Directed Independent Study on Survival Analysis and Multi State Models

by

G M Nazmus Salehin

Submitted to

Dr.Fei Heng

Department of Mathematical Science, University of North Florida

Contents

1 Introduction

The objective in survival analysis is to establish a connection between covariates and the time of an event. In this project we are going to perform different survival analysis techniques on publicly accessible dataset. For the 1st part we are mainly interested in using Kaplan- Meier plots to visualize survival curves and performing Cox proportional hazards regression to describe the effects of each variable. The 2nd part will be about multi-state models,Before the analysis, we will 1st discuss some basic concepts of Survival analysis:

2 Basics of Survival Analysis

- **Survival Function**: The survival function is a function that gives the probability that a patient, device or any other object of interest will survive beyond any specified time. Let T be a continuous random variable with cumulative distribution function $F(t)$ on the interval $[0, \text{inf})$. Its survival function or reliability function is:

$$S(t) = P(\{T > t\}) = \int_0^{\text{inf}} f(u)du = 1 - F(t)$$

- **Hazard Function**: Along with the survival function, we are also interested in the rate at which event is taking place, out of the surviving population at any given time t. In medical terms, we can define it as "out of the people who survived at time t, what is the rate of dying of those people". It is defined as:

$$\lambda(t) = \lim_{t \to \text{inf}} \frac{Prt \leq T < t + dt | T \geq t}{dt}$$

  Survival function can be derived from hazard function and vice versa, with the following equation:

$$h(t) = f(t)/S(t)$$

- **Kaplan Meier Estimator**: The Kaplan-Meier (KM) method is a non-parametric method used to estimate the survival probability from observed survival times The survival probability at time $t_i, S(t_i)$, is calculated as follow:

$$S(t_i) = S(t_{i-1}(1 - \frac{d_i}{n_i}))$$

  Where

- $S(t_i - 1)$=The probability of being alive at $t_i - 1$
- n_i=the number of patients alive just before t_i
- d_i=the number of events at t_i
- $t_0 = 0, S(0) = 1$
- **Cumulative Hazard Function**:The cumulative hazard function is the integral of the hazard function. It can be interpreted as the probability of failure at time x given survival until time x:

$$H(x) = \int_{-\inf}^{x} h(t)dt$$

The North Central Cancer Treatment Group (NCCTG) data set records the survival of patients with advanced lung cancer, together with assessments of the patients performance status measured either by the physician and by the patients themselves.The goal of the study was to determine whether patients self-assessment could provide prognostic information complementary to the physician's assessment. We will be using this dataset for our survival analysis. 1st we load the dataset. Here's how the dataset looks like: .

Index	inst	time	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
0	3	306	2	74	1	1	90	100	1175	nan
1	3	455	2	68	1	0	90	90	1225	15
2	3	1010	1	56	1	0	90	90	nan	15
3	5	210	2	57	1	1	90	60	1150	11
4	1	883	2	60	1	0	100	90	nan	0
5	12	1022	1	74	1	1	50	80	513	0
6	7	310	2	68	2	2	70	60	384	10
7	11	361	2	71	2	2	60	80	538	1
8	1	218	2	53	1	1	70	80	825	16
9	7	166	2	61	1	2	70	70	271	34
10	6	170	2	57	1	1	80	80	1025	27

Figure 1: Dataset for lungs cancer

inst: Institution code
time: Survival time in days
status: censoring status 1=censored, 2=dead
age: Age in years
sex: Male=1 Female=2

ph.ecog: ECOG performance score (0=good 5=dead)
ph.karno: Karnofsky performance score (bad=0-good=100) rated by physician
pat.karno: Karnofsky performance score as rated by patient meal.cal: Calories consumed at meals
wt.loss: Weight loss in last six months

1st we will fit Kaplan-Meier Estimator on the full data. .

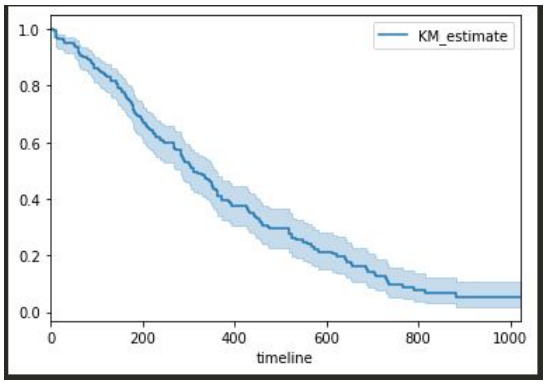


Figure 2: Survival Curve for total population

The plot gives us the survival probability at any given time for the population. Here are the result of fitting KM estimator to our dataset(Sex=1 is male, Sex=2 is female):

time	n.risk	n.event	n.censor	surv	std.err	upper	lower	strata	sex
11	138	3	0	0.9782609	0.0126898	1.0000000	0.9542301	sex=1	1
12	135	1	0	0.9710145	0.0147075	0.9994124	0.9434235	sex=1	1
13	134	2	0	0.9565217	0.0181489	0.9911586	0.9230952	sex=1	1
15	132	1	0	0.9492754	0.0196777	0.9866017	0.9133612	sex=1	1
26	131	1	0	0.9420290	0.0211171	0.9818365	0.9038355	sex=1	1
30	130	1	0	0.9347826	0.0224847	0.9768989	0.8944820	sex=1	1
31	129	1	0	0.9275362	0.0237933	0.9718155	0.8852745	sex=1	1
53	128	2	0	0.9130435	0.0262703	0.9612864	0.8672216	sex=1	1
54	126	1	0	0.9057971	0.0274522	0.9558688	0.8583483	sex=1	1
59	125	1	0	0.8985507	0.0286031	0.9503632	0.8495630	sex=1	1
60	124	1	0	0.8913043	0.0297272	0.9447781	0.8408571	sex=1	1
65	123	2	0	0.8768116	0.0319075	0.9333961	0.8236574	sex=1	1
71	121	1	0	0.8695652	0.0329690	0.9276100	0.8151525	sex=1	1
81	120	1	0	0.8623188	0.0340145	0.9217668	0.8067049	sex=1	1

88	119	2	0	0.8478261	0.0360643	0.9099232	0.7899667	sex=1	1
92	117	1	0	0.8405797	0.0370717	0.9039292	0.7816699	sex=1	1
93	116	1	0	0.8333333	0.0380693	0.8978906	0.7734176	sex=1	1
95	115	1	0	0.8260870	0.0390583	0.8918099	0.7652076	sex=1	1
105	114	1	0	0.8188406	0.0400397	0.8856890	0.7570376	sex=1	1
107	113	1	0	0.8115942	0.0410146	0.8795299	0.7489059	sex=1	1
110	112	1	0	0.8043478	0.0419837	0.8733343	0.7408108	sex=1	1
116	111	1	0	0.7971014	0.0429480	0.8671037	0.7327506	sex=1	1
118	110	1	0	0.7898551	0.0439083	0.8608395	0.7247240	sex=1	1
131	109	1	0	0.7826087	0.0448652	0.8545431	0.7167297	sex=1	1
132	108	2	0	0.7681159	0.0467716	0.8418580	0.7008332	sex=1	1
135	106	1	0	0.7608696	0.0477225	0.8354715	0.6929291	sex=1	1
142	105	1	0	0.7536232	0.0486725	0.8290569	0.6850530	sex=1	1
144	104	1	0	0.7463768	0.0496222	0.8226150	0.6772042	sex=1	1
147	103	1	0	0.7391304	0.0505722	0.8161466	0.6693819	sex=1	1
156	102	2	0	0.7246377	0.0524750	0.8031334	0.6538139	sex=1	1
163	100	3	0	0.7028986	0.0553435	0.7834316	0.6306439	sex=1	1
166	97	1	0	0.6956522	0.0563053	0.7768181	0.6229669	sex=1	1
170	96	1	0	0.6884058	0.0572708	0.7701822	0.6153123	sex=1	1
174	95	0	1	0.6884058	0.0572708	0.7701822	0.6153123	sex=1	1
175	94	1	0	0.6810823	0.0582609	0.7634689	0.6075862	sex=1	1
176	93	1	0	0.6737589	0.0592554	0.7567332	0.5998825	sex=1	1
177	92	1	0	0.6664354	0.0602549	0.7499756	0.5922008	sex=1	1
179	91	2	0	0.6517885	0.0622703	0.7363958	0.5769020	sex=1	1
180	89	1	0	0.6444650	0.0632873	0.7295743	0.5692842	sex=1	1
181	88	2	0	0.6298181	0.0653418	0.7158700	0.5541101	sex=1	1
183	86	1	0	0.6224946	0.0663803	0.7089876	0.5465533	sex=1	1
185	85	0	1	0.6224946	0.0663803	0.7089876	0.5465533	sex=1	1
188	84	0	1	0.6224946	0.0663803	0.7089876	0.5465533	sex=1	1
189	83	1	0	0.6149947	0.0674780	0.7019541	0.5388080	sex=1	1
191	82	0	1	0.6149947	0.0674780	0.7019541	0.5388080	sex=1	1
196	81	0	1	0.6149947	0.0674780	0.7019541	0.5388080	sex=1	1
197	80	1	1	0.6073072	0.0686404	0.6947607	0.5308620	sex=1	1
202	78	1	0	0.5995212	0.0698427	0.6874717	0.5228226	sex=1	1
207	77	1	0	0.5917353	0.0710555	0.6801583	0.5148075	sex=1	1
210	76	1	0	0.5839493	0.0722795	0.6728210	0.5068165	sex=1	1
212	75	1	0	0.5761633	0.0735153	0.6654600	0.4988491	sex=1	1
218	74	1	0	0.5683773	0.0747638	0.6580755	0.4909052	sex=1	1

221	73	0	1	0.5683773	0.0747638	0.6580755	0.4909052	sex=1	1
222	72	1	1	0.5604832	0.0760608	0.6505873	0.4828581	sex=1	1
223	70	1	0	0.5524763	0.0774098	0.6429911	0.4747033	sex=1	1
225	69	0	2	0.5524763	0.0774098	0.6429911	0.4747033	sex=1	1
229	67	1	0	0.5442303	0.0788570	0.6351933	0.4662937	sex=1	1
230	66	1	0	0.5359844	0.0803214	0.6273672	0.4579125	sex=1	1
237	65	0	1	0.5359844	0.0803214	0.6273672	0.4579125	sex=1	1
239	64	1	0	0.5276097	0.0818507	0.6194185	0.4494085	sex=1	1
246	63	1	0	0.5192349	0.0834000	0.6114403	0.4409341	sex=1	1
259	62	0	1	0.5192349	0.0834000	0.6114403	0.4409341	sex=1	1
267	61	1	0	0.5107229	0.0850222	0.6033320	0.4323289	sex=1	1
269	60	1	0	0.5022108	0.0866675	0.5951927	0.4237547	sex=1	1
270	59	1	0	0.4936988	0.0883374	0.5870228	0.4152113	sex=1	1
279	58	0	1	0.4936988	0.0883374	0.5870228	0.4152113	sex=1	1
283	57	1	0	0.4850374	0.0900931	0.5787122	0.4065255	sex=1	1
284	56	1	1	0.4763760	0.0918774	0.5703691	0.3978723	sex=1	1
285	54	1	0	0.4675542	0.0937596	0.5618757	0.3890664	sex=1	1
286	53	1	0	0.4587324	0.0956750	0.5533477	0.3802951	sex=1	1
288	52	1	0	0.4499107	0.0976257	0.5447854	0.3715585	sex=1	1
291	51	1	0	0.4410889	0.0996139	0.5361887	0.3628563	sex=1	1
292	50	0	1	0.4410889	0.0996139	0.5361887	0.3628563	sex=1	1
300	49	0	1	0.4410889	0.0996139	0.5361887	0.3628563	sex=1	1
301	48	1	1	0.4318995	0.1018145	0.5272874	0.3537676	sex=1	1
303	46	1	1	0.4225104	0.1041599	0.5182013	0.3444898	sex=1	1
306	44	1	0	0.4129079	0.1066669	0.5089185	0.3350103	sex=1	1
310	43	1	0	0.4033054	0.1092316	0.4995881	0.3255787	sex=1	1
320	42	1	0	0.3937029	0.1118582	0.4902103	0.3161948	sex=1	1
329	41	1	0	0.3841004	0.1145513	0.4807851	0.3068587	sex=1	1
337	40	1	0	0.3744979	0.1173160	0.4713124	0.2975705	sex=1	1
353	39	2	0	0.3552929	0.1230814	0.4522240	0.2791383	sex=1	1
363	37	1	0	0.3456903	0.1260944	0.4426077	0.2699949	sex=1	1
364	36	1	0	0.3360878	0.1292031	0.4329429	0.2609005	sex=1	1
371	35	1	0	0.3264853	0.1324152	0.4232292	0.2518556	sex=1	1
387	34	1	0	0.3168828	0.1357389	0.4134660	0.2428609	sex=1	1
390	33	1	0	0.3072803	0.1391834	0.4036527	0.2339169	sex=1	1
394	32	1	0	0.2976778	0.1427588	0.3937884	0.2250246	sex=1	1
404	31	0	1	0.2976778	0.1427588	0.3937884	0.2250246	sex=1	1
413	30	0	1	0.2976778	0.1427588	0.3937884	0.2250246	sex=1	1

428	29	1	0	0.2874130	0.1470089	0.3833899	0.2154628	sex=1	1
429	28	1	0	0.2771483	0.1514409	0.3729228	0.2059707	sex=1	1
442	27	1	0	0.2668835	0.1560732	0.3623861	0.1965495	sex=1	1
444	26	0	1	0.2668835	0.1560732	0.3623861	0.1965495	sex=1	1
455	25	1	0	0.2562082	0.1613243	0.3514896	0.1867556	sex=1	1
457	24	1	0	0.2455329	0.1668446	0.3405085	0.1770481	sex=1	1
458	23	0	1	0.2455329	0.1668446	0.3405085	0.1770481	sex=1	1
460	22	1	0	0.2343723	0.1732098	0.3291112	0.1669052	sex=1	1
477	21	1	0	0.2232117	0.1799516	0.3176084	0.1568707	sex=1	1
519	20	1	0	0.2120511	0.1871207	0.3059976	0.1469478	sex=1	1
524	19	1	0	0.2008905	0.1947771	0.2942754	0.1371402	sex=1	1
533	18	1	0	0.1897299	0.2029929	0.2824384	0.1274524	sex=1	1
558	17	1	0	0.1785694	0.2118551	0.2704819	0.1178896	sex=1	1
567	16	1	0	0.1674088	0.2214706	0.2584011	0.1084581	sex=1	1
574	15	1	0	0.1562482	0.2319723	0.2461899	0.0991653	sex=1	1
583	14	1	0	0.1450876	0.2435275	0.2338413	0.0900201	sex=1	1
613	13	1	0	0.1339270	0.2563511	0.2213475	0.0810330	sex=1	1
624	12	1	0	0.1227664	0.2707243	0.2086991	0.0722169	sex=1	1
643	11	1	0	0.1116058	0.2870236	0.1958853	0.0635875	sex=1	1
655	10	1	0	0.1004453	0.3057674	0.1828939	0.0551645	sex=1	1
689	9	1	0	0.0892847	0.3276928	0.1697108	0.0469726	sex=1	1
707	8	1	0	0.0781241	0.3538922	0.1563215	0.0390437	sex=1	1
791	7	1	0	0.0669635	0.3860690	0.1427121	0.0314207	sex=1	1
806	6	0	1	0.0669635	0.3860690	0.1427121	0.0314207	sex=1	1
814	5	1	0	0.0535708	0.4461493	0.1284373	0.0223442	sex=1	1
840	4	0	1	0.0535708	0.4461493	0.1284373	0.0223442	sex=1	1
883	3	1	0	0.0357139	0.6047445	0.1168413	0.0109164	sex=1	1
1010	2	0	1	0.0357139	0.6047445	0.1168413	0.0109164	sex=1	1
1022	1	0	1	0.0357139	0.6047445	0.1168413	0.0109164	sex=1	1
5	90	1	0	0.9888889	0.0111734	1.0000000	0.9674682	sex=2	2
60	89	1	0	0.9777778	0.0158910	1.0000000	0.9477934	sex=2	2
61	88	1	0	0.9666667	0.0195740	1.0000000	0.9302835	sex=2	2
62	87	1	0	0.9555556	0.0227331	0.9990942	0.9139143	sex=2	2
79	86	1	0	0.9444444	0.0255655	0.9929738	0.8982868	sex=2	2
81	85	1	0	0.9333333	0.0281718	0.9863173	0.8831956	sex=2	2
92	84	0	1	0.9333333	0.0281718	0.9863173	0.8831956	sex=2	2
95	83	1	0	0.9220884	0.0306689	0.9792147	0.8682947	sex=2	2
105	82	0	1	0.9220884	0.0306689	0.9792147	0.8682947	sex=2	2

107	81	1	0	0.9107045	0.0330893	0.9717245	0.8535164	sex=2	2
122	80	1	0	0.8993207	0.0353996	0.9639328	0.8390396	sex=2	2
145	79	2	0	0.8765531	0.0397733	0.9476180	0.8108176	sex=2	2
153	77	1	0	0.8651693	0.0418664	0.9391563	0.7970111	sex=2	2
166	76	1	0	0.8537855	0.0439117	0.9305216	0.7833775	sex=2	2
167	75	1	0	0.8424017	0.0459175	0.9217311	0.7698998	sex=2	2
173	74	0	1	0.8424017	0.0459175	0.9217311	0.7698998	sex=2	2
175	73	0	1	0.8424017	0.0459175	0.9217311	0.7698998	sex=2	2
177	72	0	1	0.8424017	0.0459175	0.9217311	0.7698998	sex=2	2
182	71	1	0	0.8305369	0.0480585	0.9125705	0.7558775	sex=2	2
186	70	1	0	0.8186721	0.0501663	0.9032576	0.7420075	sex=2	2
192	69	0	1	0.8186721	0.0501663	0.9032576	0.7420075	sex=2	2
194	68	1	0	0.8066328	0.0523082	0.8937184	0.7280329	sex=2	2
199	67	1	0	0.7945935	0.0544270	0.8840428	0.7141948	sex=2	2
201	66	2	0	0.7705149	0.0586155	0.8643202	0.6868904	sex=2	2
202	64	0	1	0.7705149	0.0586155	0.8643202	0.6868904	sex=2	2
203	63	0	1	0.7705149	0.0586155	0.8643202	0.6868904	sex=2	2
208	62	1	0	0.7580872	0.0608292	0.8540771	0.6728857	sex=2	2
211	61	0	1	0.7580872	0.0608292	0.8540771	0.6728857	sex=2	2
224	60	0	1	0.7580872	0.0608292	0.8540771	0.6728857	sex=2	2
226	59	1	0	0.7452383	0.0631856	0.8434878	0.6584329	sex=2	2
235	58	0	1	0.7452383	0.0631856	0.8434878	0.6584329	sex=2	2
239	57	1	0	0.7321639	0.0656178	0.8326497	0.6438050	sex=2	2
240	56	0	1	0.7321639	0.0656178	0.8326497	0.6438050	sex=2	2
243	55	0	1	0.7321639	0.0656178	0.8326497	0.6438050	sex=2	2
245	54	1	0	0.7186054	0.0682283	0.8214223	0.6286579	sex=2	2
252	53	0	1	0.7186054	0.0682283	0.8214223	0.6286579	sex=2	2
266	52	0	1	0.7186054	0.0682283	0.8214223	0.6286579	sex=2	2
268	51	1	0	0.7045151	0.0710441	0.8097727	0.6129393	sex=2	2
269	50	0	1	0.7045151	0.0710441	0.8097727	0.6129393	sex=2	2
272	49	0	1	0.7045151	0.0710441	0.8097727	0.6129393	sex=2	2
276	48	0	1	0.7045151	0.0710441	0.8097727	0.6129393	sex=2	2
285	47	1	0	0.6895254	0.0742280	0.7975047	0.5961661	sex=2	2
292	46	0	1	0.6895254	0.0742280	0.7975047	0.5961661	sex=2	2
293	45	1	0	0.6742026	0.0775554	0.7848845	0.5791287	sex=2	2
296	44	0	1	0.6742026	0.0775554	0.7848845	0.5791287	sex=2	2
305	43	1	0	0.6585235	0.0810466	0.7718951	0.5618032	sex=2	2
310	42	1	0	0.6428443	0.0845534	0.7587135	0.5446705	sex=2	2

315	41	0	1	0.6428443	0.0845534	0.7587135	0.5446705	sex=2	2
332	40	0	1	0.6428443	0.0845534	0.7587135	0.5446705	sex=2	2
340	39	1	0	0.6263611	0.0884536	0.7449320	0.5266632	sex=2	2
345	38	1	0	0.6098780	0.0923866	0.7309413	0.5088659	sex=2	2
348	37	1	0	0.5933948	0.0963640	0.7167520	0.4912680	sex=2	2
350	36	1	0	0.5769116	0.1003976	0.7023731	0.4738607	sex=2	2
351	35	1	0	0.5604284	0.1044989	0.6878120	0.4566364	sex=2	2
356	34	0	1	0.5604284	0.1044989	0.6878120	0.4566364	sex=2	2
361	33	1	0	0.5434457	0.1089357	0.6727944	0.4389650	sex=2	2
363	32	1	0	0.5264630	0.1134683	0.6575855	0.4214863	sex=2	2
364	31	0	1	0.5264630	0.1134683	0.6575855	0.4214863	sex=2	2
371	30	1	0	0.5089143	0.1184250	0.6418716	0.4034977	sex=2	2
376	29	0	1	0.5089143	0.1184250	0.6418716	0.4034977	sex=2	2
382	28	0	1	0.5089143	0.1184250	0.6418716	0.4034977	sex=2	2
384	27	0	1	0.5089143	0.1184250	0.6418716	0.4034977	sex=2	2
426	26	1	0	0.4893406	0.1247515	0.6248848	0.3831975	sex=2	2
433	25	1	0	0.4697670	0.1312616	0.6075927	0.3632055	sex=2	2
444	24	1	0	0.4501934	0.1379898	0.5900058	0.3435120	sex=2	2
450	23	1	0	0.4306198	0.1449741	0.5721318	0.3241095	sex=2	2
473	22	1	0	0.4110461	0.1522563	0.5539765	0.3049929	sex=2	2
511	21	0	2	0.4110461	0.1522563	0.5539765	0.3049929	sex=2	2
520	19	1	0	0.3894121	0.1615734	0.5344917	0.2837122	sex=2	2
524	18	1	0	0.3677781	0.1713883	0.5146024	0.2628451	sex=2	2
529	17	0	1	0.3677781	0.1713883	0.5146024	0.2628451	sex=2	2
543	16	0	1	0.3677781	0.1713883	0.5146024	0.2628451	sex=2	2
550	15	1	0	0.3432596	0.1847589	0.4930486	0.2389767	sex=2	2
551	14	0	1	0.3432596	0.1847589	0.4930486	0.2389767	sex=2	2
559	13	0	1	0.3432596	0.1847589	0.4930486	0.2389767	sex=2	2
588	12	0	1	0.3432596	0.1847589	0.4930486	0.2389767	sex=2	2
641	11	1	0	0.3120542	0.2079104	0.4690333	0.2076138	sex=2	2
654	10	1	0	0.2808487	0.2331048	0.4434980	0.1778498	sex=2	2
687	9	1	0	0.2496433	0.2612025	0.4165393	0.1496181	sex=2	2
705	8	1	0	0.2184379	0.2934006	0.3882139	0.1229094	sex=2	2
728	7	1	0	0.1872325	0.3315018	0.3585552	0.0977702	sex=2	2
731	6	1	0	0.1560271	0.3784531	0.3275969	0.0743122	sex=2	2
735	5	1	0	0.1248217	0.4395756	0.2954319	0.0527379	sex=2	2
740	4	0	1	0.1248217	0.4395756	0.2954319	0.0527379	sex=2	2
765	3	1	0	0.0832144	0.5999112	0.2696771	0.0256775	sex=2	2

821	2	0	1	0.0832144	0.5999112	0.2696771	0.0256775	sex=2	2
965	1	0	1	0.0832144	0.5999112	0.2696771	0.0256775	sex=2	2

The table shows time instants for each event, survival probability and 95% confidence interval. Next we are going to visualize survival curves for male and females separately.

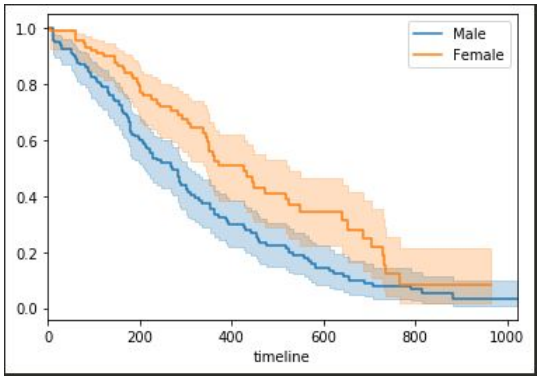


Figure 3: Separate Survival Curve for male and female

We can see from the separate survival curves that survival probability of female is greater than male. So gender may provide a role in the survival for lungs cancer patients. Here is a comparison of survival times for male and female:

```
        n events median 0.95LCL 0.95UCL
sex=1 138    112    270     212     310
sex=2  90     53    426     348     550
```

The median survival time for sex=1 (Male group) is 270 days, as opposed to 426 days for sex=2 (Female). There appears to be a survival advantage for female with lung cancer compare to male The following figure shows the 2 survival curves with the logrank test comparing the curves:

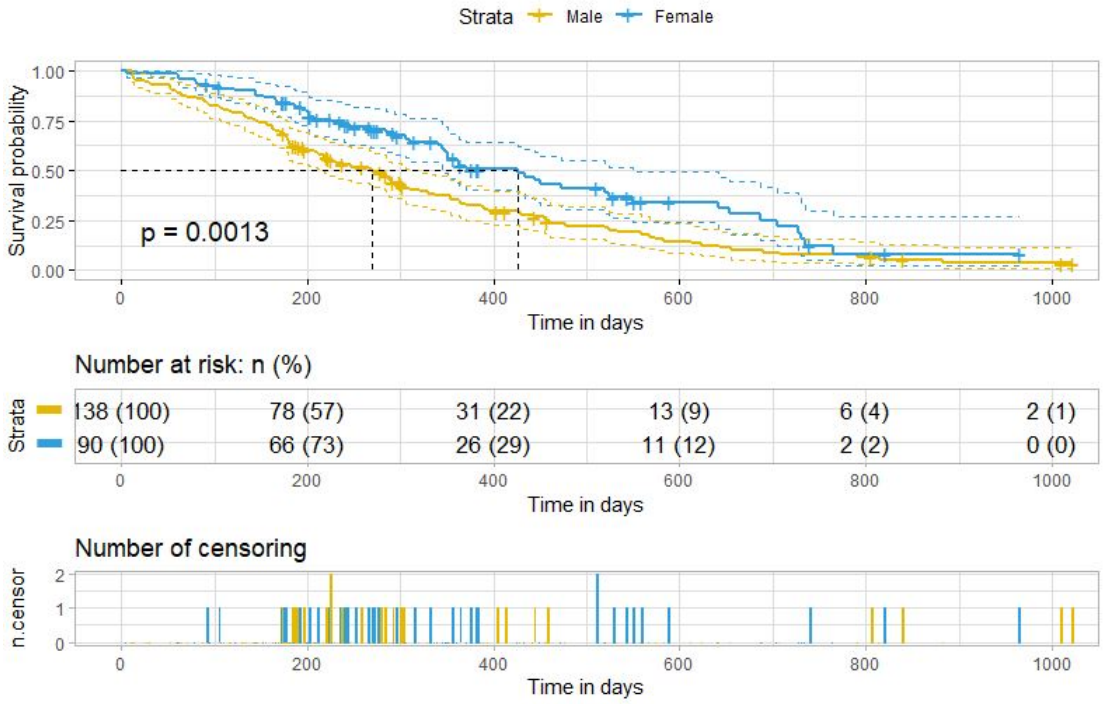


Figure 4: Comparison of Survival times for male vs female

To figure out if th . From the logrank test we get a very samll value of p (0.0013) which suggests the 2 survival curves are different.
Next we will plot the cumulative hazard function for male and female survival curves:

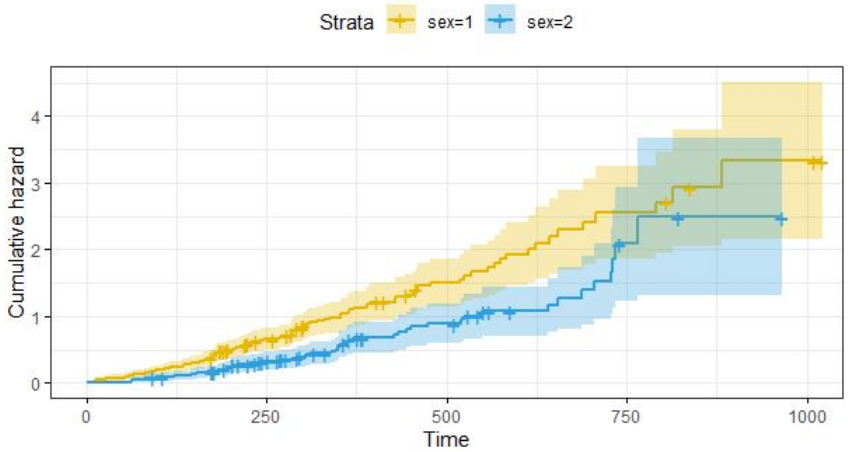
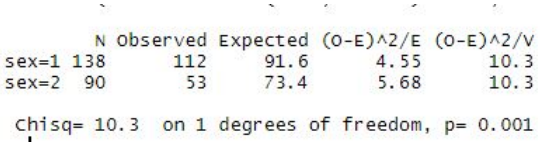


Figure 5: Comparison of Cumulative Hazard for male vs female

Comparing survival curves with Logrank test:The log-rank test is the most widely used method of comparing two or more survival curves. The null hypothesis is that there is no difference in survival between the two groups. The log rank test is a non-parametric test, which makes no assumptions about the survival distributions. Essentially, the log rank test compares the observed number of events in each group to what would be expected if the null hypothesis were true (i.e., if the survival curves were identical). The log rank statistic is approximately distributed as a chi-square test statistic. Here is the logrank test result for male vs female:



	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
sex=1	138	112	91.6	4.55	10.3
sex=2	90	53	73.4	5.68	10.3

Chisq= 10.3 on 1 degrees of freedom, p= 0.001

Figure 6: Logrank test results

The log rank test for difference in survival gives a p-value of p = 0.001, indicating that the sex groups differ significantly in survival.

Cox Proportional Hazard Model:The purpose of the model is to evaluate simultaneously the effect of several factors on survival. In other words, it allows us to examine how specified factors influence the rate of a particular event happening (e.g., infection, death) at a particular point in time. This rate is commonly referred as the hazard rate.

The Cox model is expressed by the hazard function denoted by h(t). Briefly, the hazard function can be interpreted as the risk of dying at time t. It can be estimated as follow:

$$h(t) = h_0(t) * exp(b_1x_1 + b_2x_2 + .. + b_px_p)$$

t represents the survival time
h(t) is the hazard function determined by a set of p covariates (x1,x2,...,xp)
the coefficients (b1,b2,...,bp) measure the impact (i.e., the effect size) of covariates.
the term h_0 is called the baseline hazard. It corresponds to the value of the hazard if all the xi are equal to zero (the quantity exp(0) equals 1). The 't' in h(t) reminds us that the hazard may vary over time. The Cox model can be written as a multiple linear regression of the logarithm of the hazard on the variables x_i, with the baseline hazard being an 'intercept' term that varies with time.

The quantities $exp(b_i)$ are called hazard ratios (HR). A value of b_i greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the ith covariate increases, the event hazard increases and thus the length of survival decreases.

In summary,

HR = 1: No effect HR < 1: Reduction in the hazard HR > 1: Increase in Hazard

1st we will focus on univariate cox regression model and we will consider only 1 variable (sex). Here are the results:

```
Call:
coxph(formula = Surv(time, status) ~ sex, data = lung)

       coef exp(coef) se(coef)     z       p
sex -0.5310    0.5880   0.1672 -3.176 0.00149

Likelihood ratio test=10.63  on 1 df, p=0.001111
n= 228, number of events= 165
```

The Cox regression results can be interpreted as follow:

Statistical significance The column marked "z" gives the Wald statistic value.From the output above, we can conclude that the variable sex have highly statistically significant coefficients.

The regression coefficients The second feature to note in the Cox model results is the the sign of the regression coefficients (coef). A positive sign means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable. The variable sex is encoded as a numeric vector. 1: male, 2: female. The R summary for the Cox model gives the hazard ratio (HR) for the second group relative to the first group, that is, female versus male. The beta coefficient for sex = -0.53 indicates that females have lower risk of death (lower survival rates) than males, in these data.

Hazard ratios The exponentiated coefficients (exp(coef) = exp(-0.53) = 0.59), also known as hazard ratios, give the effect size of covariates. For example, being female (sex=2) reduces the hazard by a factor of 0.59, or 41

Confidence intervals of the hazard ratios The summary output also gives upper and lower 95% confidence intervals for the hazard ratio (exp(coef)), lower 95% bound = 0.4237, upper 95% bound = 0.816.

Global statistical significance of the model Finally, the output gives p-values for three alternative tests for overall significance of the model: The likelihood-ratio test, Wald test, and score logrank statistics. These three methods are asymptotically equivalent. For large enough N, they will give similar results. For small N, they may differ somewhat. The Likelihood ratio test has better behavior for small sample sizes, so it is generally preferred. Next, the univariate cox model is applied to each variable at once. Here are the results:

```
           beta HR (95% CI for HR) wald.test p.value
age           0.019              1 (1-1)        4.1   0.042
sex           -0.53   0.59 (0.42-0.82)         10  0.0015
ph.karno     -0.016   0.98 (0.97-1)            7.9   0.005
ph.ecog        0.48    1.6 (1.3-2)             18 2.7e-05
wt.loss      0.0013      1 (0.99-1)         0.05    0.83
```

We can see that age, sex and ph.ecog are highly significant variables. Also age and ph.ecog has positive coefficients, meaning older age and higher ph.ecog coefficients are associated with poorer survival.

Next, we want to know how these variables jointly affect on survival. To answer to this question, we'll perform a multivariate Cox regression analysis. As the variable ph.karno is not significant in the univariate Cox analysis, we'll skip it in the multivariate analysis. We'll include the 3 factors (sex, age and ph.ecog) into the multivariate model.Here are the results:

```
  n= 227, number of events= 164
   (1 observation deleted due to missingness)

              coef exp(coef)  se(coef)      z Pr(>|z|)
age       0.011067  1.011128  0.009267  1.194 0.232416
sex      -0.552612  0.575445  0.167739 -3.294 0.000986 ***
ph.ecog   0.463728  1.589991  0.113577  4.083 4.45e-05 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .

          exp(coef) exp(-coef) lower .95 upper .95
age          1.0111     0.9890     0.9929     1.0297
sex          0.5754     1.7378     0.4142     0.7994
ph.ecog      1.5900     0.6289     1.2727     1.9864

Concordance= 0.637  (se = 0.025 )
```

```
Likelihood ratio test= 30.5  on 3 df,   p=1e-06
Wald test            = 29.93  on 3 df,   p=1e-06
Score (logrank) test = 30.5  on 3 df,   p=1e-06
```

, indicating that the model is significant. These tests evaluate the omnibus null hypothesis that all of the betas (β) are 0. In the above example, the test statistics are in close agreement, and the omnibus null hypothesis is soundly rejected.

In the multivariate Cox analysis, the covariates sex and ph.ecog remain significant (p<0.05). However, the covariate age fails to be significant (p = 0.23, which is grater than 0.05).

The p-value for sex is 0.000986, with a hazard ratio HR = exp(coef) = 0.58, indicating a strong relationship between the patients' sex and decreased risk of death. The hazard ratios of covariates are interpretable as multiplicative effects on the hazard. For example, holding the other covariates constant, being female (sex=2) reduces the hazard by a factor of 0.58, or 42%. We conclude that, being female is associated with good prognostic.

Similarly, the p-value for ph.ecog is $4.45e^{-}05$, with a hazard ratio HR = 1.59, indicating a strong relationship between the ph.ecog value and increased risk of death. Holding the other covariates constant, a higher value of ph.ecog is associated with a poor survival.

By contrast, the p-value for age is now p=0.23. The hazard ratio HR = exp(coef) = 1.01, with a 95% confidence interval of 0.99 to 1.03. Because the confidence interval for HR includes 1, these results indicate that age makes a smaller contribution to the difference in the HR after adjusting for the ph.ecog values and patient's sex, and only trend toward significance. For example, holding the other covariates constant, an additional year of age induce daily hazard of death by a factor of exp(beta) = 1.01, or 1%, which is not a significant contribution. Having fit a Cox model to the data, it's possible to visualize the predicted survival proportion at any given point in time for a particular risk group. The function survfit() estimates the survival proportion, by default at the mean values of covariates.

3 Multi State Models

The 2nd part of our project is concerned with multi-state models. Multi state models form a very broad class of models that includes standard survival models with an initial and final state, competing risks with multiple states and illness-death models, with an initial state and a death state. Typically, the disease or recovery process of a patient will also consist of intermediate events that can neither be classified as initial states nor final states. This model class is useful for representing movement through a discrete set of events.

The data that we will consider for analyzing multi-state models is provided by EBMT(European Group for Blood and Marrow Transplantation). The data set has 3 intermediate events: Recovery(Rec), an Adverse Event(AE) and a combination of 2 (AE and REC). It is to be expected that recovery improves the prognosis and an adverse event deteriorates it. 2279 patients were considered who were treated between 1985 and 1998. There are 4 prognostic factors. They are: donor-recipient match, prophylaxis, year of transplant and age at transplant in years. The following 6 states are considered:

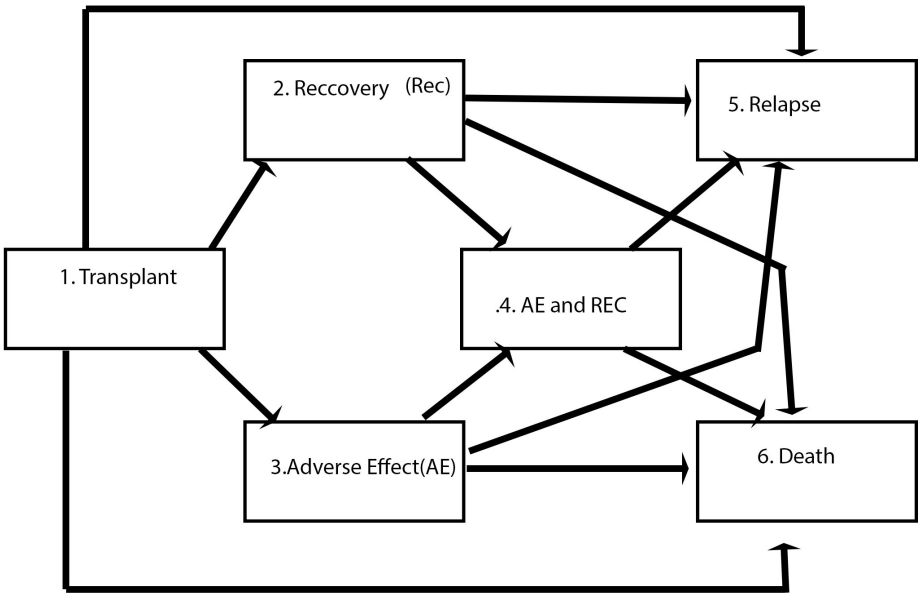


Figure 7: Multi state model for EBMT data

These 6 states are described here: 1. Alive and in remission, no recovery

or adverse event; 2. Alive in remission, recovered from the treatment; 3. Alive in remission, occurrence of the adverse event; 4. Alive, both recovered and adverse event occurred; 5. Alive, in relapse (treatment failure); 6. Dead (treatment failure).

We load the EBMT data set in R. The data are available in mstate in wide format. This means that each row in the data corresponds to a single subject. Here is the 1st 7 rows of our dataset:

id	rec	rec.s	ae	ae.s	recae	recae.s	rel	rel.s	srv	srv.s	year	agecl	proph	match
1	22	1	995.0	0	995	0	995	0	995	0	1995-1998	20-40	no	no gender mismatch
2	29	1	12.0	1	29	1	422	1	579	1	1995-1998	20-40	no	no gender mismatch
3	1264	0	27.0	1	1264	0	1264	0	1264	0	1995-1998	20-40	no	no gender mismatch
4	50	1	42.0	1	50	1	84	1	117	1	1995-1998	20-40	no	gender mismatch
5	22	1	1133.0	0	1133	0	114	1	1133	0	1995-1998	>40	no	gender mismatch
6	33	1	27.0	1	33	1	1427	0	1427	0	1995-1998	20-40	no	no gender mismatch
7	29	1	28.5	1	29	1	775	0	775	1	1995-1998	>40	no	no gender mismatch

Figure 8: EBMT dataset

The columns rec, ae, rel and srv are time variables,indicating the time measured in days post-transplant to recovery, AE, relapse and death respectively in case of an event, or last follow-up otherwise. The .s-variables are the corresponding status variables (1 for an event,0 for censoring). For instance, patient 1 had recovered after 22 days (transition from state 1 to state 2) and was censored after 995 days without a further event. Patient 2 experienced the adverse event after 12 days (transition from state 1 to state 3), then recovery after 29 days (transition from state 3 to state 4) and a relapse after 422 days (transition from state 4 to state 5). Finally, he/she died after 579 days, but this last event is not relevant to the model, because the patient had already reached an absorbing state.

Transition Matrix: Transition Matrix specifies which direct transitions are possible. This is the 1st step is multi-state model. Here is the transition matrix for the multi state model which shows the possible connections between each states:

```
    to
from     Tx Rec AE Rec+AE Rel Death
  Tx     NA   1  2     NA   3     4
  Rec    NA  NA NA      5   6     7
  AE     NA  NA NA      8   9    10
  Rec+AE NA  NA NA     NA  11    12
  Rel    NA  NA NA     NA  NA    NA
```

```
  Death  NA  NA NA     NA  NA    NA
```

In the present format, the data are not yet suitable for a multi-state analysis. First they have to be recoded into 'long format'. In this format, each subject has as many rows as transitions for which he/she is at risk. Here is the output for the 1st patient when the data has been converted to long format:

	id	from	to	trans	Tstart	Tstop	time	status	match	proph	year	agecl
1	1.00	1.00	2.00	1.00	0.00	22.00	22.00	1.00	no gender mismatch	no	1995-1998	20-40
2	1.00	1.00	3.00	2.00	0.00	22.00	22.00	0.00	no gender mismatch	no	1995-1998	20-40
3	1.00	1.00	5.00	3.00	0.00	22.00	22.00	0.00	no gender mismatch	no	1995-1998	20-40
4	1.00	1.00	6.00	4.00	0.00	22.00	22.00	0.00	no gender mismatch	no	1995-1998	20-40
5	1.00	2.00	4.00	5.00	22.00	995.00	973.00	0.00	no gender mismatch	no	1995-1998	20-40
6	1.00	2.00	5.00	6.00	22.00	995.00	973.00	0.00	no gender mismatch	no	1995-1998	20-40
7	1.00	2.00	6.00	7.00	22.00	995.00	973.00	0.00	no gender mismatch	no	1995-1998	20-40

Starting from state 1, the 1st patient is at risk for transitions 1,...,4. This means that he/she can move to states 2,3,4,5 and 6. The patient moves into state 2 at time 22. Now he/she is at a risk for a further transition to state 4,5 and 6. The patient is censored at time 995 and none of the other transitions occur. The patient has no rows from 8-12 because he/she was never risk at risk for these.

Here are the number of events and proportions of events for each state:

```
    $Frequencies
        to
from       Tx  Rec   AE Rec+AE  Rel Death no event total entering
  Tx        0  785  907      0   95   160       332            2279
  Rec       0    0    0    227  112    39       407             785
  AE        0    0    0    433   56   197       221             907
  Rec+AE    0    0    0      0  107   137       416             660
  Rel       0    0    0      0    0     0       370             370
  Death     0    0    0      0    0     0       533             533

$Proportions
        to
from             Tx        Rec         AE     Rec+AE        Rel      Death
  Tx     0.00000000 0.34444932 0.39798157 0.00000000 0.04168495 0.07020623
  Rec    0.00000000 0.00000000 0.00000000 0.28917197 0.14267516 0.04968153
  AE     0.00000000 0.00000000 0.00000000 0.47739802 0.06174201 0.21719956
  Rec+AE 0.00000000 0.00000000 0.00000000 0.00000000 0.16212121 0.20757576
  Rel    0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
  Death  0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
        to
```

```
from       no event
  Tx      0.14567793
  Rec     0.51847134
  AE      0.24366042
  Rec+AE  0.63030303
  Rel     1.00000000
  Death   1.00000000
```

Next, we will be expanding our covariates.In our example, we obtain six new covariates for each transition: one dummy variable each for donor-recipient match and prophylaxis, and two each for year of transplant and age. So in the new dataset, there will be 12 new variables for donor-recipient match, 12 new variables for prophylaxis, 24 new variables each for age and year, so total of 72 new variables. Here's what the age variable for patient 1 looks like:

	id	from	to	trans	Tstart	Tstop	time	status	year	year2.1	year2.2	year2.3	year2.4	year2.5	year2.6	year2.7	year2.8
174	24	1	2	1	0	11	11	0	1990-1994	0	0	0	0	0	0	0	0
175	24	1	3	2	0	11	11	1	1990-1994	0	0	0	0	0	0	0	0
176	24	1	5	3	0	11	11	0	1990-1994	0	0	0	0	0	0	0	0
177	24	1	6	4	0	11	11	0	1990-1994	0	0	0	0	0	0	0	0
178	24	3	4	8	11	4390	4379	0	1990-1994	0	0	0	0	0	0	0	0
179	24	3	5	9	11	4390	4379	0	1990-1994	0	0	0	0	0	0	0	0
180	24	3	6	10	11	4390	4379	0	1990-1994	0	0	0	0	0	0	0	0

	year2.9	year2.10	year2.11	year2.12
174	0	0	0	0
175	0	0	0	0
176	0	0	0	0
177	0	0	0	0
178	0	0	0	0
179	0	0	0	0
180	0	0	0	0

Figure 9: Year variable for patient 1 after expanding

Cumulative Hazard: For the estimation of the cumulative hazard, we assume we have data with independent censoring. The Nelson-Aalen estimator $\hat{A_{gh}}(t)$ of the cumulative hazard for transition $g \to h$ is given by:

$$\hat{A_{gh}}(t) = \sum_{t_i \leq t} \frac{dN_{gh}(t_i)}{Y_g(t_i)}$$

Where t_i indicate time events, $dN_{gh}(t_i)$ is the observed number of states from g to h at time t_i and $Y_g(t_i)$ is the number of subjects at risk for a transition from g to h. Here is the plot of cumulative hazard for all the different transitions:

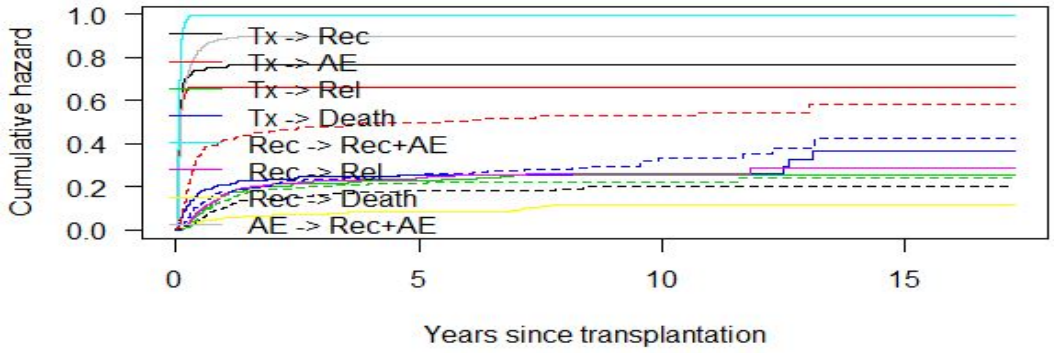


Figure 10: Cumulative Hazard for different transitions

Transition Probability Matrix The one-step transition probability is the probability of transitioning from one state to another in a single step.The transition probability matrix, P, is the matrix consisting of the one-step transition probabilities, p_{ij}. The transition probability matrix $P(s,t)$ has elements:

$$P_{gh}(s,t) = P(X(t) = h|X(s) = g)$$

denoting the transition probability from state g tot state h in time interval $(s;t]$. The transition probability matrix is estimated as:

$$P(\hat{s},t) = \sum_{u\in(s,t]} (I + \triangle\hat{A}(u))$$

where u where u indicates the event times and the elements of A are estimated already. The probtrans() function gives us the transient probability matrix where we get transient probability matrix for each of the 6 states. The transition probabilities are calculated starting from time $t = 0$. Here is how the 1st 6 output looks like for state 1:

```
    Prediction from state 1 (head and tail):
         time   pstate1      pstate2      pstate3 pstate4 pstate5      pstate6          se1
1 0.000000000 1.0000000 0.0000000000 0.0000000000       0       0 0.0000000000 0.0000000000
2 0.002737851 0.9995610 0.0000000000 0.0004389816       0       0 0.0000000000 0.0004388852
3 0.008213552 0.9978051 0.0000000000 0.0021949078       0       0 0.0000000000 0.0009805148
4 0.010951403 0.9956102 0.0000000000 0.0039508341       0       0 0.0004389816 0.0013851313
5 0.013689254 0.9907814 0.0000000000 0.0079016681       0       0 0.0013169447 0.0020023724
6 0.016427105 0.9863916 0.0004389816 0.0118525022       0       0 0.0013169447 0.0024274584
          se2          se3 se4 se5          se6
1 0.0000000000 0.0000000000   0   0 0.0000000000
2 0.0000000000 0.0004388852   0   0 0.0000000000
3 0.0000000000 0.0009805148   0   0 0.0000000000
4 0.0000000000 0.0013143406   0   0 0.0004388852
5 0.0000000000 0.0018550683   0   0 0.0007598375
6 0.0004388852 0.0022674569   0   0 0.0007598375
```

After finding transient probabilities, we can now plot transient probability plot. Here is the transient probability plot starting from time $t = 0$ The

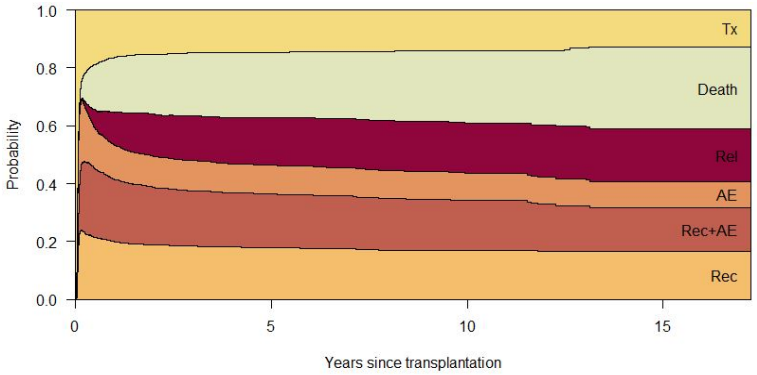


Figure 11: Transient Probability plot

distance between two adjacent curves represents the probability of being in the corresponding state. We can compare prognosis of 2 patients, with and without AE state after 100 days. Here are the plots: We can see that The

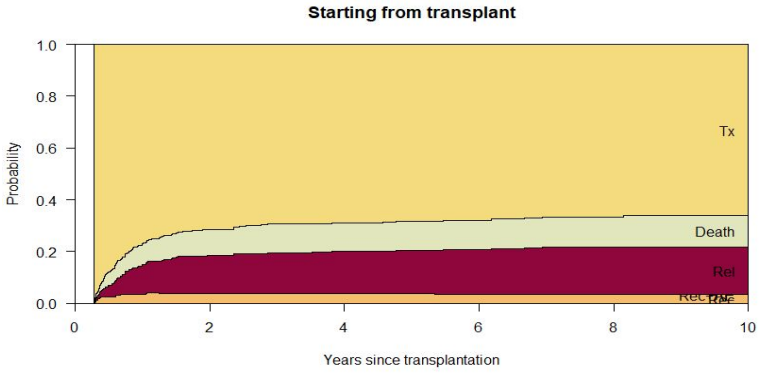


Figure 12: Transient Probability plot for patient 1 without AE after 100 days

relapse probability for patient 2 is somewhat smaller than that for patient 1 and the most likely scenario for 2 is that he/she will have no further events. The fact that patient 1 has not had any adverse event in the 1st 100 days post-transplant has improved his/her prognosis considerably.

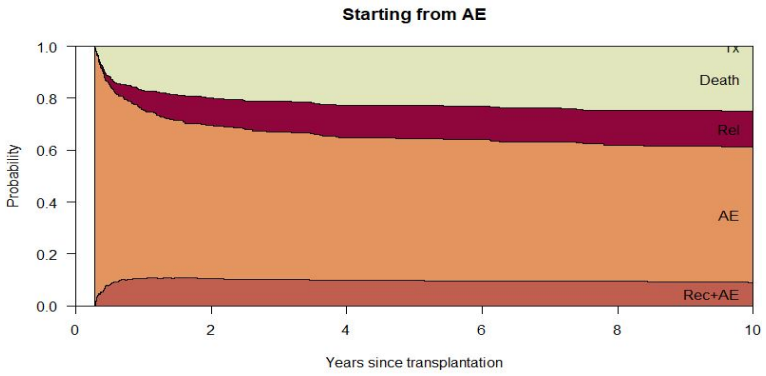


Figure 13: Transient Probability plot for patient 2 with AE after 100 days

4 Discussions

In the 1st part, we worked with lungs cancer dataset which is a record of a set of lungs cancer patients with different features.The only event in the dataset is death. We used Kaplan Meier estimator to estimate survival rate of the patients, then showed how the survival curve is different for male and female, which was also supported by log-rank test. The problem using Kaplan-Meier estimator is that we cannot consider effect of several factors, which is why we used cox proportional hazard model, which takes account of different factors. After applying univariate cox model, we found out only 3 variables: age,sex and ph.ecog were significant. However in the multivariate model age variable fails to become significant.

In the 2nd part we worked with EBMT dataset which is a record of survival of blood cancer patients after transplant.The dataset contains 3 intermediate events: recovery, adverse event and a combination of these two. We 1st converted the dataset to long format which allowed us to represent different states in different row. We then computed cumulative hazard for different state transitions. Cumulative hazards are used in calculating transition probability matrix which allows us to impacts of different transitions after certain days. We worked with non-parametric modelling which does not consider effect of covariates.

References

[1] H. Putter, M.Fiocco and R.B Geskus *Tutorial in biostatistics: Competing risks and multi-state models*. Department of Medical Statistics and Bioinformatics, Leiden University Medical Center.

[2] *Survival Analysis Basics*. http://www.sthda.com/english/wiki/survival-analysis-basics

[3] mstate: An R Package for the Analysis of Competing Risks and Multi-State Models
`http://www.jstatsoft.org/` January 2011, Volume 38, Issue 7.