and patient spectrum, including demographics. The degree of validation required of a risk prediction model depends on the setting for its intended use. Most models with some degree of external validation nearly always include temporal and geographic variation, much less so, domain variation.

Impact analysis of a prediction rule requires its testing in a clinical trial, and would typically examine several outcomes, including clinical, economic, and satisfaction. Very few prediction rules in any area has

had an impact analysis. For CRC risk prediction, Schroy and colleagues tested a decision aid alone or with a risk assessment tool for advanced neoplasia in 341 asymptomatic average-risk patients, and found no differences in test concordance between patients' preference and test ordered.⁴¹ But this is just one study in a single setting, with the outcome of test concordance, one intervention (a decision aid plus a risk assessment tool), and one significant comparator (the decision aid itself) that likely diluted the effect of the risk assessment tool itself. We need many more impact studies that consider the setting and format for how a risk prediction model should be tested and applied.

Several other reasons contribute to the non-use of risk prediction models, one of which is providers' understanding of these models, particularly their output. This barrier may be overcome with interface software that facilitates understanding and perhaps links to a preferred strategy or, in the case of CRC screening, to a preferred test for a particular level of risk. A related reason is provider trust of the model. To many providers, risk prediction models invoke a "black box" response, affecting their level of confidence in the model itself and uncertainty about the methodological quality of the research that belies it. An extension of this reason is provider fear of litigation. To assuage this concern, those models with consistent performance and high degrees of validity and generalizability should be supported by guideline organizations. Institutional factors also contribute to non-use, specifically the extent of support for and integration into the workflow of prediction models. How should these tools integrate into the electronic medical record? Should they instead be available through patient portals for patients to consider prior to seeing a provider? The answers to these questions depend both on the prediction model being considered and the goals for its use. Related to these issues, the opportunities and challenges in moving from current guidelines to personalized CRC screening have been nicely described by Robertson and Ladabaum.⁴²

AGENDA FOR RESEARCH

Despite the proliferation and promise of risk prediction models – including those for CRC screening – questions exist about whether, when, and how to use them, and whether they will improve the uptake and efficiency of screening and, most importantly, outcomes of CRC incidence, morbidity, mortality. The most important items on the research agenda are studies assessing the extent of validation of risk prediction

models, and studies assessing their impact on processes and outcomes of care. Use of large datasets may be especially helpful with validation studies, as these could be applied retrospectively if the proper factors are identifiable in those datasets. Impact analyses will require either randomized trials or quasi-randomized trials, such as before-after studies. Other research agenda items include understanding provider and patient attitudes towards risk prediction models and how these tools are best integrated into health care system. It is likely that whether and how risk prediction models are used and which ones are used will depend on the particular health care system, requiring research at the systems level.

REFERENCES:

1. Doyle J, Kannel W. Coronary risk factors: 10 year findings in 7446 Americans, Pooling Project. VI World Congress of Cardiology; 1970; London, England.
2. Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. *Journal of Clinical Epidemiology*. 1967;20(7):511-524.
3. MDCalc. Framingham Risk Score for Hard Coronary Heart Disease: Estimates 10-year risk of heart attack. Published 2005. Updated 2019. Accessed December 5, 2019.
4. Blatchford O, Murray WR, Blatchford M. A risk score to predict need for treatment for uppergastrointestinal haemorrhage. *The Lancet*. 2000;356(9238):1318-1321.
5. Rockall T, Logan R, Devlin H, Northfield T. Risk assessment after acute upper gastrointestinal haemorrhage. *Gut*. 1996;38(3):316-321.
6. Larvin M, McMahan M. APACHE-II score for assessment and monitoring of acute pancreatitis. *The Lancet*. 1989;334(8656):201-205.
7. Wu BU, Johannes RS, Sun X, Tabak Y, Conwell DL, Banks PA. The early prediction of mortality in acute pancreatitis: a large population-based study. *Gut*. 2008;57(12):1698-1703.
8. Pugh R, Murray-Lyon I, Dawson J, Pietroni M, Williams R. Transection of the oesophagus for bleeding oesophageal varices. *British Journal of Surgery*. 1973;60(8):646-649.
9. Said A, Williams J, Holden J, et al. Model for end stage liver disease score predicts mortality across a broad spectrum of liver disease. *Journal of hepatology*. 2004;40(6):897-903.
10. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute*. 1989;81(24):1879-1886.
11. Wolf AM, Fontham ET, Church TR, et al. Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. *CA: a cancer journal for clinicians*. 2018;68(4):250-281.
12. Johnston S, Wilson K, Varker H, et al. Real-world Direct Health Care Costs for Metastatic Colorectal Cancer Patients Treated With Cetuximab or Bevacizumab-containing Regimens in First-line or First-line Through Second-line Therapy. *Clinical colorectal cancer*. 2017;16(4):386-396. e381.
13. Subramanian S, Tangka FK, Hoover S, Royalty J, DeGroff A, Joseph D. Costs of colorectal cancer screening provision in CDC's Colorectal Cancer Control Program: comparisons of colonoscopy and FOBT/FIT based screening. *Evaluation and program planning*. 2017;62:73-80.
14. Fletcher RH. Personalized screening for colorectal cancer. *Medical care*. 2008;46(9):S5-S9.
15. Control DoCPa. What Are the Risk Factors for Colorectal Cancer? Colorectal (Colon) Cancer Web site. Published 2019. Updated January 30, 2019. Accessed December 5, 2019.
16. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media; 2009.
17. Cheng X, Khomtchouk B, Matloff N, Mohanty P. Polynomial regression as an alternative to neural nets. *arXiv preprint arXiv:180606850*. 2018.
18. Sullivan LM, Massaro JM, D'Agostino RB, Sr. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat Med*. 2004;23(10):1631-1660.
19. Imperiale TF, Monahan PO, Stump TE, Glowinski EA, Ransohoff DF. Derivation and Validation of a Scoring System to Stratify Risk for Advanced Colorectal Neoplasia in Asymptomatic Adults: A Cross-sectional Study. *Ann Intern Med*. 2015;163(5):339-346.

20. Usher-Smith JA, Walter FM, Emery JD, Win AK, Griffin SJ. Risk Prediction Models for Colorectal Cancer: A Systematic Review. *Cancer Prev Res (Phila)*. 2016;9(1):13-26.
21. Ma GK, Ladabaum U. Personalizing colorectal cancer screening: a systematic review of models to predict risk of colorectal neoplasia. *Clin Gastroenterol Hepatol*. 2014;12(10):1624-1634.e1621.
22. Peng L, Weigl K, Boakye D, Brenner H. Risk Scores for Predicting Advanced Colorectal Neoplasia in the Average-risk Population: A Systematic Review and Meta-analysis. *Am J Gastroenterol*. 2018;113(12):1788-1800.
23. Imperiale TF, Yu M, Monahan PO, et al. Risk of Advanced Neoplasia Using the National Cancer Institute's Colorectal Cancer Risk Assessment Tool. *J Natl Cancer Inst*. 2017;109(1).
24. Driver JA, Gaziano JM, Gelber RP, Lee IM, Buring JE, Kurth T. Development of a risk score for colorectal cancer in men. *Am J Med*. 2007;120(3):257-263.
25. Jeon J, Du M, Schoen RE, et al. Determining Risk of Colorectal Cancer and Starting Age of Screening Based on Lifestyle, Environmental, and Genetic Factors. *Gastroenterology*. 2018;154(8):2152-2164.e2119.
26. Shi Q, Gao Z, Wu P, et al. An enrichment model using regular health examination data for early detection of colorectal cancer. *Chin J Cancer Res*. 2019;31(4):686-698.
27. Kuhn L, Page K, Ward J, Worrall-Carter L. The process and utility of classification and regression tree methodology in nursing research. *J Adv Nurs*. 2014;70(6):1276-1286.
28. Venables W, Smith D. R Development Core Team (2010). An Introduction to R. R Foundation for Statistical Computing, Vienna, Austria. In.
29. Papik K, Molnar B, Schaefer R, Dombovari Z, Tulassay Z, Feher J. Application of neural networks in medicine-a review. *Medical Science Monitor*. 1998;4(3):MT538-MT546.
30. Chen PJ, Lin MC, Lai MJ, Lin JC, Lu HH, Tseng VS. Accurate Classification of Diminutive Colorectal Polyps Using Computer-Aided Analysis. *Gastroenterology*. 2018;154(3):568-575.
31. Nartowt BJ, Hart GR, Roffman DA, et al. Scoring colorectal cancer risk with an artificial neural network based on self-reportable personal health data. *PLoS One*. 2019;14(8):e0221421.
32. Misawa M, Kudo SE, Mori Y, et al. Artificial Intelligence-Assisted Polyp Detection for Colonoscopy: Initial Experience. *Gastroenterology*. 2018;154(8):2027-2029.e2023.
33. Byrne MF, Chapados N, Soudan F, et al. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. *Gut*. 2019;68(1):94-100.
34. Urban G, Tripathi P, Alkayali T, et al. Deep Learning Localizes and Identifies Polyps in Real Time With 96% Accuracy in Screening Colonoscopy. *Gastroenterology*. 2018;155(4):1069-1078.e1068.
35. Kinar Y, Kalkstein N, Akiva P, et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. *J Am Med Inform Assoc*. 2016;23(5):879-890.
36. Birks J, Bankhead C, Holt TA, Fuller A, Patnick J. Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records. *Cancer Med*. 2017;6(10):2453-2460.
37. Hornbrook MC, Goshen R, Choman E, et al. Early Colorectal Cancer Detected by Machine Learning Model Using Gender, Age, and Complete Blood Count Data. *Dig Dis Sci*. 2017;62(10):2719-2727.
38. Liu Y, Chen PC, Krause J, Peng L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *Jama*. 2019;322(18):1806-1816.
39. Doshi-Velez F, Perlis RH. Evaluating Machine Learning Articles. *Jama*. 2019;322(18):1777-1779.
40. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol*. 2008;61(11):1085-1094.

41. Schroy PC, 3rd, Duhovic E, Chen CA, et al. Risk Stratification and Shared Decision Making for Colorectal Cancer Screening: A Randomized Controlled Trial. *Med Decis Making*. 2016;36(4):526-535.
42. Robertson DJ, Ladabaum U. Opportunities and Challenges in Moving From Current Guidelines to Personalized Colorectal Cancer Screening. *Gastroenterology*. 2019;156(4):904-917.