

# SKE protokol

Aleksej Gaj\*

17/08/2023

## Obsah

<b>1</b>	<b>Zadání</b>	<b>2</b>
<b>2</b>	<b>Dataset</b>	<b>3</b>
<b>3</b>	<b>Parametrické a neparametrické modely</b>	<b>5</b>
3.1	Parametrické modely . . . . .	6
3.1.1	Pacienti léčení lékem . . . . .	6
3.1.2	Pacienti léčení placebem . . . . .	9
3.2	Neparametrické modely . . . . .	10
3.2.1	Pacienti léčení lékem . . . . .	10
3.2.2	Pacienti léčení placebem . . . . .	11
<b>4</b>	<b>Porovnání podskupin lék vs. placebo</b>	<b>11</b>
<b>5</b>	<b>Coxův regresní model</b>	<b>15</b>
5.1	Ověření předpokladů . . . . .	15
<b>6</b>	<b>Závěr</b>	<b>17</b>

---

\*email: aleksejalex@gmail.com

## 1 Zadání

**A)** Pomocí parametrických a neparametrických metod pro cenzorovaná data odhadněte vhodný spolehlivostní model pro časy dožití (survif  $T_j$ ) obou vybraných podskupin pacientů. Pro kontrolu fitu parametrické rodiny užijte Kaplan-Meierův plot nebo Nelson-Aalenův ‘hazard plot’ (nejlépe v jednom obrázku spolu s parametrickým průběhem), resp. QQ/PP při RC.

**B)** SROVNĚJTE tyto vybrané podskupiny vzhledem k jejich

- průběhu spolehlivosti (survival function)  $R(t)$ , resp.
- intenzitě poruch (survivals)  $\lambda(t)$  (IFR/DFR/CFR), resp.
- kumulativní intenzitě poruch (survivals)  $\Lambda(t)$ , resp.
- střední době života MTTF, resp.
- mediánové době života  $t_{med}$ , resp.
- ... (jiné vlastní, pokud vás něco osloví)

**C)** Graficky srovnajte log-logR ploty pro obě podskupiny a na jejich základě zdůvodněte vhodnost/nevhodnost užití Coxova PH (proportional hazard) modelu.

2em

**Skupina II.:** treat=1(standard) versus treat=2(placebo) pro cell=2(small)

## 2 Dataset

Poskytnutý dataset představuje záznam testování vlivu jistého léčiva na dobu přežití pacienta. Data se skládají ze 137 pozorování 8 proměnných, viz Tabulka 1. Cílem je modelovat dobu dožití (*survival time*), tedy **survt** je vysvětlovaná proměnná. Ta je censorována podle proměnné **cens**. Další proměnné, které jsou k dispozici, představují věk pacienta, typ buněk, Karnofsky score (představující závažnost nemoci<sup>1</sup>), trvání nemoci (proměnná **didur**) a zda pacient už absolvoval léčbu v minulosti.

V této práci se zaměříme na skupinu pacientů s typem buněk **cell=2**.

Název prom.	Komentář
treat	treatment (1 = standard/lék, 2 = test/placebo)
cell	cell type (1 = squamous, 2 = small, 3 = adeno, 4 = large)
survt	survival time (days)
cens	status (0 = censored, 1 = died)
KAR	performance status – Karnofsky score (0 = worst,..., 100 = best)
didur	disease duration from diagnosis to treatment (months)
age	age (years)
prith	prior therapy (0 = none, 10= some)

Tabulka 1: Popis proměnných v datasetu

V Tabulce 2 je uveden základní analýza na celém datasetu, v Tabulce 3 pak analýza podskupiny **cell=2**.

	treat	cell	survt	cens	KAR	didur	age	prith
count	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00
mean	1.50	2.34	121.63	0.74	58.57	8.77	58.31	2.92
std	0.50	1.07	157.82	0.44	20.04	10.61	10.54	4.56
min	1.00	1.00	1.00	0.00	10.00	1.00	34.00	0.00
25%	1.00	1.00	25.00	0.00	40.00	3.00	51.00	0.00
50%	1.00	2.00	80.00	1.00	60.00	5.00	62.00	0.00
75%	2.00	3.00	144.00	1.00	75.00	11.00	66.00	10.00
max	2.00	4.00	999.00	1.00	99.00	87.00	81.00	10.00

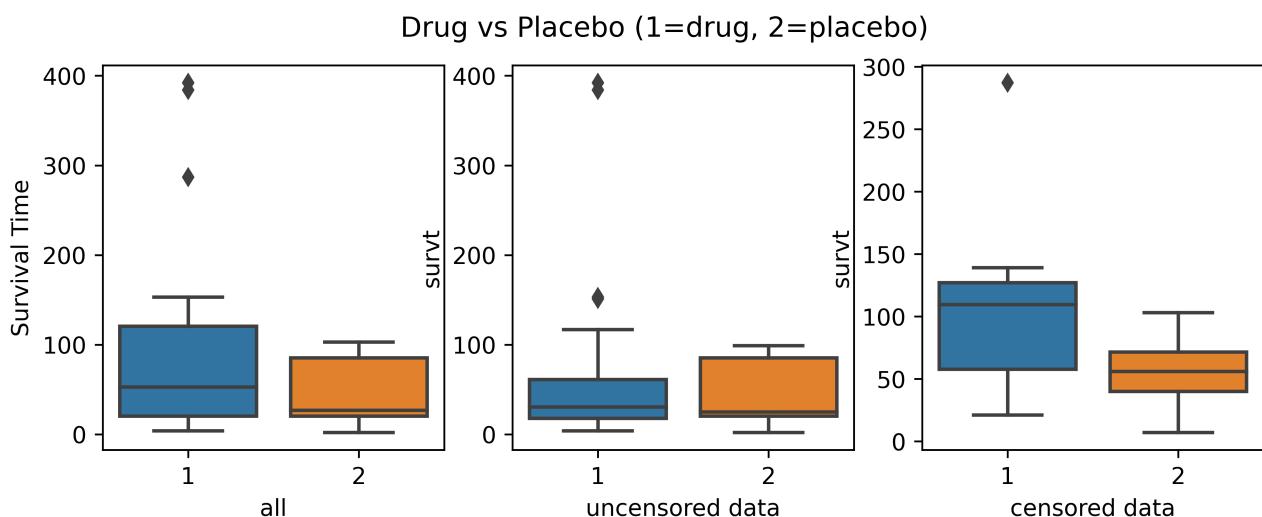
Tabulka 2: Základní analýza na celém datasetu

<sup>1</sup>Pozn: Hodnoty Karnofského skóre znamenají:  $KAR \leq 30$  – úplná hospitalizace,  $30 < KAR \leq 60$  – částečná hospitalizace,  $KAR > 60$  – vlastní péče bez hospitalizace.

	treat	cell	survt	cens	KAR	didur	age	prith
count	48.00	48.00	48.00	48.00	48.00	48.00	48.00	48.00
mean	1.38	2.00	71.67	0.75	53.54	9.25	59.88	2.29
std	0.49	0.00	85.77	0.44	19.10	13.91	9.92	4.25
min	1.00	2.00	2.00	0.00	20.00	1.00	35.00	0.00
25%	1.00	2.00	20.00	0.75	40.00	2.00	54.75	0.00
50%	1.00	2.00	51.00	1.00	60.00	4.00	62.50	0.00
75%	2.00	2.00	97.50	1.00	70.00	11.00	67.00	0.00
max	2.00	2.00	392.00	1.00	85.00	87.00	72.00	10.00

Tabulka 3: Základní analýza na podvybraném datasetu (**cell = 2**)

Podvybraný dataset obsahuje pouze 48 pozorování, jedná se o pacienty průměrně o něco starší a s výrazně kratší průměrnou dobou dožití. Dataset je rozdělen nerovnoměrně: 30 pacientů z 48 bylo léčeno skutečným lékem, a jen 18 z 48 placeboem.



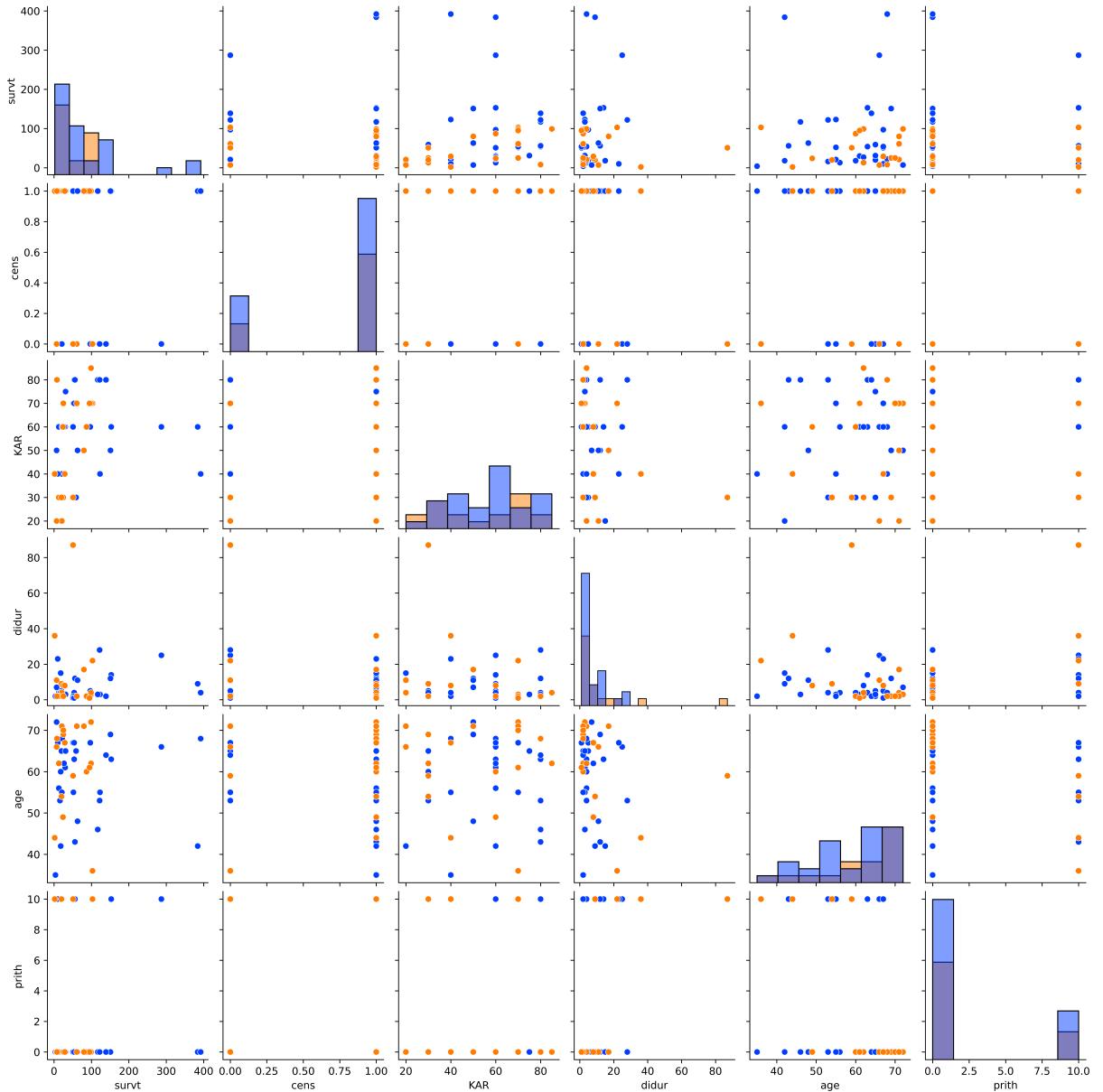
Obrázek 1: Boxplot doby dožití (nalevo - celá analyzovaná skupina, uprostřed - pouze pro necensorovaná data, napravo - pouze pro censorovaná data)

Na Obrázku 1 jsou tři boxploty, ilustrující rozdělení pacientů léčených lékem a placeboem mezi censorovanými<sup>2</sup> a necensorovanými případmi. Na všech třech boxplotech je vidět, že střední hodnota doby přežití je u pacientů léčených lékem vyšší, než u pacientů léčených placeboem. Dokonce mezi skutečně léčenými pacienty je několik outlierů s výrazně nadprůměrnou dobou dožití.

Mezi skutečně léčenými pacienty je několik outlierů s výrazně nadprůměrnou dobou dožití. Může se jednat o případy, pro které testovaný lék je obzvláště účinný nebo pouze o přirozenou výjimku. Skutečnost, že taková odlehlá pozorování jsou pouze v případě léku a pouze na jednu stranu od průměru, napovídá k první variantě.

<sup>2</sup>V kontextu spolehlivostní analýzy censorované pozorování je neúplné nebo ne zcela pozorované měření. V tomto případě censorovaná pozorování jsou například pacienti, kteří přežili po celou dobu experimentu (a tedy okamžik úmrtí nebyl zaznamenán) nebo ukončili léčbu předčasně. Necensorovaná pozorování představují pacienty, jejichž úmrtí bylo pozorováno před koncem experimentu.

Na Obrázku 2 je základní vizualizace datasetu (barevně jsou odděleny podskupiny, tj. lék a placebo). Vizualizace je pomocí pairplotu, tj. mřížka grafů, kde řádky i sloupce představují jednotlivé proměnné (sloupce) v datech. Tedy pairplot odráží, jak vypadá závislost proměnných mezi sebou, a na diagonále jsou histogramy pro danou proměnnou.



Obrázek 2: Pairplot

### 3 Parametrické a neparametrické modely

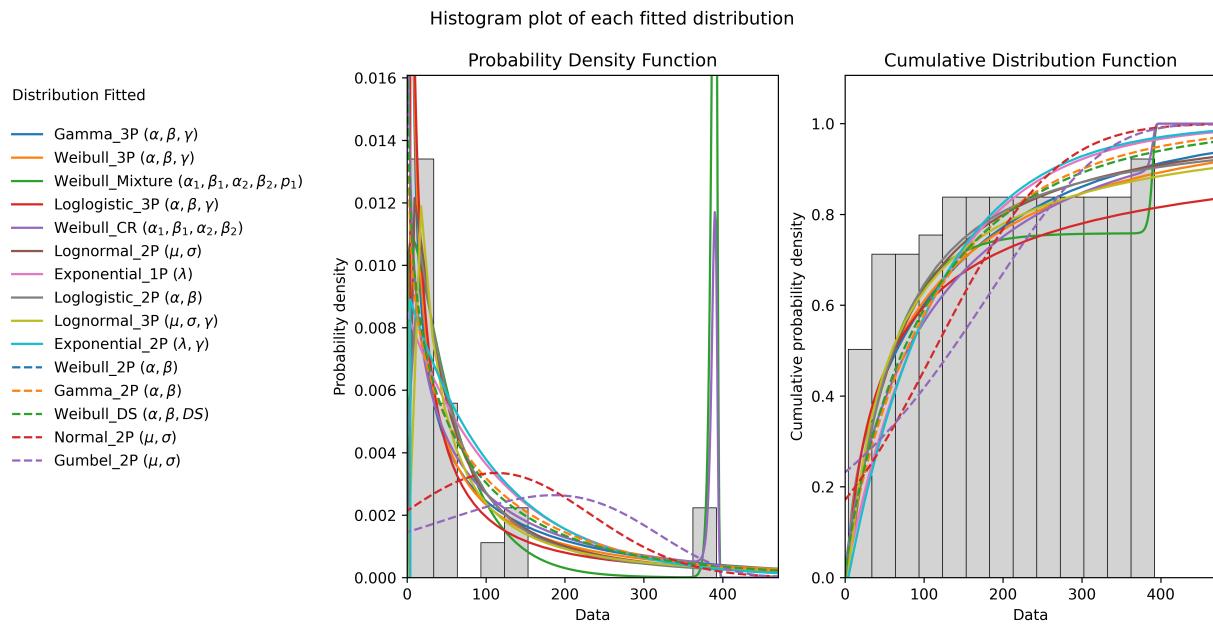
Cílem této práce je najít vhodný model popisující ve zvolené skupině doby dožití pro pacienty, kterým byl podáván lék a pro pacienty, kterým bylo podáváno placebo. Nyní se pokusíme najít nejlepší modely pro popis dat, tedy vytvoříme pro každou skupinu pacientů vlastní model. Nejprve vyzkou-

šíme pro každou podskupinu parametrický přístup, poté neparametrický a následně výsledné modely porovnáme.

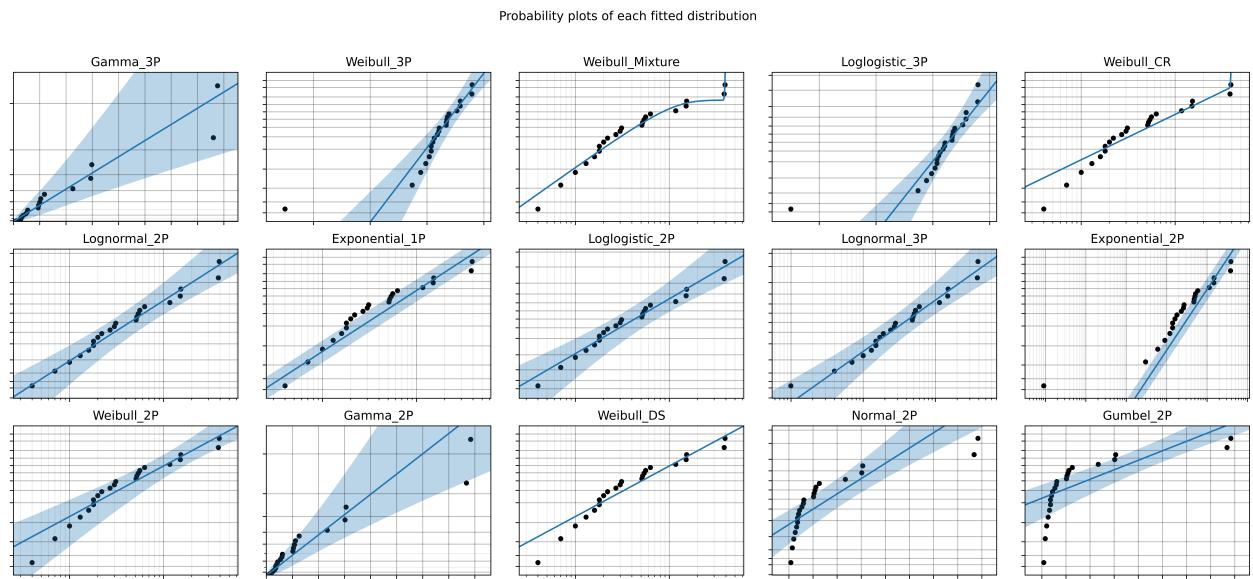
### 3.1 Parametrické modely

Ke hledání vhodného parametrického modelu použijeme existující implementaci v knihovně `reliability`. V Podsekcích 3.1.1 a 3.1.2 jsou pokaždé uvedeny grafické výstupy funkce `Fit_Everything` a stručná diskuze, vysvětlující volbu konkrétní hustoty. Funkce `Fit_Everything` proloží poskytnutá data pomocí 15 předdefinovaných hustot a pro každou odhadne hodnoty parametrů. Následně vykreslí PP ploty a probability ploty pro každý fit.

#### 3.1.1 Pacienti léčení lékem

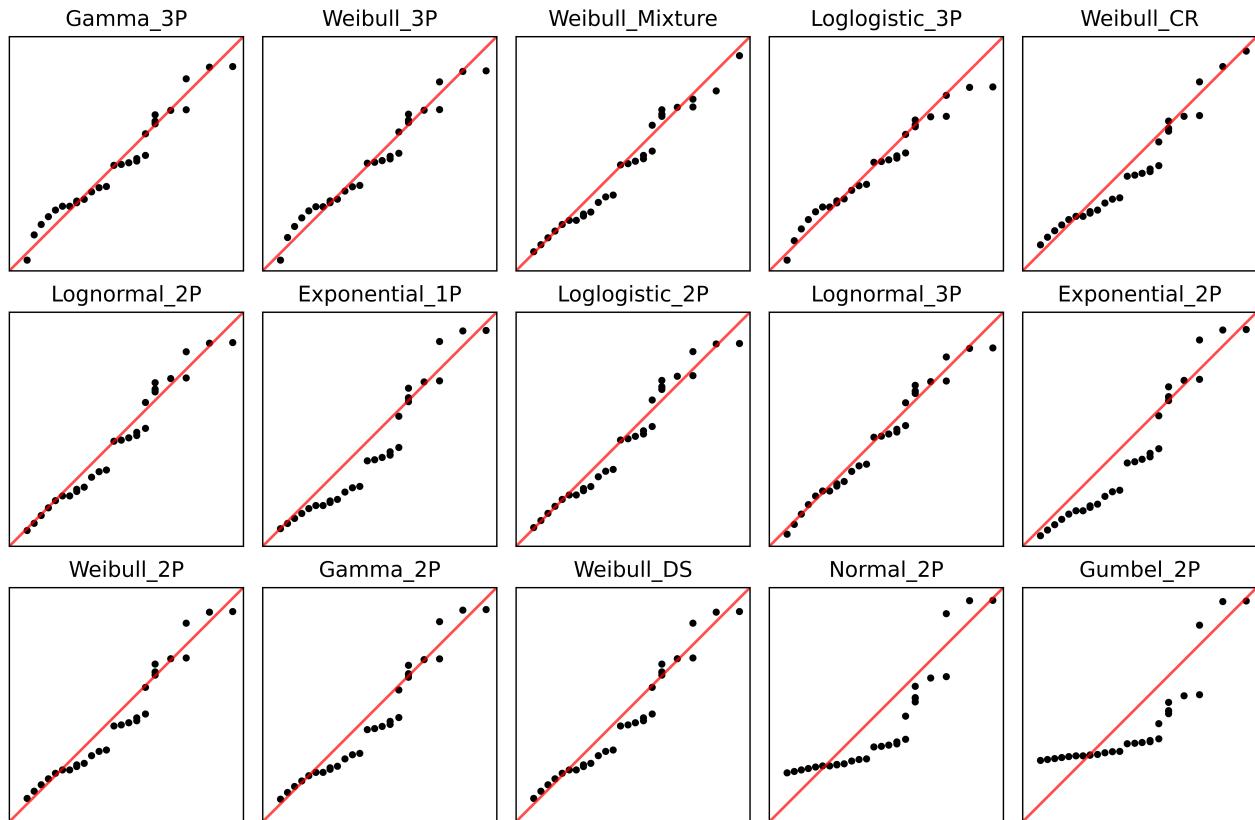


Obrázek 3: Histogram a proložené hustoty



Obrázek 4: P ploty pro jednotlivé modely

Semi-parametric Probability-Probability plots of each fitted distribution  
Parametric (x-axis) vs Non-Parametric (y-axis)



Obrázek 5: PP ploty pro jednotlivé modely

V Tabulce 4 jsou shrnuty hodnoty odhadnutých parametrů a také hodnoty log-likelihood, AIC a BIC<sup>3</sup>.

Distribution	$\alpha$	$\beta$	$\gamma$	$\alpha_1$	$\beta_1$	$\alpha_2$	$\beta_2$	Prop.	DS	$\mu$	$\sigma$	$\lambda$	Log-like.	AIC	BIC
Gamma3P	270.72	0.50	4.00										-120.39	247.71	250.99
Weibull3P	106.13	0.61	4.00										-120.93	248.79	252.07
WeibullMixt.				56.57	1.10	389.96	116.37	0.76					-118.04	248.57	253.08
Loglogistic3P	57.80	0.78	4.00										-122.66	252.25	255.53
WeibullCR				128.49	0.74	390.17	120.53						-121.13	251.86	255.86
Lognormal2P										4.13	1.39		-124.59	253.62	255.98
Exponential1P												0.01	-126.91	255.97	257.23
Loglogistic2P	60.37	1.19											-125.30	255.04	257.39
Lognormal3P			3.00							4.02	1.63		-123.98	254.87	258.15
Exponential2P			4.00									0.01	-125.87	256.18	258.54
Weibull2P	114.41	0.83											-126.20	256.85	259.21
Gamma2P	153.69	0.80											-126.49	257.43	259.79
WeibullDS	114.41	0.83						1.00					-126.20	259.33	262.61
Normal2P										112.97	119.09		-141.89	288.23	290.58
Gumbel2P										185.54	139.42		-147.79	300.02	302.38

Tabulka 4: Odhad parametrů pro podskupinu pacientů léčených lékem

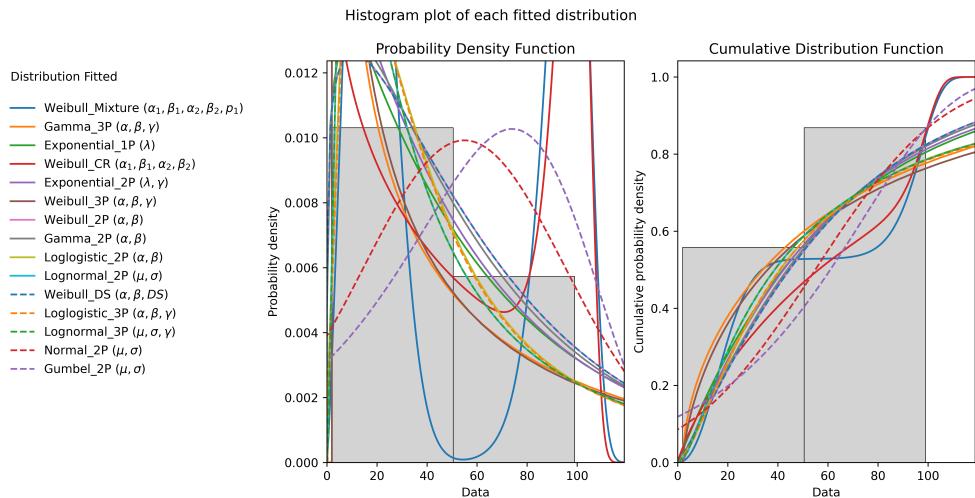
Na základě vizuálního vyhodnocení P-plotů a PP-plotů (Obr. 4 a Obr. 5)

<sup>3</sup>Zde všechna čísla jsou zaokrouhlena na dvě desetinná místa.

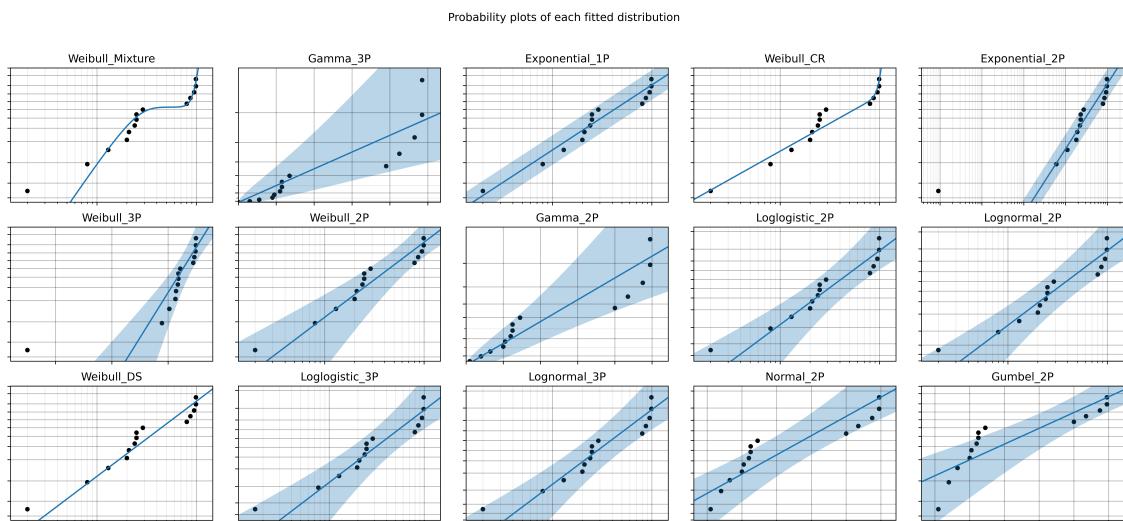
Je vidět, že oba modely (logNormal 2P a Gamma 3P) popisují data podobně dobře, a to i ve srovnání s neparametrickými metodami Kaplan-Meiera a Nelson-Aalena (viz Podsekce 3.2), a tedy zvolíme jednodušší model – dvouparametrickou lognormální hustotu s parametry  $\mu = 4.02$  a  $\sigma = 1.63$ .

### 3.1.2 Pacienti léčení placebem

V této sekci hledáme stejným postupem parametrický model pro podskupinu pacientů s placeboem.



Obrázek 6: Fit all Placebo



Obrázek 7: PP plots pro jednotlivé fity

Číselné hodnoty parametrů jsou opět uvedeny v tabulce.

Distribution	$\alpha$	$\beta$	$\gamma$	$\alpha_1$	$\beta_1$	$\alpha_2$	$\beta_2$	Prop.	DS	$\mu$	$\sigma$	$\lambda$	Log-like.	AIC	BIC
WeibullMixt.				20.74	2.23	98.21	12.93	0.53					-63.67	142.33	141.78
Gamma3P	118.42	0.55	2.00										-68.45	144.61	145.56
Exponential1P												0.02	-71.47	145.19	145.83
WeibullCR				86.32	0.86	100.28	17.23						-67.42	145.91	146.40
Exponential2P			2.00									0.02	-70.86	146.53	147.51
Weibull3P	57.46	0.69	2.00										-69.48	146.67	147.62
Weibull2P	61.35	1.15											-71.27	147.35	148.33
Gamma2P	49.03	1.21											-71.31	147.42	148.41
Loglogistic2P	40.88	1.47											-72.07	148.93	149.91
Lognormal2P										3.66	1.19		-72.11	149.02	150.00
WeibullDS	61.35	1.15						1.00					-71.27	150.26	151.22
Loglogistic3P	40.43	1.44	0.34										-72.06	151.84	152.80
Lognormal3P			0.00							3.66	1.19		-72.11	151.93	152.89
Normal2P										54.95	40.21		-74.39	153.59	154.57
Gumbel2P										74.07	35.81		-75.54	155.88	156.86

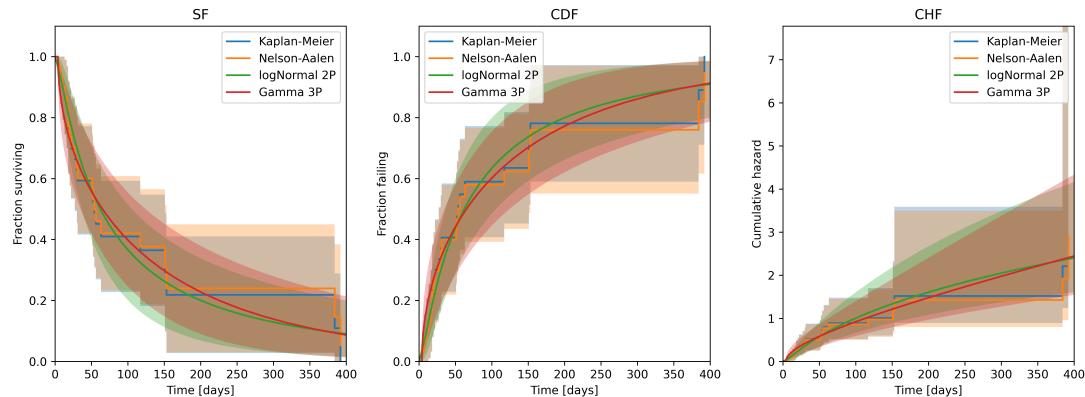
Tabulka 5: Odhadování parametrů pro podskupinu pacientů léčených placebem

Pro podskupinu pacientů, léčených placebem jsme zvolili jednoparametrickou exponenciální hustotu s hodnotou parametru  $\lambda = 0.0165$ . Hustota Gamma3P popisuje data jen nepatrne lépe, zato ale se musí odhadovat 3 parametry místo 1. Volíme tedy jednodušší model.

### 3.2 Neparametrické modely

V Podsekcích 3.2.1 a 3.2.2 jsou pro obě skupiny pacientů (tj. léčení lékem a léčení placebem) uvedeny odhadování Kaplan-Meiera a Nelson-Aalena. Na grafech jsou pro porovnání vyneseny také vybrané parametrické odhadování (viz Sekce 3.1).

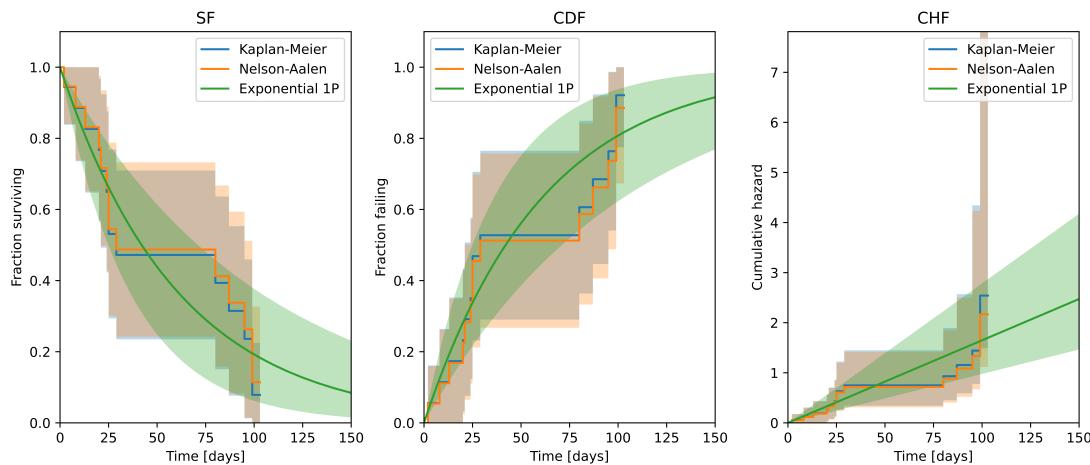
#### 3.2.1 Pacienti léčení lékem



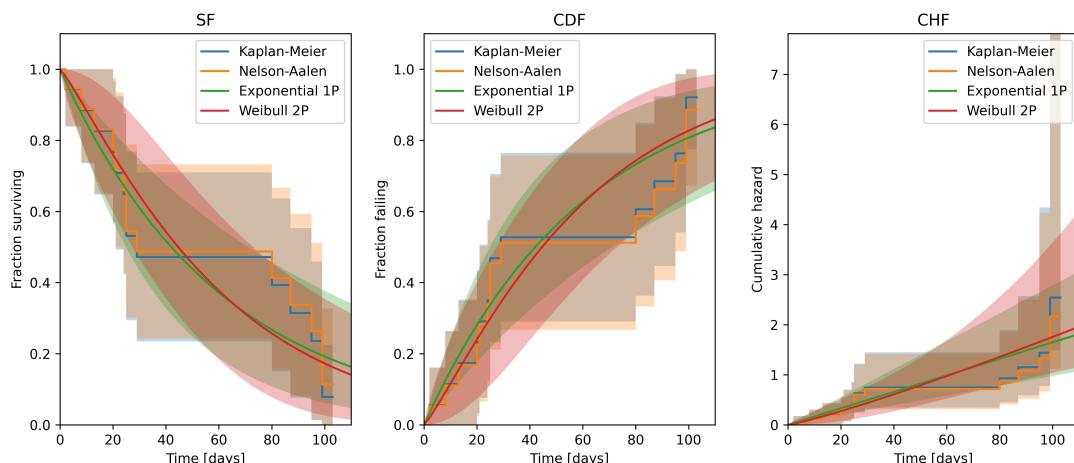
Obrázek 8: Four men drugs

Na ose x je vynesen počet dnů do hodnoty 400, jelikož maximální hodnota survit pro pacienty léčené lékem je 392.

### 3.2.2 Pacienti léčení placeboem



Obrázek 9: Four men placebo

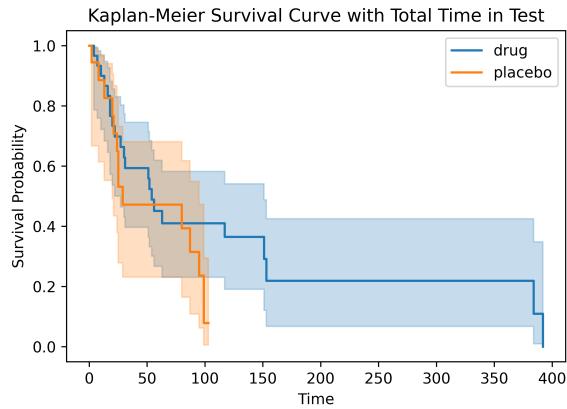


Obrázek 10: Four men placebo compared to parametrics

Zde je počet dnů vynesen pouze do hodnoty 110, jelikož maximální hodnota `survt` pro pacienty léčené placeboem je 103.

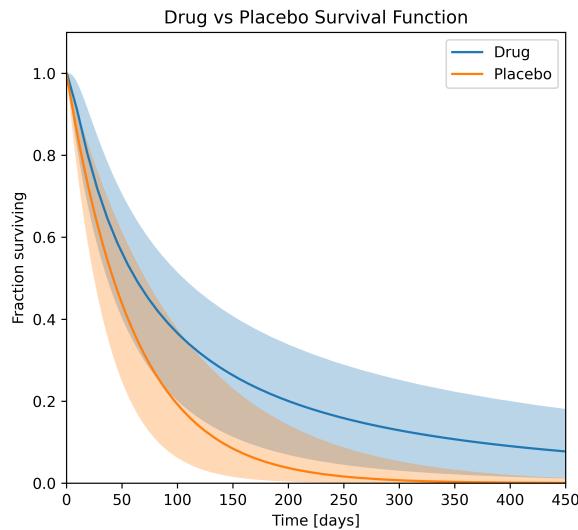
## 4 Porovnání podskupin lék vs. placebo

V této sekci uvedeme porovnání modelů, sestavených v předchozích sekcích. Na Obrázku 11 je TTTplot (total time in test), konstruovaný na základě Kaplan-Meiera. Je vidět, že pacienti, dostávající placebo, umírali výrazně rychleji. Na grafu pacientů, léčených lékem (modrá barva) je mezi 150. a 360. dnem vidět plató - patrně lze říct, že u dosud přeživších se eskalace nemoci zastavila (lék zabral).

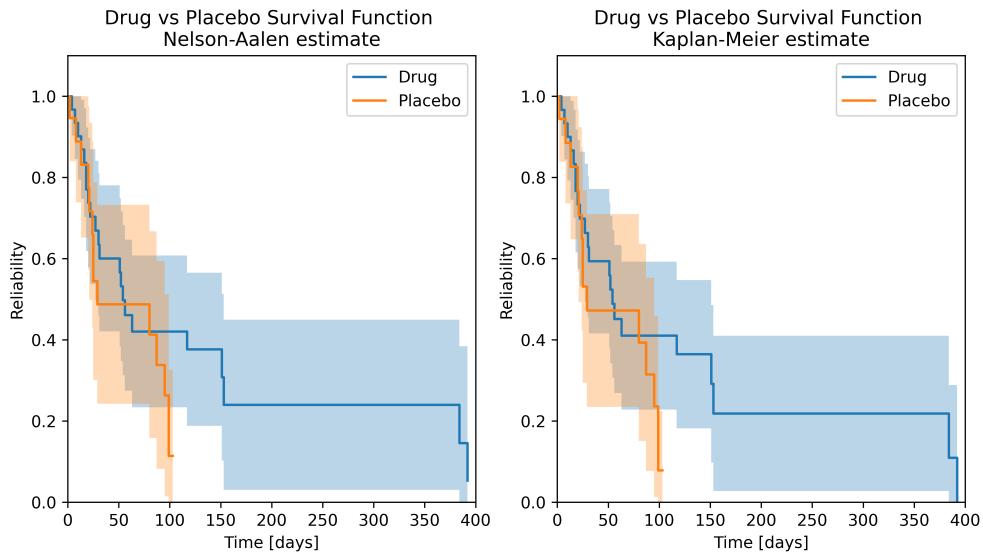


Obrázek 11: TTT plot (na základě Kaplan-Meiera)

Obrázek 12 je ve shodě s Obrázkem 11, tedy ilustruje o něco rychlejší (ve smyslu doby do úmrtí) průběh nemoci u pacientů, kteří dostávali placebo.

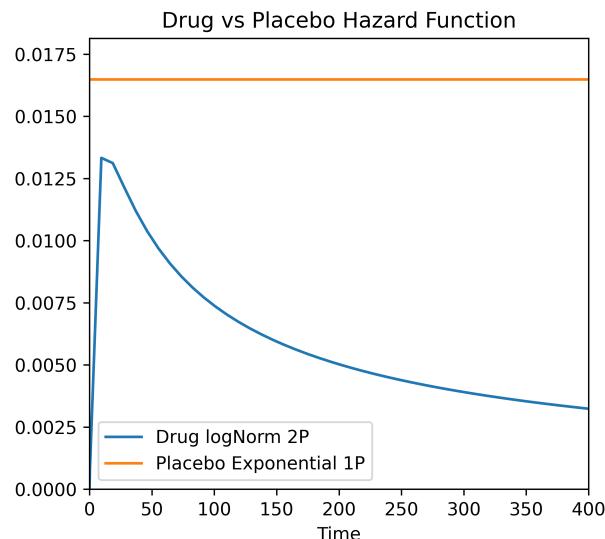


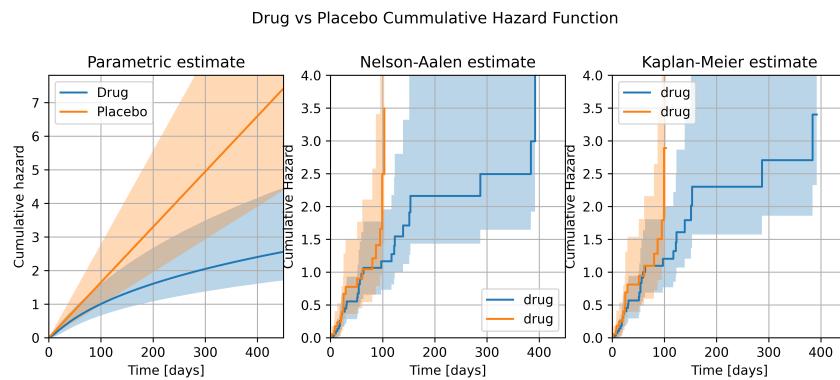
Obrázek 12: Průběh survival function pro vybrané parametrické modely



Obrázek 13: Průběh survival function pro vybrané neparametrické modely

Na Obrázku 14 je průběh intenzity poruch pro dvě zvolené parametrické hustoty: lognormální o dvou parametrech (popisující podskupinu pacientů, léčených lékem) a exponenciální o jednom parametru (popisující podskupinu pacientů, léčených placebem). Pacienti s placebem mají *constant failure rate*. U pacientů s lékem je vidět, že pokud pacient přežije prvních 30 dnů léčby, dále riziko úmrtí bude klesat.

Obrázek 14: Průběh intenzity poruch (*hazard function*) pro vybrané parametrické modely



Obrázek 15: Kumulativní intenzity poruch (*cumulative hazard function*)

Vybrané číselné charakteristiky jsou uvedeny v tabulce níže.

Podskupina	Použitý model	MTTF [dnů]	$t_{\text{med}}$	MRL [dnů]
Lék	logNormal2P	163.64	62.11	270.4
	Kaplan-Meier	129.95	54	
Placebo	Exp1P	60.64	42.03	60.64
	Kaplan-Meier	54.16	29	

Tabulka 6: MRL = střední zbytková doba života (mean residual life), MTTF = střední doba do úmrtí (mean time to failure),  $t_{\text{med}}$  = mediánová doba života

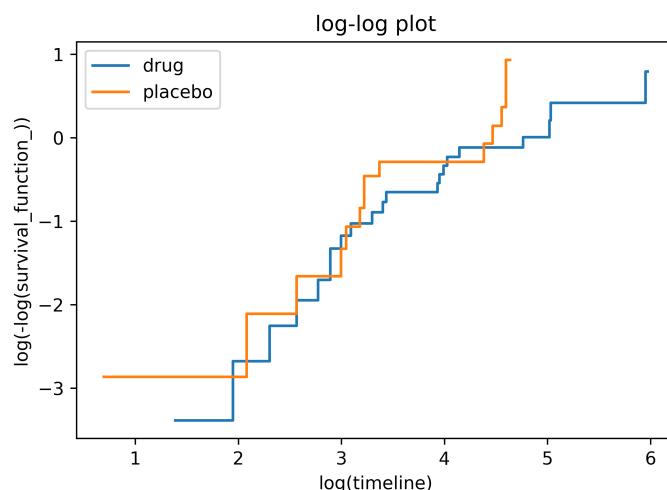
## 5 Coxův regresní model

*Coxův regresní model* je založen na Coxově proporcionalním hazardovém předpokladu, který tvrdí, že poměrné riziko dvou skupin je konstantní v čase. Tento předpoklad umožnuje odhadnout vliv různých faktorů na přežití a současně zachovává nezávislost na neovlivňujících proměnných.

Pro použití Coxova regresního modelu je nejprve potřeba mít k dispozici data o čase do události (např. úmrtí) a příslušné prediktory (faktory ovlivňující přežití). Zde byla použita existující implementace Coxova regresního modelu v knihovně **lifelines**.

### 5.1 Ověření předpokladů

Začneme vizuální kontrolou průběhu log-log  $\hat{R}_{KM}$  (Obrázek 16).



Obrázek 16: log-log plot lék vs placebo

Je vidět, že grafy nejsou “rovnoběžné”, na několika místech se kříží. Znamená to, že v tomto případě Coxův regresní model nelze použít.

**Postulate 1 (DO NOT leave).** Důvodem, proč nemohu použít Coxův proporcionalní model rizik pro mou situaci, je přítomnost několika závažných problémů ve sbírce dat a vlastnostech mého datového souboru:

1. Necenzurovaná data: Coxův model předpokládá, že data jsou cenzurována pouze zprava, což znamená, že neposkytuje adekvátní analýzu, pokud mám data cenzurována zleva nebo jsou nedostupná. Pokud je v mé datové sadě přítomnost těchto typů cenzury, mohou výsledky modelu být zkreslené a nepřesné.

2. Nelinearity in effects: Coxův model funguje nejlépe, pokud se předpokládá, že efekty proměnných jsou lineární v čase. Pokud mám důvody domnívat se, že efekty se mohou měnit nebo jsou nelineární, model nemusí poskytnout adekvátní odhad rizika.

3. Interakce mezi proměnnými: Pokud existují interakce mezi proměnnými, které ovlivňují riziko, Coxův model může být problematický. Model nefunguje dobře při komplexních vztazích mezi proměnnými.

4. Nízký počet událostí: Coxův model vyžaduje dostatečný počet událostí v porovnání s počtem proměnných, aby byl spolehlivý a stabilní. Pokud mám málo událostí ve srovnání s počtem proměnných, model může vykazovat nedostatečnou přesnost.

5. Nezávislé pozorování: Coxův model předpokládá, že pozorování jsou nezávislá. Pokud mám v datech přítomnost korelovaných nebo shlukovaných pozorování, může to model zpochybnit.

Vzhledem k těmto omezením je vhodné hledat alternativní modely, které lépe odpovídají charakteru mých dat a umožní spolehlivě odhadnout riziko ve zkoumaném kontextu.

## 6 Závěr

Tato práce