UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI SCIENZE CLINICHE
E DI COMUNITÀ

Dottorato di Ricerca in Epidemiologia, Ambiente e Sanità Pubblica

Curriculum in Biostatistica ed Epidemiologia

Ciclo XXIX

Identification of Circulating Biomarkers for the Early Diagnosis of Colorectal Cancer: Methodological Aspects

settore scientifico-disciplinare MED/01

Tesi di Dottorato di **Chiara Maura Ciniselli**

Matricola: R10676-R15

Relatore: **Dr. Paolo Verderio**

Coordinatore del Corso: **Prof. Carlo La Vecchia**

A.A. 2015-2016

Alla nostra promessa

Index

RATIONAL AND AIM

The present PhD research project starts from the need of deeply investigate some methodological aspects related to the identification, validation and application in a routine clinical setting of new non-invasive biomarkers for the (early) detection of cancer.

This research takes advantage from the ongoing "Tumor microenvironment-related changes as new tools for early detection and assessment of high-risk diseases" project at the Fondazione IRCCS Istituto Nazionale dei Tumori (INT) funded by the Associazione Italiana per la Ricerca sul Cancro (AIRC; http://www.ederaproject.it). The project aims at identifying and validating predictive, diagnostic and prognostic biomarkers in different clinical settings as well as at investigating tumour microenvironment-related changes to be used as novel tools for early detection and assessment of high-risk diseases in the major solid tumour types (lung, prostate, colorectal, breast cancer and melanoma). Specifically, circulating biomarkers have been searched in plasma and/or serum to develop molecular-based signatures that are currently been validated on large prospective series in different diseases.

In this thesis I investigated the statistical-methodological issues related to the identification and validation of molecular-based signatures, as non-invasive cancer biomarkers, detected with qPCR-based platform, using the colorectal cancer as disease model (CRC-INT study).

CHAPTER 1. INTRODUCTION

1.1 COLORECTAL CANCER

1.1.1 Clinical and Epidemiological Background

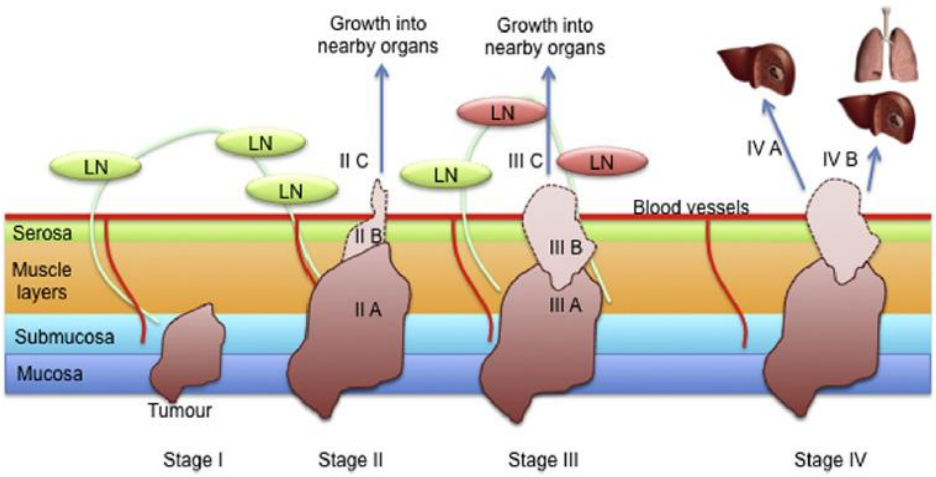
Colorectal cancer (CRC) is one of the major causes of cancer death in western countries (Jemal et al. 2011, 69-90; Mazeh et al. 2013, 281-295). Complete surgical excision of the primary tumour is the only treatment for early CRC and tumour stage at diagnosis remains the only independent predictor of survival (Bustin and Murphy 2013, 116-125; Ferlay et al. 2010, 2893-2917; Gellad and Provenzale 2010, 2177-2190). The American Joint Committee on Cancer (AJCC) identified four stages of CRC disease according to the TNM system[1]: from early cancer (stage I) to advanced metastatic (stage IV) as reported in Figure 1. Briefly, stage I-II are cancers that spread from the normal mucosa of the colon to the muscular layer (stage I) or to the serosa (stage II) without lymph node involvement, whereas stage III cancers extend through the mucosa, submucosa and muscle layers with lymph node involvement; finally stage IV cancers may grown through the wall of colon or rectum and spread through the blood and lymph nodes (Bustin and Murphy 2013, 116-125).

The lifetime incidence of CRC in the average-risk population (i.e. subjects without inflammatory bowel disease, familial adenomatous polyposis (FAP), hereditary nonpolyposis colorectal cancer (HNPCC) or positive family history of colorectal neoplasia) in North America and Western Europe is about 5%, with 90% of cases that occur after the age of 50 years (Mazeh et al. 2013, 281-295). The early detection of cancer represents the best option for reducing CRC-related mortality rate: the 5-years survival rates for CRC patients at stage I or II are 90% and 75%, respectively. These figures decrease to 65% and 5% only, for CRC patients at stage III and IV, respectively (Mazeh et al. 2013, 281-295).

[1] The TNM system is the most widely used cancer staging system. In the TNM system:

- T refers to the size and extent of the main tumor;
- N refers to the number of nearby lymph nodes that have cancer;
- M refers to whether the cancer has metastasized.

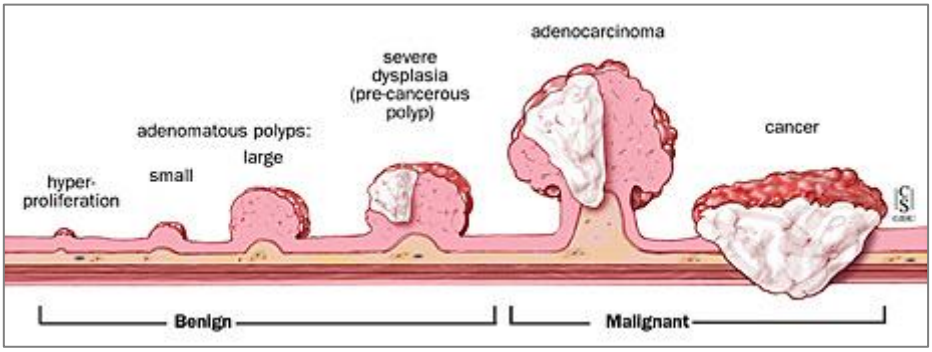
Figure 1. Colorectal cancer stages



Bustin SA and Murphy J. 2013

In most CRC patients, the progression of normal mucosa to invasive cancer requires several molecular changes with an estimated time interval ranging between 5 and 10 years. Figure 2 reports the colorectal cancer progression steps: it usually begins as benign polyps that grow from the normal mucosa and, if not surgical removed, can grow-up until become invasive cancers. Most polyps remain "*benign*" (e.g. hyperplastic polyps), whereas other, sometimes referred to pre-cancerous polyps, have a chance of becoming *"cancerous"*, if not removed. The only truly "*malignant"* polyps are those containing an invasive carcinoma inside. Adenomatous polyps tend to grow slowly over a decade, but the risk of their transformation into cancers increases as function of the size of the polyps and time they remain in the colon. In fact, less than 10% of all adenomas become cancerous polyps, however, more than 95% of colorectal cancers originate from adenomas (http://www.hopkinscoloncancercenter.org). Carcinomas can be confined to the polyps only, but in some cases they can invade one or more layers of the intestine and may thus metastasize, spreading the cancer to other organs such as liver and lungs.

Figure 2. Colorectal cancer progression



www.hopkinscoloncancercenter.org

On the basis of the natural history of CRC progression and the long time-interval of progression from normal mucosa to invasive cancer, many efforts have been focused on the development of screening programs for CRC prevention and detection at an early stage, when cancer is most likely curable. According to these consideration and the CRC age-associated risk, the current screening guidelines recommend routine testing after the age of 50 years (Mazeh et al. 2013, 281-295; Ahmed 2014, 463-485).

1.1.2 Screening for colorectal cancer

There are different methods that can be used in screening programs and they can be roughly classified into two broad categories:

- tests that look at the colon to find any abnormalities (both polyps and cancer) and allow the removal of polyps, when found;
- tests that search signs of the cancer presence in a non-invasive way.

In the first category it can be included colonoscopy (or sigmoidoscopy), that, although invasive, is one of the preferred choice for CRC screenings (Mazeh et al. 2013, 281-295). In the second group it can be found tests, based on the search of human haemoglobin in stool (i.e. faecal occult blood test) or genetic material, i.e. multitarget stool DNA (Imperiale et al. 2014, 1287-1297), or in body fluid (molecular DNA tests in blood). The latter are less invasive, and easier to carry out, but many of them demonstrate low sensitivity values for polyps identification. Therefore, many efforts have been spent during last years, and are still ongoing, for developing acceptable non-invasive tests to be used in screening programs (Mazeh et al. 2013, 281-295; Imperiale et al. 2014, 1287-1297; Bretthauer 2011, 87-98).

Currently, colonoscopy is the most accurate test for detecting early cancerous lesions and for the removal of advanced adenomas with a 95% sensitivity and 90-95% of specificity, respectively for CRCs and advanced adenoma (Bretthauer 2011, 87-98). However, procedural competence can vary among examiners and the cecum is reached in 80–98% of procedures, with the depth of penetration depending on both the experience of the endoscopist and the adequacy of bowel preparation. On the other hand, it should be mentioned that no studies examined the effectiveness of colonoscopy as a screening modality considering CRC-related mortality as primary study endpoint, and no controlled trials addressed the question of how frequently colonoscopy should be performed (Mazeh et al. 2013, 281-295).

Finally, because of its potential limitations, such as invasiveness, elevated costs and incidence of complications, the use of the faecal occult blood test (FOBT[2]) has been proposed and adopted throughout the world for large-scale population screening programs (Mazeh et al. 2013, 281-295; Bretthauer 2011, 87-98; Park et al. 2010, 2017-2025). Three clinical trials highlighted a reduction of CRC-related mortality using the FOBT as screening test that is estimated equal to 33%, among subjects who had annual rehydrated FOBT testing in a US clinical trial, and 15% in two European studies using a bi-annual non-rehydrated FOBT testing protocol (Mazeh et al. 2013, 281-295). Nevertheless, FOBT has been criticized due to its low sensitivity to cancer (11%-64%) and large adenomas (11%-41%) and also because it is a nonspecific test for gastrointestinal bleeding and human haemoglobin (Bretthauer 2011, 87-98). FOBT can in fact produce false-positive results when meats, fruits or vegetables are ingested in large quantities or when there is a bleeding in the upper gastrointestinal tract due to aspirin and non-steroidal anti-inflammatory drugs. This implies a large number of false-positive results that generate unnecessary endoscopic investigations, increasing the costs and risks of screening programs (Park et al. 2010, 2017-2025). The new automated laboratory-based method, the quantitative immunochemical fecal occult blood test (FIT)[3], is specific for human haemoglobin and eliminate the need of diet restriction. This test

[2] The fecal occult blood test (FOBT) uses chemical indicators (reagent derived from wood resin of Guajacum trees) to detect heme in stool. Heme is the iron-containing component of the blood protein haemoglobin. The idea behind the FOBT is that blood vessels at the surface of larger colorectal polyps or cancers are fragile and easily damaged by the passage of feces. The damaged vessels usually release a small amount of blood into the feces, but only rarely there is enough bleeding to be noticeable (by eye) in the stool. This test, however, cannot determine whether the blood is from the colon or from other portions of the digestive tract (such as the stomach).

[3] FIT (or iFOBT) is essentially based on the same principles of the traditional guaiac-based FOBT, but it uses antibodies to detect human haemoglobin protein in stool. The main difference is that the iFOBT uses a different technology to

provide an accurate haemoglobin quantification in stools, allowing the selection of suitable thresholds also for follow-up colonoscopy (Bretthauer 2011, 87-98; Park et al. 2010, 2017-2025). FIT has a higher sensitivity for CRC (56%-89%) with respect to FOBT, even if its specificity for advanced adenoma remains still low (27%-56%) (Bretthauer 2011, 87-98). Due to its low sensitivity in detecting precursor lesions, the potential for cancer prevention by detection and removal of adenoma is limited, achieving only modest reductions in cancer incidence among screened individuals. Another important issue related to FOBT/FIT is the limited sensitivity of 1-time testing (ranges from 11% to 64% for CRC), lower than the reported sensitivity of 1-time testing with colonoscopy (around 95% for CRC). For these reasons, FOBT/FIT must be repeated biennially with a consequent loss of adherence and a decreased effectiveness of the screening programs. Additionally little is known about the cost of ensuring adherence with annual testing and the cost related to negative colonoscopy after a false-positive faecal test.

According to the aforementioned characteristics of CRC progression and the current limits of the FOBT/FIT tests, many working groups are looking for diagnostic biomarkers, preferably present in patient's fluids, that allows a non-invasive patient's assessment. A promising technology in this field is represented by the identification of specific genomic or proteomic patterns able to discriminate patients with CRC or advanced adenoma from those without endoscopic lesions (Bustin and Murphy 2013, 116-125). With such strategy, only subjects with an increased evidence of "*malignant"* lesions may undergo further investigations, increasing the specificity and efficiency of the screening programs.

1.2 BIOMARKERS

Biomarkers are defined as biological substances, characteristics, or images that provide an indication of the biological state of an organism. The "*biomarker"* term can be referred to physiological indicators (i.e. blood pressure) and molecular markers (i.e. expression signatures) as well as radiological biomarkers (i.e. computer tomography or magnetic resonance imaging). Their proper use requires an understanding of their sensitivity and specificity, how to use them and in what context, and how to validate them properly. Nevertheless, no one has all the properties to

---

detect the presence of gastrointestinal bleeding, leading to a more accurate way to screen for blood in the stools than the traditional FOBT. Specifically, it reacts with part of the human haemoglobin protein (heme), which is found in red blood cells and it is also less likely to react to bleeding from parts of the upper digestive tract, such as the stomach.

allow the detection of the disease, to facilitate prognosis or provide evidences of response to cancer treatments (Bustin and Murphy 2013, 116-125).

1.2.1 Biomarkers validation process

Given the great number of promising biomarkers continuously proposed in literature, the need of a standardized process for biomarker validation in oncology has become increasingly relevant in the last years (Verderio et al. 2010, 62-65; Verderio et al. 2016, 1-4).

The development of a new cancer biomarker is a process that begins with biomarker discovery, followed by a rigorous definition and evaluation of the whole process of biomarker determination (analytical validation). It finally terminates with the assessment of the impact of the biomarker on clinical practice (clinical validation).

During the analytical validation the principal aim is to evaluate the performance characteristics of the biomarker assay and its optimal analytical setting to guarantee a satisfactory level of reproducibility and accuracy. The ultimate goal of this process consists in reducing the number of promising biomarkers that fail in the clinical setting as result of a lack of robust analytical validations (Verderio et al. 2010, 62-65). Table 1 summarizes the ideal phases to be followed for the analytical validation of a promising biomarker.

Table 1. Phases of biomarker analytical validation

Phase	Description
I - Operating Procedures Setting-up	Definition of operating procedures for biomarker determination
II - Operating Procedures Standardization	Validation of the operating procedures in terms of precision and accuracy according to the standards defined
III - Internal Quality Control	Evaluation of the validated standards within laboratory
IV - External Quality Assessment	Between laboratories comparison and assessment of their accuracy

As concern the clinical validation, a well-established multi-phased approach has been proposed by Pepe MS et al, as reported in Table 2 (Pepe et al. 2001, 1054-1061). Briefly, Phase I studies are exploratory studies often based on high-throughput technologies aimed at identifying potentially useful biomarkers (biomarker discovery) and in Phase II, biomarker values in cases

(individuals with disease) and controls (individuals without disease) are directly compared. In this phase, a clinical assay based on a specimen that can be obtained non-invasively should be developed and the principal aim is the estimation of the ability of the biomarker in discriminating subjects with cancer from subjects without cancer. The last Phases are focused on verifying if the biomarker is able to detect the disease before it becomes clinical and determining the extent and characteristics of the disease detected by the test.

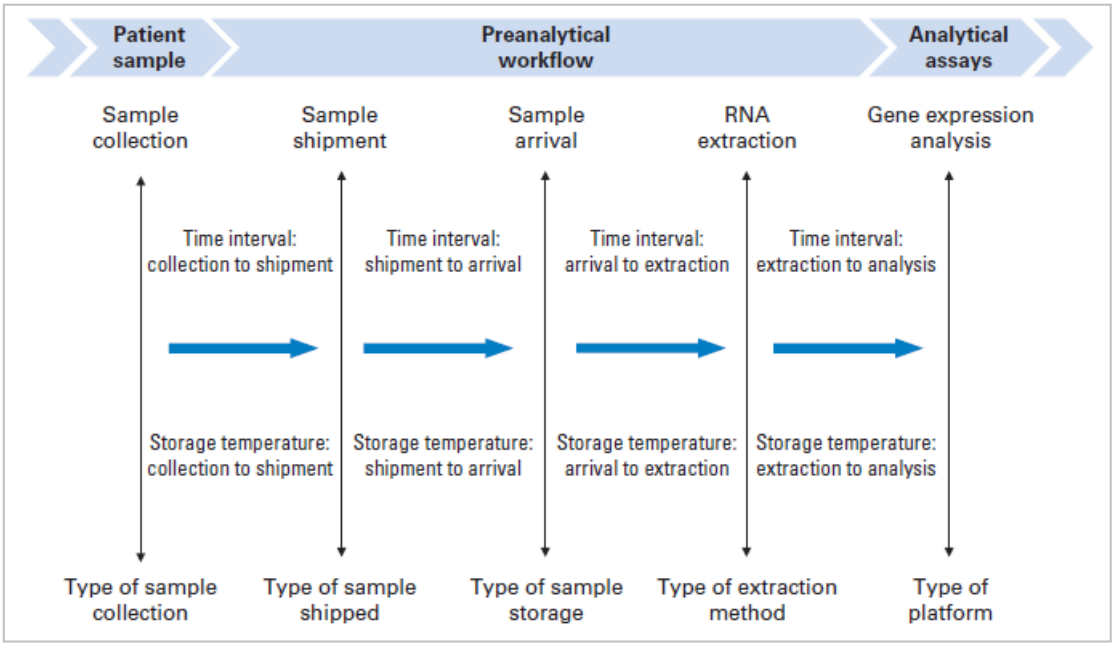
Table 2. Phases of biomarker clinical validation

Phase	Objective	Study design
I - Preclinical exploratory	Identify promising directions	Case-control
II - Clinical assay and validation	Determine if a clinical assay detects a specific disease	Case-control (population based)
III - Retrospective longitudinal	Verify if the biomarker is able to detect disease before it becomes clinical	Nested case-control in a population cohort
Iv - Prospective screening	Determine extent and characteristics of disease detected by the test	Cross-sectional cohort of people
V - Cancer control	Impact of screening on reducing the burden of disease on the population	Randomized trial (ideally)

To note that, during the analytical validation, Operating Procedures for both the pre-analytical and analytical phase of biomarker determination should be identified in order to control as much as possible factors that could affect the final results of the study. Figure 3 reports some of the most important pre-analytical factors involved in the RNA-based experiments of gene expression analysis (Verderio 2012, 1912-1915). Briefly, a first key factor could be represented by the choice of the (body fluids) biological matrices (i.e. blood, plasma or serum) and the related aspect of sample collection (i.e. time of blood collection, type of needles, type of collection tube and anti-coagulant, etc) followed by the procedure for sample processing (i.e. plasma separation). The subsequent pre-analytical step is the sample shipment, from laboratories to a centralized facility or within laboratories of the same Hospital. In this phase, times and temperatures between blood withdrawal and plasma separation, as well as those of plasma storage should be opportunely defined to minimize difference related to sample transport. The last step, before the analytical phase, is represented by the RNA extraction with the choice of the RNA isolation methods (i.e.

guanidine/phenol/chloroform-based, columns- or bead-based commercial kits). The analytical step refers to the technologies and/or assays used to detect and measure the target of interest (Verderio 2012, 1912-1915; Appierto et al. 2014, 1215-1226). Similar consideration can be done for DNA- or protein- based experiments.

Figure 3. Pre-analytical workflow from sample collection to gene expression analysis



The EU-SPIDIA project[4] (www.spidia.eu) investigated the influence of these pre-analytical factors on the quality/quantity, integrity of DNA and RNA extracted from blood or plasma highlighting that, if not properly controlled and standardized, these factors could severely influence the downstream analytical results (Malentacchi et al. 2013, 274-286; Malentacchi et al. 2014, e112293; Malentacchi et al. 2015, 205-210; Malentacchi et al. 2015, 1935-1942; Malentacchi et al. 2016, 122-128; Pazzagli et al. 2013, 20-31).

[4] SPIDIA was a 4.5 year integrated project funded by the European Commission which brings together a consortium of 16 leading academic institutions, international organisations and life sciences companies, including INT. The project was coordinated by QIAGEN GmbH and aimed to tackle the standardisation and improvement of pre-analytical procedures for in-vitro diagnostics. Activities were focused on the identification and development of evidence-based guidelines, on the creation of tools for the pre-analytical phase as well as on testing and optimisation of these tools through the development of novel assays and biomarkers..

1.2.2 Circulating microRNAs

MicroRNAs (also called miRNAs or miRs) are a class of small (18-25 nucleotides long) non-coding RNAs that act as post-transcriptional regulators of gene expression. MiRNAs have been studied intensively in the field of oncological research and many evidences suggest that altered miRNA regulation is involved in the pathogenesis of cancers, mainly by regulating the translation of oncogenes and tumor suppressors (Calin and Croce 2006, 857-866). In addition, several studies have shown that tumor-associated miRNAs are detectable in plasma and serum (Chen et al. 2008, 997-1006), suggesting that the expression profiles of circulating miRNAs could be used for diagnosing and monitoring human cancers. Presence of circulating miRNAs, for early detection of cancer, has been successfully investigated in several malignancies such as breast, lung, prostate, lymphoma, ovarian, gastric, oesophageal cancer (Cortez et al. 2011, 467-477) and also in colorectal cancer (Zanutto et al. 2014, 1001-1007; Hollis et al. 2015, 8284-8292; Verma et al. 2015, S100-6736(15)60415-9). Currently, many published studies highlight the promising role of these biomarkers as non invasive-tools for (early) detection of cancer as well as of other disease, such as cardio-vascular-disease. However, some Authors are now stressing the need of developing operative procedures (OPs) and identifying shared methods/procedures for the entire pre-analytical (i.e. sample collection, RNA extraction) and analytical phase (i.e. different experimental protocols) of miRNAs detection, in order to try to overcome the reported discrepancies about the role of specific miRNAs even within the same cancer type (Verderio et al. 2016, 1-4; Singh et al. 2016, 113-121).

1.2.2.1 Challenges in circulating microRNAs detection

As regards the pre-analytical phase of miRNAs analysis, one of the most important issue that should be taken into consideration in miRNA-based studies is represented by haemolysis, recognizable by a pink discoloration of serum or plasma due to the release of the red blood cells (RBCs) into the fluids (Kirschner et al. 2011, e24145; Pritchard et al. 2012, 492-497; Kirschner et al. 2013, 94; Yamada et al. 2014, e112481). Haemolysis commonly occurs during blood collection or sample processing and many studies showed the substantial impact of haemolysis on certain miRNAs: miR-16 and miR-456 are the most highly abundant miRNAs in RBCs and thus their levels are the most affected by haemolysis. Additional studies highlighted a similar condition also for miR-451 and miR-92a, that was also identified as a promising biomarker in several cancers (Kirschner et al. 2011, e24145). This means that altered levels of these miRNAs can reflect blood

cell based phenomena rather than the presence of cancer (Pritchard et al. 2012, 492-497; Kirschner et al. 2013, 94).

The level of haemolysis in plasma samples can be measured spectrophotometrically (Kirschner et al. 2013, 94; Yamada et al. 2014, e112481) and various haemolysis indexes, based on specific absorbance measurements, have been suggested in literature, such as absorbance peaks at 414 (Kirschner et al. 2011, e24145), haemolysis ratio (Zanutto et al. 2014, 1001-1007), HS-score (Appierto et al. 2014, 1215-1226) and haemoglobin concentration obtained by the Harboe method (MacLellan et al. 2014, 27-6890-14-27. eCollection 2014). Other pre-analytical aspects, such as the starting material (serum or plasma), the sample storage and the related miRNAs stability, as well as the RNA extraction methods are other important issues that should be considered and harmonized before the use of these biomarkers in the clinical practice [Butz, H. and Patocs, A. 2015].

As concerns the analytical phases of miRNA detection, several techniques are currently available for assessing miRNAs levels in body fluids, such as miRNA microarrays, quantitative real-time PCR (qPCR) and deep sequencing. Among these approaches, the most frequently used is qPCR-based assays in which miRNA expression data are usually provided on a continuous scale and analyzed in terms of relative quantification (Cortez et al. 2011, 467-477; Livak and Schmittgen 2001, 402-408; Deo, Carlsson, and Lindlof 2011, 795-812)

Briefly, the qPCR allows the quantification of minute amounts of nucleic acids, by using fluorescent probes that generate a signal proportional to the concentration of the amplification products. The target concentration can be determined from the fractional cycle where a threshold amount of amplified DNA/cDNA is produced. The latter is defined as threshold cycle (Ct): these Ct values are directly proportional to the amount of starting template and are the basis for the quantification of target DNA/mRNA concentration (Verderio et al. 2004, 76-79). For analyzing data from qPCR experiments two different methods are available: (i) the absolute and (ii) the relative quantification. The absolute quantification can be achieved using a standard curve, constructed by amplifying known amounts of DNA/cDNA, and using this curve as calibrator to estimate the unknown DNA/cDNA concentration using the inverse regression method (Verderio et al. 2004, 76-79). On the contrary, the relative quantification, proposed by Livak and Schmittgen, describes the changes in expression of the target gene[5] relative to a reference group (Livak and Schmittgen

[5] The method was proposed for gene-expression studies and currently applied also to miRNAs-based experiments.

2001, 402-408). This is achieved through the subtraction of the Ct values of the reference from that of target gene (i.e., normalization), allowing the comparison of the expression of target gene to each other in different samples (Livak and Schmittgen 2001, 402-408; Perkins et al. 2012, 296-2164-13-296). In addition, many Authors suggest the use of multiple reference genes for data normalization in order to reduce error and obtain a more stable relative expression (Perkins et al. 2012, 296-2164-13-296; Vandesompele et al. 2002, RESEARCH0034).

Specifically,

ΔCt_i = Ct target $gene_i$ – Ct reference $gene(s)_{i,}$ i=1,..., N with N: number of subjects [1]

$\log_2 RQ_i = -\Delta Ct_{i,}$ with RQ: relative quantity [2]

The next chapter is focused on the screening programmes ongoing in Italy with a specific focus on the colorectal cancer screening. Data reported in the following section are extracted by the 11th report of the National Centre for Screening Monitoring and by the GISCoR[6] website (http://www.osservatorionazionalescreening.it; http://www.giscor.it/).

The third chapter is focused on the statistical-methodological issues related to the identification of new molecular cancer biomarkers and in their combination into signatures (or composite scores), with details also about their validation. The fourth chapter reports the application of the investigated and developed statistical-methodological approaches to the CRC-INT study and the fifth one discusses the obtained results.

[6] GISCoR, Gruppo Italiano Screening COlonRettale

CHAPTER 2. CANCER SCREENING PROGRAMS IN ITALY

2.1 OVERVIEW

In accordance with the European Commission's 2003 Recommendation, the Italian Ministry of Health recommended the implementation of organized screening programmes for cervical, colorectal, and breast cancers. These programmes involve active invitation of the entire target population, free testing and treatment, quality assurance in all stages of the process and early outcome monitoring system. Figure 4 reports the main characteristics of the screening programs currently active in Italy, with details about the target population, the adopted test tests and the time intervals between screening rounds. As far as the cervical cancer is concerned, Italian national guidelines recommended to regions to implement organized screening programmes. The latter includes personal invitations to women aged 25 to 64 years for a Pap test[7] every three years. Regarding mammography[8] screening, the European guidelines recommend that women in the 50-69 year range are invited to undergo mammography every two years. Several programmes continue the invitation up to age 74-75 with a two year interval.

[7] The examination consists in removing, by means of a special spatula and a cotton swab, small amounts of mucus from the cervix and the cervical canal to search exfoliated cells from the tissue. The cells are then fixed to a slide and examined in the laboratory by means of staining methods and examined under a microscope by a cytologist or pathologist who will draft a report.

[8] Mammography is an x-ray exam, in which the breast is compressed between two plates to detect the presence of potential tumor formations.

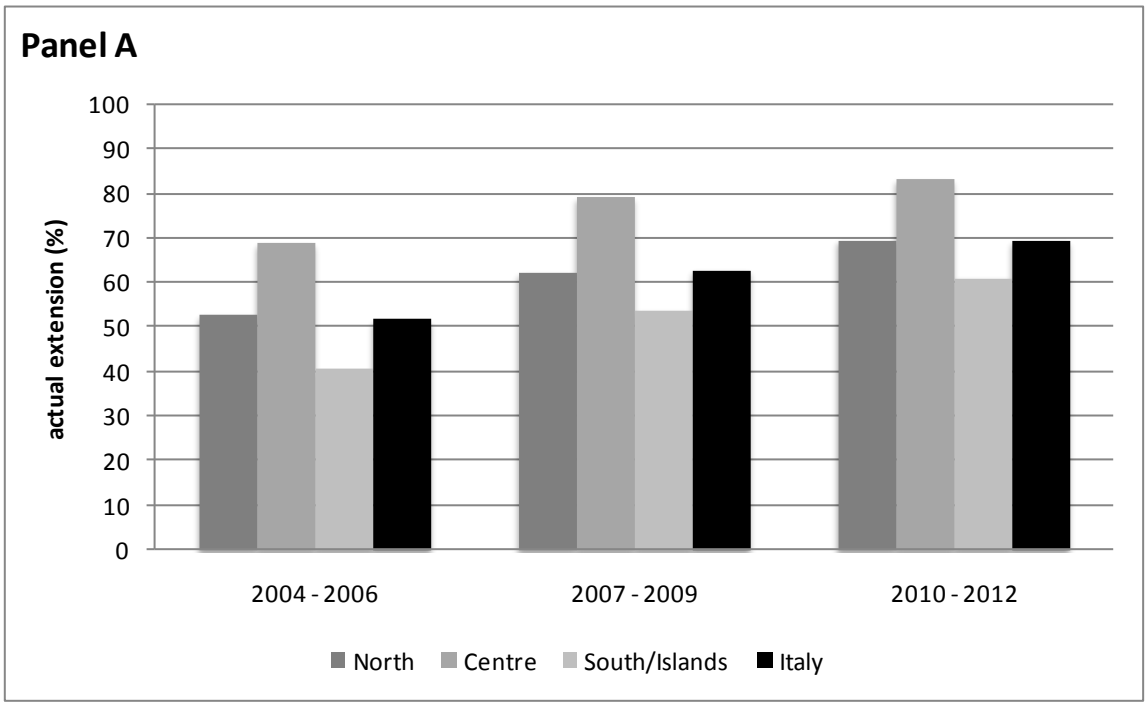
Figure 4. Main characteristics of protocols of mammographic, cervical and colorectal screening programs

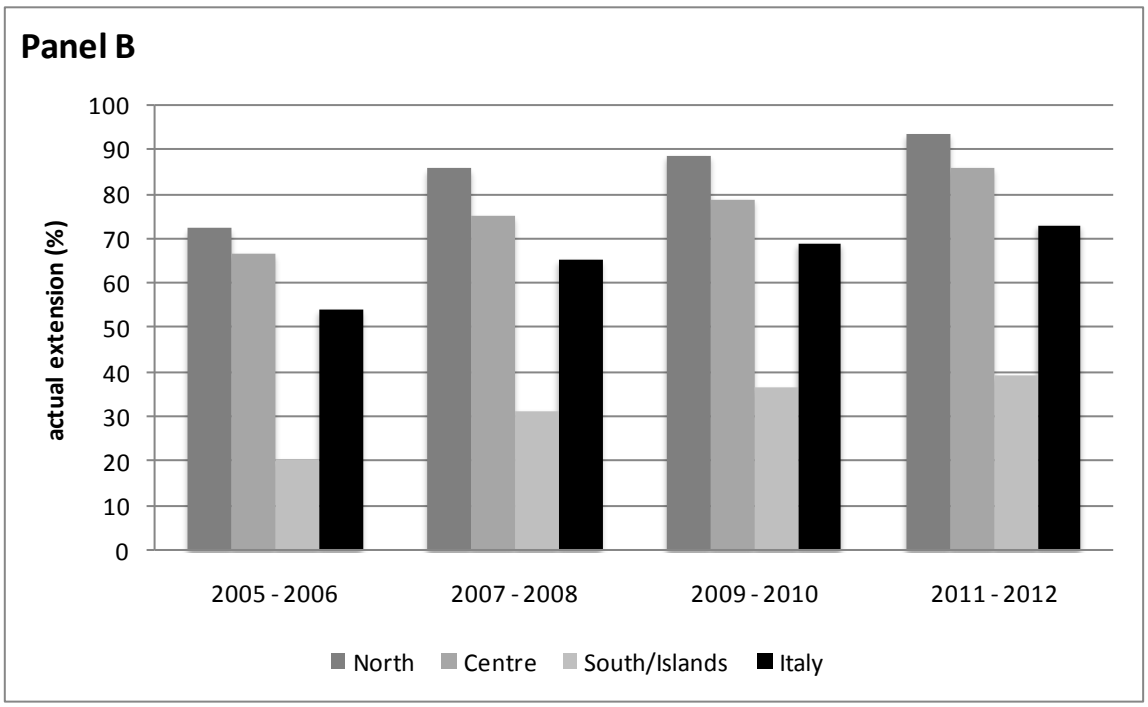
Mammographic screening	
Target population	women aged 50-69 (some regions have extended the age target from 45 to 74)
Primary test	2 views, doubling reading mammographic test
Screening interval	2 years
Cervical screening	
Target population	women aged 25-64
Primary test	Pap smear
Screening interval	3 years
Some programs have moved towards HPV testing as primary test:	
Target population	HPV: women aged 30/35-64 Pap smear: women aged 25-30/35
Primary test	HPV
Screening interval	5 years
Colorectal screening	
Primary test	fecal immunochemical test (FIT)
Target population	subjects aged 50-69 (some regions have extended the age target to 74 or 75 years)
Screening interval	2 years
Primary test	flexible sigmoidoscopy (FS) + FIT
Target population	subjects aged 58 or 60 (FS); subjects aged 59-69 (FIT)
Screening interval	flexible sigmoidoscopy once in a lifetime and FIT every 2 years for non-responders to FS

http://www.osservatorionazionalescreening.it/sites/default/files/allegati/ONS_2015_full.pdf

By looking at the data reported in Figure 5, the *actual extension* (how many women of target population receive regularly an invitation letter) of the cervical cancer screening (Panel A) in 2011-2012 was 69.5%, with an overall increase with respect to 2004-2006 (51.8%) and 2007-2009 (63%) periods. In Panel B the same figures are depicted for the mammography screening: the actual extension was 73.3%, with percentages greater than 80% in the northern and central Italy, in contrast with the 40% observed in the southern and insular regions. Although an overall increase from 2005-2006 to 2011-2012 was observed in all the three macro-areas, the goal of assuring complete breast screening coverage in Italy remains uncertain.

Figure 5. Actual extension of the cervical (Panel A) and mammographic (Panel B) screening programs in Italy (2004-2012)



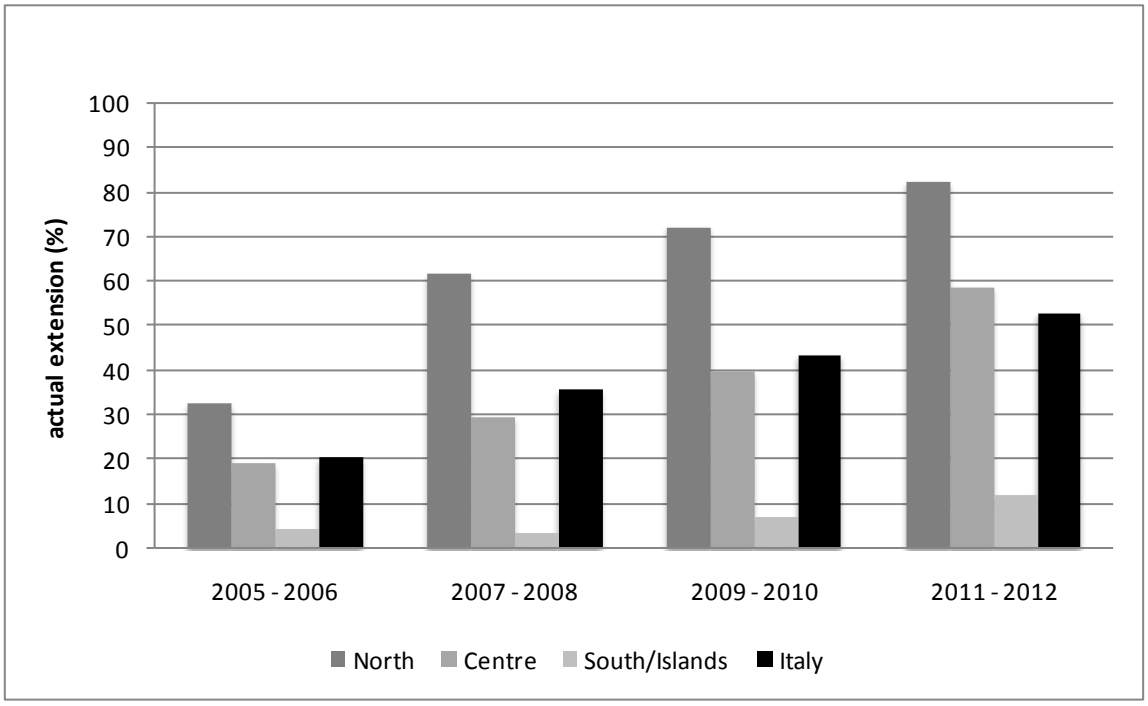


http://www.osservatorionazionalescreening.it

Finally as regards colorectal cancer screening (Figure 6), in the period 2011-2012 it was observed an overall increase of the extension of invitation (proportion of resident population who was sent a screening invitation during the study period) for the whole country. In the 2011 – 2012, the

extension was equal to 53.1% of the target population (men and women aged 50-69) with a substantial increase compared to that observed in the biennium 2005-2006 (20.7%). Unfortunately, as reported in Figure 6 differences between the North and South/Islands still remain with percentage of extension ranging from 82.5% in the North, 58.9% in the Centre, and 12.2% in the South/Islands in the last biennium.

Figure 6. Actual extension of the colorectal cancer screening programs in Italy (2005-2012)



http://www.osservatorionazionalescreening.it

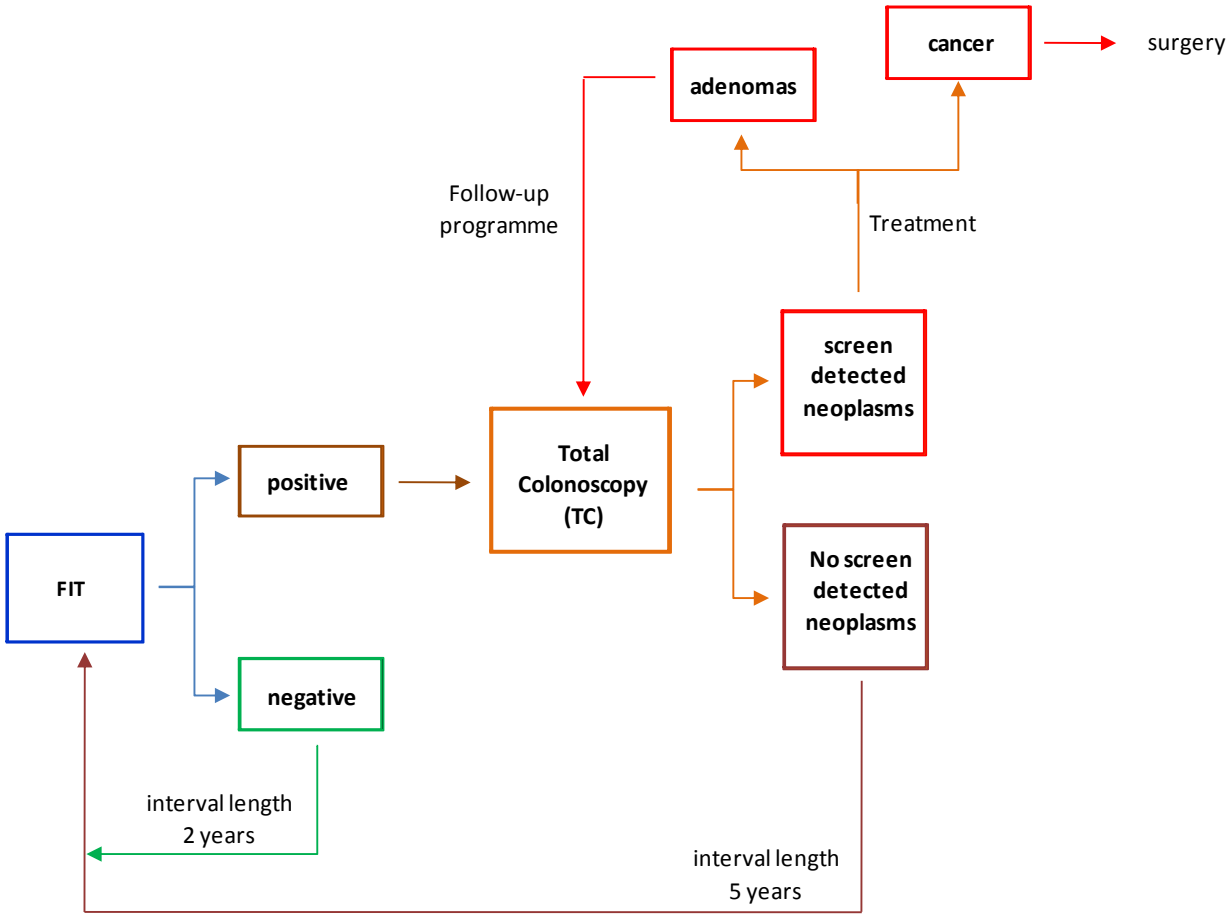
2.1.1 Colorectal Cancer Screening

Colorectal screening programmes are offered to people residents in areas covered by organized screening programmes. As reported in Figure 4, the first level tests are (i) FOBT/FIT offered every 2 years to residents of 50-69/74 years-old (FIT programmes) or (ii) flexible sigmoidoscopy (FS) to an individual age cohort, generally at 58 years-old (FS programmes). FIT programmes have different target populations as far as age is concerned: invitation to attend the screening starts for all the programs at the age of 50 years, but the maximum age ranges from 69/70 years, in most programmes, to 74/75 years in few ones. FS programmes however invites a single cohort of subjects aged between 58-60 years.

By focusing on the FIT programs, they are organized to send an invitation letter to the target population every two years to perform the FIT without any dietary restriction. Non

responders to the first invitation letter are usually re-invited after 6 months. Quantitative haemoglobin analysis is performed in a centralized reference laboratory using the threshold of 100ng/ml of faecal haemoglobin as cut-off value to determine the positivity to the test. People with a negative test are informed by mail about their results and are then invited after 2 years to repeat the test. Subjects with a positive test (FIT+) are contacted in order to perform a total colonoscopy (TC) or, when a complete colonoscopy is not possible, a double-contrast barium enema X-ray. Colonoscopies are performed at referral centres during dedicated sessions. Subjects with a screen-detected neoplasm undergo surgery or, after the neoplasm removal, are enrolled in a follow-up program. The outcomes of second level assessment can be: (a) negative, repeat FOBT/FIT after 5 years, (b) cancer, further diagnostic assessment and surgery/therapy or (c) adenoma, endoscopic surveillance programme (Figure 7).

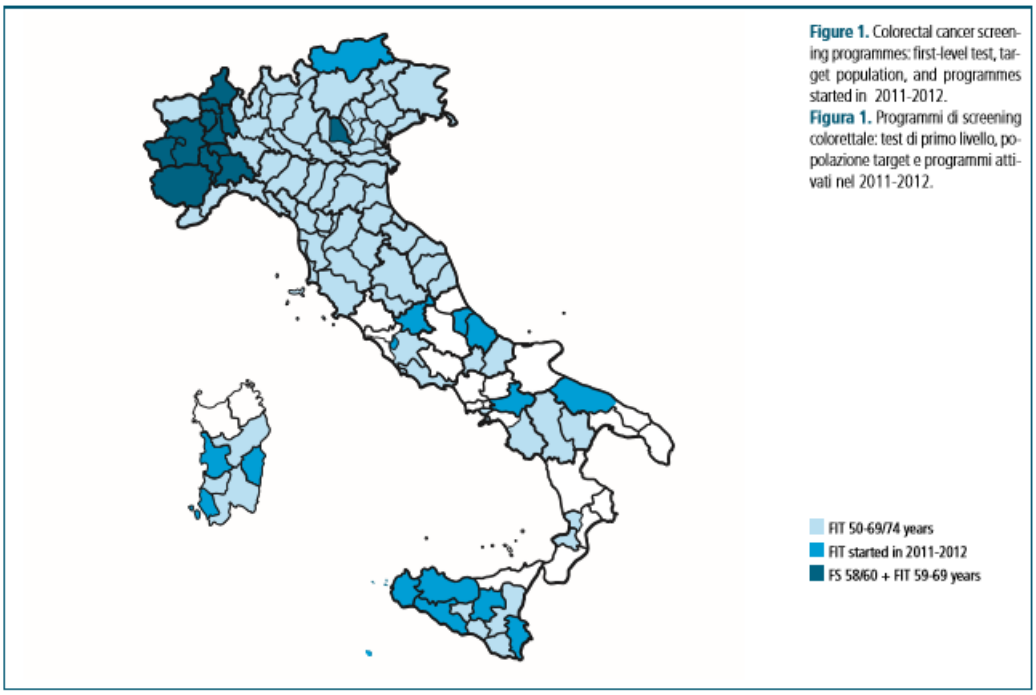
Figure 7. Colorectal Cancer screening workflow.



2.1.1.1 CRC screening – extension and compliance

As aforementioned, the extension of invitation reached the 53.1% of the target population in the 2011-2012 biennium. Figure 8 depicts the extension of different types of for CRC screening programmes at the 31/12/2012 together with the corresponding target populations. Overall, in the last biennium, twelve new programs were activated (especially in the South of Italy), whereas 7 were suspended, for a total of 112 active programmes. The majority (104/112) of the programs are based on FIT, while the remaining ones adopted flexible sigmoidoscopy once in a lifetime, and FIT for non-responders to FS.

Figure 8. Colorectal Cancer screening programmes – first level test, target population and programmes started in 2011-2012



http://www.osservatorionazionalescreening.it/sites/default/files/allegati/ONS_2015_full.pdf

The national theoretical extension (refers to eligible subjects residing in areas covered by organized screening programs) of the CRC screening programs reaches 72.3%, with a 88.5% and 80% in the northern and central regions, respectively. These figures decrease to 45.2% in the South of Italy and Islands. Compared to the previous years, there was an overall increase, moving from 64.9% observed in the 2011 to the 72.3% observed in the 2012, with the major increase registered for the South and Islands (from 25.2% to 45.2%).

As regards the compliance to the screening, the adjusted one[9] slightly decreased in the 2011-2012 compared to that observed in the previous biennium (47.1% *vs* 48%). By looking at the macro-regional areas, the adjusted compliance rate was even higher in the North/Centre of Italy with respect to the South and Islands, and a high intra-regional variability was observed, with a minimum observed value in Campania (13.7%) and the highest one in Valle d'Aosta and Veneto (> 65%). To note that overall, ~57% of the programs reached the GISCoR standard of >45% of adjusted compliance.

2.1.1.2 CRC Screening – diagnostic indicators

The most important diagnostic indicators in a screening program are the test positivity rate (PR), the detection rates (DR) and the positive predictive values (PPV), that are however influenced by the frequency of the disease in the screened population. Specifically, CRC and pre-cancerous lesions are more frequent in male than female, increase with age and disease is more frequently detected in subjects at the first screening test (prevalence test, 1st round), than in those at repeated tests (incidence rounds). All the indicators, reported in the 11th National Screening report (www.osservatorionazionalescreening.it), are thus standardized by age and gender, using the national mean as standard population and estimated separately according to the screening rounds.

As concerns the first-level test, in 2011-2012 the proportion of subjects resulted positive to the FIT is equal to 5.2% at the first screening test; the PR decreases to 4.0% for the repeat tests. Both these proportions are higher in males and progressively increase with age, particularly at first round. The second-level test (i.e. colonoscopy) was performed by the 81.1% of FIT+ subjects, in line with that observed in the previous years. The attendance rate to colonoscopy was higher in the North (83.0%), followed by the Centre (79.6%) and South/Islands (67.0%). Overall, 19.5% of the programmes met the desired standard of an attendance rate > 90%, but only 7.8% was below the cut-off of 70%. The attendance rate as well as the percentage of complete exam were higher in males than in females. Finally, the 95.5% of FIT+ subjects completed the overall diagnostic

[9] adjusted compliance is calculated as the proportion of subjects invited to attend screening (minus those with a wrong address and those excluded after invitation for a recent test) who underwent a screening test

workup[10]. As regards the percentage of complete colonoscopies, the majority of them were classified as such with 81% of the programs showing acceptable results (> 85%) and 61.5% reaching also the desired standards (> 90%).

The detection rate (DR) of invasive carcinomas, advanced adenomas (i.e., adenomas with a diameter ≥1 cm, villous/tubulo-villous type, or high-grade dysplasia) and non-advanced adenomas (smaller in size, tubular type, and low-grade dysplasia) are defined as the number of histologically confirmed lesions detected per 1000 screened subjects. Overall, the DR was 2.2‰ for carcinoma, 10.3‰ for advanced adenomas and 7.1‰ for non-advanced adenomas in subjects screened for the first time; these figures decrease in subjects undergoing repeated testing (1‰ for carcinoma, 6.8‰ for advanced adenomas and 6.1‰ for non-advanced adenomas). This reduction could be due to the removal of polyps performed in the previous rounds and the lower FIT positivity rate at the repeated tests.

By looking at the DRs at the first exam (Figure 9, Panel A), a high variability was observed among regional data: the DRs of carcinomas ranges from 1.7‰ observed in Calabria to more than 5‰ in Alto Adige. The corresponding figures for advanced and non advanced adenomas, respectively range from a 1.9‰ in Puglia to 13.7‰ in Emilia Romagna/Marche and from a 3.3‰ registered in Puglia to 14.7‰ in Friuli Venezia Giulia/Bolzano. By looking at the macro-area average data, no geographical trends were however observed (carcinomas: 2.3‰, 2.2‰, 2.2‰; advanced adenoma: 11.2‰, 10.6‰ and 7.1‰; non-advanced adenoma: 7.6‰, 7.5‰, 4.8‰ in the North, Centre and South/Islands, respectively). Higher homogeneity was observed for the repeated tests (Figure 9, Panel B). These figures were again higher in male compared to female and increased with age in both genders, irrespectively from the screening round.

As concerns the Positive Predictive Value (PPV) of FIT+ at colonoscopy[11], a diagnosis of carcinoma after colonoscopy was formulated in 5.3% of FIT+ subjects (at the first screening round) and that of advanced adenoma in a further 24.5%; among subjects at repeat screening, the corresponding values were 3.1% for carcinoma and 20.1% for advanced adenoma. Higher values of PPV for carcinoma/advanced adenoma were observed for male compared to female (31.0% vs 22.6%) and for age-cohort 65-69 compared to 50-54 (29.4% and 24.2%, respectively).

[10] proportion of subjects who underwent a second-level workup who had a complete assessment (a complete total colonoscopy and/or other exams)

[11] PPV is defined as the number of subjects with a diagnosis of carcinoma/advance carcinoma as proportion of FIT+ subjects that underwent colonoscopy

Figure 9. FIT programmes: standardized (by age and gender, utilizing the overall screened population as standard population) detection rates for carcinoma, advanced adenoma and non-advanced adenoma at first screening (Panel A) and repeat screening episodes (Panel B), by region. Years 2011-2012.

Panel A

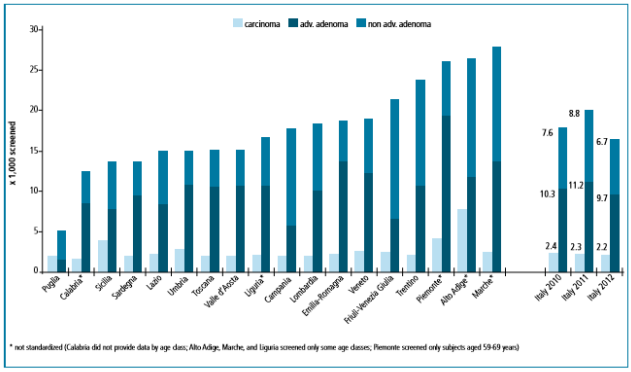


Figure 4. FIT programmes: standardized (by age and gender, utilising the overall screened population as standard population) detection rates for carcinoma, advanced adenoma and non-advanced adenoma at first screening, by region. Years 2011-2012.
Figura 4. Programmi SOF: tassi di identificazione di carcinoma, adenoma avanzato e adenoma iniziale ai primi esami, standardizzati (per età e sesso, utilizzando come riferimento l'intera popolazione screenata), per regione. Anni 2011-2012.

Panel B

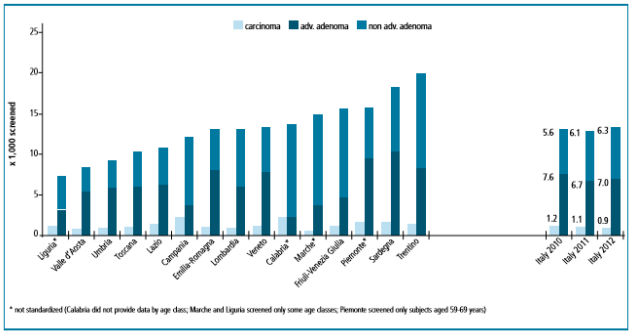


Figure 5. FIT programmes: standardized (by age and gender, utilising the overall screened population as standard population) detection rates for carcinoma, advanced adenoma and non-advanced adenoma at repeat screening episodes, by region. Years 2011-2012.
Figura 5. Programmi SOF: tassi di identificazione di carcinoma, adenoma avanzato e adenoma iniziale agli esami successivi, standardizzati (per età e sesso, utilizzando come riferimento l'intera popolazione screenata), per regione. Anni 2011-2012.

http://www.osservatorionazionalescreening.it/sites/default/files/allegati/ONS_2015_full.pdf

Finally, the time interval between the execution of the first level test (FIT) and mailing of a negative results was short with a 94% of the letters sent within 15 days after the FIT exam and a further 3% within 21 days. On the contrary, colonoscopy was carried out within 30 days after a FIT positivity only in the 53.3% of cases and only 9 programmes met the acceptable standard (>90%); in addition, the 15% of FIT+ subjects waited for more than two months for the screening colonoscopy. As regards the surgery, it was performed within 30 days after diagnosis in 52% of cases, and in a further 33% within two months.

2.1.1.3 CRC Screening – follow-up programs

As reported in Figure 7, at the end of the diagnostic workup, FIT+ subjects are recommended to specific follow-up actions according to the results of the colonoscopy output. Subjects with a negative colonoscopy should be invited for FIT test after 5-years, whereas subjects with an advanced adenomas should be recalled to colonoscopy after 1-3 years, according to the number/size of adenomas. Eighty percent of FIT+ subjects with a negative colonoscopy were invited to repeat FIT after 5 years, and a 4% after 2-years. To note that a total of 13% of subjects were invited to perform a further colonoscopy after a period ranging from 6/12 months (2.4%) to 5 years (7.3%). A similar path (FIT test after 5 years) should be followed by non-advanced adenomas (or low-risk adenomas) according to the European guidelines: this was respected only in the 8.1% of cases; in the 2.6% of cases subject were invited to repeat FIT after 2 years. The principal indication is a colonoscopy after 5 years in the 50.8%, 3 years in the 22.2% and 7.9% after 6-12 months. Surgery was performed in a 0.6% and 1.9% of subjects with a negative result or with non-advanced adenomas, respectively. Seventy-three percent of subject with an advanced adenomas (or high- risk adenoma) were invited to perform a colonoscopy after 1-3 years in line with the recommendation, even if in a 9.6% of cases the colonoscopy was anticipated only after 6-months. To note that a 2% of subjects were recalled for FIT after 2 or 5 years and 5.7% underwent surgery (in 88% of cases the treatment was exclusively endoscopic). Finally as concerns cancerized adenomas (or malignant polyps), 68.9% were sent to surgery and a 16.9% was invited to repeat colonoscopy after 6-12 months. The 85% of carcinomas underwent surgery whereas the remaining fraction underwent endoscopic resection, only.

2.1.1.4 Lombardy Screening Program

The Lombardy target population (50-69 years resident in Lombardy) for the CRC screening is grater that 2 millions according to the ISTAT estimates at 1st January 2012 and a total of 15 screening programmes are currently active. A complete theoretical extension was observed for the biennium (2011-2012), with a coverage (proportion of eligible subject screened in the biennium) equal to 45.4%. The percentage of subjects invited to the 2011-2012 screening was 97.2% and the adjusted compliance was equal to 48.5%. The corresponding 2014's percentages are equal to 44%, 96% and 49.3% for coverage, extension of invitation and adjusted compliance, respectively.

The FIT positivity rates (PR) in the 2011-2012 biennium are equal to 5.5% and 4.0% for the first screen episode and the repeated ones, respectively. In 2014 the overall PR is equal to 4.6% (5.1% and 4.4% in the first and repeated screening episodes, respectively); by stratifying by gender, this figure is equal to 6.0% in male and 4.3% in female. The attendance rate to colonoscopy is 81% in 2014 in line with that observed in the 2011-2012 at national level.

By looking at the detection rates (DR) in 2014, 1.03 carcinomas were detected per 1000 screened subjects as well as 6.1 and 8.8 adenoma at high and low risk, respectively. These figures are equal to 1.4‰ and 1‰ for carcinomas and to 7.6‰ and 6.8‰ for high risk adenomas, at the first and at repeated screening episodes, respectively.

The PPV of FIT was 22.6% for carcinomas and advanced adenoma altogether (3.5% for carcinomas and 19.1% for adenomas) at first round and 17.4% at repeated rounds (2.4% and 15% for carcinomas and adenomas, respectively). Finally, as regards the waiting time, in the 45% of cases the screening colonoscopy was scheduled within 30 days, and in a 36% of cases between 31 and 60 days ("Gli screening oncologici in Lombardia, Report 2015 su dati di attività 2014", May 2016).

CHAPTER 3. MATERIALS AND METHODS

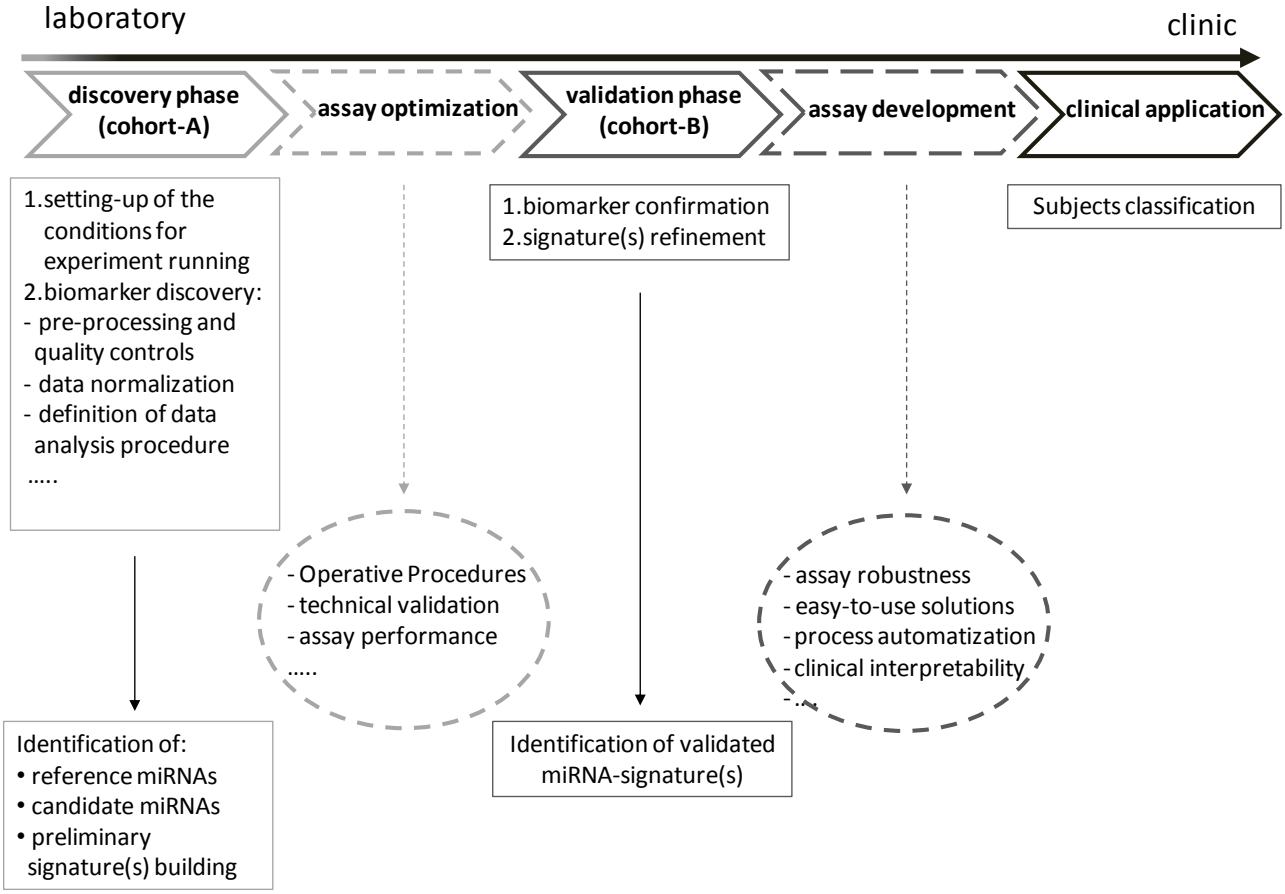
3.1 WORKFLOW FOR CANCER BIOMARKER-SIGNATURE DEVELOPMENT BASED ON MICRORNAS

Many Researches are emphasizing that, in clinical practice, novel circulating biomarker may not possess the desiderated levels of sensitivity and specificity for disease classification and outcome prediction, when alone considered. Accordingly, biomarkers are often combined in a composite score (or signatures in the miRNA context) that achieves better diagnostic accuracy (Yan, Tian, and Liu 2015, 3811-3830).

We have recently proposed (Verderio et al. 2016, 1-4) a workflow covering the major essential steps involved in the cancer biomarker-signature development, from laboratory to clinical practice. This workflow, as reported in Figure 10, tries to schematize all the key phases involved in biomarkers studies, from biomarker discovery to analytical and clinical validation including the issues related to the development of operative procedures for their analysis. The process usually begins with a discovery phase, and is followed by a validation one and then by the clinical application of the identified biomarker-signature. Two additional assay-oriented steps could be introduced in the workflow, before and/or after the validation phase. Biomarker discovery usually starts with the identification of candidate biomarkers from larger sets of tested biomarkers that may be associated with the disease under study. During this phase high-throughput platforms, allowing the evaluation of hundreds or thousands of miRNAs are usually employed; in addition, researchers should set-up all the conditions for experimental running, the pre-processing steps and identify the data normalization procedure to be used for the subsequent analysis. The output of this phase should be the identification of reference and candidate miRNAs as well as a preliminary miRNA-based signature. Validation phase(s) refers to the evaluation of the role of the biomarkers signatures previously identified in an independent cohort of subjects and in their refinement, if necessary. The final goal is the identification of a signature (or a set of signatures) to be translated in the clinical practice for subject classification and outcome prediction. Obviously, discovery and validation phases should be implemented on distinct cohorts of subjects in order to eventually evaluate the performance of the identified biomarkers. An assay optimization step, before the validation phase, could be included in the workflow, if different

assays are used in the discovery and validation steps, in order to evaluate their level of reproducibility (Verderio et al. 2016, 1-4).

Figure 10. Workflow for cancer biomarker-signature development, from laboratory to clinical practice.



Verderio P et al. 2016

As regards the assay development, this could be introduced after the validation phase(s) to set-up an easy-to-use assay to be used in the clinical setting.

This workflow was developed taking into account the peculiarity of the circulating miRNA-based study extracted from plasma/serum, but can be hypothetically applied - with appropriate changes - also to other biomarkers.

From a statistical-methodological point of view, the main key issues involved in this workflow are those related to (i) data normalization of high-throughput qPCR data, (ii) building and validation of miRNA-based signatures.

3.2. DATA NORMALIZATION OF HIGH-THROUGHPUT QPCR DATA

Data normalization represents a crucial pre-processing step aimed both at removing experimentally induced variation and at differentiating true biological changes. Inappropriate normalization strategies can affect the results of the subsequent analysis and, as a consequence, the conclusion drawn from the results. It should be mentioned that the strategies here reported for data normalization were firstly developed in the content of gene expression and then appropriately adopted to miRNA-based studies. According to these considerations, in this section I will use *reference RNA* to indicate both mRNA and miRNAs to be used for data normalization, and *target RNA* to indicate both mRNA and miRNAs of interest.

The choice of a *reference RNA* that shows variation between experimental conditions can bias the estimation of the expression of other *target RNA* within the same samples, leading to over- or underestimation of the true expression of the *target RNA* (Perkins et al. 2012, 296-2164-13-296). In the miRNA context, there are no yet verified and shared *reference RNA* in serum and plasma that can be used for data normalization. Therefore, the pre-processing step of data normalization is really a major challenge in the analysis of circulating miRNAs (Kang et al. 2012, 4-1891-3-4) especially in the analysis of high-throughput qPCR data (Deo, Carlsson, and Lindlof 2011, 795-812). A common approach for qPCR data normalization is the use of invariant *reference RNA* (also defined as endogenous controls) usually identified from a pilot study with samples representative of the experimental conditions under investigation. A suitable endogenous control should be adequately expressed in the sample specimen of interest and show minimal variability in expression between samples under the investigated experimental conditions (Silver et al. 2006, 33).

Many studies until now have used presumed stable expressed *reference RNA* according to the existing data reported in literature (e.g miR-16 and small nuclear RNAs) without proper validation of their stable expressions in the specific context (Kang et al. 2012, 4-1891-3-4). Others, reported the use of synthetic RNA or miRNAs molecules as spike-in controls for mRNA/miRNA expression, not only for monitoring the efficiency of RNA purification and reverse transcription (RT), but also for data normalization (Kang et al. 2012, 4-1891-3-4). While these kinds of spike-in controls have value in assay validation and quality control, they only correct for extraction efficiency or reverse transcription efficiency differences when used for normalization (Mestdagh et al. 2009, R64-2009-10-6-r64. Epub 2009 Jun 16). A different strategy consists in identifying suitable endogenous controls for each study through the systematic evaluation of the expression level of a set of

candidate *reference RNA* (Kang et al. 2012, 4-1891-3-4). Several statistical methods have been proposed to solve the problem of *reference RNA* selection: these methods aimed at selecting the optimal set of *reference RNA* for each experiment. Table 3 summaries the principal strategies reported in literature for the identification of the best set of *reference RNA*.

Table 3. Data normalization strategies

n	Name strategy	stability measure	Reference
1	geNorm	M-value	(Vandesompele et al. 2002, RESEARCH0034)
2	BestKeeper	BestKeeper index (BKI)	(Pfaffl et al. 2004, 509-515)
3	NormFinder	Stability value	(Andersen, Jensen, and Orntoft 2004, 5245-5250)
4	Global mean	-	(Mestdagh et al. 2009, R64-2009-10-6-r64. Epub 2009 Jun 16)
5	NqA	M-value & Stability value	(Verderio et al. 2014, 7-9)

3.2.1 Global mean method

The currently most accepted and widely used method for data normalization of miRNAs is that proposed by Mestdagh and Colleagues, based on the computation of the global mean of the expressed miRNAs (Mestdagh et al. 2009, R64-2009-10-6-r64. Epub 2009 Jun 16). Instead of using a single or a set of *reference RNA*, Authors proposed the use of the mean expression value of all expressed miRNAs as normalization factor. Their results demonstrated that the mean expression value of the expressed miRNAs is characterized by a high expression stability, according to geNorm analysis, resulting in an adequate removal of technical variability, as measured by the coefficient of variation (CV) of the normalized expression values. However, this method is obviously valid if a large number of miRNAs are profiled: typically this is the case of screening experiments performed in the initial phase of a study (discovery phase), but this is almost never applicable in validation studies focused on a limited number of miRNAs. To overcome this issue, Authors proposed to search the set of reference miRNAs that resembles the mean expression value of all the miRNAs, and use that set of *reference RNA* for data normalization.

3.2.2 geNorm strategy

Vandesompele J and Colleagues (Vandesompele et al. 2002, RESEARCH0034) firstly evaluated, in different disease-specimen, the expression level of the ten most commonly used

reference RNA (in the paper called housekeeping genes) in order to assess their presumed stable expressions. To validate that, Authors developed a gene-stability measure to establish the expression stability of *reference RNA* of non-normalized expression levels (raw data). The basic idea is that the expression ratio of two ideal *reference RNAs* is identical in all the samples, regardless of the experimental conditions or cell type. Thus, variation on the expression ratio of two *reference RNAs* implies that one or both genes are not constantly expressed among experimental conditions: thus, an increased ratio-variation means a decreased expression-stability. In details, the pair-wise variation of each *reference RNA* with all the others were computed as the standard deviation of the log-transformed ratio, and the gene-stability measure M was calculated as the average of the pair-wise variation of a particular *reference RNA* with all other *reference RNA*.

Notation:

m: number of tissue samples

n: number of *reference RNA* measured

j, k: two *reference RNA*

a_{ij}, a_{ik}: gene expression levels

A_{jk}: vector of *m* elements

V_{jk}: pair-wise variation for the j-th and k-th *reference RNA*

M_j: gene stability measure for the j-th *reference RNA*

$$\forall\, j,k \in [1,n], j \neq k$$

$$A_{jk} = \left\{ log_2\left(\frac{a_{1j}}{a_{1k}}\right), log_2\left(\frac{a_{2j}}{a_{2k}}\right), \dots, log_2\left(\frac{a_{ij}}{a_{ik}}\right) \right\}, i = 1, \dots, m \quad [3]$$

$$V_{jk} = standard\ deviation\ (A_{jk}) \quad [4]$$

$$M_j = \frac{\sum_{k=1}^{n} V_{jk}}{n-1} \quad [5]$$

According to this notation, genes with the lowest stability value (M-value) have the most stable expression. Assuming that *reference RNA* are not co-regulated, a stepwise exclusion of the gene with the highest M-value results in a combination of two consecutively expressed *reference RNA*

that have the most stable expression in the tested samples. The geNorm Microsoft Excel sheet and the R-function automatically calculate the M-value and rank the considered *reference RNA* according to their expression stability by also providing the best combination of the two most stable expressed *reference RNA*. Obviously the number of genes used for data normalization is a trade-off between practical consideration and accuracy (Vandesompele et al. 2002, RESEARCH0034).

3.2.3 BestKeeper Index

An alternative index for the identification of the most suitable *reference RNA* is the BestKeeper Index, published by Pfall MW et al. (Pfaffl et al. 2004, 509-515). In the BestKeeper Excel-based software the input expression data are the Ct values of both *reference* and *target RNA*. The software, in addition to what implemented in geNorm, enables the evaluation of the *reference RNA* and the analysis of the *target RNA*. According to the observed variability, *reference RNA* can be ranked from the most stable (lowest variation) to the least stable (highest variation).

3.2.4 NormFinder strategy

NormFinder (Andersen, Jensen, and Orntoft 2004, 5245-5250) is another common used algorithm for the identification of suitable *reference RNA*. This is a "model-based approach to estimation of expression variation" in which the inter- and intra-group variation are computed separately and then combined in a measure of stability representing the estimated systematic error: a low stability value means a low systematic error and therefore a stable expression across samples.

<u>*Notation*</u>:

k: number of genes, $i = 1, \dots, k$

G: number of groups (i.e. experimental conditions) , $g = 1, \dots, G$

n_g: number of samples within each group

j: number of sample within a group, $j = 1, \dots, n_g$

y_{igj}: log-transformed gene expression measure for the gene i in sample j within the group g

α_{ig}: expression measure of gene i on the group g to which the sample j belongs

β_{gj}: amount of template in the sample j within the group g

σ_{ig}^2: intra-group variance (variance of gene i in group g)

δ_{ig}^2: inter-group differences (differences in gene expression between groups)

Model:

$$y_{igj} = \alpha_{ig} + \beta_{gj} + \varepsilon_{igj} \quad [6]$$

where ε_{igj} is an error term with 0-mean and variance depending from gene i and group g

Intra-group variance estimation (σ_{ig}^2)

$\overline{y_{ig.}}$: gene average (average over the sample in group g)

$\overline{y_{.gj}}$: sample average (average over the gene)

$\overline{y_{.g.}}$: average over the genes and samples in group g

s_{ig}^2: sample variance for gene i in group g

$$r_{igj}^2 = y_{igj} - \overline{y_{ig.}} - \overline{y_{.gj}} + \overline{y_{.g.}}$$

$$s_{ig}^2 = \frac{\sum_{j=1}^{N} r_{igj}^2}{(n_g - 1)\left(1 - \frac{2}{k}\right)} \quad [7]$$

$$\widehat{\sigma_{ig}^2} = s_{ig}^2 - \frac{1}{k(k-1)} \sum_{v=1}^{k} s_{vg}^2 \quad [8]$$

Inter-group variation (δ_{ij})

z_{ig}: average of the gene i in group g (above defined as $\overline{y_{ig.}}$)

θ_g: average sample level in group g ($\overline{\beta_{g.}}$)

$$\delta_{ij} = \alpha_{ig} - \overline{\alpha_{i.}}$$

$$mean\,(z_{ig}) = \alpha_{ig} + \theta_g,\ var\,(z_{ig}) = \sigma_{ig}^2 / n_g$$

As θ_g is unknown (amount of template in the group g) and not estimable, even if its combination with $\overline{\alpha_{.g}}$ is estimable, it should be assumed that

$\hat{\theta} = minimize\ variation\ in\ z_{ig} - \theta_g$. In other words it means that $\overline{\alpha_{.g}}$ is independent of the group g.

$$\hat{\delta}_{ij} = d_{ij} = z_{ig} - \overline{z_{i.}} - \overline{z_{.g}} + \overline{z_{..}} \quad [9]$$

Stability value (γ^2)

$$\widehat{\gamma^2} = \frac{1}{(k-1)(G-1)} \sum_{i=1}^{k} \sum_{g=1}^{G} d_{ij} - \frac{1}{kG} \sum_{i=1}^{k} \sum_{g=1}^{G} \frac{\widehat{\sigma_{ig}^2}}{n_g} \quad [10]$$

The output of the software automatically displays the candidates *reference RNA* ranked according to the minimal inter- and intra-group variation.

As reported the Author, it should be mentioned, that the pair-wise comparison approach (Vandesompele et al. 2002, RESEARCH0034) could be problematic in presence of co-regulate genes among the candidate *reference RNA* - as it assumes that *reference RNA* are not co-regulated - and could give misleading results if the candidate *reference RNA* show systematic differences between the experimental conditions groups. On the contrary, these issues are taken into consideration in this model-based approach proposed by Andersen (Andersen, Jensen, and Orntoft 2004, 5245-5250).

3.2.5 Normalization qPCR Array (NqA) strategy

Starting from the approaches reported in literature, we developed a comprehensive data-driven normalization method for high-throughput qPCR data, which identifies a small set of miRNAs to be used as *reference* for data normalization in view of the subsequent validation studies (Verderio et al. 2014, 7-9; Pizzamiglio et al. 2014, 2016-2018). Figure 11 reports the procedure we developed. By starting from the high-throughput qPCR data, we considered the N miRNAs expressed in all the samples (1) in order to compute the mean expression value (2) according to Mestdagh et al. From the latter we computed, for each miRNA, the corresponding ΔCt value as follows:

$\Delta Ct_{miRNA[i]} = Ct_{miRNA[i]} - Ct\ mean_{[i]}$, where $mean_{[i]}$ is the mean of the N miRNAs (overall mean).

The distribution of the $\log_2 RQ_i$ ($-\Delta Ct_{miRNA[i]}$) in cases vs controls was compared by means the Kruskal-Wallis test (3) (Hollander, M., Wolfe, D.A. 1999) in order to identify a list of miRNAs differentially expressed between cases and controls (*reference list*, 4).

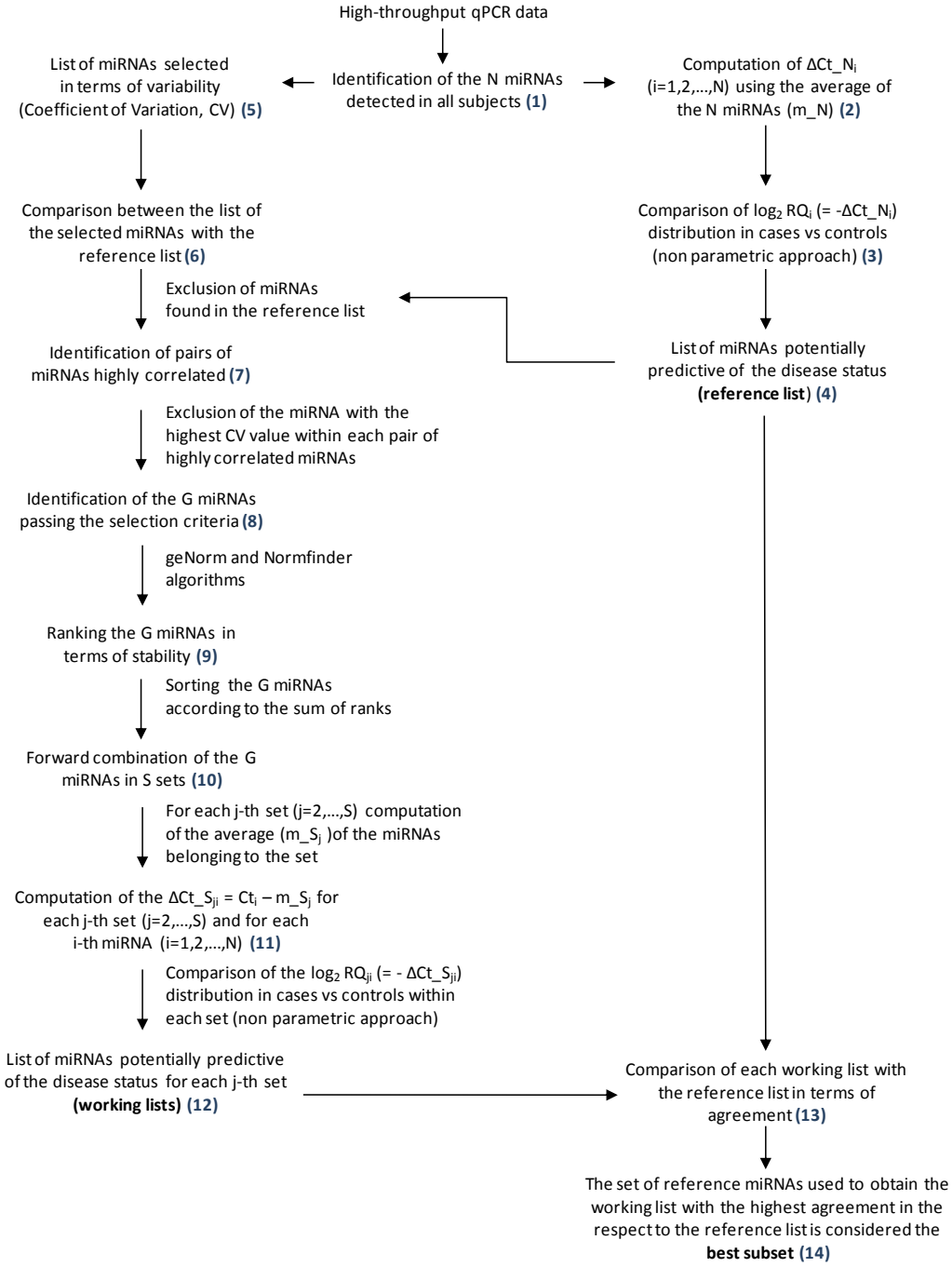
In parallel, a subset of G candidate *reference RNA* was identified according to appropriate selection criteria such as:

- variability (5): selection of miRNAs with a CV less than or equal to the 20th centile of the CV's distribution obtained by considering all the N miRNAs.

- invariance between comparison groups (6): exclusion of miRNAs showing a statistically significant difference in the $\log_2(RQ_i)$ distribution between cases and controls by using the overall mean as normalization method and by applying the Kruskal-Wallis Test. In other words this means excluding from the list of candidate reference miRNAs those reported in the *reference list*.
- co-regulation (7): identification of pairs of miRNAs highly correlated according to the Spearman correlation coefficient, with a lower limit of the Fisher Confidence Interval (Fisher 1970) greater than the value of 0.80. Within each pair, the miRNA with the highest CV is excluded.

Subsequently the identified G miRNAs were evaluated through both geNorm and NormFinder R-based function (8) and ranked according to the stability values (9). The G miRNAs were then forwardly combined in G-1=S subsets (with at least 2 miRNAs) according to their stability value (10). Once computed for each j-th set (j=1,…,S) the specific mean (m_S_j), the relative quantity of each i-th miRNA is calculated as $\log_2(RQ_{ji}) = -\Delta Ct\_S_{ji}$ where $\Delta Ct\_S_{ji} = Ct_i - Ct\ m\_S_j$ (11). The $\log_2(RQ_{ji})$ distribution was compared between groups (cases vs control) by means of Kruskal-Wallis test within each j-th set (12, *working lists*). Finally, the smallest set of reference miRNAs showing results with the highest agreement (with the upper limit of the 95% CI of the kappa statistic ≥0.80 or the highest value of the kappa statistics when no 95% CI include the threshold of 0.80, (Fleiss,J.L., Levin, B., Paik, M.C. 2004)) with those obtained when considering the overall mean (13) was identified as the best subset of reference miRNAs (14). This procedure is implemented in an R-based function, called NqA (**N**ormalization **q**PCR **A**rray) (Verderio et al. 2014, 7-9).

Figure 11. Procedure developed for the identification of the best subset of reference miRNAs



Verderio P et al. 2014

3.3 MEASUREMENTS FOR EVALUATING A DIAGNOSTIC TEST

The evaluation of a diagnostic tests is an important topic in medicine not only for confirming the presence of disease but also for excluding the disease in healthy subjects. For diagnostic tests with a dichotomous outcome (positive/negative test result), sensitivity and specificity are usually estimated to measure the accuracy of the test in comparison with the gold standard status. Starting from a 2x2 table (see Table 4), the following quantities can be estimated:

Table 4. Quantities for evaluating a diagnostic tests

		Golden standard		
		Positive	Negative	Total
Test under evaluation	Positive	TP	FP	TP + FP
	Negative	FN	TN	FN + TN
	Total	TP + FN	FP + TN	Tot

True positive (TP) are patients with the disease and for which the test resulted positive;
False positive (FP) are patients without the disease but classified as positive by the test;
False negative (FN) are patients with the disease but classified as negative by the test;
True negative (TN) are patients without the disease and correctly classified as such by the test.

Thus, sensitivity (SE) refers to the ability of the test to correctly identify those patients with the disease and can be estimated as the proportion of TP out of the total number of subjects positive with the golden standard. Technically,

$$SE = \frac{TP}{TP + FN}$$

Conversely, specificity (SP) refers to the ability of the test to correctly identify those patients without the disease.

$$SP = \frac{TN}{FP + TN}$$

Other two measurements that could be calculated from a 2x2 table are the Positive and Negative predictive value (PPV and NPV), which are useful measurements to answer the questions: “what is the probability that this patient has (has not) the disease given that the test output is positive (or negative)?”

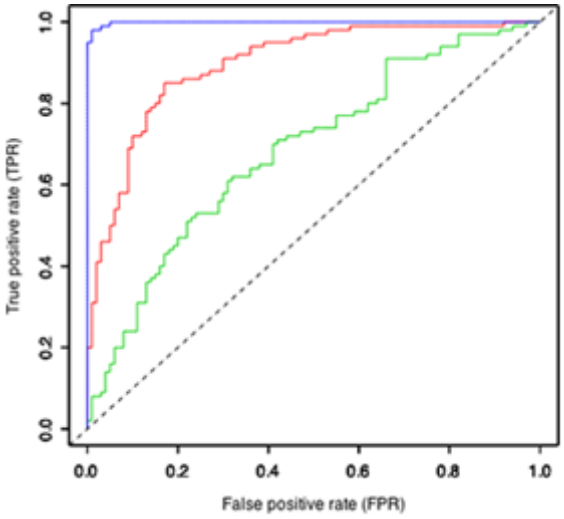
$$PPV = \frac{TP}{TP + FP}; \quad NPV = \frac{TN}{FN + TN}$$

Unlike sensitivity and specificity, the PPV and NPV are dependent on the population being tested and are influenced by the prevalence of the disease.

When the test results are recorded in ordinal scale (e.g. 5 ordinal scale: "definitely normal", "probably normal", "uncertain", "probably abnormal", "definitely abnormal") or on a continuous ones, sensitivity and specificity can be computed across all the possible threshold values. So, the sensitivity and specificity vary across the different threshold, with sensitivity inversely related to specificity. The plot of sensitivity (or True Positive Fraction) versus 1-Specificity (or False Positive Fraction) is called Receiver Operating Characteristic (ROC) curve. The latter is a monotone function from (0,0) to (1,1), with curves near to the (0,1) point associated with better discriminatory ability. By looking at the recent literature, the use of the ROC curve is extensively increased for the assessment of diagnostic ability of biomarkers (both serum/plasma markers and imaging tests) in discriminating diseased from healthy subjects.

The area under the ROC curve (AUC) is the most popular discriminatory accuracy index that summarize the "overall" location of the entire ROC curve: AUC values ranges between 0.5 and 1.0, with value equal to 1 (line blue) indicating a perfect discrimination between the two groups, and values near 0.5 (line green) meaning no discriminatory ability (Figure 12). The AUC can be in fact interpreted as the probability that a randomly chosen diseased subject is rated or ranked as more likely to be diseased than a randomly chosen non diseased subject.

Figure 12. An example of ROC curves



In this scenario, sensitivity and specificity can be estimated once identified a cut-off point of the continuous data under investigation, with obvious differences (in the SE and SP quantities) according to the identified cut-off. To identify the "optimal" cut-off value starting from a ROC curve, different alternatives are available: the Youden Index is the most common approach.

The Youden Index is defined as:

$$J = \max\{SEc + SPc - 1\}$$

where c indicates a threshold value.

The c value that corresponds to the maximum value of J is the "*optimal*" cut-point. Another approach is based on the selection of the optimal cut off value as the point (1-SP, SE) closest to (0,1) corner of the ROC curve.

3.4 COMBINATION OF MULTIPLE BIOMARKERS

As already mentioned, in medical/cancer research, single molecular biomarkers may not achieve satisfactory performance for patients classification. The linear combination of these biomarkers in a more powerful composite score could represent a suitable approach to achieve higher diagnostic performances. This is especially true in the research settings where new assays based on high-dimensional profiling are constantly developed and a wide range of weak-biomarkers (defined as those with an AUC value ranging from 0.50 to 0.70 (Yan, Tian, and Liu 2015, 3811-3830)) are incessantly identified.

As reported by Yan and Colleagues (Yan, Tian, and Liu 2015, 3811-3830), several methods, both parametric and non-parametric, could be used (see Table 5) to find the best linear combination of biomarkers in order to achieve greater discriminatory ability than those obtained using single biomarkers alone.

Table 5. Methods available in literature to find "optimal" linear combination of biomarkers

Methods	References
Su and Liu's method (SL)	(Su and Liu 1993, 1350-1355)
Pepe and Thompson (PT) method	(Pepe and Thompson 2000, 123-140)
Min-Max (MM)method	(Liu, Liu, and Halabi 2011, 2005-2014)
Stepwise (SW) method	(Kang, Liu, and Tian 2016, 1359-1380)
Pairwise (PW) method	(Yan, Tian, and Liu 2015, 3811-3830)
Logistic regression method	(Harrell 2001)

3.4.1 Su and Liu's (SL) method

The SL method (Su and Liu 1993, 1350-1355), proposed in the 1993, showed that the Fisher's discriminant coefficient allows to built the "*optimal*" linear combination of two biomarkers with the largest AUC, among all possible linear combinations under the multinormality assumption of the distribution with proportional covariance matrices. Specifically,

$$\lambda^T = (\Sigma_{\mathrm{H}} + \Sigma_{\mathrm{D}})^{-1}\mu, \text{ where } \mu = \mu_{\mathrm{D}} - \mu_{\mathrm{H}} \quad [11]$$

where μ and Σ can be estimated as the mean and variance/covariance matrix of the considered markers in the healthy (H) and diseased (D) subjects.

The optimal combined AUC can be calculated as follows:

$$AUC_{SL} = \Phi\left(\sqrt{\mu^T \lambda^T}\right) \quad [12]$$

The applicability of the SL method is however limited to the case of multinormality of the distributions (Yan, Tian, and Liu 2015, 3811-3830; Kang, Liu, and Tian 2016, 1359-1380). In addition when the size of the healthy and diseased group is not large enough, the asymptotic

results for this approach may not hold and the corresponding linear combination may not be "*optimal*" (Kang, Liu, and Tian 2016, 1359-1380).

3.4.2 Pepe and Thompson (PT) method

As concerns the Pepe and Thompson (PT) method and the group of combination-based methods described in the next three subparagraphs (MM, SW and PW method), they are all based on the empirical estimation of the AUC using the Mann-Whitney U-statistics. All these methods started from the approach reported by PT (Pepe and Thompson 2000, 123-140), which suggested to search the α–value – in the $(-\infty;+\infty)$ range - that maximizes the AUC value for the linear combination of the p-biomarkers {i.e. $LC = \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_k x_p$}.

The PT approach allows the combination of only two biomarkers, by finding the linear combination $\alpha_1 x_1 + \alpha_2 x_2$ that maximizes the AUC. This is equivalent to find the α-value that maximizes the AUC of the $x_1 + \alpha_2 x_2$ combination as the ROC curve is invariant to scale transformation (Pepe and Thompson 2000, 123-140). In their paper, the α-value that maximizes the AUC was found by dividing the (-1 ; +1) range in 201 equally spaced values. The computation of the AUC is distribution-free (AUC-DF) and thus does not depend on assumptions on the data distribution; however the AUC-DF is not a continuous function of α because this is not a derivative method but a "grid-search method" and could become computationally complex when a large number of markers are involved ($p \geq 3$).

3.4.3 min-max (MM) method

The MM method proposed by Liu et al (Liu, Liu, and Halabi 2011, 2005-2014) linearly combines the minimum and maximum values of the observed biomarkers of each subject. The "*optimal*" linear combination is obtained by finding the α-value that maximizes the AUC-DF of the following linear combination:

$$S(a,X) = X_{max} + \alpha X_{min} \qquad [13]$$

where $\alpha \in (-\infty;+\infty)$ and X_{max} and X_{min} are the minimum and maximum value of each biomarkers.

The principal advantage of this method is that only one α-value should be estimated regardless of the number of biomarkers under consideration, because only minimum and maximum values are used. On the contrary, minimum and maximum values can come from different markers for

different subjects and not all the information contained in the data are used (Kang, Liu, and Tian 2016, 1359-1380).

3.4.4 stepwise (SW) approach

The procedure underling the last two methods (SW and PW) starts with the computation of the empirical AUC, in a univariate fashion, of each considered biomarker. Then in the SW the AUC values are ranked from the highest to the lowest and the first two biomarkers are combined following the PT method. The resulting α-value that maximizes the AUC is selected and then the combination of the first three biomarkers (based on the α-value identified in the aforementioned step) is computed, until all biomarkers are included (Kang, Liu, and Tian 2016, 1359-1380).
In details,

Step 1. $S_{1,2}(\alpha,X) = X_{[1]} + \alpha_1 X_{[2]} \rightarrow \alpha_1 = \max\{AUC[S_{1,2}(\alpha,X)]\}$ [14]

Step 2. $S_{1,2,3}(\alpha,X) = X_{[1]} + \alpha_1 X_{[2]} + \alpha_2 X_{[3]} \rightarrow \alpha_2 = \max\{AUC[S_{1,2,3}(\alpha,X)]\}$ [15]

Step K-1. $S_{1,2,3,..,k}(\alpha,X) = X_{[1]} + \alpha_1 X_{[2]} + \alpha_2 X_{[3]} + ... + \alpha_{k-1} X_{[k]} \rightarrow \alpha_{k-1} = \max\{AUC[S_{1,2,3,..,k}(\alpha,X)]\}$ [16]

The final linear combination is obtained at the step K-1. This method is easier to implement with respect to PT-method also when more than 2 biomarkers are considered. Nevertheless, it should be considered that these methods are focused on the combination of well-defined clinical markers and not in a context with multiple weak biomarkers.

3.4.5 pairwise (PW) approach

The PW approach (Yan, Tian, and Liu 2015, 3811-3830) tries to overcome the drawbacks of the SW approach, by pairing one marker (anchor marker) with all the other ones, separately. The anchor marker could be that with the highest AUC or that selected according to clinical justifications.

Step 1. $S_{1,2}(\alpha,X) = X_{[1]} + \alpha X_{[2]} \rightarrow \alpha_1 = \max\{AUC[S_{1,2}(\alpha,X)]\}$ [17]

Step 2. $S_{1,3}(\alpha,X) = X_{[1]} + \alpha X_{[3]} \rightarrow \alpha_2 = \max\{AUC[S_{1,3}(\alpha,X)]\}$ [18]

Step k. $S_{1,k}(\alpha,X) = X_{[1]} + \alpha X_{[k]} \rightarrow \alpha_k = \max\{AUC[S_{1,k}(\alpha,X)]\}$ [19]

The final linear combination is obtained as follow: $X_{[1]} + \alpha_1 X_{[2]} + \alpha_2 X_{[3]} + ... + \alpha_k X_{[k]}$

The principal advantage of this method is that it uses all the biomarkers by performing all the pairwise combinations and it is not computationally complex.

All these reported methods considered the AUC as objective function. Yin J and Tian L (Yin and Tian 2014, 1426-1440) proposed on the contrary the use of the Youden Index as objective function. Authors underline the importance of the Youden Index as it offers the optimal cut-off point and also gives a direct measure of the maximum overall correct classification rate that a marker can achieve. To find the optimal linear combination Authors proposed either derivation-based method (i.e. empirical searching methods) similar to those previously described (MM and SW) or parametric (based on multivariate normality) and non-parametric method (based on kernel distribution).

3.4.6 Logistic regression model

The logistic regression model is the most widely used approach in medical and epidemiological areas to assess and predict the probability of the occurrence of a binary outcome. It allows to study how a set of predictors/covariates - denoted $X_{(i)}$ - is related to a binary response variable - denoted Y.

Specifically,

$$logit\ (P[Y_i = 1 \mid x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = \boldsymbol{\beta}\, \boldsymbol{x_i} \quad [20]$$

where Y_i is the disease status, $\boldsymbol{x_i}$ is the p-dimensional vector of the measured covariates for the i-th patient and $\boldsymbol{\beta}$ is the vector of the β coefficients chosen to maximize the log-likelihood (MLE).
In other words, it allows to assess the relationship (strength and direction of association) between the covariate(s) and the outcome, as well as to estimate the probability that a specific outcome is present (or absent), within an individual.

3.5 PREDICTION MODELS

Prediction models are usually developed to guide healthcare professionals in their decision-making regarding future management of patients (i.e. submit individuals to additional tests) or to inform individuals about their risk of having a particular disease. These models, which are inherently multivariate, are tools that combine multiple predictors by assigning the relative

weights for each predictor: the coefficients quantify the contribution of each predictor to the outcome probability or risk estimation. Technically, a regression coefficient indicates the effect of a one-unit increase in the level of the relevant predictor on the estimated outcome risk when the other predictors in the model are kept constant (Moons et al. 2012, 683-690).

Prediction models studies are usually organized in a *model development* and in a *model validation* phase or a combination of both. In the first phase the aim is to derive a multivariate prediction model by selecting the most relevant predictors and combing them into a multivariate model. The evaluation of the performance of the model on other data, not used for model development, represents the model validation phase.

The strategy that should be followed to develop a predictive model is reported in the subsequent section.

3.5.1 Model development

The development of a prediction model requires (i) the identification of the most important predictors (out of a set of pre-identified predictors) that should be considered in the model, (ii) the estimation of the relative weight for each predictor in a combined score, (iii) the estimation of the performance of the model and (iv) the assessment of its potential over-optimism using internal validation techniques. Briefly, the identification of the most important predictors can be done by using model selection strategies and the estimation of the weights of each predictor can be obtained by properly fitting a (multivariate) logistic regression model, in which the estimates of each regression coefficient are mutually adjusted for the other predictors included in the model (Moons et al. 2012, 683-690).

The followings sections provide details about the identification of candidate predictors and their combinations in a multivariate (predictive) model, as well as an overview of the quantities that should be considered for estimating the performance of a model and quantifying its optimism.

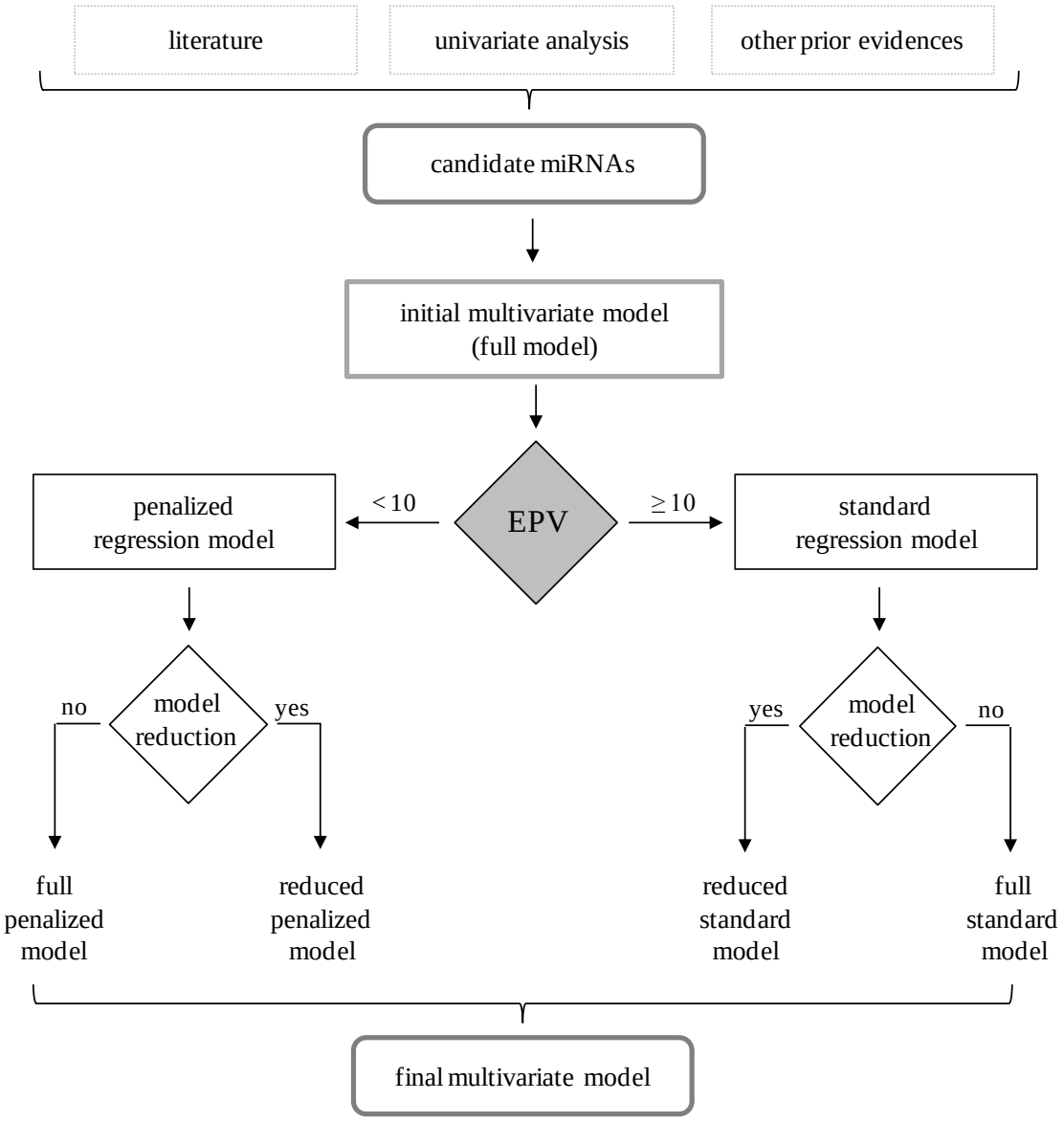
3.5.2 miRNA-based signature development

As depicted in Figure 10, a development dataset or training dataset (cohort-A) should be used for building a model (i.e. miRNA-based signature) that should be then tested on an independent cohort of subjects - validation dataset or testing dataset - to eventually evaluate the

predictive ability of the signature in discriminating subjects with the disease from subjects without the disease.

Figure 13 graphically summarizes the workflow that we proposed (Verderio et al. 2016, 1-4) for developing a miRNA-based signature. The process starts with the identification of the candidate miRNAs that should be included in the initial multivariate model. These candidates could be selected from literature or from prior evidences or can be the results of univariate analysis within the same study. Once identified these candidates, the model development phase could start, with the fitting of the initial multivariate model. It is mandatory to consider the number of event-per-variable (EPV) and the related overfitting problem to appropriately fit a multivariate model. Overfitting can in fact occur when the number of covariates is larger than the number of outcome events, so that the estimated model tends to capture not only the underlining process that generated the data, but also noise leading to an over- or under- estimation of the risk of the event in high or low risk patients (Pavlou et al. 2016, 1159-1177). As a *rule of thumb* it has been suggested that models are likely to be reliable when the EPV is at least 10 (Verderio et al. 2016, 1-4; Pavlou et al. 2016, 1159-1177). As reported in Figure 13, when EPV is less than 10 the use of penalized regression strategies (Verderio et al. 2016, e5) may represent a useful tool to reduce overfitting as much as possible (Verderio et al. 2016, 1-4), whereas when EPV > 10 standard regression strategies can be used.

Figure 13. statistical analysis flowchart for miRNAs signature development



Verderio P et al. 2016

3.5.2.1 Penalized regression models

Penalized regression models are mainly used to prevent overfitting when a large number of covariate are present in a model with respect to the number of outcome events.

As already mentioned (section 3.4.6), the estimation of the β coefficients in a standard regression model, is done by maximizing the log-likelihood function. The Newton-Raphson method is usually used to solve iteratively for the list of β that maximises the log-likelihood. The maximum log-likelihood estimations (MLE) are denoted as $\hat{\beta}$.

In the penalized regression models, $\hat{\beta}$ values are however obtained by maximising the penalized log-likelihood (and not the log-likelihood function), as follows:

$$\hat{\beta} = argmax\,\{l(\beta) - \lambda\,\mathrm{pen}(\beta)\} \quad [21]$$

where pen(β) is the penalty term and λ is the tuning parameter.

The penalty term corresponds to the functional form of the constraints, and the tuning parameter to the amount of shrinkage applied to the β. A λ=0 leads to the standard MLE estimates.
As concerns the λ tuning parameter it is usually selected using data-driven procedure, such as cross-validation (Pavlou et al. 2016, 1159-1177), whereas, for the penalty term different forms of the constraints where proposed in literature. The two most popular penalized regression approaches are the LASSO (Tibshirani 2011, 273-282) and ridge (Le Cessie and Van Houwelingen 1992, 191-201) methods.
In the first case, the constraint is imposed on the sum of the absolute value of the regression coefficients, so that the coefficients are shrunk towards zero. With this approach the shrinkage of some coefficients to exactly zero intrinsically allows model reduction.

$$\widehat{\beta_{LASSO}} = argmax\,\{l(\beta) - \lambda_1 \sum_{j=i}^{p} |\beta_j|\} \quad [22]$$

As regards the ridge method, the penalty term is proportional to the sum of squares of regression coefficients. As in LASSO, the ridge method shrinks coefficients towards zero (but not to exactly zero) and has been seen to perform well in scenario with correlated predictors.

$$\widehat{\beta_{ridge}} = argmax\,\{l(\beta) - \lambda_2 \sum_{j=i}^{p} \beta_j^2\} \quad [23]$$

Other forms of constraints are available, such as those reported by Pavlou (Pavlou et al. 2016, 1159-1177), summarized in the following subparagraphs.
Elastic net, is a hybrid between ridge and LASSO because it produces more parsimonious models with respect to ridge by performing variable selection, but also tends to select or omits highly correlated predictors as a group.

$$\widehat{\beta_{EN}} = argmax\,\{l(\beta) - \lambda_1 \sum_{j=i}^{p} |\beta_j| - \lambda_2 \sum_{j=i}^{p} \beta_j^2\} \quad [24]$$

Adaptative LASSO introduces a different weight (usually data dependent) for each parameter in the penalty term, leading to a smaller shrinkage of the coefficients of the strong parameters with respect to the weak predictors.

$$\widehat{\beta_{AL}} = argmax\,\{l(\beta) - \lambda_1 \sum_{j=i}^{p} \omega_j |\beta_j|\} \quad [25]$$

An additional variant is the Smoothly clipped absolute deviation (SCAD), which allows parameter estimation and selection. It uses a non-quadratic spline function as reported below, and applies a shrinkage similar to that of Adaptative LASSO.

$$pen(\beta_j) = \begin{cases} \lambda|\beta_j| & \text{if } |\beta_j| < \lambda \\ \frac{\lambda(\alpha - |\beta_j|/2\lambda)}{\alpha - 1} & \text{if } \lambda < |\beta_j| \leq \alpha\lambda \\ \lambda \frac{\alpha^2 \lambda}{2(\alpha-1)|\beta_j|} & \text{if } |\beta_j| > \alpha\lambda \end{cases} \quad [26]$$

In this model, there are two turning parameters (α and λ) that should be chosen. As reported by Pavlou M et al., Fan and Colleagues suggested to choose α equal to 3.7 and to estimate λ via cross-validation (Pavlou et al. 2016, 1159-1177)

Another alternative to prevent the lower accuracy of models when applied to new patients is the use of Penalized Maximum Likelihood Estimation (PMLE) (Harrell 2001; Moons et al. 2004, 1262-1270). This method, developed for logistic regression models, is a generalization of the ridge method and allows, as the other methods, the shrinkage of the coefficients directly during the fitting of the model. PMLE maximizes the penalized log-likelihood, leading to an estimation of the regression coefficients according to the following formula:

$$\widehat{\beta_{PMLE}} = argmax\,\{l(\beta) - 0.5\,\lambda \sum (s_i \beta_j)^2\} \quad [27]$$

where s_i is a scaling parameter factor for each β_j.

The optimal penalty factor could be estimated using cross-validation procedure or by maximizing the modified-AIC, defined as:

$$\chi^2_{LR} - 2\,p * d_{eff} \quad [28]$$

where χ^2_{LR} is the likelihood ratio of the penalized model and d_{eff} are the degrees of freedom after penalizing the fitted predictors (p). In the logistic regression model the degrees of freedom are equal to the number of predictors; due to penalization, the degrees of freedom effectively used in the PMLE are fewer than the actual number of predictors, decreasing the potential overfitting.

3.5.2.2 Model reduction strategies

Another important theme is the definition of the final model that could be recognized as the full initial model or as a reduced one, when the intent is to obtain a more parsimonious model without a substantial loss of information (see Figure 13) (Verderio et al. 2016, 1-4; Moons et al. 2012, 683-690). In case of no model reduction, the final multivariate model is equal to the initial multivariate model (full model), in which all the *a-priori* included predictors are considered in the model and no predictor selection is performed. In the other case, the initial multivariate model is reduced and a final model, including fewer predictors, is obtained.

Several well-established approaches for standard regression models are available, such as backward elimination or forward selection. The backward selection procedure starts from the full initial multivariate model (including all the candidate predictors) and runs a sequence of test (i.e. log-likelihood ratio test) to remove or keep variables in the models based on a predefined nominal significance level for variable exclusion. In the forward selection approach, the model is built-up in steps from the best candidate predictors. This approach however does not provide a simultaneous assessment of the effects of all the candidates in the models (Verderio et al. 2016, 1-4; Moons et al. 2012, 683-690). For PMLE, a reduced model can be obtained using the R-square method (Verderio et al. 2016, 1-4; Moons et al. 2004, 1262-1270) because the standard backward selection procedure based on ML estimations cannot be utilized. The R-square method uses the full PMLE model to estimate the linear predictor per patient. Ordinal least square regression is then used to relate all predictors to the linear predictor, obviously leading to an R^2 of 1. Then, all predictors are step wisely deleted and the R^2 after each step is estimated. If R^2 remains close to 1 after the removal of a predictor, it means that the predictor did not contribute to the prediction of the

outcome and can thus be excluded. An R^2 of 0.95 could be used as threshold to decide which predictor deleting from the full PMLE model (Moons et al. 2004, 1262-1270).

3.5.2.3 All subsets regression

As already mentioned, for clinical purposes it is important that a predictive model is based on a small number of covariates, and this is linked to the concept of parsimony from a statistical point of view. In most cases a large number of candidate variables are available and it is important to (try to) select only the most important ones. An alternative approach to the standard stepwise/backward methods, is the *all subsets regression*, which can discover combinations of variables that explain more variation in patients outcome than those obtained by using the standard stepwise/backward algorithms (Verderio et al. 2016, 1-4; Altman and Royston 2000, 453-473). This approach has several potential advantage, but also drawbacks including the possibility of selecting models that omit important predictors. In a scenario like that reported in this thesis this approach could represent a suitable option. Briefly, this implies to estimate all the possible models starting from the candidate variables.

3.5.3 Assessing the predictive performance of the model

Once developed, the performance of the developed model should be assessed by evaluating discrimination and calibration. Discrimination refers to the ability of the model to distinguish individuals with the disease from those without the disease; the c-index or the equivalent area under the ROC curve are the most widely adopted statistics indexes. Calibration refers to the agreement between the probability of developing/having the outcome of interest as estimated by the model, and the observed outcome. It is usually graphically assessed, by plotting the observed outcomes frequencies *vs* the mean predicted outcome probability (or risk), within subgroups of patients that are ranked by increasing estimated probability. Formal statistics for goodness of fit, such as Hosmer and Lemeshow test can be used for this purpose (Moons et al. 2012, 683-690; Moons et al. 2012, 691-698). In addition it should be considered that the performance (i.e. AUC values) of the developed model could be too optimistic because the same data are used for developing and testing the model. Accordingly, proper statistical tools should be adopted to both obtain bias-corrected estimates (i.e. internal validation) and validate the model (i.e. external validation).

3.5.4 Model validation

Validation of a predictive model traditionally refers to assessing the performance of the model in subjects other than those used for model development. When applied to new subjects, the performance of the model is generally lower than that observed in the sample on which the model was developed. Therefore, the performance of a developed model should be accurately evaluated in new individuals before its implementation and application in clinical practice. Two different types of validation can be adopted, according to the design of the study and the data available: internal and external validation.

In biomarker research, the term *model validation* could mean different things as reported by Taylor et al (Taylor, Ankerst, and Andridge 2008, 5977-5983). Altman and Royston (Altman and Royston 2000, 453-473) identified two types of validated models: clinically or statistically validated. In both types, an evaluation of the performance of the model is performed, but only in the statistically validated ones the aspects related to the model's goodness-of-fit are considered. According to the proposed definition of Altman and Royston "a statistically validated model is one which passes all the appropriate statistical checks, including goodness-of-fit on the original dataset and unbiased prediction on a new data set", whereas "a clinically validated model is one which performs satisfactory on a new data set according to context-dependent statistical criteria laid down for it" (p. 456). To note that a statistically validated model may be clinically invalid (i.e. weak predictive/prognostic information) as well as a clinically validated one can be statistically invalid (i.e. bias of the predictors, no fit) and that it is more difficult to obtain a statistically validated model than a clinically validated one, due to the issues related to over-optimism and bias at the model-building stage (Altman and Royston 2000, 453-473).

3.5.4.1 Internal validation

To estimate the potential optimism on the model performance, internal validation techniques can be adopted, when no other data than the study sample are used. A first approach consists in splitting the study data in a training set, used for building the model, and a testing set, to test the model's performance. The major advantage of this approach is its convenience, as it does not require the collection of additional data; on the other hand the data-splitting can reduce efficiency because not all the data are used to build the model. An extension of the training-testing splitting is to repeat the splitting of the data a large number of times: leave-one-out, k-fold and repeated random split cross-validation can be used for this purpose. In the leave-one-out

cross validation approach, each observation is used as testing set with the remaining data used as training set, so that there are as many training-testing splits as the number of observations. In the k-fold cross-validation, data are divided into k-subsets with each subset serving as the testing set for the remaining k-1 subsets pooled together. In the last procedure, the splitting of the data is randomly repeated many times. With these procedures, the model is refit to each training set and evaluated on the corresponding testing set; validation results are then reported as the average performance over all the testing sets. The cross-validation splitting is more efficient than a single training-testing split, but on the other hand, it is cumbersome and may be impossible to implement. An alternative approach is the bootstrapping especially when the development sample is relatively small and/or a large number of candidates predictors is studied (Moons et al. 2012, 683-690; Taylor, Ankerst, and Andridge 2008, 5977-5983). Briefly, bootstrap is a statistical method that aims at mimic the sampling process using only the available data, by sampling with replacement from the original samples. This means that each bootstrap set is similar but not identical to the original study sample (same sample size of the original dataset). In each bootstrap sample the data are analyzed as in the original study sample yielding potentially different models (and obviously different c-indexes) from each bootstrap sample. Then each bootstrap model is applied to the original study data, yielding a different c-index. The average of all these c-index indicates the optimism in the apparent c-index of the prediction model that was at first developed in the original study sample (Moons et al. 2012, 683-690). The bootstrapping procedure therefore allows the use of all the data for model development and provides information about the level of model overfitting and optimism as well as reflects what can be expected when the model is applied to new individuals from the same theoretical source population. Even if these internal validation methods can correctly control overfitting and optimisms, they cannot substitute external validation (Moons et al. 2012, 683-690).

3.5.4.2 External validation

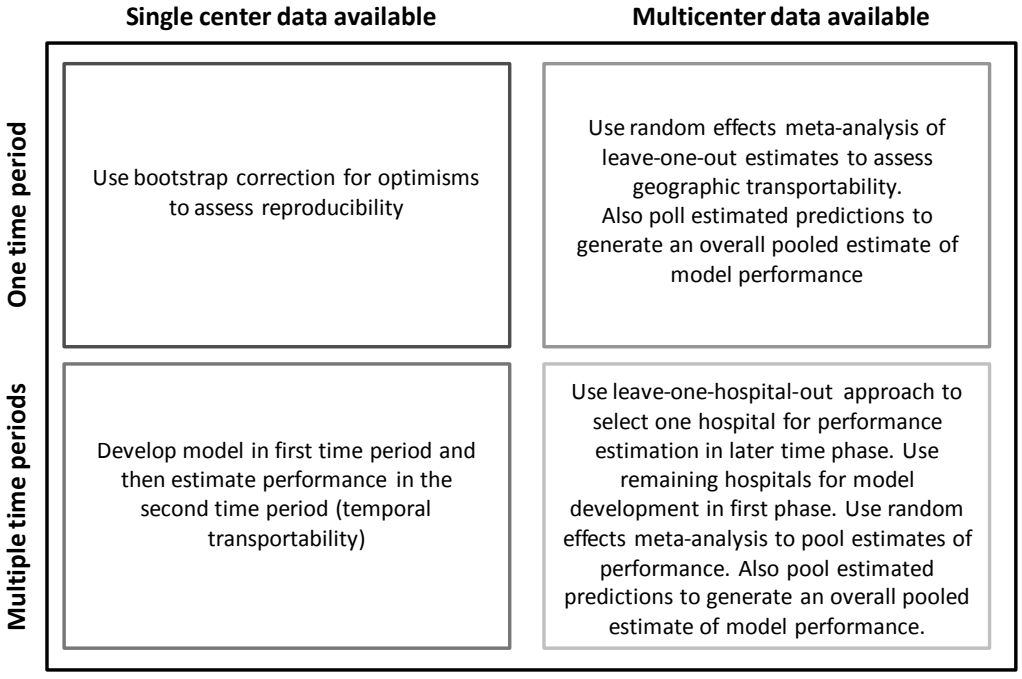
The performance of a developed and internally validated model should still be tested or validated in new individuals before its implementation and application in clinical practice. Validation on heterogeneous external data allows for an evaluation of the generalizability of the prediction tool. The objective is to apply the previous developed model to new individuals, whose data were not used in the model development, and to quantify the model's predictive performance. This means taking the original model with its predictors and relative weights (i.e.

linear predictor), applying this original model to new data and quantifying the model's predictive performance (i.e. statistical validation), without any estimation of the parameters included in the model (Moons et al. 2012, 683-690).

The new set of individuals may come from the same institution but be recruited in a different period (temporal validation), usually later than those included in the model development. They should in any case share the same inclusion/exclusion criteria as well as the same outcome definition and measurement method. However, the temporal validation cannot examine the generalizability of the model to other institutes of countries, which is the case of geographical validation. The latter can be also utilized when there are differences in the inclusion/exclusion criteria or in the outcome definition as compared with the development study. The geographical validation allows the evaluation of both the transportability and generalizability of the developed predictive model to other contexts similar to that of the development study; it may be done retrospectively - by using existing datasets of other institutes/countries - or prospectively by recruiting new individuals in a specifically designed study (Moons et al. 2012, 683-690).

Figure 14 graphically depicts the options that could be adopted for validating a prediction model with details about the different statistical procedures that can be applied according to the scheme design of the research (Austin et al. 2016, 76-85).

Figure 14. Recommendations for validating clinical prediction models



Adaptation from (Austin et al. 2016, 76-85)

3.5.5 Model updating and extension

When a poorer performance of a prediction model is obtained on new individuals compared to that observed in the development study Researchers tend to reject the model, and develop or fit a new one, sometimes by repeating the entire selection of predictors (Moons et al. 2012, 691-698). A valid alternative could be to update the existing model by adjusting or recalibrating the model to the local circumstance or setting of the validation sample at hand. In this way the updated model combines the information captured in the original model of the development dataset with information from new individuals, theoretically improving the transportability to other individuals (Vergouwe et al. 2016)

Often the differences are seen in the outcome frequencies between the development and the validation samples, which implies having predictive probabilities systematically too high or low with respect to the development samples. In this case, the adjustment by the baseline risk of the original model to individuals of the validation can improve the calibration (i.e. recalibration in large). This means fitting on the new individuals a logistic regression model including the linear predictor as an offset variable (i.e. relative weight or slope fixed to one) and the intercept as the only free parameter that should be estimated. An extension of the above method is the fitting of a logistic regression model in which both the intercept and the slope of the linear predictor covariate should be estimated (i.e. recalibration). A third method consists in updating the model by fitting the original model anew; in this case all the coefficients of the covariates included in the model (and the intercept) should be estimated (i.e. model revision). These methods also allow an improvement of the discrimination. Obviously the performance of the updated model should be tested before it can be applied in routine practice (Vergouwe et al. 2016).

To improve the performance of a clinical prediction model, new marker(s) can be incorporated in the existing model. Similarly to the updating strategies, different extending approaches can be adopted for this purpose varying from the re-estimation of all the regression coefficients on the validation samples including the new marker(s) to the Bayesian approaches or imputation (Nieboer et al. 2016, 128). Here I will focused on the re-estimation of all the regression coefficients on the validation samples including the new marker(s). A first way is to fit a logistic regression model including all the predictors considered in the original model together with the new marker(s) (i.e. model revision with extension): this means to estimate the *p+1* parameters of the original model (including intercept) plus that of the new marker. Again, if a small dataset is available the tendency to overfit can occur. A second way is to perform a model extension with

shrinkage, by shrinking the regression coefficient of the new marker to zero. A third strategy is to include the new marker in the recalibrated model (i.e. recalibration with extension). Regression coefficients are obtained by fitting a logistic regression model with the linear predictor of the original model and the new marker as predictors (Nieboer et al. 2016, 128).

The performance of an extended prediction model can be evaluated through calibration and discrimination or according to other recently developed measures that assess the added value of the new marker, such as the net reclassification index (NRI), net benefit and relative utility.

CHAPTER 4. RESULTS

4.1 THE CRC-INT STUDY

Within the AIRC 5x1000 special program, a specific task is focused on the colorectal cancer disease. The task aims at identifying plasma circulating miRNAs to be used as biomarkers for the early detection of CRC lesions in FIT positive individuals. The rationale of this study mainly starts from results obtained in two studies (Zanutto et al. 2014, 1001-1007; Reid et al. 2012, 504-515) performed at Fondazione IRCCS Istituto Nazionale dei Tumori Milano (INT) by our research group and from the hypothesized greater diagnostic performance of the genetic biomarkers compared to the currently available faecal occult blood tests.

4.1.1 Previous results

In the paper of Reid JF and Colleagues (Reid et al. 2012, 504-515), tissue specimens from 40 fresh-frozen consecutive sporadic CRCs matched with their adjacent normal mucosa were obtained from previously untreated patients lacking family history and high frequency microsatellite instability (MSI) who underwent surgical resection at the INT between 2001 and 2009. TaqMan® Array MicroRNA Cards v.2 (card A and card B, Applied Biosystem, Foster City, CA) were used to profile the CRC cohort. Each miRNA was normalized by the ΔΔCt method (Livak and Schmittgen 2001, 402-408) using the U6 for data normalization. The simultaneous 95% confidence intervals (95% SCI) of the relative quantity (RQ, $RQ = 2^{-\Delta\Delta Ct}$) of each miRNA was applied (Pizzamiglio et al. 2010, 853-860). This approach allows to consider the simultaneous determination of many miRNAs, adjusting for multiple comparisons. According to the 2-fold threshold rule ($RQ \leq 0.5$ or $RQ \geq 2$), each miRNA was considered down-regulated if the upper limit of its 95% SCI for RQ was ≤ 0.5 or up-regulated if the lower limit of its 95% SCI was ≥ 2. Results from the SCI approach identified a set of 23 miRNAs.

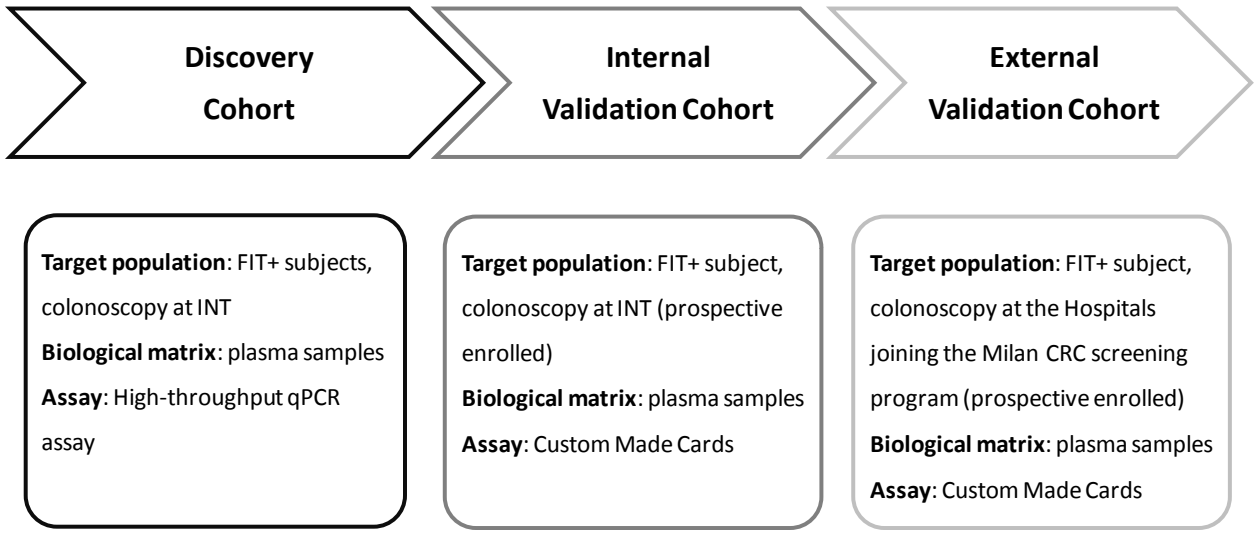
This set of miRNAs was then evaluated on a series of 65 CRC patients and 70 donors, using plasma samples as starting material. Briefly, plasma samples from patients before surgery and that of donors were collected between 2009 and 2011 at our Institute. TaqMan MicroRNA Assays were used to profile the considered cohort. Each miRNA was normalized by the ΔCt method using miR-16. The association between each miRNA (in terms of – ΔCt) and the risk of colon cancer was

evaluated by resorting to a univariate logistic regression model and by estimating the area under the ROC curve together with its 95% CI. In addition, for 46 CRC patients both the plasma of the day before surgery (T0) and that of 4–6 months after surgery (T1) were available. On this subset, the RQ value of the validated miRNAs at T0 and T1 was compared using a mixed regression model. Results from the validation set showed that 2 miRNAs (miR-21 and miR-378) were statistically associated with the disease status and able to discriminate between cases and controls. In addition, the expression of miR-378 significantly decreased after the CRC removal, suggesting a possible role of this miRNA for monitoring CRC recurrence (Zanutto et al. 2014, 1001-1007).

4.1.2 CRC-INT study: overview

The CRC-INT study enrols subjects who, after a positive FIT test, underwent a screening colonoscopy to assess the presence of CRC lesions. Each subject who signed the informed consent, before colonoscopy, was enrolled in the study. Blood collection was performed before colonoscopy and circulating miRNAs, extracted from plasma, were analyzed by using PCR assays. Dedicated databases were built to collect information related to: (i) pre-analytical workflow, (ii) screening career of each enrolled subjects by taking advantage of the Screening Program Information System of the Local Health Authority (LHA) of Milan (demographical and clinical-pathological variables), (iii) haemolysis levels of each collected sample and (iv) experimental data. Figure 15 schematizes the CRC-INT study design.

Figure 15. CRC-INT study design



The principal aim of the discovery phase was to investigate the suitability of searching miRNAs in plasma from FIT+ individuals as well as identify a set of reference and candidate miRNAs to be deeply investigated in the subsequent phases. Accordingly we adopted a high-throughput assays for discovery (TaqMan Low Density Array -TLDA, Megaplex format- MicroRNA Card by Applied Biosystem). This technology is one of the most common platform used for miRNAs profiling, as it makes possible to run hundreds even thousands PCR reactions in parallel with the same starting sample with a high enough precision (Perkins et al. 2012, 296-2164-13-296) facilitating the high-throughput profiling of miRNA expression (Deo, Carlsson, and Lindlof 2011, 795-812). The TLDA (Arrays A and B) were configured in a 384-well format microRNA assays (381 unique miRNAs and one control RNA in replicate) spotted onto a micro fluidic card; the eight sample-loading ports, each connected by a micro channel to 48 miniature reaction chambers, allow the evaluation of a total of one sample per card. For the following phases specific probes, such as custom-made cards were used. In this study, we adopted a two-phase validation: in the Internal Validation the principal aim was to generate miRNAs-based signatures using a prospective cohort of FIT+ subjects, whereas in the External Validation phase the aim was to evaluate the performance of the generated signatures in a multi-centre setting (i.e. generalization of the results). This approach was adopted also by taking into consideration the timelines related to the project and the accrual rate of FIT+ subjects in INT and in the other involved Institutes.

In the following section, results of this study that have not been already published are reported “in blind” for copyright reason. Thus, the number and name of candidate as reference miRNAs as well as those of the models developed are reported “in blind”.

Notation:

N: number of subjects enrolled in the study

R: candidate reference miRNAs

W: reference miRNAs

C: candidate miRNAs

d[i] (i=1, 2, ..., C) : ID used to identify the miRNA

m[j] (j=1,2, ..., M): name of the model developed

4.1.3 Discovery phase

During this phase, we evaluated the expression levels of human miRNAs on a cohort of already available (at the start of the study) plasma samples from FIT+ individuals who have performed a screening colonoscopy at INT. A total of N_D subjects (36.67% showing no proliferative lesions and 63.33% with lesions) where considered in this phase. For all these subjects, endoscopic lesions were classified according to the histological classification criteria adopted by the LHA of Milan in the CRC screening program: negative or no proliferative lesions (No lesions, NL), Low grade Adenoma (LgA), High grade Adenoma (HgA), cancerized adenoma or cancer (cancerous lesion, CL).

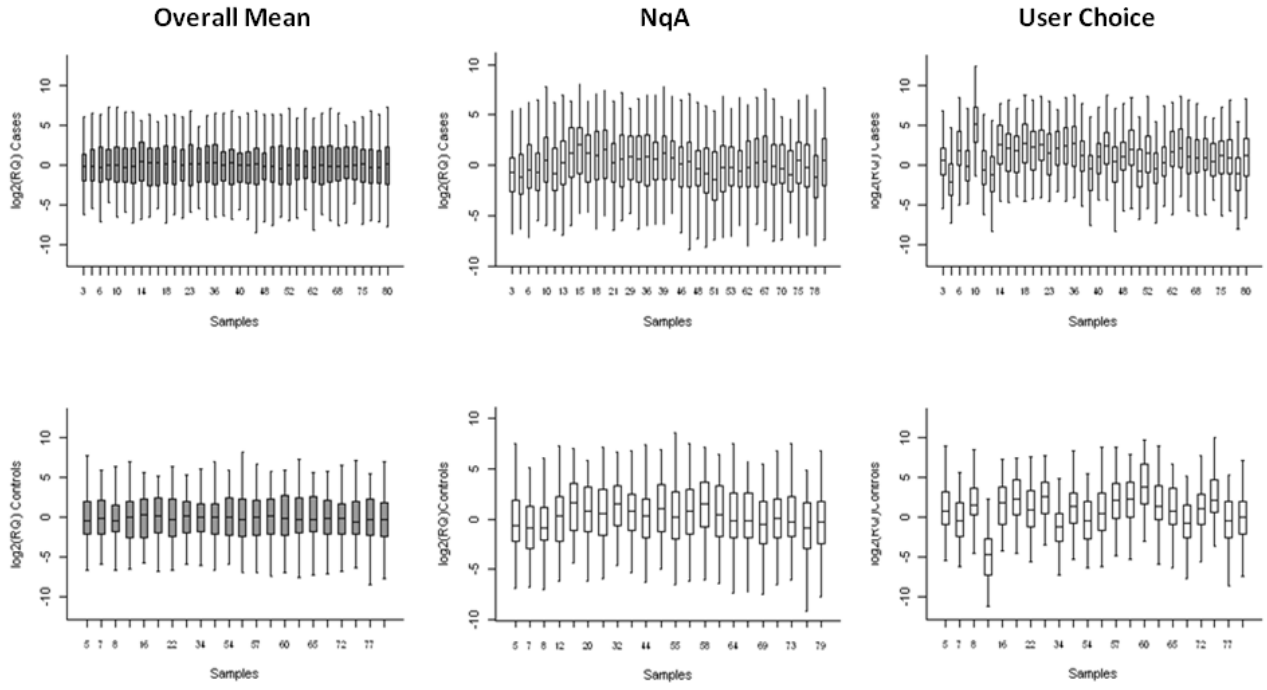
4.1.3.1 Data normalization

Data of the discovery cohort were processed with the aim of identifying (i) a set of candidate miRNAs, whose expression differs between subjects with proliferative lesions and subjects without lesions and (ii) a subset of reference miRNAs to be used in the subsequent validation steps for data normalization.

By running the NqA R-function, data were firstly normalized according to the overall mean (MO) (Mestdagh et al. 2009, R64-2009-10-6-r64. Epub 2009 Jun 16) and a set of C_{MO} candidates was identified. Among miRNAs expressed in all subjects, 4.98% of them showed a CV value ≤ 20^{th} centile of the CV's distribution (R_{CV}). Among the highly correlated miRNAs (12.5% of R_{CV} miRNAs), miRNAs showing the highest CV value were excluded as well as those resulted significantly associated with the disease status, using the OM as normalization method. miRNAs passing these steps were evaluated through both geNorm and NormFinder and ranked according to their stability values. By performing all the forward combination of these candidate reference miRNAs, we identified a first set of R candidate reference miRNAs that resembles the results obtained with overall mean. In addition, to further reduce the number of candidate reference miRNAs to be used in the validation phases, the developed procedure was updated in order to estimate all the possible combinations by starting from these R candidate reference miRNAs. A final subset of W reference miRNAs was identified as the one showing a satisfactory agreement with that obtained when considering the overall mean. Figure 16 reports the $\log_2$(RQ) distributions in cases and controls by using different data normalization strategies: overall mean method, the best set of reference miRNAs identified by NqA and according to reference miRNA(s) chosen by the user. Each boxplot reports the $\log_2$(RQ) distribution of the considered miRNAs for each considered subjects.

By looking at the log_2(RQ) distributions obtained using the overall mean method, it is possible to observe a homogeneity of the distribution within cases (or controls) indicating the appropriateness of the normalization method. The final NqA set of reference miRNAs showed a similar pattern suggesting also in this case the use of a proper method of normalization. On the contrary, the results obtained by the MammU6 (user choice) are more heterogeneous suggestive of a not fully removal of all the experimentally-induced variations.

Figure 16. NqA output – reference miRNAs selection



4.1.3.2 Identification of candidate miRNAs

By using the W identified miRNAs as reference for data normalization, a total of C non-redundant miRNAs (candidate miRNAs) were identified as showing an expression significantly different in subjects with proliferative lesions *vs* subjects without lesions or in a specific proliferative lesions *vs* subjects without lesions. Based on these results, a custom microfluidic card including the candidate miRNAs and the W reference miRNAs was designed to be used in the following Internal validation phase. Additional miRNAs that deserved further investigation have been also included in this custom-made card.

4.1.4 Technical Validation

Before moving to the custom-made assay, we performed a technical validation phase aimed at evaluating the level of reproducibility between the involved assays (TLDA megaplex format vs custom-made ones) (Verderio et al. 2015, e258-61). Megaplex TLDA, allowing the high-throughput profiling of several miRNAs, is a suitable tool for discovery purpose where the intrinsic low precision (absence of replicates) and specificity (multiple test) are balanced by the opportunity of performing a large scale screening for selecting promising miRNAs to be further investigated. In contrast customized assays ideally offer the possibility to increase the level of both precision and specificity as they are focused on a limited number of selected miRNAs assessed with replicates. To this end we analysed the samples of the discovery phase, previously analyzed with the Megaplex cards, also with the custom made ones. The association of the miRNAs differentially expressed between subjects with proliferative lesions *vs* subjects without lesions was assessed on the custom made cards by means of the Kruskal-Wallis test. The 86% of miRNAs investigated, resulted differentially expressed between cases and controls in both the Megaplex and custom-made cards. Moreover, by looking at the $\log_2$RQ values obtained with the Megaplex and custom-made cards, we observed a Concordance Correlation Coefficient (Lin 1989, 255-268; Marubini, Pizzamiglio, and Verderio 2005, 73-78) equal to 0.69 (95%CI: 0.12-0.92). By looking at the lower limits of the 95% confidence interval of CCC the two methods did not reach a fully satisfactory agreement and the wideness of this interval suggests the existence of a high variability between the two sets of compared data. This can be due to differences in both the pre-analytical (e.i. RT reaction mix, pre-amplification reaction mix) and analytical steps (PCR platform setting, number of replicates) that could influence the results. These results were discussed with the provider of the assays and quality controls steps were jointly developed.

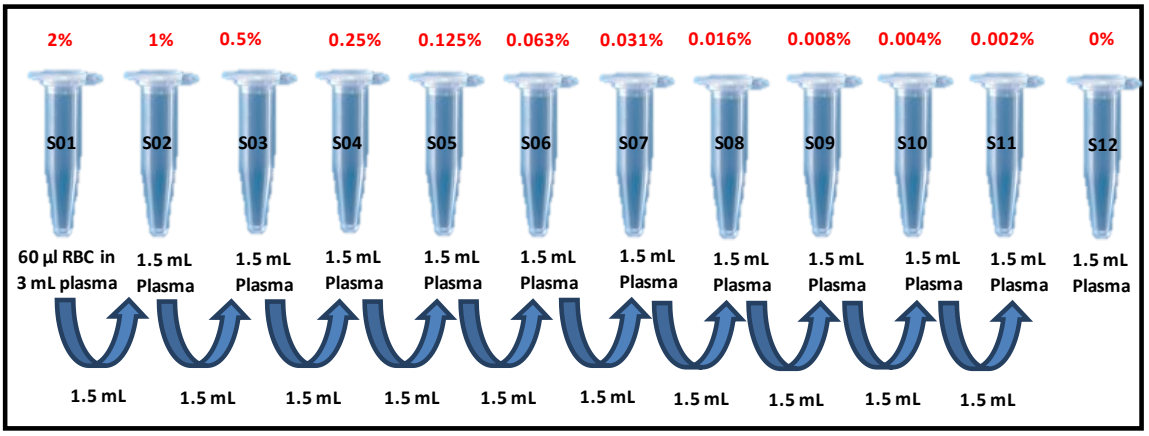
4.1.5 In-vitro controlled haemolysis experiment

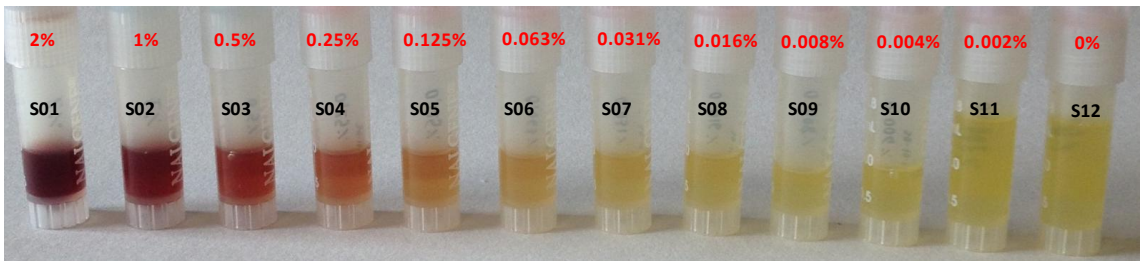
4.1.5.1 Scheme design

In order to investigate if our candidates miRNA were influenced by the haemolysis we designed and implemented an *ad-hoc* in-vitro controlled haemolysis experiment by artificially introducing different percentages of red blood cells (% RBCs) in a haemolysis-free plasma sample (Pizzamiglio et al. 2017, 315-320). Briefly, haemolysis was artificially introduced in the plasma sample by adding RBCs starting from a 2% concentration and by performing ten 1:2 serial dilutions (range: 0.002% - 1% v/v) for a total of 12 tubes (S01-S12), including the not-contaminated plasma

(0%, S12) sample. Figure 17 reports the scheme and picture of the in-vitro controlled haemolysis experiment. The level of haemolysis of each tube was computed according to the haemolysis indexes already reported in literature, such as the absorbance peak at 414 nm (Kirschner et al. 2011, e24145), the haemolysis ratio (Zanutto et al. 2014, 1001-1007), the H-score (Appierto et al. 2014, 1215-1226) and the haemoglobin concentration with the Harboe method (MacLellan et al. 2014, 27-6890-14-27. eCollection 2014). In addition, RNA was extracted from each tube (S01-S12) and qPCR was done using miRNA-specific assays. Two independent RNA extractions were performed and qPCR was performed in duplicate. Thus, for each tube both the haemolysis value and the miRNAs expression level were available.

Figure 17. Scheme and picture of the in-vitro controlled haemolysis experiment





(Pizzamiglio et al. 2007, 232-236)

4.1.5.2 Statistical analysis

The ΔCt values were used as pivotal variable to evaluate the influence of haemolysis on each miRNA of interest. Specifically, for each s-th tube (s=S1,.., S12), for each j-th replicate (j=1,2) and for each extraction (i=1,2) the ΔCt was obtained as follows: $\Delta Ct_{ijs} = Ct_{ijs} - Ct_{ijs}REF$, where $Ct_{ijs}REF$ is the Ct average of the W reference miRNAs identified in the discovery phase. Then, for

each miRNA, the Relative Quantity (RQ) was computed by subtracting the ΔCt value of each tube to the 0% tube (S12) as follow:

$RQ_{ijs}=2^{(-\Delta\Delta Ctijs)}$, where $\Delta\Delta Ct_{ijs} = \Delta Ct_{ijs}–\Delta Ct_{ij}S12$. To consider the simultaneous determination of different miRNAs on the same sample and to evaluate the relevance of the change of expression with respect to the not contaminated plasma sample (S12), we computed the 95% simultaneous confidence interval (95%SCI) of the $\log_2(RQ_{ijs})$ for each miRNA according to the bootstrap percentile method. Following the conventional two-fold threshold rule, we considered the low levels and the high levels of a specific miRNA due to RBC contamination statistically relevant if the upper limit of the 95% SCI of $\log_2(RQ)$ was ≤ -1 or the lower limit ≥ 1, respectively.

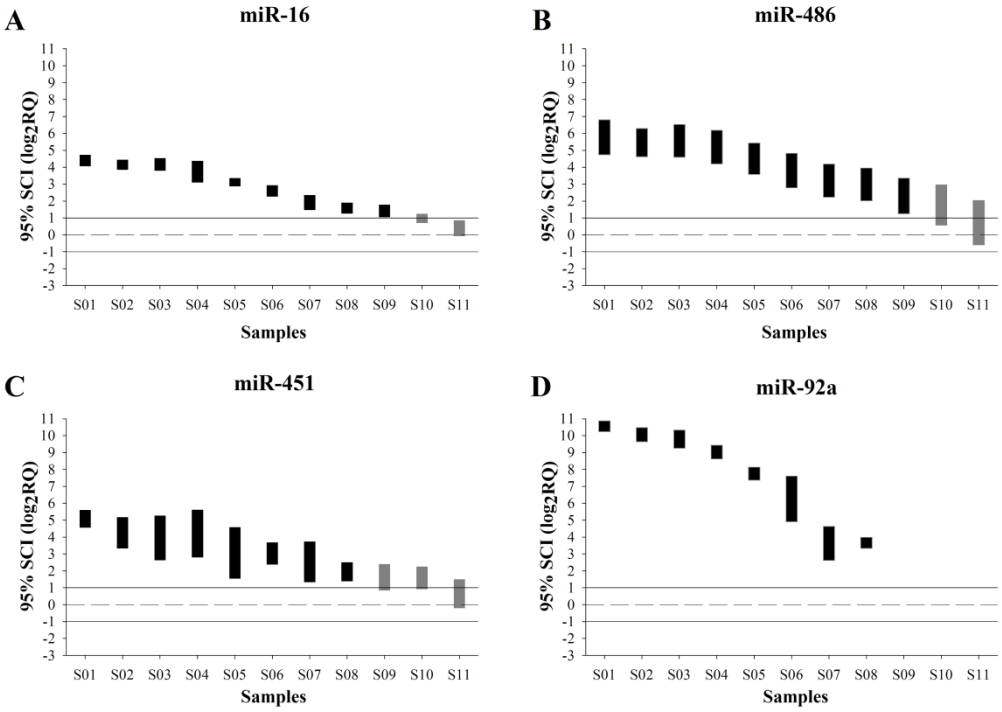
In addition, by appropriately fitting the values of the considered haemolysis indexes (computed by starting from the absorbance values) as function of the known induced percentage of RBCs of the 10 serial dilutions (samples S02-S11), a calibration haemolysis curve was generated. Finally, by applying the inverse regression method (Verderio et al. 2004, 76-79; Pizzamiglio et al. 2007, 232-236), it was possible to estimate the unknown percentage of RBCs in plasma samples collected within the study (Internal and External Validation Cohorts).

4.1.5.3 Results: miRNA expression levels vs haemolysis

According to the two-fold rule we observed that miRNAs known as haemolysis-related in literature (miR-16, miR-486, miR-451 and miR-92a) were confirmed also in our experiment as influenced by haemolysis, whereas our reference miRNAs were not influenced by haemolysis. As concerns our candidate miRNAs, some of them showed relevant changes with respect to S12, starting from S06. Accordingly, these miRNAs were not considered in the subsequent statistical analysis of the Internal Validation Phase.

In particular, as shown in Figure 18, the lower limit of the $\log_2(RQ)$ 95% SCI was above the threshold of 1 starting from sample S09 for miR-16 (Panel A) and miR-486 (Panel A), indicating a relevant increased expression starting from that percentage of RBC compared with the uncontaminated sample (S12). Similarly, for mir-451 (Panel C) the lower limit of the 95% SCI was above the threshold of 1 starting from the sample with a percentage of RBC contamination of 0.016% (S08). Finally, as regards miR-92a, the majority of Ct values were undetected between samples S12 and S09, suggesting a specific association between its expression and the presence of RBCs (Panel D). Accordingly, starting from sample S08 (0.016% contamination with RBCs) the lower limit of the 95% SCI were always above the threshold of 1 (Panel D).

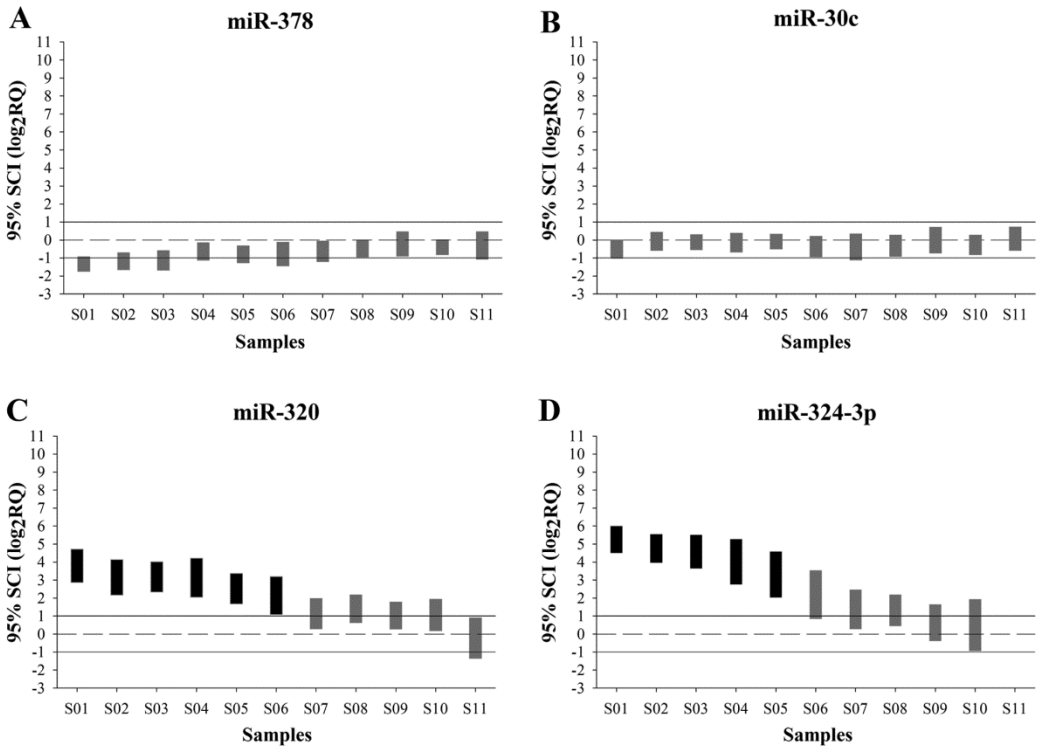
Figure 18. Graphs displaying the 95% SCIs of the log2RQ for the four haemolysis-related miRNAs



(Pizzamiglio et al. 2017, 315-320)

Figure 19 reports the SCI of two miRNAs resulted influenced by haemolysis and two not related to haemolysis. Specifically, for miR-378 and miR-30c no significant changes in expression were observed (Panel A and B). On the contrary, the lower limit of the 95% SCI was above the threshold of 1 starting from samples S06 and S05 for miR-320 and miR-324-3p (Panel C and D), respectively, indicating a relevant increased expression starting from samples with a 0.063% and 0.125% RBCs contamination compared with the un-haemolysed sample (Pizzamiglio et al. 2017, 315-320).

Figure 19. Graphs displaying the 95% SCIs of the log2RQ for the four miRNAs of interest

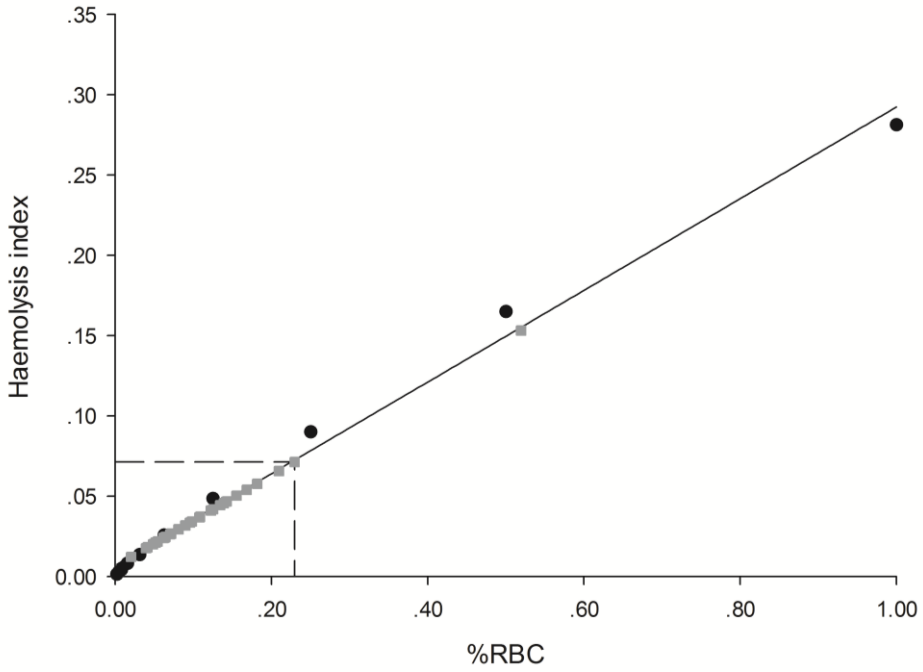


(Pizzamiglio et al. 2017, 315-320)

4.1.5.4 Results: estimation of the unknown RBC contamination in plasma samples

In order to be able to estimate the unknown percentage of RBCs in new plasma samples a calibration curve was generated. This was done by taking advantage of the availability, for each contaminated tube, of both haemolysis indexes and known RBC concentration. Finally, according to the calibration curve, the percentage of RBCs in new plasma samples was estimated by considering the regression parameters and the value of the haemolysis index obtained for each sample, using the inverse regression method (Verderio et al. 2004, 76-79; Pizzamiglio et al. 2007, 232-236). Figure 20 depicts the calibration curve obtained by fitting the Harboe index as function of the known RBC concentration: the gray dots indicate the estimated %RBCs in a subgroup of plasma samples of FIT+ individuals enrolled within our study (for illustrative purpose) and the dashed lines graphically depicts the inverse regression procedure. On the basis of this calibration curve, we estimated the percentage of RBC in all the plasma samples collected within the study.

Figure 20. Calibration curve obtained by fitting the haemolysis index as function of %RBC

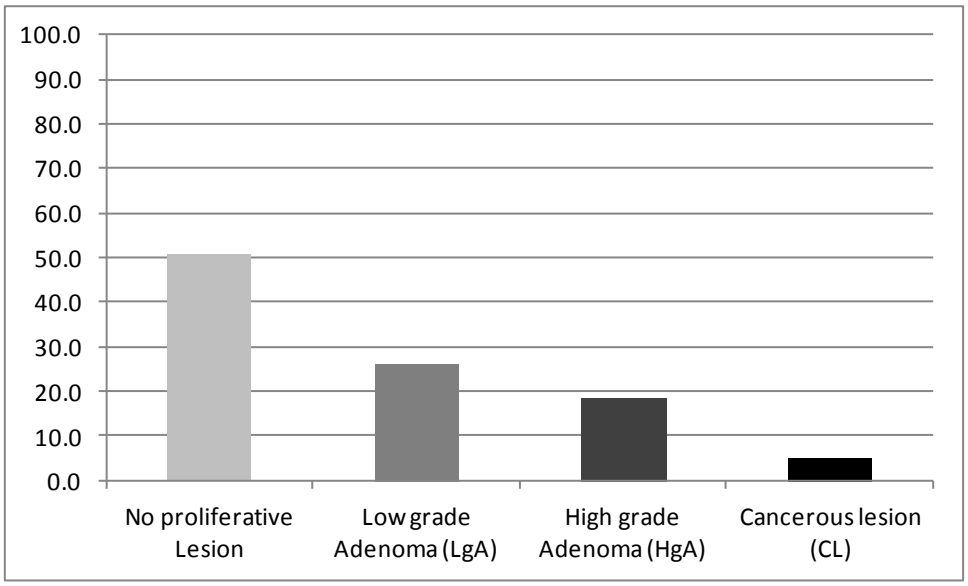


(Pizzamiglio et al. 2017, 315-320)

4.1.6 Internal Validation cohort

As reported in Figure 15 during this phase we evaluated the expression levels of our candidate miRNAs on a cohort of prospectively collected plasma samples from FIT+ individuals who underwent a screening colonoscopy at INT between February 2013 and December 2014 (Internal Validation cohort). Blood samples were collected before colonoscopy and the obtained plasma samples immediately stored at -80°C until RNA extraction. RNAs were profiled using custom made cards (Applied Biosystems), containing our candidates and reference miRNAs as well as additional mRNAs of interest. Figure 21 reports the distribution of the endoscopic lesion observed in the internal validation cohort. These figures are in line with those reported in the CRC screening survey, highlighting a decreasing trend of lesions in FIT+ subjects, moving from subjects without proliferative lesion to subjects with cancerous lesions.

Figure 21. Results of the screening colonoscopy in FIT+ subjects of the Internal Validation Cohort (IVC)



4.1.6.1 Signature building: overview

By starting from our candidate miRNAs, penalized logistic regression models (on multivariate fashion) (Harrell 2001) were implemented for each of the considered endoscopic lesions (CL, HgA, LgA). Specifically, the all-subset analysis was performed by starting from the candidate miRNAs in order to identify one or more signatures able to discriminate between subjects with a specific proliferative lesion and subjects without lesions. This strategy, allowing the estimation of multiple multivariate models, starts from the following main considerations: (i) complex research experimental scenario, i.e. searching miRNAs in plasma samples using medium throughput technologies developed for research purpose only, (ii) inability of a single biomarker to reach the desired performance for disease classification and outcome prediction and (iii) availability of multiple methods for biomarkers combination in a more powerful composite score. Next section summarizes the results obtained on the IVC.

4.1.6.2 Signature building: computational aspects

To analyze the data, an ad-hoc R-program was developed with the aim to estimate for each generated model, the AUC values obtained by fitting a (full) standard regression model (EPV ≥ 10) or a (full) PMLE model when EPV was less than 10. The *lrm* function of the rms R-package was used to fit binary logistic regression models using maximum likelihood estimation or penalized

maximum likelihood estimation. The penalty factor within the function, selects the type of estimation: a penalty of zero implies the ordinary maximum likelihood estimations, whereas a penalty different from zero implies the penalized maximum likelihood estimations (https://cran.r-project.org/web/packages/rms/rms.pdf). According to Moons et al. 2004, we chose the optimal penality factor by maximising the modified AIC using the *pentrace* function. As reported in the R package documentation, the optimization algorithm is more likely to find a reasonable solution when the starting value specified in the penalty vector is too large rather than small (http://svitsrv25.epfl.ch/R-doc/library/Design/html/pentrace.html) Accordingly, we identify a vector of penalty ranging from 0 to 50 by 0.1. AUCs and the corresponding 95%CIs were computed used the *ci.auc* functions of the pROC package (https://cran.r-project.org/web/packages/pROC/pROC.pdf). The ci.*auc* function computes the numeric value of AUCs and 95%CIs according to the "delong" or "bootstrap" method. With "delong" method the variance of the AUC is computed as defined by DeLong et al. (DeLong, DeLong, and Clarke-Pearson 1988, 837-845).

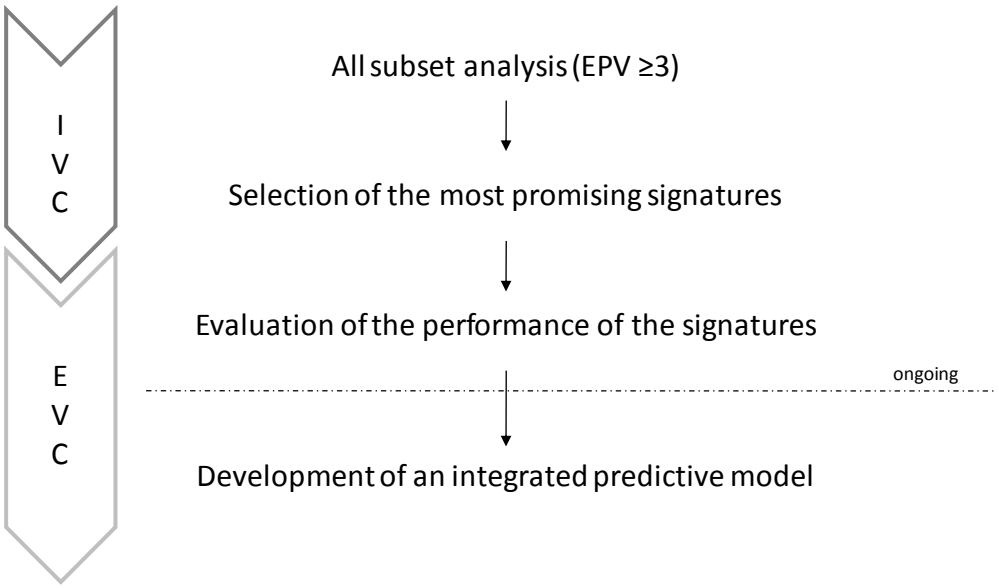
At the end of the process an M x V matrix was generated, with M indicating the number of fitted models and V the variables obtained as output of the fitting (i.e. n. cases, n. controls, EPV value, AUCs values and corresponding 95% CI as well as β estimates from MLE and PLME). Only signatures showing an EPV value ≥ 3 were fitted. This cut-off value was chosen according to simulation studies produced by Pavlou and Colleagues (Pavlou et al. 2016, 1159-1177), in which Authors compared the predictive performance of different methods in two low EPV scenarios (EPV=3 or 5).

Finally, the function *validate* in the same rms package was used to estimate, optimism and the bias-corrected indexes of the most promising models.

4.1.6.3 Signature selection

To identify the most promising signatures that were then evaluated on the External Validation Cohort, an ad-hoc procedure was developed (Figure 22). Briefly, among the possible signatures with an EPV ≥3 with a significant AUC value, the most promising signatures were selected. Signatures sharing these properties were then investigated on the EVC dataset.

Figure 22. Signature selection workflow



4.1.6.4 Signature evaluation: discrimination and calibration

Table 6 reports AUC values and the corresponding 95% Confidence Intervals of seven models (with at least 2 miRNAs) obtained among those selected as promising in the above section. These models were chosen according to the EPV values for descriptive purposes to illustrate some of the concepts reported in the Chapter 3. As above, for patent purpose, signatures are reported as S01, S02,..., S07.

Table 6. Performance of miRNA-based signatures from IVC – original data

outcome	ID model	EPV	AUC	95% CI	
CL	S01	4.0	0.772	0.651	0.892
HgA	S02	12.0	0.643	0.509	0.776
	S03	6.7	0.690	0.553	0.826
	S04	4.0	0.715	0.579	0.850
LgA	S05	16.0	0.643	0.527	0.758
	S06	9.0	0.670	0.547	0.793
	S07	3.0	0.661	0.533	0.788

As mentioned in the previous chapter, the estimation of the AUC values can be optimistic, because the same data of the IVC were used to both fit and evaluate the performance of the model. To estimate the optimism, the *validate* function within the rms package was used to obtain

bias-corrected indexes. Two hundred bootstrap datasets were generated and results are reported in Table 7.

Table 7. Performance of miRNA-based signatures from IVC – bias-corrected

outcome	ID	EPV	Type	AUC original data	AUC training	AUC testing	optimism	AUC corrected†
CL	S01	4.0	L	0.779	0.799	0.769	0.030	0.749
			P	0.772	0.799	0.770	0.030	0.742
HgA	S02	12.0	L	0.643	0.656	0.618	0.038	0.605
	S03	6.7	L	0.691	0.730	0.653	0.050	0.641
			P	0.690	0.702	0.654	0.048	0.642
	S04	4.0	L	0.703	0.745	0.657	0.088	0.615
			P	0.715	0.734	0.660	0.073	0.642
LgA	S05	16.0	L	0.643	0.651	0.632	0.019	0.624
	S06	9.0	L	0.668	0.698	0.653	0.045	0.623
			P	0.670	0.697	0.650	0.047	0.623
	S07	3.0	L	0.712	0.786	0.663	0.123	0.589
			P	0.661	0.731	0.624	0.107	0.554

† obtained by derivation from the Dxy statistics (Dxy= Somer's D) as C (or AUC)=0.5(Dxy+1); L: standard logistic regression model; P= penalized logistic regression model.

The AUC training corresponds to the average of the AUC values across the 200 fitted bootstrap datasets, whereas the testing column gives the mean of the AUC values of the models fitted to the bootstrap datasets when evaluated on the original data; differences between these two columns correspond to the optimism in the discriminatory measurement: the AUC-corrected index is finally obtained by subtracting to the index value computed on the original data the optimism value. In Table 7 I reported the results obtained by fitting both a standard regression model (even if not appropriate) and a penalized one.

As expected (i) the training AUCs are greater that the testing ones and (ii) the bias-corrected values of the AUCs are lower than those obtained on the original dataset, confirming that optimistic performances were obtained if the model is fitted only on the original data, without subsequent validations. These evidences are better marked when we look at the logistic regression results, when this method is used improperly. The highest optimism values (equal to

0.088 and 0.123) were observed of S04 and S07, that showed the lowest EPV values (4 and 3 respectively). By looking at the corresponding penalized values these figures are slightly lower, suggesting that the use of penalized regression is suitable. For S06 similar values of optimism were observed by fitting a standard (L) or a penalized regression model (P): this could be explained by the EPV value close to 10, even if lower optimism was observed when a penalized model was fitted. Model S05 and S02 showed the lowest optimism values (0.019 and 0.038), for which only a standard regression model was fitted.

By looking at the corrected calibration slopes of these models (obtained as output of the *validate* function), it is possible to confirm that, if a standard regression model is applied when EPV <10 , all these models were over-fitted. Values < 1 indicate that the range of observed probability is smaller than the range of predicted probability. In details, for S04 and S07 (models with lowest EPV values) the corrected calibration slopes were 0.56 and 0.46, respectively, whereas for S03 and S06 these quantities were 0.73 and 0.79, closer to the expected value of 1. The corrected calibration slopes moved to 0.84 and to 0.96 for the S02 and S05 standard regression models, respectively.

Particular observations should be done for the S01 model. The optimism derived from both the standard and the penalized models is 0.030 and the corrected calibration slope of the standard logistic regression is 0.89, suggesting an overall positive judgement of the model (in terms of optimism-corrected discrimination and calibration). This is a surprising result that can be possibly explained by the low number of cases (subjects with disease) considered in this cohort. I am speculating that in such a situation both the considered strategies (standard or penalized regression) could not fit appropriately and that the consequent estimation of the optimism suffer from this.

4.1.7 External validation cohort

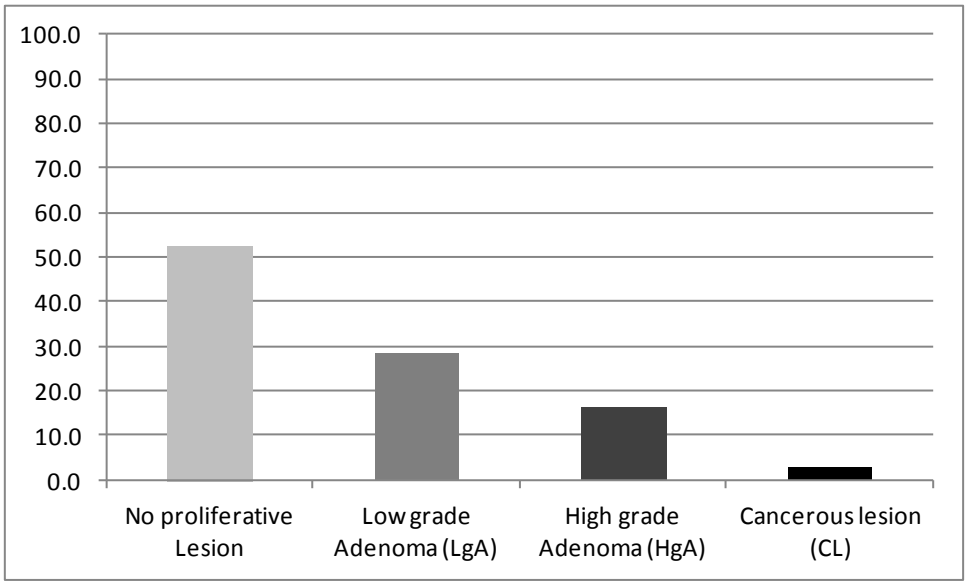
As reported in Figure 15 and Figure 22, during the External Validation phase we evaluated the expression levels of the haemolysis-free candidate miRNAs on a cohort of prospective plasma samples from FIT+ individuals who underwent a screening colonoscopy in one of the Hospitals joining the CRC screening program of the LHA of Milan. Among the 9 LHA-Hospitals (excluding INT), eight agreed to participate to the external validation of the study. The study was approved by the Ethical Committee of the LHA of Milan as well as by that of each Hospital. A study protocol was developed to describe the study and the sample handling and processing procedures to be

applied. The enrolment started in November 2013 and terminated on December 2015: overall more than 1400 subjects agreed to participate to the study.

Each participating Hospital performed the plasma separation and then stored the aliquots until shipment to INT. For each enrolled subject, a small aliquot of plasma was also prepared for haemolysis evaluation. Plasma samples were stored at -80°C until RNA extraction. Each sample was finally evaluated for haemolysis at INT, before RNA extraction. To allow the retrieval of all the clinical data from the Screening Program Information System of the LHA of Milan, dedicated registries were developed and filled-in by each participating Hospital.

A dedicated data cleaning process was implemented in order to identify those subjects for which: (i) plasma samples were processed according to the developed operative procedures, (ii) all the clinical-pathological data were available. Thus, only subjects with all these information (N_{EVC}) were considered for the External Validation Phase. An ad-hoc randomization was performed taking into account the type of lesion and the participating Hospital. RNAs were profiled using custom made cards (Applied Biosystems), containing our candidates and reference miRNAs. Figure 23 reports the distribution of the endoscopic lesions of the analyzed cohort of subjects.

Figure 23. Results of the screening colonoscopy in FIT+ subjects of the External Validation Cohort (EVC)



4.1.7.1 Signature confirmation

The selected signatures from IVC (Figure 22) were tested on the samples of the EVC. Results are preliminary as the statistical analysis of the data is still ongoing: accordingly I here report only a brief summary of the main results (in terms of model revision) on the EVC data available at this moment. For CL outcome, 37% of the promising signatures confirmed their predictive capability on the EVC; for HgA and LgA the majority (≥ 95%) of the selected promising signatures resulted confirmed on the EVC. Table 8 reports the AUC values of the selected S01,..., S07 models obtained on the EVC data.

As reported in Table 8, the AUC point estimates of these models on the EVC were lower than that obtained on the IVC, also after optimism correction (bootstrap). These results are expected and again support the need of including a validation phase on which to finally evaluate the performance of the model. In particular, S02 showed the lowest difference in terms of AUC between the IVC (bias-corrected) and the EVC, whereas S01 showed the maximum one. This is in line with the unexpected results obtained in the IVC when only an optimism of 0.030 was observed for this signature.

Table 8. AUC values of the selected models on the EVC

		IVC		**EVC**
outcome	**ID model**	**AUC (95%CI)**	**AUC corrected**	**AUC (95%CI)**
CL	S01	0.772 (0.651; 0.892)	0. 742	0.604 (0.504; 0.704)
HgA	S02	0.643 (0.509; 0.776)	0.605	0.588 (0.541; 0.635)
	S03	0.690 (0.553; 0.826)	0.642	0.615 (0.559; 0.670)
	S04	0.715 (0.579; 0.850)	0.642	0.616 (0.560; 0.673)
LgA	S05	0.643 (0.527; 0.758)	0.623	0.550 (0.510; 0.590)
	S06	0.670 (0.547; 0.793)	0.623	0.561 (0.514; 0.608)
	S07	0.661 (0.533; 0.788)	0.554	0.597 (0.552; 0.642)

Further analysis will be performed to evaluate a possible effect of the Hospital and to properly apply the different model validation techniques and the other model updating approaches as well as to evaluate the added value of miRNA-based signatures to pre-existing prediction models.

CHAPTER 5. DISCUSSION

The identification of molecular biomarkers that can be detected with non-invasive techniques has been attracting a huge interest in the research community, especially for microRNAs. Many studies have in fact highlighted relationships between the expression of specific miRNAs and the presence of different types of cancers as well as of other disease, such the cardiovascular ones.

However, results from different studies also within the same cancer type provided discordant or non-overlapping results raising some concerns about their clinical usefulness and applicability. Differences in both the pre-analytical and in the analytical phase can probably explain some of these inconsistencies, and emphasize the need of shared Operating Procedures for the entire workflow.

As regards the pre-analytical phase differences could be related to different body fluids (plasma/serum and corresponding anticoagulant) as well as to the preparation/processing protocols, while no substantial differences were seen between fresh and frozen fluids (Tiberio et al. 2015, 731479). A well recognized factor that could however influence the miRNAs expression is the haemolysis, that can occur during blood collection and sample preparation. Some of the miRNAs most influenced by haemolysis (miR-16, miR-451, miR-486 and miR-92a) were also identified as potential cancer-related biomarkers, generating concerns about their effective role as biomarkers. Following these evidences, different strategies were reported in literature to reduce as much as possible the effect of haemolysis in miRNA-bases studies by providing measurements able to recognize the haemolysed samples. Spectrophotometric measurements of single absorbance peaks or a joint combination of them allow a reliable evaluation of the RBC contamination to be used instead of using the visual evaluation alone. Once identified, the choice regards the inclusion/exclusion of these samples in the subsequent analysis or the exclusion from the list of candidate biomarkers of the miRNAs showing a relationship with the haemolysis. The exclusion of haemolysed samples from the analysis can be critical, especially if these samples come from patients with the disease of interest. An alternative option could be to correct (or adjust) for the degree of haemolysis or to set-up studies aimed at evaluating the effect of haemolysis on the miRNAs under investigation (Yamada et al. 2014, e112481; Tiberio et al. 2015,

731479). Within the CRC-INT study, we implemented an in-vitro controlled experiment aimed at identifying those candidate that suffer more from haemolysis in order to select those to be further investigated in the subsequent phases. Our results confirmed a significant different expression (between the RBC-contaminated tubes with respect to the uncontaminated one) for the four miRNAs known as haemolysis related (miR-16, miR-451, miR-486 and miR-92a), whereas all our references and the majority of our candidate miRNAs did not show a significant association with haemolysis (Pizzamiglio et al. 2017, 315-320).

As concerns the analytical phase, miRNAs can be evaluated using different detection platforms, each of them characterized by strengths and weaknesses (Mestdagh et al. 2009, R64-2009-10-6-r64. Epub 2009 Jun 16). Within the microRNA quality control study (miRQC) by Mestdagh and Colleagues, which involved the 9 vendors of miRNA profiling technologies, 12 commercially available platforms for the analysis of miRNAs were considered. Each participant had to profile 16 positive and negative controls samples and 4 additional samples from serum were included in the study as optional evaluation. Briefly, results from this study highlighted the low specificity of some platforms, the low concordance of differential expression, poor reproducibility between qPCR platforms suggesting the necessity of properly choosing the platform on the basis of the experimental setting and specific research questions (Mestdagh et al. 2014, 809-815)

From a methodological point of view, there are two main issues that could impact and probably explain the differences in miRNAs-based study, in addition to a proper study design. The latter should be implemented in order to include a discovery phase aimed at setting-up all the experimental conditions and operating procedures for the miRNA-profiling as well as to identify approaches of pre-processing and quality controls and to define the statistical analysis procedure with the final identification of reference and candidate miRNAs. During this discovery phase, technologies with a high throughput are usually used to allow the evaluation of a large number of biomarkers (hundreds or thousands) in order to select the most important ones that should be then evaluated with more specific assays (based on a limited number of probes). This phase should be followed by a validation phase based on independent cohorts of subjects not included in the first discovery phase, in order to provide information about the performance of the identified biomarkers when tested on different subjects. Within this workflow, which we recently proposed (Verderio et al. 2016, 1-4), two additional steps could be included with the aim to evaluate the performance of a medium/low throughput assay to be used (if part of the study design) on the validation cohort(s) or to set-up an easy-to-use assay to be transferred in the clinical practice.

Obviously, before their introduction in the clinical setting, independent evidences performed by different Authors in different countries, should confirm the real usefulness of these identified biomarkers.

By moving to the statistical aspect, normalization of high-throughput qPCR assays represent a major challenge especially when applied to circulating miRNAs, as no verified and shared *reference* miRNAs were identified in literature as such. *Reference RNAs* should show minimal variation between the experimental conditions under investigation and should be adequately expressed in all the experimental conditions. Different strategies were developed in the 2000's for evaluating the stability of different r*eference RNAs* in the gene-expression scenario and for confirming the stable role of some of the most used and shared housekeeping genes (i.e GAPDH). These methods were then transferred to the studies based on circulating miRNAs. For miRNA-studies based on tissues, small nuclear RNAs (i.e.U6, RNUs) can be used for data normalization. For circulating miRNAs, the global mean method is the one that is currently widely accepted for data normalization. The mean of the expressed miRNAs is characterized by a high stability but, obviously, this approach is suitable and applicable only in the discovery phase, where a large number of biomarkers are assessed. It become unfeasible when a small set of miRNAs are considered and in such scenario, Mestdagh and Colleagues, suggested to find the best combination of miRNAs that resemble the global mean (Mestdagh et al. 2009, R64-2009-10-6-r64. Epub 2009 Jun 16). Starting from these results, we proposed, within the CRC-INT study, an algorithm for the normalization of high-throughput qPCR data. The output of this algorithm is the recognition of a small set of miRNAs that should be used for data normalization in future studies based on a small number of miRNAs and medium/low throughput assays. The criteria introduced for the selection of these miRNAs are the low variation (evaluated trough the CV), the invariance between experimental condition (Kruskal-Wallis test performed on the normalized data using the global mean method), the absence of co-regulation (i.e. correlation) and the high stability values (assessed trough geNorm and Normfinder R-functions).

As regards the identification and validation of candidate miRNAs it is important to emphasize that one single circulating biomarker may not hold the desired performance for disease classification and outcome prediction. Therefore many parametric and non-parametric methods (i.e. combination-based methods) were proposed in literature to opportunely combine multiple biomarkers in a more powerful composite score (i.e. signature or linear combination). These methods can be correctly applied when a small number of well-defined clinical biomarkers (AUC >

0.70) are available (Yan, Tian, and Liu 2015, 3811-3830); in the current era of high-throughput technologies applied to relative small cohorts of subjects, the common scenario is conversely the presence of a large number of weak biomarkers (0.5 < AUC < 0.70), that pose new statistical challenges. The widely used approach for assessing the discriminatory ability of biomarkers (or a joint combination of them) in presence of a binary outcome, is the logistic regression. This allows the estimation of the relative weights of each predictor/variable included in the model and the estimation of the ability to discriminate between subjects with and without the condition under investigation, using the AUC as summary measurement.

One of the most important issue regards overfitting, that can produce an over-optimistic estimation of the discriminatory ability in the developed data set (i.e. training data) and a poorer one when applied to future observations (testing dataset). A possible solution to avoid as much as possible overfitting, is the use of proper penalized regression strategies when the number of outcome events is lower than the number of variables included in the model. All these penalized strategies maximizes the penalized-log-likelihood (instead of the log-likelihood) by introducing a penalty factor or a tuning parameter that shrinkages the regression coefficients towards zero. The specific functional form of the constraint (penalty factor) names different penalized methods, such as ridge and LASSO, which are the most common. We proposed (Verderio et al. 2016, 1-4) to use these methods for miRNA-signature building when the number of event-per-variable (EPV) is lower than 10 in order to reduce the overfitting problem. Once properly identified, the performance of these signature should be tested on an independent cohort of subjects as the performance on the generating dataset may be too optimistic. The issue in this phase is “how to validate these molecular-based signature?”. To note that in the biomarker research, validation means different things to different people (Taylor, Ankerst, and Andridge 2008, 5977-5983): Altman and Royston (Altman and Royston 2000, 453-473) distinguish in fact between two types of validation, the clinical and the statistical ones. The first one aims at evaluating the performance of the model on a new set of data, whereas the second one also takes into account formal statistical aspect, such as goodness of fit and unbiased prediction on the new dataset. Finally, different statistical approaches for external validation (i.e. single-center, multicenter, one time period, multiple time period) and different model-updating strategies are available for evaluating the performance of the developed predictive model.

For our CRC-INT study we adopted the penalized maximum likelihood estimation (PMLE) method (a generalization of the ridge regression) to estimate our miRNA-based signatures and an

all-subset analysis to evaluate all the possible combinations of our candidate mRNAs. This decision was made by mainly consider the peculiarity of the scenario under investigation where there are many weak biomarkers. We then performed a signature-selection (based on the AUC values and shrinkage) in order to select on the Internal Validation Cohort (IVC) only a few number of signatures to be tested on the External Validation Cohort (EVC). The latter includes FIT+ subjects enrolled at our Institute an also in other Hospitals joining the CRC-screening program of the LHA of Milan. Results from EVC allowed us to confirm the predictive capability of some of the identified signatures. As expected the performance of these models on the EVC was lower than that obtained on the IVC, in line with literature.

Another issue that should be mentioned regards the clinical utility of these miRNA-based signatures and their use as prediction models. Obviously the potential clinical utility of miRNAs, or in general of circulating molecular biomarkers, should be evaluated case-per-case by taking into consideration also the disease under investigation, the currently available diagnostic tools for the disease assessment and at which step the molecular prediction models could be introduced. In scenarios where the actual diagnostic tools are invasive and not hold optimal performance characteristics, the use of circulating molecular biomarkers can represent a suitable opportunity to improve the overall diagnostic workflow; this can be done if the considered circulating molecular biomarkers have higher (or at least equal) performance with respect to the standard procedure in use and have lower levels of invasiveness. In any case before its use in clinical practice, the whole pre-analytical and analytical workflow should be standardized.

By looking at the currently published prediction models for CRC (Usher-Smith et al. 2016, 13-26) the majority of the considered models are based on medical records (routinely acquired) or self-completed questionnaire data, and only a few are based on genetic biomarkers alone (SNP or genes). Overall, most models showed AUC values ranging from 0.65 to 0.75. Among the models based on routine data only, the best performing one is that of Betes et al (Betes et al. 2003, 2648-2654) based on more than 2000 asymptomatic subjects attending the CRC screening program in Spain and includes age, gender and BMI values as predictors and advanced colorectal neoplasia as primary outcome (AUC of 0.67). This model was also externally validation on two independent Chinese cohorts (Usher-Smith et al. 2016, 13-26). The three models based on genetic biomarkers alone have higher performance values (AUC > 0.75), but all of them are derived from small case-control studies and some of them are not yet externally validated. Other predictive models based on both routine clinical information and genetic data show performances similar to those obtained

by considering only routine clinical data. Obviously the increasing interest of many research groups in circulating genetic markers (that could be obtained from body fluids) could lead to both the development of new models and to the refinement of those already published by incorporating new genetic/molecular biomarkers. Finally, several issues should be taken into consideration in incorporating a prediction model into clinical practice such as the implementation of randomized controlled trials to eventually assess gains and drawbacks of introducing and using these models in the clinical routine. For molecular/genetic based models, an additional challenge is represented by the proper collection and storage of biological samples and the relative steps of processing as well as by the easy availability of molecular/genetic data detected with standardized and commercial assays (Usher-Smith et al. 2016, 13-26).

According to this recent review (Usher-Smith et al. 2016, 13-26) a future step is to evaluate the added value of our miRNAs-based signature to pre-existing prediction models (using the procedure of model extension) and to identify the best allocation of our signature within the CRC screening diagnostic workflow by eventually estimate the gain introduced by their use.

REFERENCES AND WEB REFERENCES

REFERENCES

Ahmed, F. E. (2014). miRNA as markers for the diagnostic screening of colon cancer. *Expert Review of Anticancer Therapy, 14*(4), 463-485.

Allison, J. E., Sakoda, L. C., Levin, T. R., Tucker, J. P., Tekawa, I. S., Cuff, T., et al. (2007). Screening for colorectal neoplasms with new fecal occult blood tests: Update on performance characteristics. *Journal of the National Cancer Institute, 99*(19), 1462-1470.

Altman, D. G., & Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine, 19*(4), 453-473.

Andersen, C. L., Jensen, J. L., & Orntoft, T. F. (2004). Normalization of real-time quantitative reverse transcription-PCR data: A model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Research, 64*(15), 5245-5250.

Appierto, V., Callari, M., Cavadini, E., Morelli, D., Daidone, M. G., & Tiberio, P. (2014). A lipemia-independent NanoDrop((R))-based score to identify hemolysis in plasma and serum samples. *Bioanalysis, 6*(9), 1215-1226.

Austin, P. C., van Klaveren, D., Vergouwe, Y., Nieboer, D., Lee, D. S., & Steyerberg, E. W. (2016). Geographic and temporal validity of prediction models: Different approaches were useful to examine model performance. *Journal of Clinical Epidemiology, 79*, 76-85.

Betes, M., Munoz-Navas, M. A., Duque, J. M., Angos, R., Macias, E., Subtil, J. C., et al. (2003). Use of colonoscopy as a primary screening test for colorectal cancer in average risk people. *The American Journal of Gastroenterology, 98*(12), 2648-2654.

Bretthauer, M. (2011). Colorectal cancer screening. *Journal of Internal Medicine, 270*(2), 87-98.

Bustin, S. A., & Murphy, J. (2013). RNA biomarkers in colorectal cancer. *Methods (San Diego, Calif.), 59*(1), 116-125.

Butz, H., & Patòcs, A. (2015). Technical aspects related to the analysis of circulating microRNAs. In P. Igaz (Ed.), *Circulating microRNAs in disease diagnostics and their potential biological relevance* (pp. 55-71). Basel: Springer Basel.

Calin, G. A., & Croce, C. M. (2006). MicroRNA signatures in human cancers. *Nature Reviews.Cancer, 6*(11), 857-866.

Chen, X., Ba, Y., Ma, L., Cai, X., Yin, Y., Wang, K., et al. (2008). Characterization of microRNAs in serum: A novel class of biomarkers for diagnosis of cancer and other diseases. *Cell Research, 18*(10), 997-1006.

Cortez, M. A., Bueso-Ramos, C., Ferdin, J., Lopez-Berestein, G., Sood, A. K., & Calin, G. A. (2011). MicroRNAs in body fluids--the mix of hormones and biomarkers. *Nature Reviews.Clinical Oncology, 8*(8), 467-477.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837-845.

Deo, A., Carlsson, J., & Lindlof, A. (2011). How to choose a normalization strategy for miRNA quantitative real-time (qPCR) arrays. *Journal of Bioinformatics and Computational Biology, 9*(6), 795-812.

Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C., & Parkin, D. M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International Journal of Cancer, 127*(12), 2893-2917.

Fleiss,J.L., Levin, B., Paik, M.C. (2004). Statistical methods for rates and proportions, (Third Edition ed.) Walter A. Shewart, Samuel S. Wilks.

Gellad, Z. F., & Provenzale, D. (2010). Colorectal cancer: National and international perspective on the burden of disease and public health impact. *Gastroenterology, 138*(6), 2177-2190.

Harrell, F. E. (2001). Regression modeling strategies with applications to linear models, logistic regression, and survival analysis (Springer Series in Statistics ed.) Springer-Verlag New York.

Hollis, M., Nair, K., Vyas, A., Chaturvedi, L. S., Gambhir, S., & Vyas, D. (2015). MicroRNAs potential utility in colon cancer: Early detection, prognosis, and chemosensitivity. *World Journal of Gastroenterology, 21*(27), 8284-8292.

Imperiale, T. F., Ransohoff, D. F., Itzkowitz, S. H., Levin, T. R., Lavin, P., Lidgard, G. P., et al. (2014). Multitarget stool DNA testing for colorectal-cancer screening. *The New England Journal of Medicine, 370*(14), 1287-1297.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. *CA: A Cancer Journal for Clinicians, 61*(2), 69-90.

Kang, K., Peng, X., Luo, J., & Gou, D. (2012). Identification of circulating miRNA biomarkers based on global quantitative real-time PCR profiling. *Journal of Animal Science and Biotechnology, 3*(1), 4-1891-3-4.

Kang, L., Liu, A., & Tian, L. (2016). Linear combination methods to improve diagnostic/prognostic accuracy on future observations. *Statistical Methods in Medical Research, 25*(4), 1359-1380.

Kirschner, M. B., Edelman, J. J., Kao, S. C., Vallely, M. P., van Zandwijk, N., & Reid, G. (2013). The impact of hemolysis on cell-free microRNA biomarkers. *Frontiers in Genetics, 4*, 94.

Kirschner, M. B., Kao, S. C., Edelman, J. J., Armstrong, N. J., Vallely, M. P., van Zandwijk, N., et al. (2011). Haemolysis during sample preparation alters microRNA content of plasma. *PloS One, 6*(9), e24145.

Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society.Series C (Applied Statistics), 41*(1), 191-201.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45*(1), 255-268.

Liu, C., Liu, A., & Halabi, S. (2011). A min-max combination of biomarkers to improve diagnostic accuracy. *Statistics in Medicine, 30*(16), 2005-2014.

Livak, K. J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-delta delta C(T)) method. *Methods (San Diego, Calif.), 25*(4), 402-408.

MacLellan, S. A., MacAulay, C., Lam, S., & Garnis, C. (2014). Pre-profiling factors influencing serum microRNA levels. *BMC Clinical Pathology, 14*, 27-6890-14-27. eCollection 2014.

Malentacchi, F., Pazzagli, M., Simi, L., Orlando, C., Wyrich, R., Hartmann, C. C., et al. (2013). SPIDIA-DNA: An external quality assessment for the pre-analytical phase of blood samples used for DNA-based analyses. *Clinica Chimica Acta; International Journal of Clinical Chemistry, 424*, 274-286.

Malentacchi, F., Pazzagli, M., Simi, L., Orlando, C., Wyrich, R., Gunther, K., et al. (2014). SPIDIA-RNA: Second external quality assessment for the pre-analytical phase of blood samples used for RNA based analyses. *PloS One, 9*(11), e112293.

Malentacchi, F., Ciniselli, C. M., Pazzagli, M., Verderio, P., Barraud, L., Hartmann, C. C., et al. (2015). Influence of pre-analytical procedures on genomic DNA integrity in blood samples: The SPIDIA experience. *Clinica Chimica Acta; International Journal of Clinical Chemistry, 440*, 205-210.

Malentacchi, F., Pizzamiglio, S., Verderio, P., Pazzagli, M., Orlando, C., Ciniselli, C. M., et al. (2015). Influence of storage conditions and extraction methods on the quantity and quality of circulating cell-free DNA (ccfDNA): The SPIDIA-DNAplas external quality assessment experience. *Clinical Chemistry and Laboratory Medicine, 53*(12), 1935-1942.

Malentacchi, F., Pizzamiglio, S., Wyrich, R., Verderio, P., Ciniselli, C., Pazzagli, M., et al. (2016). Effects of transport and storage conditions on gene expression in blood samples. *Biopreservation and Biobanking, 14*(2), 122-128.

Marubini, E., Pizzamiglio, S., & Verderio, P. (2005). Agreement between observers: Its measure on a quantitative scale. *The International Journal of Biological Markers, 20*(1), 73-78.

Mazeh, H., Mizrahi, I., Ilyayev, N., Halle, D., Brucher, B., Bilchik, A., et al. (2013). The diagnostic and prognostic role of microRNA in colorectal cancer - a comprehensive review. *Journal of Cancer, 4*(3), 281-295.

Mestdagh, P., Van Vlierberghe, P., De Weer, A., Muth, D., Westermann, F., Speleman, F., et al. (2009). A novel and universal method for microRNA RT-qPCR data normalization. *Genome Biology, 10*(6), R64-2009-10-6-r64. Epub 2009 Jun 16.

Mestdagh, P., N. Hartmann, L. Baeriswyl, D. Andreasen, N. Bernard, C. Chen, D. Cheo, et al. 2014. Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nature Methods* 11 (8) (Aug): 809-15.

Moons, K. G., Donders, A. R., Steyerberg, E. W., & Harrell, F. E. (2004). Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: A clinical example. *Journal of Clinical Epidemiology, 57*(12), 1262-1270.

Moons, K. G., Kengne, A. P., Woodward, M., Royston, P., Vergouwe, Y., Altman, D. G., et al. (2012). Risk prediction models: I. development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart (British Cardiac Society), 98*(9), 683-690.

Moons, K. G., Kengne, A. P., Grobbee, D. E., Royston, P., Vergouwe, Y., Altman, D. G., et al. (2012). Risk prediction models: II. external validation, model updating, and impact assessment. *Heart (British Cardiac Society), 98*(9), 691-698.

Nieboer, D., Vergouwe, Y., Ankerst, D. P., Roobol, M. J., & Steyerberg, E. W. (2016). Improving prediction models with new markers: A comparison of updating strategies. *BMC Medical Research Methodology, 16*(1), 128.

Park, D. I., Ryu, S., Kim, Y. H., Lee, S. H., Lee, C. K., Eun, C. S., et al. (2010). Comparison of guaiac-based and quantitative immunochemical fecal occult blood testing in a population at average risk undergoing colorectal cancer screening. *The American Journal of Gastroenterology, 105*(9), 2017-2025.

Pavlou, M., Ambler, G., Seaman, S., De Iorio, M., & Omar, R. Z. (2016). Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in Medicine, 35*(7), 1159-1177.

Pazzagli, M., Malentacchi, F., Simi, L., Orlando, C., Wyrich, R., Gunther, K., et al. (2013). SPIDIA-RNA: First external quality assessment for the pre-analytical phase of blood samples used for RNA based analyses. *Methods (San Diego, Calif.), 59*(1), 20-31.

Pepe, M. S., & Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics (Oxford, England), 1*(2), 123-140.

Pepe, M. S., Etzioni, R., Feng, Z., Potter, J. D., Thompson, M. L., Thornquist, M., et al. (2001). Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute, 93*(14), 1054-1061.

Perkins, J. R., Dawes, J. M., McMahon, S. B., Bennett, D. L., Orengo, C., & Kohl, M. (2012). ReadqPCR and NormqPCR: R packages for the reading, quality checking and normalisation of RT-qPCR quantification cycle (cq) data. *BMC Genomics, 13*, 296-2164-13-296.

Pfaffl, M. W., Tichopad, A., Prgomet, C., & Neuvians, T. P. (2004). Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper--excel-based tool using pair-wise correlations. *Biotechnology Letters, 26*(6), 509-515.

Pizzamiglio, S., Verderio, P., Orlando, C., & Marubini, E. (2007). Confidence interval for DNA/mRNA concentration by real-time PCR. *The International Journal of Biological Markers, 22*(3), 232-236.

Pizzamiglio, S., Cossa, G., Gatti, L., Beretta, G. L., Corna, E., Tinelli, S., et al. (2010). Simultaneous confidence intervals to compare gene expression profiles using ABC transporter TaqMan microfluidic cards. *Oncology Reports, 23*(3), 853-860.

Pizzamiglio, S., Bottelli, S., Ciniselli, C. M., Zanutto, S., Bertan, C., Gariboldi, M., et al. (2014). A normalization strategy for the analysis of plasma microRNA qPCR data in colorectal cancer. *International Journal of Cancer, 134*(8), 2016-2018.

Pizzamiglio, S., Zanutto, S., Ciniselli, C. M., Belfiore, A., Bottelli, S., Gariboldi, M., et al. (2017). A methodological procedure for evaluating the impact of hemolysis on circulating microRNAs. *Oncology Letters, 13*(1), 315-320.

Pritchard, C. C., Kroh, E., Wood, B., Arroyo, J. D., Dougherty, K. J., Miyaji, M. M., et al. (2012). Blood cell origin of circulating microRNAs: A cautionary note for cancer biomarker studies. *Cancer Prevention Research (Philadelphia, Pa.), 5*(3), 492-497.

Reid, J. F., Sokolova, V., Zoni, E., Lampis, A., Pizzamiglio, S., Bertan, C., et al. (2012). miRNA profiling in colorectal cancer highlights miR-1 involvement in MET-dependent proliferation. *Molecular Cancer Research : MCR, 10*(4), 504-515.

Silver, N., Best, S., Jiang, J., & Thein, S. L. (2006). Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Molecular Biology, 7*, 33.

Singh, R., Ramasubramanian, B., Kanji, S., Chakraborty, A. R., Haque, S. J., & Chakravarti, A. (2016). Circulating microRNAs in cancer: Hope or hype? *Cancer Letters, 381*(1), 113-121.

Su, J. Q., & Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association, 88*(424), 1350-1355.

Taylor, J. M., Ankerst, D. P., & Andridge, R. R. (2008). Validation of biomarker-based risk prediction models. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research, 14*(19), 5977-5983.

Tiberio, P., Callari, M., Angeloni, V., Daidone, M. G., & Appierto, V. (2015). Challenges in using circulating miRNAs as cancer biomarkers. *BioMed Research International, 2015*, 731479.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society.Series B (Statistical Methodology), 73*(3), 273-282.

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 Prevention Research (Philadelphia, Pa.), 9*(1), 13-26.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., et al. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology, 3*(7), RESEARCH0034.

Verderio, P., Orlando, C., Casini Raggi, C., & Marubini, E. (2004). Confidence interval estimation for DNA and mRNA concentration by real-time PCR: A new environment for an old theorem. *The International Journal of Biological Markers, 19*(1), 76-79.

Verderio, P., Mangia, A., Ciniselli, C. M., Tagliabue, P., & Paradiso, A. (2010). Biomarkers for early cancer detection - methodological aspects. *Breast Care (Basel, Switzerland), 5*(2), 62-65.

Verderio, P. (2012). Assessing the clinical relevance of oncogenic pathways in neoadjuvant breast cancer. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology, 30*(16), 1912-1915.

Verderio, P., Bottelli, S., Ciniselli, C. M., Pierotti, M. A., Gariboldi, M., & Pizzamiglio, S. (2014). NqA: An R-based algorithm for the normalization and analysis of microRNA quantitative real-time polymerase chain reaction data. *Analytical Biochemistry, 461*, 7-9.

Verderio, P., Bottelli, S., Ciniselli, C. M., Pierotti, M. A., Zanutto, S., Gariboldi, M., et al. (2015). Moving from discovery to validation in circulating microRNA research. *The International Journal of Biological* Markers, 30(2), e258-61.

Verderio, P., S. Bottelli, M. Lecchi, M. Plebani, M. Gariboldi, S. Pizzamiglio, and C. M. Ciniselli. (2016). Comment on 'circulating cell-free miRNAs as biomarker for triple-negative breast cancer'-methodological challenges in combining miRNAs as circulating biomarkers. *British Journal of Cancer,* 114 (10) (May 10): e5.

Verderio, P., Bottelli, S., Pizzamiglio, S., & Ciniselli, C. M. (2016). Developing miRNA signatures: A multivariate prospective. *British Journal of Cancer, 115*(1), 1-4.

Vergouwe, Y., Nieboer, D., Oostenbrink, R., Debray, T. P., Murray, G. D., Kattan, M. W., et al. (2016). A closed testing procedure to select an appropriate method for updating prediction models. *Statistics in Medicine,*

Verma, A. M., Patel, M., Aslam, M. I., Jameson, J., Pringle, J. H., Wurm, P., et al. (2015). Circulating plasma microRNAs as a screening method for detection of colorectal adenomas. *Lancet (London, England), 385 Suppl 1*, S100-6736(15)60415-9.

Yamada, A., Cox, M. A., Gaffney, K. A., Moreland, A., Boland, C. R., & Goel, A. (2014). Technical factors involved in the measurement of circulating microRNA biomarkers for the detection of colorectal neoplasia. *PloS One, 9*(11), e112481.

Yan, L., Tian, L., & Liu, S. (2015). Combining large number of weak biomarkers based on AUC. *Statistics in Medicine, 34*(29), 3811-3830.

Yin, J., & Tian, L. (2014). Optimal linear combinations of multiple diagnostic biomarkers based on youden index. *Statistics in Medicine, 33*(8), 1426-1440.

Zanutto, S., Pizzamiglio, S., Ghilotti, M., Bertan, C., Ravagnani, F., Perrone, F., et al. (2014). Circulating miR-378 in plasma: A reliable, haemolysis-independent biomarker for colorectal cancer. *British Journal of Cancer, 110*(4), 1001-1007.

WEB REFERENCES

- http://www.ederaproject.it
- http://www.hopkinscoloncancercenter.org
- http://www.spidia.eu
- http://www.osservatorionazionalescreening.it
- http://www.giscor.it
- https://cran.r-project.org/web/packages

ACKNOWLEDGMENT

I would like to thank all members of the PhD board and especially my Tutor, Dr Paolo Verderio, who always supported me during my research activity within the PhD course. A special thank also to my colleagues Mara, Maddalena, Sara and Stefano for their help during these years. I would also like to thank the External Reviewers, Dr.ssa Daidone and Dr.ssa Gariboldi, for their precious comments and suggestions.