# Derivation and Validation of a Predictive Model for Advanced Colorectal Neoplasia in Asymptomatic Adults

Thomas F. Imperiale, MD[1,3,4], Patrick O. Monahan, PhD[2], Timothy E. Stump, MA[2], David F. Ransohoff, MD[5]

From the Division of Gastroenterology and Hepatology, Department of Medicine[1], and Department of Biostatistics[2], Indiana University School of Medicine; The Center for Innovation, Health Services Research and Development, Roudebush VA Medical Center[3], and Regenstrief Institute[4]; and Department of Medicine[5], University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Short title: Risk model for advanced colorectal neoplasia

Word count (abstract): 250

Word count (manuscript): 3202

Grant support: This work was supported by the National Cancer Institute (R01-CA104459); the Walther Cancer Institute, Indianapolis, IN; The Indiana University Melvin and Bren Simon Cancer Center; and a project development team within the Indiana Clinical and Translational Sciences Institute (grant UL1TR001108) from the National Center for Research Resources, National Institutes of Health, Indianapolis, IN.

Key words: colorectal cancer screening, cancer prevention, colonoscopy, risk stratification, risk prediction

Corresponding author:

Thomas F. Imperiale, MD
Indiana University Medical Center
Regenstrief Institute, Inc.
1101 West 10th Street
Indianapolis, IN 46202
TEL: 317-274-9046
FAX: 317-274-9304
Email: timperia@iu.edu

---

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

Imperiale, T. F., Monahan, P. O., Stump, T. E., & Ransohoff, D. F. (2021). Derivation and validation of a predictive model for advanced colorectal neoplasia in asymptomatic adults. Gut, 70(6), 1155–1161. https://doi.org/10.1136/gutjnl-2020-321698

**ABSTRACT**

**Background:** A system to estimate risk for advanced colorectal neoplasia (AN) could help direct patients and providers to choices among the various screening tests, improving screening efficiency and uptake. We aimed to create a comprehensive risk prediction model for AN.

**Methods:** Average-risk 50-80 year-olds undergoing 1st-time screening colonoscopy were recruited from selected endoscopy units in central Indiana. We measured socio-demographic and physical features, medical and family history, and lifestyle factors, and linked these to the most advanced finding. We derived a risk equation on 2/3s of the sample, and assigned points to each variable to create a risk score. Scores with comparable risks were collapsed into risk categories. The model and score were tested on the remaining sample.

**Findings:** Among 3,025 subjects in the derivation set (mean age 57·3 (6·5) years; 52% women), AN prevalence was 9·4% (including 26 CRCs). The 12-variable model (c-statistic=0.77) resulted in 3 risk groups with AN risks of 1.5% (95% CI, 0·72-2·74%), 7·06% (CI, 5·89-8·38%) and 27·26% (CI, 23·47-31·30%) in the low-, intermediate-, and high-risk groups (P-value for trend < 0·001), with 23%, 59%, and 18% of subjects, respectively. In the validation set of 1,475 subjects with AN prevalence of 8·4%, model performance was comparable (c-statistic=0·78, with AN risks of 2·73% (CI, 1·25-5·11%), 5·57% (CI, 4·12-7·34%) and 25·79% (CI, 20·51-31·66%) in the low-, intermediate- and high-risk subgroups, respectively (P<0·001), with proportions of 23%, 59%, and 18%,

**Interpretation:** This model identifies sizeable low-risk and high-risk subgroups for which non-invasive screening and colonoscopy, respectively, may be preferable.

**Funding:** This work was supported by the National Cancer Institute (R01-CA104459); the Walther Cancer Institute, Indianapolis, IN; The Indiana University Melvin and Bren Simon Cancer Center; and a project development team within the Indiana Clinical and Translational Sciences Institute (grant UL1TR001108) from the National Center for Research Resources, National Institutes of Health, Indianapolis, IN.

## RESEARCH IN CONTEXT

### Evidence before this study

Recent studies have suggested that colorectal cancer screening be personalized based on either future risk for colorectal cancer or current risk for advanced neoplasia (the combination of colorectal cancer and advance, precancerous polyps). While several risk prediction models are available for estimating current risk for advanced neoplasia, they achieve limited risk separation, have only fair discrimination, or have limited generalizability.

### Added value of this study

We found that a combination of socio-demographic, physical, and lifestyle features provided good discrimination for risk of advanced neoplasia. In low, intermediate, and high-risk groups comprising 23%, 59%, and 18% of the study sample, respectively, risk of advanced neoplasia was 1.5%, 7.1%, and 26.3%, respectively. In the validation subset the model demonstrated good calibration (P=0.69) and good discrimination (c-statistic = 0·78).

### Implications of all the available evidence

This tool may be used to more precisely personalize an individual's risk for advanced neoplasia and to help decide how to be screened for colorectal cancer.

## INTRODUCTION

Screening for colorectal cancer reduces incidence and mortality from this third most common cancer and second most lethal cancer in the U.S.,(1) and although CRC screening is effective and cost-effective, it is expensive, costing the U.S. health care system billions of dollars annually.(2, 3) Despite having several test options, the U.S. population continues to underutilize screening,(2, 3) as just 60-65% of the screen-eligible population is current with screening.(4) Further, CRC screening is often inefficient, with low-risk persons having colonoscopy with low-yield and little benefit while high-risk persons may have non-invasive testing that may miss CRC and in particular, advanced precancerous polyps.

In the current era of personalized medicine, tailoring colorectal cancer screening based on some risk measure has potential to improve both the uptake and efficiency of screening. During the past decade or so, several instruments have been derived that predict either the future risk of CRC or current risk for advanced precancerous polyps or the combination of CRC and advanced polyps (which is referred to as advanced neoplasia).(5-7) Of those that predict current risk for advanced neoplasia, several have only fair discrimination (with c-statistics ranging from 0·62 to 0·77), a small risk difference between low-risk and high-risk groups, small sizes of the low-risk and high-risk groups, uncertain generalizability to the U.S. population, or are limited in their extent of validation.

We hypothesized that incorporating additional risk factors could improve the magnitude of discrimination, the risk gradient between groups, or both. In previous work, we published a 5-item prediction tool using the 5 most frequent factors associated with advanced neoplasia: age, sex, family history of CRC, cigarette smoking, and waist circumference. While the model had good discrimination (c-statistic of 0·77) and was validated by split-sample,(8) it is not known whether a more discriminating model is possible to create by including other factors. In the current study, we considered a larger set of candidate factors on a large population of average-risk person undergoing first-time screening colonoscopy to derive and test a predictive risk model for advanced colorectal

neoplasia. We include advanced adenomas in addition to CRC because they are believed to be the immediate precursor lesion to most CRC, despite their uncertain risk and rate of progression to CRC.(9, 10)

## METHODS

This study was conducted at the Indiana University Medical Center in Indianapolis, IN, and was approved by the Institutional Review Board of Indiana University (Indianapolis, IN). Methods and results are reported according to STROBE (strengthening the Reporting of Observational Studies in Epidemiology)(11) and TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis)(12) guidelines.

Methods for study population recruitment, study procedures, and data management have been reported previously.(8) In brief, from December 2004 to September 2011 we recruited persons ages 50 to 80 years undergoing first-time screening colonoscopy. We initially targeted two large companies providing screening colonoscopy as a preventive health benefit to their employees, retirees, and their dependents; however, due to saturated uptake, we extended recruitment to Indianapolis Gastroenterology and Hepatology, a large single-specialty practice in Indianapolis, and to Wishard Memorial Hospital, the Richard L. Roudebush Veterans Affairs Medical Center, and Margaret Mary Community Hospital, affiliates of the Indiana University Medical Center. We excluded persons with inflammatory bowel disease, those with a high-risk family history, and those reporting a history of polyps that required surveillance colonoscopy.

Prior to colonoscopy, participants were asked to complete a mailed, 12-page, 50-item survey on sociodemographic features, family history, personal medical history, lifestyle habits, and medication use. Participants were asked to measure and record their height, weight, and waist and hip circumferences, the latter using a 72-inch tape measure included with the survey. Just prior to the colonoscopy, nursing personnel recorded the four physical measures. These were used preferentially

in the analysis; patient recorded physical measures were used when nursing personnel measures were unavailable. Colonoscopy was performed in standard fashion at each site based on their individual protocols.

Trained research assistants reviewed all surveys for completeness, contacting study personnel by telephone as required for clarification and completion of study items. The research assistants also obtained, reviewed, and coded all colonoscopy and pathology reports blinded to survey information and with blinded review of coding by a study investigator (TFI) as needed. The most advanced finding and its location within the colorectum were linked to survey items, with proximal location beginning at the splenic flexure. All data were scanned into a de-identified database with unique identifier numbers. Although no endoscopist-specific colonoscopy quality metrics were recorded as part of this study, previous data from a large segment of participating endoscopists suggested very high rates of examinations to the cecum.(13)

Data were managed by the Department of Biostatistics, Indiana University School of Medicine (Indianapolis, IN). Study sample size estimate was based on ensuring an adequate number of participants with advanced neoplasia, defined as colorectal cancer or an advanced precancerous polyp. Advanced precancerous polyps included adenomas or serrated polyps ≥ 1 cm or one with villous histology or high-grade dysplasia.

**Statistical Analysis**

Subjects with complete data (96·2% of the total sample) were used for model development and validation. There were no demographic differences between subjects with complete versus incomplete data. A stratified random sample comprising two-thirds of the original sample was obtained using PROC SURVEYSELECT and its STRATA option in SAS, version 9.3 (SAS Institute).(14) Strata were created using the following variables: advanced neoplasia, sex, family history of cancer, and body mass index (BMI). Participants in two-thirds of this data set were used as the derivation set, and the other one third of participants were used as a validation set. The data set

was divided in this way to ensure a random but equal distribution of the stratification variables in both derivation and validation data subsets. The goal of the analysis was to identify the best-performing model using all candidate variables, one that would discriminate well between participants with versus without advanced neoplasia. We used age- and gender-adjusted logistic regression models on the derivation set. Alternative choices for clinically-sensible categories, when they existed for risk factors (e.g., 0 to < 20 pack-years and ≥ 20 pack-years; versus > 0 to < 30 pack-years and ≥ 30 pack years) were compared in these age- and gender-adjusted models with the omnibus likelihood ratio chi-square test to identify categories for each variable that optimized discrimination for AN. Likewise, decisions about which risk variables should be categorized versus kept continuous (e.g., number of servings of red meat per week) were made using the age-and gender-adjusted omnibus likelihood ratio chi-square test. After category optimization of each variable, the final pool included 18 candidate variables (see Appendix - Methods and Table). Best subsets multiple logistic regression(15) was then used; its output generated all possible models having 1 variable, 2 variables, 3 variables and so on up to the single model including all candidate variables. All models were then ranked from best to worst based on optimal model metrics, including the smallest Mallow's C,(16) Akaike information criterion (AIC) and the greatest frequency with which each variable appeared in the top 20 models. As a sensitivity analysis, backward deletion logistic regression was performed to determine whether its final model was the same as the best model selected using the procedure described above. Model calibration and discrimination were measured with the Hosmer-Lemeshow goodness-of-fit test and c-statistic, respectively.

Based on the model selected from the derivation subset, we created a risk scoring system using an approach described by Sullivan and colleagues.(17) A reference value for each risk variable was defined as 1 for indicator variables and as the mid-point for continuous variables. The log-odds coefficients were re-scaled into a point system in which 1 point was equivalent to the increase in risk for AN associated with a 5-year increase in age. Points were then rounded to the nearest integer. The base category for each risk variable was assigned 0 points. Consistent with previous work, we

collapsed the resulting total point score (with possible risk score range of -14 to 22 points) into a smaller number (low, intermediate, high) of risk categories with clinically similar risk estimates. The chi-square test for trend was used to test for a trend in the risk for (or prevalence of) advanced neoplasia across the three risk categories. We specified this model, its point values, the three risk categories and cutoffs as final, and then tested its performance in the validation data set, in which model calibration and discrimination were assessed, and the chi-square test for trend was performed. SAS, version 9.3 was used for all analyses.

**Role of the Funding Source**

The funding source had no role in the design, conduct, or analysis of the study or the decision to submit the manuscript for publication.

**RESULTS**

We enrolled 4500 eligible consenting persons into the study between December 2004 and September of 2011. Demographic and clinical variables of the derivation (n=3025) and validation (n=1475) subgroups were comparable, including the proportion of participants with complete data who were included in the analyses (Table 1). The study cohort was 94% white.

Variables used in the final multivariable model are shown in Table 2, which includes the criteria / definition for each variable, the prevalence associated with increased or decreased risk for advanced neoplasia, and both P-values and odds ratios (95% CI). Age, male sex, cigarette smoking, significant ethanol use, metabolic syndrome and red meat consumption were associated with increased risk for advanced neoplasia, with odds ratios ranging from 1·05 to 2·36 (Table 2). Being married or living with a significant other, advanced education, regular use of aspirin and non-steroidal anti-inflammatory drugs, regular activity, moderate and vigorous physical activity were associated with decreased risk for advanced neoplasia, with odds ratios ranging from 0·37 to 0·99. The model

demonstrated good discrimination (c-statistic = 0·77) and calibration (P=0·38). Points given to different values for each variable are shown in Table 3. Scores for the risk index ranged from -13 to +13. We computed the risk for advanced neoplasia for each score and collapsed scores with numerically similar risks into risk categories of low, intermediate, and high risk for advanced neoplasia. These three risk groups showed respective risks for advanced neoplasia of 1·5% (CI, 0·72-2·74%), 7·06% (CI, 5·89-8·38%), and 27.26% (CI, 23·47-31·30%) (P-value for trend < 0.001), and respective cohort proportions of 23%, 59%, and 18%. (Table 4). Ten low-risk participants had advanced neoplasia; however, none had colorectal cancer. Six of 10 low-risk participants with advanced neoplasia had a distal location. Based on finding a distal polyp, sigmoidoscopy would have detected 7 (70%) of these advanced polyps anywhere in the colorectum (Table 5).

In the validation subset of 1475 participants, the model demonstrated good calibration (P=0.69) and good discrimination (c-statistic = 0·78). Using the point system created from the derivation subset, risks for advanced neoplasia in the validation subset were similar to those of the derivation subset: 2·73% (CI, 1·25-5·11%) in the low-risk group, 5·57% (CI, 4.12-7·34%) in the intermediate-risk group, and 25·79% (CI, 20.51-31.66%) in the high-risk group (P-value for trend < 0·001), with respective sample proportions of 23%, 59%, and 18%, which was identical to the derivation subset. Nine low-risk subjects had advanced neoplasia, none of whom had colorectal cancer. Five of the nine subjects had distal lesions, and sigmoidoscopy for a distal polyp would have detected 6 of 9 persons with advanced neoplasia.

## DISCUSSION

In this study, we used phenotypic features to derive a risk prediction model and scoring system that estimates the current risk for advanced colorectal neoplasia among a cohort of average-risk, screen-eligible adults aged ≥ 50 years undergoing a first screening colonoscopy. We found that a combination of socio-demographic, physical, and lifestyle features provided good discrimination for

risk of advanced neoplasia; model calibration and discrimination were preserved in an independent spilt-sample validation set. This tool may be used to more precisely personalize an individual's risk for advanced neoplasia and to help decide how to be screened for colorectal cancer.

In the past ten years, risk prediction models have become more prevalent in medicine. They are used for diagnosis, prognosis, treatment, and for screening and prevention. Examples include the Wells score for diagnosis of pulmonary thromboembolism,(18, 19) the Rockall and Blatchford scores for prognosis of acute upper gastrointestinal hemorrhage,(20-22) the $CHA_2DS_2$-VASc score for deciding on anticoagulation therapy for atrial fibrillation,(23) and the Gail model for estimating breast cancer risk.(24) These tools are used by clinicians in deciding how to manage patients, and differ from decision aids, which are tools to assist patients decide whether to receive (most often) a preventive intervention.(5) While the current model was designed primarily to assist providers, it may be used by patients, but would require a careful explanation of the risk and rationale for linking a risk group to a preferred screening test.

Several predictive models are available for estimating both future risk for CRC and current risk for advanced neoplasia.(5, 7, 8, 25-30) Among the models that predict future CRC risk is the National Cancer Institute's Colorectal Cancer Risk Assessment Tool, which is based on rigorous model development and validation (https://ccrisktool.cancer.gov/).(5, 7) The tool measures demographic, familial, medical, lifestyle, and previous endoscopic variables, provides 5-year, 10-year, and lifetime-year risk for CRC, and compares these risks with persons of comparable age and sex. It is the first broadly generalizable, absolute risk model for future risk of CRC. However, it calculates a future risk (of CRC), which some may consider less actionable in the short term than an estimate of current risk of advanced neoplasia. We and others have found that the NCI's tool also accurately estimates current risk for advanced neoplasia (31, 32), although with less precision than the current model because of the greater number of risk categories (i.e., quintiles), which decreases the precision of the risk estimates.

Several models estimate current risk of advanced neoplasia in the screening setting (8, 26-30, 33). Advanced neoplasia is a desirable outcome both clinically (nearly all of advanced neoplasia is advanced adenomas, which are the precursor lesions to most CRC) and methodologically (as it is easier to study due to its higher prevalence than CRC). These models vary in several features, including country of origin and study population, risk factors, model discrimination, ability to achieve clinically-meaningful separation of risk among categories, and extent of model validation. All have limitations and tradeoffs that must be considered, including accurate and reliable measure of the risk factors, clinical importance of the risk gradient, proportions of the cohort at low and high risk, and generalizability of the findings beyond the population from which the model was derived. All of these factors may contribute to underuse of these and most predictive models.(34, 35)

Among risk prediction models developed for CRC screening is a simple, 5-variable model that includes age, sex, a first-degree relative with CRC, cigarette smoking, and waist circumference.(8) The model identifies four risk groups in which the prevalence of advanced neoplasia ranges from just under 2% for the very low-risk group (which comprised 8·2-8·7% of the cohort) to 22-25% for the high-risk group (15% of the cohort). Table 6 contrasts this previous simple model with the current, more complex, one. Model metrics and risk gradient across risk groups are comparable, although the current model's gradient across 3 risk groups is slightly greater and its lowest-risk group larger than that of the previous model. The current model's low-risk group includes no one with colorectal cancer, while the previous model's combined very-low and low-risk groups included 5 persons (0·3%) with cancer, all of which was in the distal colon or rectum. Both models improve on the degree of discrimination provided by age and sex alone, which is represented by a c-statistic of 0·66 (data not published).

Either the previous model or the current one may be used to decide how to be initially screened; that is, is invasive screening necessary, preferred, or favored because of relatively high-risk, or is non-invasive screening a good option? Because each model selects among screening strategies currently considered acceptable by the U.S. Preventive Services Task Force (36, 37) these

models provide additional information for decisions among otherwise acceptable choices. Some may also consider using the model to determine whether – in the case of low-risk – any screening is required. Given that CRC screening is recommended for all 50-75 year old persons who are "average-risk", we suggest that this and other such risk prediction models are best used to determine how, not whether, to be screened.

The main tradeoff to consider between the previous and current models is whether the complexity of the current model is worth its greater discrimination in the risk gradient and its identification of a larger low-risk group. Subsequent investigation involving both providers and patients is required to understand the value both parties place on the differences between the two models.

The model described here has strengths and limitations worth noting. Its strengths include its origin from an average-risk population; inclusion of factors that have previously been shown to associate with either or both colorectal cancer and advanced precancerous polyps; a rigorous process for model development; preliminary validation using a split sample; good model metrics of calibration and discrimination; and a clinically-important risk gradient from low-risk to high-risk group with 41% of the cohort at low- or high-risk. One limitation is the model's derivation on a predominantly white cohort undergoing a first screening colonoscopy, a factor that may limit generalizability to more racially diverse populations and to those who have been previously screened. A second limitation is inclusion of 12 variables, some of which may be difficult for users to both understand and respond to accurately. A third limitation is the model's computational complexity for which some kind of automation may be useful. A fourth limitation is the split-sample validation; subsequent completely independent validation is required to determine the robustness and generalizability of the model. Finally, the model's discrimination is imperfect; some patients with advanced neoplasia were categorized as low-risk; however, none of the low-risk persons with advanced neoplasia had CRC and most had distal lesions, which are more detectable with non-invasive testing (38) and sigmoidoscopy.

Colorectal cancer screening is effective and cost-effective, but it is also costly and underutilized. While colonoscopy is the most commonly used screening test in the U.S., there are no comparative studies that demonstrate its superiority as a screening strategy over other tests and strategies. Further, modeling studies show that alternative strategies may be as effective as colonoscopy (36). The current model may be helpful for engaging previously unscreened persons by personalizing their risk for advanced neoplasia. Further, it may help both patients and their providers in decision-making about which screening test may be most appropriate for them. Low-risk persons may be screened effectively and efficiently with tests other than colonoscopy, while colonoscopy may be preferable for those with high-risk for advanced neoplasia. Persons with intermediate risk, which is comparable to prevalence of advanced neoplasia found in population-based studies of screening colonoscopy (39-42), could choose any of the recommended options without a risk-based preference. Using technology offered by electronic medical records along with integrated risk calculators, the current model could be used in real time to facilitate shared decision making for CRC screening. If validated in other settings, this model has potential to advance the uptake and efficiency of CRC screening in the U.S.

**Declaration of interests:**

None.

## REFERENCES


1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin. 2016;66(1):7-30.
2. Joseph DA, Meester RG, Zauber AG, Manninen DL, Winges L, Dong FB, et al. Colorectal cancer screening: Estimated future colonoscopy need and current volume and capacity. Cancer. 2016;122(16):2479-86.
3. Seeff LC, Richards TB, Shapiro JA, Nadel MR, Manninen DL, Given LS, et al. How many endoscopies are performed for colorectal cancer screening? Results from CDC's survey of endoscopic capacity. Gastroenterol. 2004;127(6):1670-7.
4. Wolf AMD, Fontham ETH, Church TR, Flowers CR, Guerra CE, LaMonte SJ, et al. Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. CA Cancer J Clin. 2018;68(4):250-81.
5. Freedman AN, Slattery ML, Ballard-Barbash R, Willis G, Cann BJ, Pee D, et al. Colorectal cancer risk prediction tool for white men and women without known susceptibility. J Clin Oncol. 2009;27(5):686-93.
6. Peng L, Weigl K, Boakye D, Brenner H. Risk Scores for Predicting Advanced Colorectal Neoplasia in the Average-risk Population: A Systematic Review and Meta-analysis. Am J Gastroenterol. 2018;113(12):1788-800.
7. Park Y, Freedman AN, Gail MH, Pee D, Hollenbeck A, Schatzkin A, et al. Validation of a colorectal cancer risk prediction model among white patients age 50 years and older. J Clin Oncol. 2009;27(5):694-8.
8. Imperiale TF, Monahan PO, Stump TE, Glowinski EA, Ransohoff DF. Derivation and Validation of a Scoring System to Stratify Risk for Advanced Colorectal Neoplasia in Asymptomatic Adults: A Cross-sectional Study. Ann Intern Med. 2015;163(5):339-46.
9. Brenner H, Hoffmeister M, Stegmaier C, Brenner G, Altenhofen L, Haug U. Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840,149 screening colonoscopies. Gut. 2007;56(11):1585-9.
10. Stryker SJ, Wolff BG, Culp CE, Libbe SD, Ilstrup DM, MacCarty RL. Natural history of untreated colonic polyps. Gastroenterol. 1987;93(5):1009-13.
11. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ (Clinical research ed). 2007;335(7624):806-8.
12. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162(1):55-63.
13. Morelli MS, Miller JS, Imperiale TF. Colonoscopy performance in a large private practice: a comparison to quality benchmarks. J Clin Gastroenterol. 2010;44(2):152-3.
14. SAS Institute I. SAS®: Version 9.3 for Windows. SAS Institute, Inc. Cary (NC); 2008.
15. Hosmer DW, Jovanovic B, Lemeshow S. Best subsets logistic regression. Biometrics. 1989:1265-70.
16. Mallows CL. Some comments on C p. Technometrics. 1973;15(4):661-75.
17. Sullivan LM, Massaro JM, D'Agostino Sr RB. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. Stat Med. 2004;23(10):1631-60.
18. Wells PS, Hirsh J, Anderson DR, Lensing AW, Foster G, Kearon C, et al. Accuracy of clinical assessment of deep-vein thrombosis. Lancet. 1995;345(8961):1326-30.
19. Wells PS, Anderson DR, Bormanis J, Guy F, Mitchell M, Gray L, et al. Value of assessment of pretest probability of deep-vein thrombosis in clinical management. Lancet. 1997;350(9094):1795-8.
20. Rockall TA, Logan RF, Devlin HB, Northfield TC. Risk assessment after acute upper gastrointestinal haemorrhage. Gut. 1996;38(3):316-21.

21. Rockall TA, Logan RF, Devlin HB, Northfield TC. Selection of patients for early discharge or outpatient care after acute upper gastrointestinal haemorrhage. National Audit of Acute Upper Gastrointestinal Haemorrhage. Lancet. 1996;347(9009):1138-40.
22. Blatchford O, Murray WR, Blatchford M. A risk score to predict need for treatment for upper-gastrointestinal haemorrhage. Lancet. 2000;356(9238):1318-21.
23. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest. 2010;137(2):263-72.
24. Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst. 1999;91(18):1541-8.
25. 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.
26. Cai QC, Yu ED, Xiao Y, Bai WY, Chen X, He LP, et al. Derivation and validation of a prediction rule for estimating advanced colorectal neoplasm risk in average-risk Chinese. Am J Epidemiol. 2012;175(6):584-93.
27. 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.
28. Lin OS, Kozarek RA, Schembre DB, Ayub K, Gluck M, Cantone N, et al. Risk stratification for colon neoplasia: screening strategies using colonoscopy and computerized tomographic colonography. Gastroenterol. 2006;131(4):1011-9.
29. Yeoh KG, Ho KY, Chiu HM, Zhu F, Ching JY, 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.
30. Park HW, Han S, Lee JS, Chang HS, Lee D, Choe JW, et al. Risk stratification for advanced proximal colon neoplasm and individualized endoscopic screening for colorectal cancer by a risk-scoring model. Gastrointest Endosc. 2012;76(4):818-28.
31. Ladabaum U, Patel A, Mannalithara A, Sundaram V, Mitani A, Desai M. Predicting advanced neoplasia at colonoscopy in a diverse population with the National Cancer Institute colorectal cancer risk-assessment tool. Cancer. 2016;122(17):2663-70.
32. Imperiale TF, Yu M, Monahan PO, Stump TE, Tabbey R, Glowinski E, et al. Risk of Advanced Neoplasia Using the National Cancer Institute's Colorectal Cancer Risk Assessment Tool. J Natl Cancer Inst. 2017;109(1).
33. Tao S, Hoffmeister M, Brenner H. Development and validation of a scoring system to identify individuals at high risk for advanced colorectal neoplasms who should undergo colonoscopy screening. Clin Gastroenterol Hepatol. 2014;12(3):478-85.
34. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999;130(6):515-24.
35. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000;19(4):453-73.
36. Knudsen AB, Zauber AG, Rutter CM, Naber SK, Doria-Rose VP, Pabiniak C, et al. Estimation of Benefits, Burden, and Harms of Colorectal Cancer Screening Strategies: Modeling Study for the US Preventive Services Task Force. JAMA. 2016;315(23):2595-609.
37. Lin JS, Piper MA, Perdue LA, Rutter CM, Webber EM, O'Connor E, et al. Screening for Colorectal Cancer: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. JAMA. 2016;315(23):2576-94.
38. Zorzi M, Hassan C, Capodaglio G, Narne E, Turrin A, Baracco M, et al. Divergent Long-Term Detection Rates of Proximal and Distal Advanced Neoplasia in Fecal Immunochemical Test Screening Programs: A Retrospective Cohort Study. Ann Intern Med. 2018;169(9):602-9.
39. Regula J, Rupinski M, Kraszewska E, Polkowski M, Pachlewski J, Orlowska J, et al. Colonoscopy in colorectal-cancer screening for detection of advanced neoplasia. N Engl J Med. 2006;355(18):1863-72.

40. Ferlitsch M, Reinhart K, Pramhas S, Wiener C, Gal O, Bannert C, et al. Sex-specific prevalence of adenomas, advanced adenomas, and colorectal cancer in individuals undergoing screening colonoscopy. JAMA. 2011;306(12):1352-8.
41. Morikawa T, Kato J, Yamaji Y, Wada R, Mitsushima T, Shiratori Y. A comparison of the immunochemical fecal occult blood test and total colonoscopy in the asymptomatic population. Gastroenterol. 2005;129(2):422-8.
42. Pox CP, Altenhofen L, Brenner H, Theilmeier A, Von Stillfried D, Schmiegel W. Efficacy of a nationwide screening colonoscopy program for colorectal cancer. Gastroenterol. 2012;142(7):1460-7.e2.

**Table 1. Descriptive Data**

| Variable | Derivation set | Validation set |
|---|---|---|
| N | 3025 | 1475 |
| Age (mean ± sd) years | 57·3 ± 6·5 | 57·2 ± 7·0 |
| Women (%) | 51·6 | 51·5 |
| Caucasian (%) | 94% | 94% |
| FDR with CRC (%) | 9·6 | 9·2 |
| Advanced neoplasia (%) | 9·4 | 8·5 |
| N (%) with complete data | 2859 (94·5) | 1426 (96·7) |

CRC – colorectal cancer; FDR – first-degree relative

**Table 2. Variables in the Final Derivation Multivariable Model (N=2,859)**

| Variable | Description | Prevalence | P value | Odds Ratio (95% CI) |
|---|---|---|---|---|
| Age | Years | 50-54 yrs – 46%,<br>55-59 yrs – 25%,<br>60-64 yrs – 17%,<br>65-69 yrs – 7%,<br>≥70 yrs – 5% | <0·0001 | 1·054 (1·034-1·074) |
| Gender | Male, female | 49% (male) | <0·0001 | 1·947 (1·437-2·639) |
| Marital status | Married or living with significant other | 84% | <0·0001 | 0·460 (0·334-0·634) |
| Education | College graduate | 46% | 0·0411 | 0·733 (0·545-0·988) |
| Smoking | > 0 to < 30 pk/yr<br>> 30 pk/yr | 25%<br>15% | <0·0001<br><0·0001 | 1·995 (1·437-2·768)<br>2·364 (1·657-3·374) |
| Significant ethanol use | Men ≥ 11/wk<br>Women ≥ 6/wk | 16% | 0·0508 | 1·398 (0·999-1·957) |
| NSAID use | Daily and ≥ 1 yr | 10% | 0·0011 | 0·374 (0·207-0·675) |
| Aspirin use | Daily and ≥ 1 yr | 27% | 0·0119 | 0·666 (0·486-0·914) |
| Metabolic syndrome | ≥ 3 AJCC criteria | 17% | 0·0749 | 1·344 (0·971-1·860) |
| Red meat consumption | Servings per week | 35% - ≤ 3<br>45% - 4-5<br>20% - ≥ 6 | <0·0001 | 2·212 (1·576-3·105) |
| Regular activity (10 years) | More than once weekly for ≥2 years | 55% | 0·0169 | 0·698 (0·520-0·937) |
| Moderate activity (over last year) | hours / week (x # months) | 44% - 0-1·6<br>42% - 1·7-4·9<br>13% - ≥ 5 | 0·0539 | 0·993 (0·986-1·000) |
| Vigorous activity (over last year) | At least 1 hour per week | 71% | 0·0035 | 0·627 (0·458-0·858) |
| *Goodness-of-Fit P-value = 0·37  C-statistic = 0·77* | | | | |

Prevalence – prevalence of the risk variable categories, including prevalence by meaningful risk categories for the continuous variables (age, red meat consumption and moderate activity)

**Table 3. Points for Each Variable in the Final Derivation Model**

| Variable | Points |
|---|---|
| Age (5-year categories) | 0 to 4 |
| Gender | Women 0;<br>Men 3 |
| Married | 0, -3 (married / living with significant other) |
| Education | 0, -1 (college graduate) |
| Smoking | 0, 3 (≥30 pack-years) |
| Significant ethanol | 0, 1 (≥11 per week for men;<br>≥6 per week for women |
| NSAID use | 0, -4 for daily use and ≥1 year |
| Metabolic syndrome | 0, 1 |
| Red meat servings | 0 to 9 based on number of servings per week |
| Regular activity (10 years) | 0, -1 |
| Moderate activity (last 12 months) | 0 to -2 |
| Vigorous activity (≥1 hr/wk last 12 months) | 0, -2 |
| Aspirin use | 0, -2 for daily use and ≥1 year |

Score could range from -14 to 22 (actual range: -13 to 13)

**Table 4. Risk of Advanced Neoplasia by Risk Group**

| | | Derivation set | | Validation set | |
|---|---|---|---|---|---|
| **Risk Group** | **Score*** | **% Cohort** | **Risk AN (%) (CI)** | **% Cohort** | **Risk AN (%) (CI)** |
| Low | -13 to -5 | 23 | 1·50 (0·72-2·74) | 23 | 2·73 (1·25-5·11) |
| Intermediate | -4 to 2 | 59 | 7·06 (5·89-8·38) | 59 | 5·57 (4·12-7·34) |
| High | ≥ 3 | 18 | 27·3 (23·4–31·3) | 18 | 25·8 (20·5-31·7) |

*Scores of -13 to -5 represent risks of 0-3%; scores of -4 to 2 represents risks of 4-12%, and; scores of 3 or greater represent risks of 18-100%.

**Table 5. Advanced Neoplasia in the Low-risk Subgroup of the Derivation and Validations Sets**

| Variable | Derivation Set | Validation Set |
|---|---|---|
| N | 667 | 330 |
| % cohort | 23 | 23 |
| # (location) of CRCs | 0 | 0 |
| # Advanced adenomas | 10 | 9 |
| % AN detectable with sigmoidoscopy (colonoscopy for any distal polyp) | 70% (7 of 10) | 67% (6 of 9) |

**Table 6. Characteristics and comparison of a previous model and the current model**

| **Measure** | **Previous Model** | **Current Model** |
|---|---|---|
| Calibration | P=0·40 | P=0·37 |
| Discrimination (c-statistic) | 0·72 | 0·77 |
| Number of risk groups | four | three |
| Risk gradient[*] | 1·9% to 25% | 1·5% to 27% |
| Percent of cohort at low-risk | 8·2-8·7% | 23% |
| Percent of cohort at high-risk | 14·7-15·5% | 18% |
| Number, types of variables | 5 simple, reproducible variables | 12 variables, some complex |
| Computation | easy | complex |

[a-*]Based on derivation groups