frontiers in Big Data

ORIGINAL RESEARCH
published: 10 March 2020
doi: 10.3389/fdata.2020.00006

Robust Machine Learning for Colorectal Cancer Risk Prediction and Stratification

Bradley J. Nartowt[1], Gregory R. Hart[1], Wazir Muhammad[1], Ying Liang[2], Gigi F. Stark[3] and Jun Deng[1]*

[1] Department of Therapeutic Radiology, Yale University, New Haven, CT, United States, [2] Department of Radiation Oncology, Medial College of Wisconsin, Milwaukee, WI, United States, [3] Department of Statistics & Data Science, Yale University, New Haven, CT, United States

While colorectal cancer (CRC) is third in prevalence and mortality among cancers in the United States, there is no effective method to screen the general public for CRC risk. In this study, to identify an effective mass screening method for CRC risk, we evaluated seven supervised machine learning algorithms: linear discriminant analysis, support vector machine, naive Bayes, decision tree, random forest, logistic regression, and artificial neural network. Models were trained and cross-tested with the National Health Interview Survey (NHIS) and the Prostate, Lung, Colorectal, Ovarian Cancer Screening (PLCO) datasets. Six imputation methods were used to handle missing data: mean, Gaussian, Lorentzian, one-hot encoding, Gaussian expectation-maximization, and listwise deletion. Among all of the model configurations and imputation method combinations, the artificial neural network with expectation-maximization imputation emerged as the best, having a concordance of 0.70 ± 0.02, sensitivity of 0.63 ± 0.06, and specificity of 0.82 ± 0.04. In stratifying CRC risk in the NHIS and PLCO datasets, only 2% of negative cases were misclassified as high risk and 6% of positive cases were misclassified as low risk. In modeling the CRC-free probability with Kaplan-Meier estimators, low-, medium-, and high CRC-risk groups have statistically-significant separation. Our results indicated that the trained artificial neural network can be used as an effective screening tool for early intervention and prevention of CRC in large populations.

Keywords: colorectal cancer, risk stratification, neural network, concordance, self-reportable health data, external validation

OPEN ACCESS

Edited by:
Katrina M. Waters,
Pacific Northwest National Laboratory (DOE), United States

Reviewed by:
Maria Chan,
Memorial Sloan Kettering Cancer Center, United States
Avinash Parnandi,
Langone Medical Center, New York University, United States

****Correspondence:***
Jun Deng
jun.deng@yale.edu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *10 October 2019*
Accepted: *31 January 2020*
Published: *10 March 2020*

Citation:
Nartowt BJ, Hart GR, Muhammad W, Liang Y, Stark GF and Deng J (2020) Robust Machine Learning for Colorectal Cancer Risk Prediction and Stratification. Front. Big Data 3:6. doi: 10.3389/fdata.2020.00006

INTRODUCTION

Of all new cancer incidences in the United States, 8.1% are colorectal cancer (CRC) (Falco et al., 2018; National Cancer Institute, 2018). The 5-year survival rate for CRC ranges from 14% for a distant stage to 90% for a localized stage. CRC is responsible for 8.3% of all cancer deaths, and is especially deadly and recurrent when coincident with diabetes and hypertension (Yang et al., 2012). However, there exists little knowledge of the primary causes of CRC. Thus, current screening recommendations are only based on family history of CRC and age. Specifically, the United States Preventative Services Task Force (USPSTF) recommends screening for individuals between ages 50 and 75 while the American Cancer Society recommends screening for individuals between ages 45 and 75 (Collins et al., 2015; Bibbins-Domingo et al., 2016). Both guidelines recommend screening for anyone with one or more primary relatives who have ever had CRC. While screening according to these guidelines indisputably saves lives, high-risk individuals with no CRC family history and/or aged 18–49 would clearly benefit from a model that better detects their risk. Low-risk individuals that are flagged for screening under a new model (Collins et al., 2015; Bibbins-Domingo et al., 2016), would also be given information to help them choose whether they want to be subject to invasive, expensive, and injurious (Benard et al., 2018; National Cancer Institute, 2019) screening. Hence, it is important to develop an effective method to estimate CRC risk non-invasively and cost-effectively.

There have been a lot of previously-developed CRC-risk models that do not involve biomarkers (Usher-Smith et al., 2016). Using only professionally-collected routine data (biological sex, use of non-steroidal anti-inflammatory drugs (NSAIDs), form of recruitment, non-specific abdominal pain, bowel-habit, age, BMI, cholesterol, and triglycerides), Betes et al. achieved a concordance ∼0.7 using a multiple logistic regression model (Betes et al., 2003). Using data from a self-completed questionnaire (asking about CRC in first-degree relatives, BMI, screening, NSAID use, diet, inflammatory bowel disease, alcohol/tobacco use, and physical activity), Colditz et al. also built a multiple logistic regression model of similar concordance ∼0.7 from data on family history, obesity, screening, diet (multivitamin, alcohol, vegetables, and red meat consumption), height, physical activity, pharmaceuticals (prophylactic, post-menopausal hormone, and aspirin use), and inflammatory bowel disease (Colditz et al., 2000). Both models are externally tested, i.e., the model is built from one dataset and its performance is reported on a dataset from a separate study (Collins et al., 2015). However, compared to the simple logistic regression models, there has been no systematic study on the development of more advanced machine learning models for CRC risk prediction and stratification for a large population, in consideration of various imputation methods.

Hence in this work, we aim to identify an effective mass screening method for CRC risk based solely on personal health data. We trained and cross-tested various machine learning models with two large national databases, reporting performance in terms of the concordance, a performance metric that is biased but standard (Hanley and McNeil, 1982; Hosmer and Lemeshow, 2000; Fawcett, 2005; Hajian-Tilaki, 2013). A variety of imputation methods were explored in handling the missing data. Additionally, a component of cross-uncertainty is incorporated to the total uncertainty reported, adding stringency to our testing that to our knowledge has not been used before. Finally, we furnish some ideas on how our model can be deployed for real world applications.

MATERIALS AND METHODS

Two Datasets From Separate Studies

The National Health Interview Survey (NHIS) dataset[1] is a cross-sectional study of the overall health status of the United States. Each year, roughly 30,000 adults are interviewed on a range of current and past personal health conditions. The first survey of the NHIS after a significant revision was administered in 1997 and the next such redesign of the NHIS is scheduled to appear in 2019, so data from years 1997 to 2017 was used. Our other study is the longitudinal Pancreatic, Lung, Colorectal, Ovarian (PLCO) Cancer Screening dataset from the National Cancer Institute[2]. The PLCO dataset is a randomized, controlled longitudinal study on the efficacy of screening for prostate, lung, colorectal, and ovarian cancer. Between November 1993 and July 2001, participants were randomized, entered into the trial, answered a baseline questionnaire (BQ), and were followed for up to 14 years, exiting the trial early if they were diagnosed with any cancer or if they died. To match this PLCO data with the NHIS dataset, we assumed that answering the PLCO BQ was equivalent to participating in the NHIS's interview.

Data was marked by 7 for responses of "Refused," 8 for "Not ascertained," and 9 "Don't know" in the NHIS[1]; all these responses were assumed to indicate data missing completely at random (Little and Rubin, 2014) (MCAR). This is distinguished from data not missing at random, which is marked by the table entry being actually blank (e.g., all pregnancy data has a blank entry for male respondents). PLCO uses the same scheme of marking the missingness of data with digit-entries, while data missing not at random is actually blank.

The United States Preventative Services Task Force guidelines currently recommend anyone with family history of CRC and/or aged 50–75 years for screening (Bibbins-Domingo et al., 2016), while screening at ages 76+ is up to the individual. Thus, ages 18–49 and ages 50–75 form sub-demographics of data that are of interest. To assess performance in these sub-demographics, we trained and tested models on these age splits of the data as well as on all ages.

There are factors appearing in the NHIS dataset but missing in the PLCO dataset, and vice versa. Specifically, factors appearing in the NHIS but not in the PLCO are alcohol-use, vigorous exercise frequency, functional limitations, kidney comorbidity, and incidence of angina. Factors appearing in the PLCO but not in the NHIS are non-steroidal anti-inflammatory drug (NSAID) use, gallbladder inflammation, and incidence of diverticulitis. To ensure a TRIPOD level 3 cross-testing between separate datasets, and all its rigor and robustness, all these factors are not used in our study.

CRC vs. Never Cancer

The NHIS records each respondent's age at the time of the survey, and the age(s) at which the respondent was diagnosed with cancer of the colon and/or rectum, if at all. Respondents were counted as positive cases of CRC if their diagnosis happened <4 years prior

[1] National Health Interview Survey (NHIS) (1997–2017).

[2] Prostate/Lung/Colorectal/Ovarian (PLCO) Cancer Screening Trial (1993–2001).

to the survey. In each study, a small fraction of the respondents were recently diagnosed with CRC. We considered CRC in survey respondents ages 18–85.

In PLCO and NHIS, the following four types of respondents were discarded: (1) those diagnosed with CRC more than 4 years prior to taking the survey (NHIS) or answering the questionnaire (PLCO), (2) those non-CRC respondents diagnosed with any other cancer at any time, (3) those CRC respondents diagnosed with a cancer other than CRC at a time before their CRC diagnosis, and (4) those CRC respondents having CRC at a time before randomization (PLCO only). Members of the first group were discarded because their reported personal health data was considered irrelevant to their CRC diagnosis. Those in the second and third groups were discarded because those diagnosed with any cancer already receive heightened screening attention, defeating the purpose of assessing their risk. The fourth group is discarded because before randomization the time (in days) of CRC diagnosis is not known. Thus, the negative examples were those who were never diagnosed with any cancer and are referred to as "never-cancer" (NC) while the positive examples were those recently diagnosed with CRC and are referred to as "CRC."

To be considered a positive case of CRC in the PLCO data remaining after the deletion described above, the respondent needed to meet both of the following conditions: (1) they were diagnosed with CRC within 4 years of the BQ and (2) CRC was the first cancer they had. If both of these conditions were not met, the respondent was considered part of the non-cancer population remaining after the above discarding was carried out. Hence, the outcome variable used in both datasets was the respondent's cancer status coded to a 0 or a 1. A value of 0 indicated that the respondent was never diagnosed with cancer (CRC or any other cancer). It is assumed that a given respondent would have already been flagged for screening if previously diagnosed with any kind of cancer, defeating the purpose of risk-scoring. A value of 1 indicated that the respondent was diagnosed with CRC within four (4) years of answering either the PLCO BQ or the NHIS interview questions. All respondents who fit neither of these criteria were assumed to be data not missing at random (Little and Rubin, 2014), and thus discarded (never subject to any imputation methods).

Performance after training with such an outcome variable is not relative to the sensitivity and specificity of any gold standard. In our work, the gold standard of CRC diagnosis is colonoscopy. Unfortunately, colonoscopy data is missing not at random for a significant portion of data. Specifically, only NHIS questionnaires from years 2000, 2005, 2010, and 2015 asked the respondent if they had ever been screened by the gold standard (sigmoidoscopy, colonoscopy, or proctoscopy). We therefore assumed that neither dataset contained any false positive or false negative cases.

Data Preparation

For reproducibility, we describe how the raw data was mapped to the datasets used to train and test the machine learning algorithms (MLAs). The factors of ever having hypertension, ulcers, a stroke, any liver comorbidity, arthritis, bronchitis, coronary heart disease, myocardial infarction, and/or emphysema are binary variables and mapped to 0 for "no" and 1 for "yes." Diabetic status has one of three discrete values: not diabetic, pre-diabetic/borderline, and diabetic, respectively. These conditions were mapped to 0, 0.5, and 1, respectively. The age factor is continuous and equals the age at response to the NHIS or PLCO BQ for negative cases and the age at CRC diagnosis for positive cases. Body mass index (BMI) is likewise continuous. All such continuous factors were unitized to the interval [0, 1]. The sex factor is 0 for women and 1 for men. The variable of Hispanic ethnicity was given a value of 0 for a response of "Not Hispanic/Spanish origin" and 1 otherwise. The variable of race was set to 1 for responses of "Black/African American only," "American Indian only," "Other race," or "Multiple race," and 0 otherwise. The smoking status had a value of 1 for an everyday smoker, 0.66 for a some-day smoker, 0.33 for a former smoker, and 0 for a never smoker. The NHIS defines a "never smoker" as someone who has smoked 100 cigarettes or less over their entire lifetime, and a "former smoker" as a smoker who quit at least 6 months prior to the survey; this same definition was used to score PLCO respondents' smoking status using equivalent fields. The variable of family history represents the number of first-degree relatives who have had CRC, and was capped at 3. The family history variable values of 0, 1, 2, and 3 were mapped to 0, 0.33, 0.66, and 1, respectively.

The Levels of TRIPOD and the Cross-Testing Uncertainty

Below, we use the terms "training," "validation," and "testing" to describe increasingly-general model performances. Any portion of data designated as "training" is used to directly adjust the parameters of the model (e.g., by iterations of gradient-descent in the space of model parameters for an artificial neural network). Any portion of data designated as "validation" is not involved in direct adjustment of model parameters, but is used to stop further iterations of an algorithm based on whether overfitting happens (e.g., stopping iterations of gradient descent if the training fitting error is decreasing but the validation fitting error is increasing). Finally, any portion of data designated as "testing" is data used for neither training nor validation. In the literature, the term "validation" is sometimes used to describe what is actually testing, often by way of the term "cross-validation" (Picard and Cook, 1984). In this work, we use the term "cross-testing" to avoid any possible confusion.

We reported concordance, a performance metric, at level 3 of the hierarchy proposed by the Transparent Reporting of Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines (Collins et al., 2015). TRIPOD level 1a corresponds to testing upon the same dataset used for training (leaving any overfitting undetected). TRIPOD level 1b corresponds to n-fold cross-validation (Picard and Cook, 1984). TRIPOD levels 2a and 2b each correspond to a trained model tested upon or cross-tested between splits of the data involved neither in training nor overfitting-detection ("validation"). Level 2a corresponds to random splits of the data, accordingly yielding normally-distributed random error (Bertsekas and Tsitsiklis, 2008) in Equation (2). Level 2b corresponds to non-random splits of the data, yielding systematically-distributed error in Equation (2). TRIPOD level 3 is where a model trained by data from one study is tested upon or cross-tested between data from a separate study. TRIPOD level 4 corresponds to testing a published model on a separate dataset (**Appendix**).

Our model has a TRIPOD level of 3, as it was trained upon a dataset from a longitudinal study and tested on a dataset from a cross-sectional study and vice versa. Throughout this paper, cross-testing shall refer to training on NHIS/PLCO and testing upon PLCO/NHIS, respectively. In this case, the systematic error from Equation (2) arises from the distributional disparity [in the Bayesian perspective (Bertsekas and Tsitsiklis, 2008), where each data entry in the NHIS and PLCO is assumed to be drawn from separate probability distributions with unknown parameters] between the PLCO and NHIS datasets due to (among other things) the fact that the NHIS is cross-sectional while PLCO is longitudinal. Reporting performance at TRIPOD level 3 demonstrates generalizability of the model's predictive capacity.

Seven Machine Learning Algorithms

The MLAs used in this work are an artificial neural network (ANN), logistic regression (LR), naive Bayes (NB), decision tree (DT), random forest (RF), linear-kernel support-vector machine (SVM), and linear discriminant analysis (LDA) with automatic optimization of hyper-parameters (Fisher, 1936; Morgan and Sonquist, 1963; Rumelhart et al., 1986; Cortes and Vapnik, 1995; Hosmer and Lemeshow, 2000; Bertsekas and Tsitsiklis, 2008). The LR, NB, DT, SVM, and LDA MLAs were invoked, respectively, by the "fitglm," "fitcnb," "fitctree," "fitcsvm," and "fitcdiscr" MATLAB functions. The SVM, LDA, and DT MLAs yielded a CRC risk score using Platt scaling (Platt, 1999). The ANN used a previously developed in-house MATLAB code.

ANNs are a method of regression (Bertsekas and Tsitsiklis, 2008) as they determine function parameters that minimize fitting error using iterations of stochastic gradient descent in parameter space (Bishop, 2006; Andoni et al., 2014; Kingma and Ba, 2015) through a process called backpropagation (Rumelhart et al., 1986), and are similar to logistic regressions (LR) (Hosmer and Lemeshow, 2000). Specifically, an ANN with a logistic activation function and zero hidden layers is a logistic regression. With their hidden layers, ANNs model inter-factor coupling as logistic or hyperbolic-tangential probabilities; these probabilities are called the activation function of the ANN. ANNs that use logistic activation functions are multilinear generalizations of LRs.

The in-house MATLAB coded ANN has two hidden layers with logistic activation and deployed adaptive gradient descent via the "Adam" learning rate. It also uses both early stopping and automatic hyperparameter optimization to minimize overfitting. There is one input neuron for each factor used, and each hidden layer has one neuron for each input neuron. Each neuron is associated with a single weight W and a single bias B, which, respectively, are the slope and intercept for the linear function $z = z(X) = WX + B$ with argument X. The linear function itself is then fed into the neuron's sigmoidal activation function $\left(e^{-z} + 1\right)^{-1}$. The weights and biases are, respectively, determined by iterations of the equations $W' = W - \alpha \frac{dS}{dW}$ and $B' = B - \alpha \frac{dS}{dB}$ for fitting error $\mathrm{S} = N^{-1} \ln \prod_{i=1}^{N} \bar{Y}_i^{Y_i}(1 - \bar{Y}_i)^{1-Y_i}$ between the subject's risk-score and their actual cancer status in a total of N subjects, a process called backpropagation. In our backpropagation, we chose to iterate until $\left|W' - W\right|, \left|B' - B\right| \leq \varepsilon$ for a small ε we picked.

The NB method modeled the conditional probability of having CRC by constructing a Gaussian distribution with a conditional sample mean and conditional sample variance (Bertsekas and Tsitsiklis, 2008). This conditioning was based on whether or not each respondent was drawn from the CRC or the never-cancer population. That is, the conditional probability $P = \mathrm{P}(\mathrm{C}|\Phi)$ of the event Φ of having a set of features (e.g., hypertension, diabetes, body-mass index) resulting in the event C of having CRC was given by Bayes theorem as $\mathrm{P}(\mathrm{C}|\Phi) = \mathrm{P}(\Phi|\mathrm{C})\mathrm{P}(\mathrm{C})/\mathrm{P}(\Phi)$. The NB method thus incorporated inter-factor coupling, though as a multiplicative model that assumed the factors to be distributed independently. Despite this assumption of independence being almost always incorrect, the NB method's performance was competitive with those of more advanced MLAs (Rish, 2001).

The LDA and SVM calculate a decision boundary between the positive and negative populations that maximized a likelihood function (Fisher, 1936; Cortes and Vapnik, 1995). The method assumed homoscedasticity, multicollinearity, and that the responses were random variables drawn from completely independent Gaussian distributions. The SVM method similarly calculated a decision boundary, except without assuming the feature-values were drawn from a Gaussian distribution. In general, decision boundary methods are effective because they resist the effects of outliers.

The DT method constructed a flowchart of factors leading to CRC. The DT used the variable of lowest entropy (Shannon, 1948; Morgan and Sonquist, 1963) to construct the base of the tree, and used increasingly less informative variables at higher branches. Such a flowchart can be easily understood by a human, and is thus highly desirable in a clinical setting. Finally, we tested a bootstrap-aggregated ("bagged") collection of random trees, better known as a random forest (RF). Such RFs resist the overfitting that DT are prone to, but lack the transparency and information that DTs have in making their classifications.

Six Imputation Methods to Handle Missing Data

The datasets were subject to mean, Gaussian, Lorentzian, one-hot encoding, expectation-maximization (EM), and listwise deletion to handle data that are missing completely at random (Little and Rubin, 2014), some of which over-represented distributional moments. The six examined imputation methods have different strengths and weaknesses. Imputation by mean over-represents the mean. Imputation by drawing from a Gaussian random variable over-represents the variance about the mean. Imputation by drawing from a Cauchy random variable does not over-represent the mean or variance. Imputation by one-hot encoding (Bishop, 2006) uses the actual missingness of a data-entry as a feature. Finally, imputation by the (multivariate Gaussian) expectation-maximization (EM) iterative method over-represents the covariance between features (e.g., the covariance of diabetes with hypertension). The methods that draw from the Gaussian and Cauchy distributions used MATLAB's random number generator in invoking the function "rand," and thus are stochastic. Imputation by mean, one-hot encoding, EM algorithm, and listwise deletion, on the other hand, are deterministic.

The version of the EM algorithm that we chose assumed that all variables in each dataset were random variables drawn from a multivariate Gaussian distribution (Bishop, 2006). Iterations of the algorithm imputated the MCAR with values that overrepresented the covariance of each data column with each other. Since the multivariate Gaussian distribution is completely specified by its mean and variance, imputation by this method is incorrect only if the data is not drawn from a multivariate Gaussian or if the data is not MCAR. Because both the NHIS and PLCO datasets distinguish between data that is MCAR and data not missing at random, the effect of non-normal/Gaussian distribution of missingness remained minimized.

The multivariate-Gaussian EM algorithm is just one of many types of EM algorithms, as other data distributions (e.g., a multivariate-multinomial) may be assumed. Because our data contains a mixture of continuous and binary data-fields, and because the closed-form properties of the multivariate Gaussian are well-known (Bishop, 2006), we used Gaussian expectation-maximization for convenience. Categorical survey fields are multinomial, and a sufficiently-large number of such multinomial random variables are Gaussian by the central limit theorem (Bertsekas and Tsitsiklis, 2008). Ordinal survey fields have a distribution that in general have non-zero skewness and kurtosis, and thus are not exactly Gaussian. To avoid the calculation of the covariance of one multivariate Gaussian distribution with a non-Gaussian distribution, we just used a multivariate Gaussian for all fields. "Multivariate-Gaussian EM imputation" shall be referred to as just "EM imputation" throughout this paper.

About 1.2% of all data (795,215 respondents) was missing completely at random. However, about 16% of the 795,215 respondents had one or more of these missing entries. Listwise-deletion discards any respondent with even one missing entry, so about 16% of all data is then lost.

Model Evaluation

A popular metric of the performance in discriminating CRC incidence from non-CRC incidence is concordance, which is sometimes known as the area under the curve (AUC) of the receiver-operator characteristic (ROC) plot (Hanley and McNeil, 1982; Hosmer and Lemeshow, 2000; Fawcett, 2005). We reported concordances from training on NHIS/PLCO and testing upon PLCO/NHIS ("cross-testing"), which gives a TRIPOD level (Collins et al., 2015) of 3. Total uncertainty in concordance across cross-testing (Picard and Cook, 1984) is calculated using Equation (3).

For individuals ages 18–49 the PLCO dataset[2] has a sharply different prevalence of CRC (379 positives, 12 negatives) compared to the NHIS dataset[1] (114 positives and 76,676 negatives for family history data used; 562 positives and 398,222 negatives for family history data not used). Thus, for this age range, models were cross-tested between the 2-folds formed by the following non-random split: (1) the combination of all PLCO data with NHIS years 1997–2006 and (2) the remaining NHIS years 2007–2017. This makes the testing level for individuals ages 18–49 drop from TRIPOD 3 to TRIPOD 2b.

Stratifying CRC Risk

The ANN with EM imputation was used to stratify subjects into low-, medium-, and high CRC-risk groups. The ANN trained on NHIS data, and used this model to stratify the PLCO subjects into these risk categories. The PLCO dataset records the time in days at which the participant was diagnosed with CRC, and that time was used to build a forecast in the form of a Kaplan-Meier (KM) survival plot. Performance in risk-stratification was reported to give both an illustration of immediate clinical application and a performance metric that is not as biased as concordance is (Bertsekas and Tsitsiklis, 2008; Hajian-Tilaki, 2013).

RESULTS

Concordance Statistics of Seven Machine Learning Algorithms

Figure 1 is a ROC plot of the seven MLAs used with datasets subject to EM imputation. The standard deviation was formed from the variance from cross-testing between the NHIS and PLCO datasets, and the variance from screened/unscreened sub-populations (Hanley and McNeil, 1982) using Equation (3). Considering the mean concordance minus the total uncertainty (Equation 3) to be the metric of performance, the top performer was the ANN, with the SVM and NB as equally-performing runner-ups. LR (Hosmer and Lemeshow, 2000) offered fourth-place performance. Our ANN used the same logistic activation function (Bishop, 2006) as the LR. Our LR was our ANN with no hidden layers, suggesting the importance of inter-factor coupling possibly corresponding to complications. The good performance of the SVM came from not assuming a particular underlying distribution to the data, while LDA assumed that the NHIS and PLCO data were drawn from Gaussian distributions. The good performance of the NB came from its multiplicative incorporations of inter-factor coupling. The ANN's good performance was also roughly insensitive to which imputation method is used. The SVM and LDA perform well with one-hot encoding imputated data due to their resisting overfitting and outliers. RFs offered slightly improved performance over the DT, but worse than the ANN.

The concordance statistics for cross-testing for all combinations of MLAs and imputation methods are summarized in **Table 1**, showing relevant divisions of the datasets by age, as well as the effect of making family history data part of the model vs. leaving it out. The ANN offered performance (mean concordance minus the uncertainty) that was not only better than other MLAs but also insensitive to which imputation method was used. It can also be seen that in the group ages 18–49, among whom recent diagnosis of CRC is rarer (due in part to a greater prevalence of those bypassed for by-age screening of the USPSTF's recommendations), concordance was driven up by the increased true negative rate (or specificity). The

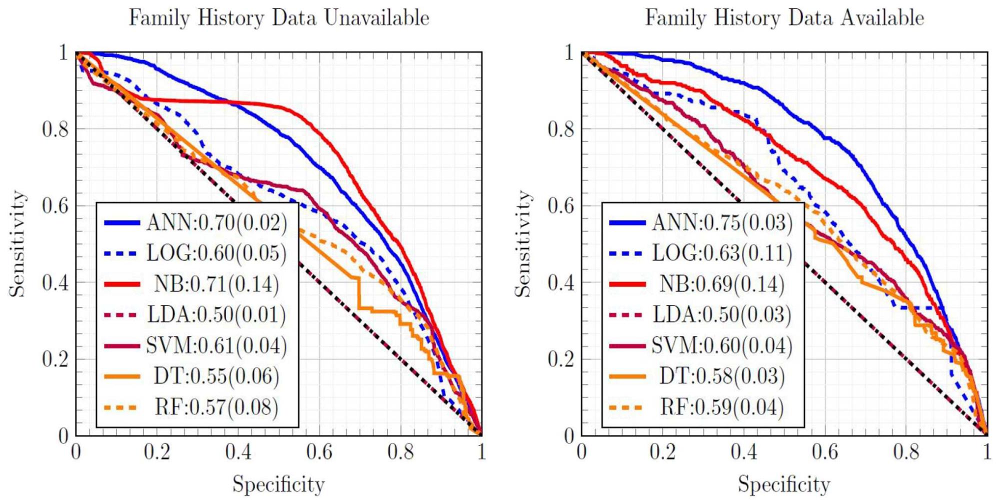
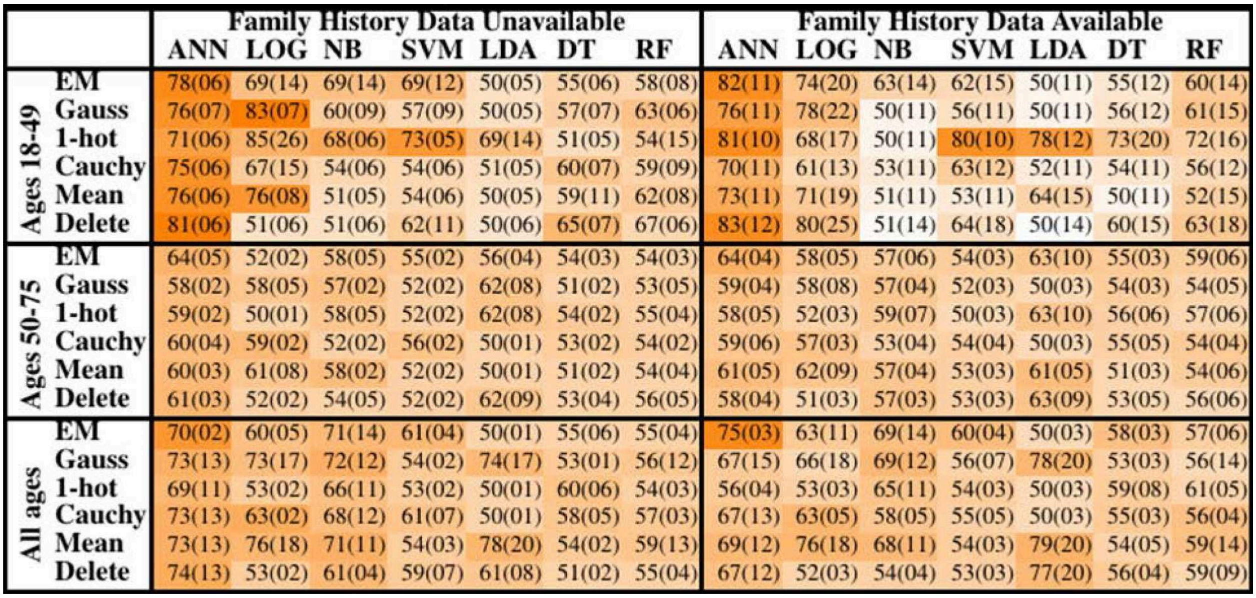


FIGURE 1 | Comparison of ROC curves of all seven MLAs, with the mean concordance and its uncertainty reported. Expectation maximization was used to impute missing data.

TABLE 1 | Mean concordance (standard deviation), multiplied by 100, for various machine learning algorithms, imputation methods, age groups, and with or without family history of CRC data.



		Family History Data Unavailable							Family History Data Available						
		ANN	LOG	NB	SVM	LDA	DT	RF	ANN	LOG	NB	SVM	LDA	DT	RF
Ages 18-49	EM	78(06)	69(14)	69(14)	69(12)	50(05)	55(06)	58(08)	82(11)	74(20)	63(14)	62(15)	50(11)	55(12)	60(14)
	Gauss	76(07)	83(07)	60(09)	57(09)	50(05)	57(07)	63(06)	76(11)	78(22)	50(11)	56(11)	50(11)	56(12)	61(15)
	1-hot	71(06)	85(26)	68(06)	73(05)	69(14)	51(05)	54(15)	81(10)	68(17)	50(11)	80(10)	78(12)	73(20)	72(16)
	Cauchy	75(06)	67(15)	54(06)	54(06)	51(05)	60(07)	59(09)	70(11)	61(13)	53(11)	63(12)	52(11)	54(11)	56(12)
	Mean	76(06)	76(08)	51(05)	54(06)	50(05)	59(11)	62(08)	73(11)	71(19)	51(11)	53(11)	64(15)	50(11)	52(15)
	Delete	81(06)	51(06)	51(06)	62(11)	50(06)	65(07)	67(06)	83(12)	80(25)	51(14)	64(18)	50(14)	60(15)	63(18)
Ages 50-75	EM	64(05)	52(02)	58(05)	55(02)	56(04)	54(03)	54(03)	64(04)	58(05)	57(06)	54(03)	63(10)	55(03)	59(06)
	Gauss	58(02)	58(05)	57(02)	52(02)	62(08)	51(02)	53(05)	59(04)	58(08)	57(04)	52(03)	50(03)	54(03)	54(05)
	1-hot	59(02)	50(01)	58(05)	52(02)	62(08)	54(02)	55(04)	58(05)	52(03)	59(07)	50(03)	63(10)	56(06)	57(06)
	Cauchy	60(04)	59(02)	52(02)	56(02)	50(01)	53(02)	54(02)	59(06)	57(03)	53(04)	54(04)	50(03)	55(05)	54(04)
	Mean	60(03)	61(08)	58(02)	52(02)	50(01)	51(02)	54(04)	61(05)	62(09)	57(04)	53(03)	61(05)	51(03)	54(06)
	Delete	61(03)	52(02)	54(05)	52(02)	62(09)	53(04)	56(05)	58(04)	51(03)	57(03)	53(03)	63(09)	53(05)	56(06)
All ages	EM	70(02)	60(05)	71(14)	61(04)	50(01)	55(06)	55(04)	75(03)	63(11)	69(14)	60(04)	50(03)	58(03)	57(06)
	Gauss	73(13)	73(17)	72(12)	54(02)	74(17)	53(01)	56(12)	67(15)	66(18)	69(12)	56(07)	78(20)	53(03)	56(14)
	1-hot	69(11)	53(02)	66(11)	53(02)	50(01)	60(06)	54(03)	56(04)	53(03)	65(11)	54(03)	50(03)	59(08)	61(05)
	Cauchy	73(13)	63(02)	68(12)	61(07)	50(01)	58(05)	57(03)	67(13)	63(05)	58(05)	55(05)	50(03)	55(03)	56(04)
	Mean	73(13)	76(18)	71(11)	54(03)	78(20)	54(02)	59(13)	69(12)	76(18)	68(11)	54(03)	79(20)	54(05)	59(14)
	Delete	74(13)	53(02)	61(04)	59(07)	61(08)	51(02)	55(04)	67(12)	52(03)	54(04)	53(03)	77(20)	56(04)	59(09)

Models for all ages and ages 50–75 were conducted at TRIPOD level 3 and models for ages 18–49 at TRIPOD level 2b. The standard deviation reported has a component of population-uncertainty from Equation (1) and a cross-uncertainty from Equation (2). The cell shading scheme was determined by subtracting the standard deviation from the mean concordance statistic, so that darker shading indicates a concordance statistic that not only had a higher mean value, but also a lower uncertainty.

opposite effect was observed in the group ages 50–75. Including family history data improved performance, but in a manner that is offset by the fact that it could only be included for a smaller data. Finally, it can be seen that the EM Gaussian algorithm tended to give the best concordance. One-hot encoding similarly performed well.

Testing the ANN at TRIPOD Level 3

In **Figure 2** the ANN with EM imputation performed consistently well as the incorporation of family history data and age range of subjects varied. The AUCs were greatest for individuals ages 18–49. The prevalence of CRC was lower in this group and **Figure 2** shows that the concordance was driven up by the low-cutoff portion of the ROC curve where the sensitivity of the ANN with EM imputation can be seen to rise sharply. This sharp rise is due to the high probability of any negative call being correct in a dataset with such low prevalence of CRC. As the cutoff increases in **Figure 2**, the sensitivity exhibits several sharp drop-offs. In the high-cutoff portion of **Figure 2**, the performance of the ANN with EM imputation becomes

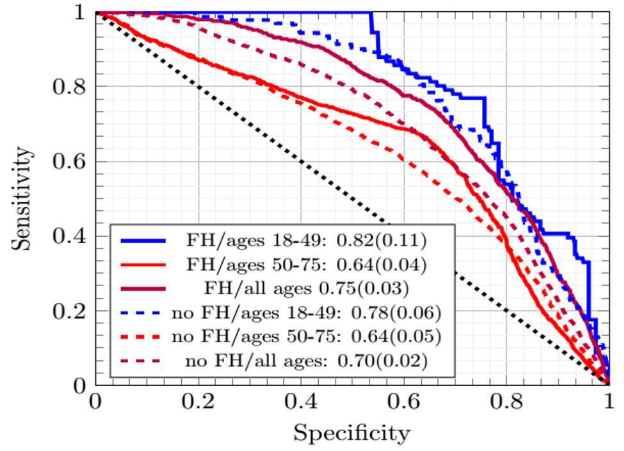


FIGURE 2 | The ROC curves of the ANNs averaged across cross-testing for different sub-demographics and with/without family history data.

insensitive to the age-demographic, or even the incorporation of family history data in the model. This trend is in sharp contrast to the low-cutoff portion of the ROC, where performance in the group of individuals ages 18–49 was significantly better than in the high-prevalence group of those ages 50–75. These results make the case that the concordance is a good measure of the performance of the ANN with EM imputation relative to other MLAs.

The improvement in concordance in **Figure 2** and **Table 1** is due to the ANN becoming increasingly insensitive when trained on data from individuals ages 18–49 and thus making more negative calls. The high number of negative calls in ages 18–49 gives a high specificity, and thus a high concordance.

Risk Stratification by ANN at TRIPOD Level 3

We stratified survey-respondents by the risk score calculated by the ANN with EM imputation. Such stratification has been completed at TRIPOD level 2b in previous work (Hart et al., 2018; Rofman et al., 2018), and was done here at TRIPOD level 3. **Figure 3** illustrates the stratification of individuals into three risk score categories. **Table 2** shows how many survey-respondents (CRC and never-cancer) ended up in each category.

In **Figure 3**, only relative (rather than absolute) values of risk are relevant, and thus the numbering of the horizontal axis is not comparable between plots. Unitizing the axis of risk to the interval [0, 1] is not done. Doing so would be misleading because differing models have differing levels of risk because the minimum and maximum values of risk needed for unitization differ between models. In stratifying NHIS/PLCO using a PLCO/NHIS-trained ANN, the risk-boundaries constructed as such were in general not equal, and the interval formed by this disparity is demarcated by vertical black dotted lines. Cumulative functions and complement-cumulative functions of negative and positive populations are plotted. Dotted lines and solid lines of the same color are complementary cumulative distributions summing to 100%.

Predicting CRC Incidence in the Never-Cancer Population

The Kaplan-Meier plots of **Figure 4** show the estimated probability of the never-cancer PLCO population getting CRC as a function of time in years, taking the CRC population as a Bayesian given. A cone of uncertainty is indicated. This cone, which widens at later times, suggests that the never-cancer population flagged as high risk (see **Figure 3** and **Table 2**) has an appreciable probability of developing CRC at a later time. Accordingly, this group regarded as "false positives" actually would benefit from screening. Because these false positives drive down the sensitivity and positive predictive value, this builds the case that concordance is better suited as a relative rather than absolute metric of performance.

In **Figure 4**, while the risk stratification into three categories is done at TRIPOD level 3, the confidence intervals are at TRIPOD level 1a. This is because they contain only a population-uncertainty calculated from an expression analogous to Equation (1).

DISCUSSION

Machine Learning Algorithms for CRC Prediction

Obtaining a concordance of 0.70 ± 0.02 on training an ANN with EM imputation gives a test that is competitive with the tests using routine data itemized in the review by Usher-Smith et al. (2016), including Betes (Betes et al., 2003) (TRIPOD level 3) and even the self-completed questionnaire used by Colditz (Colditz et al., 2000) (TRIPOD level 3). Our model of CRC risk combines routine data (involving no biomarkers) to form a score of CRC risk, and thus gives a discriminating and generalizable score of CRC risk.

Like other clinical tests, the negative calls made by the ANN with EM imputation have a greater probability of being correct compared to the corresponding probability of correctness of its positive calls. This trend can be seen by considering the model's strong performance among individuals ages 18–49 as well as its sensitivity (0.63 ± 0.06) and misclassification rate of CRC as low risk being significantly worse than its specificity (0.82 ± 0.04) and misclassification rate of non-CRC as high risk. Among individuals ages 18–49, the concordance was driven up by the increase in specificity which was probably triggered by the greater number of respondents for whom the ANN with EM imputation could make correct negative calls. Likewise, better performance was observed when testing the PLCO-trained model upon NHIS data (compared to testing the NHIS-trained model upon PLCO data). In NHIS, there are $\sim 10^6$ respondents of which $\sim 10^3$ have CRC, whereas in PLCO there are $\sim 10^5$ respondents of which $\sim 10^3$ have CRC. Thus, for the PLCO dataset, there were (an order of magnitude) fewer specificity-boosting negative calls. In typical clinical practice such a test that makes a negative call gives a

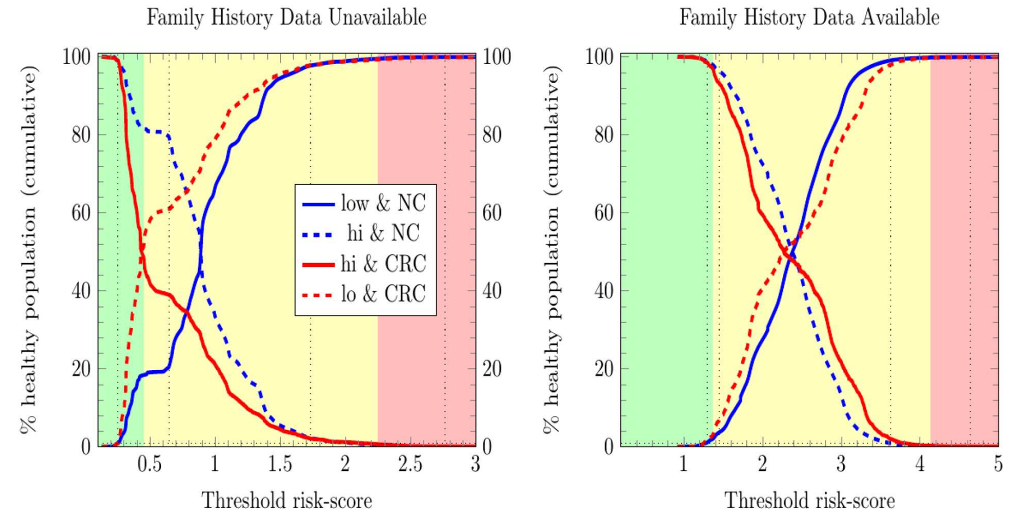
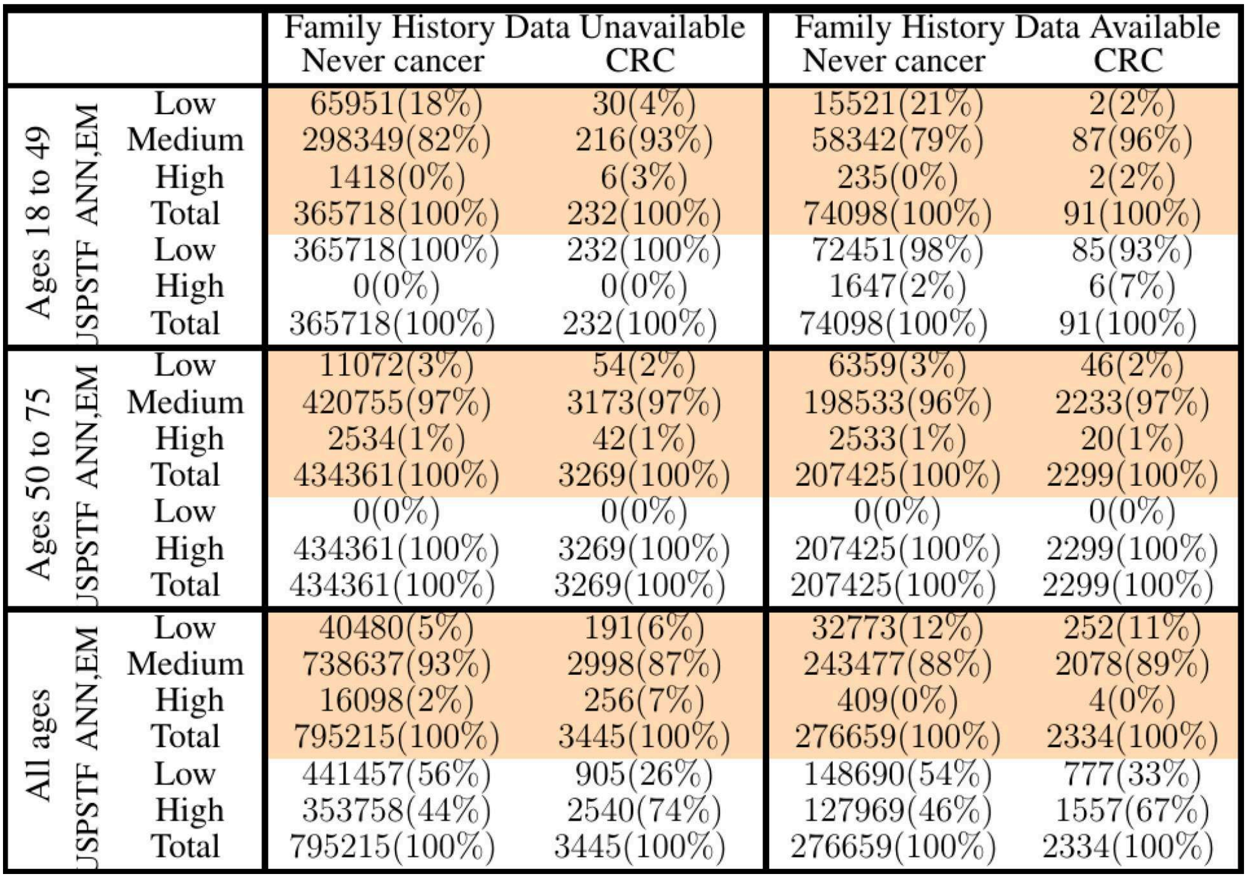


FIGURE 3 | Stratification of individuals into low-, medium-, and high CRC-risk groups by the ANN with EM imputation. Risk categories are defined by the requirement that no more than 1% of positive cases be classified as low risk, and no more than 1% of negative cases be classified as high risk.

TABLE 2 | Comparison of our ANN with EM imputation with USPSTF screening guidelines in stratifying PLCO and NHIS respondents into low-, medium-, and high CRC-risk groups.



			Family History Data Unavailable		Family History Data Available	
			Never cancer	CRC	Never cancer	CRC
Ages 18 to 49	ANN,EM	Low	65951(18%)	30(4%)	15521(21%)	2(2%)
		Medium	298349(82%)	216(93%)	58342(79%)	87(96%)
		High	1418(0%)	6(3%)	235(0%)	2(2%)
		Total	365718(100%)	232(100%)	74098(100%)	91(100%)
	USPSTF	Low	365718(100%)	232(100%)	72451(98%)	85(93%)
		High	0(0%)	0(0%)	1647(2%)	6(7%)
		Total	365718(100%)	232(100%)	74098(100%)	91(100%)
Ages 50 to 75	ANN,EM	Low	11072(3%)	54(2%)	6359(3%)	46(2%)
		Medium	420755(97%)	3173(97%)	198533(96%)	2233(97%)
		High	2534(1%)	42(1%)	2533(1%)	20(1%)
		Total	434361(100%)	3269(100%)	207425(100%)	2299(100%)
	USPSTF	Low	0(0%)	0(0%)	0(0%)	0(0%)
		High	434361(100%)	3269(100%)	207425(100%)	2299(100%)
		Total	434361(100%)	3269(100%)	207425(100%)	2299(100%)
All ages	ANN,EM	Low	40480(5%)	191(6%)	32773(12%)	252(11%)
		Medium	738637(93%)	2998(87%)	243477(88%)	2078(89%)
		High	16098(2%)	256(7%)	409(0%)	4(0%)
		Total	795215(100%)	3445(100%)	276659(100%)	2334(100%)
	USPSTF	Low	441457(56%)	905(26%)	148690(54%)	777(33%)
		High	353758(44%)	2540(74%)	127969(46%)	1557(67%)
		Total	795215(100%)	3445(100%)	276659(100%)	2334(100%)

recommendation for no further testing, while a positive call gives a recommendation for further testing by a more accurate (and costly) test (Simundic, 2017).

This paper reported uncertainty at TRIPOD level 3 wherever possible. Reporting uncertainty is crucial to determine optimal performance because concordance can be misleadingly high even

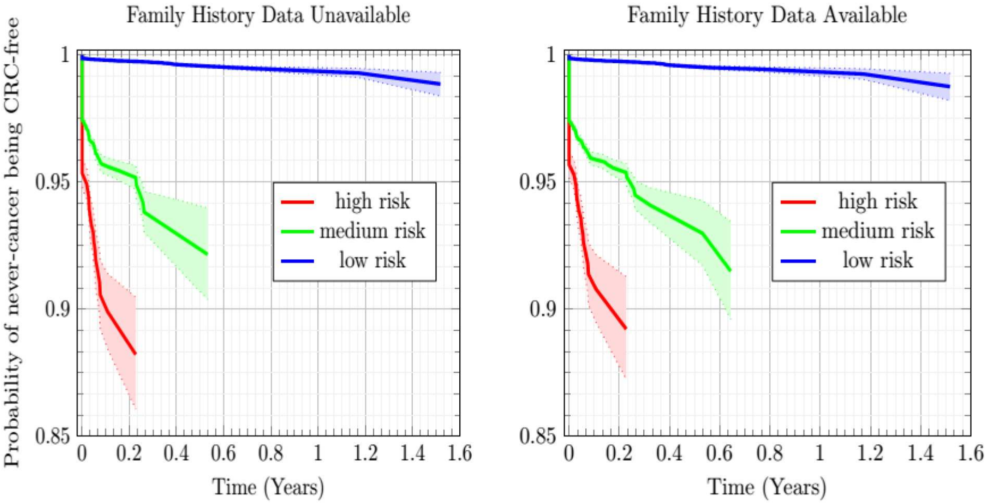


FIGURE 4 | Plot of the Kaplan-Meier estimator of the CRC-free probability of the PLCO respondents vs. time for the low-, medium-, and high CRC-risk groups stratified by ANN with EM imputation model. The shaded regions are 95% confidence intervals at TRIPOD level 1a.

when averaged across cross-testing. For instance, imputation of missing data with the average of a data-field gave a concordance that was almost 0.80, but with an accompanying uncertainty of 0.20. The performance for which mean minus uncertainty was greatest was 0.70 ± 0.02 (see **Table 1**) when ANN with EM imputation were used.

Improving Performance With Additional Relevant Factors

The input predictors to the MLAs were selected based on availability in both the NHIS and PLCO datasets, what Rubin calls the file-matching problem (Little and Rubin, 2014). Because of this selection criteria, some of the stronger factor-correlations with CRC (e.g., NSAIDs, such as aspirin and ibuprofen; Rodriguez and Huerta-Alvarez, 2001; Betes et al., 2003) needed to be omitted from the model, as data on use of NSAIDs was only available in the NHIS dataset for years 2000, 2005, 2010, and 2015. The risk-stratification demonstrated in **Table 2** would likely be even more effective if these stronger predictors were used. Indeed, a data-driven approach to detecting CRC risk in the general public would put priority on recording these strong predictors more regularly.

ANN With EM Imputation

The concordance of 0.70 ± 0.02 of our ANN with EM imputation is competitive with previous externally-tested (TRIPOD 3) risk models using routine data (Betes et al., 2003; Usher-Smith et al., 2016) as input. To our knowledge, calculating an uncertainty by the law of total variance (Bertsekas and Tsitsiklis, 2008) so as to incorporate both the population-uncertainty (Hanley and McNeil, 1982; Fawcett, 2005) of Equation (1) and the cross-uncertainty due to variance in performance across cross-testing (Picard and Cook, 1984) of Equation (2) has never been done before. Incorporating this additional component of cross-uncertainty demonstrates the advantage of using the ANN. The advantage of the ANN over LR is not in having a high mean concordance, but rather in having a much lower uncertainty, which demonstrates the generalizability of the model. Because of better generalizability, the ANN with EM imputation is considered the best among all the model/imputation configurations.

Clinical Deployment

In this work, the developed ANN with EM imputation is used to predict the colorectal cancer risk for individuals based on their personal health data. The output of the model, the colorectal cancer risk score, can be used to help the clinicians make screening decisions. Generally speaking, true positives require further screening and true negatives require no screening. False positives still stand to benefit from our model, which offers this population their individual cancer risk as a function of personal health habits they have at different times. Drops in an individual's risk score in response to better personal health habits, such as quitting smoking and treatment of diabetes will provide positive feedback for that individual in the form of a reduced risk-score. Furthermore, high-risk never-cancer false positives warrant heightened screening attention, as demonstrated by the sharply decreasing Kaplan-Meier probability of high-risk never-cancer individuals remaining free of CRC over time. In general, the temporal trend of cancer risk will determine the next step for the individuals.

CONCLUSION

In this comparative study, we have evaluated seven machine learning algorithms in combination with six imputation methods for missing data, all trained and cross-tested with the NHIS and PLCO datasets. Among various machine learning algorithms using different imputation methods, the artificial neural network with Gaussian expectation-maximization imputation was found to be optimal, with a concordance of 0.70 ± 0.02, a sensitivity of 0.63 ± 0.06, and a specificity of 0.82 ± 0.04. In CRC risk stratification this optimal model had a never-cancer misclassification rate of only 2%, and a CRC misclassification rate of only 6%. Being a TRIPOD level 3 study, our model with low uncertainty suggests that it can be used as a non-invasive and cost-effective tool to screen the CRC risk in large populations effectively using only personal health data.

DATA AVAILABILITY STATEMENT

The code used in this study is not publicly available due to a concern of intellectual property proprietary to Yale University. Requests to access the NHIS datasets should be directed to the Centers for Disease Control and Prevention (CDC) at https://www.cdc.gov/nchs/nhis/. Requests to access the PLCO datasets should be directed to the National Cancer Institute (NCI) at https://biometry.nci.nih.gov/cdas/plco/.

AUTHOR CONTRIBUTIONS

BN analyzed the data, produced the results, and wrote the technical details. GH, WM, YL, and GS produced the technical details, and reviewed the manuscript. JD generated the research ideas and reviewed the manuscript.

FUNDING

Research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number R01EB022589.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the facilities provided by the Yale Department of Therapeutic Radiology at which this work was carried out.

REFERENCES

Andoni, A., Panigrahy, R., Valiant, G., and Zhang, L. (2014). Learning polynomials with neural networks. *JMLR* 32.

Benard, F., Barkun, A. N., Martel, M., and von Renteln, D. (2018). Systematic review of colorectal cancer screening guidelines for average-risk adults: Summarizing the current global recommendations. *World J. Gastroenterol.* 24, 124–138. doi: 10.3748/wjg.v24.i1.124

Bertsekas, D. P., and Tsitsiklis, J. N. (2008). *Introduction to Probability*. Belmont, MA: Athena Scientific.

Betes, M., Munoz-Navas, M. A., Duque, J., Angos, R., Macias, E., Subtil, J. C., et al. (2003). Use of colonoscopy as a primary screening test for colorectal cancer in average risk people. *Am. J. Gastroenterol.* 98, 2648–2654. doi: 10.1111/j.1572-0241.2003.08771.x

Bibbins-Domingo, K., Grossman, D. C., Curry, S. J., Davidson, K. W., Epling, J. W. Jr., García, F. A. R., et al. (2016). Screening for colorectal cancer us preventive services task force recommendation statement. *J. Am. Med. Assoc.* 315, 2564–2575. doi: 10.1001/jama.2016.5989

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Colditz, G. A., Atwood, K. A., Emmons, K., Monson, R. R., Willett, W. C., Trichopoulos, D., et al. (2000). Harvard report on cancer prevention volume 4: Harvard cancer risk index. *Cancer Causes Control* 11, 477–488. doi: 10.1023/A:1008984432272

Collins, G. S., Reitsma, J. B., Altman, D. G., and Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *Br. J. Cancer* 162, 55–63. doi: 10.1161/CIRCULATIONAHA.114.014508

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Falco, M., Wyant, T., and Simmons, K. (2018). *What is Colorectal Cancer?* American Cancer Society.

Fawcett, T. (2005). An introduction to receiver-operator characteristic analysis. *Pattern Recogn. Lett.* 27, 861–874.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Mach. Learn.* 20, 273–297.

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J. Intern. Med.* 4, 627–635.

Hanley, J. A., and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 143, 29–36. doi: 10.1148/radiology.143.1.7063747

Hart, G. R., Roffman, D. A., Decker, R., and Deng, J. (2018). A multi-parameterized artificial neural network for lung cancer risk prediction. *PLoS ONE*. 13:e0205264. doi: 10.1371/journal.pone.0205264

Hosmer, D. W., and Lemeshow, S. (2000). *Applied Logistic Regression*. New York, NY: Wiley. doi: 10.1002/0471722146

Kingma, D. P., and Ba, J. L. (2015). Adam: a method for stochastic optimization. *ICLR 2015* arXiv:1412.6980v9.

Little, R. J. A., and Rubin, D. B. (2014). *Statistical Analysis with Missing Data*. Wiley.

Morgan, J. N., and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Am. Statist. Assoc.* 58, 415–434. doi: 10.1080/01621459.1963.10500855

National Cancer Institute (2018). *Cancer Stat Facts: Colorectal Cancer*. National Cancer Institute.

National Cancer Institute (2019). *Tests to Detect Colorectal Cancer and Polyps*. National Cancer Institute.

Picard, R. R., and Cook, R. D. (1984). Cross-validation of regression models. *J. Am. Stat. Assoc.* 79, 575–583. doi: 10.1080/01621459.1984.10478083

Platt, J. C. (1999). "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," in *Advances in Large Margin Classifiers* (Cambridge, MA: MIT Press), 61–74.

Rish, I. (2001). An empirical study of the naive bayes classifier. *IJCAI 2001 Work Empir. Methods Artif. Intell.* 3, 41–46.

Rodriguez, L. G., and Huerta-Alvarez, C. (2001). Reduced risk of colorectal cancer among long-term users of aspirin and nonaspirin nonsteroidal antiinflammatory drugs. *Epidemiology* 12, 88–93. doi: 10.1097/00001648-200101000-00015

Rofman, D., Hart, G., Girardi, M., Ko, C. J., and Deng, J. (2018). Predicting non-melanoma skin cancer via a multi-parameterized artifcial neural network. *Sci. Rep.* 8, 1–7. doi: 10.1038/s41598-018-19907-9

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Techn. J.* 27, 623–656. doi: 10.1002/j.1538-7305.1948.tb00917.x

Simundic, A. M. (2017). Extent of diagnostic agreement among medical referrals. *EJIFCC* 19, 203–211. doi: 10.1111/jep.12747

Usher-Smith, J. A., Walter, F. M., Emery, J. D., Win, A. K., and Griffin, S. J. (2016). Risk prediction models for colorectal cancer: a systematic review. *Cancer Prev. Res.* 9, 13–26. doi: 10.1158/1940-6207.CAPR-15-0274

Yang, Y., Mauldin, P. D., Ebeling, M., Hulsey, T. C., Liu, B., Thomas, M. B., et al. (2012). Effect of metabolic syndrome and its components on recurrence and survival in colon cancer patients. *Cancer* 119, 1512–1520. doi: 10.1002/cncr.27923

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

The Levels of TRIPOD and the Cross-Testing Uncertainty

Differing levels of TRIPOD have cross-testing uncertainty from different sources of distributional disparity. There is always a component of uncertainty due to the finitude of the dataset used, and that of the concordance is well-known (Hanley and McNeil, 1982; Fawcett, 2005). For a population of N respondents C of whom have cancer and a MLA giving a concordance of AUC, what shall be called the "population-uncertainty" Π^2 is a function of AUC, C, and N alone and given as,

$$\Pi^2 = \frac{1}{C(N-C)} \left(\begin{array}{l} AUC\,(1-AUC) + (C-1)\left(\dfrac{AUC}{2-AUC} - AUC^2\right) \\ + (N-C-1)\left(\dfrac{2AUC^2}{1+AUC} - AUC^2\right) \end{array} \right) \tag{1}$$

For TRIPOD level 1b or higher, an additional uncertainty component we call the "cross-uncertainty" (τ^2) arises from cross-validation or cross-testing. This uncertainty, unlike the population-uncertainty, depends explicitly upon the disparity of the distribution of the two datasets. If the split of the data is random (TRIPOD levels 1b and 2a), the cross-uncertainty is normal or Gaussian (Bishop, 2006; Bertsekas and Tsitsiklis, 2008). If the split of the data is non-random (TRIPOD levels 2b and 3), the cross-uncertainty indicates the difference between the distributions of the data of each group. In the case of cross-testing between NHIS and PLCO, the cross-uncertainty indicates the difference between the underlying distributions of each dataset.

For splitting of the data into n_f-folds the cross-uncertainty τ^2 is estimated as the sample variance resulting from the concordance $\mathrm{AUC_i}$ from testing or validating upon the ith-fold of data summed over all n_f-folds, which is done as:

$$\tau^2 = \frac{1}{n_f - 1} \sum_{i=1}^{n_f} \left(AUC_i - \overline{AUC}\right)^2; \;\; \overline{AUC} = \frac{1}{n_f} \sum_{i=1}^{n_f} AUC_i; \tag{2}$$

A true positive rate or sensitivity (TPR) and a true negative rate or specificity (SPC) determine a concordance or area under the (receiver-operator characteristic) curve AUC. Taking TPR and SPC to be random variables that take on differing values over differing folds of cross-testing, we used the law of total variance (Hosmer and Lemeshow, 2000) to form the uncertainty σ^2. Both the population-uncertainty Π^2 and the cross-uncertainty τ^2 between folds of data were incorporated as the following sum of variances conditioned upon a specific TPR and SPC:

$$\begin{aligned} \sigma^2 &= E[\mathrm{var}\, AUC|TPR, SPC] + \mathrm{var}\, E[AUC|TPR, SPC] \\ &= \frac{1}{n_f} \sum_{i=1}^{n_f} \Pi_i^2 + \tau^2 = \overline{\Pi_i^2} + \tau^2 \end{aligned} \tag{3}$$

Throughout this paper, we reported the square root of the total uncertainty $\sqrt{\sigma^2} = \sigma > 0$ from Equation (3), formed from summing the mean population-uncertainty $\overline{\Pi_i^2}$ from Equation (1) and the cross-uncertainty τ^2 from Equation (2). Through the cross-uncertainty τ^2 the disparity between the distributions of the folds of cross-testing or cross-validation appear explicitly in the total uncertainty σ^2.