A Clinical Trial

J.A.Hartigan Yale University
September 2013

1 The Courage Trial

This trial ran from 1999 until 2006 at 50 U.S. and Canadian hospitals.
From *Optimal Medical Therapy with or without PCI for Stable Coronary Disease N Engl J Med 2007; 356:1503-1516April 12, 2007 Boden et al*

We conducted a randomized trial involving 2287 patients who had objective evidence of myocardial ischemia and significant coronary artery disease at 50 U.S. and Canadian centers. Between 1999 and 2004, we assigned 1149 patients to undergo PCI with optimal medical therapy (PCI group) and 1138 to receive optimal medical therapy alone (medical-therapy group). The primary outcome was death from any cause and nonfatal myocardial infarction during a follow-up period of 2.5 to 7.0 years (median, 4.6). There were 211 primary events in the PCI group and 202 events in the medical-therapy group. The 4.6-year cumulative primary-event rates were 19.0% in the PCI group and 18.5% in the medical-therapy group (hazard ratio for the PCI group, 1.05; 95% confidence interval (CI), 0.87 to 1.27; P=0.62). There were no significant differences between the PCI group and the medical-therapy group in the composite of death, myocardial infarction, and stroke (20.0% vs. 19.5%; hazard ratio, 1.05; 95% CI, 0.87 to 1.27; P=0.62); hospitalization for acute coronary syndrome (12.4% vs. 11.8%; hazard ratio, 1.07; 95% CI, 0.84 to 1.37; P=0.56); or myocardial infarction (13.2% vs. 12.3%; hazard ratio, 1.13; 95% CI, 0.89 to 1.43; P=0.33). As an initial management strategy in patients with stable coronary artery disease, PCI did not reduce the risk of death, myocardial infarction, or other major cardiovascular events when added to optimal medical therapy. (ClinicalTrials.gov number, NCT00007657.)
See also, for a critical review:
http://content.onlinejacc.org/cgi/content/full/50/16/1598

2 Reading Courage Data

The data were available in the Courage directory in the files "data/CourageD.csv" and "data/COURAGEVariableDescriptions.txt". (For confidentiality reasons, the file CourageD:csv is not accessible. I include this code only to show how it was done).
To check that the required file is in the directory:

```
wd <- getwd()
list.files(paste(wd, "data", sep="/"))
```

```
[1] "Courage.csv"
[2] "COURAGEVariableDescriptions.txt"
```

Read variable descriptions:

```
vdescr<- scan("data/COURAGEVariableDescriptions.txt",
              what="",sep="\n")
```

```
Read 38 items
```

See what you have:

```
head(vdescr)
```

```
[1] "Trt:     Randomized treatment assignment:
1=PCI+OMT,2=OMT"
[2] "Nid:     patient id: 4 digits"
[3] "Age:     age: Years "
[4] "White:   white:   "
[5] "Male:    male sex: 0=No, 1=Yes"
[6] "Fam_hx:  family history of coronary artery disease: 0=no,
1=yes"
```

Read courage data:

```
courage <- read.csv("data/Courage.csv", header=T,
                    na.string=" ", as.is=T)
```

See what you have:

```
dim(courage)
```

```
[1] 2285   38
```

3 First Look at courage

Show properties of variables after description:

```
look <- FirstLook(courage)
```

Truncate descriptions to 25 characters, add to look.

```
look$descr <- substr(vdescr,10,21)
print(look)
```

	type	distinct	missing	min	max	descr
trt	char	2	0	Meds	PCI+	Randomized t
nid	numeric	2285	0	2001	4285	patient id:
age	numeric	58	0	31	89	age: Years
white	binary	2	0	0	1	white:
male	binary	2	0	0	1	male sex: 0=
fam_hx	binary	2	235	0	1	family histo
htn	binary	2	25	0	1	history of h
smoker	binary	2	1	0	1	current smok
diabetes	binary	2	37	0	1	history of d
hx_mi	binary	2	38	0	1	prior MI: 0
cabg	binary	2	0	0	1	prior corona
hx_pci	binary	2	6	0	1	prior(>6mont
chf	binary	2	15	0	1	congestive h
cerebro	binary	2	0	0	1	history of s
pvd	binary	2	20	0	1	peripheral v
numvess	numeric	4	0	0	3	numver of di
plad	binary	2	0	0	1	proximal lef
ef	numeric	183	4	23	90	ejection fra
ccs	numeric	5	5	0	4	Canadian car
exer	binary	2	45	0	1	exercise 5 o
bmi	numeric	257	7	15.2	54	body mass in
sbp	numeric	113	8	80	210	systolic blo
dbp	numeric	73	8	31	120	diastolic bl
ldl	numeric	442	10	21	245	low density
hdl	numeric	202	5	15.4	99.8	high density
tg	numeric	662	5	31	1263	triglyceride
fpg	numeric	304	28	47	463	fasting bloo
gfr	numeric	732	206	15.9	256.8	glomular fil
hct	numeric	217	265	6.2	58.4	hematocrit:
can	binary	2	0	0	1	Canadian hos
va	binary	2	0	0	1	Veterans aff
hcs	numeric	3	0	1	3	health care
tdth	numeric	1194	0	0	2559	time to deat
dth	binary	2	0	0	1	death: 0=no,
tmi	numeric	1260	0	0	2559	time to mi:
mi	binary	2	0	0	1	MI: 0=no, 1=
tdmi	numeric	1260	0	0	2559	time to deat
dmi	binary	2	0	0	1	death or MI

4 Which treatment has the fewer dmis?

Make factors of the treatment and hospital system variables:

```
Meds <- courage$trt=="Meds"
PCI <- courage$trt=="PCI+"
courage$hcsF <- factor(courage$hcs,
                       labels=c("CAN","US","VA"))
courage$trtF <- factor(courage$trt,
                       labels=c("Meds","PCI+"))
courage$dmiF <- factor(courage$dmi,
                       labels=c("surv", "d/mi"))
table(courage$dmiF,courage$trtF)
```

```
       Meds PCI+
  surv  934  935
  d/mi  203  213
```

```
summary(table(courage$dmi,courage$trt))
```

```
Number of cases in table: 2285
Number of factors: 2
Test for independence of all factors:
     Chisq = 0.19, df = 1, p-value = 0.7
```

```
fisher.test(table(courage$dmi,courage$trt))
```

```
     Fisher's Exact Test for Count Data
data:  table(courage$dmi, courage$trt)
p-value = 0.7044
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.84 1.30
sample estimates:
odds ratio
         1
```

The Fisher test p-values are accurate, but hardly differ from the traditional chi-square test p-values when all the counts in the table exceed 5. The two tests report no significant difference between the two treatments. Did we just waste the millions of dollars the trial cost? Well, not really, because PCI (percutaneous coronary intervention) is an expensive and intrusive surgical procedure, costing about 10 billion dollars a year in the U.S., and the trial suggests that for many patients with moderate symptoms, it does not improve clinical outcomes.

5 Censoring, hazard, and survival

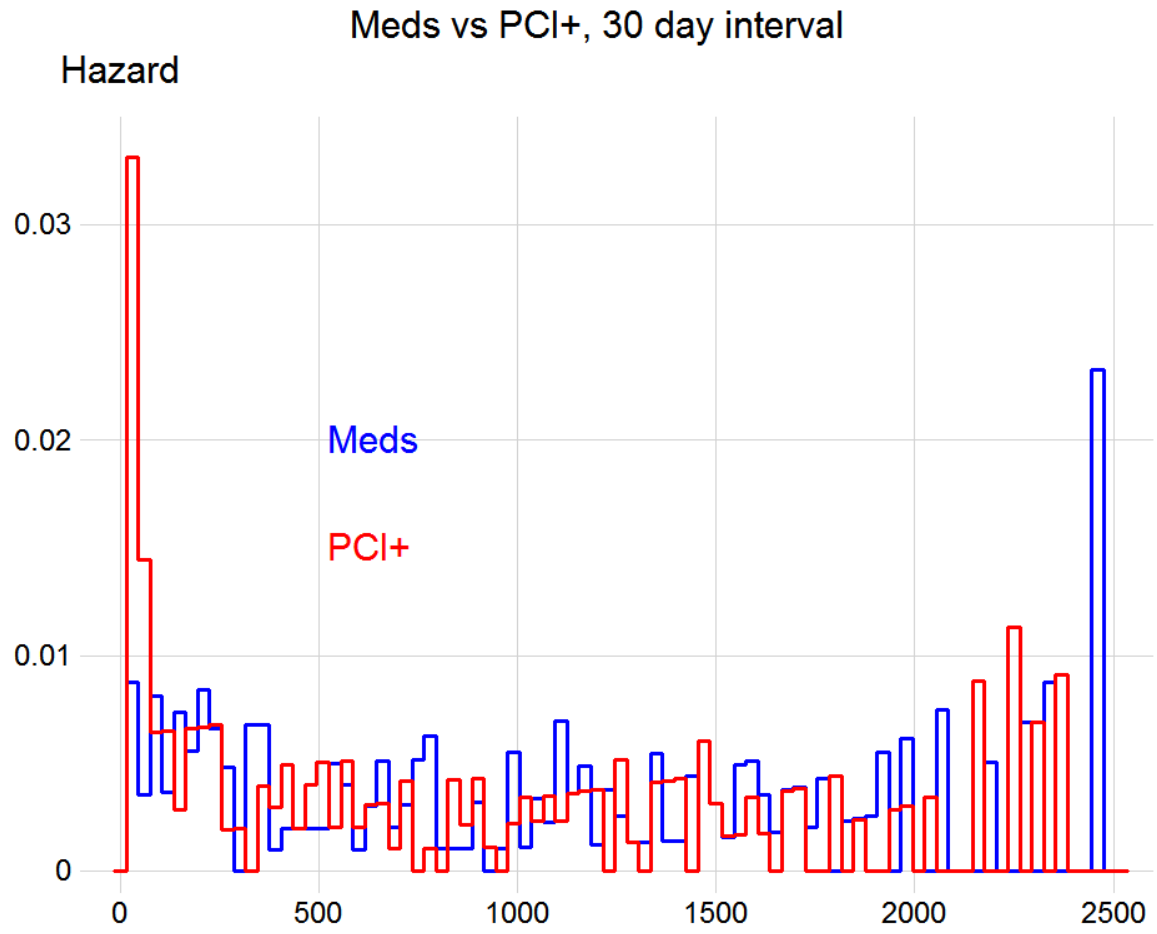
Censoring of outcomes occurs because people leave the study, either by choice during the study, or because follow-up terminates. Thus some death or mi (dmi) times are uncensored, known to the day, whereas others are censored, with dmi times known only to be greater than the censoring day.
One way to handle censoring is to assume that the censoring event is uninformative, as if the people leaving the study were selected at random. Survival to day t means reaching day t without a dmi.
Define a survival function, and a hazard function:
survival : S(t) = Probability of surviving until day t,
hazard : h(t) = Probability of dmi on day t given survival to day t.

The two functions are related by: $S(t) = \prod s|s<t\ (1-h(s))$.
In words, you survive past day t if you don't dmi on any of the previous days, so the probability of surviving until day t is the product of the probabilities of no dmi on all the days before day t.

The maximum likelihood estimate of the hazard:
h(t) = (number of dead on day t)/(number survived until day t).
The number survived until day t includes patients censored after day t, but excludes patients censored before day t, since they are not known to have survived to day t. Thus censoring is handled by excluding previously censored patients from the count of patients exposed to hazard on day t.
The Kaplan-Meier estimate of survival uses the maximum likelihood estimate of hazard to compute the survival function by the product above. The complement to the survival function is the distribution function F(t), the probability of survival no greater than t. But the survival function works better in clinical trials because it handles censoring more conveniently. Similarly, the hazard function works better than the histogram for display of densities. However, the hazard function is more informative than the survival function.

5.1 Hazard Plots

```
tiff("pictures/Hazard.tif", w=1000, h=400)
par(mar=c(5,5,4,2))
hm30<-Hazard(courage$tdmi[Meds],1-courage$dmi[Meds],30)
hp30 <- Hazard(courage$tdmi[PCI], 1-courage$dmi[PCI], 30)

Grid(c(-100, seq(0, 2500, 500), 2600),
     c(-.001, seq(0, 0.03, .01), .035),  at=c(1200, 0),
      ylab="Meds vs PCI+, 30 day interval/Hazard")

points(30*(0:length(hm30))-15, c(0,hm30), type="s",
       col="blue", lwd=3)
points(30*(0:length(hp30))-15, c(0,hp30), type="s",
       col="red", lwd=3)
text(pos=4, 500, .02, "Meds", col="blue", cex=2.5)
text(pos=4, 500, .015, "PCI+", col="red", cex=2.5)
dev.off()
```

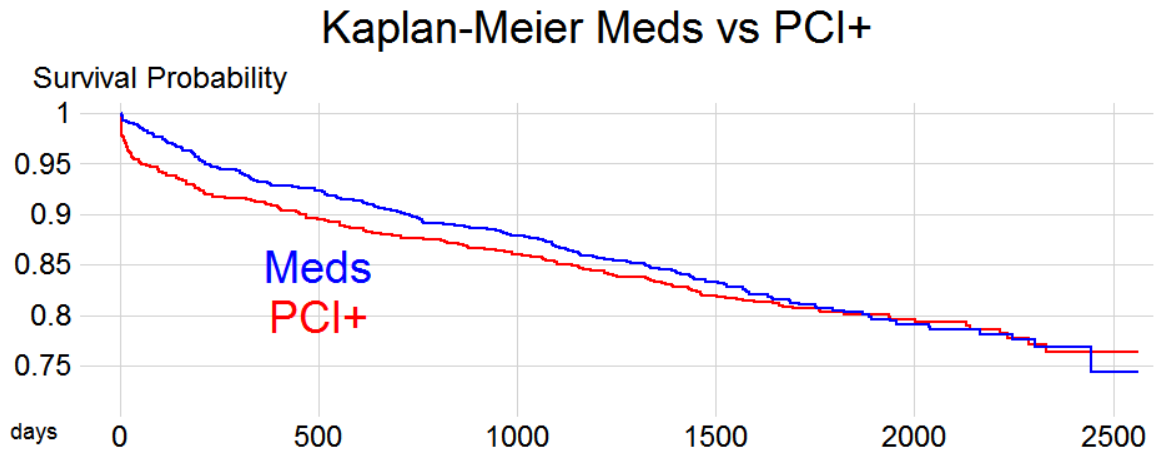


The early large PCI values indicate increased hazard for PCI patients in the first 60 days after the intervention. The hazard rates for the two treatments are fairly constant and equal thereafter, up till about 2000 days, and then there is an apparent increase in hazard. The anomalous large hazard for Meds at the extreme right of the plot is caused by the last person with dmi, on day 2443, when only 67 patients remained in the study.

5.2 Kaplan-Meier Survival curves

Compute survival from the product of 1-hazard:

```
hm <- Hazard(courage$tdmi[Meds], 1-courage$dmi[Meds])
hp <- Hazard(courage$tdmi[PCI], 1-courage$dmi[PCI])
fit1 <- exp( c(0, cumsum( log(1-hm) )) )
fit2 <- exp( c(0, cumsum( log(1-hp) )) )

tiff("pictures/Kaplan-Meir.tif", w=1000, h=400)

Grid(c(-100,seq(0,2500,500),2600),c(seq(0.7,1,.05),1.01
), ylab="Kaplan-Meier Meds vs PCI+/Survival Probability",
at=c(1200,100), cex=c(3,2))

text(500,.85,"Meds", col="blue", cex=3)
text(500,0.8,"PCI+", col="red", cex=3)
text(-300, .685, "days", pos=4, xpd=T, cex=1.5)
points(fit2, col="red", type="s", lwd=2)
points(fit1, col="blue", type="s", lwd=2)
dev.off()
```

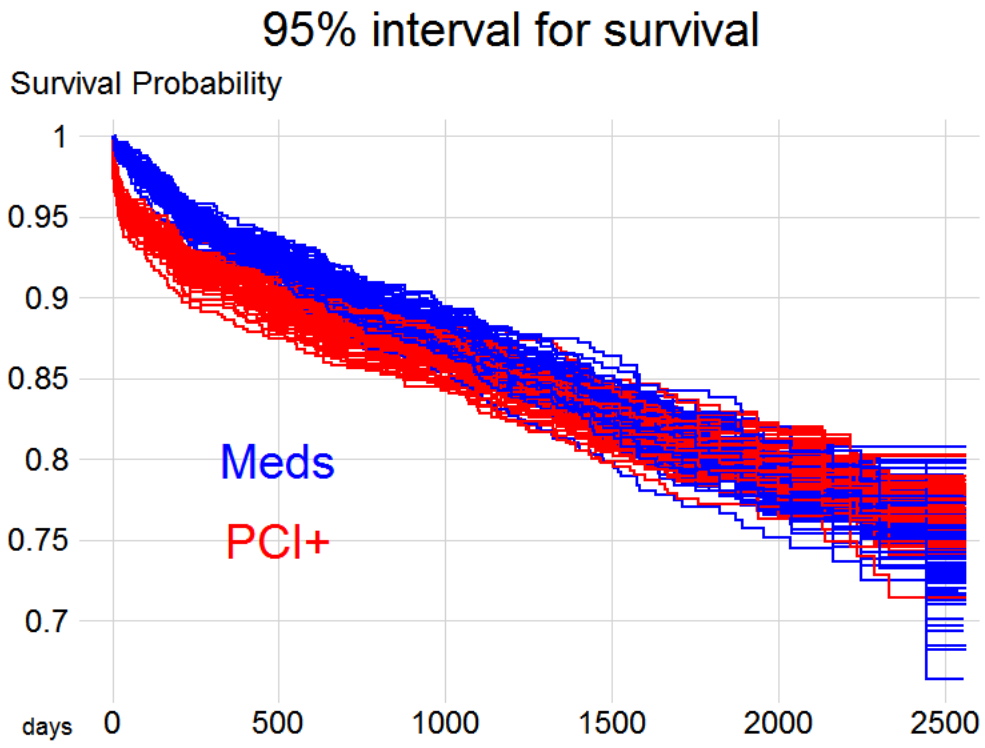


We see the big drop in PCI survival in the first two months, corresponding to the high hazard in that period, with the curve coming back to equality with Meds after 6 or 7 years. The curves are more variable towards the end because the estimates there are based on hazard functions for fewer people. Only people who are enrolled early in the study will be observed for 6 or 7 years.

5.3 95% intervals for survival curves

We generate these curves by subsampling. In each subsample, each patient is selected with probability 0.5.

```
tiff("pictures/0.95 percent survival.tif", w=800, h=600)
par(mfrow=c(1,1))
np <- dim(courage)[1]

# mess about a bit here
nrow <- dim(courage)[1]
rsample <- sample(1:nrow, nrow)
for( s in 1:79){

  subs <- rbinom(np, 1, 0.5) == 1
  msubs <- Meds & subs
  psubs <- PCI & subs
  hm <- Hazard(courage$tdmi[msubs],1-courage$dmi[msubs])
  hp <- Hazard(courage$tdmi[psubs],1-courage$dmi[psubs])
  fit1 <- exp( c(0, cumsum( log(1-hm) )) )
  fit2 <- exp( c(0, cumsum( log(1-hp) )) )

  if (s == 1){
    Grid(c(-100,seq(0,2500,500),2600),
         c(seq(0.65,1,.05),1.01),
   ylab="95% interval for survival/Survival Probability",
         at=c(1200,100), cex=c(3,2))

    text(500,.8,"Meds", col="blue", cex=3)
    text(500,0.75,"PCI+", col="red", cex=3)
    text(-300, .635, "days", pos=4, xpd=T, cex=1.5)
  }

  points(fit2, col="red", type="s", lwd=2)
  points(fit1, col="blue", type="s", lwd=2)
}

dev.off()
```



The 95% interval for survival at any time is obtained by dropping the smallest and largest random subsample value at that time. We see established differences in the first two or three years, with PCI having lower survival. The curves converge thereafter. The longer times have larger errors because fewer patients are available for the hazard calculation.

6 Censoring and binomial regression

The simplest model, justified by the random assignment to treatment, is the multinomial model for two binary variables, treatment and the outcome of dmi. Still, we can't spend 35 million dollars, and just produce a 2 by 2 table to explain the results. Nope, we need to look more closely at other contributors to the outcome... for example, maybe the health care systems change the treatment outcomes, because perhaps the University hospitals have more expert well-paid enthusiastic fee-receiving interventionists than the VA and Canada, which are more like European public health care systems, resistant to expensive operations. Or perhaps take account of age in the VA system.

There is a way to handle censoring with a binomial model that looks at outcomes one day at a time, and assumes constant hazard for each individual. Consider an individual with base variables B say. It is assumed that for this individual, a dmi occurs independently on each day, with a probability depending on B, p(B), say.

If this individual has an event on day T, then dmi=1 for that individual and the binomial with probability p(B) has T-1 failures and 1 success. If the individual is censored at day T, then dmi=0 and the binomial with probability p(B) has T failures and no successes. The likelihood of the data is then the product over all individuals of the quantity $(1-p(B))^{T-dmi}\, p(B)^{dmi}$.

Now the logistic model:

$$\log [p(B)/(1-p(B)] = \text{linear function of variables in B,}$$

can be applied using the general linear model function in R.

A similar but fancier model is the Cox proportional hazards model, which can be made to handle probabilities of events varying over time. In the present case, with patients being in the study about 4.5 years, it seems plausible to imagine event rates not varying much with time. (Although indeed, the PCI patients have a larger event rate at the time of the procedure.)

6.1 dmi dependence on health care system

Bind event to times, to get successes and failures in the binomial response:

```
options(digits=2)
use <- courage[, c("dmi", "tdmi", "trtF", "hcsF")]
```

The function glmZs gives the regression summary under the constraint that effects for each factor add to zero. It includes terms erroneously dropped in the regular regression summary.

```
coef <-  glmZS(fdata=use, cbind(dmi, tdmi - dmi) ~ trtF +
hcsF,  family=binomial)$coef
coef <- data.frame(coef[, c(1,3)])
coef$relrate <- exp(coef[, 1])
round(coef, 3)
```

```
            Estimate  z.value relrate
(Intercept)   -9.039 -163.291    0.00
hcsF.CAN      -0.208   -2.787    0.81
hcsF.US        0.005    0.058    1.00
hcsF.VA        0.203    2.958    1.23
trtF.Meds      0.000    0.001    1.00
trtF.PCI+      0.000   -0.001    1.00
```

Note that the coefficients for health care system add to zero, as is assumed in the zero sum model.
We see no treatment effect, but the VA has significantly higher dmi rate, and the Canadian system has significantly lower dmi rate; the two systems may be quite different in the initial health conditions of their patients.

6.2 Interactions between treatment and health care system

```
use <- courage[, c("dmi", "tdmi", "trtF", "hcsF")]
coef <-  glmZS(fdata=use, cbind(dmi, tdmi - dmi) ~
               trtF * hcsF,  family=binomial)$coef
coef <- data.frame(coef[, c(1,3)])
coef$relrate <-  exp(coef[, 1])

round(coef, 3)
```

	Estimate	z.value	relrate
(Intercept)	-9.045	-161.691	0.00
hcsF.CAN	-0.206	-2.741	0.81
hcsF.US	-0.003	-0.033	1.00
hcsF.VA	0.209	3.032	1.23
trtF.Meds	0.043	0.767	1.04
trtF.PCI+	-0.043	-0.767	0.96
trtF.Meds:hcsF.CAN	-0.159	-2.114	0.85
trtF.Meds:hcsF.US	0.206	2.257	1.23
trtF.PCI+:hcsF.CAN	0.159	2.114	1.17
trtF.PCI+:hcsF.US	-0.206	-2.257	0.81
trtF.Meds:hcsF.VA	-0.047	-0.684	0.95
trtF.PCI+:hcsF.VA	0.047	0.684	1.05

The main effects are not much changed by including the interaction terms, one of the benefits of using zero sum constraints.

Each estimate is approximately the percentage increase in dmis for the corresponding combination of treatment and health care system. For example, a person in Canadian health care on Meds would have -21% dmi for CAN, +4% for Meds, -16% for the interaction between CAN and Meds. We conclude that the CAN patients do relatively well on Meds vs PCI+, whereas the US patients do relatively poorly on Meds vs PCI+.

Rather than this fancy analysis, we could just as well look at a table:

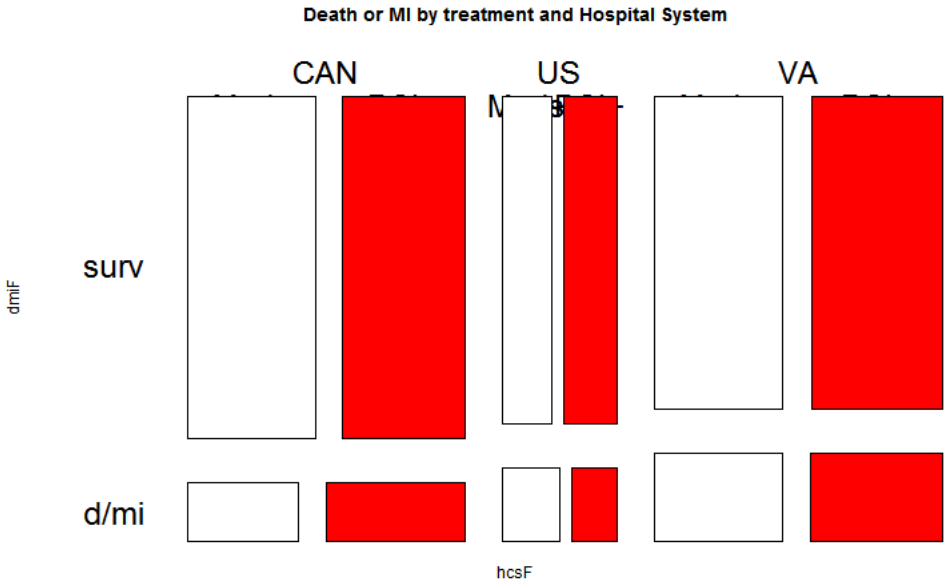
```
table(courage$dmiF, courage$hcsF, courage$trtF)

, ,  = Meds

       CAN  US  VA
  surv 407 152 375
  d/mi  60  39 104
, ,  = PCI+

       CAN  US  VA
  surv 389 164 382
  d/mi  75  31 107
```

From this, we conclude, yes, a pretty hefty dmi rate in the VA, and yes, the US does slightly better using PCI rather than Meds, and Canada does slightly better using Meds rather than PCI.

6.3 Tables and Mosaics

We use a mosaic to display the cross tabulation by treatment, hospital system, and death or mi.

```
tiff("pictures/mosaic.tif", w=800, h=500)
mosaicplot( ~hcsF+dmiF+trtF, data=courage,las=1,
col=c("white", "red"), main= "", cex.axis=2)
title("Death or MI by treatment and Hospital System")
dev.off()
```



The area of each rectangle is proportional to the number of people in the corresponding cell of the tables. For example, the upper left rectangle has area proportional to the number of people who survived on Meds in the Canadian health care system. The most noticeable effect is the difference between the Canada and US dmi rates. The higher dmi rates for the VA may be due to the VA patients being sicker to start with. Note the slightly higher dmi rates for PCI patients in Canada, and slightly lower dmi rates for PCI patients in the US (non VA) hospitals.

7 Including Age

Since Age is always a good predictor of mortality, it may be the explanation for the VA increased mortality:

```
use <- courage[, c("dmi", "tdmi", "trtF", "hcsF", "age")]
coef <-  glmZS(fdata=use, cbind(dmi, tdmi - dmi) ~
(trtF + hcsF + age)^2,  family=binomial)$coef
coef <- data.frame(coef[, c(1,3)])
coef$relrate <-  exp(coef[, 1])

round(coef, 3)
```

	Estimate	z.value	relrate
(Intercept)	-10.261	-28.40	0.00
age	0.019	3.47	1.02
hcsF.CAN	-0.535	-1.07	0.59
hcsF.US	0.223	0.39	1.25
hcsF.CAN:age	0.006	0.71	1.01
hcsF.US:age	-0.004	-0.43	1.00
hcsF.VA	0.312	0.69	1.37
hcsF.VA:age	-0.002	-0.25	1.00
trtF.Meds	0.081	0.24	1.08
trtF.Meds:age	-0.001	-0.12	1.00
trtF.PCI+	-0.081	-0.24	0.92
trtF.PCI+:age	0.001	0.12	1.00
trtF.Meds:hcsF.CAN	-0.160	-2.12	0.85
trtF.Meds:hcsF.US	0.206	2.25	1.23
trtF.PCI+:hcsF.CAN	0.160	2.12	1.17
trtF.PCI+:hcsF.US	-0.206	-2.25	0.81
trtF.Meds:hcsF.VA	-0.046	-0.67	0.95
trtF.PCI+:hcsF.VA	0.046	0.67	1.05

The age interactions are insignificant, so we can drop them from the equation. Redoing the regression without the age interaction term:

```
coef <-  glmZS(fdata=use, cbind(dmi, tdmi - dmi) ~
(trtF + hcsF)^2 + age,  family=binomial)$coef
coef <- data.frame(coef[, c(1,3)])
coef$relrate <-  exp(coef[, 1])
round(coef, 3)
```

	Estimate	z.value	relrate
(Intercept)	-10.274	-30.63	0.00
age	0.020	3.77	1.02
hcsF.CAN	-0.184	-2.44	0.83
hcsF.US	-0.018	-0.20	0.98
hcsF.VA	0.202	2.93	1.22
trtF.Meds	0.041	0.74	1.04
trtF.PCI+	-0.041	-0.74	0.96
trtF.Meds:hcsF.CAN	-0.159	-2.12	0.85
trtF.Meds:hcsF.US	0.205	2.25	1.23
trtF.PCI+:hcsF.CAN	0.159	2.12	1.17
trtF.PCI+:hcsF.US	-0.205	-2.25	0.81
trtF.Meds:hcsF.VA	-0.046	-0.67	0.95
trtF.PCI+:hcsF.VA	0.046	0.67	1.05

Thus age is a significant predictor of increased dmi's, but does not explain the increased dmi's in the VA. The coefficient .02 for age means that the dmi rate during the study increased 2% for each one year increase in age. The main effects and interaction effects are much the same as they were without including age in the regression.

8 Conclusions

The treatments Percutaneous Coronary Intervention(PCI) and Optimal Medical Therapy(MED) on 2285 patients were evaluated in a randomized clinical trial conducted at 50 hospitals in Canada, US private hospitals, and Veteran's Administration hospitals. The outcome considered was death or myocardial infarction (dmi) in the period of the trial, which had an average follow up time of 4.5 years.

Methods:
We used simple tables and mosaics to display the numbers of patients classified by health care system, intervention, and outcome. We did a standard survival analysis to handle censoring, and compared hazard functions and survival functions for the two treatments. We carried out a binomial-logistic regression in order to assess the effect on death or MI of treatment, health care system, and age.

The conclusions from the different techniques are similar:

There is no statistically significant difference between treatments in death or MI's. The VA has significantly lower overall dmi survival rates. The US hospitals had somewhat better rates for PCI vs MED and Canada had somewhat better rates for MED vs PCI. Adjusting for age had no effect on the conclusions.

9 Acknowledgement:

Pamela Hartigan, Department of Veteran Affairs, assisted in the analysis.

10 Functions

```
FirstLook <- function(data){
# Report summary properties of variables in data frame
if ( !is.data.frame(data) ) return(" data not data.frame")

# define data frame for summary variables
nrows <- dim(data)[1]
ncols <- dim(data)[2]
look <- data.frame( matrix(0,nrow=ncols, ncol=5) )
names(look) <- c("type", "distinct", "missing", "min",
"max")
rownames(look) <- names(data)

# compute summaries for each variable
for (col in 1:ncols){
  v <- data[, col]
# type
  if  ( is.numeric(v) )  look$type[col] <- "numeric"
  if  ( length(unique(v[!is.na(v)])) == 2)
look$type[col] <- "binary"
  if  ( is.factor(v) )  look$type[col] <- "factor"
  if  ( is.character(v) )  look$type[col] <- "char"
# distinct values
  look$distinct[col] <- length(unique(v[!is.na(v)]))
# missing val
  look$missing[col] <- sum(is.na(v))
# min
  look$min[col] <- min(v, na.rm=T)
# max
  look$max[col] <- max(v, na.rm=T)
}
return(look)
}
```

```
Hazard <- function(t, s, interval=1 ){
# hazard function for t event times, s=1 for censored values
# interval is the length of time in which the event is
counted; a larger interval gives a coarser more stable
hazard function; return the proportion of deaths in each
interval among those who  survived to reach that interval

if  (interval <= 0) return("interval must be positive")
if  (sum(t < 0) > 0) return(" times must be non-negative")

# count t events in intervals, including censored data
tr <- round(t/interval)
maxtr <- max(tr)
ttc <- rep(0,  maxtr+1)
tt <- table( round(t/interval) )
ttc[ as.numeric(names(tt)) + 1 ] <- tt

# count t events excluding censored data
tsc <- rep(0, maxtr+1)
ts <- table( round(t[!s == 1]/interval) )
tsc[ as.numeric(names(ts)) + 1 ] <- ts
distn <- c(0, cumsum(ttc))[-maxtr -2]
hazard <- (tsc/(length(t) - distn))[-maxtr-1]

return(hazard)
}
```

```
Grid <- function(xticks, yticks, ylab="",
        at=(min(xticks)+ mean(xticks))/2, cex=2.5){
# background for plot using grid of light grey lines
par(mar=c(3,3,6,2))
plot(1, 1,  xlim=range(xticks),  ylim = range(yticks),
           xlab="", ylab="", axes=F, pch="")

# use only interior values of tick ranges in plots
usey <- rep( T, length(yticks) )
usey[c( 1, length(yticks) )] <- F
usex <- rep( T, length(xticks) )
usex[c( 1, length(xticks) )] <- F

# grey lines in both directions
for ( row in yticks[usey] )
lines(range(xticks), c(row, row), col="light grey")
for ( col in xticks[usex] )
lines(c(col, col), range(yticks), col="light grey")

# put ylab on left top, using / to split long expressions
ylabs <- unlist(strsplit(ylab,"/"))

# identify tick marks on both axes
if (length(yticks) > 2)
  text(pos=2, rep(min(xticks), length(yticks)-2 ),
       yticks[usey], yticks[usey], cex=2, xpd=T)
if (length(xticks)>2)
  text(pos=1, xticks[usex],  rep(min(yticks),
      length(xticks)-2), xticks[usex], cex=2, xpd=T)

lylabs <- min(5, length(ylabs))
if(lylabs > 0){
mtext(ylabs, side=3,line = (5/lylabs)*(lylabs-1):0,
      at = at,cex=cex)
}

par(mar=c(5, 4, 4, 2))

invisible()
}
```

```
glmZS <-  function(fdata, ...)
{
# fixes up labelling and missing terms in categorical
models to handle contrast sum
# fdata is a data matrix including all variables in the
regression, corresponding to data =

# use options contrast so that effects sum to zero
options(contrasts = c("contr.sum","contr.sum"))

data <- fdata
ncol <- dim(data)[2]

# pick out factors in data
fl <- rep(F, ncol)
for ( col in  1:ncol) fl[col] <- is.factor(data[, col])

n <- sum(fl)
if( n==0) return(" no factors in data")

fl <- which(fl)

# run over 2^n choices of factor level patterns to be
omitted
binmat <- matrix(0, nrow =2^n, ncol=n)
for(i in 2:2^n) {
  binmat[i, ] <- binmat[i-1, ]
  for (j in 1:n) {
    if (binmat[i, j] == 0) {
      binmat[i, 1:j] <- 0
      binmat[i, j] <- 1
       break
    }
  }
}

# construct initial levels for factors
llevels <- list(1:n)
for (i in 1:n) {
```

```
  llevels[[i]] <- levels(data[, fl[i]])
}

# define different factors for each pattern of missing
levels and iterate through each choice
for (iter in 1:2^n) {
  for(i in 1:n) {
    nlevels <- length(llevels[[i]])

# first case return to original levels
    if(binmat[iter, i] == 0)
      data[, fl[i]] <-
      factor(data[, fl[i]],llevels[[i]])

# second case interchange last two levels
    if(binmat[iter,i] ==1 ){
      if(nlevels == 2)
        data[, fl[i]] <-
        factor(data[, fl[i]], llevels[[i]][2:1])
      if(nlevels > 2)
        data[, fl[i]] <- factor(data[, fl[i]],
        llevels[[i]][c(1:(nlevels-2), nlevels,
        nlevels-1)])
     }
  }


# run regression with this choice of missing levels
   fn <- names(data)[fl]

   lm.f <- glm(data=data, ...)
   sc=summary(lm.f)$coef

if(sum(is.na(lm.f$coef)))
    return(" cant handle Na's in coef")

# get level names for these missing levels
  for (i in 1:length(fn)) {
    levelnames <- levels(data[,fl[i]])
    nlevels <- length(levelnames)
    for (j in 1:nlevels ) {
      newname <-
```

```
      paste(fn[i], levelnames[j], sep = ".")
      oldname <-
      paste(fn[i], as.character(j), sep = "")

# substitute meaningful newname for obscure oldname
 row.names(sc) <- gsub(oldname,newname,row.names(sc))
    }
  }

# combine all the lists of coefficients
if(iter == 1) coef <- sc
  if(iter > 1){
    use=!row.names(sc) %in% row.names(coef)
    if(sum(use)>0){
      rn <- c(row.names(coef),row.names(sc)[use])
      coef <-  rbind(coef, sc[use,])
    row.names(coef)=rn

    }
  }
}

# order by main variables
rn <- row.names(coef)
rn <- gsub(" ", "", rn)
rn1 <- rn[-1]
lv <- rep(0, length(rn1))

# pick out variable names before the dot, if a dot
# couldnt figure out how to use regexpr to find the "."
for( i in 1:length(rn1)){
  wheredot <- which(unlist(strsplit(rn1[i],""))==".")
  if(length(wheredot) == 0) lv[i] <- nchar(rn[i])+1
  if(length(wheredot) > 0) lv[i] <- min(wheredot)
}

use <- c("", substr(rn1, 1, lv-1))
coef <- coef[order(use),]
rn <-  row.names(coef)

# make sure lower order interactions come first
low <- rep(0, length(rn))
```

```
for( i in 1:length(rn))
 low[i]<-sum(unlist(strsplit(rn[i],split=""))==".")
coef <- coef[order(low),]

ss <- summary(lm.f)
ss$coef <- coef

return(ss)
}
```