UNIVERSITÀ DEGLI STUDI DI MILANO

Facoltà di Medicina e Chirurgia

Dipartimento di Scienze Cliniche e di Comunità

Scuola di Dottorato in Scienze Biomediche Cliniche e Sperimentali

Dottorato in Statistica Biomedica – XXV ciclo

IMPACT OF LIFESTYLE FACTORS ON SCREENING-DETECTED COLORECTAL NEOPLASIA

Dott. Edoardo Botteri

Matricola R08579

Tutor: Dott. Vincenzo Bagnardi

Coordinatore del Dottorato: Ch.mo Prof. Adriano Decarli

Anno Accademico 2011-2012

Index

1. Abstract……page 3

2. The project……page 6

2.1 Background and aims……page 6

2.2 Patients ……page 9

2.3 Statistical methods……page 15

3. Risk factors at the first colonoscopy……page 17

3.1 Descriptive and univariate analysis……page 17

3.2 Multinomial multivariable logistic regression……page 24

3.3 “Spike at zero” function……page 32

3.4 Linear predictor - risk score……page 41

3.5 Nomogram……page 49

4. Risk factors at the second colonoscopy……page 53

3.1 Doctor’s care scheme ……page 54

3.2 Multistate Markov Model……page 70

5. Conclusions……pag 81

6. References……pag 83

1. Abstract

Background and aims: The detection and removal of precancerous lesions through colorectal cancer (CRC) screening, and the intervention on modifiable risk factors for CRC - such as smoking habits, physical activity, red meat consumption and alcohol intake - represent the two possible ways for reducing CRC incidence and mortality. The aim of this project was to investigate whether lifestyle factors, gender, family history and daily low-dose Aspirin use are important factors in predicting endoscopy findings at a first round screening level and whether they can have a significant impact on the natural history of the disease in screened patients during their follow-up (second round screening level).

Patients and methods: Me and my work team identified and selected a study population of 870 men and women of age 50-74 years who underwent a screening colonoscopy at the European Institute of Oncology (IEO) between the years 2007-2009 after a positive Fecal Occult Blood Test (FOBT+). We set up a telephone questionnaire in order to retrieve information on smoking habits, BMI, physical activity, diet, alcohol consumption, family history and usage of low-dose Aspirin at the time of the first colonoscopy. All patients were then interviewed by me by telephone. Ninety-five individuals were not

interviewed for various reasons, making the final population size n=775. Patients who could answer the questionnaire were similar to the unreached individuals in terms of outcome of the first colonoscopy.

Results: At first colonoscopy, we observed 415 patients presenting with a high-risk neoplasia (i.e. 3 or more adenomas or at least one adenoma bigger than 10 mm / with villous component / with high-grade dysplasia or invasive tumor). At the univariate analysis, gender, family history, physical activity, smoking habits, alcohol intake, fruit and vegetable intake and daily low-dose Aspirin were associated with the prevalence of high-risk neoplasia. Using a "Spike at zero function", we showed that light drinkers (<5 grams per day) seemed to have a lower risk of high-risk neoplasia compared to non-drinkers. We concluded that a proportion of non-drinkers might avoid alcohol because of some health conditions linked to the endpoint of interest. At a multivariable level, all those factors remained statistically significantly associated with the outcome of interest. We therefore combined the information of lifestyle factors, gender, family history and daily low-dose Aspirin use to obtain a reliable individual risk score (i.e. linear predictor) and build a nomogram.

The second colonoscopy visit date was fixed in advanced at the time of first colonoscopy, based on the outcome of the first colonoscopy, following a typical example of Doctor's care scheme of examinations. After adjusting for

the severity of the outcome of the first colonoscopy and for the time from first to second colonoscopy, we obtained a statistically significant association between the linear predictor and the risk of high-risk neoplasia detected at the second colonoscopy.

We then applied homogeneous Markov Models to simultaneously model the disease process over time. The effect of the linear predictor on the transitions – from one disease stage to the other – resulted statistically significant. Moreover, as the linear predictor increased, the probability of getting better decreased. In other words, the worse the lifestyle, the lower the probability for the intestinal mucosa to heal. On the other hand, the estimated parameter for the effect of linear predictor on the aggravation transition resulted positive: the worse the lifestyle, the higher the probability to find new high-risk polyps.

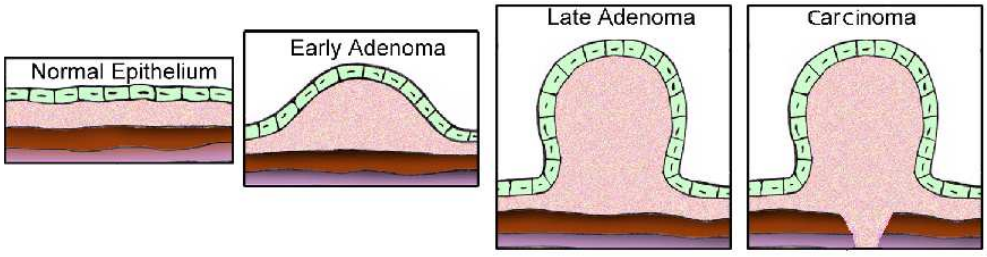
Conclusions: Lifestyle should be considered in the planning of population CRC screenings, because the identification of different risk groups can lead to more tailored screening policies, and accordingly to more efficient and cost-effective interventions.

2. The project

2.1 Background and aims

Colorectal cancer (CRC) is the third most commonly diagnosed cancer in males and the second in females, with over 1.2 million new cancer cases and 608,700 deaths estimated to have occurred in 2008 worldwide[1]. The detection and removal of precancerous lesions through CRC screening, and the intervention on modifiable risk factors for CRC, such as smoking, physical activity, red and processed meat consumption, and alcohol intake, represent the two possible ways for reducing CRC incidence and mortality. Regarding the modifiable factors associated with CRC risk, through the past 2 decades, while a consistent association between cigarette smoking and colorectal adenomatous polyps, recognized precursor lesions of CRC[2,3] (Figure 2.1), has been shown, the smoking-CRC link remained controversial until the very recent years.

Figure 2.1 The adenoma-carcinoma sequence, from normal epithelium to tumor infiltration of the basement membrane



Our group provided strong evidence on the detrimental effect of cigarette smoking on the development of both colorectal adenomatous polyps and CRC, based on two systematic reviews of the literature and meta-analyses[4,5]. In the first study on adenomas, we also showed that the smoking-related increase of risk was significantly greater for high-risk adenomas (villous component or size >10 mm or severe dysplasia) compared to low-risk adenomas, suggesting that smoking may be important for both the formation and aggressiveness of adenomas[4]. In the second study on cancer, we showed how cigarette smoking is significantly associated with both CRC incidence and mortality[5].

Besides smoking, strong evidence on the association between gender, body mass index (BMI), family history, physical activity and the incidence of CRC is well reported in the literature[7-10]. Moreover, there is emerging indication that alcohol consumption, diet, and daily low-dose Aspirin are possible additional factors associated with the risk of CRC[11-13].

Since all these associations could have important implications on future screening policies[6,14], I here present a project aiming at showing that lifestyle-related factors - smoking, alcohol consumption, diet, physical activity and BMI - together with gender, family history and daily low-dose Aspirin use,

could be important factors in predicting endoscopy findings at a first round screening level and that the same factors may also impact on the natural history of the disease in screened patients. In other words, our goal is to show that lifestyle, together with gender, family history and daily low-dose Aspirin use, could possibly represent an important factor to consider when a) deciding on the age at which CRC screening should begin, either by lowering the age in individuals with a poor lifestyle or increasing the age in individuals with a healthy lifestyle and b) deciding how much time should pass from the first screening colonoscopy to the second control colonoscopy, basing the future indications on the finding of the primary colonoscopy as well as on the patients' characteristics.

2.2 Patients

Since 2005, the Italian National Health System has implemented a screening program for CRC for all citizens of 50 years of age or more. Screening tests are free for the target population (so-called Minimal Care Level guaranteed for all Italian citizens). Invitees are asked to take an immunological test for Fecal Occult Blood (FOBT) every two years. Individuals with a positive FOBT test are invited to undergo a total colonoscopy in an SSN-accredited Endoscopy Department.

The identification of the present study population was performed using the Database of the Division of Endoscopy of the European Institute of Oncology (IEO), which collects data on all the IEO patients receiving any health service for the diseases of the gastrointestinal tract and data on their endoscopic findings. We identified and selected a study population of 870 men and women of age 50-74 years who underwent a screening colonoscopy at the IEO between the years 2007-2009 after a positive FOBT. All the patients were participants to the Colorectal Cancer Screening Program of the Lombardy Region.

We decided to select a high-risk population (FOBT positive) in order to work with a larger number of events and consequently gain an adequate power for the study.

Since the relation lifestyle-colorectal neoplasia could be biased by different behavioral correlates of lifestyle, such as tendency for people who a poor lifestyle to delay seeking medical care, we decided to include only asymptomatic patients presenting for their first screening colonoscopy. Patients who had undergone a colonoscopy before the first screening colonoscopy were excluded. Furthermore, patients with any previous or present disease that could affect the lifestyle-related adenoma risk, such as hereditary CRC syndromes, chronic inflammatory bowel disease, history of colorectal polyps or cancer, or previous bowel resection, were excluded. Presence of symptoms and related comorbidities has always been collected in the database.

Then we set up a telephone questionnaire in order to retrieve information on smoking habits, BMI, physical activity, diet, alcohol consumption, family history and usage of low-dose Aspirin. A data-manager created an *ad hoc* database using Microsoft Access 2007.

All patients were then interviewed by me by telephone in order to collect information on their lifestyle, family history of colorectal neoplasia and use of low-dose Aspirin. Patients were also asked if they had undergone an endoscopy before their first screening colonoscopy.

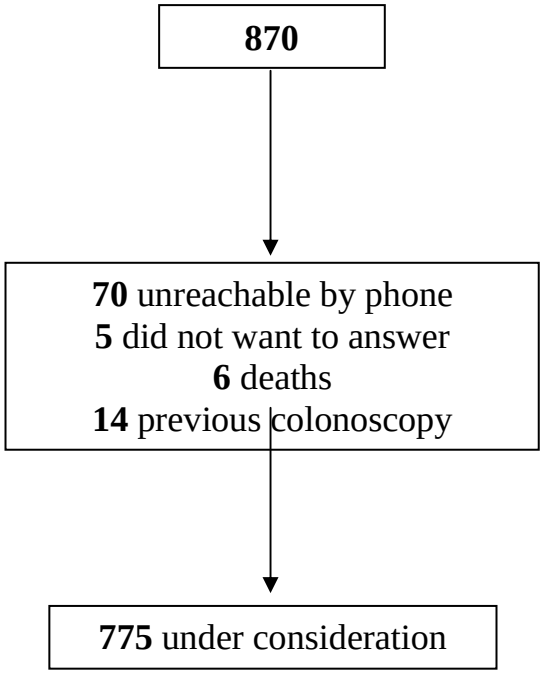
In Table 2.1 the exact telephone questionnaire (originally in Italian) used to collect information and fill in the database is reported. The average duration of a phone-call was about 3.5 minutes, and the average number of attempts to reach an individual was 1.3.

Table 2.1 Telephone questionnaire

1) Height in cm
2) Weight in kg (preferably at the time of colonoscopy or typical weight otherwise)
3) How would you define your job (or your typical work day)?
Sedentary
Necessitating mild-to-moderate physical activity
Necessitating strong physical activity
4) In your spare time, do you regularly practice physical activity / sport?
Yes
No
5) With regard to smoking, at the time of your first colonoscopy, you were:
Current smoker (at the colonoscopy o up to 12 months before)
Ex-smoker (stopped at least 12 months before colonoscopy)
Never smoker
For smokers and ex-smokers only:
6) During your smoking years, what was your average number of cigarettes smoked?
7) When did you start smoking?
For ex-smokers only:
8) When did you stop smoking?
9) Have you ever drunk an alcoholic beverage?
Yes
No
If 9 is yes:
10) Think about the last 7 days: how many days did you happen to drink any alcoholic beverage?
Values 0-7
If 10 is 1 or more:
11) How many drinks of alcoholic beverages have you drunk on average in those drinking days?
Value
12) How many meals of fruit or vegetables do you usually eat?
Value
13) Has any of your relatives ever been diagnosed with a colorectal neoplasia?
Indicate the most severe among the following:
I grade CCR (<60 years old)
I grade CCR (≥ 60 years old)
II grade CCR
Adenomas only
No family history
14) At the time of your first colonoscopy, were you using daily low-dose aspirin?
Yes
No
If 14 is yes:
15) Since when?
Age at start

Ninety-five patients out of 870 (10.9%) were not interviewed for various reasons: 70 (8.0) cases were not reachable by telephone, 5 (0.6%) refuse to answer the questionnaire, 6 (0.7%) died between the date of their first colonoscopy and the date of phone call. Moreover, 14 (1.6%) individuals stated that they had undergone one or more colonoscopy before their first screening colonoscopy and the questionnaire was discontinued (Figure 2.2).

Figure 2.2 Flowchart



Individuals who answered the questionnaire were similar to the individuals who did not in terms of outcome of the first colonoscopy, as shown in Table 2.2.

Table 2.2. Most severe outcome at first colonoscopy

	IN	**OUT**	
Outcome	**No. (%)**	**No. (%)**	**P-value**[a]
No adenoma	227 (29.3)	23 (24.2)	0.28
Low-risk Adenoma	133 (17.2)	34 (35.8)	
High-risk Adenoma	351 (45.3)	33 (34.7)	
Invasive tumour	64 (8.3)	5 (5.3)	
Total	775	95	

[a]Mantel-Haenszel Chi-square test for trend

2.3 Statistical methods

We used a variety of statistical methods to analyze the data for this project. We will describe in details each of the methods - and report the appropriate literature references - when its applications and corresponding results will be reported along the next chapters. Briefly, we used the following statistical methods.

- Simple **descriptive and univariate analyses**: both the Chi-square test and the Chi-square test for trend were used to explore the associations between outcome (endoscopic finding) and individuals' characteristics.
- Multivariable logistic regression and its extension to the **multinomial multivariable logistic regression** were used to identify independent risk factors.
- **"Spike at zero" functions**, based on fractional polynomials**,** were used to estimate the dose–response function for continuous exposures in circumstances where there was a certain percentage of unexposed individuals (i.e. never smokers).
- A multivariable linear predictor was computed to assign to each patient a **Risk Score** predicting the probability of poor endoscopic outcome.

- A **Nomogram** was built to map the predicted probabilities of poor endoscopic outcome into points on a scale from 0 to 100 in a user-friendly graphical interface.
- We finally applied **Homogeneous Markov Models** to simultaneously model the disease process over time, and evaluate the effect of lifestyle and other characteristics on each transition from one state to another.

Descriptive, univariate and multivariable logistic regression analyses were carried out with the SAS software (SAS Institute, Cary, NC). The Nomogram and the Markov Models analysis were computed using the R (http://cran.r-project. org/) software. For the “Spike at zero” functions we used both the R software and the STATA (College Station, TX, USA) software.

3. Risk factors at the first colonoscopy

In this chapter we will evaluate whether lifestyle factors, gender, family history and low-dose Aspirin have an impact on the detected neoplasia at the first screening colonoscopy. If a statistically significant association between those factors and the detected neoplasia is demonstrated, important conclusions will be drawn. Those factors should indeed be considered when deciding on the age at which CRC screening should begin, either lowering the age in the bad prognostic group (high-risk of neoplasia) or increasing the age in the good prognostic group (low-risk of neoplasia).

3.1 Descriptive and univariate analysis

In order to synthesize the information regarding all the observed outcomes, from normal mucosae to invasive tumors (Table 3), we identified and grouped the endoscopic findings which should be considered at high risk of developing a CRC (high-risk adenomas). The 2006 guideline on postpolypectomy surveillance of the United States Multi-Society Task Force (MSTF) on CRC will be used to distinguish two main types of adenomas: (1) *low-risk adenomas*, defined as 1–2 tubular adenomas < 10 mm, and (2) *high-*

risk adenomas, defined as adenoma with villous histology, high-grade dysplasia, ≥ 10 mm, or 3 or more adenomas[15]. We then grouped the *invasive tumors* together with the *high-risk adenoma* to form the high-risk neoplasia category.

Table 3.1. Most severe outcome at first colonoscopy in details

	Outcome	No. (%)	
	Normal Mucosae	43 (5.5)	
	Non-oncological alteration	148 (19.1)	
	Non-adenomatous polyp	36 (4.6)	
	Low-risk Adenoma <10 mm [a]	133 (17.2)	
High-risk adenoma	High-risk Adenoma <10 mm	99 (12.8)	High-risk neoplasia
High-risk adenoma	Adenoma 10-19 mm	193 (24.9)	High-risk neoplasia
High-risk adenoma	Adenoma ≥ 20 mm	59 (7.6)	High-risk neoplasia
	Invasive tumour	64 (8.3)	High-risk neoplasia
	Total	**775**	

[a] One or two adenomas < 10 mm with no villous component and no evidence of high-grade dysplasia

The outcomes reported in Table 3.1 are the detailed colonoscopy outcomes of the 775 patients who answered the questionnaire. One-hundred and ninety-one individuals (24.6%) had no polyps (non-oncologic alterations were mainly hemorrhoids and diverticula); 36 had non-adenomatous (hyperplastic in the vast majority) polyps. A patient was classified in the

category “low-risk adenoma” if his/her endoscopic finding was one or two adenomas < 10 mm in diameter with no villous component and no evidence of high-grade dysplasia; patients with 3 or more adenomas, or at least one adenoma bigger than 10 mm or with villous component or with high-grade dysplasia were classified in the category “high-risk adenoma” (n=351, 45.3%). Finally, 64 (8.3%) patients were found with an invasive neoplasia, 20 (2.6%) of them were diagnosed as adenocarcinoma in a polyp (i.e. cancerous polyp), while 44 (5.7%) were proper invasive tumors. Forty-two of them were adenocarcinomas, 1 was a neuroendocrine tumor and one was a spinocellular carcinoma. All the 44 patients with invasive tumors underwent radical surgery and received adjuvant treatment according to the stage of the disease.

Table 3.2. Characteristics of population and prevalence of high-risk neoplasia at first colonoscopy

	Categories	No. (col %)	High-risk Neoplasia No. (row %)	P
All individuals		**775 (100.0)**	**415 (53.5)**	
Gender	Male	400 (51.6)	250 (62.5)	<0.01
	Female	375 (48.4)	165 (44.0)	
Age	50-60	268 (34.6)	133 (49.6)	0.11
	61-67	296 (38.2)	162 (54.7)	
	68-74	211 (27.2)	120 (56.9)	
Family history [a]	None	667 (87.6)	358 (52.9)	0.08 0.02 [b]
	2nd grade - CRC	23 (3.0)	10 (43.5)	
	1st grade - CRC ≥ 60 years	53 (6.9)	30 (56.6)	
	1st grade - CRC < 60 years	20 (2.6)	16 (80.0)	
Physical activity	Weak	303 (39.1)	180 (59.4)	<0.01
	Moderate	206 (26.6)	111 (53.9)	
	Strong	266 (34.3)	124 (46.6)	
BMI	< 25.0 (Normal weight)	366 (47.2)	194 (53.0)	0.50
	25.0-29.9 (Overweight)	311 (40.1)	177 (56.9)	
	≥ 30.0 (Obese)	98 (12.7)	44 (44.9)	
Smoking status	Never smoker	353 (45.6)	172 (48.7)	<0.01
	Former smoker	228 (29.4)	124 (54.4)	
	Current smoker	194 (25.0)	119 (61.3)	
Smoking (pack-years)	0	353 (45.6)	172 (48.7)	<0.01
	1-15	76 (9.8)	35 (46.1)	
	16-30	133 (17.2)	76 (57.1)	
	31-40	94 (12.1)	55 (58.5)	
	> 40	119 (15.4)	77 (64.7)	
Alcohol intake [a] (grams/day)	0	316 (41.0)	151 (47.8)	<0.01
	0.1-12.4	145 (18.8)	67 (46.2)	
	12.5-24.9	133 (17.2)	77 (57.9)	
	≥ 25.0	177 (23.0)	117 (66.1)	
Fruit and vegetable intake (meals per day) [a]	≤ 2	225 (29.5)	137 (60.9)	<0.01
	3-4	368 (48.2)	206 (56.0)	
	≥ 5	171 (22.4)	62 (36.6)	
Daily low-dose Aspirin [a]	Never user	661 (86.4)	360 (54.5)	0.08 0.02 [c]
	≤ 5 years	49 (6.4)	30 (61.2)	
	> 5 years	55 (7.2)	21 (38.2)	

[a] Some patients had missing values; [b] 1st grade - CRC < 60 years *vs* others; [c] > 5 years *vs* others.

We reported in Table 3.2 the characteristics of the 775 individuals who answered the questionnaire. The study population was divided almost equally between men and women. The majority of the patients declared no family history and only 20 patients out of 775 (2.6%) reported a family history of CRC in a first-degree relative who was diagnosed under the age of 60 years. Typical work day and sports were combined in one variable, *physical activity*, expressed in three categories: weak if the individual declared to lead a sedentary life; moderate if the individual declared to do mild to moderate physical activity during his/her work day and no sports; strong otherwise. A few patients were obese. With regard to smoking, 353 individuals declared they never smoked (45.6%), while 194 declared to be "*current smokers*" at the time of their colonoscopy. With regard to alcohol drinking, 316 individuals declared to be teetotalers (41.0%) while 177 (23.0%) to drink at least 2 drinks per day. We used grams per day (g/day) as a standard measure of ethanol intake, 12.5 grams being the standard alcohol intake per drink of any alcoholic beverage. Moreover 225 (29.5%) people reported a low intake of fruit and vegetables (2 or less meals per day). Finally, 55 (7.2%) individuals had been taking low-dose Aspirin for more than 5 years before the colonoscopy. Five years is the length of time that has been recognized to have a clear protective effect on CRC[13].

We observed 415 patients presenting with a high-risk neoplasia (high-risk adenoma or invasive neoplasia). At the univariate analysis, male gender, high-risk family history (first grade relative diagnosed with CRC at a young age), low physical activity and low fruit/vegetable intake were associated with a higher risk of high-risk neoplasia. A long-term use of daily low-dose Aspirin was associated with a low prevalence of high-risk neoplasia, while a short-term use of daily low-dose Aspirin was associated with prevalence of high-risk neoplasia similar to the one observed for the never users. As for the smoking habit and alcohol intake, the association was statistically significant, but a clear increase in risk was not observed for low consumption of neither tobacco nor alcohol.

Prevalence of high-risk neoplasia did not increase significantly as the age increased. This could be explained by the fact that, despite age is a well-known risk factor for CRC, the age range in our population was quite narrow (50-74 years) and only little variation in risk could be observed. Neither BMI was statistically associated with the risk of high-risk neoplasia. The lack of association between BMI and high-risk neoplasia was quite surprising, since high BMI is a well known risk factor for CRC. So how can we possibly explain the absence of association reported in this analysis? We hypothesized that patients with a high BMI (e.g. >25) are at higher risk of hemorrhoids and

diverticula compared to patients with a lower BMI. Moreover, we must remember that all the individuals of this study population had a previous positive FOBT, which can be associated with the presence of hemorrhoids and diverticula. All this could have lead to an over-representation of the population with high values of BMI in the “No polyps” reference outcome category, this causing to a dilution of the effect of BMI on the risk of high-risk neoplasia.

3.2 Multinomial multivariable logistic regression

Multinomial logistic regression is the extension for the binary logistic regression when the categorical dependent outcome has more than two levels[16]. Consider a random variable Y_i that may take one of several discrete values, which we index 1,2,…,J. In our case, the response is a recategorization of "most severe outcome at first colonoscopy" (see Table 3.3) taking the values J=1 for the categories "Normal mucosae", "Non-oncological alteration" and "Non adenomatous polyp", J=2 for "Low-risk adenoma" and J=3 for "High-risk adenoma" and "Invasive neoplasia".

Table 3.3. Recategorization of the outcome in three categories

Category (J)	Finding at colonoscopy
1	Normal mucosae Non-oncological alteration Non-adenomatous polyp
2	Low-risk adenoma
3	High-risk adenoma Invasive neoplasia
	Total

This recategorization of the outcome led to comparable frequencies among the categories and has a clinical significance.

Let

$$\pi_{ij} = \Pr\{Y_i = j\}$$

denote the probability that the i-th response falls in the j-th category. In the example π_{i1} is the probability that the i-th respondent resulted in a normal mucosae or non-oncological alteration or non-adenomatous polyp.

We now consider models for the probabilities π_{ij}. In particular, we would like to consider models where these probabilities depend on a vector X_i of covariates associated with the *i*-th individual or group. We nominate one of the response categories as a baseline or reference cell, calculate log-odds for all other categories relative to the baseline, and then let the log-odds be a linear function of the predictors. We pick the first category as a baseline (normal mucosae or non-oncological alteration or non-adenomatous polyp). In the multinomial logit model we assume that the log-odds of each response follow a linear model

$$\eta_{ij} = \log \frac{\pi_{ij}}{\pi_{i1}} = \alpha_j + \mathbf{x}_i' \boldsymbol{\beta}_j,$$

where α_j is a constant and β_j is a vector of regression coefficients, for j = 2, 3…J (in our analysis for j=2 and 3, as 1 is the reference category).

This model is analogous to a logistic regression model, except that the probability distribution of the response is multinomial instead of binomial and we have J-1 equations instead of one.

The multinomial logit model may also be written in terms of the original probabilities π_{ij} rather than the log-odds.

$$\pi_{ij} = \frac{\exp\{\eta_{ij}\}}{\sum_{k=1}^{J} \exp\{\eta_{ik}\}}.$$

Note that the convention $\eta_{i1} = 0$ makes this formula valid for all j.

To describe smoking, we used the pack years of smoking for the tobacco exposure, calculated as the mean number of packs smoked per day multiplied by the number of years that the patient smoked. We used a standard categorization of pack years, i.e. 0, 1-15 and >15, because 15 years of smoking (corresponding to 15 years of pack years on average) are thought to be necessary to cause major DNA damages that lead to polyps[14].

In Table 3.4 we reported the results from the multivariable multinomial logistic regression analysis

Table 3.4. Multivariable multinomial logistic regression analysis for high-risk neoplasia

		Type of neoplasia	OR (95% CI)	P[a]
Gender	**Male vs female**	**Low-risk**	**1.04 (0.65 - 1.66)**	**<0.01**
		High-risk	**1.79 (1.20 - 2.68)**	
Family history	1st grade - CRC < 60 years *vs* others	Low-risk	1.08 (0.15 - 7.89)	0.14
		High-risk	3.32 (0.72 - 15.35)	
Physical activity	Moderate vs Weak	Low-risk	0.69 (0.39 - 1.23)	0.91
		High-risk	0.71 (0.44 - 1.16)	
	Strong vs Low	Low-risk	0.75 (0.44 - 1.28)	0.37
		High-risk	0.61 (0.38 - 0.97)	
Pack-years of smoking	15.1-30 vs ≤ 15	Low-risk	1.77 (0.99 - 3.17)	0.43
		High-risk	1.46 (0.87 - 2.44)	
	>30 vs ≤ 15	Low-risk	2.23 (1.25 - 3.99)	0.96
		High-risk	2.21 (1.33 - 3.66)	
Alcohol (grams/day)	12.5-24.9 vs <12.5	Low-risk	1.10 (0.60 - 2.03)	0.35
		High-risk	1.41 (0.83 - 2.38)	
	≥ 25 vs <12.5	**Low-risk**	**0.97 (0.53 - 1.80)**	**0.03**
		High-risk	**1.73 (1.05 - 2.87)**	
Fruit and vegetable intake (meals per day)	3-4 vs ≤ 2	Low-risk	0.87 (0.50 - 1.50)	0.91
		High-risk	0.89 (0.55 - 1.43)	
	> 4 vs ≤ 2	Low-risk	0.48 (0.26 - 0.90)	0.18
		High-risk	0.33 (0.19 - 0.57)	
Daily low-dose Aspirin usage	≤ 5 years vs Never user	Low-risk	0.51 (0.19 - 1.38)	0.36
		High-risk	0.77 (0.36 - 1.65)	
	> 5 years vs Never user	Low-risk	0.42 (0.19 - 0.91)	0.40
		High-risk	0.30 (0.15 - 0.58)	

[a]Wald test, testing homogeneity of odds ratios between low and high-risk

The Wald test evaluates whether or not the independent variable is statistically significant in differentiating between the two categories in each of the embedded binary logistic comparisons. For example, when we compared high intake (>4 meals of fruits and vegetables) versus low intake (≤2 meals) we obtained β_2=-0.729 for low-risk adenomas and β_3=-1.1152 for high-risk adenomas. Given β_2-β_3=0.386, VAR(β2)=0.102, VAR(β_3)=0.078 and COV(β_2,β_3)=0.049 we can calculate VAR(β_2,β_3)= 0.102+0.078-2(0.049)=0.082 and a standardized normal empirical value of 1.349, which corresponds to a 2 sided p-value of 0.177. We accept the null hypothesis that high versus low intake of fruits and vegetables has the same protective effect on the prevalence of low-risk adenomas (category 2) and high-risk adenomas (category 3).

A very interesting result is that high intakes of alcohol (2 drinks per day or more) and male gender have a differential association with low-risk adenomas and high-risk adenomas. Drinking 2 drinks per day or more and being a man seem to decrease the time latency from normal mucosae to high-risk adenoma or from low-risk adenoma to high-risk adenoma. If we had used only the classical binary logistic regression we would have missed this important information (see Table 3.5).

Table 3.5. Multivariable logistic regression analysis

Variable	Comparison	High-risk neoplasia OR (95% C.I.)
Gender	Male vs female	1.76 (1.27 - 2.45)
Family history	1st grade - CRC < 60 years *vs* others	3.24 (1.04 - 10.12)
Physical activity	Moderate vs Low	0.86 (0.59 - 1.27)
	High vs Low	0.71 (0.48 - 1.03)
Smoking (pack-years)	15.1-30 vs ≤ 15	1.10 (0.73 - 1.65)
	>30 vs ≤ 15	1.46 (1.00 - 2.14)
Alcohol (grams/day)	12.5-24.9 vs <12.5	1.33 (0.87 - 2.01)
	≥ 25 vs <12.5	1.73 (1.16 - 2.59)
Fruit and vegetable (meals per day)	3-4 vs ≤ 2	0.97 (0.67 - 1.40)
	> 4 vs ≤ 2	0.46 (0.29 - 0.73)
Daily low-dose Aspirin usage	≤ 5 years vs Never user	1.05 (0.55 - 1.99)
	> 5 years vs Never user	0.44 (0.24 - 0.80)

The final multivariable model is reported in Table 3.5. We went back to the simple binary outcome, because the primary aim of our project is to evaluate which factors are associated with the risk of high-risk neoplasia. The model showed that men had a 76% risk increase of having a high-risk neoplasia compared to women. As expected, individuals with a high-risk profile of family history (i.e. individuals who had a first grade relative diagnosed with CRC at a young age) were characterized by a more than three-fold increase in risk of high-risk neoplasia when compared to individuals with no family history or low-risk family history. A long-term consumption of low-dose Aspirin was associated with a statistically significantly reduced risk of

high-risk neoplasia, confirming recent published evidence[13,14]. All modifiable lifestyle factors were statistically associated with the risk of high-risk neoplasia (with the exception of physical activity which showed a borderline statistically estimate; OR=0.71 with 95% confidence interval 0.48 - 1.03). All these significant associations represent probably the most important finding in the first phase of our analysis, because we provided strong evidence that supports the role in CRC risk of physical activity, diet, smoke and alcohol habits, and medication use, which all are potentially modifiable factors. Moreover, all these factors should probably be considered in the decision process about the age at which CRC screening should begin, either by lowering the age in individuals with a poor lifestyle or increasing the age in individuals with a healthy lifestyle.

3.3 “Spike at zero” functions

A common task in epidemiology is to estimate the dose–response function for a continuous exposure. Spike at zero functions[17] can be used when there is a certain percentage of unexposed individuals. Typical examples are cigarette consumption, alcohol intake, or occupational exposures. The subjects who are not exposed may be characterized by unknown or uncollected factors which might be associated to outcome in the study. A classical example is represented by the association between alcohol and cardiovascular diseases: a percentage of non drinkers might avoid alcohol because of their health conditions, this leading to slightly decreased risk of disease in moderate drinkers compared to non drinkers. For this reason it is useful to analyze separately exposed and not exposed, albeit within the same model. Any model of continuous exposure variables – i.e. fractional polynomials (FP) and spline functions - could be extended to allow for a proportion of unexposed individuals.

Fractional Polynomial-based "Spike at zero" function

Royston and Altman[18] introduced and formalized the fractional polynomial (FP) models in 1994. A first-order fractional polynomial (FP1) is written as:

$$\beta_0 + \beta_1 x^{p_1},$$

while a second-order fractional polynomial (FP2) is written as:

$$\beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2},$$

and so on. The powers p are chosen from a restricted set, S = {−2, −1, −0.5, 0, 0.5, 1, 2, 3}.

Assume that x≥0 for all individuals. In order for FP functions of x to be defined at x=0, the origin of x is shifted by adding a small constant, c, before analysis. By default, we take c as the smallest difference between successive positive values of x[17]. Consider a model whose linear predictor, η, is given by:

$$\eta = \begin{cases} \beta, & x=0 \\ \beta_0 + \mathrm{FP}m(x+c;\ p_1, p_2, \ldots p_m), & x>0 \end{cases}$$

The linear predictor η is a FPm function of x+c when x>0 and a constant (β) when x=0. Thus η is a discontinuous function of x with a possible jump at x=0. For a first-grade FP, we can re-write the expression for η as:

$$\eta = \beta_0 + (\beta - \beta_0)z + (1-z)\mathrm{FP}1^{+}(x+c;\ p_1)$$

where:

$$z = \begin{cases} 1, & x=0 \\ 0, & x>0 \end{cases},$$

$$\mathrm{FP}1^{+}(x+c;\ p_1) = \begin{cases} 0, & x=0 \\ \mathrm{FP}1(x+c;\ p_1), & x>0 \end{cases}$$

As reported by Royston et al. in their paper[17]:

"The FSP-spike procedure for selecting a model has two stages. 1. In the first stage, the most complex model comprising z and the best $FP2^{+}(x+c;p1,p2)$ is compared with the null model on 5 d.f. (4 d.f. from the best FP2 model plus one from the binary z term). If the test is significant, the steps of the FSP for selecting an FP function are followed, but with z always included in the model. If the test is not significant, stop, concluding that the effect of x is 'not significant' at the alfa level. Otherwise continue. Test $FP2^{+}(x+c;p1,p2)$ against the best straight line at the alfa level using 3 d.f. If the test is not significant, stop, the final model being a straight line. Otherwise continue. Test $FP2^{+}(x+c;p1,p2)$ against the best $FP1^{+}(x+c;p1)$ at the alfa level using 2 d.f. If the test is not significant, the final model is FP1, otherwise the final model is FP2. End of the procedure.

2. In the second stage (performed separately), z and the remaining FP or linear component are each tested for removal from the model. If both parts are significant, the final model includes both; if one or both parts are non-significant, the one with the smaller deviance difference is removed. In the latter case, the final model comprises either the binary dummy variable or the selected FP function. If only an FP function is selected, the spike at zero plays no further part. Since the selection of an FP function may be affected by the presence of the binary dummy variable, the resulting model may differ from that from a standard FP analysis".

FSP-spike procedure for alcohol

First Stage: function selection procedure: determine the 'best' function from the FP class

	Deviance	Distance to Dev(FP2+Z)	d.f.	P	Power(s)
null	1065.187	25.397	5	0.000	
lin+Z	1048.904	9.114	3	0.028	1
FP1+Z	1043.364	3.574	2	0.167	0
FP2+Z	1039.790	-			1, 3

FP2+Z was not statistically better than FP1+Z. Therefore, at the first stage we chose FP1+Z.

Second Stage: z and the chosen FP are each tested

	Deviance	Distance to Dev(FP1+Z)	d.f.	P	
FP1+Z	1043.364	-			0
FP1 (Dropping Z)	1048.872	5.508	1	0.019	0
Z (Dropping FP1)	1058.311	14.947	2	0.001	

Both terms were significant. We accepted to keep FP1+z as the final model

If we compare AIC of the selected spike function with the one of the best FP1 we obtain:

Model	AIC	d.f.	Power
FP1 + z	1048.4	3	0
$FP1^{+}$	1050.6	2	0.5

AIC is a criterion for selecting an optimum model in a class of nested and non-nested models or models fitted on different samples. It takes into account both the binomial deviance and the degrees of freedom of each model and was defined as:

$$AIC(m) = -2L(m) + 2k(m)$$

where $L(m)$ is the maximum log-likelihood for the m model and $k(m)$ is the number of predictors for the m model. Better models have smaller AIC.

So the best model was the spike function model, which can be written as:

Logit[(P(Outcome=High Risk Neoplasia)]= -0.8901+ (Z) 0.8014 + (1-Z) 0.4495 log(Grams/day)

So, for example:

a) Non drinkers

LOGIT[P(Outcome=High Risk Neoplasia)]= - 0.0887

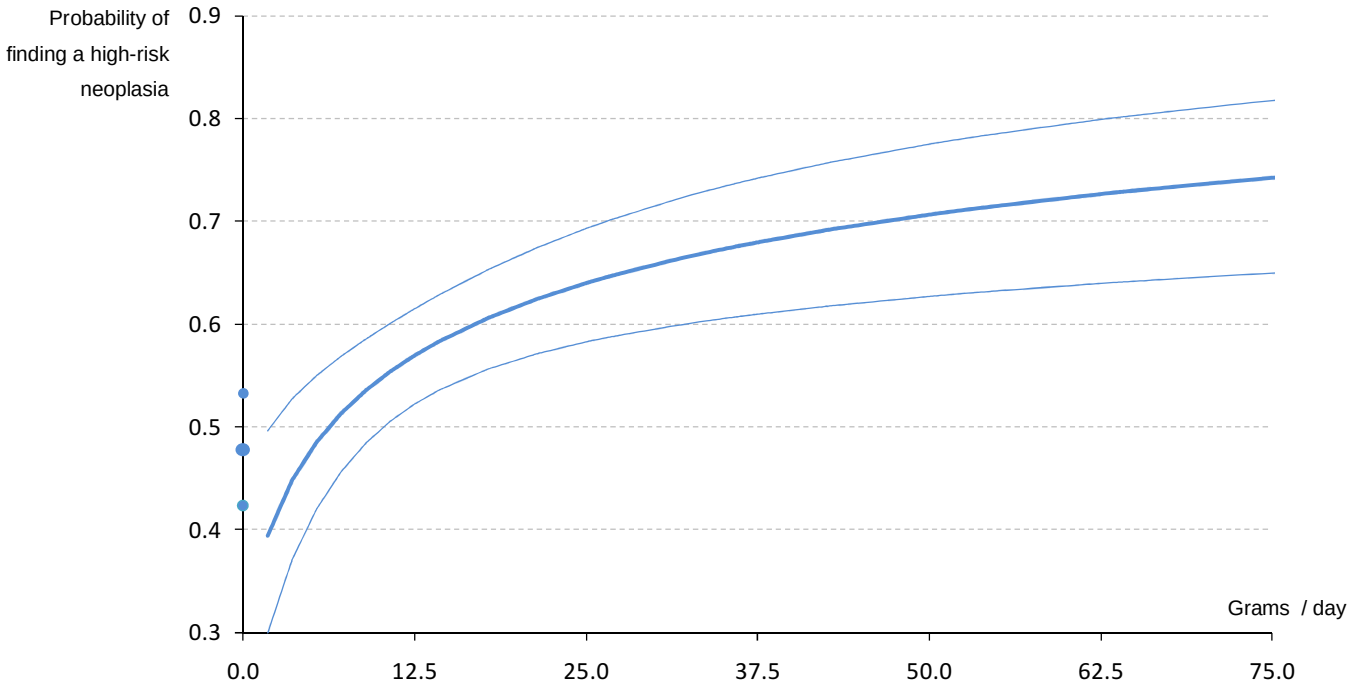
P(Outcome=High Risk Neoplasia)=exp(-0.0887)/1+exp(-0.0887)=0.48

b) Drinkers of 40 grams /day

LOGIT[P(Outcome=High Risk Neoplasia)]= -0.8901+0.4495 LN(40)=0.768

P(Outcome=High Risk Neoplasia)= exp(0.768)/1+exp(0.768)=0.68

Figure 3.1. Association between alcohol and high-risk neoplasia; spike at zero function, $FP1^{+}$ and p1=0



For non-drinkers, there was a probability of 0.48 (95% CI 0.42-0.53) of having a colorectal high-risk neoplasia detected at their first colonoscopy. Light drinkers (<12.5 grams/day i.e. 1 drink per day) did not seem to have a higher risk of colorectal neoplasia compared to non-drinkers. When considering the lower doses, the FP1 function was steeply increasing with increasing number of grams/day, then a lessening increase rate is shown (Figure 3.1).

Noteworthy, the chosen spike at zero function (FP1+z, p1=0) had a better relative goodness of fit (AIC=1048.4) when compared with the best plain FP1 model (p=0.5; AIC = 1050.6). We can therefore *hypothesize* that a proportion of non-drinkers might avoid alcohol because of some health conditions linked to the endpoint of interest.

FSP-spike procedure for smoking

	Deviance	Distance to Dev(FP2+Z)	d.f.	P	Power(s)
null	1068.937	14.145	5	0.015	
lin+Z	1063.142	8.35	3	0.039	1
FP1+Z	1058.538	3.746	2	0.154	0
FP2+Z	1054.792	-			1, 2

FP2+Z was not statistically better than FP1+Z. Therefore, from the first stage we chose FP1+Z.

	Deviance	Distance to Dev(FP1+Z)	d.f.	P	Power(s)
FP1+Z	1058.538	-			0
FP1 (Dropping Z)	1060.846	2.308	1	0.129	0
Z (Dropping FP1)	1066.133	7.595	2	0.022	

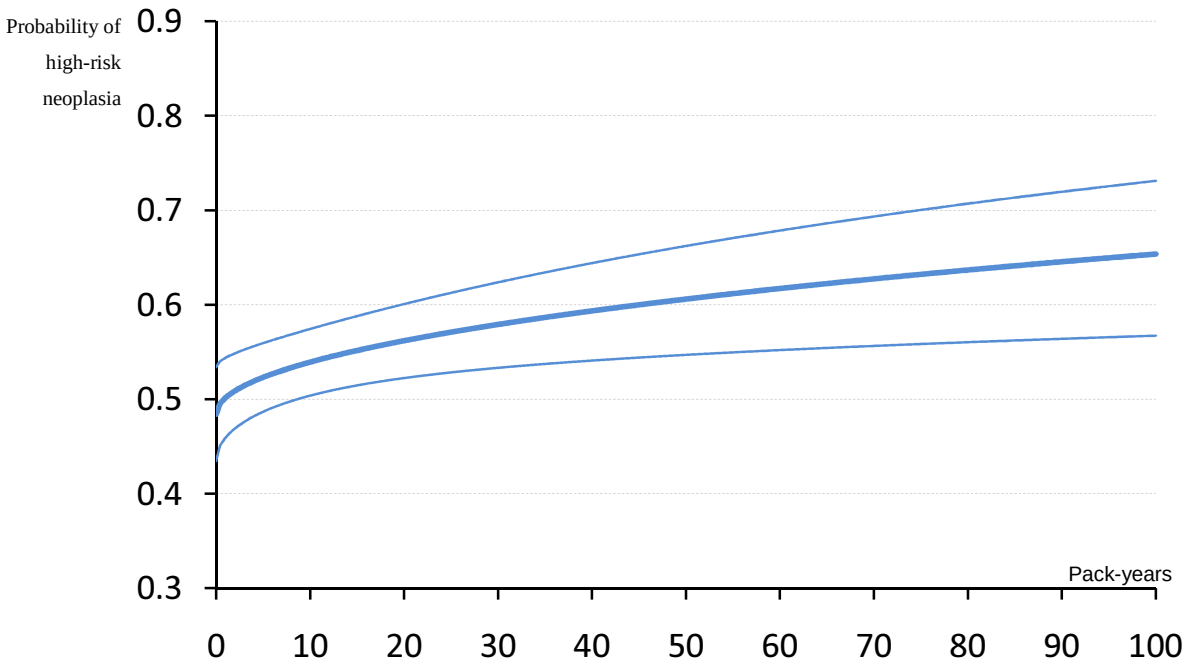
FP1+Z was not better than FP1, therefore we drop the spike at zero model. Only the FP function was selected, as the spike at zero plays no further part. In confirmation of this, the AIC of the spike model is higher than the AIC of the simple best FP1.

Model	AIC	d.f.	Power
Best FP1 + z	1064.5	3	0
Best $FP1^{+}$	1063.5	2	0.5

So the best model was the $FP1^{+}$ function model with power=0.5, which can be written as:

$$\text{Logit}[P(\text{Outcome=High Risk Neoplasia})] = -0.0635 + 0.0699\ (\text{Pack-years})^{0.5}$$

Figure 3.2. Association between pack-years of smoking and high-risk neoplasia. $FP1^{+}$, p1=0.5



As shown in the Figure 3.2, there was an increasing prevalence of high-risk neoplasia with increasing number of pack-years.

3.4 Linear predictor - risk score

We then wanted to build a unique linear predictor, an individual risk score, based on the information deriving from all the studied variables. At the multivariable level, we modeled age and fruit/vegetables consumption as continuous variables and assumed a linear relationship between those covariates and the log-odds of high-risk neoplasia. Then, since the advantage of a spike at zero function for alcohol was no longer significant at a multivariable level, we used the best first-order fractional polynomial function (i.e. the one with power=0.5) to evaluate the association between alcohol and the log-odds of high-risk neoplasia. The same first-order fractional polynomial function was used for smoking. Gender was used as dichotomous variable, as well as family history, as described in the table below. Physical activity was used in three categories and therefore two dummy variables were used in the model. Since in the previous analysis the Aspirin effect was evident after a long-term consumption (see Table 7), we dichotomized the variable in > 5 years of consumption versus ≤ 5 years (the latter category including the never users).

We hereafter report the final linear predictor of the log-odds of high-risk neoplasia:

Logit [Pr(Oucome=High Risk Adenoma)] =

Variable	Description	Parameter estimates	P-value
Intercept		-0.083	0.9238
Age	*Continuous*	+0.009	0.4933
Gender	*M=1; F=0*	+0.557	0.0009
Family History	*1st grade < 60 yrs = 1; others=0*	+1.211	0.0391
Moderate phisical activity	*Yes=1; No=0*	-0.217	0.2846
Strong phisical activity	*Yes=1; No=0*	-0.376	0.0535
Smoking (pack-years)	*Continuous*	+0.038	0.0649
Alcohol (grams/day)	*Continuous*	+0.375	0.0076
Fruit/vegetables meals/day	*Continuous*	-0.221	0.0007
Low-dose aspirin	*> 5 years = 1; others=0*	-0.859	0.0036

The risk of finding a high-risk neoplasia at the first screening colonoscopy significantly increased with increasing alcohol consumption, pack-years of smoking (borderline significant) and decreased with increasing fruit and vegetables consumption. Male gender and high-risk profile of family history were associated with an increased risk of high-risk neoplasia, while long-term consumption of low-dose Aspirin and strong physical activity were associated with decreased risk of high-risk neoplasia. Age was not statistically associated with the risk of high-risk neoplasia.

Model accuracy

Evaluating the model accuracy, that is assessing the model's ability to accurately fit the data, is a critical step in the modelling process to guarantee robust estimates calculations. We used an internal validation of predictive logistic regression models for the decision-making based on the evaluation of both discrimination and calibration[19]. Discrimination refers to the correct relative ranking of predicted probabilities of a specific event, whereas calibration describes whether predicted probabilities are too high or too low relative to true population values.

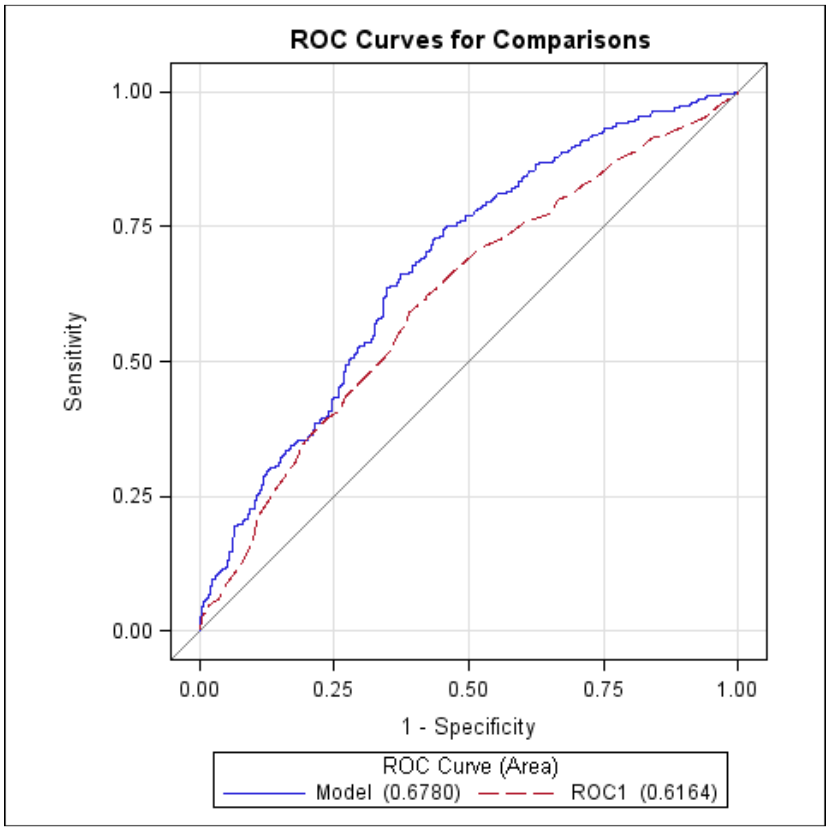
Discrimination

A widely accepted measure of discrimination ability of a predictive model is the *c*-index (for concordance), which applies to predictions that are continuous, dichotomous, ordinal, and censored time-to-event outcome predictions[20]. In binary cases, *c*-index is equivalent to the area under the Receiver Operating Characteristic (ROC) curve, which is a common method of measuring the predictive ability of logistic regression models.

The curve is constructed by varying the cut-point that determines which estimated event probabilities are considered to predict the event. The curve plots the proportion of incorrectly predicted outcomes (1-specificity) on the x-axis and the proportion of correctly predicted outcomes (sensitivity) at a given cut-point on the y-axis. The area under a ROC curve (*c*-index), which ranges from zero to one, provides a measure of the model's ability to discriminate between those subjects who experience the outcome of interest (high-risk neoplasia) versus those who did not. The greater the area under the ROC curve the better the model's discriminatory power.

We used the ROCCONTRAST statement in SAS to implement the non-parametric approach of DeLong, DeLong, and Clarke-Pearson to compare ROC curves[21]. When two curves are constructed based on regression models performed on the same individuals, statistical analysis on differences between curves must take into account the correlated nature of the data. DeLong *et al.* presented a nonparametric approach to the analysis of areas under correlated ROC curves.

Figure 3.3. ROC curves



We built two models, the first named “ROC1” and the other “Model”. The first one derived from a multivariable logistic regression model including age, gender and family history as covariates. These three variables are the most recognized risk factors of CRC[10]. The second one derived from a multivariable logistic regression model including age, gender, family history plus all the lifestyle factors and low-dose Aspirin use.

We then calculated the area under the ROC curve for the two models and built a test to compare ROC curves.

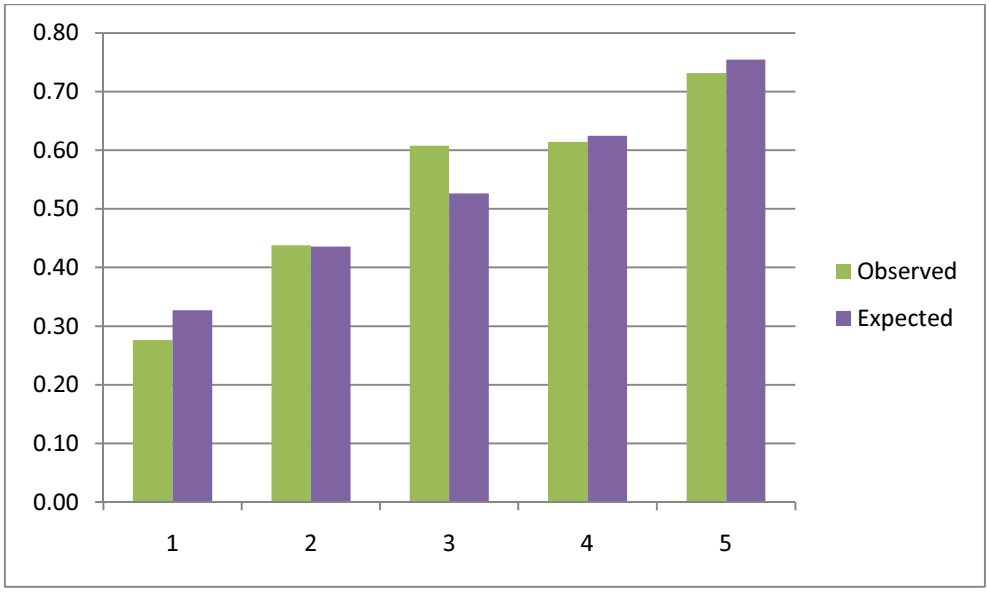
1) C-index for all variables ("Model"): 0.678
2) C-index for known risk factors ("ROC1"): 0.616
3) ROC Contrast Test Results; Chi-Square=12.7; P<0.001

ROC values of around 0.7 are considered to indicate a good discriminating model[18]. Therefore both models had good discrimination ability. But, given the results from the ROC Contrast Test, we could conclude that the modifiable lifestyle factors adds additional information to the model with only age, gender, and family history (ROC1) in distinguishing between patients who were diagnosed with high-risk neoplasia and those who were not.

Calibration

Calibration refers to whether the predicted probabilities agree with the observed probabilities.

Figure 3.4. Probability of high-risk neoplasia at first screening colonoscopy according to quintiles of the linear predictor: observed versus expected



We evaluated the calibration of the logistic regression models using the Hosmer–Lemeshow test[16]. We used quintiles to re-categorize the distribution of expected and observed probabilities.

Hosmer–Lemeshow test calculation

Quintile	N	Obs Events	Exp Events	Obs Probability	Exp Probability	W= Exp Probability / (1-Exp Probability)	Obs Events - Exp Events	(Obs Events - Exp Events) ^ 2	[(Obs Events - Exp Events) ^ 2] / W
1	152	42	49.8	0.28	0.33	0.22	-7.78	60.53	1.81
2	153	67	66.7	0.44	0.44	0.25	0.32	0.10	0.00
3	153	93	80.6	0.61	0.53	0.25	12.45	154.89	4.06
4	153	94	95.5	0.61	0.62	0.23	-1.50	2.26	0.06
5	153	112	115.5	0.73	0.75	0.19	-3.48	12.14	0.43
TOTAL	**764**	**408**	**408**					**Chi-square**	**6.36**
								P-value	**0.10**

Since Hosmer–Lemeshow test was not significant we could not reject the null hypothesis that observed and expected values are the same, so we were lead to conclude that the model fits the data well.

The test has several limitations[22]. The test can be very sensitive to small fit discrepancies observed in very large samples, but it was not our case. Also, the results of the test depend on the number of groups specified (five in the example) as well as the distribution of the linear predictor values within this group. Therefore, we tried to overcome this problem by repeating the test using 4, 8 and 10 categories, and the Hosmer–Lemeshow test was never significant.

3.5 Nomogram

Nomograms are widely used for cancer prognosis, primarily because of their ability to reduce statistical predictive models into a single numerical estimate of the probability of an event, such as death or recurrence, that is tailored to the profile of an individual patient. We transferred the use of the nomogram to a screening setting because, to our opinion, it might provide practical and useful information to the general practitioner and gastroenterologist/endoscopist in order to decide whether a patient should undergo such an invasive exam as the colonoscopy or could be submitted to other less invasive exams, such as sigmoidoscopy or rectoscopy.

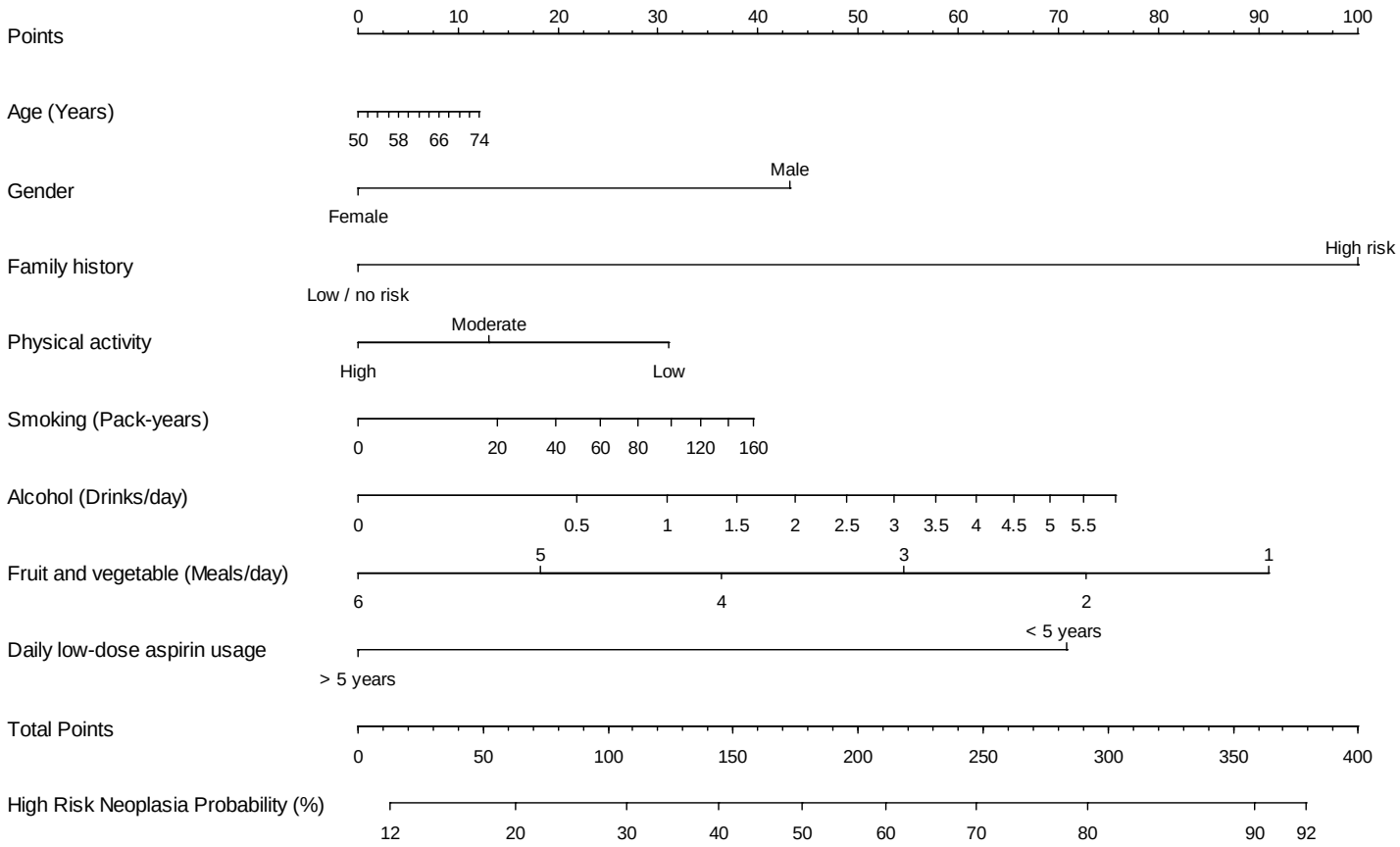
Nomograms may convey the results of a variety of statistical models. In our case, the intention was to predict a binary outcome, i.e. the presence/absence of high-risk neoplasia at colonoscopy, by using gender, physical activity, diet, smoking (pack years), alcohol consumption, use of low-dose Aspirin as independent variables.

A guide on how to build and interpret a nomogram can be found in the 2008 article by Alexia Iasonos et al.[23] and the R (http://cran.r-project.org/) software recently provided the function called *nomogram* in the *rms* package,

which easily conveys the results of any statistical model to a graphical representation.

The usefulness of a nomogram is that it maps the predicted probabilities into points on a scale from 0 to 100 in a user-friendly graphical interface. The total points accumulated by the various covariates correspond to the predicted probability of event for a patient (Figure 3.5).

Figure 3.5. The nomogram



To explain the use of it, we can say that a patient who smokes 1 pack/day for 20 years acquires around 15 points, whereas a patient who has never smoked got 0 points. Males acquire 10 points, while women acquire 0 points. And so on. By summing the *points* of all the characteristics, one gets the individual *total points*, which can be converted to the predicted probability of finding a high-risk neoplasia for a patient.

For example, a 70 year-old woman who has never smoked or drunk, often practices sports, takes low-dose Aspirin and eats 4 meals of /fruits vegetables every day, had a total of 45 points, corresponding to a predicted probability of high-risk neoplasia of 15%. This woman resulted FOBT positive probably because of some acute and not serious intestinal issue, such as hemorrhoids. On the contrary, a 50-year man who drinks and smokes, eats a few vegetables and never practices sports obtains a large total of “risk points” and should have probably begun his screening program earlier than 50 years old (Figure 3.5).

4. Risk factors at the second colonoscopy

After we demonstrated that lifestyle factors, gender, family history and low-dose Aspirin have an impact on the probability of finding a high-risk neoplasia at the first screening colonoscopy, we wanted to evaluate whether these factors have an impact on the probability of finding a high-risk neoplasia at the second screening colonoscopy. We used the linear predictor calculated in Chapter 3.4 in order to evaluate the association of the endoscopic finding with an individual risk score, rather than with all the single variables.

If a statistically significant association between the individual risk score and the outcome of the second screening colonoscopy is demonstrated, additional important conclusions will be drawn. The risk score should in fact possibly be considered when deciding how much time should pass from the first screening colonoscopy to the second control colonoscopy, basing future indications on the outcome of the primary colonoscopy as well as on the patients' characteristics.

4.1 Doctor's care scheme

Only patients diagnosed with adenoma at the first colonoscopy were included in the following analyses, because individuals with no adenoma and patients with invasive neoplasia were automatically excluded from the following colonoscopy screening process. The first group of individuals should have repeated FBOT after 5 years and the second group underwent a radical surgery and eventually an adjuvant therapy followed by a tight follow-up.

We focused on the follow-up of the patients and especially on the effect of lifestyle and other patients' characteristics on all the possible neoplastic events between the first screening colonoscopy and the second control colonoscopy. Since many of the included patients were followed in time by IEO clinicians, according to a precise schedule based on the severity of clinical findings, we have the opportunity to study the evolution of the patients' conditions. There is no general consensus on the timing of follow-up colonoscopies. This is what IEO clinicians recommend to screened patients, in accordance to the findings at colonoscopy.

- Normal mucosae or non-oncological alteration or non-adenomatous polyp: repeat FOBT at 5 years
- 1 or 2 adenomas: repeat colonoscopy at 5 years
- 3 or 4 adenomas: repeat colonoscopy at 3 years
- 5 or more adenomas: repeat colonoscopy at 1 year
- Invasive tumour: in general, surgery plus visits every 6 months after treatment

Table 4.1. Years from the 1st to the 2nd colonoscopy by severity of the 1st colonoscopy finding in 484 patients diagnosed with adenoma

		Years form the first to the second colonoscopy			
		1 year	**3 years**	**5 years**	**Total**
First colonoscopy	**1-2 low-risk adenomas**	1 (2.6)	7 (18.4)	30 (78.9)	38
	1-2 high-risk adenomas	10 (9.2)	97 (89.0)	2 (1.8)	109
	3-4 adenomas	30 (30.3)	65 (65.7)	4 (4.0)	99
	> 4 adenomas	75 (88.2)	10 (11.8)	0 (0.0)	85

Row percentages are reported in parentheses

Unfortunately, 153 patients did not come back for a second colonoscopy, making the number of patients analyzed in this phase 331. The

one depicted in Table 4.1 is a typical example of *Doctor's care scheme* of examinations[24]. The second colonoscopy date was fixed in advanced at the time of first colonoscopy, based on the outcome of the first colonoscopy. Thirty patients out of 38 (78.9%) with 1 or 2 low-risk adenomas came back after 5 years; 97 out of 109 (89.0%) with 1 or 2 high-risk adenomas came back after 3 years; 75 out of 85 (88.2%) with 4 or more adenomas came back after 1 year.

The *Doctor's care scheme* is highly relevant for many clinical studies because it allows the doctor monitoring the patient's progress to choose the next examination time for that patient depending on the state the patient is in at the current examination. In particular, patients with more advanced disease could be monitored more closely than those in whom disease was less advanced[24]. With regards to this, what we observed in our data is reported in Table 4.2.

Table 4.2. Outcome of the 2nd colonoscopy by severity of the 1st colonoscopy finding in 331 patients who had two colonoscopies

		2nd colonoscopy			
		No adenoma	**Low-risk adenoma**	**High-risk neoplasia**	**Total**
1st colonoscopy	**1-2 low-risk adenomas**	28 (73.7)	6 (15.8)	4 (10.5)	38
	1-2 high-risk adenomas	53 (48.6)	40 (36.7)	16 (14.7)	109
	3-4 adenomas	45 (45.5)	34 (34.3)	20 (20.2)	99
	> 4 adenomas	17 (20.0)	34 (40.0)	34 (40.0)	85

Row percentages are reported in parentheses

Despite the Doctor's care scheme, with differential visit times according to the severity of the first outcome, the outcome of the second colonoscopy was highly associated with the outcome of the first colonoscopy. The probability of finding a high-risk neoplasia at the second colonoscopy increased with the increasing severity of the outcome of the first. This was reasonable because a damaged mucosa remains damaged even after the removal all the polyps during a previous colonoscopy, hence a highly damaged mucosa tends to form new polyps more often than a less damaged mucosa.

If we want to evaluate the effect of lifestyle on the outcome of the second colonoscopy, can we apply simple regression models by adjusting for the outcome of the first colonoscopy? In other words: is the doctor's scheme a noninformative scheme?

In our case, information is incomplete in the sense that it is known only that an individual has been in certain disease states at several time points. Also, examination schemes are highly dependent on the outcome of the previous colonoscopy (see Table 5). The question is: can we apply a simple multivariable logistic model predicting the outcome of the second colonoscopy by using the finding of the primary colonoscopy plus the patients' characteristics as covariates? Grüger et al.[24], as I will show hereafter, demonstrated that it is possible to do so.

In order to interpolate models to longitudinal data with observation arbitrary visit times, one should consider the reason for which observations have been made in the time data. Possible schemes of observation are:

- **Fixed**: each patient is observed at fixed times, which are specified in advance;

- **Random**: observation times vary randomly, regardless of the current state of the

disease;

- **At the discretion of the physician** or **Doctor's care**: the observations for the sickest patients are more frequent. the next observation time is chosen on the basis of the current status of the disease;

- **Auto-selection of the patient**: a patient decides to pay a visit to the doctor because he feels bad.

Grüger and al. have discussed the conditions under which the observation times are informative. When considering a multi-state, ignoring the information contained in the observation time points could lead to a biased inference, because the times of observation should also be considered as random variable and modeled along with the observed process X(t). The ideal situation would be one in which the joint likelihood of time points and process is found to be proportional to the likelihood obtained in the case of time observation established a priori. In this way the parameters of the process can be estimated independently of the parameters of the sampling scheme. In

particular, the authors show that for the fixed, random and Doctor's care schemes, observation times can be considered as non-informative, while, on the other hand, the self-selection of the patient leads to informative observation times.

Suppose the disease process X(t) for a particular patient is observed at a finite number of fixed examination times $t_0 < t_1 < \ldots < t_m$ to be in states s_0, s_1, ..., s_m. The likelihood is then given by

$$L_0 = \Pr(X(t_0) = s_0 ,\ldots, X(t_m) = s_m)$$

This is the likelihood that inferences are usually based on. However, in practical applications, examination times are seldom fixed in advance, but are subject to random fluctuations. In fact, not only are the examination times T_0, T_1 , . . . , T_m, random, but also their number *M* is a random variable. So one should instead consider the likelihood:

$$L_0 = \Pr(X(t_0) = s_0 ,\ldots, X(t_m) = s_m; T_0 = t_0,\ldots, T_m = t_m; M = m)$$

Our aim is to make inferences about the probability that a patient will be in a particular disease state at time t, regardless of whether an examination is performed at this and past times or not. So even if examination times are random, L_0 is the likelihood we would like to analyze, because it contains the relevant transition probabilities. We therefore introduce the following definition.

Definition An examination scheme ($T_0 = t_0,\ldots, T_m = t_m; M = m$) is called non-informative for the disease process X, if the full likelihood on the event { $T_0, \ldots, T_m; M = m$ } is proportional to the likelihood obtained, if the number of examinations and their times were fixed in advanced, i.e.,

$$L = const * L_0$$

where the constant might depend on { $T_0, \ldots, T_m; M = m$ }, but not on X.

A straightforward application of the definition of conditional probabilities yields a factorization of the full likelihood into

$$L = \Pr(X(t_0) = s_0 ,\ldots, X(t_m) = s_m \mid T_0 = t_0,\ldots, T_m = t_m; M = m) \; x$$

$$\Pr(T_0 = t_0,\ldots, T_m = t_m; M = m)$$

Thus, any examination scheme that is stochastically independent of the process under observation is a noninformative examination scheme, because then the condition in the first factor of (3) can be ignored and the second factor is a constant with respect to the parameters of L_0.

Still this is not satisfactory, however, because often the independence assumption will be violated. […] In the "doctor's care" examination scheme the next examination time is chosen on the basis of the current observed disease state. For patients in the critical stage with an increased risk of dying, this time will be chosen in the very near future, whereas for patients in the stable stage, time intervals between successive examinations will be longer.

We can cope with this situation by factoring the full likelihood in a dynamic fashion, which reflects the accumulation of information about the disease process in time. To this end we define the history $H_0 = \{T_0 = t_0, X(t_0) = s_0\}$ and for j=1,…,m,

$$H_j = \{T_0 = t_0, X(t_0), \ldots, T_j = t_j, X(tj)\};$$

$$H_{j-} = \{T_0 = t_0, X(t_0), \ldots, T_j = t_j\}.$$

H_j contains all the information about the disease process up to and including the *jth* examination, whereas H_{j-} includes only the time but not the result of the jth examination. Then by successively conditioning on the past we get

$$L = \Pr(H_m) = \Pr(H_m \mid H_{m-1}) \times \Pr(H_{m-} \mid H_{m-1}) \times \Pr(H_{m-1})$$

$$= \Pr(H_0) \prod_{j=1}^{m} \mathbf{Pr}(\boldsymbol{H_j} \mid \boldsymbol{H_{j-}}) \times \prod_{j=1}^{m} \mathbf{Pr}(\boldsymbol{H_{j-}} \mid \boldsymbol{H_{j-1}})$$

$$= \Pr(H_0) \prod_{j=1}^{m} \mathbf{Pr}(\boldsymbol{X(t_j)} = \boldsymbol{s_j} \mid \boldsymbol{Tj} = \boldsymbol{t_j}, \boldsymbol{H_{j-1}}) \prod_{j=1}^{m} \mathbf{Pr}(\boldsymbol{T_j} = \boldsymbol{t_j} \mid \boldsymbol{H_{j-1}})$$

From this we can derive the following conditions for the examination scheme to be noninformative in the sense of the above definition:

1. $\Pr(X(t_j) = s_j \mid T_j = t_j, H_{j-1}) = \Pr(X(t_j) = s_j \mid X(t_0) = s_0, \ldots, X(t_{j-1}) = s_{j-1})$
2. The conditional distribution of the *jth* examination time T_j, i.e., $\Pr(T_j = t_j \mid H_{j-1})$ is functionally independent of parameters governing the transition intensities of X.

The first of these two conditions is the important one, since it guarantees that what we can estimate from the data [i.e., $\Pr(X(t_j) = s_j \text{ I } T_j = t_j, H_{j-1})$, the probability of being in state s_j, given that examinations take place at

$t_0,\ldots, t_j$] is identical to what we are interested in [i.e., Pr($X(t_j) = s_j$ I $X(t_0) = s_0$,. . . , $X(t_{j-1}) = s_{j-1}$), the probability of being in state s_j, irrespective of whether an examination has taken place or not]. So past examinations should not exert any effect on the future behavior of the process. However, examination times may be based on all information available up to the last examination, i.e., the times of examinations and the disease states observed (quotes from Grüger[24]).

Therefore, having demonstrated that the *Doctor's care scheme* is noninformative, as it is not dependent on the status of the patient, we can use a standard logistic regression to model the outcome of the second colonoscopy, adjusting for the outcome of the primary colonoscopy as well as for the patients' characteristics.

Table 4.3. Multivariable logistic regression analysis modeling the outcome of second colonoscopy

		High-risk neoplasia **OR (95% C.I.)**
Linear predictor	One unit increase	1.54 (1.02-2.42)
Outcome of first colonoscopy	1-2 low-risk adenomas *vs* > 4 adenomas	0.22 (0.05-1.07)
	1-2 high-risk adenomas *vs* > 4 adenomas	0.27 (0.11-0.71)
	3-4 adenomas *vs* > 4 adenomas	0.43 (0.20-0.94)
Time from 1st to 2nd colonoscopy	One year increase	0.85 (0.61-1.18)

As expected, the severity of the outcome of the first colonoscopy resulted statistically significantly associated with the outcome of the second colonoscopy: patients with 1-2 low-risk adenomas at first colonoscopy had a much more lower risk of high-risk neoplasia at second colonoscopy compared to patients with more than 4 adenomas (OR=0.22; Table 4.3). Moreover, patients with 1-2 high-risk adenomas or 3-4 adenomas at first colonoscopy still had a much more lower risk of high-risk neoplasia at second colonoscopy compared to patients with 4 or more adenomas (OR=0.27 and 0.43, respectively).

After adjusting for the severity of the outcome of the first colonoscopy and for the time from first to second colonoscopy, we obtained a statistically significant OR associated with the linear predictor. For each unit of increase, the risk of finding a high-risk neoplasia at the second colonoscopy increased by 54%. The poorer the lifestyle, the higher the probability of finding a high-risk neoplasia at the second round, irrespective of the outcome of the first round.

These findings represented the most important in this second phase of our analysis, because we provided again – as we did before in the "First colonoscopy" chapter – strong evidence supporting the role in CRC carcinogenesis of physical activity, diet, smoke and alcohol habits, and medication use. Therefore, all these factors should be considered when deciding how much time should pass from the first screening colonoscopy to the second control colonoscopy, basing the future indications on the outcome of the primary colonoscopy as well as on the patients' characteristics.

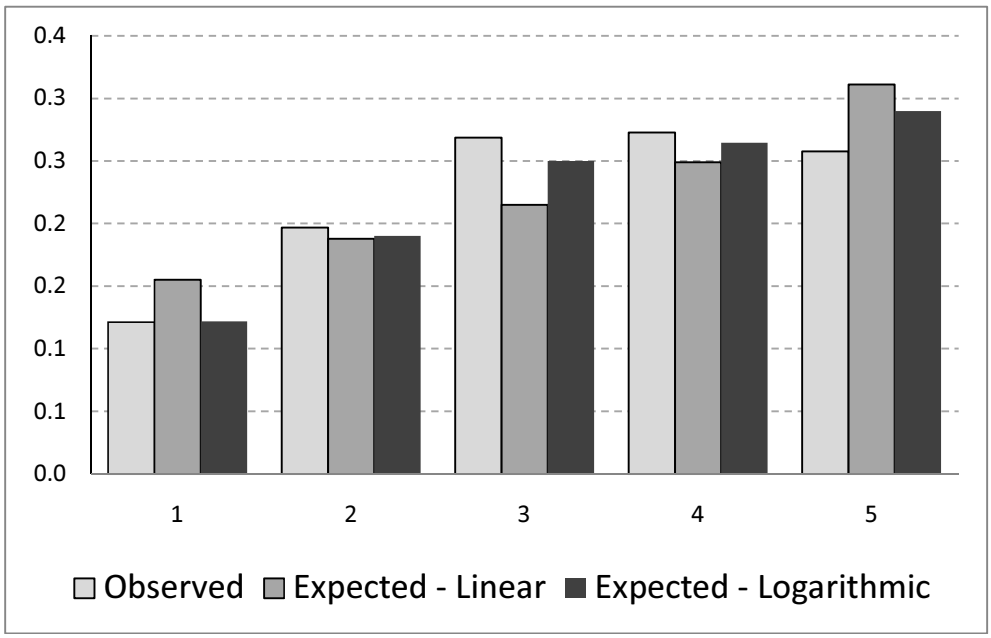
In a sensitivity analysis, we used the best FP1 (logarithmic transformation of the linear predictor, Table 4.4) instead of the simple linear function of the linear predictor (Table 4.3) and we obtained:

Table 4.4. Multivariable logistic regression analysis modeling the outcome of second colonoscopy

		High-risk neoplasia **OR (95% C.I.)**
Log(Linear predictor)	One unit increase	1.98 (1.10-3.56)
Outcome of first colonoscopy	1-2 low-risk adenomas *vs* > 4 adenomas	0.21 (0.04-1.03)
	1-2 high-risk adenomas *vs* > 4 adenomas	0.27 (0.10-0.71)
	3-4 adenomas *vs* > 4 adenomas	0.43 (0.20-0.95)
Time from 1st to 2nd colonoscopy	One year increase	0.87 (0.62-1.21)

After adjusting for the severity of the outcome of the first colonoscopy and for the time from first to second colonoscopy, we obtained a highly statistically significant OR associated with the log(linear predictor). For each unit of increase, the risk of finding a high-risk neoplasia at the second colonoscopy doubles. Again, the more your lifestyle is bad, the higher the probability of finding a high-risk neoplasia at the second round, irrespective of the outcome of the first round.

Figure 4.1. Probability of high-risk neoplasia at second screening colonoscopy according to quintiles of the linear predictor: observed *versus* expected



- AIC using linear predictor as covariate: 349.9. Hosmer–Lemeshow test P-value=0.42
- AIC using the logarithmic transformation of the linear predictor as covariate (Best FP1): 345.9. Hosmer–Lemeshow test P-value=0.92

Since Hosmer–Lemeshow test was not significant we could not reject the null hypothesis that observed and expected values are the same, so we concluded that both models fitted the data well. Nevertheless, by comparing the two AICs, we could conclude that the model using the logarithmic transformation was better than the one using the linear predictor as covariate.

4.2 Multistate Markov Model

The previous reported statistical analysis relies on simple two-state models, where one single event (the high-risk neoplasia) is taken as the outcome of interest. On the other hand, more than one endpoint can be defined in our case, such as no adenoma, low-risk adenoma and high-risk neoplasia. If one wants to take into account the different types of outcome, separate analyses are usually carried out for each of the endpoints and particular subgroups (i.e. multinomial logistic regression and competing risk survival analysis). These analyses are not completely satisfactory, since they fail to highlight the relations between different types of outcomes.

Recently, methods that simultaneously model the disease process over time have been developed, like the multi-state models[25]. In particular, in recent times, some interesting applications of Multistate Markov Models to screening programs have been developed[26]. In our study, such models allowed us to evaluate the effect of lifestyle and other factors (summarized in the linear predictor) on each transition, and make some final conclusions 1) on the age at which CRC screening should begin, either by lowering the age in people leading an unhealthy lifestyle or increasing the age in people leading a

healthy lifestyle and 2) deciding whether the second control colonoscopy should be anticipated or delayed according to patients' characteristics.

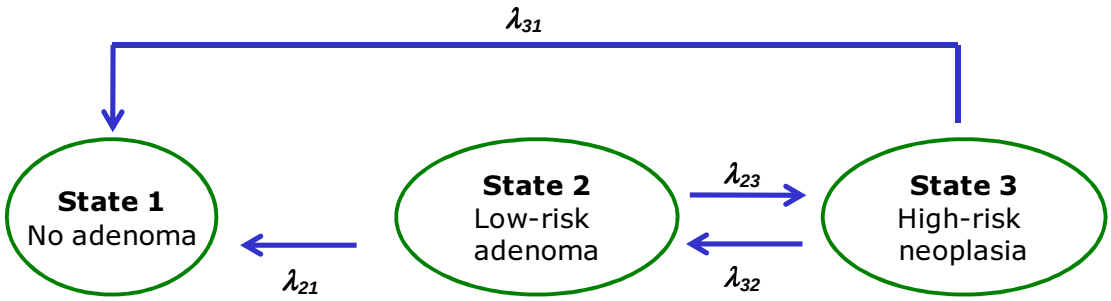
Despite such models may require specialized and quite complicated analytical tools, they give a better insight on the disease progression mechanisms and on the evaluation of the influence of prognostic factors on the transition rates from one state of the disease to another[27-29].

A multi-state model is defined as a model that describes a stochastic process{X(t), $\boldsymbol{t} \in \boldsymbol{T}$}, where for stochastic process we intend a family of random variables X indexed by t in T. Usually the parameter t is the time and the set T the temporal space. The sample space S of X (t) refers to the state space, with elements of S states. In the context of stochastic processes, the space S can be either discrete, consisting of a finite or denumerable infinity of states that the random variable X can assume, or consisting of continuous and non-countable infinity of states. Similarly T can be discrete or continuous. A multi-state model defines stochastic processes with discrete and finite sample space S and T continuous space-time.

In our case we deal with a 3-state process: the first is the “No adenoma” state, the second is the “Low-risk adenoma” and the third one is the “High-risk adenoma”. The time between the first and the second colonoscopy is kept continuous.

The structure of the states specifies which transitions from state to state are possible, and it can be represented graphically. The complete statistical model is defined by the stages structure matrix and the rule that governs the process.

Figure 4.2. Complete statistical model



None of the patients started from State 1, as we selected only patients having at least one adenoma at baseline. For each patient, the disease stage at time t is a variable X(t) which assumes values in {1,2,3}, having 3 stages. Stages structure matrix specifies the stages and the possible transitions from stage to stage.

Table 4.5. Outcome of the 2nd colonoscopy by severity of the 1st colonoscopy finding

		2nd colonoscopy			
		No adenoma	**Low-risk adenoma**	**High-risk adenoma**	**Total**
1st colonoscopy	**No adenoma**	-	-	-	
	Low-risk adenoma	28 (73.7)	6 (15.8)	4 (10.5)	38
	High-risk adenoma	115 (39.2)	108 (36.9)	70 (23.9)	293

Row percentages are reported in parentheses

Definition and formulations of general multi-state models can be found in Hougaard[28], Commenges[29] and Andersen and Keiding[30]. As we have seen, a multi-state process is a stochastic process in continuous time *X(t)* which can

take a finite number of states in the set S = {1, 2, ..., K}. For a given n and $t_0 < t_1 < ... < t_n$, the set of observed values $\{X(t_0), X(t_1), ..., X(t_n)\}$ of $X(t)$ at times $\{t_0, t_1, ..., t_n\}$ is called path or history of the process and is indicated by Ψ_t. The history of the process is continuous on the right, such that $X(t^+) = X(t)$.

The law which governs the multi-state process can be given in terms of both matrix of transition probabilities $P(s,t)$ with generic element:

$$\boldsymbol{p_{hj}(s,t) = \Pr(X(t) = j \mid X(s) = h, \Psi_s)}$$

for $h, j \in S$, $s, t \in T$, $s<t$, or in terms of transition intensity matrix $\boldsymbol{\Lambda}(t)$, whose generic element is the derivative:

$$\boldsymbol{\lambda_{hj}(t, \Psi_t) = \lim_{\Delta t \to 0} \frac{p_{hj}(t, t + \Delta t)}{\Delta t} \quad for\ h \neq j}$$

To guarantee that the sum of transition probability from one specific state to any other state (including the same starting state) is one, we constrain the row sum in the transition matrix $\boldsymbol{\Lambda}(t)$ to be equal to zero, i.e. that $\lambda_{hh}(t) = -\sum_{j \neq h} \lambda_{hj}$

The intensity of transition λ_{hj} can be interpreted as the instantaneous rate of change from the state h to the state j at time t, and $\lambda_h(t) = -\lambda_{hh}(t)$ as the rate of exit from the state *h* at time *t*.

Table 4.6. Transition intensity matrix (Δt=1 year)

	State 1	State 2	State 3
State 1	0	0	0
State 2	0.215	- 0.742	0.527
State 3	0.254	1.043	- 1.297

-2(log-likelihood): 677.67

Multi-state models and Markov models are not equivalent, but both share the concept of state. In short, the Markovian assumption implies that the future evolution of a condition depends only on the current state: in other words, all the information on the previous history of the disease process is contained in the state at time t.

Markovian assumption: $\boldsymbol{\lambda_{hj}(t, \Psi_{t-}) = \lambda_{hj}(t)}$

This assumption defines a non-homogeneous Markov model, because the intensity of transition may vary over time.

In our analysis we will use **homogeneous Markov Models**. it is assumed that the intensity of transition is not time-dependent:

$$\boldsymbol{p_{hj}(s,t) = \mathbf{Pr}[\mathbf{X(t) = j \mid X(s) = h}] = \mathbf{Pr}[\mathbf{X(t-s) = j \mid X(0) = h}]}$$

$$\boldsymbol{\leftrightarrow \quad \lambda_{hj}(t) = \lambda_{hj}}$$

If it is assumed that the heterogeneity could partly be explained by a vector of explanatory variables $\mathbf{Z}_i$ that characterizes the subject, one can write:

$$\boldsymbol{\lambda^{i}_{hj}(t) = \lambda^{i}_{hj}(t, Z_i)}$$

In this case, the subjects all share the function $\boldsymbol{\lambda_{hj}}(\cdot,\cdot)$, and the population can be defined homogeneous conditionally on $\mathbf{Z}_i$, i = 1, ..., M. An assumption that greatly simplifies the process of estimating the parameters of the model is that the values of the intensities conditioned to z_i are proportional to a basal intensity:

$$\boldsymbol{\lambda_{hj}(t, Z_i) = \lambda_{hj\,0}(t)\, f(Z_i)}$$

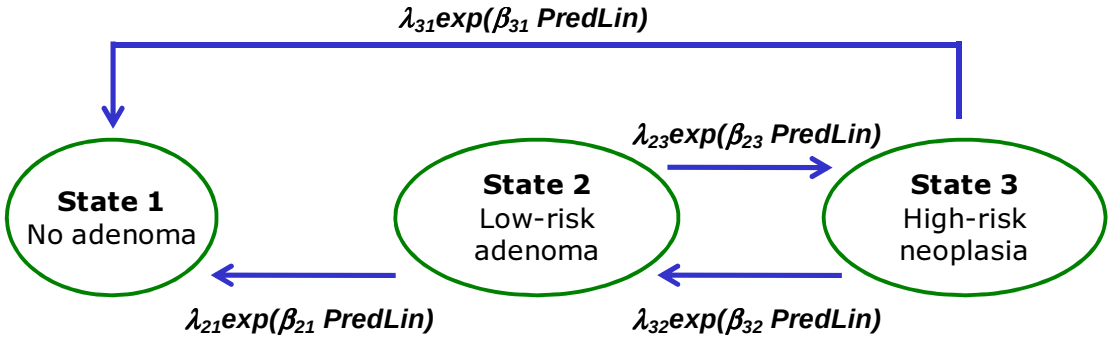
For analytical convenience and for an immediate interpretation of results in terms of usual relationship between risks, the model is reparameterized introducing covariates as a factor proportional to the intensity of the basic transition $\boldsymbol{\lambda}$, so that we obtain a log-linear model for the intensities of transition. The regression for the element *(h, j)* of the transition matrix $\boldsymbol{\Lambda}$ is thus indicated:

$$\boldsymbol{\lambda_{hj}(Z) = \lambda_{hj}\, exp\left(\beta'_{hj} Z\right)}$$

with $\boldsymbol{\beta'_{hj}}$ vector of regression coefficients associated with the vector of covariates **z** for transitions between states *h* and *j*. It can then redefine the matrix of intensity transition in terms of the parameters $\boldsymbol{\lambda}$ **and** $\boldsymbol{\beta}$, and indicate it as $\boldsymbol{\Lambda}$**(z)**.

One can test whether the covariate **z** affects the transition intensities by comparing the likelihoods of the restricted model and the unrestricted model, through the likelihood ratio test. With regards to the practical implementation, we used the *msm* package of functions for multi-state modelling using the R statistical software[31].

Figure 4.3. Possible transitions with *linear predictor* as covariate



PredLin= linear predictor as calculated in chapter 3.4

Table 4.7. Log-linear effects of linear predictor on transitions

	State 1	State 2	State 3
State 1	-	0	0
State 2	- 0.121	-	0.261
State 3	- 0.380	- 0.545	-

-2(log-likelihood): 662.63

The one reported in Table 4.7 was a very interesting result. The estimated parameters for the effects of linear predictor on ***improvement*** transitions (State 2 → State 1 or State 3 → State 2 or State 3 → State 1) were negative,

which means that as the linear predictor increased, the probability of getting better decreased. The interpretation is that, in other words, the worse the lifestyle, the lower the probability for the intestinal mucosa to heal.

On the other hand, the estimated parameter for the effect of linear predictor on ***aggravation*** transition (State 2 → State 3) was positive, which means that as the linear predictor increased, the probability of getting worse increased. In other words, the worse the lifestyle, the higher the probability for the intestinal mucosa to worsen.

Moreover, we tested whether the introduction of the lifestyle as covariate added significant information to the simple transition model. The difference between *-2 (log-likelihood)* of the null model and the *-2 (log-likelihood)* of the model with lifestyle was given by 677.67-662.63=15.04 which distributes as a Chi-square with 4 degrees of freedom. Since the table value of a Chi-square with 4 degrees and α=0.05 is 9.49, we rejected the null hypothesis of no impact of lifestyle on the transition model.

We also estimated the effects of the linear predictor on transitions by restraining the effects of lifestyle on contiguous transitions to be the same.

Table 4.8. Log-linear effects of linear predictor on transitions with restraints

	State 1	State 2	State 3
State 1	-	0	0
State 2	- 0.320	-	0.523
State 3	-0.044	- 0.320	-

-2(log-likelihood): 664.99

Since we obtained similar results, we chose the simpler model without restraints (Table 13).

Looking at the results of this analysis, we can say that the combination of lifestyle factors plus gender, family history and low-dose Aspirin use was significantly and independently associated not only with the probability of finding a high-risk neoplasia at the second colonoscopy, but also with the transitions from different disease states.

5. Conclusions

All these results allow us to draw two main important conclusions:

1) Besides gender and family history, which are well-known features associated with screening-detected colorectal neoplasia, also lifestyle factors – such as physical activity, smoking habits, alcohol consumption and diet – are associated with the outcome of the first screening colonoscopy. Also, long-term low-dose Aspirin use was an additional significant factor in predicting the outcome. All these factors may soon change the clinical practice about the age at which CRC screening should begin, either by lowering the age in individuals with a poor lifestyle or increasing the age in individuals with a healthy lifestyle.

2) Lifestyle factors, with gender, family history and use of low-dose Aspirin, should be taken in consideration when deciding how much time should pass from the first screening colonoscopy to the second control colonoscopy, basing future indications not only on the outcome of the primary colonoscopy but also on the patients' characteristics.

Therefore, our findings increased the evidence that lifestyle is substantially associated with the carcinogenesis of the colorectal cancer. Having said that, two types of interventions are now possible, the first referring to the primary prevention (i.e. modification of lifestyle), and the second referring to the secondary prevention (i.e. modification of screening policies). Firstly, avoidance of smoking and heavy alcohol use, high consumption of fruits and vegetables, the maintenance of a reasonable level of physical activity and the use of low-dose Aspirin can each have a beneficial impact on the risk of colorectal cancer (primary prevention). Secondly, lifestyle should be considered in the planning of population colorectal cancer screenings (secondary prevention), because the identification of different risk groups can lead to more tailored screening policies and, accordingly, to more efficient and cost-effective interventions.

6. References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011 Mar-Apr;61(2):69-90.
2. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell 1990; 61:759–767.
3. Hill MJ, Morson BC, Bussey HJ. Aetiology of adenoma—carcinoma sequence in large bowel. Lancet 1978;1:245–247.
4. Botteri E, Iodice S, Raimondi S, Maisonneuve P, Lowenfels AB. Cigarette smoking and adenomatous polyps: a meta-analysis. Gastroenterology. 2008; 134(2):388-395.
5. Botteri E, Iodice S, Bagnardi V, Raimondi S, Lowenfels AB, Maisonneuve P. Smoking and colorectal cancer: a meta-analysis. JAMA. 2008 Dec 17; 300(23):2765-78.
6. Maisonneuve P, Botteri E, Lowenfels AB. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps. Gastroenterology. 2008;135(2):710
7. Nguyen SP, Bent S, Chen YH, Terdiman JP. Gender as a risk factor for advanced neoplasia and colorectal cancer: a systematic review and meta-analysis. Clin Gastroenterol Hepatol. 2009 Jun;7(6):676-81.e1-3.

8. Ning Y, Wang L, Giovannucci EL. A quantitative analysis of body mass index and colorectal cancer: findings from 56 observational studies. Obes Rev. 2010 Jan;11(1):19-30. Epub 2009 Jun 16.
9. Qureshi N, Carroll JC, Wilson B, Santaguida P, Allanson J, Brouwers M, Raina P. The current state of cancer family history collection tools in primary care: a systematic review. Genet Med. 2009 Jul;11(7):495-506.
10. Chan AT, Giovannucci EL. Primary prevention of colorectal cancer. Gastroenterology. 2010 Jun;138(6):2029-2043.e10.
11. Seitz HK, Maurer B, Stickel F. Alcohol consumption and cancer of the gastrointestinal tract. Dig Dis. 2005;23(3-4):297-303.
12. Magalhães B, Peleteiro B, Lunet N. Dietary patterns and colorectal cancer: systematic review and meta-analysis. Eur J Cancer Prev. 2012 Jan;21(1):15-23.
13. Rothwell PM. Aspirin in prevention of sporadic colorectal cancer: current clinical evidence and overall balance of risks and benefits. Recent Results Cancer Res. 2012;191:121-42.
14. Chan AT, Giovannucci E. Primary Prevention of Colorectal Cancer. Gastroenterology. 2010 June ; 138(6): 2029–2043.e10.
15. Winawer SJ, Zauber AG, Fletcher RH, et al. Guidelines for colonoscopy surveillance after polypectomy: a consensus update by the

US Multi-Society Task Force on colorectal cancer and the American Cancer Society. Gastroenterology 2006;130:1872–1885.

16. Hosmer DW, Lemeshow S. Applied Logistic Regression, New York: Wiley, 2000.
17. Royston P, Sauerbrei W, Becher H. Modelling continuous exposures with a 'spike' at zero: a new procedure based on fractional polynomials. Stat Med. 2010 May 20;29(11):1219-27.
18. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with Discussion). Applied Statistics 1994; 43(3):429--467.
19. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001 Aug;54(8):774-81.
20. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984;3:143–52.
21. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988 Sep;44(3):837-45.

22. Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models (Statistics for Biology and Health). 2011, Springer.
23. Iasonos A, Schrag D, Raj GV, Panageas KS. How To Build and Interpret a Nomogram for Cancer Prognosis. J Clin Oncol. 2008 Mar 10;26(8):1364-70.
24. Grüger J, Kay R, Schumacher M. The Validity of Inferences Based on Incomplete Observations in Disease State Models. Biometrics 1991 June; 47, 595-605.
25. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007 May 20;26(11):2389-430.
26. Uhry Z, Hédelin G, Colonna M, Asselain B, Arveux P, Rogel A, et al. Multi-state Markov models in cancer screening evaluation: a brief review and case study. Stat Methods Med Res. 2010 Oct;19(5):463-86.
27. Dancourt V, Quantin C, Abrahamowicz M, Binquet C, Alioum A, Faivre J. Modeling recurrence in colorectal cancer. J Clin Epidemiol. 2004; 57: 243-51
28. Hougaard P. Multi-state models: a review. Lifetime Data Anal; 1999, 5: 239-64.
29. Commenges D. Multi-state models in epidemiology. Lifetime Data Anal. 1999; 5: 315-27.

30. Andersen PK, Keiding N. Multi-state models for event history analysis. Stat Methods Med Res; 2002, 11: 91-115

31. Jackson CH. Multi-State Models for Panel Data: The msm Package for R. Journal of Statistical Software; 2011, 38:8.