Survival Models for the Social and Political Sciences
Week 1: Introduction and Overview

JEFF GILL
Professor of Political Science
Professor of Biostatistics
Professor of Surgery (Public Health Sciences)
Washington University, St. Louis

Background

- Common questions in social sciences and epidemiology include: how long will a process continue, how long until termination/death, and do some groups persevere longer than others.
- In some clinical trials we are interested in the time from randomization to some "critical event."
- Usually: death, remission, gestation, full healing, dissolution, and so on.
- Example: Pearson *et al.* (2007) run a randomized clinical trial in children with neuroblastoma comparing two chemotherapy regimes. Endpoints of interest are then survival types following a relapse: response, progression, and death.

Two Views of Time

- Calender Time: reflects the actual dates of events in the study, meaning that actor of interest enters and exits are recorded by calendar date.
- Patient Time: the schedule that an individual actor of interest sees as he/she/it enters and exits the study. Starts at zero.
- Patient time is generally more important, especially in the regression sense.
- This distinction can be important when not all subjects start at the same exact time, which is relatively common.

Censored Data

- Right Censoring occurs when the unit of analysis leaves the study for reasons unrelated to the study:
  - a sample is lost or spoiled,
  - the patient lives to the end of the trial and is no longer followed,
  - the patient drops out of the study,
  - the patient dies from a cause unrelated to the study,
  - there is no more data collected on this country/group/person.
- Left censoring occurs when the event happens before the start date of the study, and is usually less of an issue.
- Censoring is critical to almost all survival models.

Quick Example

- Chant *et al.* (1984) enrolled 108 appendectomy patients comparing two drugs intended to minimize postoperative wound infection, *metronidazole* and *ampicilin*, from the date of surgery to the date of fever resolution.
- There was no censoring and durations were summarized with the geometric mean of each group, $g(\mathbf{X}) = \sqrt[n]{x_1 x_2 \cdots x_n}$, which is more robust when the log distribution of the data is more normal-like than the distribution of the data.
- Results: $g(metronidazole) = 3.5, \qquad g(ampicilin) = 3.0, \qquad t = 2.45,\ df = 106,\ p = 0.014.$

Tool: The Kaplan-Meier Curve

- Basic idea: count the number of terminations at some point of interest and divide by the number still unterminated.
- Graph looks like stairs.
- KM is most useful when comparing two regimes, i.e. treatment and control.
- Example: Hawthorne *et al.* (1992) conducted a randomized clinical trial of 67 ulcerative colitis (inflammatory bowel disease that affects the large intestine and rectum) patients with 2 months of remission while taking *azathiprine*.
- Randomization: continue with *azathioprine* or placebo.
- The following figure shows that at every time-point the treatment group has a higher proportion in remission.

Two Treatment Regimes

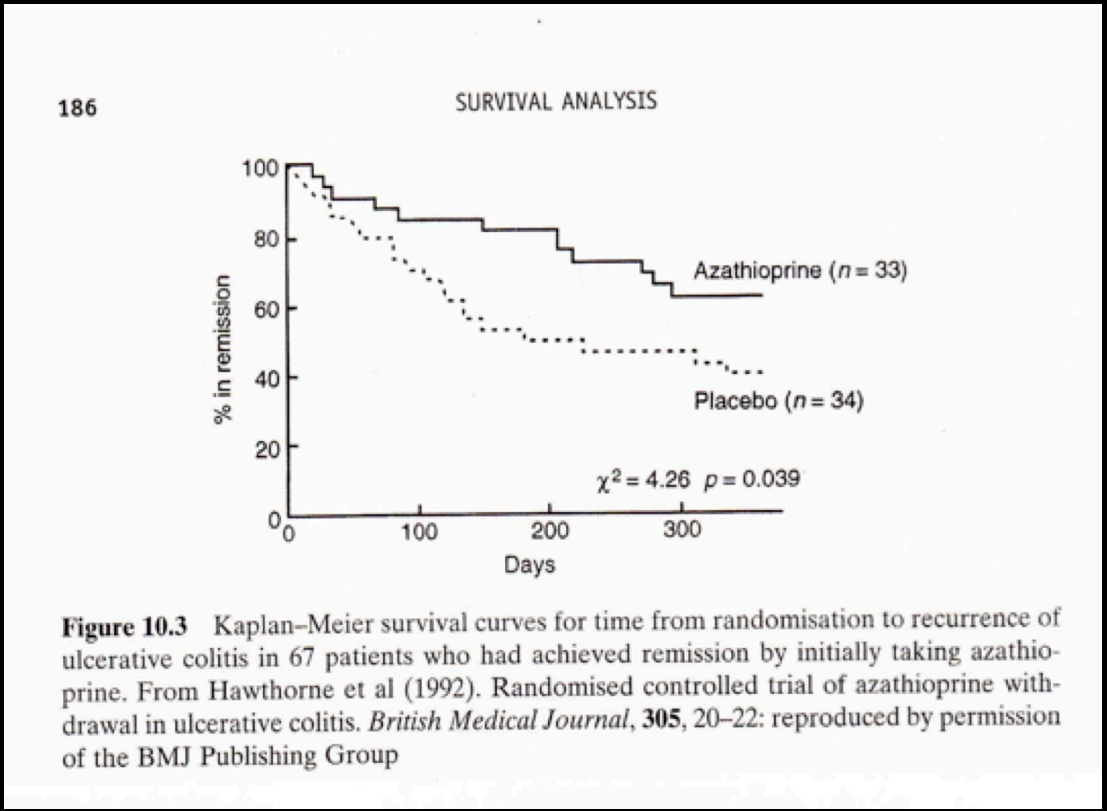


Figure 10.3 Kaplan–Meier survival curves for time from randomisation to recurrence of ulcerative colitis in 67 patients who had achieved remission by initially taking azathioprine. From Hawthorne et al (1992). Randomised controlled trial of azathioprine withdrawal in ulcerative colitis. *British Medical Journal*, **305**, 20–22: reproduced by permission of the BMJ Publishing Group

Kaplan-Meier Overview

- If there are no "left-censored" cases then the KM curve for n subjects starts at time 0 with value 1 (100% alive).
- It then continues horizontally until the first event, dropping by k/n, for k deaths at that time.
- This continues until the "curve" hits 0 or the study ends.
- The value k is recorded when the measured time is less granular, say days instead of minutes, for time of death.
- If there are censored values in the middle of the recorded period (these are not included in k), then these only affect the denominator n, giving uneven stairstep values.

Computational Details

- Define a set of consecutive times at which the data are observed:

$$t_0, t_1, t_2, \ldots, t_{\text{finish}}$$

- The time t_0 is the starting point of the study.
- Here T is the *random variable* for the point of interest indicating termination.
- There is an important quantity to consider:

$$p(T \geq t_i | T > t_{i-1})$$

  which is read: "the probability of survival to time i or later, given survival past time $i-1$."

- The conditional is easy to observe since it is the number observed alive at current $t-1$.
- We will now make extensive use of the definition of conditional probability:

$$p(A|B) = \frac{p(A,B)}{p(B)} \qquad\qquad p(A|B)p(B) = p(A,B)$$

Computational Details

- Suppose we want the *unconditional* probability of survival at a specific time: $p(T \geq t_i)$.
- We can use conditional probability to get this unconditional:

$$p(T \geq t_i | T > t_{i-1}) p(T > t_{i-1}) = p(T \geq t_i, T > t_{i-1}) = p(T \geq t_i)$$

since the joint probability has redundant information.

- This is the probability of surviving to t_i or later *given* the case has survived past t_{i-1} times the probability of surviving past t_{i-1}.
- We can obviously take this back in time one step recursively:

$$p(T \geq t_i | T > t_{i-1}) p(T \geq t_{i-1} | T > t_{i-2}) p(T > t_{i-2}) = p(T \geq t_i)$$

since

$$p(T > t_{i-1}) = p(T \geq t_{i-1} | T > t_{i-2}) p(T > t_{i-2})$$

by the same logic.

Computational Details

- Why not take it back in time to the beginning of the study?

$$
\begin{aligned}
& p(T \geq t_i | T > t_{i-1}) p(T \geq t_{i-1} | T > t_{i-2}) p(T \geq t_{i-2} | T > t_{i-3}) p(T \geq t_{i-3} | T > t_{i-4}) \\
& \times p(T \geq t_{i-4} | T > t_{i-5}) p(T \geq t_{i-5} | T > t_{i-6}) p(T \geq t_{i-6} | T > t_{i-7}) p(T \geq t_{i-7} | T > t_{i-8}) \\
& \vdots \\
& \times p(T \geq t_2 | T > t_1) p(T \geq t_1 | T > t_0) p(T > t_0) = p(T \geq t_i)
\end{aligned}
$$

  meaning that the case has to survive consecutive time periods to be considered in the the probability calculation for the subsequent time point.

- This gives us RESULT 1:
  the unconditional probability of interest is the product of all cumulative previous conditionals.

- Plus we know that $p(T > t_0) = 1$, unless the study is badly designed or there is left censoring (to be worried about later).

Computational Details

- So at each time t_i, define:

$$\begin{aligned} d_i &= \text{ the number of deaths/terminations} \\ c_i &= \text{ the number of censorings/drop-outs} \end{aligned}$$

- So the cases for whom the unconditional makes sense are:

$$\begin{aligned} n_i &= n_{i-1} - d_{i-1} - c_{i-1} \\ &= \text{ number in previous period minus deaths minus drop-outs} \end{aligned}$$

- Importantly n_i is called "the number at risk at time i."

Computational Details

- The probability of terminating at time t_i given the case made it to t_i then is:

$$p(T = t_i) = p(T < t_{i+1} | T > t_{i-1}) = \frac{d_i}{n_i}$$

note the use of strictly "greater than" and "less than."

- Subjects necessarily either survive or terminate at time t_i, so:

$$p(T < t_{i+1} | T > t_{i-1}) + p(T \geq t_i | T > t_{i-1}) = 1$$

$$\text{probability of termination} + \text{ probability of survival} = 1$$

$$\frac{d_i}{n_i} + p(T \geq t_i | T > t_{i-1}) = 1$$

- Which gives us RESULT 2:

$$p(T \geq t_i | T > t_{i-1}) = 1 - \frac{d_i}{n_i}$$

Computational Details

- RESULT 1: $p(T \geq t_i)$ is the product of all cumulative conditionals that come before t_i.
- RESULT 2: $p(T \geq t_i | T > t_{i-1}) = 1 - \frac{d_i}{n_i}$.
- Putting the two results together produces:

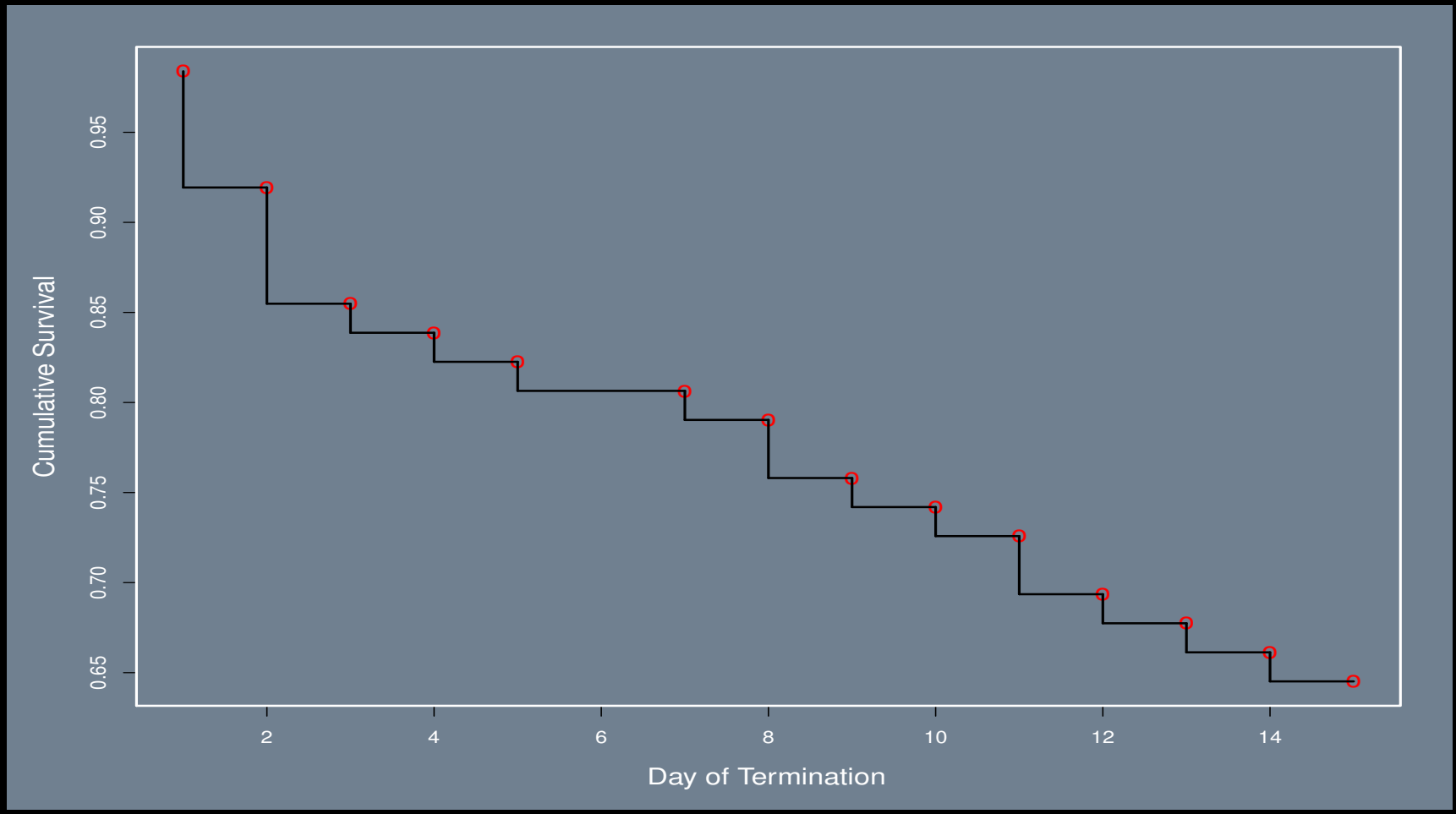
$$p(T \geq t_i) = \prod_{j=1}^{i} \left(1 - \frac{d_j}{n_j}\right)$$

- So we have accounted for the list of conditional distributions specified before in detail.
- Note that this is the logic behind the Kaplan-Meier curve.

Example Data: Duration of a Government

t_i	d_i	n_i	$1 - d_i/n_i$	cumulative product
1	1	62	0.984	0.984
2	4	61	0.934	0.919
3	4	57	0.930	0.855
4	1	53	0.981	0.839
5	1	52	0.981	0.823
7	1	51	0.980	0.806
8	1	50	0.980	0.790
9	2	49	0.966	0.758
10	1	47	0.979	0.742
11	1	46	0.978	0.726
12	2	45	0.956	0.694
13	1	43	0.977	0.677
14	1	42	0.976	0.661
15	1	41	0.976	0.645

Kaplan-Meier Curve, Unconditional Survival Probabilities



Example Data: Duration of a Government

- These calculations are done serially over time:

t_i	d_i	n_i	$1 - \frac{d_i}{n_i}$	cumulative product
1	1	62	0.984	0.984
2	4	61	0.934	0.919
3	4	57	0.930	0.855
:	:	:	:	:

- Sometimes authors want to do this in the other direction with a sum:

t_i	d_i	n_i	$\frac{d_i}{n_i}$	cumulative sum
1	1	62	0.0161	0.0161
2	4	61	0.0655	0.0816
3	56	57	0.9824	1.0640
:	:	:	:	:

R Code:

```
tj <- seq(1,15,length=15)[-6]   # NO EVENT AT TIME PERIOD 6
d <- c(1,4,4,1,1,1,1,2,1,1,2,1,1,1)
n <- c(62,61,57,53,52,51,50,49,47,46,45,43,42,41)
risk <- 1 - d/n
survivor <- cumprod(risk)
postscript("./Images/km-fig.ps")
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white", col="white",bg="slategray", cex.lab=1.3)
plot(tj,survivor,pch="o",col="red",cex=1.3,
    xlab="Day of Termination",ylab="Cumulative Survival")
for (i in 2:length(tj))  {
    segments(tj[i-1],survivor[i-1],tj[i-1],survivor[i])
    segments(tj[i-1],survivor[i],tj[i],survivor[i])
}
dev.off()
```

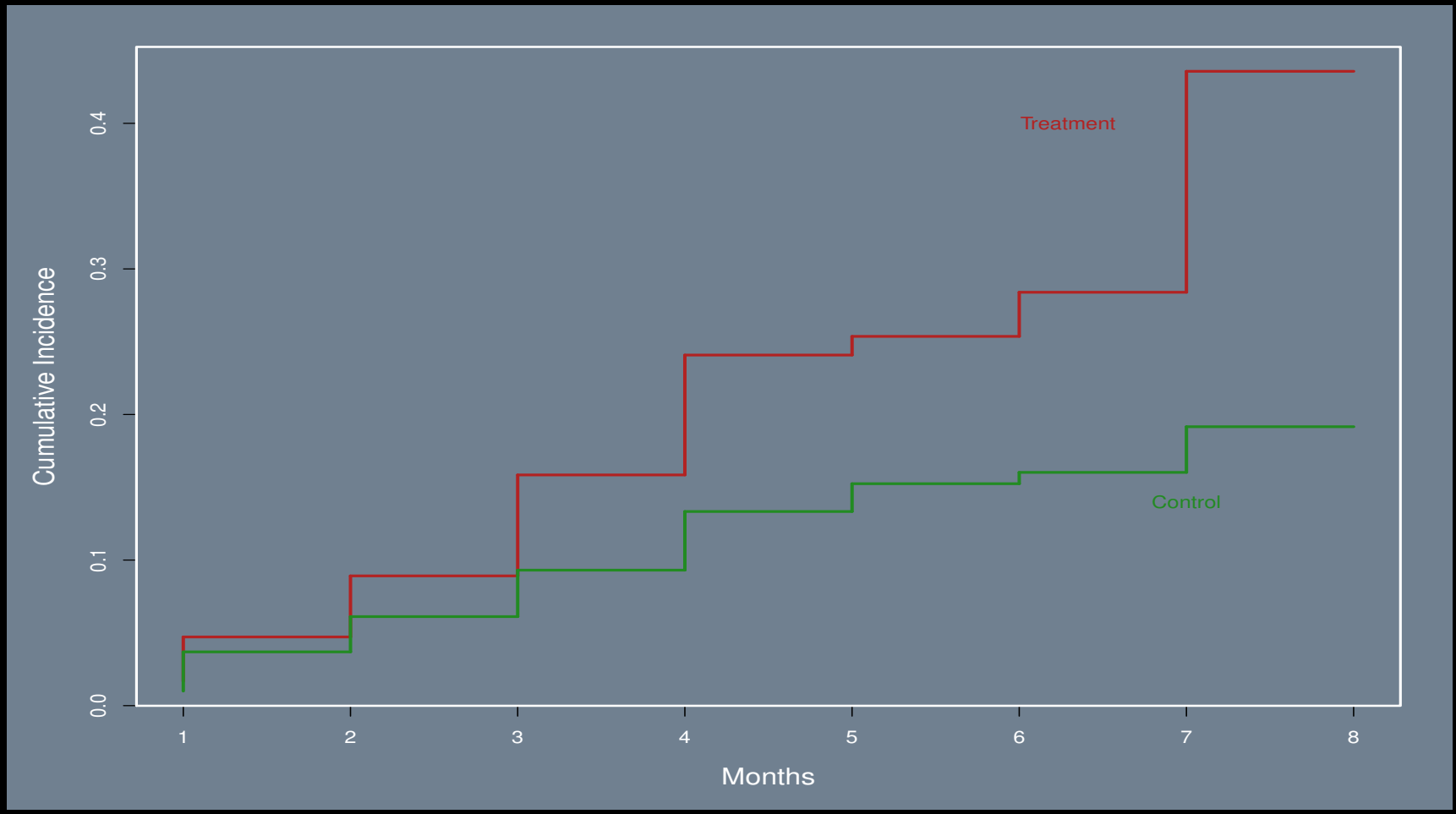
Another Direction: Events Accumulate

- Suppose we track terminations in the other direction for two groups: treatment (new drugs) and control (old drugs).
- Small dataset, onset of AIDS from HIV patients:

t_i	$d_i(t,c)$	$n_i(t,c)$	$\frac{d_i}{n_i}(t,c)$	cumulative incidence(t,c)
1	$5,3$	$300,300$	$0.017, 0.010$	$0.017, 0.010$
2	$9,8$	$295,297$	$0.031, 0.027$	$0.047, 0.037$
3	$12,7$	$286,289$	$0.042, 0.024$	$0.089, 0.061$
4	$19,9$	$274,282$	$0.069, 0.032$	$0.158, 0.093$
5	$21,11$	$255,273$	$0.082, 0.040$	$0.241, 0.133$
6	$3,5$	$234,262$	$0.013, 0.019$	$0.254, 0.152$
7	$7,2$	$231,257$	$0.030, 0.008$	$0.284, 0.160$
8	$34,8$	$224,255$	$0.152, 0.031$	$0.436, 0.192$

- Where (t,c) denotes treatment and control.

Kaplan-Meier Curve, Cumulative Incidences



R Code:

```
tj <- seq(1,8,length=8); d <- c(5,9,12,19,21,3,7,34); n <-  300-cumsum(c(0,d[-8]))
incidence <- d/n; cum.incidence <- cumsum(incidence)
postscript("./km-fig-up.ps")
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white", col="white",bg="slategray", cex.lab=1.3)
plot(tj,cum.incidence,type="n",xlab="Months",ylab="Cumulative Incidence",cex=2)
for (i in 2:length(tj))  {
    segments(tj[i-1],cum.incidence[i-1],tj[i-1],cum.incidence[i],
    col="firebrick",lwd=2.5)
    segments(tj[i-1],cum.incidence[i],tj[i],cum.incidence[i],
    col="firebrick",lwd=2.5)
}
```

R Code

```
text(6.3,0.4,"Treatment",col="firebrick")
tj <- seq(1,8,length=8); d <- c(3,8,7,9,11,5,2,8); n <-  300-cumsum(c(0,d[-8]));
incidence <- d/n; cum.incidence <- cumsum(incidence)
for (i in 2:length(tj))  {
    segments(tj[i-1],cum.incidence[i-1],tj[i-1],cum.incidence[i],
    col="forest green",lwd=2.5)
    segments(tj[i-1],cum.incidence[i],tj[i],cum.incidence[i],
        col="forest green",lwd=2.5)
}
text(7,0.14,"Control",col="forest green");  dev.off()
```

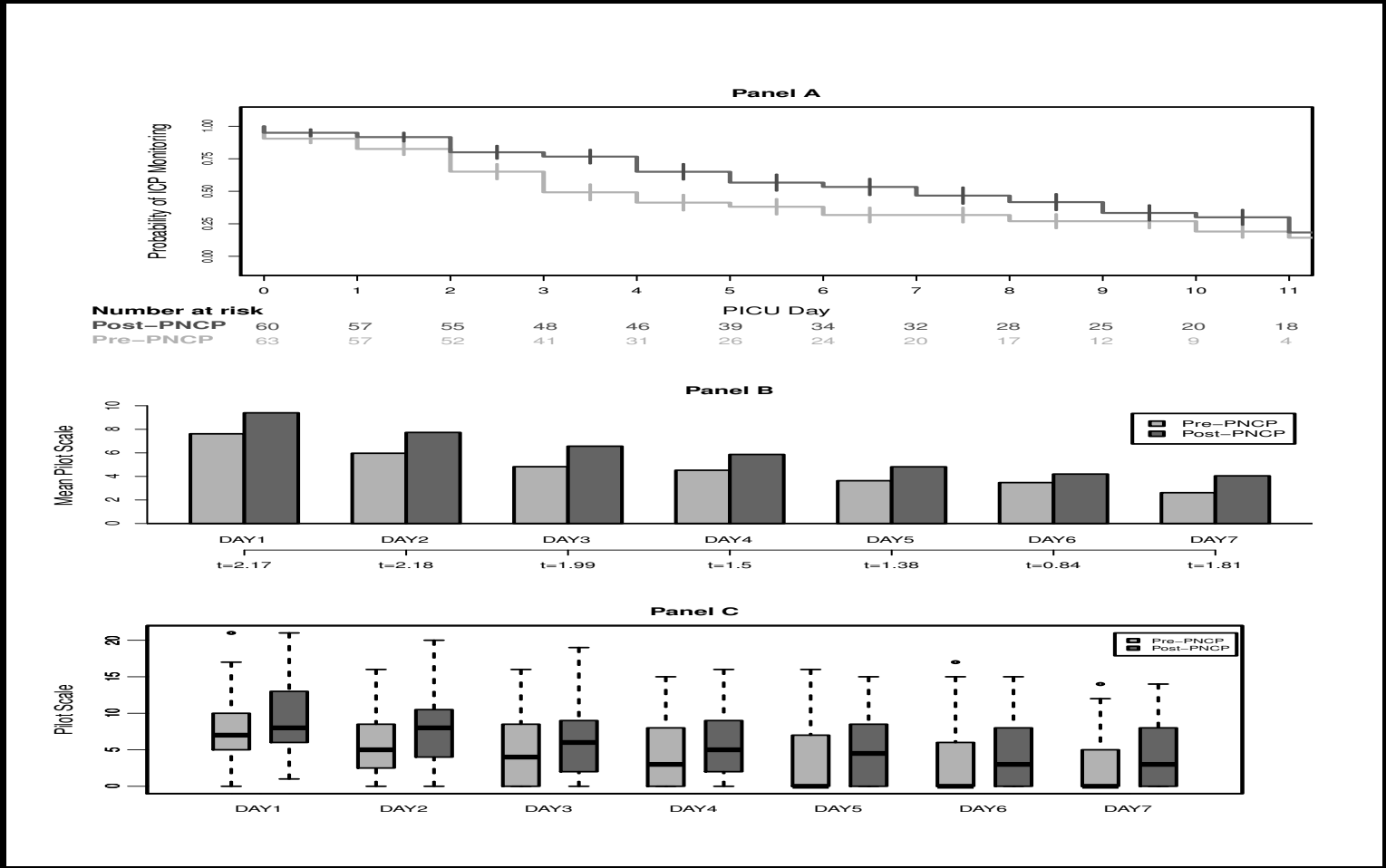
More R Code

```
nf <- layout(matrix(c(1,2,3,4,5,6),6,1,byrow=TRUE), widths=c(1,1),
    heights=c(0.35,0.10,0.05,0.25,0.05,0.40), respect=TRUE)
par(mar=c(4,3,1,1),lwd=2,oma=c(0,0,0,0), col.axis="black",col.lab="black",
    col.sub="black", col="black",bg="white", cex.lab=1.0)

# PANEL 1A: KAPLAN-MEIER SURVIVAL CURVES
library(survival)
icp.mon <- read.table("Article.PTBI/icp.mon.txt",header=TRUE)
surv.summary <- summary(survfit(Surv(time=icp.mon$Day.ICP.removed,
    origin=icp.mon$Day.ICP.placed) ~ icp.mon$Group),
    col = c("grey70","grey40"), lwd=3,xaxt="n", xlim=c(1,19))
par(mar=c(1,10,3,3),xaxs="r");
plot(survfit(Surv(time=icp.mon$Day.ICP.removed,origin=icp.mon$Day.ICP.placed)
     ~ icp.mon$Group), col = c("grey70","grey40"), lwd=2.5, xaxt="n", yaxt="n",
     xlim=c(-0.25,11.25),ylim=c(-0.1,1.1))
```

More R Code

```
axis(side=1,seq(0,11,by=1))
axis(side=2,c(0,0.25,0.50,0.75,1.0),cex.axis=0.8)
legend(16,0.95, legend=c("Pre-PNCP","Post-PNCP"), col=c("grey70","grey40"),
    lwd=3, box.col="black")
mtext(side=1,"PICU Day",cex=0.8,line=3)
mtext(side=2,"Probability of ICP Monitoring",cex=0.8,line=3)
title(main = list("Panel A"),cex=0.5, line=1)
for (i in 1:11)  {
      if (i <8) lines(c(i-.5,i-.5),surv.summary$surv[i]
            + c(-1,1)*surv.summary$std.err[i],col="grey70")
      if ((i==8) | (i==9)) lines(c(i-.5,i-.5),surv.summary$surv[(i-1)]
            + c(-1,1)*surv.summary$std.err[(i-1)],col="grey70")
      if ((i==10) | (i==11)) lines(c(i-.5,i-.5),surv.summary$surv[(i-2)] +
        c(-1,1)*surv.summary$std.err[(i-2)],col="grey70")
      lines(c(i-.5,i-.5),surv.summary$surv[(i+14)]
        + c(-1,1)*surv.summary$std.err[(i+14)],col="grey30")
}
```

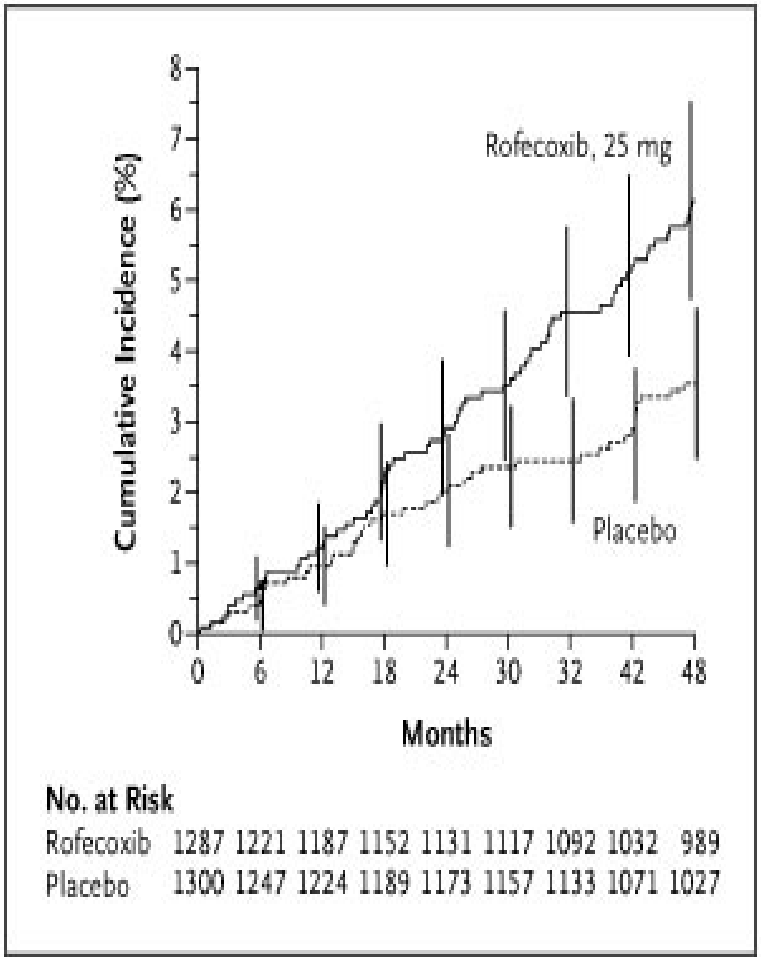
Kaplan Meier Example

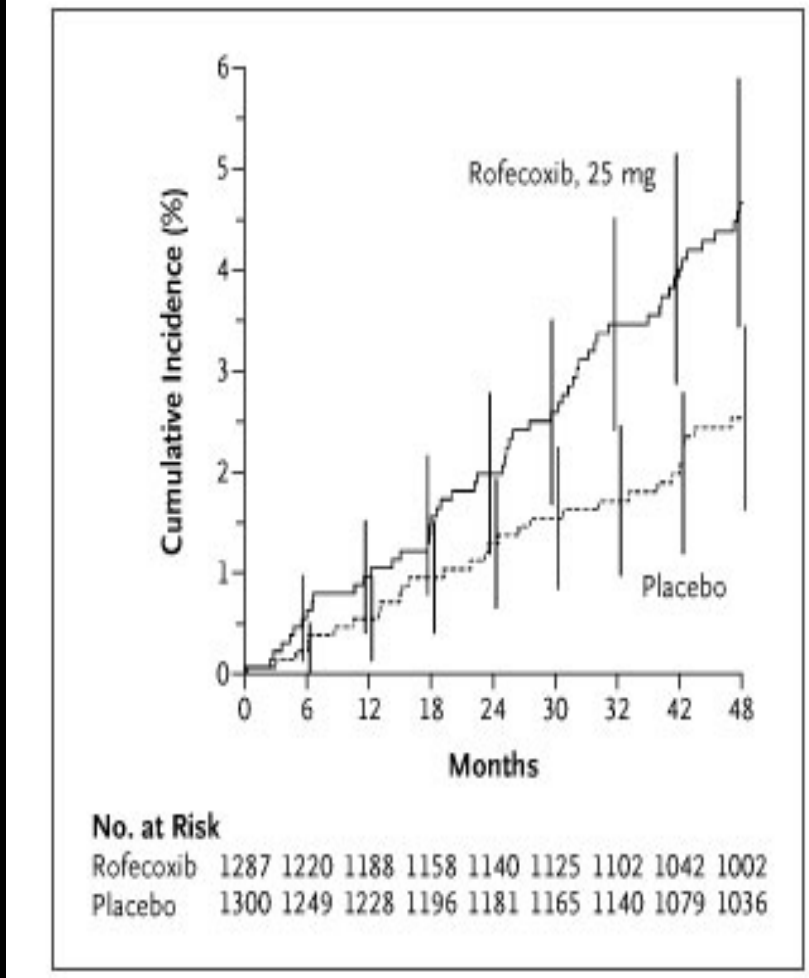


High Profile Controversy Involving Kaplan-Meier Curves

- In 2006 the NEJM published a correction to an original study of cardiovascular events associated with rofecoxib (Vioxx) versus a placebo published in 2005.
- Vioxx is a nonsteriodal anti-inflammatory for pain, particularly from arthritis.
- The incorrect sentence that was corrected:
  "The increased relative risk became apparent after 18 months of treatment; during the first 18 months, the event rates were similar in the two groups."
- This correction illustrates how Kaplan-Meier curves can be misleading and how they differ with various censoring assumptions.
- Problem: in dropping (censoring) data that occurred 14 or more days *after* subjects discontinued the study, the Kaplan-Meir curves for thrombotic events (formation of a blood clot inside a blood vessel) did not separate until 18 months.
- This censoring missed: 12 (8 treatment, 4 control) thrombotic events that occurred 14 days or more past the end of the clinical trial but within 36 months of the start of the trial (randomization).

High Profile Controversy Involving Kaplan-Meier Curves





High Profile Controversy Involving Kaplan-Meier Curves

- Merck Response to NEJM:

  Merck Stands Behind Original APPROVe Study Results Increased Relative Risk Observed Beginning After 18 Months

  WHITEHOUSE STATION, N.J., June 26, 2006 - Merck & Co., Inc. reaffirmed in an open letter today to the scientific community that it stands behind the original results of the APPROVe study published in the New England Journal of Medicine (NEJM) in 2005, in which there was an increased relative risk for confirmed thrombotic cardiovascular events for VIOXX compared to placebo beginning after 18 months of continuous daily treatment. The Company's letter follows a correction notice published by the NEJM editors.

- The implication is that physicians can prescribe VIOXX safely up to 18 months, which is not supported from the uncensored Kaplan-Meier graph.

- Result: billions of dollars in litigation.

Acute Myelogenous Leukemia Data

- AML, also known as acute nonlymphocytic leukemia, represents a group of clonal hematopoietic stem cell disorders in which both failure to differentiate and overproliferation into the stem cell compartment result in the accumulation of myeloblasts.
- The patients were also factored into 2 groups according to the presence or absence of a morphologic characteristic of white blood cells. Patients termed `ag` positive were identified by the presence of Auer rods (an abnormal, needle-shaped or round, pink-staining inclusion) and/or significant granulation of the leukemic cells in the bone marrow at the time of diagnosis.
- `wbc`: white blood count (usually between 4,300 and 10,800 cells per cubic millimeter of blood).
- Low white blood cell count is called leukopenia, and high white blood cell count is termed leukocytosis.
- `time`: survival time in weeks.

Acute Myelogenous Leukemia Data

wbc	ag	time	wbc	ag	time
2300	present	65	750	present	156
4300	present	100	2600	present	134
6000	present	16	10500	present	108
10000	present	121	17000	present	4
5400	present	39	7000	present	143
100000	present	1	52000	present	5
100000	present	65	4400	absent	56
3000	absent	65	4000	absent	17
1500	absent	7	9000	absent	16
5300	absent	22	10000	absent	3
19000	absent	4	27000	absent	2
28000	absent	3	31000	absent	8
26000	absent	4	21000	absent	3
79000	absent	30	100000	absent	4
100000	absent	43			

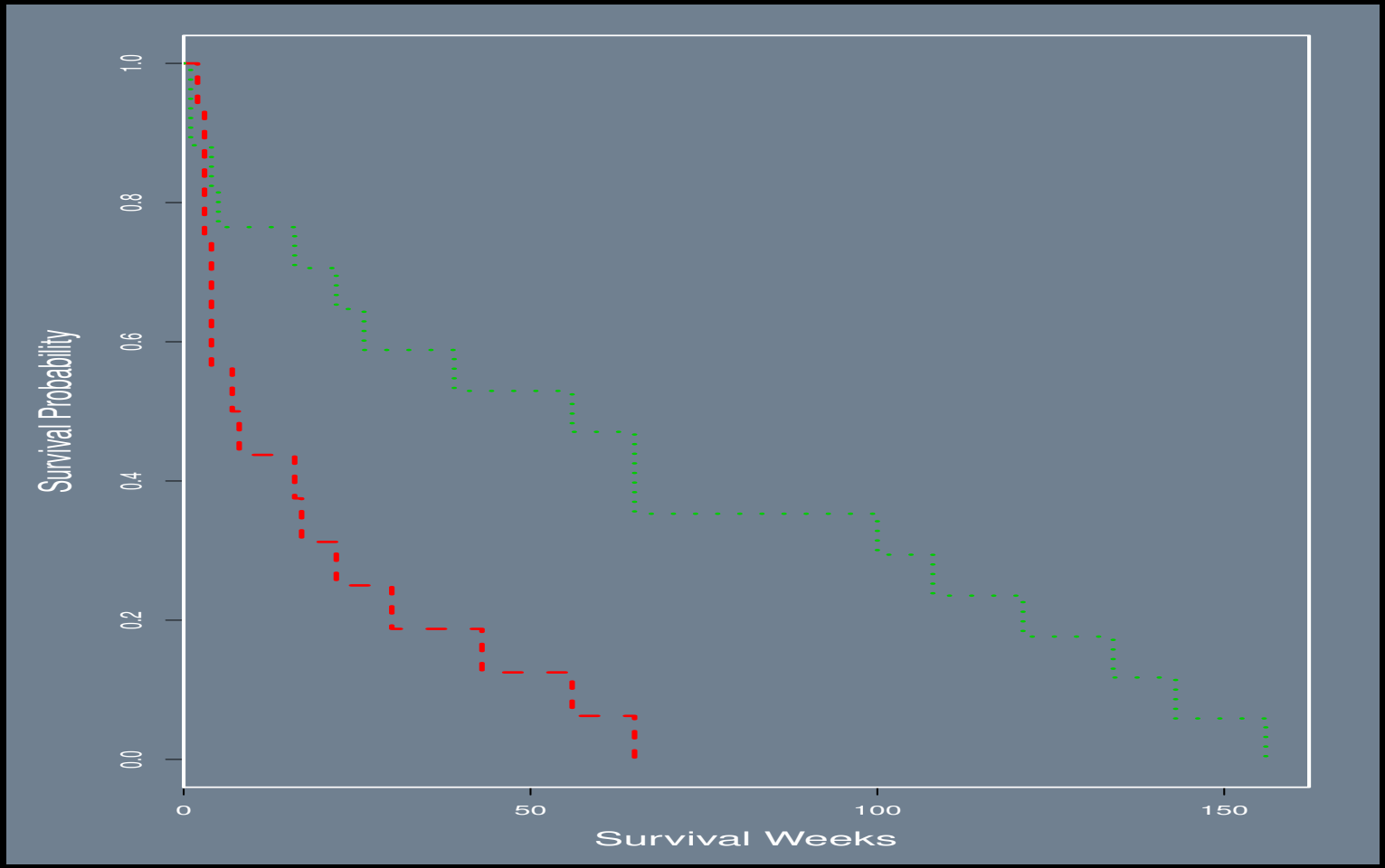
R Code

- Using regular R:

```
library(survival)
leuk <- read.table("http://jgill.wustl.edu/data/leuk.dat")
postscript("./Images/leuk-fig.ps")
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white", col="white",bg="slategray", cex.lab=1.3)
plot(survfit(Surv(time) ~ ag, data = leuk), lty = 2:3, col = 2:3, lwd=3)
legend(110, 0.9, legend=levels(leuk$ag), lty=3:2, col=3:2, lwd=3, box.col="white")
mtext(side=1,"Survival Weeks",cex=1.5,line=3)
mtext(side=2,"Survival Probability",cex=1.5,line=3)
dev.off()
```

Note the use of the modeling statement: `survfit(Surv(time) ~ ag, data = leuk)`.

Leukemia Survival



The Logrank Test

- So far we have just been looking at figures for survival data.
- The logrank test is a formal comparison of two KM curves, with a null hypothesis of no difference between group A and group B.
- Steps:
  - O_A is the total for group A, and O_B is the total for group B,
  - Under H_0 the expected deaths at time t_i for group A is: $e_{A_i} = d_i n_{A_i}/n_i$, where n_{A_i} is the number at risk in group A and n_i is the total number at risk (both at time i).
  - The total number of deaths for group A under the null hypothesis is $E_A = \sum_T e_{A_i}$.
  - The total number of deaths for group B under the null hypothesis is $E_B = \sum_T d_i - E_A$.
  - The Chi-Square statistics with $df = 1$ is:

$$\chi^2_{\text{logrank}} = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B}$$

Leukemia Survival Data in Analysis Form

	time	ag	wbc	di	ni	prob.surv	cum.prob	risk.1	exp.1
1	0	0	0	0	33	1.00000	1.00000	17	0.000000
14	1	1	100000	2	33	0.93939	0.93939	15	0.909091
15	1	1	100000	NA	32	NA	NA	15	NA
26	2	0	27000	0	31	1.00000	0.93939	15	0.000000
24	3	0	10000	0	31	1.00000	0.93939	15	0.000000
27	3	0	28000	0	31	1.00000	0.93939	15	0.000000
30	3	0	21000	0	31	1.00000	0.93939	15	0.000000
8	4	1	17000	1	31	0.96774	0.90909	14	0.451613
25	4	0	19000	0	30	1.00000	0.90909	14	0.000000
29	4	0	26000	0	30	1.00000	0.90909	14	0.000000
32	4	0	100000	0	30	1.00000	0.90909	14	0.000000
16	5	1	52000	1	30	0.96667	0.87879	13	0.433333
21	7	0	1500	0	29	1.00000	0.87879	13	0.000000
28	8	0	31000	0	29	1.00000	0.87879	13	0.000000
5	16	1	6000	1	29	0.96552	0.84848	12	0.413793
22	16	0	9000	0	28	1.00000	0.84848	12	0.000000
20	17	0	4000	0	28	1.00000	0.84848	12	0.000000

Leukemia Survival Data in Analysis Form

	time	ag	wbc	di	ni	prob.surv	cum.prob	risk.1	exp.1
13	22	1	35000	1	28	0.96429	0.81818	11	0.392857
23	22	0	5300	0	27	1.00000	0.81818	11	0.000000
12	26	1	32000	1	27	0.96296	0.78788	10	0.370370
31	30	0	79000	0	26	1.00000	0.78788	10	0.000000
9	39	1	5400	1	26	0.96154	0.75758	9	0.346154
33	43	0	100000	0	25	1.00000	0.75758	9	0.000000
11	56	1	9400	1	25	0.96000	0.72727	8	0.320000
18	56	0	4400	0	24	1.00000	0.72727	8	0.000000
110	65	1	2300	2	24	0.91667	0.66667	6	0.500000
17	65	1	100000	NA	23	NA	NA	6	NA
19	65	0	3000	0	22	1.00000	0.66667	6	0.000000
3	100	1	4300	1	22	0.95455	0.63636	5	0.227273
6	108	1	10500	1	21	0.95238	0.60606	4	0.190476
7	121	1	10000	1	20	0.95000	0.57576	3	0.150000
4	134	1	2600	1	19	0.94737	0.54545	2	0.105263
10	143	1	7000	1	18	0.94444	0.51515	1	0.055556
2	156	1	750	1	17	0.94118	0.48485	0	0.000000

Logrank Test with the Acute Myelogenous Leukemia Data

```
( E.A <- sum(leuk$exp.1, na.rm=TRUE) )
[1] 4.8658
( E.B <- sum(leuk$di, na.rm=TRUE) - E.A )
[1] 12.134
( O.A <- sum(leuk$ag*leuk$di, na.rm=TRUE) )
[1] 17
( O.B <- sum(leuk$di, na.rm=TRUE) - O.A )
[1] 0
( chi2.logrank <- ((O.A-E.A)^2/E.A) + ((O.B-E.B)^2/E.B) )
[1] 42.394
pchisq(chi2.logrank, df=1, lower.tail=FALSE)
[1] 7.4604e-11
```

Hazard Ratio For Comparing Two Groups

- Observe O_A and O_B.
- H_O : no difference between groups, H_A : groups are different.
- Calculate E_A and E_B, as before:

$$E_A = \sum_T e_{A_i} \qquad E_B = \sum_T d_i - E_A.$$

- O_A/E_A is the relative death rate in group A, and O_B/E_B is the relative death rate in group B.
- The Hazard Ratio is:

$$HR = \frac{O_A/E_A}{O_B/E_B},$$

which is near 1 under the null hypothesis of no difference.

Confidence Interval for the Hazard Ratio

- The HR is skewed to the right and bounded by $[0 : \infty]$, so to make it more normal (and therefore easier to test), it is often treated as: $\log(HR)$.
- The standard error of the $\log(HR)$ is given by:

$$SE(\log(HR)) = \sqrt{\frac{1}{E_A} + \frac{1}{E_B}},$$

  and generally requires relatively large sample size.

- Therefore the 95% confidence interval for the $\log(HR)$ is:

$$[\log(HR) - 1.96 \times SE(\log(HR)) : \log(HR) + 1.96 \times SE(\log(HR))],$$

  and the 95% confidence interval for the HR is:

$$[\exp\{\log(HR) - 1.96 \times SE(\log(HR))\} : \exp\{\log(HR) + 1.96 \times SE(\log(HR))\}].$$

More Theoretical Details

- We will use a *proportional hazards model* for the critical event.
- The *hazard function* gives the proportion of cases who fail just after time t given that they have survived until time t.
- From a PDF for the event over time, $f(t)$, define the distribution function (CDF) of time t and the *survival function*:

$$F(t) = \int_0^t p(T < t)dt \qquad\qquad S(t) = p(T \geq t) = 1 - F(t).$$

The hazard function is created by:

$$h(t) = \lim_{\delta t \to 0} \left[\frac{p(t \leq T < t + \delta t | T \geq t)}{\delta t} \right]$$

(also called the *instantaneous hazard rate*, the *instantaneous death rate*, the *intensity rate*, and the *force of mortality*).

- Note that:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}(\log S(t)), \qquad S(t) = \exp(-H(t)), \quad \text{where } H(t) = \int_0^t h(u)du$$

A Regression Model for Survival Data

- The Cox Proportional Hazards Model gives a regression where the outcome variable is the instantaneous hazard rate, $h(t)$.
- For individual i in the study at time t this links $h_i(t)$ to a baseline hazard rate and covariates according to:

$$\log[h_i(t)] = \log[h_0(t)] + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}.$$

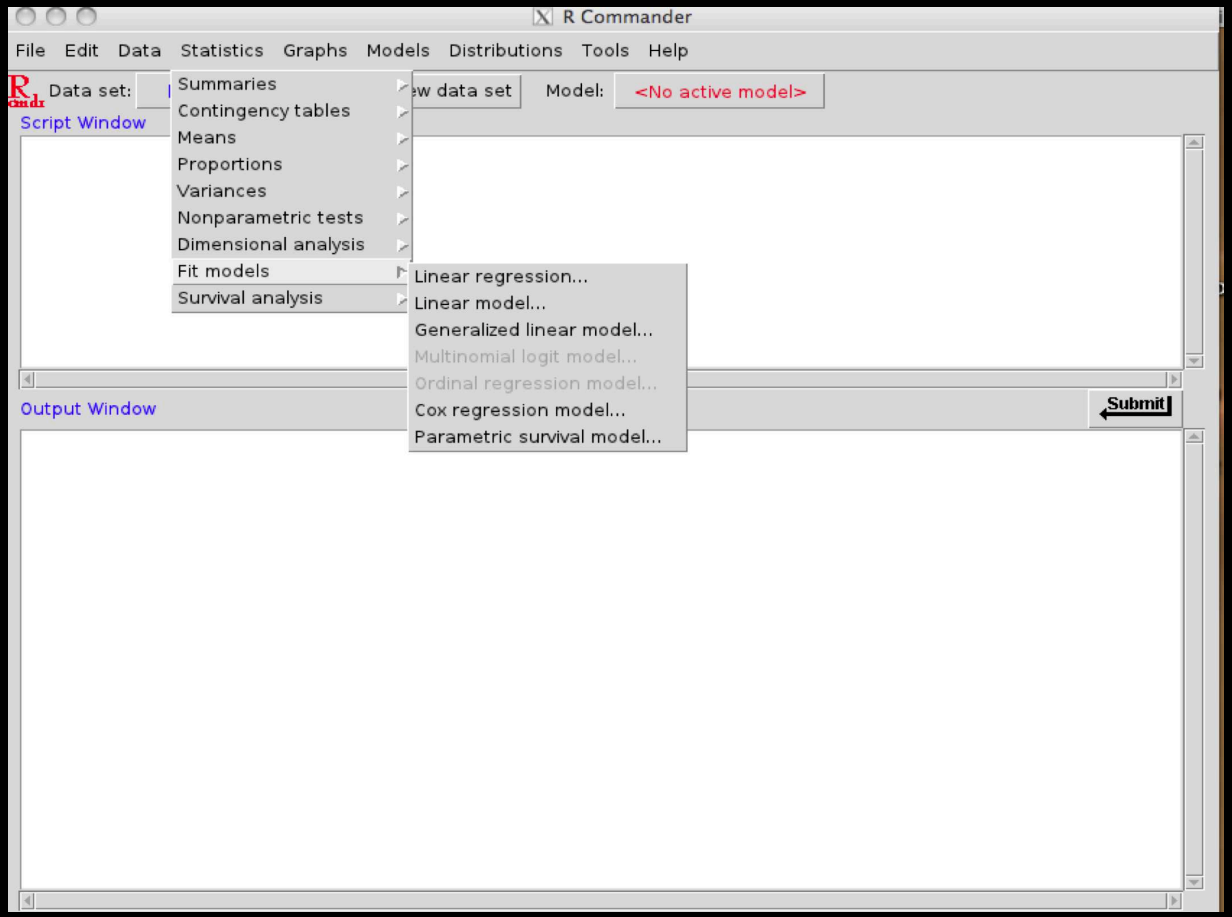
- This model can also be expressed with exponentiation:

$$h_i(t) = h_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$$

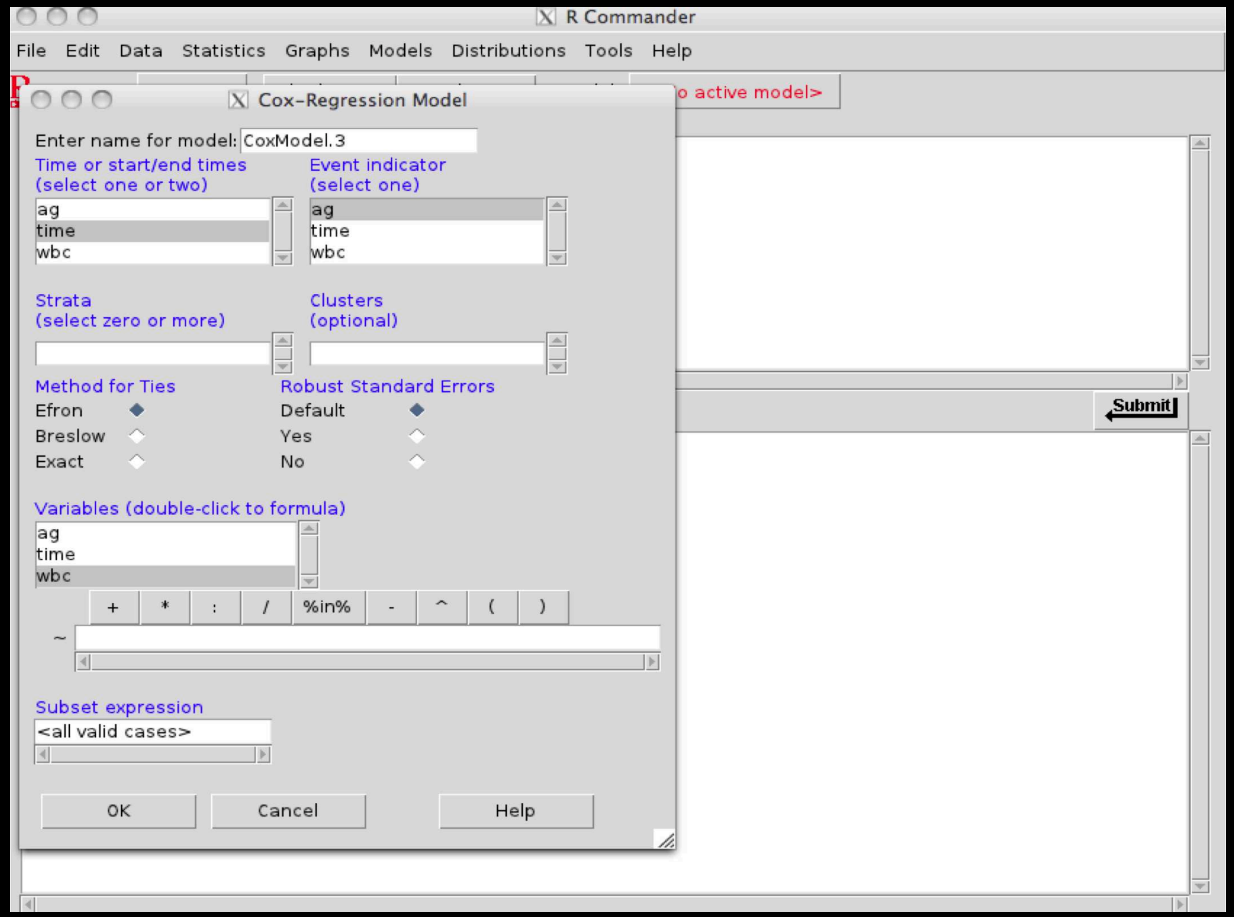
$$\frac{h_i(t)}{h_0(t)} = \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}).$$

- This latter form reveals the proportional nature of the model more directly.
- Finally, note that there is an assumption that hazard ratios between individuals are constant over time: $\frac{h_i(t)}{h_j(t)} = k$.

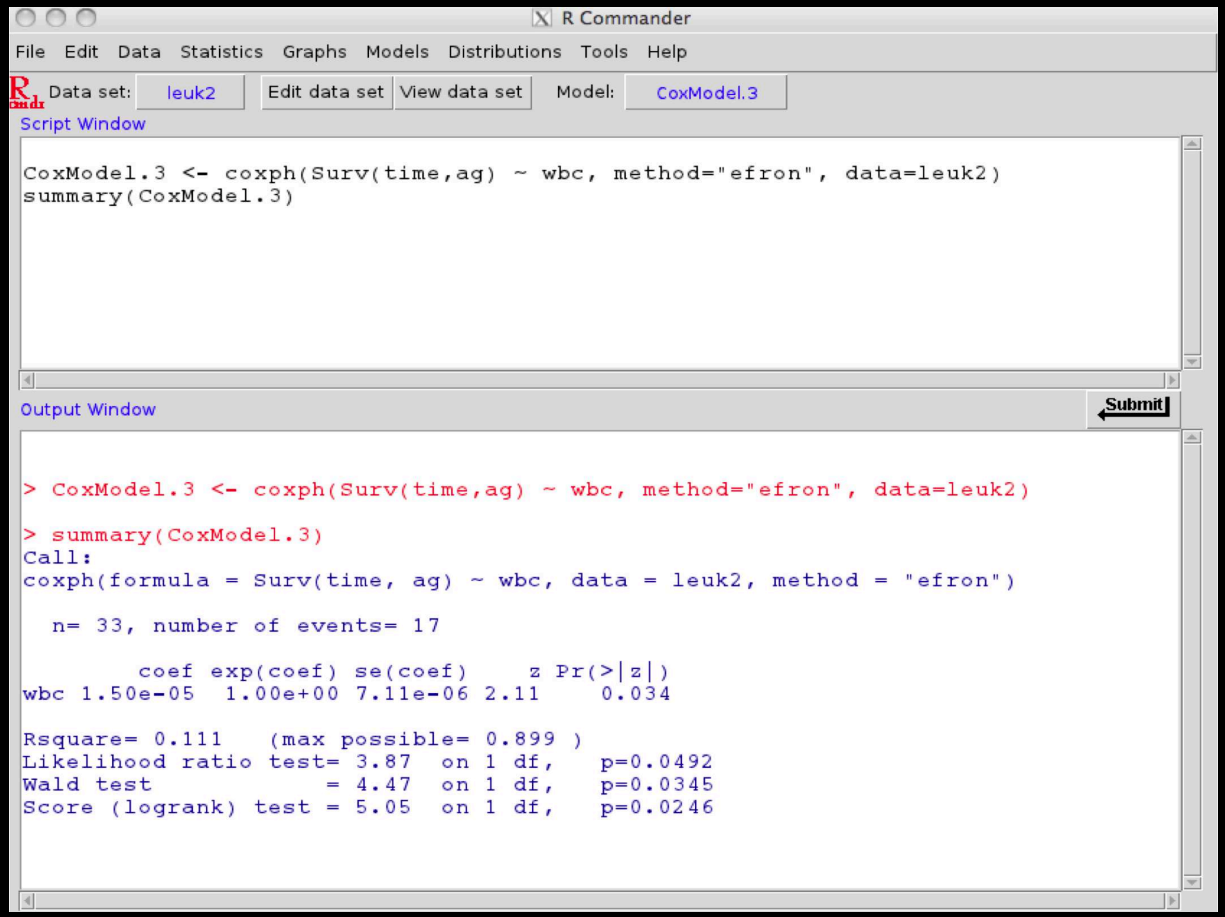
The Cox Proportional Hazards Model In `Rcmdr`, Acute Myelogenous Leukemia



The Cox Proportional Hazards Model In `Rcmdr`, Acute Myelogenous Leukemia

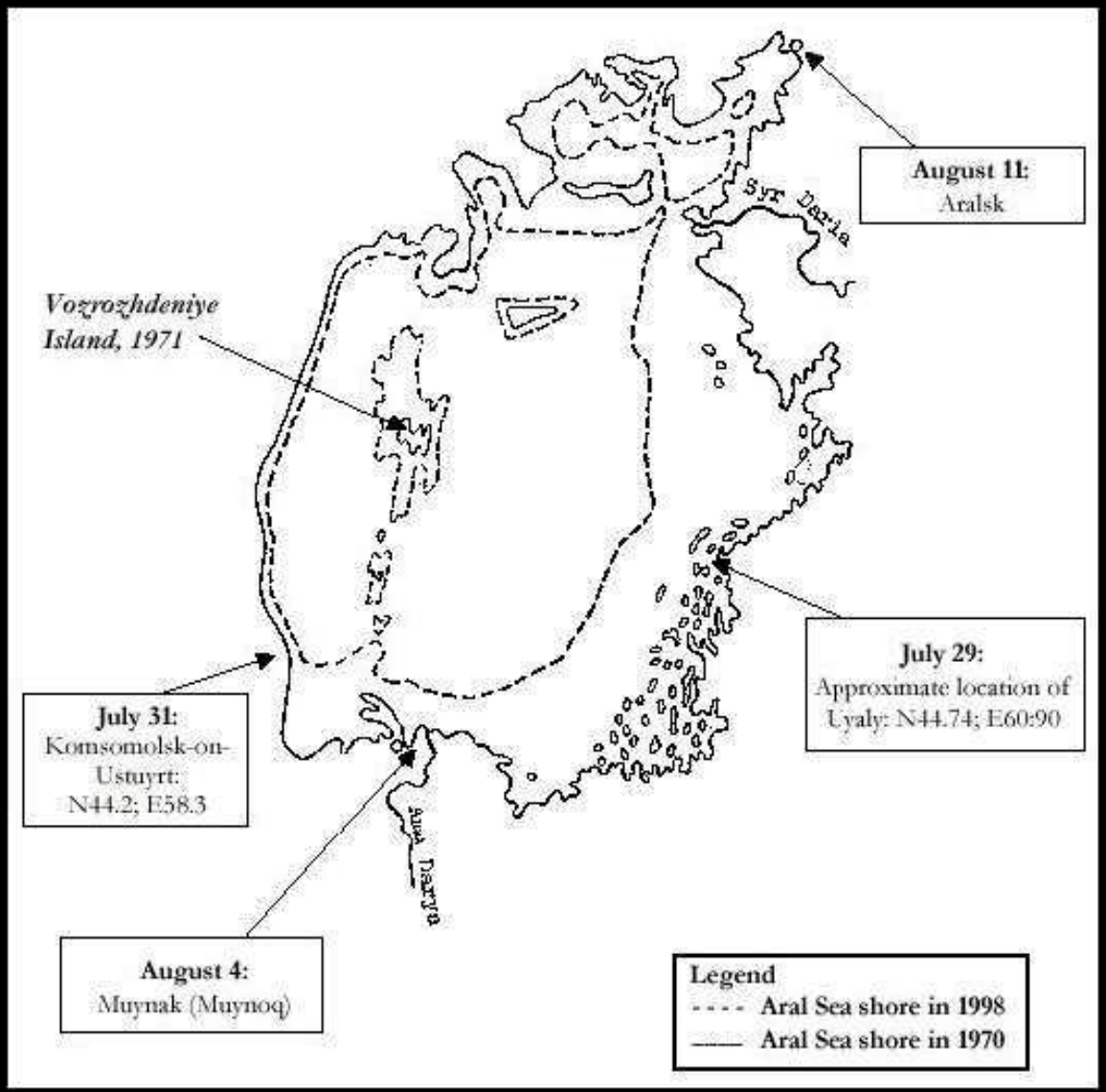


The Cox Proportional Hazards Model In `Rcmdr`, Acute Myelogenous Leukemia



The Soviet Biological Weapons Program/Smallpox Outbreak

- On July 15, 1971 the research vessel *Lev Berg* set sail from Aralsk (Kazakhstan) to survey the Aral Sea, then the 4th largest freshwater lake in the world.
- The Soviet Union had been steadily draining the Aral for agricultural purposes since the 1950s and the *Lev Berg* was to measure the ecological damage.
- This trip included passing by the island Vozrozhdeniye on the South side.



The Soviet Biological Weapons Program/Smallpox Outbreak

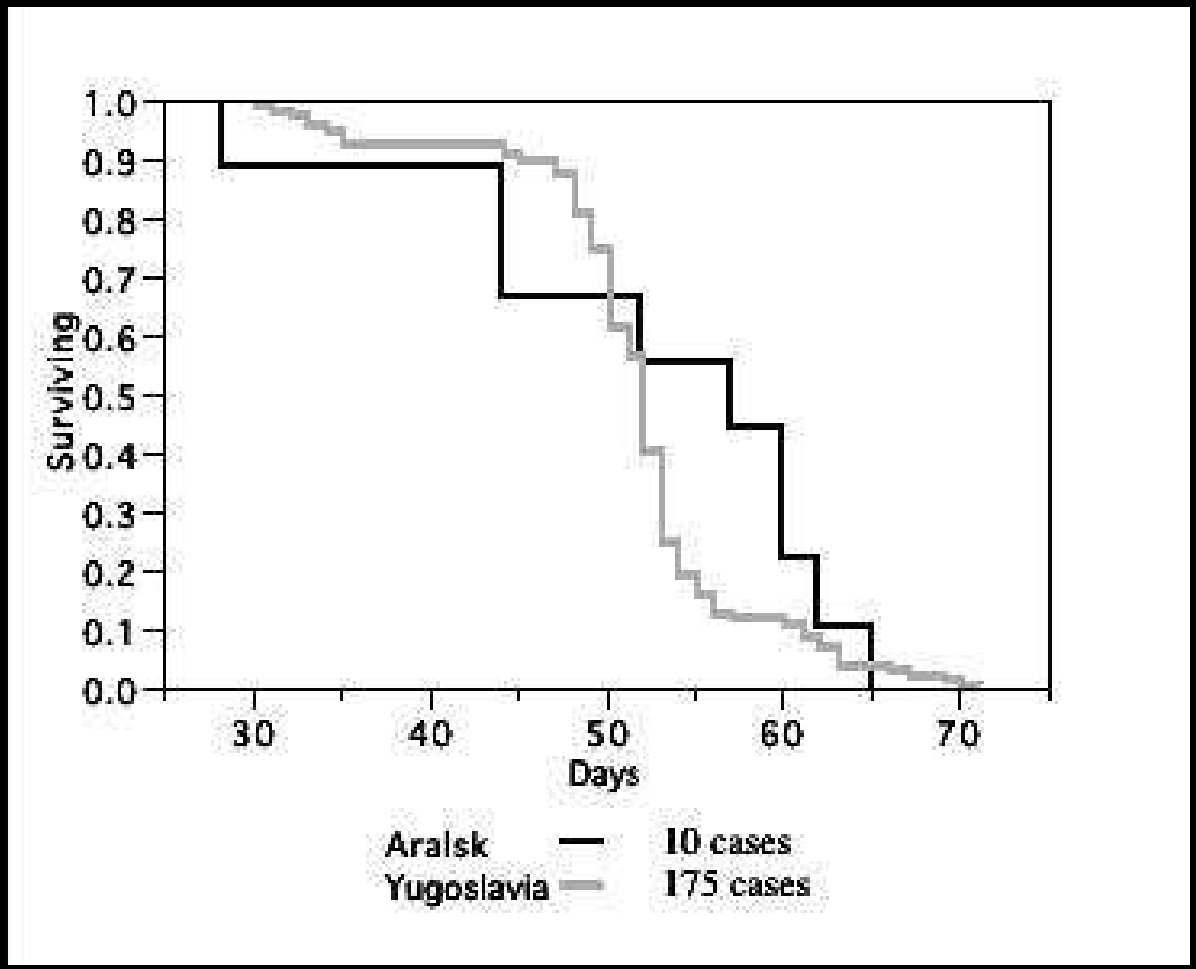
- Vozrozhdeniye was an ideal site for the main Soviet bio-weapons field testing because it was in a remote area, easily secured as an island, and had reliable winds from the North to the South allowing "safe" testing on the South end and and housing on the North end.
- The site was active from 1936 until 1990 when Yeltsin publicly denounced the program and had it shut down.
- This is despite the Soviet Union having signed the *1972 Biological and Toxin Weapons Convention* outlawing such research.
- Shortly after the *Lev Berg* returned to Aralsk, there was an unusual outbreak of smallpox there, starting with a young researcher who had been on-board.

The Soviet Biological Weapons Program/Smallpox Outbreak

	AGE	FEMALE	DAYS1971	RASH	VACC	MORT
1	24.00	1.00	224.00	2.00	1.00	0.00
2	9.00	0.00	234.00	2.00	1.00	0.00
3	23.00	1.00	253.00	3.00	0.00	1.00
4	36.00	1.00	253.00	2.00	1.00	0.00
5	5.50	0.00	261.00	2.00	1.00	0.00
6	38.00	0.00	267.00	1.00	1.00	0.00
7	0.80	0.00	269.00	3.00	0.00	1.00
8	60.00	1.00	269.00	1.00	1.00	0.00
9	33.00	0.00	271.00	1.00	1.00	0.00
10	0.33	1.00	275.00	3.00	0.00	1.00

The Soviet Biological Weapons Program/Smallpox Outbreak

- Comparison Case: in 1972 a Muslim man from Kosovo went on a pilgrimage to Mecca, returning through Baghdad where he was infected with smallpox.
- This was the first reported smallpox case in Kosovo since 1930 and it went undiagnosed for six weeks producing 175 cases and 35 deaths.
- A good comparison since rates of vaccination were similar as were socio-economic conditions.
- Kaplan-Meier graph with time-to-event = onset of illness:



The Soviet Biological Weapons Program/Smallpox Outbreak

- Key difference: all 3 Aralsk deaths were from hemorrhagic smallpox and only $5/175 = 0.0286$ in Kosovo were.
- Baseline for naturally occurring hemorrhagic smallpox: Rao's study in Madras, India had $10,857$ cases with only 240 hemorrhagic, prevalence $240/10857 = 0.0221$.
- Only two possible explanations accredited for the differences:
  1. host conditions (nutrition, genetic resistance, environment, public health) were *very* different.
  2. Aralsk strain was an unusual type of smallpox.

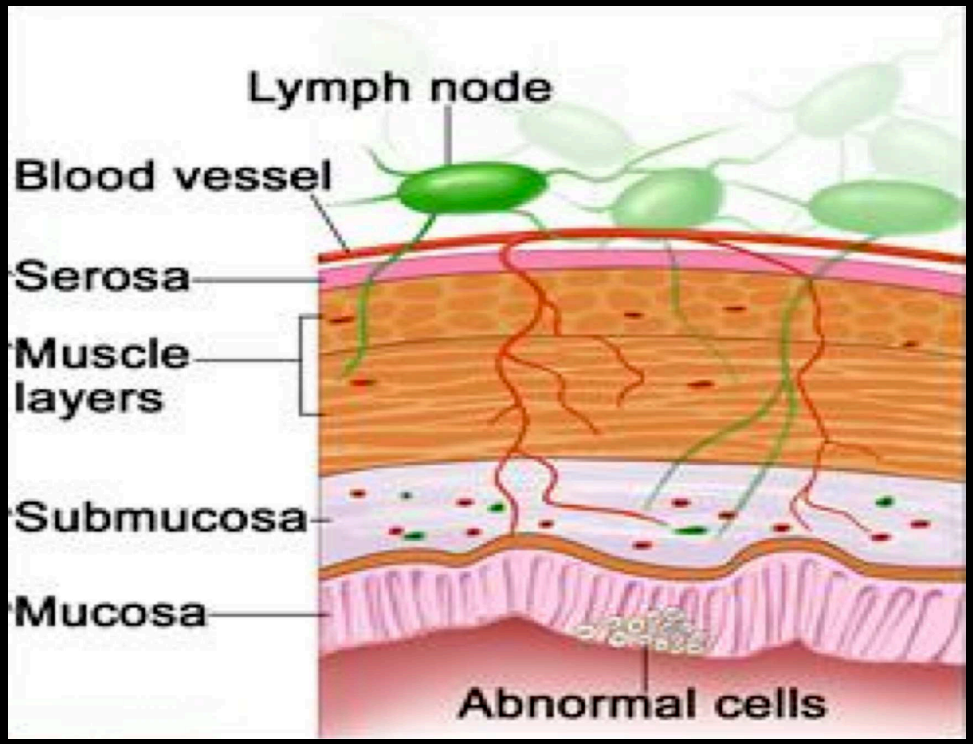
Limitations of Kaplan-Meier Models

- Descriptive rather than inferential.
- Does not include the effects of explanatory variables.
- Requires categorical outcomes.
- Omits time-dependent effects.

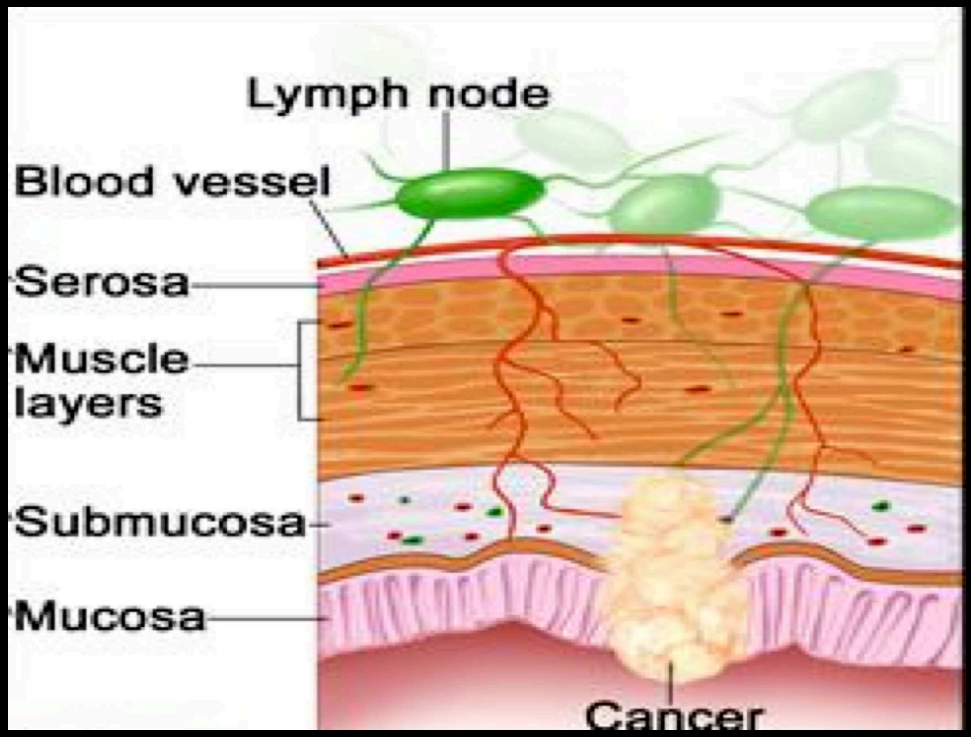
Chemotherapy for Stage 2-3 Colon Cancer

- These are data from one of the first successful trials of adjuvant chemotherapy for colon cancer: Duke's Stage C patients, enrolled from March 1985 to October 1987.
- See: Charles G. Moertel, M.D., Thomas R. Fleming, Ph.D., John S. Macdonald, M.D., Daniel G. Haller, M.D., John A. Laurie, M.D., Phyllis J. Goodman, M.S., James S. Ungerleider, M.D., William A. Emerson, M.D., Douglas C. Tormey, M.D., John H. Glick, M.D., Michael H. Veeder, M.D., and James A. Mailliard, M.D. "Levamisole and Fluorouracil For Adjuvant Therapy of Resected Colon Carcinoma." *New England Journal of Medicine* **322**, 352-8.
- Levamisole is a low-toxicity compound previously used to treat worm infestations in animals.
- 5-Fluorouracil is a moderately toxic (as these things go) chemotherapy agent.
- There are two records per person, one for recurrence and one for death.

Colon Cancer Stages

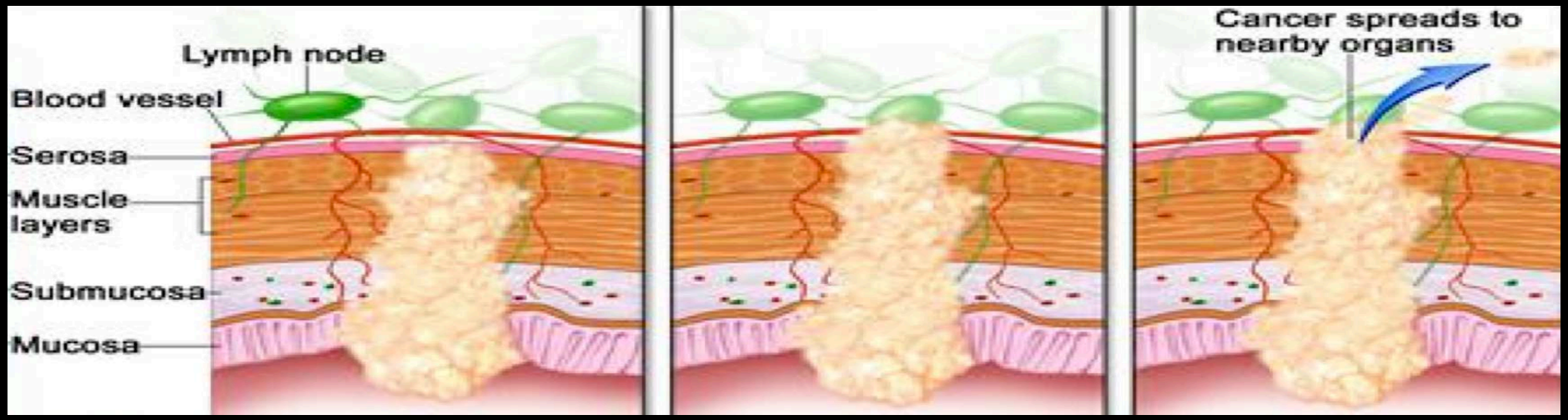


Stage 0



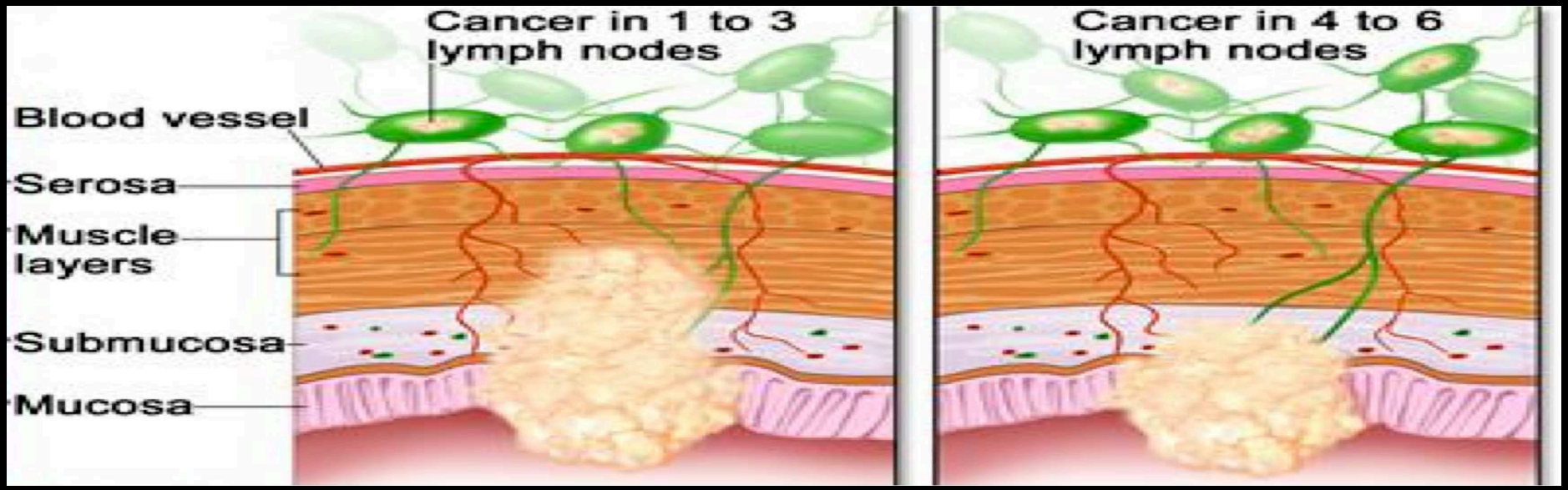
Stage I

Colon Cancer Stages



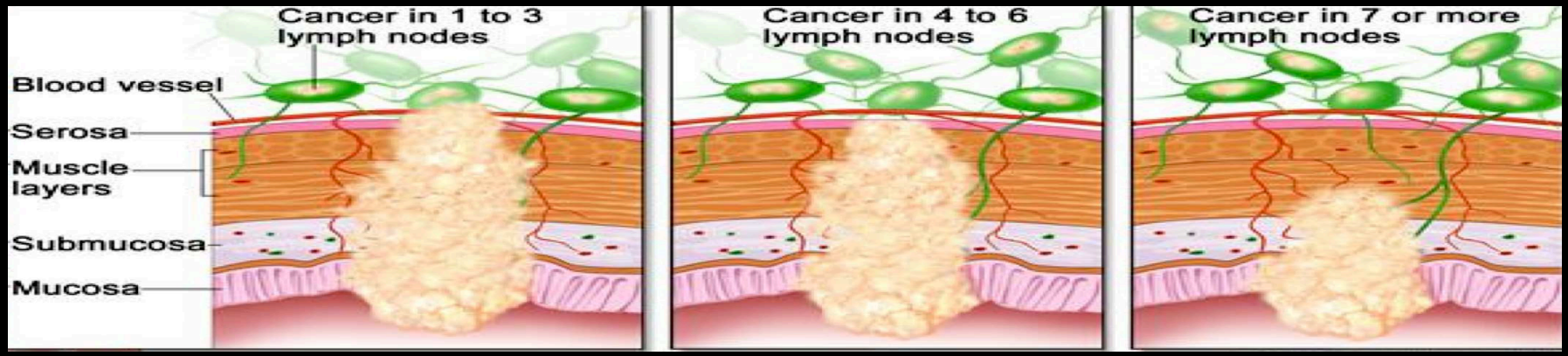
Stage IIA Stage IIB Stage IIC

Colon Cancer Stages



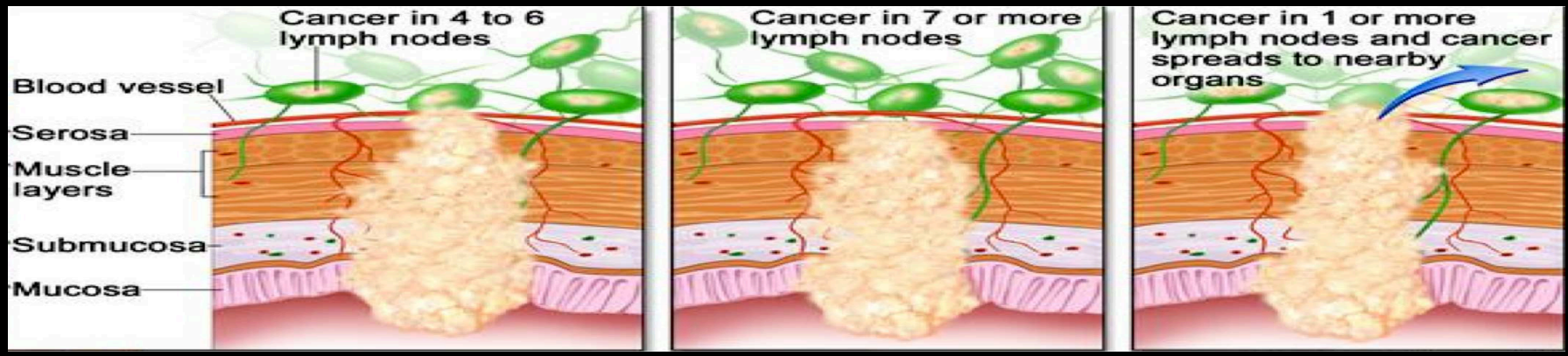
Stage IIIA

Colon Cancer Stages



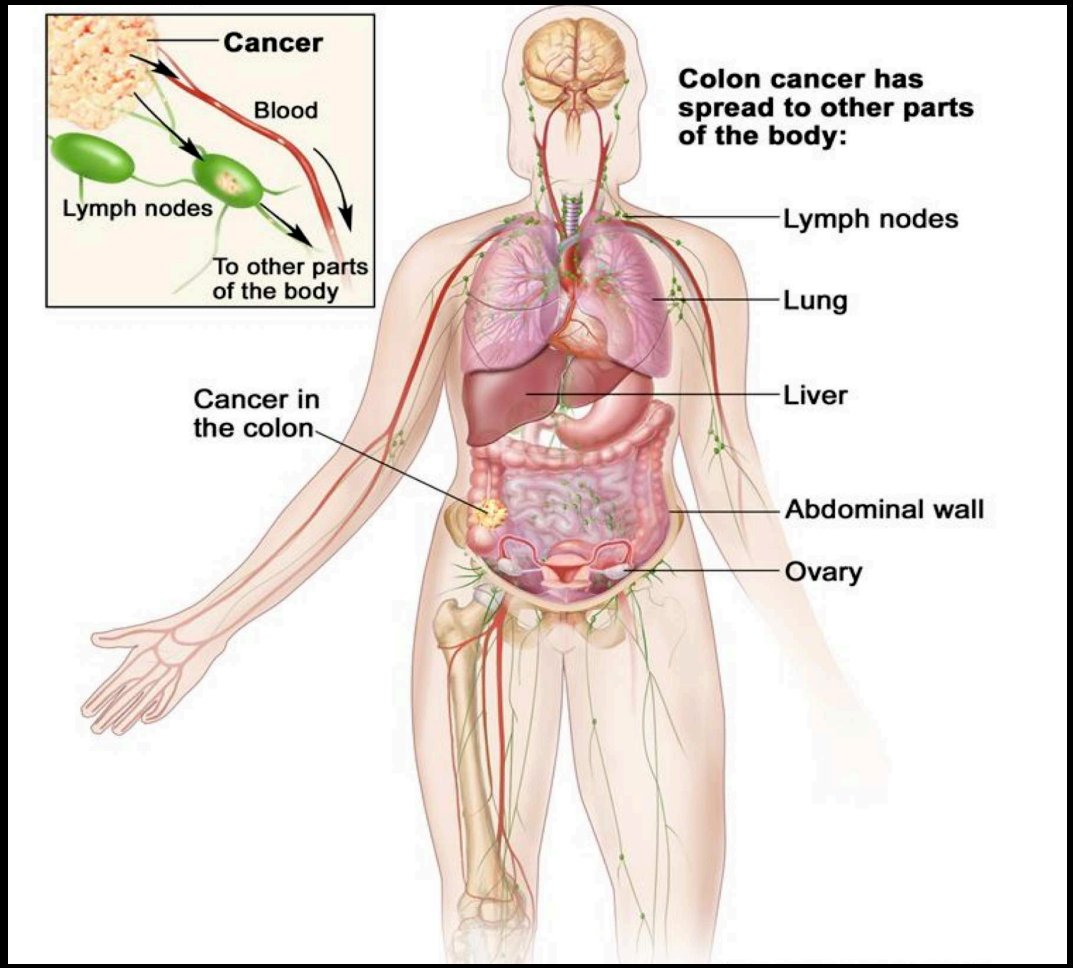
Stage IIIB

Colon Cancer Stages



Stage IIIC

Colon Cancer Stages



Stage IV

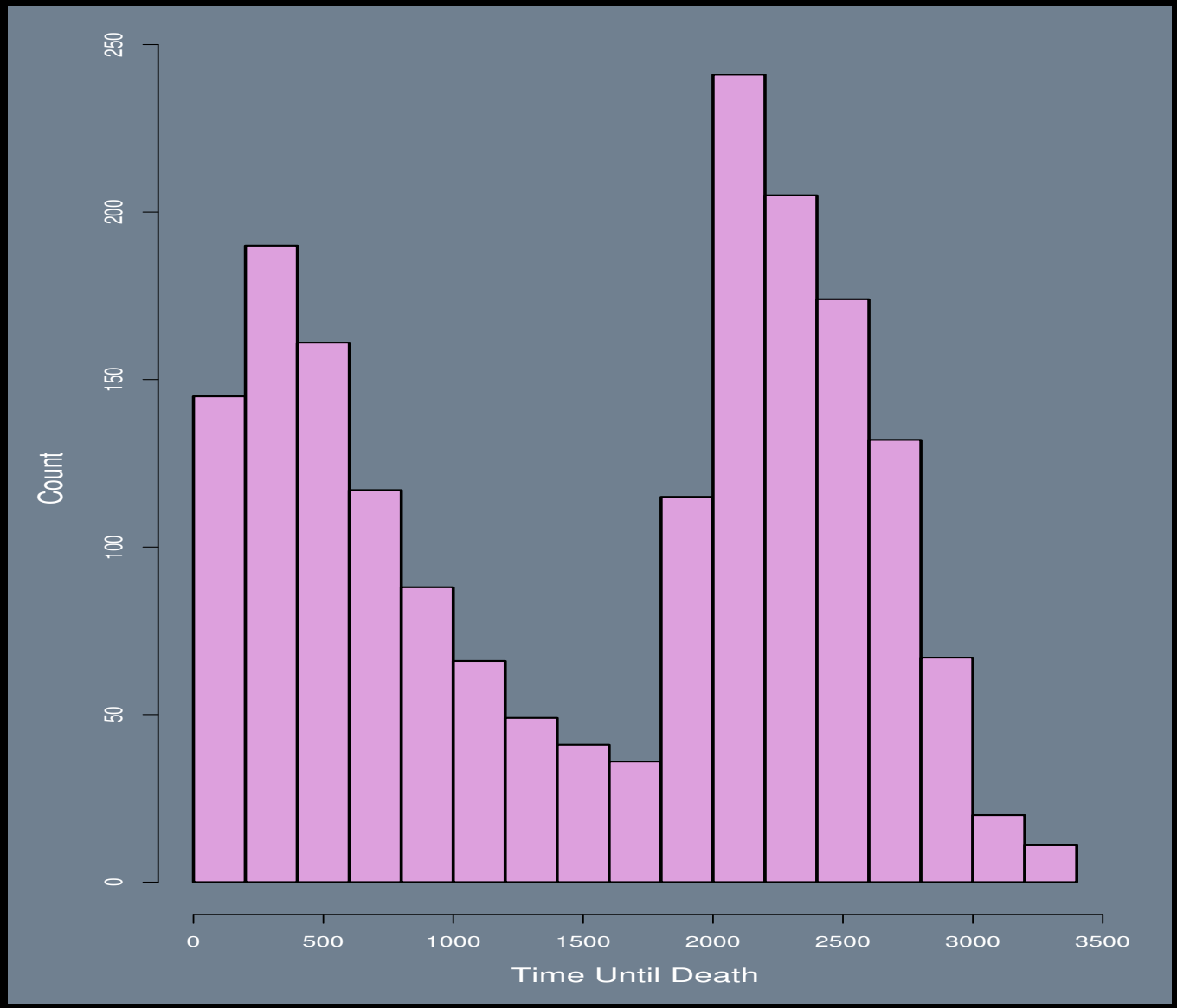
Chemotherapy for Stage 2-3 Colon Cancer

- `id`: id
- `study`: 1 for all patients
- `rx`: Treatment - Just Observation (1), Levamisole (2), Levamisole + 5-Fluorouracil (3)
- `sex`: 1=male
- `age`: in years
- `obstruct`: obstruction of colon by tumor
- `perfor`: perforation of colon
- `adhere`: adherence to nearby organs
- `nodes`: number of lymph nodes with detectable cancer
- `status`: censoring status
- `differ`: differentiation of tumor (1=well, 2=moderate, 3=poor): extent to which a tumor resembles its tissue of origin
- `extent`: extent of local spread (1=submucosa, 2=muscle, 3=serosa (serous membrane), 4=contiguous structures)
- `surg`: time from surgery to registration (0=short, 1=long)
- `node4`: more than 4 positive lymph nodes
- `time`: days until death
- `etype`: 1=recurrence,2=death

Chemotherapy for Stage 2-3 Colon Cancer

```
colon <- read.table("http://jgill.wustl.edu/data/colon.dat")

dim(colon)
[1] 1858   16

names(colon)
 [1] "id"       "study"    "rx"       "sex"      "age"
 [6] "obstruct" "perfor"   "adhere"   "nodes"    "status"
[11] "differ"   "extent"   "surg"     "node4"    "time"
[16] "etype"

colon$rx <- as.numeric(colon$rx)

postscript("./Images/colon.hist.ps")
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white", col="white",bg="slategray", cex.lab=1.3)
hist(colon$time,prob=FALSE,xlab="Time Until Death", ylab="Count",main="",col="plum")
dev.off()
```

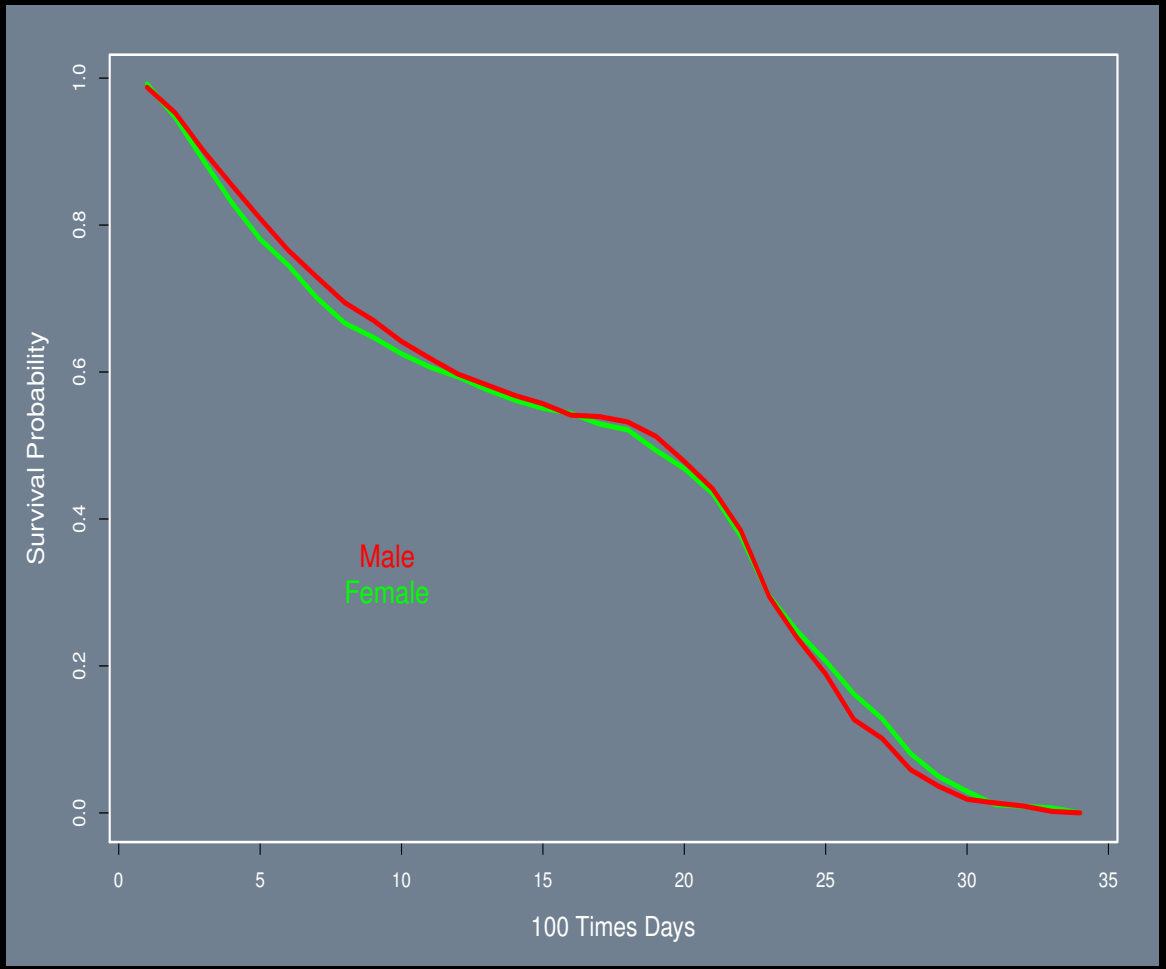
Chemotherapy for Stage 2-3 Colon Cancer, Histogram of `time`



Chemotherapy for Stage 2-3 Colon Cancer, Difference by Sex

```
fit0  <- survfit(Surv(round(colon$time/100)) ~ colon$sex)
postscript("./Images/colon0.ps")
female <- summary(fit0)$surv[1:34]
male <- summary(fit0)$surv[35:69]
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white", col="white",bg="slategray", cex.lab=1.3)
plot(female,type="l",col="green",xlab="100 Times Days",ylab="Survival Probability",
     lwd=5)
lines(male,col="red",lwd=5)
text(9.5 ,0.35,"Male",col="red",cex=2.5)
text(9.5 ,0.30,"Female",col="green",cex=2.5)
dev.off()
```

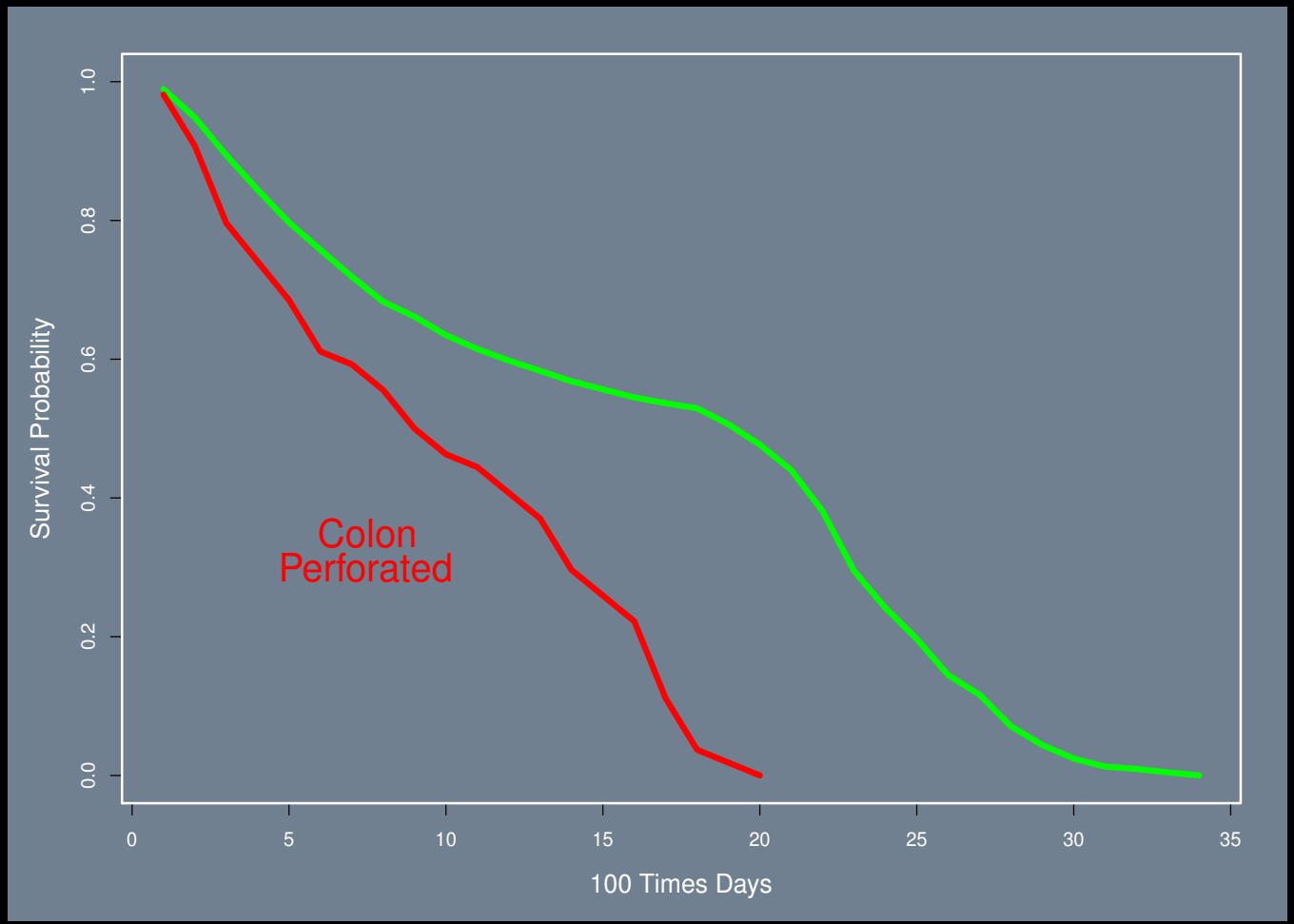
Chemotherapy for Stage 2-3 Colon Cancer, Difference by Sex



Chemotherapy for Stage 2-3 Colon Cancer, Difference by Colon Perforation

```
fit1  <- survfit(Surv(round(colon$time/100)) ~ colon$perfor)
postscript("./Images/colon1.ps")
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white", col="white",bg="slategray", cex.lab=1.3)
perfor1 <- summary(fit1)$surv[1:34]
perfor2 <- summary(fit1)$surv[35:54]
plot(perfor1,type="l",col="green",xlab="100 Times Days",ylab="Survival Probability",
     ylim=c(0,1),lwd=5)
lines(perfor2,col="red",lwd=5)
text(7.5,0.35,"Colon",col="red",cex=2.0)
text(7.5,0.30,"Perforated",col="red",cex=2.0)
dev.off()
```

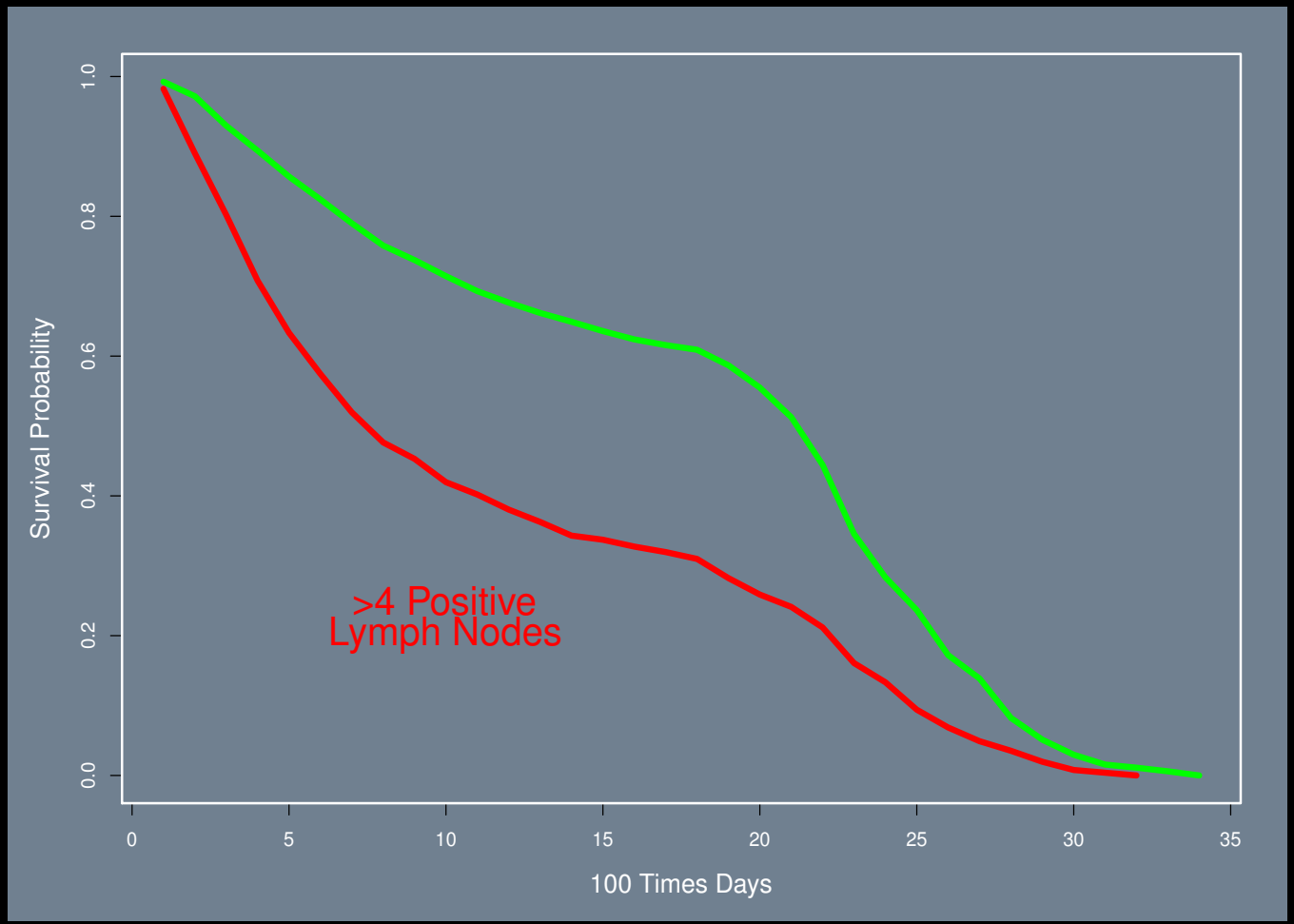
Chemotherapy for Stage 2-3 Colon Cancer, Difference by Colon Perforation



Chemotherapy for Stage 2-3 Colon Cancer, More Than 4 Positive Lymph Nodes

```
fit2  <- survfit(Surv(round(colon$time/100)) ~ colon$node4)
postscript("./Images/colon2.ps")
node4.1 <- summary(fit2)$surv[1:34]
node4.2 <- summary(fit2)$surv[35:66]
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white", col="white",bg="slategray", cex.lab=1.3)
plot(node4.1,type="l",col="green",xlab="100 Times Days",ylab="Survival Probability",
     lwd=5)
lines(node4.2,col="red",lwd=5)
text(10 ,0.25,">4 Positive",col="red",cex=2)
text(10 ,0.20,"Lymph Nodes",col="red",cex=2)
dev.off()
```

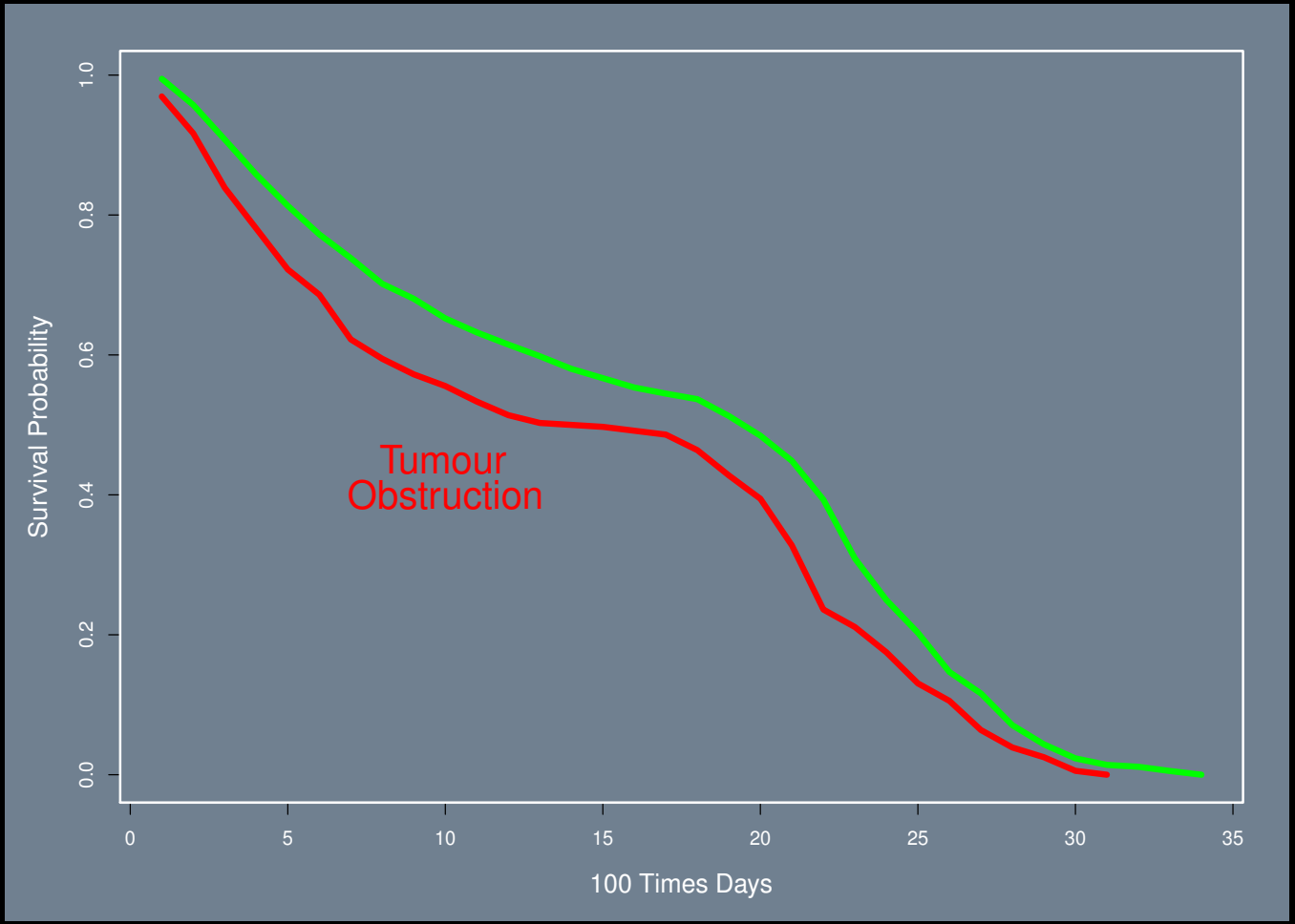
Chemotherapy for Stage 2-3 Colon Cancer, More Than 4 Positive Lymph Nodes



Chemotherapy for Stage 2-3 Colon Cancer, Difference by Obstruction of Colon by Tumor

```
fit3  <- survfit(Surv(round(colon$time/100)) ~ colon$obstruct)
postscript("./Images/colon3.ps")
obstruct.1 <- summary(fit3)$surv[1:34]
obstruct.2 <- summary(fit3)$surv[35:65]
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white", col="white",bg="slategray", cex.lab=1.3)
plot(obstruct.1,type="l",col="green",xlab="100 Times Days",
     ylab="Survival Probability", lwd=5)
lines(obstruct.2,col="red",lwd=5)
text(10 ,0.45,"Tumor",col="red",cex=2)
text(10 ,0.4,"Obstruction",col="red",cex=2)
dev.off()
```

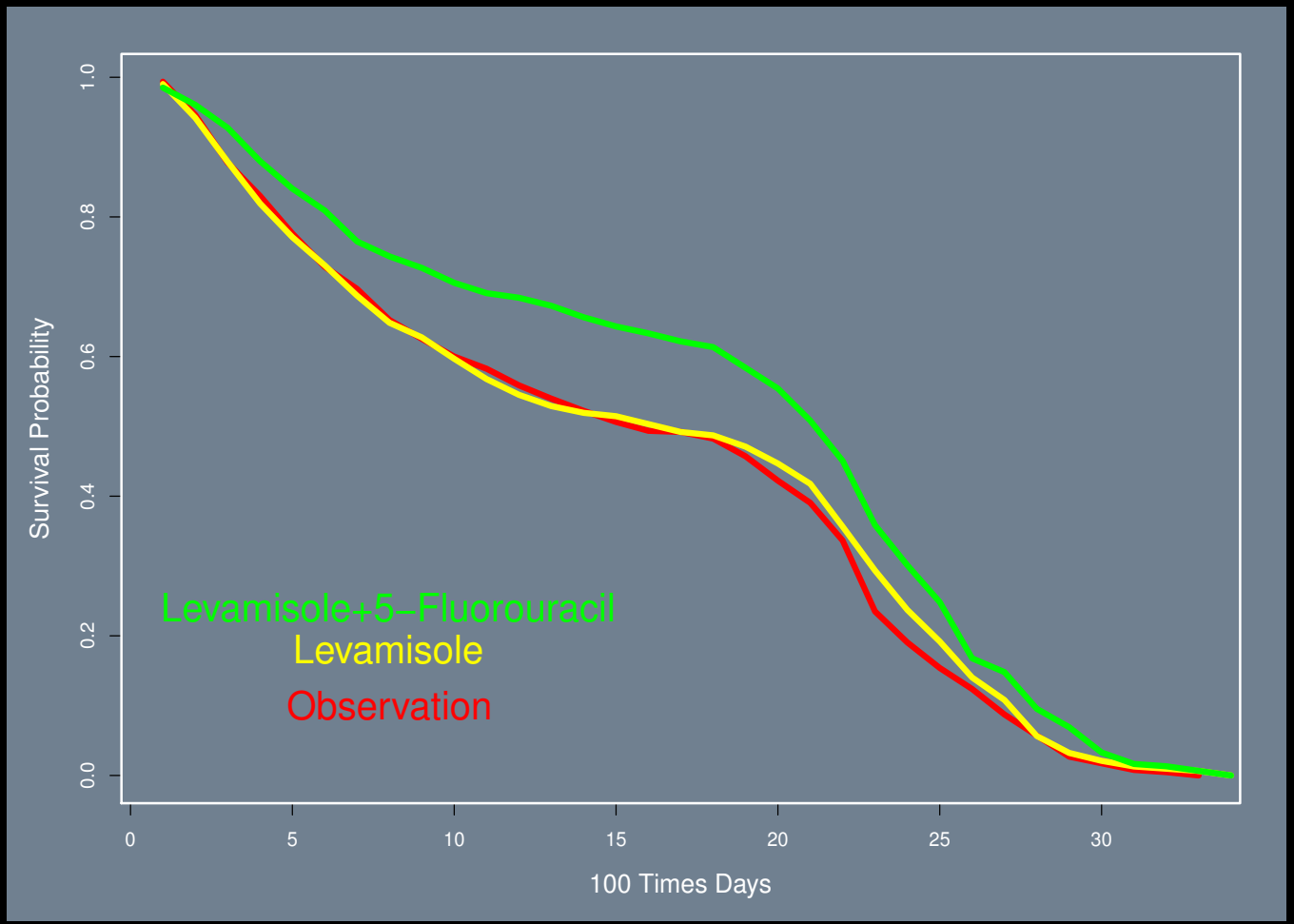
Chemotherapy for Stage 2-3 Colon Cancer, Difference by Obstruction of Colon by Tumor



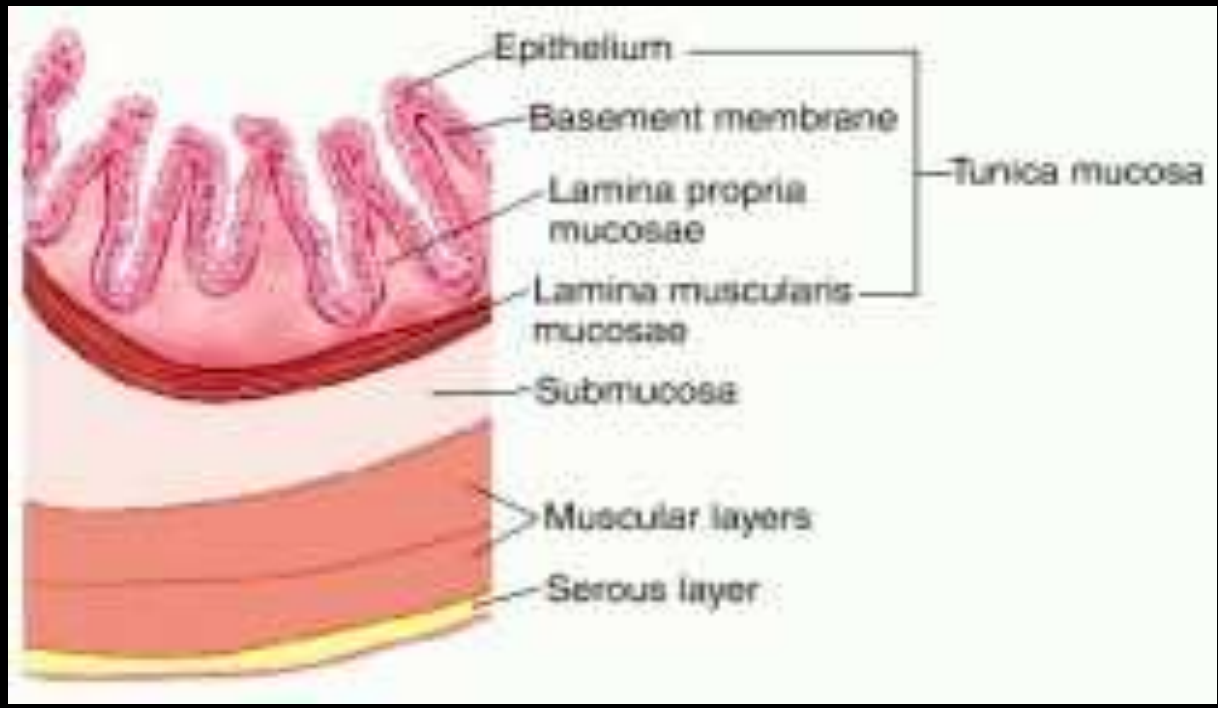
Chemotherapy for Stage 2-3 Colon Cancer, Difference by Chemotherapy Treatment

```
fit4  <- survfit(Surv(round(colon$time/100)) ~ colon$rx)
postscript("./Images/colon4.ps")
rx.1 <- summary(fit4)$surv[1:33]
rx.2 <- summary(fit4)$surv[34:67]
rx.3 <- summary(fit4)$surv[68:101]
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white", col="white",bg="slategray", cex.lab=1.3)
plot(rx.1,type="l",col="red",xlab="100 Times Days",ylab="Survival Probability",
     lwd=5)
lines(rx.2,col="yellow",lwd=5)
lines(rx.3,col="green",lwd=5)
text(8 ,0.10,"Observation",col="red",cex=2.0)
text(8 ,0.18,"Levamisole",col="yellow",cex=2.0)
text(8 ,0.24,"Levamisole+5-Fluorouracil",col="green",cex=2.0)
dev.off()
```

Chemotherapy for Stage 2-3 Colon Cancer, Difference by Chemotherapy Treatment



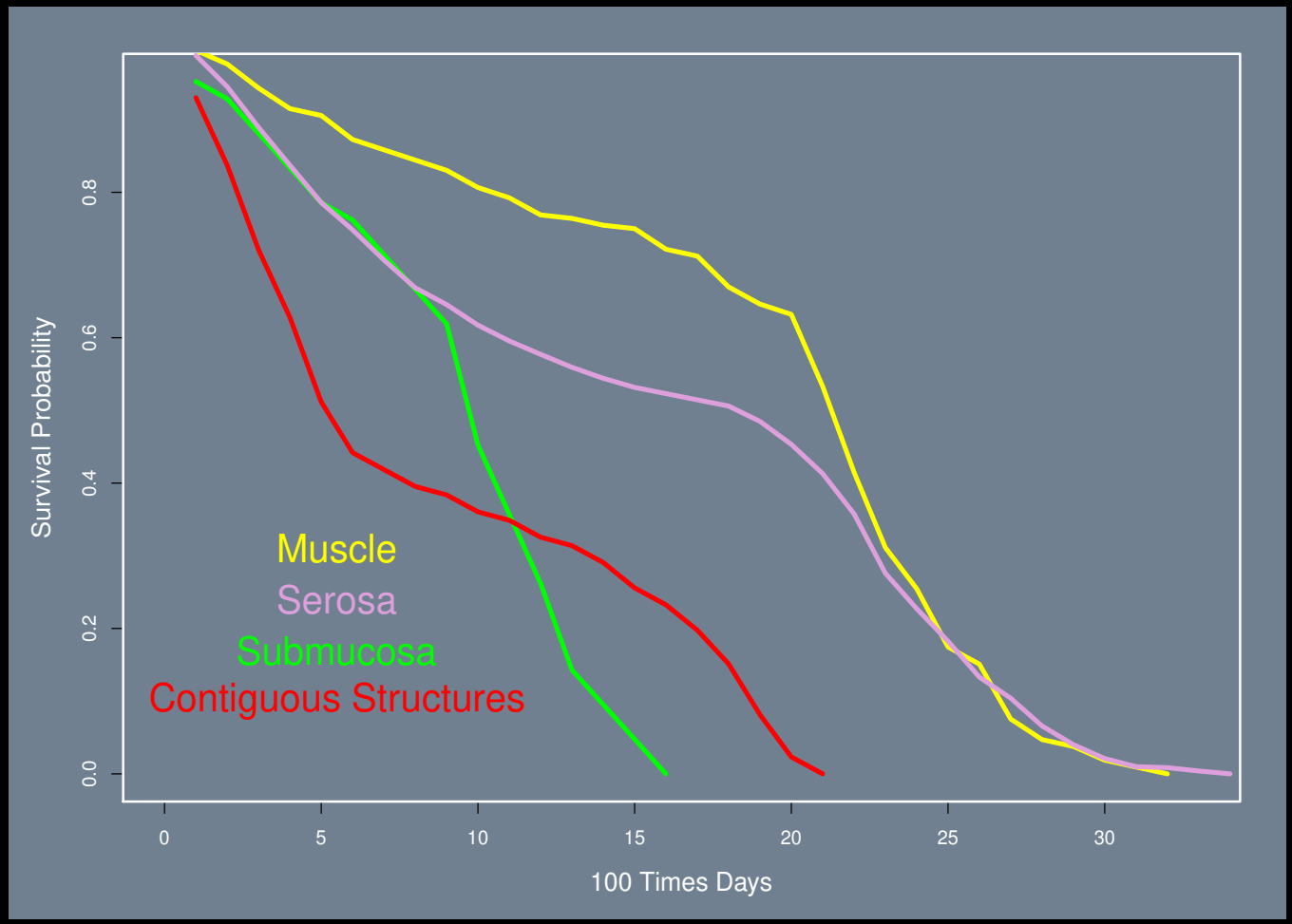
Colon Cross-Section



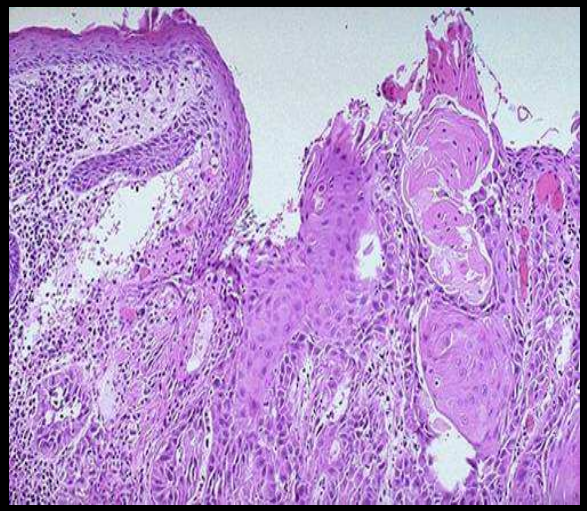
Chemotherapy for Stage 2-3 Colon Cancer, Difference by Extent of Growth

```
fit5  <- survfit(Surv(round(colon$time/100)) ~ colon$extent)
postscript("./Images/colon5.ps")
extent.1 <- summary(fit5)$surv[1:16]
extent.2 <- summary(fit5)$surv[17:48]
extent.3 <- summary(fit5)$surv[49:82]
extent.4 <- summary(fit5)$surv[83:103]
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white", col="white",bg="slategray", cex.lab=1.3)
plot(extent.1,type="l",col="green",xlab="100 Times Days",ylab="Survival Probability",
     lwd=4,xlim=c(0,33))
lines(extent.2,col="yellow",lwd=4)
lines(extent.3,col="plum",lwd=4)
lines(extent.4,col="red",lwd=4)
text(5.5 ,0.31,"Muscle",col="yellow",cex=2.0)
text(5.5 ,0.24,"Serosa",col="plum",cex=2.0)
text(5.5 ,0.17,"Submucosa",col="green",cex=2.0)
text(5.5,0.10,"Contiguous Structures",col="red",cex=2.0)
dev.off()
```

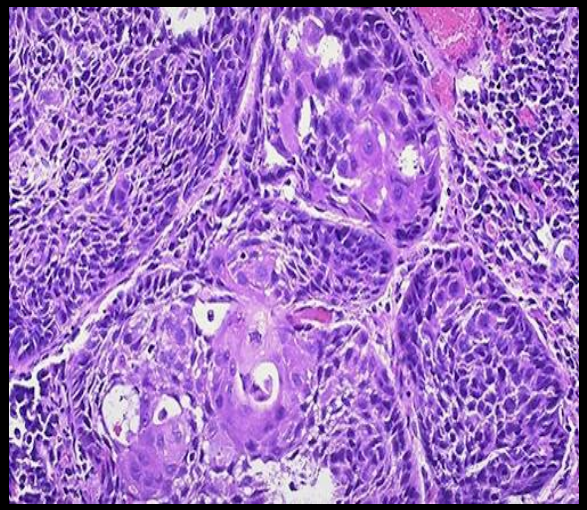
Chemotherapy for Stage 2-3 Colon Cancer, Difference by Extent of Growth



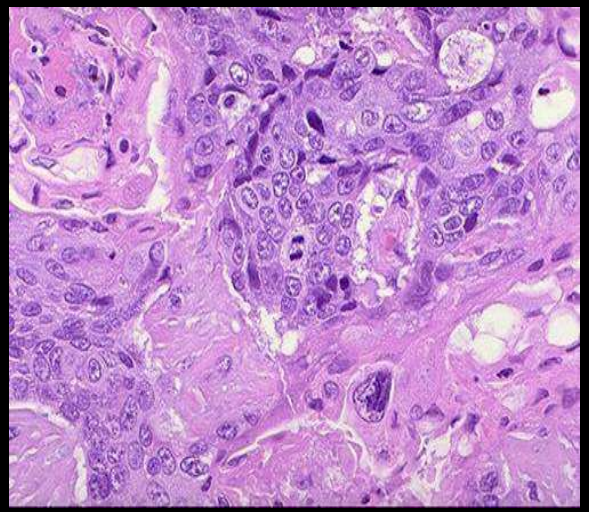
Chemotherapy for Stage 2-3 Colon Cancer, Structure of Tumor



Well Differentiated



Moderately Differentiated

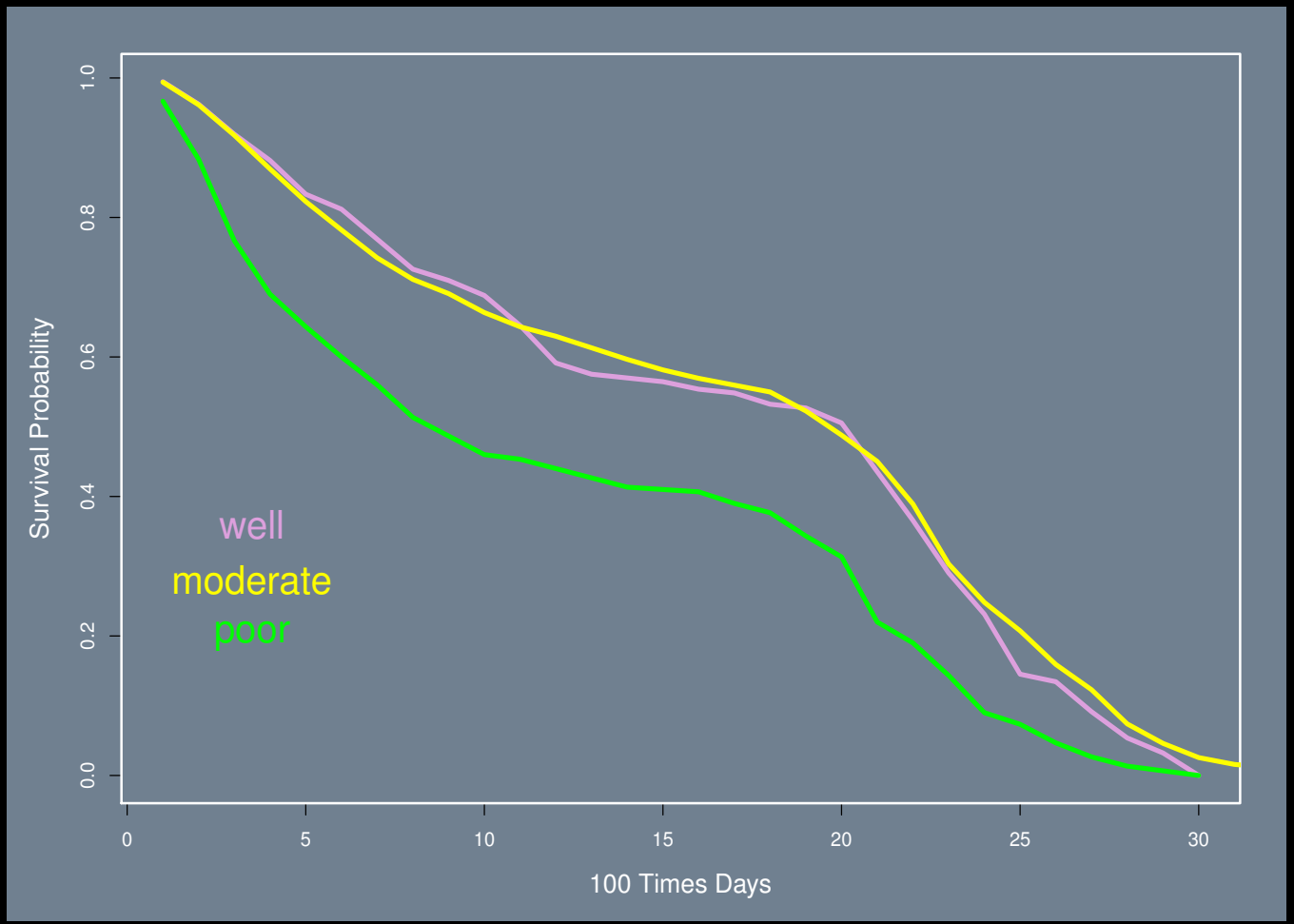


Poorly Differentiated

Chemotherapy for Stage 2-3 Colon Cancer, Differentiation of Tumor

```
fit6  <- survfit(Surv(round(colon$time/100)) ~ colon$differ)
postscript("./Images/colon6.ps")
differ.1 <- summary(fit6)$surv[1:30]
differ.2 <- summary(fit6)$surv[31:64]
differ.3 <- summary(fit6)$surv[65:94]
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white", col="white",bg="slategray", cex.lab=1.3)
plot(differ.1,type="l",col="plum",xlab="100 Times Days",ylab="Survival Probability",
     lwd=4)
lines(differ.2,col="yellow",lwd=4)
lines(differ.3,col="green",lwd=4)
text(3.5 ,0.36,"well",col="plum",cex=2.0)
text(3.5 ,0.28,"moderate",col="yellow",cex=2.0)
text(3.5 ,0.20,"poor",col="green",cex=2.0)
dev.off()
```

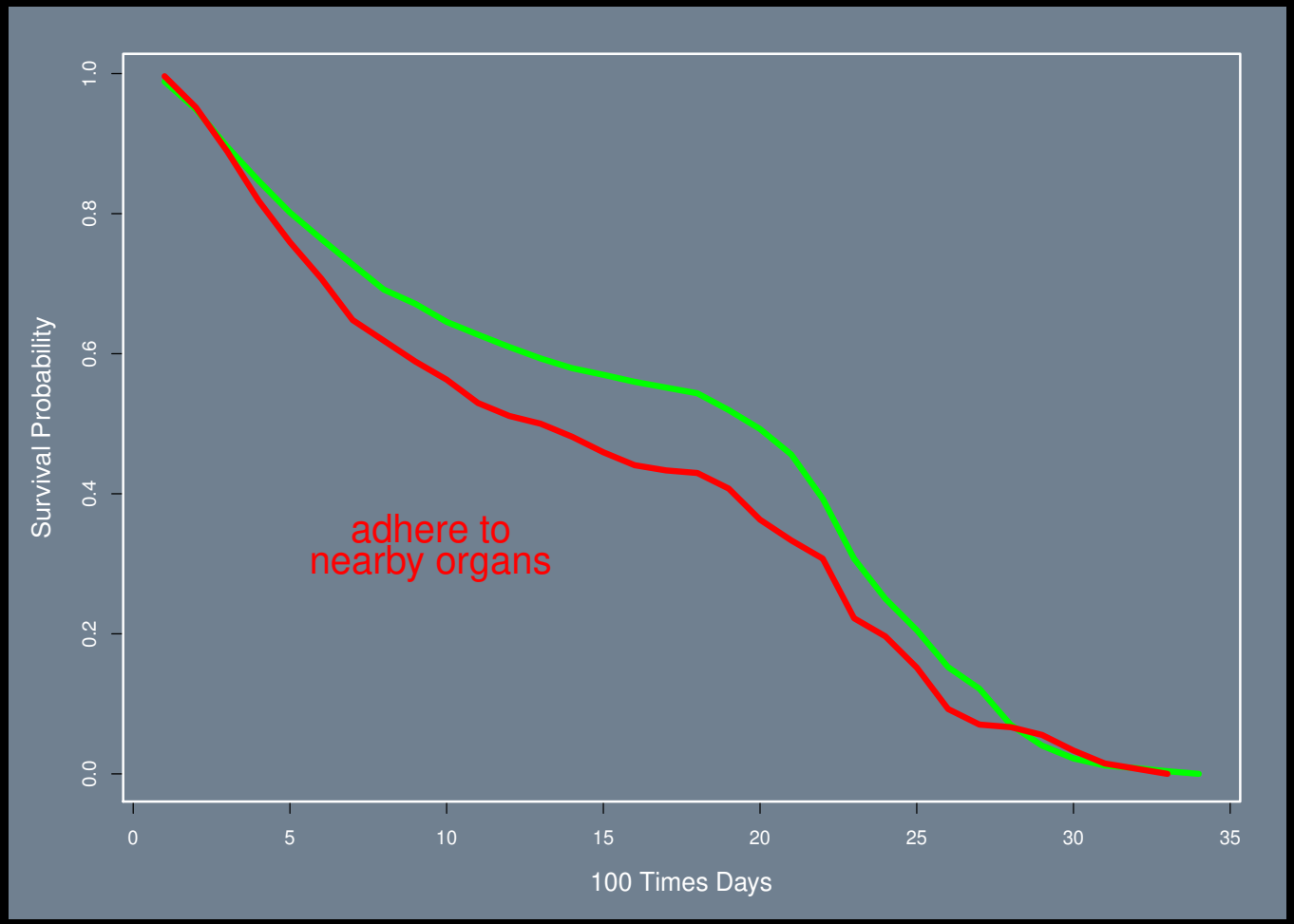
Chemotherapy for Stage 2-3 Colon Cancer, Differentiation of Tumor



Chemotherapy for Stage 2-3 Colon Cancer, Difference by Adherence to Organs

```
fit7  <- survfit(Surv(round(colon$time/100)) ~ colon$adhere)
postscript("./Images/colon7.ps")
adhere.1 <- summary(fit7)$surv[1:34]
adhere.2 <- summary(fit7)$surv[35:67]
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white", col="white",bg="slategray", cex.lab=1.3)
plot(adhere.1,type="l",col="green",xlab="100 Times Days",ylab="Survival Probability",
    lwd=5)
lines(adhere.2,col="red",lwd=5)
text(9.5 ,0.35,"adhere to",col="red",cex=2.0)
text(9.5 ,0.30,"nearby organs",col="red",cex=2.0)
dev.off()
```

Chemotherapy for Stage 2-3 Colon Cancer, Difference by Adherence to Organs



Cox Proportional Hazards Model

```
cox1.fit <- coxph(Surv(time) ~ rx + sex + age + obstruct + perfor + adhere
                          + differ + extent + node4 + surg, data=colon)
summary(cox1.fit)
               coef exp(coef)  se(coef)      z Pr(>|z|)
rx        -0.149313  0.861300  0.028739 -5.195 2.04e-07
sex        0.042153  1.043054  0.047399  0.889  0.37383
age        0.005928  1.005945  0.002084  2.845  0.00444
obstruct   0.162754  1.176747  0.061407  2.650  0.00804
perfor     0.151289  1.163333  0.142743  1.060  0.28920
adhere     0.025918  1.026257  0.069657  0.372  0.70984
differ     0.127173  1.135614  0.048852  2.603  0.00923
extent     0.196991  1.217733  0.050057  3.935 8.31e-05
node4      0.598202  1.818845  0.054457 10.985  < 2e-16
surg       0.046655  1.047761  0.053253  0.876  0.38098
```