MATEMATIKOS IR INFORMATIKOS FAKULTETAS

IŠGYVENAMUMO ANALIZĖ (SEMIPARAMTERINĖ REGRESIJA)

Laboratorinis darbas

Atliko: Simona Gelžinytė,
Laineda Morkytė,
duomenų mokslas 3 k. 2gr.

Vilnius, 2023

TURINYS

1. ĮVADAS

1.1 Tyrimo tikslas

Pritaikyti Kokso semiparamterinės regresijos modelį pasirinktiems duomenims.

1.2 Tyrimo uždaviniai

- Atlikti pirminę duomenų analizę;
- Patikrinti modelio prielaidas;
- Sukonstruoti modelį;
- Įvertinti modelio tinkamumą;
- Pateikti gauto modelio interpretacijas;
- Apibendrinti gautus rezultatus, pateikti išvadas.

1.3 Duomenys ir programinė įranga

Pasirinktas duomenų rinkinys - veteranų administracijos plaučių vėžio gydymo tyrimas, kuriame atsitiktinai parinktiems veteranams atliktas dviejų tipų gydymas. Rinkinys turi 8 kovariantės:

- Gydimo tipas (1 = standartinis arba 2 = testinis);
- Naviko histologinis tipas (plokščias, mažų ląstelių, adeno arba didelis);
- Išgyvenamumo laikas (mėnesiais);
- Cenzūros statusas (1 – mirė arba 0 – cenzūruotas);
- Karnofskio efektyvumo balas, apibūdinantis bendrą pacientų būklę tyrimo pradžioje (0-100 balų, 100 = gera būklė);
- Laikas nuo diagnozės nustatymo iki atsitiktinės atrankos (mėnesiais)
- Amžius (metais)
- Ankstesnis gydimas (0 – ne, 10 - taip)

Cenzūruota iš dešinės.

Iš viso yra 137 stebėjimai, praleistų stebėjimų nėra. Tyrimo metu naudota „R“ programinė įranga.

2. SEMIPARAMTERINĖ KOKSO REGRESIJA NAUDOJANT R

Parinkus semiparametrinės Kokso regresijos modelį pereiname visus modelio parinkimo etapus:

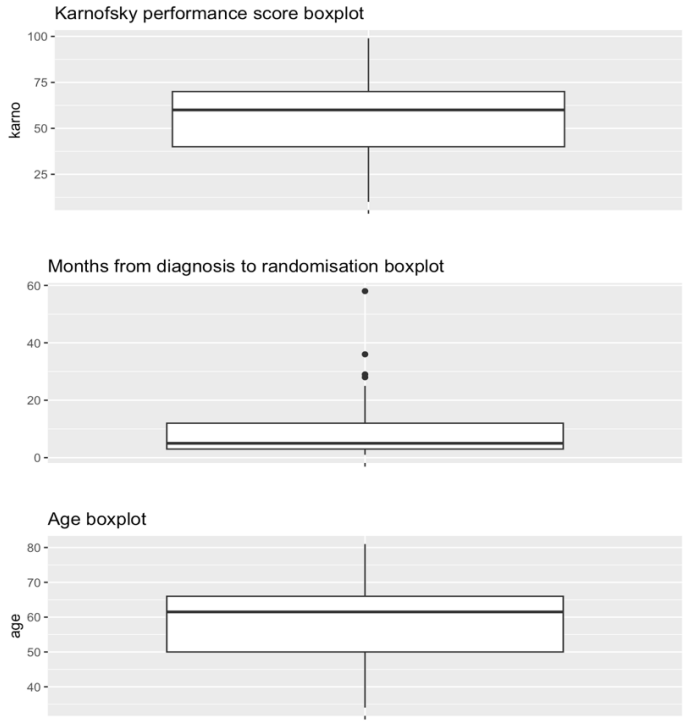
1. Pradinė analizė;
2. Prielaidų tikrinimas – proporcingosios rizikos, išskirtys ir netiesiškumas;
3. Parametrų įvertinimas, interpretacija.

Pirminis Kokso regresijos modelis atrodo taip:

$H(t) = H_0(t) \cdot \exp [b_1 \cdot gydimo\ tipas +\ b_2 \cdot$ naviko histologinis tipas $+\ b_3 \cdot Karnofskio\ efektyvumo\ balas + b_4 \cdot$ laikas nuo diagnozės nustatymo iki atsitiktinės atrankos $+\ b_5 \cdot amžius +\ b_6 \cdot ankstesnis\ gydymas]$

2.1 Pradinė analizė

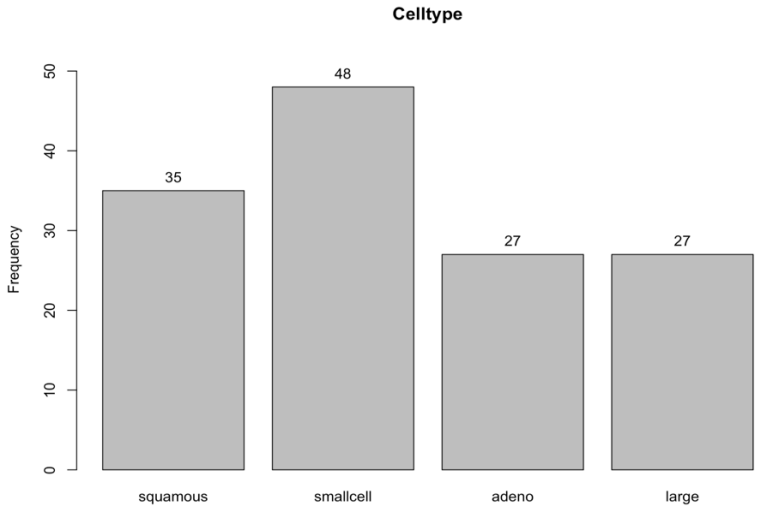
Prieš sudarant Kokso regresijos modelį, vizualiai pasižiūrime į duomenis. Iš stačiakampių diagramų matome, jog išskirtys yra fiksuojamos tik viename požymyje – laikas nuo diagnozės nustatymo iki atsitiktinės atrankos. Iš duomenų aprašomosios statistikos matome, jog Karnofskio efektyvumo balo medianinė reikšmė – 60, žemiausias balas – 10, o aukščiausias – 99, Laiko nuo diagnozės nustatymo iki atsitiktinės atrankos reikšmė – 5 mėnesiai, amžius - 62. Matome, jog 48 pacientams iš 137 naviko tipas buvo nustatytas - mažųjų ląstelių, 35 pacientams - plokščiasis. 69 pacientams buvo paskirtas standartinis gydymo būdas, likusiems – testinis. 97 pacientams prieš atranką nebuvo taikomas joks gydymo būdas.



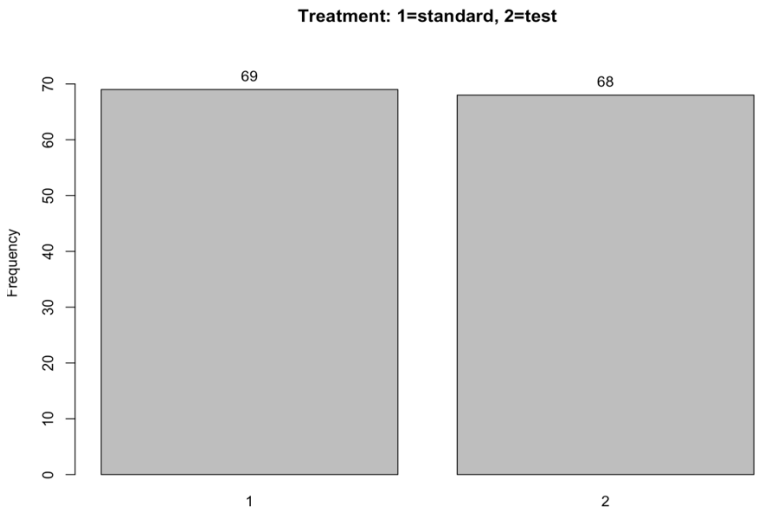
1 pav. Stačiakampių diagramų panelė

1 lentelė. Pradinė duomenų aprašomoji statistika

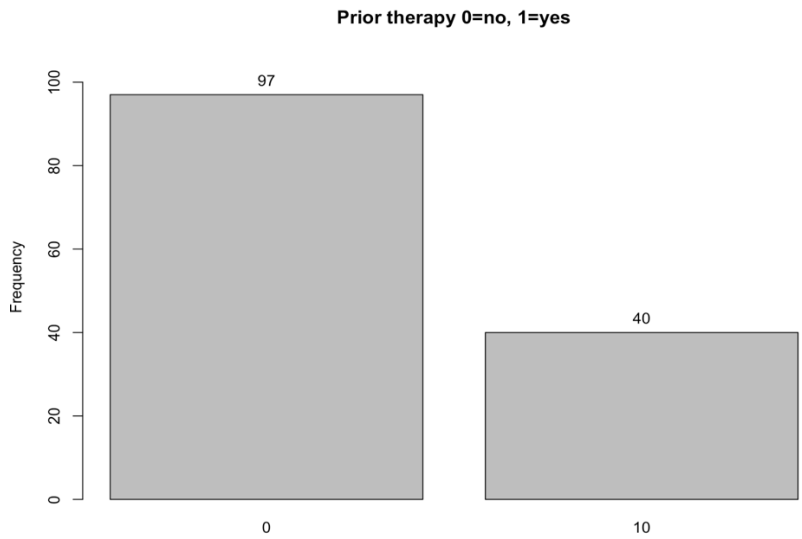
	vidurkis	**mediana**	**min**	**max**
Karnofskio efektyvumo balas	58,57	60	10	99
Laikas nuo diagnozės nustatymo iki atsitiktinės atrankos	9	5	1	87
Amžius	58	62	34	81



2 pav. Histograma apie naviko tipą



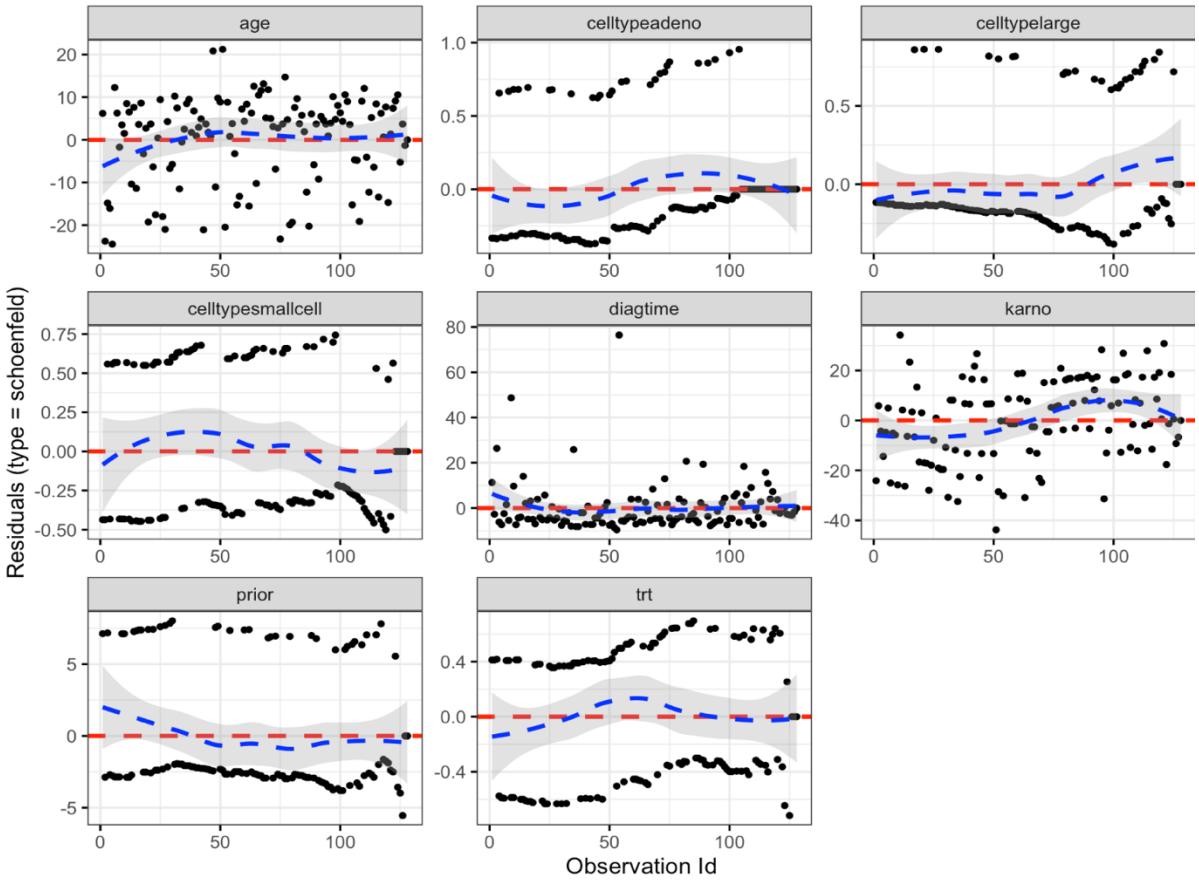
3 pav. Histograma apie gydymo tipą



4 pav. Histograma apie ankstesnį gydymą

2.2 Prielaidų tikrinimas

Atlikę pradinę duomenų analizę pritaikome Kokso regresijos modelį ir tikriname prielaidas. Pirmiausia patikrinsime proporcingosios rizikos prielaidą, naudojant Schoenfield liekanų grafiką ir proporcingų rizikos funkcijų statistinį testą.



5 pav. Schoenfield liekanų grafikas

2 lentelė. Proporcingų rizikos funkcijų statistinio testo rezultatai

	chisq	df	p
trt	0.2644	1	0.60712
celltype	15.2274	3	0.00163
karno	12.9352	1	0.00032
diagtime	0.0129	1	0.90961
age	1.8288	1	0.17627
prior	2.1656	1	0.14113
GLOBAL	34.5525	8	3.2e-05

Iš grafikų bei lentelės, galime matyti, jog ne visos kovariantės tenkina proporcingosios rizikos prielaidą. Esant reikšmingumo lygmeniui 0,05, kintamieji naviko histologinis tipas bei Karnofskio efektyvumo balas šios prielaidos netenkina. Tačiau pažiūrėjus į modelio reikšmingų kovariančių lentelę, galime pastebėti, jog tos pačios kovariantės yra reikšmingos, o tai reiškia, šioms kovariantėms reiks įtraukti laiko sąveiką

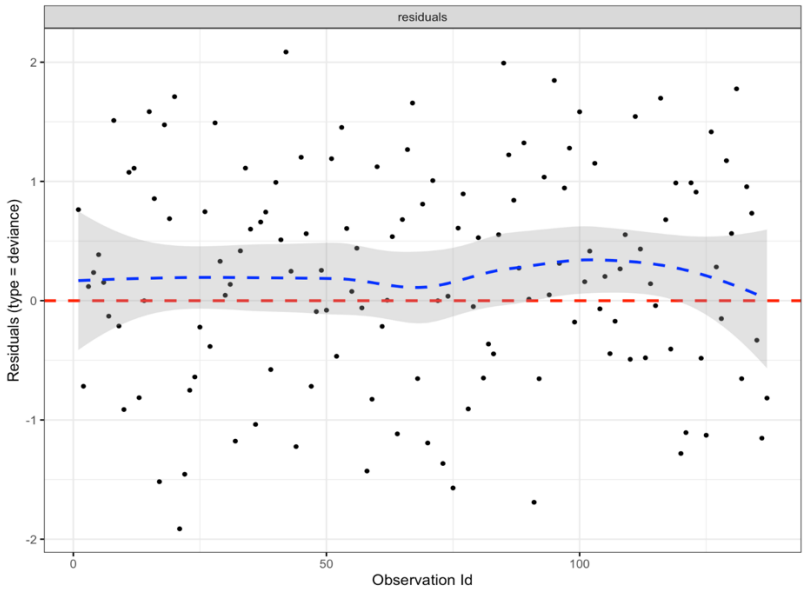
3 lentelė. Reikšmingų kovariančių lentelė

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
trt	2.946e-01	1.343e+00	2.075e-01	1.419	0.15577	
celltypesmallcell	8.616e-01	2.367e+00	2.753e-01	3.130	0.00175	**
celltypeadeno	1.196e+00	3.307e+00	3.009e-01	3.975	7.05e-05	***
celltypelarge	4.013e-01	1.494e+00	2.827e-01	1.420	0.15574	
karno	-3.282e-02	9.677e-01	5.508e-03	-5.958	2.55e-09	***
diagtime	8.132e-05	1.000e+00	9.136e-03	0.009	0.99290	
age	-8.706e-03	9.913e-01	9.300e-03	-0.936	0.34920	
prior	7.159e-03	1.007e+00	2.323e-02	0.308	0.75794	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

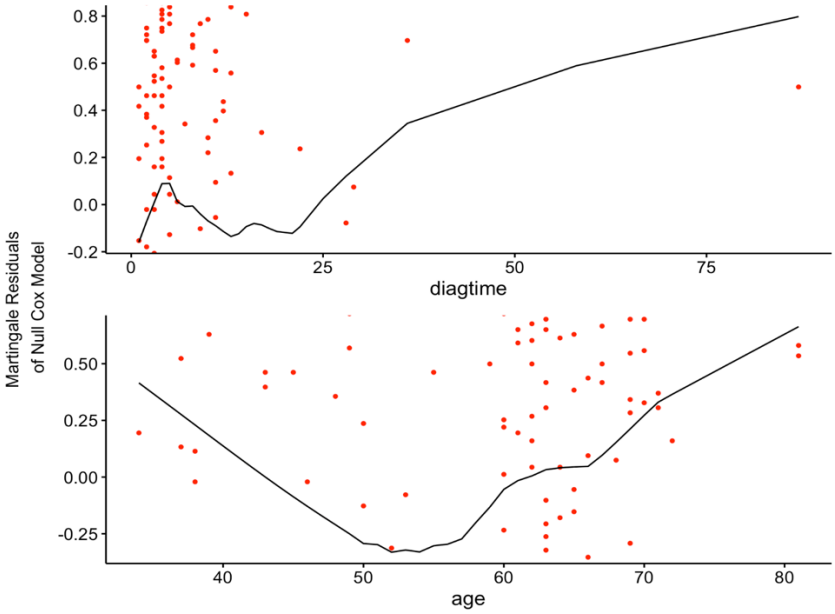
Paminėtoms kovariantėms įtraukiame sąveiką su laiku ir iš naujo užrašome modelį, kuris jau tenkina proporcingosios rizikos prielaidą.

Kadangi pirmoji prielaida patikrinta ir sutvarkyta, dabar galime patikrinti išskirtis. Tam naudosime deviacijos liekanas.



6 pav. Deviacijos liekanos

Kaip galime matyti iš paveikslėlio, visos liekanos yra tarp -2 ir 2, o tai reiškia, jog neturime statistiškai reikšmingų išskirčių. Toliau patikrinsime netiesiškumą, tam naudosime martingalių liekanų ir kovariančių reikšmių sklaidos diagramas, skirtas tik kiekybiniams kintamiesiems.



7 pav. Martingalų liekanų ir kovariančių reikšmių sklaidos diagramos

Kaip galime matyti iš paveikslėlio, netiesiškumo sąlyga tenkinama, todėl galime pereiti prie modelio interpretacijos.

2.3 Parametrų įvertinimas, interpretacija

4 lentelė. Reikšmingos kovariantės

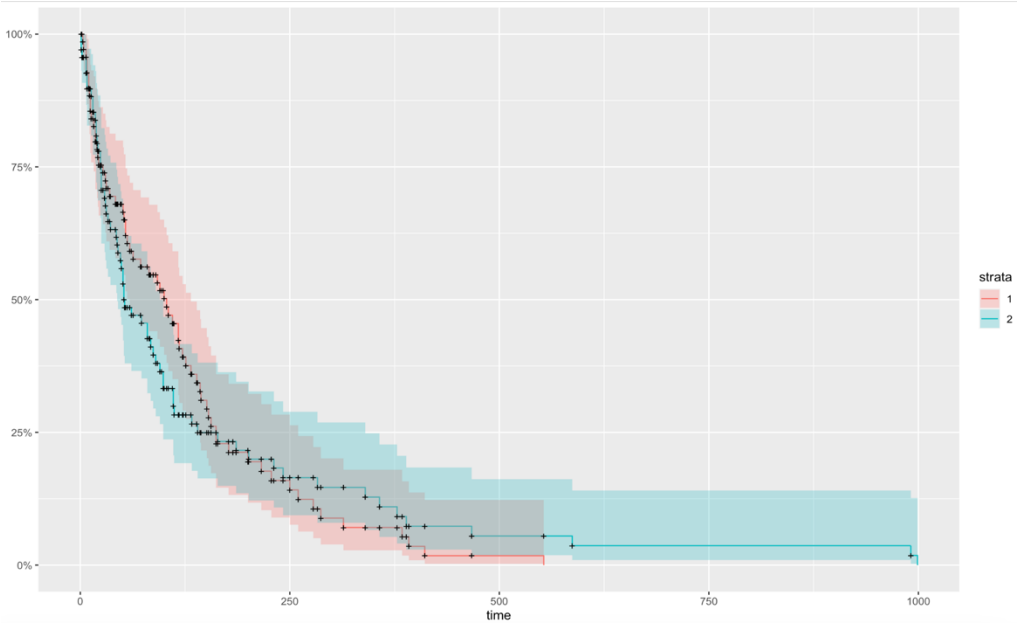
	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
trt	0.240984	1.272501	0.213989	1.126	0.260102	
celltypesmallcell	-0.601377	0.548056	1.413990	-0.425	0.670614	
celltypeadeno	-4.132083	0.016049	1.921668	-2.150	0.031535	*
celltypelarge	-3.621742	0.026736	1.697041	-2.134	0.032830	*
prior	0.015289	1.015407	0.023418	0.653	0.513834	
karno	-0.120268	0.886683	0.032083	-3.749	0.000178	***
diagtime	-0.004100	0.995908	0.008853	-0.463	0.643289	
age	-0.010553	0.989503	0.009605	-1.099	0.271905	
celltypesquamous:stop	-0.836877	0.433061	0.345744	-2.421	0.015499	*
celltypesmallcell:stop	-0.522203	0.593212	0.364112	-1.434	0.151520	
celltypeadeno:stop	0.392409	1.480543	0.482438	0.813	0.415995	
celltypelarge:stop	NA	NA	0.000000	NA	NA	
stop:karno	0.019946	1.020147	0.007506	2.657	0.007874	**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Įtraukus sąveikas, gavome, jog tik dviejų tipų navikai yra reikšmingi (koeficientas neigiamas), t.y. naviko histologinio tipo įtaka laikui bėgant mažėja. Mirties rizika tiems, kurie turėjo adeno naviką sudaro tik 2 proc., o tiems, kurie turėjo didelį naviką – 3 proc. rizikos, lyginant su tais, kurie turėjo kitokio tipo naviką, o skirtumas yra 98 ir 97 proc., t.y. rizika mirti tiems, kurie turėjo adeno naviką yra 98 proc., o tie, kurie turėjo didelį naviką – 97 proc. mažesnė lyginat su kitokio tipo navikais.

Kita reikšminga kovariante - Karnofskio efektyvumo balas, ((0,886-1) * 100 = -11,4), padidėjus Karnofskio balui per vienetą, rizika mirti sumažėja net 11,4%.

Dabar pažiūrėkime išgyvenamumo kreivę.



8 pav. Išgyvenamumo kreivė pagal gydymo tipą

Kaip galime matyti, gydymo tipai gan skirtingi. Naudojant standartinį gydymo tipą, išgyvenamumas po maždaug 100 savaičių siekia 50 proc., kai tuo tarpu naudojant testinį gydymo tipą, po tiek pat laiko, išgyvenamumas siekia tik 35 proc., tačiau vėliau, po maždaug 185 savaičių, testinis gydymo tipas pasirodo daug pranašesnis nei standartinis.

3. IŠVADOS

Tyrime siekta ištirti pacientų išgyvenamumą po plaučių vėžio diagnozės, remiantis turimais duomenimis.

Sudarant Cox semiparametrinį modelį nepagrindinėms kovariantėms, kurioms negaliojo proporcingų rizikos funkcijų prielaida, naudotas laiko sąveikos įtraukimas. Šiuo modeliu gautos statistiškai reikšmingos kovariantės: Karnofskio efektyvumo balas, adeno ir didelis naviko tipai. Buvo pastebėta, jog mirties rizika tiems, kurie turėjo adeno naviką sudaro tik 2 proc., o tiems, kurie turėjo didelį naviką – 3 proc. rizikos, lyginant su tais, kurie turėjo kitokio tipo naviką taip pat, jog padidėjus Karnofskio balui per vienetą, rizika mirti sumažėja net 11,4%. Iš išgyvenamumo kreivės buvo rasta, jog po ilgesnio laiko (~185 savaičių) testinis gydymas tampa efektyvesnis nei standartinis. Gautas galutinis modelis:

$$H(t) = H_0(t) \cdot \exp\left[-4.132 \cdot \text{adeno navikas} - 3.622 \cdot \text{didelis navikas} - 0.120 \cdot Karnofskio\ efektyvumo\ balas\right]$$