Applications in medical statistics - meta-analysis, nonparametric testing, and power calculations

Malcolm Hudson*
Professor, Department of Statistics
Macquarie University

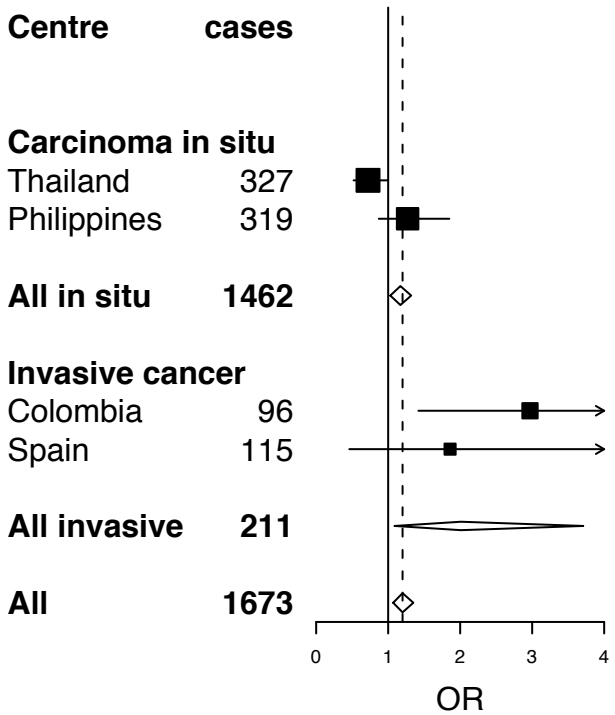
malcolm@ctc.usyd.edu.au

June, 2008

*Thank NHMRC CTC: supporting R-Workshop & ASC08 attendance

Meta-analysis graphics

Meta-analysis graphics



I. Meta-analysis graphics

Malcolm Hudson[1], Victor DeGruttola[2], Carol Hargreaves and Val Gebski[3]

[1] Macquarie University
[2] Harvard School of Public Health
[3] NHMRC Clinical Trials Centre

Womens Health Initiative study of HRT

Early stopping by the SDMC raised questions

Ethical issues Weigh individual risk of trial participants vs. community benefit

Statistical interpretation of findings *Over-estimate* risk of the adverse treatment effect (breast cancer) that led to stopping the trial;

- Statistical estimation of odds ratios requires adjusting for multiple outcomes
- Stopping rule based on a mix of outcomes (1 primary, 7 adverse) implies limited information about each.
- Should we adjust the OR of breast cancer down?

Specific effects (inducing trial-specific bias) apply to randomized trials

This talk: Graphic synthesis

Aim: to review variation in published HRT trial results and the potential for combining risk estimates from RCTs with those of cohort and case-control studies

Cross-design synthesis (CDS): synthesis of evidence from multiple (trial) sources and designs (RCTs, observational)
identify sources of variation in reported outcomes
appropriate identification, adjustments for bias
statistical model and methods
evaluation in meta-analysis of 28+ HRT studies

Issues bias- variance compromise
selection criteria for study inclusion in meta-analysis

Scope & Limitations uses reported summary statistics not IPD;
known within trial measurement uncertainty

Sources of bias in observational studies

"Observational evidence is clearly better than opinion, but it is thoroughly unsatisfactory." (Archibald Cochran)

Therapy is chosen to affect outcome.

Treatment imbalances: Confounding. *Why* did the patient get treatment?
Time origin: Time since study enrolment? Subject age?
Temporal change

In observational studies estimating HRT effect on breast cancer, necessary to allow for biases:

- earlier diagnosis, differential reporting of use,
- potential confounders: time since menopause, BMI, delay starting HRT after menopause, years of HRT
- lead to substantial underestimation of risk of breast cancer associated with the use of HRT[4].

[4] Collaborative Group on Hormonal Factors in Breast Cancer (HFBC) Lancet, 1997

Reducing bias

Exclusion strategy: In a meta-analysis Peto[5] excluded trials:

. . . "treatment assignment was not by strict randomisation"

Sources of bias in RCTs Lack of treatment concealment

Outcome evaluation not double blind

Study quality

In observational studies Stratification and model adjustment for confounders

[5] Stampfer, Goldhaber, Yusuf, Peto and Henneken (NEJM 307)

Meta analysis models and weighting

Single true meta effect (fixed effect) versus Inhomogeneity (random effects).

RE model (DerSimonian and Laird[6])

$$Y_j = \delta + u_j + e_j,$$

e_j, measurement error in the estimated treatment effect in study j, is distributed N(0, V_{0j}^2).

- Y_j is the apparent effect,
- δ – average (meta) effect of treatment,
- u_j, mean 0, variance σ_1^2, varies treatment effect due to specific study effects
- V_{0j} – measurement variance in the estimate of effect in study j.

Weightings of trial estimates are inverse to their variance: $V_{0j} + \sigma_1^2$

[6] DerSimonian & Laird, 1986

Cross-design RE models

Stratified binary outcomes: e.g. DerSimonian-Laird method with log odds-ratio estimates Y_j.

Study classes: e.g. randomised R, non-randomised NR. Postulate LME model:

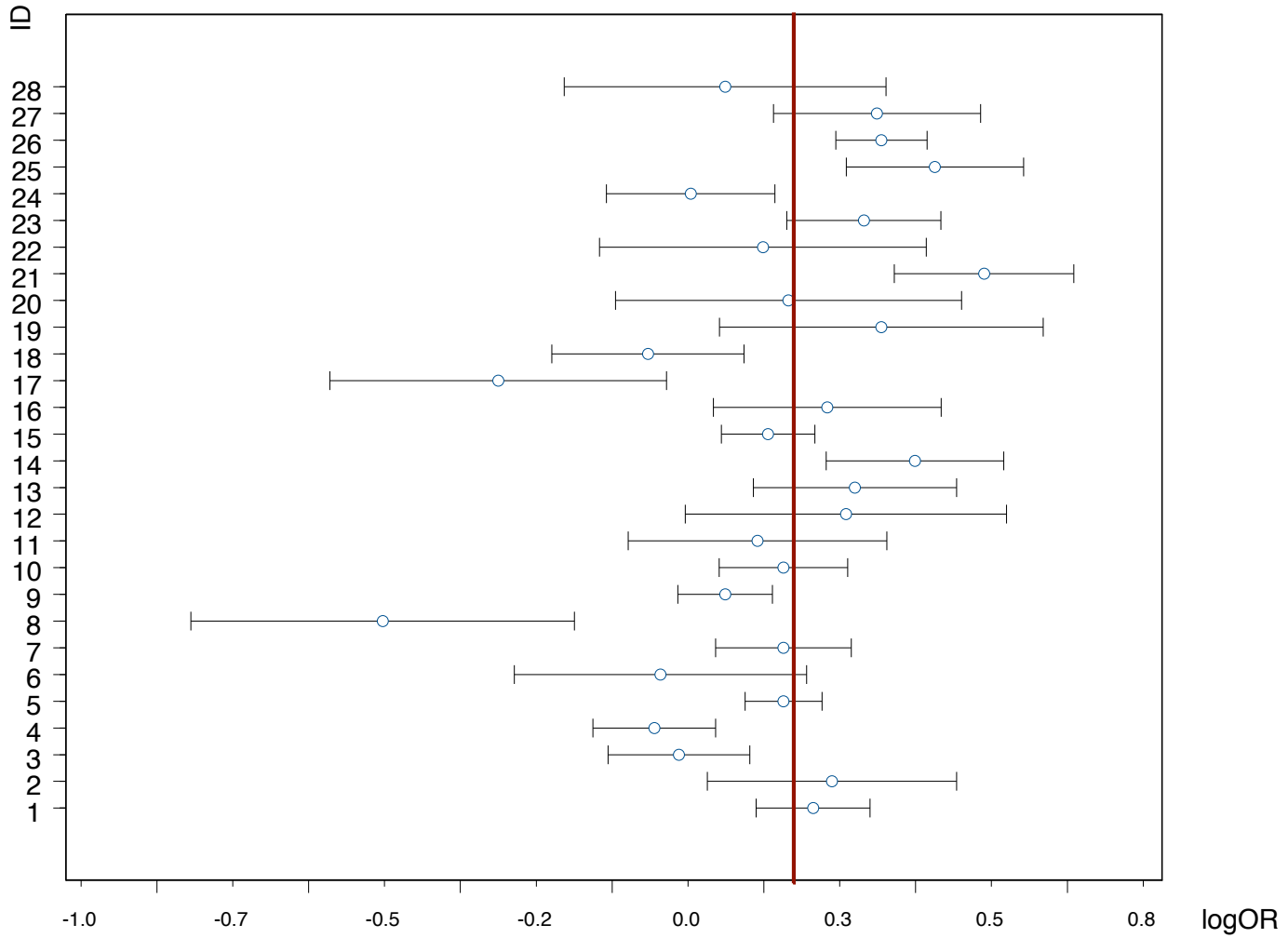
$$\begin{aligned} E(Y_j|u) = \mu + u_{j1} &\sim N(0, \sigma_1^2), \quad \text{for } j \in R \\ E(Y_j|u) = \mu + \delta + u_{j1} + u_{j2} &\sim N(0, \sigma_1^2 + \sigma_2^2), \quad \text{for } j \in NR. \end{aligned}$$

Notes:

- Introduces an extra source of variation in NR studies
- If $\delta = 0$, pooling class meta-estimates is legitimate.
- Not covered by DerSimonian-Laird theory.

Meta-Analysis: HRT studies up to WHI 2002

1. Included:
   - all studies included in the HFBC (1997) meta-analysis (RCTs 0);
   - published papers since this date (n=4, RCTs 2). Total N=28 estimates.
2. Goal: meta-estimate and display
3. Outcome: HRT effect on invasive breast cancer incidence
   Odds-ratio (adjusted) comparing HRT (ever) vs HRT never.
4. Trial types case-control (hospital controls; community based controls), prospective/ cohort,
   two recent randomized clinical trials.



EM Algorithm


```
{ # For two groups, R (n1 trials, type=1), NR (n2 trials, type != 1)
  # [snip: skip E-step and outer iteration loop]
  # M-step: update variance components d1 (sigma_1^2) and d2 (sigma_2^2)

  V <- V0 + d1 * rep(1, n)
  V[type!=1] <- V[type!=1] + d2 * rep(1, n2)
  w <- 1/V

  res <- (y-mu)
  ss1 <- sum(w^2* res^2)
  d1 <- ( d1^2 * ss1 + d1* (n - d1 * sum(w)) )/n
  d1var[itn] <- d1

  ss2 <- sum(w[type!=1]^2 * res[type!=1]^2)
  d2 <- ( d2^2 * ss2 + d2 * (n2- d2* sum(w[type!=1])) ) /n2
  d2var[itn] <- d2

  mu <- mu + sum(w * res)/sum(w0)
  means[itn] <- mu
  }

EM.Searle.100 <- EM3(y,V0, nitn=100, d1=0.0001,d2=0.0004, type=Study.Type)
```

Model estimates

log-Likelihood statistics; after 1000 EM iterations

Model	Parameter estimate	$-2l$	df
Homogeneous model	$\hat{\mu} = 0.186$		
no random effects	$\sigma_1^2 = 0$		
	$\sigma_2^2 = 0$	37.73	27
Heterogeneity but shared mean	$\hat{\mu} = 0.188$		
non-randomised studies only,	$\sigma_1^2 = 0$		
	$\hat{\sigma}_2^2 = 0.00684$	27.407	26
Heterogeneity but shared mean	$\hat{\mu} = 0.188$		
in both RCTs and NRCTs,	$\hat{\sigma}_1^2 = 0.00011$		
shared mean	$\hat{\sigma}_2^2 = 0.00672$	27.405	25

Findings

1. Odds ratios of risk of invasive breast cancer were generally consistent over the 28 studies, once stratified by age, parity, age at first child, years since menopause and BMI.
2. Exceptions can either be seen as 'outlier' trials, or as providing support for extra variation (or over-dispersion) in OR estimates among non-randomized studies (of any design class).
3. Outlier trials were indicated by a discrepancy between the naive variance $(1/a + 1/b + 1/c + 1/d)$ and the correct pooled variance after stratification.
4. In either case, there is extra variation but no statistical evidence of consistent bias when studies are classified by their design class.
5. The data is generally consistent with an average log OR comparing (HRT ever use) with (HRT never use) between 0.16 and 0.22 with 95% confidence.

R-notes

Meta-analysis in R: references

1. Paul Murrell, R Graphics, 2005, Chapman & Hall
2. Brian Everitt & Torsten Hothorn, A handbook of statistical analyses using R. 2006, Boca Raton: Chapman & Hall/CRC
3. Maindonald and Braun, Data Analysis and Graphics Using R, Second Edition, Cambridge
4. `http://cran.rproject.org/doc/vignettes/HSAUR/Ch_meta_analysis.pdf`
5. MiMa function, `http://www.wvbauer.com/downloads.html` to fit Meta-Analytic Mixed-, Random-, and Fixed-Effects Models.

Error Bar Plot via S-Plus Object Oriented Graphics (non R-compatible)

```
Data <- HRT5[, c("Y","ID","SE.Y")]
Data[,3] <- 1.96*Data[,3]

guiPlot( PlotType = "Horiz Error Bar",
    GraphSheet="ErrorBarPlot",
    DataSetValues = Data) #HRT5[, c("Y","ID","SE.Y")]

guiCreate("ReferenceLine", Name = "ErrorBarPlot$1$1",
  LineColor = "Black", LineStyle="Short Dash",
  Orientation = "Vertical", Position = 0)
```

⋮

```
guiModify( "Graph2D", Name = "ErrorBarPlot$1",
  PanelType = "Condition",
  ConditionColumns = "TT",
  ConditionType = "Discrete")
```

Single population, home grown code, snippet

```
"EM1" = function(y, V0, maxitn = 1, mu = sum(y/V0)/sum(1/V0), d1 = 0.2, cc1 =
  cc2 = 0.001)
{
  # Searle's algorithm (8.15)
  # d1 variance component
  # input logOR (unscaled) for specified subgp as y
  # V0 measurement variances
  # Searle's cgence criterion
```

Current environment

Mandriva Linux 2007.1

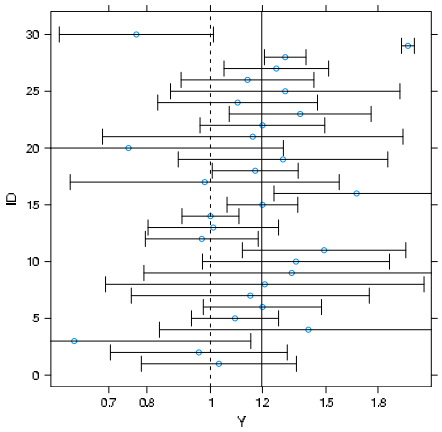
R ver 2.6

KDE 3.5.6

RKWard 0.4.6 R GUI interface (fantastic)

kile LaTeX editor

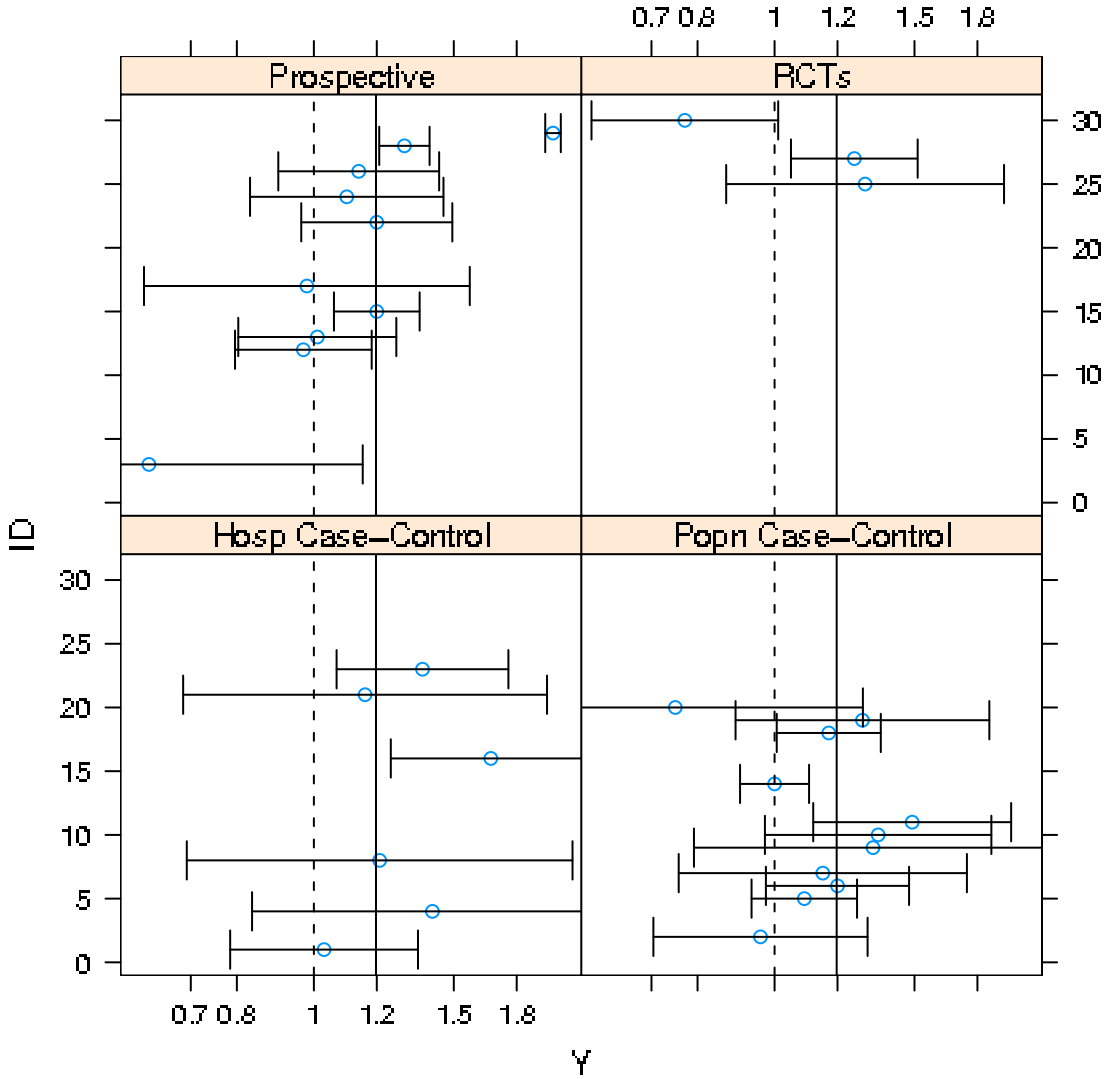
Lattice Display




```
library(lattice)
yscale <- round(exp(c(-0.4,-0.2,0,0.2,0.4,0.6)),1) # 1 decimal place

# basic plot, no grouping
xyplot(ID~Y, data=hrt5s, sd=hrt5s$SE.Y,
  panel=function(x,y,subscripts,sd,...){
  panel.xyplot(x,y,...)
  larrows(x-sd[subscripts],y,
    x+sd[subscripts],y,
    angle=90,code=3,len=0.1,#lwd=1/sd[subscripts])/4,
    ...)
  panel.abline(v=0,lty=2)
  panel.abline(v=0.18)
  },
  scales=list(x=list(at=log(yscale),labels=yscale))
  )
```

Lattice Display: grouped by study type



Meta Plot via Lattice Graphic


```
xyplot(ID~Y|TT, data=hrt5s, sd=hrt5s$SE.Y,
  panel=function(x,y,subscripts,sd,...){
  panel.xyplot(x,y,...)
  larrows(x-sd[subscripts],y,
    x+sd[subscripts],y,
    angle=90,code=3,len=0.1,#lwd=1/sd[subscripts])/4,
    ...)
  panel.abline(v=0,lty=2)
  panel.abline(v=0.18)
  },
  scales=list(x=list(at=log(yscale),labels=yscale))
  )
```

R packages: meta and rmeta

metabin(meta)	Meta-analysis of binary outcome data
metacont(meta)	Meta-analysis of continuous outcome data
metacum(meta)	Cumulative meta-analysis
metagen(meta)	Generic inverse variance meta-analysis
metainf(meta)	Influence analysis in meta-analysis
trimfill(meta)	Trim and fill method for meta-analysis
plot(meta)	meta-analysis plots

Type 'help(FOO, package = PKG)' to inspect entry 'FOO(PKG) TITLE'

rmeta:

der Simmonian and Laird RE, produces some nice graphs

Power comparisons

II. Power comparisons

Context: parametric and rank tests: grouped outcomes with zero-spike.
Survival trade-off outcomes:

- In cancer studies, preferences between treatments may depend on trading off discomfort and inconvenience for enhanced survival
- Two forms of outcome measure:
  - time trade-off (TTO): offer extra survival time
  - probability trade-off (PTO): offer higher probability of survival
  - minimum outcome necessary to make treatment worthwhile

Statistical analysis of survival trade-offs

STOs

- T: survival gain required for treatment to be worthwhile
- 50-70% of women judged a 1% improvement in 5 year survival rates or a 3 month improvement in life expectancy to make either 6 cycles of CMF or 4 cycles of AC worthwhile. [7]
- Analysis perspectives
  - underlying/latent continuous outcome?
  - ordinal discrete (esp. survival categories, e.g, 'low-realistic')?
  - mixture distribution?
    - both non-traders ($T = 0$, discrete) and continuous ($T > 0$) outcomes

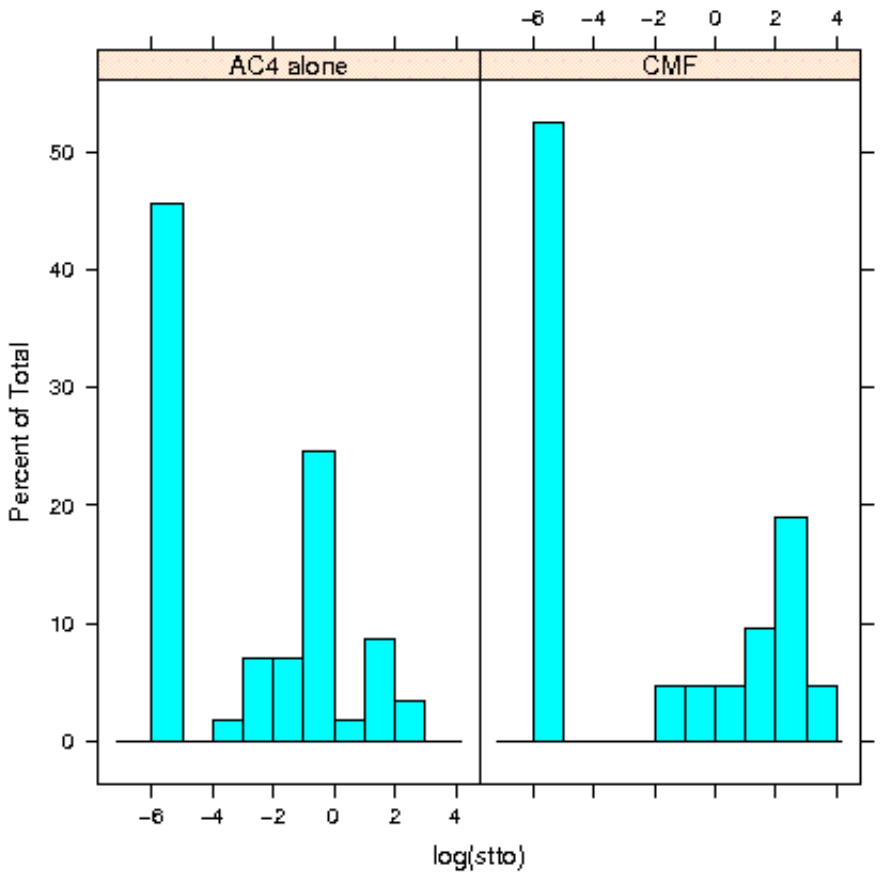
[7] Duric et al, Annals of Onc, 2005

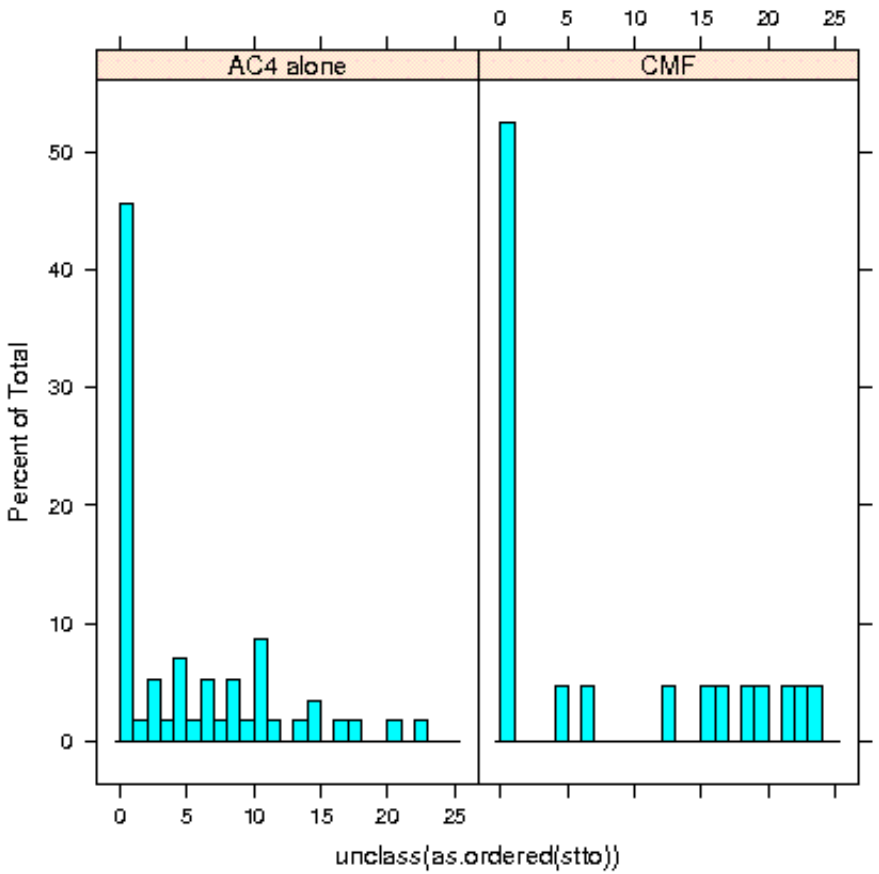
Continuous TTO inference

- T - time required for ACT to be worthwhile
  - t- test, 'log'-transformation (ad hoc)?
  - rank tests?
    - Wilcoxon-Mann-Whitney
    - Normal scores (common choice, underlying lognormal)?
    - rank tests are invariant to (monotone) transformation
  - discrete distributions (binning)?
    - observed outcomes are discrete (1 day, 1 month, 3 mths, . . .
    - pre-assign 'scores'
      - t-test, score STO levels using log
      - rank tests scores are the o.s. under a distribution

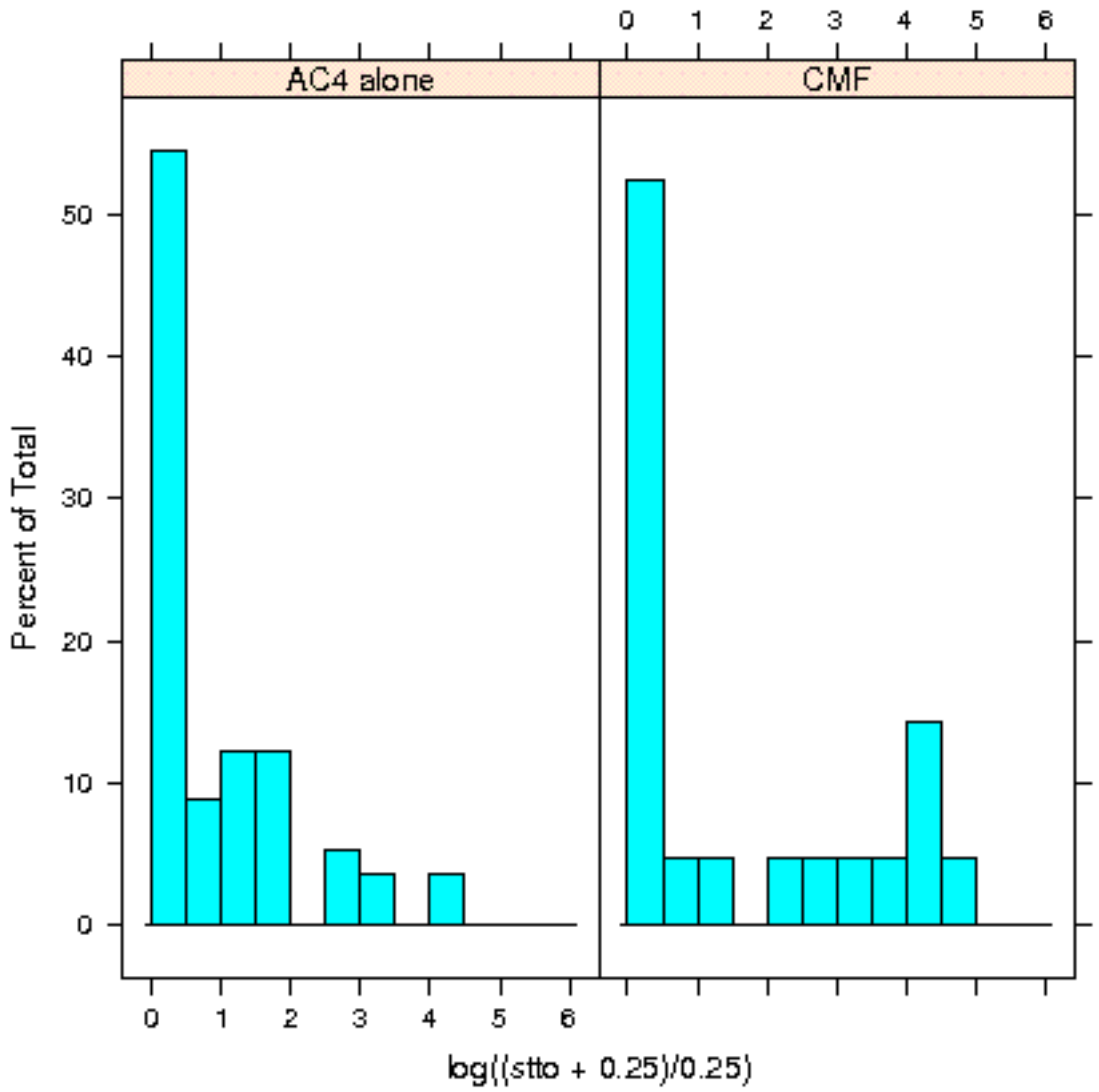
Comparisons by scores vs. ranks

AC4 vs CMF



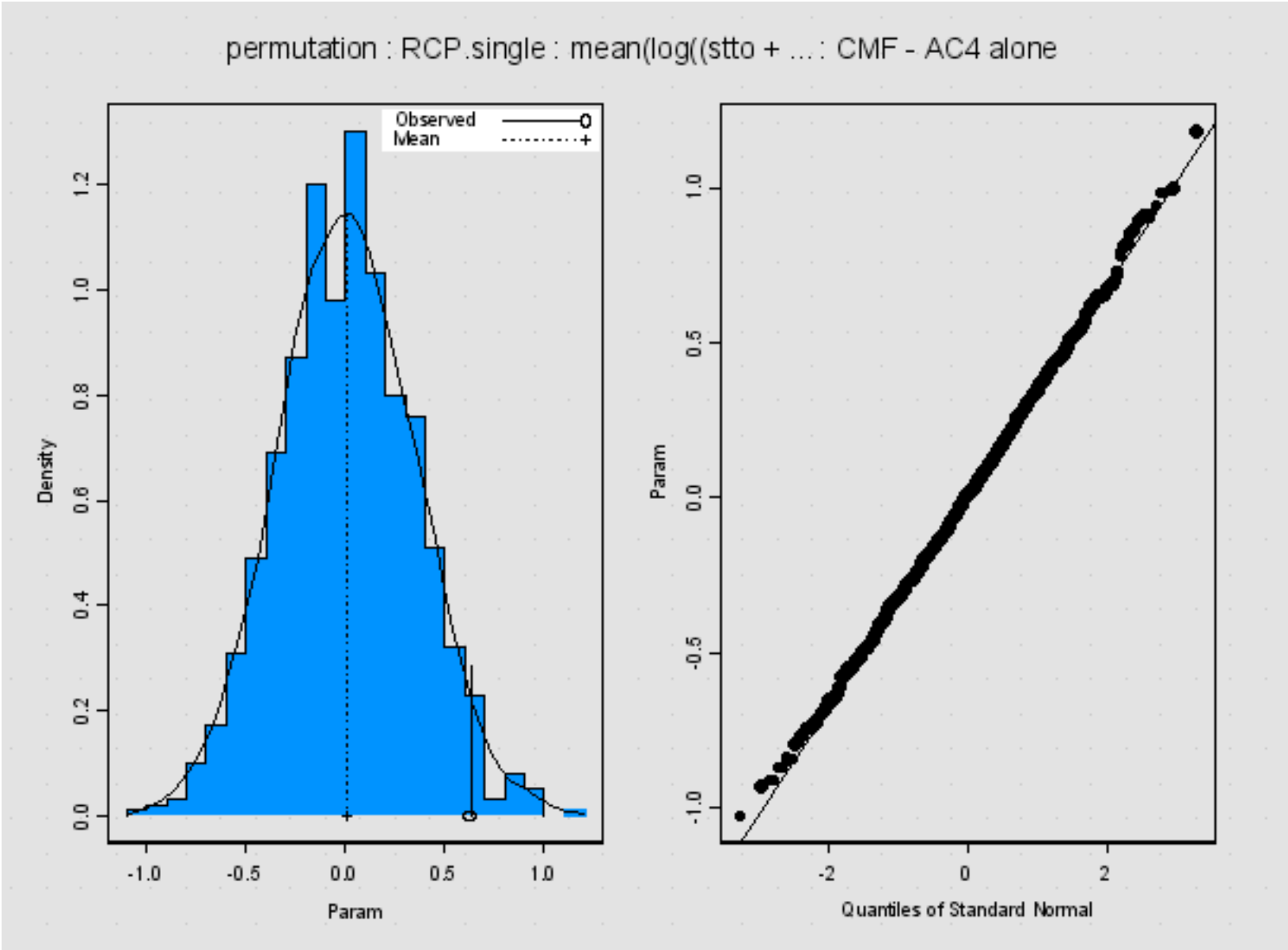


AC4 vs CMF



Effect of ties on P-values

statistic = mean difference in log((STTO+0.25)/0.25)



```
P=0.07

*** Permutation Test Results ***


Number of Replications: 999

Summary Statistics:
      Observed     Mean     SE alternative p.value
Param   0.6302 0.006444 0.3365   two.sided    0.07

Percentiles:
             2.5%          5%       95%     97.5%
Param -0.6539397 -0.5460052 0.5544847 0.6529697
```

Further analysis

- Consistency: logrank (CoxPH) with other tests
- Kruskal-Wallis, Normal scores, ordinal regression
- Logrank test P-values
- Effect of ties in Cox PH models?
- Ad hoc analysis by jittering to break ties
- Ad hoc analysis by t-test of log(1+TTO/0.25)

Simulation study goals

- Validity of P-values reported in discrete TTO data
  - based on asymptotic normality (finite sampling theory)
  - permutation distribution P-values are gold standard
- Power comparisons
  - location-shift alternatives to latent log-normal TTOs
  - alternative: multiplicative factor changes latent TTO
  - grouped in fixed intervals to form the discrete distributions
- Tests considered
  - log- scores (permutation t-test)
  - Wilcoxon (rank) test
  - Normal scores (rank) test
  - Exponential scores (Savage rank) test

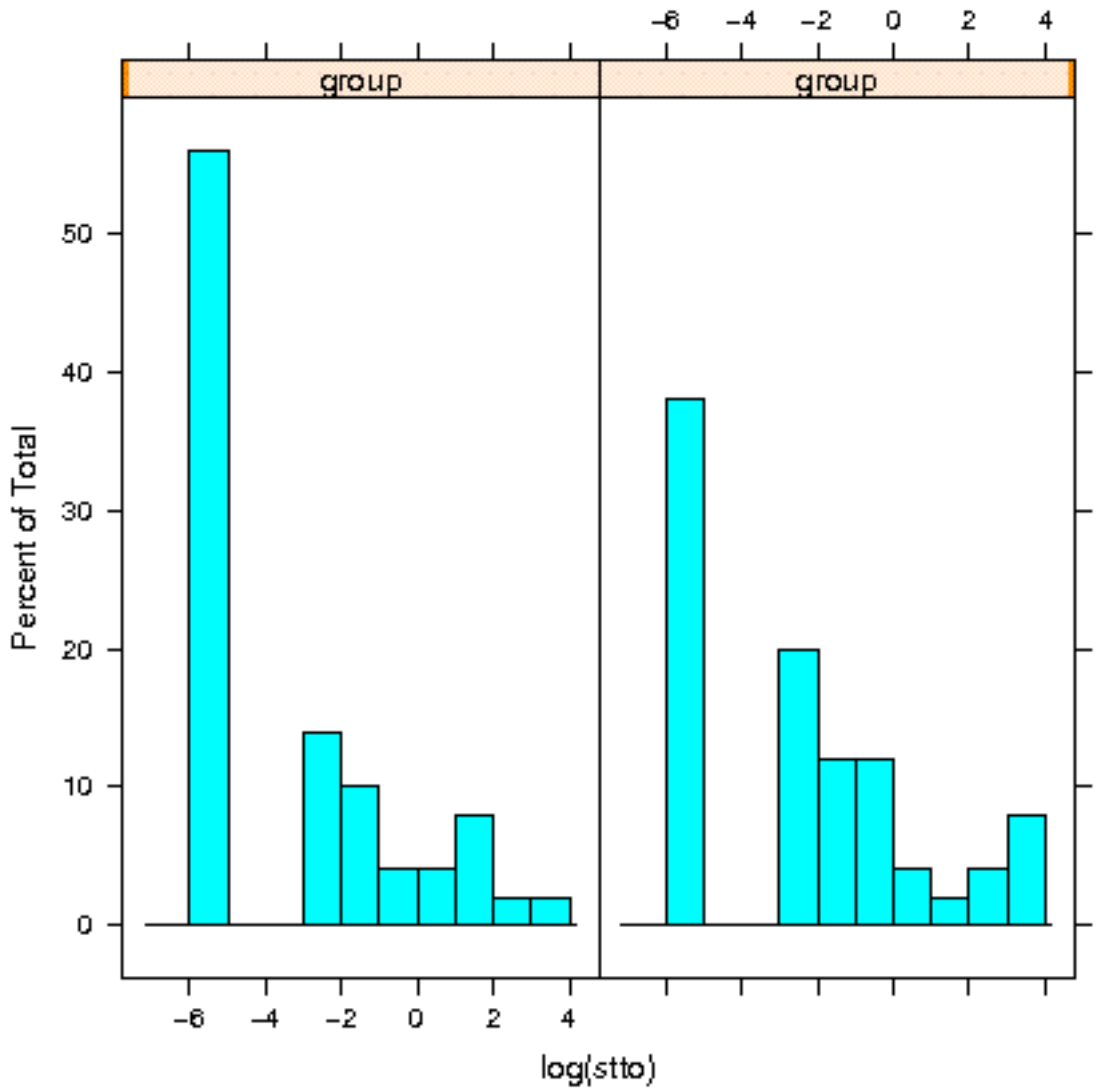
Simulated data: NULL effect

Null effect: type 1 error rates

Equal sample sizes	Effect: NULL			
N=100	Rejection rate			
Test	%	%	%	%
alpha	0.1	1	5	10
	%	%	%	%
Wilcoxon RS	0.06	0.92	5.0	9.9
Normal scores	0.07	0.90	5.0	9.9
(unconditional)	0.10	0.98	5.0	9.8
Logrank (exponential scores)	0.08	1.00	4.7	9.5
t-test (permutation)	0.02	0.69	4.7	9.8
(unconditional)	0.02	0.70	4.6	9.7

Table 1: Rejection rates

Location shift (log) alternative

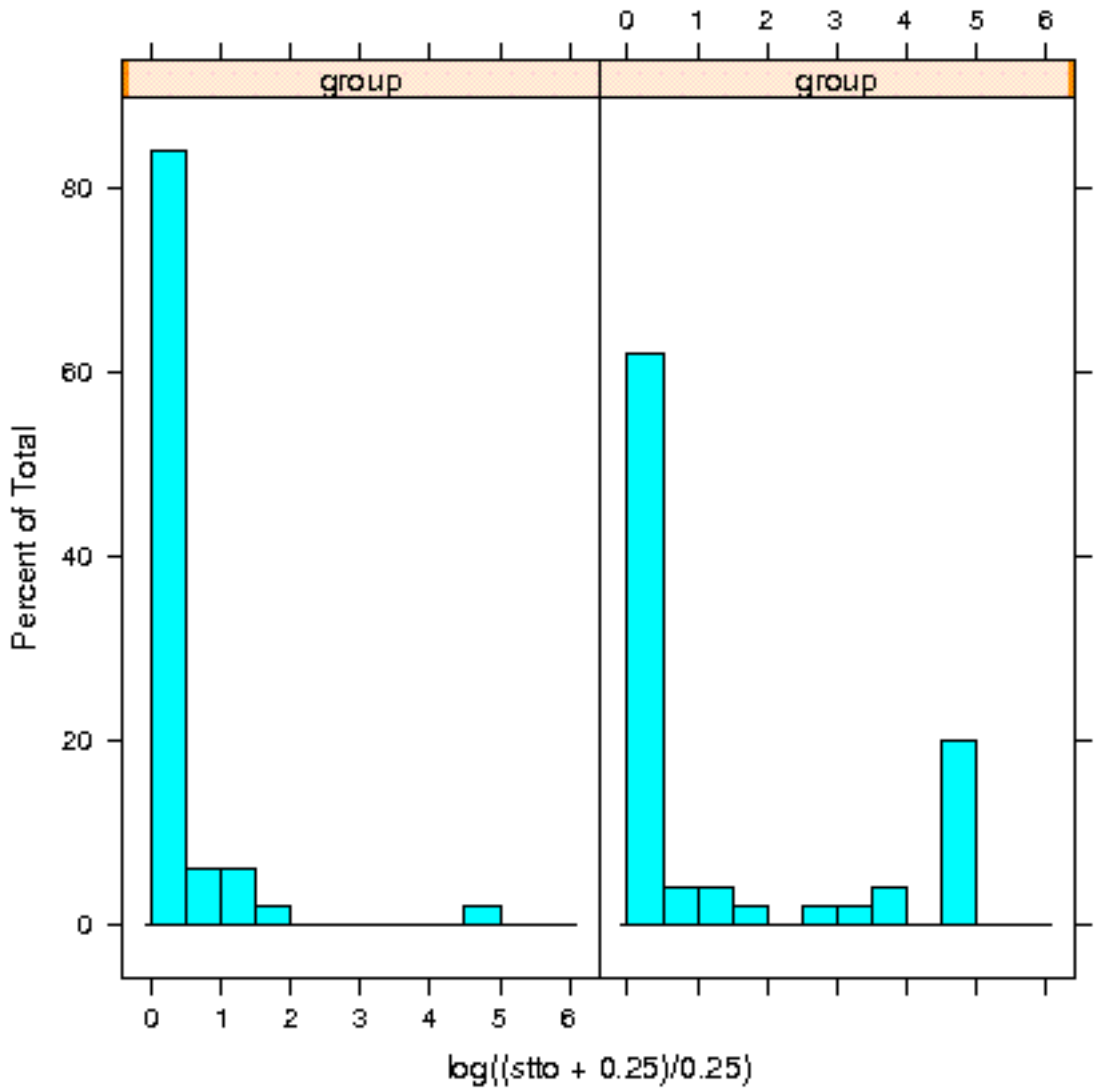


Effect: location shift

Equal sample sizes	Effect: SHIFT	0.5 SD		
N=100	Rejection rate*			
Test	%	%	%	%
alpha	0.1	1	5	10
	%	%	%	%
Wilcoxon RS	14	36	62	73
Normal scores	14	37	63	74
(unconditional)	15	38	63	74
Logrank (exponential scores)	13	32	57	68
t-test (permutation)	6	25	50	63
(unconditional)	7	25	50	63
*N=10000 replicated data sets				

Table 2: Power: SHIFT alternative

Polarisation



Effect: polarisation

Equal sample sizes	Effect: POLARISE		2.0*SD	
N=100	Rejection rate*			
Test	%	%	%	%
alpha	0.1	1	5	10
	%	%	%	%
Wilcoxon RS	0.6	5	15	24
Normal scores	1.2	8	22	32
(unconditional)	2	8	22	32
Logrank (exponential scores)	5	21	43	57
t-test (permutation)	6	25	50	63
(unconditional)	10	36	63	75
*N=10000 replicated data sets				

Table 3: Power: POLAR alternative

Conclusion

- Nominal type 1 error rates (finite sample asymptotics) are reliable for STO data
- Standard method, normal scores tests, Wilcoxon share good performance under translation shift alternatives
- Very poor power in heterogeneous groups, relative to permutation t-test and logrank test
- mixture model analysis
- log rank tests for TTO and STO data!
- agrees with ad hoc analysis: $\log(1 + T/0.25)$.

R-notes

coin: Conditional Inference

Exact and asymptotic permutation distribution probabilities:

T. Horthorn **R News**, Vol 1/1, January 2001, p11

oneway_test	two- and K-sample permutation test
wilcox_test	Wilcoxon-Mann-Whitney rank sum test
normal_test	van der Waerden normal quantile test
ansari_test	Ansari-Bradley test
fligner_test	Fligner-Killeen test
chisq_test	Pearsons χ^2 test
cmh_test	Cochran-Mantel-Haenszel test
lbl_test	linear-by-linear association test
surv_test	two- and K-sample logrank test
spearman_test	Spearmans test
wilcoxsign_test	Wilcoxon-Signed-Rank test

Related R bootstrap packages

boot: This package incorporates quite a wide variety of bootstrapping tricks.

bootstrap: A package of relatively simple functions for bootstrapping and related techniques.

coin: A package for permutation tests (discussed above).

MChtest: This package is for Monte Carlo hypothesis tests, that is, tests using some form of resampling. This includes code for sampling rules where the number of samples taken depend on how certain the result is.

permtest: A package containing a function for permutation tests of microarray data.

resper: A package for doing restricted permutations.

scaleboot: This package produces approximately unbiased hypothesis tests via bootstrapping.

simpleboot: A package of a few functions that perform (or present) bootstraps in simple situations, such as one and two samples, and linear regression.

Power Calculation code snippets

```
Nscores.2 <- normal.scores(stto.2)
nscores.out2 <- t.test(Nscores.2[group==0],Nscores.2[group==1])
nscores.out2$p.value
test.NS.2 <- sum(Nscores.2[group==1])
?replicate
sum(replicate(10000,sum(Nscores.2[sample(n,n1)]))>=test.NS.2)/10000
sum(replicate(10000,sum(Nscores.2[sample(n,n1)]))<=test.NS.2)/10000
Nscores.2.rep <- apply(stto.2.rep,2,normal.scores)
nscoresP <- t.test(normal.scores(stto.2)[1:50],normal.scores(stto.2)[51:100])$p.value
nscores.P.2 <- apply(Nscores.2.rep,2, function(x){t.test(x[1:50],x[51:100])$p.value})
summary(nscores.P.2)
qqplot(nscores.P.2,unif.os)
for(alpha in c(0.001,0.01,0.05,0.10)){
  print(sum(nscores.P.2 <= alpha)/10000)
  }
## more precise P
```

Power Calculation code snippets

```
norm.approx <- function(obs,scores,n1,N)
{ # approx permutation P-value from sampling without replacement mean, var
  # many scores are tied, but jittering leaves unchanged mu, V and sum of second group
  # test conditional on values observed, no continuity correction
mu <- n1* mean(scores)
s2 <- var(scores)
f <- n1/N
V <- n1*s2*(1-f)
z <- (obs-mu)/sqrt(V)
P1 <- pnorm(z)
P2 <- 1-P1
P <- ifelse(P1<=0.5, 2*P1, 2*P2)
list(z,P,P1,P2)
}
```

Power Calculation code snippets

```
nscores.P.2b <- apply(Nscores.2.rep,2, function(x){
          norm.approx(sum(x[51:100]),x,50,100)[[2]]
          })
summary(nscores.P.2b)
qqplot(nscores.P.2b,unif.os)
for(alpha in c(0.001,0.01,0.05,0.10)){
  print(sum(nscores.P.2b <= alpha)/10000)
  }
## Exponential scores rank test
```

Appendix

Women's Health Inititiative Study of HRT

Key features

- Second large RCT[8] on Estrogen/progestin vs. placebo
- First of a pair of RCTs conducted by WHI with different HRT treatments
- Primary outcome CHD, primary adverse outcome invasive breast cancer
- Healthy post-menopausal women aged 50-79 yrs
- Population sample (direct mailing campaign)
- Multiple outcomes – CHD, colorectal cancer, hip fractures, ...
- Global index of monitored outcomes: balancing risks and benefits

Controversial

- Settled advice to women
- Trial was stopped early (5 yrs vs 8.5 yrs) by the SDMC
- Stopping rule based on mix of outcome boundaries (1 positive, 8 adverse)
- Adverse boundaries were for breast cancer and 7 other outcomes
  the latter employed 1-sided $\alpha = 0.05/7$ boundaries
- Compliance: treatment non-compliance 25%-30% at 5 yrs

[8] WHI Investigators, JAMA 2002

Annals of Internal Medicine
Established in 1927 by the American College of Physicians

LETTER

Large-Database Research

Complement to Randomized Trials?

Jose A. Sacristan, MD; Javier Soto, MD, PhD; and Ines Galende, MD

15 May 1998 | Volume 128 Issue 10 | Page 875

To the Editor: We are disappointed by the emphasis that the articles on database research in the supplement published on 15 October 1997 place on sophisticated new mathematical models to control for confounding factors and by the classic commonplace that clinical database studies are "attractive alternatives to randomized trials" [1]. Using large databases to compare therapies remains controversial [2]. By design, databases record observations made in clinical practice. Because treatment decisions are not randomly allocated, any observed therapeutic effect may be due to unrecognized factors affecting the treatment allocation rather than the treatment itself.

It is surprising that a supplement focused on the future of databases did not mention new research methods, such as cross-design synthesis [3], directed toward the generation of results with an acceptable balance between internal and external validity. Specifically, cross-design synthesis proposes the assessment, adjustment, and combination of treatment effects obtained with randomized studies and database analyses.

Random effects/ Variance Component Model

Searle's random effects model: $Y = X\mu + \sum_{i=1}^{2} Z_i u_i + e$,
where $u_1 = (u_{11}, \ldots, u_{1N})^T$ and $u_2 = (u_{2N_1+1}, \ldots, u_{2N})^T$, X is an arbitrary design matrix for fixed effects, Z_1 is an NxN identity matrix and $Z_2 = \left[0^T : I\right]^T$ is $N \times N_2$, with $N = N_1 + N_2$.
The log-likelihood l is conveniently expressed as

$$-2l = \sum_j \log(V_j) + \sum_j \frac{(y_j - \mu)^2}{V_j}$$

where V_j represents the variance of the treatment summary outcome in trial j according to the model. For example, in the model with two strata:

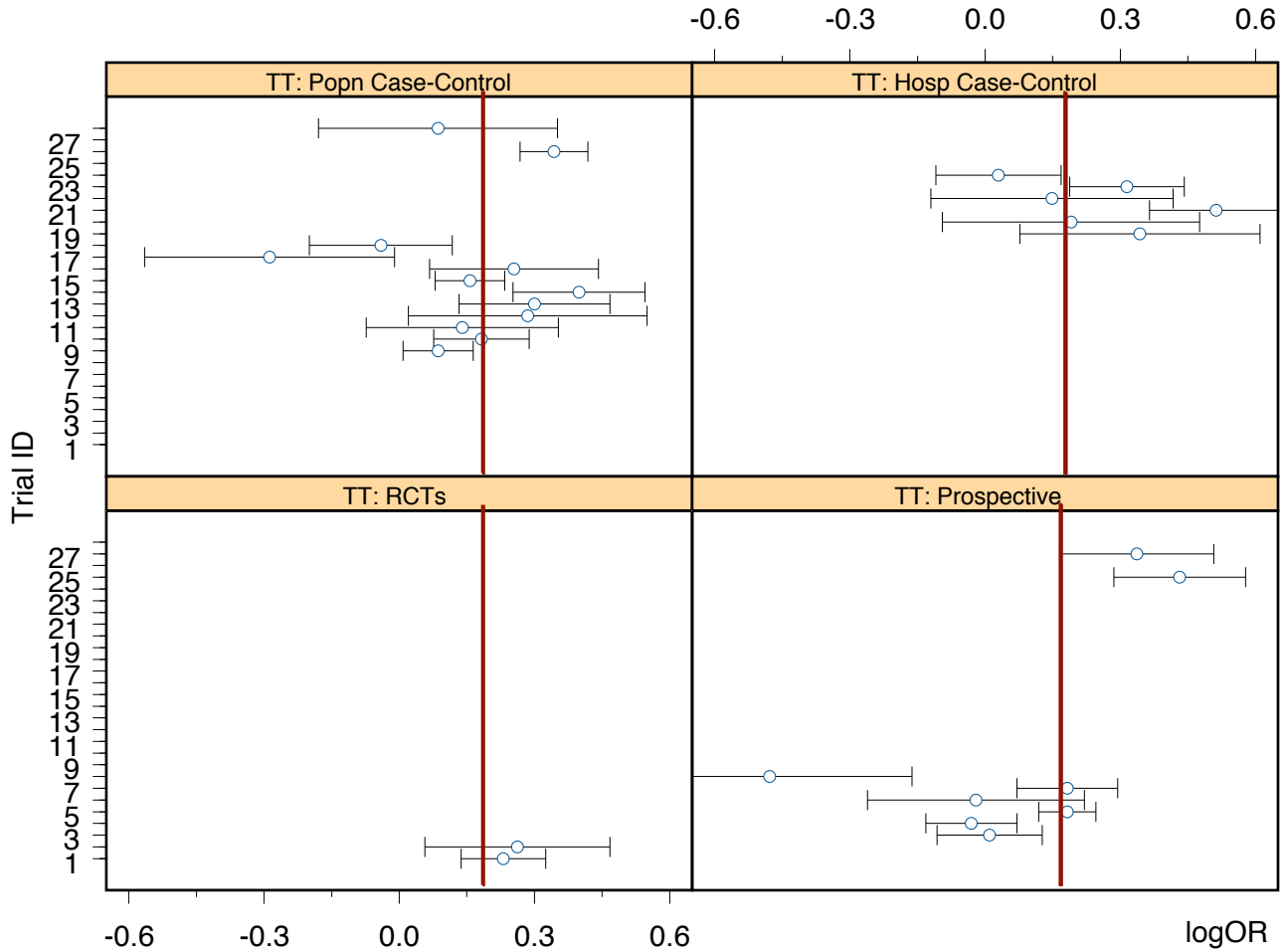
$$\begin{aligned} V_j &= V_{j0} + \sigma_1^2 \qquad \text{for } j = 1, \ldots, N_1 \\ &= V_{j0} + \sigma_1^2 + \sigma_2^2 \qquad \text{for } j = (N_1 + 1), \ldots, N \end{aligned} \tag{1}$$

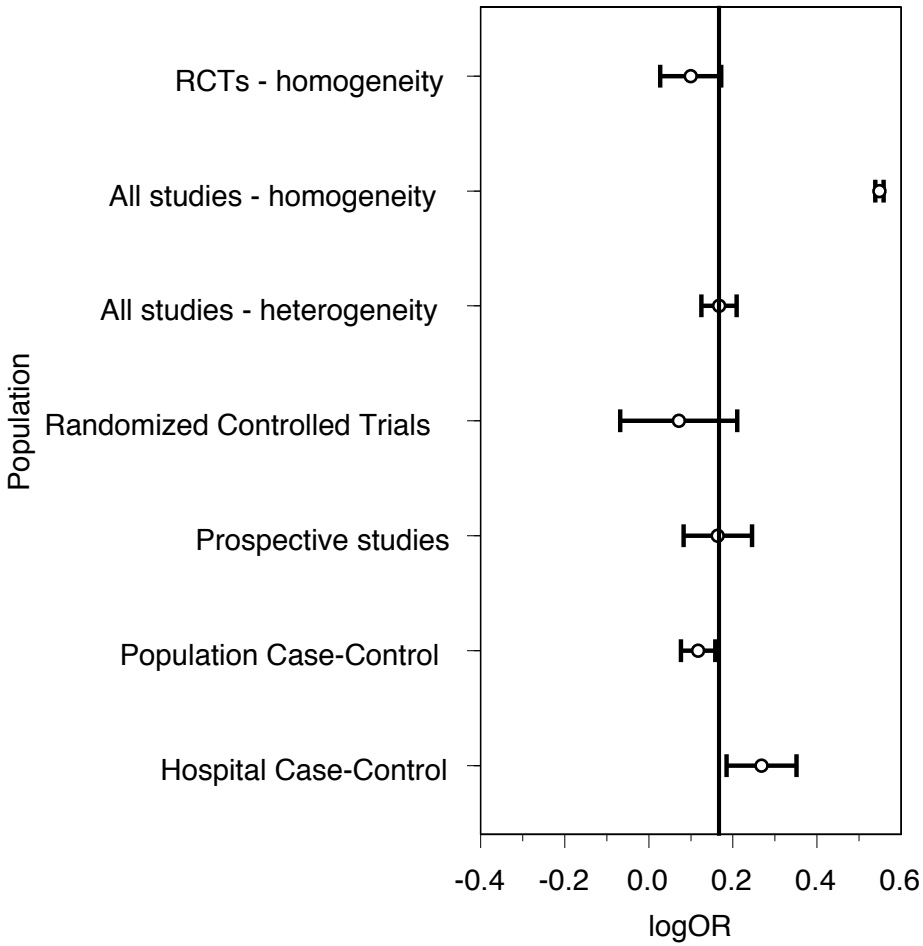
Inference

Nested models may readily be compared by difference in log-likelihoods, once variance parameters are estimated.

Differences in twice log-likelihood $-2\Delta l$ should be compared with *half* the tabled value for chi-square with degrees of freedom the number of extra variance parameters [9].

[9] Stram, Biometrics, 1994





References

- Hajek, Sidak, Sen Theory of rank tests, Wiley, 1998.
- Best Non-parametric comparisons of two histograms, Biometrics, 1994
- Buning et al. Power of generalised Wilcoxon test, Communications in Statistics
- Tanizaki Power comparisons of non-parametric tests: small-sample properties from Monte-Carlo experiments, 1997
- Varice, Weil, Exact non-null distributions of rank statistics, Communications in Statistics, 2001

Power Calculation code snippets

```
simul2 <- data.frame(stto=stto.2,group)

## R graphics
library(lattice)
histogram(~log(stto)|group,data=simul2,breaks=(-7:4))
dev.set (2)
dev2bitmap ("simul2plot1.png", type="png256", res=72.00000000)
histogram(~log((stto+0.25)/0.25)|group,data=simul2,breaks=(seq(0,6,by=0.5)))
```