The Role of Covariable Adjustment in the Analysis of Clinical Trials

Frank E Harrell Jr
Department of Biostatistics
Vanderbilt University School of Medicine
Nashville, Tennessee USA
f.harrell@vanderbilt.edu
biostat.mc.vanderbilt.edu

ABBOTT LABORATORIES N. CHICAGO IL 22 OCT 2010

Contents

1 Covariable Adjustment in Linear Models

- Model: $E(Y|X) = X\beta + \epsilon$
- Continuous response variable Y, normal residuals
- Statistical testing for baseline differences is scientifically incorrect (Altman & Doré 1990, Begg 1990, Senn 1994, Austin *et al.*2010)
- If we are worried about baseline imbalance we need to search patient records for counter–balancing factors
- $\rightarrow$ imbalance is not the reason to adjust for covariables
- Adjust to gain efficiency by subtracting explained variation
- Relative efficiency of unadjusted treatment comparison is $1 - \rho^2$
- Unadjusted analyses yields unbiased treatment effect estimate

2 Hidden Assumptions in 2×2 Tables

- Traditional presentation of 2–treatment clinical trial with a binary response: 2×2 table
- Parameters: P_1, P_2 for treatments 1, 2
- Test of goodness of fit: H_0: all patients in one treatment group have same probability of positive response (P_j constant)
- $\rightarrow H_0$: no risk factors exist
- Need to account for patient heterogeneity

3 Covariable Adjustment in Nonlinear Models

3.1 Models for Binary Response

- Model for probability of event must be nonlinear in predictors unless risk range is tiny
- Useful summary of relative treatment effect is the odds ratio (OR)
- Use of binary logistic model for covariable adjustment will result in an **increase** in the S.E. of the treatment effect (log odds ratio) (Robinson & Jewell, 1991)
- But even with perfect balance, adjusted OR $\neq$ unadjusted OR
- Adjusted OR will be greater than unadjusted OR

Example from GUSTO–I

- Steyerberg, Bossuyt, Lee[41]
- Endpoint: 30–day mortality (0.07 overall)
- 10,348 patients given accelerated t–PA
- 20,162 patients given streptokinase (SK)
- Means and Percentages

Characteristics of 30,000 GUSTO Patients

Baseline Characteristic	t–PA	SK
Age	61.0	60.9
Female	25.3	25.3
Weight	79.6	79.4
Height	171.1	171.0
Hypertension	38.2	38.1
Diabetes	14.5	15.1
Never smoked	29.8	29.6
High cholesterol	34.6	34.3
Previous MI	16.9	16.5
Hypotension	8.0	8.3
Tachycardia	32.5	32.7
Anterior MI	38.9	38.9
Killip class I	85.0	85.4
ST elevation	37.3	37.8

Unadjusted / Adj. Logistic Estimates

- With and without adjusting for 17 baseline characteristics

Unadjusted and Adjusted GUSTO Analyses

Type of Analysis	Log OR	S.E.	χ^2
Unadjusted	-0.159	0.049	10.8
Adjusted	-0.198	0.053	14.0

- Percent reduction in odds of death: 15% vs. 18%
- -0.159 (15%) is a biased estimate
- Increase in S.E. more than offset by increase in treatment effect
- Adjusted comparison based on 19% fewer patients would have given same power as unadjusted test

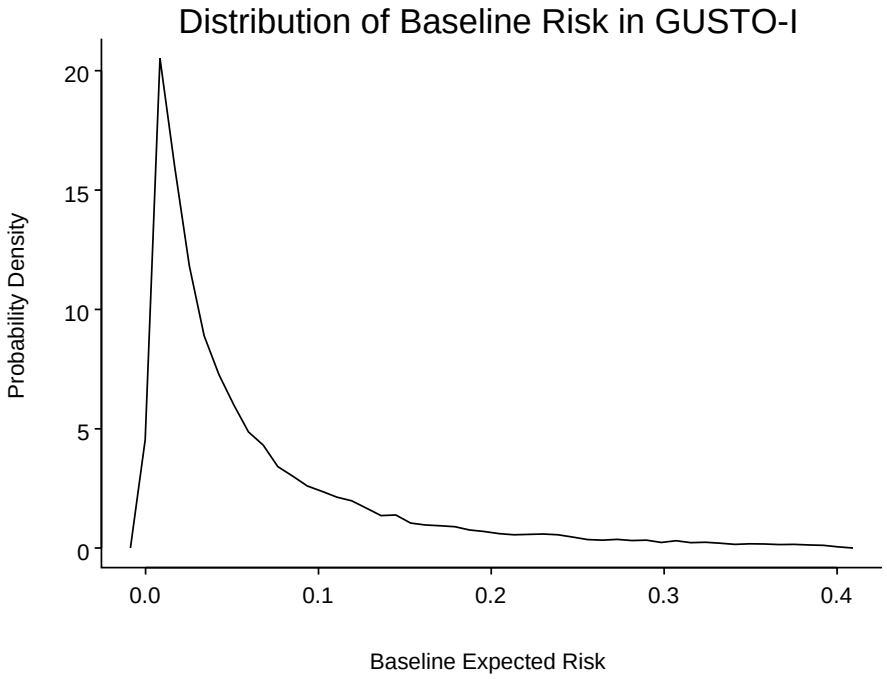

Figure 1: *Kernel density estimate of risk distribution for SK treatment. Average risk is 0.07. See also [21].*

- Robinson & Jewell: “It is always more efficient to adjust for predictive covariates when logistic models are used, and thus in this regard the behavior of logistic regression is the same as that of classic linear regression.”

Simple Logistic Example – Gail 1986

$X = -1$

	Treatment A	Treatment B
$Y = 1$	500	100
$Y = 0$	500	900
	1000	1000

Odds Ratio: 9

$X = 1$

	Treatment A	Treatment B
$Y = 1$	900	500
$Y = 0$	100	500
	1000	1000

Odds Ratio: 9

Pooled

	Treatment A	Treatment B
$Y = 1$	1400	600
$Y = 0$	600	1400
	1000	1000

Odds Ratio: 5.44

3.2 Nonlinear Models, General

- Gail, Wieand, Piantadosi 1984 showed that for unadjusted treatment estimates to be unbiased, regression must be linear or exponential
- Gail 1986 showed that for logistic, Cox, and paired survival models unadjusted treatment effects are asymptotically biased low in absolute value
- Gail also studied normal, exponential, additive risk, Poisson
- Senn [38, p. 3747] summarizes the problem as follows:

  > "Part of the problem with Poisson, proportional hazard and logistic regression approaches is that they use a single parameter, the linear predictor, with no equivalent of the variance parameter in the Normal case. This means that lack of fit impacts on the estimate of the predic-

tor."

4 Cox / Log–Rank Test for Time to Event

- Lagakos & Schoenfeld 1984 showed that type I error is preserved if don't adjust
- If hazards are proportional conditional on covariables, they are not proportional if omit covariables
- Morgan 1986 derived asymptotic relative efficiencies (ARE) of unadjusted log–rank test if a binary covariable is omitted
- If prevalence of covariable X is 0.5:

Efficiency of Unadjusted Log–Rank Test

$X = 1 : X = 0$ Hazard Ratio	ARE
1.0	1.00
1.5	0.95
2.0	0.88
3.0	0.72

- Ford, Norrie, Ahmadi 1995: Treatment effect does not have the same interpretation under unadjusted and adjusted models
- No reason for the two hazard ratios to have the same value
- Akazawa, Nakamura, Palesch 1997: Power of unadjusted and stratified log–rank test

Power With and Without Adjustment

Number of Strata	Range of Log Hazards	Power Unadj.	Power Adjusted
1	0	.78	–
2	0–0.5	.77	.78
	0–1	.67	.78
	0–2	.36	.77
4	0–3	.35	.77
8	0–3.5	.33	.77

4.1 Sample Size Calculation Issues

- Schoenfeld 1983 implies that covariable adjustment can only ↑ sample size in random-

ized trials

- Need to recognize ill–definition of unadjusted hazard ratios

5 Why are Adjusted Estimates Right?

- Hauck, Anderson, Marcus [17], who have an excellent review of covariable adjustment in nonlinear models, state:

> "For use in a clinician–patient context, there is only a single person, that patient, of interest. The subject-specific measure then best reflects the risks or benefits for that patient. Gail has noted this previously [ENAR Presidential Invited Address, April 1990], arguing that one goal of a clinical trial ought to be to predict the direction and size of a treatment benefit for a patient with specific covariate values. In contrast, population–averaged estimates of treatment effect compare outcomes in groups of patients. The groups being compared are determined by whatever covariates are included in the model.

The treatment effect is then a comparison of average outcomes, where the averaging is over all omitted covariates."

6 How Many Covariables to Use?

- Try to adjust for the bulk of the variation in outcome[17, 42]
- Neuhaus[29]: "to improve the efficiency of estimated covariate effects of interest, analysts of randomized clinical trial data should adjust for covariates that are strongly associated with the outcome"
- Raab *et al.*[31] have more guidance for choosing covariables and provide a formula for linear model that shows how the value of adding a covariable depends on the sample size

7 Other Uses of Modeling

7.1 Interactions

- Assessing differential treatment effect best done with formal interaction tests rather than subgroup analysis
- Pre–specify sensible effect modifiers
  - interactions between treatment and extent of disease
  - "learned" interventions: interaction between treatment and duration of use by physician
- Interactions with center are not necessarily sensible
- Need to use penalized estimation (e.g., interaction effects as random effects) to get sufficient precision of differential treatment effects, if # interaction d.f. > 4 for example [33, 44]

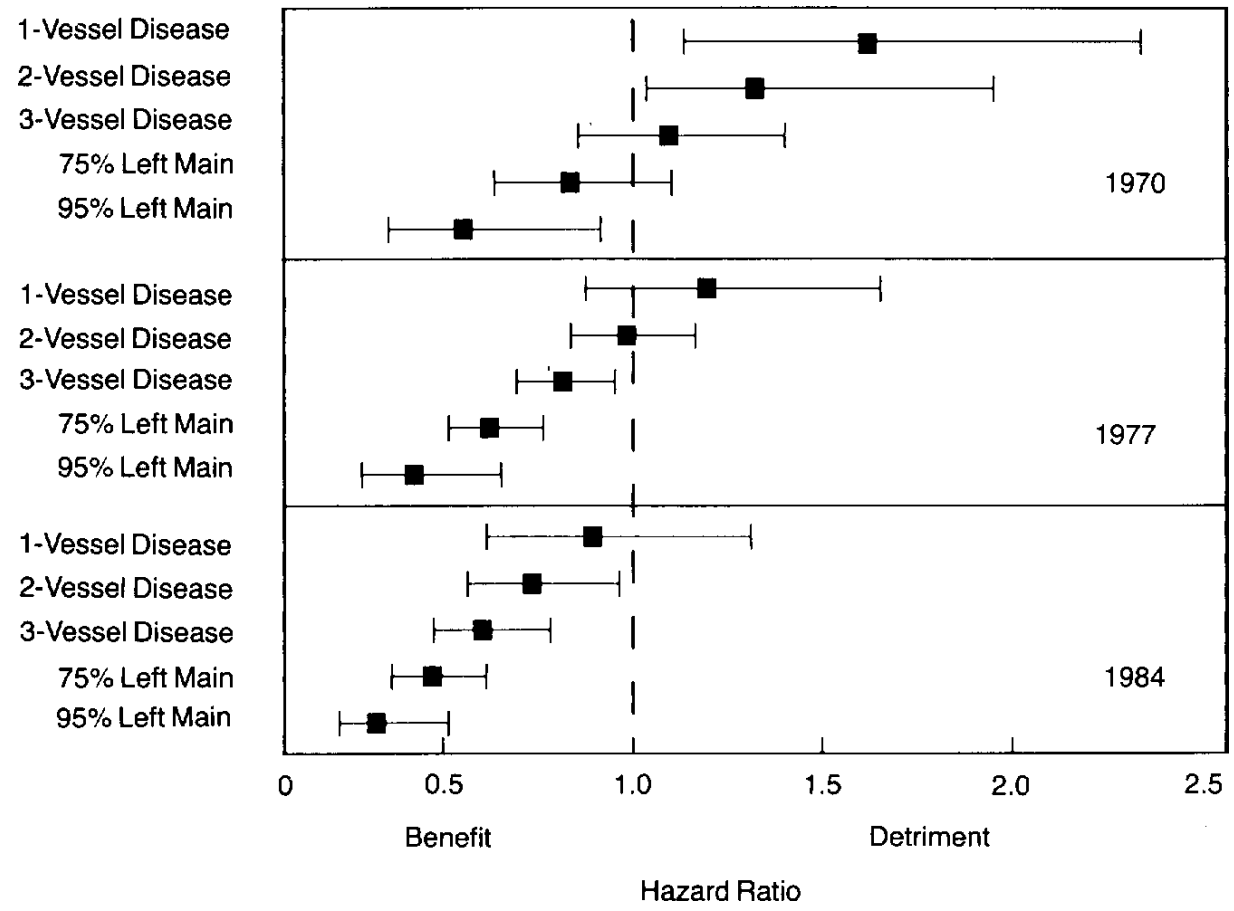

Figure 2: *A display of an interaction between treatment, extent of disease, and calendar year of start of treatment*[6]

7.2 Estimating Absolute Treatment Effects

- Absolute efficacy measures:
  - Risk difference (δ)
  - number needed to treat (reciprocal of risk difference)
  - Years of life saved
  - Quality–adjusted life years saved
- Binary response, no interactions with treatment, risk for control patient P:
  $\delta = P - \frac{P}{P+(1-P)/OR}$
- δ is dominated by P

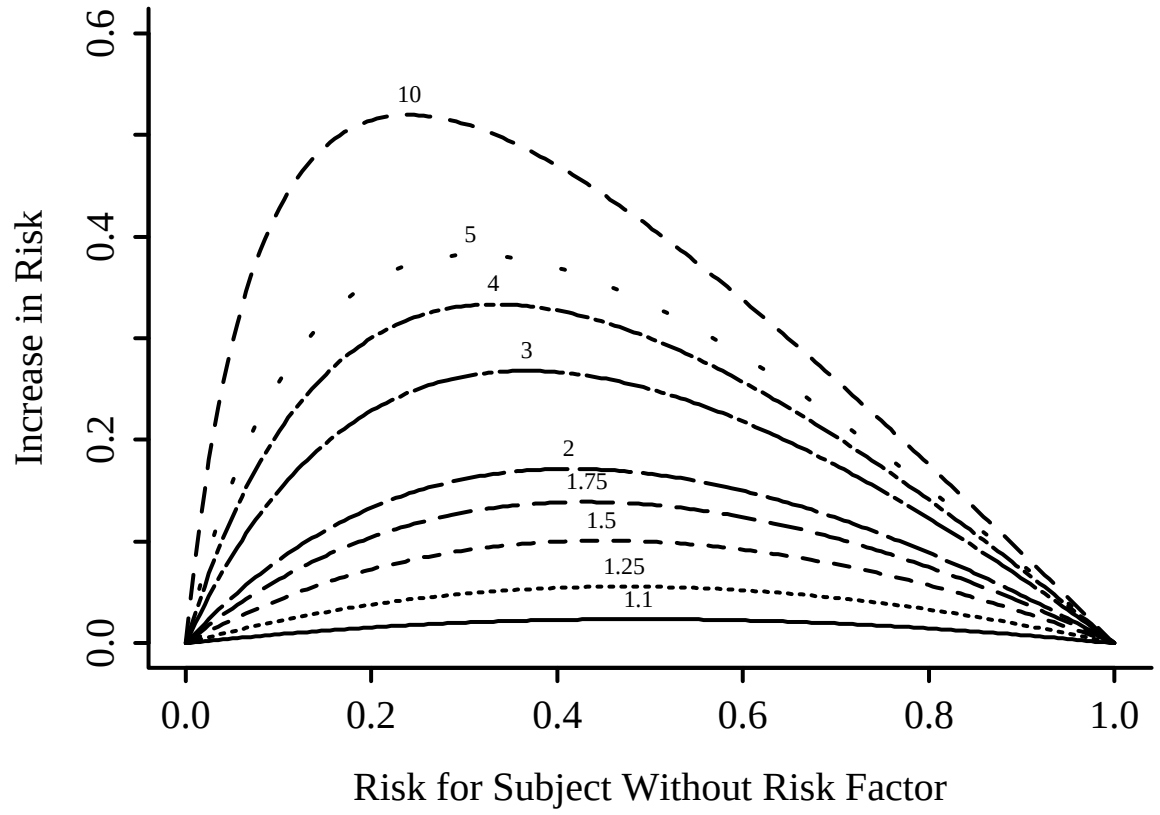

Figure 3: *Absolute risk reduction as a function of risk for control subject. Numbers on curves are treatment:control odds ratios.*

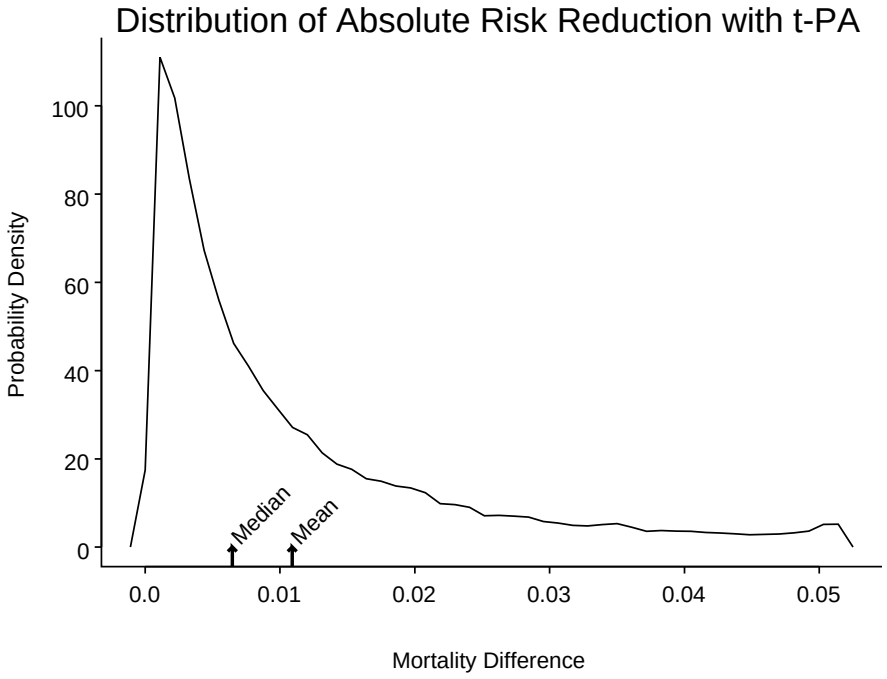

Figure 4: *Distribution of absolute benefit of t–PA vs. SK*

Absolute Treatment Effects for GUSTO–I

- No evidence for interactions with treatment
- Misleading subgroup analysis showed that elderly patients not benefit from t–PA; result of strong age $\times$ Killip class interaction
- Wide variation in absolute benefit of t–PA
- Overall mortality difference of 0.011 dominated by high–risk patients

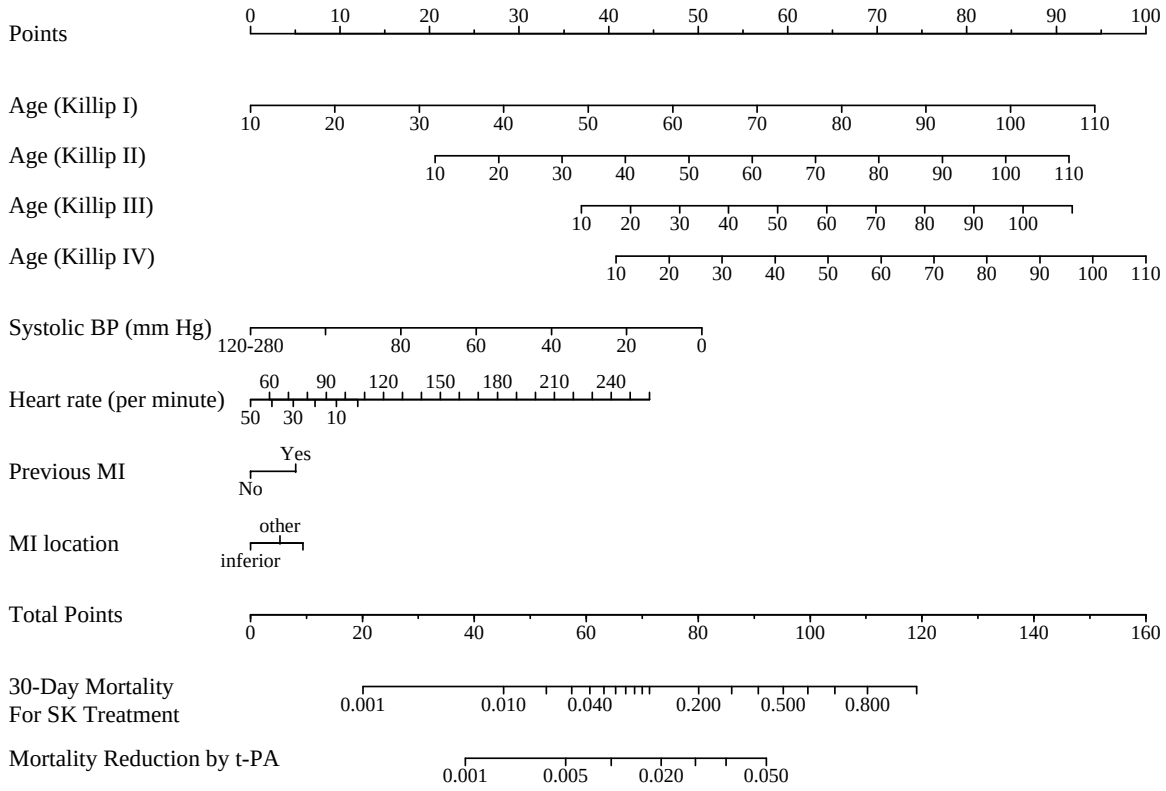

Figure 5: *Nomogram to predict SK - t–PA mortality difference, based on the difference between two binary logistic models.*

Absolute Benefit on Survival Prob.

- Cox PH model
- Modeling can uncover time course of treatment effect
- X_1 = treatment, $A = X_2, \ldots, X_p$ adjustment variables
- Survival difference is
  $S(t|X_1 = 1, A) - S(t|X_1 = 0, A)$
  $= S(t|X_1 = 0, A)^{HR} - S(t|X_1 = 0, A)$
- See also[23].

8 Cost–Effectiveness Ratios

- Effectiveness E (denominator of C–E ratio) is always absolute
- Absolute treatment effectiveness varies greatly with patient characteristics
- $\rightarrow$ C–E ratio varies greatly

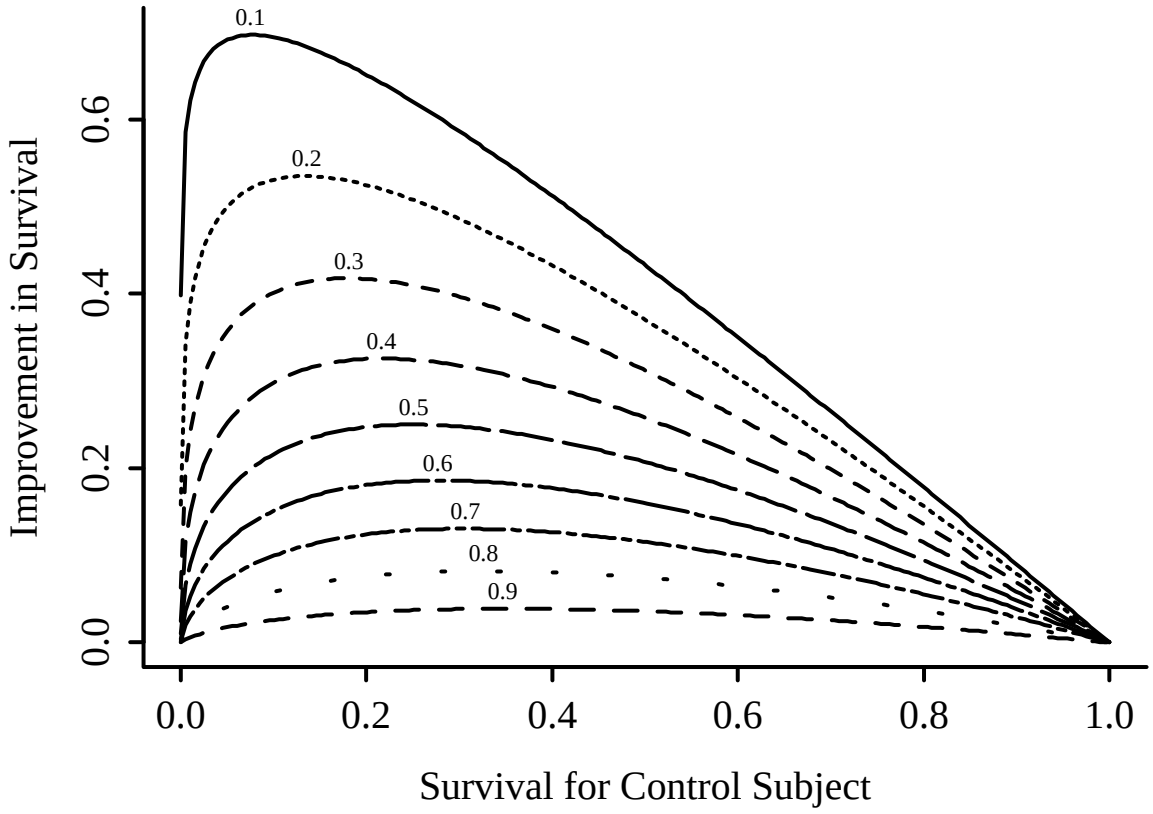

Figure 6: *Relationship between baseline risk, relative treatment effect (hazard ratio — numbers above curves) and absolute treatment effect.*

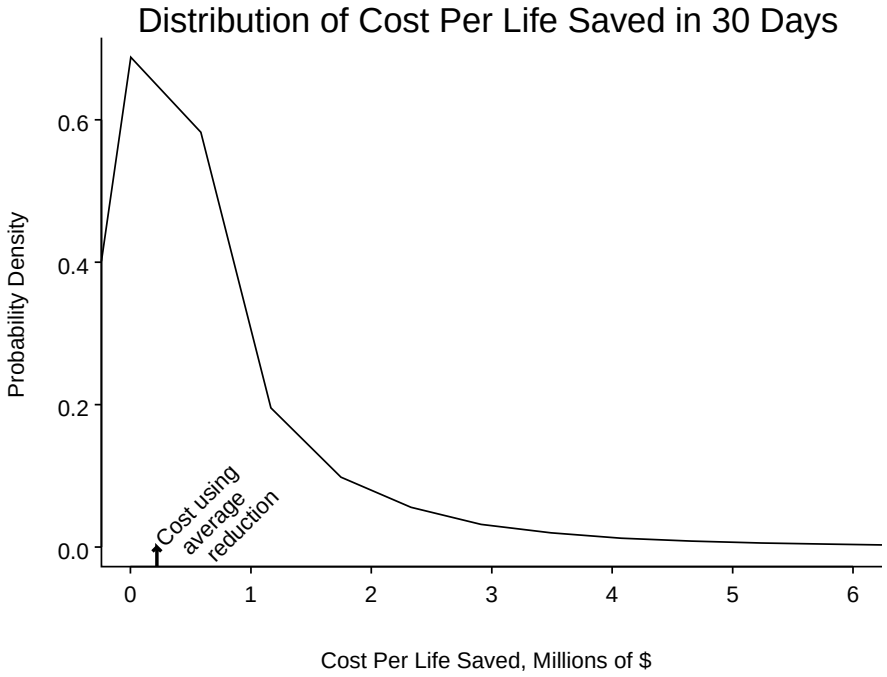

Figure 7: *Distribution of cost per life saved in GUSTO–I*

- A C–E ratio based on average E and average C may not apply to any existing patient!
- Need a model to estimate E
- C may also depend on patient characteristics

9 Treatment Contrasts for Multi–Site Randomized Trials

- Primary model: covariables, treatment, site main effects
- Planned secondary model to assess consistency of treatment effects over sites (add site $\times$ treatment interactions)
- Advantages for considering sites as random effects (or use penalized MLE to shrink site effects, especially for small sites). See [3] for a random effects Cox model and a demonstration that treatment effects may be inconsistent when non–zero site main effects are ignored in the Cox model. See also [44].
- Types of tests / contrasts when interactions are included[35]:
  - Type I: not adjusted for center
  - Type II: average treatment effect, weighted by size of sites

    S-PLUS or `R Design` library command:

```
sites ← levels(site)
contrast(fit, list(treat='b', site=sites),
              list(treat='a', site=sites),
         type='average', weights=table(site))
```

– Type III: average treatment effect, unweighted

```
contrast(fit, list(treat='b', site=sites),
              list(treat='a', site=sites), type='average')
# Built-in to S-Plus: anova(fit from lm, ssType=3)
```

Low precision; studies are not powered for Type III tests.

- Another interesting test: combined effects of treatment and site $\times$ treatment interaction; tests whether treatment was effective at *any* site.

10 Statistical Plan for Randomized Trials

- When a relevant dataset is available before the trial begins, develop the model from the dataset and use the predicted value as a single adjustment covariable in the trial (Knaus et al. 1993)
- Otherwise: CPMP Working Party: Finalize choice of model, transformations, interactions before merging treatment assignment into analysis dataset.
  Edwards[9]: Pre–specify family of models that will be used, along with the strategy for selecting the particular model.
  Masked model derivation does not bias treatment effect.
- New CPMP guidance[7]
  - "Stratification may be used to ensure balance of treatments across covariates; it may also be used for administrative reasons. The factors that are the basis of stratification should normally be included as covariates in the primary model.
  - Variables known a priori to be strongly, or at least moderately, associated with the primary outcome and/or variables for which there is a strong clinical rationale for such an association should also be

considered as covariates in the primary analysis. The variables selected on this basis should be pre-specified in the protocol or the statistical analysis plan.

- Baseline imbalance observed post hoc should not be considered an appropriate reason for including a variable as a covariate in the primary analysis.
- Variables measured after randomization and so potentially affected by the treatment should not normally be included as covariates in the primary analysis.
- If a baseline value of a continuous outcome measure is available, then this should usually be included as a covariate. This applies whether the primary outcome variable is defined as the 'raw outcome' or as the 'change from baseline'.
- Only a few covariates should be included in a primary analysis. Although larger data sets may support more covariates than smaller ones, justification for including each of the covariates should be provided. (???)
- In the absence of prior knowledge, a simple functional form (usually either linearity or dichotomising a continuous scale) should be assumed for the relationship between a continuous covariate and the outcome variable. (???)
- The validity of the model assumptions must be checked when assessing the results. This is particularly important for generalized linear or non-linear models where mis-specification could lead to incorrect estimates of the treatment effect. Even under ordinary linear models, some attention should be paid to the possible influence of extreme outlying values.
- Whenever adjusted analyses are presented, results of the treatment effect in subgroups formed by the covariates (appropriately categorised, if relevant) should be presented to enable an assessment of the validity of the model assumptions. (???)
- Sensitivity analyses should be pre-planned and presented to investigate the robustness of the primary results. Discrepancies should be discussed and explained. In the presence of important differences that cannot be logically explained-for example, between the results of adjusted and unadjusted analyses-the inter-

pretation of the trial could be seriously affected.

– The primary model should not include treatment by covariate interactions. If substantial interactions are expected a priori, the trial should be designed to allow separate estimates of the treatment effects in specific subgroups.
– Exploratory analyses may be carried out to improve the understanding of covariates not included in the primary analysis, and to help the sponsor with the ongoing development of the drug.
– A primary analysis, unambiguously pre-specified in the protocol or statistical analysis plan, correctly carried out and interpreted, should support the conclusions which are drawn from the trial. Since there may be a number of alternative valid analyses, results based on pre-specified analyses will carry most credibility."

"In confirmatory trials, a model is pre-specified, and it is necessary to pretend that it is true. In most other statistical applications, the choice of model is data–driven, but it is necessary to pretend that it is not."[9]

See also Siqueira and Taylor[40].

- Choose predictors based on expert opinion
- Impute missing values rather than discarding observations
- Keep all pre–specified predictors in model, regardless of P–value

- Use shrinkage (penalized maximum likelihood estimation) to avoid over–adjustment
- Detailed strategy in REGRESSION MODELING STRATEGIES, Springer, 2001.

10.1 Sites vs. Covariables

- Site effects (main or interaction) are almost always trivial in comparison with patient-specific covariable effects
- It is not sensible to include site in the model when important covariables are omitted
- The most logical and usually the most strong interaction with treatment is not site but is the severity of disease being treated

10.2 Covariable Adjustment vs. Allocation Based on Covariates

As Senn[38] states (see also[39]),

> ”The decision to fit prognostic factors has a far more dramatic effect on the precision of our inferences than the choice of an allocation based on covariates or randomization approach and one of my chief objections to the allocation based on covariates approach is that trialists have tended to use the fact that they have balanced as an excuse for not fitting. This is a grave mistake.” (p. 3748)
>
> ”My view . . . was that the form of analysis envisaged (that is to say, which factors and covariates should be fitted) justified the allocation and *not vice versa*.” (p. 3747)

11 Summary

As Senn [38, p. 3741] said

> ”The point of view is sometimes defended that analyses that ignore covariates are superior because they are simpler. I do not accept this. A value of $\pi = 3$ is a simple one and accurate to one significant figure . . . However very few would seriously maintain that if should generally be adopted by engineers.”

12 Notes

From a posting by Harrell to the Medstats google group on 19Jan09: I think it is most important to decide what it is you want to estimate, and then formulate a model that will accomplish that. Unlike ordinary linear models, which provide unbiased treatment effects if balanced covariates are mistakenly omitted from the model in an RCT, most models (such as the Cox PH model) result in biased treatment effects even when there is perfect balance in covariates, if the covariates have nonzero effects on the outcome. This is another way of talking about residual outcome heterogeneity.

If you want to estimate the effect of variable X on survival time, averaging over males and females in some strange undocumented way, you can get the population averaged effect of X without including sex in the model. Recognize however this is like comparing some of the males with some of the females when estimating the X effect. This is seldom of interest. More likely we want to know the effect of X for males, the effect for females, and if there is no interaction we pool the two to more precisely estimate the effect of X conditional on sex.

Another way to view this is that the PH assumption is more likely to hold when you condition on covariates than when you don't.

No matter what happens though, if PH holds for one case, it cannot hold for the other, e.g., if PH holds after conditioning, it cannot hold when just looking at the marginal effect of X.

References

[1] K. Akazawa, T. Nakamura, and Y. Palesch. Power of logrank test and Cox regression model in clinical trials with heterogeneous samples. *Stat Med*, 16:583–597, 1997. []

[2] D. G. Altman and C. J. Doré. Randomization and baseline comparison in clinical trials. *Lancet*, 335:149–153, 1990. []

[3] P. K. Andersen, J. P. Klein, and M. Zhang. Testing for centre effects in multi-centre survival studies: A monte carlo comparison of fixed and random effects tests. *Stat Med*, 18:1489–1500, 1999. [23]

[4] P. C. Austin, A. Manca, M. Zwarenstein, D. N. Juurlink, and M. B. Stanbrook. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journal. *J Clin Epi*, 63:142–153, 2010. []

[5] C. B. Begg. Significance tests of covariate imbalance in clinical trials. *Controlled Clin Trials*, 11:223–225, 1990. []

[6] R. M. Califf, F. E. Harrell, K. L. Lee, J. S. Rankin, et al. The evolution of medical and surgical therapy for coronary artery disease. *JAMA*, 261:2077–2086, 1989. [15]

[7] Committee for Proprietary Medicinal Products. Points to consider on adjustment for baseline covariates. *Stat Med*, 23:701–709, 2004. [25]

[8] CPMP Working Party. Biostatistical methodology in clinical trials in applications for marketing authorizations for medicinal products. *Stat Med*, 14:1659–1682, 1995. []

[9] D. Edwards. On model pre-specification in confirmatory randomized studies. *Stat Med*, 18:771–785, 1999. [25, 27]

[10] I. Ford and J. Norrie. The role of covariates in estimating treatment effects and risk in long-term clinical trials. *Stat Med*, 21:2899–2908, 2002. []

[11] I. Ford, J. Norrie, and S. Ahmadi. Model inconsistency, illustrated by the Cox proportional hazards model. *Stat Med*, 14:735–746, 1995. []

[12] M. Gail, S. Wieand, and S. Piantadosi. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71:431–444, 1984. []

[13] M. H. Gail. Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In S. H. Moolgavkar and R. L. Prentice, editors, *Modern Statistical Methods in Chronic Disease Epidemiology*, pages 3–18. Wiley, New York, 1986. []

[14] M. H. Gail, W. Y. Tan, and S. Piantadosi. Tests for no treatment effect in randomized clinical trials. *Biometrika*, 75:57–64, 1988. []

[15] F. E. Harrell. *Regression Modeling Strategies, with Applications to Linear Models, Survival Analysis and Logistic Regression*. Springer, New York, 2001. []

[16] F. E. Harrell, K. L. Lee, and D. B. Mark. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*, 15:361–387, 1996. []

[17] W. W. Hauck, S. Anderson, and S. M. Marcus. Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Controlled Clin Trials*, 19:249–256, 1998. [12, 13]

[18] W. W. Hauck, J. M. Neuhaus, J. D. Kalbfleisch, and S. Anderson. A consequence of omitted covariates when estimating odds ratios. *J Clin Epi*, 44:77–81, 1991. []

[19] A. V. Hernández, E. W. Steyerberg, and J. D. F. Habbema. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epi*, 57:454–460, 2004. []

[20] International Conference on Harmonisation. Guideline E-9: Statistical Principles for Clinical Trials, 1998. []

[21] J. P. A. Ioannidis and J. Lau. The impact of high-risk patients on the results of clinical trials. *J Clin Epi*, 50:1089–1098, 1997. [6]

[22] A. Källén. Treatment-by-center interaction: What is the issue? *Drug Info J*, 31:927–936, 1997. []

[23] D. M. Kent and R. Hayward. Limitations of applying summary results of clinical trials to individual patients. *JAMA*, 298:1209–1212, 2007. [20]

[24] W. A. Knaus, F. E. Harrell, C. J. Fisher, D. P. Wagner, S. M. Opan, J. C. Sadoff, E. A. Draper, C. A. Walawander, K. Conboy, and T. H. Grasela. The clinical evaluation of new drugs for sepsis: A prospective study design based on survival analysis. *JAMA*, 270:1233–1241, 1993. []

[25] S. W. Lagakos and D. A. Schoenfeld. Properties of proportional hazards score tests under misspecified regression models. *Biometrics*, 40:1037–1048, 1984. []

[26] J. A. Lewis. Statistical issues in the regulation of medicines. *Stat Med*, 14:127–136, 1995. []

[27] D. B. Macnaughton. Which sums of squares are best in unbalanced analysis of variance? `http://www.matstat.com/ss/`, 1998. []

[28] T. M. Morgan. Omitting covariates from the proportional hazards model. *Biometrics*, 42:993–995, 1986. []

[29] J. M. Neuhaus. Estimation efficiency with omitted covariates in generalized linear models. *J Am Stat Assoc*, 93:1124–1129, 1998. [13]

[30] S. J. Pocock, S. E. Assmann, L. E. Enos, and L. E. Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med*, 21:2917–2930, 2002. []

[31] G. M. Raab, S. Day, and J. Sales. How to select covariates to include in the analysis of a clinical trial. *Controlled Clin Trials*, 21:330–342, 2004. [13]

[32] L. D. Robinson and N. P. Jewell. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review*, 59:227–240, 1991. []

[33] D. J. Sargent and J. S. Hodges. A hierarchical model method for subgroup analysis of time-to-event data in the Cox regression setting. Presented at the Joint Statistical Meetings, Chicago, 1996. [14]

[34] D. A. Schoenfeld. Sample size formulae for the proportional hazards regression model. *Biometrics*, 39:499–503, 1983. []

[35] G. Schwemer. General linear models for multicenter clinical trials. *Controlled Clin Trials*, 21:21–29, 2000. [23]

[36] S. Senn. Testing for baseline balance in clinical trials. *Stat Med*, 13:1715–1726, 1994. []

[37] S. Senn. *Statistical Issues in Drug Development*. Wiley, Chichester, England, 1997. []

[38] S. Senn. Controversies concerning randomization and additivity in clinical trials. *Stat Med*, 23:3729–3753, 2004. [9, 28, 29]

[39] S. Senn, V. V. Anisimov, and V. V. Fedorov. Comparisons of minimization and Atkinson's algorithm. *Stat Med*, 29:721–730, 2010. [28]

[40] A. L. Sigueira and J. M. G. Taylor. Treatment effects in a logistic model involving the Box-Cox transformation. *J Am Stat Assoc*, 94:240–246, 1999. [27]

[41] E. W. Steyerberg, P. M. M. Bossuyt, and K. L. Lee. Clinical trials in acute myocardial infarction: Should we adjust for baseline characteristics? *Am Heart J*, 139:745–751, 2000. Editorial, pp. 761-763. [4]

[42] J. W. Tukey. Tightening the clinical trial. *Controlled Clin Trials*, 14:266–285, 1993. [13]

[43] P. Verweij and H. C. van Houwelingen. Penalized likelihood in Cox regression. *Stat Med*, 13:2427–2436, 1994. []

[44] T. Yamaguchi and Y. Ohashi. Investigating centre effects in a multi-centre clinical trial of superficial bladder cancer. *Stat Med*, 18:1961–1971, 1999. [14, 23]