Extensions to the Cox Model: Stratification

David M. Rocke

October 23, 2025

Anderson Data

Remission survival times on 42 leukemia patients, half on new treatment, half on standard treatment. This is the same data as the drug6mp data from KMsurv, but with two other variables and without the pairing.

Name	Description
treat	"standard", "new"
sex	"female", "male"
lwbc	log of white blood count
time	time to relapse or censoring
status	0 = censored, 1 = relapsed

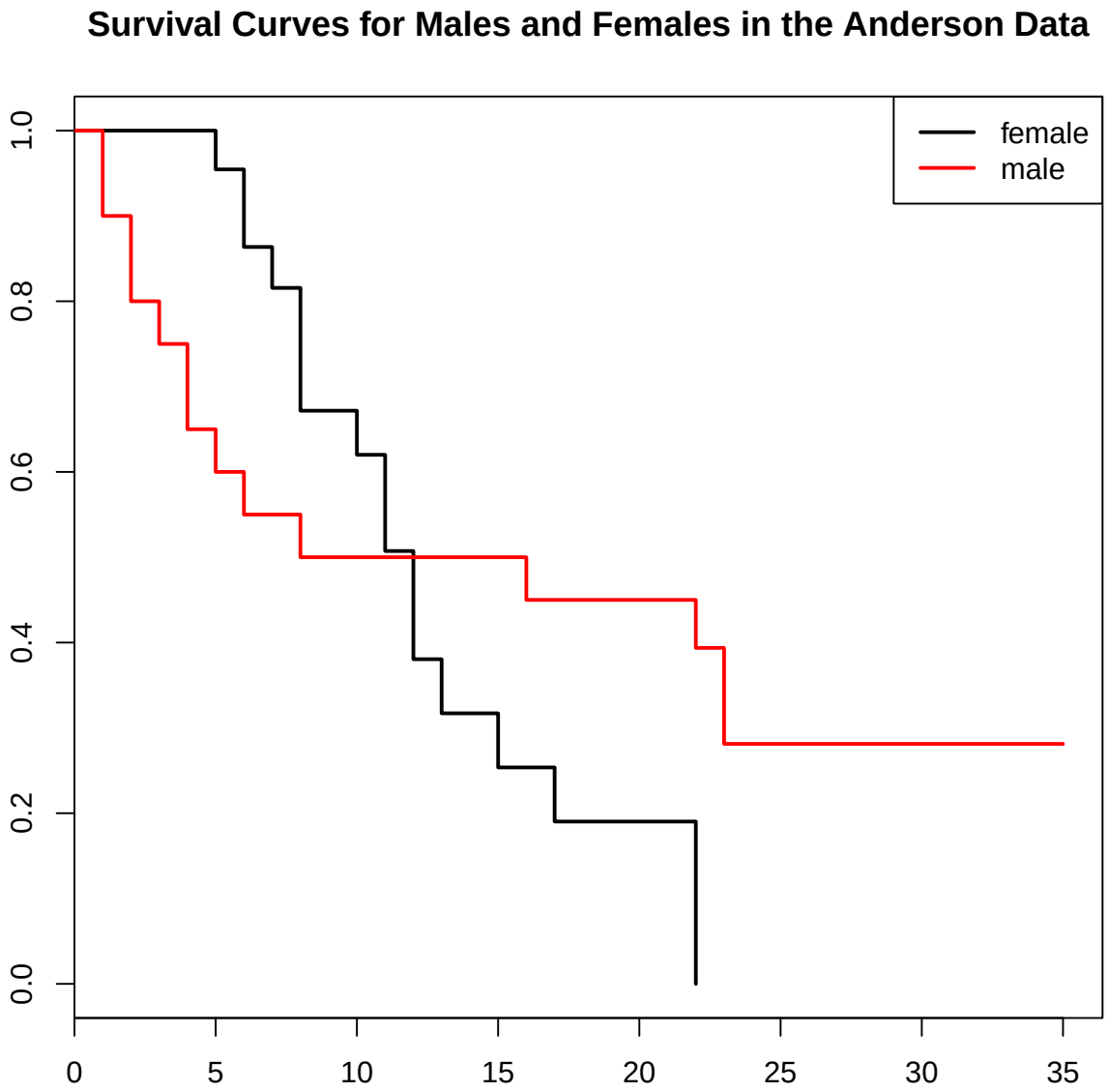
```
require(survival)
vars <- c("time","status","sex","lwbc","treat")
anderson <- read.table("anderson.dat",header=F,col.names=vars)
anderson$treat <- factor(anderson$treat,labels=c("new","standard"))
anderson$sex <- factor(anderson$sex,labels=c("female","male"))

anderson.surv <- with(anderson,Surv(time,status))
anderson.cox1 <- coxph(anderson.surv~treat+sex+lwbc,data=anderson)

> anderson.cox1

               coef exp(coef) se(coef)    z       p
treatstandard 1.504     4.498    0.462 3.26  0.0011
sexmale       0.315     1.370    0.455 0.69  0.4887
lwbc          1.682     5.376    0.337 5.00 5.8e-07

Likelihood ratio test=47.2  on 3 df, p=3.17e-10
n= 42, number of events= 30
> cox.zph(anderson.cox1)
       chisq df    p
treat  0.036  1 0.85
sex    5.420  1 0.02
lwbc   0.142  1 0.71
GLOBAL 5.879  3 0.12
```



The survival curves cross, which indicates a problem in the proportionality assumption by sex. This can be fixed by using strata or possibly by other model alterations.

The Stratified Cox Model

- In a stratified Cox model, each stratum, defined by one or more factors, has its own base survival function $h_0(t)$.
- But the coefficients for each variable not used in the strata definitions are assumed to be the same across strata.
- To check if this assumption is reasonable one can include interactions with strata and see if they are significant (this may generate a warning and NA lines but these can be ignored).
- Since the sex variable shows possible non-proportionality, we try stratifying on sex.

Stratified Model

```
> summary(coxph(anderson.surv~treat+lwbc+strata(sex),data=anderson))

  n= 42, number of events= 30

                coef exp(coef) se(coef)     z Pr(>|z|)
treatstandard 0.9981    2.7131   0.4736 2.108   0.0351 *
lwbc          1.4537    4.2787   0.3441 4.225 2.39e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

              exp(coef) exp(-coef) lower .95 upper .95
treatstandard     2.713     0.3686     1.072     6.864
lwbc              4.279     0.2337     2.180     8.398

Concordance= 0.812  (se = 0.093 )
Rsquare= 0.534   (max possible= 0.967 )
Likelihood ratio test= 32.06  on 2 df,   p=1.092e-07
Wald test            = 22.75  on 2 df,   p=1.15e-05
Score (logrank) test = 30.8  on 2 df,   p=2.052e-07
```

Separate Models

```
> summary(coxph(anderson.surv~treat+lwbc,data=anderson,sub=(sex=="male")))

  n= 20, number of events= 14

               coef exp(coef) se(coef)     z Pr(>|z|)
treatstandard 1.9779    7.2275   0.7392 2.676  0.00746 **
lwbc          1.7428    5.7132   0.5358 3.253  0.00114 **

> summary(coxph(anderson.surv~treat+lwbc,data=anderson,sub=(sex=="female")))

  n= 22, number of events= 16

               coef exp(coef) se(coef)     z Pr(>|z|)
treatstandard 0.3113    1.3652   0.5636 0.552   0.5807
lwbc          1.2061    3.3406   0.5035 2.396   0.0166 *
```

The coefficients of treatment look different. Are they statistically different?

Interaction Model

```
> summary(coxph(anderson.surv~(treat+lwbc)*strata(sex),data=anderson))
  n= 42, number of events= 30

                                coef exp(coef) se(coef)     z Pr(>|z|)
treatstandard                 0.3113    1.3652   0.5636 0.552   0.5807
lwbc                          1.2061    3.3406   0.5035 2.396   0.0166 *
treatstandard:strata(sex)male 1.6666    5.2942   0.9295 1.793   0.0730 .
lwbc:strata(sex)male          0.5366    1.7102   0.7352 0.730   0.4655
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(coxph(anderson.surv~treat+lwbc+strata(sex),data=anderson),
        coxph(anderson.surv~(treat+lwbc)*strata(sex),data=anderson),test="Chisq")
Analysis of Deviance Table
 Cox model: response is  anderson.surv
 Model 1: ~ treat + lwbc + strata(sex)
 Model 2: ~ (treat + lwbc) * strata(sex)
   loglik  Chisq Df P(>|Chi|)
1 -55.735
2 -53.852 3.7659  2    0.1521
```

Stratified Model for Anderson Data

- We chose to use a stratified model because of the apparent non-proportionality of the hazard for the sex variable.
- When we fit interactions with the strata variable, we did not get an improved model (via the likelihood ratio test).
- So we use the stratifed model with coefficients that are the same across strata.

Another Modeling Approach

- We used an additive model without interactions and saw that we might need to stratify by sex.
- Instead, we could try to improve the model—maybe the interaction of treatment and sex is real, and after fitting that we might not need separate hazard functions.
- Either approach may work.

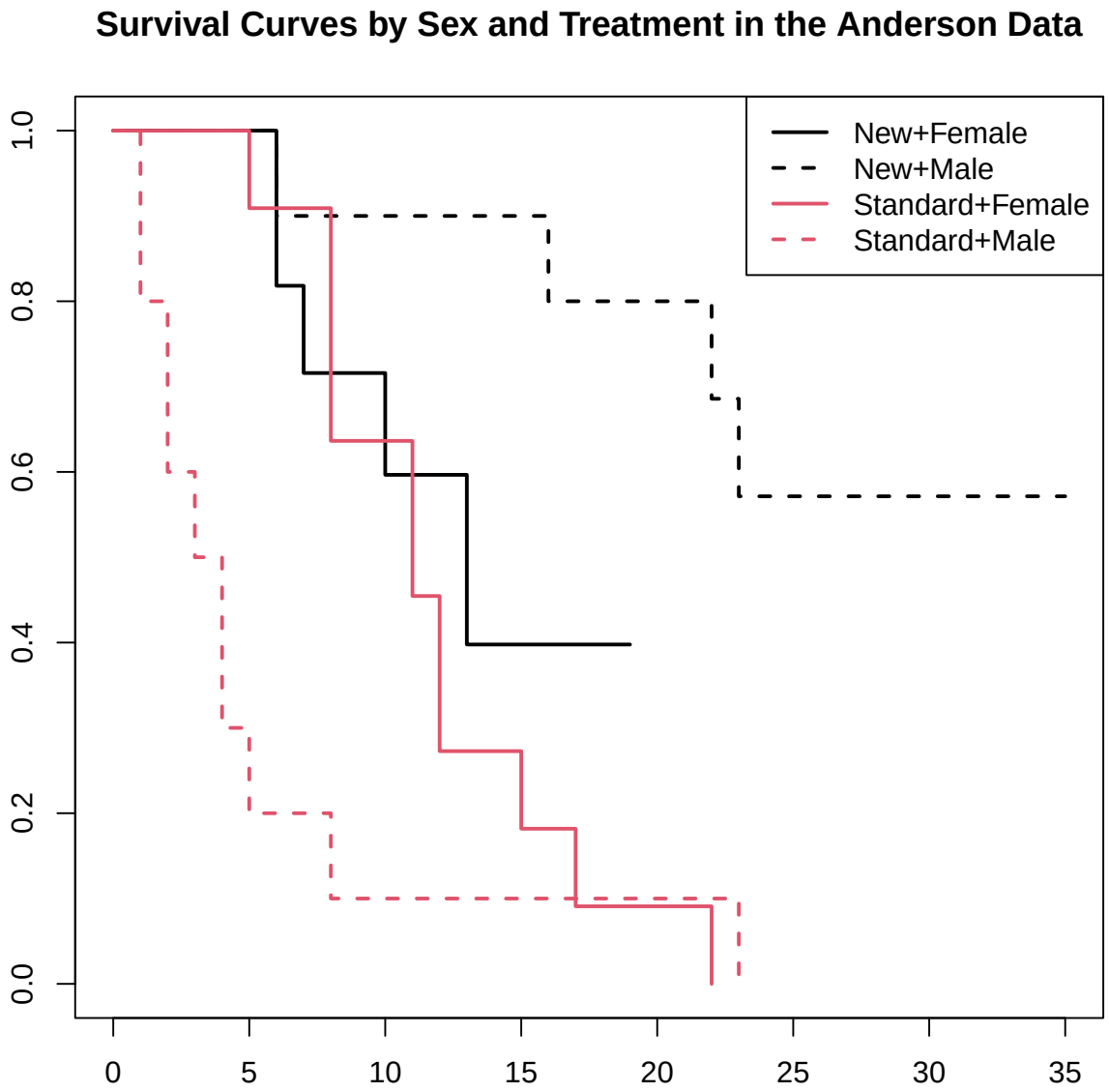
```
> coxph(anderson.surv~treat+lwbc+sex+lwbc:sex+treat:sex,data=anderson)

                          coef exp(coef) se(coef)     z     p
treatstandard          0.37481   1.45471  0.55452  0.68 0.499
lwbc                   1.06370   2.89707  0.47261  2.25 0.024
sexmale               -4.98338   0.00685  2.11360 -2.36 0.018
lwbc:sexmale           1.23031   3.42230  0.63008  1.95 0.051
treatstandard:sexmale  2.17816   8.83008  0.91095  2.39 0.017

Likelihood ratio test=57  on 5 df, p=5.18e-11
n= 42, number of events= 30

> cox.zph(coxph(anderson.surv~treat+lwbc+sex+lwbc:sex+treat:sex,data=anderson))
          chisq df    p
treat     0.136  1 0.71
lwbc      1.652  1 0.20
sex       1.266  1 0.26
lwbc:sex  0.102  1 0.75
treat:sex 1.637  1 0.20
GLOBAL    3.747  5 0.59
```

```
> pdf("SexTreatmentPlot.pdf")
> plot(survfit(anderson.surv~treat+sex,data=anderson),
      col=c(1,1,2,2),lty=c(1,2,1,2),lwd=2)
> legend("topright",c("New+Female","New+Male","Standard+Female","Standard+Male"),
      col=c(1,1,2,2),lty=c(1,2,1,2),lwd=2)
> title("Survival Curves by Sex and Treatment in the Anderson Data")
> dev.off()
```


Survival Curves by Sex and Treatment in the Anderson Data
New+Female
New+Male
Standard+Female
Standard+Male
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35

Cox Regression with the drug6mp Data

```
time12 <- c(drug6mp$t1,drug6mp$t2)
cens12 <- c(rep(1,21),drug6mp$relapse)
treat12 <- rep(1:2,each=21)
pairs12 <- rep(1:21,2)

drug6mp.surv <- Surv(time12,cens12)
drug6mp.cox1 <- coxph(drug6mp.surv~treat12)
drug6mp.cox2 <- coxph(drug6mp.surv~treat12+strata(pairs12))

print(survdiff(Surv(time12,cens12)~treat12))
print(survdiff(Surv(time12,cens12)~treat12+strata(pairs12)))
print(summary(drug6mp.cox1))
print(summary(drug6mp.cox2))
```

```
Call:
survdiff(formula = Surv(time12, cens12) ~ treat12)

            N Observed Expected (O-E)^2/E (O-E)^2/V
treat12=1 21       21     10.7      9.77      16.8
treat12=2 21        9     19.3      5.46      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4e-05
Call:
survdiff(formula = Surv(time12, cens12) ~ treat12 + strata(pairs12))

            N Observed Expected (O-E)^2/E (O-E)^2/V
treat12=1 21       21     13.5      4.17      10.7
treat12=2 21        9     16.5      3.41      10.7

 Chisq= 10.7  on 1 degrees of freedom, p= 0.001
```

```
coxph(formula = drug6mp.surv ~ treat12)

  n= 42, number of events= 30

           coef exp(coef) se(coef)      z Pr(>|z|)
treat12 -1.5721    0.2076   0.4124 -3.812 0.000138 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
treat12    0.2076      4.817   0.09251    0.4659
------------------------------------------------------------------------
coxph(formula = drug6mp.surv ~ treat12 + strata(pairs12))

  n= 42, number of events= 30

           coef exp(coef) se(coef)      z Pr(>|z|)
treat12 -1.7918    0.1667   0.6236 -2.873  0.00406 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
treat12    0.1667          6   0.04909    0.5658
```

```
coxph(formula = drug6mp.surv ~ treat12)

Concordance= 0.69  (se = 0.041 )
Likelihood ratio test= 16.35  on 1 df,   p=5e-05
Wald test            = 14.53  on 1 df,   p=1e-04
Score (logrank) test = 17.25  on 1 df,   p=3e-05
```

```
coxph(formula = drug6mp.surv ~ treat12 + strata(pairs12))

Concordance= 0.857  (se = 0.108 )
Likelihood ratio test= 11.89  on 1 df,   p=6e-04
Wald test            = 8.26  on 1 df,   p=0.004
Score (logrank) test = 10.71  on 1 df,   p=0.001
```

- With all of the logrank test in `survdiff`, the Wald coefficient test in `coxph`, and the full model tests by LR, Wald, and score, the tests without the conditioning on pairs via the `strata` term has too small a p-value by an order of magnitude.
- The original analysis was via a sequential procedure in which the fraction of pairs in which the placebo failed first was tracked until (as happened in this case) the fraction exceeded 0.75.

Real Study/"Toy" Data Sets

- The `drug6mp` data and the `anderson` data were both derived from a 1963 article in the leading journal **Blood** by Emil J. Freireich and colleagues.
- The original purpose of the study, the study design, and the conclusions about the outcome are not necessarily represented in the data sets used to illustrate methods.
- There is nothing wrong with altering the background of a data set to use it to illustrate methods, but it may be useful sometimes to look back to the original study.

Freireich et al. 1963

- This paper is focused on palliative care, which is meant to relieve symptoms and may extend life, rather than curative protocols meant to eliminate disease.
- For leukemia, extending remission may be palliative but may also extend life and may give more opportunity for potential curative therapy.

Freireich et al. 1963

The Effect of 6-Mercaptopurine on the Duration of Steroid-induced Remissions in Acute Leukemia: A Model for Evaluation of Other Potentially Useful Therapy

Emil J. Freireich, Edmund Gehan, Emil Frei III, Leslie R. Schroeder, Irving J. Wolman, Rachad Anbari, E. Oman Burgert, Stephen D. Mills, Donald Pinkel, Oleg S. Selawry, John H. Moon, B. R. Gendel, Charles L. Spurr, Robert Storrs, Farm Haurani, Barth Hoogstraten and Stanley Lee.
Blood, **21**, No. 6, June 1963, pp 699–716.

The existence of effective palliative therapy for acute leukemia has hampered the evaluation of new and potentially more effective therapeutic agents. Therapeutic trials with new agents are usually reserved for patients who have been treated with and have become refractory to the agents of proven value. Such patients have active acute leukemia at the onset of study. Because agents are studied for their ability to induce remissions and are not always effective, many patients expire during treatment and thus the number of patients that can receive a new agent is greatly diminished. Moreover, the study of the therapeutic and toxic effects of agents in such patients is frequently confused by the manifestations of the active leukemic process.

To overcome these problems, a study was designed to test the ability of a therapy to prolong the duration of a remission. A higher proportion of patients would be available early in the course of their illness for such a study. Moreover, the treatment of patients in whom the leukemic process is in remission would permit objective evaluation of pharmacologic and toxic properties of the agent. Finally, a study of remission maintenance uses a continuous variable, namely duration of remission compared to remission induction where a yes or no variable is used, and allows for the quantitative evaluation of an agent. Such a quantitative evaluation could be a basis for ranking of agents in man. This ranking could be of great aid to those concerned with the synthesis of new compounds and the testing of compounds in animal systems. 6-MP was selected as a known active agent to test such an experimental design.

RESULTS

A total of 97 patients with acute leukemia were entered into the study by the 11 participating institutions. Of these patients, 92 (95 per cent) were considered acceptable for analysis. . . . The other five (5 per cent) patients were rejected for the following reasons: two had different treatment administered in Phase I, one had no bone marrow at end of Phase I, one had drug error, and one was lost to follow-up in Phase I. The first patient was entered in April 1959 and the last one in April 1960. The decision to terminate the study was based on the analysis of the duration of remissions of 21 pairs of patients—this number resulting in the sample path crossing a boundary line of the restricted sequential procedure. The sequential design is explained in the Appendix.

SUMMARY

The effect of 6-MP therapy on the duration of remissions induced by adrenal corticosteroids has been studied as a model for testing of new agents. Ninety-two patients under age 20 entered the study and were accepted for analysis. Sixty-two (67 per cent) had complete or partial remissions induced by corticosteroids. Patients in remission were randomly assigned to maintenance therapy with either 6-MP or placebo. The median duration of 6-MP-maintained complete remissions was 33 weeks and for placebo, 9 weeks. A sequential experimental design was used to analyze remission times while the study was in progress. This resulted in the study being stopped after analysis of the remission times of 21 pairs of patients (42 patients). Overall survival was not significantly different for the two treatment programs, since patients maintained on placebo were treated with 6-MP when relapse occurred.

APPENDIX

The comparison of the lengths of remission maintained on 6-MP and placebo was made using a restricted sequential procedure originally developed by Armitage (1960). The patients at each institution were paired according to remission status, complete or partial, one patient receiving 6-MP and the other placebo by a random allocation. As the patients relapsed from remission, a preference was recorded for 6-MP or placebo depending upon which therapy resulted in the longest remission. The purpose of the sequential design was to enable the trial to be stopped as soon as it could be established that one of the treatments was superior to the other.

The trial was designed to be sensitive to a proportion of preferences of 0.75 favoring either therapy, i.e., if the true proportion of preferences for 6-MP or placebo was 0.75, the probability of concluding that the proper therapy was in fact superior was 0.95. This was roughly equivalent to trying to detect a 20-week difference in average remission times. If there was really no difference between the therapies the probability was 0.95 that the trial would end showing no real difference.

Sequential Analysis Plot

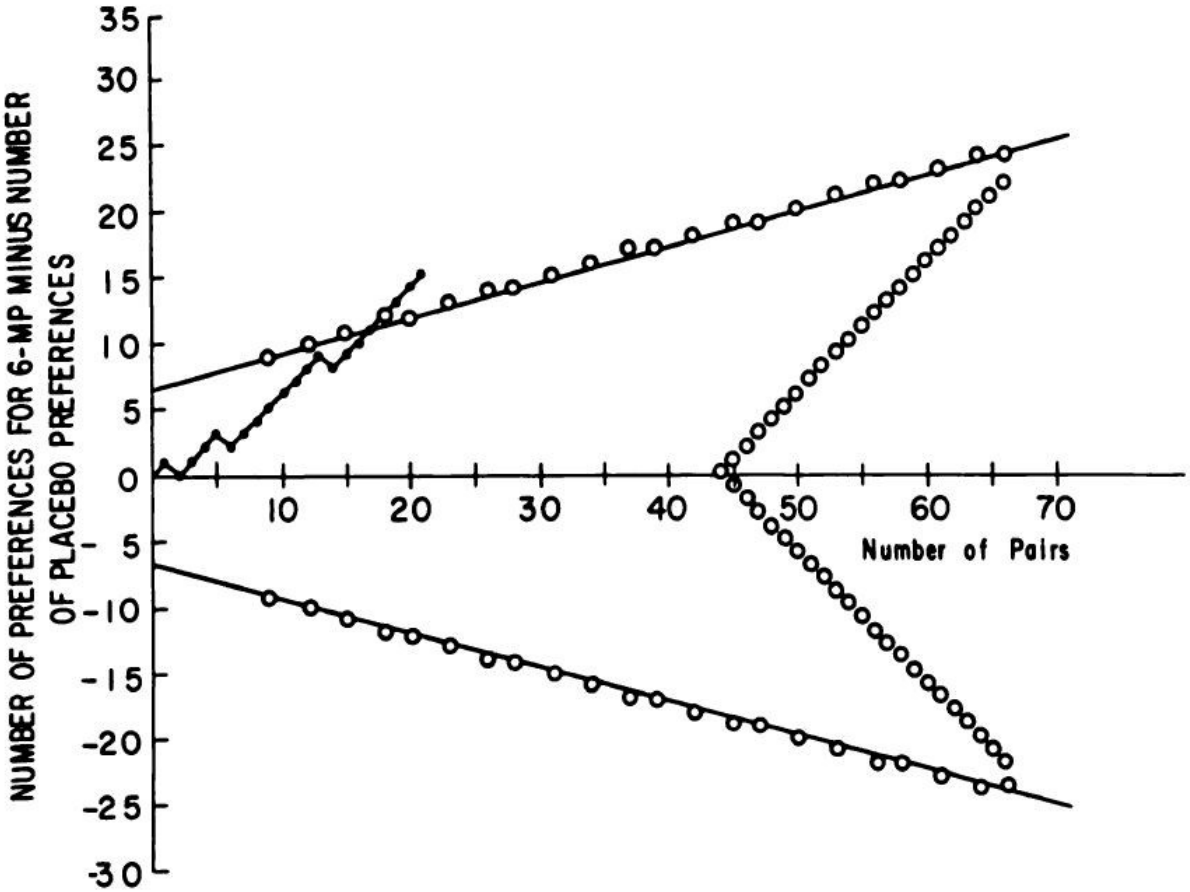


Fig. 8.—Chart for restricted sequential procedure applied to preferences designed to be sensitive to proportion of 6-MP or placebo preferences = 0.75.

When a preference was available from a pair of patients, a point was plotted on the chart, one unit to the right and one unit up from the zero point for a 6-MP preference and one unit to the right and one unit down for a placebo preference. Thus, the difference between the number of 6-MP and placebo preferences could be read from the vertical axis after a given number of preferences had been recorded. If the sample path crosses the upper (or lower) boundary, a decision is made favoring 6-MP (or placebo) . If the boundary to the right is crossed, then it is concluded that there is no real difference between treatments. The design provided a fixed upper limit to the number of patients entered in the trial. Thus, the maximum number of pairs of patients was 66 and the minimum number was nine.

The results from the pairs of patients entered in the trial are given .... The lengths of remissions are those that were available at the time the decision was made to halt the trial in April 1960. The decision was made to stop the trial after 21 preferences were recorded. Actually, the trial could have been stopped after 18 preferences were recorded, but data on remission times were gathered only about every 3 months, just prior to a group meeting. Note that 12 patients were still in remission at the time the study was stopped, though a preference could be recorded for each of the 21 pairs of patients. The entry of patients into the study was stopped while these 12 patients were still being studied. Of course, it is inherent in the design of such a trial that reliable data be submitted by each investigator. There is the danger that some of the patients will not be treated according to protocol and hence their data will be invalid. In this trial, the results from one pair of patients were later invalidated so it was fortunate that a number of extra pairs of patients was available.

Some Key Practical Points

- There were 31 pairs (62 patients) in the study, but only 21 pairs were reported, and one of those might not have been valid.
- A pair could be reported only when either 6-MP or placebo "won" which required that at least one patient had relapsed.
- Pairs (10) where neither have relapsed carry little or no information about the effect of the drug.
- It is not clear what would be done if the patients had a tied relapse time (months). Probably, this pair would be omitted unless the order could be determined.

- No patients died in this part of the study.
- Data were gathered only every 3 months and time to relapse or censoring is in months.
- Because of the paired randomization, analysis should be conditional on the pair (with `strata`). Alternatively, at least remission status should be used.

- This was a pediatric study of patients under 20:
  - 44 patients aged 0–4,
  - 25 patients aged 5–9,
  - 14 patients aged 10–14, and
  - 9 patients aged 15–19.

  Probably sex is not an important variable since most of the patients are pre-adolescent.