UNIVERSITÀ DEGLI STUDI DI MILANO

PhD Course in Epidemiology, Environment and Public Health

Innovative biostatistical and bioinformatic approaches in the analysis of breast cancer: competing risk survival analysis through pseudo-values and comprehensive evaluation of methods for the tumor microenvironment dissection available at the present day.

CO-SUPERVISOR: Prof. Elia Biganzoli

CO-SUPERVISOR: Patrizia Boracchi

PhD thesis of

Davide De Bortoli

Matr. R11731

Academic Year 2018 – 2019

Acknowledgments

I would like to thank my two co-supervisors Elia Biganzoli and Patrizia Boracchi, along with the members of my group Giuseppe Marano, Annalisa Orenti and Federico Ambrogi.

A very special thanks to Christine Desmedt, leader of an international research group focused on breast cancer, which provided the data on which my projects have been developed and welcomed me in Bruxelles to work to Tumor microenvironment dissection project together with the other members of the group composed by Iris Nederlof, Yacine Bareche, Bastien Nguyen, Michiel de Maaker, Gerrit K.J. Hooijer, Laurence Buisseret, Marleen Kok, Marcel Smid, Gert van den Eynden, Arjen Brinkman, Jan Hudecek, Jan Koster, Christos Sotiriou, Denis Larsimont, John W.M. Martens, Marc J. van de Vijver, Hugo M. Horlings and Roberto Salgado.

General introduction

Since my personal original background was quite distant from the statistical bioinformatic approaches for data analysis, having a master degree in Sanitary Biotechnology and Molecular Medicine, my PhD fellowship was spent in building my skills in this field while studying and trying to contribute to the development of statistical and bioinformatic approaches to be applied in clinic, with a special focus on oncology, in the optic to contribute to the field of personalized medicine.

Personalized medicine is indeed the ultimate goal for life sciences, particularly for oncology, and, in my opinion, a key aspect of the future wellness of humanity. Personalized medicine in the idea of developing the ability to identify the best therapeutic strategy for each unique person and its efficacy relies on having accurate diagnostic tests that identify patients who can benefit from targeted therapies. A striking example consists in the determination of the overexpression of the human epidermal growth factor receptor type 2 (HER2) in the routinely diagnosis of Breast Cancer (BC). HER2 is indeed associated with a worse prognosis but also predicts a better response to the medication trastuzumab; a test for HER2 was approved along with the drug (as a “companion diagnostic”) so that clinicians can better target patients' treatment [1].

In the following pages I present two projects that have mainly characterized my fellowship. Both projects rely on breast cancer (BC) and the objective of understanding the effects of chronic low inflammation, which has been studied in my projects as the leucocyte infiltration and the body mass index.

The focus on BC derives from a practical aspect and an epidemiological aspect. The practical aspect consists on the fact that my group is part of a European research group, led by Christine Desmedt from Belgium, which allowed me to obtain unique data and to interact with experts of BC and bioinformatics from different countries. The epidemiological aspect is represented by the fact that breast cancer is actually a hot topic, being the second most common cancer worldwide and the first among women [2], but still open to investigations, since the complexity and variability of BC, reflected both at histopathological [3,4,5] and molecular level, have proven challenging to classify and therefore to effectively treat to the present day.

The first project presented, the tumor microenvironment (TME) dissection project, occupied the first part of my fellowship. This project was conducted as a joint effort of numerous colleagues that composed a big European collaboration focused on BC and leaded by Christine Desmedt. This particular project was focused on the comparison of different and already existing tools and approaches used to analyze breast cancer TME. My role here was to review the methodological issues and properties of the methods proposed in literature by exploring the specific algorithms and codes according to the statistical and bioinformatic skill I acquired during my PhD. This work is considered of essential importance since one of the major issues related to the application of these methods is essentially a blind use without going into details of the quantitative statistical implication of the methods.

This project consisted in a big European collaboration which tried to establish the reliability of bioinformatic tools in retrieving the TME composition by analyzing and comparing the obtained results to standard approaches, as the pathologist evaluation, and emerging methods as digital image analysis.

This project led to the preparation of a paper and submission of the paper “Comprehensive evaluation of methods to 3 assess overall and cell-specific immune 4 infiltrates in breast cancer”, of which I am first co-author together with Iris Nederlof, that has been recently accepted by Breast Cancer Research journal.

The second project presented, the competing risk analysis through pseudo-values project, which characterized the third year of my PhD, is more focused on the statistical aspects of clinical data analysis and represent the arrival point of my studies of statistical methodology. The project consisted in the exploration of a forefront approach to the analysis of survival data based on pseudo-values, which has the desirable feature to generate measures with a clear and direct interpretation at a clinical level, becoming an invaluable tool for clinical decision making. This project represents a first step in a longer-term project that will led to the preparation of several papers in the future.

The two projects are presented separately, with their own introduction, materials and methods, results and conclusion parts; only references are put together at the end of the manuscript.

Tumor microenvironment dissection project

Introduction

Personalized medicine is actually the ultimate goal for life sciences, and particularly for oncology. The efficacy of personalized medicine relies on having accurate diagnostic tests that identify patients who can benefit from targeted therapies. A striking example consists in the determination of the overexpression of the human epidermal growth factor receptor type 2 (HER2) in the routinely diagnosis of Breast Cancer (BC). HER2 is indeed associated with a worse prognosis, but also predicts a better response to the medication trastuzumab; a test for HER2 was approved along with the drug (as a “companion diagnostic”) so that clinicians can better target patients' treatment [6].

The complexity and variability of BC, reflected both at histopathological [7, 8, 9] and at molecular level, have proven challenging to classify to the present day.

In this context, in the past decades the role of tumor microenvironment (TME), composed of non-cancerous cells, steadily gained interest and has been studied along with a variety of methods to classify the immune infiltrate. The composition of the tumor and TME have proven to have a relevant clinical impact, as it can be observed studying the relationship between the TME and disease progression or therapy benefit in breast cancer patients. Tumor immunogenicity is defined by infiltration of tumor-reactive immune cells (or the lack), visibility of the tumor by presentation of (neo)antigens and sensitivity of the tumor to tumor-reactive immune cells [10].

The role of innate immune cells, like macrophages and neutrophils, have been associated with tumor progression and metastasis. In BC the presence of immune infiltrate and its composition have been showed to affect prognosis and treatment efficacy, including the response to novel immunotherapies [11,12,13,14,15]. Particularly, increased levels of stromal tumor infiltrating lymphocytes (sTIL) are associated with response to neoadjuvant chemotherapy and prognosis in triple negative BC (TNBC) patients [16, 17, 18, 19, 20, 21]. Moreover, sTIL has recently been recognized as a valid prognostic biomarker by the expert panel of the 16th St. Gallen Breast Cancer Conference.

Reliable methods to estimate the amount and composition of the immune infiltrate are therefore critical, for cross study comparisons and further biomarker development. This would improve the read-out of TIL estimates, provide better information on the cell phenotypes contributing to TIL, and would ultimately lead to better stratification of patients in the future.

Along with the development of methods to study the immune composition through the microscope [22, 23], the constant decrease of the costs of next-generation sequencing (NGS) technologies in the last years [10], enabled their use in routine oncology, promoting collaborations like the cancer genome atlas (TCGA) [24] and gaining access to an unprecedented amount of RNA sequencing (RNA-seq) data describing the tumor microenvironment composition, which, to the present day, has been studied mainly through immunohistochemistry (IHC), immune fluorescence (IF), and flow cytometry.

In recent years, many groups have been working on the development of computational approaches capable to infer the composition of tumor-infiltrating immune cells from bulk tumor RNA-seq data, using well defined groups of immune-specific marker genes or expression signatures [25].

Two main approaches for TME dissection exist. The most famous one, based on the analysis of marker genes, is gene set enrichment analysis (GSEA) [26]. Methods based on GSEA calculate an enrichment score which is higher if among the most expressed genes in the sample there are many specific for a certain cell type (i.e. the cell type is enriched in the sample) and low otherwise (Figure 1a).

Unlike GSEA-based approaches, which score is just semi-quantitative, describing the enrichment of specific cell types in a sample, deconvolution methods are capable of calculating quantitative estimates of the relative fractions of the cell types of interest. The gene expression profile of a heterogeneous sample is considered by deconvolution algorithms as the convolution of gene expression levels of different cells and thus estimate the unknown cell fractions leveraging on a signature matrix describing the cell-type-specific expression profiles (Figure 1b) [25].

More recently, deconvolution methods that rely on methylation data have been presented; the usage of methylation data could be more reliable since specific cell type methylome is more stable than transcriptome, which is prone to fluctuations or could be little informative in case of non-highly transcribing cells, like secreting plasma cells.

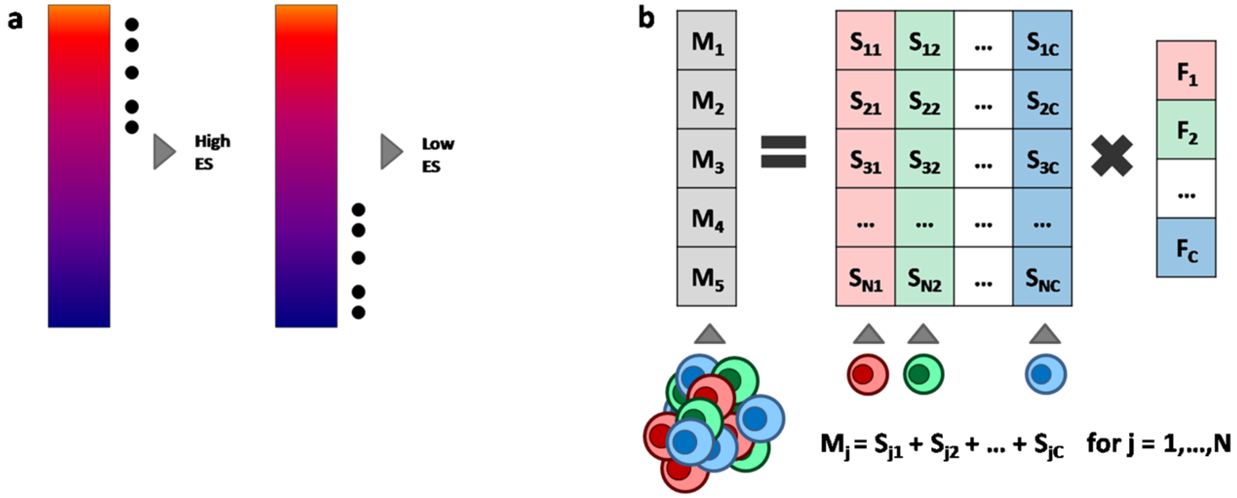

a
High ES
Low ES
b
M_1
M_2
M_3
M_4
M_5
S_{11} S_{12} ... S_{1C}
S_{21} S_{22} ... S_{2C}
S_{31} S_{32} ... S_{3C}
... ...
S_{N1} S_{N2} ... S_{NC}
F_1
F_2
...
F_C
$M_j = S_{j1} + S_{j2} + ... + S_{jC}$ for j = 1,...,N

Figure 1: ***a*** *GSEA approaches rank the genes according to their expression in a sample and then calculate an enrichment score (ES) based on the position of a set of cell-type-specific marker genes (black dots) in the ranked list. The ES is high when the marker genes are among the top highly expressed genes (red) and low otherwise (blue).* ***b*** *In the deconvolution approaches the expression of a gene in a mixture M is modeled as a linear combination of the expression of that gene in the different cell types, whose average expression profiles are stored in a signature matrix S, weighted by the relative fractions F of the cell types in the mixture.*

Since a comprehensive comparison of microscopic and computational methods is still missing, the primary objective of the current study was therefore to compare the estimations of overall and cell-specific immune infiltration obtained by microscopic, transcriptomic and methylomic methods in the ICGC breast cancer cohort [27, 28]. The secondary objective of this work was then to evaluate the ability of the different methods to classify tumors as “hot”, highly infiltrated, or “cold”, poorly or not infiltrated [20].

Materials and methods

Dataset

The cohort used in the study is the ICGC BASIS breast cancer cohort [26], which provided sequencing, expression and methylation data. Briefly, for 560 breast cancer patients with clinic-pathological data available, DNA and RNA were analyzed using WGS, DNA methylation profiling and RNA sequencing. The breast cancer dataset is hosted by the International Cancer Genome Consortium.

From the original dataset of 560 patients, we excluded males (n=4), local relapses (n=7) and metastasis (n=1). Of the remaining 548 remaining subjects, 257 patients had available RNA sequencing data (all with more than 5 million counts assigned, 5 subject with RNA data available were excluded because having less aligned reads) and 318 have methylation data available

Gene expression values were retrieved as FPKM and transformed to TPM (transcripts per million), providing expression data for 44558 genes.

Besides sequencing, expression and methylation data, comprised in the original dataset, imaging analyses were performed for the first time by my group for this project. Overall infiltration was evaluated on hematoxylin and eosin (H&E) whole slides (WS, n=243), while immunohistochemistry was performed on either TMA's (n=254) or whole slides (n=82).

The TMA's used in this work were constructed using formalin-fixed, and paraffin-embedded primary invasive breast cancer samples from the BASIS cohort. From the selected (donor) blocks three cores of tumor with thickness of 0.6 mm were collected using a TMA instrument (Beecher Instruments, Silver Springs MD, USA) and inserted in a recipient block. Each block was sectioned at 4 µm, and dried overnight at 37ºC. For the Whole Slides, 4 µm sections were cut and dried overnight in a 37ºC incubator. All of the staining procedures were performed on a Ventana Benchmark XT automated staining instrument (Ventana Medical Systems). Sections were scanned at 20× magnification using a Aperio Scanscope Scanner.

Whole slides were incubated with primary antibodies specific for CD3 (SP7, Thermo Scientific), CD4 (SP35, Roche), CD8 (C8/144B, DAKO), CD20 (L26, DAKO), CD68 (PG-M1, DAKO), FOXp3 (236A/E7, AbCam), and PD-L1 (22C3, DAKO).

Assessment by pathologists

Stromal and intratumoral overall immune infiltration (sTIL and itTIL respectively) was independently evaluated by two experienced pathologists (Hugo Horlings, Roberto

Salgado) according to published guidelines [22] using SlideScore (www.slidescore.com), an online platform, on the 243 available H&E's WS. Briefly, the relative proportion of stromal area to tumor area was determined from the microscopy slide of a given tumor region. TIL were reported for the stromal compartment (= percent sTIL) and intratumoral compartment (= percent itTIL). The denominator used to determine the percentage of sTIL and itTIL was the area of stromal tissue or intratumoral tissue, respectively.

The same methodology was applied for evaluation of stromal and intratumoral infiltration of specific immune cell types on TMA (n=254) and WS (n=82).

These values were used for the inter-observer evaluation. Hereafter, the pathologists re-evaluated all highly discordant stromal IHC (arbitrarily defined as >10% difference between pathologists) to strive for a better accuracy of immune characterization for further analysis. The geometric mean of the two scores was used.

Assessment by digital pathology

In addition, all TMA cores and the 82 WS were evaluated by digital imaging analysis using Visiopharm Integrator System software. Markers were quantified as the percentage of positive area on the total tissue area. Empty space was automatically excluded to avoid contributions from regions of no interest. Cores with artifacts were excluded post-processing and not included in the comparative analysis.

Tumor microenvironment dissecting tools

Gene set enrichment

The original GSEA approach evaluates if a certain group of genes is differentially expressed between two biological states [26]. Briefly, the genes in the expression data set are ordered on the basis of the correlation between their expression and the condition considered. Thus, for each gene in the ranked list, a running-sum statistic is increased when one of the genes belonging to the query set is encountered and decreased otherwise. In the end, the Enrichment Score (ES) is calculated as the maximum deviation from zero of the running-sum statistic.

Single-sample GSEA (ssGSEA), differently to the original GSEA method, orders the genes considering their absolute expression in a specific sample and, by integrating the differences between the empirical cumulative distribution functions of the gene ranks, computes ESs which represent the degree to which genes in a particular gene set are uniformly up- or down-regulated [29]. On this approach is based xCell, a recently published method which estimates the abundance scores of 64 immune cell types, through the use of a compendium of 489 gene sets collected from large-scale expression data from different projects and studies: FANTOM5 [30], ENCODE [31], Blueprint [32], Immune Response In Silico (IRIS) [33], Human Primary Cell Atlas (HPCA) [34], and Novershtern et al. [35].

For each cellular type considered, xCell calculates the abundance scores through four steps: (I) ssGSEA is used independently for each gene of the 489 gene sets using the GSVA R package [36]; (II) ES are averaged across all gene sets related to a certain cell type; (III) ES are converted in abundance scores in a platform-specific way; (IV) correlations between related cell types are corrected using a "spillover" approach similar to that used for flow cytometry data analysis. Abundance scores provided by xCell cannot be used directly as cell fractions, however a high correlation between them and the true cell proportion has been seen [37].

Among the 64 cell types considered in xCell, only those that could be compared with IHC markers were considered; Table 1 shows which of them were used for the comparisons.

MCP-counter is a recently published GSEA method. Developed by the group of Becht [38], available as an R-package, MCP-counter consists in a method for the quantification of tumor-infiltrating immune cells, fibroblasts and epithelial cells based on a robust set of marker genes. For every cell type and sample, an abundance score is calculated as the geometric mean of the expression values of cell-type-specific genes. Because arbitrary units are used for the abundance scores, they cannot be directly interpreted as cell fractions, nor used to compare different cell types, although quantitative validation using well-defined cell mixtures showed high correlation between the estimated scores and the true cell fractions. In Table 2 the cell types used for the comparisons.

Table 1*: Cell types taken into consideration among the xCell ones.*

Cell types used for comparisons	Corresponding cell types from xCell
CD3_Tcells	CD4 memory T cells CD4 naive T cells CD4 T cells CD4 Tcm CD4 Tem CD8 naive T cells CD8 T cells CD8 Tcm CD8 Tem Tgd cells Th1 cells Th2 cells Tregs
CD4_Tcells	CD4 memory T cells CD4 naive T cells CD4 T cells CD4 Tcm CD4 Tem
CD8_Tcells	CD8 naive T cells CD8 T cells CD8 Tcm CD8 Tem
CD20_Bcells	B cells Memory B cells naive.B.cells
CD68_macrophages	Macrophages, Macrophages M1, Macrophages M2
FOXp3	Tregs

Table 2*: Cell types taken into consideration among the ones used by MCPCounter.*

Cell types used for comparisons	Corresponding cell types from MCPCounter
CD3_Tcells	T cells
CD4_Tcells	NA
CD8_Tcells	CD8 T cells
CD20_Bcells	B lineage
CD68_macrophages	Monocytic lineage
FOXp3	NA

Deconvolution

The deconvolution problem can be thought as a system formed by linear equations, each describing the expression of a gene as a linear combination of the expression levels of that gene in the different cell types present in the sample, weighted by their relative cell fractions (Figure 1b) [25]. Although the relationship between the expression levels of pure and heterogeneous samples is not strictly linear, the linearity assumption has been shown reasonable [39].

Abbas et al. proposed an approach based on linear least square regression to solve the deconvolution problem, force all negative estimates to zero and re-normalize the cell fractions to sum up to one [40]. Linear least square regression approach is comprised as well in CellMix R package, selecting the *lsfit* method.

Gong et al. presented a method based on constrained least squares and quadratic programming to identify the deconvolution solution with the lowest error while forcing the cell fractions to be non-negative and to sum up to one [41]. This approach is comprised in CellMix R package as *qprog* method.

Moreover, CellMix package offer two more methods, discussed below: *deconf* and *DSA*.

CellMix's signature matrix comprise 17 cell types; in the following Table 3 are shown the specific cell types that were taken into consideration.

***Table3**: Cell types taken into consideration among the ones used by CellMix.*

Cell types used for comparisons	Corresponding cell types from CellMix
CD3_Tcells	T helper, T helper activated, T cytotoxic, T cytotoxic activated
CD4_Tcells	T helper, T helper activated
CD8_Tcells	T cytotoxic, T cytotoxic activated
CD20_Bcells	B cells, B cells activated
CD68_macrophages	monocytes, monocytes activated
FOXp3	NA

CIBERSORT takes into consideration 22 immune cell types, including also different functional states for the same phenotype, summarized in a signature matrix built from microarray data [42]. CIBERSORT estimates the cell fractions using nu support vector regression (ν-SVR), in Table 4 the cell types taken into consideration. For each sample, v-SVR is run with three different ν values (0.25, 0.5, and 0.75) and the solution providing the lowest root-mean-square error (RMSE) between the true expression and the estimated expression $\widehat{M} = S * \widehat{F}$ is selected. Also in this approach, the coefficients are forced to non-negative values and normalized to sum up to one. CIBERSORT is freely available online at https://cibersort.stanford.edu/.

***Table 4**: Cell types taken into consideration among the ones used by CIBERSORT.*

Cell types used for comparisons	Corresponding cell types from CIBERSORT
CD3_Tcells	T cells CD4 naïve, T cells CD4 memory resting, T cells CD4 memory activated, T cells CD8, T cells follicular helper, regulatory T cell, T cells gamma delta
CD4_Tcells	T cells CD4 naïve, T cells CD4 memory resting, T cells CD4 memory activated, T cells follicular helper,
CD8_Tcells	T cells CD8,
CD20_Bcells	B cells naive B cells memory
CD68_macrophages	Monocytes, Macrophages M0, Macrophages M1, Macrophages M2
FOXp3	regulatory T cell

Racle et al. recently developed a tool to Estimate the Proportion of Immune and Cancer cells (EPIC). EPIC is based on constrained least square regression, integrating in this way the non-negativity constraint into the deconvolution problem, and to impose that the sum of all cell fractions in each sample is lower than one. Is therefore possible to retrieve the estimate of the proportion of uncharacterized cells in the mixture, as the difference between 1 (i.e., 100% of the cells in the mixture) and the sum of the deconvoluted cell fractions. These uncharacterized cells are the cell types not taken into consideration by the signature matrix used for deconvolution and, in RNA-seq data from bulk tumors, represents the tumor content [43].

EPIC provides two RNA-seq-derived signature matrices, one describes the expression signature of six blood-circulating immune cell types, the second of five tumor-infiltrating immune cell types plus endothelial cells and cancer-associated fibroblasts (CAF), whose expression signatures were extracted from melanoma single-cell RNA-seq data [44]. Our analyses have been performed using the second signature matrix; specific cell types considered for the comparisons are shown in Table 5.

***Table 5**: Cell types taken into consideration among the ones used by EPIC.*

Cell types used for comparisons	Corresponding cell types from EPIC
CD3_Tcells	CD4 T cells CD8 T cells
CD4_Tcells	CD4 T cells
CD8_Tcells	CD8 T cells
CD20_Bcells	B cells
CD68_macrophages	Macrophages
FOXp3	NA

quanTIseq is, currently, one of the most recent published deconvolution tool and is specifically developed for RNA-seq data [45]. quanTIseq is based on constrained least square regression (to consider the non-negativity and sum-to-one constraints) and on a novel signature matrix derived from a compendium of 51 RNA-seq data sets from purified or enriched immune cell types, including also Treg cells and classically (M1) and alternatively (M2) activated macrophages. Moreover, quanTIseq implements a series of functions to perform all the steps in the analysis of RNA-seq data, from read pre-processing to deconvolution of cell fractions in order to inconsistencies between the mixture and the signature matrix. In Table 6 are shown the specific cell types considered in the analysis.

***Table 6**: Cell types taken into consideration among the ones used by quantiSeq.*

Cell types used for comparisons	Corresponding cell types from quantiSeq
CD3_Tcells	T cells CD4, T cells CD8
CD4_Tcells	T cells CD4
CD8_Tcells	T cells CD8
CD20_Bcells	B cells
CD68_macrophages	Macrophages M1 Macrophages M2 Monocytes
FOXp3	T regulatory

Beyond these, which are called partial deconvolution methods, complete deconvolution approaches exist, which estimate relative cell fractions and simultaneously disentangle their expression profiles.

Starting from the pioneering work of Venet et al. [46], several methods have leveraged on non-negative matrix factorization (NMF) to alternate least-square estimation of the cell proportions and expression profiles [47, 48, 49].

NMF is a completely unsupervised approach and thus it might decompose the mixture matrix into components that are not related to the cell types of interest. By using well-defined mixtures of four hematological cancer cell lines, Gaujoux and colleagues were able to demonstrate that incorporating prior knowledge from cell-specific marker genes into NMF-based methods can dramatically improve the results of complete deconvolution. Deconf, a method based on this approach, is implemented in CellMix R package.

DSA is another complete deconvolution approach that uses quadratic programming to calculate cell fractions and expression profiles in complex tissues leveraging on a set of marker genes that are highly expressed in specific cell types [50]. Also DSA is implemented in CellMix, but it gives back the same results as the "qprog" method, which is comprised in the same package.

Methylation data

Recently, a new approach based on methylation data has been proposed. MethylCIBERSORT is an R package which is capable to create a mixture matrix starting from normalized methylation data that can been then uploaded to the previously discussed online tool CIBERSORT, along with a signature matrix chosen between some provided within the package itself. A breast cancer specific signature matrix was available and used. Raw methylation data was normalized using *preprocessQuantile()* function in the R package *minfi*. To be noted, MethylCIBERSORT provides signature matrices that take into consideration also the tumor cell fraction, allowing also the evaluation of tumor purity [51]. The cell types taken into consideration using MethylCIBERSORT are shown in Table 7.

***Table 7**: Cell types taken into consideration among the ones used by MethylCIBERSORT.*

Cell types used for comparisons	Corresponding cell types from MethylCIBERSORT
CD3_Tcells	T cells CD4 effector, T cells CD8, T regulatory
CD4_Tcells	T cells CD4 effector
CD8_Tcells	T cells CD8
CD20_Bcells	CD19 positive cells
CD68_macrophages	NA
FOXp3	T regulatory

Addititional methods based on signatures

To provide potential validation for the immune infiltration, gene signatures specific for inflammation or immune cell activity, including cytolytic activity (Cyt Act) 12, interferon-gamma signaling (IFNg) 13, lymphocyte signature 14, STAT1 immune signature (STAT1) 15, and two additional immune signatures ImmPerez 16 and TFH 17, [52, 53, 54, 55, 56, 57], were used.

In addition, the published methylome TIL (MethylTIL) signature 11 [58] to estimate TIL abundance from methylation level was used.

Calculations of these latter additional methods were conducted by the colleague Bareche Yacine, one of the co-author with which the paper relative to this work is published, and therefore details are not provided.

Statistical analyses

Software

All statistical analyses were performed using R version 3.5.1.

Concordance Correlation Coefficient

Also known as Lin's coefficient [59], the Concordance Correlation Coefficient (CCC) is a correction of the Pearson's correlation coefficient, which provides a measure of the extent to which the points in the scatter plot conform to the best fitting line. CCC modifies the Pearson's correlation coefficient by assessing not only how close the data are about the line of best fit, but also how far that line is from the 45-degree line through the origin, which represent perfect agreement.

CCC can be calculated as follow:

$$CCC = \frac{2rs_x s_y}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2}$$

Where r is the Pearson's correlation coefficient, s_x^2 and s_y^2 are the estimated variances of x and y respectively, $\bar{x}$ and $\bar{y}$ are the means of x and y respectively.

CCC is equal to 1 when all the point lies on the 45-degree line and there is perfect agreement.

CCC was used as a summary measure of reproducibility. CCC was calculated using *epi.ccc()* function in the *epiR* package.

Bland and Altman diagram

The Bland and Altman diagram is a display of the differences between numeric pairs of readings that can offer an insight into the pattern and extent of the agreement.

It is displayed by plotting the differences between pairs on the vertical axis against the mean between the pairs on the horizontal axis; moreover the average of the differences is displayed as an horizontal line, along with an upper and lower lines lying at $\bar{d} \pm 2s_d$ (where $\bar{d}$ is the mean of the differences, and s_d is the standard deviation) representing the limits of agreement [60].

We used log transformed values, which modify the Bland and Altman diagram: on the vertical axis are plotted the ratios of the pairs and on the horizontal axis the geometric mean. The log transformation has been used to display more clearly the differences between low values (as the majority of observations) and it is the only transformation giving back transformed differences which are easy to interpret [61].

Bland and Altman diagram was used for the comparisons between TMA and WS and between microscopic (assessed by pathologists) and digital analysis.

Passing-Bablok

Passing-Bablok [62] is an approach for the comparison of two measurement methods that should give the same values. Passing-Bablok regression is a robust, nonparametric method for fitting a straight line to two-dimensional data where both variables, X and Y, are measured with error.

The comparison of the two methods is performed by estimating a linear regression line and testing whether the intercept is zero and the slope is one. The Passing-Bablok regression procedure fits the intercept (β_0) and the slope (β_1) of the linear equation $Y = \beta_0 + \beta_1 X$.

The slope's estimate (β_1) is calculated as the median of all slopes that can be formed from all possible pairs of data points, except those pairs that result in a slope of 0/0 or -1. To correct for estimate bias caused by the lack of independence of these slopes, the median is shifted by a factor K which is the number of slopes that are less than -1. This creates an approximately unbiased estimator. The estimate of the intercept (β_0) is the median of $\{Y_i - \beta_1 X_i\}$.

The intercept represents the systematic bias (difference) between the two methods. The slope measures the amount of proportional bias (difference) between the two methods.

Passing-Bablok was used along with Bland and Altman diagram for the comparisons between TMA and WS and between microscopic (assessed by pathologists) and digital analysis, but also for the comparison between global, intratumoral and stromal TIL. Passing-Bablok regression was performed through the *PBreg()* function in *MethComp* package.

Cohen's kappa

Cohen's kappa is a statistical approach to evaluate agreement between two methods that measure binary variables [63].

Cohen's kappa is calculated as follow:

$$k = \frac{Observed\ agreement - Chance\ agreement}{Maximum\ agreement - Chance\ agreement} = \frac{\rho_0 - \rho_E}{1 - \rho_E}$$

Where ρ_0 is the sum of the observed frequencies along the diagonal of the contingency table divided for the number of the subjects, while ρ_E is the sum of the expected frequencies along the diagonal of the contingency table divided for the number of the subjects.

When the agreement is perfect Cohen's kappa is equal to 1, while a value of 0 suggest that agreement is no better than which that would be obtained by chance alone.

Also for ordinal data with more than two categories, weighted Cohen's kappa is used, which takes into consideration not only the agreement between the pairs of results, but also the extent to which there is disagreement between them: the disagreement is greater if for a subject the response of the two methods differ by two categories than by one category.

Weighted kappa is calculated assigning weights to the frequencies in the non-diagonal cells of the contingency table according to their distance from the diagonal, with the magnitude of the weight diminishing the further the cell is from the diagonal [64].

Cohen's kappa was used when comparing infiltration measures relative to Macrophages, since scored as 4 levels categorical variable.

Cohen's kappa was calculated through the *ckap()* function in package *rel*.

Spearman's rank correlation

Spearman's rank correlation is a non-parametric method to measure the linear association of two variables X_1, X_2 without making assumption on their distribution. It is calculated as following:

$$\rho_s = \frac{Cov(rk_{X_1},\ rk_{X_2})}{\sigma_{rk_{X_1}}\sigma_{rk_{X_2}}}$$

where $Cov(rk_{X_1},\ rk_{X_2})$ is the covariance between the ranks of the two variables,

rk_{X_1}, rk_{X_2} are the ranks of the two variables X_1, X_2 respectively and $\sigma_{rk_{X_1}}\sigma_{rk_{X_2}}$ are the standard deviations of the two rank variables.

Spearman's rank correlation was used to compare cell-specific infiltration estimations by various methods and was calculated using the cor() function in R specifying the usage of only complete observations and the method "spearman".

Results

In following pages, the results of the numerous comparisons are showed.

First, in the paragraph titled "Assessment of immune cells on whole slides" the comparisons of the scores calculated from the two pathologists on WS (H&E and IHC) are showed.

Secondly, in the "Comparison of tumor immune infiltration estimates between tissue micro arrays and whole slides" section, the scores calculated on TMA are compared to the ones from WS, for both pathologists' and digital evaluations, however only for CD3 in the last case. Moreover, the comparison between pathologists' and digital scores calculated on TMA is presented.

Thirdly, in the "Comparison of microscopic, transcriptomic and methylomic evaluation of overall tumor immune infiltration" part, methods (microscophy-, transcriptomic-, methylomic-, immune signature-based) capable of estimating global infiltration are compared considering all samples and subgroups based on the presence of ER.

Fourthly, in the "Comparison of microscopic, transcriptomic and methylomic evaluation of cell-specific immune infiltration", the scores evaluating the infiltration of specific cell types produced by the various methods are compared.

Finally, in the fifth section, "Classifying cold and hot tumors", the capacity of the methods to distinguish between "cold" (i.e. with sTIL≤ 10%) and "hot" (i.e. sTIL≥60%) tumors is explored.

Assessment of immune cells on whole slides

Two experienced pathologists (HH and RS) scored itTIL, sTIL and 6 immune cell-types (CD3+, CD4+, CD8+, CD20+, CD68+, FOXP3+) in the International Cancer Genomics Consortium BC cohort using hematoxylin and eosin (H&E) whole slides (WS, n=243), immunohistochemistry (IHC)-stained tissue microarray (TMA, n=254) and WS (n=82).

Regarding the overall infiltration, in line with previous reports [65, 66, 67, 68, 69], a high inter-observer CCC was observed both for sTIL and itTIL (0.84 and 0.85, respectively, Figure 2). The limits of agreement showed a fair relative precision between measurements, and no major constant (intercept) or proportional (slope) drift between the two pathologists (Figure 3).

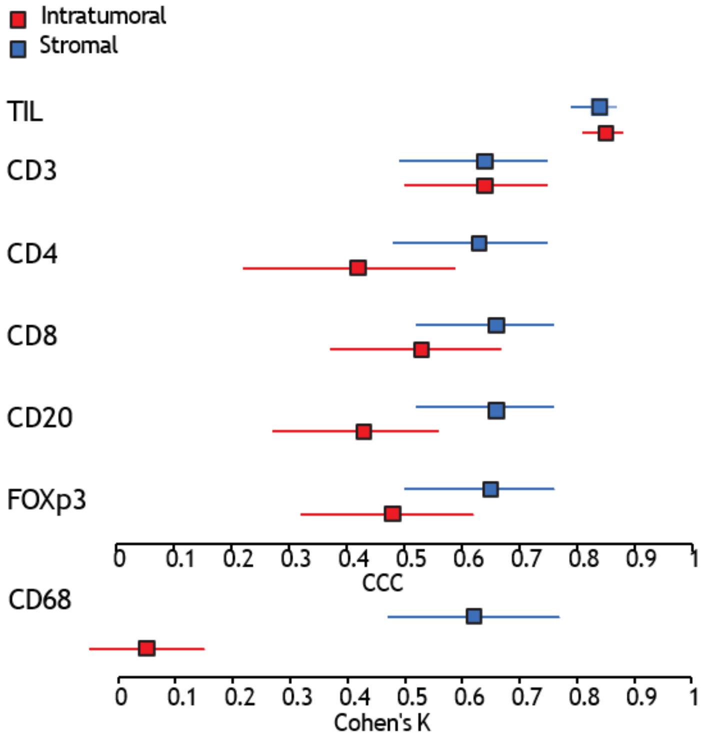


Figure 2*: Forest plots representing estimated concordance correlation coefficients with 95% confidence interval (CI) for each pairing between pathologists (inter-observer agreement) on H&E and Whole Slides (WS). Cohen's K for the macrophages staining (CD68) based on 4 infiltration categories (nil, mild, moderate and severe).*

Concerning cell specific infiltration evaluation, the inter-observer analysis demonstrated a fair CCC for the immunohistochemical (IHC) assessments, where stromal scoring performed overall better and more precisely than intratumoral scoring (Figure 2). All stromal IHC CCC values were lower for immune cell subtypes than for overall TIL assessment on H&E, ranging 0.63-0.66. The intratumoral IHC methods had a CCC below 0.6, except for itCD3 (CCC= 0.63). We therefore only considered the more reliable stromal estimates for further analysis. The Passing-Bablok regression analysis further showed a rather fair slope for the more abundant stromal infiltrated markers (CD3+, CD4+, CD8+), while, the less abundant cells (FOXP3+ and CD20+) show a drifted slope and were therefore deemed less reliable (Figure 3).

To evaluate the contribution of the itTIL and sTIL values to the global infiltration, a "global" TIL score was calculated as the arithmetic mean of the sTIL and itTIL scores for each sample. A good concordance was observed between stromal and global scores (CCC 0.84, Figure 4 a), while, on the other hand, a lower concordance between the intratumoral and global TIL scores was observed (CCC 0.37, Figure 4 b). These results show that immune infiltration in BC is mainly localized in the stromal compartment and is not greatly interfered by intratumoral infiltration.

Since a fair overall concordance between the two pathologists was observed, a unique score for human evaluation, calculated as the mean of the two, was used in the comparison with other methods.

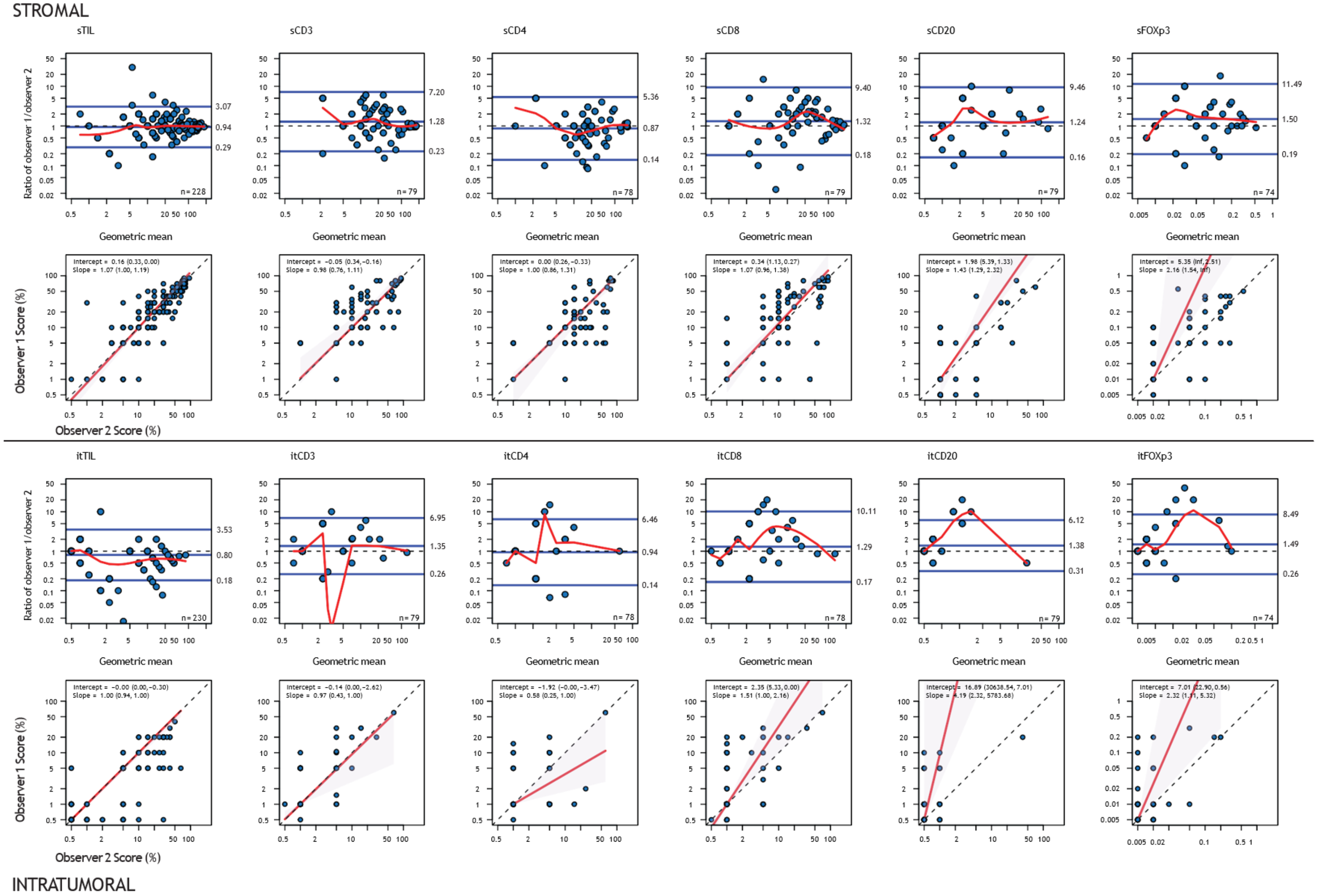


***Figure 3**: For the inter-observer analysis, the stromal and intratumoral values are reported separately. Bland-Altman Plots show the limits of agreement and are used to compare two observers (RS and HH), for the same variable. Passing-Bablok Regression show a nonparametric regression line for all markers (CD3, CD4, CD8, CD20, FOXp3 and CD68). The intercept is interpreted as the systematic bias (difference) between the two observers. The slope measures the amount of proportional bias (difference) between the two observers.*

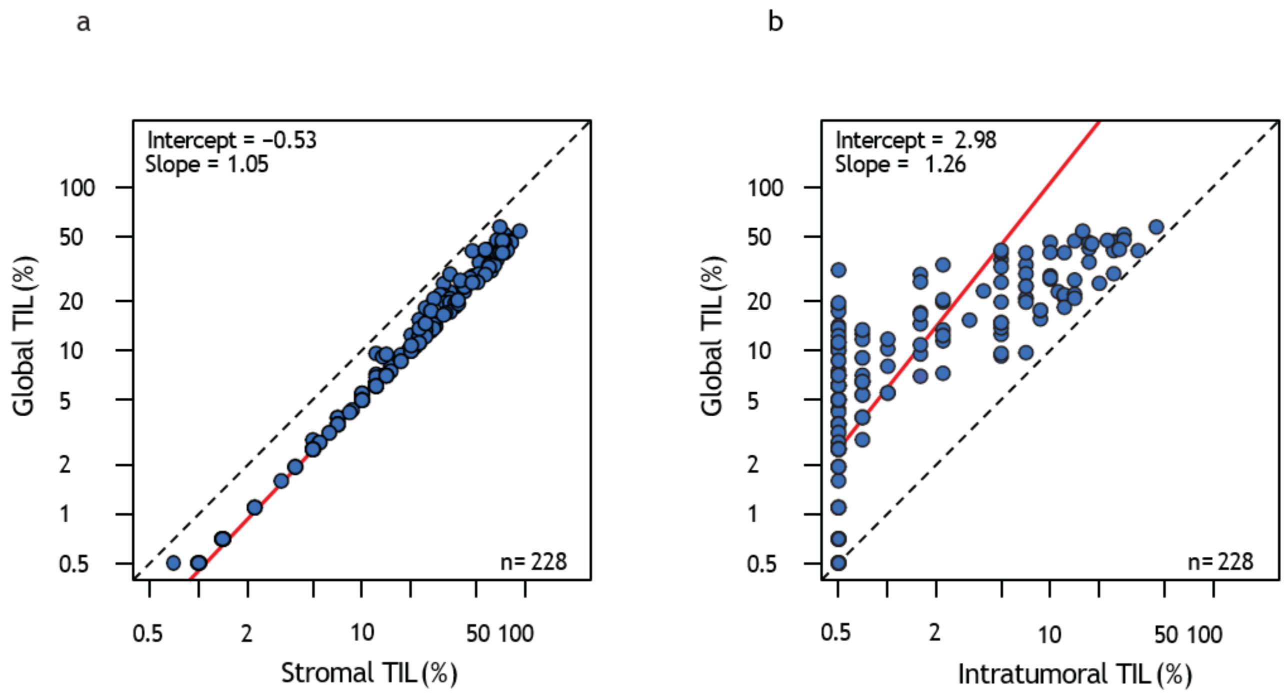


***Figure 4**: The Passing-Bablok regression plot representing the comparisons between the global TIL and sTIL.scores (a) and between the global TIL and itTIL (b), respectively.*

Comparison of tumor immune infiltration estimates between tissue micro arrays and whole slides

To investigate if the immune infiltrate is comparable when assessed on TMA or WS, infiltration of 6 immune cell-types (CD3+, CD4+, CD8+, CD20+, CD68+, FOXP3+) was characterized with both methods.

The immune infiltrate scores were systemically higher on WS as compared to TMAs, as depicted in Figure 5 for all markers but CD68. The CCCs were globally low and ranged between 0.21 for CD4+ and 0.43 for FOXp3+ cells as summarized in figure 6. Concerning CD68, a Cohen’s kappa of 0.33 [0.16; 0.51] was calculated.

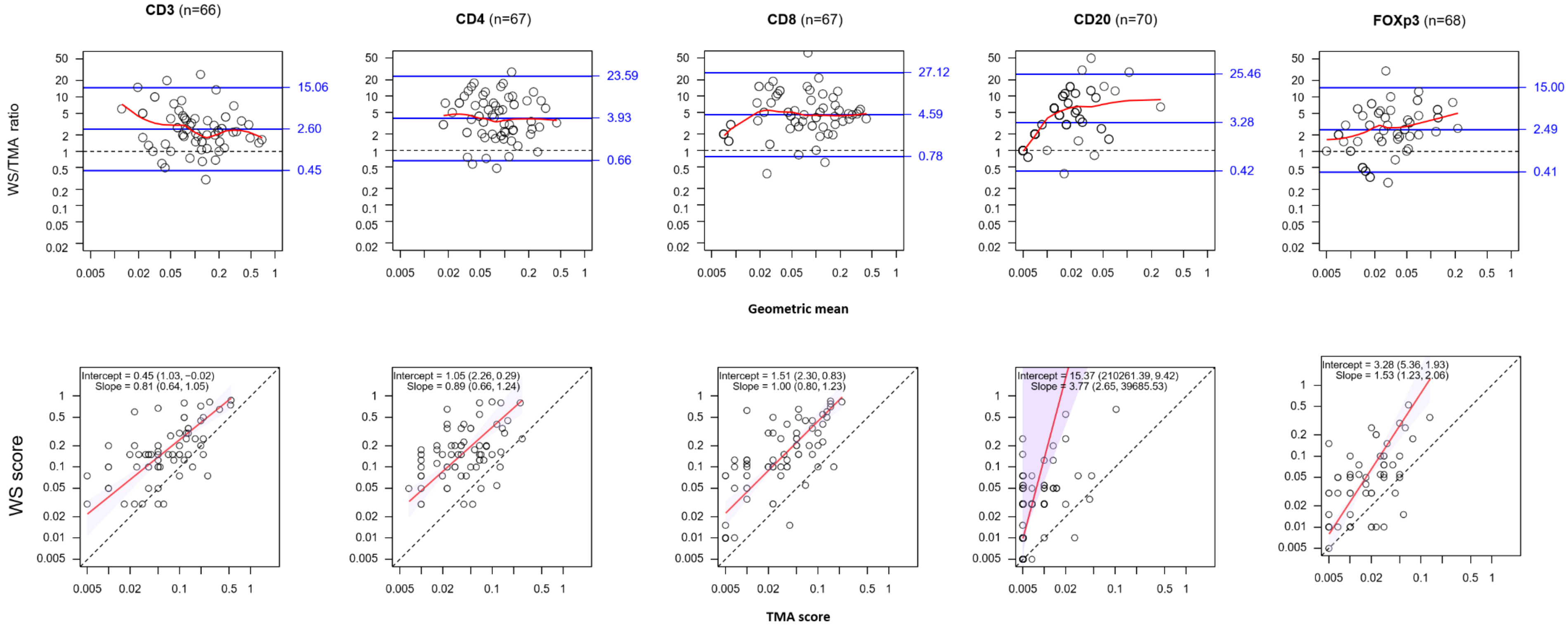


Figure 5*: Bland and Altman diagrams (above) and Passing-Bablok plots (below) comparing the scores generated by the two pathologists on WS and TMAs for, from the left, CD3, CD4, CD8, CD20 and FOXp3.*

.

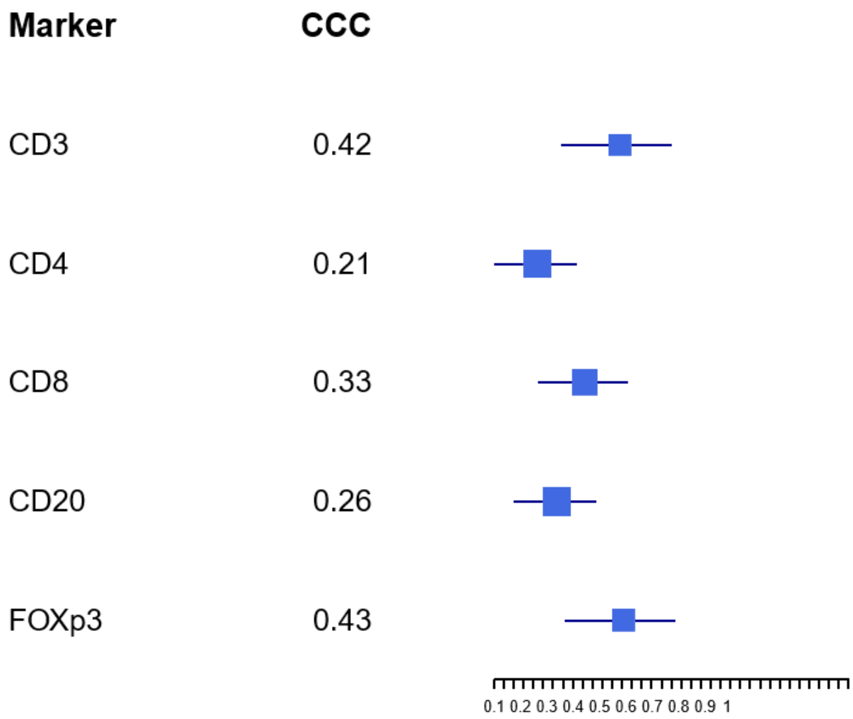


Figure 6*: Forest plots representing estimated concordance correlation coefficients with 95% confidence interval (CI) for each pairing between WS and TMA scores by pathologists.*

WS and TMA were further evaluated through digital pathology. The evaluation of CD3 on WS by digital pathology was performed only on CD3 for technical reasons, the comparison of the estimates generated by digital pathology on WS and TMA for CD3 are showed in Figure 7, confirming the higher level of immune infiltration observed on WS as compared to TMAs by the pathologists and highlighting the spatial heterogeneity of the tumor immune microenvironment. To be noted, the linear subset that appear in the plots may be explained considering that if the sample is low infiltrated, it is more difficult to distinguish the real signal from the noise and produce precise estimations; as consequence observations appear to be aligned when one method scores zeros and the other not. Moreover, should be considered the fact that for very low infiltrated samples, the evaluation of the smaller area through TMAs may prevent to appreciate the presence of few cells that instead are observed in WS.

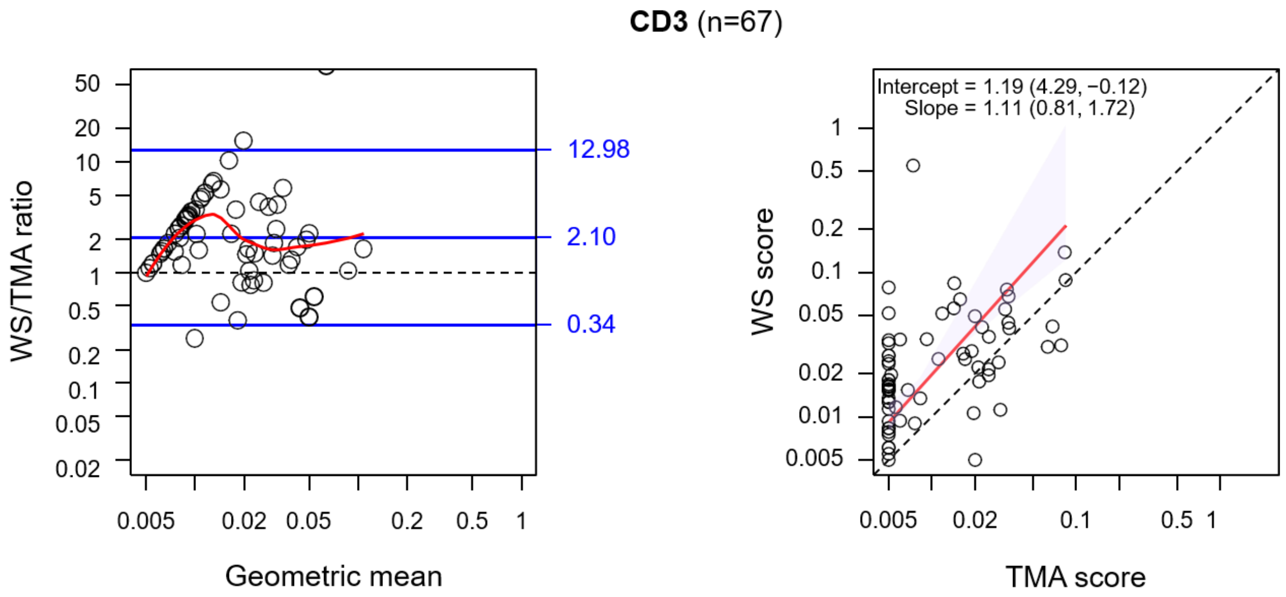


***Figure 7**: Bland and Altman diagram (left) and Passing-Bablok plot (right) showing the comparison between digital scores calculated on WS and TMA.*

Comparing human and digital estimation on TMAs, in line with previous reports [66], a fair concordance was observed when comparing human and digital assessment, as showed in Figure 8 for CD3, CD4 and CD8 (regarding the remaining markers, calculations were impossible to complete since digital pathology scored no or very low infiltration for the lower abundant (CD20+, CD68+, FOXp3+) cells types. Also for these comparisons a linear subset appears for lightly infiltrated samples, confirming, for both human and digital evaluation, the difficulty to make precise estimation when the signal is low.

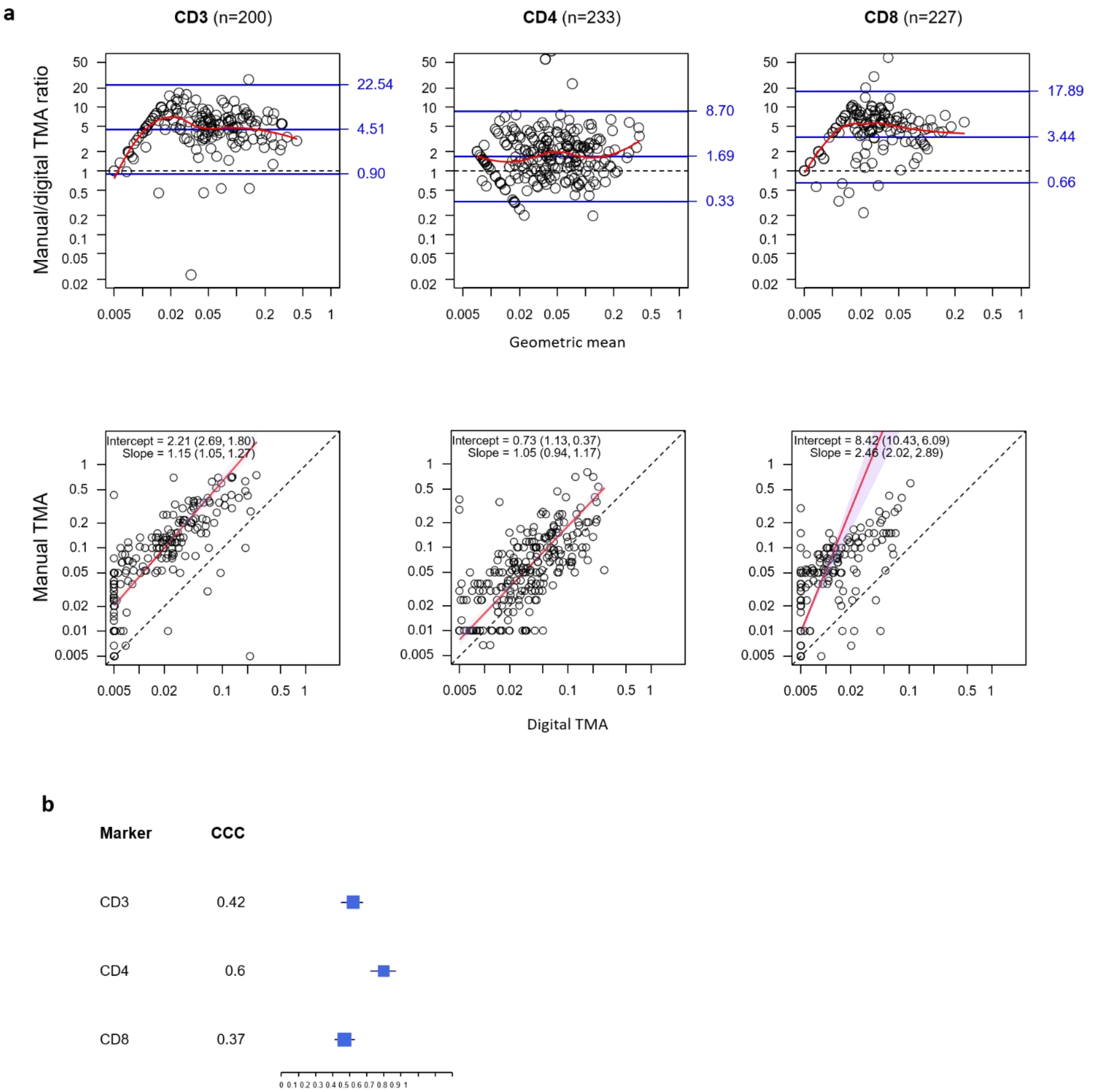


***Figure 8**: **a**) Balnd and Altman diagrams (above) and Passing-Bablok plots (below) showing the comparison of manual and digital TMA scores. **b**) Forest plot representing estimated concordance correlation coefficients with 95% confidence interval (CI) for the comparion of manual and digital TMA scores.*

Comparison of microscopic, transcriptomic and methylomic evaluation of overall tumor immune infiltration

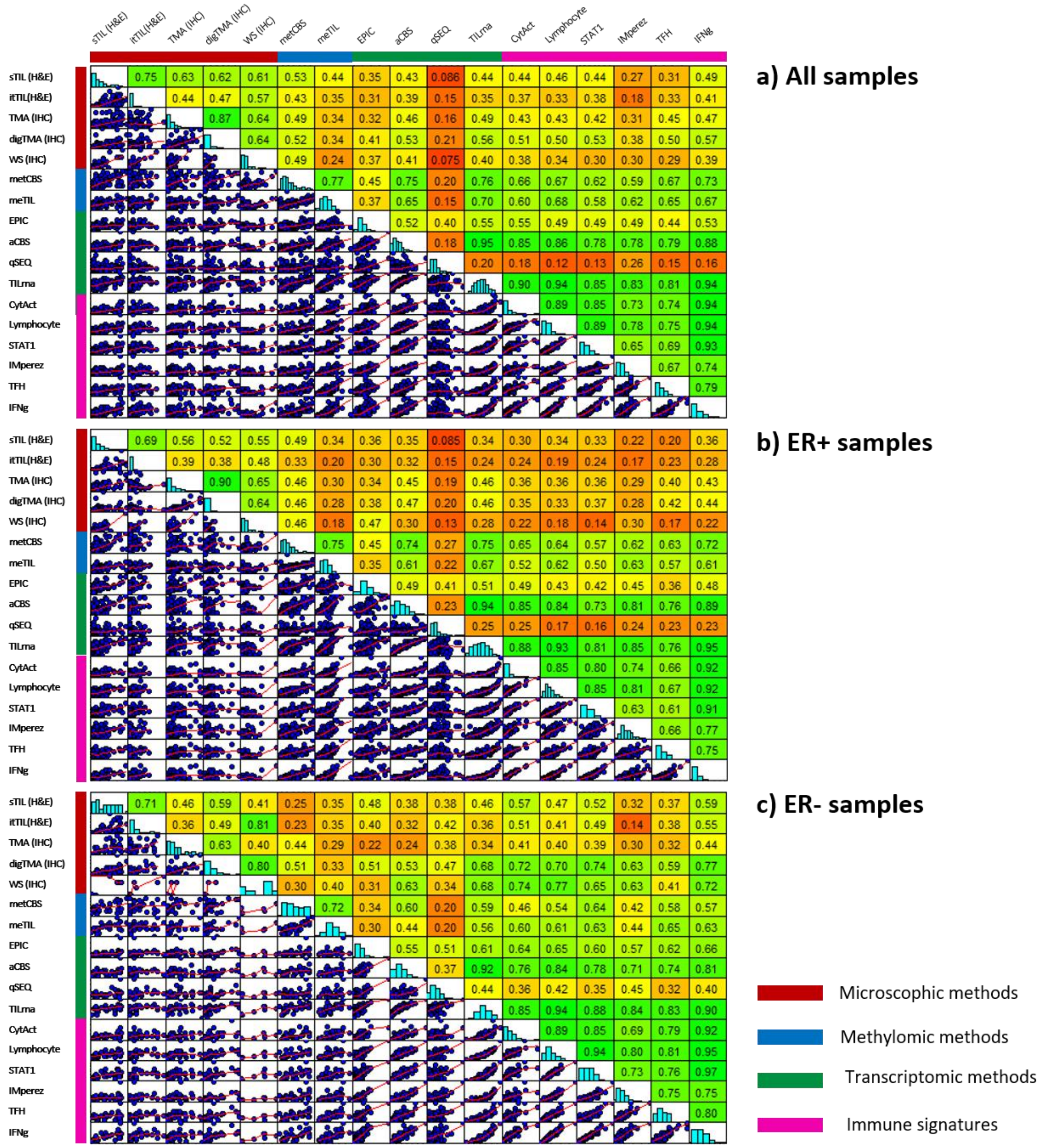
To evaluate if different data types could estimate overall immune infiltration consistently, several microscopic-, transcriptomic- and methylomic-based methods, listed in Table 8, were compared.

Many important observations can be retrieved from this method comparison. First, while the Spearman correlations between microscopic and all methylomic or transcriptomic estimates were weak to moderate, our analysis showed that stromal infiltration correlates better with all other methods, including transcriptomic and methylomic methods, as compared to the intratumoral microscopy scores (Figure 9a). The possibility that higher intratumoral infiltration may lead to more pronounced inflammatory gene expression than stromal infiltration as intratumoral infiltration may have a crucial anti-tumor role as suggested by several studies [70, 71, 72, 73, 74] was investigated, but no higher correlation between intratumoral infiltration and inflammation or lymphocyte associated signatures scores was observed. Secondly, only a fair correlation was observed between stromal assessment on H&E and IHC-stained WS, where the sum of T (CD3+) and B (CD20+) cells were considered (r=0.61). Thirdly, a slightly improvement was observed when the infiltrate was scored with digital pathology (Figure 9a), most probably because both methods do not distinguish between stromal and intratumoral infiltrates. Fourthly, as expected, methods using the same approach showed fair correlations. For example, several transcriptomic estimates were strongly correlated with each other (r >0.85), and methylCIBERSORT and the methylomic TIL score (Jeschke et al., 2017) showed a reassuring agreement (r=0.77). The correlation between transcriptomic and methylomic was variable, but methylCIBERSORT showed good correlations with CIBERSORT (r=0.75) and the TILrna signature (0.76), and similar observations could be made for meTIL and CIBERSORT (r=0.65) and TILrna(r=0.70). Fifthly, among the transcriptomic and methylomic methods, methylCIBERSORT and absoluteCIBERSORT showed the highest correlations with imaging scores, though the correlations were still not impressing with the highest being 0.53 considering sTIL on WS (methylCIBERSORT) and 0.53 with TMAs (absoluteCIBERSORT). These results were however in line with a recently published lung cancer study [75].

Finally, among the methods capable to predict global immune infiltration based on the transcriptome (green label in Figure 9a), absoluteCIBERSORT showed the highest correlations with the various immune gene signatures (pink label, ranging from 0.78 till 0.94), while quanTIseq showed the poorest correlations (ranging from 0.12 till 0.26). Similar analyses were further carried on separately for ER-negative and ER-positive tumors (Figure 9 b and c). The correlations for the microscopic versus methylomic and transcriptomic-based methods were in general slightly higher in the ER-negative compared to the ER-positive subgroup. Nevertheless, compared to all samples, the ER-negative tumors did not necessarily show higher correlations.

Method	Overall infiltration	T cells	CD4+ T cells	CD8+ T cells	Tregs	B cells	Macrophages
TMA manual	CD3+CD20	CD3	CD4	CD8	FOXp3	CD20	CD68
TMA digital	CD3+CD20	CD3	CD4	CD8	FOXp3	CD20	CD68
WS	CD3+CD20	CD3	CD4	CD8	FOXp3	CD20	CD68
WS digital	NA	CD3	NA	NA	NA	NA	NA
Absolute CIBERSORT (aCBS)	T + B cells	CD8, CD4.naive, CD4.memory.resting, memory.activated, follicular.helper, regulatory, gamma.delta	CD4.naive, memory.resting, memory.activated, follicular.helper	CD8	Regulatory	naive, memory	Monocytes, Macrophages.M0, Macrophages.M1, Macrophages.M2
quanTIseq, lsfit (qSEQ)	CD4+ CD8 + Treg +CD19	T.cells.CD4 ,T.cells.CD8, Tregs	T.cells.CD4	T.cells.CD8	Tregs	B.cells	Macrophages.M1, Macrophages.M2, Monocytes
MCP-counter (MCP)	NA	T.cells	Cytotoxic.lymphocytes	CD8.T.cells	NA	B.lineage	Monocytic.lineage
xCell	NA	CD4.memory, CD4.naive, CD4.T.cells, CD4.Tcm, CD4.Tem, CD8.naive, CD8.T.cells, CD8.Tcm, CD8.Tem, Tgd.cells, Th1.cells, Th2.cells, Tregs	CD4.memory, CD4.naive, CD4.T.cells, CD4.Tcm, CD4.Tem	CD8.T.cells, CD8.Tcm, CD8.Tem	Tregs	B.cells, Memory.B.cells, naive.B.cells	Macrophages, Macrophages.M1, Macrophages.M2
EPIC	Bcells, CD4.Tcells, CD8.Tcells	CD4.Tcells ,CD8.Tcells	CD4.Tcells	CD8.Tcells	NA	Bcells	Macrophages
MethylCIBERSORT (metCBS)	CD8, CD19, CD4.Eff, Treg	CD4.Eff ,CD8 ,Treg	CD4.Eff	CD8	Treg	CD19	NA
Cell signatures Davoli et al.	NA	NA	CD4.Tcells	CD8.Tcells	Tregs	B.cells	Macrophages
Cell signatures Danaher et al.	NA	NA	Th1	CD8.Tcells	Tregs	B.cells	Macrophages
Cell signatures , Azizi et al.	NA	NA	CD4 EM	CD8 EM	Treg	B cells	Macrophages
Cell signatures, Tamborero et al.	NA	T EM	T helper cells	Activated CD8 T cells	Regulatory T cells	B cells	Macrophages

Table 8*: Overview of used cell fractions for overall immune infiltration and specific immune subtypes. If multiple cell types are listed for a method, these were summed.*



a) All samples

	sTIL (H&E)	itTIL(H&E)	TMA (IHC)	digTMA (IHC)	WS (IHC)	metCBS	meTIL	EPIC	aCBS	qSEQ	TILrna	CytAct	Lymphocyte	STAT1	IMperez	TFH	IFNg
sTIL (H&E)		0.75	0.63	0.62	0.61	0.53	0.44	0.35	0.43	0.086	0.44	0.44	0.46	0.44	0.27	0.31	0.49
itTIL(H&E)			0.44	0.47	0.57	0.43	0.35	0.31	0.39	0.15	0.35	0.37	0.33	0.38	0.18	0.33	0.41
TMA (IHC)				0.87	0.64	0.49	0.34	0.32	0.46	0.16	0.49	0.43	0.43	0.42	0.31	0.45	0.47
digTMA (IHC)					0.64	0.52	0.34	0.41	0.53	0.21	0.56	0.51	0.50	0.53	0.38	0.50	0.57
WS (IHC)						0.49	0.24	0.37	0.41	0.075	0.40	0.38	0.34	0.30	0.30	0.29	0.39
metCBS							0.77	0.45	0.75	0.20	0.76	0.66	0.67	0.62	0.59	0.67	0.73
meTIL								0.37	0.65	0.15	0.70	0.60	0.68	0.58	0.62	0.65	0.67
EPIC									0.52	0.40	0.55	0.55	0.49	0.49	0.49	0.44	0.53
aCBS										0.18	0.95	0.85	0.86	0.78	0.78	0.79	0.88
qSEQ											0.20	0.18	0.12	0.13	0.26	0.15	0.16
TILrna												0.90	0.94	0.85	0.83	0.81	0.94
CytAct													0.89	0.85	0.73	0.74	0.94
Lymphocyte														0.89	0.78	0.75	0.94
STAT1															0.65	0.69	0.93
IMperez																0.67	0.74
TFH																	0.79
IFNg																	

b) ER+ samples

	sTIL (H&E)	itTIL(H&E)	TMA (IHC)	digTMA (IHC)	WS (IHC)	metCBS	meTIL	EPIC	aCBS	qSEQ	TILrna	CytAct	Lymphocyte	STAT1	IMperez	TFH	IFNg
sTIL (H&E)		0.69	0.56	0.52	0.55	0.49	0.34	0.36	0.35	0.085	0.34	0.30	0.34	0.33	0.22	0.20	0.36
itTIL(H&E)			0.39	0.38	0.48	0.33	0.20	0.30	0.32	0.15	0.24	0.24	0.19	0.24	0.17	0.23	0.28
TMA (IHC)				0.90	0.65	0.46	0.30	0.34	0.45	0.19	0.46	0.36	0.36	0.36	0.29	0.40	0.43
digTMA (IHC)					0.64	0.46	0.28	0.38	0.47	0.20	0.46	0.35	0.33	0.37	0.28	0.42	0.44
WS (IHC)						0.46	0.18	0.47	0.30	0.13	0.28	0.22	0.18	0.14	0.30	0.17	0.22
metCBS							0.75	0.45	0.74	0.27	0.75	0.65	0.64	0.57	0.62	0.63	0.72
meTIL								0.35	0.61	0.22	0.67	0.52	0.62	0.50	0.63	0.57	0.61
EPIC									0.49	0.41	0.51	0.49	0.43	0.42	0.45	0.36	0.48
aCBS										0.23	0.94	0.85	0.84	0.73	0.81	0.76	0.89
qSEQ											0.25	0.25	0.17	0.16	0.24	0.23	0.23
TILrna												0.88	0.93	0.81	0.85	0.76	0.95
CytAct													0.85	0.80	0.74	0.66	0.92
Lymphocyte														0.85	0.81	0.67	0.92
STAT1															0.63	0.61	0.91
IMperez																0.66	0.77
TFH																	0.75
IFNg																	

c) ER- samples

	sTIL (H&E)	itTIL(H&E)	TMA (IHC)	digTMA (IHC)	WS (IHC)	metCBS	meTIL	EPIC	aCBS	qSEQ	TILrna	CytAct	Lymphocyte	STAT1	IMperez	TFH	IFNg
sTIL (H&E)		0.71	0.46	0.59	0.41	0.25	0.35	0.48	0.38	0.38	0.46	0.57	0.47	0.52	0.32	0.37	0.59
itTIL(H&E)			0.36	0.49	0.81	0.23	0.35	0.40	0.32	0.42	0.36	0.51	0.41	0.49	0.14	0.38	0.55
TMA (IHC)				0.63	0.40	0.44	0.29	0.22	0.24	0.38	0.34	0.41	0.40	0.39	0.30	0.32	0.44
digTMA (IHC)					0.80	0.51	0.33	0.51	0.53	0.47	0.68	0.72	0.70	0.74	0.63	0.59	0.77
WS (IHC)						0.30	0.40	0.31	0.63	0.34	0.68	0.74	0.77	0.65	0.63	0.41	0.72
metCBS							0.72	0.34	0.60	0.20	0.59	0.46	0.54	0.64	0.42	0.58	0.57
meTIL								0.30	0.44	0.20	0.56	0.60	0.61	0.63	0.44	0.65	0.63
EPIC									0.55	0.51	0.61	0.64	0.65	0.60	0.57	0.62	0.66
aCBS										0.37	0.92	0.76	0.84	0.78	0.71	0.74	0.81
qSEQ											0.44	0.36	0.42	0.35	0.45	0.32	0.40
TILrna												0.85	0.94	0.88	0.84	0.83	0.90
CytAct													0.89	0.85	0.69	0.79	0.92
Lymphocyte														0.94	0.80	0.81	0.95
STAT1															0.73	0.76	0.97
IMperez																0.75	0.75
TFH																	0.80
IFNg																	

Figure 9*: Methods to assess overall infiltration (A) Matrix plot of Spearman correlations for the methods providing information on overall immune infiltration; Tumor Infiltrating Lymphocytes (TIL) or T cells and B cells (the sum was taken to derive a TIL fraction). (B) Matrix plot for Spearman correlations of the methods providing information on overall immune infiltration in ER positive Tumors and (C) ER negative tumors.*

Abbreviations: sTIL = stromal TIL on H&E, itTIL= intratumoral TIL on H&E, TMA= tissue micro array scored by pathologists, digTMA= tissue microarray scored by Visiopharm (digital analysis), WS = whole slide immunohistochemistry by pathologists, metCBS= MethylCIBERSORT(Chakravarthy et al., 2018), MethylTIL = methyl TIL score(Jeschke et al., 2017), aCBS=absolute CIBERSORT(Newman et al., 2015), TILrna= TIL score based on transcriptome(Massink et al., 2015), qSEQ=quanTIseq(Finotello et al., 2017), MCP = MCP-counter(Becht et al., 2016), EPIC(Racle et al., 2017).

Comparison of microscopic, transcriptomic and methylomic evaluation of cell-specific immune infiltration

Considering only those methods that could provide a quantitative measure of specific immune cell infiltration (see Table 8), several observations can be retrieved by the comparison trough spearman correlation (Figure 10), generated separately for each cell type.

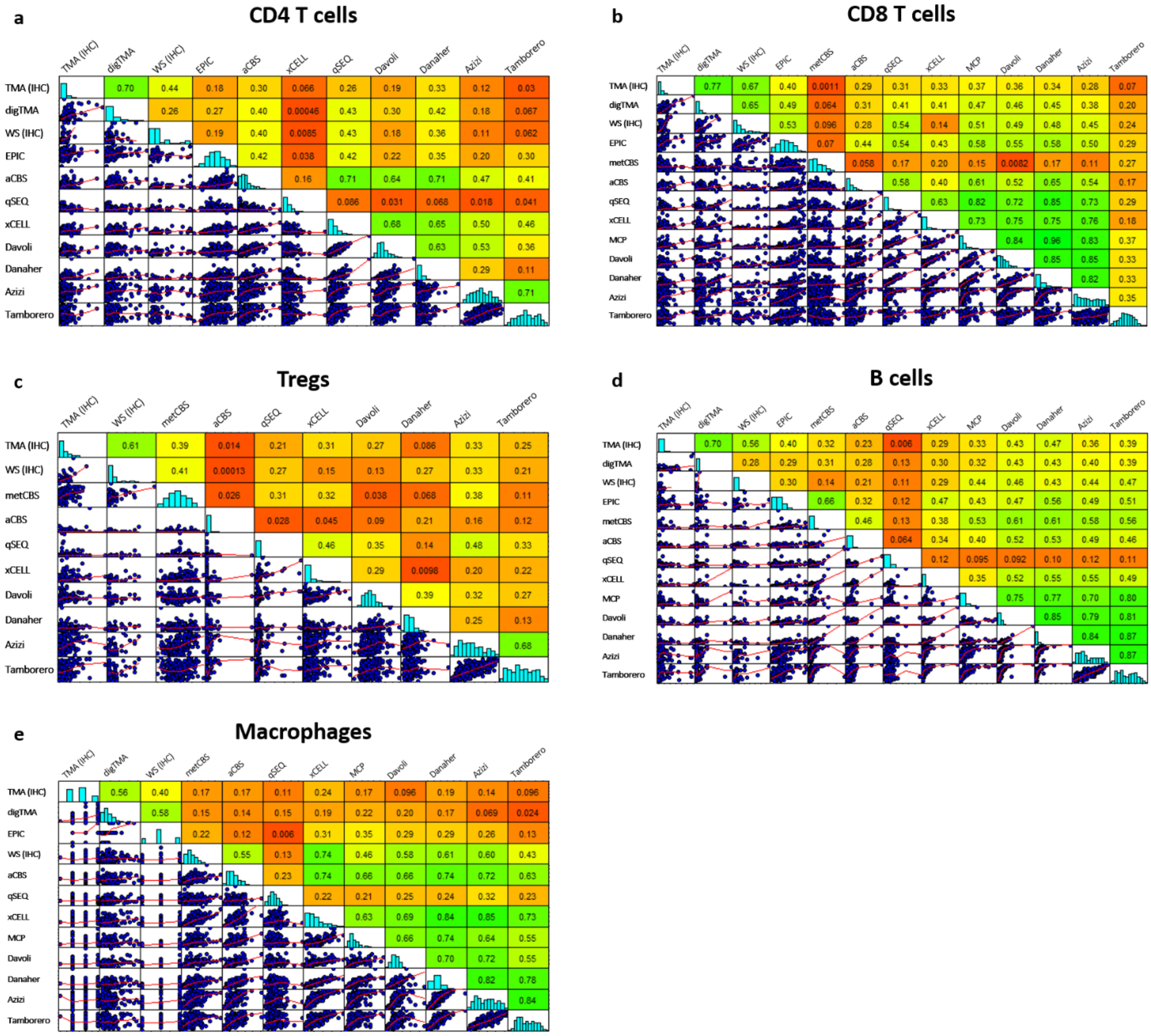


***Figure 10**: The matrix correlation plots for CD4 T cells (a), CD8 T cells (b), regulatory T cells (Tregs) (c), B cells (d), and Macrophages (e).*

Firstly, the correlations between the microscopic evaluations and the transcriptomic and methylomic data were always below 0.60, with CD8+ T cells showing the highest correlation between microscopic and methylomic or transcriptomic methods (Figure 10b), while macrophages the lowest (Figure 10e). Interestingly, no systematic increase in the correlation coefficients when considering WS instead of TMA scores was observed. To be noted, when CCC between methods for specific cell types was considered, very weak concordance (<0.3) was found for all methods (data not shown). Secondly, most omics-derived methods showed large inconsistencies between cell types regarding their correlation with microscopy. For example, while quanTIseq showed a poor correlation with sTIL (0.09; Figure 9a) and CD4+ cells (0.01, WS evaluation; Figure 10b), the correlations for CD8+ were better (0.54, WS evaluation, Figure 10a). Altogether these results underline the fluctuating correlation between the different omics-derived methods and visual assessments, which further varies according to the immune cell-type.

Classifying cold and hot tumors

Finally, considering the classification and thresholds developed by Denker and colleagues that showed to have a relevant clinical impact on the response of BC patients to immunotherapy [20], the ability of the methods to identify cold (sTIL ≤10%) and hot (sTIL≥60%) tumors, was investigated through a receiver operating characteristic (ROC) curve analysis (Figure 11 a, b), which results are summarized in the forest plot depicted in Figure 12.

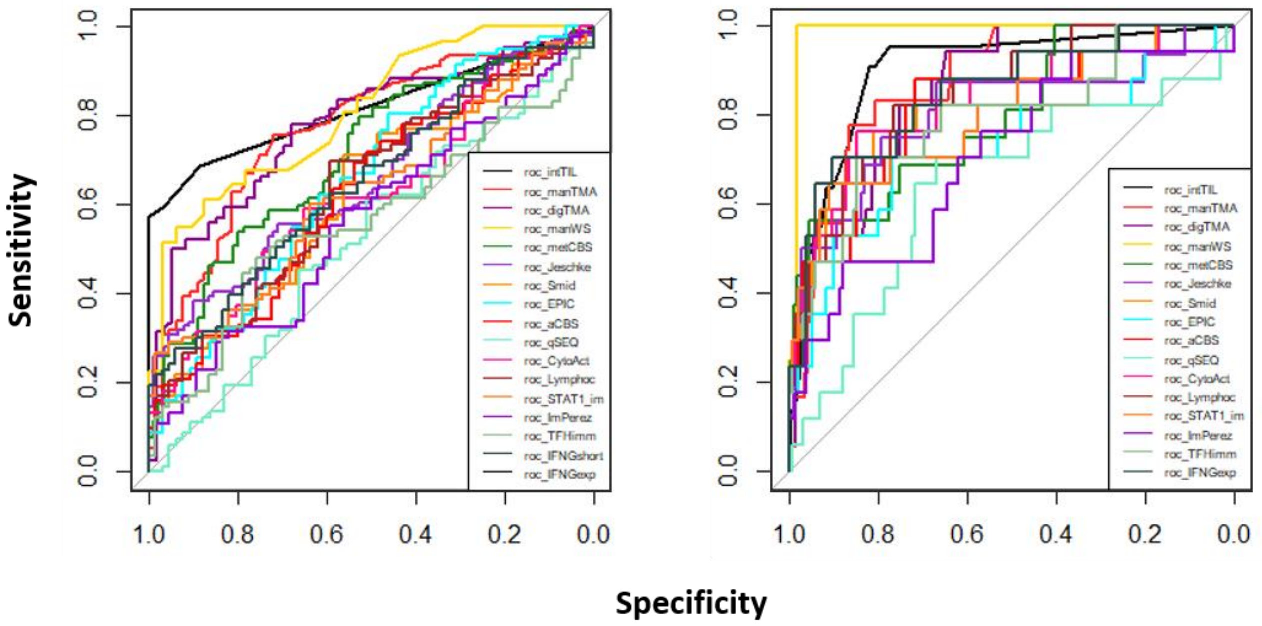


__Figure 11__: The ROC curves for the different TIL methods to classify (__a__) lowly infiltrated/cold, tumors compared to the intermediate/hot tumors, and (__b__) to classify hot tumors compared to the intermediate/cold tumors.

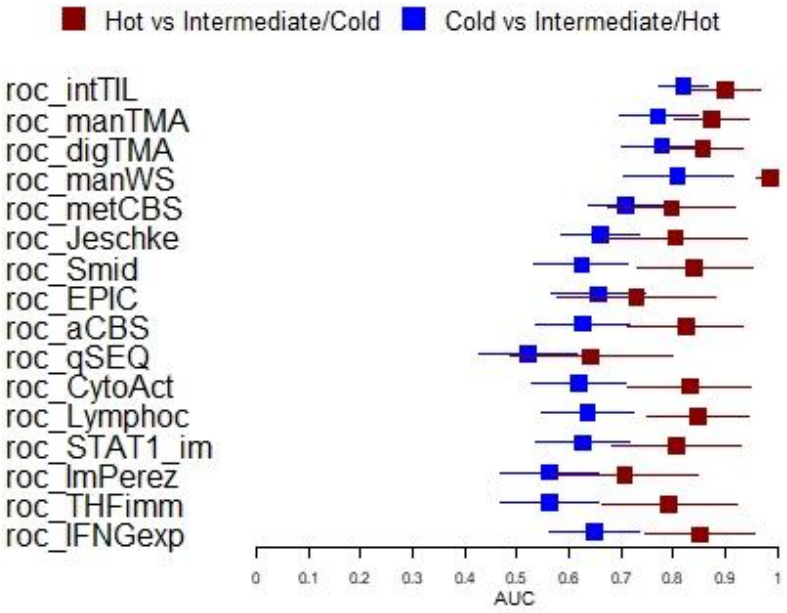


***Figure 12**: Forest plot showing the AUCs (and relative 95%CI) corresponding to curves depicted in Figure 11.*

AUC can be interpreted as the probability of a test to correctly classify an individual; an AUC of 0.5, which is represented in Figure 11 by the grey straight line with 45° slope, indicate a method which is correct only in 50% of the cases and therefore not useful. As showed in Figure 11a and by the blue boxes in Figure 12, most methods performed poorly in recognizing cold versus intermediate/hot tumors. The itTIL score showed the highest area under the curve (AUC, Figure 12 blue boxes), and the methylomic-based TIL scores had a slightly higher AUC as compared to the transcriptomic-based methods, although their confidence intervals are overlapping. On the other hand, ROC analyses showed that most methods were capable to more accurately identify hot tumors as compared to cold tumors. Here, the highest AUCs were still retrieved from the microscopy methods, but also methylomic and transcriptomic-based methods showed fair to high AUCs (Figre 11b, Figure 12 red boxes).

When examining the distribution of the various estimates according the cold, intermediate and hot categories (Figure 13), it is of interest to note for all but not the pathology-based methods, the spread of the various estimates in the cold tumors. Although hot tumors have the highest expression of inflammatory immune signatures, including interferon-g and cytolytic activity, high signature values could be observed in “cold” tumors. An important spread of estimates provided by the omics methods was also observed in hot tumors (Figure 13).

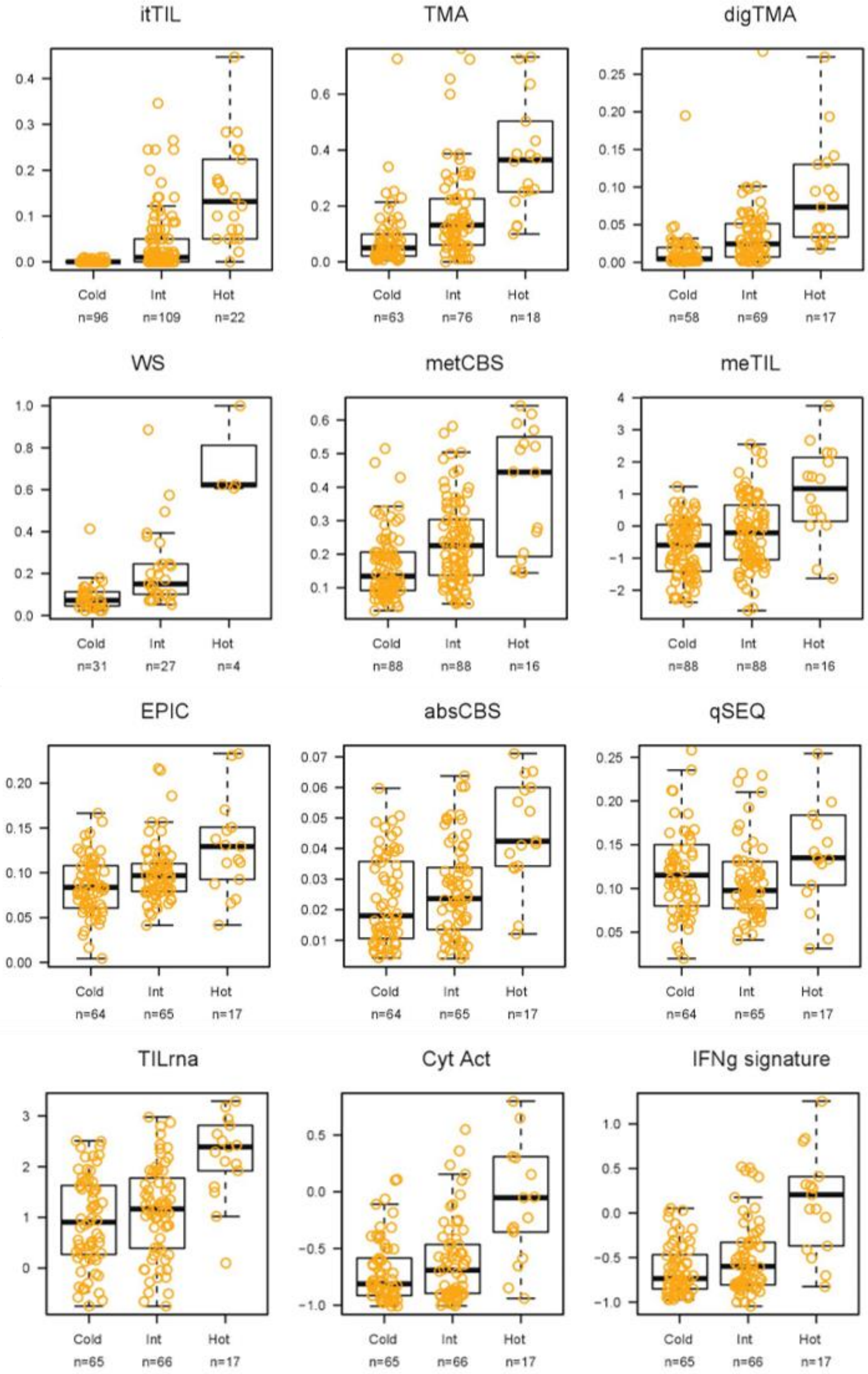


Figure 13*: The distribution of TIL scores for the TIL methods and inflammatory signatures is depicted in (D) for the cold (≤10%), intermediate(11-59%) and hot (≥60%) tumors according to stromal TIL scores.*

Discussion

The aim of this study was to evaluate the reliability of methods already existing in retrieving the composition of TME.

This study demonstrated that methods of the same modality (microscopy, transcriptomic or methylomic-based) show fair correlations when estimating overall infiltration, however correlations drop down when different modalities are compared. Moreover, beyond quantifying overall immune infiltration, specific immune cell types infiltration was quantified using microscopic, transcriptomic and methylomic-based methods. A strong heterogeneity in the correlations between the microscopic and omics-based estimates was observed, as well as between the different omics-based estimates. This work clearly highlights that different transcriptomic and methylomic methods have limitations in estimating immune infiltration, as correlations with microscopy-based methods do not exceed 0.6. Since the pathology assessment of TIL has reached level 1B evidence with new studies confirming its value for patients [15], the fact that correlations are not high may call for extra caution when using non-pathology methods as these may measure different characteristics of the tumor immune infiltrate. These results might be explained by the fact that transcriptomic and methylomic data are not perfectly representing protein expression of immune cells. Moreover, regular bulk transcriptome or methylome analysis does not detect heterogeneity within a sample and ignore the localization of specific cells [76]. To be noted, transcriptomic or methylomic methods may be biased towards a specific cell-state of immune cells, since they are often derived from cells stressed with experimental processes (e.g. tissue digestion or flow cytometry) or from different origin (e.g. peripheral blood or different tumors), and this may partially explain the differences observed with microscopy.

A deeper understanding of immune cell profiles in BC [77, 78, 79] may improve the assessment of immune cells. However inferring which genetic information from bulk sequencing descents from which cell type, might still prove difficult and remain a mere estimation. Nevertheless, the information from bulk sequencing techniques, may guide the profiling of tumors into immunologically and clinically relevant subtypes, as described by a recent study for six immune subtypes [80]. More research focused on the contribution and spatial distribution of specific immune cells is needed to understand beneficial and deleterious immune cell profiles in the context of clinical outcome. Spatially resolved omics methods, measuring genetic and phenotypic diversity, would support advances for clinical studies [81]. In addition, deep learning approaches may be useful, in the future, to identify spatial features of immune-cells identified by pathologists. Integration of deep learning approaches with morphological features identified by pathologists, in conjunction with genomic-derived data will probably be needed to derive a full comprehensive evaluation of the immune-environment in solid tumors.

This work is based on an extensive microscopy-based characterization of a large BC cohort, and this data provided insightful observations that can guide future research using (digital) image analysis. Stromal evaluation of several immune cell types, including regulatory T cells, macrophages and CD8$^+$ T cells, was showed to be reliable with an acceptable concordance observed between pathologists.

Percentages may also be estimated through a digital approach which could be practical for high-throughput analysis in the context of larger clinical cohorts. However, in line with other studies, a lower estimation was systematically observed for digital imaging methods compared to standard pathology estimates [66]. Stromal infiltration showed a higher accuracy by pathologists compared to intratumoral assessment of immune cells. To be noted, small punches of the tumor may lack of some information of infiltration as WS showed overall a higher infiltration than TMA; however, this could partly be explained by the fact that these TMAs were not constructed to specifically investigate the tumor microenvironment. This phenomenon was previously reported and it was concluded that the cores in the TMAs may not be taken close enough to the invasive tumor front, suggesting WS are more informative for assessing immune parameters [82]. In addition, also intra-tumor heterogeneity most probably contributes to the observations. These results therefore call for caution when evaluating immune markers on TMA (or biopsies) and are in contrast with several studies which support that TMAs are reliable for the evaluation of several prognostic epithelial-based tumor markers [83, 84, 85, 86, 87].

The second objective of this project was to evaluate the capability of the methods in identifying hot and cold tumors. Although immune infiltration is a continuous variable, we used here cutoff points previously described to investigate the extreme categories, the hot and cold tumors. In the previous sections was showed that the majority of the methods are much better in recognizing hot as compared to cold tumors. The lower accuracy in identifying cold tumors may be problematic as the majority of the breast tumors will have infiltration above 0% but far below 60%, as shown by Loi et al. [15] in early TNBC. These results should be considered when developing inclusion and stratification criteria, as well as endpoints in the context of clinical trials where patients are treated with immunotherapies. Moreover, despite accuracy for the identification of hot tumors, a higher immune cell presence did not necessarily mean anti-tumor activity, as not all tumor infiltrating T cells harbor a T-cell receptor repertoire with intrinsic capacity to recognize tumor antigens [88, 89] or the tumor specific T cells can be suppressed [90] or exhausted [91]. Therefore, being able to distinguish tumors that are cold by suppression or cold by absence or exclusion represents important ongoing research topic that will support personalized medicine.

An important limitation of this work is represented by the fact that tissue analysis and DNA/RNA isolation were not performed on the exact same area from the tumor, as nucleic acids extractions were performed using the frozen sample and H&E and IHC staining using the FFPE samples. Infiltration may be heterogeneous and this partially explains the correlations not exceeding 0.6. However, multiple cores spread throughout the tumor in the TMA and large sections of the tumors in WS were analyzed, far exceeding the area usually evaluated with biopsies. Moreover, a recent study suggested that a single WS was sufficient to assess the spatial immune heterogeneity of a breast tumor using anti-CD3$^+$, CD8$^+$ and CD20$^+$ stained slides[92]. Nevertheless, correlation with various omics-based methods did not systematically increase when considering WS versus TMA. Another limitation is that the omics-based methods do not consider the localization of the cells in the tumors, which may be crucial as demonstrated in a previous study in which was showed that for instance CD8$^+$ T cells may need to be localized at the core of the tumor for immunotherapy to be effective [93]. In this context, a study suggesting that transcriptomics data might have the potential to derive this spatial information [94] was recently published. Last, this cohort does not provide follow-up or therapy response data. Future studies may test superiority of specific measures and guide precision medicine strategies to enhance clinical response.

In conclusion, these analyses bring out an important heterogeneity in the estimates of immune infiltrates in BC provided by various methods and calls for caution when used in clinical context. At the present day, there is an urgent need for the development of international guidelines to categorize breast tumors according to their immune infiltrate in both a quantitatively and a qualitatively manner. Combining the valuable information from multiple approaches, e.g. the spatial information from pathology and transcriptomic information on cellular activity or the transcriptomic and methylomic information, may elucidate the role of immune infiltration in disease progression in a more accurate manner.

Pseudo-values survival analyses project

Introduction

When performing survival analyses on advanced diseases, usually the end-point of interest is the time to death by any cause. However, if the topic of the study is a non-aggressive disease occurred in middle-aged patients where long follow-ups are available, patients may die for causes not related to the disease. In this case, the event of interest is usually the time to disease progression; the occurrence of death without the occurrence of disease progression (i.e. for causes not related to the disease) prevents the observation of the event of interest. In oncology, besides these circumstances, which are particularly frequent (e.g. breast or colon cancer), researchers may be interested in a particular type of event, which could be masked by the occurrence of other failure event such as local recurrences, distant metastases or other primary tumors in different sites. A usual approach in this context is to focus the attention on the first occurring event, considering the first evidence of treatment failure, because the subsequent clinical management complicates the interpretation of the following patient's clinical history [95]. To investigate treatment effect or prognostic role of clinical and pathological variables, regression models for cause-specific hazards (CSH) for each event through Cox approach, treating competing causes of failure as censored times, is a commonly used method [96, 97]. However, although CSH is useful when investigating biological hypothesis on the disease dynamics, the crude cumulative incidence (CCI) is a more suitable measure when the purpose is to provide an aid to clinical decision making. CCI is the probability of the occurrence of a certain event of interest as the first in presence of competing risks. Regression modeling techniques related to CCI are based on sub-distribution hazard [95, 96]. The most common approaches to study CCI and to perform hypothesis testing are based on non-parametric estimates and on the semi-parametric proportional subdistribution hazard model proposed by Fine and Gray, similar to the Cox model [95]. A different, more general approach is based on transformation models [100,101], such that a function of CCI, the so-called link function, is linearly related to a model predictor as in generalized linear models. An alternative and interesting estimation procedure based on the usage of pseudo-observations has been proposed by Klein and Andersen. This approach is based on the usage of pseudo-values obtained from a jackknife non-parametric estimate of the CCI in a generalized linear model [102]. The most common link function is "complementary log log" [102]. An extension of this approach, regarding the usage of different link functions, has been proposed in paper published by my research group [95]. This approach has many advantages derived from the properties of the pseudo-observations, but above all, it allows to directly interpret regression coefficients as measures of clinical relevance, such as relative risk (RR), cumulative subdistribution hazard ratio (CSDHR), odds ratio (OR) and absolute risk reduction (ARR), through the application of suitable link functions.

The primary aim of this work is indeed the exploration of the capabilities of this method to easily retrieve clinical useful measures in presence of censored data and competing risks, as previously proposed in the paper of my research group [95]. To this goal, the method is applied to a breast cancer dataset, on which survival analyses were performed in order to explore the recurrence dynamics of breast cancer (BC) in another work of my group [103]. The secondary aim of this work is to clarify the role of relevant covariates in BC and especially of body mass index (BMI), which is of particular interest.

High BMI is indeed an important risk factor for many pathologies such as diabetes, cardiovascular and kidney diseases, but has also been linked to the development of breast cancer, particularly the estrogen receptor positive tumors in postmenopausal women [104].

Many studies (reviewed in Jiralerspong and Goodwin [105]) have investigated the role of BMI in BC considering various end-points and, although discrepancies have emerged, which may be explained by the heterogeneity of the studies, most of them have reported an adverse association between elevated BMI and survival.

The work presented in the following pages represents a first step in a wide project that will be carried out by my group after the conclusion of my PhD.

Materials and methods

Dataset

The dataset used for these analyses is composed of patients from the phase III Belgian study, which was designed to compare three treatment regimens: high dosage of epirubicin-cyclophosphamide (HDE), low dosage of epirubicin-cyclophosphamide (SDE) and cyclophosphamide, methotrexate and fluorouracil (CMF), the standard treatment.

The original dataset included 777 patients [106] with age 70 or younger and node-positivity, who were randomly assigned to one of the three treatment regimens between March 1988 and December 1996. The original study did not highlight an advantage in the usage of high dosage of epirubicin-cyclophosphamide over classical CMF in the adjuvant therapy of node-positive pre- and postmenopausal women with breast cancer. In addition, the study confirmed that the response to epirubicin is dose dependent.

Of the original 777 patients, only 734 (95%) had weight and height information and, among these, one subject had missing failure time and was therefore not considered. Because the effect of BMI was of primary interest and the failure time was necessary to the estimation of the cumulative incidence function, the dataset considered for the analyses in this study was composed of 733 patients. Moreover, the complete model included also included treatment, ER, menopausal status and tumor dimension; complete data on these variables was available only for 535 subjects. The characteristics for all 733 patients are shown in Table 9, while for the subset of 535 patients in Table 10; these two tables shows that no particular unbalancing was generated by using the subset with complete data.

Baseline BMI was calculated as the weight in kilograms divided by the square of the height in meters at the beginning of the follow-up and divided in categories following the World Health Organization guidelines [107], although a two classes division, which contraposed underweight and normal-weight patients on one side and overweight and obese on the other side was considered in the following analyses.

Drug dosages were calculated from patients' body surface without any cap. As for the analyses described by Biganzoli et al [103], ER status was re-established with respect of the original study, considering ER-positive tumors as those that show any expression of ER.

Various types of failure event, such as controlateral tumors, distant metastasis, locoregional metastasis and second primary tumors were considered in the original study [106]. However, for these analyses a unique composite event indicating treatment failure was considered, opposed to the competing event of death without evidence of disease (NEDdeath).

Survival analysis

Survival analysis is a group of statistical techniques specifically developed to analyze survival data, which is mainly constituted of random variables T, which represent the time between the entrance of a subject in the study until the observation of the event of interest; this kind of random variables is non-negative and called as "time to event", "failure time" or "survival time".

This kind of data is particular and requires specific analytical techniques. The distinctive trait of this type of variable derives from the way it is generated. While for other types of data, the responses are measured instantaneously and independently of the size of the response, the event time variable is measured sequentially from the beginning of the study, meaning that retrieving large responses (i.e. long times to event) requires more time than smaller ones. As a consequence of this trait is the

***Table 9**: Clinical features regarding the complete dataset composed of 733 patients.*

		All n (%)	BMI<25 n (%)	25≤BMI<30 n (%)	BMI≥30 n (%)
Age (n=733)	<50 y	396 (54)	262 (65)	94 (44)	40 (29)
	≥50 y	337 (46)	141 (35)	119 (56)	77 (71)
Menopausal status (n=732)	Pre	428 (58)	275 (68)	106 (50)	47 (34)
	Post	304 (42)	127 (32)	107 (50)	70 (66)
Positive lymph nodes (n=733)	1-3	432 (59)	250 (62)	122 (57)	60 (51)
	>3	301 (41)	153 (38)	91 (43)	57 (49)
Tumor size (n=614)	<2 cm	238 (39)	141 (42)	69 (38)	28 (28)
	≥2 cm	376 (61)	192 (58)	113 (62)	71 (72)
ER status (n=629)	ER-	166 (26)	98 (28)	37 (20)	31 (31)
	ER+	463 (74)	247 (72)	146 (80)	70 (69)
PgR status (n=627)	PgR-	203 (32)	115 (33)	50 (28)	38 (38)
	PgR+	424 (68)	231 (67)	130 (72)	63 (62)
Neu-Fish (n=330)	non contributive	264 (80)	149 (81)	67 (74)	48 (89)
	weakly amplified	16 (5)	7 (4)	8 (9)	1 (2)
	strongly amplified	50 (15)	29 (16)	16 (18)	5 (9)
Histological subtype (n=293)	Ductal	200 (68)	124 (73)	50 (62)	26 (62)
	Lobular	73 (25)	37 (22)	23 (28)	13 (31)
	other	20 (7)	9 (5)	8 (10)	3 (7)
Grade (n=624)	I	143 (23)	84 (24)	34 (18)	25 (26)
	II	329 (53)	186 (54)	96 (52)	47 (49)
	III	152 (24)	74 (22)	54 (29)	24 (25)
Treatment arm (n=733)	CMF	242 (33)	141 (35)	72 (34)	29 (25)
	SDE	251 (34)	131 (33)	78 (37)	42 (36)
	HDE	240 (33)	131 (33)	63 (30)	46 (39)

***Table10**: Clinical features regarding the subset of 535 patients with complete data.*

		All	BMI<25	25≤BMI<30	BMI≥30
		n (%)	n (%)	n (%)	n (%)
Age (n=535)	<50 y	282 (53)	185 (64)	72 (46)	25 (29)
	≥50 y	253 (47)	106 (36)	85 (54)	62 (71)
Menopausal status (n=534)	pre	305 (57)	194 (67)	80 (51)	31 (36)
	post	229 (43)	96 (33)	77 (49)	56 (64)
Positive lymph nodes (n=535)	1-3	303 (57)	174 (60)	85 (54)	44 (51)
	>3	232 (43)	117 (40)	72 (46)	43 (49)
Tumor size (n=535)	<2cm	208 (39)	126 (43)	58 (37)	24 (28)
	≥2cm	327 (61)	165 (57)	99 (63)	63 (72)
ER status (n=535)	ER-	138 (26)	81 (28)	32 (20)	25 (29)
	ER+	397 (74)	210 (72)	125 (80)	62 (71)
PgR status (n=534)	PgR-	168 (32)	95 (33)	41 (27)	32 (37)
	PgR+	364 (68)	196 (67)	113 (73)	55 (63)
Neu-Fish (n=277)	Non-contributive	220 (79)	120 (79)	60 (75)	40 (89)
	weakly amplified	14 (5)	5 (3)	8 (10)	1 (2)
	strongly amplified	43 (16)	27 (18)	12 (15)	4 (9)
Histological subtype (n=208)	Ductal	143 (69)	91 (74)	33 (60)	19 (63)
	Lobular	53 (25)	26 (21)	18 (33)	9 (30)
	other	12 (6)	6 (5)	4 (7)	2 (7)
Grade (n=465)	I	110 (24)	65 (25)	24 (18)	21 (28)
	II	244 (52)	138 (54)	71 (53)	35 (47)
	III	111 (24)	54 (21)	38 (29)	19 (25)
Treatment arm (n=535)	CMF	179 (33)	106 (36)	52 (33)	21 (24)
	SDE	181 (34)	85 (29)	63 (40)	33 (38)
	HDE	175 (33)	100 (34)	42 (27)	33 (38)

occurrence of censored data. An event time is said censored if it is not observed during the follow-up, but all it is known is that it occurred before the beginning of the study (left censoring, not common) or after the end of the follow-up (right censoring, more common). An important assumption for all survival analyses is that censoring is independent or non-informative, which means that, at each time point, subjects who remain in the study have the same future risk of the occurrence of the event as those subjects that have been censored, as if losses to follow-up were random and thus non-informative.

In the presence of right censoring, survival data may be described by a pair of variables (T, δ), where T is the observed time since the entrance in the study and δ is an indicator of failure, taking values of $\delta = 1$ if the event of interest is observed and $\delta = 0$ if the time is censored. Formally, suppose V indicates the true time to event that one would like to investigate but not always observed, and W is the potential censoring time. Then, the observed time is:

$$T = \min(V, W) \text{ and } \delta = 1 \; if \; V \leq \; W \; or \; \delta = 0 \; if \; V > W$$

that is, $T = V$ only when the observation is not censored. To be noted, here it is assumed that every subject would experience the event if observed for long enough time.

In our study to apply the standard survival analysis a composite event i.e. all disease progression evidence and death without evidence of disease progression should be considered being censoring (i.e lost to follow-up without event or survived without disease progression at the end of the study) considered as independent.

To analyze survival data in basic settings, the following functions are commonly used:

1) Cumulative distribution

$$F(t) = P(T \leq t)$$

It describes the probability of subjects to have the event in the time interval $[0, t]$, which is also called (cumulative) incidence. It is a non-decreasing function such that:

$$F(0) = \; 0$$
$$\lim_{t \to \infty} F(t) = 1$$

2) Survival function

$$S(t) = P(T > t) = 1 - F(t)$$

It gives the probability for a subject to survive without event up to time *t*. It is a non-increasing function such that:

$$S(0) = 1$$
$$\lim_{t \to \infty} S(t) = 0$$

Survival function can be estimated non parametrically from data through the Kaplan-Meier estimator:

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - \; d_i}{n_i}$$

where t_i is the *i*-th event time, n_i is the risk set, i.e the decreasing number of subjects at risk just before t_i , and d_i is the number of events at t_i.

An estimate of the incidence function can be obtained as the complement to 1 of the Kaplan-Meier estimate $\hat{S}(t)$.

3) Hazard function

$$h(t) = \; \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \; \Delta t \,|\, T \geq \; t)}{\Delta t}$$

It describes the instantaneous risk of an event at time t, given that it has not occurred prior to t, and the dynamics of the occurrence of the event, in the sense that describes the changing of the hazard of the occurrence of the event with respect to time.

Although the quantity $h(t)\Delta t$ can be thought as an approximate conditional probability of an event in the interval $[\mathrm{t}, \mathrm{t}+\Delta t)$, the hazard function is not a probability itself, but it can be thought as the rate for which the risk of an event changes with time.

Competing risks

In survival analyses, particularly when concerning a biomedical subject, competing risks often occur. A competing risk is an event whose occurrence prevents the observation of the event of interest [108]. For example, in a study examining the time to tumor recurrence, death due to non-treatment/non-tumor related death is a competing event.

Conventional statistical approaches do not consider competing risks and treat the occurrence of a competing event by censoring. However, this strategy is not correct because it may violate the assumption of non-informative censoring and lead to an overestimation of the incidence of the event of interest [108].

Suitable approaches are adopted to analyze survival data in presence of competing risks. A common strategy, used particularly in biomedical studies, is to focus the attention on the first event occurring, since the following clinical management may complicate the clinical interpretation of the subject's subsequent event history [95]. From a mathematical/theoretical point of view it is useful to think that, at the beginning of the study, each patients is at risk of K failure events at potential times of occurrence $(V_1, \dots, V_k, \dots, V_K)$ called latent failure times, assuming that each event would be occurring if other types of event are removed and the subject is observed for long enough time. Formally, adapting the previous definition of the event time variable, suppose *V* is the true event time one wishes to investigate, given that $V = min(V_1, \dots, V_k, \dots, V_K)$ and *W* is the potential censoring time. Then, the observed time is:

$$T = \min(V, W) = \min(V_1, \dots, V_k, \dots, V_K, W)$$

$$\text{and } \delta = k \; if \; V = V_k \leq \; W \; or \; \delta = 0 \; if \; V > W$$

The most important functions for analyzing competing risks data are:

1) Cause-specific hazard (CSH) function

$$h_k^{cs}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \; \Delta t, D = k \; |T \geq t)}{\Delta t}$$

where D is the cause of failure, i.e. $T = V_k$. It describes the instantaneous rate of occurrence of the event *k* in subjects which are event-free right before time *t*; this means that all subject that have not experienced any event are considered at risk.

2) Subdistribution hazard function

$$h_k^{sd}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, D = k \,|T \geq t \, \cup (T < t \, \cap D \, \neq k)\,)}{\Delta t}$$

It describes the instantaneous rate of occurrence of the event *k* in subjects which had not experienced the event *k*; this means that subjects for which an event different from the *k*-th has occurred are considered at risk.

3) Crude cumulative incidence (CCI)

$$F_k(t) = P(T \leq t, D = k)$$

It represents the probability for a patient to experience the event *k* before time *t* or before the occurrence of another event $\neq k$. The relationship between the subdistribution hazard function and the crude cumulative incidence is $h_k^{sd}(t) = -d \log\{1 - F_k(t)\}/dt$. The CCI can be thought as an analogous of the cumulative distribution function, but in this case $F_k(t) \neq 1 - S_k^*(t)$, where $S_k^*(t) = \, exp\left[- \int_0^t h_k^{cs}(u)du\right]$ is the cause specific (i.e. the survival function where all events but *k* are considered censures) and $\lim_{t \to \infty} F_k(t) \neq 1$, because the occurrence of competing events prevent the observation of the *k*-th event for all the patients.

An important propriety of CCI is that the sum of all the CCIs calculated for each competing events equals the CCI calculated for the composite event defined as any event.

The cumulative incidence function can be estimated through the Aalen-Johansen estimator, which is approximately unbiased [109]

$$\hat{F}_k(t) = \int_0^t \prod_{v<u} \left(1 - \frac{\sum_{D=1}^K dN_k(v)}{Y(v)}\right) d\, \hat{H}_k(u)$$

where $Y(t)$ describe the risk set at time *t, $N_k(t)$* is a counting process that returns the number of individuals that fail for cause *k* and $\hat{H}_k(\cdot)$ is the Nelson-Aalen estimator for the cause-specific cumulative hazard function.

It is important to notice that, although the $F_k(t)$ can be written as $F_k(t) = \int_0^t h_k^{cs}(u)S(u)du,$ cause-specific hazard function has no direct connection with the incidence function and the quantity $1 - F_k(t) \neq exp\left[- \int_0^t h_k^{cs}(u)du\right]$ is difficult to interpret, and surely not as survival function. This means that, while cause-specific hazard function is useful to quantify the instantaneous risk for subjects still alive, it is not adequate to calculate summary measures [110]. On the other hand, a direct connection exists for the subdistribution hazard function, particularly:

$$1- \; F_k(t) = \exp\left[-\int_0^t h_k^{sd}(u)du\right]$$

In our study only two times to event are considered: time to the composite event of disease progression (V_P) and time to NEDdeath (V_n), besides to time to right censoring (W). We are interested in V_P.

$$T = \min(V, W) = \min(V_p, V_N, W)$$

$$\text{and } \delta = P \; if \; V = V_P \leq \; W \; or \; \delta = 0 \; if \; V > W$$

$$h_P^{sd}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \; \Delta t, D = P \; |T \geq t \; \cup (T < t \; \cap D \; \neq P)\;)}{\Delta t}$$

$$F_P(t) = P(T \leq t, D = P)$$

Regression models

The main objective of survival analyses is often to understand the effect of covariates (for example the usage of different treatments or some clinical characteristic of patients, such age, BMI, smoke habits, etc) on the survival expectance.

To this aim, different regression models, based on different functions are available.

One of the most popular is the Cox proportional hazard model, which relates the hazard function to a set of covariates. In the absence of competing risks, it can be written as

$$\log[h(t, \mathbf{x})] = \log[h_0(t)] + \boldsymbol{\beta}^T\mathbf{x}$$

where h_0 indicates the baseline hazard function, $\boldsymbol{\beta}$ the regression coefficient vector and **x** the covariate vector. Because of the logarithmic transformation, the covariates have a multiplicative effect on the hazard function and regression coefficients are interpreted as log-hazard ratios (i.e. hazard ratios log-transformed).

In the presence of competing risks, the hazard function can be substituted with the CSH function or the SDH function. Considering the vector $\mathbf{x_0}$ as baseline level for covariates, two different models can be defined:

1) Cause-specific hazard models

$$\log[h_k^{cs}(t, \mathbf{x})] = \log[h_k^{cs}(t, \mathbf{x_0})] + \; \boldsymbol{\beta}_{h^{cs}}^T(\mathbf{x} - \mathbf{x_0})$$

which has been suggested to be more useful when one is interested in studying the etiology of diseases [111].

2) Subdistribution hazard models

$$\log\left[h_k^{sd}(t, \mathbf{x})\right] = \log\left[h_k^{sd}(t, \mathbf{x_0})\right] + \; \boldsymbol{\beta}_{h^{sd}}^T(\mathbf{x} - \mathbf{x_0})$$

which allow to extend the results of the effect of the covariate on the CCI. For this reason, the subdistribution hazard model is more useful in predicting individual risk [111] and prognosis [112] and thus is better suited to develop clinical prediction models and risk-scoring systems [108].

A different and more general approach, related to transformation models, has been proposed by Fine [113, 114] to allow inference directly from the CCIs. Particularly, these are nonlinear models which use a linear relationship between covariates and a nonlinear transformation *g( )* of the CCI.

The model can be written as:

$$g[F_k(t,\mathbf{x})] = \alpha(t) + \boldsymbol{\eta}^T\mathbf{x}$$

where $\alpha(t)$ is the failure probability of the reference category, **η** is the regression coefficient vector and **x** is the covariate vector.

In our study the model should be

$$\mathrm{g}[\mathrm{F_P}(\mathrm{t},\mathbf{x})] = \alpha(\mathrm{t}) + \boldsymbol{\eta}^{\mathrm{T}}\mathbf{x}$$

where $\mathbf{x}$ is the vector of patient covariates: treatment (dummy variables $\mathrm{x_{CS}}$ i.e.SDE treatment vs CMF, $\mathrm{x_{CH}}$ i.e. the HDE treatment vs CMF), BMI ($\mathrm{x_{BMI}}$ i.e. overweight and obese vs normal weight), hormonal status ($\mathrm{x_{ERpre}}$ i.e. ER positivity and premenopausal status vs ER negativity, $\mathrm{x_{ERpost}}$ i.e. ER positivity and postmenopausal status vs ER negativity), tumor dimension ($\mathrm{x_{dim}}$, i.e. tumor with diameter greater than 2 cm vs smaller tumor) , first order interactions between variables and time dependent effects.

In particular when $\mathrm{g}(\,) = \log\left(-\log\left(\mathrm{F_P}(\mathrm{t},\mathrm{x})\right)\right)$ the proportional subdistribution hazard model can be retrieved.

Pseudo-values

Pseudo-values were used in regression analyses.

Pseudo-values were proposed by Andersen [115] as an extension of the jackknife method introduced by Miller [116], to be used in generalized regression analyses for survival data.

In the original proposal by Fine, the estimation procedure does not allow to consider time-varying effects. Recently, Klein and Andersen [114] have proposed an alternative estimating procedure based on pseudo-values derived from the CCI.

The definition of pseudo-values is the following:

Let X_1, ..., X_n be independent and identically distributed survival time random variables and $\hat{\theta}(X)$ be an unbiased (or approximately unbiased) estimator of the parameter $\theta = \theta(X) = E[\phi(X)]$, for a certain function ϕ. For each X_j the pseudo-observation is defined by

$$\hat{\theta}_j(X) = n\hat{\theta}(X) - (n-1)\hat{\theta}^{-j}(X)$$

where $\hat{\theta}^{-j}(\cdot)$ is an estimator similar to $\hat{\theta}_j(\cdot)$ based on the observations i ≠ j.

Practically, the pseudo-observation $\hat{\theta}_j$ may be interpreted as the contribution of the *j*-th observation to the estimate of $E[\phi(X)]$ based on the entire sample of size *n*.

Andersen's intuition of using pseudo-values in regression comes from the fact that the expected values of the pseudo observation $\hat{\theta}_j(X)$ as function of the covariates Z_j is equal to the conditional mean $\theta_j(Z_j)$, where $Z_j = [Z_{j1}, \dots, Z_{jp}]^T$ is the covariate vector for subject *j* and $\theta_j(Z_j) = E_x[\phi(X)|Z_j]$.

Pseudo-values in competing risks survival analyses

In competing risks survival analyses, the parameter $\theta = \theta(X) = E[\phi(X)]$ to be estimated is the cumulative incidence function for the specific cause *k* $F_k(t)$ and the $\phi(\cdot)$ function is given by:

$$\phi(X) = \phi_{tk}(X) = 1[X \le t, D = k]$$

$\hat{F}_{jk}(t)$ is the unbiased estimator of $F_k(t)$, as described in [126].

The *j*-th pseudo-value corresponding to $F_k(\cdot)$ at time t is then given by

$$\hat{\theta}_{jk}(t) = n\hat{F}_k(t) - (n-1)\,\hat{F}_k^{-j}(t), \qquad k = 1, \dots, K$$

In our model the crude cumulative incidence is $\widehat{F_P}$ as above described.

In practice, pseudo-values were calculated as following:

- The cumulative incidence function for the composite event at 15 years was calculated through the *cuminc()* function in R package *cmprsk*
- To allow good performance of the program without losing information, estimates were then extracted at specific timepoints selected (on complete dataset) so that at least 10 failure events occurred in each interval as suggested by Ambrogi et al. [95]
- Pseudo-values were then calculated for each subject at each timepoint as the estimate from the total dataset minus the estimate obtained without that subject, as described above

Since 32 timepoints were considered, 23,456 pseudo-values were calculated when considering the total dataset (n=733) and 17,120 pseudo-values when considering the subset with complete data (n=535).

Regression models based on pseudo-values

Regression models based on pseudo values has been used to overcome some limitation of the method based on the hazard function.

Consider the generalized linear model

$$g(\theta_j) = \alpha + \boldsymbol{\beta}^T \mathbf{Z}_j$$

where $g(\cdot)$ is some link function and $\theta_j(\mathbf{Z}_j) = E_x[\phi(X)|Z_j]$.

Andersen et al. [115] proposed to replace the function $\phi(\cdot)$ by a pseudo-value and then estimate the unknown parameters through the generalized estimation equation (GEE).

To be noted, the parameter θ_j may be multivariate, that is $\theta_j = [\theta_{j1}, \dots, \theta_{jL}]^T$, i.e. for each subject, there is a distinct parameter for each timepoint. Thus, for each θ_{jl}, *l* = 1 , … , L , one may specify a model as

$$g(\theta_{jl}) = \alpha_l + \boldsymbol{\beta}^T \mathbf{Z}_j$$

Where the notation α_l indicates that the intercept may depend on the time t_l.

As an example the proportional subdistribution hazard model can be obtained considering complementary log log as link function:

$$\log(-\log(\theta(t)) = \log(-log(F_K(t))) = \alpha_l + \boldsymbol{\beta}^T \mathbf{Z}_j$$

which is Fine and Grey subdistribution hazard model. Other link functions are described below.

Parameter estimation

Following the proposal by Andersen, generalized estimating equations (GEE) were used to estimate regression parameters.

The GEE approach was developed by Liang and Zeger to estimate the parameters of a generalized linear model with a possible unknown correlation between the response variables [117], which may arise with longitudinal studies or when data is collected as repeated measures.

When analyzing data with an intrinsic correlation, if the parameter estimation is done without accounting for this correlation, the estimates could result biased and inefficient. GEE estimates are more correct and efficient because the method allows to specify a correlation structure.

GEEs belong to a class of regression techniques that are called semiparametric because they rely on specification of only the first two moments. They constitute an alternative to the likelihood–based generalized linear models which are more sensitive to variance structure specification [118]. Their usage is common in large epidemiological studies, especially multi-site cohort studies, because they can handle many types of unmeasured dependencies between outcomes.

An attractive feature of GEE method is that, under non strictly regularity conditions, its estimates are consistent even when the covariance structure is not correctly specified [117].

After defining the model of interest, it is possible to specify estimating equations which are consistent estimators of regression coefficients **β**.

The generic form of an estimating equation is:

$$\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}\right)^T \mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}) = 0$$

where **μ** is a vector of length *n* of the mean expected responses, whose element $\mu_i = E(Y_i)$ is given by $g^{-1}(\boldsymbol{\beta}^T \mathbf{x}_i)$, and **V** is the working covariance matrix, i.e. the element accounting for the covariance.

The GEE can be solved through the following iterative procedure:

- Step 0: Computing an initial estimate of **β,** $\widehat{\boldsymbol{\beta}}^{(1)}$**,** using for example GLM
- Step 1: Compute the working covariance matrix
- Step 2: Update $\widehat{\boldsymbol{\beta}}^{(step)}$ with:

$$\widehat{\boldsymbol{\beta}}^{(step+1)} = \widehat{\boldsymbol{\beta}}^{(step)} - \left[\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}\right)^T \mathbf{V}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}\right]^{-1} \left[\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}\right)^T \mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right]$$

- Step 3: Repeat steps 1 and 2 until convergence or reaching the maximum number of iterations.

In the case of pseudo-values the GEE to estimate regression parameters can be rewritten as:

$$\sum_{j=1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\beta}} g^{-1}(\boldsymbol{\beta}^T \mathbf{Z}_j)\right)^T \mathbf{V}_j^{-1} \left(\widehat{\boldsymbol{\theta}}_j - g^{-1}(\boldsymbol{\beta}^T \mathbf{Z}_j)\right) = 0$$

where $\widehat{\boldsymbol{\theta}}_j = \left(\hat{\theta}_{j1}, \dots, \hat{\theta}_{jl}, \dots, \hat{\theta}_{jL}\right)$ is the vector of pseudo-values for subject *j* at each time point, while Z_j is the covariate matrix for subject *j* including the spline bases $Z_{j1}, \dots, Z_{jL}$ and V_j is the working covariance matrix.

In their original proposal, Klein and Andersen [102] proposed three feasible covariance matrices: the identity matrix; an exact working covariance, which cannot be easily obtained in standard softwares; and the empirical working covariance matrix. They did not observe a substantial difference in the usage of the three possible matrices and therefore the simple independence working covariance matrix has been used in this work.

QIC

To compare and select models, and particularly to choose the number of knots of the spline functions, the quasi-likelihood information criterion (QIC) was used.

QIC was proposed as an extension of the well-known Akaike's Information Criterion (AIC) [119] by Pan [120]. AIC is common approach that measures the quality of an estimation, taking into account the goodness of fit and the complexity of the model (which can be though as the number of the covariates considered).

Given a model *M* with *m* estimable parameters, its AIC can be defined as:

$$AIC_M = -2l\left(\widehat{\boldsymbol{\beta}}_M\right) + 2m$$

where $l\left(\widehat{\boldsymbol{\beta}}_M\right)$ is the value of the maximized log-likelihood for the model *M*, and *2m* is a penalty term which account for the number of parameters considered.

Since AIC is based on likelihood and the asymptotic properties of the maximum likelihood estimator, it cannot be applied to GEE because in this approach no distribution is assumed and therefore no likelihood can be computed.

Pan's proposal consist in substituting the likelihood function $l\left(\widehat{\boldsymbol{\beta}}_M\right)$ with the quasi-likelihood function constructed from the estimating equation [121], along with a modification of the penalty term.

Pan's QIC can be therefore defined as:

$$QIC_M = -2Q\left(\widehat{\boldsymbol{\beta}}_M\right) + 2\, trace(\Gamma\boldsymbol{\varphi})$$

where $Q\left(\widehat{\boldsymbol{\beta}}_M\right)$ is the value of the quasi-likelihood under the independence assumption, computed on the GEE estimate of $\boldsymbol{\beta}$, while the second term represent the effective degrees of freedom of the model taking into account the clustered structure of the repeated observation for each subject. For the technical details please refer to the original publication of Pan.

In practice, the $Q(\widehat{\boldsymbol{\beta}}_M)$ value was obtained as

$$Q(\widehat{\boldsymbol{\beta}}_M) = -0.5\,(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^2$$

where $\boldsymbol{\theta}$ is the vector containing the observed pseudo-values for each patients at each time point considered and $\widetilde{\boldsymbol{\theta}}$ is the vector containing the corresponding prediction by the regression model. On the other hand $trace(\Gamma\boldsymbol{\varphi})$ is the trace of the matrix obtained by multiplicating the naïve variance matrix Γ with the robust variance matrix $\boldsymbol{\varphi}$, both calculated by GEE.

Splines

The proposed method allows to evaluate the presence of time-varying effects by including interactions terms between covariates and time [102]. To include a reasonable number of regressors in the model, the usage of a smoothing function (for example a spline) is recommended to model the failure probability in the reference class. To this aim a common procedure, used in this work, is to include into regression model spline functions.

Splines are particular piecewise polynomials used to approximate complex curves. The domain of the variable of interest (in this case time) is divided in intervals delimited by the so called knots and allowing the curve to be a different polynomial of degree *d* in each interval, although the condition is imposed that at each knot the value of polynomials and their derivatives up to the *d-1* order agree.

The simplest spline function is a linear spline function that can be represented as a broken line with angles at the knots. However, although linear splines are simple and can be used to approximate many common relationships, they are not smooth and do not fit well highly curved functions.

Cubic splines have been found to have useful properties and can fit sharply curved shapes.

With *k* knots there are *k+1* cubic polynomials and *4(k+1)* coefficients. Imposing the 3k continuity conditions up to the second derivative, the number of independent coefficients to be estimated is restricted to k+4. Thus, chosen *k* knots $\xi_1, \dots, \xi_k$, each cubic spline with can be expressed as a linear combination of a basis of k+4 independent spline functions.

The simplest spline basis is the "truncated power" one, with elements: $1, x, x^2, x^3, (x - \xi_1)^3_+, \dots, (x - \xi_k)^3_+,$

where:

$$(x-\xi_i)_+^3 = \begin{cases} 0 & if\ x < \xi_i \\ (x-\xi_i)^3 & if\ x \geq \xi_i. \end{cases}$$

The choice of the number and position of the knots is main characteristic of the spline. Concerning the position, has been suggested [125] that truncated power splines shapes is not lightly influenced by the knot's position and the strategy used is to place knots in the percentiles of the time to event distribution: the outer knots are placed at the 5th and 95th percentiles and the remaining at equidistant percentiles.

Cubic B-spline bases are more complex: bases are defined by a recursive formula and each basis function spans on 4 consecutive intervals [122]. Additionally, "boundary knots" need to be specified, even if they are generally placed at the minimum and maximum of the observed values. B-splines are in principle more numerically stable than truncated power splines.

The difference between truncated power and B-Splines basis is the definition of boundary knots, while inner knots are placed equidistantly as for truncated power.

In this study, cubic B-splines and Restricted Cubic Splines (RCS) were taken into consideration.

To overcome the problem of the bad behavior of cubic splines in the tails i.e. before the first knot and after the last, a linear restriction in these regions is often adopted.

The generic RCS function with *k* knots ξ_1 … ξ_k, can be written as:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1}$$

where $X_1 = X$ and for $j = 1, \ldots, k-2$,

$$X_{j+1} = (X - t_j)_+^3 - \frac{(X-\xi_{k-1})_+^3(\xi_k - \xi_j)}{(\xi_k - \xi_{k-1})} + \frac{(X-\xi_k)_+^3(\xi_{k-1} - \xi_j)}{(\xi_k - \xi_{k-1})}.$$

To include the spline function for time, the model can be therefore rewritten as:

$$g(\theta_{jl}) = \boldsymbol{\delta}^T \mathbf{b}_l + \boldsymbol{\beta}^T \mathbf{Z}_j$$

where $\mathbf{b}_l$ is the vector of the bases of the spline and $\boldsymbol{\delta}$ is the corresponding vector of regression coefficients.

The model can be furthermore extended to include time-varying effects by adding interaction terms between covariates and spline bases ($\mathbf{s}_{b_l x_j}$):

$$g(\theta_{jl}) = \boldsymbol{\delta}^T \mathbf{b}_l + \boldsymbol{\beta}^T \mathbf{Z}_j + \boldsymbol{\xi}^T \mathbf{s}_{b_l x_j}$$

B-splines were modelled through the *bs()* function, while the RCS with the *rcs()* function, implemented respectively in the *splines* and *rms* R packages.

As the number of the knots is more crucial than their position, we focused our attention in the choice of the latter.

As restricted and natural cubic splines with different number of knots are not hierarchical is not possible to compare different models with statistical formal tests and thus inflation criteria, in these case we used QIC, need to be used.

As truncated powers are more easily to implement and to interpret, when results of the truncated power smoothed hazard function did not show particularly differences in performances (measured through the QIC discussed above) compared to B-splines, truncated power bases functions results were reported.

Link functions

One of the most relevant features of this approach is the possibility to easily retrieve clinically useful measures regarding the impact of covariates by applying specific link functions and eventually exponentiating the estimated coefficients.

For the sake of clarity, let $\mathbf{x}_1$ and $\mathbf{x}_0$ be two covariate vectors, with $\mathbf{x}_0$ the reference category and $F_r(\mathbf{x}_0, t) \geq F_r(\mathbf{x}_1, t)$ for any t (i.e. $\mathbf{x}_1$has a protective effect, since the relative incidence function is less than $F_r(\mathbf{x}_0, t)$); moreover it is supposed that no time-dependent effects are present and therefore the model

$$g(\theta_{jl}) = \boldsymbol{\delta}^T \mathbf{b}_l + \boldsymbol{\beta}^T \mathbf{Z}_j$$

is used.

By applying a different link function *g(·)* it is possible to interpret the estimated coefficients (eventually after exponentiation, depending on the link function used) as measures of the clinical impact of the covariates.

The measures taken into consideration in this project, as functions of the CCIs, are showed in Table 11, along with the link functions *g(·)* to be used in the transformation model to retrieve the estimates of the measures directly from the coefficients.

Applying the cloglog link function allows to interpret a regression coefficient, after exponentiation, as the cumulative subdistribution hazard ratio (CSDHR), which is the same quantity estimated by the Fine and Gray model and can be therefore be compared.

The relative risk (RR), which represents the reduction of incidence of patients with covariates x_1 with respect to patients have covariates x_0, is obtained by applying the log link function. From the RR is it possible to obtain the relative risk reduction (RRR) which can be thought as the proportional reduction of risk for patients having covariates x_1 with respect to the reference and can be calculated as 1-RR.

***Table 11**: Definition of the clinically useful measure in term of CCI. The link functions to be used in the model and the coefficient transformation to obtain the corresponding clinically useful measures are showed. CSDHR, cumulative subdistribution hazard ratio; RR, relative risk; OR, odds ratio; ARR, absolute risk reduction. x_0 represent the reference covariate pattern [Table 2 from Ambrogi F et al. - Statist. Med. 2008]. In our study $F_r = F_P$ i.e. the crude cumulative incidence function of disease progression.*

Measure	Definition	Link function $g(\cdot)$	Coefficient transformation
CSDHR	$\frac{\int_0^t h_r^{sd}(u,\mathbf{x}_1)du}{\int_0^t h_r^{sd}(u,\mathbf{x}_0)du}$	cloglog $\log[-\log(1-F_r(t,\mathbf{x}))]$	$\exp[\boldsymbol{\gamma}^T(\mathbf{x}_1-\mathbf{x}_0)]$
RR	$F_r(t,\mathbf{x_1})/F_r(t,\mathbf{x_0})$	log $\log(F_r(t,\mathbf{x}))$	$\exp[\boldsymbol{\gamma}^T(\mathbf{x}_1-\mathbf{x}_0)]$
OR	$\left[\frac{F_r(t,\mathbf{x}_1)}{1-F_r(t,\mathbf{x}_1)}\right]\Big/\left[\frac{F_r(t,\mathbf{x}_0)}{1-\ F_r(t,\mathbf{x}_0)}\right]$	logit $\log\left[\frac{F_r(t,\mathbf{x})}{1-\ F_r(t,\mathbf{x})}\right]$	$\exp[\boldsymbol{\gamma}^T(\mathbf{x}_1-\mathbf{x}_0)]$
ARR	$F_r(t,\mathbf{x}_0)-\ F_r(t,\mathbf{x}_1)$	Identity $F_r(t,\mathbf{x})$	$\boldsymbol{\gamma}^T(\mathbf{x}_1-\mathbf{x}_0)$

The odds ratio (OR), which is a relative measure of the odds, commonly used in case-control studies, is obtained by applying the logit transformation.

The absolute risk reduction (ARR) measures the incidence difference between patients with $\mathbf{x}_1$ and $\mathbf{x}_0$. The reciprocal of the ARR is the average number of patients need to be treat (NNT) to prevent one failure event, which is a common indicator to evaluate the treatment effects (in this case, the measure is called NNTB, number need to treat to benefit); if the NNT refers to a deleterious factor, NNT can be interpret as the number of subjects to be exposed to have one more failure event (NNTH, number need to treat to harm).

The 95% confidence intervals are calculated for CSDHR, RR and OR as follows from the exponentiation of the estimated coefficients:

$$95\%\, IC = [exp(l_{lower}), exp(l_{upper})]$$

where $l_{lower}, l_{upper} = \hat{\boldsymbol{\gamma}}^T(\mathbf{x}_1-\mathbf{x}_0) \pm 1.96\, st.error(\hat{\boldsymbol{\gamma}}^T(\mathbf{x}_1-\mathbf{x}_0))$ and $\hat{\boldsymbol{\gamma}}$ is the regression coefficient vector estimated by the model comparing $\mathbf{x}_1$ and $\mathbf{x}_0$.

In the case of ARR, since obtained through identity link, the confidence interval is derived without exponentiation:

$$95\%\, IC = [exp(l_{lower}), exp(l_{upper})]$$

A different situation arises for NNT. As suggested by Altman in a comment on a paper by Hutton [123], the confidence interval for NNT should be interpreted as an alternative to present results of ARR. A NNT value of 1 represent the largest possible effect (corresponding to $ARR \rightarrow \infty$), while if the treatment has no effect, NNT is $\pm\infty$; on the other side, the most harmful effect is expressed by NNT as -1. When treatment or the covariate has a significant ARR, confidence interval can be built as

$$95\%\, IC = \left[\frac{1}{l_{upper}}, \frac{1}{l_{lower}}\right]$$

while, when ARR is about 0 and, therefore, NNT approaches infinite, the confidence interval for NNT, as outlined by Altman [124], should be written as

$$95\%\, IC_{NNT} = \left[-\infty, \frac{1}{l_{lower}}\right] \cup \left[\frac{1}{l_{upper}}, +\infty\right]$$

where l_{lower} and l_{upper} refer to confidence limits for ARR.

Wald test

Following the procedure proposed in the paper of my colleagues [95], Wald test was used, along with the QIC, to evaluate the presence of interactions between time and other covariates in the model; interaction between BMI and tumor dimension was also evaluated.

Wald test is used to evaluate the presence of an effect of the covariate, i.e. if there is a relevant statistical connection between the response variable and the independent variable. For example, if in a regression analysis β represents the coefficient of the effect of the independent variable X_1 on the response variable Y, the estimated $\hat{\beta}$ can be tested against the value proposed by the null hypothesis, typically 0, or 1 in the case of a ratio.

In practice, under the assumption that the estimate and the difference between the former and the value proposed by the null hypothesis are normally distributed, is it possible to calculate the following test statistic

$$\frac{\left(\hat{\beta} - \beta_0\right)^2}{Var(\hat{\beta})} \sim \chi_1^2$$

which has an approximately chi-squared distribution with 1 degree of freedom under the null hypothesis [125].

The robust variance calculated in the GEE step was used in the calculations.

The agreement between Wald test and QIC is not guarantee. Particularly in the results parts relative to logit and identity link functions, QIC and Wald test did not agree on the presence of interaction between BMI and tumor dimension; in these cases, results relative to both the model with and without interaction are presented.

Software used

All analyses were conducted using R software (https://www.r-project.org/) version 3.5.2.

CIFs were calculated through the *cuminc()* function in the R package *cmprsk*. The *cuminc()* function requires as input two vectors: one with event times and one with the corresponding status, indicating the type of event or the occurrence of censoring. The function gives back the estimated incidence for each of the competing event at each time unit, along with an estimation of the variance.
Two competing events were considered:

- A composite event defined as any event indicating treatment failure; in particular controlateral tumors(CT), distant metastasis (DM), locoregional metastasis (LM) and second primary tumors (SP) were considered.
- Death without evidence of disease (NEDdeath)

GEE parameter estimation was performed using the *gee()* function in the R package *gee*. Starting estimates of the parameter were provided through the *glm()* function using the same formula and parameters.

RCS and B-splines were calculated through *rcs()* function from package *rms* and bs() function from package *splines* respectively.

Results

Complete information at 15 years (the median follow-up calculated by reverse Kaplan-Meier in complete data) was available for 535 subjects. At 15 years, 23 CTs, 185 DMs, 37 LMs and 21 SPs were observed, for a total of 266 composite events, along with 25 NEDdeaths and 244 censored observations.

In the following pages, clinically useful measures obtained by applying the four different link functions to the complete model are showed separately.

Means of fitted incidence values retrieved by the model were calculated at the 32 timepoints, overall and by treatment arm, and used for a visual comparison with the observed incidence as a non parametric CCI curve (obtained through the *cuminc()* function).

The following strategy was used for each link function separately.

1_Definition of baseline model

As a first step, regression models with only the spline functions for time were built separately for each treatment group for a preliminary evaluation of the type (RCS vs B-splines) and complexity (number of knots) of the spline needed for an adequate model fitting and to evaluate a possible difference in the shape among treatments.

As a second step, spline complexity was re-evaluated considering all dataset (i.e. all treatments arms together) and only RCS was used, since more simple and no relevant difference was observed compared to B-splines. The number of knots of the splines were selected by minimum QIC among models with splines function having from 3 to 7 knots. Models with 7 knots emerged as the best behaving for log, cloglog and logit link functions, while for identity function the model with 4 knots spline resulted the most efficient.

As a third step of this first part, a model including the selected spline function for time, dummy variables for treatment and the interaction between spline for time and treatment was evaluated. The presence of the interaction was evaluated both by QIC and Wald test.

2_Inclusion of BMI, hormonal status and tumor dimension into the "baseline model"

Separately for the four link functions, more complex models were evaluated. Covariates included comprehend ER and menopausal status, BMI, treatment and tumor dimension; these variables were included since commonly considered feature in the evaluation of BC.

BMI was considered as a 2 levels categorical variable, comparing underweight/normalweight (BMI < 25) and overweight/obese (BMI $\geq$ 25) patients; this decision was made because underweight patients consisted in few patients and because in preliminary analyses (data not shown) emerged that the usage of two classes led to more clear results. Hormonal status considered 3 classes: ER-negative, ER-positive and pre-menopausal, ER-positive and postmenopausal, while tumor dimension as 2 levels variable, comparing tumors below and above 2 cm of diameter.

Interaction between BMI, hormonal status, tumor dimension and time and between BMI e tumor dimension were evaluated by both QIC and Wald test.

3_Graphical Evaluation of the goodness of fit

To evaluate the ability of the final models to predict the incidence, graphical evaluation of the goodness of fit was considered. Non parametric incidence curve, obtained through the cuminc() function was plotted against the fitted values retrieved by each model.

Link cloglog

Table 12 shows the QICs obtained for RCS and B-splines separately in the three arms of treatment, while Table 13 shows the QICs obtained considering all dataset and only RCS.

***Table 12**: QICs obtained for link cloglog ($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) in the three arms of treatment separately for RCS and B-splines considering from 3 to 7 knots*

Spline_type	Nr. of knots	CMF	SDE	HDE
RCS	3	476.121	571.381	404.875
RCS	4	474.987	569.051	404.587
RCS	5	474.799	569.031	404.484
RCS	6	474.519	569.083	404.516
RCS	7	474.522	569.128	404.525
B-splines	3	474.535	569.117	404.529
B-splines	4	474.509	569.154	404.569
B-splines	5	474.537	569.185	404.562
B-splines	6	474.565	569.213	404.586
B-splines	7	474.564	569.240	NA

Table 13*: QICs obtained for link cloglog ($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) in the complete dataset considering from 3 to 7 knots.*

Spline_type	Nr. of knots	QIC
RCS	3	4178.451
RCS	4	4165.136
RCS	5	4162.908
RCS	6	4162.762
RCS	7	4162.733

RCS with 7 knots was used to model time when using cloglog link function.

Wald test did not highlight an interaction between time and treatment (p=0.11); coherently, the model considering the interaction between time and treatment was associated to a QIC of 4159.407, while QIC associated to the model without this interaction was 4157.146.

No interaction between covariates was observed in the complete model, therefore the model is a proportional hazard ratio model and regression coefficients, can be used to calculate, through exponentiation, hazard ratios, showed in Table 14 along with the corresponding 95%ICs. For the sake of completeness, although without interpretation, coefficients associated to the intercept and to spline bases are reported, without exponentiation, in Table 15. Figure 14 shows the graphical evaluation of the goodness of fit of the model considering the whole dataset and the three treatments separately.

Although not significant, as confirmed by the inclusion of value 1 in the confidence interval, regimen treatment SDE shows a tendency to a worse prognosis compared to CMF, while HDE shows a tendency in the opposite direction. For what concerns the hormonal status, generally having a tumor positive to ER seams to lead to a better prognosis with respect to ER-negative, although only the ER-positive and postmenopausal status results significant, having both the boundaries of confidence interval below 1. Concerning BMI and tumor dimension, having a BMI above 25 and having a tumor diameter at diagnosis greater than 2 cm are both significant risk factors.

CSDHRs obtained through pseudo-values, correspond to estimates obtainable through the Fine and Gray model, which are showed in Table 16.

It is easy to see that results obtained through the two methods are quite similar and consistent.

Table 14*: Subdistribution hazard ratios obtained through the cloglog link function. HRs were obtained through the exponentiation of regression coefficient calculated after the application of the cloglog ($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link function. 95% confidence intervals (CIs) are also showed.*

	Estimate [95% CI]
Treatment: SDE vs CMF	1.17 [0.85; 1.59]
Treatment: HDE vs CMF	0.84 [0.60; 1.17]
ERpos_premenopusal vs ERneg	0.83 [0.61; 1.14]
ERpos_postmenopusal vs ERneg	0.64 [0.45; 0.91]
BMI: Overweight/Obese vs Normalweight	1.37 [1.05; 1.80]
dimTum: >= 2cm vs <2 cm	1.46 [1.10; 1.93]

Table 15*: Regression coefficient associated to the intercept and to spline bases when applying the cloglog ($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link function.*

	Estimate
(Intercept)	-5.42
splinetime	0.17
splinetime'	-3.56
splinetime''	7.4
splinetime'''	-3.5
splinetime''''	-0.28
splinetime'''''	-0.06

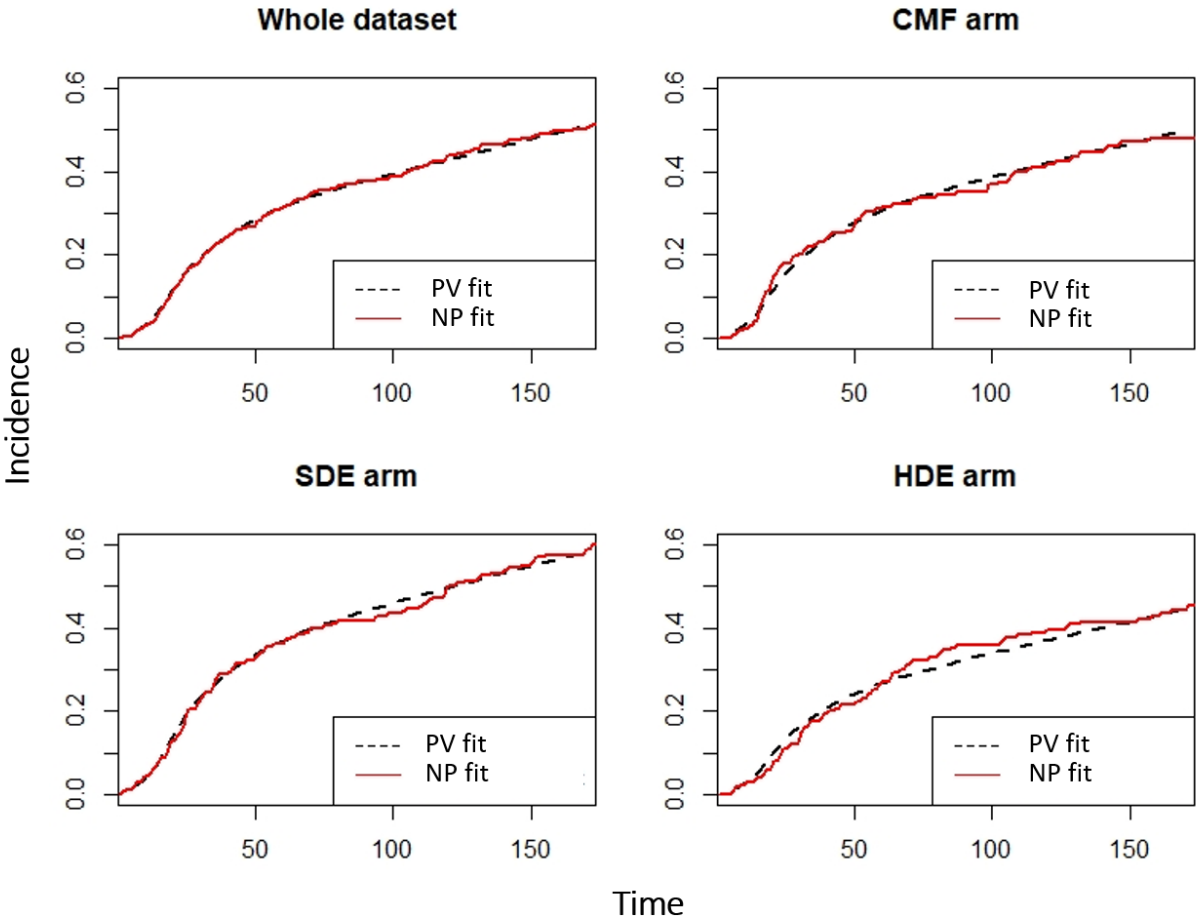


***Figure 14**: Graphical evaluation of the goodness of fit in the whole dataset and separately in the three arms of treatment for cloglog($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link model. PV fit: Pseudo-values fit; NP fit: Non Parametric fit.*

***Table 16**: CSDHR of disease progression obtained through the Fine and Gray model.*

	exp(coef) [95% CI]
Treatment: SDE vs CMF	1.25 [0.93; 1.67]
Treatment: HDE vs CMF	0.88 [0.64; 1.20]
ERpos_premenopusal vs ERneg	0.90 [0.67; 1.21]
ERpos_postmenopusal vs ERneg	0.69 [0.50; 0.96]
BMI: Overweight/Obese vs Normalweight	1.24 [0.97; 1.60]
dimTum: >= 2cm	1.46 [1.12; 1.88]

Link log

Table 17 shows the QICs obtained for RCS and B-splines separately in the three arms of treatment, while Table 18 shows the QICs obtained considering all dataset and only RCS.

RCS with 7 knots was used to model time when using log link function.

***Table 17**: QICs obtained for link log($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) in the three arms of treatment separately for RCS and B-splines considering from 3 to 7 knots.*

Spline_type	Nr. of knots	Treatment arm		
		CMF	SDE	HDE
RCS	3	475.966	571.035	404.815
RCS	4	474.943	569.019	404.575
RCS	5	474.781	569.03	404.479
RCS	6	474.511	569.083	404.516
RCS	7	474.519	569.127	404.525
B-splines	3	474.537	569.118	404.529
B-splines	4	474.511	569.154	404.569
B-splines	5	474.537	569.186	404.562
B-splines	6	474.565	569.213	404.586
B-splines	7	474.564	569.24	NA

***Table 18**: QICs obtained for link log($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) in the complete dataset considering from 3 to 7 knots.*

Spline_type	Nr. of knots	QIC
RCS	3	4180.898
RCS	4	4165.642
RCS	5	4162.981
RCS	6	4162.793
RCS	7	4162.737

Wald test did not highlight an interaction between time and treatment (p=0.12); coherently, the model considering the interaction between time and treatment was associated to a QIC of 4159.404, while QIC associated to the model without this interaction was 4157.449.

As in the previous case, no interaction between covariates was observed in the complete model, therefore the model is a proportional relative risk model and regression coefficients can be used to calculate relative risks, showed in Table 19 along with the corresponding IC95. Table 20 shows, without exponentiation, the the coefficients associated to the intercept and to the spline bases. Figure 15 shows the graphical evaluation of the goodness of fit of the model considering the whole dataset and the three treatments separately.

Table 19*: Relative risks obtained through the log(*$F_P(t)$*) (where (*$F_P(t)$ *is the crude cumulative incidence of disease progression) link function. RRs were obtained through the exponentiation of regression coefficient calculated after the application of the log link function. IC95 burdens are also showed.*

	Estimate [95% CI]
Treatment: SDE vs CMF	1.13 [0.90; 1.42]
Treatment: HDE vs CMF	0.87 [0.67; 1.13]
ERpos_premenopusal vs ERneg	0.89 [0.71; 1.11]
ERpos_postmenopusal vs ERneg	0.72 [0.56; 0.94]
BMI: Overweight/Obese vs Normalweight	1.27 [1.04; 1.55]
dimTum: >= 2cm	1.34 [1.07; 1.67]

Table 20*: Regression coefficient associated to the intercept and to spine bases when applying the log(*$F_P(t)$*) (where (*$F_P(t)$ *is the crude cumulative incidence of disease progression) link function.*

	Estimate
(Intercept)	-5.33
splinetime	0.16
splinetime'	-3.57
splinetime''	7.49
splinetime'''	-3.61
splinetime''''	-0.25
splinetime'''''	-0.06

As with the cloglog function, regimen treatment SDE shows a tendency to a worse prognosis compared to CMF, while HDE shows a tendency in the opposite direction, although without being significant, since their IC95 comprehend the value 1. Concerning the hormonal status, positivity to ER seems to be associated to a risk reduction, although only the ER-positive and postmenopausal status results significant, having both the boundaries of confidence interval below 1. Relative to BMI and tumor dimension, having a BMI above 25 and having a tumor diameter at diagnosis greater than 2 cm are both significant risk factors.

From RR estimates, RRR measures, showed in Table 21, were calculated.

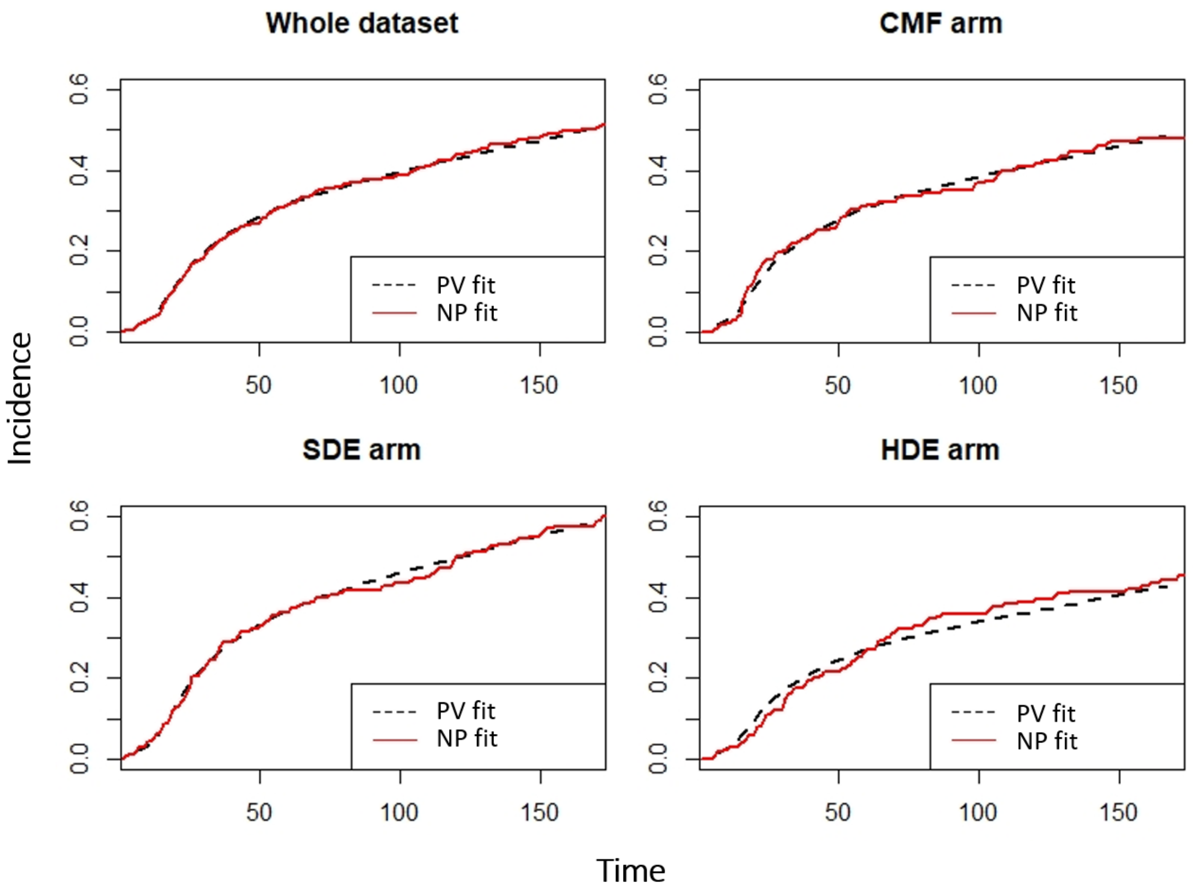


***Figure 15**: Graphical evaluation of the GoF in the whole dataset and separately in the three arms of treatment for log($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link model. PV fit: Pseudo-values fit; NP fit: Non Parametric fit*

***Table 21**: Relative risk reduction of disease progression measures obtained from RRs.*

	Estimate [95% CI]
Treatment: SDE vs CMF	-0.13 [-0.42; [0.10]
Treatment: HDE vs CMF	0.13 [-0.13; 0.33]
ERpos_premenopusal vs ERneg	0.11 [-011; 0.29]
ERpos_postmenopusal vs ERneg	0.28 [0.06; 0.44]
BMI: Overweight/Obese vs Normalweight	-0.27 [-0.55; -0.04]
dimTum: >= 2cm	-0.34 [-0.67; -0.07]

Link logit

Table 22 shows the QICs obtained for RCS and B-splines separately in the three arms of treatment, while Table 23 shows the QICs obtained considering all dataset and only RCS.

RCS with 7 knots was used to model time when using logit link function.

In the case of link logit, QIC and Wald test did not agree regarding the presence of interaction between BMI and tumor diameter: Wald test couldn’t detect the presence of interaction (p=0.16) but QIC relative to the model with interaction was slightly lesser than the QIC relative to the model without interaction ($2987.040 < 2987.045$).

Table 22*: QICs obtained for link logit($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) in the three arms of treatment separately for RCS and B-splines considering from 3 to 7 knots.*

Spline_type	Nr. of knots	CMF	QDE	HDE
RCS	3	475.966	571.035	404.815
RCS	4	474.943	569.019	404.575
RCS	5	474.781	569.03	404.479
RCS	6	474.511	569.083	404.516
RCS	7	474.519	569.127	404.525
B-splines	3	474.537	569.118	404.529
B-splines	4	474.511	569.154	404.569
B-splines	5	474.537	569.186	404.562
B-splines	6	474.565	569.213	404.586
B-splines	7	474.564	569.24	NA

Table 23*: QICs obtained for link logit($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) in the complete dataset considering from 3 to 7 knots.*

Spline_type	Nr. of knots	QIC
RCS	3	4176.386
RCS	4	4164.706
RCS	5	4162.846
RCS	6	4162.737
RCS	7	4162.730

Table 24 shows the regression coefficients obtained through the model without interaction after exponentiation, which can be therefore be interpreted as OR, while Table 25 shows the coefficient associated to the intercept and spline bases not exponentiated. Figure 16 shows the graphical evaluation of the goodness of fit of the model considering the whole dataset and the three treatments separately.

When not considering the interaction, results are consistent with the previous cases.

Still not significant, including the value 1 in the confidence interval, regimen treatment SDE shows a tendency to a worse prognosis compared to CMF, while HDE shows a protective tendency. For what concerns the hormonal status, having a tumor positive to ER confirms its protective role, although once again only the status which consider ER positivity and being postmenopausal results significant. Concerning BMI and tumor dimension, having a BMI above 25 and having a tumor diameter at diagnosis greater than 2 cm are both significant risk factors, having OR and their whole confidence intervals over 1.

***Table 24**: Odds ratios obtained through the logit($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link function and the model without interaction. ORs were obtained through the exponentiation of regression coefficients calculated after the application of the logit link function to the model without interaction. IC95burdens are also shown.*

	Estimate [95% CI]
trtmSDE	1.20 [0.81; 1.79]
trtmHDE	0.80 [0.53; 1.21]
subgrER-positive_pre	0.78 [0.52; 1.17]
subgrER-positive_post	0.56 [0.36; 0.88]
BMI_Nw_vs_OwOb	1.49 [1.05; 2.10]
dimTUM>=2cm	1.60 [1.13; 2.26]

***Table 25**: Regression coefficient associated to the intercept and to spine bases when applying the logit($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link function to the model without interaction.*

	Estimate
(Intercept)	-5.54
splinetime	0.18
splinetime'	-3.56
splinetime''	7.29
splinetime'''	-3.34
splinetime''''	-0.33
splinetime'''''	-0.06

If the model with interaction is considered, results showed in Table 26 are obtained, along with the coefficient associated to intercept and spline bases, showed in Table 27.

***Table 26**: Odds ratios obtained through the logit($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link function and the model with interaction. ORs were obtained through the exponentiation of regression coefficients calculated after the application of the logit link function to the model with interaction.*

	Estimate [95% CI]
Treatment: SDE vs CMF	1.19 [0.80; 1.78]
Treatment: HDE vs CMF	0.78 [0.52; 1.19]
ERpos_premenopusal vs ERneg	0.78 [0.52; 1.16]
ERpos_postmenopusal vs ERneg	0.57 [0.36; 0.89]
BMI: Overweight/Obese vs Normalweight	1.05 [0.59; 1.88]
dimTum: >= 2cm	1.26 [0.79; 2.00]
interaction_BMI:dimTum	1.65 [0.82; 3.31]

***Table 27**: Regression coefficient associated to the intercept and to spine bases when applying the logit($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link function to the model with interaction.*

	Estimate
(Intercept)	-5.35
splinetime	0.17
splinetime'	-3.51
splinetime''	7.2
splinetime'''	-3.3
splinetime''''	-0.31
splinetime'''''	-0.06

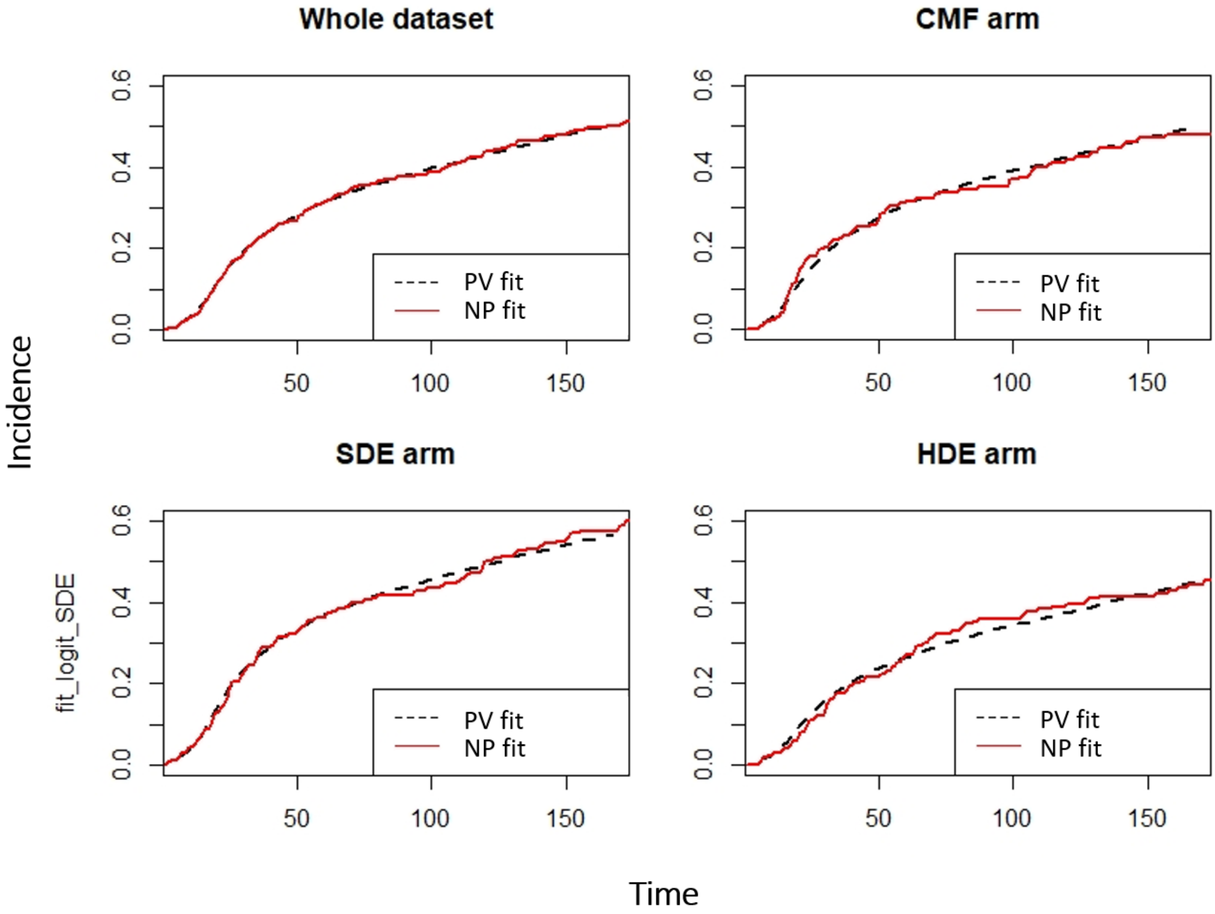


***Figure 16**: Graphical evaluation of the GoF in the whole dataset and separately in the three arms of treatment for logit($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link model without interaction. PV fit: Pseudo-values fit; NP fit: Non Parametric fit*

If interaction is considered, for a patient with BMI > 25, having also a tumor diameter greater than 2 cm results in an OR of $\exp(0{,}23 + 0{,}50) \cong 2{,}08$ $(95\%\ CI = [1{,}01\,; 3{,}14])$, where 0,23 and 0,50 are the raw coefficients (i.e. not exponentiated) estimated by the model for tumor dimension greater than 2 cm and interaction respectively. Changing perspective, for a patient with a tumor with diameter greater than 2 cm, being also overweight or obese is associated with an OR of $\exp(0{,}05\ +\ 0{,}50) \cong 1{,}73\ (95\%\ CI = [0{,}52\ ; 2{,}95])$, where 0,05 and 0,50 are the raw coefficients for BMI>25 and interaction respectively.

Figure 17 shows the graphical evaluation of the goodness of fit of the model with interaction.

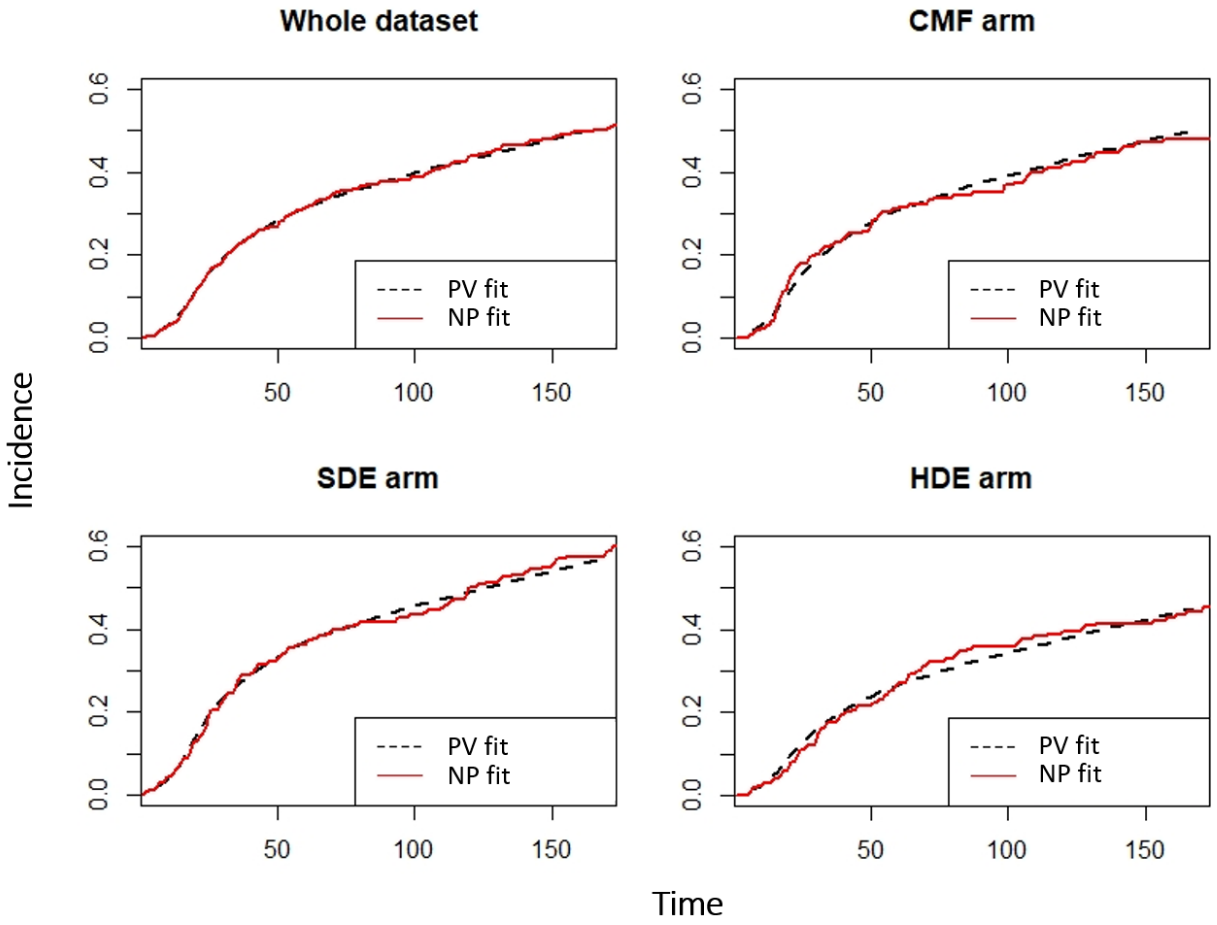


Figure 17*: Graphical evaluation of the GoF in the whole dataset and separately in the three arms of treatment for logit($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link model with interaction. PV fit: Pseudo-values fit; NP fit: Non Parametric fit.*

Link identity

Table 28 shows the QICs obtained for RCS and B-splines separately in the three arms of treatment, while Table 29 shows the QICs obtained considering all dataset and only RCS.

***Table 28**: QICs obtained for link identity($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) in the three arms of treatment separately for RCS and B-splines considering from 3 to 7 knots.*

Spline_type	Nr. of knots	CMF	SDE	HDE
RCS	3	475.024	569.494	404.378
RCS	4	474.556	569.012	404.424
RCS	5	474.565	569.057	404.471
RCS	6	474.458	569.076	404.518
RCS	7	474.494	569.126	404.525
B-splines	3	474.54	569.118	404.536
B-splines	4	474.519	569.148	404.567
B-splines	5	474.534	569.186	404.562
B-splines	6	474.565	569.213	404.586
B-splines	7	474.564	569.24	NA

***Table 29**: Regression coefficient associated to the intercept and to spine bases when applying the identity($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link function to the model without interaction.*

Spline_type	Nr. of knots	QIC
RCS	3	4165.032
RCS	4	4162.563
RCS	5	4162.876
RCS	6	4162.880
RCS	7	4162.825

RCS with 4 knots was used to model time when using identity link function.

In the case of link identity, QIC and Wald test did not agree on the presence of interaction between BMI e tumor diameter: Wald test couldn't detect the presence of interaction ($P = 0.085$) but QIC relative to the model with interaction was lesser than the QIC relative to the model without interaction ($2988.039 < 2991.373$).

Table 30 shows the regression coefficients obtained through the model without interaction, which can be directly interpreted as ARR, while Table 31 shows the coefficients associated to the spline bases. Figure 18 shows the graphical evaluation of the goodness of fit of the model considering the whole dataset and the three treatments separately.

***Table 30**: Absolute risk reductions obtained through the identity($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link function. ARRs were obtained directly without exponentiation from the regression coefficient calculated after the application of the identity link function to the model without interaction.*

	Estimate [95% CI]
trtmSDE	0.03 [-0.04; 0.10]
trtmHDE	-0.04 [-0.11; 0.03]
subgrER-positive_pre	-0.05 [-0.12; 0.02]
subgrER-positive_post	-0.1 [-0.18; -0.02]
BMI_Nw_vs_OwOb	0.07 [0.01; 0.13]
dimTUM>=2cm	0.08 [0.02; 0.13]

***Table 31**: Regression coefficient associated to the intercept and to spine bases when applying the identity($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link function to the model without interaction.*

	Estimate
(Intercept)	-0.08
splinetime	0.01
splinetime'	-0.03
splinetime''	0.05

As for the previous link functions, measures relative to treatments are not significant, being their ARRs approximately around 0; the tendency of SDE treatment to lead to a worse prognosis and the tendency of HDE to a better prognosis, compared to reference CMF, are visible also in this case. Once again, patients with tumor positive to ER show a reduction of incidence, although only in the case of the postmenopausal subgroup this reduction is significant, having both burdens of IC below 0. Tumor dimension and BMI confirm their significant deleterious effect.

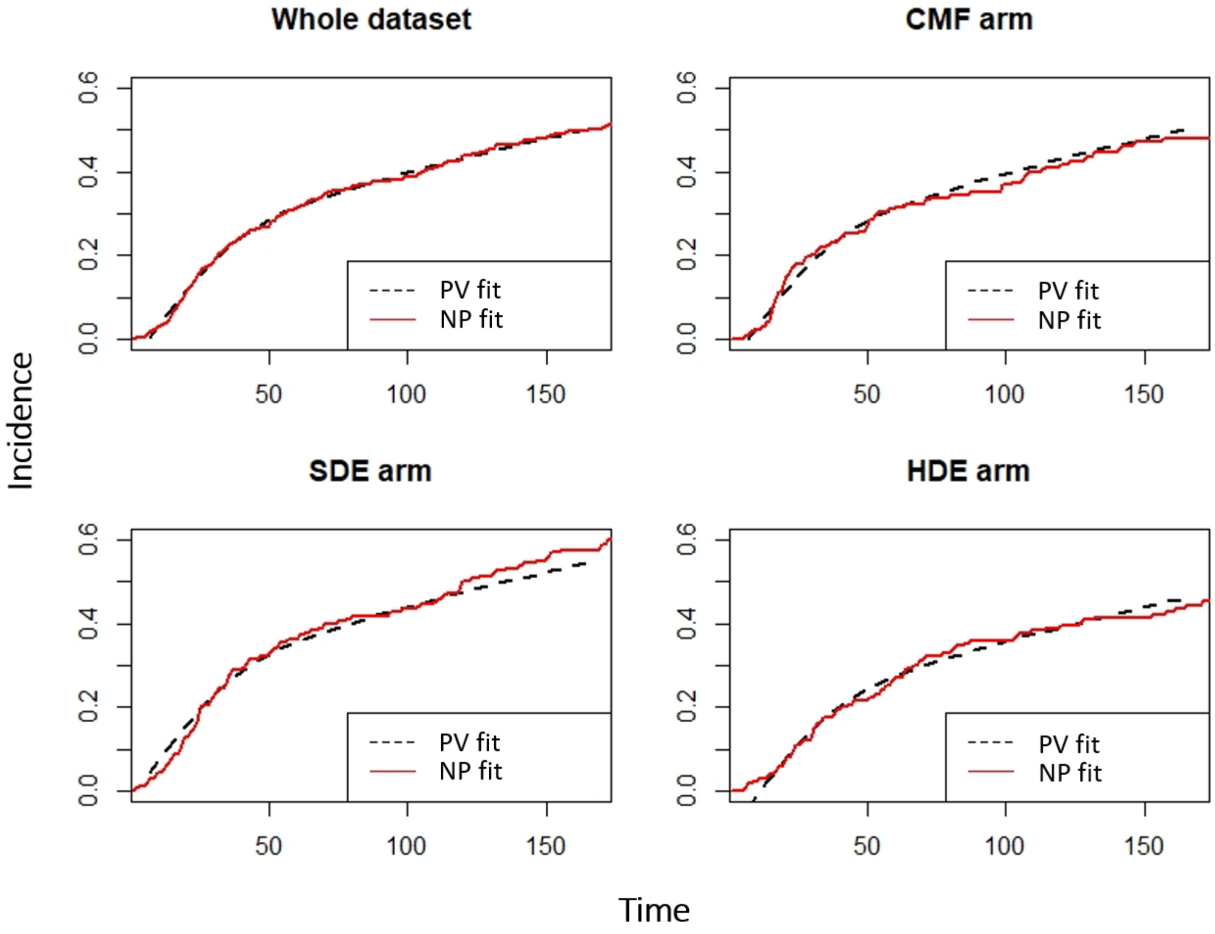


***Figure 18**: Graphical evaluation of the GoF in the whole dataset and separately in the three arms of treatment for identity($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link model without interaction. PV fit: Pseudo-values fit; NP fit: Non Parametric fit.*

From ARR measures, it is possible to retrieve NNTs, which are shown in Table 32.

In the cases of treatment SDE, treatment HDE and subgroup "ER-positive and premenopausal", since their ARR was around 0, NNT estimates cannot be interpreted.

For subgroup "ER-positive and postmenopausal", since its protective effect, NNT can be interpreted as NNTB, indicating that every around 10 patients with ER positive tumor and postmenopausal, 1 adverse event can be prevented on average. On the other hand, since their deleterious effect, NNTs relative to BMI and tumor dimension can be interpreted as NNTH: every around 15 patients with BMI above 25 or every around 13 patients with tumor diameter bigger than 2 cm, 1 more failure is observed on average.

Table 32*: Number of patients Need to Treat to prevent disease progression and Number of patients Need to be exposed to have one more harmful event.*

	Estimate [95% CI]
trtmSDE vs CMF	37.57 [10.30; ∞]
trtmHDE vs CMF	24.70 [9.27; ∞]
subgrER-positive_pre vs ER-negative	20.03 [8.12; ∞]
subgrER-positive_post vs ER-negative	9.89 [5.60; 42.30]
BMI_Nw_vs_OwOb	15.01 [7.98; 126.68]
dimTUM<2cm vs >=2cm	12.76 [7.43; 45.38]

If the interaction between BMI and tumor diameter is considered, the results obtained relative to the coefficients and to the intercept and spline bases are showed in Table 33 and Table 34 respectively. Figure 19 shows the graphical evaluation of the goodness of fit of the model with interaction.

Table 33*: Absolute risk reductions obtained through the identity($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link function. ARRs were obtained directly without exponentiation from the regression coefficient calculated after the application of the identity link function to the model with interaction.*

	Estimate [95% CI]
trtmSDE	0.03 [-0.04; 0.10]
trtmHDE	-0.04 [-0.11; 0.02]
subgrER-positive_pre	-0.05 [-0.12; 0.02]
subgrER-positive_post	-0.10 [-0.18; -0.02]
BMI_Nw_vs_OwOb	0.00 [-0.08; 0.09]
dimTUM>=2cm	0.04 [-0.04; 0.11]
BMI_Nw_vs_OwOb:dimTUM>=2cm	0.10 [-0.01; 0.21]

Table 34: *Regression coefficient associated to the intercept and to spine bases when applying the identity($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link function to the model with interaction.*

	Estimate
(Intercept)	-0.05
splinetime	0.01
splinetime'	-0.03
splinetime''	0.05

Considering the interaction leads in this case to results difficult to interpret. Particularly interesting is that most of the effect of BMI and tumor dimension covariates is taken by the interaction coefficient; IC95s relative to both original variables now comprise the 0.

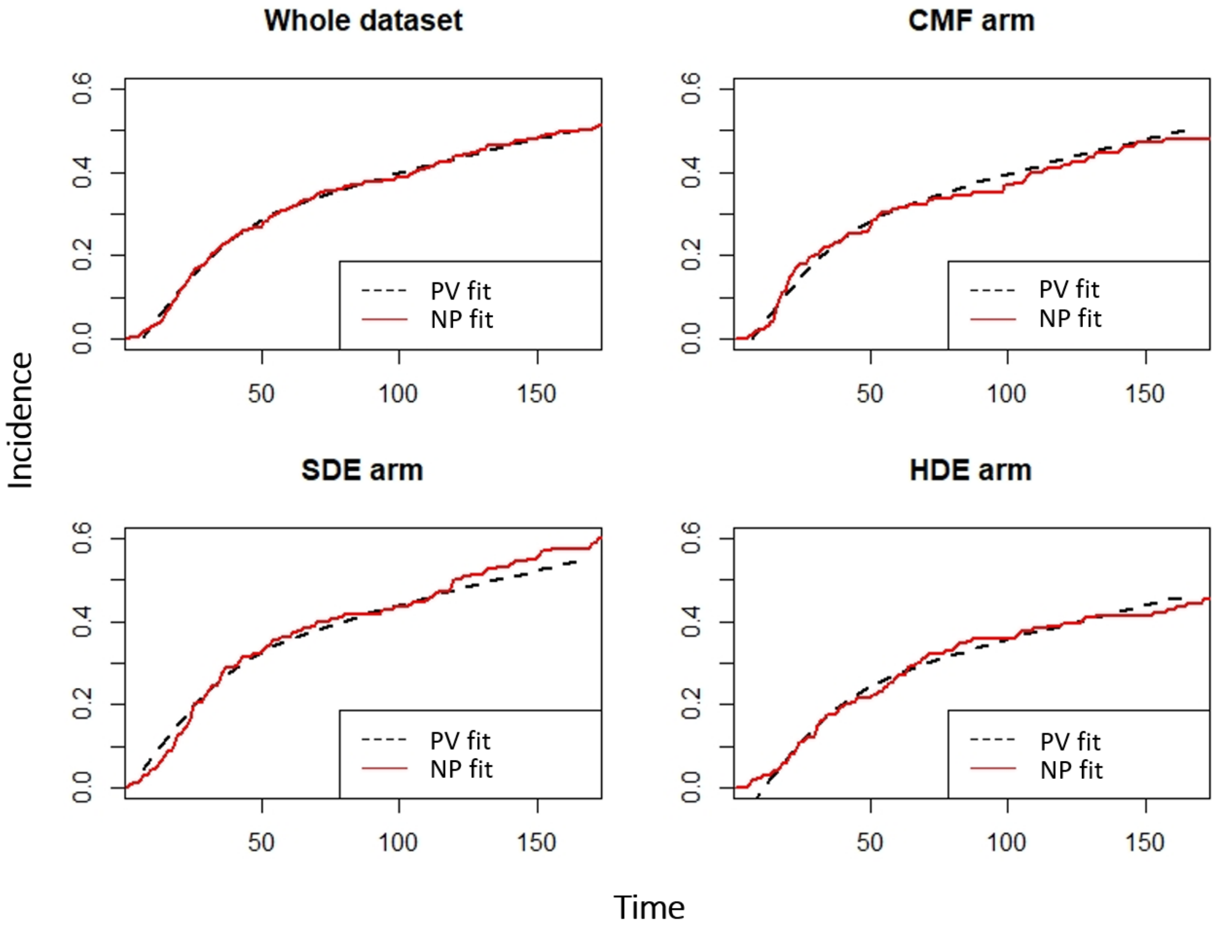


Figure 19*: Graphical evaluation of the GoF in the whole dataset and separately in the three arms of treatment for identity($F_P(t)$) (where ($F_P(t)$ is the crude cumulative incidence of disease progression) link model with interaction. PV fit: Pseudo-values fit; NP fit: Non Parametric fit.*

Since the two covariates showed a deleterious effect through the analyses, one can expect that is difficult that they could have a protective effect; thus the interaction can be interpreted as follows. For a patient with BMI > 25, having also a tumor diameter greater than 2 cm results in a risk augmentation (since >0) of $0{,}04 + 0{,}10 \cong 0{,}14$ ($95\% \ CI = [-0{,}04 \ ; \ 0{,}30]$), while for a patients which already have a big tumor, being also overweight or obese lead to a risk augmentation of $0{,}00 + 0{,}10 \cong 0{,}10$ ($95\% \ CI = [\ -0{,}08 \ ; \ 0{,}29]$).

Conclusions

Differentially from standard regression survival models, the approach presented in this work allow to obtain measures of covariate effects which can be directly interpreted in clinic.

Traditional regression models are based on sub-distribution hazard, which are not useful to describe disease dynamics and have no direct clinical interpretation. Even when considering the cloglog transformation, which model cumulative sub-distribution hazard, the only obtainable information is whether two cumulative incidence curves are different or not, but one cannot quantify this difference.

The pseudo-values based model allows the usage of different link functions in order to quantify the difference, retrieving different clinical useful measures.

One of the most used clinical measure in longitudinal studies, like the one presented, is RR, which can be calculated by applying the log link function. RRs obtained showed a relative risk for patients who received SDE of 1.13, meaning an augmented risk, compared to the standard treatment with CMF, of 13% to fail within 15 years, while HDE showed a reduction of the risk, having a RR of 0.87. Having a tumor positive for ER, which constitute an important therapy target, has a protective effect in both premenopausal and postmenopausal women, which have a 0.89 and 0.72 RR respectively, although only the latter is a significant result. BMI confirms its well-known deleterious effect in the context of BC, which in this case correspond to an estimated RR of 1.27. Lately, tumor dimension brings a RR of 1.34 which is the strongest effect measured; this is not surprisingly, since a big tumor at diagnosis indicates that it is in already advanced status and then more difficult to treat.

Visual comparisons showed that each model, despite the different link functions adopted, was capable to predict correctly the incidence, since curves drawn with fitted values were almost overlapping the non-parametric cause-specific CCI. This aspect represents a first proof demonstrating that the pseudo-values approach is coherent with common methods.

This can be seen also through the comparison of CSDHR calculated through the pseudo-values method and the cloglog link function with the estimates obtained through the Fine and Gray model, showing that the measures estimated are quite similar: for example CSDHRs calculated for treatments effect through pseudo-values are 1.17 for SDE and 0.84 for HDE, while the corresponding measures from the Fine and Gray model are 1.25 and 0.88; the estimates for the effect of tumor dimension are identical.

Covariate effects estimated are coherent through the different link functions: all models are concordant in evaluating the protective effects of having an ER-positive tumors and being in menopause and the deleterious effect of being overweight or obese and having a tumor with diameter above 2 cm. Coherently, treatment regimen and having ER–positive tumor while being pre-menopausal remain not significant throughout the analyses.

All together, these results show that the pseudo-values based method is reliable to directly retrieve measures and to quantify covariates effects on them, thus proposing itself as a powerful and useful instrument to guide and support clinical decisions.

References

[1] Margaret A. Hamburg, M.D., and Francis S. Collins, M.D., Ph.D. The Path to Personalized Medicine. N Engl J Med 2010; 363:301-304.

[2] https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics

[3] Blows FM, Driver KE, Schmidt MK, et al. Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. PLoS Med. 2010;7(5):e1000279. Published 2010 May 25. doi:10.1371/journal.pmed.1000279

[4] E.H. Lips, L. Mulder, J.J. de Ronde, I.A.M. Mandjes, B.B. Koolen, L.F.A. Wessels, S. Rodenhuis, J.Wesseling. Breast cancer subtyping by immunohistochemistry and histological grade outperforms breast cancer intrinsic subtypes in predicting neoadjuvant chemotherapy response. Breast Cancer Res Treat (2013) 140: 63. https://doi.org/10.1007/s10549-013-2620-0.

[5] Andrew H. Beck, Ankur R. Sangoi, Samuel Leung, Robert J. Marinelli, Torsten O. Nielsen, Marc J. van de Vijver, Robert B. West, Matt van de Rijn and Daphne Koller. Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival. Science Translational Medicine 09 Nov 2011:Vol. 3, Issue 108, pp. 108ra113. DOI: 10.1126/scitranslmed.3002564.

[6] Margaret A. Hamburg, M.D., and Francis S. Collins, M.D., Ph.D. The Path to Personalized Medicine. N Engl J Med 2010; 363:301-304.

[7] Blows FM, Driver KE, Schmidt MK, et al. Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. PLoS Med. 2010;7(5):e1000279. Published 2010 May 25. doi:10.1371/journal.pmed.1000279

[8] E.H. Lips, L. Mulder, J.J. de Ronde, I.A.M. Mandjes, B.B. Koolen, L.F.A. Wessels, S. Rodenhuis, J.Wesseling. Breast cancer subtyping by immunohistochemistry and histological grade outperforms breast cancer intrinsic subtypes in predicting neoadjuvant chemotherapy response. Breast Cancer Res Treat (2013) 140: 63. https://doi.org/10.1007/s10549-013-2620-0.

[9] Beck J, Urnovitz HB, Mitchell WM, Schütz E. Next generation sequencing of serum circulating nucleic acids from patients with invasive ductal breast cancer reveals differences to healthy and nonmalignant controls. Mol Cancer Res. 2010 Mar;8(3):335-42. doi: 10.1158/1541-7786.MCR-09-0314. Epub 2010 Mar 9.

[10] Blank MAB1, Antaki JF2. Breast Lesion Elastography Region of Interest Selection and Quantitative Heterogeneity: A Systematic Review and Meta-Analysis. Ultrasound Med Biol. 2017 Feb;43(2):387-397. doi: 10.1016/j.ultrasmedbio.2016.09.002. Epub 2016 Oct 14.

[11] Ali, H. R., Chlon, L., Pharoah, P. D. P., Markowetz, F. & Caldas, C. Patterns of Immune Infiltration in Breast Cancer and Their Clinical Implications: A Gene-Expression-Based Retrospective Study. PLOS Med. 13, e1002194 (2016).

[12] Emens, L. A. et al. Long-term Clinical Outcomes and Biomarker Analyses of Atezolizumab Therapy for Patients With Metastatic Triple-Negative Breast Cancer. JAMA Oncol. (2018). doi:10.1001/jamaoncol.2018.4224

[13] Savas, P. et al. Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. Nat. Med. 24, 986–993 (2018).

[14] Schmid, P. et al. Atezolizumab and Nab-Paclitaxel in Advanced Triple-Negative Breast Cancer. N. Engl. J. Med. NEJMoa1809615 (2018). doi:10.1056/NEJMoa1809615

[15] Loi, S. et al. Tumor-Infiltrating Lymphocytes and Prognosis: A Pooled Individual Patient Analysis of Early-Stage Triple-Negative Breast Cancers. J. Clin. Oncol. JCO1801010 (2019). doi:10.1200/JCO.18.01010

[16] Teschendorff, A. E., Miremadi, A., Pinder, S. E., Ellis, I. O. & Caldas, C. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biol. 8, R157 (2007).

[17] Desmedt, C. et al. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clin. Cancer Res. 14, 5158–65 (2008).

[18] Ignatiadis, M. et al. Gene modules and response to neoadjuvant chemotherapy in breast cancer subtypes: a pooled analysis. J. Clin. Oncol. 30, 1996–2004 (2012).

[19] Solinas, C. et al. Targeting immune checkpoints in breast cancer: an update of early results. ESMO Open 2, e000255 (2017).

[20] Denkert, C. et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. Lancet Oncol. 19, 40–50 (2018).

[21] Dieci, M. V et al. Association of tumor-infiltrating lymphocytes with distant disease-free survival in the ShortHER randomized adjuvant trial for patients with early HER2+ breast cancer. Ann. Oncol. (2019). doi:10.1093/annonc/mdz007

[22] Salgado, R. et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. Ann. Oncol. 26, 259–271 (2015).

[23] Hendry, S. et al. Assessing Tumor-infiltrating Lymphocytes in Solid Tumors: A Practical Review for Pathologists and Proposal for a Standardized Method From the International Immunooncology Biomarkers Working Group: Part 1: Assessing the Host Immune Response, TILs in Invasive Breast Carcinoma and Ductal Carcinoma In Situ, Metastatic Tumor Deposits and Areas for Further Research. Adv Anat Pathol 24, (2017).

[24] Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013 Oct;45(10):1113-20. doi: 10.1038/ng.2764.

[25] Finotello, F. & Trajanoski, Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. Cancer Immunol. Immunother. 67, 1031–1040 (2018).

[26] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences Oct 2005, 102 (43) 15545-15550; DOI: 10.1073/pnas.0506580102

[27] Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534, 47–54 (2016).

[28] Smid et al., Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration. Nature Communicationsvolume 7, Article number: 12910 (2016)

[29] Barbie et al., Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Naturevolume 462, pages108–112 (2009)

[30] Lizio et al. Update of the FANTOM web resource: high resolution transcriptome of diverse cell types in mammals. Nucleic Acids Research, Volume 45, Issue D1, January 2017, Pages D737–D743, https://doi.org/10.1093/nar/gkw995

[31] ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004 Oct 22;306(5696):636-40.

[32] Fernández JM, de la Torre V, Richardson D, et al. The BLUEPRINT Data Analysis Portal. Cell Syst. 2016;3(5):491–495.e5. doi:10.1016/j.cels.2016.10.021

[33] A R Abbas, D Baldwin, Y Ma, W Ouyang, A Gurney, F Martin, S Fong, M van Lookeren Campagne, P Godowski, P M Williams, A C Chan & H F Clark. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. Genes & Immunityvolume 6, pages319–331 (2005)

[34] Mabbott NA, Baillie JK, Brown H, Freeman TC, Hume DA. An expression atlas of human primary cells: inference of gene function from coexpression networks. BMC Genomics. 2013;14:632. Published 2013 Sep 20. doi:10.1186/1471-2164-14-632

[35] Novershtern N, Subramanian A, Lawton LN, et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell. 2011;144(2):296–309. doi:10.1016/j.cell.2011.01.004

[36] Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7. Published 2013 Jan 16. doi:10.1186/1471-2105-14-7.

[37] Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18(1):220. Published 2017 Nov 15. doi:10.1186/s13059-017-1349-1.

[38] Becht E, Giraldo NA, Lacroix L, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression [published correction appears in Genome Biol. 2016 Dec 1;17 (1):249]. Genome Biol. 2016;17(1):218. Published 2016 Oct 20. doi:10.1186/s13059-016-1070-5

[39] Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Curr Opin Immunol. 2013;25(5):571–578. doi:10.1016/j.coi.2013.09.015

[40] Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS One. 2009 Jul 1;4(7):e6098. doi: 10.1371/journal.pone.0006098.

[41] Gong T, Hartmann N, Kohane IS, Brinkmann V, Staedtler F, Letzkus M, Bongiovanni S, Szustakowski JD. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. PLoS One. 2011;6(11):e27156. doi: 10.1371/journal.pone.0027156. Epub 2011 Nov 16.

[42] Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods 12, 453–457 (2015).

[43] Racle, J., Jonge, K. de, Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife 6, (2017).

[44] Tirosh I, Venteicher AS., Hebert C. Escalante LE. Patel AP, Yizhak K. Fisher JM, Rodman C, Mount C, Filbin MG., Neftel C. Desai N, Nyman J, Izar B, Luo CC, Francis JM. Patel AA, Onozato ML, Riggi N, Livak KJ, Gennert D, Satija R, Nahed BV, Curry WT, Martuza RL, Mylvaganam R, Iafrate AJ, Frosch MP, Golub TR, Rivera MN. Getz G. Rozenblatt-Rosen O, Cahill DP, Monje M, Bernstein BE. Louis DN, Regev A. Suvà ML. Single-cell RNA-

seq supports a developmental hierarchy in human oligodendroglioma. Nature. 2016 Nov 10;539(7628):309-313. doi: 10.1038/nature20123. Epub 2016 Nov 2.

[45] Francesca Finotello, Clemens Mayer, Christina Plattner, Gerhard Laschober, Dietmar Rieder, Hubert Hackl, Anne Krogsdam, Wilfried Posch, Doris Wilflingseder, Sieghart Sopper, Marieke IJsselsteijn, Douglas Johnson, Yaomin Xu, Yu Wang, Melinda E. Sanders, Monica V. Estrada, Paula Ericsson-Gonzalez, Justin Balko, Noel de Miranda, Zlatko Trajanoski. quanTIseq: quantifying immune contexture of human tumors. BioRxiv 223180; doi: https://doi.org/10.1101/223180

[46] D. Venet, F. Pecasse, C. Maenhaut, H. Bersini, Separation of samples into their constituents using gene expression data , Bioinformatics, Volume 17, Issue suppl_1, June 2001, Pages S279–S287, https://doi.org/10.1093/bioinformatics/17.suppl_1.S279

[47] Repsilber D, Kern S, Telaar A, Walzl G, Black GF, Selbig J, Parida SK, Kaufmann SH, Jacobsen M. Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. BMC Bioinformatics. 2010 Jan 14;11:27. doi: 10.1186/1471-2105-11-27.

[48] Lähdesmäki H, Shmulevich L, Dunmire V, Yli-Harja O, Zhang W. In silico microdissection of microarray data from heterogeneous cell populations. BMC Bioinformatics. 2005 Mar 14;6:54.

[49] Gaujoux R, Seoighe C. Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study. Infect Genet Evol. 2012 Jul;12(5):913-21. doi: 10.1016/j.meegid.2011.08.014. Epub 2011 Sep 10.

[50] Zhong Y, Wan YW, Pang K, Chow LM, Liu Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC Bioinformatics. 2013 Mar 7;14:89. doi: 10.1186/1471-2105-14-89.

[51] Chakravarthy A, Furness A, Joshi K, Ghorani E, Ford K, Ward MJ, King EV, Lechner M, Marafioti T, Quezada SA, Thomas GJ, Feber A, Fenton TR. Author Correction: Pan-cancer deconvolution of tumour composition using DNA methylation. Nat Commun. 2018 Nov 2;9(1):4642. doi: 10.1038/s41467-018-07155-4.

[52] Alexe G, Dalgin GS, Scanfeld D, Tamayo P, Mesirov JP, DeLisi C, Harris L, Barnard N, Martel M, Levine AJ, Ganesan S, Bhanot G. High expression of lymphocyte-associated genes in node-negative HER2+ breast cancers correlates with lower recurrence rates. Cancer Res. 2007 Nov 15;67(22):10669-76.

[53] Teschendorff, A. E., Miremadi, A., Pinder, S. E., Ellis, I. O. & Caldas, C. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biol. 8, R157 (2007).

[54] Desmedt C1, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, Bontempi G, Delorenzi M, Piccart M, Sotiriou C. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clin Cancer Res. 2008 Aug 15;14(16):5158-65. doi: 10.1158/1078-0432.CCR-07-4756.

[55] Perez EA, Thompson EA, Ballman KV, Anderson SK, Asmann YW, Kalari KR, Eckel-Passow JE, Dueck AC, Tenner KS, Jen J, Fan JB, Geiger XJ, McCullough AE, Chen B, Jenkins RB, Sledge GW, Winer EP, Gralow JR, Reinholz MM. Genomic analysis reveals that immune function genes are strongly linked to clinical outcome in the North Central Cancer Treatment Group n9831 Adjuvant Trastuzumab Trial. J Clin Oncol. 2015 Mar 1;33(7):701-8. doi: 10.1200/JCO.2014.57.6298. Epub 2015 Jan 20.

[56] Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. Cell 160, 48–61 (2015).

[57] Ayers M, Lunceford J, Nebozhyn M, Murphy E, Loboda A, Kaufman DR, Albright A, Cheng JD, Kang SP, Shankaran V, Piha-Paul SA, Yearley J, Seiwert TY, Ribas A, McClanahan TK. IFN-γ-related mRNA profile predicts

clinical response to PD-1 blockade. J Clin Invest. 2017 Aug 1;127(8):2930-2940. doi: 10.1172/JCI91190. Epub 2017 Jun 26.

[58] Jeschke J, Bizet M,,, Desmedt C, Calonne E, Dedeurwaerder S, Garaud S5, Koch A6, Larsimont D, Salgado R, Van den Eynden G, Willard Gallo K5, Bontempi G,, Defrance M,, Sotiriou C, Fuks F. DNA methylation-based immune response signature improves patient diagnosis in multiple cancers. J Clin Invest. 2017 Aug 1;127(8):3090-3102. doi: 10.1172/JCI91095. Epub 2017 Jul 17.

[59] Lin LH, Hedayat A, W. W. Statistical Tools for Measuring Agreement. (Springer, 2012)

[60] Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. Theriogenology. 2010 Jun;73(9):1167-79. doi: 10.1016/j.theriogenology.2010.01.003.

[61] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986 Feb 8;1(8476):307-10.

[62] Passing H, Bablok W (1983). "A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in Clinical Chemistry, Part I". Journal of Clinical Chemistry & Clinical Biochemistry. 21 (11): 709–20. doi:10.1515/cclm.1983.21.11.709. PMID 6655447

[63] Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". Educational and Psychological Measurement. 20 (1): 37–46. doi:10.1177/001316446002000104.

[64] Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. Theriogenology. 2010 Jun;73(9):1167-79. doi: 10.1016/j.theriogenology.2010.01.003.

[65] Swisher SK, Wu Y, Castaneda CA, Lyons GR, Yang F, Tapia C, Wang X, Casavilca SA, Bassett R, Castillo M, Sahin A, Mittendorf EA. Interobserver Agreement Between Pathologists Assessing Tumor-Infiltrating Lymphocytes (TILs) in Breast Cancer Using Methodology Proposed by the International TILs Working Group.Ann Surg Oncol. 2016 Jul;23(7):2242-8. doi: 10.1245/s10434-016-5173-8. Epub 2016 Mar 10.

[66] Buisseret L, Desmedt C, Garaud S, Fornili M, Wang X, Van den Eyden G, de Wind A, Duquenne S, Boisson A, Naveaux C, Rothé F, Rorive S, Decaestecker C, Larsimont D, Piccart-Gebhart M, Biganzoli E, Sotiriou C, Willard-Gallo K. Reliability of tumor-infiltrating lymphocyte and tertiary lymphoid structure assessment in human breast cancer. Mod Pathol. 2017 Sep;30(9):1204-1212. doi: 10.1038/modpathol.2017.43. Epub 2017 Jun 16.

[67] Khoury, T., Peng, X., Yan, L., Wang, D. & Nagrale, V. Tumor-Infiltrating Lymphocytes in Breast Cancer. Am. J. Clin. Pathol. 150, 441–450 (2018). [68] O'Loughlin et al., 2018.

[69] Tramm T, Di Caterino T, Jylling AB, Lelkaitis G, Lænkholm AV, Ragó P, Tabor TP, Talman MM, Vouza E; Scientific Committee of Pathology, Danish Breast Cancer Group (DBCG). Standardized assessment of tumor-infiltrating lymphocytes in breast cancer: an evaluation of inter-observer agreement between pathologists. Acta Oncol. 2018 Jan;57(1):90-94. doi: 10.1080/0284186X.2017.1403040. Epub 2017 Nov 23.

[70] Zhang L, Conejo-Garcia JR, Katsaros D, Gimotty PA, Massobrio M, Regnani G, Makrigiannakis A, Gray H, Schlienger K, Liebman MN, Rubin SC, Coukos G. Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer. N Engl J Med. 2003 Jan 16;348(3):203-13.

[71] Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pagès C, Tosolini M, Camus M, Berger A, Wind P, Zinzindohoué F, Bruneval P, Cugnenc PH, Trajanoski Z, Fridman WH, Pagès F. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science. 2006 Sep 29;313(5795):1960-4.

[72] Chen Z, Chen X, Zhou E, Chen G, Qian K, Wu X, Miao X, Tang Z. Intratumoral CD8$^+$ cytotoxic lymphocyte is a favorable prognostic marker in node-negative breast cancer. PLoS One. 2014 Apr 17;9(4):e95475. doi: 10.1371/journal.pone.0095475. eCollection 2014.

[73] Catacchio, I., Scattone, A., Silvestris, N. & Mangia, A. Immune Prophets of Lung Cancer: The Prognostic and Predictive Landscape of Cellular and Molecular Immune Markers. Transl. Oncol. 11, 825–835 (2018).

[74] Catacchio, I. et al. Intratumoral, rather than stromal, CD8+ T cells could be a potential negative prognostic marker in invasive breast cancer patients. Transl. Oncol. 12, 585–595 (2019).

[75] Rosenthal R, Cadieux EL, Salgado R, Bakir MA, Moore DA, Hiley CT, Lund T, Tanić M, Reading JL, Joshi K, Henry JY, Ghorani E, Wilson GA, Birkbak NJ, Jamal-Hanjani M, Veeriah S, Szallasi Z, Loi S, Hellmann MD, Feber A, Chain B, Herrero J, Quezada SA, Demeulemeester J, Van Loo P, Beck S, McGranahan N,, Swanton C,; TRACERx consortium. Neoantigen-directed immune escape in lung cancer evolution. Nature. 2019 Mar;567(7749):479-485. doi: 10.1038/s41586-019-1032-7. Epub 2019 Mar 20.

[76] Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, Mollbrink A, Linnarsson S, Codeluppi S, Borg Å, Pontén F, Costea PI, Sahlén P, Mulder J, Bergmann O, Lundeberg J, Frisén J. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016 Jul 1;353(6294):78-82. doi: 10.1126/science.aaf2403.

[77] Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, Nainys J, Wu K, Kiseliovas V, Setty M, Choi K, Fromme RM, Dao P, McKenney PT, Wasti RC, Kadaveru K, Mazutis L, Rudensky AY, Pe'er D. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. Cell. 2018 Aug 23;174(5):1293-1308.e36. doi: 10.1016/j.cell.2018.05.060. Epub 2018 Jun 28.

[78] Keren L, Bosse M, Marquez D, Angoshtari R, Jain S, Varma S, Yang SR, Kurian A, Van Valen D, West R, Bendall SC, Angelo M. A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. Cell. 2018 Sep 6;174(6):1373-1387.e19. doi: 10.1016/j.cell.2018.08.039.

[79] Wagner J, Rapsomaniki MA, Chevrier S, Anzeneder T, Langwieder C, Dykgers A, Rees M, Ramaswamy A, Muenst S, Soysal SD, Jacobs A, Windhager J, Silina K, van den Broek M, Dedes KJ, Rodríguez Martínez M, Weber WP, Bodenmiller B. A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. Cell. 2019 May 16;177(5):1330-1345.e18. doi: 10.1016/j.cell.2019.03.005. Epub 2019 Apr 11.

[80] Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, Ziv E, Culhane AC, Paull EO, Sivakumar IKA, Gentles AJ, Malhotra R, Farshidfar F, Colaprico A, Parker JS, Mose LE, Vo NS, Liu J, Liu Y, Rader J, Dhankani V, Reynolds SM, Bowlby R, Califano A, Cherniack AD, Anastassiou D, Bedognetti D, Mokrab Y, Newman AM, Rao A, Chen K, Krasnitz A, Hu H, Malta TM, Noushmehr H, Pedamallu CS, Bullman S, Ojesina AI, Lamb A, Zhou W, Shen H, Choueiri TK, Weinstein JN, Guinney J, Saltz J, Holt RA, Rabkin CS; Cancer Genome Atlas Research Network, Lazar AJ, Serody JS, Demicco EG, Disis ML, Vincent BG, Shmulevich I. The Immune Landscape of Cancer. Immunity. 2018 Apr 17;48(4):812-830.e14. doi: 10.1016/j.immuni.2018.03.023. Epub 2018 Apr 5.

[81] Crosetto, N., Bienko, M. & van Oudenaarden, A. Spatially resolved transcriptomics and beyond. Nat. Rev. Genet. 16, 57–66 (2015).

[82] Sobral-Leite, M. et al. (2018) 'Assessment of PD-L1 expression across breast cancer molecular subtypes, in relation to mutation rate, BRCA1 -like status, tumor-infiltrating immune cells and survival', OncoImmunology, 7(12), p. e1509820. doi: 10.1080/2162402X.2018.1509820.

[83] Camp, R. L., Charette, L. A. and Rimm, D. L. (2000) ‘Validation of tissue microarray technology in breast carcinoma.’, Laboratory investigation; a journal of technical methods and pathology, 80(12), pp. 1943–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11140706 (Accessed: 20 February 2019).

[84] Kyndi M, Sørensen FB, Knudsen H, Overgaard M, Nielsen HM, Andersen J, Overgaard J. Tissue microarrays compared with whole sections and biochemical analyses. A subgroup analysis of DBCG 82 b&c.

[85] Khouja MH, Baekelandt M, Sarab A, Nesland JM, Holm R. Limitations of tissue microarrays compared with whole tissue sections in survival analysis. Oncol Lett. 2010 Sep;1(5):827-831. Epub 2010 Sep 1. Acta Oncol. 2008;47(4):591-9. doi: 10.1080/02841860701851871.

[86] Kündig P, Giesen C, Jackson H, Bodenmiller B, Papassotirolopus B, Freiberger SN, Aquino C, Opitz L, Varga Z. Limited utility of tissue micro-arrays in detecting intra-tumoral heterogeneity in stem cell characteristics and tumor progression markers in breast cancer. J Transl Med. 2018 May 8;16(1):118. doi: 10.1186/s12967-018-1495-6.

[88] Simoni Y, Becht E, Fehlings M, Loh CY, Koo SL, Teng KWW, Yeong JPS, Nahar R, Zhang T, Kared H, Duan K, Ang N, Poidinger M, Lee YY, Larbi A, Khng AJ, Tan E, Fu C, Mathew R, Teo M, Lim WT, Toh CK, Ong BH, Koh T, Hillmer AM, Takano A, Lim TKH, Tan EH, Zhai W, Tan DSW, Tan IB, Newell EW. Bystander CD8+ T cells are abundant and phenotypically distinct in human tumour infiltrates. Nature. 2018 May;557(7706):575-579. doi: 10.1038/s41586-018-0130-2. Epub 2018 May 16.

[89] Scheper W, Kelderman S, Fanchi LF, Linnemann C, Bendle G, de Rooij MAJ, Hirt C, Mezzadra R, Slagter M, Dijkstra K, Kluin RJC, Snaebjornsson P, Milne K, Nelson BH, Zijlmans H, Kenter G, Voest EE, Haanen JBAG, Schumacher TN. Low and variable tumor reactivity of the intratumoral TCR repertoire in human cancers. Nat Med. 2019 Jan;25(1):89-94. doi: 10.1038/s41591-018-0266-5. Epub 2018 Dec 3.

[90] Binnewies M, Roberts EW, Kersten K, Chan V, Fearon DF, Merad M, Coussens LM, Gabrilovich DI, Ostrand-Rosenberg S, Hedrick CC, Vonderheide RH,0, Pittet MJ, Jain RK, Zou W, Howcroft TK, Woodhouse EC, Weinberg RA, Krummel MF. Understanding the tumor immune microenvironment (TIME) for effective therapy. Nat Med. 2018 May;24(5):541-550. doi: 10.1038/s41591-018-0014-x. Epub 2018 Apr 23.

[91] Thommen DS, Koelzer VH, Herzig P, Roller A, Trefny M, Dimeloe S, Kiialainen A, Hanhart J, Schill C, Hess C, Savic Prince S, Wiese M, Lardinois D, Ho PC, Klein C, Karanikas V, Mertz KD, Schumacher TN, Zippelius A. A transcriptionally and functionally distinct PD-1+ CD8+ T cell pool with predictive potential in non-small-cell lung cancer treated with PD-1 blockade. Nat Med. 2018 Jul;24(7):994-1004. doi: 10.1038/s41591-018-0057-z. Epub 2018 Jun 11.

[92] Mani NL, Schalper KA,, Hatzis C, Saglam O, Tavassoli F, Butler M, Chagpar AB, Pusztai L, Rimm DL. Quantitative assessment of the spatial heterogeneity of tumor-infiltrating lymphocytes in breast cancer. Breast Cancer Res. 2016 Jul 29;18(1):78. doi: 10.1186/s13058-016-0737-x.

[93] Herbst RS, Soria JC, Kowanetz M, Fine GD, Hamid O, Gordon MS, Sosman JA, McDermott DF, Powderly JD, Gettinger SN, Kohrt HE, Horn L, Lawrence DP, Rost S, Leabman M, Xiao Y, Mokatrin A, Koeppen H, Hegde PS, Mellman I, Chen DS, Hodi FS. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. Nature. 2014 Nov 27;515(7528):563-7. doi: 10.1038/nature14011.

[94] Gruosso T, Gigoux M, Manem VSK, Bertos N, Zuo D, Perlitch I, Saleh SMI, Zhao H, Souleimanova M, Johnson RM, Monette A, Ramos VM, Hallett MT, Stagg J, Lapointe R, Omeroglu A, Meterissian S, Buisseret L, Van den Eynden G, Salgado R, Guiot MC, Haibe-Kains B, Park M. Spatially distinct tumor immune microenvironments stratify triple-negative breast cancers. J Clin Invest. 2019 Apr 1;129(4):1785-1800. doi: 10.1172/JCI96313. Epub 2019 Mar 18.

[95] Ambrogi F, Biganzoli E, Boracchi P. Estimates of clinically useful measures in competing risks survival analysis. Stat Med 2008; 27: 6407–6425.

[96] Kay R. Treatment effects in competing-risks analysis of prostate cancer. Biometrics 1986; 42:203–211.

[97] Klein JP. Modelling competing risks in cancer studies. Statistics in Medicine 2006; 25:1015–1034.

[98] Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association 1999; 94:496–509.

[99] Scheike T. A flexible semiparametric transformation model for survival data. Lifetime Data Analysis 2006; 12:461–480.

[100] Fine JP. Analysing competing risks data with transformation models. Journal of the Royal Statistical Society, Series B 1999; 61(4):817–830.

[101]. Fine JP. Regression modeling of competing crude failure probabilities. Biostatistics 2001; 2:85–97.

[102] Klein JP, Andersen PK. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. Biometrics 2005; 61:223–229.

[103] Recurrence dynamics of breast cancer according to baseline body mass index Biganzoli, Elia et al., European Journal of Cancer 2017, Volume 87, 10 – 20

[104] Lauby-Secretan B, Scoccianti C, Loomis D, Grosse Y, Bianchini F, Straif K. Body fatness and cancereviewpoint of the IARC working group. N Engl J Med 2016;375:794e8.

[105]] Jiralerspong S, Goodwin PJ. Obesity and breast cancer prognosis: evidence, challenges, and opportunities. J Clin Oncol 2016;34: 4203e16.

[106] Piccart MJ, Di Leo A, Beauduin M, Vindevoghel A, Michel J, Focan C, et al. Phase III trial comparing two dose levels of epirubicin combined with cyclophosphamide with cyclophosphamide, methotrexate, and fluorouracil in node-positive breast cancer. J Clin Oncol 2001;19:3103e10.

[107] Organization WH. Report of a WHO Expert Consultation: World Health Organization Technical Report. Series number 854. Geneva: World Health Organization; 1995. Physical status: the use and interpretation of anthropometry. 1995

[108] Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation. 2016;133(6):601–609. doi:10.1161/CIRCULATIONAHA.115.017719

[109] Andersen et al. 1993

[110] Fine, J.P.. (2001). Regression modeling of competing crude failure probabilities. Biostatistics (Oxford, England). 2. 85-97. 10.1093/biostatistics/2.1.85.

[111] Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. Am J Epidemiol. 2009;170:244–256. doi: 10.1093/aje/kwp107.

[112] Koller MT, Raatz H, Steyerberg EW, Wolbers M. Competing risks and the clinical community: irrelevance or ignorance? Stat Med. 2012;31:1089– 1097. doi: 10.1002/sim.4384.

[113] Fine JP. Analysing competing risks data with transformation models. Journal of the Royal Statistical Society, Series B 1999; 61(4):817–830.

[114] Fine JP. Regression modeling of competing crude failure probabilities. Biostatistics 2001; 2:85–97.

[115] Andersen PK, Klein JP, Rosthøj S. Generalized linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika 2003; 90:15–27.

[116] Miller, Rupert G. “The Jackknife--A Review.” Biometrika, vol. 61, no. 1, 1974, pp. 1–15. JSTOR, www.jstor.org/stable/2334280.

[117] Kung-Yee Liang and Scott Zeger (1986). "Longitudinal data analysis using generalized linear models". Biometrika. 73 (1): 13–22. doi:10.1093/biomet/73.1.13.

[118] Fong, Y; Rue, H; Wakefield, J (2010). "Bayesian inference for generalized linear mixed models". Biostatistics. 11 (3): 397–412. doi:10.1093/biostatistics/kxp053. PMC 2883299. PMID 19966070

[119] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Proceedings of the Second International Symposium on Information Theory, B. N. Petrov and F. Csaki (eds), 267-281. Budapest: Akademiai Kiado.

[120] Pan, W. (2001), Akaike's Information Criterion in Generalized Estimating Equations. Biometrics, 57: 120-125. doi:10.1111/j.0006-341X.2001.00120.x

[121] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika 61, 439-447.

[122] Rosenberg, Philip S. “Hazard Function Estimation Using B-Splines.” Biometrics, vol. 51, no. 3, 1995, pp. 874–887. JSTOR, www.jstor.org/stable/2532989.

[123] Hutton JL. Number needed to treat: properties and problems. Journal of the Royal Statistical Society, Series A 2000; 163:381–402.

[124] . Altman DG. Confidence intervals for the number needed to treat. British Medical Journal 1998; 317:1309–1312.

[125] Frank E Harrell Jr (2001), Regression modeling strategies, Springer-Verlag, Sections 9.2, 9.6, 10.5

[126] J. D. Kalbfleisch, Ross Prentice. The Statistical Analysis of Failure Time Data. John Wiley & Sons, New York, 980, pag 168.