Preventive Medicine Reports 25 (2022) 101678

Contents lists available at ScienceDirect

Preventive Medicine Reports

journal homepage: www.elsevier.com/locate/pmedr

Five-year absolute risk estimates of colorectal cancer based on CCRAT model and polygenic risk scores: A validation study using the Quebec population-based cohort CARTaGENE

Rodolphe Jantzen [a,b,*], Yves Payette [a], Thibault de Malliard [a], Catherine Labbé [a], Nolwenn Noisel [a,b], Philippe Broët [a,b,c,d]

[a] CARTaGENE, Research Center, CHU Sainte-Justine, Montreal, Quebec, Canada
[b] Université de Montréal, Montréal, Québec, Canada
[c] University Paris-Saclay, CESP, INSERM, Villejuif, France
[d] Assistance Publique-Hôpitaux de Paris (AP-HP), Hôpitaux Universitaires Paris-Sud, Hôpital Paul Brousse, 12 Avenue Paul Vaillant Couturier, 94807 Villejuif, France

ARTICLE INFO

Keywords:
Polygenic risk score (PRS)
CCRAT
Model calibration
Model discrimination
Accuracy
Colorectal cancer occurrence

ABSTRACT

The objective was to evaluate the predictive performance of the Colorectal Cancer Risk Assessment Tool (CCRAT) and three polygenic risk scores (Hsu et al., 2015; Law et al., 2019, Archambault et al., 2020) to predict the occurrence of colorectal cancer at five years in a Quebec population-based cohort. By using the CARTaGENE cohort, we computed the absolute risk of colorectal cancer with the CCRAT model, the polygenic risk scores (PRS) and combined clinico-genetic models (CCRAT + PRS). We also tailored the CCRAT model by using the marginal age-specific colorectal incidence rates in Canada and the risk score distribution. We reported the calibration and the discrimination. Performances of the PRSs, combined and tailored CCRAT models were compared to the original CCRAT model. The expected-to-observed ratio of the original CCRAT model was 0.54 [0.43–0.68]. The c-index was 74.79 [68.3–80.5]. The tailored CCRAT model improved the expected-to-observed ratio (0.74 [0.59–0.94]) and c-index (76.39 [69.7–82.1]). All PRS improved the expected-to-observed ratios (around 0.83, confidence intervals including one). PRSs' c-indexes were not significantly different from CCRAT models. Results from the combined models were close to those from the PRS models, Archambault combined model's c-index being significantly higher than the original and tailored CCRAT models (78.67 [70.8–86.5]; $p < 0.001$ and $p = 0.028$, respectively). In this Quebec cohort, CCRAT model has a good discrimination with a poor calibration. While the tailored CCRAT provides some gain in calibration, clinico-genetic models improved both calibration and discrimination. However, better calibrations must be obtained before a practical use among the inhabitants of Quebec province.

1 Introduction

Colorectal cancer is the third diagnosed cancer in Canada (Canadian Cancer Statistics Advisory Committee, 2019). Despite decreasing incidence and death rates — partly due to screening, resection of pre-malignant lesions — about 50% of colorectal cancers are still detected at a late stage (Canadian Cancer Statistics Advisory Committee, 2019; Edwards et al., 2010). The latest Canadian Task Force guidelines (2016) (Canadian Task Force on Preventive Health Care, 2016) recommend screening adults aged 50 to 74 years for colorectal cancer with fecal occult blood testing every two years. These guidelines do not apply to those at high risk for colorectal cancer (i.e., previous colorectal cancer or polyps, inflammatory bowel disease, signs or symptoms of colorectal cancer, family history of colorectal cancer, or hereditary syndromes predisposing to colorectal cancer). However, due to the increasing incidence of colorectal cancer with age, the biennial fecal occult blood screening among people aged 50–59 years leads to a lower absolute risk reduction compared to the 60–74 years group (Canadian Task Force on Preventive Health Care, 2016).

In this context, cancer risk assessment tools might be used to identify

* Corresponding author at: 3175 Chemin de la Côte-Sainte-Catherine, Montréal, QC H3T 1C5, Canada.
E-mail address: rodolphe.jantzen@gmail.com (R. Jantzen).

https://doi.org/10.1016/j.pmedr.2021.101678
Received 22 August 2021; Received in revised form 21 October 2021; Accepted 24 December 2021
Available online 27 December 2021

the at-risk population, to refine the screening regimens or to select individuals for a preventive therapy. As an example, the Breast Cancer Risk Assessment Tool is used in the United-States to identify women who may be candidates for breast cancer chemo-prevention (Moyer, 2013). For colorectal cancer, various predictive models have been developed, but they are less used in practice (Smith et al., 2019). Among these predictive models, the Colorectal Cancer Risk Assessment Tool (CCRAT) (Freedman et al., 2009) is an interactive tool proposed by the National Cancer Institute to estimate an individual's risk of developing colorectal cancer over a specific period (National Institutes of Health (NIH) — ccrisktool.cancer.gov). CCRAT includes the clinical variables age, race, sigmoidoscopy/colonoscopy and polyps history, vigorous activity, servings of vegetables, relatives with a colorectal cancer, non-steroidal anti-inflammatory drugs and estrogen use, cigarettes smoked, and body mass index. The relative risks and attributable risks used to compute the absolute risk were derived from an American population-based case-control data (Utah, Minnesota and Northern California), while the estimation of the marginal age-specific cancer hazard rates were obtained from the National Cancer Institute's Surveillance Epidemiology and End Results (SEER) Program. This predictive model was validated in a large prospective cohort study from the NIH (Park et al., 2009) and was recently revalidated (Smith et al., 2019), with an Area Under the receiver operating characteristic (ROC) Curve (AUC) of 0.61 for both studies. According to a systematic review of risk prediction models for colorectal cancer, the CCRAT model did not provide the best AUC among the models also based on self-completed questionnaire (Smith et al., 2019; Usher-Smith et al., 2016). However, the CCRAT model was updated in March 2019 with the incidence rates of the SEER18 (2000–2015) and the mortality rates of the 2010–2015 data. This updated version makes it possible to compute the absolute risk for people between 40 and 85 years old, whereas it was previously limited to 50–85 years old. Moreover, as the CCRAT model is based on self-completed questionnaires, it can be easily implemented into clinical practice.

Besides predictive models that rely on routine data or self-completed questionnaires, other models were developed and provide good discriminatory power, with AUCs higher than 0.80 (Han et al., 2008; Marshall et al., 2009). Some of these predictive models are based on polygenic risk scores (PRS), derived from published genome-wide association studies. The Hsu et al. (Hsu et al., 2015), Law et al. (Law et al., 2019) and Archambault et al. (2020) studies (Hsu et al., 2015; Law et al., 2019; Archambault et al., 2020) are among the few to provide enough publicly available information to compute the PRS (e.g., alleles' risk), with single-nucleotide polymorphism's (SNP) odds ratio specific to the European population. These three PRSs include 27 (Hsu et al., 2015); 40 (Law et al., 2019) and 95 SNPs (Archambault et al., 2020), respectively, with few overlaps: four SNPs are in common between Archambault and Hsu PRSs, three between Archambault and Law PRSs and none between Law and Hsu PRSs. The Archambault PRS contained the 40 SNPs reported in the Huyghe *et al.* study (Huyghe et al., 2019). The PRS proposed by Hsu *et al.* was trained on a large sample with more than 12,000 participants, but provided moderate discriminatory power, the best AUC being of 0.60. Moreover, these genetic-based predictive models were not combined with clinical-based models such as the CCRAT model.

These clinical and genetic-based colorectal cancer risk assessment tools could be helpful for primary care physicians to enhance screening uptake among those less likely to participate in organized screening. Moreover, it would be of great interest for physicians to know the performance of predictive models relying on lifestyle risk factors, as it may represent a way to promote behavioral intentions (e.g., diet, physical activity, screening). The performance of these predictive models usually trained on US populations may vary across populations. Thus, it is useful to evaluate CCRAT model in Quebec since the French-Canadians constitute the majority of the Quebec's population that has specific genetic patterns, as compared to the general European population, together with lifestyle/exposure risk factors that are at the intersection between those from Europe and North America.

In this study, we report the predictive abilities of the CCRAT model and the PRS proposed by Hsu *et al.*, Law *et al.* and Archambault *et al.* for the occurrence of colorectal cancer at five years in the population-based cohort CARTaGENE from Quebec.

2 Materials & methods

2.1. Participant selection and outcome

This study used participants (men and women) from the CARTaGENE cohort, which is composed of 43,037 Quebec residents aged between 40 and 69 years (Awadalla et al., 2013). Briefly, participants were randomly selected to be broadly representative of the population recorded on the Quebec administrative health insurance (RAMQ) registries (about 98% of Quebec residents (RAMQ, 2017). Participants have been recruited during two phases (Phase A: 2009–2010; Phase B: 2013–2014) in metropolitan areas, where nearly 70% of Quebecers live. At the recruitment date, each participant filled a health questionnaire.

As the CCRAT model cannot accurately estimate the risk of colorectal cancer for people with certain health conditions, the exclusion criteria were an age under 40 years, a colorectal cancer before the inclusion date and a history of ulcerative colitis or Crohn disease. Some variables' missing values were not supported by the CCRAT model (vigorous activity, vegetables servings and body mass index). Therefore, individuals with missing values for any of these variables were excluded. Other missing values were coded according to the CCRAT model (i.e., coding missing family history as zero relative and unknown colonoscopy/sigmoidoscopy history as a separate category, see Supplementary File 1). Based on genetic data, we did not include participants with non-European ancestries (Supplementary File 1). Information about familial adenomatous polyposis and Lynch syndrome were not available in our cohort.

A fraction of the CARTaGENE cohort (n = 12,062) was genotyped. These participants were selected to be genotyped through various scientific subprojects unrelated to colorectal cancer (Akçimen et al., 2019; Hodgkinson et al., 2014; Hussin et al., 2015). Imputation and quality control are described in the Supplementary File 1. The availability of genotyping information and clinical variables led us to consider two subcohorts for evaluating the predictive models (cf. "Predictive scores" section). The validation of the CCRAT model was done using individuals with a computable CCRAT absolute risk (hereinafter referred as "CCRAT cohort"). The validation of the PRS models was done using individuals from the CCRAT cohort with genotyping information (hereinafter referred as "PRS cohort") (Fig. 1.

For identifying cases with a colorectal cancer (invasive or *in situ*), we used the MED-ECHO administrative health database with the Tonelli *et al.* algorithm (Tonelli et al., 2016): individuals with at least two claims in two years or one hospitalization related to a colorectal cancer. The incidence date was the date of first hospital discharge or first claim with the appropriate International Classification of Diseases (ICD) (ICD-9: 153, 154, 2303, 2304; ICD-10: C18, C19, C20, C21, D010, D011). Data were available from January 1st, 1998 to March 31st, 2016. Dates of death were also retrieved from the RAMQ.

The outcome was the time before occurrence of a colorectal cancer from the enrollment in the cohort. Patients without colorectal cancer were censored at five years, at death or on March 31st, 2016 (administrative censoring).

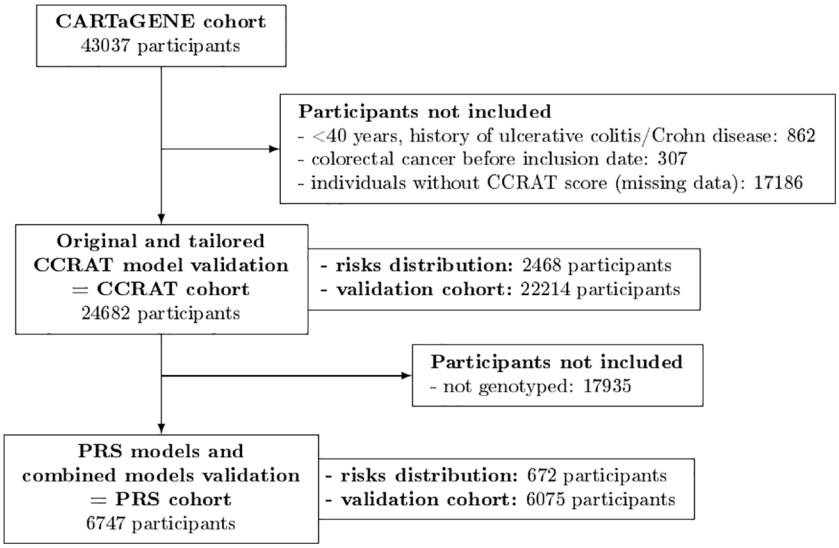

Fig. 1. Flow-chart. PRS: polygenic risk score. Tailored CCRAT model: original CCRAT model updated considering the risk factors distribution and the Canadian baseline age-specific hazard rate and age-specific mortality hazard rate.

2.2. Predictive scores

2.2.1 Absolute risk of colorectal cancer at five years

The absolute risk of developing colorectal cancer over the next 5 years, for a participant free of colorectal cancer at age t_0 (date of enrollment) and with a risk score S, is the probability that he/she will be diagnosed over the period $[t_0,t_1]$, where $t_1 = t_0 + 5$ years. Under the assumption of a multiplicative proportional hazard model (or Cox model), the absolute risk (denoted $AR(t_0,t_1;S)$) can be written such as:

$$AR(t_0,t_1;S) = \int_{t_0}^{t_1} \lambda_0(t)e^S exp\left[-\int_t^{t_0} \lambda_0(u)e^S + \gamma(u)du\right]dt$$

where $\lambda_0(t)$ and $\gamma(t)$ are the baseline age-specific hazard rate for the colorectal cancer and the age-specific mortality hazard rate from other causes (competing risks), respectively. In practice, the absolute risk is computed using piece-wise constant hazard rates.

These baseline age-specific hazard rates are calculated using marginal (or composite) hazard rates obtained from registries, together with either the attributable hazard function or the risk factor distribution. In this study, we used Statistics Canada (Statistics Canada, 2019; Statistics Canada, 2019) to retrieve the 2017 marginal age-specific hazard rates, i.e., the Canadian colorectal incidence for each five years from 40 to 89 years (see Supplementary File 2). The colorectal incidence rates was not available for the Quebec province after 2010.

2.2.2 Original and tailored colorectal cancer risk Assessment Tool models

We considered two different approaches to estimate the absolute risk: (i) the original CCRAT model as implemented by the National Cancer Institute and (ii) a tailored CCRAT model that uses the Canadian marginal age-specific hazard rates and the distribution of the risk score.

To compute the original CCRAT model's absolute risks at five years, at the inclusion date, we extracted the CCRAT's variables from the CARTaGENE questionnaire (Table A.1 of the Supplementary File 1) and used the NIH SAS macro, version 3.0 updated in January 30th, 2019 (dceg.cancer.gov/tools/risk-assessment/ccratsasmacro) with the R language.

For the tailored CCRAT model, we used the Individualized Coherent Absolute Risk Estimator (iCARE) package (Choudhury et al., 2020). The distribution of the risk scores was obtained by the sampling at random of 10% of the individuals from the CCRAT cohort, for each gender, with a small probability for the cancer cases to be selected. We reported the results using the 90% remaining individuals (hereinafter referred as "CCRAT validation cohort", Fig. 1.

2.2.3 Genetic models and combined models (clinico-genetic models)

In this study, genotyping information was used for computing the PRSs, a weighted linear combination of the risk-conferring variant alleles. Weights are the log odd ratio of each at-risk allele. We obtained the loci and corresponding SNPs' weight associated with colorectal cancer from the Hsu *et al.* (27 SNPs), Law *et al.* (40 SNPs) and Archambault *et al.* (95 SNPs) studies (Hsu et al., 2015; Law et al., 2019; Archambault et al., 2020). With our data, we retrieved all the SNPs of the Hsu PRS, 39 SNPs of the Law PRS and 93 SNPs of the Archambault PRS. We considered a weight of zero for the three missing SNPs (rs6928864, rs755229494 and rs377429877). In a sensitivity analysis, we replaced the missing SNPs with surrogate SNPs using high-linkage disequilibrium: rs6904092, rs1801155 and rs9537756, respectively. More information about SNPs included can be found in Supplementary File 3.

To compute the absolute risk, we also used the iCARE package. The distribution of the risk scores were obtained by the sampling at random of 10% of the individuals from the PRS cohort, for each gender, with small probability weights for the cancer cases. We reported the results using the 90% remaining individuals (hereinafter referred as "PRS validation cohort", Fig. 1.

For estimating the absolute risk of colorectal cancer with a combination of both clinical (CCRAT) and genetic data (PRS), we used the same procedure as described in the sub-section "PRS models". In practice, the combination was simply the sum of the PRS and the CCRAT relative risks (hereinafter referred as "combined models").

The same methodology was used to compute the tailored CCRAT model in the PRS validation cohort. Then, we compared the original and tailored CCRAT models with the PRS and combined models in the PRS validation cohort.

2.3. Statistical analysis

To show the distribution of the absolute risks, we plotted predictiveness curves and rug plots, with the cumulative concentration of predictions as a function of cumulative percentage of individuals.

2.3.1 Calibration

The calibration was assessed by computing the expected-to-observed ratio (E/O) obtained as the sum of the estimated risk, divided by the number of observed cases. The 95% confidence interval (95%CI) was calculated assuming a Poisson distribution by the formula (Rockhill et al., 2001):

$$\frac{expected}{observed} \times e^{\pm 1.96 \times \sqrt{1/observed}}$$

We also compared graphically the predicted and observed proportion of colorectal cancers in three absolute risk groups, defined as the tertile of each model. The observed proportion at five-year in each risk group was calculated using a Kaplan-Meier estimator. Using these three groups, we computed a Pearson's chi-squared goodness of fit test under the null hypothesis of no difference for assessing the discrepancy between observed and expected proportion.

We also assessed the overall calibration by reporting estimates of the intercept and slope, fitted from a logistic regression model to observed outcomes with the logit of the predicted probabilities as the independent variable. A good calibration should have an intercept and slope close to zero and one, respectively. The calibration slope evaluates the spread of the estimated risks. A slope greater than one indicates that risk estimates are too moderate. A slope less than one indicates the opposite. Negative intercept values indicate overestimation whereas positive values indicate the opposite.

2.3.2 Discrimination

The discrimination was assessed by the c-index (equivalent to the AUC for survival analysis) with an Inverse Probability of Censoring Weighting estimation of cumulative time-dependent ROC curve, with their 95%CI computed by bootstrap (Uno et al., 2007; Blanche et al., 2013; Blanche et al., 2012). ROC curves were generated.

All statistical analyses were performed using R software, version 3.6 (R Core Team, 2021).

2.4. Ethics approval and consent to participate

This project met the institution's guidelines for protection of human subjects concerning their safety and privacy as it has been approved by the Research Ethics Board of the Sainte-Justine University Hospital Center under the reference 2020–2427. In addition, CARTaGENE has obtained ethics approval by the Sainte-Justine University Hospital Center under the reference: MP-21-2011-345, 3297. Written informed consent was obtained from all the participants.

3 Results

Overall, 24,682 individuals were included in the CCRAT cohort for the validation of the original and tailored CCRAT models. Genotyped data was available for 6,747 individuals (Fig. 1. The median age at enrollment was 53.2 years [Q1-Q3 47.3–60.1] for participants of the CCRAT cohort and 53.8 years [48.3–61.0] for the PRS cohort, with 80 (0.32%) and 24 (0.36%) individuals having a colorectal cancer during the five years of follow-up in the CCRAT cohort and the PRS cohort, respectively. The median time of follow-up in each cohort was of 3.1 [2.5–5] and 5 years [3.1–5], respectively. The Table A.2 in Supplementary File 1 compare the baseline characteristics between the CCRAT and PRS cohorts. The most notable differences are reported here. Compared to the CCRAT cohort, the PRS cohort had more individuals from the first enrollment phase (69.4% from phase A versus 41.2%). Therefore, as the variables "history of sigmoidoscopy/colonoscopy" and "family history of colorectal cancer" were not available for individuals included during the phases A and B, respectively, the proportion of missing data differed between the CCRAT and the PRS cohorts for these two variables. Men of the PRS cohort smoked more (35.0% vs 19.8% smoked more than 11 cigarettes per day). The CCRAT absolute risk median did not differ between the cohorts (0.20%).

3.1. Original and tailored colorectal cancer risk Assessment Tool models

In the CCRAT validation cohort, the highest absolute risk computed with the original CCRAT model was 2.4% (Fig. 2A). The overall calibration showed a global underestimation, with an E/O of 0.54 [95%CI 0.43–0.68]. The goodness of fit test was non-significant ($p = 0.17$) (Table 1). The intercept and slope were not significantly different from zero and one, respectively. The calibration curve shows this overall underestimation, with the E/O significantly lower than one for the second and third tertiles (0.49 [0.33–0.75] and 0.54 [0.40–0.72], respectively) (Fig. 2B). The original CCRAT model provided a c-index of 74.79 [68.3–80.5] (Fig. 2C and Table 1).

Compared to the original CCRAT model, results from the tailored CCRAT model showed a significant improvement of the calibration ($p = 0.03$). Nevertheless, the E/O remained significantly lower than one (0.74 [0.59–0.94]). The goodness of fit test was non-significant ($p = 0.63$). The intercept and slope were not significantly different from zero and one, respectively. Only the E/O of the third tertile was significantly lower than one (0.74 [0.56–0.99]) (Fig. 2B). The discriminatory power was slightly improved, but the confidence intervals overlapped (c-index of 76.39 [69.7–82.1]) (Fig. 2C and Table 1).

3.2. Genetic and combined models

In the PRS validation cohort, the three PRS models showed E/O non-significantly different from one (around 0.83), with non-significant goodness of fit tests (Fig. 3 and Table 2). The intercepts and slopes were not significantly different from zero and one, respectively. None of the E/O in each risk group were significantly different from one. None of the PRS models' c-index was significantly different from the original and tailored CCRAT models (Fig. 4 and Table 2).

Combined models had similar E/O than genetic-based models, with non-significant goodness of fit tests. The intercepts were also not significantly different from zero. Slopes slightly decreased but remained not significantly different from one. None of the E/O in each risk group were significantly different from one (Fig. 3 and Table 2). All the combined models produced a higher c-index than the original and tailored CCRAT models, the Archambault combined model being significantly higher (78.67 [70.8–86.5]; compared with the original CCRAT: $p < 0.001$, tailored CCRAT: $p = 0.028$) (Fig. 4 and Table 2).

4 Discussion

In this work, we assessed the predictive performance of the CCRAT model and three PRSs for predicting the occurrence of colorectal cancer at five years in a Quebec population. Our results showed that in our population the original CCRAT model had a good discriminatory but poor calibration with a global underestimation of risks. However, the use of the Canadian marginal age-specific hazard rate and age-specific mortality hazard rate improved these results. While the discriminatory abilities of the PRS models did not significantly differ from those obtained with the original and tailored CCRAT models, the Archambault's combined model significantly improved the discrimination. Calibration was improved for both PRS and combined models compared to the original CCRAT model.

Interestingly, the original CCRAT model had a better c-index than those obtained in the validation study of Smith *et al.* (Smith et al., 2019; Park et al., 2009). However, it is worth noting that we used the latest version of the CCRAT model with updated American colorectal cancer incidences and a five years follow-up, while the predictions in the original study were over ten-years. In our cohort, compared to colorectal

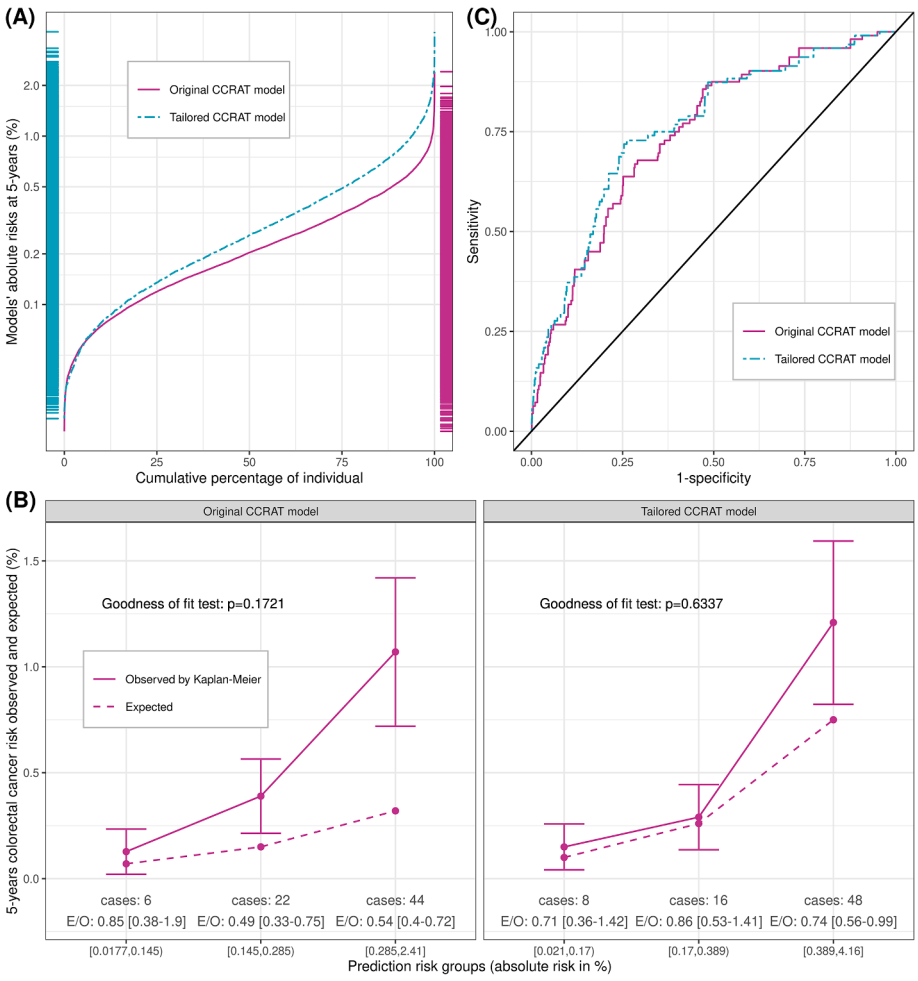

Fig. 2. Risk distribution and performance of the original and tailored CCRAT models (CCRAT validation cohort). Tailored CCRAT model: original CCRAT model updated considering the risk factors distribution and the Canadian baseline age-specific hazard rate and age-specific mortality hazard rate. CCRAT validation cohort: 90% of the CCRAT cohort for validating the models (n = 22,214). (A) Distribution of the original CCRAT and tailored CCRAT models' predictions as a function of cumulative percentage of individuals. Rug plot on the y-axis. (B) Calibration according to the original CCRAT and tailored CCRAT models' predictions groups (tertile). P values were computed using a goodness of fit test statistic compared to the critical value from the chi-squared distribution. E/O: expected-to-observed cases. (C) Discrimination power of the original CCRAT and tailored CCRAT models according to sensitivity and specificity. C-indexes were calculated using the Inverse Probability of Censoring Weighting estimation of cumulative time-dependent ROC curve.

Table 1
Comparison of the original CCRAT model and tailored CCRAT model in the CCRAT validation cohort.

	Original CCRAT model	Tailored CCRAT model
C-index	74.79 [68.3, 80.5]	76.39 [69.7, 82.1]
Global E/O	0.54 [0.43, 0.68]	0.74 [0.59, 0.94]
Goodness of fit	p = 0.17	p = 0.63
Intercept	0.4 [-1.6, 2.3]	−0.9 [−2.5, 0.6]
Slope	1 [0.7, 1.4]	0.8 [0.6, 1.1]

CCRAT: Colorectal Cancer Risk Assessment Tool; E/O: expected-to-observed ratio. Tailored CCRAT model: CCRAT model updated considering the risk factors distribution in our population and the Canadian baseline age-specific hazard rate and age-specific mortality hazard rate. CCRAT validation cohort: 90% of the CCRAT cohort for validating the models. 95% confidence intervals in square brackets.

cancer risk prediction models using "self-completed questionnaire" variables (Usher-Smith et al., 2016), the original CCRAT model in our cohort had the best c-index (74.8% versus 71%, the highest c-index of the systematic review). However, the original CCRAT model is not well calibrated with a global underestimation of the absolute risk that might be explained by a higher colorectal incidence among Canadian (Statistics Canada, 2019). In contrast, Smith *et al.* used the previous version of the CCRAT model with higher incidence rates and found an overestimation of the CCRAT model for individuals in the highest 20% of predicted risk group (Smith et al., 2019).

The use of the Canadian incidences improved the calibration and slightly improved the discriminative capacity. It should be noted that the improvement in overall calibration was not shown with the intercept and slope analysis. Even though the original CCRAT model seemed to have a better intercept and slope, the large 95%CIs make the comparison difficult. These results underline the interest of adapting a predictive model to a new population using publicly available information. The iCARE R package (Choudhury et al., 2020) facilitates this process.

The calibrations of all the PRS models provided E/O quite similar to those obtained from the tailored CCRAT. Using only clinical (original and tailored CCRAT models) or genetic data (PRS models) produced similar discriminating capacity. In contrast, Hsu *et al.* (Hsu et al., 2015) reported a c-index significantly higher when using only genetic data as compared to the one using only clinical data. However, their clinical model had fewer variables than the CCRAT model (sex, age, family history and history of endoscopic examinations) without taking into account the smoking status, which might explain their lower c-index as compared to the CCRAT model (52 versus 74.8). In addition, Hsu *et al.* found a significantly higher c-index when adding PRS with family history (51% versus 59% for men). In our study, while the Hsu combined model and the Law combined model moderately improved the discriminatory power and remained non-significantly different from the original and tailored CCRAT models, the Archambault combined model significantly improved the c-index up to 78.7%, event compared to the tailored CCRAT model. It is worth noting that the Archambault PRS was the most recent PRS with the highest number of SNPs.

In our study, the better calibration of the PRS and combined models compared to the original CCRAT model might be explained by the use of Canadian cancer incidence rates, since the tailored CCRAT model also

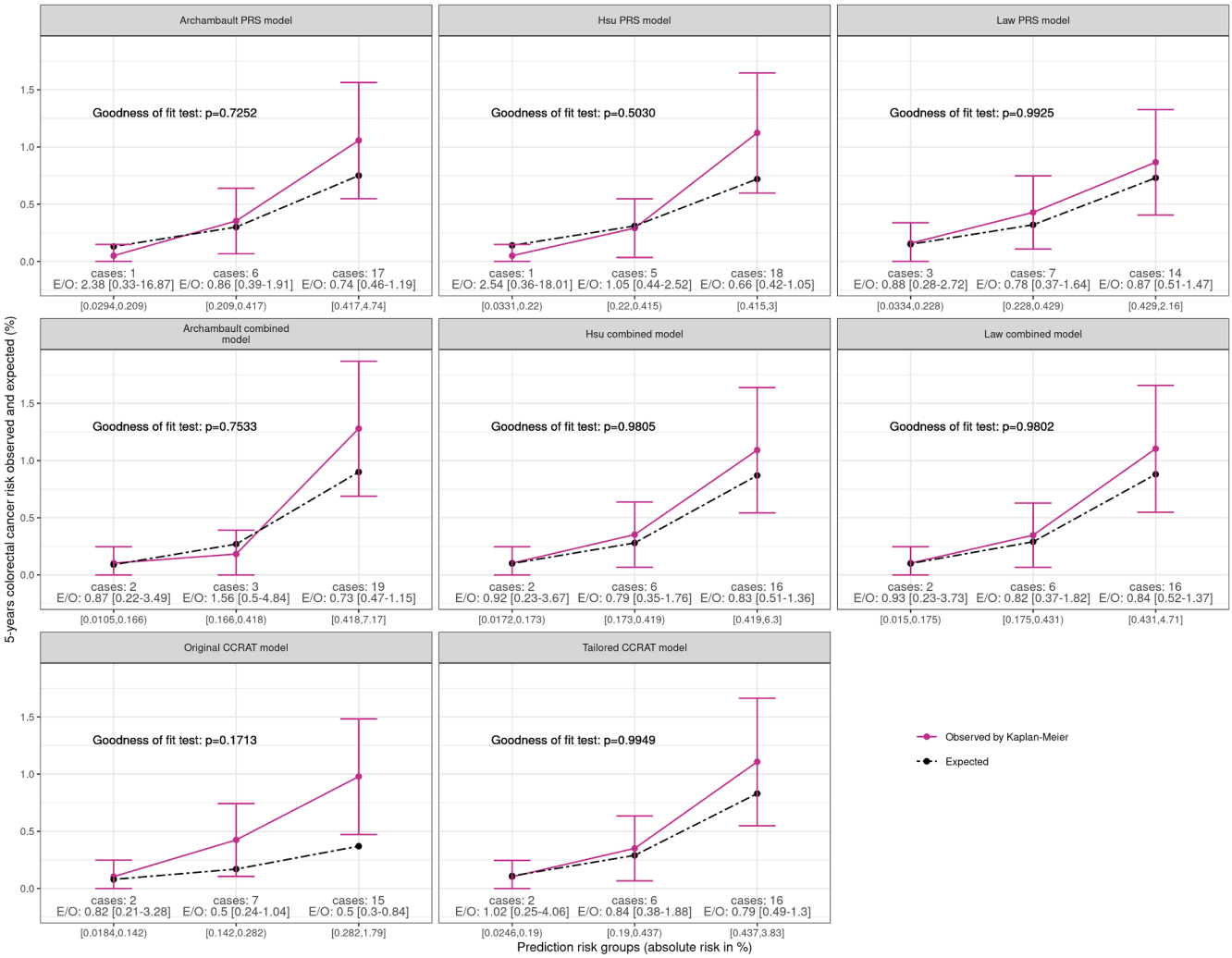

Fig. 3. Calibration of the original CCRAT model, PRS models and combined models (polygenic risk score validation cohort). E/O: expected-to-observed ratio. PRS: polygenic risk score. Combined models: models combining CCRAT and PRS models' absolute risks. PRS validation cohort: 90% of the PRS cohort for validating the models (n = 6,075). P values were computed using a goodness of fit test statistic compared to the critical value from the chi-squared distribution. Expected points are the absolute risks median in each group. Groups are the absolute risk tertiles of each model.

produced a better calibration. It is worth noting that predictive models using biomarkers led to higher c-indexes (Han et al., 2008; Marshall et al., 2009), the highest being of 88%, but none of the SNPs used in the three PRS models is located in these biomarkers' genes.

Strengths and limitations of this work should be discussed. This study represents a contribution to the efforts made to evaluate risk cancer assessment tools across various populations (e.g., (Shin et al., 2014; Ma et al., 2010). A main strength is that it relies on a large population-based cohort representative of the Quebec urban population of middle-aged and older adults. Moreover, the linkage with administrative health databases improves the outcome quality and accuracy. Nevertheless, there are some limitations linked to the data collection. As we only included participants with European ancestries, results may be obviously less accurate for other populations. The information regarding family history of colorectal cancer was only available for participants enrolled during the first phase, while the colonoscopy/sigmoidoscopy history was only available for participants enrolled during the second phase. However, it is worth noting that the CCRAT model can handle missing data for these two variables. The date of the last colonoscopy/sigmoidoscopy was not available. When an individual had no medication, we could not know if he/she had no treatment or if the question was not answered. In this latter case, we considered the variable as missing. For the number of vegetables served per week, if the precise quantity per day was unavailable (exact quantity for each vegetable), we used the number of serving per day (one, two, etc.), which was less precise. For the genetic-based models, three SNPs were not available in our cohort, one of the Law PRS (odds ratio of 1.13) and two of the Archambault PRS (odds ratio of 1.87 and 1.05). However, replacing the missing SNPs with surrogate SNPs using high-linkage disequilibrium did not change the results. Moreover, the predictive accuracy seemed to be unaffected, as the Archambault PRS and combined models were the best predictive models. Finally, the good discrimination but poor calibration of CCRAT is clearly an issue that jeopardizes its practical implementation in Quebec. Even though, the use of Canadian cancer incidence rates improve CCRAT's calibration, more works should be done to better understand the sources of this miscalibration and correct it. In particular, it might be useful to update the parameter of the CCRAT model to the population of Quebec.

5 Conclusions

To the best of our knowledge, this study is the first to evaluate the CCRAT model in a Quebec population for predicting colorectal cancer at five years. We found that CCRAT model has a good discrimination with a poor calibration. While the tailored CCRAT provides some gain in calibration using the Canadian cancer incidences, clinico-genetic models

Table 2
Comparison of the original CCRAT model with the polygenic risk score models and the combined models in the polygenic risk score validation cohort.

	Original CCRAT model/tailored CCRAT model	Hsu PRS model	Law PRS model	Archambault PRS model
C-index	72.89 [64.9, 80.8] 74.48 [66.0, 83.0]	72.53 [64.9, 80.2]	70.52 [62.3, 78.8]	76.24 [69.1, 83.4]
C-index comparison	Original CCRAT model Tailored CCRAT model	p = 0.93 p = 0.63	p = 0.50 p = 0.20	p = 0.35 p = 0.61
Global E/O	0.53 [0.35, 0.79] 0.82 [0.55, 1.23]	0.82 [0.55, 1.23]	0.84 [0.57, 1.26]	0.84 [0.56, 1.25]
Goodness of fit	p = 0.17 p = 0.99	p = 0.50	p = 0.99	p = 0.73
Intercept	−0.5 [-3.9, 2.8] −1.5 [-4.1–1]	−0.7 [-3.8, 2.4]	−0.7 [-4, 2.4]	0.4 [-2.5, 3.3]
Slope	0.8 [0.3, 1.4] 0.7 [0.2–1.2]	0.9 [0.3, 1.5]	0.9 [0.3, 1.5]	1.1 [0.5, 1.6]
	Original CCRAT model/tailoredCCRAT model	Hsu combined model	Law combined model	Archambault combined model
C-index	72.89 [64.9, 80.8] 74.48 [66.0, 83.0]	75.80 [67.7, 83.9]	74.84 [66.4, 83.3]	78.67 [70.8, 86.5]
C-index comparison	Original CCRAT model Tailored CCRAT model	p = 0.15 p = 0.61	p = 0.20 p = 0.78	p < 0.001 p = 0.028
Global E/O	0.53 [0.35, 0.79] 0.82 [0.55, 1.23]	0.83 [0.56, 1.24]	0.84 [0.57, 1.26]	0.85 [0.57, 1.27]
Goodness of fit	p = 0.17 p = 0.99	p = 0.98	p = 0.98	p = 0.75
Intercept	−0.5 [-3.9, 2.8] −1.5 [-4.1–1]	−1.8 [-4.1, 0.5]	−1.8 [-4.3, 0.5]	−1.1 [-3.4, 1]
Slope	0.8 [0.3, 1.4] 0.7 [0.2–1.2]	0.7 [0.2, 1.1]	0.7 [0.2, 1.1]	0.8 [0.4, 1.2]

CCRAT: Colorectal Cancer Risk Assessment Tool; E/O: expected-to-observed ratio; PRS: polygenic risk score. Combined models: models combining CCRAT and PRS models' absolute risks. Tailored CCRAT model: CCRAT model updated considering the risk factors distribution in our population and the Canadian baseline age-specific hazard rate and age-specific mortality hazard rate. PRS validation cohort: 90% of the PRS cohort for validating the models. 95% confidence intervals in square brackets.

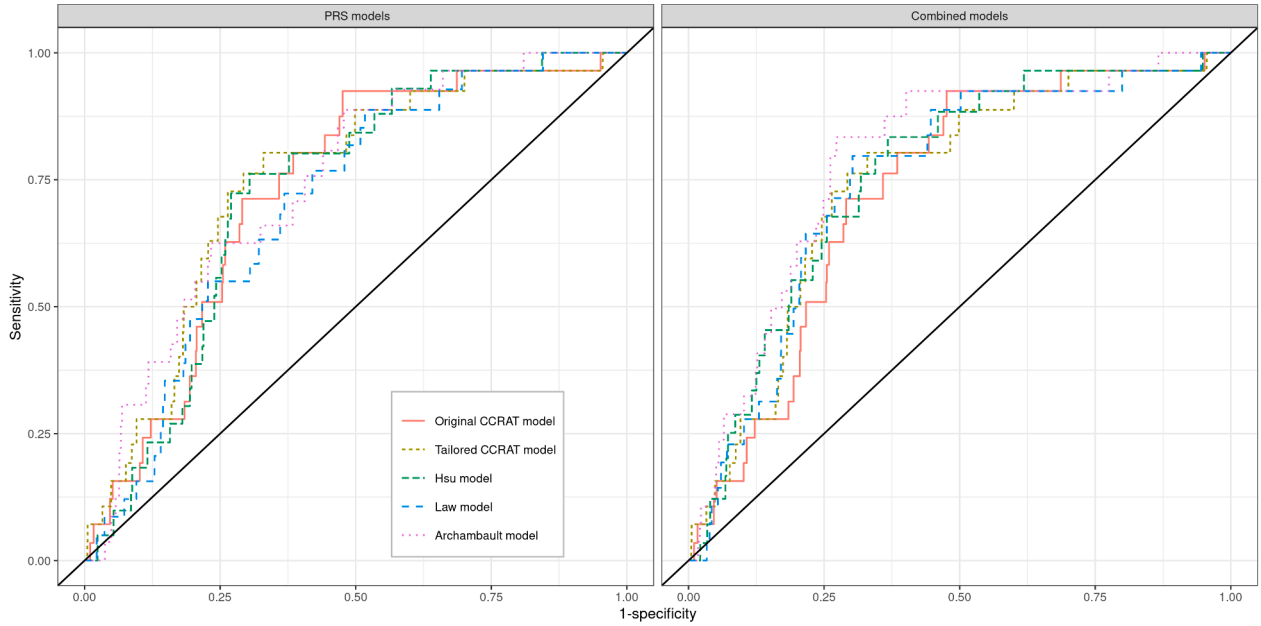

Fig. 4. Discrimination power of the original CCRAT model, PRS models and combined models (polygenic risk score validation cohort). PRS: polygenic risk score. PRS validation cohort: 90% of the PRS cohort for validating the models (n = 6,075). Combined models: models combining CCRAT and PRS models' absolute risks. C-indexes were calculated using the Inverse Probability of Censoring Weighting estimation of cumulative time-dependent ROC curve.

improved both calibration and discrimination. However, better calibrations must be obtained before a practical use of these predictive tools among the inhabitants of Quebec province.

CRediT authorship contribution statement

Rodolphe Jantzen: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Yves Payette:** Data curation,

Software, Writing – review & editing. **Thibault de Malliard:** Data curation, Software. **Catherine Labbé:** Resources. **Nolwenn Noisel:** Conceptualization, Resources, Writing – review & editing. **Philippe Broët:** Conceptualization, Formal analysis, Methodology, Project administration, Supervision, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank all the CARTaGENE participants for their generous investments in health research. We would also like to thank the RAMQ and the Commission d'accès à l'information (CAI) for their support in obtaining the data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pmedr.2021.101678.

References

Canadian Cancer Statistics Advisory Committee. Canadian Cancer Statistics 2019. Toronto, ON: Canadian Cancer Society [Internet]. 2019 Sep [cited 2019 Nov 13]; Available from: cancer.ca/Canadian-Cancer-Statistics-2019-EN.

Edwards, B.K., Ward, E., Kohler, B.A., Eheman, C., Zauber, A.G., Anderson, R.N., et al. 2010. Annual Report to the Nation on the Status of Cancer, 1975–2006, Featuring Colorectal Trends and Impact of Interventions (Risk Factors, Screening, and Treatment) to Reduce Future Rates. Cancer 116(3), 544–573.

Canadian Task Force on Preventive Health Care. 2016. Recommendations on screening for colorectal cancer in primary care. CMAJ 188(5), 340–348.

Moyer, V.A. 2013. Medications for risk reduction of primary breast cancer in Women: U. S. Preventive services task force recommendation statement. Ann. Intern. Med. [Internet]. 2013 Sep 24 [cited 2019 Nov 28]; Available from: http://annals.org/article.aspx?doi=10.7326/0003-4819-159-10-201311190-00718.

Smith, T., Muller, D.C., Moons, K.G.M., Cross, A.J., Johansson, M., Ferrari, P., Fagherazzi, G., Peeters, P.H.M., Severi, G., Hüsing, A., Kaaks, R., Tjonneland, A., Olsen, A., Overvad, K., Bonet, C., Rodriguez-Barranco, M., Huerta, J.M., Barricarte Gurrea, A., Bradbury, K.E., Trichopoulou, A., Bamia, C., Orfanos, P., Palli, D., Pala, V., Vineis, P., Bueno-de-Mesquita, B., Ohlsson, B., Harlid, S., Van Guelpen, B., Skeie, G., Weiderpass, E., Jenab, M., Murphy, N., Riboli, E., Gunter, M.J., Aleksandrova, K.J., Tzoulaki, I., 2019. Comparison of prognostic models to predict the occurrence of colorectal cancer in asymptomatic individuals: a systematic literature review and external validation in the EPIC and UK Biobank prospective cohort studies. Gut 68 (4), 672–683.

Freedman, A.N., Slattery, M.L., Ballard-Barbash, R., Willis, G., Cann, B.J., Pee, D., Gail, M.H., Pfeiffer, R.M., 2009. Colorectal cancer risk prediction tool for white men and women without known susceptibility. J. Clin. Oncol. 27 (5), 686–693.

Park, Y., Freedman, A.N., Gail, M.H., Pee, D., Hollenbeck, A., Schatzkin, A., Pfeiffer, R. M., 2009. Validation of a colorectal cancer risk prediction model among white patients age 50 years and older. J. Clin. Oncol. 27 (5), 694–698.

Usher-Smith, J.A., Walter, F.M., Emery, J.D., Win, A.K., Griffin, S.J., 2016. Risk prediction models for colorectal cancer: a systematic review. Cancer Prev. Res. 9 (1), 13–26.

Han, M., Liew, C.T., Zhang, H.W., Chao, S., Zheng, R., Yip, K.T., Song, Z.-Y., Li, H.M., Geng, X.P., Zhu, L.X., Lin, J.-J., Marshall, K.W., Liew, C.C., 2008. Novel blood-based, five-gene biomarker set for the detection of colorectal cancer. Clin. Cancer Res. 14 (2), 455–460.

Marshall, K.W., Mohr, S., Khettabi, F.E., Nossova, N., Chao, S., Bao, W., Ma, J., Li, X.-J., Liew, C.-C., 2009. A blood-based biomarker panel for stratifying current risk for colorectal cancer. Int. J. Cancer. https://doi.org/10.1002/ijc.24910.

Hsu, L.i., Jeon, J., Brenner, H., Gruber, S.B., Schoen, R.E., Berndt, S.I., Chan, A.T., Chang-Claude, J., Du, M., Gong, J., Harrison, T.A., Hayes, R.B., Hoffmeister, M., Hutter, C. M., Lin, Y.i., Nishihara, R., Ogino, S., Prentice, R.L., Schumacher, F.R., Seminara, D., Slattery, M.L., Thomas, D.C., Thornquist, M., Newcomb, P.A., Potter, J.D., Zheng, Y., White, E., Peters, U., 2015. A model to determine colorectal cancer risk using common genetic susceptibility loci. Gastroenterology 148 (7), 1330–1339.e14.

Law, P.J., Timofeeva, M., Fernandez-Rozadilla, C., Broderick, P., Studd, J., Fernandez-Tajes, J., Farrington, S., Svinti, V., Palles, C., Orlando, G., Sud, A., Holroyd, A., Penegar, S., Theodoratou, E., Vaughan-Shaw, P., Campbell, H., Zgaga, L., Hayward, C., Campbell, A., Harris, S., Deary, I.J., Starr, J., Gatcombe, L., Pinna, M., Briggs, S., Martin, L., Jaeger, E., Sharma-Oates, A., East, J., Leedham, S., Arnold, R., Johnstone, E., Wang, H., Kerr, D., Kerr, R., Maughan, T., Kaplan, R., Al-Tassan, N., Palin, K., Hänninen, U.A., Cajuso, T., Tanskanen, T., Kondelin, J., Kaasinen, E., Sarin, A.-P., Eriksson, J.G., Rissanen, H., Knekt, P., Pukkala, E., Jousilahti, P., Salomaa, V., Ripatti, S., Palotie, A., Renkonen-Sinisalo, L., Lepistö, A., Böhm, J., Mecklin, J.-P., Buchanan, D.D., Win, A.-K., Hopper, J., Jenkins, M.E., Lindor, N.M., Newcomb, P.A., Gallinger, S., Duggan, D., Casey, G., Hoffmann, P., Nöthen, M.M., Jöckel, K.-H., Easton, D.F., Pharoah, P.D.P., Peto, J., Canzian, F., Swerdlow, A., Eeles, R.A., Kote-Jarai, Z., Muir, K., Pashayan, N., Harkin, A., Allan, K., McQueen, J., Paul, J., Iveson, T., Saunders, M., Butterbach, K., Chang-Claude, J., Hoffmeister, M., Brenner, H., Kirac, I., Matošević, P., Hofer, P., Brezina, S., Gsur, A., Cheadle, J.P., Aaltonen, L.A., Tomlinson, I., Houlston, R.S., Dunlop, M.G., 2019. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. Nat. Commun. 10 (1) https://doi.org/10.1038/s41467-019-09775-w.

Archambault, A.N., Su, Y.-R., Jeon, J., Thomas, M., Lin, Y., Conti, D.V., et al., 2020. Cumulative burden of colorectal cancer-associated genetic variants is more strongly associated with early-onset vs late-onset cancer. Gastroenterology 158 (5), 1274–1286.e12.

Huyghe, J.R., Bien, S.A., Harrison, T.A., Kang, H.M., Chen, S., Schmit, S.L., et al., 2019. Discovery of common and rare genetic risk variants for colorectal cancer. Nat. Genet. 51 (1), 76–87.

Awadalla, P., Boileau, C., Payette, Y., Idaghdour, Y., Goulet, J.-P., Knoppers, B., et al. 2013. Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. Int. J. Epidemiol. 42(5):1285–1299.

RAMQ. Table PA.01 - Nombre de personnes inscrites et admissibles au régime d'assurance maladie du Québec selon le sexe, le groupe d'âge et la région sociosanitaire [Internet]. 2017 [cited 2019 Nov 25]. Available from: https://www4.prod.ramq.gouv.qc.ca/IST/CD/CDF_DifsnInfoStats/CDF1_CnsulInfoStatsCNC_iut/DifsnInfoStats.aspx?ETAPE_COUR=3&IdPatronRapp=8&Annee=2017&Per=0&LANGUE=en-CA.

Akçimen, F., Ross, J.P., Sarayloo, F., Liao, C. 2019. De Barros Oliveira R, Ruskey JA, et al. Genetic and epidemiological characterization of restless legs syndrome in Québec. Sleep [Internet]. 2019 [cited 2020 Apr 16];43(4). Available from: https://academic.oup.com/sleep/article/43/4/zsz265/5610251.

Hodgkinson, A., Idaghdour, Y., Gbeha, E., Grenier, J.-C., Hip-Ki, E., Bruat, V., Goulet, J.-P., de Malliard, T., Awadalla, P., 2014. High-resolution genomic analysis of human mitochondrial RNA sequence variation. Science 344 (6182), 413–415.

Hussin, J.G., Hodgkinson, A., Idaghdour, Y., Grenier, J.-C., Goulet, J.-P., Gbeha, E., Hip-Ki, E., Awadalla, P., 2015. Recombination affects accumulation of damaging and disease-associated mutations in human populations. Nat. Genet. 47 (4), 400–404.

Tonelli, M., Wiebe, N., Fortin, M., Guthrie, B., Hemmelgarn, B.R., James, M.T., Klarenbach, S.W., Lewanczuk, R., Manns, B.J., Ronksley, P., Sargious, P., Straus, S., Quan, H., 2016. Methods for identifying 30 chronic conditions: application to administrative data. BMC Med. Informatics Decision Making 15 (1). https://doi.org/10.1186/s12911-015-0155-5.

Statistics Canada. Table 13-10-0111-01 Number and rates of new cases of primary cancer, by cancer type, age group and sex [Internet]. Government of Canada; [cited 2019 Nov 10]. Available from: doi.org/10.25318/1310011101-eng.

Statistics Canada. Table 13-10-0392-01 Deaths and age-specific mortality rates, by selected grouped causes [Internet]. Government of Canada; [cited 2019 Nov 10]. Available from: https://doi.org/10.25318/1310039201-eng.

Choudhury, P.P., Maas, P., Wilcox, A., Wheeler, W., Brook, M., Check, D., et al. 2020. iCARE: an R package to build, validate and apply absolute risk models. PLoS One 15 (2), e0228198.

Rockhill, B., Spiegelman, D., Byrne, C., Hunter, D.J., Colditz, G.A. 2001. Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. J. Natl. Cancer Inst. 93(5), 358–366.

Uno, H., Cai, T., Tian, L., Wei, L.J., 2007. Evaluating prediction rules for t-year survivors with censored regression models. J. Am. Stat. Assoc. 102 (478), 527–537.

Blanche, P., Dartigues, J.-F., Jacqmin-Gadda, H. 2013. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Statist. Med. 32(30), 5381–5397.

Blanche, P., Latouche, A., Viallon, V. 2012. Time-dependent AUC with right-censored data: a survey study. arXiv:12106805 [statME] [Internet]. 2012 Oct 25 [cited 2017 Sep 26]; Available from: http://arxiv.org/abs/1210.6805.

R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/.

Shin, A., Joo, J., Yang, H.-R., Bak, J., Park, Y., Kim, J., et al. 2014. Risk Prediction Model for Colorectal Cancer: National Health Insurance Corporation Study, Korea. Zhang Z, editor. PLoS One 9(2), e88079.

Ma, E., Sasazuki, S., Iwasaki, M., Sawada, N., Inoue, M., 2010. 10-Year risk of colorectal cancer: Development and validation of a prediction model in middle-aged Japanese men. Cancer Epidemiol. 34 (5), 534–541.