

Základy analýzy přežití

Tomáš Pavlík
Institut biostatistiky a analýz



Lékařská fakulta
Masarykova univerzita



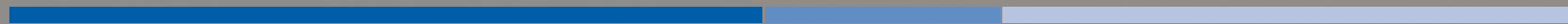
Přírodovědecká fakulta
Masarykova univerzita



MASARYK UNIVERSITY
INSTITUTE OF BIostatISTICS AND ANALYSES

www.iba.muni.cz

ZÁKLADNÍ POJMY ANALÝZY PŘEŽITÍ

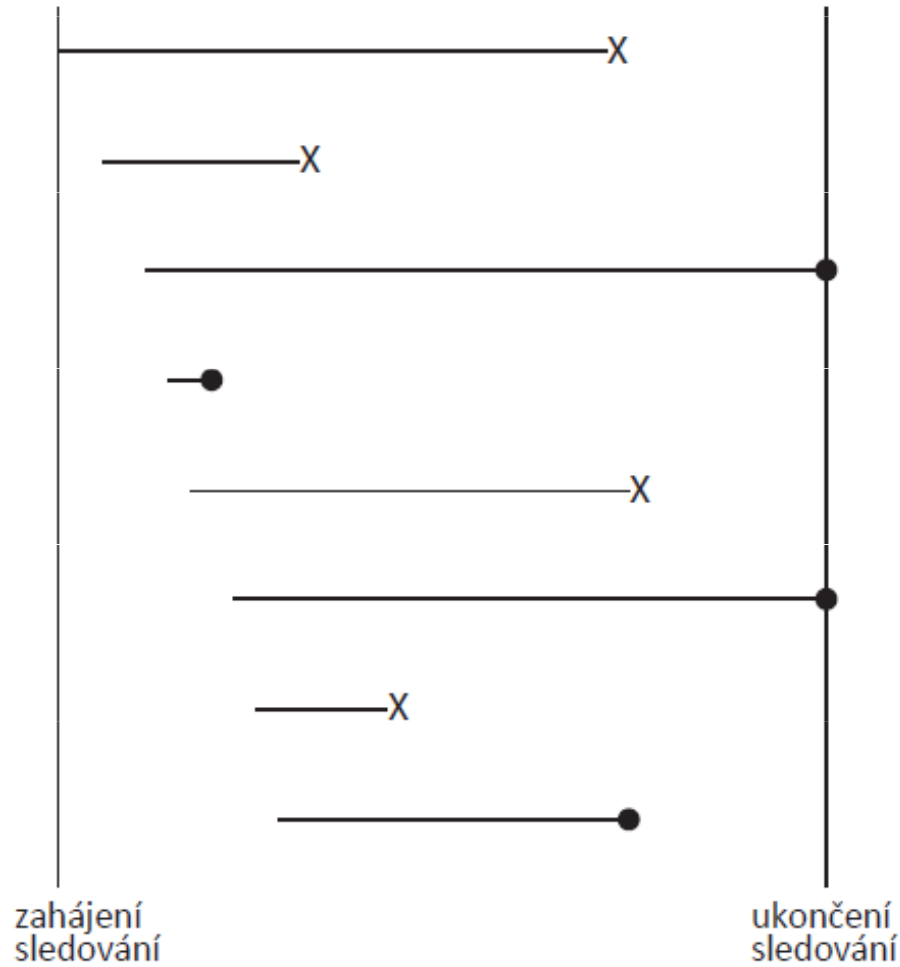


- Hlavní charakteristikou odlišující **data o přežití** (*survival data*) od ostatních typů dat, např. od klasického vnímání mortality jako podílu zemřelých pacientů v klinických aplikacích, je jejich **časová složka**.
- Data o přežití totiž odráží nejen informaci o počtu, respektive podílu sledovaných událostí, ale zároveň nás také informují, kdy k dané události došlo.
- Analýza přežití zahrnuje matematicko-statistické metody pro hodnocení **času do výskytu sledované události**.
- Klíčovým prvkem analýzy přežití je definice **sledované události** (*event of interest*). Ta musí být stanovena jednoznačně a měla by být také snadno pozorovatelná či zjistitelná.

- **Popisné metody** (*descriptive methods*). Cílem použití těchto metod je popsat časový průběh výskytu sledované události u daného souboru subjektů či objektů a identifikovat pravděpodobnosti přežití bez výskytu dané události v jednotlivých časových bodech.
- **Srovnávací metody** (*comparative methods*). Srovnávací metody jsou používány tehdy, chceme-li zjistit, zda se daný soubor subjektů či objektů liší ve výskytu sledované události od předpokládané hodnoty, případně zda se jednotlivé skupiny subjektů liší mezi sebou.
- **Modely přežití** (*survival models*). Použití stochastického modelování v analýze přežití nám pomáhá řešit otázku, zda a jakým způsobem závisí délka přežití subjektu nebo skupiny subjektů na jedné nebo více sledovaných proměnných, případně jestli se tato závislost nějak vyvíjí v čase. Jinými slovy, jde nám o identifikaci proměnných, které v případě sledované události ovlivňují pravděpodobnost, že tato nastane dříve či později.

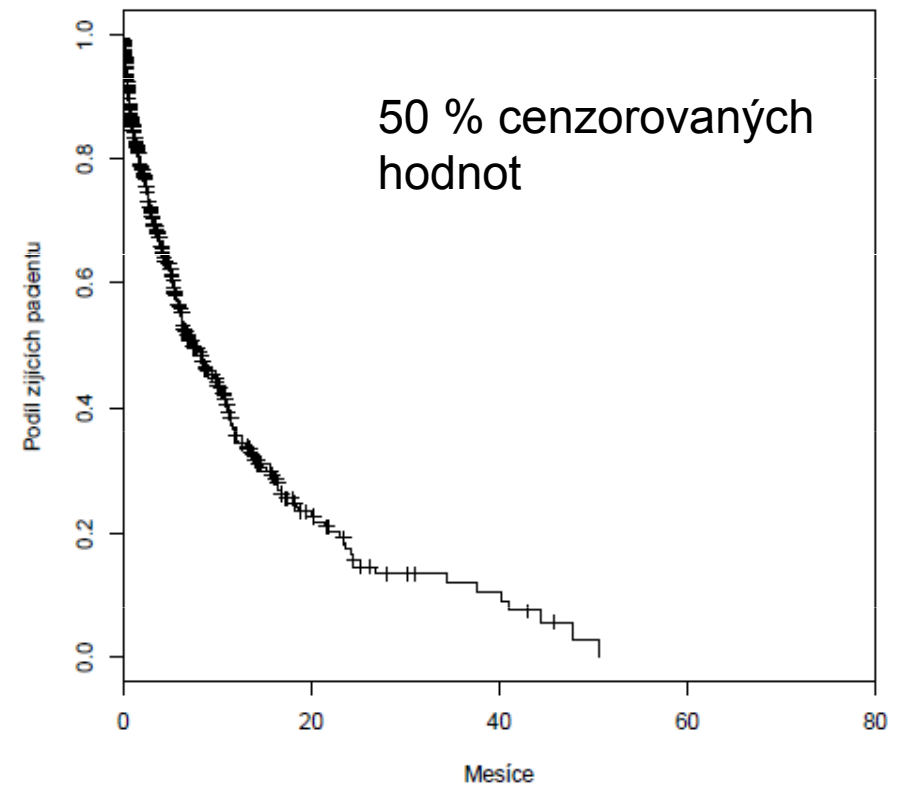
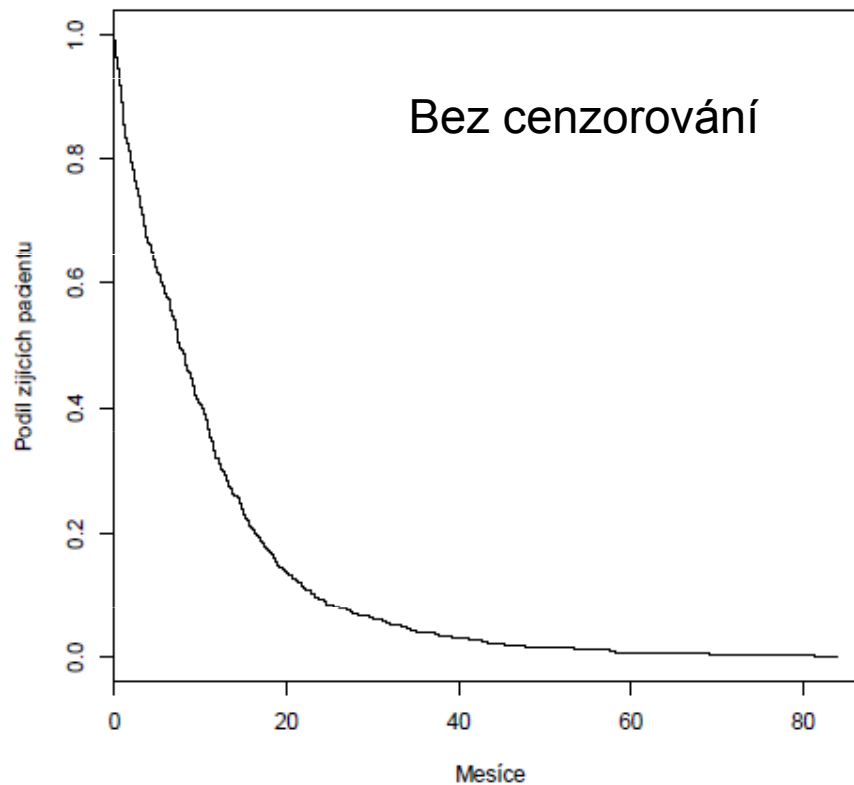
- Definovaná událost se nemusí v průběhu sledování vyskytnout u všech subjektů (pozorování není kompletní). Čas přežití subjektů, u nichž v průběhu sledování nenastala definovaná událost, označujeme jako **cenzorovaný** (*censored*).
- **Subjekty bez sledované události nelze v žádném případě z hodnocení vyloučit, neboť nepřítomnost události lze velmi často považovat za pozitivní ukazatel.**
- Příčiny cenzorování: **ukončení sledování** (např. uzavření databáze), **ztráta kontaktu s pacientem**, výskyt události, která výskyt námi definované události jednoznačně vylučuje: **kompetitivní událost** (*competing event*).
- **cenzorování zprava** (*right censoring*),
- **cenzorování zleva** (*left censoring*),
- **intervalové cenzorování** (*interval censoring*).

Délka sledování a výskyt události



● = cenzorování
X = výskyt události

- Procento cenzorovaných subjektů ve studii bývá měřítkem kvality sledování daného souboru, protože **výrazně ovlivňuje kvalitu odhadů přežití**.



- Cenzorování vnímáme jako ztrátu informace.
- Vysoké procento cenzorovaných časů přežití může vypovídat o **nedostatečné délce sledování**, kdy jsme nebyli schopni pozorovat dostatečné množství událostí, které hodnotíme.
- **Cenzorování tak má vliv na požadovanou velikost souboru** hodnocených subjektů.
- Cenzorované časy přežití nejsou rovnocenné s kompletními časy přežití a v přítomnosti cenzorování je nutné navýšit velikost souboru tak, abychom zajistili dostatečný počet kompletních časů přežití.
- Předpoklad nezávislosti cenzorování a výskytu sledované události, tedy tzv. **neinformativního cenzorování** (*non-informative censoring*).



MASARYK UNIVERSITY
INSTITUTE OF BIostatISTICS AND ANALYSES

www.iba.muni.cz

HLAVNÍ CHARAKTERISTIKY V ANALÝZE PŘEŽITÍ



- **Čas přežití** (*survival time*) neboli dobu do výskytu sledované události reprezentujeme nezápornou náhodnou veličinou T , která představuje buď skutečný čas přežití daného subjektu, nebo cenzorovaný čas přežití.
- **Funkce přežití** (*survival function*), označme ji $S(t)$, vyjadřuje pravděpodobnost, že se náhodná veličina T realizuje na reálné ose až za danou hodnotou t , což znamená, že čas přežití daného subjektu bude větší, než je zvolený čas t .
- Funkci přežití lze tedy zapsat jako:

$$S(t) = P(T > t) = 1 - P(T \leq t)$$

- **Riziková funkce** (*hazard function*) vyjadřuje intenzitu výskytu sledované události v čase t za podmínky, že subjekt přežil do času t :

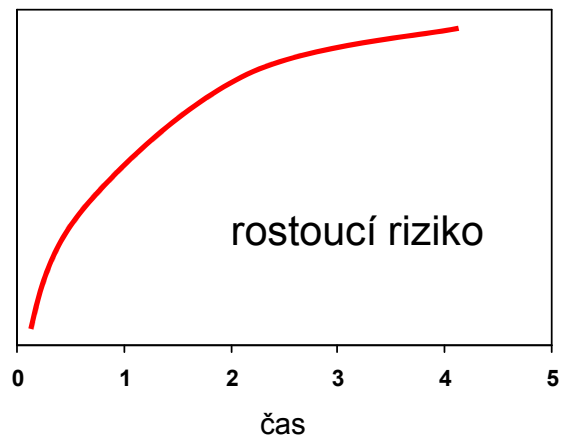
$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(t < T \leq t + \Delta t | T > t).$$

- **Kumulativní rizikovou funkci** (*cumulative hazard function, integrated hazard*) získáme integrací rizikové funkce podle času:

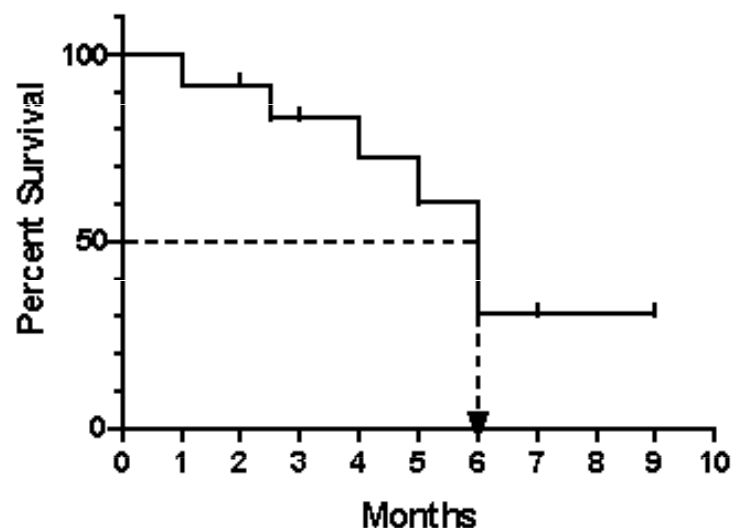
$$H(t) = \int_0^t h(x) dx.$$

- Kumulativní riziková funkce odpovídá celkovému riziku výskytu sledované události od začátku sledování až do času t . Vzhledem k tomu, že se jedná o riziko a nikoliv o pravděpodobnost, není funkce $H(t)$ na rozdíl od funkce přežití $S(t)$ shora omezena číslem 1.

Různý průběh rizikové funkce



- **Medián přežití** (*median survival time*) je definován jako čas, ve kterém má funkce přežití hodnotu 0,5, tedy jako čas pro který platí $S(t_{0,5}) = 0,5$.
V klinických studiích zaměřených na hodnocení přežití pacientů se výpočet mediánu přežití stal standardem, který je reportován jako hlavní výsledek.



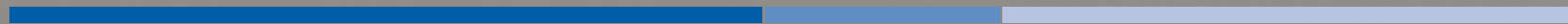
- Obdobně jako medián přežití jsme schopni definovat i další kvantily rozdělení pravděpodobnosti náhodné veličiny T . Čas, který bude $100 \cdot p$ -procentním kvantilem náhodné veličiny T , označme ho t_p , je definován jako čas, pro který platí, že $S(t_p) = 1 - p$.

- **Průměrná doba přežití** (*mean survival time*) představuje střední hodnotu náhodné veličiny T .
- Průměrná doba přežití je jednoduše definována jako integrál z funkce přežití $S(t)$ na intervalu od nuly do nekonečna, tedy jako **plocha pod křivkou přežití**.

$$E(T) = \int_0^{\infty} tf(t)dt = \int_0^{\infty} S(t)dt$$

- Aby byla střední hodnota náhodné veličiny T definovaná, předpokládáme, že pravděpodobnost přežití bez výskytu sledované události jde s rostoucím časem k nule. Tento předpoklad je logický u studií, kde je sledovanou událostí například úmrtí nebo jiná událost jednoznačně spjatá se sledovanými subjekty.

POZOR NA MEDIÁN PŘEŽITÍ

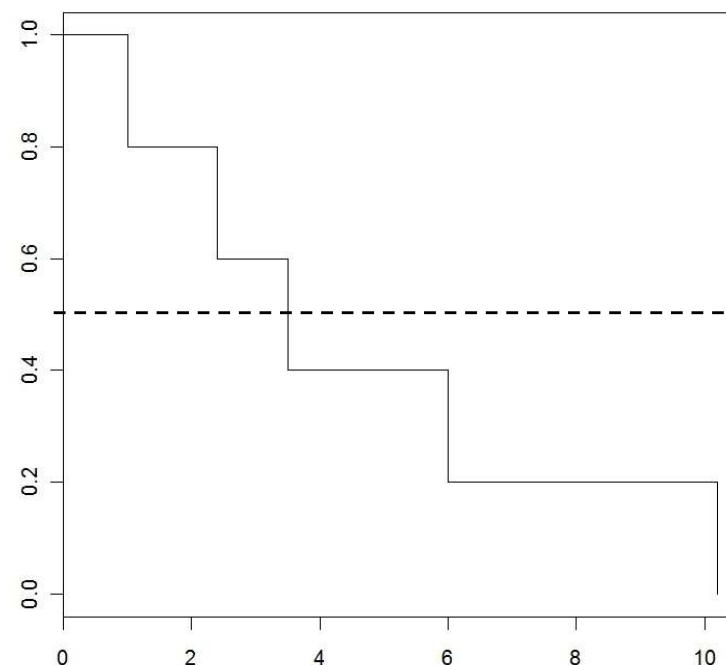


- Odhad mediánu přežití může být velmi ovlivněn cenzorováním časů přežití!
- **Přesnost odhadu mediánu přežití souvisí s délkou sledování pacientů, ale medián přežití a medián délky sledování nemusí být nutně stejné!**

Pacient	Čas sledování	Událost sledována
1	1,0 měsíce	Ano
2	2,4 měsíce	Ano
3	3,5 měsíce	Ano
4	6,0 měsíců	Ano
5	10,2 měsíců	Ano

**Medián délky sledování
= 3,5 měsíce**

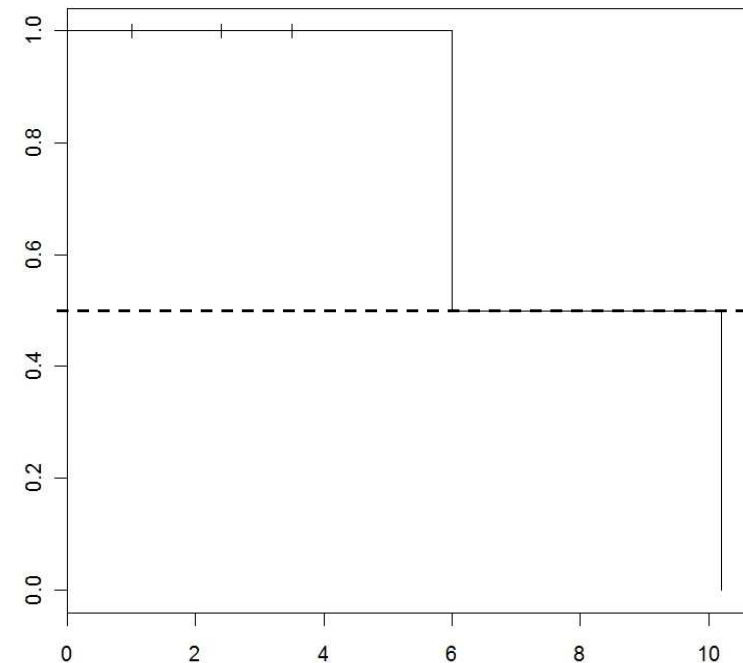
**Medián OS
= 3,5 měsíce**



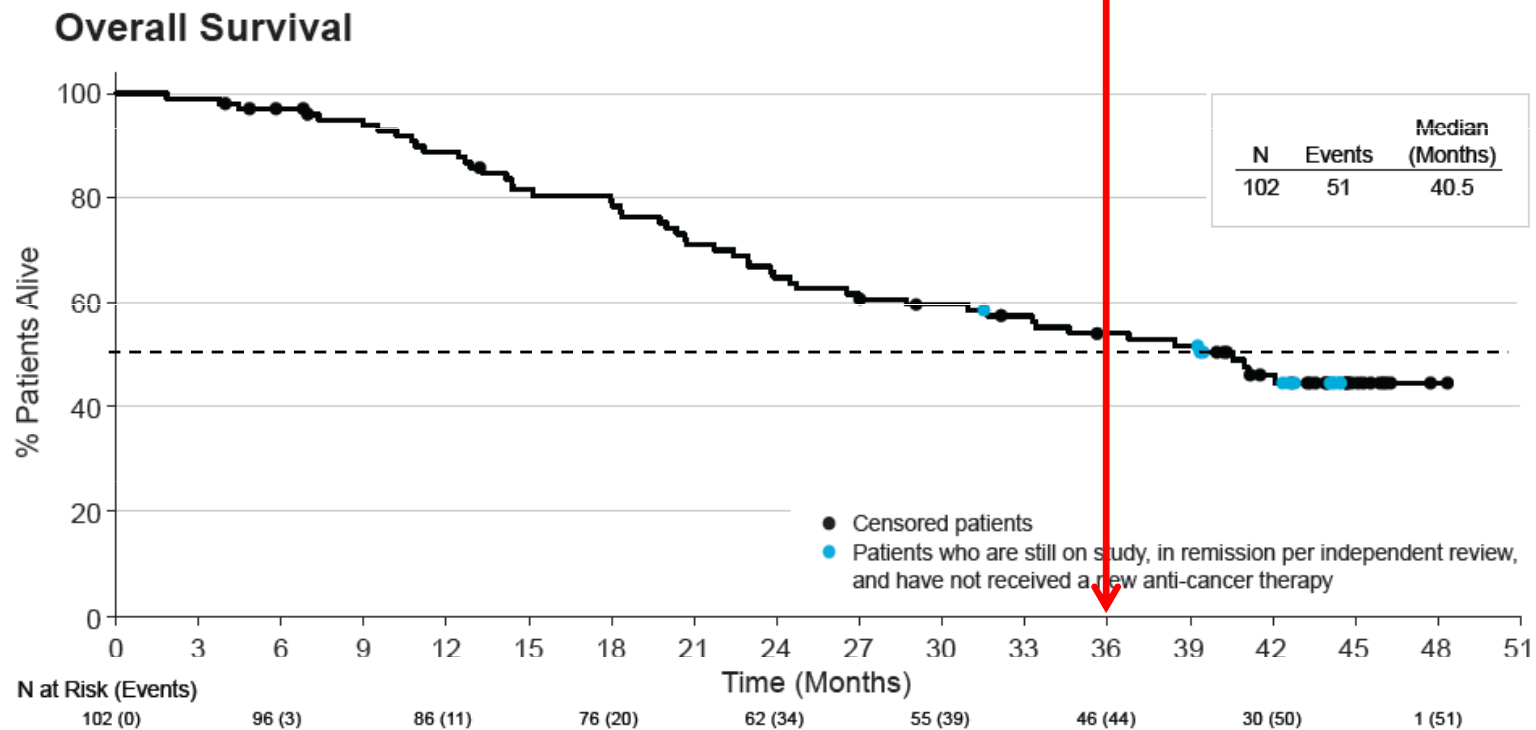
Pacient	Čas sledování	Událost sledována
1	1,0 měsíce	Ne (cenzorování)
2	2,4 měsíce	Ne (cenzorování)
3	3,5 měsíce	Ne (cenzorování)
4	6,0 měsíců	Ano
5	10,2 měsíců	Ano

**Medián délky sledování
= 3,5 měsíce**

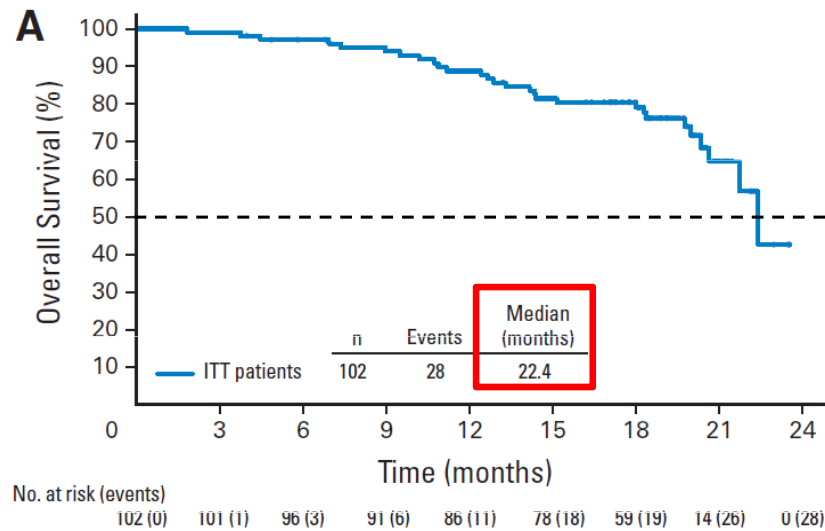
**Medián OS
= 6,0 měsíců**



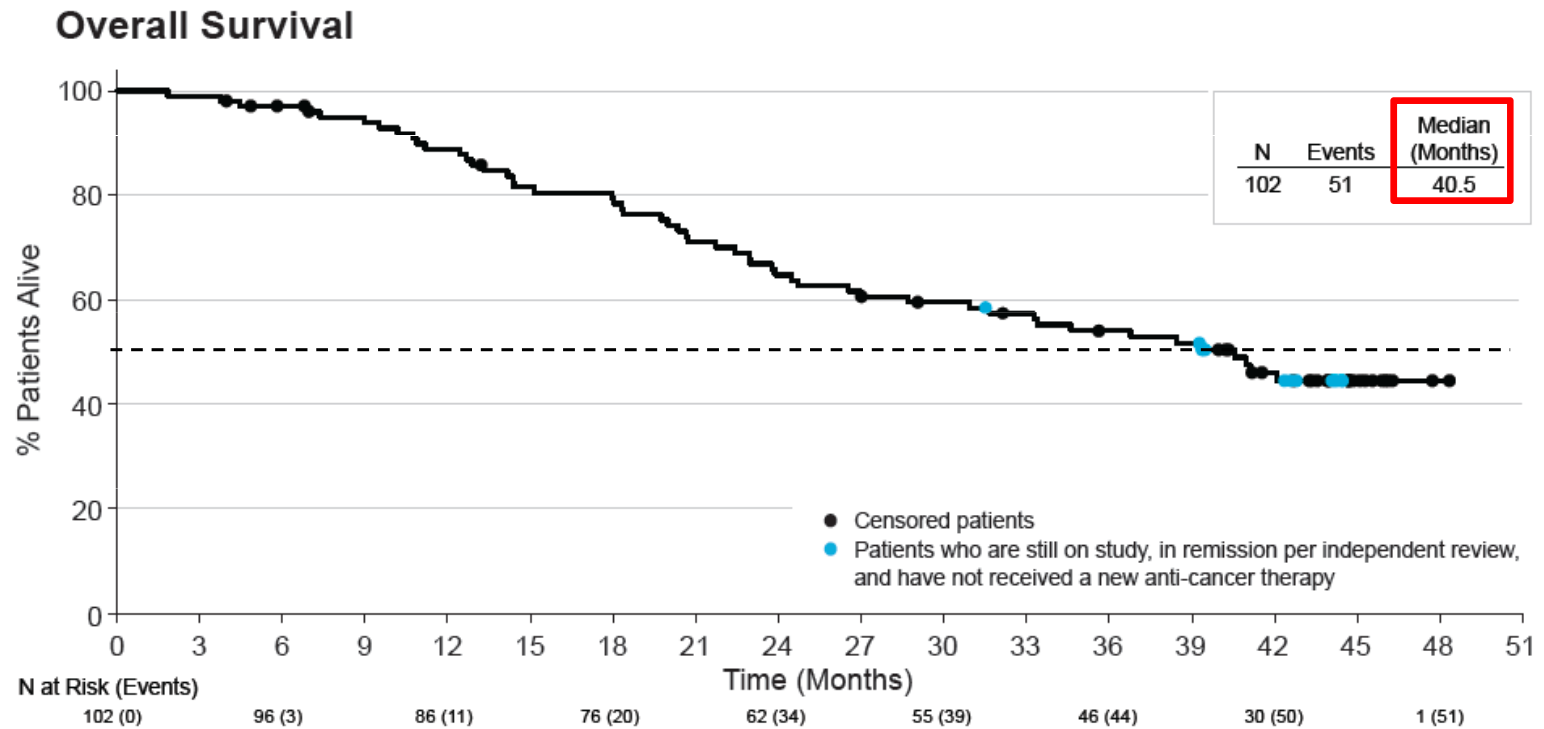
- Je tedy možné, že medián délky sledování byl 32,7 měsíce a zároveň 3leté OS = 0,54?



Výsledky 2012

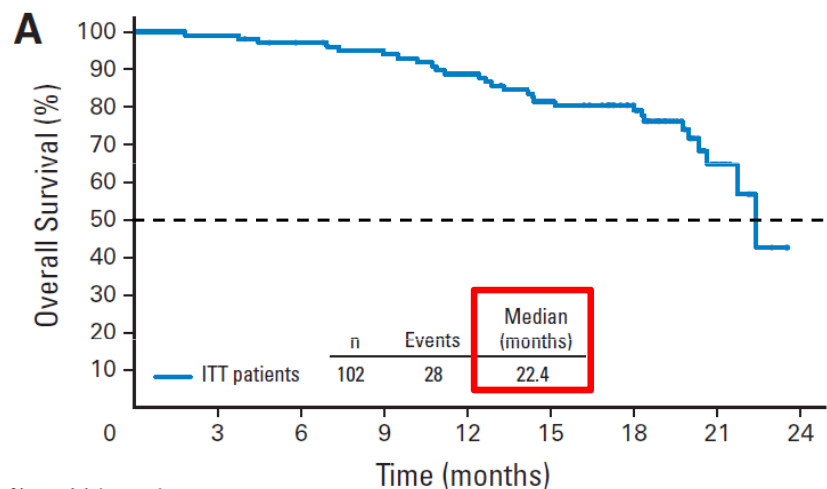


Výsledky 2013



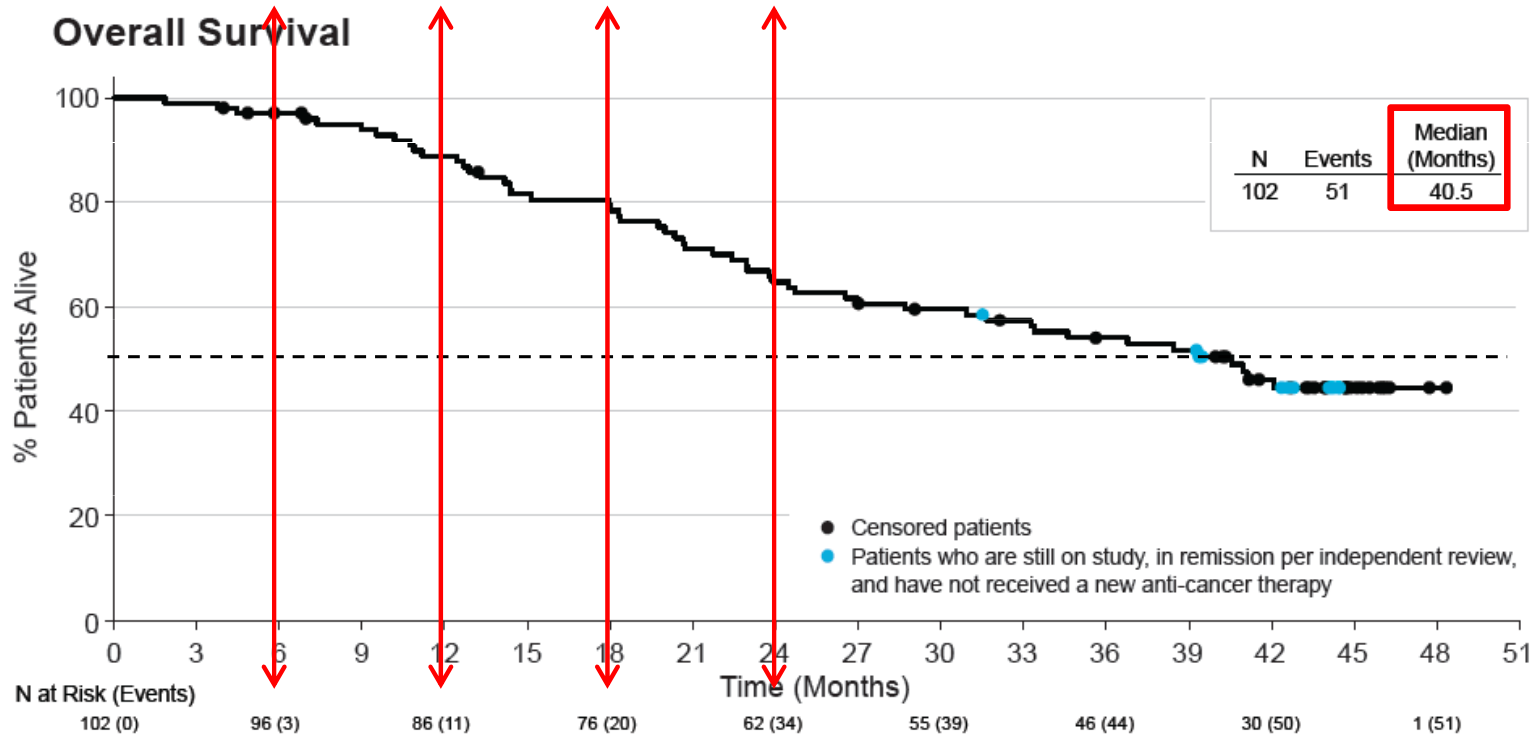
- I když medián OS a medián délky sledování nemusí být nutně stejné, medián OS je délkou sledování jednotlivých subjektů jednoznačně ovlivněn.
- **Kvalita odhadu OS dále úzce souvisí s počtem pacientů ve sledování (tzv. v riziku sledované události) v čase!**

Výsledky 2012

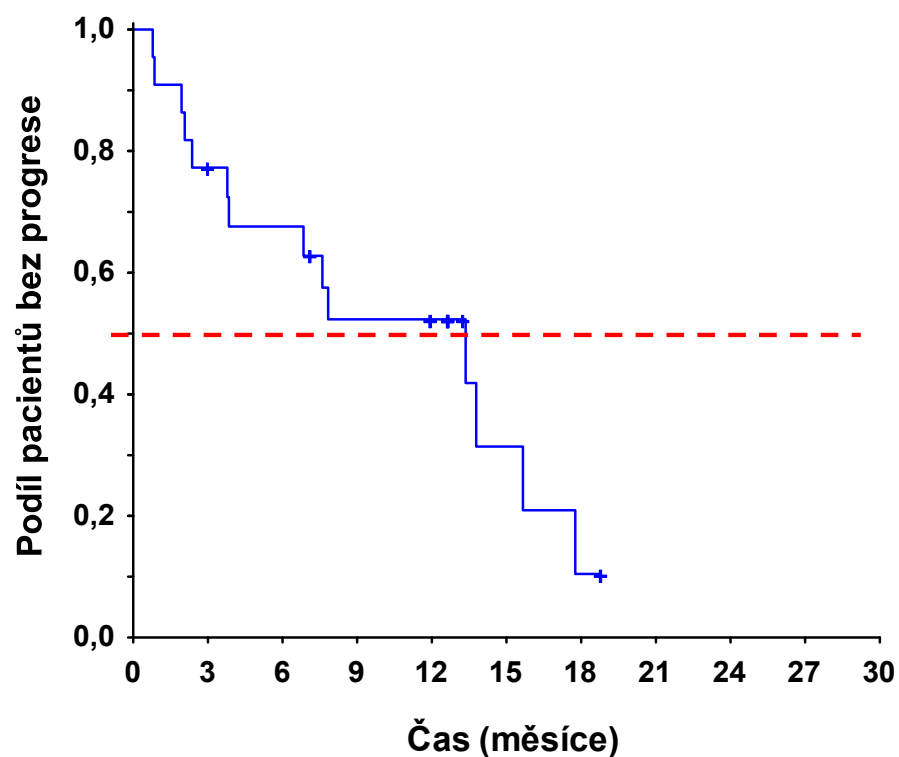


No. at risk (events)
102 (0) 101 (1) 96 (3) 91 (6) 86 (11) 78 (18) 59 (19) 14 (26) 0 (28)

Výsledky 2013

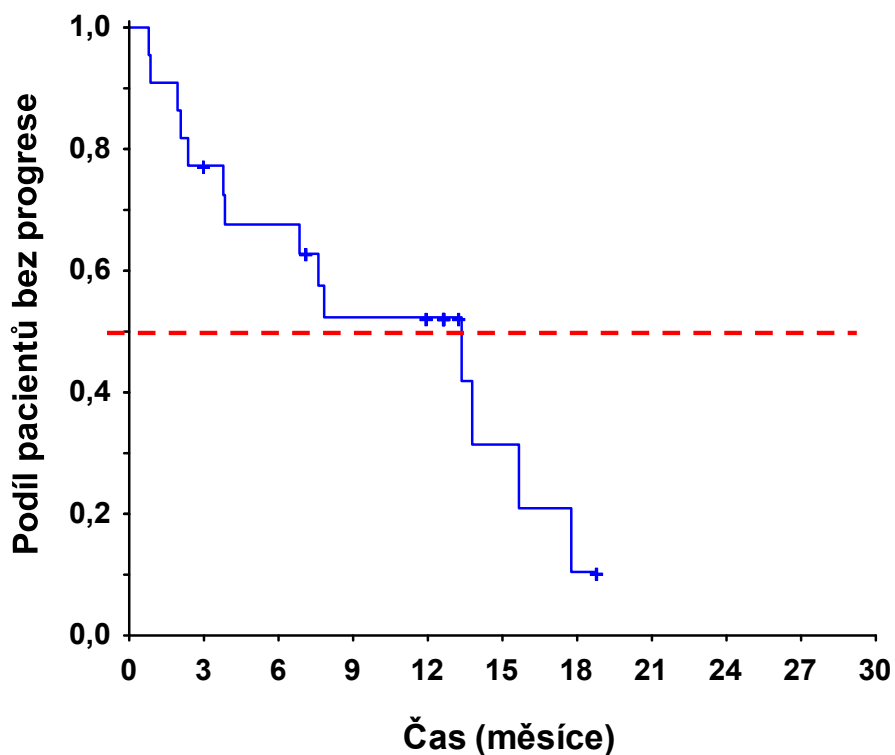


Přežití bez známek progresu od data zahájení léčby Iressou

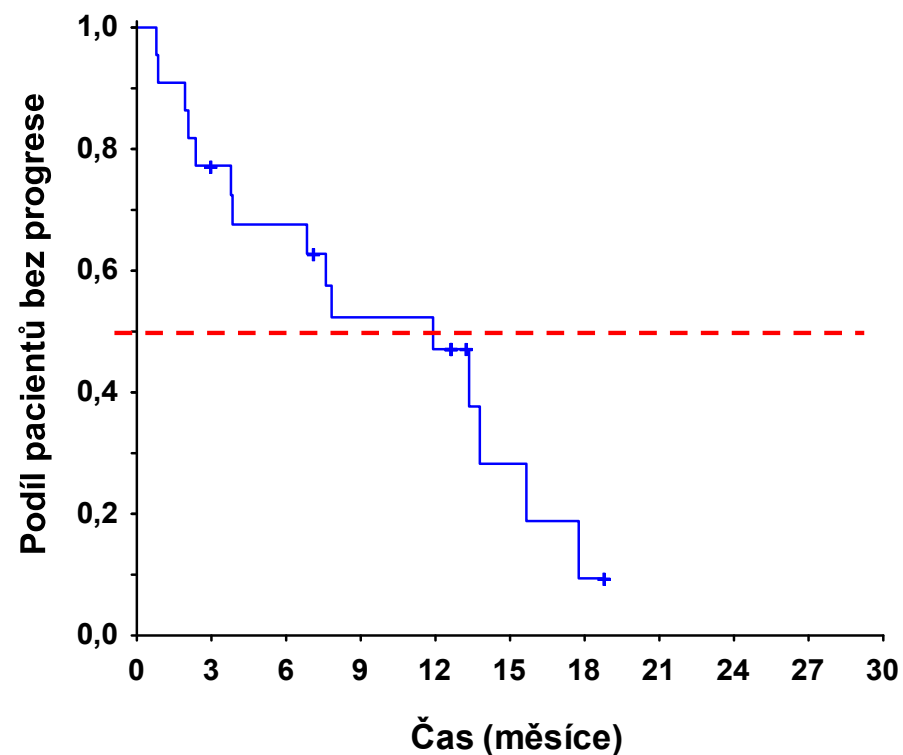


PFS	
Medián PFS (95% IS)	13,4 měsíce (4,1; 22,6)

Přežití bez známek progresu od data zahájení léčby Iressou



	PFS
Medián PFS (95% IS)	13,4 měsíce (4,1; 22,6)

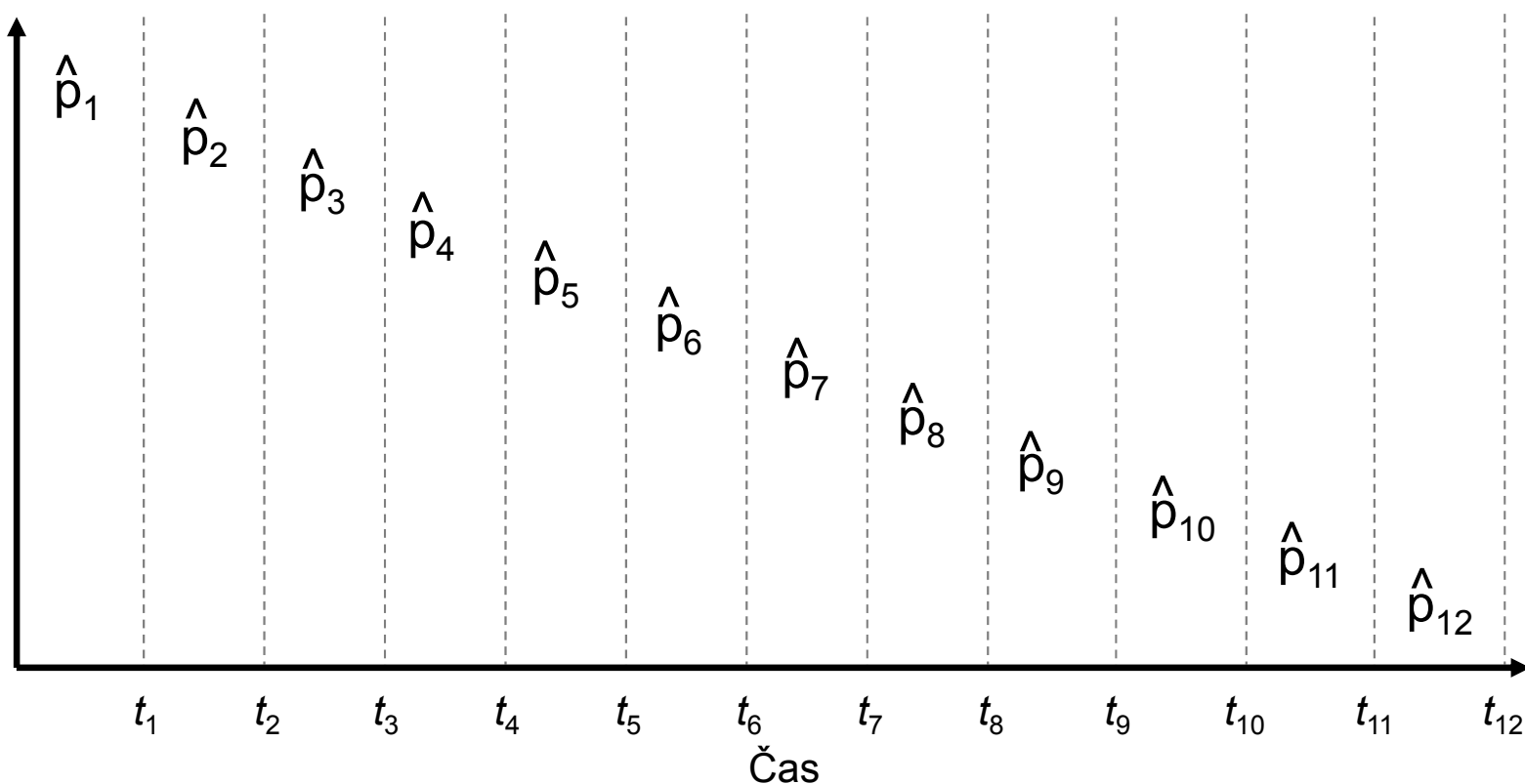


	PFS
Medián PFS (95% IS)	11,9 měsíce (5,6; 18,2)

NEPARAMETRICKÉ ODHADY V ANALÝZE PŘEŽITÍ

- Statistické metody lze obecně rozdělit na základě jejich předpokladu o charakteru pozorovaných dat na parametrické a neparametrické.
- **Parametrické metody** (*parametric survival analysis*) vyžadují specifikaci konkrétního rozdělení náhodné veličiny T .
- **Neparametrické metody** (*nonparametric survival analysis*) žádné zvláštní předpoklady ohledně rozdělení pravděpodobnosti náhodné veličiny T nevyžadují.
- **Korektní specifikace rozdělení pravděpodobnosti je důležitá.**
- V analýze přežití definujeme ještě další skupinu metod označovanou jako **semiparametrické** (*semiparametric survival analysis*). Jedná se o modelovací přístupy, které nejsou plně parametrické, protože nevyžadují předpoklad o znalosti rozdělení veličiny T , nicméně jakožto modely s parametry, respektive regresními koeficienty pracují (Coxův model).

- Je nejnámějším a nejpoužívanějším neparametrickým odhadem funkce přežití, který se také stal standardem pro hodnocení přežití v klinických studiích.
- **Myšlenka výpočtu:** aby byl subjekt v čase t bez sledované události (aby se např. pacient s nádorovým onemocněním dožil času t), nesmí se u něj událost vyskytnout v žádném čase t^* takovém, pro nějž platí, že $t^* < t$.



- Při odhadu pravděpodobností přežití jednotlivých časů t_i je třeba adekvátně zohlednit cenzorování. Cenzorované časy přežití totiž nelze hodnotit stejně jako kompletní pozorování, neboť nepřispívají k počtu událostí (d_i), ale zároveň je nelze z hodnocení vyřadit.
- Kaplanův-Meierův odhad pracuje s cenzorováním tak, že tato pozorování vypadávají ze skupiny subjektů v riziku ihned po zaznamenaném čase cenzorování.
- **Vzorec pro Kaplanův-Meierův odhad funkce přežití:**

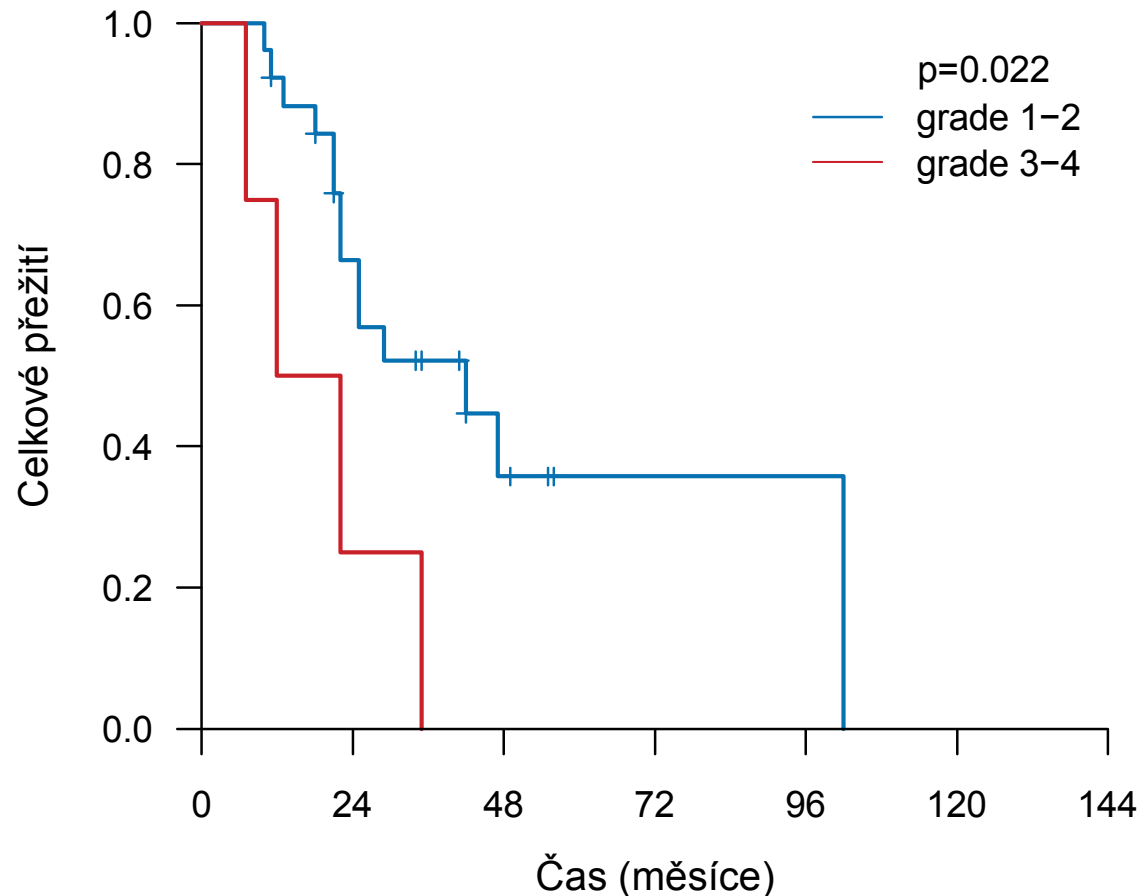
$$\hat{S}(t) = \prod_{t_i \leq t} \hat{p}_i = \prod_{t_i \leq t} \left(1 - \frac{d_i}{R_i} \right)$$

Počet sledovaných událostí v čase t_i

Intervalově specifický odhad přežití od času t_{i-1} do času t_i

Počet subjektů v riziku sledované události v čase t_i

- Výsledkem Kaplanova-Meierova odhadu je **schodovitá funkce s poklesem v časech sledované události**. Cenzorování odhad přežití nemění (mění ho pouze zprostředkovaně).



- Data o přežití 15 pacientů s nemalobuněčným karcinomem plic léčených v 1. linii chemoterapií.
- Data **bez cenzorování hodnot** (kompletní záznamy):

Pacient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Čas přežití (měsíce)	2,9	2,9	4,8	5,9	6,3	6,9	7,8	8,3	8,7	9,8	10,9	11,1	12,4	12,6	17,1

- Data **s cenzorováním hodnot** (nekompletní záznamy):

Pacient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Čas přežití (měsíce)	2,9	2,1+	4,8	4,9+	6,3	6,9	7,0+	8,3	8,7	9,8	10,9	10,5+	11,2+	12,6	17,1

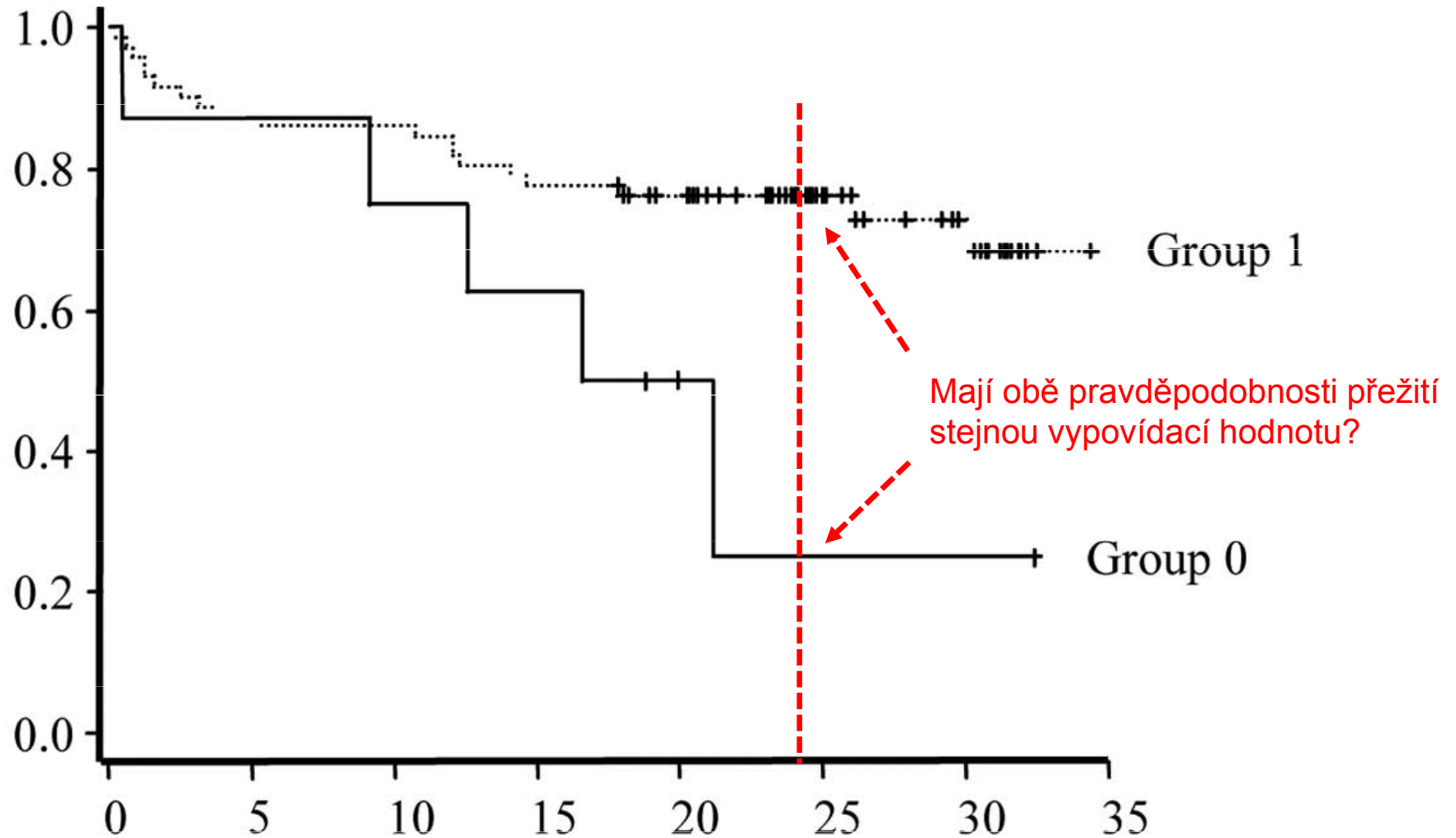
Kaplanův-Meierův odhad - příklad

Bez přítomnosti cenzorovaných pozorování					S přítomností cenzorovaných pozorování				
t_i	R_i	d_i	$1 - \frac{d_i}{R_i}$	$\hat{S}(t)$	t_i	R_i	d_i	$1 - \frac{d_i}{R_i}$	$\hat{S}(t)$
2,9	15	2	0,867	0,867	2,9	14	1	0,929	0,929
4,8	13	1	0,923	0,800	4,8	13	1	0,923	0,857
5,9	12	1	0,917	0,733	5,9	11	0	1,000	0,857
6,3	11	1	0,909	0,667	6,3	11	1	0,909	0,779
6,9	10	1	0,900	0,600	6,9	10	1	0,900	0,701
7,8	9	1	0,889	0,533	7,8	8	0	1,000	0,701
8,3	8	1	0,875	0,467	8,3	8	1	0,875	0,613
8,7	7	1	0,857	0,400	8,7	7	1	0,857	0,526
9,8	6	1	0,833	0,333	9,8	6	1	0,833	0,438
10,9	5	1	0,800	0,267	10,9	4	1	0,750	0,329
11,1	4	1	0,750	0,200	11,1	3	0	1,000	0,329
12,4	3	1	0,667	0,133	12,4	2	0	1,000	0,329
12,6	2	1	0,500	0,067	12,6	2	1	0,500	0,165
17,1	1	1	0,000	0,000	17,1	1	1	0,000	0,000

Kaplanův-Meierův odhad - příklad

Bez přítomnosti cenzorovaných pozorování					S přítomností cenzorovaných pozorování				
t_i	R_i	d_i	$1 - \frac{d_i}{R_i}$	$\hat{S}(t)$	t_i	R_i	d_i	$1 - \frac{d_i}{R_i}$	$\hat{S}(t)$
2,9	15	2	0,867	0,867	2,9	14	1	0,929	0,929
4,8	13	1	0,923	0,800	4,8	13	1	0,923	0,857
5,9	12	1	0,917	0,733	5,9	11	0	1,000	0,857
6,3	11	1	0,909	0,667	6,3	11	1	0,909	0,779
<p>Bez cenzorování je K-M odhad dán vztahem $S(t_i) = 1 - (\sum d_j / R_1)$.</p> <p>V přítomnosti cenzorování je K-M odhad zkreslený.</p>									
8,7	7	1	0,857	0,400	8,7	7	1	0,857	0,526
9,8	6	1	0,833	0,333	9,8	6	1	0,833	0,438
10,9	5	1	0,800	0,267	10,9	4	1	0,750	0,329
11,1	4	1	0,750	0,200	11,1	3	0	1,000	0,329
12,4	3	1	0,667	0,133	12,4	2	0	1,000	0,329
12,6	2	1	0,500	0,067	12,6	2	1	0,500	0,165
17,1	1	1	0,000	0,000	17,1	1	1	0,000	0,000

- Samotný bodový odhad pravděpodobnosti přežití je nedostatečný.



- Pro konstrukci $100(1 - \alpha)\%$ intervalu spolehlivosti pro odhad $S(t)$ potřebujeme získat jeho rozptyl
- Standardem pro odhad variability Kaplanova-Meierova odhadu funkce přežití je tzv. **Greenwoodův vzorec**:

$$\text{var}(\hat{S}(t)) = (\hat{S}(t))^2 \sum_{t_i \leq t} \frac{d_i}{R_i(R_i - d_i)}$$

- Alternativní odhad dle autorů Peto a kol. (1977) – pro časy, kdy se odhad $S(t)$ blíží hodnotám 1 nebo 0, při nichž by odhad pomocí Greenwoodova vzorce mohl skutečnou variabilitu podhodnocovat:

$$\text{var}(\hat{S}(t)) = (\hat{S}(t))^2 \frac{1 - \hat{S}(t)}{R_t}$$

- Nejpoužívanějším postupem pro **konstrukci $100(1 - \alpha)\%$ intervalu spolehlivosti pro odhad $S(t)$ je využití aproximace normálním rozdělením** (podmínky dobré aproximace souvisí především s dostatečným množstvím subjektů zahrnutých do analýzy)

- $100(1 - \alpha)\%$ interval spolehlivosti:

$$\left(\hat{S}(t) - z_{1-\alpha/2} \sqrt{\text{var}(\hat{S}(t))}; \hat{S}(t) + z_{1-\alpha/2} \sqrt{\text{var}(\hat{S}(t))} \right)$$

- **Komplementární logaritmická transformace** odhadu funkce přežití $\ln(-\ln(S(t)))$ umožňuje transformovat odhad $S(t)$ na hodnoty z intervalu $(-\infty, \infty)$ a vyhnout se tak problémům výše uvedeného symetrického intervalu spolehlivosti v okolí hodnot odhadu funkce přežití 1 a 0:

$$\left[\left(\hat{S}(t) \right)^{\exp\left(\frac{z_{1-\alpha/2}}{\ln \hat{S}(t)} \sqrt{\text{var}(\hat{S}(t))}\right)}, \left(\hat{S}(t) \right)^{\exp\left(\frac{-z_{1-\alpha/2}}{\ln \hat{S}(t)} \sqrt{\text{var}(\hat{S}(t))}\right)} \right].$$

- Metoda pro hodnocení populačních dat. Hlavní myšlenka odhadu zůstává stejná. Na rozdíl od Kaplanova-Meierova odhadu, kde byly časové intervaly určeny pozorovanými hodnotami časů přežití, zde **pracujeme s předem definovanou sadou J časových intervalů**.
- Pracujeme s delšími časovými intervaly, pro odhad $S(t)$ nám stačí pouze agregovaná data, tedy souhrnné údaje pro jednotlivé časové intervaly:

$$\hat{S}(J) = \prod_{j=1}^J \hat{p}_j = \prod_{j=1}^J \left(1 - \frac{d_j}{R_j - \frac{c_j}{2}} \right)$$

- kde d_j je počet sledovaných událostí v j -tém intervalu, kde $j = 1, \dots, J$, dále R_j je počet subjektů v riziku výskytu sledované události na začátku intervalu j a c_j je počet subjektů s časem přežití cenzorovaným v průběhu j -tého intervalu.

- **Nelsonův-Aalenův odhad** je základní neparametrickou metodou odhadu kumulativní rizikové funkce, která stejně jako Kaplanův-Meierův odhad pracuje pouze se souborem n pozorovaných hodnot časů přežití takových, že $t_1 < t_2 < \dots < t_n < t$:

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{R_i}$$

- kde d_i značí počet sledovaných událostí zaznamenaných v čase t_i a R_i je počet subjektů v riziku výskytu sledované události v čase t_i .
- Odhad rozptylu Nelsonova-Aalenova odhadu kumulativní rizikové funkce:

$$\text{var}(\hat{H}(t)) = \sum_{t_i \leq t} \frac{d_i}{R_i^2}$$

- 100(1 - α)% interval spolehlivosti:

$$\left[\hat{H}(t) - z_{1-\alpha/2} \sqrt{\text{var}(\hat{H}(t))}, \hat{H}(t) + z_{1-\alpha/2} \sqrt{\text{var}(\hat{H}(t))} \right].$$

- **Pro připomenutí** – vztah mezi funkcí přežití a kumulativní rizikovou funkcí:

$$\hat{Q}(t) = -\ln \hat{R}(t)$$

$$\hat{R}(t) = \exp(-\hat{Q}(t))$$

- **Breslowův odhad** funkce přežití využívá předchozího vztahu a neparametrického Nelsonova-Aalenova odhadu kumulativní rizikové funkce

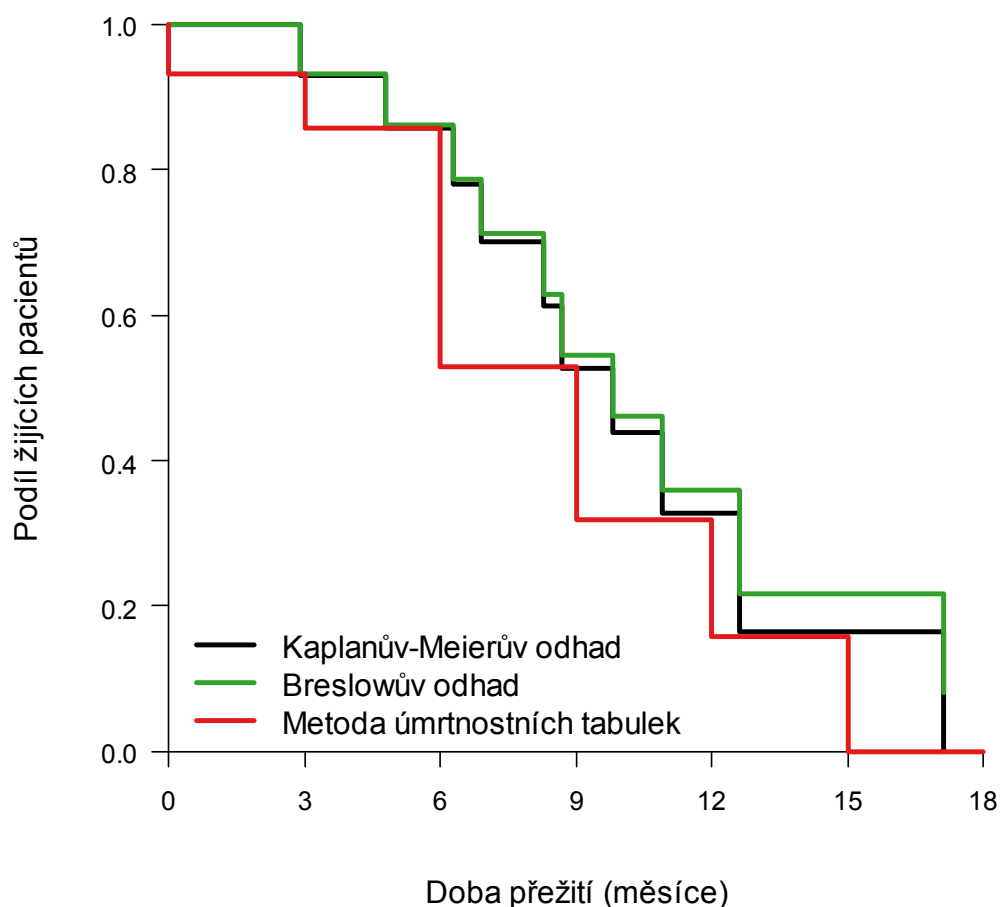
$$\hat{R}(t) = \exp(-\hat{Q}(t)) = \exp\left(-\sum_{t_j \leq t} \frac{d_j}{n_j}\right)$$

- Pro konstrukci intervalu spolehlivosti je opět třeba odhad rozptylu odhadu funkce přežití:

$$\text{var}(\hat{R}(t)) = (\hat{R}(t))^2 \sum_{t_j \leq t} \frac{d_j}{n_j^2}$$

- **Příklad:** data o přežití 15 pacientů s nemalobuněčným karcinomem plic léčených v 1. linii chemoterapií – data s cenzorovaním hodnot

Pacient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Čas přežití (měsíce)	2,9	2,1+	4,8	4,9+	6,3	6,9	7,0+	8,3	8,7	9,8	10,9	10,5+	11,2+	12,6	17,1



- Odhad pomocí **metody úmrtnostních tabulek** respektuje časové intervaly (v tomto případě 3 měsíce)
- Při **dostatečné velikosti souboru** jsou odhady **téměř shodné**

PARAMETRICKÉ ODHADY V ANALÝZE PŘEŽITÍ

- **Parametrické odhady** předpokládají konkrétní funkční vyjádření rozdělení pravděpodobnosti náhodné veličiny T
- **Korektní specifikace rozdělení pravděpodobnosti je důležitá** (předpoklad konkrétního rozdělení pravděpodobnosti náhodné veličiny T je zvláště v analýze přežití silný a může být zrádný)
- **Hlavní výhody parametrických odhadů:**
 1. **Jednodušší odhad kvantilů $S(t)$** (mediánu přežití a střední doby dožití),
 2. **Možnost vyjádření $S(t)$, $h(t)$ a $H(t)$ pomocí **spojité funkce**,**
 3. **Přesnější odhad funkce přežití** než s pomocí Kaplanova-Meierova odhadu,
 4. **Nižší variabilita** (resp. standardní chyba) odhadů hlavních charakteristik náhodné veličiny T .

- V klasické statistice hraje hlavní roli normální rozdělení pravděpodobnosti, případně diskrétní rozdělení pravděpodobnosti (binomické a Poissonovo)
- ~~Diskrétní rozdělení~~ → v analýze přežití předpokládáme spojitou náhodnou veličinu T
- ~~Normální rozdělení~~ → časy přežití mají většinou kladně sešikmené rozdělení (to znamená, že většina osob má kratší či střední doby přežití a osob s delšími až extrémními časy přežití je relativně málo)
- **Nejčastěji používaná rozdělení pravděpodobnosti v analýze přežití:**
 - Exponenciální rozdělení
 - Weibullovo rozdělení
 - Logaritmicko-normální rozdělení
 - Logaritmicko-logistické rozdělení

- Spojité rozdělení pravděpodobnosti, které popisuje délky časových intervalů mezi výskyty jednotlivých událostí tzv. **Poissonova procesu**
- Popisuje délku časových intervalů mezi jednotlivými událostmi, když se tyto události vyskytují vzájemně nezávisle a s konstantní intenzitou
- **Jeden parametr:** λ popisuje intenzitu neboli míru rizika v čase

Hustota:

$$f(t) = \lambda e^{-\lambda t}$$

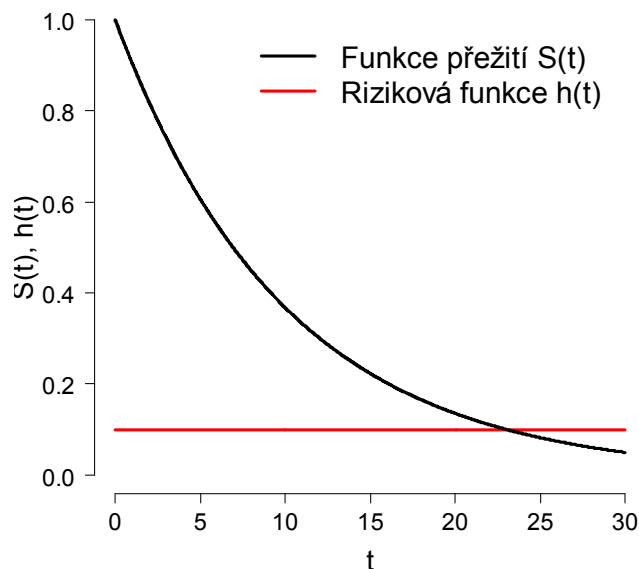
Riziková funkce:

$$h(t) = \lambda$$

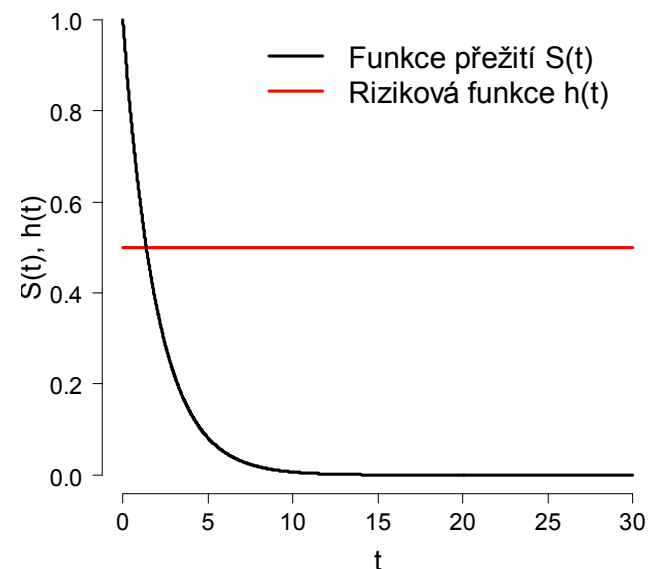
Funkce přežití:

$$F(t) = e^{-\lambda t}$$

$$\lambda = 0,1$$



$$\lambda = 0,5$$



- Zobecnění exponenciálního rozdělení, které navrhl Weibull (1951) pro popis životnosti materiálů
- Nepředpokládá konstantní riziko výskytu sledované události v čase, ale uvažuje monotónní rizikovou funkci (tedy s časem monotónně rostoucí nebo klesající funkci)
- **Dva parametry:** γ určuje tvar hustoty pravděpodobnosti a λ škálu hodnot

Hustota:

$$f(t) = \frac{\gamma}{\lambda} \left(\frac{t}{\lambda}\right)^{\gamma-1} e^{-\left(\frac{t}{\lambda}\right)^\gamma}$$

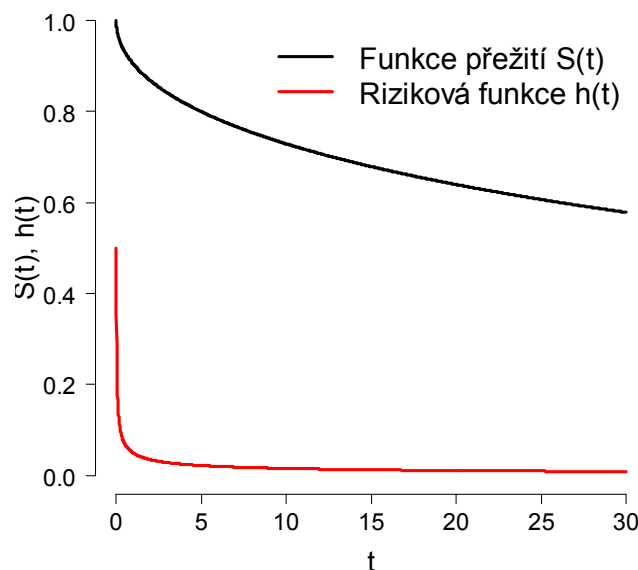
Riziková funkce:

$$h(t) = \gamma \left(\frac{t}{\lambda}\right)^{\gamma-1} e^{-\left(\frac{t}{\lambda}\right)^\gamma}$$

Funkce přežití:

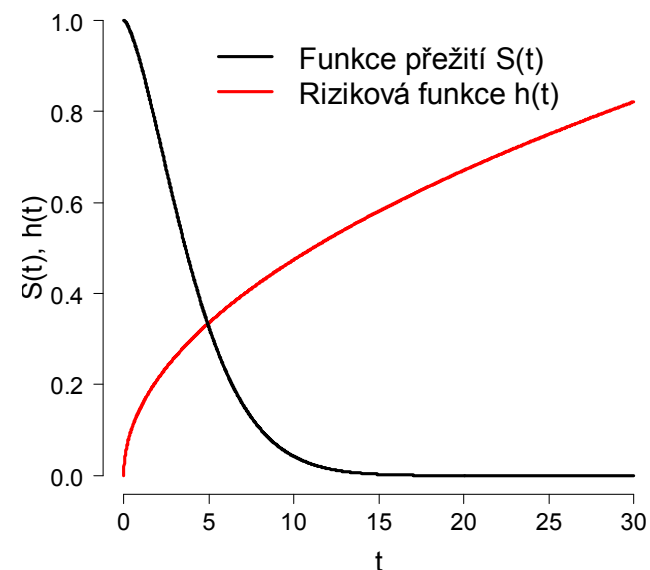
$$S(t) = e^{-\left(\frac{t}{\lambda}\right)^\gamma}$$

$\gamma = 0,1$; $\lambda = 0,5$



$\gamma < 1$... $h(t)$ je **monotónně klesající**

$\gamma = 0,1$; $\lambda = 1,5$



$\gamma > 1$... $h(t)$ je **monotónně rostoucí**

$\gamma = 1$... $h(t)$ je konstantní = exponenciální rozdělení

Logaritmicko-normální rozdělení:

- Veličina T má logaritmicko-normální rozdělení, pokud veličina $Y = \ln(T)$ má normální rozdělení
- Vhodné v případech, kdy můžeme v období bezprostředně po zahájení sledování očekávat nárůst rizika úmrtí (např. po chirurgickém zákroku), které však po dosažení maximální hodnoty opět klesá (pacienti, kteří se zotaví ze srdečního selhání).
- **Dva parametry:** μ a σ^2 , které mají význam střední hodnoty a rozptylu normálního rozdělení korespondující náhodné veličiny $Y = \ln(T)$

Hustota:

$$f(t) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln t - \mu)^2}{2\sigma^2}\right)$$

Riziková funkce:

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

Funkce přežití:

$$F(t) = 1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)$$

Pozn.: $\Phi(\cdot)$ je distribuční funkce standardizovaného normálního rozdělení s parametry $\mu = 0$ a $\sigma^2 = 1$

Logaritmicko-logistické rozdělení:

- Lze jej chápat jako **transformaci Weibullova rozdělení**
- **Riziková funkce** je rozšířena o člen (jmenovatel), což umožňuje rizikové funkci větší flexibilitu, která na rozdíl od Weibullova rozdělení **nemusí být monotónní**

Riziková funkce:

$$h(t) = \frac{(\ln t)^{\alpha-1}}{1 + (\ln t)^\alpha}$$

- Parametrické odhady $S(t)$ jsou v analýze přežití založeny na **metodě maximální věrohodnosti** (*maximum likelihood estimation*)
- Zavádí tzv. **věrohodnostní funkci**, která je použita pro odhad neznámých parametrů vybraného rozdělení pravděpodobnosti

$$L(\theta; t_1, \dots, t_n) = L(t_1, \dots, t_n | \theta) = \prod_{i=1}^n L(t_i | \theta)$$

- Principem je nalezení odhadu parametru θ , který maximalizuje pravděpodobnost, že pozorované hodnoty pocházejí z předpokládaného rozdělení
- **Hlavní myšlenka:** sdružená hustota jako funkce vektoru parametrů θ (při pevně daných hodnotách t_1, t_2, \dots, t_n) \rightarrow ze všech možných hodnot θ vybírá metoda maximální věrohodnosti takové, aby věrohodnostní funkce nabývala svého maxima

- **Věrohodnostní funkce přežití** (v přítomnosti cenzorování):

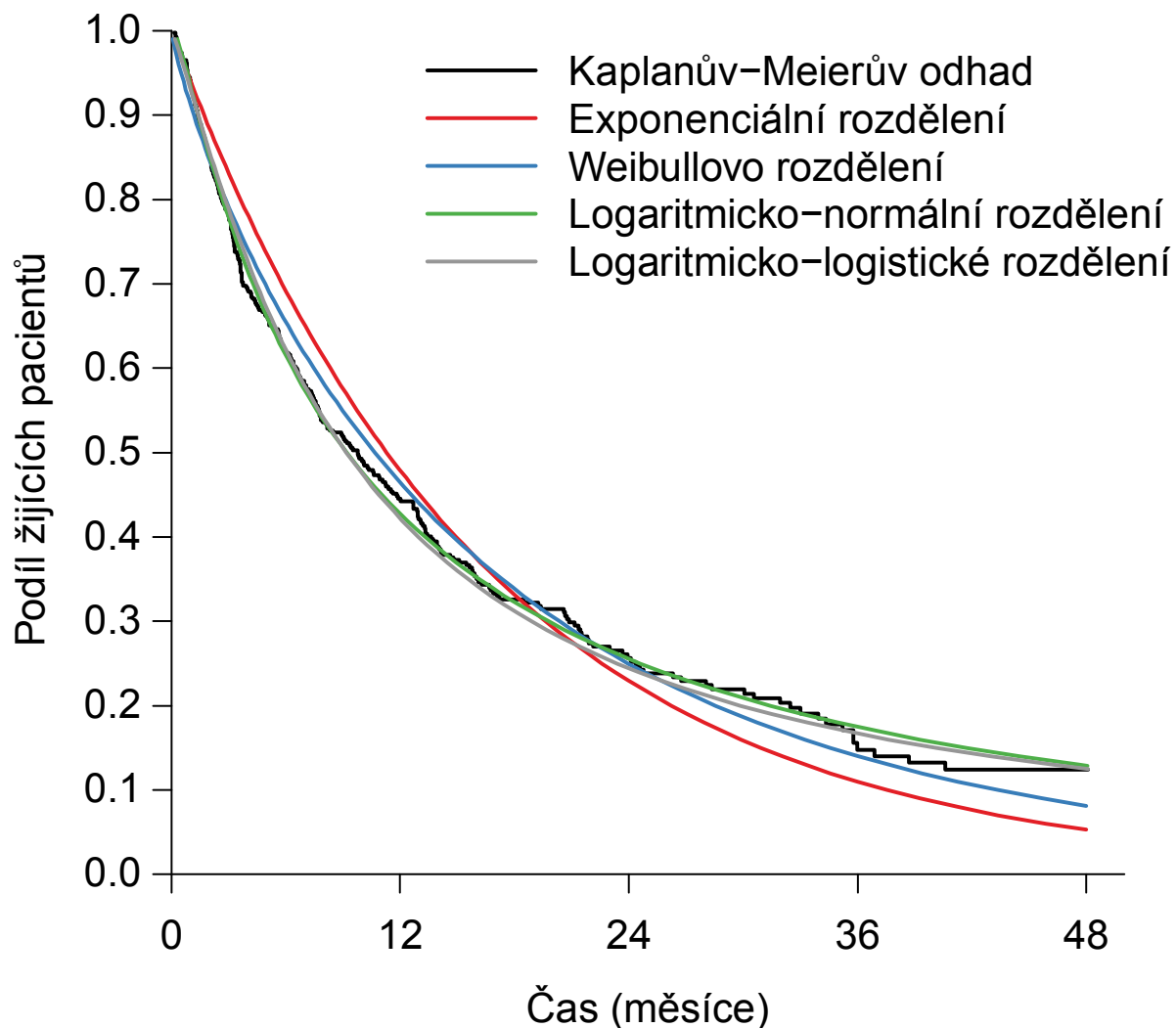
$$L(\lambda | (t_1, \delta_1), (t_2, \delta_2), \dots, (t_n, \delta_n)) = \prod_{i=1}^n h(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

- **Logaritmická věrohodnostní funkce přežití:**

$$\begin{aligned} \ln L(\lambda | (t_1, \delta_1), (t_2, \delta_2), \dots, (t_n, \delta_n)) &= \ln \prod_{i=1}^n h(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \\ &= \sum_{i=1}^n (\delta_i \ln h(t_i) + (1-\delta_i) \ln S(t_i)) \\ &= \sum_{i=1}^n (\delta_i \ln h(t_i) - (1-\delta_i) \ln h(t_i)) \end{aligned}$$

- **Úplné pozorování:** příspěvek *i*tého pacienta k věrohodnostní funkci lze vyjádřit jako $f(t_i) = h(t_i)S(t_i)$ – pravděpodobnost, že se subjekt dožil času t_i bez události a zároveň u něj v čase t_i událost nastala
- **Cenzorované pozorování:** příspěvek *i*tého pacienta k věrohodnostní funkci lze zjednodušit pouze na $f(t_i) = S(t_i)$, neboť jediné, co víme, je, že se subjekt bez události dožil času t_i

- **Příklad:** data pacientů s metastatickým karcinomem plic z registru TULUNG, kteří byli léčeni protinádorovou terapií



- Logaritmicke-normální a logaritmicke-logistické rozdělení vystihují pozorované hodnoty přežití s drobnými výjimkami, které však mohou být způsobeny způsobem sběru dat
- Odhady pro exponenciální a Weibullovo rozdělení jsou méně přesné, neboť méně kopírují Kaplanův-Meierův odhad

- **Ověření zvoleného rozdělení pravděpodobnosti = klíčový prvek** pro použití parametrických modelů v hodnocení přežití
- Tento krok samozřejmě není jednoduchý a může být do značné míry subjektivním, zvláště srovnáváme-li např. neparametrický Kaplanův-Meierův odhad s proloženou parametrickou křivkou
- Pro exponenciální a Weibullovo rozdělení však existují jednoduchá pravidla pro ověření vhodnosti těchto rozdělení pravděpodobnosti

• **Ověření předpokladu exponenciálního rozdělení:**

Platí: $h(t) = \lambda e^{-\lambda t}$ $\xrightarrow{\text{dle definičních vztahů}}$ $\hat{H}(t) = \frac{1}{\lambda}$ $\xrightarrow{\text{splnění předpokladu}}$ Nelsonův-Aalenův odhad $H(t)$ by měl přibližně tvořit **přímku**

• **Ověření předpokladu Weibullova rozdělení:**

Platí: $\hat{H}(t) = \frac{1}{\lambda} e^{-\lambda t^\alpha}$ $\xrightarrow{\text{úprava transformací}}$ $\ln(-\ln \hat{H}(t)) = \ln \frac{1}{\lambda} e^{-\lambda t^\alpha} = \ln \frac{1}{\lambda} - \lambda t^\alpha = \ln \frac{1}{\lambda} - \lambda t^\alpha$

$\xrightarrow{\text{splnění předpokladu}}$ Logaritmus $H(t)$ je lineárně závislý na logaritmu času

$\xrightarrow{\text{ověření předpokladu}}$ Pomocí Kaplanova-Meierova odhadu $S(t)$ – znázornění $\ln(-\ln \hat{S}(t))$ proti $\ln(t)$

METODY PRO SROVNÁNÍ ODHADŮ PŘEŽITÍ

- Pro hodnocení přežití však z důvodu přítomnosti cenzorování nelze použít standardní statistické testy.
- Máme dvě možnosti, jak při srovnání přežití dvou a více skupin subjektů postupovat:
 1. První možností je **srovnání odhadnutých pravděpodobností přežití v daném časovém bodě**, například v 1 roce od začátku sledování.
 2. Druhou možností je tak použití **statistického testu, který by umožňoval vypořádat se s cenzorováním časů přežití** (tedy s nekompletní informací) a zároveň zohlednit časový vývoj přežití ve srovnávaných skupinách.

- I v testování hypotéz o přežití pracujeme s nulovou a alternativní hypotézou, jejichž platnost dále ověřujeme pomocí statistického testu, tedy rozhodovacího pravidla, které každé množině pozorovaných hodnot náhodné veličiny T přiřadí právě jedno ze dvou možných rozhodnutí: nulovou hypotézu nezamítáme nebo naopak, nulovou hypotézu zamítáme.
- Nulová hypotéza:
 1. Srovnání odhadnutých pravděpodobností přežití v daném časovém bodě:

$$H_0: S_1(t_0) = S_2(t_0)$$

2. Zohlednění časového vývoje přežití ve srovnávaných skupinách:

$$H_0: \lambda_1(t) = \lambda_2(t)$$

- Specifikace **alternativní hypotézy** (*alternative hypothesis*), značíme ji H_1 , která popírá platnost nulové hypotézy a vymezuje, jaká situace nastává, když nulová hypotéza neplatí, se liší v závislosti na typu použitého statistického testu.
- Nejpoužívanější neparametrický test pro srovnání přežití dvou a více skupin subjektů, tzv. **log-rank test** (*log-rank test*), předpokládá, že v případě rozdílného přežití sledovaných skupin subjektů jsou jejich rizikové funkce vzájemně proporcionální, což znamená, že jednu lze vyjádřit pomocí druhé násobením kladnou konstantou různou od 1.
- Alternativní hypotéza log-rank testu:

$$H_1: \lambda_1(t) = \theta \lambda_2(t)$$

- Připustíme-li i neproporcionální rizikové funkce, je možné alternativní hypotézu zapsat obdobně jako v případě standardního testování hypotéz:

$$H_1: \lambda_1(t) \neq \lambda_2(t)$$

- Použijeme-li index 1 pro první skupinu subjektů a index 2 pro druhou skupinu subjektů, můžeme testovou statistiku zapsat jako

$$X^2 = \frac{(\log(-\log \hat{S}_1(t)) - \log(-\log \hat{S}_2(t)))^2}{\hat{\sigma}_1(t)^2 / (\log \hat{S}_1(t))^2 + \hat{\sigma}_2(t)^2 / (\log \hat{S}_2(t))^2}$$

kde

$$\hat{\sigma}(t)^2 = \sum_{t_i \leq t} \frac{d_i}{R_i(R_i - d_i)}$$

což je výraz z rozptylu odhadu $S(t)$.

- Log-rank test se také stal standardem v hodnocení klinických studií a byl navržen Mantelem a Haenszelem jako modifikace testu pro analýzu stratifikovaných kontingenčních tabulek.
- V každém z n pozorovaných časů přežití můžeme sestavit kontingenční tabulku shrnující pozorované přežití v obou skupinách:

Skupina	Počet subjektů v riziku v čase t_i	Počet událostí v čase t_i	Počet subjektů bez události v čase t_i
1	R_{1i}	d_{1i}	$R_{1i} - d_{1i}$
2	R_{2i}	d_{2i}	$R_{2i} - d_{2i}$
Celkem	R_i	d_i	$R_i - d_i$

- R_{1i} a R_{2i} jsou počty subjektů v riziku v čase t_i ve skupině 1 a 2 a obdobně d_{1i} a d_{2i} počty sledovaných událostí v čase t_i ve skupině 1 a 2.
- Za platnosti nulové hypotézy (mezi skupinami 1 a 2 není rozdíl v přežití) bychom měli v čase t_i pozorovat přibližně stejně událostí v obou skupinách.

Skupina	Počet subjektů v riziku v čase t_i	Počet událostí v čase t_i	Počet subjektů bez události v čase t_i
1	R_{1i}	d_{1i}	$R_{1i} - d_{1i}$
2	R_{2i}	d_{2i}	$R_{2i} - d_{2i}$
Celkem	R_i	d_i	$R_i - d_i$

- Stejný počet událostí vyhodnocujeme pomocí tzv. očekávaného počtu událostí

$$E(d_{1i}) = R_{1i} \frac{d_i}{R_i}$$

- Testová statistika je pak založen na rozdílu mezi pozorovanými počty událostí a očekávanými počty v jednotlivých časových bodech:

$$U = \sum_{i=1}^n (d_{1i} - E(d_{1i}))$$

- Má-li statistika U přibližně normální rozdělení pravděpodobnosti s nulovou střední hodnotou, pak statistika $U/\sqrt{\text{var}(U)}$ má přibližně normální rozdělení pravděpodobnosti s nulovou střední hodnotou a jednotkovým rozptylem (aproximace je tím lepší, čím více máme pozorovaných událostí).
- Náhodná veličina d_{1i} má hypergeometrické rozdělení. Z toho plyne

$$\text{var}(d_{1i}) = \left[R_{1i} \frac{d_i}{R_i} \left(1 - \frac{d_i}{R_i} \right) \right] \left(\frac{R_i - R_{1i}}{R_i - 1} \right) = \frac{R_{1i} R_{2i} d_i (R_i - d_i)}{R_i^2 (R_i - 1)}$$

- Rozptyl U lze za předpokladu nezávislosti jednotlivých pozorovaných časů přežití vyjádřit jako

$$\text{var}(U) = \sum_{i=1}^n \text{var}(d_{1i})$$

- Z teorie pravděpodobnosti pak plyne, že statistika

$$Q_{M-H} = \frac{U^2}{\text{var}(U)} = \frac{\left\{ \sum_{i=1}^n (d_{1i} - E(d_{1i})) \right\}^2}{\sum_{i=1}^n \text{var}(d_{1i})}$$

má přibližně chí-kvadrát rozdělení pravděpodobnosti s jedním stupněm volnosti.

- Pozorované či cenzorované časy přežití (v měsících) pacientů s nemalobuněčným karcinomem plic a jejich performance status (PS 0 – skupina 1, PS 1 a 2 – skupina 2)

Pacient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Čas přežití (měsíce)	2,9	2,1+	4,8	4,9+	6,3	6,9	7,0+	8,3	8,7	9,8	10,9	10,5+	11,2+	12,6	17,1
Skupina podle PS	2	2	1	2	1	2	2	2	1	2	1	1	1	2	1

- Pozorované či cenzorované časy přežití (v měsících) pacientů s nemalobuněčným karcinomem plic a jejich performance status (PS 0 – skupina 1, PS 1 a 2 – skupina 2)

t_i	R_i	R_{1i}	R_{2i}	d_i	d_{1i}	d_{2i}	$E(d_{1i})$	$\text{var}(d_{1i})$
2,9	14	7	7	1	0	1	0,500	0,250
4,8	13	7	6	1	1	0	0,538	0,249
6,3	11	6	5	1	1	0	0,545	0,248
6,9	10	5	5	1	0	1	0,500	0,250
8,3	8	5	3	1	0	1	0,625	0,234
8,7	7	5	2	1	1	0	0,714	0,204
9,8	6	4	2	1	0	1	0,667	0,222
10,9	4	3	1	1	1	0	0,750	0,188
12,6	2	1	1	1	0	1	0,500	0,250
17,1	1	1	0	1	1	0	1,000	0,000
Suma				10	5	5	6,339	2,095

$$Q_{M-H} = \frac{(5 - 6,339)^2}{2,095} = 0,856$$



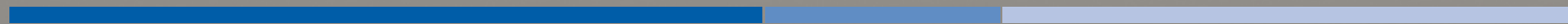
Srovnáme s hraniční hodnotou pro statistickou významnost. Jak to asi zde dopadne?



MASARYK UNIVERSITY
INSTITUTE OF BIostatISTICS AND ANALYSES

www.iba.muni.cz

POZOR NA LOG-RANK TEST

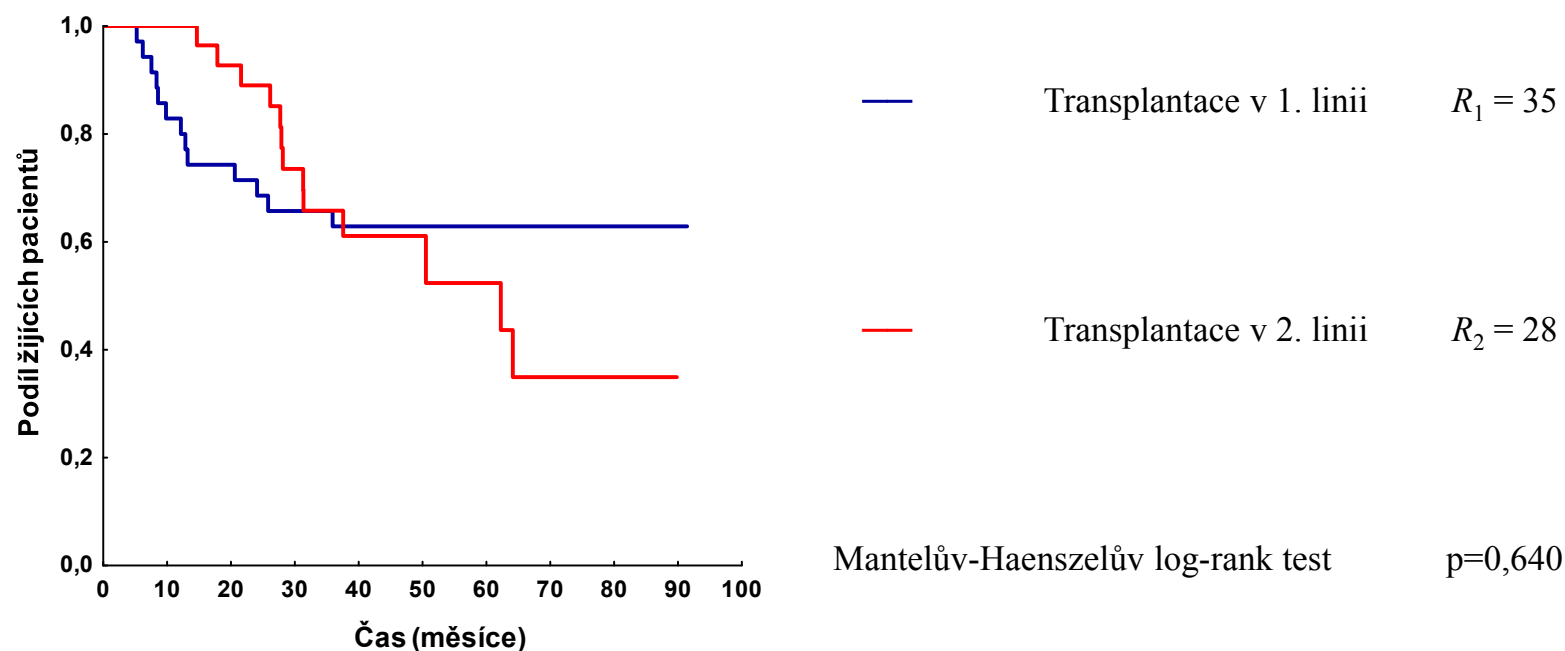


Pozor na výsledek log-rank testu

- **Log-rank test může dát statisticky nevýznamný výsledek v případě křížení křivek přežití.**

- Srovnání celkového přežití pacientů s chronickou myeloidní leukémií podle linie léčby, v níž byli pacienti transplantováni.

CML – celkové přežití podle linie transplantace



Opravdu v tom celkovém přežití není rozdíl?

- Log-rank test bere všechny pozorované události v průběhu doby sledování jako sobě rovné. Alternativou jsou testy, které dávají větší váhu na pozorování na začátku sledování – zde jsou totiž odhady přežití přesnější než na konci sledování.
- Obecně lze testovou statistiku pro výše popsanou skupinu testů zapsat jako

$$Q = \frac{\left\{ \sum_{i=1}^n w_i (d_{1i} - E(d_{1i})) \right\}^2}{\sum_{i=1}^n w_i^2 \text{var}(d_{1i})}$$

Test	Testová statistika	Váhy v čase t_j
Mantelův-Haenszelův log-rank	Q_{M-H}	1
Gehanův	Q_G	R_i
Taroneho-Wareův	Q_{T-W}	$\sqrt{R_i}$