PLOS ONE

RESEARCH ARTICLE

Scoring colorectal cancer risk with an artificial neural network based on self-reportable personal health data

Bradley J. Nartowt[1], Gregory R. Hart[1], David A. Roffman[2], Xavier Llor[3], Issa Ali[1], Wazir Muhammad[1], Ying Liang[1], Jun Deng[1]*

1 Department of Therapeutic Radiology, School of Medicine, Yale University, New Haven, Connecticut, United States of America, **2** Sun Nuclear Corporation, Melbourne, FL, United States of America, **3** Department of Digestive Diseases, School of Medicine, Yale University, New Haven, Connecticut, United States of America

* jun.deng@yale.edu

OPEN ACCESS

Citation: Nartowt BJ, Hart GR, Roffman DA, Llor X, Ali I, Muhammad W, et al. (2019) Scoring colorectal cancer risk with an artificial neural network based on self-reportable personal health data. PLoS ONE 14(8): e0221421. https://doi.org/10.1371/journal.pone.0221421

Editor: Frank T. Kolligs, University of Munich, GERMANY

Received: March 5, 2019

Accepted: August 6, 2019

Published: August 22, 2019

Data Availability Statement: All data used in this paper is publicly available through the CDC's website. At the time of submission the URL: https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm, goes directly to the webpage from which each year of NHIS data can be found.

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

Abstract

Colorectal cancer (CRC) is third in prevalence and mortality among all cancers in the US. Currently, the United States Preventative Services Task Force (USPSTF) recommends anyone ages 50–75 and/or with a family history to be screened for CRC. To improve screening specificity and sensitivity, we have built an artificial neural network (ANN) trained on 12 to 14 categories of personal health data from the National Health Interview Survey (NHIS). Years 1997–2016 of the NHIS contain 583,770 respondents who had never received a diagnosis of any cancer and 1409 who had received a diagnosis of CRC within 4 years of taking the survey. The trained ANN has sensitivity of 0.57 ± 0.03, specificity of 0.89 ± 0.02, positive predictive value of 0.0075 ± 0.0003, negative predictive value of 0.999 ± 0.001, and concordance of 0.80 ± 0.05 per the guidelines of Transparent Reporting of Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) level 2a, comparable to current risk-scoring methods. To demonstrate clinical applicability, both USPSTF guidelines and the trained ANN are used to stratify respondents to the 2017 NHIS into low-, medium- and high-risk categories (TRIPOD levels 4 and 2b, respectively). The number of CRC respondents misclassified as low risk is decreased from 35% by screening guidelines to 5% by ANN (in 60 cases). The number of non-CRC respondents misclassified as high risk is decreased from 53% by screening guidelines to 6% by ANN (in 25,457 cases). Our results demonstrate a robustly-tested method of stratifying CRC risk that is non-invasive, cost-effective, and easy to implement publicly.

Introduction

Colorectal adenocarcinomas are the result of unregulated growth in the colon mucosa that commonly starts with polypoid lesions progressing into advanced cancers [1]. Of all new cancer cases in the US, 8.0% are colorectal. Colorectal cancer (CRC) claims 8.4% of all cancer-deaths and the overall 5-year survival rate is 66% [2]. Early stage (localized) CRC has a 5-year

R01EB022589. DAR was initially funded by R01EB022589 when writing the first working version of the ANN. BJN et al. took over and continued this study with the funding of R01EB022589 while DAR went on to be employed and supported with salary by Sun Nuclear Corporation. Sun Nuclear Corporation provided support in the form of salaries for authors DAR, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section. The specific role of DAR is articulated in the 'author contributions' section. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or of Sun Nuclear Corporation.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: DAR was employed and supported with salary by Sun Nuclear Corporation. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

survival rate close to 90%, while that of distant/metastatic CRC survival is less than 14% [2]. Treatment of colorectal cancer even in its metastatic state is fairly standardized, typically with bevacizumab [3] which has a side-effect of high blood pressure dependent upon the dose [4,5].

There are multiple personal health factors correlating moderately with incidence of advanced colorectal neoplasia [6] and colorectal cancer which are both self-reportable, easily gathered, and usable in scoring CRC risk [7–10]. For example, CRC is more frequent in men than women and African Americans have the highest incidence in the US [2,11]. Environmental factors, socio-economic features, and co-morbidities additionally influence CRC risk [2]. In a study of subjects from Wisconsin and Minnesota, risk factors for long-term colorectal-cancer mortality were found to be age, sex, and higher body-mass index (BMI) [12], though colorectal-cancer incidence was only marginally increased with higher BMI. In a meta-study more specialized to BMI as a factor of CRC risk, those of lower BMI were at risk for colorectal cancer as well [13]. It is possible to build a CRC risk score from these many factors, and many authors have done so [14–16].

Efforts to create risk indexes or prediction models for CRC [14–16] have used a variety of data of three types: (1) routine, (2) reportable by self-completed questionnaire, (3) genetic biomarkers as recently summarized [7,17]. While logistic regression [18] is a popular method of scoring CRC risk [15], this work opts to use an artificial neural network (ANN) trained with professionally-collected routine data [19] and shown in Fig 1. While an ANN is not strictly superior to logistic regression [20], an ANN incorporates complicated inter-factor coupling [21]. Given the complexity of human biology, inter-factor coupling is likely to be important in the predicting CRC from personal health data.

The United States Preventative Services Task Force [22] (USPSTF) and various medical societies [23] currently recommend screening by age and family history only, despite models incorporating additional risk factors. Specifically, those with no family history of cancer aged 50–75 are recommended for screening [22,24], meaning that United States citizens living to age 50 and beyond are flagged for screening. Of the incidences of CRC in the National Health Interview Survey (NHIS) dataset [19] within 4 years of the survey, 65.6% occur in ages 50–75 and 80.7% are ages 50 and older, leaving 19.3% under age 50 and outside the USPSTF

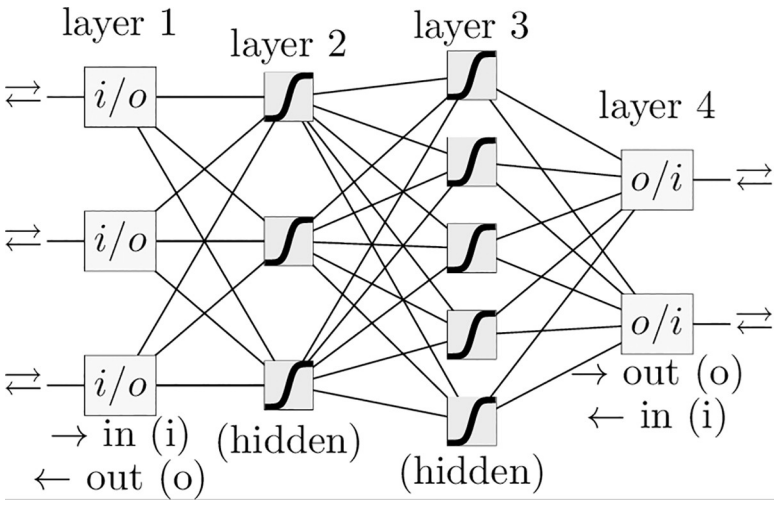

Fig 1. Schematic of example ANN. A schematic of an ANN with four layers and a logistic activation function. The ANN in this paper has one input neuron for each of the 12 to 14 factors, the same number of neurons in each hidden layer, and a single output neuron for the prediction. The upper arrow indicates forward-propagation and the lower arrow indicates back-propagation.

https://doi.org/10.1371/journal.pone.0221421.g001

screening guidelines. The USPSTF's screening guidelines saves many lives [25] but at the expense of many false-positives, which could lead to unnecessary, expensive, and occasionally injurious screening. There is also a remainder of false-negatives (specifically, CRC occurring under age 50 in those with no family history [11]) that could be flagged for screening by an appropriate model of risk.

Current screening procedures for CRC (by colonoscopy [24,26,27] every 10 years or sigmoidoscopy [28,29] / colonography [22] every 5 years) are often invasive with suboptimal accuracy, and always expensive. For example, perforation of the colon and/or bleeding during both colonoscopy and flexible sigmoidoscopy have been reported, demonstrating the need for non-invasive techniques [22]. Hence, even at the expense of lowered positive/negative predictive value (PPV/NPV), there has been a push for less expensive and less invasive screening methods. From the 1990s to the 2010s, this push has yielded the yearly fecal occult blood test (FOBT) [29] and the yearly fecal immunochemical test (FIT) [30] possibly combined [31] with the SEPT9-methylation [32,33] test. An efficient and non-invasive method that can help clinicians identify appropriate CRC screenings for individuals (e.g., colonoscopy/sigmoidoscopy for high risk and stool tests for low risk) is desired [34,35].

In this work, we use an artificial neural network (ANN) to score an individual's risk of CRC as a means of assisting screening recommendations. Previously, the ANN has been used to assist diagnosis from expertly-collected data (e.g., distinguishing between sporadic colon adenomas and cancers vs. inflammatory bowel disease-related dysplasia or cancer) from high-dimensional genetic datasets [14]. This application is limited to use by professionals. In contrast, a comparatively-large portion of the population can self-report their habits of smoking, exercise, their hypertension, diabetes, emphysema, and other personal-health data. Our ANN is trained with this latter data specifically because of the ease of self-reporting per TRIPOD guidelines [36]. This makes it inappropriate for use as a diagnostic tool, but able to be mass-implemented. Members of the general public can make personal decisions to be screened [22] based on their score of CRC risk, and clinicians can use this same score to assist their screening recommendations.

Materials and methods

Data

The data to train and validate the ANN was the 1997–2016 responses to the NHIS sample adult questionnaire from the Centers for Disease Control and Prevention (CDC) [19]. About 20% of respondents were discarded in our study due to missing entries in the NHIS questionnaire. The NHIS data inquired about both colon and rectal cancer. Therefore, we are counting anyone that had colon cancer, rectal cancer, or both colon and rectal cancer as a single incidence of colorectal cancer. Subjects who were diagnosed with CRC more than 4 years prior to the time of survey were discarded from the dataset and thus not used. While the NHIS dataset records the age at which the respondent was professionally diagnosed with CRC (if at all), the dataset does not record the time at which diagnoses of other predictors (e.g., diabetes) was given. Therefore, to increase the probability that the predictors arose prior to the cancer we only include those that were recently diagnosed with cancer (within four years prior to taking the NHIS).

For the default model, 525,394 and 58,376 subjects were used to train and test the ANN respectively, among whom 1,269 and 140 respondents were told by a doctor or other health professional that they had CRC recently; that is, within 4 years of their taking the NHIS. (This is referred to as "recently and professionally diagnosed with CRC" throughout the manuscript) recently and professionally diagnosed with CRC. Variants of this model whose resulting ROCs

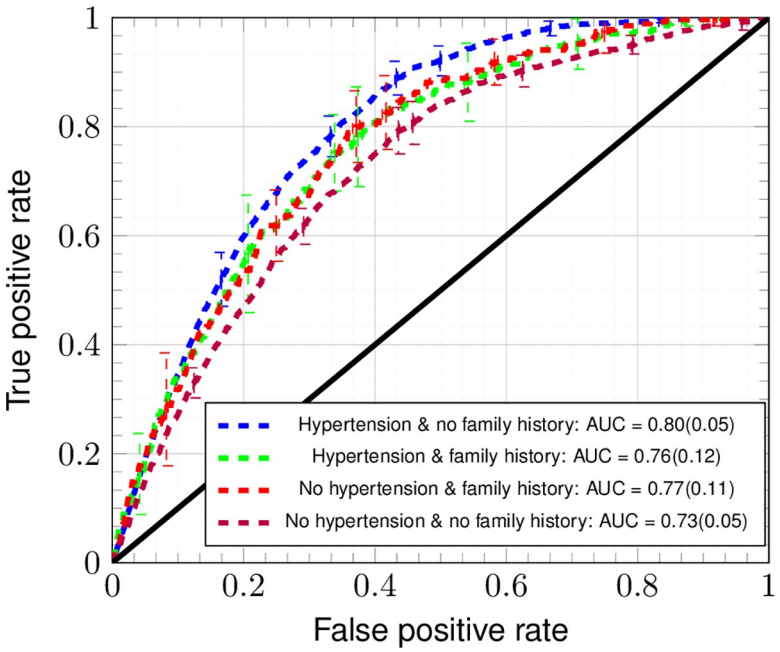

Fig 2. ROC curves of the ANN for ten-fold cross-testing dataset (TRIPOD 2a). The ANN trained with the factors marked "default model" in Table 2 with (blue line) and without hypertension (purple line). The ANN was also trained on a reduced dataset that included family history with (green line) and without (red line) hypertension. Error bars denote the standard deviation of the TPR and FPR across ten folds of stratified cross-testing (TRIPOD level 2a).

https://doi.org/10.1371/journal.pone.0221421.g002

are studied in Fig 2 significantly change the number of subjects in each population. Each population is given in Table 1.

From the NHIS data we obtained the factors appearing in Table 2. Heart conditions are pooled into 1 factor, and training is done with and without data on hypertension and CRC family history, leaving 12 to 14 factors. These factors were selected because they correlate strongly with CRC incidence and also have either "ever" as their time of incidence (see Discussion), are "permanent" (e.g., age, ethnicity), or are "per week" (vigorous exercise) frequency.

Table 1. Number of NHIS respondents in final dataset when certain factors are chosen.

Ever screened? †	Family history†	Age (years)	Hyper--tension	Training	Training & CRC	Testing	Testing& CRC
Unused◊	Unused◊	18–85◊	Used◊	525,394◊	1,269◊	58,376◊	140◊
Unused	Used	18–85	Used	105,950	245	11,772	27
Unused	Used	18–85	Unused	105,760	245	11,723	27
Unused	Unused	18–85	Unused	525,394	1,269	58,376	140
Unused	Unused	18–49	Used	298,085	162	33,120	18
Unused	Unused	50–75	Used	227,310	1,107	25,255	122
Used	Unused	18–85	Used	10,261	72	217,774	446
Used	Used	18–85	Used	66,938	300	161,098	218
Used	Used	18–85	Unused	9,002	59	219,176	459
Used	Unused	18–85	Unused	13,755	71	212,995	443

† Data appearing in NHIS years 2000, 2005, 2010, and 2015 only when a set of supplementary questions were asked.

◊ Data in the default model.

https://doi.org/10.1371/journal.pone.0221421.t001

Table 2. All factors in the NHIS datasets used to train the ANN in scoring CRC risk, in descending order of correlation magnitude.

Name of Factor	Correlation with Recent CRC, $\times 10^{-2}$	Type of Factor	# of Unique Values of Factor	Time of Incidence, Frequency, or Duration
Current or Cancer Age†	+4.907	Continuous	68	Permanent
Hypertension†	+3.045	Ordinal	2	Ever
Number of first-degree relatives with CRC (NHIS years 2000, 2005, 2010, and 2015 only)	+2.906	Ordinal	4	Permanent
Coronary heart disease	+2.349	Ordinal	2	Ever
Pooled heart conditions†	+2.063	Ordinal	2	Ever
Myocardial infarction	+2.060	Ordinal	2	Ever
Diabetes (non-gestational) †	+2.056	Ordinal	3	Ever
Heart condition/disease	+1.972	Ordinal	2	Ever
Vigorous exercise frequency†	-1.971	Continuous	33	Per week
Angina pectoris	+1.769	Ordinal	2	Ever
Ulcer (stomach, duodenal, peptic)†	+1.540	Ordinal	2	Ever
Hispanic ethnicity†	-1.269	Categorical	2	Permanent
Stroke†	+1.218	Ordinal	2	Ever
Emphysema†	+1.220	Ordinal	2	Ever
American Indian, African American, other, or multiple race †	-0.494	Categorical	2	Permanent
Sex (male) †	-0.350	Categorical	2	Permanent
Body-mass index†	+0.234	Continuous	4223	Current
Smoking frequency†	+0.0461	Ordinal	4	Current

† Denotes factors that are part of the model referred to as "default" throughout this paper.

https://doi.org/10.1371/journal.pone.0221421.t002

The relative importance of these factors (in terms of Pearson correlation [37] with CRC) are presented in Table 2.

One of the 14 factors we would like to use is hypertension, but the approval [38] of bevacizumab as a treatment for CRC in 2004 can result in many people having hypertension because of the treatment of CRC instead of being a risk factor for CRC itself. Therefore, we determined the correlation of hypertension and CRC prior to 2004 (3.16×10^{-2}) and after 2004 (2.86×10^{-2}). This is an unexpected post-approval [3,38] decrease of 3×10^{-3}. Thus, we decided bevacizumab-induced hypertension was unimportant and we used hypertension in all models unless indicated otherwise.

Raw data in the NHIS dataset was mapped to the interval [0,1] to be input to the ANN in two ways depending on whether the data is categorical or ordinal. Referring to Table 2, the factors of having hypertension, ulcers, a stroke, and emphysema are binary variables and naturally to map to 0 for "no", 1 for "yes". Diabetic status can have one of three discrete values: not diabetic, pre-diabetic/borderline, and diabetic. These were mapped to 0, 0.5, and 1, respectively. The age factor is continuous and is the age at the time of responding to the NHIS if the respondent never had any cancer or the age at which they were recently and professionally diagnosed with CRC otherwise. The factors of weekly frequency of vigorous exercise and BMI are also continuous. The NHIS defines vigorous exercise as lasting 10 minutes or more and resulting in one or more of: heavy sweating, breathing, or elevated heart rate. All such continuous factors are unitized to the interval [0,1] using the replacement $x \Rightarrow \frac{x-\min x}{\max x-\min x}$ for factor x. The factor sex 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 (avoiding one-hot encoding to reduce overfitting). 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 one who has smoked 100 cigarettes or less over their entire life, and a "former smoker" as a smoker who has quit at least 6 months. The variable of family history is the number of first-degree relatives with CRC but is capped at 3, which is then mapped to values of 0, 1/3, 2/3, 1. Finally, any answer of "yes" to coronary heart disease, myocardial infarction, heart disease, and angina contributes 0.25 to the "Pooled heart conditions" field. These mapped values (rather than the raw data) are used in Table 2 for the correlation calculation.

An artificial neural network (ANN)

An artificial neural network is a network of neurons, with each neuron being equivalent to a logistic regression. A neuron's inputs, the outputs of the preceding layer's neurons, are combined in a weighted sum with a bias term. This linear function is fed into a sigmoidal ("activation") function to produce the neuron's output. The input layer consists of the model's input data (the NHIS data). The output layer returns model predictions to the user. In this work, our ANN has only one neuron in the output layer, representing an individual's risk score for CRC. Layers between the input and output layer are known as hidden layers. Therefore, an ANN is essentially a statistical regression that is nonlinear with respect to the model parameters. Note that with zero hidden layers the ANN is equivalent to logistic regression due to the logistic activation function.

Using weights W_1, W_2, W_3 and biases B_1, B_2, B_3, the ANN forward-propagates from input data X to a risk score $\bar{Y}$ by the following three compositions (indicated by parentheses) of the logistic activation function $\sigma = \sigma(z)$ having argument z (e.g., in the first logistic function, $z = B_1 + W_1X$),

$$\bar{Y} = [\text{risk score}] = \sigma(B_3 + W_3\sigma(B_2 + W_2\sigma(B_1 + W_1X))); \ \sigma(z) \equiv 1/[e^{-z} + 1]; \quad (1)$$

We used an in-house MATLAB code to minimize fitting error of (or "train") our four-layered ANN. Fig 1 shows an example of a four-layer ANN. Our ANN had 12 to 14 inputs, and each hidden layer had the same number of neurons as the input layer. The cross-entropy loss function [37] comparing the model's predictions, $-Y_i$ for the i^{th} NHIS-respondent's CRC, with the actual CRC status, Y_i (0 for never-cancer and 1 for CRC), for N respondents is,

$$[\text{loss}] = \frac{1}{N}\ln\prod_{i=1}^{N}\bar{Y}_i^{\,Y_i}(1-\bar{Y}_i)^{1-Y_i} = \frac{1}{N}\sum_{i=1}^{N}(Y_i\ln\bar{Y}_i + (1-Y_i)\ln(1-\bar{Y}_i)) \quad (2)$$

Backpropagation minimizes the fitting error numerically. It involves the chain rule derivative of Eq (1) with respect to weights W_1, W_2, W_3 and biases B_1, B_2, B_3 in iterative gradient decent with the Adam learning rate. Because it numerically minimizes the fitting error Eq (2), backpropagation plays the role of setting slope/intercept-derivatives of the sum-of-squares error equal to zero and solving for slope and intercept in one-variable linear regression.

During training we used ten-fold cross-testing to test our model, while an additional test set (NHIS year 2017 dataset) was held out for developing the stratification scheme after the ANN was trained. Because there is zero intersection [37] between the training and testing datasets, our model has TRIPOD levels [36] of 2a and 2b. The training was done with the standard backpropagation algorithm with the "Adam" learning rate. Rather than selecting a decision boundary to determine what output is considered a positive or negative cancer prediction, we just use the raw output of the ANN of Fig 1 and refer to it, loosely, as a person's risk of having colorectal cancer.

Model evaluation

The output of the ANN is a number from 0 to 1 which we treat as a risk-score. Using this output, one can decide on a score-threshold above which the result is considered a positive for CRC and otherwise is a negative result. To evaluate the performance of our ANN, we parametrically plotted the sensitivity and specificity as a function of the risk-score threshold to produce an ROC plot [39,40]. We created an analogous plot with the positive predictive value (PPV) and negative predictive value (NPV).

Stratified cross-testing [41] with ten random folds (having zero intersection [37] with the training dataset) was used to attain a TRIPOD level [36] of 2a. Two sources give statistical variance [37] in the concordance (C) statistic: the first is the number of cancer-cases within the dataset (the population variance) and the second is the number of folds of stratified cross-testing. Hanley and McNeil model the population variance as due to the risk-score being distributed normally (or as a Gaussian). The variance σ^2 is across n_{xv} stratified cross-validations. The i^{th} population-variance σ_i^2 and C-statistic A_i is assumed to have a Gaussian distribution and is thus estimated (with maximum likelihood) within a $\alpha\times100\%$ confidence interval for a population of D diseased subjects and H healthy subjects by [40],

$$\sigma_i^{\,2} = \frac{(2\,\mathrm{erf}^{-1}\alpha)^2}{HD}\left(A_i(1-A_i) + (D-1)\left(\frac{A_i}{2-A_i} - A_i^{\,2}\right) + (H-1)\left(\frac{2A_i^{\,2}}{1+A_i} - A_i^{\,2}\right)\right) \quad (3)$$

The total variance σ, assuming Gaussian distribution of population-variance [37] and a number $n_{xv} = 10$ of folds of cross-testing, is $\sigma^2 = \sum_{i=1}^{n_{xv}} (\sigma_i/n_{xv})^2 = \sigma_0^{\,2}/n_{xv}$, the average of the variances. Due to the tradeoff of having high variance [37] in the limits $n_{xv}\to N = D+H$ ("leave-one-out" stratified cross-testing, where the testing set is just a single data point and thus $D\to\{0,1\} = 1-H$) and high bias [37] in the limit $n_{xv} = 2$ ("split-sample" stratified cross-testing), the appropriate number of stratified cross-validations must be determined empirically. We thus decided upon ten-fold stratified cross-testing ($n_{xv} = 10$).

Stratifying by risk-score

To demonstrate the potential application of our ANN model in the clinic, we stratified 2017 NHIS respondents [19] into risk-categories. Based on the calculated CRC risk-score from our model, we renormalized and selected a low/medium risk boundary and a medium/high risk boundary to divide the calculated CRC risk into 3 categories. The two boundaries were chosen such that no more than 1% of 1997–2016 NHIS respondents with cancer are categorized as low risk, and no more than 1% of 1997–2016 NHIS respondents without cancer are classified as high risk.

Results

Model performance

In Fig 2, the receiver operating characteristic curves (ROC) [39,40] are plotted for the ANN trained with data from Table 2. Cross-testing with ten folds is used to emphasize the generalizability of the training. The sensitivity, specificity, and concordance [18,40] of only the testing across ten random folds (TRIPOD 2a) is reported in Fig 2. The crossed error bars in each ROC are the standard deviation across the ten folds of cross-testing. Inclusion of family history data requires removing a large number of respondents for whom this information is missing, yet gives a performance comparable to the case where all NHIS years are included. It can be seen that deviation from the mean performance of the default model is within 2 standard deviations

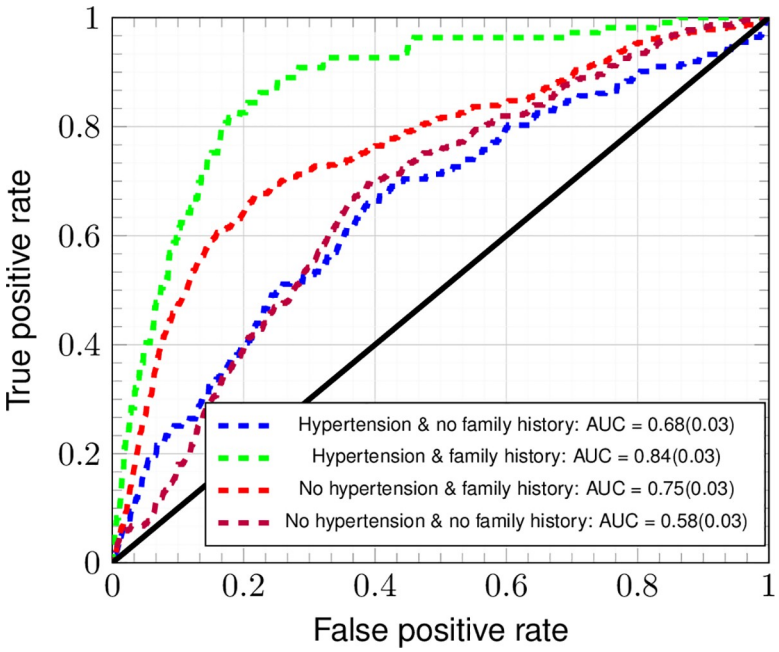

Fig 3. Cross-testing ROC curves of ANN for data non-randomly split between screened and non-screened NHIS respondents (TRIPOD 2b). Using the reduced dataset the ANN was cross-tested between the group of survey respondents (NHIS years 2000, 2005, 2010, and 2015 only) screened for CRC by colonoscopy/sigmoidoscopy and the remaining year group.

https://doi.org/10.1371/journal.pone.0221421.g003

in the error bars of Fig 2, showing insensitivity to addition or removal of the factors most strongly correlating with being recently and professionally diagnosed with CRC.

At the risk-cutoff value where their sum is maximized, the sensitivity is of 0.57 ± 0.03 and the specificity is of 0.89 ± 0.02 (mean value ± the standard deviation across the ten folds of cross-testing). Uncertainty is larger in the sensitivity compared to the specificity because of the low prevalence of CRC within the dataset. The uncertainty in the concordance has a contribution due to the population [39,40] and a contribution due to the standard deviation across folds of stratified cross-testing [41]. The latter contribution is normally distributed [37] as a result of random shuffling of the data before partitioning into folds of testing.

Cross-testing between the screened and unscreened

In Fig 3 we repeat the analysis of Fig 2, but only use respondents for whom family history data was available. This eliminates any advantage the models without family history gain from having a larger sample size. In addition, we show the performance resulting from cross-testing between the group of those who were ever screened by colonoscopy/sigmoidoscopy and the group formed by the remaining population. The green ROC (hypertension and family history) with an AUC of 0.84 outperforms the blue ROC (hypertension but no family history) with an AUC of 0.68. Similarly, the red ROC (no hypertension with family history) has an AUC of 0.75, better than the 0.58 AUC for the purple ROC (no hypertension no family history). Due to the strong correlation of CRC family history with screening history (see Table 2), inclusion of family history data sharply improves performance by greater than just a standard deviation in testing upon a population not screened by colorectal exam after training on the examined population, as reported in Fig 3. Without family history data, performance worsens by about one standard deviation of true positive rate.

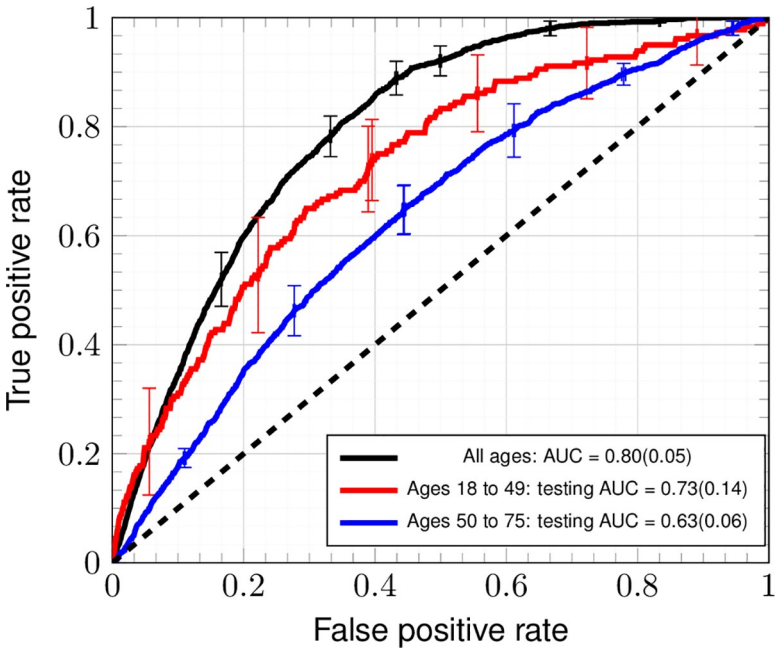

Fig 4. Cross-testing ROC curves of ANN for age groups formed by USPSTF screening guidelines. The ANN is trained by and tested upon 3 datasets: ages 18–49, ages 50–75, and all ages for the full dataset. Error bars denote the standard deviation of the TPR and FPR across ten-fold stratified cross-testing (TRIPOD level 2a).

https://doi.org/10.1371/journal.pone.0221421.g004

Performance in age groups

In Fig 4, ROC plots are given for an ANN trained and tested upon only ages 18–49, ages 50–75, and all ages for tenfold random cross-testing dataset (TRIPOD 2a). These age groups were formed because in ages 50–75 and ages 18–49, USPSTF screening guidelines [22] have respective sensitivity and specificity each of 100% (TRIPOD level 4) in flagging and not flagging a person for screening. There is good [18] mean discrimination in ages 18–49, in which there has been a recent rise in colorectal cancer incidence [11]. In contrast, in ages 50–75 the mean discrimination is only acceptable [18]. The mean is across ten folds of stratified cross-testing [41]. Due to the accompanying standard deviation across folds of cross-testing and the population-variance [40], the performance in ages 18–49 extends down into being merely acceptable, and the performance in ages 50–75 extends down into being indiscriminate [18]. Clearly, performance in cross-testing within ages 18–49 is better than the performance in cross-testing within ages 50–75.

Predictive value of the ANN

In Fig 5, the positive predictive value and false omission rate is reported for a wide variety of risk-cutoffs. The positive predictive value of the trained ANN is much lower than its negative predictive value at almost all values of the risk-cutoff, meaning that a negative call by the ANN is far more meaningful than a positive call [37]. Due to the NHIS dataset being a cross-sectional study with no follow-up, it remains possible that the false positives that drive the PPV down to such a low value are those who are of high risk and have CRC that has not yet been detected (whom it is highly desirable to screen). Barring this possibility, the ANN is better suited for making screening recommendations than functioning as a diagnostic tool, and thus is next demonstrated to calculate risk.

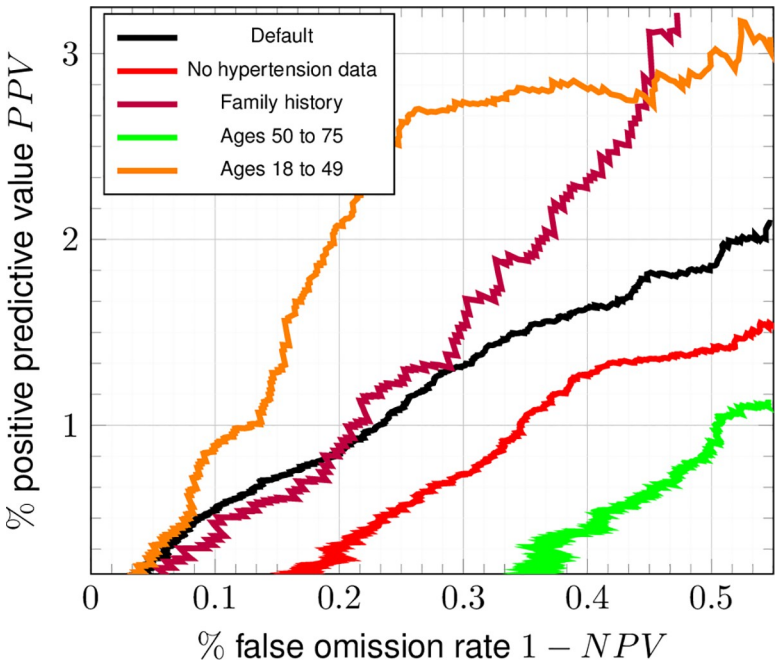

Fig 5. Diagnostic performance of the ANN for the random testing dataset (TRIPOD level 2a). Positive predictive value PPV and false omission rate were parametrically plotted in analogy to Fig 2.

https://doi.org/10.1371/journal.pone.0221421.g005

3-category risk stratification

To test its potential application in stratifying colorectal cancer risk, we ran the developed ANN with the 2017 NHIS dataset [19]. As shown in Fig 6, three categories of risk (green: low risk; yellow: medium risk; red: high risk) are stratified. The solid-lined cumulative distributions are the respective cancer/non-cancer 2017 populations correctly classified as high/low risk. The

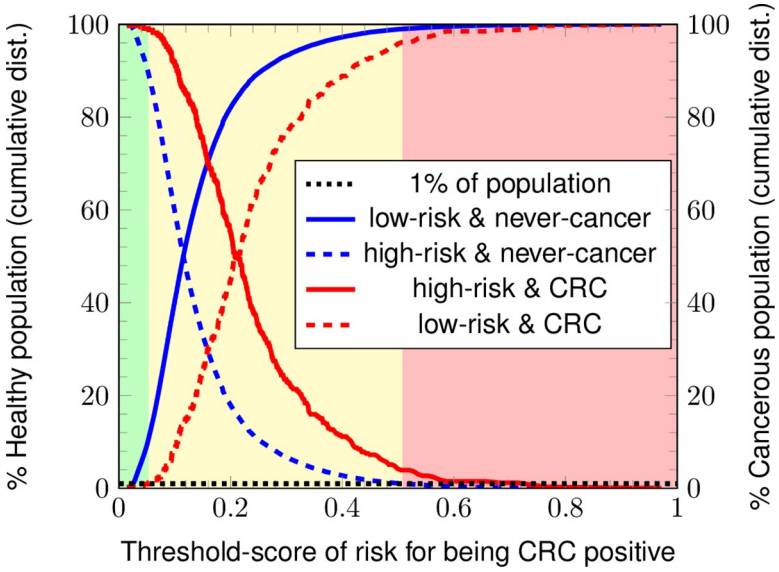

Fig 6. Risk stratification into three categories. The 2017 NHIS respondents are stratified by the ANN into three categories for CRC risk: green (low risk), yellow (medium risk), and red (high risk).

https://doi.org/10.1371/journal.pone.0221421.g006

dashed-line cumulative distributions are the respective cancer/non-cancer 2017 populations misclassified as low/high risk. The low/medium and medium/high risk boundaries are defined by requiring a 1% misclassification rate for the 1997–2016 NHIS respondents. The CRC and never-cancer 2017 NHIS respondents are misclassified at respective rates of 5% and 6%.

A summary of the results of stratifying the populations into three risk categories is given in Table 3. Our ANN classifies 8% of the cancerous population as high score and 87% of the same as medium risk with a misclassification rate of 5%. For the non-cancerous population, our ANN classifies 12% of the non-cancer population as low score and 82% of the same as medium risk with a misclassification rate of 6%. In comparison, the USPSTF guidelines misclassify 35% of the cancerous population as low risk and 53% of the non-cancerous population as high risk at TRIPOD level 4.

Discussion

Due to the low value of its positive calls, the ANN will not replace current screening methods for CRC. This can be seen by the comparison in Table 4 of the PPVs among conventional screening methods with that of the trained ANN found in Fig 5. The gold standards of colonoscopy and sigmoidoscopy are the only tests with PPV close to 1, but these tests are invasive and sometimes injurious [22]. Their lower-PPV counterparts, FIT, FOBT, and SEPT9, have varied test accuracy depending on the CRC stage [42] (a feature shared by colonoscopy and sigmoidoscopy, albeit to a lesser extent). Different levels of risk [43] as in Fig 6 could be assigned different screening methods, depending on the judgment of the clinician and the availability of resources (e.g., in some countries, the FIT is the only screening test available). An example of such a scheme for concreteness: a clinician might preferentially give colonoscopies (most expensive) to those of high risk, SEPT9 and/or FOBT tests (moderate expense) to those of medium risk, and FIT (least expensive) to those of low risk (see Table 4) [44,45]. Given the low cost, ease of mass implementation, and low invasiveness of the trained ANN, it emerges as highly attractive for stratifying risk of CRC, despite its inability to perform screening.

There are several aspects of our chosen methods that minimize the effects of potential sources of bias in the calculated risk. While the ANN is a powerful statistical tool [46], the ANN is only as good as the data used to train it, so a discussion of these biases is called for.

First, the outcome variable is defined as those NHIS respondents [19] recently and professionally diagnosed with CRC, and not necessarily those cases of CRC diagnosed by the gold predictive standard of colonoscopy (see Table 4). Unfortunately, NHIS years 2000, 2005, 2010, and 2015 only contain data on whether the respondents have ever been screened by sigmoidoscopy, colonoscopy, or proctoscopy. In lieu of such screening data for every year, Fig 3 reports testing-performance of a model trained on the screened population and tested on the remaining population. Performance without family history is within a standard deviation of true

Table 3. Comparison of ANN risk-scoring with USPSTF screening guidelines on 2017 NHIS dataset for 3-category risk-score stratification.

	# Respondents	# Low Score	% Low Score	# Medium Score	% Medium Score	# High Score	% High Score
Our ANN							
CRC (2017)	60	3	5%	52	87%	5	8%
Never-cancer (2017)	25,457	2,932	12%	20,998	82%	1,527	6%
USPSTF Guidelines							
CRC (2017)	60	21	35%	n/a	n/a	39	65%
Never-cancer (2017)	25,457	11,845	47%	n/a	n/a	13,612	53%

https://doi.org/10.1371/journal.pone.0221421.t003

Table 4. Comparison of ANN to conventional screening methods.

Screening method	Sensitivity, Specificity and/or PPV	Advantages	Disadvantages
Artificial neural network (ANN) trained with NHIS data years 1997–2016 tested on ten random splits	• Sensitivity ~ of 0.57 ± 0.03 • Specificity ~ of 0.89 ± 0.02 • PPV ~ of 0.0075 ± 0.0003	• Better performance w/more training data • Privacy • Inexpensive • Stage-independent • Can stratify risk	• Low PPV • Assumes integrity of data • Only correlation • Cannot be used for screening
Guaiac or immunoassay fecal occult blood test (gFOBT or iFOBT)	• Sensitivity ~ 0.9 • Specificity ~ 0.9 • PPV ~ 0.02	• No pre-test colon-cleansing • Privacy • Non-invasive	• Low PPV • Pre-test diet • False-positives • Depends on CRC stage • Moderately expensive
• Fecal immunochemical test (FIT) • Fecal immunochemical DNA test (FIT-DNA)	(1) For FIT: • Sensitivity ~ 0.1 • Specificity ~ 0.9 • PPV ~ 0.4 (2) For FIT-DNA: • Sensitivity ~ 0.2 • Specificity ~ 0.9 • PPV ~ 0.5	• No pre-test colon-cleansing • Privacy • Inexpensive ($14) • Non-invasive	• Adenoma insensitivity • False-positives • Low PPV • Depends on CRC stage
Methylated SEPT9 gene test	• Sensitivity ~ 0.6 at Stage I. • Sensitivity ~ 0.9 at Stage IV.	• No pre-test colon-cleansing • Privacy • Noninvasive	• Moderately expensive • Depends on CRC stage
Flexible sigmoidoscopy	• Sensitivity ~ 0.6 • Specificity ~ 0.7 • PPV ~ 0.8	• Able to perform biopsy/polypectomy • Less colon-cleansing • No sedation	• Only rectum, lower-colon • Dieting, bowel cleansing • Invasive • Expensive
Virtual colonoscopy	• Sensitivity ~ 0.6 • Specificity ~ 0.7 • PPV ~ 0.8	• Noninvasive • Sedation unneeded • Better at identifying advanced adenomas.	• Colon-cleansing • Ionizing radiation • Expensive (~$8000 in costs and charges)

https://doi.org/10.1371/journal.pone.0221421.t004

positive rate poorer, and is significantly greater with training by family history (which correlates extremely strongly with having been screened). The decrease in performance without family history data is attributable to the biases and confounders we discuss.

Another concern is that the predictor variables shown in Table 2 are marked as "ever" having occurred, while the age at which the NHIS respondent was professionally diagnosed with CRC is recorded. This is the purpose of the 4 years cutoff beyond which an instance of CRC is regarded as too long ago from the date of taking the NHIS, and is discarded, thus decreasing (if not eliminating) the probability that the predictors came before the CRC diagnosis. Previous work scoring risk of lung and skin cancer [43,47] have found their ANN insensitive to this cutoff.

The correlation of a given factor with CRC incidence (see Table 2) is a necessary but insufficient condition for that factor to be definitely causal of CRC. This is why the effect of training with and without hypertension data in the training of the ANN is indicated in Fig 2. Patients taking bevacizumab [3] for CRC often develop high blood pressure [4,5], and if this was largely responsible for the high correlation found in Table 2 it would be inappropriate to train the ANN with this. As it turns out, bevacizumab was FDA-approved [38] as a second-line treatment of metastatic CRC in 2004 and the NHIS dataset [19] of personal health data extends from years 1997–2017. The correlation of hypertension with CRC in years 1997–2003 is 3.25×10^{-2} and in years 2004–2017 decreases (unexpectedly) to 2.97×10^{-2}. It is thus possible that the predictor variables in Table 2 have their confounding with those NHIS respondents

recently and professionally diagnosed with CRC controlled to some degree. This allows inclusion of hypertension as a predictor, which is important due to the role of hypertension in CRC recurrence and mortality [48,49].

Another source of bias lies in what age the NHIS respondent was recently and professionally diagnosed with CRC. There is a bias towards screening at that age interval because this age interval was selected by the USPSTF [22], and this is seen on a histogram of CRC incidence vs. age of diagnosis where it is observed that about 60% - 70% of CRC cases occur in ages 50–75. Clearly, the diagnosis came at the time of screening, so the screening time could be far after the time the CRC first nucleated. This may call for an augmentation of the NHIS survey with a question about the extent of the CRC's advancement when it was first detected by screening.

The last source of bias is due to discarding any NHIS respondent with one or more blank entries as a result of answers of "refused", "not ascertained", or "don't know" (see NHIS documentation [19]). It is speculated that responses of "not ascertained" and "don't know" belong to data missing completely at random, while the response of "refused" belongs to data missing at random [50]. For instance, an answer of "refused" to the factor of a subject's smoking habits may mean smoking of illegal substances. If this is true, then our model would carry the bias [37] of excluding (and thus less accurately scoring CRC risk in) the corresponding members of the population. A similar bias results from the model excluding those having CRC more than 4 years from the year they answered the NHIS, which are also discarded. Future work would draw upon imputation techniques allowing for systematic treatment of such missing entries, thus avoiding the loss of an entire survey-respondent. This is especially critical in fields with larger quantities of missing data, such as family history of CRC heavily relied upon by clinicians to make screening recommendations.

We now discuss the ANN architecture. An ANN with two hidden layers is selected due to its being the smallest ANN that can learn low-degree polynomial functions [51]. This allows the ANN to deal with noise. The ANN's good performance in spite of the bias and confounding discussed above could be attributable to it viewing these as noise. The cross-tested C-statistic of logistic regression [18] (ROC not shown) is 0.60 ± 0.03, suggesting the importance of inter-factor coupling whose incorporation is made possible by a second hidden layer [21]. In the ANN, factors are fed into any one trained neuron of the first hidden layer, linearly combined by the weights, and composed with the sigmoid function. Each neuron of the first hidden layer thus takes on a value that is the pooled effect of all factors of the input-layer. Each neuron of the second hidden layer receives these various pooled values, and thus is the layer where the trained weights incorporate inter-factor couplings before being fed into the single output-neuron (CRC risk score). If the second hidden layer is made of a single neuron, it becomes equivalent to the output neuron and the calculated risk incorporates no inter-factor coupling. This gives a C-statistic of ~0.5, which means the ANN is almost maximally indiscriminant [18]. We interpret this (as well as the insensitivity of the C-statistic to presence/absence of data on hypertension and family history reported in Fig 2 and Fig 5) as indicative of how much more important inter-factor coupling in one's risk of CRC is compared to factors acting in isolation (which the process of naïve one-by-one variable selection assumes). The insensitivity of the ANN to being trained or not trained with cancer-family-history data (NHIS years 2000, 2005, 2010, and 2015 only) shown in Fig 2 might also be attributable to inter-factor coupling [21].

A model of CRC risk within the NHIS dataset [19] has been developed thusly. The model shows predictive power in a general demographic in Fig 2. The resulting concordance is 0.80 ± 0.05, and thus is competitive with that of Kaminski et al.[15] (which incorporates family history of CRC as well as regular aspirin use [52]) and with that of the highest-performing models (among 11) [15] in a recent review of MEDLINE, Scopus, and Cochrane Library

databases from January 1990 through March 2013 [17]. In Fig 3, because performance improves by greater than one standard deviation of true positive rate on including family history data and worsens by one standard deviation of true positive rate, the worsening of performance can be attributable to the sources of bias described above (taking colorectal exams to be the gold standard of diagnosis as in Table 4). The trained ANN retains its predictive power even within ages 18–49: in Fig 4, the standard deviation of true positive rate approximately reaches the performance at all ages. This suggests promise in performance in the demographic not flagged for screening by age. Although the ANN's positive calls are almost meaningless (see Fig 5), it is usable as a risk stratification tool (see Fig 6 and Table 3) to assist clinicians.

Future work will study generalizability. The present work omits a calibration plot, and just report discrimination due to this being a development (rather than validation) study. Future work will study training/testing with the NHIS data and testing/training upon a separate dataset to determine the effect of disparate risks of CRC between datasets upon model performance. Advanced techniques such as validation-based early stopping and/or dropout are also to be used.

A future aim is to implement this system of risk scoring in a software-application (an "app") that a smartphone can run. The application will be along the lines of CT Gently [53], an application previously developed. The user of the application will answer NHIS survey questions and immediately receive the scoring of their risk for CRC, in much the same manner as a current website [54] hosting an algorithm underpinning the corresponding published [55] model. Moreover, the application will simulate immediate adjustments in the user's personal health habits: for instance, the user will be able to drag a sliding-bar representing their smoking habits from high to low and see their score of CRC-risk drop in response to quitting smoking. Alternatively, the ANN could retrieve electronic medical records in a clinical setting and provide immediate risk-scoring during consultation.

Conclusion

A multi-parameterized artificial neural network was developed to score risk of colorectal cancer based solely on personal health data. The trained ANN has been robustly tested per TRIPOD level 2a and level 2b protocols. The concordance of the ANN is comparable to that of current methods of scoring CRC risk (including those using biomarkers). The ANN outperforms logistic regression, suggesting the importance of inter-factor coupling. The low positive predictive value indicates unsuitability of the ANN to replace conventional screening methods. Nevertheless, in comparison to USPSTF guidelines, the trained ANN can stratify individual's colorectal cancer risk more accurately for more effective screening and intervention. As the ANN is built from self-reportable personal health data, it can be easily implemented on a mobile platform for more widespread applications.

Acknowledgments

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. DAR was funded by R01EB022589 while writing the first working version of the ANN, which BJN et al continued under R01EB022589 while DAR went on to be supported with salary by Sun Nuclear Corporation. Sun Nuclear Corporation did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific role of DAR is articulated in the 'author contributions' section. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or of Sun Nuclear Corporation.

Author Contributions

Conceptualization: Gregory R. Hart, David A. Roffman, Xavier Llor, Issa Ali, Wazir Muhammad, Ying Liang, Jun Deng.

Formal analysis: Bradley J. Nartowt.

Funding acquisition: Jun Deng.

Investigation: Bradley J. Nartowt, Gregory R. Hart, Xavier Llor, Issa Ali, Wazir Muhammad, Ying Liang.

Methodology: Gregory R. Hart, Wazir Muhammad, Ying Liang.

Project administration: Jun Deng.

Resources: Jun Deng.

Software: Bradley J. Nartowt, David A. Roffman.

Supervision: Jun Deng.

Validation: Bradley J. Nartowt.

Visualization: Bradley J. Nartowt.

Writing – original draft: Bradley J. Nartowt.

Writing – review & editing: Bradley J. Nartowt, Gregory R. Hart, David A. Roffman, Xavier Llor, Issa Ali, Wazir Muhammad, Ying Liang, Jun Deng.

References

1. What is Colorectal Cancer? [Internet]. 2018 [cited 2018 Oct 12]. p. 2. Available from: https://www.cancer.org/cancer/colon-rectal-cancer/about/what-is-colorectal-cancer.html
2. National Cancer Institute. Cancer Stat Facts: Colorectal Cancer [Internet]. 2018. p. 2. Available from: http://seer.cancer.gov/statfacts/html/colorect.html
3. Hurwitz H, Fehrenbacher L, Novotny W, Cartwright T, Hainsworth J, Heim W, et al. Bevacizumab plus Irinotecan, Fluorouracil, and Leucovorin for Metastatic Colorectal Cancer. N Engl J Med [Internet]. 2004; 350(23):2335–42. Available from: https://www.ncbi.nlm.nih.gov/pubmed/15175435 https://doi.org/10.1056/NEJMoa032691 PMID: 15175435
4. Scartozzi M, Galizia E, Chiorrini S, Giampieri R, Berardi R, Pierantoni C, et al. Arterial hypertension correlates with clinical outcome in colorectal cancer patients treated with first-line bevacizumab. Ann Oncol. 2009; 20(2):227–30. https://doi.org/10.1093/annonc/mdn637 PMID: 18842611
5. Syrigos KN, Karapanagiotou E, Boura P, Manegold C, Harrington K. Bevacizumab-induced hypertension: Pathogenesis and management. BioDrugs. 2011; 25(3):159–69. https://doi.org/10.2165/11590180-000000000-00000 PMID: 21627340
6. Aleksandrova K, Pischon T, Jenab M, Bueno-de-Mesquita HB, Fedirko V, Norat T, et al. Combined impact of healthy lifestyle factors on colorectal cancer: A large European cohort study. BMC Med. 2014; 12(1).
7. Usher-Smith JA, Walter FM, Emery JD, Win AK, Griffin SJ. Risk prediction models for colorectal cancer: A systematic review. Cancer Prev Res. 2016; 9(1):13–26.
8. Betés M, Muñoz-Navas MA, Duque JM, Angós R, Macías E, Súbtil JC, et al. Use of Colonoscopy as a Primary Screening Test for Colorectal Cancer in Average Risk People. Am J Gastroenterol. 2003; 98(12):2648–54. https://doi.org/10.1111/j.1572-0241.2003.08771.x PMID: 14687811
9. Chen G, Mao B, Pan Q, Liu Q, Xu X, Ning Y. Prediction rule for estimating advanced colorectal neoplasm risk in average-risk populations in southern Jiangsu Province. Chin J Cancer Res [Internet]. 2014; 26(1):4–11. Available from: papers://5aecfcca-9729-4def-92fe-c46e5cd7cc81/Paper/p95692 https://doi.org/10.3978/j.issn.1000-9604.2014.02.03 PMID: 24653621
10. 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):257–63. https://doi.org/10.1016/j.amjmed.2006.05.055 PMID: 17349449

11. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017. CA Cancer J Clin [Internet]. 2017; 67(1):7–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28055103 https://doi.org/10.3322/caac.21387 PMID: 28055103

12. Shaukat A, Dostal A, Menk J, Church TR. BMI Is a Risk Factor for Colorectal Cancer Mortality. Dig Dis Sci. 2017; 62(9):2511–7. https://doi.org/10.1007/s10620-017-4682-z PMID: 28733869

13. Doleman B, Mills KT, Lim S, Zelhart MD, Gagliardi G. Body mass index and colorectal cancer prognosis: a systematic review and meta-analysis. Tech Coloproctol. 2016; 20(8):517–35. https://doi.org/10.1007/s10151-016-1498-3 PMID: 27343117

14. Ahmed FE. Artificial neural networks for diagnosis and survival prediction in colon cancer. Mol Cancer. 2005; 4:1–12. https://doi.org/10.1186/1476-4598-4-1

15. Kaminski MF, Polkowski M, Kraszewska E, Rupinski M, Butruk E, Regula J. A score to estimate the likelihood of detecting advanced colorectal neoplasia at colonoscopy. Gut. 2014; 63(7):1112–9. https://doi.org/10.1136/gutjnl-2013-304965 PMID: 24385598

16. Yeoh KG, Ho KY, Chiu HM, Zhu F, Ching JYL, Wu DC, et al. The Asia-Pacific Colorectal Screening score: A validated tool that stratifies risk for colorectal advanced neoplasia in asymptomatic Asian subjects. Gut. 2011; 60(9):1236–41. https://doi.org/10.1136/gut.2010.221168 PMID: 21402615

17. Ma GK, Ladabaum U. Personalizing Colorectal Cancer Screening: A Systematic Review ofModels to Predict Risk of Colorectal Neoplasia. Clin Gastroenterol Hepatol [Internet]. 2014; 12(10):1624–34. Available from: https://doi.org/10.1016/j.cgh.2014.01.042 PMID: 24534546

18. Hosmer, David W.; Lemeshow S. Applied Logistic Regression. 2000.

19. National Health Interview Survey dataset of personal health (1997–2017) [Internet]. 2018. Available from: https://www.cdc.gov/nchs/nhis/

20. Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. Cancer [Internet]. 2001; 91 Suppl 8:1636–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11309761

21. Jack V. Tu. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol. 1996; 49(11):1225–31. PMID: 8892489

22. Bibbins-Domingo K, Grossman D, Curry S, Davidson K, Epling, JW J, García F, et al. Screening for colorectal cancer. Semin Oncol. 2017; 44(1):34–44. https://doi.org/10.1053/j.seminoncol.2017.02.002 PMID: 28395761

23. American Cancer Society. What Is Colorectal Cancer? [Internet]. [cited 2018 Dec 10]. Available from: https://www.cancer.org/cancer/colon-rectal-cancer/about/what-is-colorectal-cancer.html

24. Bnard F, Brkun AN, Martel M, Von Renteln D. Systematic review of colorectal cancer screening guidelines for average-risk adults: Summarizing the current global recommendations. World J Gastroenterol. 2018; 24(1):124–38. https://doi.org/10.3748/wjg.v24.i1.124 PMID: 29358889

25. Bujanda L, Sarasqueta C, Hijona E, Hijona L, Cosme A, Gil I, et al. Colorectal cancer prognosis twenty years later. World J Gastroenterol. 2010; 16(7):862–7. https://doi.org/10.3748/wjg.v16.i7.862 PMID: 20143465

26. Pineau BC, Paskett ED, Chen GJ, Durkalski VL, Espeland MA, Vining DJ. Validation of virtual colonoscopy in the detection of colorectal polyps and masses: Rationale for proper study design. Int J Gastrointest Cancer [Internet]. 2001; 30(3):133–40. Available from: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L36306728%5Cnhttp://sfx.library.uu.nl/utrecht?sid=EMBASE&issn=01694197&id=doi:&atitle=Validation+of+virtual+colonoscopy+in+the+detection+of+colorectal+polyps+and+masses%3A+Rationale+ https://doi.org/10.1385/IJGC:30:3:133 PMID: 12540025

27. Pineau BC, Paskett ED, Chen GJ, Espeland MA, Phillips K, Han JP, et al. Virtual colonoscopy using oral contrast compared with colonoscopy for the detection of patients with colorectal polyps. Gastroenterology. 2003; 125(2):304–10. https://doi.org/10.1016/s0016-5085(03)00885-0 PMID: 12891529

28. Irvine EJ, O'Connor J, Frost RA, Shorvon P, Somers S, Stevenson GW, et al. Prospective comparison of double contrast barium enema plus flexible sigmoidoscopy v colonoscopy in rectal bleeding: Barium enema v colonoscopy in rectal bleeding. Gut. 1988; 29(9):1188–93. https://doi.org/10.1136/gut.29.9.1188 PMID: 3273756

29. Mandel JS, Bond JH, Church TR, Snover DC, Bradley GM, Schuman LM, et al. Reducing Mortality from Colorectal Cancer by Screening for Fecal Occult Blood. N Engl J Med [Internet]. 1993; 328(19):1365–71. Available from: https://doi.org/10.1056/NEJM199305133281901 PMID: 8474513

30. Allison JE, Sakoda LC, Levin TR, Tucker JP, Tekawa IS, Cuff T, et al. Screening for colorectal neoplasms with new fecal occult blood tests: Update on performance characteristics. J Natl Cancer Inst. 2007; 99(19):1462–70. https://doi.org/10.1093/jnci/djm150 PMID: 17895475

31. Imperiale TF, Ransohoff DF, Itzkowitz SH, Levin TR, Lavin P, Lidgard GP, et al. Multitarget Stool DNA Testing for Colorectal-Cancer Screening. N Engl J Med [Internet]. 2014; 370(14):1287–97. Available from: https://doi.org/10.1056/NEJMoa1311194 PMID: 24645800

32. Lofton-Day C, Model F, DeVos T, Tetzner R, Distler J, Schuster M, et al. DNA methylation biomarkers for blood-based colorectal cancer screening. Clin Chem. 2008; 54(2):414–23. https://doi.org/10.1373/clinchem.2007.095992 PMID: 18089654

33. Aaltonen LA, Peltomäki P, Leach FS, Sistonen P, Pylkkänen L, Mecklin JP, et al. Clues to the pathogenesis of familial colorectal cancer. Science (80-). 1993; 260(5109):812–6. https://doi.org/10.1126/science.8484121 PMID: 8484121

34. Sharara N, Nolan S, Sewitch M, Martel M, Dias M, Barkun AN. Assessment of a Colonoscopy Triage Sheet for Use in a Province-Wide Population-Based Colorectal Screening Program. Can J Gastroenterol Hepatol. 2016;2016.

35. Paterson WG, Depew WT, Paré P, Petrunia D, Switzer C, Veldhuyzen van Zanten SJ, et al. Canadian consensus on medically acceptable wait times for digestive health care. Can J Gastroenterol. 2006; 20(6):411–23. https://doi.org/10.1155/2006/343686 PMID: 16779459

36. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. Eur Urol. 2015; 67(6):1142–51. https://doi.org/10.1016/j.eururo.2014.11.025 PMID: 25572824

37. Bertsekas DP, Tsitsiklis JN. Introduction to Probability. 2nd ed. Athena Scientific; 1998.

38. Cohen MH, Gootenberg J, Keegan P, Pazdur R. FDA Drug Approval Summary: Bevacizumab Plus FOLFOX4 as Second-Line Treatment of Colorectal Cancer. Oncologist [Internet]. 2007; 12(3):356–61. Available from: https://doi.org/10.1634/theoncologist.12-3-356 PMID: 17405901

39. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Casp J Intern Med. 2013; 4(2):627–35.

40. Hanley JA, Mcneil J. under a Receiver Characteristic. Radiology [Internet]. 1982; 143:29–36. Available from: http://radiology.rsna.org/content/143/1/29.full.pdf https://doi.org/10.1148/radiology.143.1.7063747 PMID: 7063747

41. Picard RR, Cook RD. of Regression Models Cross-Validation. J Am Stat Assoc. 2012; 79(387):575–83.

42. Song L-L, Li Y-M. Current noninvasive tests for colorectal cancer screening: An overview of colorectal cancer screening tests. World J Gastrointest Oncol [Internet]. 2016; 8(11):793. Available from: http://www.wjgnet.com/1948-5204/full/v8/i11/793.htm https://doi.org/10.4251/wjgo.v8.i11.793 PMID: 27895817

43. Roffman D, Hart G, Girardi M, Ko CJ, Deng J. Predicting non-melanoma skin cancer via a multi-parameterized artificial neural network. Sci Rep [Internet]. 2018; 8(1):1–7. Available from: https://doi.org/10.1038/s41598-017-17765-5

44. Quintero E, Castells A, Bujanda L. Colonoscopy versus Fecal Immunochemical Testing in Colorectal-Cancer Screening. Gastroenterol Endosc. 2012; 54(4):1510.

45. Tests to Detect Colorectal Cancer and Polyps [Internet]. 2018. Available from: https://www.cancer.gov/types/colorectal/screening-fact-sheet

46. Cruz JA, Wishart DS. Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Inform [Internet]. 2006; 2:59–77. Available from: https://www.ncbi.nlm.nih.gov/pubmed/19458758

47. Hart GR, Roffman DA, Decker R, Deng J. A multi-parameterized artificial neural network for lung cancer risk prediction. PLoS One. 2018; 13(10):1–13.

48. Lin CC, Huang KW, Luo JC, Wang YW, Hou MC, Lin HC, et al. Hypertension is an important predictor of recurrent colorectal adenoma after screening colonoscopy with adenoma polypectomy. J Chinese Med Assoc [Internet]. 2014; 77(10):508–12. Available from: http://dx.doi.org/10.1016/j.jcma.2014.03.007

49. Van De Poll-Franse L V., Haak HR, Coebergh JWW, Janssen-Heijnen MLG, Lemmens VEPP. Disease-specific mortality among stage I-III colorectal cancer patients with diabetes: A large population-based analysis. Diabetologia. 2012; 55(8):2163–72. https://doi.org/10.1007/s00125-012-2555-8 PMID: 22526616

50. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Vol. 43, Population (French Edition). 1988. 1174 p.

51. Andoni A, Panigrahy R, Valiant G, Zhang L. Learning polynomials with neural networks. Proc 31st Int Conf Mach Learn. 2014;(32).

52. Rothwell PM, Wilson M, Elwin CE, Norrving B, Algra A, Warlow CP, et al. Long-term effect of aspirin on colorectal cancer incidence and mortality: 20-year follow-up of five randomised trials. Lancet. 2010; 376(9754):1741–50. https://doi.org/10.1016/S0140-6736(10)61543-7 PMID: 20970847

53. Deng J, Ming X, Zhang Y, Zhou L, Zhang Y, Wu H, et al. CT Gently: Personalizing CT and CBCT Imaging for the Children. SAJ Cancer Sci [Internet]. 2014; 1(1):1–5. Available from: http://fulltext.scholarena.com/CT-Gently-Personalizing-CT-and-CBCT-Imaging-for-the-Children.php

54. Hippisley-Cox J, Coupland C. QCancer [Internet]. 2017. Available from: https://www.qcancer.org/

55. Chiang PPC, Glance D, Walker J, Walter FM, Emery JD. Implementing a QCancer risk tool into general practice consultations: an exploratory study using simulated consultations with Australian general practitioners. Br J Cancer [Internet]. 2015; 112(s1):S77–83. Available from: http://dx.doi.org/10.1038/bjc.2015.46