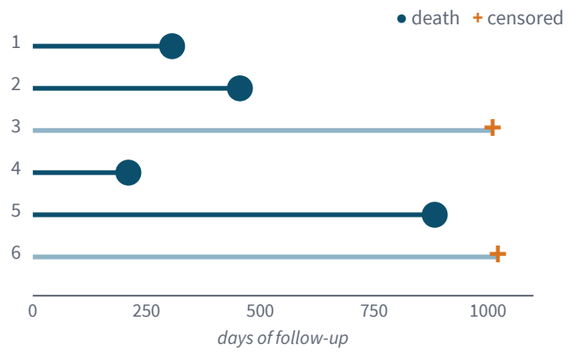
Survival analysis with survival :: CHEAT SHEET

The survival package models time until an event — death, relapse, discharge — when some subjects are followed only partway.

1 The survival object

Each subject gives a **time** and an **event indicator**. If the event was not seen, the time is **censored** (+): we only know survival lasted at least that long.




```
library(survival)
Surv(lung$time, lung$status)
#> 306   455   1010+   210   883   1022+ ...
```

CODING THE EVENT (status =)

`0 / 1`	1 = event (the usual)
`1 / 2`	2 = event (as in lung)
`FALSE / TRUE`	TRUE = event

OTHER LAYOUTS

`Surv(start, stop, event)` counting process: time-varying data, delayed entry

`Surv(time, time2, type = "interval2")` interval censoring (NA = open end)

TIP Censoring must be **non-informative**: a patient lost because they got sicker breaks every method on this sheet. Check why people leave follow-up.

TIME FROM DATES

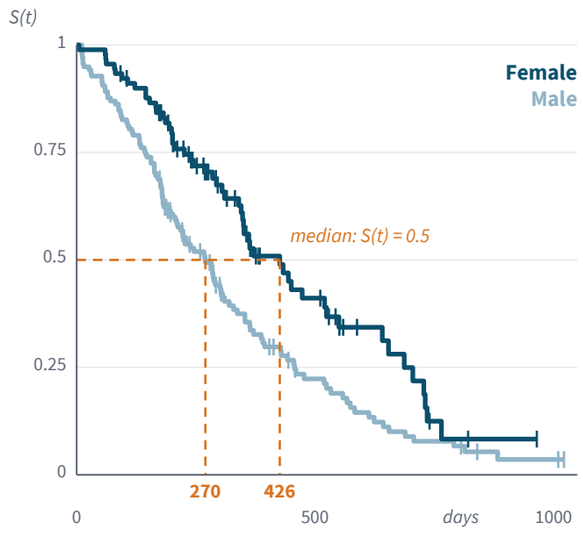
```
df$time <- as.numeric(difftime(
  df$end, df$start, units = "days"))
```

end = event date, or last contact if censored.

2 Kaplan–Meier curves

```
lung$sex <- factor(lung$sex, 1:2,
  labels = c("Male", "Female"))
km <- survfit(Surv(time, status) ~ sex,
              data = lung)
km
```

	n	events	median	95% CI
Male	138	112	**270**	212–310
Female	90	53	**426**	348–550



Steps mark event times; ticks mark censored times. **Overall curve: `~ 1`** (median 310 days).

```
plot(km, col = 1:2, conf.int = TRUE,
     mark.time = TRUE, xlab = "Days")
```

Quick base plot. For publication: survminer or ggsurvfit.

CUMULATIVE INCIDENCE 1 − S(t)

```
plot(km, fun = "event")   # curves rise
```

Use with one event type only; with competing events see page 2.

CHECK PROPORTIONAL HAZARDS BY EYE

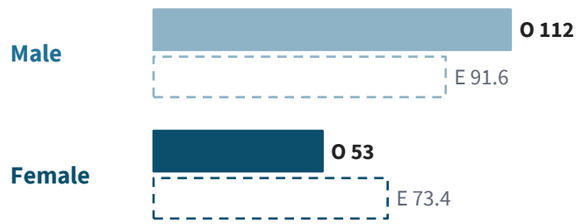
```
plot(km, fun = "cloglog")
```

log(−log S) vs log(time): roughly parallel curves support a Cox model.

3 Compare groups

```
survdiff(Surv(time, status) ~ sex,
         data = lung)
```

The test compares observed deaths with those expected if both groups shared one survival curve:



Men: more deaths than expected · women: fewer → **$\chi^2(1) = 10.3$, p = 0.001**

CHOOSE THE WEIGHTING: rho =

`rho = 0` χ^2 10.3

Log-rank (default). Weighs all times equally; most power when hazards are proportional.

`rho = 1` χ^2 12.7

Peto–Peto (Gehan–Wilcoxon type). More weight to early events.

ADJUST FOR A CONFOUNDER

```
survdiff(Surv(time, status) ~ sex +
  strata(ph.ecog), data = lung)
```

Stratified log-rank: compares sex within each ECOG level (χ^2 10.8).

TIP Curves that **cross** break proportional hazards: the log-rank may miss a real difference. Report survival at fixed times or the RMST instead (see 4).

MORE THAN TWO GROUPS

```
survdiff(Surv(time, status) ~ ph.ecog,
         data = lung)
```

One global test ($\chi^2(3) = 22$, $p < 0.001$). Pairwise: survminer::pairwise_survdiff().

4 Pull out the numbers

SURVIVAL AT CHOSEN TIMES

```
summary(km, times = c(180, 365))
```

	day 180	day 365
Male	64% (57–73)	34% (26–43)
Female	84% (77–92)	53% (42–66)
at risk	89 / 71	35 / 30

MEDIAN AND OTHER QUANTILES

```
quantile(km, probs = c(0.25, 0.5, 0.75))
```

Median Male 270 (212–310) · Female 426 (348–550). *No median if S(t) never reaches 0.5* — report it as "not reached".

RESTRICTED MEAN SURVIVAL (RMST)

```
print(km, rmean = 365)
```

Mean days lived in the first year (area under S(t) to 365): Male 241, Female 297. Valid even when curves cross.

AS A DATA FRAME

```
summary(km)$table   # n, events, median, CI
broom::tidy(km)     # one row per time
```

USEFUL survfit() ARGUMENTS

`conf.type = "log-log"` CI that stays in 0–1

`conf.int = 0.90` CI level

`start.time = 30` condition on surviving 30 days

TIP Always show the **number at risk** below a KM plot: the right tail rests on few subjects and is unstable.

MEDIAN FOLLOW-UP (REVERSE KM)

```
survfit(Surv(time, status == 1) ~ 1,
        data = lung)   # censoring = event
```

→ **588 days** (511 – 1010). Report it with every survival analysis.

CC BY SA Anthony Romero-Cerdán • anthonyromero.netlify.app • Learn more at `vignette("survival")` • survival version 3.8-12 • Updated: 2026-10

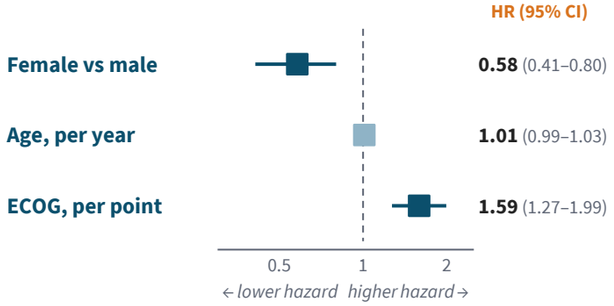
Survival analysis with survival :: CHEAT SHEET

From curves to models: hazard ratios adjusted for covariates, their assumptions, and the extensions you will meet in practice.

5 Cox regression

```
cx <- coxph(Surv(time, status) ~
  sex + age + ph.ecog, data = lung)
summary(cx)
```

The hazard ratio (HR) compares the instantaneous risk of the event, at any time, between groups:



Read it: *at any time, women had a 42% lower hazard of death than men (HR 0.58, 95% CI 0.41–0.80), adjusting for age and ECOG.*

ALSO IN summary(cx)

Likelihood ratio test 30.5 on 3 df, p < 0.001

Concordance (C) 0.64: discrimination

n / events 227 / 164 (1 row with NA dropped)

TIDY TABLE OF HRs

```
broom::tidy(cx, exponentiate = TRUE,
            conf.int = TRUE)
```

ADJUSTED SURVIVAL CURVES

```
nd <- data.frame(sex = c("Male", "Female"),
                 age = 62, ph.ecog = 1)
survfit(cx, newdata = nd)
```

→ **1-year survival 33% vs 53%** for a 62-year-old with ECOG 1.

REFERENCE LEVEL & OVERALL TESTS

`relevel(sex, ref = "Female")` flips the HR (1/0.58 = 1.7)

`anova(cx)` sequential LR test per term, e.g. for a factor with > 2 levels

6 Check the assumptions

Cox assumes **proportional hazards**: the HR is constant over follow-up. Test it with scaled Schoenfeld residuals.

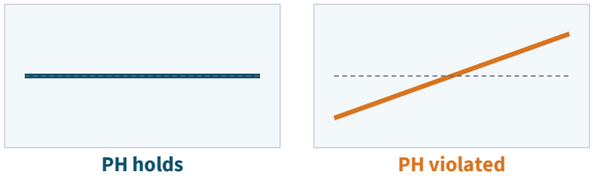
```
zph <- cox.zph(cx)
zph
```

	chisq	df	p
sex	2.31	1	0.13
age	0.19	1	0.66
ph.ecog	2.05	1	0.15
GLOBAL	**4.46**	**3**	**0.22**

All p > 0.05 → no evidence against PH. A small p flags that variable.

LOOK, DON'T ONLY TEST

```
plot(zph[1])    # residuals of sex vs time
```



Schematic: a flat smooth line means a constant effect; a trend means the HR changes with time.

IF PH FAILS FOR A VARIABLE

- `strata(ph.ecog)` separate baseline hazard per level; no HR for it, the others stay adjusted.
- `tt(age)` let the effect change with time, e.g. `tt = function(x, t, ...) x * log(t)`
- `survSplit(..., cut = 180)` split follow-up and estimate one HR per period.

```
coxph(Surv(time, status) ~ sex + age +
  strata(ph.ecog), data = lung)
```

7 Beyond the basics

TIME-VARYING COVARIATES

Values that change during follow-up (lab results, treatment) need one row per interval: (tstart, tstop].

```
temp <- subset(pbc, id <= 312,
  select = c(id:sex, stage))
pbc2 <- tmerge(temp, temp, id = id,
  death = event(time, status == 2))
pbc2 <- tmerge(pbc2, pbcseq, id = id,
  bili = tdc(day, bili))
coxph(Surv(tstart, tstop, death) ~
  log(bili), data = pbc2)
```

→ **HR 3.94** per log-unit of current bilirubin.

NON-LINEAR EFFECTS

```
coxph(Surv(time, status) ~ sex +
  pspline(age) + ph.ecog, data = lung)
```

Tests linear and non-linear parts (age: non-linear p = 0.41).

COMPETING RISKS

If another event prevents the one you study (death before progression), use a **factor** as status: first level = censored.

```
etime <- with(mgus2, ifelse(pstat == 0,
  futime, ptime))
event <- with(mgus2, ifelse(pstat == 0,
  2 * death, 1))
event <- factor(event, 0:2, labels =
  c("censor", "pcm", "death"))
cr <- survfit(Surv(etime, event) ~ 1,
              data = mgus2)
summary(cr, times = 120)$pstate
```

→ **10-year CIF: PCM 6.4%, death 53%** (Aalen–Johansen).

Regression on the CIF (Fine–Gray): `finegray()` then `coxph(..., weight = fgwt)`.

CLUSTERED OR RECURRENT DATA

```
coxph(Surv(time, status) ~ sex + age,
      cluster = inst, data = lung)
```

Adds robust (sandwich) SEs for patients nested in centres or repeated events per subject.

8 AFT models & pitfalls

ACCELERATED FAILURE TIME (survreg)

```
wb <- survreg(Surv(time, status) ~
  sex + age + ph.ecog, data = lung,
  dist = "weibull")
exp(coef(wb))   # time ratios
```

→ **Time ratio 1.49:** women's survival time is stretched by 49%. Here > 1 is protective (opposite to HR).

`dist =` "weibull" · "exponential" · "lognormal" · "loglogistic"

COMMON PITFALLS

- **Too many covariates.** Aim for ≥ 10 events per parameter: 164 events support about 16 parameters (count events, not patients).
- **HR ≠ RR.** An HR of 0.58 is a hazard (rate) ratio, not a 42% lower cumulative risk.
- **Immortal time bias.** Never classify exposure using information from after time zero (e.g. "ever treated"). Use a time-varying covariate.
- **1 − KM with competing risks** overestimates the cumulative incidence. Use the factor status (see 7).
- **Silent row drops.** Rows with missing covariates are dropped: check n and events in summary().
- **Ties.** coxph() uses the Efron method by default; Stata defaults to Breslow, so results differ slightly.

CHECKLIST Report: n, events, median follow-up, KM curves with numbers at risk, HRs with 95% CI, and how PH was checked.

COMPANION PACKAGES

survminer · ggsurvfit KM plots with risk tables

gtsummary tbl_regression(cx, exponentiate = TRUE)

cmprsk · tidycmprsk Gray's test, CIF regression

flexsurv flexible parametric models

CC BY SA Anthony Romero-Cerdán • anthonyromero.netlify.app • Learn more at `vignette("survival")` • survival version 3.8-12 • Updated: 2026-10