Lebanese University
Faculty of Science, Applied Mathematics

January 3, 2023

Survival Analysis Project:

Exploring Colorectal Cancer Progression

Mean Sojourn Time, Lead Time Bias, Survival Correction, Factors Impact and Hazard on Survival.

Ali Karbala

Supervised by Dr. Abbass Mourad

Abstract: Determining the survival time and risk factors of colorectal cancer needs the survival time function which is subjected to lead time bias. Sojourn time is essential for determining the lead time, calculated for non-aggressive, mid-aggressive and aggressive types of cancer and assumed to have exponential distribution. The author employs two correction approaches for survival time, Schwartz and Duffy, and the available factors tested whether it have impact on the survival of patients using Kaplan-Meir and Cox model. The study is demonstrated using a cohort of 550 patients exposed to colorectal cancer with 16 distinct factors.

1 Introduction

Cancer is a worldwide concern, due to its harmfulness, cure difficulty and cost, specially at advanced stages. Cancer is a generic term for a large group of diseases that can affect any part of the body. Other terms used are malignant tumours and neoplasms. One defining feature of cancer is the rapid creation of abnormal cells that grow beyond their usual boundaries, and which can then invade adjoining parts of the body and spread to other organs, it arises from the transformation of normal cells into tumour cells in a multi-stage process that generally progresses from a pre-cancerous lesion to a malignant tumour. (WHO,2022).

The poor outcomes for cancers diagnosed at an advanced stage have been the driver behind research into techniques to detect disease before symptoms are manifest. Statistical techniques and methods, have a solid role in cancer research, where researchers depend on, covers helpful ideas in early cancer detection. In order to help researchers and epidimiologists to asses guide lines for investigating cancers progress for high risk patients, this project covers estimating sojourn time and lead time for cancer patients using statistical methods and tools in order to correct survival time for patients from lead time bias, testing the factors that have impact on survival time and estimating hazard ratios and odds ratios for factors. The methodology employed in this project is applicable to various cancer types, however, the

empirical statistics were specifically derived from a cohort of patients diagnosed with colorectal cancer.

1.1 Literature Review

There is extensive research articles covers sojourn time, lead time and survival corrections for various cancers, including lung, cervical, breast, and colorectal cancer. Numerous articles dive into this topic, employing a diverse array of statistical methods and models including Marcov chains, Cox model Kaplan Meir and others. Researchers frequently leverage cohort studies and survival analysis, with the help of advanced screening methods, imaging methods like tomography (CT scans) and magnetic resonance imaging (MRI) and laboratory tests like cancer marker identification to understand the details. Gill Lawrence and partners published in 2008 an article about estimating survival in women taking into account lead time and length bias. (Lawrence,2008). Another aproach in 2008, CR Chien, THH Chen worked on estimating mean sojourn time for lung cancer screening with computed tomograpy (Chein,2008).

2 Materials and Methods

This cohort is a study that collected data on the clinical features and treatment allocations of patients with newly diagnosed

CRC. We took advantage of this large prospective cohort to compare the clinical features at diagnosis, the therapeutic allocations, and the outcomes from alcohol-related and non-alcohol-related CRC.

2.1 Descriptive Statistics

A cohort study was conducted involving 550 patients, incorporating both qualitative and quantitative variables as shown in tables (1) and (2), respectively. The description include graphical representation of tumor size and survival days, presented through boxplots in figure (1) and histograms in figure (2). Mortality rates were determined, as well as rates for screened and non-screened patients. Additionally, relative risk assessments for factors influencing survival were conducted, detailed in table (3), along with an examination of the impact of tumor size on patient survival, represented through a fitted logistic model in figure (3) and equation (6).

Factor	Category 1	Category 2
Gender	Female: 18	Male: 82
Alcohol	Non-Alcoholic: 30	Alcoholic: 70
Vegetarian	No: 91	Yes: 9
Over Weight	No: 88	Yes: 12
Screening	No: 76	Yes: 24
Diabetes	No: 68	Yes: 32
Smoker	No: 60	Yes: 40
Metastatic	No: 85	Yes: 15
Curative Treatment	No: 79	Yes: 21
Status	Alive: 23	Dead: 77

Factor	Category 1	Category 2	Category 3	Category 4
Marital Status	Divorced: 1.4	Married: 85	Single: 1.8	Widowed: 12
Cancer Stage	A: 14	B: 5	C: 46	D: 35

Table 1: Percentage of qualitative variables according to categories per 100 patients

Variable	Min	1st Q	Median	3rd Q	Max	Mean	sd
Age	22	60	68	76	99	67.42	10.82
Tumor Size (mm)	10	28	45	77	900	58.81	58.88
Survival Days	1	51	78	406	2155	315	391.47

Table 2: Quantitative variables representation.

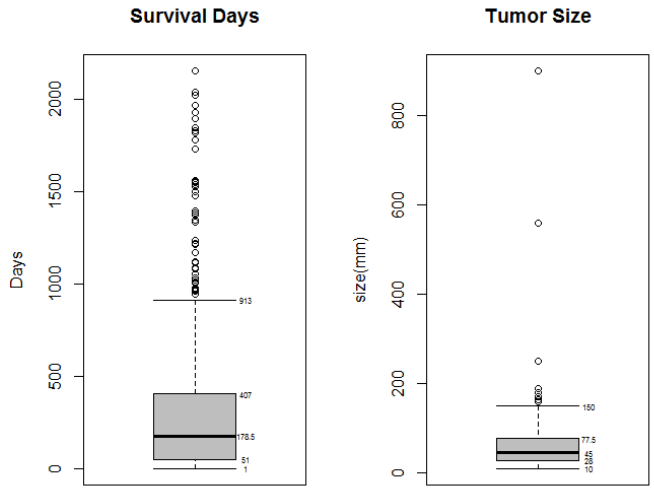

Figure 1: Boxplots for Survival days and Tumor size (mm).

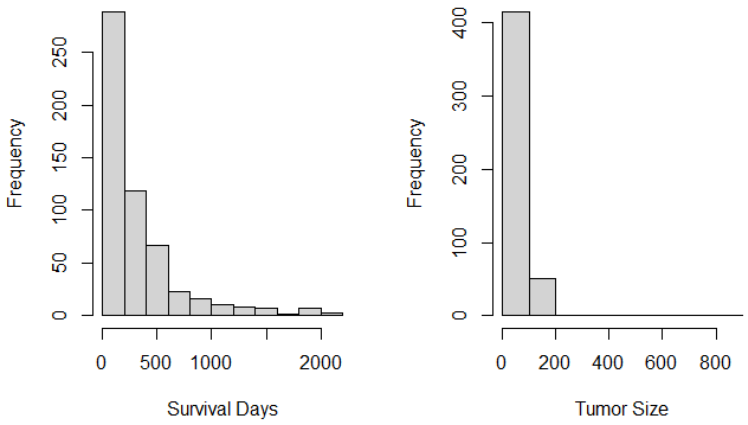

Figure 2: Histograms for Survival days and Tumor size (mm).

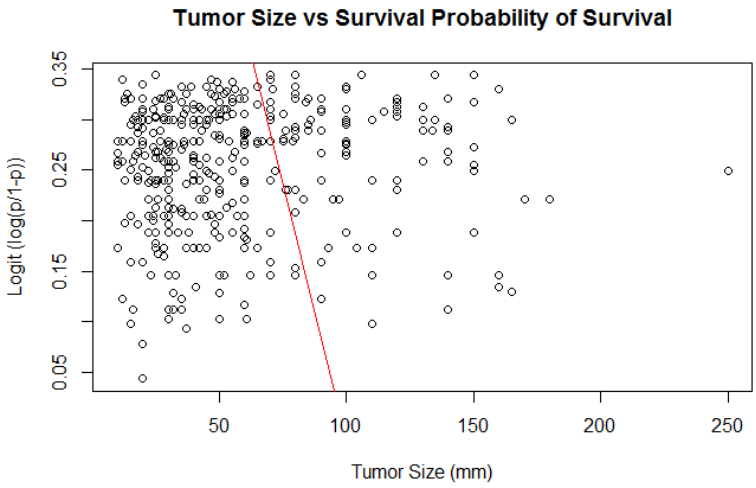

Figure 3: Impact of Tumor Size on Survival.

Factor	R1	R0	Relative Risk (R1/R0)
GenderF	0.73	0.78	0.93
AlcoolNo	0.71	0.79	0.90
VegetarianNo	0.76	0.72	1.05
OverweightNo	0.77	0.71	1.08
ScreeningNo	0.82	0.61	1.35
DiabetesNo	0.79	0.72	1.09
SmokerNo	0.79	0.72	1.08
Curative TreatmentNo	0.85	0.49	1.77

Table 3: Relative Risk for factors on Survival

2.2 Sojourn time, Lead time and Doubling time

Sojourn time, Lead time and Doubling time are key terms in survival analysis, a definition for every term is provided below in addition to mathematical formulation in the next part. Sojourn time is the length of the pre-clinical screen-detectable phase, a period when a test can detect asymptomatic disease. Mean sojourn time (MST) is an important factor in determining appropriate screening intervals (Zheng,2012). A long mean sojourn time will indicate a good potential for screening. The shorter the mean sojourn time, the more frequently screening will have to take place in order to be effective. If mean sojourn time is very short, then it may not be worth screening at all. Lead time is the time by which the diagnosis is anticipated by screening or surveillance with respect to the clinical presentation of a disease. It represents an artificial addition of time to survival of cases detected during screening, leading to a specious improvement in prognosis(Cucchetti, 2014) . Doubling time is the time the size of the tumor will be the double. (DT) is widely used for quantification of tumor growth rate. DT is usually determined from two volume estimations with measurement time intervals comparable with or shorter than DT (Mehrara, 2007). article

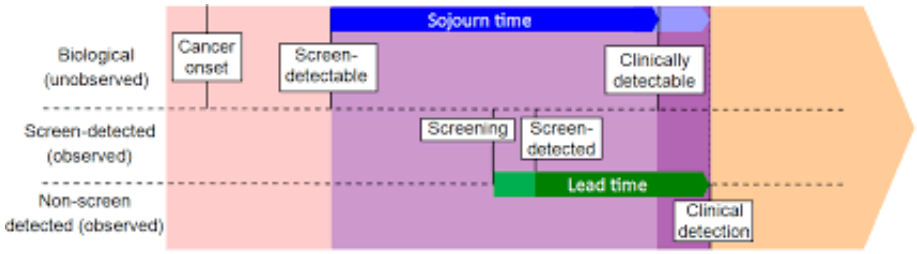

Figure 4: Sojourn time and Lead time.

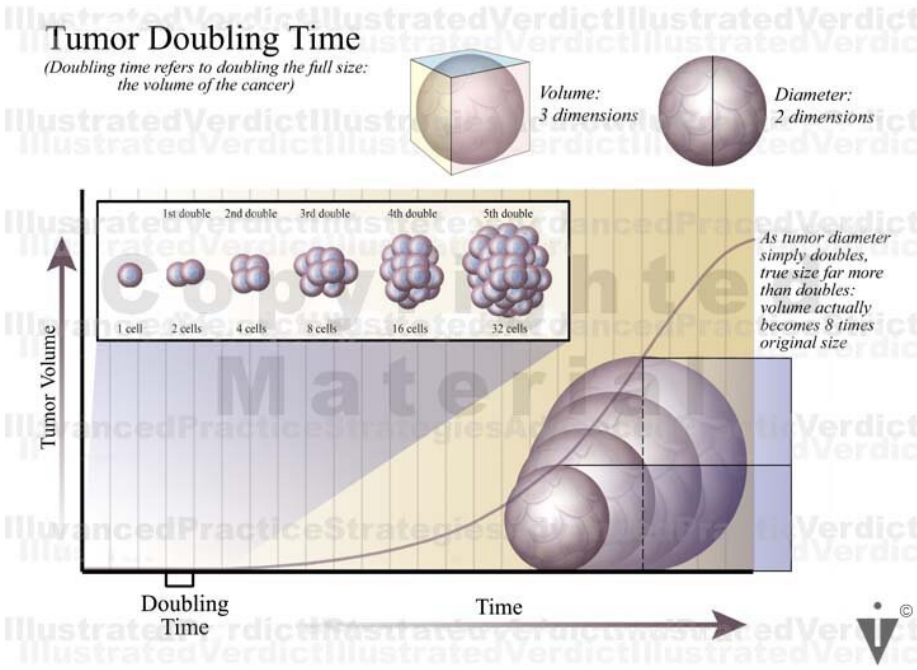

Figure 5: Tumor doubling time.

2.3 Mathematical Formulation

The prerequisites for correcting survival time, the core aim in this project, is the sojourn time and lead time. There are two ap-

proaches to lead time estimation, Mordecai Schwartz approach in his paper "A BIOMATHEMATICAL APPROACH TO CLINICAL TUMOR GROWTH" (Schwartz,1961) and Stephen W. Duffy and partners approach in their paper "Correcting for Lead Time and Length Bias in Estimating the Effect of Screen Detection on Cancer Survival" (Duffy,2008). Both approaches used the mean sojourn time for estimating the lead time bias where it is assumed to have a certain distribution or estimated by the following formula:

$$ST = 3 \cdot DT \cdot \frac{\log\left(\frac{dt1}{dt0}\right)}{\log 2} \tag{1}$$

where DT is the doubling time, dt1 is the tumor size of at time t and dt0 is the minimum tumor size detected.

2.3.1 Schwartz Approach

Schwartz states the following formula:

$$t_d = \frac{t \cdot \log_2}{3 \cdot \log\left(\frac{D_T}{D_0}\right)} \tag{2}$$

where td equals the doubling time, representing the time for the whole tumor, t equals the time of the final or second measurement of the tumor size, which represents the time lapse between the 2 measurements, Do equals the initial diameter of the tumor and DT equals the tumor diameter at time t. Lead time can be extracted from the formula, which is t, when DT and D0 be

the clinical detected measurements and screening detected measurements respectively.

2.3.2 Duffy Approach

Duffy and partners formulate the expected additional follow up time due to lead time bias as the expectation of the lead time conditional on its being less than t where t is the time of death. lambda in the equation below is the mean sojourn time assuming that it is exponentially distributed.

$$E(s) = \frac{1 - e^{-\lambda t} - \lambda t e^{-\lambda t}}{\lambda(1 - e^{-\lambda t})} \tag{3}$$

In case the patient still alive at time t, then

$$E(s) = \frac{1 - e^{-\lambda t}}{\lambda} \tag{4}$$

2.4 Statistical Tools

R studio 2022.07.2 Build 576 was used for descriptive statistics, inference, modeling and simulations.

2.5 Corrected Survival Time.

$$CorrectedSurvivalTime = SurvivalTime - LeadTimeBias \tag{5}$$

Corrected survival time is the empirical survival time for screened patients subtracted from lead time bias. Below is the steps for estimating lead time to determine the corrected survival time. Simulations for distributions are tested for exponential family distributions, Gamma, Weibull, Exponentail and Lognormal, the best distribution is taken according to LogRank test and Kolmogrov-Smirnov test, this is detailed in table 4 and represented graphically in figure 6.

Distribution	Patients Tumor Size		Survival Time
	Screened	Non-Screened	
Gamma	-530.54	-1672.75	-955.16
	0.569	0.459	0.370
Weibull	-539.40	-1679.48	-955.25
	0.88	0.57	0.372
Exponential	-570.39	-1740.97	-955.84
	0.005	0.005	0.131
Log-Normal	-524.24	-1673.95	-965.68
	0.25	0.186	0.008

Table 4: LogRank test - LR and Kolmogorov-Smirnov test - KS p-value for screened and non-screened patients and survival time distribution.

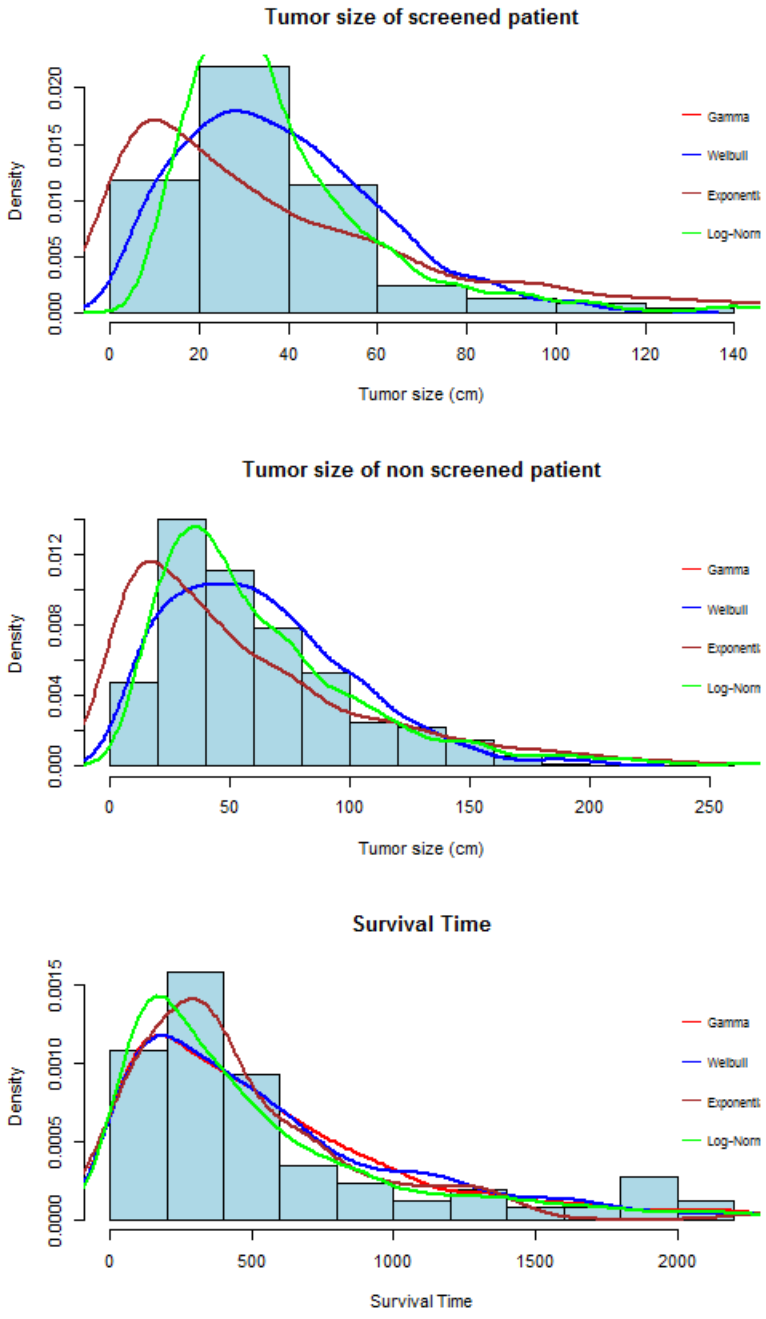

Figure 6: Simulations of Survival Time t and tumor size dt0 and dt1 distributions for screened and non screened patients.

2.5.1 Doubling Time

The doubling time DT were taken for three levels of cancer aggressiveness, 112, 211, 404 days for aggressive, mid-aggressive non-aggressive tumor respectively.

2.5.2 MST Estimation

Equation (1) is used for estimating MST, tumor size dt1 taken equal to non-screened patients distribution and simulated to better fit exponential distribution, the initial tumor size dt0 assumed to follow triangular distribution (1;1.2;1.4).

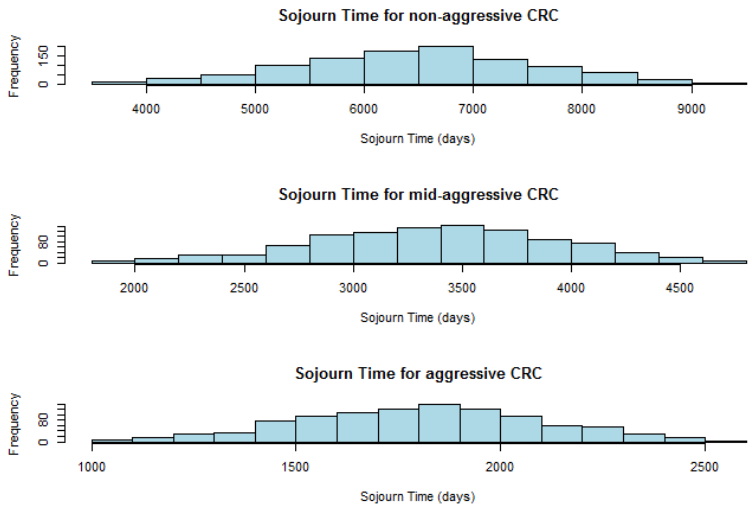

Figure 7: Sojourn time for three types of cancer aggressiveness.

2.5.3 Lead Time Estimation

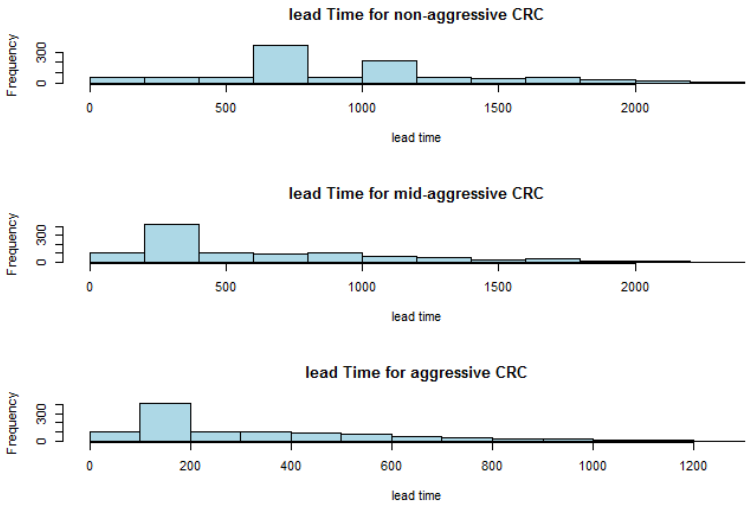

Figure 8: Schwartz lead time for three types of cancer aggressiveness.

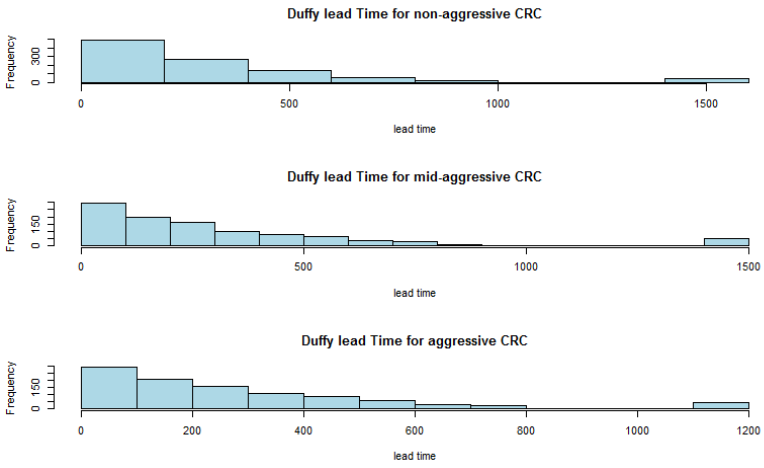

Figure 9: Duffy lead time for three types of cancer aggressiveness.

Two lead times estimated each for equation (2) and (3) and (4). In Shwartz approach using equation (2), the tumor size at detectable level DT assumed to fit the distribution of tumor size for non-screened patients (exponential), and the tumor size at screen level Do taken equal to screened patients distribution and simulated to better fit exponential distribution. In Duffy approach using equation (3) and (4), lambda is the expected value of the sojourn time which is 1 divided by MST, t is the survival time simulated to fit lognormal distribution.

2.5.4

Non-Aggressive CRC Case

	Min	1st Q	Median	3rd Q	Max	Mean	sd
MST	3700	5737	6515	7206	9082	6483	1051.448
Lead Time (Schwartz)	7.321	683.181	683.181	1621.714	4479	1154.07	839.67
Lead Time (Duffy))	0.467	83.99	182.64	378.97	1591.56	294.76	336.49

Mid-Aggressive CRC Case

	Min	1st Q	Median	3rd Q	Max	Mean	sd
MST	1933	2996	3403	3764	4743	3386	549.1476
Lead Time (Schwartz)	3.824	356.81	356.81	846.98	2339.29	602.748	438.54
Lead Time (Duffy)	0.4674	83.6613	181.04	371.95	1411.52	282.55	304.40

Survival Correction

Aggressive CRC Case

	Min	1st Q	Median	3rd Q	Max	Mean	sd
MST	1026	1590	1806	1998	2518	1797	291.4907
Lead Time (Schwartz)	2.03	189.4	189.4	449.58	1241.71	319.94	232.78
Lead Time (Duffy)	0.467	83.04	178.08	359.01	1146.66	262.91	257.58

Table 5: Mean sojourn time (MST) and lead time for colorectal cancer (CRC).

2.5.5 Survival Correction

Equation (5) is used to determine the corrected survival time using Schwartz and Duffy methods, determined for the screened patients assuming three types of cancer, detailed below in Table (5).

2.5.6

Non-Aggressive CRC Case

Survival	Min	1st Q	Median	3rd Q	Max	Mean	sd
Original	3.0	228.8	388.5	661.0	2155.0	574.0	538.825
Corrected (Schwartz)	0.289	138.163	307.730	584.925	3816.173	427.157	420.04
Corrected (Duffy))	0.467	85.682	190.686	414.298	3352.697	307.975	340.68

Mid-Aggressive CRC Case

Survival	Min	1st Q	Median	3rd Q	Max	Mean	sd
Original	3.0	228.8	388.5	661.0	2155.0	574.0	538.825
Corrected (Schwartz)	0.935	118.088	265.626	546.634	4142.544	415.509	463.20
Corrected (Duffy)	0.4674	85.0625	187.7314	401.360	3087.8303	288.3345	300.751

Survival Correction

Aggressive CRC Case

Survival	Min	1st Q	Median	3rd Q	Max	Mean	sd
Original	3.0	228.8	388.5	661.0	2155.0	574.0	538.825
Corrected (Schwartz)	0.167	99.13	235.44	544.20	4309.957	426.753	519.31
Corrected (Duffy)	0.467	84.73	186.13	390.94	2907.78	276.12	277.321

Table 6: Quartiles for corrected and non-corrected survival time (days) for different aggressiveness levels of colorectal cancer.

2.6 Estimating The Survival Function and The Covariates That have Impact On Survival

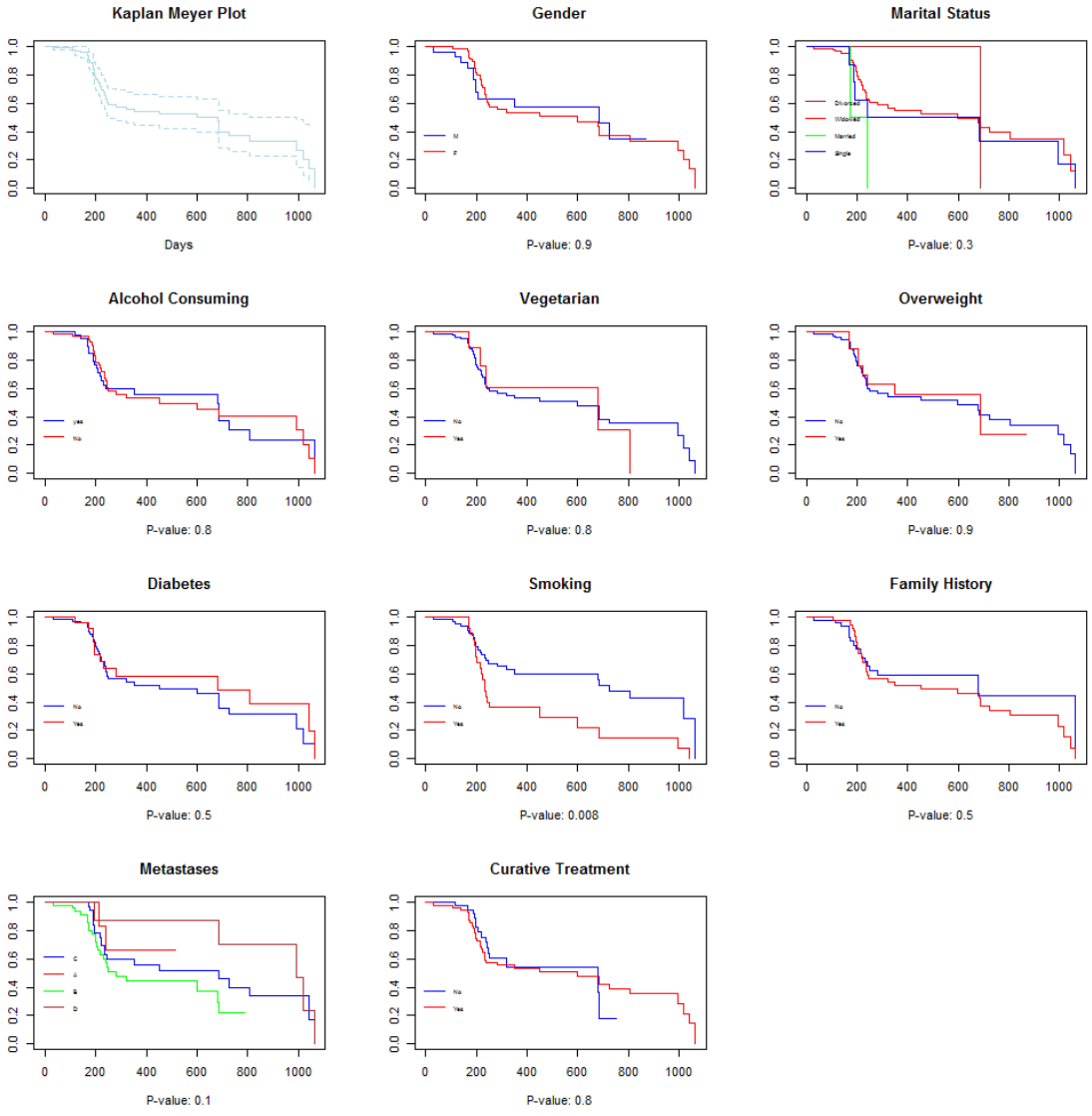

Figure 10: Kaplan Meir plot showing the impacts of different factors on survival time of screened patients.

Survival data is determined using survival packages in R, then Kaplan Meir model applied to estimate survival function. Different factors tested weather it affects patients survival, using Chi-squared test as detailed in figure(10).

2.7 Hazard of Factors on Survival.

Cox model fitted to the survival function, the exponential of the coefficients of the model are defined as the hazard ratio. Table (6) detailed the hazard ratio for every factor on the survival of screened patients.

Factor	Hazard Ratio	P-value
GenderM	0.72	0.44
AlcoolYes	0.52	0.24
VegetarianYes	1.27	0.70
OverweightYes	0.51	0.27
Cancer Stage B	1.2	0.82
Cancer Stage C	2.54	0.013
Cancer Stage D	0.78	0.71
DiabetesYes	0.73	0.48
SmokerYes	2.47	0.013
Curative TreatmentYes	1.27	0.53
Whole Model		0.07

Table 7: Hazard ratios of factors affecting survival time of screened patients.

In fitting Kaplan Meir and Cox models, the used corrected survival data are the Duffy corrected aggressive case.

2.8 Discussion

The majority of the participants are males (82%), alcoholic (70%), non vegetarian (91%), not over-weighted (88%), they do screening (76%), non diabetic (86%), non smokers (60%), have no metastatic case (85%), and go for curative treatment (79%). 77% of the participants died at the end of the study. The mean age of the participants and the standard deviation are (67.42, 10.82), which shows that the participants selected are near the geriatric age. The tumor size and survival factor mean and standard deviation are 58.81,58.88 and 315,391.47 respectively. High values for standard deviation indicates that there is a dispersion of tumor sizes and survival days among patients. Box plots and histograms of tumor size and survival days show a positive skewness, indicates that the majority of the participants had a smaller tumor sizes and died early. Moreover, outliers for both factors are detected. The mortality rate confidence interval is [0.73,0.80], indicates that 73% to 80% of the population died from colorectal cancer. Relative risk analysis claim that patients that do not screen are 35% more likely to die than patients that do screen and patients that dont go through curative treatments are 77% more likely to die than patients that go through curative treatments. The fitted logistic model equation showing the impact of tumor size on

survival is :

$$logit(Survival) = -0.546338 - 0.010155 \times TumorSize. \quad (6)$$

The intercept and the TumorSize parameters is significant (p-value ¡ 0.001) and the AIC for the model is 520. The model shows a negative impact of tumor size on survival, as tumor size increases the survival time decreases. The mean sojourn time and standard deviation for non-aggressive, mid-aggressive and aggressive cases are (17.7, 2.87), (9.27, 1.5) and (4.92, 0.8) years respectively. This shows that approxiamtely for non-aggressive, mid-aggressive and aggressive cases the time for the cancer to shift from screened detected to clinically detected phase are 15-20, 8-11 and 4-6 years for colorectal cancer patients. The Schwartz lead time and standard deviation for non-aggressive, mid-aggressive and aggressive cases are (3.15, 2.29), (1.65, 1.2) and (0.87, 0.63) years and Duffy lead time and standard deviation are (0.71, 0.7), (0.77, 0.83) and (0.8, 0.92) years respectively. This results conclude that Schwartz correction is sensitive to the type of cancer while Duffy correction are not. Focusing on the aggressive case and taking Duffy correction the mean and standard deviation of corrected survival are (0.84, 0.93) years, Kaplan Meir model present that that the only significant impact on survival is the smoking factor, smokers are more likely to die than non-smokers. Hazard ratios shows that smokers are 2.47 times more likely to die than non-smokers and patients with cancer stage C are 2.54 times more likely to die than patients who don't have cancer stage C.

This study's findings are valuable for epidemiologists and physicians in evaluating and directing recommendations for patients with colorectal cancer. They provide essential insights into the timing of screening and risk factors. Moreover, the analytical approach employed here can be extrapolated to other cancer types with distinct cohorts. In conclusion, this research contributes crucial information that can guide clinical decisions and benefit a broader range of cancer patients.

A References

World Health Organization. (2023).

Lawrence, G., Wallis, M., Allgood, P. et al. Population estimates of survival in women with screen-detected and symptomatic breast cancer taking account of lead time and length bias. Breast Cancer Res Treat 116, 179–185 (2009). https://doi.org/10.1007/s10549-008-0100-8

C.R. Chien, T.H. Chen Mean sojourn time and effectiveness of mortality reduction for lung cancer screening with computed tomographyInt J Cancer, 122 (11) (2008), pp. 2594-2599,.

Zheng, Wenying, and Carolyn M. Rutter. "Estimated mean so-

journ time associated with hemoccult SENSA for detection of proximal and distal colorectal cancer." Cancer epidemiology, biomarkers prevention 21.10 (2012): 1722-1730.

A. Cucchetti, F. Trevisani, A. Pecorelli, V. Erroi, F. Farinati, F. Ciccarese, et al. Estimation of lead-time bias and its impact on the outcome of surveillance for the early diagnosis of hepatocellular carcinoma J Hepatol, 61 (2) (2014), pp. 333-341

Mehrara, Esmaeil, et al. "Specific growth rate versus doubling time for quantitative characterization of tumor growth rate." Cancer research 67.8 (2007): 3970-3975.

SCHWARTZ M. A biomathematical approach to clinical tumor growth. Cancer. 1961 Nov-Dec;14:1272-94. doi: 10.1002/1097-0142(196111/12)14:6¡1272::aid-cncr2820140618¿3.0.co;2-h. PMID: 13909709.

Duffy SW, Nagtegaal ID, Wallis M, Cafferty FH, Houssami N, Warwick J, Allgood PC, Kearins O, Tappenden N, O'Sullivan E, Lawrence G. Correcting for lead time and length bias in estimating the effect of screen detection on cancer survival. Am J Epidemiol. 2008 Jul 1;168(1):98-104. doi: 10.1093/aje/kwn120. Epub

2008 May 25. PMID: 18504245