SKE protokol

Michaela Mašková

15. července 2021

Tento protokol je závěrečnou prací předmětu SKE vypracovanou v akademickém roce 2020/2021.

Poznámka: Analýza byla prováděna v programovacím jazyce Python a pracovní Jupyter notebook s celým kódem pro vytvoření tohoto protokolu je dostupný na adrese:

https://github.com/masenka31/fjfi_projects/tree/SKE/2020-2021/SKE/python_code

Nejedná se o zcela upravenou verzi, nicméně může sloužit jako dobrý doplněk pro získání náhledu do výsledků. Obsahuje i SKE.HTML dokument, který je jednodušší pro prohlížení.

Obsah

1 Zadání

1. Pomocí parametrických a neparametrických metod pro cenzorovaná data odhadněte vhodný spolehlivostní model pro časy dožití (survt T_j) vybraných podskupin pacientů. Pro kontrolu fitu parametrické rodiny užijte Nelson-Aalenův 'hazard plot', resp. QQ/PP při RC.

2. Zhodnoťte shodu/rozdílnost pro tyto vybrané podskupiny vzhledem k jejich

   - průběhu spolehlivosti (survival function) $R(t)$, resp.
   - intenzitě poruch (survivals) $r(t)$ (IFR/DFR/CFR), resp.
   - kumulativní intenzitě poruch (survivals) $\Lambda(t)$, resp.
   - střední době života MTTF, resp.
   - mediánové době života t_{med}, resp.
   - střední zbytkové době života MRL($t_0 = 90$ dnů), resp. ...

3. Graficky srovnejte log-log $\widehat{R}_{KM}$ ploty pro obě podskupiny a na jejich základě zdůvodněte vhodnost/nevhodnost užití Coxova PH (proportional hazard) modelu.

   Dobrovolně nepovinně:

4. Pokud to log-log $\widehat{R}_{KM}$ ploty dovolí, můžete zkusit aplikovat Coxův PH model semiparametrické regrese (při zvolených doprovodných kovariátách X) a najít jejich efekty na vysvětlení časů dožití (survt).

5. Zkusit si, jak se na odhadech $R(t)$ projeví vliv cenzorovaných dat v podskupině [treat=1 & cell=2]. Tedy srovnejte odhady $R(t)$ v této podskupině, pokud:

   (a) ignorujete příznak censored (úplný soubor n=30),

   (b) vypustíte řádky s příznakem cens=0 (úplný soubor n=22),

   (c) uvažujete řádný cenzorovaný soubor n=22(poruch)+8(cenzorů).

2 Data

2.1 Základní charakteristiky

Data se skládají z 8 sloupců, jejichž popis je následující:

```
Column 1 = treatment (1 = standard/lek, 2 = test/placebo)
Column 2 = cell type (1 = squamous, 2 = small, 3 = adeno, 4 = large)
Column 3 = survival time (days)
Column 4 = status (0 = censored, 1 = died)
Column 5 = performance status - Karnofsky score (0 = worst,...,100 = best)
Column 6 = disease duration from diagnosis to treatment (months)
Column 7 = age (years)
Column 8 = prior therapy (0 = none, 10 = some)
```

Máme tedy 8 proměnných, modelovat chceme *survival time* (spojitá proměnná), což je však censorovaná proměnná podle sloupce `cens`, proměnná *status*. Můžeme si všimnout, že máme pouze tři spojité proměnné, již zmíněný čas dožití, dále pak KAR skóre a věk pacienta. Zbylé proměnné jsou kategorické, pouze typ rakovinových buněk má více úrovní.

Celkem bylo ve studii přítomno 137 pacientů. Souhrné charakteristiky datasetu jsou následující:

	treat	cell	survt	cens	KAR	didur	age	prith
mean	1.5	2.34	121.63	0.74	58.57	8.77	58.31	2.92
std	0.5	1.07	157.82	0.44	20.04	10.61	10.54	4.56
min	1.0	1.00	1.00	0.00	10.00	1.00	34.00	0.00
25%	1.0	1.00	25.00	0.00	40.00	3.00	51.00	0.00
50%	1.0	2.00	80.00	1.00	60.00	5.00	62.00	0.00
75%	2.0	3.00	144.00	1.00	75.00	11.00	66.00	10.00
max	2.0	4.00	999.00	1.00	99.00	87.00	81.00	10.00

Vidíme zde například, že průměrný věk pacientů je 58 let. Pro analýzu jsem si vybrala podskupinu I) podle zadání:

`treat=1(standard) versus treat=2(placebo) pro cell=1(squamous),`

ve které se vyskytuje již pouze 35 pacientů. Souhrné charakteristiky tohoto zmenšeného datasetu jsou následující:

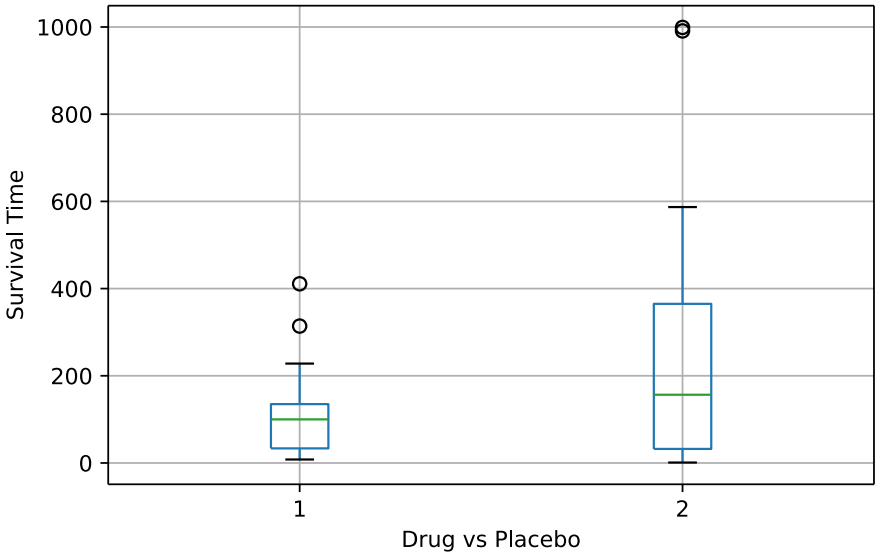
	treat	cell	survt	cens	KAR	didur	age	prith
mean	1.57	1.0	200.20	0.74	60.86	11.03	58.46	4.00
std	0.50	0.0	248.23	0.44	20.49	11.53	10.37	4.97
min	1.00	1.0	1.00	0.00	20.00	1.00	35.00	0.00
25%	1.00	1.0	31.50	0.50	50.00	4.00	51.50	0.00
50%	2.00	1.0	111.00	1.00	60.00	7.00	62.00	0.00
75%	2.00	1.0	262.50	1.00	75.00	12.50	64.50	10.00
max	2.00	1.0	999.00	1.00	90.00	58.00	81.00	10.00

Můžeme například vidět, že průměrný věk zůstává zhruba stejný jako v celém datasetu, že je zhruba stejný počet pacientů, kteří dostali lék a těch, kteří dostali placebo. Zvýšila se nám také střední hodnota *survival time*. Nicméně, zde se stále bavíme o censorovaných datech. Když budeme zkoumat dál, zjistíme, že v našem subdatasetu 35 pacientů je 9 censorovaných zápisů.

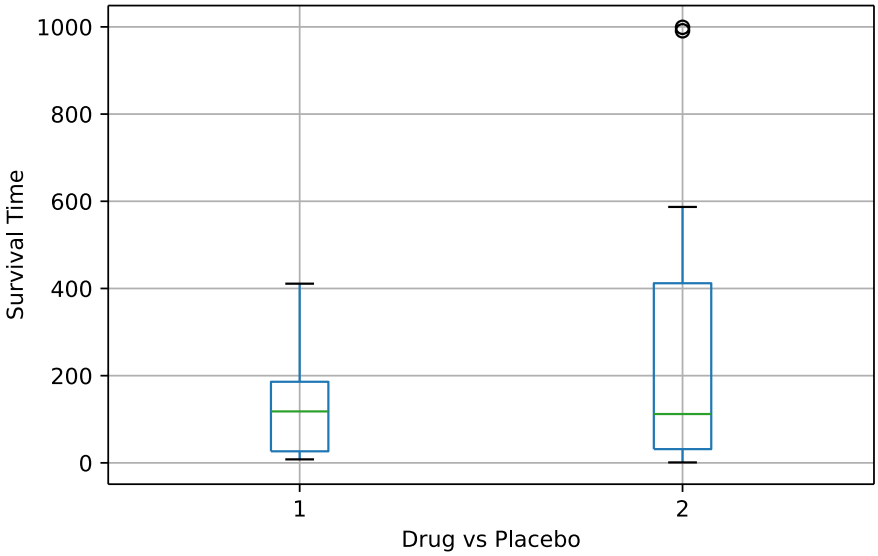
2.2 Deskriptivní analýza

Naším cílem bude zjistit, jestli je jiné riziko (jiná doba dožití) pro pacienty, kteří byli léčeni a těmi, kteří dostávali pouze placebo. Můžeme se podívat na boxploty pro tyto dvě skupiny. Na prvním boxplotu na Obr. 1 vidíme výsledek pro všech 35 pacientů (včetně cenzorovaných dat). Na obrázku 2 pak můžeme

vidět druhý boxplot, tentokrát pouze pro necensorovaná data. Z obou obrázků si můžeme všimnout, že obecně se zdá doba přežití delší pro pacienty dostávající placebo než reálný lék. Pacienti dostávající placebo mají větší rozptyl v hodnotách, ovšem medián pro necensorovaná data je zhruba stejný jako pro pacienty dostávající lék. Vidíme také několik outlierů, kdy se časy dožití pohybují kolem 1000 dní. Jedná se konkrétně pro dva pacienty dostávající placebo, jejichž čas dožití je 991 a 999 dní (necensorované hodnoty).



Obrázek 1: Boxplot pro čas dožití podle lék vs placebo pro všechna data



Obrázek 2: Boxplot pro čas dožití podle lék vs placebo pouze pro necensorovaná data

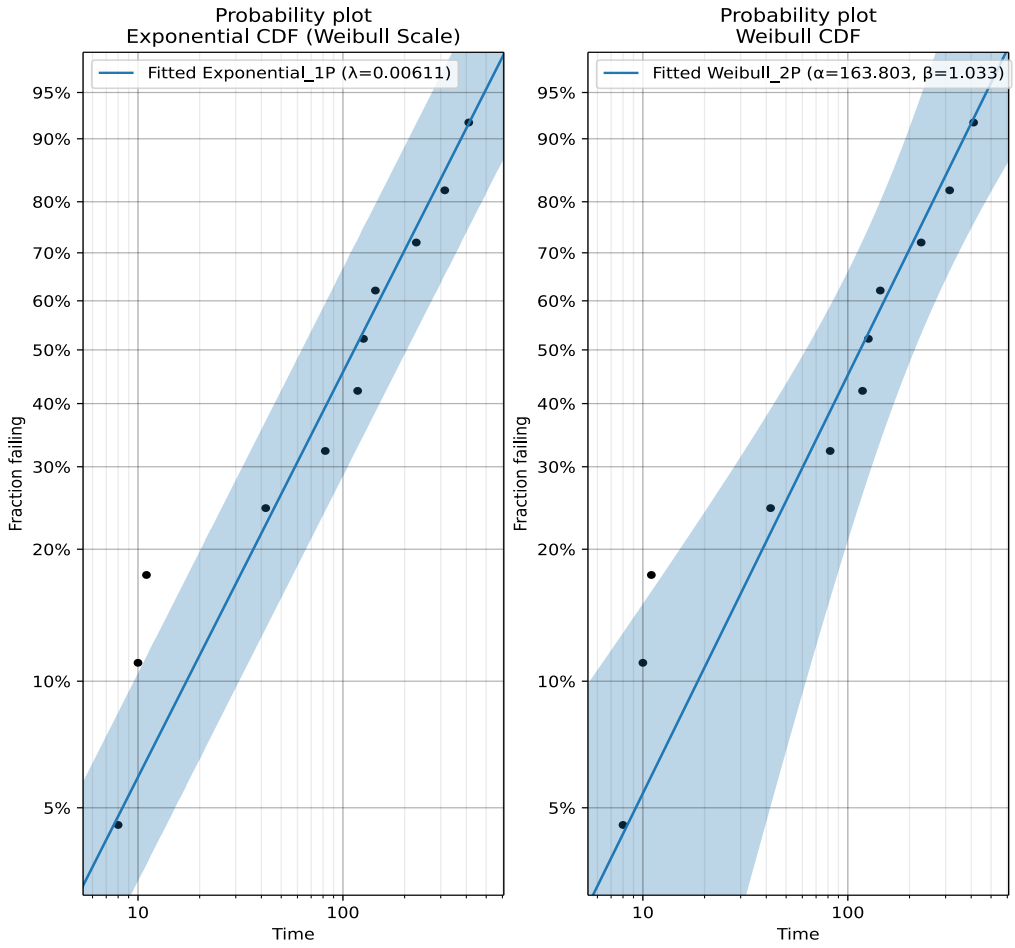
3 Parametrické a neparametrické modely

V této sekci se pokusíme najít nejvhodnější modely pro popis dat. Pro každou podskupinu pacientů (lék vs placebo) vytvoříme vlastní model. Pro analýzu použijeme balíček `reliability`, který umožňuje rychlou analýzu s již implementovanými metodami právě pro spolehlivostní analýzu.

3.1 Pacienti - lék

První skupinou jsou pacienti, kteří byli léčeni. Pro zjištění, které rozdělení pravděpobodnosti bude pro data nejvhodnější, použijeme funkci `Fit_Everything`, která automaticky nafituje několik rozdělení a vytvoří celkovou analýzu všech modelů. Modely jsou seřazené podle hodnoty log-likelihood.

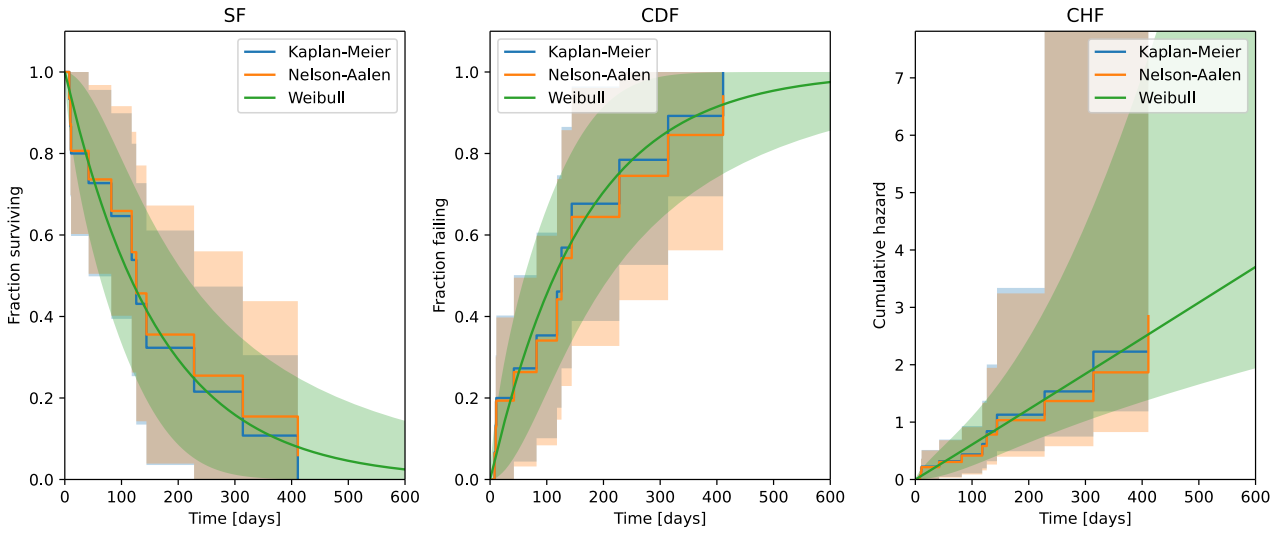
Z vizuálního porovnání výsledných QQ plotů a PP plotů se nejlepší zdají modely Exponenciálního rozdělení s jedním parametrem a Weibullova rozdělení se dvěma parametry. Na Obr. 3 jsou vidět Probability ploty pro obě rozdělení.



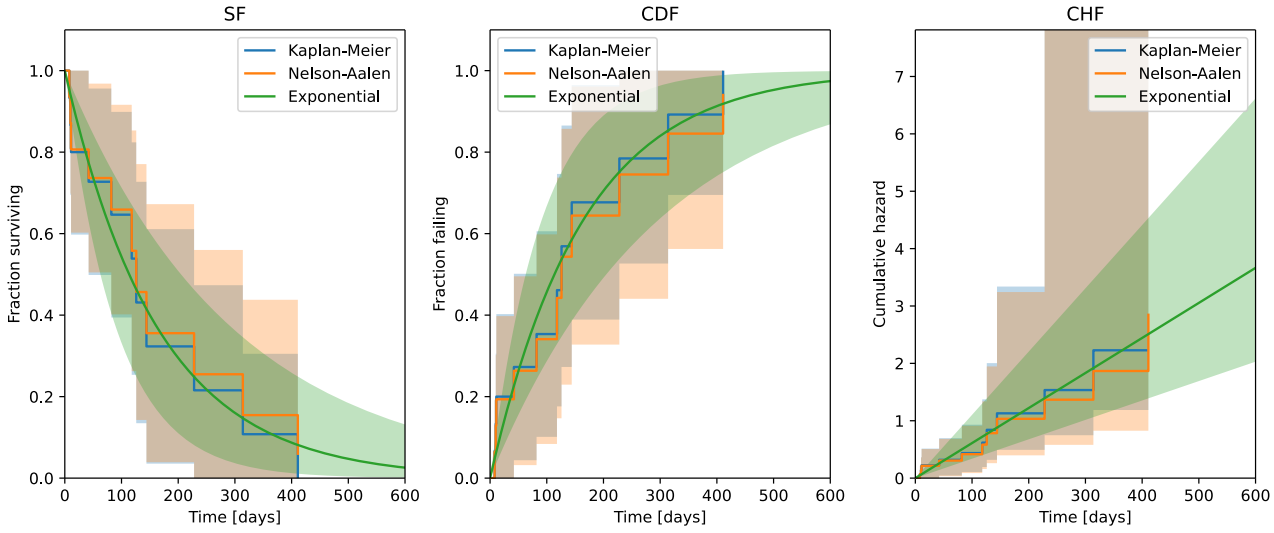
Obrázek 3: Probability plot pro Exponenciální a Weibullovo rozdělení pro skupinu pacientů 1.

Dále se můžeme podívat na survival function, cummulative distribution function a cummulative hazard function a jejich porovnání s neparametrickými odhady pomocí Kaplan-Meier odhadu a Nelson-Aalen odhadu. Vizualizace pro Weibullovo rozdělení je na Obr. 4, pro Exponenciální rozdělení pak na Obr. 5. Z obrázků je také vidět, že obě rozdělení jsou výsledkem velmi podobná, pouze Weibullovo rozdělení

disponuje širším konfidenčním intervalem. Nezdá se, že by Weibullovo rozdělení přinášelo výrazně lepší model, proto budeme ctít poučku, že jednodušší model je lepší a zvolíme jako parametrický model pro tuto skupinu pacientů model Exponenciální.



Obrázek 4: Porovnání parametrického odhadu Weibullova rozdělení s neparametrickými odhady.

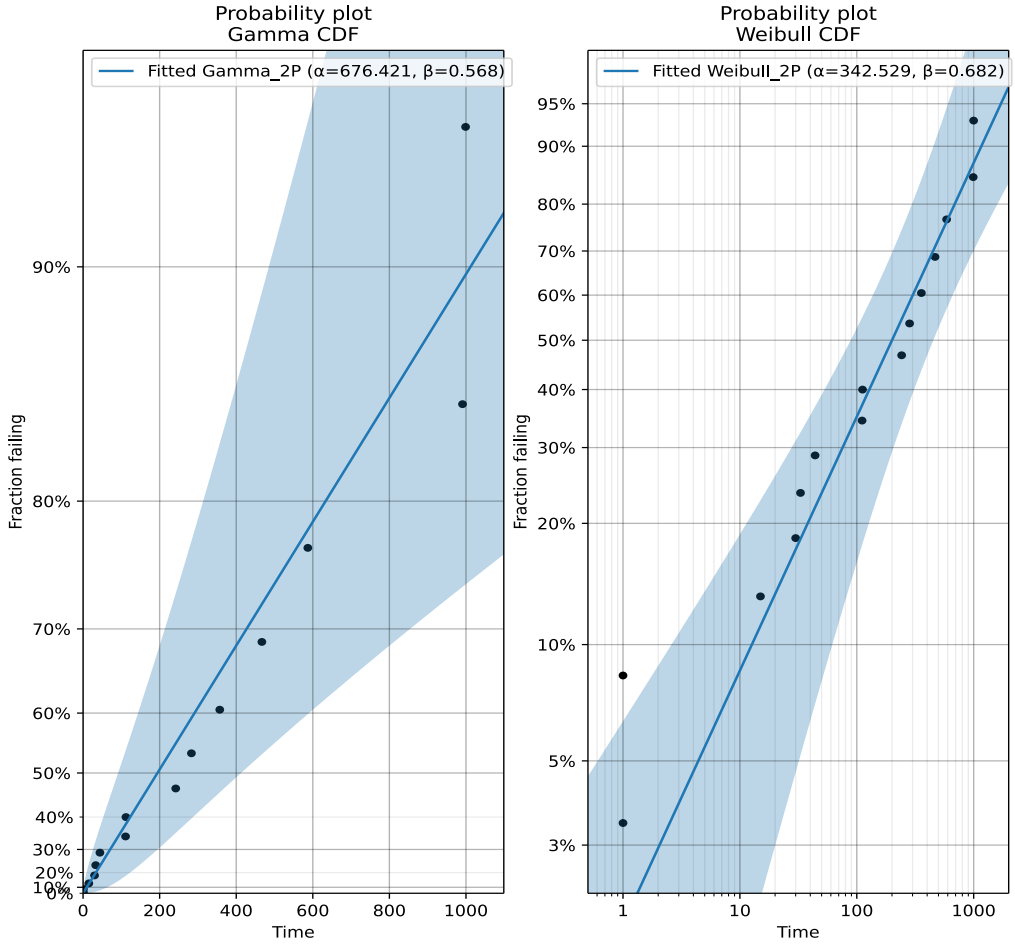


Obrázek 5: Porovnání parametrického odhadu Exponenciálního rozdělení s neparametrickými odhady.

3.2 Pacienti - placebo

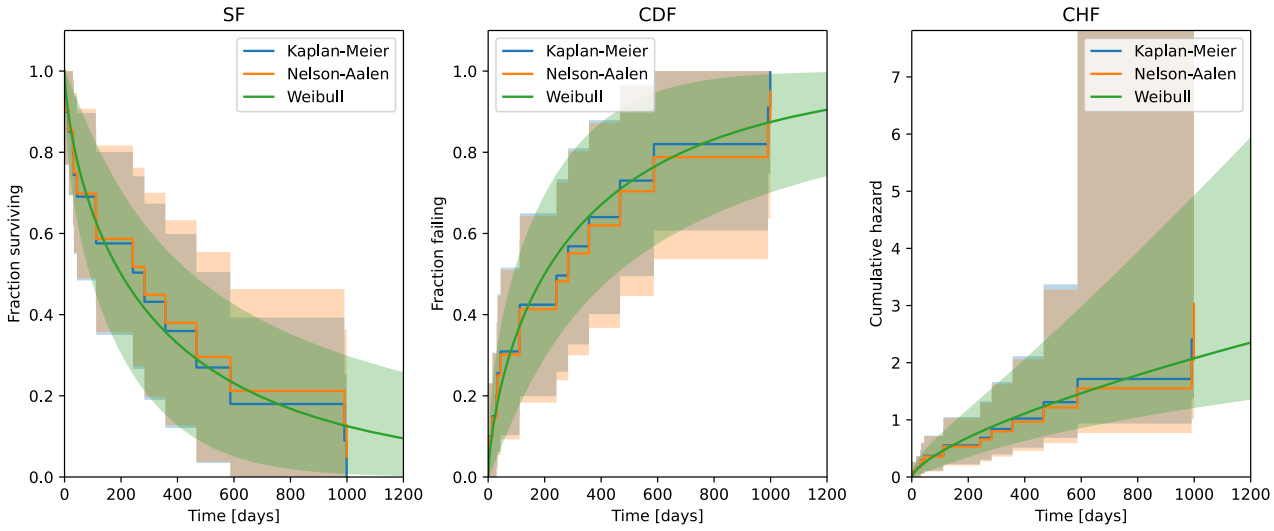
Pro pacienty dostávající placebo byla také použita funkce `Fit_Everything` a závěrem byly po vizuální inspekci QQ a PP plotů zvolena dvě rozdělení jako kandidáti: Weibullovo rozdělení se dvěma parametry a Gamma rozdělení se dvěma parametry.

Na Obr. 6 jsou vidět Probability ploty pro obě rozdělení. Je vidět, že Gamma rozdělení má pro zvyšující se čas dožití širší a širší konfidenční intervaly, oproti tomu Weibullovo rozdělení nikoli.

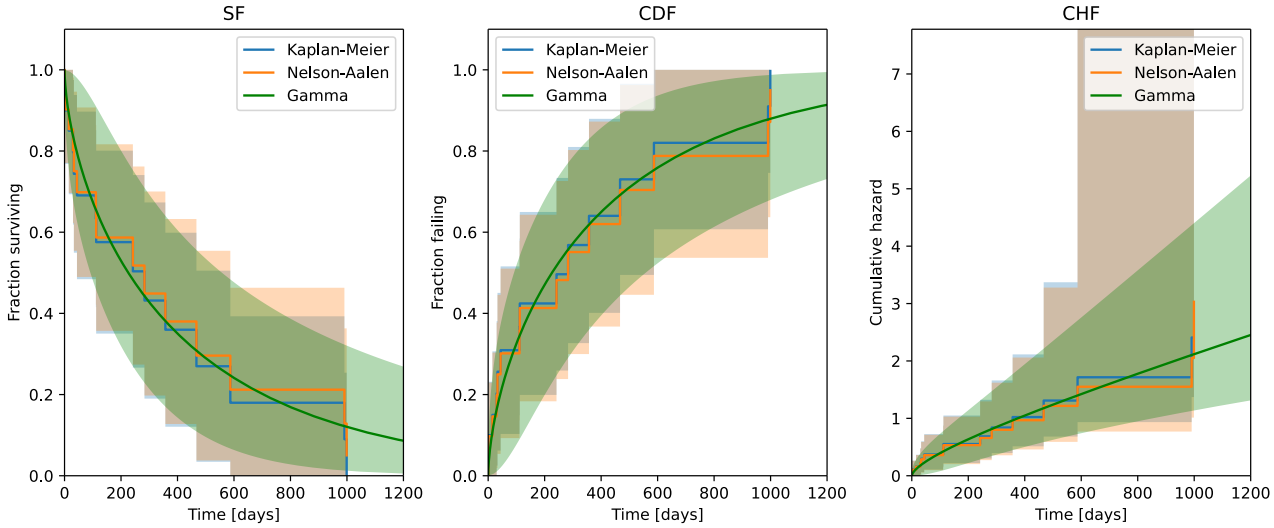


Obrázek 6: Probability plot pro Weibullovo a Gamma rozdělení pro skupinu pacientů 2.

Opět pro obě rozdělení porovnáme s neparametrickými odhady K-M a N-A. Výsledky pro Weibullovo rozdělení jsou na Obr. 7, pro Gamma rozdělení pak na Obr. 8. Z porovnání se zdá, že Gamma rozdělení lépe aproximuje naměřená data, proto volíme Gamma model jako parametrický model pro pacienty dostávající placebo.



Obrázek 7: Porovnání parametrického odhadu Weibullova rozdělení s neparametrickými odhady.



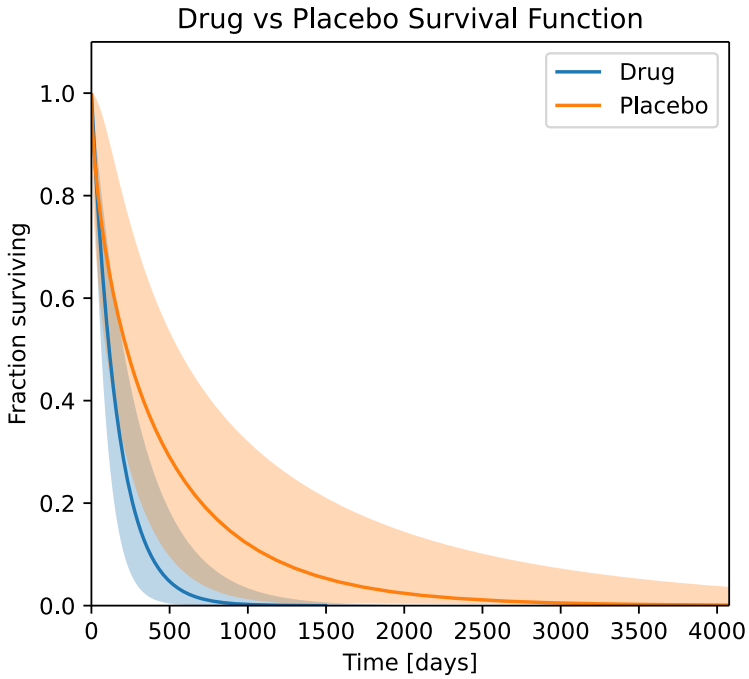
Obrázek 8: Porovnání parametrického odhadu Gamma rozdělení s neparametrickými odhady.

4 Porovnání podskupin lék vs placebo

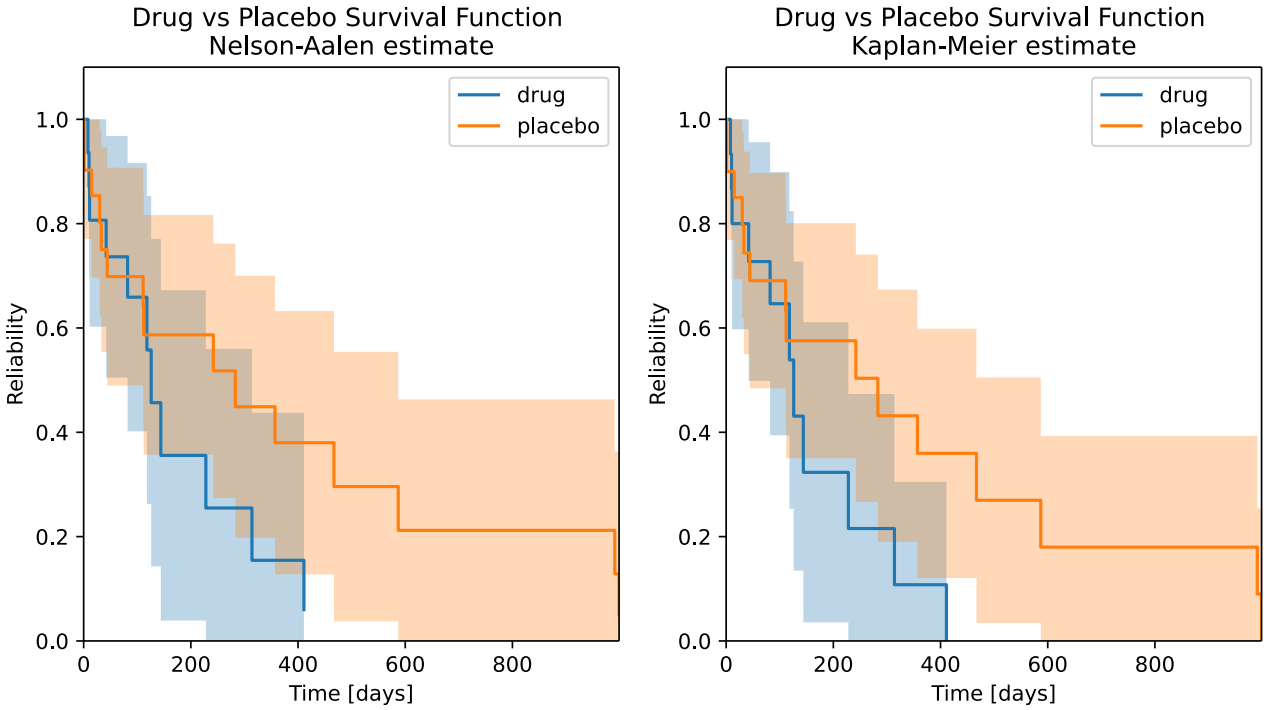
Teď, když máme pro obě skupiny pacientů vybrané parametrické modely i neparametrické odhady, můžeme se podívat na porovnání obou skupin.

4.1 Survival function

Nejdříve se podíváme na intenzitu poruch, resp. tedy na survival function. Na Obr. 9 vidíme porovnání pro parametrické modely, na Obr. 10 pak porovnání pro K-M a N-A odhady. Jak vidíme ze všech tří grafů, funkce klesá rychleji pro pacienty, kteří dostávali lék.



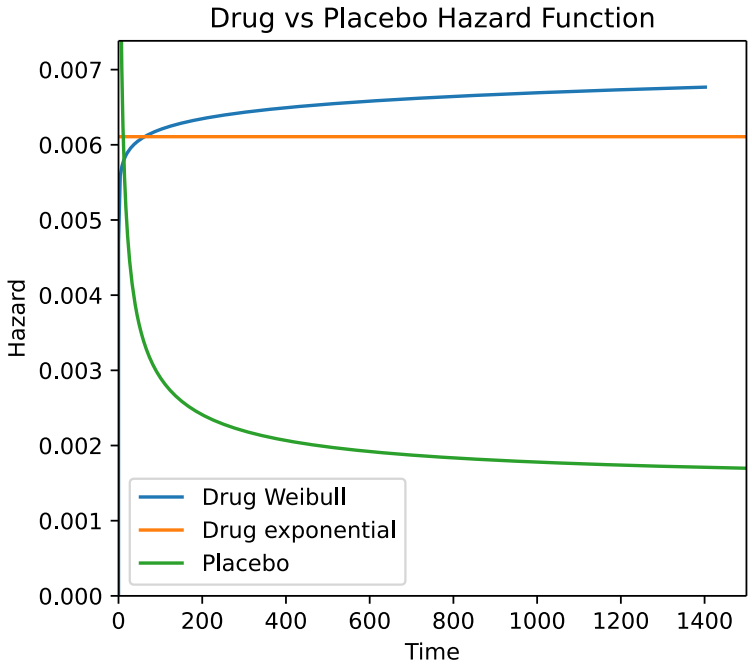
Obrázek 9: Průběh survival funkce pro parametrické modely.



Obrázek 10: Průběh survival funkce pro neparametrické odhady.

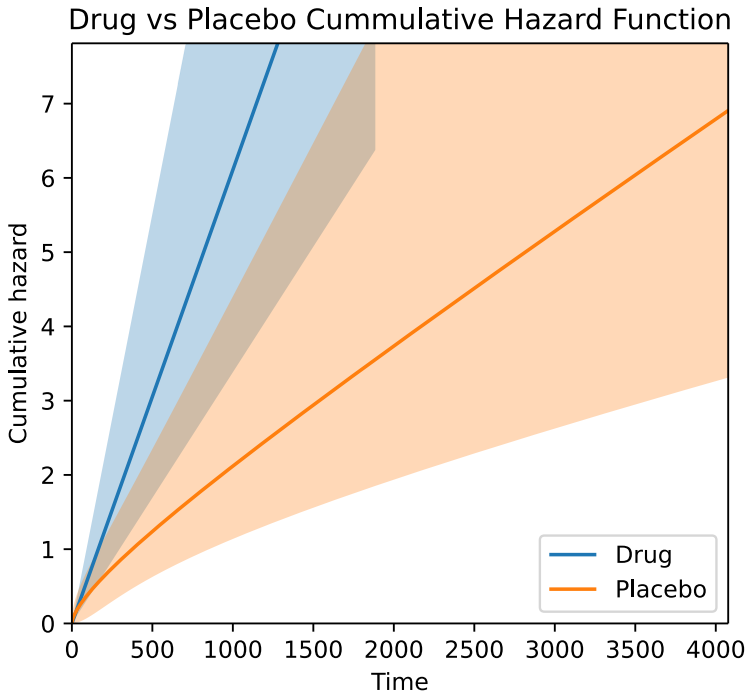
4.2 Hazard rate & cummulative hazard rate

Teď už se můžeme podívat na průběh intenziry poruch (hazard rate) a kumulativní intenzitu poruch. Klasickou intenzitu poruch můžeme získat pouze pro parametrické modely, výsledek je vidět na Obr. 11. Pro zajímavost zde je uvedená i intenzita poruch pro pacienty léčené lékem a Weibullovo rozdělení. Jak vidíme z obrázku, intenzita poruch pro pacienty dostávající placebo je klesající, v limitě se ustálí na hodnotě > 0. Naopak pro pacienty, kterým byl podáván lék je intenzita poruch konstantní (pro zvolené Exponenciální rozdělení), případně lehce rostoucí (pokud by bylo zvoleno Weibullovo rozdělení). Je zde tedy jasné, že obě skupiny se výrazně liší.

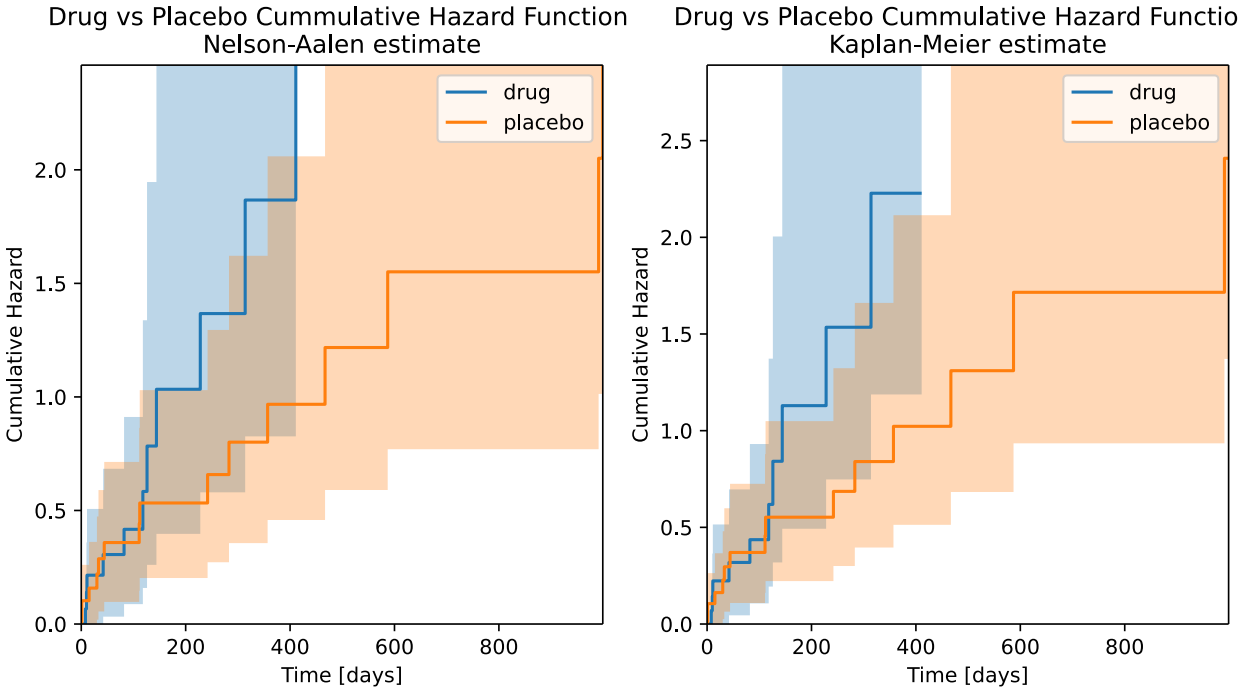


Obrázek 11: Průběh intenzity poruch (hazard function) pro parametrické modely.

Co se týče kumulativní intenzity poruch, můžeme vidět na Obr. 12 parametrické křivky a to, že pro pacienty dostávající lék roste funkce výrazně rychleji než pro pacienty dostávající placebo. Stejná situace je pro neparametrické odhady na Obr. 13.



Obrázek 12: Průběh kumulativní intenzity poruch (cummulative hazard function) pro parametrické modely.



Obrázek 13: Průběh kumulativní intenzity poruch (cummulative hazard function) pro neparametrické modely.

4.3 Numerické charakteristiky

Nakonec se můžeme podívat na další charakteristiky jako jsou MTTF nebo mediánová doba přežití. Všechny získané hodnoty jsou uvedené v tabulce 1. Můžeme zde přehledně vidět, že léčení pacienti mají výrazně nižší střední dobu dožití, stejně jako medián. I průměrná doba dožití při již dožitých 90 dnech je výrazně nižší než pro pacienty, kteří dostávali placebo.

Při porovnání parametrických a neparametrických metod můžeme vidět, že pro pacienty dostávající lék vycházejí získané hodnoty přibližně stejně. Pro pacienty, kteří dostávali placebo, se odhady poněkud liší, neparametrické modely například odhadují vyšší mediánovou dobu dožití.

Opět je zde jasně vidět, že skupiny pacientů se liší. Střední doba dožití i medián jsou výrazně vyšší pro pacienty, kteří brali placebo. Významný rozdíl vidíme i s MRL po čase 90 dní.

model	MTTF	t_{med}	MRL $t_o = 90$ dní
lék			
Exponenciální	163.7	113.5	163.7
Weibull	163.2	114.0	162.0
Kaplan-Meier	156.1	126.0	
Nelson-Aalen	168.4	126.0	
placebo			
Gamma	384.4	194.3	476.5
Weibull	444.6	200.2	557.8
Kaplan-Meier	356.6	283.0	
Nelson-Aalen	379.3	283.0	

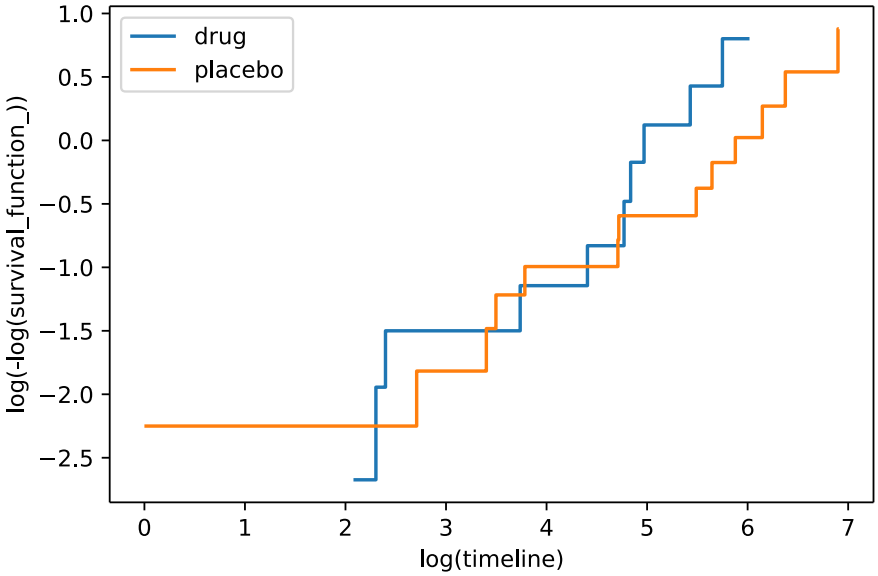
Tabulka 1: Tabulka charakteristik pro parametrické i neparametrické modely pro obě skupiny pacientů (lék vs placebo).

5 Coxův regresní model

Posledním úkolem je podívat se na skupiny pacientů a odhadnout, jestli je možné použít Coxův model.

5.1 Test předpokladů

Jako první bychom se chtěli podívat na log-log $\widehat{R}_{KM}$ plot, abychom zjistili, jestli jsou linie pro skupiny pacientů přibližně rovnoběžné a je tedy splněna podmínka $R_{lek} = R^{\gamma}_{placebo}$. Log-Log plot můžeme vidět na Obr. 14. Z obrázku je vidět, že grafy nejsou rovnoběžné, dokonce se ve středu kříží. Rovnoběžnost nastává až pro vyšší časy dožití, což je až zhruba poslední třetina (čtvrtina) grafu.



Obrázek 14: Log-Log Kaplan-Meier plot pro skupiny pacientů lék vs placebo.

Můžeme také provést Log-Rank test, který testuje hypotézu $H_0 : R_{lek} = R_{placebo}$. Pokud použijeme funkci `logrank_test` z balíku `lifelines`, dostaneme výsledek uvedený v tabulce 2. Vidíme, že test nezamítá nulovou hypotézu na hladině $\alpha = 0.05$ (ani na hladině $\alpha = 0.1$).

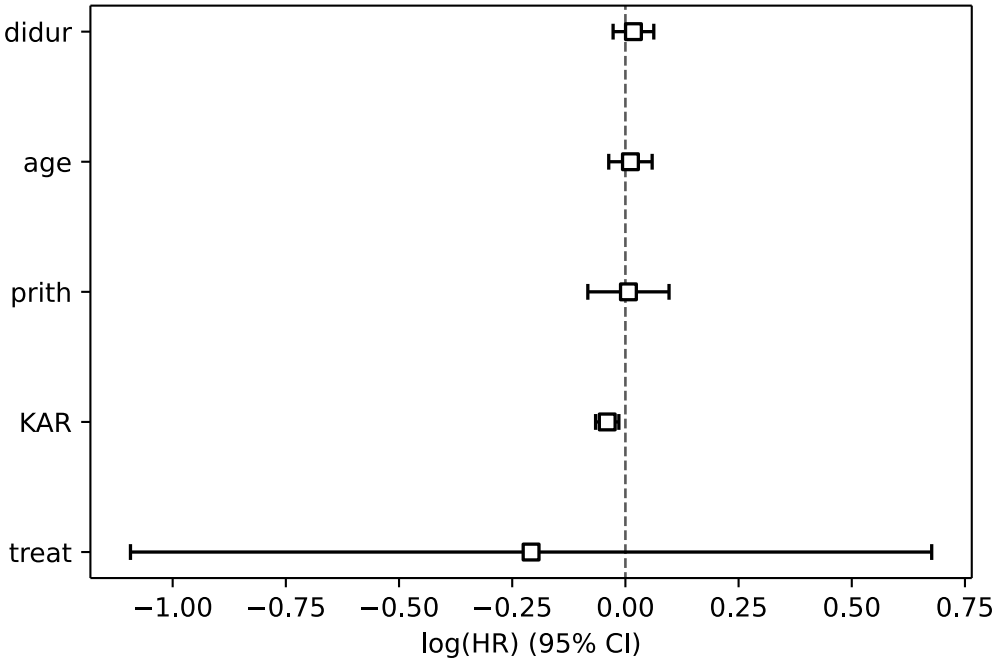
	test_statistic	p	-log2(p)
0	2.40839	0.120687	3.05066

Tabulka 2: Výslek log-rank testu pro data lék vs placebo.

Přesto se však můžeme pokusit naladit Coxův model a podívat se, jestli další vysvětlující proměnné budou mít vliv na výsledek.

5.2 Coxův model

Naše data obsahují momentálně 7 proměnných – treat (lék vs placebo), survival time, censored, KAR, didur, age a prith. Do Coxova modelu bude vstupovat survival time jako vysvětlovaná proměnná, používáme censorovaná data.



Obrázek 15: Log-Log Kaplan-Meier plot pro skupiny pacientů lék vs placebo.

Coxův model byl nafitován bez problémů, funkce vrací i indikaci, jestli jsou hazard ratios v pořádku, podle modelu jsou předpoklady OK. Podívejme se na výsledky přehledně v tabulce 3. Pro vizuálně lepší představu o parametrech se můžeme podívat na obrázek , kde můžeme vidět efekt jednotlivých parametrů na model. Jak vidíme, proměnná treat (lék vs placebo) je výrazněji posunutá od nuly, ovšem její konfidenční interval je velmi široký a i proto je vysoká p-hodnota pro tento parametr. Nízkou p-hodnotu vidíme pro parametr KAR, který by se mohl zdát významnější proměnnou v modelu. Zbylé proměnné se zdají nevýznamné pro naši skupinu pacientů.

	β_i	e^{β_i}	std(β_i)	β_{lw} 95%	β_{up} 95%	$e^{\beta_{lw}}$ 95%	$e^{\beta_{up}}$ 95%	z	p	$-\log 2(p)$
treat	-0.21	0.81	0.45	-1.09	0.68	0.34	1.97	-0.46	0.64	0.63
KAR	-0.04	0.96	0.01	-0.07	-0.01	0.94	0.99	-3.06	<0.005	8.82
didur	0.02	1.02	0.02	-0.03	0.06	0.97	1.06	0.77	0.44	1.19
age	0.01	1.01	0.02	-0.04	0.06	0.96	1.06	0.46	0.65	0.62
prith	0.01	1.01	0.05	-0.08	0.10	0.92	1.10	0.15	0.88	0.18

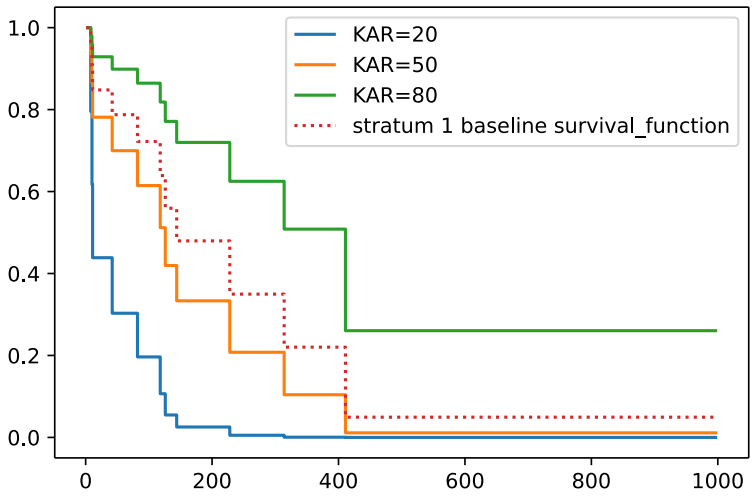
Tabulka 3: Výsledná tabulka pro koeficienty Coxova modelu.

Vyzkoušíme tedy ještě jeden model – budeme modelovat cenzorovaný survival time pomocí spojité proměnné KAR, proměnnou treat ponecháme v modelu jako *stratification variable*, tedy proměnnou, která nesplňuje podmínky proporčních rizik. To znamená, že budou naladěny dva modely, jeden pro pacienty léčené a jeden pro pacienty dostávající placebo. V tabulce 4 vidíme výsledek.

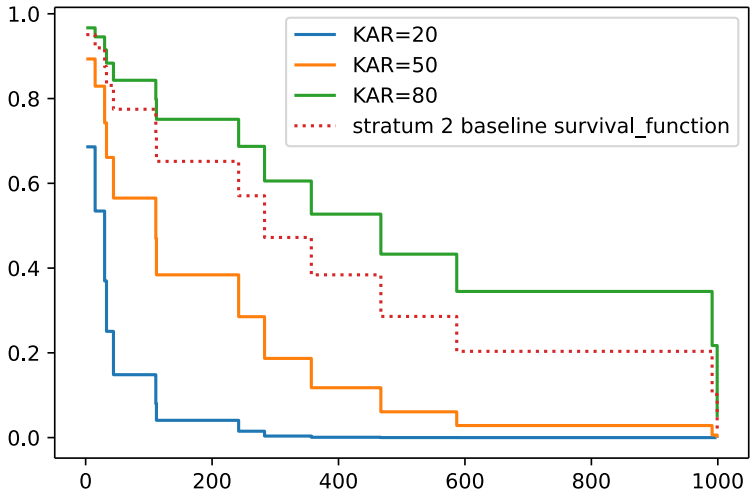
Můžeme se pak podívat na průběh survival funkce pro různé hodnoty Karnoffskyho skóre. Na Obr. 16 vidíme výsledek pro léčené pacienty, na Obr. 17 pak výsledek pro pacienty dostávající placebo. Porovnáme-li *baseline*, tedy survival funkci podle proměnné treat, můžeme si všimnout, že pro placebo klesá funkce pomaleji.

	β_i	e^{β_i}	std(β_i)	β_{lw} 95%	β_{up} 95%	$e^{\beta_{lw}}$ 95%	$e^{\beta_{up}}$ 95%	z	p	$-\log 2(p)$
KAR	-0.04	0.96	0.01	-0.07	-0.02	0.94	0.99	-3.13	<0.005	9.16

Tabulka 4: Výsledná tabulka pro koeficient proměnné KAR Coxova modelu.



Obrázek 16: Plot survival funkcí pro různé hodnoty Karnoffskyho skóre pro proměnnou treat = 1 (lék).



Obrázek 17: Plot survival funkcí pro různé hodnoty Karnoffskyho skóre pro proměnnou treat = 2 (placebo).

6 Závěr

V protokolu byla sledována skupina pacientů v klinickém experimentu. Jedna skupina dostávala lék, druhá pouze placebo. Cílem bylo porovnat, jestli existuje rozdíl mezi těmito dvěma skupinami v čase dožití.

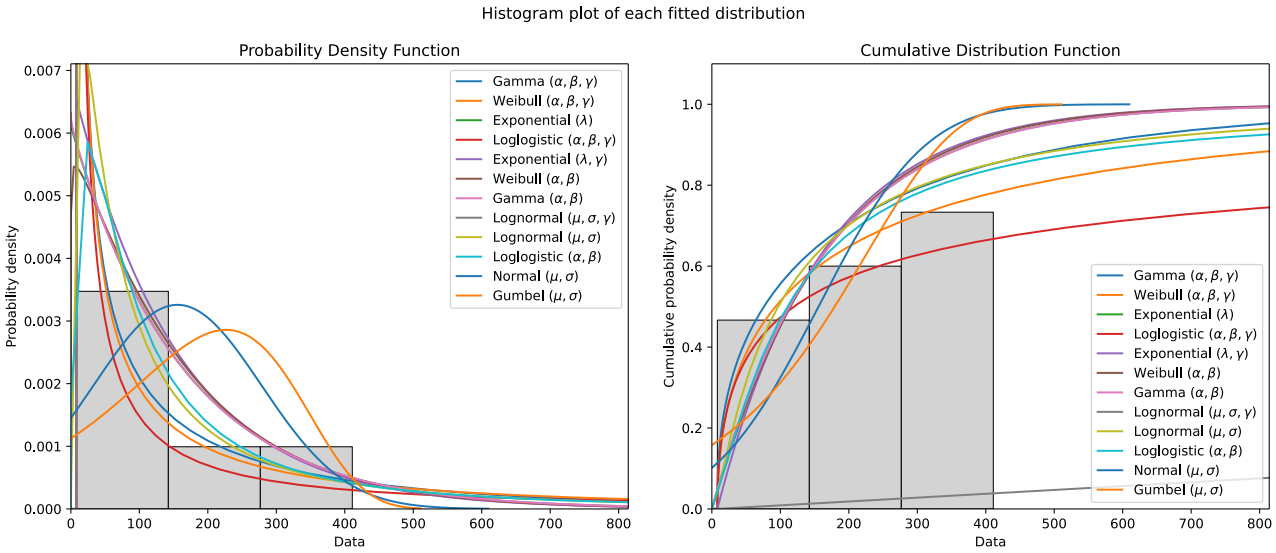
Pro obě skupiny byly vytvořeny jak parametrické, tak neparametrické modely. Bylo zjištěno, že vizuálně se průběhy rizikových funkcí pro obě skupiny liší a že skupina pacientů, kteří dostávali placebo, měla vyšší dobu dožití než skupina dostávající pravý lék.

Nakonec byl vytvořen i Coxův model proporčních rizik, bylo ovšem zjištěno, že obecně podle log-log plotu to nevypadá, že by zde byla proporční závislost. Ukázalo se, že v regresním modelu je významná pouze proměnná KAR, u které pak můžeme vidět výrazné rozdíly v průběhu survival funkce mezi různými hodnotami, obecně platí, že vyšší hodnota Karnofského skóre má za následek vyšší dobu dožití.

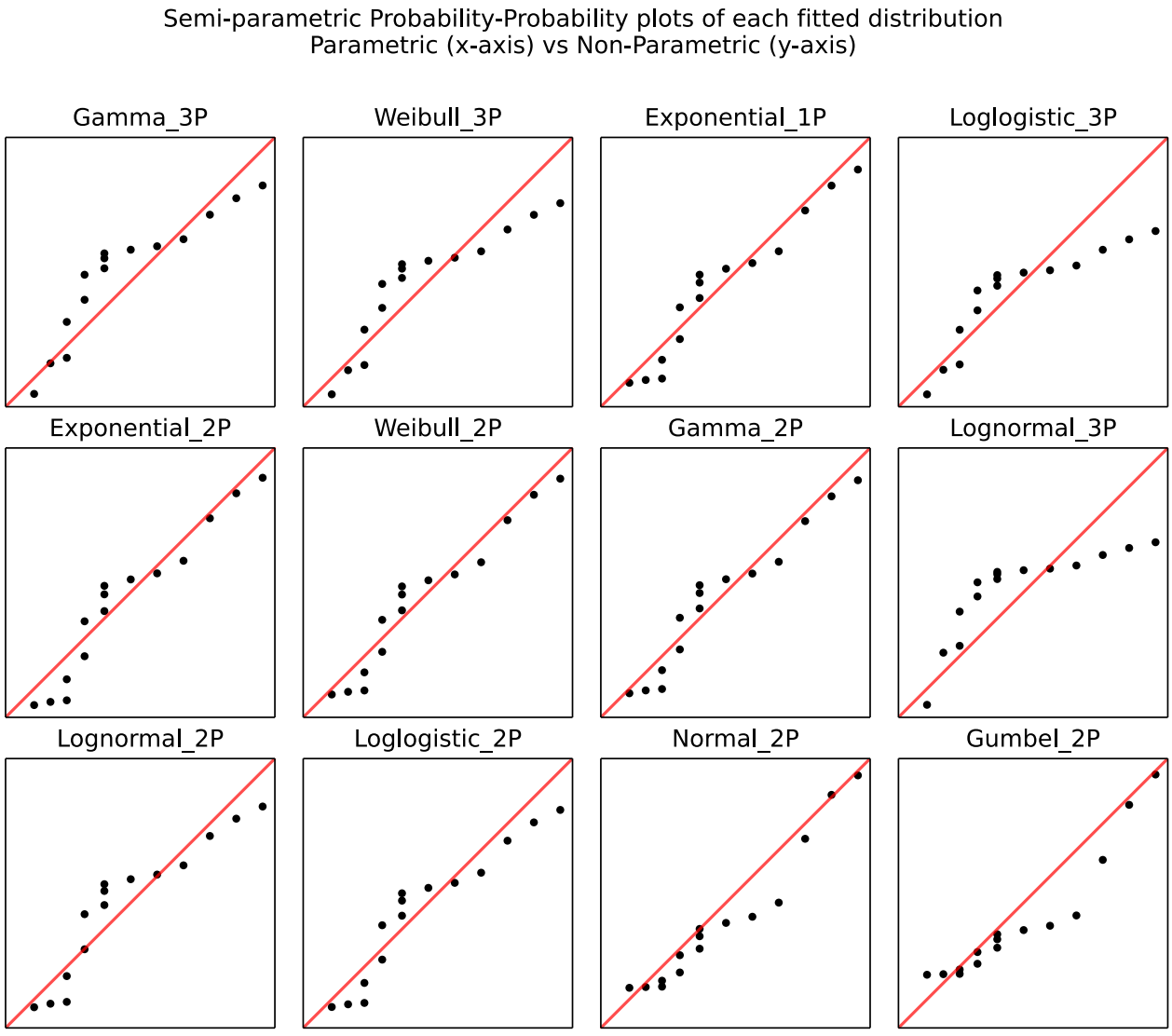
Celkový závěr je tak následující: Nebylo prokázáno, že by měl lék pozitivní dopad na čas dožití pacientů ve skupině `cell = squamous`. Dopad se zdá z naměřených dat spíše negativní.

7 Příloha

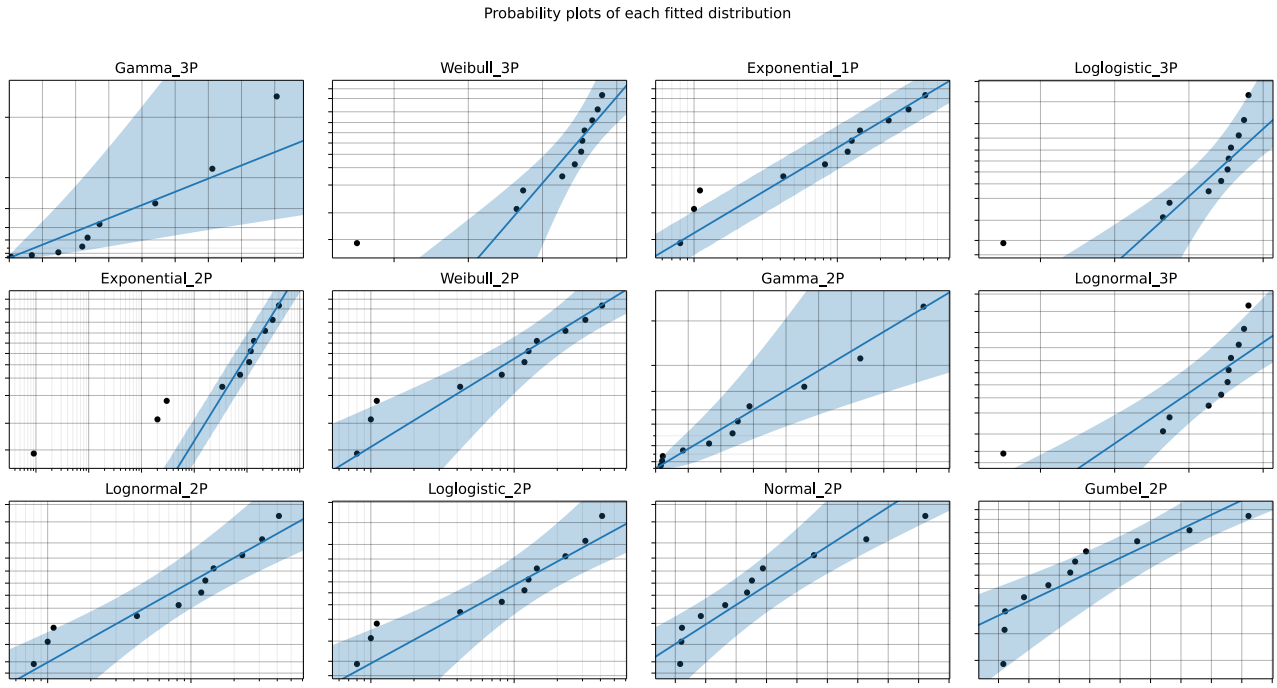
V příloze máme výsledné ploty z funkce `Fit_Everything`. Pro každou skupinu máme nejdříve obrázek histogramu hodnot společně s nafitovanými hustotami pravděpodobnosti a distribučními funkcemi pro různá rozdělení. Následují PP ploty a nakonec probability ploty.



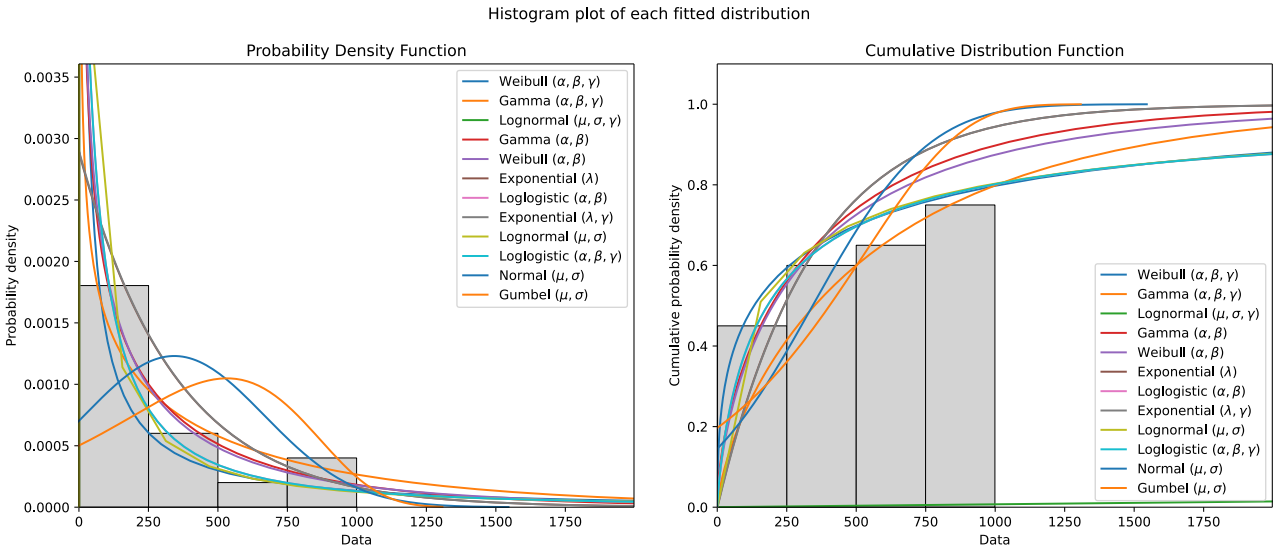
Obrázek 18: Histogram + PDF a CDF pro různá rozdělení pro léčenou skupinu pacientů.



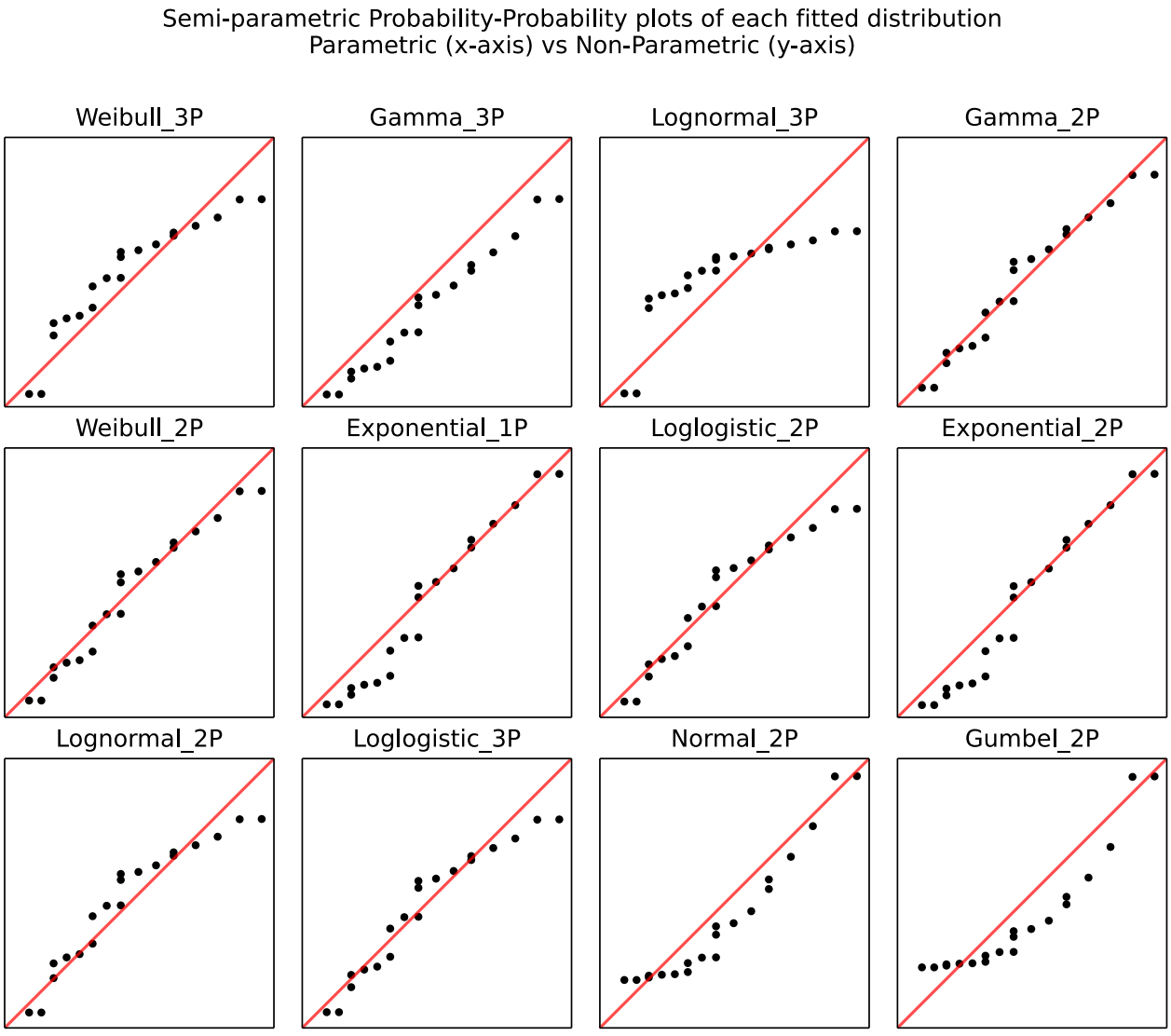
Obrázek 19: PP ploty pro různá rozdělení pro léčenou skupinu pacientů.



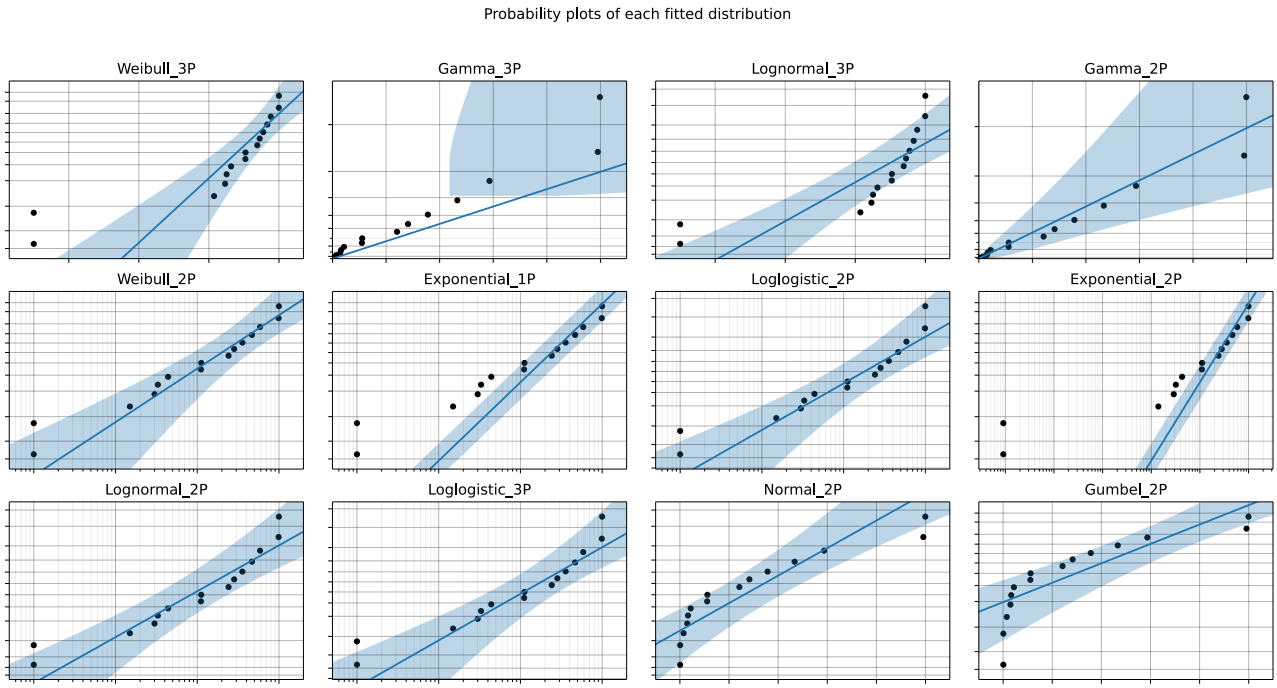
Obrázek 20: Probability ploty pro různá rozdělení pro léčenou skupinu pacientů.



Obrázek 21: Histogram + PDF a CDF pro různá rozdělení pro skupinu pacientů dostávající placebo.



Obrázek 22: PP ploty pro různá rozdělení pro skupinu pacientů dostávající placebo.



Obrázek 23: Probability ploty pro různá rozdělení pro skupinu pacientů dostávající placebo.