Estimate Complete the Survival Function for Real Data of Lung Cancer Patients

Abbas N. Salman
Ibtehal H. Farhan
Dept. of Mathematics / College of Education for Pure science (Ibn AL-Haitham) / University of Baghdad

Received in :8 June 2014 , Accepted in: 1 September 2014

Abstract

In this paper, we estimate the survival function for the patients of lung cancer using different nonparametric estimation methods depending on sample from complete real data which describe the duration of survivor for patients who suffer from the lung cancer based on diagnosis of disease or the enter of patients in a hospital for period of two years (starting with 2012 to the end of 2013). Comparisons between the mentioned estimation methods has been performed using statistical indicator mean squares error, concluding that the survival function for the lung cancer by using shrinkage method is the best .

Keywords: Nonparametric Estimation , Lung Cancer Disease, Complete Real Data, Empirical Estimator, Borkowf –type estimator, Thompson– type estimator, and Nelson-Aalen Estimator, Mean Squares Error

1. Introduction

1.1: Preface and History

Survival analysis is one of the widely used techniques in medical statistics; its importance also arises in diverse fields such as medicine, engineering, epidemiology, biology, economics, physics, public health and or event history analysis in sociology. survival analysis involves the modeling of time to event data; in this context, death or failure is considered an "event" in the survival analysis literature – traditionally only a single event occurs for each subject, after which the organism or mechanism is dead or broken[6]. Cancer is a class of diseases when a cell or group of cells display uncontrolled growth, invasion and sometimes spread to other locations in the body via lymph or blood (metastasis)[2]. Lung cancer is the most common cancers in the world and the cause of cigarette smoking most types of lung cancer, the more the number of cigarettes smoked per day more and more beginning was in the habit of smoking in the age of the youngest whenever the risk of lung cancer the biggest, as well as the high levels of air pollution and exposure radiation and asbestos may also increase the risk of lung cancer.

Nelson, W. [11] presents theory and applications of a simple graphical method, called hazard plotting for the analysis of multiply censored life data consisting of failure times of failed units intermixed with running times on un failed units. Applications of the method are given to multiply censored data on service life of equipment, for strength data on an item with different failure modes, and for biological data multiply censored on both sides from paired comparisons. Theory for the hazard plotting method, which is based on the hazard function of a distribution, is developed from the properties of order statistics from Type II multiply censored samples.

Petrson ,A.V.[12] proved that the Kaplan-Meier estimator has consistency property and proposed an estimator for the cumulative hazard function.

Haifa, K. [5] estimated the reliability function for the tools of 14 Ramadan factory of tissues with Non-parameters Kaplan and Meier methods . She made a comparison between Kaplan and Meier method and the reliability function when failure data is exponential distribution and concluded that no differences had been significant between the two estimations.

Al-Qurashi,I.K.[1] suggested two formulas for estimating the reliability function whatever of the size of their data especially with small size data without access in the theoretical probability distributions, and comparing the proposed formulas with other parametric and non-parametric estimation methods.

Borkowf, C. B. [3] proposed a survival function under the framework of the Kaplan-Meier survival function which is called Shrunken Kaplan-Meier survival function. The Shrunken Kaplan-Meier survival function having *n* number of cases in the study and proved that these estimators performed better as compared to the Greenwood and Peto's estimators. Borkowf in his study analyzed only the variance estimators.

Mei-C. W. [9] presented some Non-Parametric estimation methods in Survival analysis and gave summary notes for survival analysis in Biostatistics

Basher, F .M.[4] presented some Non-parametric methods to estimate the reliability function with the practical application using eight estimation methods , Empirical (EM), Product Limit (PLEM) Empirical Kaplan Meier (EKMEM), Empirical Weighted Kaplan-Meier (WEKM) ,Modified Kaplan-Meier (MKMM),a weighted for reliability function(WMR) , Modified One(MMO) , Modified Two(MMT) and reached best Non-parametric method is an Empirical (EM) method using statistical indicator , integral mean square error (IMSE).

The knowledge of statistics is one of the important measurements in the pivotal trial and the method of data analysis and evaluation of results [10] In this paper, we rely on real data for patients with lung cancer was the size of the sample (118), the number of males (68) and the number of females (50) for the years 2012 and 2013 may have got more types of cancers in humans killed .

The aim of this paper is to estimate the survival function for the mentioned complete real data. Comparisons between the proposed estimation methods has been performed using statistical indicator mean squares error, concluding that the survival function for the lung cancer by using shrinkage method is the best.

Kaplan, E.L. & Meier, P. [6] suggested estimating the conditional probability of failure of time t by the observed proportion of failures of time t , and combined these estimates in the usual manner to obtain an estimate of the underlying survival distribution S(t).They studied the properties of proposed estimates and concluded the maximum likelihood estimate was strongly consistent and asymptotically normal.

1.2 Basic Concepts

1.2.1 Survival Function

The object of primary interest is the survival function, conventionally denoted by S, which is defined as [7]:

$$S(t) = \Pr(T > t) \qquad (1)$$

Where , T is a r.v. , t is the time of death, .

The survival function $S(t)$ is the probability that the patient will survive till time t.

Survival probability is usually assumed to approach zero as age increases. i.e.;

1. $S(0) = 1$.
2. $\lim_{t\to\infty} S(t) = 0$.
3. $S(t)$ is decreasing and continuous from right side.

Another characteristic of survival data is that the survival time cannot be negative[13]. See figure (1).

2. Nonparametric Estimation Method

Nonparametric method is often very easy and simple to understand as compared to parametric method .Furthermore, nonparametric analyses are more widely used in situation, where there is doubt about the exact form of distribution [13]. In this research we use some nonparametric methods like Empirical Survival Function (EM), Borkowf –type (BE), Nelson

-type (NE), and Thompson– type(TB) estimators for estimating the survival function for the patient of lung cancer based on complete censored data .

2.1 Empirical survival Function Estimation Method(EM)

Let F (t) denotes the life distribution for a certain type of items. We want to estimate the distribution function F (t) and the survivor function S (t) = 1 - F (t) from a complete data set of **n** independent lifetimes. Let $t_{(1)} \leq t_{(2)} \leq . . . \leq t_{(n)}$ be the data set arranged in ascending order. The empirical distribution function is defined as [6],[14]

$$F(t) = \frac{(\text{Number of life time} \leq t)}{n} \quad (2)$$

If we assume that there are no ties in the data set, the empirical distribution function may be written

$$F(t) = \begin{cases} 0 & \text{for} \quad t < t_{(1)} \\ \frac{i}{n} & \text{for } t_i \leq t \leq t_{(i+1)} \quad i = 1, \dots\dots\dots\dots, n \\ 1 & \text{for } t_n \leq t \end{cases} \quad (3)$$

The corresponding empirical survivor function is

$$S(t) = 1 - F(t) = \frac{\text{Number oflife time} > t}{n} \quad (4)$$

If there are no ties in the data set, the Empirical survivor function may also be written

$$\hat{S}(t)_{EM} = \begin{cases} 1 & \text{for} \quad t < t_{(1)} \\ 1 - \frac{i}{n} & \text{for } t_i \leq t \leq t_{(i+1)} \quad i = 1, \dots\dots, n \\ 0 & \text{for } t_n \leq t \end{cases} \quad (5)$$

The variance of Empirical survivor function is [14]

$$\text{Var}(\hat{S}(t))_{EM} = \frac{\hat{S}(t)_{EM}(1 - \hat{S}(t)_{EM})}{n} \quad (6)$$

If all observations are distinct, S(t) is a step function that decreases by l/n just before each observed failure time[6]. A simple adjustment accommodates any ties present in the data. S(t) as a function of t is illustrated, so we have

$$\hat{S}(t)_{EM} = 1 - \frac{i}{n} \qquad i = 1, \dots\dots\dots\dots, n \quad (7)$$

2.2 Borkowf –Type Estimator Method (BE)

Borkowf proposed a survival function under the framework of the Empirical survival function[3] .The Borkowf Empirical survival function having *n* number of cases in the study which is defined by the expression

$$\hat{S}(t)_{BE} = \frac{(n-1)\hat{S}(t)_{EM}}{n} + \frac{1}{2n} \quad (8)$$

Borkowf proved that the Greenwood's variance of the proposed estimator $\hat{S}(t)_{BE}$ less than the Greenwood's variance of Empirical survival function ($\hat{S}$ (t)$_{EM}$). The standard error of $\hat{S}(t)$ is usually found from Greenwood's formula. The variance of Borkowf proposed a survivor function is[3],[14]

$$\mathrm{Var}(\hat{S}(t)_{BE}) = \left(\frac{n-1}{n}\right)^2 \frac{\hat{S}(t)_{EM}{}^2 (1-\hat{S}(t)_{EM})}{n} \qquad (9)$$

2.3: Thompson– Type Estimator Method (TE)

The shrinkage estimation method is the Bayesian approach depending on prior information regarding the value of the specific parameter θ from past experiences or previous studies.

In this section we have to estimate S(t) when a prior information about S(t) available as initial value $S_0(t)$.

Thus, Thompson- type shrinkage estimator have the following form [15]

$$\tilde{S}(t)_{TE} = \xi\, \hat{S}(t)_{EM} + (1-\xi) S_\circ(t), \quad 0 \le \xi \le 1 \qquad (10)$$

Where ξ is a shrinkage factor, $0 < \xi < 1$. Here, $S_\circ(t)$ is selected based on Wald test statistic for $H_0 : S(t) = S_0(t)$, against H_A: $S(t) \neq S_0(t)$ with Level of significance equal to$\alpha = 0.05$.

Where test statistic is

$$Z = \frac{\hat{S}_{EM} - S_0}{[\mathrm{var}\,\hat{S}_{EM}]^{1/2}}$$

In this paper, we put forward the shrinkage weight function ξ as Exp(-10/n) ∙

2.4:Nelson-Aalen Estimator Method (NE)

The Nelson-Aalen Estimator[11], an alternative estimate of the survival function which is based on the individual event times and of cumulative hazard rate H(t) at time t as below:-

$$\hat{H}(t) = \sum_{t_i \le t} \frac{d_i}{n} \quad \text{for } t > 0 \qquad (11)$$

Suppose that there are n individuals with observed survival times $t_1, t_2, \ldots, t_n$. The ordered death times $t_{(i)}$,i=1,2,...,n. Where d_i is the number of individuals who die at time $t_{(i)}$.

$$H(t) = \int_0^t h(x)dx = -\mathrm{Ln}S(t)$$

Where h(t) refers to hazard rate at time t.

Thus ,we can write the survival Nelson estimation as following

12)( ) $\hat{S}(t)_{NL}$=Exp (-$\hat{H}$(t)

3. Estimation of Survival Function Methods

The results of the estimation for the Survival Function using four mentioned methods under complete data using the MATLAB (2012a) program [8,14] are shown in Table(1) .

4.Numerical Result and Conclusions

1. As an expected the values of survival function of all estimation methods which are proposed in this paper has been decreasing gradually with increasing failure times for lung cancer patients , that is means there is an opposite relationship between failure times and survival function. This shows that the value of survival function for patients was high when the patients were alive in the hospital and became low otherwise [14].
2. The mean squares error [14], for proposed estimation methods of the survival function are given in table (2).

Where;

$$\mathrm{MSE}\left[\hat{S}(t_i)\right] = \frac{\sum_{i=0}^{n}\left[\hat{S}(t_i) - S(t_i)\right]^2}{n} \qquad (13)$$

Where $S(t_i)$ is the Median rank survival function, $\hat{S}(t_i)$ is the specific estimated survival function and n refer to the sample size of the patient .

3. As a consequence, the computations of mentioned statistical indicators which are shown in table (2) above, leads to the result that the mean squares error(MSE) for Thompson estimator (TE) method are less than those of the EM, BE and NE methods, so the shrinkage Method is the best estimation method.
4. By observing figure (4) below, one can note the matching of the proposed estimation methods in this paper and the extent of convergence resulting accuracy of these methods, especially to real Median rank survival function methods S (t) .See figure (2).

References

1. AL- Qurashi , A . K. (2001). Estimate survival function for nonparametric methods Ph.D. thesis in statistics submitted to the Faculty of Management and Economics at University of Mustansiriya.

2. American Cancer Society (December 2007). "Report sees 7.6 million global 2007 cancer deaths". Reuters. Retrieved 2008-08-07.

3. Borkowf, C. B. (2005). A simple hybrid variance estimator for the Kaplan-Meier survival function. *Statistics in Medicine*, . 24; 827-851.

4. Basher, F. M.(2010).Some Of The Parametric Methods And Nonparametric to Estimate the reliability function With The Practical Application. Master thesis, Baghdad University, College of Administration and Economics

5. Haifa, K. (1987).Use Non-Parametric Method to Find Reliability Function For the Tools of 14 Ramadan factory - Department of textile . Master Thesis, Baghdad University, College of Administration and Economics.

6.Kaplan, E.L. and Meier, P. (1958). Non Parametric Estimation from Incomplete Observations .*Journa1 of the American Statistical Association*. 53. 457-481.

7. Marvin. R and Arnljot .H. (2004) . System Reliability Theory Models, Statistical Methods and Applications . Second Edition.

8.Mathews J. H. And Fink K. D. (2003)," Numerical Method Using MATLAB", Third Edition, Prentice Hall, USA.

9. Mei-C,W. (2006) .Summary Notes for Survival Analysis. Department of Biostatistics. Johns Hopkins University.

10. National Collaborating Centre for Cancer (2011), " The diagnosis and treatment of lung cancer (update) ".
http://www.nice.org.uk/nicemedia/live/13465/54199/54199.pdf

11. Nelson, W.(1972).Theory and application of hazard plotting for censored failure Data ." *Techno metrics* 14:945-966.

12. Petrson , A.V.(1977) . Expressing the Kaplan-Meier estimator as a function of empirical sub-survival function . *JASA*.72,854-858.

13.Qamruz, Z. and Karl, P. (2011) ,Survival Analysis Medical Research.
http://interstat.statjournals.net/YEAR/2011/abstracts/1105005.php.

14. Taha , A. T (2013)," Estimate the Parameters and Related Probability Functions for Data of the Patients of Lymph Glands Cancer via Birnbaum- Saunders Model", M.Sc., Baghdad University, Education College for Pure Sciences(Ibn Al-Haitham) .

15. Thompson , J.R. (1968) . Some Shrinkage Techniques for Estimating the Mean . *J. Amer. Statist. Assoc*.63..113-122

Table No. (1): Estimated Values for the Survival Function

No.	Time/d	$\hat{S}\_EM$	$\hat{S}\_BE$	$\hat{S}\_NT$	$\hat{S}\_NL$
1	3	0.9915	0.9874	0.9917	0.9916
2	37	0.9831	0.9790	0.9833	0.9832
3	72	0.9746	0.9706	0.9748	0.9749
4	75	0.9661	0.9622	0.9663	0.9667
5	91	0.9576	0.9537	0.9578	0.9585
6	100	0.9492	0.9453	0.9494	0.9504
7	103	0.9407	0.9369	0.9409	0.9424
8	121	0.9322	0.9285	0.9324	0.9345
9	127	0.9237	0.9201	0.9240	0.9266
10	140	0.9153	0.9117	0.9155	0.9187
11	154	0.9068	0.9033	0.9070	0.9110
12	156	0.8983	0.8949	0.8985	0.9033
13	164	0.8898	0.8865	0.8901	0.8957
14	186	0.8814	0.8781	0.8816	0.8881
15	211	0.8729	0.8697	0.8731	0.8806
16	212	0.8644	0.8613	0.8646	0.8732
17	213	0.8559	0.8529	0.8562	0.8658
18	217	0.8475	0.8445	0.8477	0.8585
19	218	0.8390	0.8361	0.8392	0.8513
20	221	0.8305	0.8277	0.8308	0.8441
21	221	0.8220	0.8193	0.8223	0.8370
22	233	0.8136	0.8109	0.8138	0.8299

23	240	0.8051	0.8025	0.8053	0.8229
24	241	0.7966	0.7941	0.7969	0.8160
25	243	0.7881	0.7857	0.7884	0.8091
26	249	0.7797	0.7773	0.7799	0.8022
27	254	0.7712	0.7689	0.7715	0.7955
28	266	0.7627	0.7605	0.7630	0.7888
29	273	0.7542	0.7521	0.7545	0.7821
30	276	0.7458	0.7437	0.7460	0.7755
31	277	0.7373	0.7353	0.7376	0.7690
32	278	0.7288	0.7269	0.7291	0.7625
33	281	0.7203	0.7185	0.7206	0.7560
34	290	0.7119	0.7101	0.7121	0.7497
35	301	0.7034	0.7017	0.7037	0.7433
36	301	0.6949	0.6933	0.6952	0.7371
37	301	0.6864	0.6849	0.6867	0.7308
38	302	0.6780	0.6765	0.6783	0.7247
39	304	0.6695	0.6681	0.6698	0.7186
40	304	0.6610	0.6597	0.6613	0.7125
41	306	0.6525	0.6512	0.6528	0.7065
42	307	0.6441	0.6428	0.6444	0.7005
43	307	0.6356	0.6344	0.6359	0.6946
44	308	0.6271	0.6260	0.6274	0.6887
45	313	0.6186	0.6176	0.6190	0.6829
46	313	0.6102	0.6092	0.6105	0.6772
47	314	0.6017	0.6008	0.6020	0.6715
48	318	0.5932	0.5924	0.5935	0.6658
49	330	0.5847	0.5840	0.5851	0.6602
50	331	0.5763	0.5756	0.5766	0.6546
51	332	0.5678	0.5672	0.5681	0.6491
52	332	0.5593	0.5588	0.5596	0.6436
53	334	0.5508	0.5504	0.5512	0.6382
54	334	0.5424	0.5420	0.5427	0.6328
55	335	0.5339	0.5336	0.5342	0.6274
56	335	0.5254	0.5252	0.5258	0.6221
57	335	0.5169	0.5168	0.5173	0.6169
58	335	0.5085	0.5084	0.5088	0.6117
59	338	0.5000	0.5000	0.5003	0.6065
60	341	0.4915	0.4916	0.4919	0.6014
61	342	0.4831	0.4832	0.4834	0.5963
62	345	0.4746	0.4748	0.4749	0.5913
63	349	0.4661	0.4664	0.4665	0.5863
64	354	0.4576	0.4580	0.4580	0.5814
65	357	0.4492	0.4496	0.4495	0.5765
66	363	0.4407	0.4412	0.4410	0.5716
67	364	0.4322	0.4328	0.4326	0.5668
68	364	0.4237	0.4244	0.4241	0.5620
69	366	0.4153	0.4160	0.4156	0.5572
70	367	0.4068	0.4076	0.4071	0.5525
71	368	0.3983	0.3992	0.3987	0.5479
72	371	0.3898	0.3908	0.3902	0.5433
73	373	0.3814	0.3824	0.3817	0.5387

74	374	0.3729	0.3740	0.3733	0.5341
75	380	0.3644	0.3656	0.3648	0.5296
76	387	0.3559	0.3572	0.3563	0.5252
77	392	0.3475	0.3488	0.3478	0.5207
78	393	0.3390	0.3403	0.3394	0.5163
79	397	0.3305	0.3319	0.3309	0.5120
80	399	0.3220	0.3235	0.3224	0.5076
81	400	0.3136	0.3151	0.3140	0.5034
82	400	0.3051	0.3067	0.3055	0.4991
83	401	0.2966	0.2983	0.2970	0.4949
84	402	0.2881	0.2899	0.2885	0.4907
85	407	0.2797	0.2815	0.2801	0.4866
86	409	0.2712	0.2731	0.2716	0.4825
87	419	0.2627	0.2647	0.2631	0.4784
88	421	0.2542	0.2563	0.2546	0.4744
89	421	0.2458	0.2479	0.2462	0.4704
90	422	0.2373	0.2395	0.2377	0.4664
91	422	0.2288	0.2311	0.2292	0.4625
92	423	0.2203	0.2227	0.2208	0.4586
93	427	0.2119	0.2143	0.2123	0.4547
94	428	0.2034	0.2059	0.2038	0.4509
95	430	0.1949	0.1975	0.1953	0.4471
96	446	0.1864	0.1891	0.1869	0.4433
97	450	0.1780	0.1807	0.1784	0.4395
98	454	0.1695	0.1723	0.1699	0.4358
99	461	0.1610	0.1639	0.1615	0.4321
100	463	0.1525	0.1555	0.1530	0.4285
101	470	0.1441	0.1471	0.1445	0.4249
102	477	0.1356	0.1387	0.1360	0.4213
103	481	0.1271	0.1303	0.1276	0.4177
104	481	0.1186	0.1219	0.1191	0.4142
105	483	0.1102	0.1135	0.1106	0.4107
106	483	0.1017	0.1051	0.1021	0.4073
107	497	0.0932	0.0967	0.0937	0.4038
108	511	0.0847	0.0883	0.0852	0.4004
109	512	0.0763	0.0799	0.0767	0.3970
110	512	0.0678	0.0715	0.0683	0.3937
111	516	0.0593	0.0631	0.0598	0.3904
112	517	0.0508	0.0547	0.0513	0.3871
113	519	0.0424	0.0463	0.0428	0.3838
114	533	0.0339	0.0378	0.0344	0.3806
115	534	0.0254	0.0294	0.0259	0.3774
116	535	0.0169	0.0210	0.0174	0.3742
117	540	0.0085	0.0126	0.0090	0.3710
118	550	0	0.0042	0.0005	0.3679

Table No. (2): Comparing between four Non-parametric Methods

Methods	**MSE[$\tilde{S}(t_i)$]**
EM	0.000018
NE	0.0292
BK	0.000019
TH	0.000015

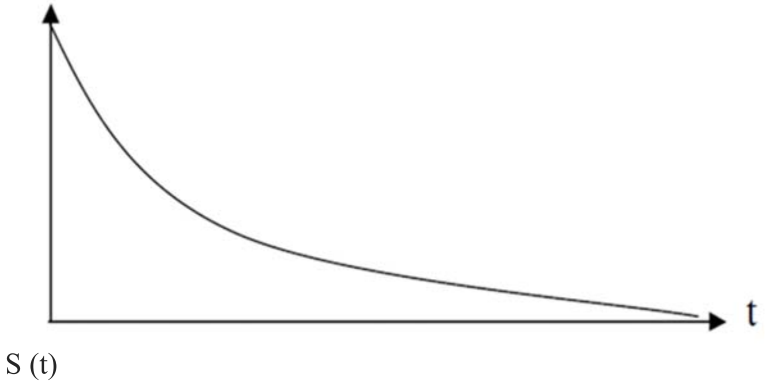

Figure No.(1): Shows the curve of the survival function

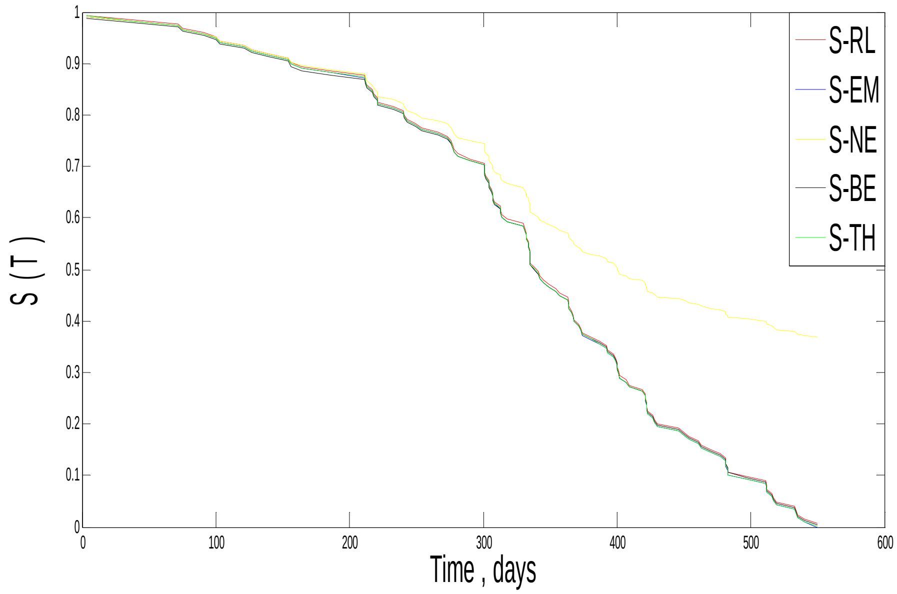

Figure No.(2): Shows the curve of four used estimation methods for the survival function

تقدير دالة البقاء لبيانات حقيقية كاملة لمرضى سرطان الرئة

عباس نجم سلمان
ابتهال حسين فرحان
قسم الرياضيات / كلية التربية للعلوم الصرفة (ابن الهيثم) / جامعة بغداد

استلم في :8 حزيران 2014 ، قبل في :1 ايلول 2014

الخلاصة

في هذا البحث قدرت دالة البقاء لمرضى سرطان الرئة باستخدام طرائق تقدير لامعلمية مختلفة اعتمادا على بيانات حقيقية كاملة التي تصف مدة البقاء للمرضى الذين يعانون من سرطان الرئة والمعتمد على تشخيص المرض، أو دخول المرضى في المستشفى مدة سنتين (تبدأ من بداية عام 2012 إلى نهاية عام 2013). وقد أجريت مقارنات بين طرائق التقدير المقترحة باستخدام مؤشر إحصائي متوسط مربعات الخطأ، وخلصت الدراسة إلى أن تقدير دالة البقاء على قيد الحياة لمرضى سرطان الرئة باستخدام طريقة التقلص هي الفضلى.

الكلمات المفتاحية : مقدرات غير معلمية, مرض سرطان الرئة، البيانات الحقيقية الكاملة، المقدر التجريبي,قدر بوركوف, مقدر نلسون، مقدر التقلص ومتوسط مربعات الخطأ.