Heterogeneity of patients in clinical trials

Subgroup analysis and covariate adjustment in cardiovascular and neurosurgical trials

Adrián V. Hernández

Acknowledgements

The work presented in this thesis was conducted at the Department of Public Health, Department of Neurosurgery, and the Department of Cardiology of the Erasmus MC-University Medical Centre Rotterdam, Rotterdam, The Netherlands.

The studies described in this thesis were supported by the Netherlands Organization for Scientific Research (ZON/MW) project No. 908-02-117, and by the National Institutes of Health, USA (Clinical Trial Design and Analysis in TBI Project, Grant 1 R01 NS-042691-001A1).

The author gratefully acknowledges the collaboration with the Coronary Care Unit, Department of Cardiology, Hospital Universitario Gregorio Marañón, Madrid, Spain (Héctor Bueno), and the Department of Hygiene and Epidemiology, University of Ioannina, Ioannina, Greece (John P. A. Ioannidis).

The Department of Public Health of the Erasmus Medical Center, Rotterdam, Centocor B.V. Leiden, and the Nederlandse Hartstichting, Den Haag, The Netherlands, provided financial support for the publication of this thesis.

Cover design and layout: Adrián V. Hernández, Madrid, Spain
adrianhernandezdiaz@gmail.com

Machu Picchu walls were built with stones of different and unique shapes. It is difficult to find two equal stones. Patients recruited for clinical trials are also unique.

Printed by: Optima Grafische Communicatie, Rotterdam, The Netherlands.
ISBN: 90-8559-130-9

Heterogeneity of patients in clinical trials

Subgroup analysis and covariate adjustment in cardiovascular and neurosurgical trials

Heterogeniteit van patiënten in gerandomiseerde klinische studies

Subgroep en covariaat analyse in cardiovasculaire en neurochirurgische studies

Proefschrift

Ter verkrijging van de graad van doctor
aan de Erasmus Universiteit Rotterdam
op gezag van de rector magnificus
Prof. dr. S.W.J. Lamberts
en volgens besluit van het College voor Promoties

De openbare verdediging zal plaatsvinden op
22 Februari 2006 om 15:45 uur

door

Adrián V. Hernández
Geboren te Lima, Perú

Promotiecommissie

Promotor:	Prof. Dr. J. Dik F. Habbema
Overige leden:	Prof. Dr. Theo Stijnen
	Prof. Dr. Peter J. Koudstaal
	Dr. Eric Boersma
Copromotor:	Dr. Ewout W. Steyerberg

Financial support by the Netherlands Heart Foundation and the Department of Public Health, Erasmus MC for the publication of this thesis is gratefully acknowledged.

A la memoria de mi madre Luz y de mi padre Edgardo

A Daniela

A mis hermanos y familiares

A mis amigos de ayer, hoy y siempre

CONTENTS

This thesis is based on the following publications and manuscripts:

Chapter 2.1
Hernández AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. J Clin Epidemiol 2004; 57: 454-60

Chapter 2.2
Hernández AV, Eijkemans MJC, Steyerberg EW. Randomized controlled trials with time-to-event outcomes: how much does pre-specified covariate adjustment increase power?. Ann Epidemiol 2006; 16: 41-8

Chapter 3.1
Hernández AV, Steyerberg EW, Habbema JDF. Inappropriate Use of Baseline Characteristics in Clinical Trials: Assessment of High Impact Medical Journals. Submitted

Chapter 3.2
Hernández AV, Boersma E, Murray GD, Habbema JDF, Steyerberg EW. Subgroup analysis in therapeutic cardiovascular clinical trials: Are most of them misleading?. Am Heart J 2006 (in press)

Chapter 3.3
Hernández AV, Steyerberg EW, Taylor GS, Marmarou A, Habbema JDF, Maas AIR. Subgroup analysis and covariate adjustment in randomized clinical trials of traumatic brain injury: A systematic review. Neurosurgery 2005; 57: 1244-53

Chapter 4.1
Hernández AV, Steyerberg EW, Mushkudiani N, Murray GD, Taylor GS, Butcher I, Marmarou A, Choi SC, Lu J, Habbema JDF, Maas AIR. Adjustment for strong predictors of unfavorable outcome reduces sample size requirements by over 25% in traumatic brain injury trials. Submitted

Chapter 4.2
Hernández AV, Westerhout CM, Steyerberg EW, Ioannidis JPA, Bueno H, White H, Theroux P, Moliterno DJ, Armstrong PW, Califf RM, Wallentin LC, Simoons ML, Boersma E. Effects of platelet glycoprotein IIb/IIIa receptor blockers in non-ST-segment elevation acute coronary syndromes: Benefit and harm in different age subgroups. Submitted

Chapter 4.3

Westerhout CM, Hernández AV, Steyerberg EW, Bueno H, White H, Theroux P, Moliterno DJ, Armstrong PW, Califf RM, Wallentin LC, Simoons ML, Boersma E. Predictors of stroke at 30 days in patients with non-ST-segment elevation acute coronary syndromes. Submitted

General Introduction

1

1.1 Heterogeneity of patients in clinical trials

Randomized clinical trials (RCTs) are essential to evaluate the usefulness of treatments and interventions [1]. Trials inform and influence clinical practice [2]. Clinicians rely increasingly on RCTs to distinguish between worthwhile, useless and harmful interventions.

RCTs are often performed in patients who are heterogeneous with respect to prognosis. Heterogeneity occurs in many clinical fields, such as traumatic brain injury (TBI), acute coronary syndromes (ACS), stroke, and cancer. Even with strict inclusion criteria, patient prognosis can vary according to baseline characteristics (e.g. age, gender, disease severity). For instance, the 6-month mortality for TBI patients older than 65 years is around 72%, in contrast to 21% for patients younger than 35 years [3]. Likewise, patients with unstable angina/non-ST-segment elevation acute coronary syndromes and 7 strong prognostic variables have a nuch higher risk to develop an adverse outcome than those without prognostic variables (41% vs. 5%) [4].

1.2 Problems related to heterogeneity of patients

Heterogeneity may lead to imbalance of randomized groups with respect to prognosis. This imbalance is due to pure chance when a proper randomization procedure was followed [5]. Stratified randomization reduces the chances of imbalance between treatment groups for known factors [6]. This is a two-stage procedure in which patients who enter a clinical trial are first grouped into strata according to predictors. Within each stratum, patients are then assigned to a treatment according to separate randomization schedules. For example, patients over 65 years may be randomized separately from patients under 65.

Heterogeneity may dilute the beneficial effect of treatments in some subgroups of patients satisfying the inclusion criteria. For instance, TBI patients with good prognosis (<20% of 6-month unfavorable outcome) or bad prognosis (>80% of 6-month unfavorable outcome) may not receive benefit from a treatment, because they are too healthy or too sick. It can be hypothesized that patients with an intermediate prognosis (e.g. between 20% and 80%) may benefit most from a treatment [7.] Such a prognostic category can be defined by a combination of baseline characteristics. Clinical staging in cancer is a way to combine patients characteristics, and it is frequently used in oncology clinical trials. Moreover, a treatment may be more effective in subgroups of patients with some characteristics, especially those related to a mechanism of action of the treatment. For example, it could be that TBI patients with mass

lesions may benefit more from certain treatments than others [7].

1.3 Possibilities of analysis in trials with heterogeneous patients

The heterogeneity of patients offers some possibilities to deal with the cited problems. Two methods are often used in the analysis phase of a trial: covariate adjustment and subgroup analysis. Covariate adjustment leads to adjusted estimates of the treatment effects, in contrast to unadjusted estimates. An unadjusted estimate of the treatment effect may be interpreted as relating to an "average" patient with a certain disease. Adjusting for one predictor (e.g. age) results in an estimate for "a patient with a certain age". Adjusting for all known predictive baseline characteristics results in a treatment effect in patients with "a certain risk profile" [8]. Covariate adjustment also corrects for imbalance, and increases the statistical power to detect a significant treatment effect [8, 9]. Thus, it potentially reduces the sample size requirements. For instance, adjustment of the treatment effect for 17 predictors of 30-day mortality of patients with acute myocardial infarction (MI) enrolled in the GUSTO-I trial reduced the required sample size by 15% [8]. In a simulation study using a TBI registry, adjustment for age and Glasgow Coma Scale (GCS) motor score reduced the sample size requirements by 30% [10]. However, the quantification of the increase in power and the potential reductions in sample size requirements in empirical RCTs has been insufficiently studied.

Subgroup analysis assesses differences in treatment effect between different subpopulations of patients [11-13]. A subgroup is a group of patients with a common set of baseline characteristics. Journals have devoted a large number of pages describing treatment effects in particular groups of patients [11]. However, subgroup analyses have been generally misused and overinterpreted in the last two decades [12], and it is not known if their use has improved in recent years. Trials of a size adequate to detect an overall effect cannot be expected to provide reliable effect estimates within smaller groups [11]. This problem can be tackled by a meta-analysis of RCTs, which may substantially increase the power to detect differential treatment effects across subgroups of patients [13].

1.4 Heterogeneity in traumatic brain injury and acute coronary syndrome trials

Heterogeneity is a major problem in the field of TBI, where the risks of a poor 6-month outcome vary widely between patients (e.g. risks varying from <10% to >90%). All recent RCTs in TBI have shown negative results, i.e. that the

new treatment was not significantly better than placebo [14]. Heterogeneity was insufficiently considered in the design and analysis of the treatment effect in these studies. Adjustment for important predictors of the outcome may decrease the sample size requirements in TBI trials. The reduction is sample size will be larger where the inclusion criteria are less strict, and hence the population is more heterogeneous.

Patients with non-ST-segment elevation acute coronary syndromes essentially differ in their short and long term outcome, and prognostic categories are commonly used to allocate a less or more aggressive treatment [15]. Platelet glycoprotein IIb/IIIa (GP IIb/IIIa) receptor blockers decrease the risk of death or MI at 30 days in patients with non-ST-segment elevation acute coronary syndromes not routinely scheduled for early revascularization [16]. Elderly patients (>80 years) may be undertreated with these drugs because they were underrepresented in previous RCTs, and because the drugs can produce more harmful effects than in younger patients. There has been no demonstration that these drugs are equally efficacious in old and young patients [17]. A meta-analysis of large international RCTs may help to define the benefits and risks of GP IIb/IIIa receptor blockers in elderly patients.

Stroke is an uncommon and serious event in patients with non-ST-segment elevation acute coronary syndromes [18]. However, few papers have been performed to establish which patients are at higher risk to develop this serious complication. This group of patients can potentially benefit from preventive interventions after the ACS, such as statins [19]. An evaluation of predictors in a population with a larger number of stroke events is required to define predictors accurately.

1.5 Scope of the Thesis

This thesis describes both methodological aspects and clinical applications of covariate adjustment and subgroup analysis in RCTs. The following research questions are addressed in this thesis:

1. Which are the pros and the cons of adjustment of the treatment effect in RCTs for baseline covariates?

2. How well are covariate adjustment and subgroup analysis used, reported and interpreted in current internal medicine, oncology, cardiology, and neurosurgery trials?

3. How much reduction in sample size can be obtained from adjustment for important predictors of unfavorable outcome in TBI trials?

4. Are the effects of GP IIb/IIIa receptor blockers similar in the young and the elderly in a meta-analysis of non-ST-segment elevation acute coronary syndrome trials?

5. Which subgroups of patients are at higher risk to develop a stroke after a non-ST-segment elevation acute coronary syndrome?

These issues have been insufficiently studied and quantified in theoretical and empirical data. The first part of this thesis considers simulation studies to quantify the gain in power and the reduction in sample size requirements in trials with dichotomous outcomes, with the use of the logistic regression model (Chapter 2) and the Cox proportional hazards models (Chapter 3). The reduction in sample size indicates that adjusted analyses might give the same power as an unadjusted analysis but with smaller sample size.

The second part describes the current practices of reporting and interpretation of covariate adjustment and subgroup analysis in 84 trials from high Impact Factor journals in General Internal Medicine, Cardiology and Oncology (Chapter 4), in 63 therapeutic cardiovascular clinical trials (Chapter 5), and in 18 trials of moderate and severe traumatic brain injury with more than 100 patients (Chapter 6).

The third part describes clinical applications of relevant methods in heterogeneous trial populations: an analysis of the reduction in sample size requirements with covariate adjustment for strong predictors in 8 trials of moderate-severe traumatic brain injury (n=6298) (Chapter 7), a subgroup analysis of the effects of the platelet glycoprotein IIb/IIIa receptors blockers by age in a meta-analysis of 6 trials of patients with non-ST-segment elevation acute coronary syndromes (n=31402) (Chapter 8), and an analysis to find predictors of all-cause stroke in heterogeneous patients with non-ST-segment elevation acute coronary syndromes (Chapter 9).

This thesis ends with a discussion of the theorethical and practical results, and gives recommendations for appropriate analysis of RCTs with heterogeneous populations.

References

1. DeMets DL. Clinical trials in the new millennium. Stat Med 2002; 21: 2779-87.
2. Yusuf S. Randomized controlled trials in cardiovascular medicine: past and achievements, future challenges. BMJ 1999; 319: 564-68.
3. Hukkelhoven CW, Steyerberg EW, Rampen AJ, et al. Patients age and outcome following severe traumatic brain injury: an analysis of 5600 patients. J Neurosurg 2003; 99: 666-73.
4. Antman EM, Cohen M, Bernink PJLM, et al. The TIMI risk score for unstable angina/non-ST-elevation MI. A method of prognostication and therapeutic decision making. JAMA 2000; 284: 835-42.
5. Roberts C, Torgerson DJ. Baseline imbalance in randomised trials. BMJ 1999; 319: 185.
6. Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. J Clin Epidemiol 1999; 52: 19-26.
7. Machado SG, Murray GD, Teasdale GM. Evaluation of designs for clinical trials of neuroprotective agents in head injury. J Neurotrauma 1999; 16: 1131-8.
8. Steyerberg EW, Bossuyt PM, Lee KL. Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics?. Am Heart J 2000; 139: 745-51.
9. Hauck WW, Anderson S, Marcus SM: Should we adjust for covariates in non-linear regression analyses of randomized trials?. Control Clin Trials 1998; 19: 248-56.
10. Choi SC. Sample size in clinical trials with dichotomous endpoints: use of covariables. J Biopharm Stat 1998; 8: 367-75.
11. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991; 266: 93-8.
12. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other mis(uses) of baseline data in clinical trials. Lancet 2000; 355: 1064-9.
13. Brookes SJ, Whitley E, Egger M, Davey Smith G, Mulheram PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol 2004; 57: 229-36.
14. Maas AIR, Steyerberg EW, Murray GD, et al. Why have recent trials of neuroprotective agents in head injury failed to show convincing efficacy? A pragmatic analysis and theoretical considerations. Neurosurgery 1999; 44: 1286-98.
15. Braunwald E. Application of current guidelines to the management of unstable angina and non-ST-segment elevation myocardial infarction. Circulation 2003; 108(suppl 1): III28-III33.
16. Schulman SP. Antiplatelet therapy in non-ST-segment elevation acute coronary syndromes. JAMA 2004; 292: 1875-82.
17. Alter DA, Manuel DG, Gunraj N, Anderson G, Naylor CD, Laupacis A. Age, risk-benefit trade-offs, and the projected effects of evidence-based therapies. Am J Med 2004; 116: 540-5.

18. Sabatine MS, Januzzi JL, Snappin S, Theroux P, Jang I-K. A risk score system for predicting adverse outcomes and magnitude of benefit with glycoprotein IIb/IIIa inhibitor therapy in patients with unstable angina pectoris. Am J Cardiol 2001; 88: 488-92.
19. West MJ, White HD, Simes RJ, et al. Risk factors for non-hemorrhagic stroke in patients with coronary artery disease and the effect of lipid-modifying therapy with pravastatin. J Hypertens 2002; 20: 2513-7.

Simulation Studies

2

Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements

2.1

ABSTRACT

Background and Objective: Randomized controlled trials (RCTs) with dichotomous outcomes may be analyzed with or without adjustment for baseline characteristics (covariates). We studied type I error, power and potential reduction in sample size with several covariate adjustment strategies.

Methods: Logistic regression analysis was applied to simulated data sets (n=360) with different treatment effects, covariate effects, outcome incidences and covariate prevalences. Treatment effects were estimated with or without adjustment for a single dichotomous covariate. Strategies included always adjusting for the covariate ('pre-specified'), or only when the covariate was predictive or imbalanced.

Results: We found that the type I error was generally at the nominal level. The power was highest with pre-specified adjustment. The potential reduction in sample size was higher with stronger covariate effects (from 3% to 46%, at 50% outcome incidence and covariate prevalence) and independent of the treatment effect. At lower outcome incidences and/or covariate prevalences, the reduction was lower.

Conclusion: We conclude that adjustment for a predictive baseline characteristic may lead to a potentially important increase in power of analyses of treatment effect. Adjusted analysis should hence be considered more often for RCTs with dichotomous outcomes.

Introduction

Randomized controlled trials (RCTs) have emerged as the principal research tool to inform and influence clinical practice [1, 2] . Clinicians rely increasingly on efficient, well designed RCTs to distinguish between worthwhile, useless or harmful interventions [2]. Efforts have been made to improve quality in reporting RCTs [3], but the way to analyze RCTs properly is still under discussion [4-7].

In particular, the treatment effect in a RCT can be analyzed and shown either as an average effect, as an adjusted effect, or both [8]. Adjusted effect estimates attempt to take the heterogeneity of patients in RCTs into account. The heterogeneity of patients is related to their prognostic baseline characteristics [4], which may be used to obtain adjusted treatment effects. Common methods of adjustment are baseline imbalance adjustment [5, 9], subgroup analysis [5, 10-12], stratified randomization plus adjustment [13], and covariate adjustment (post-stratification) [14-18].

Covariate adjustment provides more individualized effect estimates, especially in non-linear models such as logistic regression and Cox proportional hazards regression [8, 13-20]. Further, adjusted effect estimates take into account chance differences in baseline characteristics between treatment arms [14, 15] and improve the power, i.e. the ability to identify treatment effects when they really exist [15-22].

The use of covariate adjustment in the current literature is not consistent [5, 11], probably because the strategies have not been fully developed and tested [23, 24]. A key aspect of the adjustment strategies is the way of selection of the covariate to be adjusted for [25]. Moreover, the effects on power and type I error after adjustment have not been studied thoroughly.

We used various strategies for covariate adjustment in simulated logistic regression models with one dichotomous covariate. Our aim was to identify the pros and the cons of each covariate adjustment strategy, with a focus on changes in statistical power. We expressed any increase in statistical power as the decrease in sample size that gives the same power as an unadjusted analysis.

Methods

Treatment effects and adjustment strategies

Logistic regression models were used to analyze the effects of treatment on a

dichotomous outcome (e.g. 30-day mortality). A dichotomous baseline characteristic was entered as covariate to achieve adjustment of the treatment effect. The formula is: log odds (outcome) = $\beta 0 + \beta 1$*Treatment + $\beta 2$*Covariate. The logistic regression coefficients and their standard errors (SE) were estimated with standard maximum likelihood procedures. Statistical significance was based on the Wald statistic (coefficient/SE), with two-sided p-values < 5% considered significant.

We used three approaches for covariate adjustment. In practice, they should be clearly described in the protocol. 'Adjustment' refers to a pre-specified adjusted analysis. Pre-specified analysis uses a known predictive covariate, e.g. derived from previous research. This leads to a stratified estimate of the treatment effect, which is also corrected for imbalance [14]. Further, the predictive effect of a baseline characteristic can be used for adjustment if statistically significant using chi-square test (p< 5%) and this is referred to as 'Predictor p<5%'. This chi square is not adjusted for treatment. The imbalance according to a baseline characteristic may also be used for adjustment if statistically significant according to a chi-square test (p<5%) and it is referred to as 'Imbalance p<5%'. Moreover, 'No adjustment' refers to an unadjusted analysis of the treatment effect. We report the results with the default significance level of 5%. We also explored the effects of applying a more liberal criterion for covariate selection (p<20%).

Data sets and simulations

The various approaches to estimation of the treatment effect were applied in different data sets. First, we simulated RCTs of 360 patients, 180 randomized to treatment A and 180 to treatment B, based on a previously discussed hypothetical example (Table 1), using mortality as outcome (number of events equal to 180 on average) [14]. The distributions of gender and outcome were generated by sampling with replacement from the data set underlying the same table. Here, the prognostic effect of gender was very strong, with a 16% mortality rate in men compared to 84% in women (Odds Ratio [OR] = 30). The treatment effect for treatment B showed a mortality OR of 1.43 in comparison with treatment A in the unadjusted analysis, which used the total table and ignored heterogeneity (p=0.09). In contrast, the mortality OR was 2.0 in the strata formed by gender, meaning that treatment A reduced the odds of death in men and women by half (p=0.02). This example illustrates that the unadjusted OR may substantially differ from the adjusted OR.

We analyzed more hypothetical data sets using combinations of different unadjusted covariate effects (No covariate effect [OR=1], moderate covariate

	Total: OR=1.43			Males: OR=2.0		Females: OR=2.0	
Treatment	Dead	Survive	Total	Dead	Survive	Dead	Survive
A	82	98	180	10	80	72	18
B	98	82	180	18	72	80	10
% of death	(50%)			(16%)		(84%)	

Table 1. Hypothetical example of stratification in a randomized clinical trial, where treatment A is compared to treatment B, while gender is balanced.

effect [OR = 2], strong covariate effect [OR = 5] and very strong covariate effect [OR = 30]). Unadjusted treatment effects were varied from no treatment effect [OR=1], weak treatment effect [OR = 1.4] to mild treatment effect [OR = 1.7]. These effects are within the range of 1.1 to 1.9 that we observed in a survey of 13 large cardiovascular RCTs reported in The Lancet, The New England Journal of Medicine and JAMA during 2001 and 2002.

We simulated RCTs of 360 patients (180 to treatment A and 180 to treatment B, on average) and their distributions were generated by sampling with replacement from each hypothetical data set. The number of outcomes considered in the main analysis was 180 on average (50% of total number of patients). Because the outcome incidence and the covariate prevalence may be lower in RCTs, we also considered situations with a lower outcome incidence and a lower covariate prevalence (25%, 12.5% and 6.25% of the total number of patients in both cases). Simulations were performed with 20,000 repetitions when a treatment effect was truly present and 100,000 when no treatment effect was present (i.e. under the null hypothesis).

Table 2 shows the 12 logistic regression models considered. For example, the model with weak treatment effect and very strong covariate effect was made from a data set such as shown in Table 1. Unadjusted covariate and treatment ORs are the univariate ORs (either covariate or treatment introduced in the logistic model). Adjusted covariate and treatment ORs are the multivariate ORs (both covariate and treatment introduced in the logistic model). The covariate and treatment ORs are higher in adjusted analyses, specially when the covariate effects are stronger. We repeated the analyses using double sample size (i.e. 720 patients, with 360 in each treatment arm). When the outcome incidences and

Model		OR			
		Covariate		Treatment	
		Unadjusted	Adjusted	Unadjusted	Adjusted
No covariate effect:					
Strong TE:	- 0.268+ 0*Covariate+ 0.54*Tx	1.0	1.0	1.7	1.7
Moderate TE:	- 0.178+ 0*Covariate+ 0.36*Tx	1.0	1.0	1.4	1.4
No TE:	0+ 0*Covariate+ 0*Tx	1.0	1.0	1.0	1.0
Moderate covariate effect:					
Strong TE:	- 0.64+ 0.73*Covariate+ 0.55*Tx	2.0	2.1	1.7	1.7
Moderate TE:	- 0.55+ 0.72*Covariate+ 0.36*Tx	2.0	2.1	1.4	1.4
No TE:	- 0.36+ 0.71*Covariate+ 0*Tx	2.0	2.0	1.0	1.0
Strong covariate effect:					
Strong TE:	- 1.12+ 1.63*Covariate+ 0.63*Tx	4.9	5.1	1.7	1.9
Moderate TE:	- 1.01+ 1.61*Covariate+ 0.42*Tx	4.9	5.0	1.4	1.5
No TE:	- 0.79+ 1.59*Covariate+ 0*Tx	4.9	4.9	1.0	1.0
Very strong covariate effect:					
Strong TE:	- 2.32+ 3.58*Covariate+ 1.08*Tx	29.4	35.9	1.7	2.9
Moderate TE:	- 2.08+ 3.47*Covariate+ 0.69*Tx	29.4	32.1	1.4	2.0
No TE:	- 1.69+ 3.38*Covariate+ 0*Tx	29.4	29.4	1.0	1.0

OR: Odds ratio; TE: Treatment effect; Tx: treatment

Table 2. Characteristics of the models used for adjusted and unadjusted analyses of the treatment effect.

covariate prevalences were small (12.5% and 6.25%), we used bigger sample sizes (n=1200) to facilitate the construction of adequate data sets.

Evaluation

We studied the actual type I error (α) when there was no treatment effect (OR=1),

and power (1-β, where β indicates type II error) when there truly was a treatment effect (OR>1), in each of the strategies of covariate adjustment. The 95% CI for the type I error of 5% ranged from 4.87% and 5.14% with 100 000 simulations. The formula used to calculate the type I error and the power for each strategy was: 100* (number of simulations with statistically significant treatment effect / total number of simulations), where the statistical significance was established according to the Wald statistic. We calculated the reduction of sample size to express the gain in power with each of the adjustment strategies. The formula used was: 100 - 100*[(mean of Z score of unadjusted strategy)/(mean of Z score of the adjusted strategy)]² (see appendix), where Z score is equal to the Wald statistic of the treatment effect coefficient [6]. We used S-plus software (version 2000, Insightful Inc, Seattle, WA) for all calculations.

Results

Table 3 shows the main results of our simulations. When there was no treatment effect (OR=1), SEs from adjusted analysis were larger in direct relation to the strength of the covariate effect. The type I error was rather similar for most adustment strategies and for all covariate effects. The results were mainly slightly below 5%, especially when the covariate effect was strong and the significant imbalance strategy was used (type I error 3.8%). When there were no covariate effects, covariate adjustment strategies did not markedly change the treatment effect coefficients and SEs. The power was slightly reduced and the required sample size was slightly increased (<0.1%). This indicates that very limited damage was done when a noise covariate was included in the adjusted analysis.

When there were treatment effects and covariate effects (OR>1), the pre-specified covariate adjustment strategy yielded on average a more extreme treatment effect coefficient compared to no adjustment. Consequently, power increased, as expected. Adjustment for one statistically significant predictor generated similar results to pre-specified adjustment, since the latter was nearly always statistically significant. Statistically significant imbalance adjustment generated on average only a slightly higher coefficient than the unadjusted coefficients, leading to a small increase in power.

A potential reduction in sample size was evident when covariate adjustment was pre-specified or based on the predictive significance of the covariate. The reduction in sample size was largest when the covariate effect was strongest. The range of reductions in sample size was from 3% to 46%, corresponding to unadjusted covariate OR from 2 to 30. When the significant imbalance strategy was used, the reduction in sample size was less than 3%. Remarkably, we observed

	No treatment effect		Unadjusted treatment effect OR = 1.4			Unadjusted Treatment effect OR = 1.7		
	Coef ± SE	Type I error (%)	Coef ± SE	1-type II error (%)	Reduction in sample size(%)	Coef ± SE	1-type II error (%)	Reduction in sample size(%)
No covariate effect (OR= 1)								
No adjustment	0.00 ± 0.21	5.0	0.36 ± 0.21	40.0	----	0.54 ± 0.21	72.3	----
Adjustment	0.00 ± 0.21	4.4	0.36 ± 0.21	39.7	-0.04	0.54 ± 0.21	72.1	-0.05
Predictor effect <5%	0.00 ± 0.21	4.6	0.36 ± 0.21	39.9	-0.08	0.54 ± 0.21	72.3	-0.04
Imbalance p<5%	0.00 ± 0.21	4.6	0.36 ± 0.21	39.9	-0.07	0.54 ± 0.21	72.3	-0.08
Moderate unadjusted covariate effect (OR= 2)								
No adjustment	0.00 ± 0.21	5.2	0.36 ± 0.21	39.5	----	0.54 ± 0.21	72	----
Adjustment	0.00 ± 0.22	5.0	0.37 ± 0.22	40.1	3.3	0.56 ± 0.22	73.1	2.9
Predictor effect <5%	0.00 ± 0.22	5.0	0.37 ± 0.22	40.1	3.1	0.56 ± 0.22	73.0	2.7
Imbalance p<5%	0.00 ± 0.21	5.1	0.36 ± 0.21	39.5	0.2	0.54 ± 0.21	72.2	0.1
Strong unadjusted covariate effect (OR= 5)								
No adjustment	0.00 ± 0.21	5.2	0.36 ± 0.21	40.3	----	0.54 ± 0.21	73	----
Adjustment	0.00 ± 0.23	5.1	0.43 ± 0.23	44.9	13.8	0.64 ± 0.23	78.8	13.5
Predictor effect <5%	0.00 ± 0.23	5.1	0.43 ± 0.23	44.9	13.8	0.64 ± 0.23	78.8	13.5
Imbalance p<5%	0.00 ± 0.21	4.9	0.37 ± 0.21	40.3	0.8	0.55 ± 0.21	73.4	0.8
Very strong unadjusted covariate effect (OR= 30)								
No adjustment	0.00 ± 0.21	5.1	0.36 ± 0.21	39.7	----	0.54 ± 0.21	71.9	----
Adjustment	0.00 ± 0.30	4.6	0.71 ± 0.31	63.9	45.5	1.10 ± 0.34	94	43.6
Predictor effect <5%	0.00 ± 0.30	4.6	0.71 ± 0.31	63.9	45.5	1.10 ± 0.34	94	43.6
Imbalance p<5%	0.00 ± 0.21	3.8	0.38 ± 0.22	40.4	3.4	0.57 ± 0.25	73.9	3.3

Table 3. Results of simulations of various covariate adjustment procedures.

the same reduction in sample size when different treatment effects were used, given a certain covariate effect. For example, when pre-specified adjustment was used, the reduction in sample size was 13.8% if the treatment effect was weak (OR=1.4) and 13.6% if it was mild (OR=1.7).

When either the outcome incidence or the covariate prevalence or both were lower, the reduction in sample size was also lower (Table 4). For example, with a covariate OR of 5, an outcome incidence of 25% and a covariate prevalence of 50%, the average reduction in sample size was 10.4% instead of 13.7%. The reduction in sample size was the same when the outcome incidence was 50% with a covariate prevalence of 25% and when the outcome incidence was 25% with covariate prevalence of 25%. Smaller outcome incidences and/or covariate prevalences yielded even lower reductions in sample (e.g. 3.0% when outcome incidence 6.25% and covariate prevalence 6.25%). The reductions in sample size were similar when the sample size was increased, and when we considered $p<20\%$ instead of $p<5\%$ as covariate selection limit.

Discussion

Covariate adjustment increased the power of statistical analyses of a treatment effect in the context of a randomized trial, without inflation of type I error. The increase in power was translated into moderate potential reductions in sample size, indicating that adjusted analyses might give the same power as an unadjusted analysis but with a smaller sample size. We found that pre-specified covariate and significant predictor covariate adjustment strategies were the most statistically efficient, with reductions in sample size between 3% and 46%, depending on the covariate strength. The potential reduction in sample size was independent of the magnitude of the treatment effect or sample size, which makes it an attractive summary measure to express the benefit of adjustment strategies. There was very limited damage when adjustment was performed if there was no true covariate effect.

Adjusted and unadjusted estimates in RCTs with homogeneous groups show on average similar results, as confirmed by our simulations with no covariate effects. However, several problems arise when dealing with heterogeneous populations. For example, adequate randomization sometimes produces unbalanced groups (statistically or non-statistically different at the 5% level) with respect to the baseline characteristics [4]. Moreover, the effect of non-statistically significant imbalance between groups is not negligible, especially if the prognostic effect is strong [4, 9, 14, 17]. Further, the overall clinical trial result is not directly applicable to individuals [7, 8].

Covariate adjustment may overcome some of the problems of heterogeneous populations [4, 5, 8, 13-17, 24]. The reported potential benefits of adjustment for covariates include the removal of confounding due to imbalance [4, 8, 14, 19], the acquisition of a more subject-specific estimate instead of a population-averaged estimate [8, 14], and the gain in statistical power [8, 14, 15, 17, 19, 24]. The gain in statistical power in logistic regression is explained by a larger increase in expected size of the adjusted treatment effect than the increase in the SE [16, 19, 24].

We found that adjustment for a strong covariate (either pre-specified or tested) led to a more extreme estimate of the treatment effect, as expected [5]. The relative difference between unadjusted and adjusted estimates on the logistic scale was approximately constant for a given covariate effect and independent of the treatment effect. We also confirmed that the variability (SE) of the adjusted estimates was larger in direct relation to the strength of the covariate [14-18]; the opposite of what happens in linear regression [19, 25]. We obtained a conservative type I error when the covariate effect was strong and the imbalance strategy was used. This conservatism is explained by the fact that imbalance constrains the outcome variability between treatment groups [17].

The power to detect a true treatment effect increased with the pre-specified and the statistically significant predictor strategies, which was translated into reductions in sample size up to 46%. In practical situations, the covariate effect may be either moderate or strong (OR from 2 to 5, corresponding to Spearman's correlations (ρ) between 0.17 and 0.38 [9, 17]), and the treatment effect may be weak (OR of 1.4). Therefore, the potential reductions in sample size would be moderate, maximally between 3% and 14% with a 50% outcome incidence and 50% covariate prevalence.

Our results agree with Pocock et al. [17] who found a reduction in sample size of 10% with adjustment for a covariate with ρ= 0.3, in the context of RCTs with continuous outcomes. When the covariate was very strong (OR: 30; ρ= 0.69) the reduction of sample size was highest (46%), in agreement with a previously found 50% reduction at ρ= 0.7. These similarities in the reduction in sample size are attributed to the similarity of effects of the covariates in the linear and logistic regression models with 50% outcome incidence.

When we chose a covariate with the significant imbalance strategy the gain in power was not observed. Therefore, this procedure is not advised, in concordance with others [4, 5, 8]. The results of the combination of strategies ('predictor and imbalance' and 'predictor or imbalance') were similar to the significant imbalance

Situations	Outcome incidence(%)	Covariate prevalence(%)	Reduction in sample size(%)
Decreasing event incidence			
	50	50	13.7
	25	50	10.4
	12.5	50	5.5
	6.25	50	2.9
Decreasing covariate prevalence			
	50	25	10.4
	50	12.5	5.5
Decreasing event incidence and covariate prevalence			
	25	25	10.5
	12.5	12.5	5.7
	6.25	6.25	3.0
Decreasing event incidence and increasing covariate prevalence			
	12.5	75	2.8
	12.5	87.5	1.6

Table 4. Reduction in sample size with pre-specified covariate adjustment, considering different outcome incidences and covariate prevalences (Unadjusted covariate OR = 5).

and significant predictor strategies, respectively (results not shown).

In contrast to the scenario with 50% outcome incidence and 50% covariate prevalence, we found that the potential reduction in sample size was lower when the event incidence or the covariate prevalence were lower than 50%. For example, this situation is commonly seen in published cardiovascular RCTs, where the outcome incidences range typically from 5% to 25% [26-28]. Remarkably, the reduction in sample size was similar given a value of either the event incidence or the covariate prevalence or both. Moreover, when the event incidence was lower and the covariate prevalence was higher than 50%, the reduction in sample size was the lowest (1.6 % in the example given in table 4).

The unadjusted strategy yields an average treatment effect, without any consideration of heterogeneity in prognosis among patients. Covariate adjustment and subgroup analyses both consider heterogeneity and attempt to provide more individualized estimates of the treatment effect. They are however substantially

different. Covariate adjustment obtains a single more individualized treatment effect estimate, which is assumed to be applicable to all patients [15, 17]. Subgroup analyses provide multiple treatment effect estimates, assuming that treatment effects differ between particular groups of patients [12]. Covariate adjustment commonly is achieved with regression models [14-18] while adequately performed subgroup analyses apply tests of interaction [10, 12, 17]. Covariate adjustment is sometimes used in the primary analysis of RCTs. Remarkably, unadjusted analyses receive generally more emphasis. This may be caused by a suspicion of data-driven adjustment, or because adjustment is more difficult for readers to understand [5]. Subgroup analyses are often performed [17], but they rarely have enough power to detect differential treatment effects. Tests of interaction are underused and subgroup analyses results are commonly over-interpreted [12].

We advise to perform adjusted analyses because of several advantages, including a more individual-oriented treatment effect corrected for imbalance, a gain in power and, as a consequence, a potential reduction in sample size. Even when the event incidence and covariate prevalence are low, there is a potential reduction in sample size. If a strong predictive covariate is known before the analysis of the RCT, for example, after a literature review [4, 8, 9, 14, 24], the best strategy of covariate adjustment is to pre-specify adjustment. If a number of covariates are potentially predictive, a second best option is to use the 'significant predictor' strategy. Here, bias may be introduced by the selection procedure, particularly in the context of small sample sizes [29]. This theoretical objection was not found important in our simulations. Moreover, Edwards [30] pointed out that if the model is chosen with blinding of the treatment indicator variable, the type I error may be controlled.

We did not present calculations for covariate adjusted sample size in RCT design. To quantify any anticipated sample size gains with covariate adjustment, we would need to specify covariate effects and covariate distributions [31]. In practice, the study would then have to meet these assumptions to achieve the calculated power. Therefore, we advise to perform an unadjusted sample size calculation, which needs less assumptions.

Our work has several limitations. We presented results from analyses of simulated models, using hypothetical treatment effects and covariate effects, without real patient data. We did not study the effect of covariate adjustment using continuous covariates. We used the Wald test for statistical significance, although the Likelihood Ratio test may yield better results. We chose only one covariate to simplify the simulation process and to clarify the presentation. When more prognostic covariates are pre-specified for adjusted analysis, the same conclusions are expected on treatment effect estimation, power and potential reduction

in sample size [8, 17, 18]. However, when many potential predictors are available, a complication may be which variables to choose. A selection procedure, e.g. stepwise backward elimination, might be considered to choose a limited number of covariates, but the effects of such a procedure on effect estimation and power require further study.

In conclusion, adjustment for a predictive baseline characteristic increased the power of statistical analyses of a treatment effect, without inflation of Type I error and with potentially important reductions in sample size. The covariate ideally has to be pre-specified in the RCT protocol. Alternatively, testing for a statistically significant predictive effect is suggested. Adjustment for important predictors should be considered more often in the analysis of RCTs with dichotomous outcomes.

Appendix

Deduction of the Reduction of Sample Size formula

In the context of a RCT, we defined nu and na as the unadjusted and adjusted sample sizes respectively, Z_α is Z value at $\alpha = 0.05$, $Z_{\beta u}$ is the Z value of the unadjusted power 1-βu, $Z_{\beta a}$ is the Z value of the adjusted power 1-βa, Z_u is the mean standardized Z score of the estimated unadjusted coefficient defined as coefficient /SE for each simulation, and, Z_a is the mean standardized Z score of the estimated adjusted coefficient defined as coefficient /SE for each simulation.

We expressed any increase in statistical power of the adjusted analysis as the decrease in the sample size that gives the same power of an unadjusted analysis (i.e. $Z_{\beta u} = Z_{\beta a}$). This assumption gives approximately:

$$n_a / n_u = (Z_u / Z_a)^2 \quad (1)$$

The potential reduction in sample size (RSS) between unadjusted and adjusted strategies is expressed in percentage as:

$$[(n_u - n_a)/ n_u]*100 \text{ , or}$$
$$100 - [100*(n_a /n_u)] \quad (2)$$

Replacing (1) in (2):

$$RSS = 100 - 100* (Z_u / Z_a)^2 \quad (3)$$

References

1. De Mets DL, Califf RM. Lessons learned from recent cardiovascular clinical trials: Part I. Circulation 2002; 106: 746-751.
2. Yusuf S. Randomised controlled trials in cardiovascular medicine: past achievements, future challenges. BMJ 1999; 319: 564-568.
3. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001; 357: 1191-1194.
4. Roberts C, Torgerson DJ. Baseline imbalance in randomised trials. BMJ 1999; 319: 185.
5. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355:1064-1069.
6. Piantadosi S. Clinical Trials: A Methodologic Perspective. New York: John Wiley & Sons Inc; 1997.
7. Rothwell PM. Can overall results of clinical trials be applied to all patients? Lancet 1995: 345: 1616-1619.
8. Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? Control Clin Trials. 1998; 19: 249-256.
9. Senn S. Testing for baseline balance in clinical trials. Stat Med 1994; 13: 1715-1726.
10. Moreira ED, Stein Z, Susser E. Reporting on methods of subgroup analysis in clinical trials: a survey of four scientific journals. Braz J Med Biol Res 2001; 34: 1441-1446.
11. Parker AB, Naylor CD. Subgroups, treatment effects and baseline risks: Some lessons from major cardiovascular trials. Am Heart J 2000; 139: 952-961.
12. Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technol Assess 2001; 5: 1-56.
13. Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. J Clin Epidemiol 1999; 52: 19-26.
14. Steyerberg EW, Bossuyt PM, Lee KL. Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics? Am Heart J 2000; 139: 745-751.
15. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. International Statistical Review 1991; 58: 227-240.
16. Canner PL. Covariate adjustement of treatment effects in clinical trials. Control Clin Trials 1991; 12: 359-366.
17. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. Stat Med 2002; 21: 2917-2930.
18. Ford I, Norrie J. The role of covariates in estimating treatment and risk in long

term clinical trials. Stat Med 2002; 21: 2899-2908.
19. Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. Control Clin Trials 2000; 21: 330-342.
20. Hauck WW, Neuhaus JM, Kalbfleisch JD, Anderson S. A consequence of omitted covariates when estimating odds ratios. J Clin Epidemiol 1991; 44: 77-81.
21. Begg MD, Lagakos S. Loss of efficiency caused by omitting covariates and misspecifying exposure in logistic regression models. J Am Stat Association 1993; 88: 166-170.
22. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with non-linear regressions and omitted variables. Biometrika 1984; 71: 431-444.
23. Collins R, MacMahon S. Reliable assessment of the effects of treatment on mortality and major morbidity, I: Clinical trials. Lancet 2001; 357: 373-380.
24. Bailey KR. Clinical trials in acute myocardial infarction: when should we adjust for baseline characteristics? Am Heart J 2000; 139: 761-763.
25. Beach ML, Meier P. Choosing covariates in the analysis of clinical trials. Control Clin Trials 1989; 10: 161S-175S.
26. Serruys PW, de Feyer P, Macaya C, et al. Fluvastatin for prevention of cardiac events following successful first percutaneous coronary intervention: a randomized controlled trial. JAMA 2002; 287: 3215-3222.
27. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20536 high-risk individuals: a randomized placebo-controlled trial. Lancet 2002; 360: 7-22.
28. van Es RF, Jonker JJ, Verheugt FW, Deckers JW, Grobbee DE. Aspirin and coumadin after acute coronary syndromes (the ASPECT-2 study): a randomised controlled trial. Lancet 2002; 360: 109-113.
29. Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 1999; 52: 935-942.
30. Edwards D. On model pre-specification in confirmatory randomized studies. Stat Med 1999; 18: 771-785.
31. Choi SC. Sample size in clinical trials with dichotomous endpoints: Use of covariables. J Biopharm Stat 1998; 8: 367-375.

Randomized controlled trials with time-to-event outcomes: how much does pre-specified covariate adjustment increase power?

2.2

ABSTRACT

Purpose: We evaluated the effects of various strategies of covariate adjustment on type I error, power and potential reduction in sample size, in randomized controlled trials (RCTs) with time-to-event outcomes.

Methods: We used Cox models in simulated datasets with different treatment effects (Hazard Ratio [HR]= 1, 1.4, 1.7), covariate effects (HR=1, 2, 5), covariate prevalences (10%, 50%) and censoring levels (No, low, high). The treatment and a single covariate were dichotomous. We examined the sample size that gives the same power as an unadjusted analysis, for three strategies: pre-specified, significant predictive and significant imbalance.

Results: Type I error was generally at the nominal level. The power to detect a true treatment effect was higher with adjusted than with unadjusted analyses, especially with pre-specified and significant predictive strategies. Potential reductions in sample size with covariate HR between 2 and 5 were between 15% and 44% (covariate prevalence 50%) and between 4% and 12% (covariate prevalence 10%). The significant imbalance strategy yielded small reductions. The reduction was higher with stronger covariate effects, but was independent of the treatment effect, sample size and censoring level.

Conclusions: Adjustment for one predictive baseline characteristic yields higher power to detect a true treatment effect than unadjusted analysis, without inflation of type I error and with potentially moderate reductions in sample size. The analysis of RCTs with time-to-event outcomes should adjust for predictive covariates.

Introduction

Randomized controlled trials (RCTs) are important research tools to evaluate the usefulness of treatments and interventions [1]. Heterogeneity is common among patients participating in RCTs with time-to-event outcomes [2]. Prognosis commonly varies according to patient baseline characteristics, which are routinely recorded in RCTs. After proper randomization, imbalance in patient characteristics may arise by chance [3].

Covariate adjustment for prognostic baseline characteristics is usually performed with Cox proportional hazards model in RCTs with time-to-event outcomes [3-10]. The inclusion of a strongly predictive covariate in addition to the treatment variable in a Cox model provides three important benefits: correction for imbalance [3, 4, 6, 9], acquisition of more individualized treatment effects [3, 7, 9] and increase in statistical power, i.e., the ability to detect a treatment effect when it really exists [2, 5-9]. Moreover, omission or misspecification of prognostic covariates in the analysis produces deviations from the proportional hazards assumptions [5, 10-15].

The power of covariate adjustment strategies in RCTs with time-to-event outcomes depends on various characteristics: strength of treatment effect, strength of covariate effect, covariate prevalence and censoring level [2, 5-7, 9, 11-13, 22]. The effects of the covariate adjustment strategies on statistical power and type I error, using plausible clinical scenarios, have been insufficiently studied [16, 17]. Some examples of covariate adjustment in RCTs with survival outcomes are available in the medical literature, specially in oncology and cardiology [18-21].

We used various strategies for choice of covariates (pre-specified, predictive and imbalance strategies) in simulated Cox proportional hazards models with one dichotomous covariate, using different treatment effects, covariate effects, covariate prevalences and censoring levels. We aimed to identify the pros and the cons of each covariate adjustment strategy, with a focus on the quantification of changes in statistical power. We expressed the gain in statistical power in the decrease in the sample size that gives the same power as an unadjusted analysis.

Methods

Models and adjustment strategies

Cox proportional hazards models were used to analyze the effects of treatment on a time-to-event outcome (e.g. time to death). For simplicity, a single dichotomous baseline characteristic was considered as covariate to adjust the treatment effect.

We expected that our results were generalizable to more complex cases (i.e. more covariates included), as demonstrated by others [2, 5-7, 9, 11, 13, 22]. The Cox model coefficients and standard errors (SE) were estimated with standard maximum likelihood procedures.

We used three strategies for choice of covariates. The 'adjustment' strategy refers to a pre-specified adjusted analysis, which uses a known predictive covariate. This strategy should be written in advance in the protocol. This leads to a stratified estimate of the treatment effect, which is also corrected for imbalance. 'Predictor p<5 percent' refers to a strategy in which the predictive effect of a baseline characteristic can be used for adjustment if statistically significant using Pearson's chi-square test (p<5 percent). The treatment effect was not included in this predictive model. The imbalance according to a baseline characteristic may also be used for adjustment if statistically significant ('Imbalance p<5 percent') [16, 22]. Statistically significant imbalance was based on Pearson's chi-square test with 1 degree of freedom.

Data and simulations

The various approaches to estimate the treatment effect were applied in different hypothetical data sets. We simulated RCTs of 200 patients, 100 randomized to treatment A and 100 to treatment B on average. Simulations were performed with 20,000 repetitions when a treatment effect was truly present and 100,000 when no treatment effect was present (i.e. under the null hypothesis).

In each simulation, survival times were generated from an exponential distribution with hazard rate $\exp(\gamma U + \beta V)$, where γ was the coefficient of the treatment effect, β the coefficient of the covariate effect, U the random dichotomous value of treatment with 50 percent prevalence for treatment A on average and V the random dichotomous value of covariate at a given prevalence [7, 8]. The coefficients (γ and β) are equal to the natural logarithm of the Hazard Ratio (HR) of the treatment effect and covariate effect, respectively. Censoring times were independently generated from an exponential distribution given a pre-defined censoring hazard rate and a randomly selected censoring prevalence (i.e. censoring time = -log [1-(censoring prevalence)] /censoring hazard rate) during each simulation. The survival and censoring prevalences followed a uniform distribution.

Multivariable models were constructed using different covariate effects: No covariate effect (HR = 1), moderate covariate effect (HR = 2) and strong covariate effect (HR = 5). Treatment effects varied from no treatment effect (HR = 1), weak treatment effect (HR = 1.4) to mild treatment effect (HR = 1.7). The

covariate prevalence (proportion of patients with a positive covariate value) was set at 50 percent and 10 percent. Three censoring levels were defined according to censoring hazard rates: No censoring (hazard rate 0), low level censoring (censoring between 8 percent and 18 percent, corresponding to a censoring hazard rate of 0.22) [7] and high level censoring (censoring between 25 percent and 50 percent, corresponding to a censoring hazard rate of 1). We repeated the analyses using sample sizes of 500 and 1000 patients.

An example of one of the simulations with treatment HR of 1.7, 50 percent covariate prevalence, and no censoring is displayed in Figures 1a and 1b. We observe that the average treatment and placebo survival curves (Figure 1a) are in between the survival curves fitted with the covariate with strong effect (HR=5), (Figure 1b).

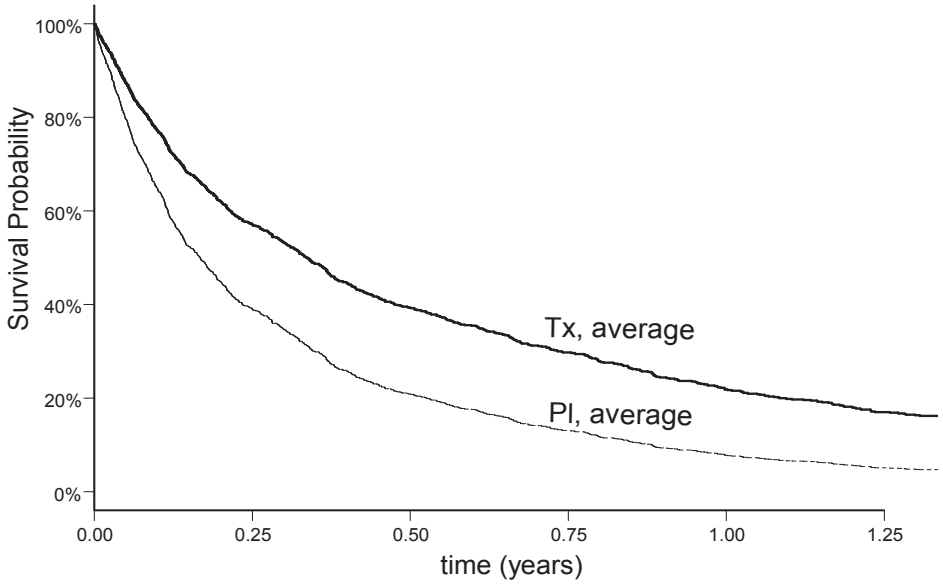

Figure 1a: Average survival curves for the treatment (Tx) and placebo (Pl) arms in a hypothetical RCT (Treatment HR= 1.7, covariate HR=5, covariate prevalence 50% , No censoring, n=1000).

Evaluation

We studied the actual type I error (α) when there was no treatment effect, and power (1-β, where β indicates type II error) when there was a true treatment effect, in each of the strategies of covariate adjustment. Statistical significance

was based on the Wald statistic (coefficient/SE) (16), with p-values < 5 percent considered significant. We expressed the gain in statistical power with adjusted strategies as the potential decrease in the required sample size that gives the same power of an unadjusted analysis. The formula of the potential reduction in sample size is shown in the Annex I [23]. We report the results with the default significance level of 5 percent. We also explored the effects of applying a more liberal criterion for selection (p<20 percent). We used S-PLUS software (version 2000, Insightful Inc, Seattle, WA, USA) for all calculations.

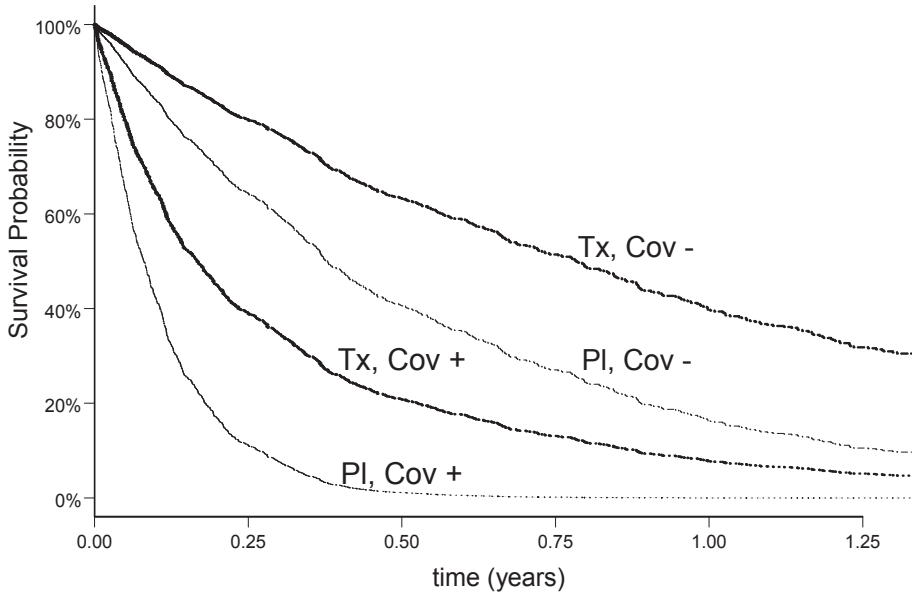

FIGURE 1b: Survival curves for the treatment and placebo arms in the same hypothetical RCT as Fig. 1a. The covariate (Cov) was included in the Cox model. The upper 2 curves correspond to each arm in the absence of the covariate. The lower 2 curves correspond to each arm in the presence of the covariate. The difference between the upper and lower curves corresponds to a HR=5.

Results

Covariate prevalence 50 percent and no censoring

When there was no treatment effect, the type I error was rather similar in all adjustment strategies and with all covariate effects (Table 1). The type I error was mainly slightly below 5 percent, especially when the covariate effect was very

Covariate effect and adjustment strategy	Treatment Effect (HR=1.0)†		Treatment Effect (HR=1.4)			Treatment effect (HR=1.7)		
	Coeff (SE)‡	Type I error (%)	Coeff (SE)	Power (%)	Reduction in sample size(%)	Coeff (SE)	Power (%)	Reduction in sample size(%)
Moderate covariate effect (HR =2):								
No adjustment	0.00 (0.09)	5.1	0.33 (0.15)	62.9	----	0.49 (0.15)	92.7	----
Adjust/Predict p<5% §	0.00 (0.09)	5.0	0.36 (0.15)	70.3	15.1	0.55 (0.15)	96.2	14.9
Imbalance p<5%	0.00 (0.09)	4.8	0.33 (0.15)	63.4	1	0.50 (0.15)	93.2	0.7
Strong covariate effect (HR =5):								
No adjustment	0.00 (0.09)	5.2	0.27 (0.15)	45.1	----	0.40 (0.15)	79.5	----
Adjust/Predict p<5%	0.00 (0.09)	4.9	0.36 (0.15)	70.1	45.3	0.55 (0.15)	95.9	43.4
Imbalance p<5%	0.00 (0.09)	4.2	0.27 (0.15)	46.4	2.9	0.41 (0.15)	80.8	2.6

† HR=Hazard ratio, defined as exp(regression coefficient); ‡ Coeff= Coefficient; § Adjustment or predictor <5% strategy.

Table 1. Simulations with covariate prevalence 50% and No censoring (n=200).

strong (3.8 percent for imbalance adjusted strategy at covariate HR=10). This conservative estimate implies that fewer false-positive effects were identified. When there was no covariate effect, covariate adjustment strategies on average did not change the treatment effect coefficients and slightly reduced the power and increased the required sample size (around 0.1 percent).

When there were treatment and covariate effects (Table 1), the pre-specified covariate adjustment ('adjustment') strategy yielded on average a more extreme treatment effect compared to no adjustment, as theoretically expected. Adjustment for one statistically significant predictor ('predictor p<5 percent') generated similar results to pre-specified adjustment since nearly always the pre-specified covariate was a statistically significant predictor. Significant imbalance adjustment ('imbalance p<5 percent') on average generated coefficients smaller than both pre-specified and significant predictive coefficients and nearly similar to the unadjusted coefficients.

With both the pre-specified and the significant predictor strategies, power was higher than with unadjusted analyses. In contrast, power with the significant imbalance strategy was only slightly increased. The increase in power was directly related to the strength of the covariate effect. A potential reduction in sample size was evident when pre-specified and significant predictor strategies were performed (between 15 percent and 44 percent, corresponding to covariate HR between 2 and 5) as shown in Table 1. The imbalance adjustment yielded smaller reductions (between 1 percent and 3 percent, covariate HR 2 to 5).

Remarkably, we observed a similar reduction in sample size when different treatment effects were used given a certain covariate effect (e.g. 44 percent if treatment effect was weak and 43 percent if treatment effect was mild, with covariate HR=5). The potential reduction in sample size was similar in simulations with larger sample sizes (n=500, n=1000) and when we considered p<20 percent instead of p<5 percent as covariate selection limit. The same similarities were also found in all the following scenarios.

Covariate prevalence 50 percent and censoring

Compared to covariate prevalence 50 percent and no censoring, the treatment effect coefficients and type I error were similar. In the same comparison, power was slightly reduced with all unadjusted and adjusted strategies with low level censoring, but moderately reduced with high level censoring (e.g. from 96 percent to 84 percent, with covariate HR=2, treatment HR=1.7, n=200). The potential reduction in sample size was similar with the pre-specified and significant predictor strategies (between 12 percent and 44 percent, covariate HR 2 to 5). There

Covariate effect and adjustment strategy	Treatment effect (HR=1)†		Treatment Effect (HR=1.4)			Treatment effect (HR=1.7)		
	Coeff (SE)	Type I error (%)	Coeff (SE)	Power (%)	Reduction in sample size(%)	Coeff (SE)	Power (%)	Reduction in sample size(%)
Moderate covariate effect (HR =2):								
No adjustment	0.00 (0.09)	5.1	0.35 (0.15)	67.5	----	0.53 (0.15)	95.2	----
Adjust/Predict p<5% §	0.00 (0.09)	5.0	0.36 (0.15)	69.5	3.8	0.54 (0.15)	95.8	3.7
Imbalance p<5%	0.00 (0.09)	4.6	0.35 (0.15)	69.4	0.2	0.53 (0.15)	95.2	0.2
Strong covariate effect (HR =5):								
No adjustment	0.00 (0.09)	5.2	0.33 (0.15)	64	----	0.50 (0.15)	93.7	----
Adjust/Predict p<5%	0.00 (0.09)	5.1	0.36 (0.15)	70	11.9	0.54 (0.15)	95.8	11.6
Imbalance p<5%	0.00 (0.09)	5.0	0.34 (0.15)	65	0.8	0.51 (0.15)	94	0.5

† HR=Hazard ratio, defined as exp(regression coefficient); ‡ Coeff= Coefficient; § Adjustment or predictor <5% strategy.

Table 2. Simulations with covariate prevalence 10% and No censoring (n=200).

	No Censoring		High level censoring	
	50% CP†	10% CP	50% CP	10% CP
Moderate covariate effect (HR =2)‡:				
Adjustment/ Predictor p<5%	15	4	12	4
Imbalance p<5%	1	0.4	1	0.1
Strong covariate effect (HR =5):				
Adjustment/ Predictor p<5%	44	12	44	14
Imbalance p<5%	3	0.6	3	0.8

† CP= covariate prevalence; ‡ HR= Hazard ratio.

Table 3. Summary of the percentage of reduction in sample size according to censoring, covariate prevalence and covariate effect.

was only a small reduction in sample size with the imbalance strategy (between 1 percent and 3 percent, covariate HR 2 to 5).

Covariate prevalence 10 percent and no censoring

Adjusted treatment effect coefficients had slightly higher values than the unadjusted ones, in contrast to the situation with covariate distribution 50 percent and no censoring, where the difference was larger. Type I error was close to the nominal level, without important conservative estimates (Table 2). Power estimates of adjusted strategies were closer to the unadjusted power estimate, because of the similarity of unadjusted and adjusted treatment effects. Consequently, the potential reduction in sample size was lower with adjustment for pre-specified or significant predictor covariates (between 4 percent and 12 percent, covariate HR between 2 and 5) as shown in Table 2. The imbalance strategy reduced the sample size only slightly (between 0.4 percent and 0.6 percent).

Covariate prevalence 10 percent and censoring

Compared to covariate distribution 10 percent and no censoring, the treatment effect coefficients were similar. Type I error was mainly slightly below 5 percent, without a clear relation among adjustment strategies. The power was slightly lower for all unadjusted and adjusted strategies with low level censoring, but mildly reduced with high level censoring. The potential reduction in sample size with pre-specified and significant predictive strategies was between 4 percent and

14 percent (covariate HR 2 to 5). The imbalance strategy only slightly reduced the sample size (0.1 percent to 0.8 percent). Table 3 shows a summary of the potential reductions in sample size within plausible scenarios.

Discussion

Covariate adjustment yields a higher statistical power than an unadjusted analysis of a randomized controlled trial with time-to-event outcomes. In our study, the gain in power was translated into potential reductions in sample size, without inflation of the type I error. We used several approaches of covariate adjustment in simple simulated Cox models with different treatment effects, covariate effects, covariate prevalences and censoring levels. We found that pre-specified and significant predictor adjustment strategies, with covariate HR between 2 and 5, yielded a potential reduction of sample size between 15 percent and 44 percent (covariate prevalence 50 percent) and between 4 percent and 12 percent (covariate prevalence 10 percent). This reduction was greater if the covariate effect was higher. The reduction in sample size was independent of the treatment effect, sample size and censoring level. The significant imbalance adjustment strategy showed smaller potential reductions in sample size. There was very limited damage when adjustment was used in the absence of a true covariate effect.

The use of covariate adjustment strategies has been previously studied in RCT with different outcomes: continuous [22], dichotomous [23, 24], and time-to-event [7-9, 16]. However, it is not known what truly constitutes the best strategy for covariate-adjusted analyses [22]. The significant predictor, significant imbalance and predictor plus imbalance adjustment strategies were previously described by Beach and Meier in RCTs with time-to-event outcomes analyzed with Cox proportional hazards models [16]. The imbalance strategy for one covariate yielded similar values of treatment effect than the predictive strategy, probably related to the low number of events or low predictive significance of the variable included in their example. However, they did not give a clear recommendation about the most efficient strategy to perform covariate adjustment. Pocock et al. [22] described the pre-specified, significant predictor and significant imbalance adjustment strategies in RCTs analyzed with linear and non-linear regression models, and they advised to use pre-specified predictors in the analysis. Our results agree with this recommendation. We did not report the results of the combination of strategies ('predictor and imbalance' and 'predictor or imbalance') because they were similar to the significant imbalance and significant predictor strategies, respectively.

A potential problem of covariate adjustment when analyzing RCTs with Cox

models is that we may misspecify a prognostic variable (e.g. by categorizing a continuous variable or by linear modeling of a non-linear relationship). This situation produces deviations of the proportional hazards assumption and understimation of the treatment effect [5, 10-15]. This leads to a loss of power of the test of no treatment effect. In contrast, the misspecification of predictive covariates in linear models does not produce biased estimates of the treatment effect. It however leads to a loss of precision of the estimate [3, 5]. Consequently, the recommendation to adjust for a predictive covariate in Cox models is based on really different arguments than in the classical linear models [6, 10].

We found that type I error was maintained closely below to the nominal level with the pre-specified and predictive adjustment strategies, using the Wald test. Li [7] found type I error values very close to the nominal level of 5 percent in the context of the covariate-adjusted Cox score test. The imbalance adjustment strategy was conservative, since this strategy constrains the outcome variability between treatment groups [22]. We also confirmed that there was only a small loss in power when adjustment was used in the absence of covariate effect, as suggested by others [7, 9, 22, 25]. Further, we quantified a substantial gain in power when a strong predictive covariate was included in the Cox model, especially with 50 percent covariate prevalence and independently of censoring.

We expressed the gain in statistical power as a potential reduction in the required sample size. The potential reductions in sample size with pre-specified and significant predictor strategies were moderate (between 12 and 44 percent) in plausible scenarios (treatment HRs between 1.4 and 1.7, covariate HRs between 2 and 5 [Spearman's correlation between 0.29 and 0.57], covariate prevalence of 50 percent). The imbalance strategy yielded lower reductions (below 3 percent). With 10 percent covariate prevalence, the reductions in sample size were below 13 percent in the pre-specified/predictor strategies, and below 1 percent in the significant imbalance strategy. Furthermore, censoring levels in RCT populations in some medical fields are high (e.g. Oncology). Remarkably, the proposed potential reduction in sample size in the high level censoring situation was similar to the no or low censoring scenarios. Hence, censoring was not relevant to the choice of adjustment strategy.

Moreover, the potential reduction in the sample size was independent of the treatment effect. We used treatment effects (HRs 1.4 and 1.7) in the range suggested in simulations by others (range of treatment effects HRs from 1 to 2.7) [2, 5-7, 16] and similar to the treatment HR of some recently published RCTs in Oncology and Cardiology, which range from 1.1 to 1.5 [18-21]. Different sample sizes of the simulated RCTs did not change the benefit in the potential reduction in sample size, as suggested previously [22]. Nevertheless, low covariate prevalence

(10 percent) yielded lower reductions in sample size. The stability of the potential reduction of sample size calculations in common clinical scenarios makes it an attractive summary measure to express the benefit of adjustment strategies, as was suggested in RCTs analyzed with logistic models [23].

Reductions in sample size given a covariate effect are not directly comparable between a RCT with dichotomous outcomes [23] and a RCT with survival outcomes, using covariate prevalence of 50 percent and no censoring. Values of Odds Ratios (OR) and HR are related to each other through the formula: OR= HR*(p2/p1), where p2 is the survival in the covariate group with the best prognosis and p1 is the survival in the covariate group with the worst prognosis (see Annex II). So, a given value of HR corresponds to a higher OR. For example, a reduction of sample size of 16 percent can be achieved after adjustment for a covariate effect with HR = 2. A comparable reduction of sample size (14 percent) can be obtained after adjustment for a covariate effect with OR = 5 [ref. 23]. Conversely, for the same numerical values of OR and HR (e.g. 2), we obtained reductions in sample size after adjustment of 3 percent and 16 percent, respectively.

We did not present calculations for covariate adjusted sample size in RCT design. To quantify any anticipated sample size gains with covariate adjustment, we would need to specify covariate effects and covariate distributions. In practice, the study would then have to meet these assumptions to achieve the calculated power. Therefore, we advise to perform an unadjusted sample size calculation, which needs fewer assumptions. The actual analysis will have more power if the pre-specified or significant predictor strategy is followed.

The role of covariate adjustment in nonlinear analyses of randomized controlled trial is a matter of debate [3, 13, 22, 24]. Hauck et al. recommended that the primary analysis should be adjusted for important prognostic covariates in order to come as close as possible to the clinically most relevant subject-specific measure of treatment effect [3]. However, this practice is not common [22]. Further, Chastang et al. also noted that if important differences are found between unadjusted and adjusted analyses, it is probably preferable to base inference on the latter [13]. However, a recent survey indicated that the number of trials with a difference between adjusted and unadjusted analyses sufficient to affect the conclusions is very low [22]. The importance to pre-specify the covariates to adjust for has been also highlighted [3, 13, 22, 26]. Nevertheless, this recommendation may be unrealistic because prognostic covariates might not be available in the study design stage. Despite of the criticisms about their data-driven nature and potential for manipulation, variable selection procedures may have a useful role on formulating covariate adjustment, especially in large trials [22, 26]. Further, the variable selection procedures may be applied to select covariates in a way that

does not affect the type I error for the test for no treatment effect [26].

Finally, if a covariate is not predictive of an outcome, a statistically significant imbalance of this covariate is irrelevant [22]. What it is useful is to adjust for predictive covariates, independently of their imbalance status. Thus, some researchers have suggested the use of a limited number of pre-specified, prognostic covariates, and the avoidance of the assessment of and the adjustment for imbalance [22-25].

Our work has several limitations. We only included results of simulations with plausible covariate and treatment effects and not real data. We only considered covariate prevalences of 10 percent and 50 percent. Nevertheless, results with covariate distribution of 90 percent are rather similar to results of covariate distribution of 10 percent [5, 12]. We did not study the effect of covariate adjustment using a continuous covariate. We used the Wald test for statistical significance, although the Likelihood Ratio test may yield statistically more appropriate results. Also, we did not address the consequences of covariate adjustment when more than one covariate is available. If we had ten covariates, five of them were predictive and five were not but allowed the choice to vary randomly from simulation to simulation, then a strategy of pre-specification would work less well. A selection procedure, e.g. stepwise backward elimination, might be considered to choose a limited number of the covariates, but the effects of this procedure would require further study. Edwards [26] has suggested a method of blinded selection of covariates, in which a family of possible models and a model selection criterion are pre-specified.

We recommend use of the adjusted analysis in time-to-event RCTs because we obtained more appropriate treatment effects (corrected for imbalance, individually oriented), a higher power than the unadjusted effects and consequently a reduction in sample size. The best strategy of covariate adjustment should be pre-specified adjustment if a strong predictive covariate is known before the RCT analysis [7, 9, 22]. An alternative is to use the significant predictor strategy, when a strong predictive covariate is not known. The analysis of RCTs with time-to-event outcomes should adjust for predictive covariates.

Annexes

I. Deduction of the Reduction in Sample Size (RSS) formula

In the context of a RCT, we defined nu and na as the unadjusted and adjusted sample sizes respectively, Z_α is Z value at α =0.05, $Z_{\beta u}$ is the Z value of the unadjusted power 1-βu, $Z_{\beta a}$ is the Z value of the adjusted power 1-βa, Z_u is the mean standardized Z score of the estimated unadjusted coefficient defined as coefficient /SE for each simulation, and, Z_a is the mean standardized Z score of the estimated adjusted coefficient defined as coefficient /SE for each simulation.

We expressed any increase in statistical power of the adjusted analysis as the decrease in the sample size that gives the same power of an unadjusted analysis (i.e. $Z_{\beta u} = Z_{\beta a}$). This assumption gives approximately:

$$n_a / n_u = (Z_u / Z_a)^2 \quad (1)$$

The potential reduction in sample size (RSS) between unadjusted and adjusted strategies is expressed in percentage as:

$$[(n_u - n_a)/ n_u]*100 \text{ , or}$$
$$100 - [100*(n_a /n_u)] \quad (2)$$

Replacing (1) in (2):

$$RSS = 100 - 100* (Z_u / Z_a)^2 \quad (3)$$

II. Relation between Hazard Ratio (HR) and Odds Ratio (OR)

Let us consider a randomized controlled trial with survival outcomes, one dichotomous covariate, covariate prevalence of 50 percent and no censoring. Each covariate group has different prognosis, according to its baseline characteristics. Survival curves of each covariate group (group 1 and group 2) go from time t0 until time t_1 ($t_1>t_0$). The survival proportions at t_1 are p_1 (group 1) and p_2 (group 2). The number of patients is the same for each covariate group ($n=n_1=n_2$).

Let us assume that group 1 has worse prognosis and that the hazard rates (h) are constant from t_0 until t_1. So, the h for each covariate group (g_1 and g_2) are defined as:

h_1= (number events g_1)/(total follow-up1) = $[(1-p_1)*n_1]/ t_1*n_1 = (1-p_1)/t_1$

h_2= (number events g_2)/(total follow-up2) = $[(1-p_2)*n2]/ t1*n_2 = (1-p_2)/t_1$

The hazard ratio (HR) of the covariate (g_1 vs. g_2) is defined as:

$$HR = h_1/h_2 = [(1-p_1)/t_1]/[(1-p_2)/t_1] = (1-p_1)/(1-p_2) \quad (4)$$

At time t_1, the Odds Ratio (OR) of death of g_1 vs. g_2 is defined as:

$$OR = [(\text{death } g_1)*(\text{alive } g_2)]/[(\text{death } g_2)*(\text{alive } g_1)] = [(1-p_1)*p_2]/[(1-p_2)*p_1] \quad (5)$$

Then, replacing (4) in (5), we have:

$$OR = HR* (p_2/p_1) \quad (6)$$

which implies that OR is always larger than HR in this particular situation.

References

1. DeMets DL. Clinical trials in the new millennium. Stat Med 2002; 21: 2779-87.
2. Schumacher M, Olschewski M, Schmoor C. The impact of heterogeneity on comparison of survival times. Stat Med 1987; 6: 773-84.
3. Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in non-linear regression analyses of randomized trials?. Control Clin Trials 1998; 19: 249-56.
4. Fleming TR, Lin DY. Survival analysis in clinical trials: past developments and future directions. Biometrics 2000; 56: 971-83.
5. Schmoor C, Schumacher M. Effects of covariate omission and categorization when analysing randomized trials with the Cox model. Stat Med 1997; 16: 225-37.
6. Ford I, Norrie J, Ahmadi S. Model inconsistency, illustrated by the Cox proportional hazards model. Stat Med 1995; 14: 735-46.
7. Li Z. Covariate adjustment for non-parametric tests for censored survival data. Stat Med 2001; 20: 1843-53.
8. Akazawa K, Nakamura T, Palesch Y. Power of logrank test and Cox regression model in clinical trials with heterogeneous samples. Stat Med 1997; 583-97.
9. Ford I, Norrie J. The role of covariates in estimating treatment effects and risk in long-term clinical trials. Stat Med 2002; 21: 2899-2908.
10. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with non-linear regression and omitted covariates. Biometrika 1984; 71: 431-44.
11. Lagakos SW, Schoenfeld DA. Properties of proportional-hazards score tests under misspecified regression models. Biometrics 1984; 40: 1037-48.
12. Morgan TM. Omitting covariates from the proportional hazards model. Biometrics 1986; 42: 993-5.

13. Chastang C, Byar D, Piantadosi S. A quantitative study of the bias in estimating the treatment effect caused by omitting a balanced covariate in survival models. Stat Med 1988; 7: 1243-55.
14. Struthers CA, Kalbfleisch JD. Misspecified proportional hazards models. Biometrika 1986; 73: 363-9.
15. Anderson GL, Fleming TR. Model misspecification in proportional hazards regression. Biometrika 1995; 82: 527-41.
16. Beach ML, Meier P. Choosing covariates in the analysis of clinical trials. Control Clin Trials 1989; 10: 161S-175S.
17. Canner PL. Covariate adjustment of treatment effects in clinical trials. Control Clin Trials 1991: 12: 359-66.
18. Brem H, Piantadosi S, Burger PC et al. Placebo-controlled trial of safety and efficacy of intraoperative controlled delivery by biodegradable polymers of chemotherapy for recurrent gliomas. Lancet 1995; 345: 1008-12.
19. The West of Scotland Coronary Prevention Study Group. Baseline risk factors and their association with outcome in the West of Scotland coronary prevention study. Am J Cardiol 1997; 79: 756-62.
20. Dickstein K, Kjekshus J, and the OPTIMAAL Steering Committee. Effects of losartan and captopril on mortality and morbidity in high-risk patients after acute myocardial infarction: the OPTIMAAL randomised trial. Lancet 2002: 360:752-760.
21. Bonnefoy E, Lapostolle F, Leizorovicz A, et al. Primary angioplasty versus prehospital fibrinolysis in acute myocardial infarction: a randomised study. Lancet 2002: 360: 825-829.
22. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med 2002; 21: 2917-30.
23. Hernández AV, Steyerberg EW, Habbema JDF. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power, and reduces sample size requirements. J Clin Epidemiol 2004; 57: 454-60.
24. Steyerberg EW, Bossuyt PM, Lee KL. Clinical trials in acute myocardial infarction: should we have adjust for baseline characteristics? Am Heart J 2000; 139: 745-51.
25. Senn S. Testing for baseline balance in clinical trials. Stat Med 1994; 13: 1715-26.
26. Edwards D. On model prespecification in confirmatory randomized studies. Stat Med 1999; 18: 771-85.

Current practice of reporting clinical trials

3

Inappropriate use of baseline characteristics in clinical trials: Assessment of high impact medical journals

3.1

ABSTRACT

We aimed to describe current reporting of uses of patient baseline characteristics (baseline comparability, covariate adjustment, and subgroup analysis) in randomized controlled trials (RCTs), and to determine whether reporting violations decreased since the introduction of the revised CONSORT (Consolidated Standards of Reporting Trials) statement. Main reports of phase III RCTs with more than 100 patients, published in high impact factor clinical journals between September 1 and November 30, 2002 were used. Our main outcome measure was the appropriateness of reporting of baseline comparability, covariate adjustment and subgroup analysis. We identified 84 RCTs (general medicine=46, cardiology=21 and oncology=17). Remarkably, 34 RCTs (40%) tested baseline imbalances, which is methodologically unjustified. Forty-four RCTs (52%) performed covariate adjustment as the primary analysis, and they used mainly predictive covariates. Of these, 30 RCTs gave more emphasis to the adjusted analysis. Forty-seven RCTs (56%) reported subgroup analyses. Appropriate statistical interaction tests were used in only 20 of these trials, while subgroup differences were emphasized in 22. Adherence to methodological standards was not better in CONSORT adopting journals. Inappropriate analyses of baseline characteristics, especially subgroup analyses, are still frequent. Recommendations on appropriate reporting need to be further implemented in current practice, since conclusions of RCTs may be misleading otherwise.

Introduction

In the current era of evidence-based medicine, randomized controlled trials (RCTs) are the key to guide clinical decision making [1]. Adequate reporting of RCTs provide clinicians with valuable information that helps them to accept or reject treatments or interventions, and, therefore, to improve their practice [2]. Several patient characteristics are usually recorded at inclusion. These baseline characteristics may serve several purposes, including demonstration of balance between treatment groups (baseline comparability) [3], and implementation of more elaborate analyses (such as covariate adjustment and subgroup analyses) [4].

The revised recommendations of the Consolidated Standards of Reporting Trials (CONSORT) statement has condensed some appropriate and inappropriate uses of baseline characteristics in RCTs (Table 1) [5, 6]. Baseline characteristics among trial arms should be displayed without testing for imbalances, since these are the product of chance if a proper randomization method was followed [5]. Important covariates might be used for covariate adjustment. Such adjustment should be pre-specified, and be focused on a limited number of covariates. The reasons for choosing covariates should be clearly described [7]. Likewise, subgroup analyses should be pre-specified and be restricted to a small number of subgroups. Interaction tests should be applied when differences in treatment effect are claimed. Subgroup analyses should generally be considered as secondary analyses [4, 7].

The reporting of the use of baseline characteristics in RCTs has been found inappropriate: overuse of baseline comparisons, misuse of significance tests for baseline comparisons, inconsistencies in the use of covariate adjustment, underuse of tests of interactions, and overinterpretation of subgroup analyses [7]. These problems were demonstrated in RCTs from general medical journals during the past two decades [4, 8, 9]. The use of the revised CONSORT statement led to improvements in the quality of reporting [10, 11], but whether the reporting of the use of baseline characteristics has also improved has not yet been evaluated. Moreover, the quality of reporting on RCTs may be better in general medical journals than in specialist journals [12].

We studied adherence of the reporting of baseline data in RCTs to methodological standards. Secondarily, we studied differences between CONSORT-adopting and non-adopting journals, and explored differences according to type of journal and overall result.

Topic	Revised CONSORT Item	Appropriate	Inappropriate
Baseline comparability	15	Describe comparison of baseline characteristics between arms.	Test for imbalance of baseline characteristics between arms.
		Do not show p values in table 1.	Show p values in table 1.
Covariate adjustment	12, 18	Adjust for a limited number of covariates.	Examine many models and discuss only the model with the largest treatment effect.
		State clearly the reasons to choose covariates.	Do not state the reasons to choose covariates.
Subgroup analysis	12, 18	Perform a limited number of pre-specified subgroup analyses.	Perform a large number, post hoc subgroup analyses.
		Use interaction tests.	Use separate group p values.
		Consider as secondary analyses.	Consider as important as or more important than primary analysis

RCTs denotes randomized controlled trials; CONSORT denotes Consolidated Standards of Reporting Trials.

Table 1. Appropriate and inappropriate use of baseline characteristics in RCTs, according to the revised CONSORT statement.

Methods

Selection of clinical trial reports

We handsearched RCTs from 16 journals with the highest Impact Factor in their categories (general medicine, cardiology, and oncology), published between September 1 and November 30, 2002. This period was arbitrarily chosen, and the 3-month period resembled the one used in a previous study [4]. The journals were at the top of the list in their respective subject category (ISI Journal Citation Reports 2002) and published RCTs regularly.

General medicine journals were N Engl J Med, JAMA, Lancet, Ann Intern Med, BMJ, and Arch Intern Med. Cardiology journals included Circulation, J Am Coll Cardiol, Eur Heart J, Am Heart J, and Am J Cardiol. Oncology jour-

nals included J Nat Cancer Inst, J Clin Oncol, Cancer, Brit J Cancer, and Eur J Cancer. We included main reports of phase III parallel RCTs, with individual randomization [13], and excluded cross-over trials, cluster-randomized trials, equivalence/non-inferiority trials, trials with less than 50 patients per arm, and factorial trials [7, 14]. We excluded secondary trial reports designed to address additional hypotheses and subgroup comparisons specifically.

Assessment of trial reports and definitions

We examined RCTs and collected information in standard formats. Information about reporting of background information, baseline comparability, covariate adjustment and subgroup analyses was similar to a previous study [4] (Table 2). We measured the correspondence between the reporting of uses of baseline characteristics and methodological standards (Table 1). Our main outcome measure was the appropriateness of reporting of baseline comparability, covariate adjustment and subgroup analysis.

For consistency of evaluation across RCT reports, one of us (A.V.H.) selected the RCTs, checked inclusion and exclusion criteria and collected the information. To examine the reproducibility of the information, another reviewer (E.W.S.) also assessed the RCTs while blinded to the initial assessments. We found small differences in the extracted data, which were resolved by consensus.

An interaction test directly assesses differences in treatment effect between complementary subgroups and involves one statistical test irrespective of the number of subgroups. A significant interaction test means that the treatment effect in one subgroup is different from the treatment effect in the other subgroup. In contrast, the separate subgroup p value method assesses treatment effects in each group independently, which is inappropriate from a methodological point of view [5].

Secondary analyses

We compared baseline use between RCTs from journals which adopted the CONSORT statement (JAMA, Lancet, Ann Intern Med, BMJ, Arch Intern Med, J Nat Cancer Inst, Eur J Cancer) and journals which did not (N Engl J Med, Circulation, J Am Coll Cardiol, Am Heart J, Am J Cardiol, J Clin Oncol, Brit J Cancer) [15]. Our focus was on baseline data (revised CONSORT item 15), covariate adjustment and subgroup analyses (revised CONSORT items 12 and 18). We hypothesized that CONSORT-adopting journals would report these issues better than non-adopting journals [10, 11].

Background information
Number of patients
Number of treatment arms
Length of follow-up
Number of centers
Primary outcome
Baseline comparability
Number of baseline characteristics
Use of significance tests
Number of imbalances reported
Reporting of p values in table of baseline characteristics (“Table1”)
Covariate adjustment
Pre-specification
Number of covariates
Selection strategy (e.g. prognostic, center, imbalanced, etc)
Statistical method
Primary use of covariate adjustment
Emphasis of adjusted analysis with respect to unadjusted analysis
General judgement
Subgroup analyses
Pre-specification
Number of subgroup factors
Number of subgroup outcomes
Number of subgroup analyses (product of factor by outcomes)
Statistical method (interaction test/separate subgroup p value)
Claiming of subgroup effects
Emphasis of subgroups effects with respect to overall effect
General judgment

RCT denotes randomized controlled trial.

Table 2. Data collected from RCT reports

The baseline use between RCTs published in general medicine journals and those in specialty journals (cardiology and oncology) was also compared. The reporting of baseline characteristics would be better in general medicine journals than specialty journals, based on a recent paper comparing the quality of

reporting (12). Finally, we searched whether covariate adjustment and subgroup analyses were related to the overall trial results (positive or negative). A positive trial reports a statistically significant treatment effect. Negative RCTs might report covariate adjustment and/or subgroup analyses more frequently [7].

Statistical analyses

Differences between continuous variables were compared with t-tests or Mann-Whitney U tests. Categorical variables were tested with Chi-square tests [16]. We used tests for trend (one degree of freedom) for testing differences in time and differences in ordered variables. Statistical significance was defined as $p<0.05$. The software used was EXCEL and SPSS 10.0 (SPSS Inc., Chicago, USA, 1999).

Results

General characteristics of trial reports

Eighty-four RCTs were chosen from 14 journals (46 in general medicine, 21 in cardiology and 17 in oncology journals) [17, 99]. They included 44 treatment, 27 management, 11 prevention, and 2 screening RCTs. One N Engl J Med paper reported two trials [97]. Reasons for exclusion were: phase II trials (n=5), substudies (n=23), cross-over trials (n=1), equivalence/non-inferiority trials (n=6), small trials (n=25) and factorial trials (n=8). The RCTs mainly had 2 arms (n=68). The number of patients per RCT ranged from 103 to 266,064 (median 505) [26, 62], and 76 RCTs were multicentric (range 2 to 519, median 15 centers). The follow-up time ranged from 3 days to 11 years (median 1 year).

Baseline comparability

Fifty-five RCTs (65 percent) presented comparability of 10 or more variables (range 0 to 31) (Table 3). Thirty-four tested for baseline differences, and 9 noted baseline imbalances. There were 11 significant imbalances at $p<0.05$, with a total of 145 reported tests (7 percent of tests).

Covariate adjustment

Covariate adjustment was reported in 44 (52 percent) RCTs, using mainly logistic or Cox regression analyses. The RCTs included between 1 and 12 baseline characteristics (Table 4). When both adjusted and unadjusted results were given (n=24), covariate adjustment did not alter the conclusions of the unadjusted

analyses, and unadjusted analyses received more emphasis (n=14). Twenty RCTs only gave adjusted results. Eight RCTs with covariate adjustment did not state any reason to choose a covariate. Half of the adjusted analyses (22 out of 44) were motivated by prognostic characteristics, which we consider appropriate.

Subgroup analysis

Forty–seven (55 percent) RCTs reported subgroup analysis, which were fully pre-specified in 22 (Table 5). The median number of subgroup factors was three (range 1 to 23). Twenty-three RCTs used more than two outcomes. Sixteen and six RCTs performed ≥12 and ≥25 subgroup analyses, respectively. Only 20 RCTs used the appropriate tests of interaction. Overall, 40 significant subgroups were found among 482 tests performed (9 percent), and 16 significant results among 235 reported interaction tests (7 percent). Twenty-two RCTs overemphasized the subgroup findings.

	n=84	%
Number of baseline characteristics compared in "table 1"		
0	1	1
1 to 4	3	4
5 to 9	25	30
10 to 19	36	43
20 to 29	15	18
30 or more	4	5
Significance tests for baseline differences?		
Number of trials that performed tests	34	41
Significant imbalance per trial	9	26
Number of significance tests	145	
Significant imbalance per test	11	7

Table 3. Reporting of comparability of baseline chracteristics

Secondary analyses

The use of baseline characteristics was slightly better in CONSORT-adopting in comparison with non-adopting journals: testing for imbalances (13/36 vs. 18/48), reporting of covariate adjustment (16/36 vs. 28/48), number of subgroup analyses (median 5 vs. 9), and use of interaction tests (11/25 vs. 9/22). Although reporting of subgroup analysis was more frequent in CONSORT-adopting journals (25/36

	n=44	%
Which analysis received more emphasis?		
Unadjusted	14	32
Adjusted	20	45
Equal emphasis	10	23
Number of covariates		
1	8	18
2	4	9
3	7	16
4	6	14
5 to 9	10	23
10 or more	1	2
Unclear	8	10
Reasons for choice of covariates*		
No reason given	8	18
Covariates were/expected to be prognostic	22	50
Were prognostic (tested)	6	14
Were expected to be (pre-specified)	16	36
Imbalances between groups	5	11
Center or country	12	27
Baseline value of quantitative outcome	8	18
Other treatment factor in a factorial trial	NA	
Covariates used in stratified randomization	8	18

* more than one reason in some trials.

Table 4. Reporting of covariate adjustment in 44 clinical trials.

	n=47	%
Number of baseline characteristics		
1 to 3	29	62
4 to 6	9	19
7 or more	9	19
Number of outcomes for subgroups		
1 or 2	24	51
3 to 5	17	36
6 or more	4	13
Total number of subgroup analysis		
1 to 2	8	17
3 to 5	12	26
6 to 8	4	9
9 to 11	7	15
12 to 24	10	21
25 or more	6	13
Statistical method used		
Descriptive only	0	0
Subgroup p values	25	53
Interaction test	20	43
Subgroups differences claimed?		
Yes	22	47
No	25	53
Subgroup findings overemphasized?		
Yes	20	43
No	27	57

Table 5. Reporting of subgroup analysis in 47 clinical trials.

vs. 22/48; p=0.03), it was mainly pre-specified (14/25 vs. 8/22).

RCTs from general medicine and cardiology journals compared more baseline characteristics than oncology journals (≥ 10 characteristics: 67 percent

vs. 90 percent vs. 29 percent; p=0.03). Reporting of other aspects of baseline comparability, covariate adjustment, and subgroup analysis was similar among specialties.

Negative and positive trials reported covariate adjustment in similar frequency (22/39 vs. 22/45), as well as subgroup analyses (22/39 vs. 25/45). Negative trials reported post-hoc subgroups more frequently (15/22 vs. 10/25), but tended to use a lower number of subgroups than positive trials (median 7 vs. 9). Subgroup differences were claimed in similar frequency in both groups (10/22 vs. 12/25). These differences were not significant.

Discussion

Despite the introduction of the revised CONSORT statement and publication of a number of methodological papers [4, 9], the reporting of baseline comparability, covariate adjustment and subgroup analysis was not appropriate in RCTs from top journals. A third of the 84 RCTs tested for baseline imbalance, which has no methodological justification [4, 5]. From the 47 RCTs that performed subgroup analyses, half were pre-specified, and only 43 percent used the appropriate interaction tests.

The first CONSORT statement suggested that appropriate reporting of design, conduct, and analysis of parallel RCTs should enable the reader to make informed judgements regarding the internal and external validity of the trial [100]. The revised CONSORT statement sharpened these recommendations further [5, 6]. The adoption of the first CONSORT guidelines led to improvements in quality of reporting [10, 11]. We provide empirical evidence that, after the introduction of the revised CONSORT statement, reporting of baseline comparisons, covariate adjustment and subgroup analysis is still a problem.

Baseline comparisons increased over time. A previous review of 80 RCTs published in 1987 in Ann Intern Med, BMJ, Lancet,and N Engl J Med showed that 39 percent of the RCTs compared ≥10 baseline characteristics [9]. In 1997 these comparisons were found in 62 percent (31 out of 50) of the RCTs [4], and we found 65 percent (55 out of 84) in 2002. A review of 206 RCTs from 10 surgical journals, published between 1997 and 1999, found that only one-third tested baseline differences [101].

The proportion of significant tests for baseline differences was around the 5 percent expected by chance. Testing for baseline differences is inappropriate, and without practical value, especially if multiple variables are compared [3, 8].

Covariate adjustment is especially advantageous when strong predictors are used [106, 108, 109]. Prognostic covariates were preferentially used in the reviewed RCTs. For example, the SoS investigators properly adjusted for five pre-specified prognostic characteristics (age, angina grade, diabetes mellitus, ejection fraction, and angiographic severity) [84]. However, a fifth of RCTs did not specify the number of covariates, or whether adjustment changed the conclusion of the unadjusted analysis. Recommendations to perform covariate adjustment include the use of a limited number of pre-specified, prognostic covariates, and the avoidance to use non-prognostic, imbalanced covariates [4, 6, 106, 109]. Pre-specification aims to preclude that investigators perform multiple analyses before arriving at the "final" set of adjustment variables that best support their conclusions [110]. However, the choice of prognostic covariates and the emphasis on adjusted analyses has not been fully agreed upon. Reviewers may sometimes have suggested to adjust for baseline variables that may be important, but were over-looked by the authors.

Subgroup analysis was overused and overinterpreted. Examples are easy to find in cardiology journals, where authors searched subgroups effects in separate publications intensively (e.g. pravastatin, metoprolol and estrogen/progestin trials) [111-114]. This may be related to the fact that investigators re-analyze data in response to findings from other studies, and suggestions from reviewers to examine additional subgroups. Likewise, the total number of subgroups evaluated was high. For instance, among the reviewed reports, 4 RCTs reported between 28 and 33 subgroups [44, 47, 71, 82]. As expected, some did not succeed [44, 47], and some found subgroup effects and claimed differences [71, 82]. Subgroups effects were usually consistent with the overall trial results [9, 115]. Importantly, only 22 out of 47 reported pre-specified subgroups, and probably the majority were post-hoc subgroups [116, 117].

Interaction testing is the appropriate method to analyze subgroups [7, 118, 119], but it was amazingly underutilized in trial reports (only 20 out of 47). However, the power of the interaction test is low [118-120]. In a trial with 80 percent power for the overall effect, the interaction test only has at most 29 percent power to detect an interaction effect of the same magnitude of the overall effect [119]. For smaller interactions effects or unequal distribution of covariates, its power is even more limited. In our review, only 7 percent of the interactions tests yielded significant results. This finding suggests that substantial true interactions are rare. Moreover, RCTs of a size adequate to detect an overall effect cannot be expected to provide reliable effect estimates within smaller groups [118]. For example, for interactions of the same size and the same power of the overall effect, sample sizes should at least be inflated fourfold [119].

The evaluation of treatment effects in separate groups is misleading since false positive subgroup effects can often be found [118-121]. Subgroup-specific tests are unreliable: A significant effect in one subgroup can be observed in 7 to 64 percent, depending on trial characteristics [119]. This problem worsens when post-hoc subgroups and multiple comparisons are performed [4, 116, 118]. For example, one RCT evaluated the effect of antiplatelet therapy following percutaneous coronary intervention in the composite end-point of death, myocardial infarction or stroke [69]. The authors used 3 baseline factors (gender, diabetes, acute coronary syndrome) and 10 outcomes, used separate subgroups, and found 5 significant subgroups among 60 tests. Likely, some of these subgroup effects are the product of chance.

Subgroups therefore should not be overinterpreted, and should be reported with caution. A recent RCT and its editorial stressed the finding that vasopressin was more effective than epinephrine for cardiopulmonary resuscitation only in asystolic patients, and claimed a change in treatment guidelines in these patients [122, 123]. This seems an overinterpretation. Subgroup findings must be replicated in other populations before they are accepted, even if an interaction test is significant in a limited number of subgroups. For example, the PRAISE II trial [124] could not replicate the subgroup effect found in the PRAISE trial [125], which showed a beneficial treatment effect of amlodipine in a subgroup of patients with nonischemic heart failure. Subgroups should also be interpreted in the context of known biological mechanisms and patient prognosis [116, 117]. Thus, some researchers consider it reasonable to rely only on main effects [119]. This reasoning is supported by the finding of only 9 percent of significant subgroups among the 482 assessed, and 7 percent significant results of 235 interaction tests.

A previous paper found that RCTs from CONSORT-adopting journals (JAMA, Lancet, BMJ) were reported significantly better than those from a non-adopting journal (N Engl J Med) [10]. Another paper showed that reports from 10 CONSORT-adopting journals improved the presentation of 11 methodological factors (including description of prognostic baseline characteristics) in comparison with 16 non-adopting journals [11]. However, these two papers evaluated the adoption of the first CONSORT guidelines [100]. We found no differences according to the CONSORT adoption. This finding may be related to the fact that we included RCTs from late 2002 only, a year and a half after the publication of the revised CONSORT guidelines. Moreover, editors and reviewers of RCTs may be aware of the CONSORT guidelines, and may apply them informally in their practice. Thus, less differences in reporting can be expected.

We noted few differences in trial reporting between general medicine and specialty journals. A previous paper found subgroup analysis in 58 of 67 (87

percent) cardiology RCTs (n>1000) published between 1980 and 1997, addressing unstable angina, myocardial infarction, left ventricular dysfunction and congestive heart failure [126]. Only 24 RCTs had fully pre-specified subgroups, and 27 reported interaction tests. These reporting deficiencies have the same direction as our findings.

Our study has some limitations. First, we reviewed a limited number of RCTs (n=84), although this number is larger than in two previous studies (38 and 50, respectively) [4, 8]. This aspect may not affect the general conclusions on reporting, but it makes comparisons between CONSORT adoption, among types of journal, and between overall results of limited power. Second, the information provided in an RCT may not necessarily be the same as what was planned in the study protocol [127]. Finally, it is possible that the influence of the CONSORT guidelines on reporting requires more time to be adopted [10].

In conclusion, the reporting of baseline characteristics in RCTs from highly influential journals can be improved considerably. Despite the introduction of the revised CONSORT recommendations, a number of important problems remain, particularly with regard to testing for imbalance, the use of large numbers of subgroup analyses, the underuse of interaction tests, and the overinterpretation of subgroup effects. The CONSORT guidelines may need to further emphasise appropriate reporting recommendations of subgroup analysis, since conclusions of RCTs may be misleading otherwise.

References

1. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA 2003;290:1624-32.
2. DeMets DL. Clinical trials in the new millennium. Stat Med 2002;21:2779-87.
3. Senn S. Testing for baseline balance in clinical trials. Stat Med 1994;13:1715-26.
4. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000;355:1064-9.
5. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663-94.
6. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001;357:1191-4.
7. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate

adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med 2002;21:2917-30.
8. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. N Engl J Med 1987;317:426-32.
9. Altman DG, Dore CJ. Randomisation and baseline comparisons in clinical trials. Lancet 1990;335:149-53.
10. Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA 2001;285:1992-5.
11. Devereaux PJ, Manns BJ, Ghali WA, Quan H, Guyatt GH. The reporting of methodological factors in randomized controlled trials and the association with a journal policy to promote adherence to the Consolidated Standards of Reporting Trials (CONSORT) checklist. Control Clin Trials 2002;23:380-8.
12. Huwiler-Muntener K, Juni P, Junker C, Egger M. Quality of reporting of randomized trials as a measure of methodologic quality. JAMA 2002;287:2801-4.
13. Green SB. Design of randomized trials. Epidemiol Rev 2002;24:4-11.
14. Gomberg-Maitland M, Frison L, Halperin JL. Active-control clinical trials to establish equivalence or noninferiority: methodological and statistical concepts linked to quality. Am Heart J 2003;146:398-403.
15. The CONSORT statement. Available from: http://www.consort-statement.org. Accessed 24 May 2004.
16. Dawson B, Trapp RG. Basic & clinical biostatistics. 3rd ed. New York, NY: Mc Graw-Hill; 2001.
17. Allen JK, Blumenthal RS, Margolis S, Young DR, Miller ER 3rd, Kelly K. Nurse case management of hypercholesterolemia in patients with coronary heart disease: results of a randomized clinical trial. Am Heart J 2002;144:678-86.
18. Elbaz M, El Mokhtar E, Khalife K, et al. Is direct coronary stenting the best strategy for long-term outcome? Results of the multicentric randomized benefit evaluation of direct coronary stenting (BET) study. Am Heart J 2002;144:E7.
19. Mitchell RG, Stoddard MF, Ben-Yehuda O, et al. Esmolol in acute ischemic syndromes. Am Heart J 2002;144:E9.
20. White WB, Sica DA, Calhoun D, Mansoor GA, Anders RJ. Preventing increases in early-morning blood pressure, heart rate, and the rate-pressure product with controlled onset extended release verapamil at bedtime versus enalapril, losartan, and placebo on arising. Am Heart J 2002;144:657-65.
21. Dujovne CA, Ettinger MP, McNeer JF, et al. Efficacy and safety of a potent new selective cholesterol absorption inhibitor, ezetimibe, in patients with primary hypercholesterolemia. Am J Cardiol 2002;90:1092-7.
22. Freed MI, Ratner R, Marcovina SM, et al. Effects of rosiglitazone alone and in combination with atorvastatin on the metabolic abnormalities in type 2 diabetes mellitus. Am J Cardiol 2002;90:947-52.
23. Gagne C, Bays HE, Weiss SR, et al. Efficacy and safety of ezetimibe added to

ongoing statin therapy for treatment of patients with primary hypercholesterolemia. Am J Cardiol 2002;90:1084-91.
24. Hoffmann R, Herrmann G, Silber S, et al. Randomized comparison of success and adverse event rates and cost effectiveness of one long versus two short stents for treatment of long coronary narrowings. Am J Cardiol 2002;90:460-4.
25. Mauri L, Bonan R, Weiner BH, et al. Cutting balloon angioplasty for the prevention of rest enosis: results of the Cutting Balloon Global Randomised Trial. Am J Cardiol 2002;90:1079-83.
26. Schaefer EJ, McNamara JR, Tayler T, et al. Effects of atorvastatin on fasting and postprandial lipoprotein subclasses in coronary heart disease patients versus control subjects. Am J Cardiol 2002;90:689-96.
27. Tsuchikane E, Kobayashi T, Takeda Y, Otsuji S, Sakurai M, Awata N. Debulking and stenting versus debulking only of coronary artery disease in patients treated with cilostazol (final results of ESPRIT). Am J Cardiol 2002;90:573-8.
28. Whelton A, White WB, Bello AE, Puma JA, Fort JG. Effects of celecoxib and rofecoxib on blood pressure and edema in patients > or =65 years of age with systemic hypertension and osteoarthritis. Am J Cardiol 2002;90:959-63.
29. Francis CW, Davidson BL, Berkowitz SD, et al. Ximelagatran versus warfarin for the prevention of venous thromboembolism after total knee arthroplasty. A randomized, double-blind trial. Ann Intern Med 2002;137:648-55.
30. Kremer JM, Genovese MC, Cannon GW, et al. Concomitant leflunomide therapy in patients with active rheumatoid arthritis despite stable doses of methotrexate. A randomized, double-blind, placebo-controlled trial. Ann Intern Med 2002;137:726-33.
31. Pavelka K, Gatterova J, Olejarova M, Machacek S, Giacovelli G, Rovati LC. Glucosamine sulfate use and delay of progression of knee osteoarthritis: a 3-year, randomized, placebo-controlled, double-blind study. Arch Intern Med 2002;162:2113-23.
32. Samama CM, Vray M, Barre J, et al. Extended venous thromboembolism prophylaxis after total hip replacement: a comparison of low-molecular-weight heparin with oral anticoagulant. Arch Intern Med 2002;162:2191-6.
33. DAFNE Study Group. Training in flexible, intensive insulin management to enable dietary freedom in people with type 1 diabetes: dose adjustment for normal eating (DAFNE) randomised controlled trial. BMJ 2002;325:746.
34. Hermiz O, Comino E, Marks G, Daffurn K, Wilson S, Harris M. Randomised controlled trial of home based care of patients with chronic obstructive pulmonary disease. BMJ 2002;325:938.
35. Meade T, Zuhrie R, Cook C, Cooper J. Bezafibrate in men with lower extremity arterial disease: randomised controlled trial. BMJ 2002;325:1139.
36. Moore S, Corner J, Haviland J, et al. Nurse led follow up and conventional medical follow up in management of patients with lung cancer: randomised trial. BMJ 2002;325:1145.
37. Rowland D, DiGuiseppi C, Roberts I, et al. Prevalence of working smoke alarms

in local authority inner city housing: randomised controlled trial. BMJ 2002;325:998-1001.
38. Walker Z, Townsend J, Oakley L, et al. Health promotion for adolescents in primary care: randomised controlled trial. BMJ 2002;325:524.
39. Bonneterre J, Roche H, Monnier A, et al. Docetaxel vs. 5-fluorouracil plus vinorelbine in metastatic breast cancer after anthracycline therapy failure. Br J Cancer 2002;87:1210-5.
40. Hanks GW, Robbins M, Sharp D, et al. The imPaCT study: a randomised controlled trial to evaluate a hospital palliative care team. Br J Cancer 2002;87:733-9.
41. de Jongh S, Ose L, Szamosi T, et al. Efficacy and safety of statin therapy in children with familial hypercholesterolemia: a randomized, double-blind, placebo-controlled trial with simvastatin. Circulation 2002;106:2231-7.
42. Hodis HN, Mack WJ, LaBree L, et al. Alpha-tocopherol supplementation in healthy individuals reduces low-density lipoprotein oxidation but not atherosclerosis: the Vitamin E Atherosclerosis Prevention Study (VEAPS). Circulation 2002;106:1453-9.
43. Holmes DR, Jr., Savage M, LaBlanche JM, et al. Results of Prevention of REStenosis with Tranilast and its Outcomes (PRESTO) trial. Circulation 2002;106:1243-50.
44. Packer M, Fowler MB, Roecker EB, et al. Effect of carvedilol on the morbidity of patients with severe chronic heart failure: results of the carvedilol prospective randomized cumulative survival (COPERNICUS) study. Circulation 2002;106:2194-9.
45. Stone AF, Mendall MA, Kaski JC, et al. Effect of treatment for Chlamydia pneumoniae and Helicobacter pylori on markers of inflammation and cardiac events in patients with acute coronary syndromes: South Thames Trial of Antibiotics in Myocardial Infarction and Unstable Angina (STAMINA). Circulation 2002;106:1219-23.
46. Taylor AJ, Kent SM, Flaherty PJ, Coyle LC, Markwood TT, Vernalis MN. ARBITER: Arterial Biology for the Investigation of the Treatment Effects of Reducing Cholesterol: a randomized trial comparing the effects of atorvastatin and pravastatin on carotid intima medial thickness. Circulation 2002;106:2055-60.
47. Zanchetti A, Bond MG, Hennig M, et al. Calcium antagonist lacidipine slows down progression of asymptomatic carotid atherosclerosis: principal results of the European Lacidipine Study on Atherosclerosis (ELSA), a randomized, double-blind, long-term trial. Circulation 2002;106:2422-7.
48. Colozza M, Bisagni G, Mosconi AM, et al. Epirubicin versus CMF as adjuvant therapy for stage I and II breast cancer: a prospective randomised study. Eur J Cancer 2002;38:2279-88.
49. von Minckwitz G, Loibl S, Brunnert K, et al. Adjuvant endocrine treatment with medroxyprogesterone acetate or tamoxifen in stage I and II endometrial cancer--a multicentre, open, controlled, prospectively randomised trial. Eur J Cancer 2002;38:2265-71.
50. Deanfield JE, Detry JM, Sellier P, et al. Medical treatment of myocardial ischemia in coronary artery disease: effect of drug regime and irregular dosing in the CAPE II trial. J Am Coll Cardio 2002;40:917-25.

51. Faxon DP, Gibbons RJ, Chronos NA, Gurbel PA, Sheehan F. The effect of blockade of the CD11/CD18 integrin receptor on infarct size in patients with acute myocardial infarction treated with direct angioplasty: the results of the HALT-MI study. J Am Coll Cardiol 2002;40:1199-204.
52. Carmichael J, Popiela T, Radstone D, et al. Randomised comparative study of tegafur/uracil and oral leucovorin versus parenteral fluorouracil and leucovorin in patients with previously untreated metastatic colorectal cancer. J Clin Oncol 2002;20:3617-27.
53. El-Sayed S, Nabid A, Shelley W, et al. Prophylaxis of radiation-associated mucositis in conventionally treated patients with head and neck cancer: a double-blind, phase III, randomized, controlled trial evaluating the clinical efficacy of an antimicrobial lozenge using a validated mucositis scoring system. J Clin Oncol 2002;20:3956-63.
54. Fisher B, Bryant J, Dignam JJ, et al. Tamoxifen, radiation therapy, or both for prevention of ipsilateral breast tumor recurrence after lumpectomy in women with invasive breast cancers of one centimeter or less. J Clin Oncol 2002;20:4141-9.
55. Hersey P, Coates AS, McCarthy WH, et al. Adjuvant immunotherapy of patients with high-risk melanoma using vaccinia viral lysates of melanoma: results of a randomized trial. J Clin Oncol 2002;20:4181-90.
56. Kaiser U, Uebelacker I, Abel U, et al. Randomised study to evaluate the use of high-dose therapy as part of primary treatment for "aggressive" lymphoma. J Clin Oncol 2002;20:4413-9.
57. Kosmidis P, Mylonakis N, Nicolaides C, et al. Paclitaxel plus carboplatin versus gemcitabine plus paclitaxel in advanced non-small-cell lung cancer: a phase III randomized trial. J Clin Oncol 2002;20:3578-85.
58. Nachman JB, Sposto R, Herzog P, et al. Randomised comparison of low-dose involved-field radiotherapy and no radiotherapy for children with Hodgkin's disease who achieve a complete response to chemotherapy. J Clin Oncol 2002;20:3765-71.
59. Ramanathan RK, Potter DM, Belani CP, et al. Randomised trial of influenza vaccine with granulocyte-macrophage colony-stimulating factor or placebo in cancer patients. J Clin Oncol 2002;20:4313-8.
60. Scagliotti GV, De Marinis F, Rinaldi M, et al. Phase III randomized trial comparing three platinum-based doublets in advanced non-small-cell lung cancer. J Clin Oncol 2002;20:4285-91.
61. Shepherd FA, Giaccone G, Seymour L, et al. Prospective, randomized, double-blind, placebo-controlled trial of marimastat after response to first-line chemotherapy in patients with small-cell lung cancer: a trial of the national cancer institute of Canada-clinical trials group and the European organization for research and treatment of cancer. J Clin Oncol 2002;20:4434-9.
62. Smith TJ, Staats PS, Deer T, et al. Randomised clinical trial of an implantable drug delivery system compared with comprehensive medical management for refractory cancer pain: impact on pain, drug-related toxicity, and survival. J Clin Oncol 2002;20:4040-9.

63. Saad F, Gleason DM, Murray R, et al. A randomized, placebo-controlled trial of zoledronic acid in patients with hormone-refractory metastatic prostate carcinoma. J Natl Cancer Inst 2002;94:1458-68.
64. Thomas DB, Gao DL, Ray RM, et al. Randomised trial of breast self-examination in Shanghai: final results. J Natl Cancer Inst 2002;94:1445-57.
65. Ball K, Berch DB, Helmers KF, et al. Effects of cognitive training interventions with older adults: a randomized controlled trial. JAMA 2002;288:2271-81.
66. Buchbinder R, Ptasznik R, Gordon J, Buchanan J, Prabaharan V, Forbes A. Ultrasound-guided extracorporeal shock wave therapy for plantar fasciitis: a randomized controlled trial. JAMA 2002;288:1364-72.
67. Burgio KL, Goode PS, Locher JL, et al. Behavioral training with and without biofeedback in the treatment of urge incontinence in older women: a randomized controlled trial. JAMA 2002;288:2293-9.
68. Gerritsen AA, de Vet HC, Scholten RJ, Bertelsmann FW, de Krom MC, Bouter LM. Splinting vs. surgery in the treatment of carpal tunnel syndrome: a randomized controlled trial. JAMA 2002;288:1245-51.
69. Hodnett ED, Lowe NK, Hannah ME, et al. Effectiveness of nurses as providers of birth labor support in North American hospitals: a randomized controlled trial. JAMA 2002;288:1373-81.
70. Randolph AG, Wypij D, Venkataraman ST, et al. Effect of mechanical ventilator weaning protocols on respiratory outcomes in infants and children: a randomized controlled trial. JAMA 2002;288:2561-8.
71. Steinhubl SR, Berger PB, Mann JT 3rd, et al. Early and sustained dual oral antiplatelet therapy following percutaneous coronary intervention: a randomized controlled trial. JAMA 2002;288:2411-20.
72. Weinberger M, Murray MD, Marrero DG, et al. Effectiveness of pharmacist care for patients with reactive airways disease: a randomized controlled trial. JAMA 2002;288:1594-602.
73. Ashton HA, Buxton MJ, Day NE, et al. The Multicentre Aneurysm Screening Study (MASS) into the effect of abdominal aortic aneurysm screening on mortality in men: a randomised controlled trial. Lancet 2002;360:1531-9.
74. Boerma D, Rauws EA, Keulemans YC, et al. Wait-and-see policy or laparoscopic cholecystectomy after endoscopic sphincterotomy for bile-duct stones: a randomised trial. Lancet 2002;360:761-5.
75. Bonnefoy E, Lapostolle F, Leizorovicz A, et al. Primary angioplasty versus prehospital fibrinolysis in acute myocardial infarction: a randomised study. Lancet 2002;360:825-9.
76. Dickstein K, Kjekshus J. Effects of losartan and captopril on mortality and morbidity in high-risk patients after acute myocardial infarction: the OPTIMAAL randomised trial. Optimal Trial in Myocardial Infarction with Angiotensin II Antagonist Losartan. Lancet 2002;360:752-60.
77. Fox KA, Poole-Wilson PA, Henderson RA, et al. Interventional versus conser-

vative treatment for patients with unstable angina or non-ST-elevation myocardial infarction: the British Heart Foundation RITA 3 randomised trial. Randomised Intervention Trial of unstable Angina. Lancet 2002;360:743-51.
78. MAGIC Study Group. Early administration of intravenous magnesium to high-risk patients with acute myocardial infarction in the Magnesium in Coronaries (MAGIC) Trial: a randomised controlled trial. Lancet 2002;360:1189-96.
79. Molyneux A, Kerr R, Stratton I, et al. International Subarachnoid Aneurysm Trial (ISAT) of neurosurgical clipping versus endovascular coiling in 2143 patients with ruptured intracranial aneurysms: a randomised trial. Lancet 2002;360:1267-74.
80. Mwinga A, Nunn A, Ngwira B, et al. Mycobacterium vaccae (SRL172) immunotherapy as an adjunct to standard antituberculosis treatment in HIV-infected adults with pulmonary tuberculosis: a randomised placebo-controlled trial. Lancet 2002;360:1050-5.
81. O'Grady JG, Burroughs A, Hardy P, Elbourne D, Truesdale A. Tacrolimus versus microemulsified ciclosporin in liver transplantation: the TMC randomised controlled trial. Lancet 2002;360:1119-25.
82. Shepherd J, Blauw GJ, Murphy MB, et al. Pravastatin in elderly individuals at risk of vascular disease (PROSPER): a randomised controlled trial. Lancet 2002;360:1623-30.
83. Singh RB, Dubnov G, Niaz MA, et al. Effect of an Indo-Mediterranean diet on progression of coronary artery disease in high risk patients (Indo-Mediterranean Diet Heart Study): a randomised single-blind trial. Lancet 2002;360:1455-61.
84. SoS Investigators. Coronary artery bypass surgery versus percutaneous coronary intervention with stent implantation in patients with multivessel coronary artery disease (the Stent or Surgery trial): a randomised controlled trial. Lancet 2002;360:965-70.
85. Sulo J, Chimpeni P, Hatcher J, et al. Chlorproguanil-dapsone versus sulfadoxine-pyrimethamine for sequential episodes of uncomplicated falciparum malaria in Kenya and Malawi: a randomised clinical trial. Lancet 2002;360:1136-43.
86. Van Damme L, Ramjee G, Alary M, et al. Effectiveness of COL-1492, a nonoxynol-9 vaginal gel, on HIV-1 transmission in female sex workers: a randomised controlled trial. Lancet 2002;360:971-7.
87. de Gans J, van de Beek D. Dexamethasone in adults with bacterial meningitis. N Engl J Med 2002;347:1549-56.
88. Fried MW, Shiffman ML, Reddy KR, et al. Peginterferon alfa-2a plus ribavirin for chronic hepatitis C virus infection. N Engl J Med 2002;347:975-82.
89. Gill TM, Baker DI, Gottschalk M, Peduzzi PN, Allore H, Byers A. A program to prevent functional decline in physically frail, elderly persons who live at home. N Engl J Med 2002;347:1068-74.
90. Harding GK, Zhanel GG, Nicolle LE, Cheang M. Antimicrobial treatment in diabetic women with asymptomatic bacteriuria. N Engl J Med 2002;347:1576-83.
91. Holmberg L, Bill-Axelson A, Helgesen F, et al. A randomized trial comparing

radical prostatectomy with watchful waiting in early prostate cancer. N Engl J Med 2002;347:781-9.
92. Hulscher JB, van Sandick JW, de Boer AG, et al. Extended transthoracic resection compared with limited transhiatal resection for adenocarcinoma of the esophagus. N Engl J Med 2002;347:1662-9.
93. Hurlen M, Abdelnoor M, Smith P, Erikssen J, Arnesen H. Warfarin, aspirin, or both after myocardial infarction. N Engl J Med 2002;347:969-74.
94. Konstantinides S, Geibel A, Heusel G, Heinrich F, Kasper W. Heparin plus alteplase compared with heparin alone in patients with submassive pulmonary embolism. N Engl J Med 2002;347:1143-50.
95. Koutsky LA, Ault KA, Wheeler CM, et al. A controlled trial of a human papillomavirus type 16 vaccine. N Engl J Med 2002;347:1645-51.
96. Lassen MR, Borris LC, Nakov RL. Use of the low-molecular-weight heparin reviparin to prevent deep-vein thrombosis after leg injury requiring immobilization. N Engl J Med 2002;347:726-30.
97. Stanberry LR, Spruance SL, Cunningham AL, et al. Glycoprotein-D-adjuvant vaccine to prevent genital herpes. N Engl J Med 2002;347:1652-61.
98. Thakar R, Ayers S, Clarkson P, Stanton S, Manyonda I. Outcomes after total versus subtotal abdominal hysterectomy. N Engl J Med 2002;347:1318-25.
99. Weaver LK, Hopkins RO, Chan KJ, et al. Hyperbaric oxygen for acute carbon monoxide poisoning. N Engl J Med 2002;347:1057-67.
100. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276:637-9.
101. Hall JC, Hall JL. Baseline comparison in surgical trials. ANZ J Surg 2002;72:567-9.
102. Burgess DC, Gebski VJ, Keech AC. Baseline data in clinical trials. Med J Aust 2003;179:105-7.
103. Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. Stat Med 1992;11:1685-704.
104. Ford I, Norrie J. The role of covariates in estimating treatment effects and risk in long-term clinical trials. Stat Med 2002;21:2899-908.
105. Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials?. Control Clin Trials 1998;19:248-56.
106. Steyerberg EW, Bossuyt PMM, Lee KL. Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics?. Am Heart J 2000;139:745-51.
107. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. International Statistical Review 1991;58:227-40.
108. Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. Control Clin Trials 2000;21:330-42.
109. Hernández AV, Steyerberg EW, Habbema JDF. Covariate adjustment in RCTs with dichotomous outcomes increased statistical power and reduced sample size requirememts. J Clin Epidemiol 2004;57:454-60.

110. Moreira ED Jr, Stein Z, Susser E. Reporting on methods of subgroup analysis in clinical trials: a survey of four scientific journals. Braz J Med Biol Res 2001;34:1441-6.
111. Sacks FM, Tonkin AM, Shepherd J, et al. Effect of pravastatin on coronary disease events in subgroups defined by coronary risk factors: the prospective pravastatin pooling project. Circulation 2000;102:1893-1900.
112. Wedel H, DeMets D, Deedwania P, et al. Challenges of subgroup analyses in multinational clinical trials: Experiences from the MERIT-HF trial. Am Heart J 2001;142:502-11.
113. Furberg CD, Vittinghoff E, Davidson M, et al. Subgroup interactions in the Heart and Estrogen/Progestin Replacement Study. Lessons learned. Circulation 2002;105:917-22.
114. Bouzamondo A, Hulot J-S, Sanchez P, Lechat P. Beta-blocker benefit according to severity of heart failure. Eur J Heart Fail 2003;5:281-9.
115. DeMets DL, Califf RM. Lessons learned from recent cardiovascular clinical trials: Part I. Circulation 2002;106:746-51.
116. Oxman AD, Guyatt GH. A Consumer's guide to subgroup analyses. Ann Intern Med 1992;116:78-84.
117. Adams KF Jr. Post hoc subgroup analysis and the truth of a clinical trial. Am Heart J 1998;136:753-8.
118. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991;266:93-8.
119. Brookes ST, Whitley E, Egger M, Davey Smith G, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol 2004;57:229-36.
120. Moyé LA, Deswal A. Trials within trials: Confirmatory subgroup analyses in controlled clinical experiments. Control Clin Trials 2001;22:605-19.
121. van der Windt DAWM, van Poppel MNM. Mistakes in methodology. IX. De interpretation of subgroup analyses [in Dutch]. Ned Tijdschr Geneeskd 1998;142:2245-7.
122. Wenzel V, Krismer AC, Arntz HR, Sitter H, Stadlbauer KH, Lindner KH, for the European Resuscitation Council Vasopressor during Cardiopulmonary Resuscitation Study Group. A comparison of vasopressin and epinephrine for out-of-hospital cardiopulmonary resuscitation. N Engl J Med 2004;350:105-13.
123. McIntyre KM. Vasopressin in asystolic cardiac arrest. N Engl J Med 2004;350:179-80.
124. Carson PA, O`Connor CM, Miller AB, et al. Circadian rhythm and sudden death in heart failure: results from the Prospective Randomized Amlodipine Survival Trial. J Am Coll Cardiol 2000;36:541-6.
125. Packer M, O`Connor CM, Ghali JK, et al. Effect of amlodipine on morbidity and mortality in severe chronic heart failure. Prospective Randomized Amlodipine Evaluation Study Group. N Engl J Med 1996;335:1107-14.

126. Parker AB, Naylor CD. Subgroups, treatment effects and baseline risks: some lessons from cardiovascular journals. Am Heart J 2000;139:952-61

127. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials. Comparison of protocols to published trials. JAMA 2004;291:2457-65.

Subgroup analyses in therapeutic cardiovascular clinical trials: Are most of them misleading?

3.2

ABSTRACT

Background: Treatment decisions in clinical cardiology are directed by results from randomized clinical trials (RCTs). We studied the appropriateness of the use and interpretation of subgroup analysis in current therapeutic cardiovascular RCTs.

Methods: We reviewed main reports of phase III cardiovascular RCTs, with at least 100 patients, published in 2002 and 2004, and from major journals (Circulation, JACC, Am Heart J, Am J Cardiol, N Engl J Med, Lancet, JAMA, BMJ, Ann Intern Med). Information on subgroups included: pre-specification, number, interaction test use, significant subgroups found, and emphasis on findings. We examined appropriateness of reporting, and differences according to sample size, overall trial result, and CONSORT (Consolidated Standards for Reporting of Trials) adoption.

Results: We selected 63 RCTs, with a median of 496 patients (range: 100–15245). Thirty-nine RCTs were reported with subgroup analyses, and 26 with >5 subgroups. No trial was specifically powered to detect subgroups effects, and only 14 RCTs were reported with fully pre-specified subgroups. Only 11 RCTs were reported with interaction tests. Furthermore, 21 RCTs were reported with claims of significant subgroups, and 15 with equal or more emphasis to subgroups than to the overall results. Subgroup analyses in large RCTs (>500 patients) were reported more often than in small ones (24/30 vs. 15/33; P=0.005). No differences were found according to overall result (positive/negative) or CONSORT adoption.

Conclusions: Subgroup analyses in recent cardiovascular RCTs were reported with several shortcomings, including a lack of pre-specification, and testing of a large number of subgroups, without the use of the statistically appropriate test for interaction. Reporting of subgroup analysis needs to be substantially improved, since emphasis on these secondary results may mislead treatment decisions.

Introduction

Treatment decisions in clinical cardiology are driven by results from randomized clinical trials (RCTs). As the number of drugs, devices, and treatment strategies has proliferated during the last two decades, so have the trials evaluating them [1]. Indeed, some of these RCTs have provoked major changes in clinical practice (e.g. GISSI-1 [2], SOLVD [3], 4S [4], NINDS [5], MERIT-HF [6], SIRIUS [7]).

A result of a RCT represents the effect of a particular intervention in an 'average' patient. However, clinicians treat individual patients with complex combinations of characteristics [8]. Thus, RCT results may not directly be applicable to one patient. Therefore, many investigators are interested to know treatment effects in specific subpopulations of patients ('subgroup analysis') [9]. Subgroups are based on patient characteristics measured before randomization [10]. Treatment effects in subgroups might offer clinicians more insight into treating individual patients.

Main RCT reports or separate publications have investigated effects of treatments in subgroups. For instance, a substantial number of separate publications have given special attention to gender subgroups in acute coronary syndromes [11-14] or heart failure [15-17]. Usually the overall treatment effect was consistent with that in subgroups. However, differential subgroup effects may be identified by chance, because multiple tests are performed in many subgroups, and by the use of inadequate methods [8, 10, 18].

Even though subgroup analyses are popular and occupy a large number of pages in the cardiovascular literature, little attention has been directed to their adequate reporting and interpretation. We aimed to review the appropriateness of reporting of subgroup analysis in RCTs recently published in major cardiology and internal medicine journals.

Methods

Selection of RCT reports

We handsearched therapeutic cardiovascular RCTs from journals with high impact factor in the ISI Journal Citation Report 2002 (categories cardiology and general internal medicine). Cardiology journals included Circulation, J Am Coll Cardiol, Am Heart J, and Am J Cardiol. General internal medicine journals included N Engl J Med, Lancet, JAMA, Br Med J, and Ann Intern Med. Two periods of 3 months were arbitrarily chosen: September 1 to November 30 2002, and May 1 to July 31 2004.

Therapeutic RCTs included interventions such as drugs, devices or therapeutic strategies. The diseases included acute coronary syndromes, left ventricular dysfunction, heart failure, arrhythmia, cardiac valve disease, hypertension, stroke, atherosclerosis of peripheral arteries (carotid, renal, lower limb), pulmonary emolism and venous thromboembolism, venous insufficiency, diabetes mellitus, and hypercholesterolemia. We included main reports of phase III parallel RCTs, with individual randomization, and we excluded cross-over trials, cluster-randomized trials, equivalence/non-inferiority trials, factorial trials, and trials with less than 100 patients. Because it was not the focus of our review, we excluded secondary trial reports designed to specifically address post-hoc hypotheses and subgroup comparisons.

Subgroup analysis reporting

Information about subgroup analysis was extracted, which was considered essential to evaluate the appropriateness of the method, and the validity of the results [19]. This information included: Pre-specification of subgroup analysis (as reported in methods), number of subgroup factors (patient baseline characteristics measured before randomization) [10], number of subgroup outcomes (primary or secondary endpoints), total number of subgroups (the product of the number of subgroup factors and the number of subgroup outcomes), statistical method (interaction test or separate analysis), number of significant subgroups found, and emphasis in subgroup findings (subgroup results mentioned in abstract and/or conclusion).

An interaction test directly assesses differences in treatment effect between complementary subgroups, and involves one statistical test irrespective of the number of subgroups [9, 19]. A significant interaction test means that the treatment effect in one subgroup is significantly different from the treatment effect in the other subgroup. In contrast, the separate subgroup method assesses treatment effects in each group independently, and it involves two or more tests. This is inappropriate from a statistical point of view [9].

For consistency in the gathering of information, one of us (A.V.H.) selected the RCT reports, checked the inclusion and exclusion criteria, and collected the information. Another of us (E.W.S.) independently assessed the same RCT reports, and was blinded to the assessments of the first reviewer. There were few differences, which were resolved by consensus.

Secondary analyses

We examined differences in reporting between relatively small (<500 patients)

and large trials (≥500 patients). We expected that researchers in large trials would feel more confident to search for differential treatment effects in subgroups. We also examined subgroup reporting between negative (non-significant overall result) and positive trials. Researchers might try to explore subgroup effects if the overall effect was non-significant. Finally, we examined differences according to adoption of the CONSORT (Consolidating Standards for Reporting Trials) Statement [20], which might improve reporting of subgroup analysis.

Statistical analysis

Descriptive statistics included percentages and medians. Categorical variables were tested with Chi-square tests. The software used was EXCEL and SPSS 10.0 (SPSS Inc., Chicago, USA, 1999).

Results

RCT reports

In total, 63 RCTs [21-83] satisfied the inclusion criteria (Table 1). Fifty-three (84%)

	Year 2002	Year 2004	Total
Number of RCTs	33	30	63
Median number of patients (range)	680 (103 – 11484)	477 (100 – 15245)	496 (100-15245)
RCTs with >500 patients (%)	17 (52)	13 (43)	30 (48)
Median time of follow up, months	6	6	6
Median number of centers (range)	34 (1 – 432)	4 (1 – 507)	20 (1 – 507)
RCTs with two arms (%)	27 (82)	26 (87)	53 (84)
Negative results (%)	13 (39)	20 (67)	33 (54)
CONSORT adoption (%)	9 (27)	9 (30)	18 (29)

RCTs denotes randomised clinical trials; CONSORT denotes Consolidated Standards of Reporting Trials guidelines.

Table 1. Characteristics of the randomized clinical trials included in this review.

RCTs had two arms. Two thirds (n= 42) of the RCTs evaluated drugs, and one third (n=21) devices or strategies. Half of the RCTs had more than 500 patients, and the follow-up ranged between 3 days and 4,5 years. Most of the RCTs were multicenter. Approximately half of the RCTs yielded negative overall results. One third of the RCTs were published in journals which adopted the CONSORT statement.

Subgroup reporting

Two-thirds (39 out of 63) of the RCTs were reported with subgroup analyses (Table 2), while 24 RCTs only performed crude analyses. According to what was described, full pre-specification of all subgroups was only done in 14 reports, and 4 others included both pre-specified and non-pre-specified subgroups. Subgroups in the other 21 RCTs were reported without a rationale, and these may represent after-analysis subgroups.

The median number of patient baseline characteristics used as subgroup factors was three. However, some RCTs included higher numbers, with a maximum of 23 [28]. Two thirds (26 out of 39) of the RCTs were reported with >5 subgroups. The main factors were: gender (n=20), age (n=16), comorbidities other than diabetes (n=15), severity of disease (n=13), and diabetes (n=11). Likewise, the median number of subgroup outcomes was two, and some reported as many as seventeen [48]. Therefore, the number of total subgroups (product of factors by outcomes) was high, ranging from 7 to 60 [55]. We found 63 significant subgroups among the 508 subgroups assessed (12%): 25 significant subgroups were pre-specified, and 38 were post-specified.

The interaction test was used in 11 out of 39 RCTs, and separate tests in 28. When interaction tests were used, the authors identified 10 (6%) significant subgroup effects among 175 tests performed. With separate subgroups, the authors identified 53 (16%) significant subgroup effects among 333 tests performed. Nearly half (21 out of 39) of the RCTs claimed differences between subgroup treatment effects and the average treatment effects, and showed them in their result sections. Moreover, fifteen reports emphasized their subgroup findings by highlighting them either in the main conclusions or in the abstract. Only 6 RCTs reports included the appropriate interaction test and did not put emphasis on subgroup results.

Secondary analysis

Large trials (n>500) more often reported subgroup analyses than small trials (24/30

Topic	n
Reported subgroup analysis	39
Pre-specified subgroup analysis (%)	
Fully	14 (36)
Partially	4 (10)
Median number of subgroup factors (range)	3 (1 – 23)
Median number of subgroup outcomes (range)	2 (1 – 17)
Median number of subgroups (range)	7 (1 – 60)
RCTs claimed subgroup effects (%)	21 (54)
using interaction tests	7/11
using separate subgroups	14/28
RCTs emphasized subgroup results in abstract or conclusions (%)	15 (39)
using interaction tests	5/11
using separate subgroup tests	10/28

RCTs denotes randomized clinical trial.

Table 2. Description of reporting of subgroup analysis.

vs. 15/33, p=0.005). However, we did not observe differences in pre-specification, number of subgroups, statistical methods or emphasis in results. Moreover, these features of subgroup reporting did not differ between CONSORT-adopting and non-adopting trials, and between negative and positive trials.

Discussion

We found several shortcomings in the use and reporting of subgroup analysis in recent cardiovascular RCTs from major general medicine and cardiology jour-

nals. Many subgroups, frequently not pre-specified, were reported. The test of interaction was underused, and subgroup findings were commonly emphasized. These shortcomings may mislead treatment decisions when particular subgroups of patients are being evaluated.

During the last two decades the number of RCTs in cardiology has dramatically increased: from 5,410 in the period 1980-1989 to 14,845 in the period 1990-2000 [1]. Simultaneously, the interest to explore treatment effects in subgroups of patients has increased [8, 84]. In the last two decades, many pages have been devoted to look at differential treatment effects in particular subgroups such as elderly people, female patients, diabetic patients, and severely-ill patients in diseases such as acute coronary syndromes, heart failure, hypertension, stroke, and hypercholesterolemia.

Although subgroups should be clearly defined in the RCT protocol ('pre-specified subgroups') [18], researchers may decide to perform subgroup analysis when the RCT is running ('pre-analysis subgroups') or when the RCT analysis has been completed ('after-analysis subgroups'). Frequently, the RCT reports do not clearly make this differenciation.

One review evaluated 67 large RCTs in unstable angina, myocardial infarction, left ventricular dysfunction or heart failure [84]. Some shortcomings in subgroup reporting were noticed: little supporting rationale, lack of pre-specification, overuse of separate analyses, and underuse of the formal statistical test of interaction. However, the authors limited the review to large (>1000 patients) drug RCTs, from 1980 to 1997. We included more recent RCTs (2002 and 2004), with medium size (two-thirds <1000 patients) and a broader perspective (evaluation of drugs, devices and strategies). Our review may reflect the scope of RCTs reviewed by cardiologists nowadays.

Two thirds of the RCTs in our study reported pre-analysis, after-analysis, or partially pre-specified subgroups. These findings may be related to the fact that investigators re-analyze data in response to preliminary results, results from other studies, and suggestions from reviewers. It is possible that investigators evaluated an unplanned, large number of subgroups: two thirds of the RCTs reported more than 5 subgroups. Remarkably, seven RCTs reported more than 25 subgroups [29, 30, 37, 39, 55, 61, 79]. Three of these RCTs [30, 55, 79] found a high number of statistically significant subgroups, ranging from 5 to 18. Moreover, the proportion of significant subgroups was higher than the 5% expected (Z test= 4.19, $p<0.01$), especially for separate tests. Most of these subgroups were based on non-prognostic baseline characteristics, and may hence be considered

with suspicion [85, 86]. Pre-specification of a limited number of subgroups based on predictive baseline characteristics may decrease the probability of spurious subgroup effects (false positives) [19, 20, 84, 86].

Interaction testing is the appropriate method to analyze subgroups [8, 9, 19, 20, 86]. Remarkably, interaction tests were only used in around 30% of cardiovascular RCTs which reported subgroups. This method decreases the risk to find false positive subgroups, but its power to detect true subgroups is low [20, 86]. For instance, in a trial with 80% power for the overall treatment effect, the interaction test has at most 29% power to detect an interaction effect of the same magnitude of the overall effect [9]. The sample size should be increased at least four-fold to achieve the same power [10]. Likewise, the evaluation of treatment effects in separate groups is misleading since false positive subgroup effects can be found: A significant effect in one subgroup can be expected in 7%, when the observed overall effect was nonsignificant [9].

The interpretation of subgroups is important for treatment decisions in cardiology [84, 85]. Thus, some treatments may be withhold for some patients who require them, and some other patients may be treated with drugs that are not needed [84]. For instance, aspirin was studied in few women in primary and secondary prevention RCTs of coronary heart disease, and a non-significant effect of aspirin was found in subgroup analysis in women [87]. An overinterpretation of this subgroup may have led to the undertreatment of women for years, although we now know that aspirin is effective in women. Other classic examples include the use of thrombolytic therapy and beta blockade only in patients with anterior myocardial infarctions [18].

Half of the cardiovascular RCTs claimed significant subgroup effects, and 40% gave equal emphasis to subgroup and overall results. This is worrying, because subgroup analysis is a secondary, hypothesis-generating exercise to stimulate further research [18, 84, 86]. Thus, less emphasis should be placed on subgroup results. The best estimate of treatment effect to be expected for a patient treated outside the trial may be still the overall effect [18]. Rather, subgroup analyses should measure the consistency of beneficial and harmful treatment effects across different risk and demographic groups. This was shown in recent RCTs on beneficial treatment effects for heart failure, coronary intervention, coronary reperfusion and secondary prevention with statins, and on harmful effects of inotropic drugs for heart failure and oral glycoprotein IIb/IIIa inhibitors for unstable angina [8].

An example of satisfactory reporting and interpretation of subgroup analysis is the MATCH trial [57]. This RCT compared the use of aspirin and clopidogrel

versus clopidogrel alone after ischemic stroke or transient ischemic attack in high–risk patients (n=7599). The overall treatment effect was not significant. The authors used 13 pre-specified patient baseline characteristics to define subgroups. They used interaction tests to explore differential treatment effects by baseline characteristics on one composite outcome, and found that the treatment effect differed by patient age (p=0.012). They considered this finding with scepticism, and showed the overall effect as the main result.

On the other hand, the ACST trial reported its subgroup analyses less satisfactorily [61]. The authors evaluated the prevention of disabling and fatal strokes by successful carotid endarterectomy in patients without recent neurological symptoms (n=3120). They extensively reported 27 partially pre-specified subgroup analyses, based on 9 baseline characteristics and 3 outcomes. However, it is not clear how many analyses really were performed. Moreover, they analyzed these subgroups independently, resulting in 72 p-values. Although the overall effect was positive, three subgroup findings were emphasized: effects in women, in patients with hypercholesterolemia, and in patients older than 75 years. These subgroups were of limited sample size and had few events. In particular, the investigators concluded that both men and women had a significant benefit with carotid endarterectomy on the risk of stroke, although the effect in women was 'not as definite as in men' [61].

We summarize some recommendations to appropriately perform and interpret subgroup analysis in cardiovascular RCTs, based on current recommendations 20 and methodologic papers (Table 3) [8-10, 18, 19, 84, 85, 86]. Apart from looking at subgroup results of prior RCTs and independent subgroup confirmation, meta-analyses can be used to study subgroups. Meta-analyses have higher power to detect significant subgroup effects in comparison to individual RCTs [9], and hence give readers more confidence to believe in a potential subgroup. Nice examples are readily available from the literature [88-90]. For instance, a small meta-analysis of the results of the ACST and the ACAS trials has highlighted a significant effect of endarterectomy in men and a nonsignificant effect in women on the combined risk of any stroke and operative death [90]. If a subgroup is still not completely reliable and there is a special interest on it, probably the best, but less feasible alternative is to design a RCT specifically targeted to study the subgroup. One example is to design a confirmatory subgroup analysis within a RCT, which aims to find explicit effects of interventions in a prospectively defined subgroup stratum of interest [91]. This method usually defines a subgroup sample size between 40 to 60% of the total sample size [86, 91].

Our paper has some limitations. We reviewed a limited number of RCTs. However, we used a broad sample of major journals, which publish the most

Specification
Specify a limited number of subgroups in advance, with a clear rationale.
Analysis
Use statistical interaction tests in the full RCT population.
Interpretation
a. Be sceptical if subgroups were not pre-specified, not biologically plausible or no interaction tests were applied.
b. Interpret in context, e.g. look at prior findings, and independent confirmation.
c. See subgroup analysis as a hypothesis-generating exercise to stimulate further research.
d. Put emphasis on overall results, which may be considered better estimates of treatment effects than the subgroup effects.

Table 3. Suggestions to appropriately perform and interpret subgroup analysis.

influential RCTs in cardiology. Moreover, our main aim was to highlight the appropriateness of use and interpretation of subgroup analysis, and this sample may well be considered representative of current practice. Finally, we only used information that was published in the RCT reports. The information provided in the reports may not necessarily be the same as in the RCT protocol, nor comprise all analyses that were performed.

In conclusion, subgroup analyses in recent cardiovascular RCTs were reported with several shortcomings, such as a lack of pre-specification, and testing of a large number of subgroups, without the use of the statistically appropriate test for interaction. Reporting of subgroup analysis needs to be substantially improved in cardiovascular RCTs, since emphasis in these secondary results may mislead treatment decisions for particular groups of patients.

References

1. Lader EW, Cannon CP, Ohman EM, et al. The clinician as investigator. Participating in clinical trials in the practice setting. Circulation 2004; 109: 2672-9.
2. Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. Lancet 1986; 1: 397-402.
3. The SOLVD investigators. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. N Engl J Med 1991; 325: 293-302.
4. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). Lancet 1994; 344: 1383-9.
5. Tissue plasminogen activator for acute ischemic stroke: the National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. N Engl J Med 1995; 333: 1581-7.
6. Effect of metoprolol CR/XL in chronic heart failure: metoprolol CR/XL randomised intervention trial in congestive heart failure (MERIT-HF). Lancet 1999; 353: 2001-7.
7. Moses JW, Leon MB, Popma JJ, et al. Sirolimus-eluting stents versus standard stents in patients with stenosis in a native coronary artery. N Engl J Med 2003; 349: 1315-23.
8. DeMets DL, Califf RM. Lessons learned from recent cardiovascular clinical trials: Part I. Circulation 2002; 106: 746-51.
9. Brookes ST, Whitely E, Egger M, Davey Smith G, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol 2004; 57: 229-36.
10. Pieper KS, Tsiatis AA, Davidian M, et al. Differential treatment benefit of platelet glycoprotein IIb/IIIa inhibition with percutaneous coronary intervention versus medical therapy for acute coronary syndromes: Exploration of methods. Circulation 2004; 109: 641-6.
11. Lincoff AM, Califf RM, Ellis SG, et al. Thrombolytic therapy for women with myocardial infarction: Is there a gender gap?. J Am Coll Cardiol 1993; 22: 1780-7.
12. Hochman JS, McCabe CH, Stone PH, et al. Outcome and profile of women and men presenting with acute coronary syndromes: A report of the TIMI IIIB. J Am Coll Cardio 1997; 30: 141-8.
13. Jacobs AK, Kelsey SF, Brooks MM, et al. Better outcome for women compared with men undergoing coronary revascularization. A report from the Bypass Angioplasty Revascularization Investigation (BARI). Circulation 1998; 98: 1279-85.
14. Glaser R, Herrman HC, Murphy SA, et al. Benefit of an early invasive management strategy in women with acute coronary syndromes. JAMA 2002; 288: 3124-9.
15. Ghali JK, Pina IL, Gottlieb SS, Deedwania PC, Wikstrand JC. Metoprolol CR/XL in female patients with heart failure. Analysis of the experience in Metoprolol

Extended-Release Randomised Intervention Trial in Heart Failure (MERIT-HF). Circulation 2002; 105: 1585-91.
16. Rathore SS, Wang Y, Krumholz HM. Sex-based differences in the effect of digoxin for the treatment of heart failure. N Engl J Med 2002; 347: 1403-11.
17. Ghali JK, Krause-Steinrauf HJ, Adams KF, et al. Gender differences in advanced heart failure: insights from the BEST study. J Am Coll Cardiol 2003; 42: 2128-34.
18. Sleight P. Debate: Subgroup analyses in clinical trials- fun to look at, but don't believe them!. Curr Control Trials Cardiovasc Med 2000; 1: 25-7.
19. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355: 1064-9.
20. Altman DG, Schultz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001; 134: 663-94.
21. Lassen MR, Borris LC, Nakov RL. Use of low-molecular-weight heparin reviparin to prevent deep venous thrombosis after leg injury requiring immobilization. N Engl J Med 2002; 347: 726-30.
22. Hurlen M, AbdelnoorM, Smith P, Erikssen J, Arnesen H. Warfarin, aspirin or both after myocardial infarction. N Engl J Med 2002; 347: 969-74.
23. Konstantinides S, Geibel A, Heusel G, Heinrich F, Kasper W, for the Management Strategies and Prognosis of Pulmonary Embolism-3 Trial Investigators. N Engl J Med 2002; 347: 1143-50.
24. Fox KAA, Poole-Wilson PA, Henderson RA et al. Interventional versus conservative treatment for patients with unstable angina or non-ST-elevation myocardial infarction: the British Heart Foundation RITA 3 randomised trial. Lancet 2002; 360: 743-51.
25. Dickstein K, Kjekshus J, and the OPTIMAAL Sterering Committee, for the OPTIMAAL Study Group. Effects of losartan and captopril on mortality and morbidity in high-risk patients after acute myocardial infarction: the OPTIMAAL randomised trial. Lancet 2002; 360: 752-60.
26. Bonnefoy E, Lapostolle F, Leizorovicz A, et al. Primary angioplasty versus prehospital fibrinolysis in acute myocardial infarction: a randomised study. Lancet 2002; 360; 825-9.
27. The SoS Investigators. Coronary artery bypass surgery versus percutaneous coronary intervention with stent implantation in patients with multivessel coronary artery disease (the Stent or Surgery trial): a randomised controlled trial. Lancet 2002; 360: 965-70.
28. The Magnesium in Coronaries (MAGIC) Trial Investigators. Early administration of intravenous magnesium to high-risk patients with acute myocardial infarction in the Magnesium in Coronaries (MAGIC) Trial: a randomised controlled trial. Lancet 2002; 360: 1189-96.
29. Shepherd J, Blauw GJ, Murphy MB, et al. Pravastatin in elderly individuals at

risk of vascular disease (PROSPER): a randomised controlled trial. Lancet 2002; 360: 1623-30.
30. Steinhubl SR, Berger PB, Mann III JT, et al. Early and sustained dual oral antiplatelet therapy following percutaneous coronary intervention. A randomized controlled trial. JAMA 2002; 288: 2411-20.
31. Meade T, Zuhrie R, Cook C, Cooper J, on behalf of the MRC General Practice Research Framework. Benafibrate in men with lower extremity arterial disease: ramdomised controlled trial. BMJ 2002; 325: 1139.
32. Francis CW, Davidson BL, Berkowitz SD, et al. Ximelagatran versus warfarin for the prevention of venous thromboembolism after total knee arthroplasty. A randomized, double-blind trial. Ann Intern Med 2002; 137: 648-55.
33. Holmes DR, Savage M, LaBlanche J-M, et al. Results of Prevention of rest enosis with Tranilast and its Outcomes (PRESTO) trial. Circulation 2002; 106: 1243-50.
34. Stone AFM, Mendal MA, Kaski JC, et al. Effect of treatment for Chlamydia pneumoniae and Helicobacter pylori on markers of inflammation and cardiac events in patients with acute coronary syndromes. South Thames Trial on Antibiotics in Myocardial Infarction and Unstable Angina (STAMINA). Circulation 2002; 106: 1219-23.
35. Hodis HN, Mack WJ, LaBree L, et al. Alpha-thocopherol supplementation in healthy individuals reduces low-density lipoprotein lipoprotein oxidation but not atherosclerosis. The Vitamin E Atherosclerosis Prevention Study (VEAPS). Circulation 2002; 106: 1453-9.
36. Taylor AJ, Kent SM, Flaherty PJ, Coyle LC, Markwood TT, Vernalis MN. ARBITER: Arterial Biology for the Investigation of the Treatment Effects of Reducing Cholesterol. A randomized trial comparing the effects of atorvastatin and pravastatin on carotid intima medial thickness. Circulation 2002; 106: 2055-66.
37. Packer M, Fowler MB, Roecker EB, et al. Effect of carvedilol on the mortality of patients with severe chronic heart failure. Results of the Carvedilol Prospective Randomized Cumulative Survival (COPERNICUS) Study. Circulation 2002; 106: 2194-9.
38. de Jongh S, Ose L, Szamosi T, et al. Efficacy and safety of statin therapy in children with familial hypercholesterolemia. A randomized, double-blind, placebo-controlled trial with simvastatin. Circulation 2002; 106: 2231-7.
39. Zanchetti A, Bond MG, Hennig M, et al. Calcium antagonist lacidipine slows down progression of asymptomatic carotid atherosclerosis. Principal results of the European Lacidipine Study on Atherosclerosis (ELSA), a randomized, double-blind, long-term trial. Circulation 2002; 106: 2422-7.
40. Deanfield JE, Detry J-M, Sellier P, et al. Medical treatment of myocardial ischemia in coronary artery disease: effect of drug regime and irregular dosing in the CAPE II trial. J Am Coll Cardiol 2002; 40: 917-25.
41. Faxon DP, Gibbons RJ, Chronos NAF, Gurbel PA, Sheehan F, for the HALT-MI

Investigators. The effect of blockade of the CD11/CD18 integrin receptor on infarct size in patients with acute myocardial infarction treated with direct angioplasty: the results of the HALT-MI study. J am Coll Cardiol 2002; 40: 1199-204.
42. White WB, Sica DA, Calhoum D, Mansoor GA, Anders RJ. Preventing increases in early-morning blood pressure, heart rate, and the rate-pressure product with controlled onset of extended release verapamil at bedtime versus enalapril, losartan, and placebo on arising. Am Heart J 2002; 144: 657-65.
43. Elbaz M, El Mokhtar E, Khalifé K, et al. Is direct coronary stenting the best strategy for long-term outcome?. Results of the multicentric randomized benefit evaluation of direct coronary stenting (BET) study. Am Heart J 2002; 144: e7.
44. Allen JK, Blumenthal RS, Margolis S, Young DR, Miller III, ER, Kelly K. Nurse management of hypercholesterolemia in patients with coronary heart disease: results of a randomized clinical trial. Am Heart J 2002; 144: 678-86.
45. Mitchell EG, Stoddard MF, Ben-Yehuda O, et al. Esmolol in acute ischemic syndromes. Am Heart J 2002; 144: e9.
46. Hoffmann R, Herrmann G, Silber S, et al. Randomised comparison of success and adverse event rates and cost effectiveness of one long versus two short stents for treatment of long coronary narrowings. Am J Cardiol 2002; 90: 460-4.
47. Tsuchikane E, Kobayashi T, Kobayashi T, et al. Debulking and stenting versus debulking only of coronary artery disease in patients treated with cilostazol (final results of ESPRIT). Am J Cardiol 2002; 90: 573-8.
48. Schaefer EJ, McNamara JR, Taylor T, et al. Effects of atorvastatin on fasting and postprandial lipoprotein subclasses in coronary heart disease patients versus control patients. Am J Cardiol 2002; 90: 689-96.
49. Whelton A, White WB, Bello AE, Puma JA, Fort JG, and the SUCCESS-VII Investigators. Effects of celecoxib and rofecoxib on blood pressure and edema in patients ≥65 years of age with systemic hypertension and osteoarthritis. Am J Cardiol 2002; 90: 959-63.
50. Freed MI, Ratner R, Marcovina SM, et al. Effects of rosiglitazone alone and in combination with atorvastatin on the metabolic abnormalities in type 2 diabetes. Am J Cardiol 2002; 90: 947-52.
51. Dujovne CA, Ettinger MP, McNeer F et al. Efficacy and safety of a potent new selective cholesterol absorption inhibitor, exetimibe, in patients with primary hypercholesterolemia. Am J Cardiol 2002; 90: 1092-7.
52. Gagné C, Bays HE, Weiss SR, et al. Efficacy and safety of ezetimibe added to ongoing statin therapy for treatment of patients with primary hypercholesterolemia. Am J Cardiol 2002; 90: 1084-91.
53. Mauri L, Bonan R, Weiner BH, et al. Cutting balloon angioplasty for the prevention of restenosis: Results of the Cutting Balloon Global Randomised Trial. Am J Cardiol 2002; 90: 1079-83.
54. Lange H, Suryapranata H, De Luca G, et al. Folate therapy and in-stent restenosis after coronary stenting. N Engl J Med 2004; 350: 2673-81.

55. Bristow MR, Saxon LA, Boehmer J, et al. Cardiac-resyncronization therapy with or without an implantable defribillator in advanced chronic heart failure. N Engl J Med 2004; 350: 2140-50.
56. Kadish A, Dyer A, Daubert JP, et al. Prophylactic defribillator implantation in patients with nonischemic dilated cardiomiopathy. N Engl J Med 2004; 350: 2151-8.
57. Diener H-C, Bogousslavsky J, Brasss LM, et al. Aspirin and clopidogrel compared with clopidogrel alone after recent ischaemic stroke or transient ischemic attack in high-risk patients in high-risk patients (MATCH): randomised, double-blind, placebo-controlled trial. Lancet 2004; 364: 331-7.
58. Anand I, McMurray J, Cohn JN, et al. Long term effects of danusentan on left-ventricular remodelling and clinical outcomes in the EndothelinA Receptor Antagonist Trial in Heart Failure (EARTH): randomised, double-blind, placebo-controlled trial. Lancet 2004; 364: 347-54.
59. Julius S, Kjeldsen SE, Weber M, et al. Outcomes in hypertensive patients at high cardiovascular risk treated with regimens based on valsartan or amlodipine: the VALUE randomised trial. Lancet 2004; 363: 2022-31.
60. Barwell JR, Davies CE, Deacon J, et al. Comparison of surgery and compression with compression alone in chronic venous ulceration (ESCHAR study): randomised controlled trial. Lancet 2004; 363: 1854-9.
61. MRC Asymptomatic Carotid Surgery Trial (ACST) Collaborative Group. Prevention of disabling and fatal strokes by successful carotid endarterectomy in patients without recent neurological symptoms: randomised controlled trial. Lancet 2004; 363: 1491-502.
62. Verrier ED, Sherman SK, Taylor KM, et al. Terminal complement blockade with pexelizumab during coronary artery bypass graft surgery requiring cardiopulmonary bypass. A randomized trial. JAMA 2004; 291: 2319-27.
63. Wiegman A, Hutten BA, de Grootm E, et al. Efficacy and safety of statin therapy in children with familial hypercholesterolemia. A randomized controlled trial. JAMA 2004; 292: 331-7.
64. The SYNERGY investigators. Enoxaparin vs. unfractionated heparin in high-risk patients with non-ST-segment elevation acute coronary syndromes managed with an intended early invasive strategy. Primary results of the SYNERGY randomized trial. JAMA 2004; 292: 45-54.
65. Yancy WS, Olsen MK, Guyton JR, Bakst RP, Westman EC. A low-carbohydrate, ketogenic diet versus low-fat diet to treat obesity and hyperlipidemia. Ann Intern Med 2004; 140: 769-77.
66. Lockhart PB, Brennan MT, Kent ML, Norton HJ, Weinrib DA. Impact of amoxicillin prophylaxis on the incidence, nature and duration of bacteremia in children after intubation and dental procedures. Circulation 2004; 109: 2878-84.
67. Bilgin YM, van de Watering LMG, Eijsman L, et al. Double-blind, randomized controlled trial on the effect of leucocyte-depleted erythrocyte transfusions in cardiac valve surgery. Circulation 2004; 109: 2755-60.

68. Sutton AGC, Campbell PG, Graham R, et al. A randomized trial of rescue angioplasty versus a conservative approach for failed fibrinolysis in ST-segment elevation myocardial infarction. The Middlesbrough Early Revascularization to Limit INfarction (MERLIN) trial. J Am Coll Cardiol 2004; 44: 287-96.
69. Valgimigli M, Percoc G, Barbieri D, et al. The additive value of tirofiban administered with the high-dose bolus in the prevention of ischemic complications during high-risk coronary angioplasty, The ADVANCE trial. J Am Coll Cardiol 2004; 44: 14-9.
70. van der Heijden DJ, Westerdorp ICD, Riezebos RK, et al. Lack of efficacy of clopidogrel pre-treatment in the prevention of myocardial damage after elective stent implantation. J Am Coll Cardiol 2004; 44: 20-4.
71. Besterhorn H-P, Neumann F-J, Buttner HJ, et al. Evaluation of the effect of oral verapamil on clinical outcome and angiographic restenosis after percutaneous coronary intervention. The randomized, double-blind, placebo-controlled, multicenter Verapamil Slow-Release for Prevention of Cardiovascular Event After Angioplasty (VESPA) trial. J Am Coll Cardiol 2004; 43: 2160-5.
72. Hueb W, Soares PR, Gersh BJ, et al. The medicine, angioplasty, or surgery study (MASS-II): a randomized, controlled clinical trial of three therapeutic strategies for multivessel coronary artery disease. One-year results. J Am Coll Cardiol 2004; 43: 1743-51.
73. Gattis WA, O'Connor CM, Gallup DS, Haseelblad V, Gheorghiade M, on behalf of the IMPACT-HF Investigators and Coordinators. Predischarge initiation of carvedilol in patients hospitalized for decompensated heart failure. Results of the Initiation Management Predischarge: Process fro Assessment of Carvedilol Therapy in Heart Failure (IMPACT-HF) trial. J Am Coll Cardiol 2004; 43: 1534-41.
74. Hillis Gs, Pearson CV, Harding SA, et al. Effects of a brief course of azithromycin on soluble cell adhesion molecules and markers of inflammation in survivors of an acute coronary syndrome: a double-blind, randomized, placebo-controlled study. Am Heart J 2004; 148: 72-9.
75. Thompson PL, Meredith I, Amerena J, et al. Effect of pravastatin compared with placebo initiated withing 24 hours of onset of acute myocardial infarction or unstable angina: The Pravastatin in Acute Coronary Treatment (PACT) trial. Am Heart J 2004; 148: e2.
76. Pache J, Kastrati A, Mehilli J, et al. A randomized evaluation of the effects of glucose-insulin-potassium infusion on myocardial salvage in patients with acute myocardial infarction treated with reperfusion therapy. Am Heart J 2004; 148: e3.
77. Cleland JGF, Findlay I, Jafri S, et al. The Warfarin/Aspirin Study in Heart Failure (WASH): A randomized trial comparing antithrombotic strategies for patients with heart failure. Am Heart J 2004; 148: 157-64.
78. Rahel BM, Suttorp MJ, Laarman GJ, et al. Primary stenting of occluded native coronary arteries: Final results of the Primary Stenting of Occluded Native Coronary Arteries (PRISON) study. Am Heart J 2004; 147: e22.

79. Yu C-M, Li L S-W, Lam M-F, Siu D C-W, Miu R K-M, Lau C-P. Effect of a cardiac rehabilitation program on left ventricular diastolic function and its relationship to exercise capacity in patients with coronary heart disease: Experience from a randomized, controlled study. Am Heart J 2004; 147: e24.
80. Danzi GB, Sesana M, Capuano C, Mauri L, Centurini PB, Baglini R. Comparison in patients having primary coronary angioplasty of abciximab versus tirofiban on recovery of left ventricular function. Am J Cardiol 2004; 94: 35-9.
81. Sick PB, Gelbrich G, Kalnins U, et al. Comparison of early and late results of a carbofilm-coated stent versus a pure high-grade stainless steel stent (the Carbostent-Trial). Am J Cardiol 2004; 93: 1351-6.
82. Hanekamp C, Koolen J, Bonnier H, et al. Randomized comparison of balloon angioplasty versus silicon carbon-coated stent implantation for de novo lesions in small coronary arteries. Am J Cardiol 2004; 93: 1233-7.
83. Lincoff AM, Bittl JA, Kleiman NS, Sarembock IJ, et al. Comparison of bivalirudin versus heparin during percutaneous coronary intervention (the Randomised Evaluation of PCI Linking Angiomax to Reduced Clinical Events [REPLACE]-1 trial). Am J Cardiol 2004; 93: 1092-6.
84. Parker AB, Naylor CD. Subgroups, treatment effects, and baseline risks: some lessons from major cardiovascular trials. Am Heart J 2000; 139: 952-61.
85. Adams KF Jr. Post hoc subgroup analysis and the truth of a clinical trials. Am Heart J 1998; 136: 753-8.
86. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991; 266: 93-8
87. Shaw LJ, Miller DD, Romeis JC, et al. Gender differences in the non-invasive evaluation and management of patients with suspected coronary artery disease. Ann Intern Med 1994; 120: 559-66.
88. Gueyffier F, Bulpitt C, Boissel J-P, et al. Antihypertensive drugs in very old people: a subgroup meta-analysis of randomised controlled trials. Lancet 1999; 353: 793-96.
89. Rothwell PM, Eliasziw M, Gutnikov SA, Warlow CP, Barnett HJM, for the Carotid Endarterectomy Trialists Collaboration. Endarterectomy for symptomatic carotid stenosis in relation to clinical subgroups and timing of surgery. Lancet 2004; 363: 915:24.
90. Rothwell PM. ACST: which subgroups will benefit most from carotid endarterectomy. Lancet 2004; 364: 1122-3.
91. Moyé LA, Deswal A. Trials within trials: Confirmatory subgroup analyses in controlled clinical experiments. Control Clin Trials 2001; 22: 605-19.

Subgroup analysis and covariate adjustment in randomized clinical trials of traumatic brain injury: A systematic review

3.3

ABSTRACT

Objective: Few randomized clinical trials (RCTs) in the field of traumatic brain injury (TBI) have shown a significant treatment benefit. We aimed to critically review the uses of two types of secondary analyses, covariate adjustment and subgroup analysis, which may be common in TBI trials.

Methods: We performed a systematic review of therapeutic phase III RCTs, including acute, adult, moderate-to-severe TBI patients. Glasgow Outcome Scale (GOS) at ≥3 months as outcome, and ≥50 patients per arm were required. We compared the actual reporting of covariate adjustment and subgroup analyses with the CONSORT (Consolidated Standards of Reporting Trials) recommendations. Likewise, we reviewed six protocols of large multicenter RCTs, and compared planned and reported subgroups.

Results: We identified 18 RCTs (n= 6439). Sixteen trials used GOS at 6 months as outcome. Five RCTs reported covariate adjustment. The number of covariates was limited (≤5), most frequently including age. Many covariates were outcome predictors. Four RCTs reported only adjusted treatment effects as the main efficacy parameter. Eleven RCTs reported subgroup analyses. Several subgroup factors (≤7, mainly outcome predictors) and outcomes (≤4) were included. The highest total number of subgroups was fifteen, and only three RCTs completely pre-specified subgroups. Notably, 10 out of 11 RCTs performed inappropriate separate subgroup analyses. Five out of 11 RCTs gave subgroups the same emphasis as the overall effect. Reported subgroup analyses were insufficiently described, and clearly differed from those planned in the protocol.

Conclusions: The reported covariate adjustment and subgroup analyses from TBI trials had several methodological shortcomings. Appropriate performance and reporting of covariate adjustment and subgroup analysis should be considerably improved in future TBI trials, since interpretation of treatment benefits may be misleading otherwise.

Introduction

Acute traumatic brain injury (TBI) is a leading cause of premature death and disability and remains a major public health problem around the world [30, 35]. Brain injury management is primarily aimed at measures to prevent and limit the development of secondary brain damage [4, 19, 40]. Many therapeutic randomized clinical trials (RCTs) have failed to demonstrate significant improvement in outcomes in patients with moderate and severe TBI [18, 39, 40]. These failures have been attributed to many causes, including insufficient pre-clinical and initial clinical work-up, small sample size, inadequate use of the outcome measure, too optimistic expectations, ineffective therapies, inappropriate target mechanism, and heterogeneity of patients [7, 11, 17, 18, 32, 39, 47].

The heterogeneity of the TBI patients remains despite strict inclusion and exclusion criteria [28, 32, 33]. Patient baseline characteristics, collected at entry to a RCT, give information on prognosis and type of injury, and reflect this heterogeneity. Proposals for dealing with heterogeneity in the design of RCTs include: Stratified randomization [7, 10, 32], block randomization [7], inclusion of patients with similar types of injury [7], and targeting patients most likely to benefit from the treatment (e.g. exclude patients with very good and very bad prognosis) [32, 33, 39]. Other proposals have focused on the analysis of RCTs: covariate adjustment [9, 10, 24, 25, 27, 44], and subgroup analysis [6, 10, 13, 39, 50].

The CONSORT (Consolidated Standards of Reporting Trials) statement includes recommendations to improve the analysis and reporting of covariate adjustment and subgroup analyses in RCTs [1]. Recently, the misuse of baseline data, especially the overinterpretation of subgroup analyses, has been noted in general medicine RCT reports [2]. It is likely that such misuse may have been more common in TBI trials because of the significant problems of heterogeneity. No formal evaluation of covariate adjustment and subgroup analysis in TBI trials has been performed so far.

We aimed to describe the appropriateness of reporting covariate adjustment and subgroup analysis in published phase III RCTs including acute, closed, and moderate to severe TBI adult patients, with a substantial sample size and with clinically relevant outcomes. We further reviewed a sample of the protocols of the largest multicenter TBI trials, in order to compare the concordance between planned and reported subgroup analyses.

Methods

RCT report search strategy

We searched MEDLINE (1966 to 1 April 2004), EMBASE (1974 to 1 April 2004), the Cochrane Controlled Trials Register and the Cochrane Database of Systematic Reviews (Cochrane Library, Issue 2, 19 April 2004) for RCTs in TBI. All searches used the key words traumatic brain injury or traumatic head injury or traumatic cranial injury in conjunction with each of the following words: severe, moderate, acute, treatment, randomized controlled trials, and randomized clinical trials. The searches were limited to RCTs in English, French, German and Spanish. We reviewed the bibliographies of relevant studies (RCTs and non-RCTs) to search for additional eligible RCTs. Only data accessible in peer-reviewed journals were included, and we were not masked with regard to authors or journal.

Inclusion and exclusion criteria of RCTs

We selected RCTs that met the inclusion and exclusion criteria (Box). One of us (A.V.H.) screened the titles and abstracts to exclude non-human studies and review articles, retrieved potentially relevant manuscripts for detailed evaluation, and selected publications compliant with the inclusion and exclusion criteria. Jointly with another of us (A.I.R.M.), both researchers reassessed inclusion and exclusion criteria. Reports that did not meet all these criteria were excluded. Differences were resolved by discussion until consensus was reached.

Data extraction

The eighteen eligible trials underwent data extraction from the full-text papers by one of us (A.V.H.) using a pre-designed extraction form. This form was previously used in another related review by two of the authors (A.V.H. and E.W.S.) and the information obtained was highly reproducible. Patient attributes measured before randomization were considered as baseline characteristics [13]. The information retrieved included general information, such as sample size, number of centers, and primary outcome. Information on covariate adjustment included number of covariates, pre-specification, selection strategy, statistical method, primary use of covariate adjustment, and emphasis with respect to unadjusted analysis. Information on subgroup analyses included number of subgroup factors, number of subgroup outcomes, number of subgroup analyses (product of factors by outcomes), pre-specification, statistical method (interaction test or separate subgroup p value), and emphasis given to subgroups.

Inclusion criteria

a. Prospective, parallel-groups, phase III RCT with random assignment to either
   a new medication/intervention or placebo/best intervention available.
b. Main RCT report.
c. Patients with non-missile, closed, clinically diagnosed TBI.
d. Acute TBI presentation (<24 hours between injury and treatment).
e. Moderate or severe TBI (i.e. Glasgow Coma Scale [GCS] ≤ 12).
f. Primary outcome expressed as Glasgow Outcome Score (GOS) at ≥ 3 months.
g. Patients older than 15 years.
h. More than 50 patients per treatment arm.

Exclusion criteria

a. Phase II RCTs.
b. Mild TBI (GCS: 13 to 15).
c. Chronic TBI treatments (e.g. rehabilitation).
d. Pediatric TBI patients
e. Primary outcome other than GOS.

Box. Inclusion and exclusion criteria for the systematic review.

Methodological standards

The CONSORT statement established standards of trial reporting, based on a large amount of methodological and applied information regarding appropriate design and analysis of trials [1]. This document proposed 22 items that should be reported in every RCT in order to allow the readers to judge the validity of the findings, and to be more confident when applying these findings in their practice. The CONSORT items 12, 15, and 18 refer to covariate adjustment and subgroup analysis. Covariate adjustment should use a limited number of covariates, and should clearly establish the reasons to choose covariates [1, 2, 13, 44]. Subgroup analysis should use a limited number of pre-specified subgroups, use interaction tests, and be considered a secondary analysis [1, 2, 6, 13, 50].

Definitions

An interaction test directly assesses differences between complimentary subgroups by studying treatment*subgroup factors. It involves one statistical test irrespective of the number of subgroups [6, 13]. In contrast, the separate subgroup p value method evaluates treatment effects in each complementary group independently [50]. This is inappropriate from a methodological point of view [1, 13].

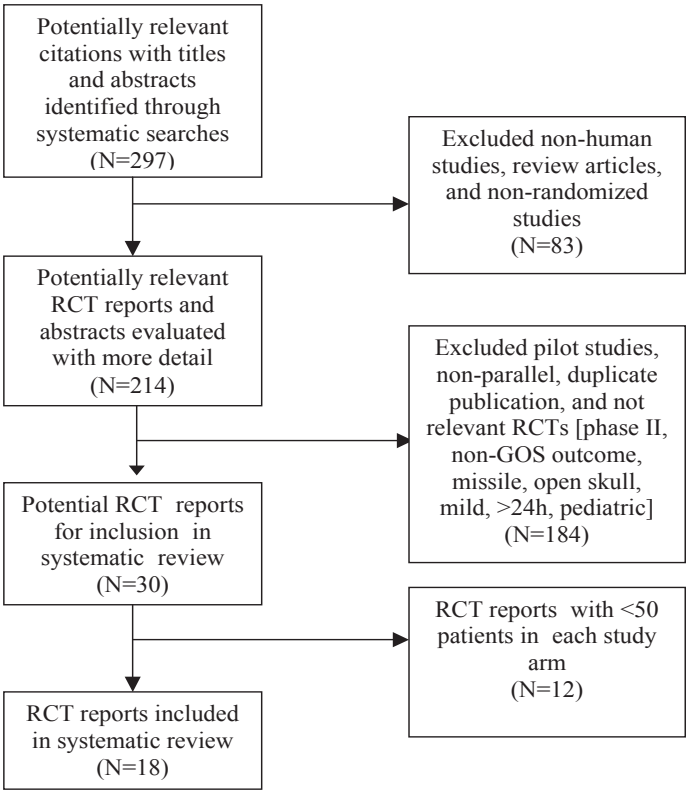

Figure. Flow diagram of literature search and selection of trials. RCT denotes randomized controlled trial.

Subgroups were pre-specified if they were clearly established in the methods part or if they were clearly labeled as pre-specified.

Subgroups were post-hoc if they were shown only in results and/or discussion or if they were clearly labeled as post-hoc [13]. Emphasis on subgroups was classified as similar to the overall effect if subgroups were reported in the abstract or main conclusion of the paper [43]. If the subgroups were only presented in the results and/or discussion, the overall effect had more emphasis.

RCT Protocols

We hypothesized that results on subgroup analyses might differ from subgroup analyses as specified in the protocol. Six protocols of some of the largest multicenter TBI trials were reviewed. We focused on the number of subgroup factors

and outcomes, whether subgroups were pre-specified, and statistical method of subgroup analysis. Thus, we compared the planned subgroup analyses with the reported subgroup analyses, in order to evaluate their concordance and discrepancies.

Statistical analyses

Descriptive statistics included percentages and medians. The software used was EXCEL and SPSS 10.0 (SPSS Inc., Chicago, USA, 1999).

Results

General RCT characteristics

The systematic review of the literature identified 297 potentially relevant citations with titles and abstracts. We finally included 18 RCTs [3, 5, 12, 14, 15, 16, 20, 21, 22, 31, 34, 36, 41, 42, 45, 48, 49, 51] (Figure). All RCT reports were published in English, from 1981 to 2004 (Table 1). The number of patients included ranged between 100 and 1120 (median: 265), originating from 1 to 95 centers (median: 8) mainly from developed countries [34, 36].

Eleven trials investigated drugs, and they were published before the year 2000. The time of follow up ranged between 6 and 14 months (median 6 months). The primary end point was mainly GOS at 6 months (16 out of 18 trials), which was dichotomized in 9 trials. Notably, from six trials that yielded positive results [14, 16, 31, 48, 49, 51], five were single center studies, and used a therapy other than a drug— i.e. surgery, hypothermia, mannitol and hyperbaric oxygen [15, 31, 41].

Covariate adjustment

Five trials published before 1997 reported covariate adjustment [20, 22, 45, 48, 49] (Table 2). The number of covariates included ranged from one to five. Covariates from 2 trials were not clearly defined [45, 48]. Age was the most commonly chosen covariate [22, 45, 48, 49]. Most of the covariates were well-known outcome predictors (e.g. age, initial Glasgow Coma Scale [GCS], motor score, pupilary reactivity, and initial Computed Tomography [CT] abnormalities) but this reason was not clearly established in the RCT reports. Center was considered as a covariate in two older trials [20, 48], and imbalance in age and GCS was reported as the reason for adjustment in one trial [22]. Logistic regression was the main method of adjustment in 3 trials [22, 45, 48]. Four RCTs gave only adjusted results, and one RCT [22] reported adjusted and unadjusted results. In this trial, the unadjusted treatment

Author [reference], year	Country*	Centers	n (arms)	Treatment	Target population†	GOS¶
Positive trials (n=6)						
Lu [31], 2003	CHN	1	230 (2)	Craniotomy	Severe	6m
Zhi [50], 2003	CHN	1	396 (2)	Hypothermia	Severe	6m
Cruz [16], 2002	BRA	1	141 (2)	Mannitol	Coma, Acute IPH	6m
Cruz [15], 2001	BRA	1	178 (2)	Mannitol	Coma, Acute SDH	6m
Harders [22], 1996	GER	21	123 (2)	Nimodipine	SAH	6m
Rockswold [40], 1992	USA	1	168 (2)	Hyperbaric O_2	Severe	12, 6, 18m
Negative trials (n=12)						
Cooper [14], 2004	AUS	12	226 (2)	Hypertonic saline	Severe	6m (usual & extended)
Clifton [12], 2001	USA	11	392 (2)	Hypothermia	Severe	6m
Morris [36], 1999	USA, ISR, EUR, CAN, AUS, ARG.	95	693 (2)	NMDA antagonist (Selfotel)	Severe	6m
Marshall [34], 1998	EUR, AUS, ISR.	50	1120 (2)	Tirilazad	Severe (85%), Moderate (15%)	6m
Young [48], 1996	USA	29	463 (3)	Pegorgotein	Severe	3m
Grumme [21], 1995	GER, AUT	9	396 (2)	Triamcinolone	Severe	Discharge & 1y
European Study Group Nimodipine [44], 1994	EUR	21	852 (2)	Nimodipine	Severe, Not obeying commands	6m
Gaab [20], 1994	GER	10	300 (2)	Dexamethasone	Severe Moderate	12m (modified)
Wolf [47], 1993	USA	2	149 (2)	Tromethamine	Severe	3, 6, 12m
Bailey [3], 1991	UK, FIN	6	351 (2)	Nimodipine	Not obeying commands	6m
Braakman [5], 1983	NED	2	161 (2)	Dexamethasone	Coma, Severe	6m & Survival 1y
Saul [41], 1981	USA	1	100 (2)	Methylpredni-solone	Severe	6m

*AUS: Australia; CHN: China; BRA: Brazil; USA: United States of America; ISR: Israel; EUR: Europe; GER; Germany; AUT: Austria; UK: United Kingdom; FIN: Finland; NED: Netherlands.
†Severe TBI: Glasgow Coma Score (GCS) <9; Moderate TBI: GCS 9 to 12; IPH: Intraparenchymal hemorrhage; SDH: Subdural hematoma; SAH: Subarachnoid hemorrhage.
¶GOS: Glasgow Outcome Scale; extended: GOSE, Glasgow Outcome Score Extended; modified: mGOS, Modified GOS; m: months; y: year.

Table 1. General characteristics of trial reports included in the review (n=18).

Author [reference]	Covariates included		Reason to include	Method used in covariate adjustment	Emphasis in results	Change in conclusion
	n	Definition*				
Young [48], 1996	1	Age (<45; ≥45)	Not given	Stratified Mantel Haenzel	Adjusted**	NA††
Harders [22], 1996	4	Age (continuous), GCS† at entry (≤12, >12); amount of blood in initial CT (ordinal, 3); Fisher grade in initial CT (ordinal, 4)	Imbalance (age, GCS); rest not given	Logistic regression	Unadjusted	No
European Study Group Nimodipine [44], 1994	5	Age (ns‡); gender (ns), craniotomy (ns); pupilary reaction (ns); GCS motor responses (ns)	Not given	Logistic regression	Adjusted**	NA
Gaab [20], 1994	1	Center (categorical, 10)	Center	Stratified Mantel Haenzel	Adjusted**	NA
Wolf [47], 1993	5	Age (ns); GCS motor score (ns); strata of neurological insult (ns); elevated ICP§ (ns); center (categorical, 2)	Center; rest not given	Logistic regression	Adjusted**	NA

* Definition refers to the way they were considered in analyses: continuous, categorical or ordinal. In case of categorical or ordinal covariates, it is written the number of categories considered.
† GCS: Glasgow Coma Scale.
‡ ns: Non specified.
§ ICP: Intracranial pressure.
** Only adjusted effect given.
†† Not available.

Table 2. Covariate adjustment on five trial reports.

Results of trial	TBI trials			Internal Medicine trials‡		
	n	Covariate adjustment	Subgroup analyses	n	Covariate adjustment	Subgroup analyses
Negative	12	4 (33%)†	9 (75%)	39	22 (56%)	22 (56%)
Positive	6	1 (17%)	2 (33%)	45	22 (49%)	25 (56%)
Total	18	5 (28%)	11 (61%)	84	44 (52%)	47 (56%)

* No significant differences between positive and negative trials, and between TBI and internal medicine trials.
† Percentages are separated by topic (TBI or Internal Medicine), and correspond to: number of trials with the analyses / number of trials (n) in the same row.
‡ Internal Medicine includes: General Internal Medicine, Cardiology and Oncology (reference 26).

Table 3. Number of reported covariate adjustment and subgroup analyses in positive and negative trials: Comparison between traumatic brain injury and internal medicine trials*

Author [reference]	Subgroup factors	Subgroup outcomes	Total ¶	Pre-specified?**	Statistical method	Subgroup found? (n)††	Emphasis in results
Cooper [14], 2004	GCS	GOSE†6m	1	Yes	Separate test	No	Overall
Clifton [12], 2001	Age, GCS, compressed cisterns, surgical hematoma, initial hypothermia	Poor outcome‡, death, ICP>30 mmHg	15	No	Separate test	Yes (2)	Overall
Morris [36], 1999	EDH*, SDH, IPH, GCS, secondary ischemic events	Favorable outcome‡, death	10	Yes	Separate test	No	Overall
Marshall [34], 1998	Gender, SAH, GCS.	Favorable outcome, death	6	Partially	Separate test	Yes(3)	Equal subgroup & overall
Young [48], 1996	GCS.	Good outcome‡, favorable outcome (3m & 6m)	4	No	Separate test	No	Overall
Harders [22], 1996	GCS, amount of blood-CT, Fisher grade-CT	Favorable outcome	3	No	Separate test	No	Overall
Grumme [21], 1995	Age, focal lesions, GCS, EDH, SDH, admission <1h, initial steroid treatment	GOS at discharge, GOS 1y	14	Partially	Separate test	Yes(1)	Equal subgroup & overall
European Study Group Nimodipine [44], 1994	Brain damage, SAH, others (unknown)	Favorable outcome	>2	Partially	Separate test	Yes(1)	Equal subgroup & overall
Rockswold [40], 1992	GCS, mass lesion, contusion, pupilary reactivity	Favorable outcome, death (1y)	8	Partially	Separate test	Yes(2)	Equal subgroup & overall
Bailey [3], 1991	Age, motor GCS, intracerebral lesion, motor response, others (unknown)	Favorable outcome, death	>8	Partially	Interaction test	No	Overall
Braakman [5], 1983	Probability of survival§, center	Survival 1m & 6m	4	Yes	Separate test	No	Equal subgroup & overall

* EDH: Epidural hemorrhage.
† GOSE categories (8): dead, vegetative, lower severe disability, upper severe disability, lower moderate disability, upper moderate disability, lower good recovery, upper good recovery.
‡ GOS at 6 months, unless otherwise indicated. Categories (5); Poor outcome=severe disability, vegetative state and death; Favorable outcome= good recovery and moderate disability; Good outcome= Good recovery.
§ Probability of survival was calculated at admission, based on age, pupil reactivity to light and best motor response of the arms.
¶ Total number of subgroup analyses= number of factors * number of outcomes.
** Yes: Subgroups described in methods; No: Subgroups only described in results/discussion; Partially: some subgroups described in methods and some others only in results/discussion. †† n: number of significant ($p<0.05$) subgroups.

Table 4. Subgroup analyses in trial reports.

effect received more emphasis than the adjusted one, but it did not affect the overall conclusion. One positive trial reported a covariate adjusted treatment effect [22] (Table 3).

Subgroup analysis

Eleven trials reported subgroup analyses [3, 5, 12, 14, 21, 22, 34, 36, 41, 45, 49] (Table 4). The maximum number of subgroup factors was seven, including mainly GCS and mass lesions. The maximum number of subgroup outcomes was four, mainly favorable outcome of GOS at 6 months (both good recovery and moderate disability). Two trials did not report the complete number of subgroup factors [3, 45]. The highest total number of examined subgroups was fifteen [12]. Three trials did not report pre-specified subgroups [12, 22, 49]. Some of the trials reported partially pre-specified subgroups [3, 21, 34, 41, 45]. Ten out of 11 trials performed separate subgroup analyses, and only one performed the statistically appropriate interaction test [3]. Five trials found subgroup effects [12, 21, 34, 41, 45], and five trials gave equal emphasis to the subgroup effects and overall effects [5, 21, 34, 41, 45]. Two positive trials reported subgroup analyses [22, 41] (Table 3).

Subgroup analysis in protocols and reports

The protocols corresponded to six RCT reports [3, 12, 34, 36, 45, 49] (Table 5). All six reported on subgroup effects, and the reported analyses clearly differed from those specified in the protocols. Two protocols did not define any subgroup factors, while they were considered in the reports [3, 12]. In the three protocols that pre-specified outcomes, these were also considered in the subgroup analyses [34, 36, 45]. Subgroup analyses were not pre-specified in three protocols [3, 12, 49]. When subgroup analyses were pre-specified in the protocols [34, 36, 45], only one trial reported them as planned [36]. Finally, five trials reported separate subgroup tests, but protocols specified an interaction test, a combination of separate and interaction test, or nothing [12, 34, 36, 45, 49]. These findings supported our hypothesis that trial reports included different subgroups than those planned in the protocols.

Discussion

Many RCTs have been published in acute TBI, but most of these have failed to yield convincing treatment benefits [18, 32, 33, 39, 40]. Various explanations for these failures have been put forward: Most of the TBI trials have insufficient sample size, and are poorly designed to detect or refute treatment benefits [17]. Others have suggested that the hypothesized absolute treatment benefit was too large and too optimistic (e.g. 10-15%), that investigating a more modest absolute benefit (e.g. 5-8%) would be desirable, and that larger trials are necessary [32, 39]. The largest

Author [reference]	Subgroup factors‡		Subgroup outcomes		Pre-specified?		Method	
	Protocol	Report	Protocol	Report	Protocol	Report¶	Protocol	Report
Clifton [12], 2001	Not defined	Age, GCS, cisterns, surgical hematoma, hypothermia	Not defined	Poor outcome, death, ICP>30 mmHg	No	No	Not defined	Separate test
Morris [36], 1999	Time between injury and treatment*	EDH, SDH, IPH, severity, secondary ischemic events	Favorable outcome, death, GOS 4 categories, DRS§(all at 6m &3m)	Favorable outcome, death (both 6m)	Yes	Yes	Interaction test	Separate test
Marshall [34], 1998	GCS, SBP†, center	Gender, SAH, GCS.	Favorable outcome, death.	Favorable outcome, death	Yes	Partially	Interaction test/Separate test	Separate test
Young [48], 1996	Age	GCS	Not defined	Good + favorable outcome (3m &6m)	No	No	Not defined	Separate test
European Study Group Nimodipine [44], 1994	Age, sex, motor GCS, pupilary reactivity, surgery for occupying lesion	Brain damage, SAH, others (unknown)	Favorable outcome	Favorable outcome	Yes	Partially	Not defined	Separate test
Bailey [3], 1991	Not defined	Age, motor GCS, intra-cerebral lesion.	Not defined	Favorable outcome, death	No	Partially	Not defined	Interaction test

* Defined as subgroup factor in protocol, but not considered baseline characteristic in our review (see methods).
† Systolic blood pressure
‡ Abbreviations as in previous tables.
§ Disability Rating Scale.
¶ Yes: Subgroups described in methods; No: subgroups only described in results/discussion; Partially:some subgroups described in methods and some others only in results/discussion.

Table 5. Subgroup analysis in selected traumatic brain injury trials: protocol vs. report

phase III trial in our review had 1120 patients (34). Here, most of the patients had severe TBI (85%), and hence more than 20% of baseline risk of unfavorable outcome. Even, in this study, the power to find an absolute benefit of less than 8% was limited [17].

These disappointing results have led to an increased interest in subgroup

analyses, trying to identify more homogeneous subgroups who may benefit from the intervention. This stems from the realization that heterogeneity of patients is a major confounding factor in the design and analysis of TBI trials, and that heterogeneity is related to injury patterns and prognostic risk [32, 33]. Approaches for dealing with heterogeneity relate to both the design and analysis phases. In the design phase stratified randomization can be employed. Alternatively, treatment can be targeted to patients with a defined type of pathology or to patients with an intermediate prognosis, e.g. between 20 and 80% probability of unfavorable outcome [27, 32, 33, 39]. A reduction of the sample size of 30% may be achievable, for the same power as when the whole population is considered [33]. In the analysis phase, covariate adjustment and subgroup analysis may be employed.

Covariate adjustment uses covariates in order to provide a more individual-oriented treatment effect, that is also corrected for imbalance. It increases power, and reduces the required sample size [9, 10, 24, 25, 27, 44]. Subgroup analysis assesses differences in treatment effect across different subpopulations of patients [6, 13, 25, 50]. Covariate adjustment and subgroup analysis are common in RCTs, but their reporting has shown many flaws in trials from general medicine journals, especially for subgroup analysis [2]. The revised CONSORT statement recommended guidelines to improve the reporting of the RCTs, and facilitates informed judgments regarding the validity of the trials [1]. We found that reporting of covariate adjustment and subgroup analysis in TBI clinical trials has several shortcomings. Further, reported subgroups differed substantially from those planned in the protocols.

Covariate adjustment was reported in trials that mainly had negative treatment benefits and were multicenter. Some papers have recommended using covariate adjustment in TBI trials, especially for prognostic factors [9, 10, 27, 32, 44]. An appropriate small number of baseline characteristics were used in the trials, including age [22, 45, 48, 49], GCS [22, 45, 48], pupilary reactivity [45], elevated intracranial pressure [48], CT abnormalities [22], gender [45], and center [20, 48]. Most of these baseline characteristics are known prognostic factors [30, 35]. Although covariates were appropriately few, the reasons to select these were not clearly stated. We also found that covariate definitions (i.e. dichotomized, ordinal or continuous) were insufficiently reported. We consider that covariate adjustment using a limited number of pre-specified, clearly defined prognostic factors is a valid procedure.

Age is the strongest predictor of unfavorable outcome in TBI, as demonstrated in data from epidemiological studies [23, 29, 37] and RCTs [37]. Age is continuously associated with unfavorable outcome, and should be used as a continuous variable in covariate adjustment [29]. Motor GCS is another strong predictor, and should be

used as covariate [9, 32]. Indeed, adjusting for age and motor GCS could reduce the sample size by 25 to 30% in TBI trials [9, 10]. Other severity characteristics (e.g. pupilary reactivity, CT severity, hypoxia, hypotension) may also be included [32]. Gender has not been suggested as a variable to adjust for [32], and its potential predictive value could be related to its association with other strong predictors. Center/country also may be considered [10, 44], because differences in patient baseline characteristics have been demonstrated across continents and regions [28].

A moderate number of subgroup analyses were noted in TBI reports. Limiting the number of subgroups is strongly recommended, as this decreases problems arising from multiplicity and helps define valid statistical tests [50]. The subgroup factors were mainly predictors of unfavorable outcome [30, 35, 39]. TBI trials might focus on appropriate subgroups of patients, defined by predictors: age, gender, GCS, pupilary reactivity, hypotension, CT severity and SAH [10, 39]. Moreover, interaction tests are appropriate to assess differences between complementary subgroups. The separate subgroup p value method evaluates treatment effects in independent subgroups, which is inappropriate from a methodological point of view [1, 13].

Remarkably, the reporting of subgroups in TBI trials had important methodological shortcomings [1, 2, 13, 50]: most trials reported partially or did not report pre-specified subgroups, used separate subgroup testing, and performed post-hoc subgroup analyses. Moreover, nearly half of the trials overemphasized any statistically significant effect found.

For example, the Tirilazad trial [34] and the hyperbaric oxygen trial [41] reported partially pre-specified subgroups, used separate analyses, and emphasized their subgroup findings. The former found a lower mortality in males, moderately head injured and non-SAH patients; the latter found lower mortality in severely head injured (GCS 4 to 6) and in patients with surgical mass lesions. However, without a good a priori rationale for subgroup differences, the overall treatment effect should provide a reasonable estimate for each subgroup of TBI patients, unless confirmatory evidence of treatment differences become available. These subgroups may be unreliable (i.e. based on small number of patients), should be considered exploratory, and should only serve to motivate further trials rather than drawing definite conclusions [13, 43].

It can be argued that good or poor subgroup analysis reporting is not completely related to a good or poor practice in the analysis of TBI trials. It is clear that most physicians have only access to trial reports, and take decisions based on reports. Hence, the evaluation of reporting in TBI trials is worthwhile [46]. Moreover, the

CONSORT guidelines based their recommendations on papers dealing with design, conduct, analysis, and reporting of trials. These reporting guidelines may therefore include a part of good practice on trial analysis. However, it is also possible that an analysis plan may differ from a reported analysis. Results from Table 5 show that the most plausible explanation for poor subgroup analysis reporting is a poor analysis plan, especially for pre-specification of subgroups and statistical methods.

Differences in subgroup definitions between the TBI protocols and trial reports were substantial. We observed differences regarding subgroup factors, pre-specification, and method of analysis. This seems an indication of post-hoc analyses. Thus, the protocol stated a particular definition of subgroups, but the results only became statistically significant when an alternative definition of the subgroups was used [50]. Further, post-hoc subgroups should be treated with skepticism, as they test data-derived hypotheses rather than hypotheses stated a priori [50, 13]. A recent paper found that trial reporting of outcomes was not only frequently incomplete, but also biased and inconsistent with protocols [8]. Our results confirm the inconsistency and incompleteness of subgroup analyses between trials and protocols.

We have recently studied the reporting of covariate adjustment and subgroup analyses in high impact internal medicine (IM) journals (general medicine, cardiology and oncology) [26]. As shown in Table 3, TBI trials reported less covariate adjustment than IM trials, independently of their final result (positive or negative). Moreover, TBI trials adjusted for fewer covariates ($\leq$5) than IM trials (mainly between 5 and 9). However, reasons to include covariates were poorly reported in TBI trials in comparison to IM trials (36 out of 44 reports). TBI and IM trials reported subgroup analyses in a similar frequency. However, negative TBI trials reported more subgroup than negative IM trials, and positive TBI trials reported fewer subgroups than IM trials. Moreover, TBI reports fully pre-specified subgroup analyses less commonly (3/11 vs. 22/47), and used interaction testing less frequently (1/11 vs. 20/47) than IM trials. This is an additional indication that TBI trials inappropriately reported subgroup analyses.

Future design of TBI trials should incorporate covariate adjustment for important predictors. Such an adjustment should be pre-specified in the trial protocol, including the coding of predictors. Alternatively, inclusion can be restricted to certain prognostic groups, defined by predictors. For example, patients with intermediate prognosis (20%-80% of favorable outcome) or patients with focal injury may be hypothesized to benefit from treatment. Such a targeting approach may decrease the sample size requirements by 30% [33]. If subgroup effects are studied, the number of subgroups should be pre-specified with a clear rationale,

and should be analyzed with the appropriate test of interaction. Treatment effects can also be tested in a pre-specified subgroup within a TBI trial, such as the focal injury subgroup ('confirmatory subgroup analysis') [38].

Our paper has some limitations. We did not include phase II TBI trials. However, we observed that most of them were small (n<100) and did not report relevant clinical outcomes (data not shown). We evaluated only reported results, and it is possible that there may be differences between the planned and the reported analysis. For subgroup analysis, we found clear discrepancies between protocols and reports. No evaluation was possible for covariate adjustment. It is also possible that investigators explored, for instance, more subgroups than those planned and reported. We could not evaluate this possibility. Finally, we used only a limited number of protocols.

In conclusion, reporting of covariate analyses and subgroup analyses in TBI clinical trials had several shortcomings, particularly for subgroup analyses. Likewise, reported subgroup analyses clearly differed from those planned in the protocols. The appropriate reporting of these secondary analyses should be considered in future TBI trials, since interpretation of treatment benefits may be misleading otherwise.

References

1. Altman DG, Schultz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, Lang T; CONSORT GROUP (Consolidated Standards of Reporting Trials): The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 134:663-694, 2001.

2. Assmann SF, Pocock SJ, Enos LE, Kasten LE: Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 355:1064-1069, 2000.

3. Bailey I, Bell A, Gray R, Gullan R, Heiskanan O, Marks PV, Marsh H, Mendelow DA, Murray G, Ohman J, Quaghebeur G, Sinar J, Skene A, Teasdale G, Waters A: A trial of the effect of nimodipine on outcome after head injury. Acta Neurochir (Wien) 110:97-105, 1991.

4. Bayir H, Clark RSB, Kochanek PM: Promising strategies to minimize secondary brain injury after head trauma. Crit Care Med 31[Suppl.]:S112-S117, 2003.

5. Braakman R, Schouten HJA, Blaauw-van Dishoeck M, Minderhoud JM: Megadose steroids in severe head injury. Results of a prospective double blind clinical trial. J Neurosurg 58:326-330, 1983.

6. Brookes ST, Whitley E, Egger M, Davey Smith G, Mulheran PA, Peters TJ: Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol 57:229-236, 2004.

7. Bullock MR, Merchant RE, Choi SC, Gilman CB, Kreutzer JS, Marmarou A, Teasdale GM: Outcome measures for clinical trials in neurotrauma. Neurosurg Focus 13, 2002; Accessible at http://www.aans.org/education/journal/neurosurgical/july02/13-1-nsf-toc.asp
8. Chan A, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG: Empirical evidence for selective reporting of outcomes in randomized trials. Comparison of protocols to published results. JAMA 291:2457-2465, 2004.
9. Choi SC: Sample size in clinical trials with dichotomous endpoints: use of covariables. J Biopharm Stat 8:367-375, 1998.
10. Choi SC, Bullock R: Design and statistical issues in multicenter trials of severe head injury. Neurol Res 23:190-192, 2001.
11. Choi SC, Clifton GL, Marmarou A, Miller ER: Misclassification and treatment effect on primary outcome measures in clinical trials of severe neurotrauma. J Neurotrauma 19:17-22, 2002.
12. Clifton GL, Miller ER, Choi SC, Levin HS, McCauley S, Smith KR, Muizelaar JP, Wagner FC, Marion DW, Luerssen TG, Chesnut RM, Schwartz M: Lack of effect of induction of hypothermia after acute brain injury. N Engl J Med 344:556-563, 2001.
13. Cook DI, Gebski VJ, Keech AC: Subgroup analysis in clinical trials. Med J Aust 180:289-291, 2004.
14. Cooper DJ, Myles PS, McDermott FT, Murray LJ, Laidlaw J, Cooper G, Tremayne AB, Bernard SS, Ponsford J, for the HTS Study Investigators: Prehospital hypertonic saline resuscitation of patients with hypotension and severe traumatic brain injury. A randomized controlled trial. JAMA 291:1350-1357, 2004.
15. Cruz J, Minoja G, Okuchi K: Improving clinical outcomes from acute subdural hematomas with the emergency preoperative administration of high doses of mannitol: a randomized trial. Neurosurgery 49:864-871, 2001.
16. Cruz J, Minoja G, Okuchi K: Major clinical and physiological benefits of early high doses of mannitol for intraparenchymal temporal lobe hemorrhages with abnormal pupilary widening: a randomized trial. Neurosurgery 51:628-638, 2002.
17. Dickinson K, Bunn F, Wentz R, Edwards P, Roberts I: Size and quality of randomised controlled trials in head injury: review of published studies. BMJ 320:1308-1311, 2000.
18. Doppenberg EMR, Choi SC, Bullock R: Clinical trials in traumatic brain injury: lessons for the future. J Neurosurg Anesthesiol 16:87-94, 2004.
19. Dutton RP, MacCunn M: Traumatic brain injury. Curr Opin Crit Care 9:503-509, 2003.
20. Gaab MR, Trost HA, Alcantara A, Karimi-Nejad A, Moskopp D, Schultheiss R, Bock WJ, Piek J, Klinge H, Scheil F, Osterwald P, Samii M, Brawanski A, Meixensberger J, Schurmann K, Schubert R, Arnold H, Kehler U, Deisenroth K, Benker G, Vester JC, Dietz H: "Ultrahigh" dexamethasone in acute brain injury. Results from a prospective randomized double-blind multicenter trial (GUDHIS). Zentralbl Neurochir 55:135-143, 1994.

21. Grumme T, Baethmann A, Kolodziejczyk D, Krimmer J, Fischer M, Eisenhart Rothe B, Pelka R, Bennefeld H, Pollauer E, Kostron H, Leheta F, Necek S, Neeser G, Sachsenheimer W, Sommerauer J, Verhoeven F: Treatment of patients with severe head injury by triamcinolone: a prospective, controlled multicenter clinical trial of 396 cases. Res Exp Med (Berl) 195:217-229, 1995.
22. Harders A, Kakarieka A, Braakman R, and the German tSAH Study Group: Traumatic subarachnoid hemorrhage and its treatment with nimodipine. J Neurosurg 85:82-89, 1996.
23. Harris C, DiRusso S, Sullivan T, Benzil DL: Mortality risk after head injury increases at 30 years. J Am Coll Surg 197:711-716, 2003.
24. Hauck WW, Anderson S, Marcus SM: Should we adjust for covariates in non-linear regression analyses of randomized trials?. Control Clin Trials 19:248-256, 1998.
25. Hernández AV, Steyerberg EW, Habbema JDF: Clinical trials with dichotomous end-points: covariate adjustment increased power and potentially reduces sample size. J Clin Epidemiol 57:454-460, 2004.
26. Hernández AV, Steyerberg EW, Habbema JDF: Inappropriate use of baseline characteristics in clinical trials: Assessment of high impact factor medical journals. Med Decis Making 2004; 24: 426.
27. Hukkelhoven CWPM, Steyerberg EW, Maas AIR: Quality of randomised controlled trials in head injury. BMJ 321:704, 2000.
28. Hukkelhoven CWPM, Steyerberg EW, Farace E, Habbema JDF, Marshall LF, Maas AIR: Regional differences in patient characteristics, case management, and outcomes in traumatic brain injury: experience from tirilazad trials. J Neurosurg 97:549-557, 2002.
29. Hukkelhoven CWPM, Steyerberg EW, Rampen AJJ, Farace E, Habbema JDF, Marshall LF, Murray GD, Maas AIR: Patient age and outcome following severe traumatic brain injury: an analysis of 5600 patients. J Neurosurg 99:666-673, 2003.
30. Jennett B: Epidemiology of head injury. J Neurol Neurosurg Psychiatry 60:362-369, 1996.
31. Lu LQ, Jiang JY, Yu MK, Hou LJ, Chen ZG, Zhang GJ, Zu C: Standard large trauma craniotomy for severe traumatic brain injury. Chin J Traumatol 6:302-304, 2003.
32. Maas AIR, Steyerberg EW, Murray GD, Bullock R, Baethmann A, Marshall LF, Teasdale GM:Why have recent trials of neuroprotective agents in head injury failed to show convincing efficacy? A pragmatic analysis and theoretical considerations. Neurosurgery 44:1286-1298, 1999.
33. Machado SG, Murray GD, Teasdale GM: Evaluation of designs for clinical trials of neuroprotective agents in head injury. J Neurotrauma 16:1131-1138, 1999.
34. Marshall LF, Maas AIR, Marshall SB, Bricolo A, Fearnside M, Iannotti F, Klauber MR, Lagarrigue J, Lobato R, Persson L, Pickard JD, Piek J, Servadei F, Wellis GN, Morris GF, Means ED, Musch B: A multicenter trial on the efficacy of using tirilazad mesylate in cases of head injury. J Neurosurg 89:519-525, 1998.

35. Masson F: Epidemiology of severe cranial injuries [in French]. Ann Fr Anesth Reanim 19:261-269, 2000.
36. Morris GF, Bullock R, Marshall SB, Marmarou A, Maas A, the Selfotel Investigators, Marshall LF: Failure of the competitive N-methyl-D-aspartate antagonist Selfotel (CGS 19755) in the treatment of severe head injury: results of two phase III clinical trials. J Neurosurg 91:737-743, 1999.
37. Mosenthal AC, Lavery RF, Addis M, Kaul S, Ross S, Marburger R, Deitch EA, Livingston DH: Isolated traumatic brain injury: Age is an independent predictor of mortality and early outcome. J Trauma 52:907-911, 2002.
38. Moyé LA, Deswal A. Trials within trials: confirmatory subgroup analyses in controlled clinical experiments. Control Clin Trials 22:605-619, 2001
39. Narayan RK, Michel ME, and the Clinical Trials in Head Injury Study Group: Clinical trials in head injury. J Neurotrauma 19:503-557, 2002.
40. Reinert MM, Bullock R: Clinical trials in head injury. Neurol Res 21:330-338, 1999.
41. Rockswold GL, Ford SE, Anderson DC, Bergman TA, Sherman RE: Results of a prospective randomized trial for treatment of severely brain-injured patients with hyperbaric oxygen. J Neurosurg 76:929-934, 1992.
42. Saul TG, Ducker TB, Salcman M, Carro E: Steroids in severe head injury. A prospective randomized clinical trial. J Neurosurg 54:596-600, 1981.
43. Simes RJ, Gebski VJ, Keech AC: Subgroup analysis: application to individual patient decisions. Med J Aust 180:467-469, 2004.
44. The European Agency for the Evaluation of Medicinal Products. Committee for Proprietary Medicinal Products (CPMP): Points to consider on adjustment for baseline characteristics. Document CPMP/EWP/2863/99, EMEA, London, UK, 22 May 2003.
45. The European Study Group on Nimodipine in Severe Head Injury: A multicenter trial of the efficacy of nimodipine on outcome after severe head injury. J Neurosurg 80:797-804, 1994.
46. Vranos G, Tatsioni A, Polyzoidis K, Ioannidis JPA: Randomized trials of neurosurgical interventions: a systematic appraisal. Neurosurgery 55:18-26, 2004.
47. Wilson JTL: Assessing outcome in head injury trials. Curr Pharm Des 7:1537-1552, 2001.
48. Wolf AL, Levi L, Marmarou A, Ward JD, Muizelaar PJ, Choi S, Young H, Rigamonti D, Robinson WL: Effect of THAM upon outcome in severe head injury: a randomized prospective clinical trial. J Neurosurg 78:54-59, 1993.
49. Young B, Runge JW, Waxman KS, Harrington T, Wilberger J, Muizelaar JP, Boddy A, Kupiec JW: Effects of pegorgotein on neurologic outcome of patients with severe head injury. A multicenter, randomized controlled trial. JAMA 276:538-543, 1996.
50. Yusuf S, Wittes J, Probstfield J, Tyroler HA: Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 266:93-98, 1991.

Clinical applications

4

Adjustment for strong predictors of outcome reduces sample size requirements by 25% in traumatic brain injury trials

4.1

ABSTRACT

Purpose: To quantify the potential reduction in sample size that can be achieved by adjustment for predictors of outcome in traumatic brain injury (TBI) trials.

Methods: We used individual patient data from eight therapeutic phase III randomized clinical trials (RCTs, n=6292) in moderate or severe TBI, and three TBI surveys (n=2238). The primary outcome was the dichotomized Glasgow Outcome Scale at six months (favorable/unfavorable). Baseline predictors of outcome considered were age, motor score, pupillary reactivity, CT classification, traumatic subarachnoid hemorrhage, hypoxia, hypotension, glycemia, and hemoglobin. We calculated the potential sample size reduction obtained by adjustment of a hypothetical treatment effect for one to seven predictors with logistic regression models.

Results: The distribution of predictors was more heterogeneous in surveys than in trials. Adjustment of the treatment effect for the strongest predictors (age, motor score and pupillary reactivity) yielded a reduction in sample size between 16% and 23% in RCTs, and between 28% and 35% in surveys. Adjustment for seven predictors yielded a reduction of about 25% in most studies: between 20% and 28% in RCTs, and between 32% and 39% in surveys.

Conclusions: A major reduction in sample size can be obtained with covariate adjustment in TBI trials. The design and analysis of future clinical trials should adjust for important predictors of outcome.

Introduction

Traumatic Brain Injury (TBI) is an important public health problem worldwide (Jennett, 1996; Masson 2000). Unfortunately the search for effective treatments has not been very successful to date because of limitations in clinical development programs and methodological weaknesses in phase III trials. Problems have included limited pre-clinical evaluation, inadequate sample sizes, insensitive outcome measures, over-optimistic expectations, ineffective therapies, inappropriate selection of patients, and heterogeneity of the patient population (Narayan et al., 2002; Doppenberg et al., 2004; Maas et al., 1999; Dickinson et al., 2000; Choi et al., 2002; Wilson, 2001).

Adjustment of the treatment effect for important predictors of outcome ('covariates') can increase the statistical power to identify an efficacious treatment or intervention in a randomized clinical trial (RCT) (Gail et al., 1984; Hauck et al., 1998; Steyerberg et al., 2000; Hernández et al., 2004; Hernández et al., in press a). Covariate adjustment also provides a more subject-specific treatment effect and corrects for any imbalance in measured patient characteristics. Many predictors of outcome are known, including age, Glasgow Coma Scale (GCS) motor score, pupillary reactivity, and computed tomography (CT) abnormalities (Maas et al., 1999). In spite of this, covariate adjustment is not a standard procedure in TBI trials (Hernández et al., in press b).

An adjusted analysis needs fewer patients than an unadjusted analysis for the same statistical power. Therefore, one way to quantify the increase in power with covariate adjustment is by calculation of the reduction in sample size (RSS) (Hernández et al., 2004; Hernández et al., in press a; Choi, 1998). A previous investigation to quantify the RSS was performed with data from a Medical College of Virginia TBI survey (Choi, 1998). Adjustment of the treatment effect for age and GCS motor score yielded a reduction in sample size of 30%. However, a survey population would be expected to be more heterogeneous than a typical RCT population, where strict inclusion and exclusion criteria are applied. Thus this previous study may have over-estimated the potential reduction achievable in the context of a RCT.

No further evaluation of covariate adjustment has been reported in TBI. As part of the work of the IMPACT (International Mission for Prognosis of Head Injury and Analysis of Clinical Trials) Group we aimed to quantify the RSS achievable by adjustment for predictors of outcome for TBI patients included in RCTs and in surveys.

Methods

RCTs and Surveys

IMPACT links researchers based in the Netherlands, the UK, and the USA in a project addressing methodological problems in the design and the analysis of TBI trials. We have access to individual data from eight large phase III RCTs (Tirilazad International [TINT], Tirilazad USA [TIUS], Selfotel, SAPHIR, PEGSOD, HIT-I, HIT-II, SKB) and three TBI surveys (TCDB, EBIC, UK4).

Outcome and predictors

The primary outcome was the Glasgow Outcome Scale (GOS) at 6 months, dichotomized as unfavorable (death, vegetative state, and severe disability) or favorable (moderate disability, and good recovery). Age is a strong predictor of outcome in moderate to severe TBI patients (Hukkelhoven et al., 2003), as are GCS motor score and pupillary reactivity (Choi, 1998; Mushkudiani et al., 2004). Most of these prognostic studies were performed before the general availability of CT scans. Hukkelhoven et al. (in press) recently developed a model including seven early accessible clinical features and CT abnormalities: age, GCS motor score, pupillary reactivity, Marshall CT classification, traumatic subarachnoid hemorrhage (tSAH), hypoxia, and hypotension. Other novel potential predictors that we secondarily considered in our analyses were biochemical values: hyperglycemia (Jeremitsky et al., 2005) and anemia (Ariza et al., 2004).

Imputation of missing data

We selected patients with complete data on outcome, age and motor score (including a 'not testable' category for the motor score). Further, only patients aged 14 years and above were included in this analysis. Some of the other potential predictors were incomplete (Table 1). The SAPHIR and SKB trials did not record pupillary reactivity. The Tirilazad trials had few missing values (<14% over 9 predictors). The PEGSOD trial did not record CT classification, hypoxia, and hypothension. Glycemia and hemoglobin were not available for the Selfotel and HIT-II trials, and the three surveys. We performed a two-stage single imputation technique where we estimated the expected values for the patients with missing covariate values based on correlations with other variables (Little, 1992). Values were randomly imputed to reflect the variability around the expected values using the aregImpute function in the Hmisc library of S-plus (Little, 2004).

Predictors	Trials								Surveys		
	TINT N=1118	TIUS N=1041	Selfotel N=409	SAPHIR N=919	PEGSOD N=1510	HIT-I N=350	SKB N=126	HIT-II N=819	UK4 N=812	TCDB N=604	EBIC N=822
Age	100	100	100	100	100	100	100	100	100	100	100
Motor score	100	100	100	100	100	100	100	100	100	100	100
Pupils	93	95	97	0	97	98	0	98	90	100	94
CT class	99	99	100	99	0	0	100	99	0	0	98
tSAH	97	95	99	99	100	73	100	100	0	87	95
Hypoxia	88	89	100	93	0	0	67	0	98	100	99
Hypotension	97	97	0	93	0	82	83	100	99	100	99
Glucose	96	99	0	95	96	85	98	0	0	0	0
Hemoglobin	99	98	0	90	30	97	93	0	0	0	0

Motor score: GCS motor score; Pupils: Pupillary reactivity; CT class: Marshall CT classification; tSAH: traumatic subarachnoid hemorrhage.

Table 1. Proportion of available predictors across studies.

Prognostic Models

A reference model including only the treatment variable was used as the comparator for other models. Five increasingly complex logistic regression models were then used to quantify the RSS, with a treatment indicator being included in every model. The first regression model included age. Then GCS motor score was added, followed by pupillary reactivity to give the 'Core model'. CT classification, tSAH, hypoxia, and hypotension were then added to give the '7-predictor model', followed by glucose and hemoglobin to give the 'Full model'. Each model was fitted for each of the studies.

Age, glucose and hemoglobin were treated as continuous variables, and the remainder as categorical variables: GCS motor score (5 categories – see Table 2), pupillary reactivity (3 categories – see Table 2); Marshall CT classification (Marshall et al., 1991) (3 categories: normal CT [I] /cisterns present –shift<5mm [II]; cisterns compressed or absent-shift<5mm [III] /shift>5mm [IV]; evacuated mass lesion [V] /non-evacuated mass lesion [VI]); tSAH (yes/no); hypoxia (yes/no); and hypotension (yes/no). The independent prognostic effects of the covariates were examined on the log-likelihood scale, providing chi-square (χ^2) statistics.

Model performance was evaluated using the c-statistic and Nagelkerke's R^2. The c-statistic quantifies the discrimination of the model, i.e. the ability to distinguish unfavorable from favorable outcome. For binary outcomes the c-statistic is identical to the area under the receiver operating characteristic (ROC) curve. The c-statistic lies between 0.5 and 1 and indicates better model discrimination if closer to 1 (Steyerberg et al., 2001). Nagelkerke's R^2 measures the variability of the outcome explained by the model (i.e. the strength of association between the covariates and outcome). Nagelkerke's R^2 ranges from 0 to 1, with larger values indicating a better fit (Nagelkerke, 1991). The full models included glucose and hemoglobin, which were totally missing in the Selfotel and HIT-II trials and in the three surveys (Table 1). Therefore, we showed performance measures up to the 7-predictor models across all trials and surveys.

Simulations

None of the RCTs included had demonstrated a significant treatment effect. We simulated a positive treatment effect that gave an unadjusted odds ratio (OR) of 0.57 (coefficient: -0.557, corresponding to an average absolute risk reduction of 10% in unfavorable outcome) (Machado et al., 1999). Fifty percent of the patients were randomly allocated to the hypothesized treatment. A new outcome variable

was generated per study and per simulation, based on the comparison of a random uniform distribution (from 0 to 1) and the probability of poor outcome. This probability was based on the combination of the predictors of the full model and the treatment effect. Each regression model was applied in turn for this outcome to compare estimates of treatment effect. One thousand simulations were run using the original sample size for each study.

Reduction in sample size

We calculated the RSS to express the gain in power for each of the adjusted models. The formula used was: 100 - 100*[(mean of Z score for reference model) / (mean of Z score for adjusted model)]^2 (Hernández et al., 2004), where Z score is equal to the Wald statistic of the treatment effect coefficient. We only calculated RSS when data were available on the predictors included in the adjusted models (Table 1). Data on pupillary reactivity was not available in the SAPHIR and SKB trials, and the RSS was not calculated for their core models. Likewise, RSS was not calculated for the full models on trials or surveys with 100% of missings on biochemical values. We used S-plus 6 software (Insightful Inc, Seattle, WA).

Results

Predictors of outcome

Age, GCS motor score and pupillary reactivity were the strongest predictors of outcome and had a χ^2 of around 400. Other important predictors were tSAH (χ^2=175), CT classification (χ^2=102), hypotension (χ^2=82), and hypoxia (χ^2=30). Glucose and hemoglobin were weaker predictors.

The three strongest predictors were distributed differently across the datasets (Table 2). Patients enrolled in TBI trials (median age [25th-75th percentile]: 30 [21-43]) were younger than patients enrolled in surveys (32 [22-53]). The UK4 and EBIC surveys had the oldest patients, and the TCDB survey the youngest.

A substantial proportion of patients from surveys had untestable GCS motor scores (n=375, 17%). However, a higher proportion of patients with GCS motor score ≥4 was seen in trials (56%) in comparison with surveys (46%). A lower proportion of patients from trials had bilateral non-reactive pupils (21%) than patients from surveys (32%). Thus, patients from trials had lower risks of unfavorable outcome than patients from surveys (Table 2).

	Trials								Surveys		
	TINT	**TIUS**	**Selfotel**	**SAPHIR**	**PEGSOD**	**HIT-I**	**SKB**	**HIT-II**	**UK4**	**TCDB**	**EBIC**
	N= 1118	N=1041	N=409	N=919	N=1510	N=350	N=126	n=819	n=812	N=604	N=822
Age (Median, 25th-75th percentile)	30 (21-45)	30 (23-41)	28 (21-43)	32 (23-47)	27 (20-38)	34 (21-47)	27 (20-39)	33 (22-49)	36 (22-55)	26 (21-40)	38 (24-59)
Motor score (n, %)											
No response/extension	141 (13)	152 (15)	55 (13)	264 (29)	655 (43)	163 (47)	56 (44)	280 (34)	200 (25)	243 (40)	230 (28)
Flexion abnormal	237 (21)	132 (13)	91 (22)	143 (16)	165 (11)	45 (13)	14 (11)	92 (11)	37 (5)	74 (12)	55 (7)
Flexion withdrawal	327 (29)	300 (29)	127 (31)	223 (24)	334 (22)	56 (16)	16 (13)	181 (22)	142 (17)	122 (20)	113 (14)
Localizes pain/obeys	413 (37)	457 (44)	136 (33)	286 (31)	356 (24)	77 (22)	23 (18)	207 (25)	232 (29)	134 (22)	281 (34)
Non-testable	0	0	0	3 (.3)	0	9 (3)	17 (13)	59 (7)	201 (25)	31 (5)	143 (17)
Pupillary reactivity (n, %)											
Both responsive	813 (73)	709 (68)	316 (77)	612 (67)*	779 (52)	235 (67)	67 (53)*	583 (71)	445 (55)	300 (50)	527 (64)
One unresponsive	170 (15)	122 (12)	79 (19)	161 (18)*	160 (11)	51 (15)	35 (28)*	101 (12)	116 (14)	55 (9)	87 (11)
Both unresponsive	135 (12)	210 (20)	14 (3)	146 (16)*	571 (38)	64 (18)	24 (19)*	135 (16)	251 (31)	249 (41)	208 (25)
Unfavorable outcome (n,%)	456 (41)	395 (38)	177 (43)	378 (41)	774 (51)	171 (49)	70 (56)	328 (40)	518 (64)	393 (65)	422 (51)

* Imputed data

Table 2. Distribution of three main predictors and primary outcome across studies.

	Trials								Surveys		
	TINT	**TIUS**	**Selfotel**	**SAPHIR**	**PEGSOD**	**HIT-I**	**SKB**	**HIT-II**	**UK4**	**TCDB**	**EBIC**
Core model											
c-statistic	0.72	0.76	0.73	0.69	0.76	0.79	0.81	0.73	0.77	0.82	0.76
Nagelkerke-R^2	0.20	0.27	0.24	0.16	0.26	0.33	0.36	0.21	0.29	0.39	0.26
RSS (%)	16	22	18	16	23	23	18	17	28	35	30
7-predictor model											
c-statistic	0.81	0.81	0.78	0.74	0.78	0.84	0.84	0.80	0.82	0.87	0.81
Nagelkerke-R^2	0.36	0.35	0.31	0.24	0.29	0.44	0.44	0.34	0.38	0.51	0.37
RSS (%)	24	27	24	20	25	28	24	24	32	39	36

Core model included age, motor score, and pupillary reactivity; 7-predictor model included age, motor score, pupillary reactivity, Marshall CT classification, tSAH, hypoxia, and hypotension.

Table 3. Performance and reduction in sample size (RSS) for two logistic regression models across studies.

Performance of the prognostic models

The performance of most models derived from surveys was better than for models derived from trials (Table 3). The c-statistics for the core model ranged between 0.69 and 0.82, and increased substantially in the 7-predictor model, ranging between 0.74 and 0.87. Likewise, Nagelkerke's R^2 of the core model ranged between 16% and 39%, and between 24% and 51% for the 7-predictor model.

Reduction in sample size

The RSS for each model and across all studies is illustrated in Figure 1. We showed the RSS where data were available for the predictors included in the adjusted models. Adjustment of the treatment effect for age, motor score and pupillary reactivity yielded a RSS between 16% and 23% in trials, and between 28% and 35% in surveys (Table 3). Adjustment for seven predictors yielded even larger reductions: between 20% and 28% in trials, and between 32% and 39% in surveys. Further adjustment for 9 predictors (full model) increased the reductions in sample size limitedly (not shown). The positive relationship between both the c-statistic and Nagelkerke's R^2 with the RSS is shown in Figures 2a and 2b for the core and 7-predictor models.

Discussion

A substantial reduction in the sample size requirements for TBI trials can be obtained with adjustment for important predictors. Adjustment of treatment effect for seven predictors (clinical, radiological and hypoxia) gave a reduction in sample size requirements between 20% and 28%. Adjustment for just three core clinical predictors showed a reduction between 16% and 23%. As expected, adjustment for predictors in population-based TBI surveys usually gave larger reductions in sample size requirements than trials.

The first formal attempt to explore the potential reduction in sample size with covariate adjustment in the context of clinical trials with dichotomous outcomes was made by Choi (1998). He simulated a trial from data of a registry of 555 patients of the Medical College of Virginia. The registry was divided into 2 independent treatment groups (A and B): group A retained its observed probability of favorable outcome, and group B was reconstructed by changing a proportion of patients (10%) from unfavorable to favorable outcome. A 30% reduction in

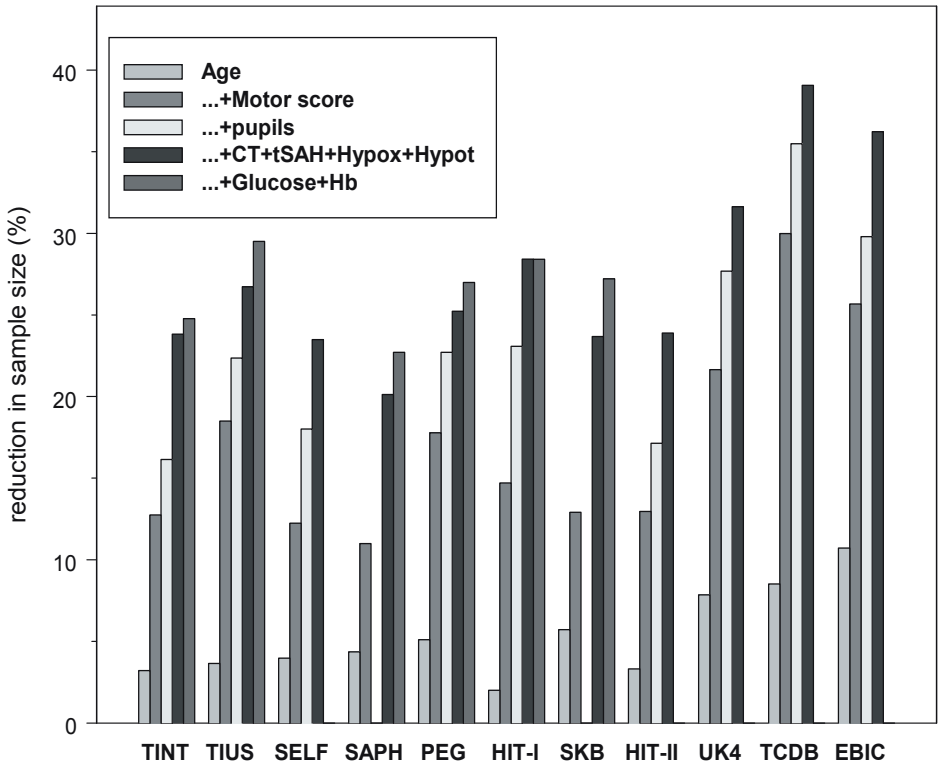

Figure 1. Reduction in sample size across studies for different logistic regression models. For each trial/survey: the first column to the left represents the model with age adjustment, the column in the middle the core model, and the last column the full model. Larger reductions in sample size are achieved in surveys (UK4, TCDB and EBIC) than in trials. Motor denotes GCS motor score; Pupils: pupillary reactivity; CT: Marshall CT classification; tSAH: traumatic subarachnoid hemorrhage; Hypox: hypoxia; Hypot: hypotension; Hb: hemoglobin.

sample size was found with adjustment for two strong predictors (age and GCS motor score).

The main concern with Choi's procedure is that the change to outcomes of group B patients was rather artificial. In clinical practice, a patient in a vegetative state is less likely to move to a good recovery state than a patient with severe disability with a given treatment. An alternative approach was described by Machado et al. (1999). They incorporated a positive treatment effect (OR= 0.57) into a logistic model with 3 predictors (age, GCS motor score, and Marshall CT

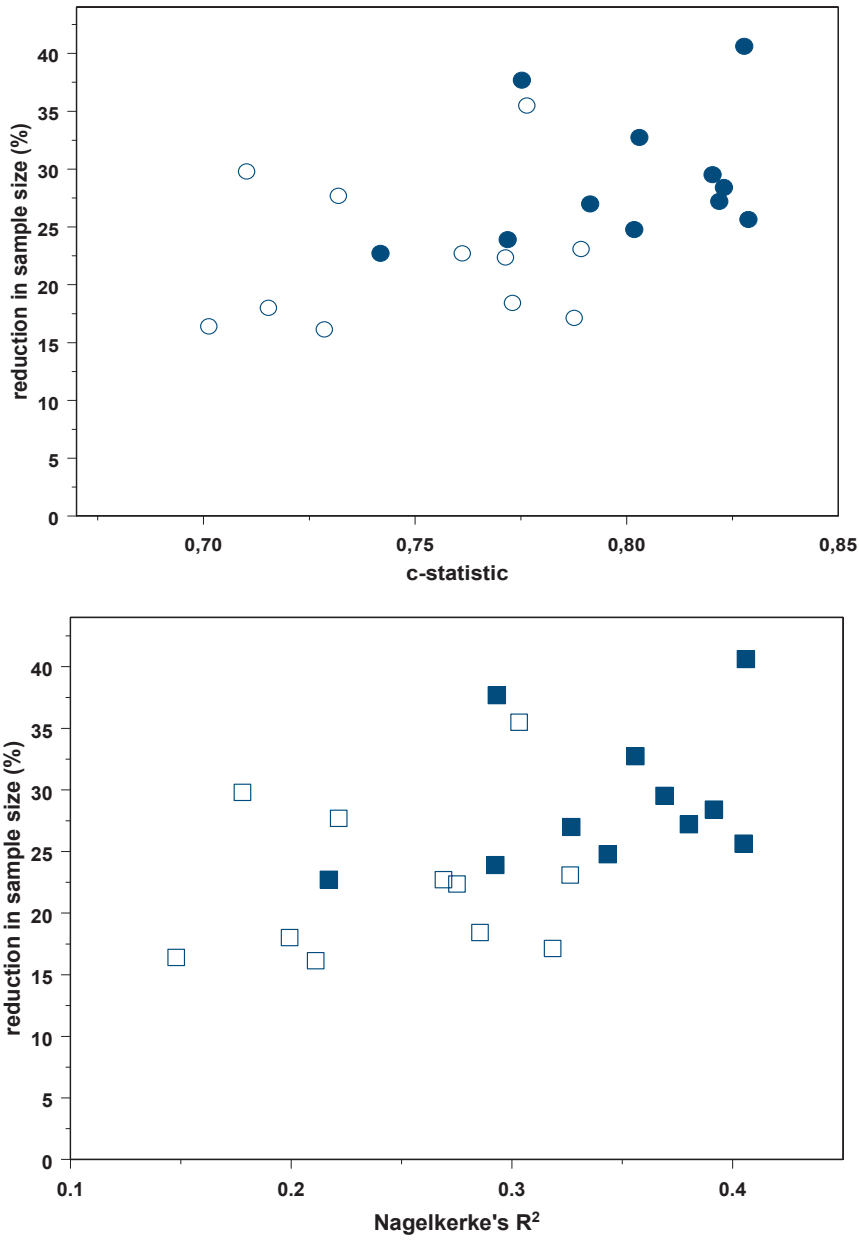

Figure 2. Relation between the reduction in the sample size, the c-statistic (Figure 2a), and Nagelkerke's R^2 (Figure 2b). Empty and solid symbols represent values for the core model and 7-predictor model, respectively. A positive relationship between the RSS and both the c-statistic and Nagelkerke's R^2 is observed.

classification) to create a new outcome. Thus, this outcome incorporated information from predictors, and patients with better prognosis had a greater chance of moving from an unfavorable to a favorable outcome in comparison to patients with a worse prognosis. We followed the latter approach, but found that Choi's approach gave very similar results (data not shown).

In our simulations the core model (age, GCS motor score, pupillary reactivity) achieved good discrimination, which improved with the inclusion of more covariates. This is consistent with findings that it is possible to make better predictions of GOS outcome with more predictors (Maas et al., 1999; Machado et al., 1999). Using the same seven predictors as in Hukkelhoven et al. (in press), improved the predictive ability of the core model substantially, but a further extension with glucose and hemoglobin yielded only minor added value.

A similar pattern was observed in the RSS calculations. A larger RSS was observed when the 7-predictor model was used in comparison to the 3-predictor model. Reductions in sample size of approximately 20% with the core model and around 24% with the full model in TBI trials are greater than those obtained in trials of other fields in medicine. For instance, adjustment for 17 predictors of 30-day mortality in patients with acute myocardial infarction included in the GUSTO-I trial reduced the sample size requirements by only 15% (Steyerberg et al., 2000).

Other approaches to achieve a RSS have been suggested. Machado et al. (1999) used data from the EBIC survey, and defined models on different TBI prognostic groups, based on three predictors. They quantified the power that could be achieved when the analysis was targeted to some prognostic groups (e.g. defined by clinical or radiological characteristics), and when different treatment effects were defined for different prognostic groups (e.g. treatment benefit only in patients with intermediate prognosis). For instance, for a moderate/severe TBI population and with a uniform treatment effect, a trial with 344 patients with intermediate prognosis per arm had the same 90% power as a trial with 500 unselected patients per arm –that is, the sample size could be reduced by 31% with no decrease in the power.

More recently, Murray et al. (2005) proposed another approach to analyze TBI trials with ordered outcomes (such as GOS), called the 'sliding dichotomy'. Using data from the Tirilazad International and Tirilazad USA trials, this approach increased the treatment OR from 1.38 to 1.71, and increased the statistical power to detect a significant treatment effect. These increases in power are equivalent to halving the trial sample size in a conventional analysis. Another approach is to use a proportional odds model. A formal comparison of these approaches to

quantify their benefits is required.

Our work has several limitations. First, information was incomplete for seven of the nine predictors, particularly glucose and hemoglobin. However, it was possible to replace missing values using a well-accepted imputation technique based on correlations with other variables (Little, 1992). Second, the RSS may vary in other trials according to their inclusion criteria. More or less restrictive criteria will produce less or more heterogeneous trial populations, respectively. Trials with more homogeneous patients may have a more limited benefit from covariate adjustment. Mega-trials, such as CRASH (n=10008), have considerable heterogeneity (CRASH trial collaborators, 2004, 2005). It is expected that covariate adjustment might yield a substantial RSS, although completeness of data for predictors may be of concern. Third, we used trials conducted between 1984 and 1997, and more recent trials may contain more or less heterogeneous populations. Finally, further work should indicate the incremental benefit to the RSS for the combination of the covariate adjustment with newer analysis approaches that exploit the ordinal nature of the outcome, such as the sliding dichotomy (Murray et al., 2005).

In conclusion, adjustment of the treatment effect for seven strong predictors yielded a reduction in sample size requirements by 25% in TBI trials, whilst adjustment for the three strongest predictors achieved a reduction of sample size of about 20%. Covariate adjustment for strong predictors should be incorporated in the design and analysis of future TBI trials.

References

ARIZA M., MATARO M., POCA M.A., et al. (2004). Influence of extraneurologic insults on ventricular enlargement and neuropsychological functioning after moderate and severe traumatic brain injury. J. Neurotrauma 21, 864-876.

CHOI S.C. (1998). Sample size in clinical trials with dichotomous endpoints: use of covariables. J. Biopharm. Stat. 8, 367-375.

CHOI S.C., CLIFTON G.L., MARMAROU A., MILLER E.R. (2002). Misclassification and treatment effect on primary outcome measures in clinical trials of severe neurotrauma. J. Neurotrauma 19, 17-22.

CRASH TRIAL COLLABORATORS. (2004). Effect of intravenous corticosteroids on death within 14 days in 10 008 adults with clinically significant head injury (MRC CRASH trial): randomised placebo-controlled trial. Lancet 364, 1321-1328.

CRASH TRIAL COLLABORATORS. (2005). Final results of MRC CRASH, a randomised placebo-controlled trial of intravenous corticosteroid in adults with head injury —outcomes at 6 months. Lancet 365, 1957-1959.

DICKINSON K., BUNN F., WENTZ R., EDWARDS P., ROBERTS I. (2000). Size and quality of randomised controlled trials in head injury: review of published studies. BMJ 320, 1308-1311.

DOPPENBERG E.M.R., CHOI S.C., BULLOCK R. (2004). Clinical trials in traumatic brain injury: lessons for the future. J. Neurosurg. Anesthesiol. 16, 87-94.

GAIL M.H., WIEAND S., PIANTADOSI S. (1984). Biased estimates of treatment effect in randomized experiments with non-linear regression and omitted covariates. Biometrika 71, 431-444.

HAUCK W.W., ANDERSON S., MARCUS S.M. (1998). Should we adjust for covariates in non-linear regression analyses of randomized trials?. Control. Clin. Trials 19, 248-256.

HERNANDEZ A.V., STEYERBERG E.W., HABBEMA J.D.F. (2004). Clinical trials with dichotomous end-points: covariate adjustment increases power and potentially reduces sample size. J. Clin. Epidemiol. 57, 454-460.

HERNANDEZ A.V., EIJKEMANS M.J.C., STEYERBERG E.W. (in press a). Randomized controlled trials with time-to-event outcomes: How much does pre-specified covariate adjustment increase power?. Ann. Epidemiol.

HERNANDEZ A.V., STEYERBERG E.W., TAYLOR G.S., MARMAROU A., HABBEMA J.D.F., MAAS A.I.R., for the IMPACT (International Mission for Prognosis of Head Injury and Analysis of Clinical Trials) Study Group. (in press b). Subgroup analysis and covariate adjustment in randomized clinical trials of traumatic brain injury: A systematic review. Neurosurgery.

HUKKELHOVEN C.W.P.M, STEYERBERG E.W., RAMPEN A.J.J., et al. (2003). Patient age and outcome following severe traumatic brain injury: an analysis of 5600 patients. J. Neurosurg. 99, 666-673.

HUKKELHOVEN C.W.P.M., STEYERBERG E.W., HABBEMA J.D.F., et al. (in press). Outcome after severe or moderate traumatic brain injury: development and validation of a prognostic score based on admission characteristics. J. Neurotrauma.

JENNETT B. (1996). Epidemiology of head injury. J. Neurol. Neurosurg. Psychiatry 60, 362-369.

JEREMITSKY E., OMERT L.A., DUNHAM C.M., WILBERGER J., RODRIGUEZ A. (2005). The impact of hyperglycemia on patients with severe brain injury. J. Trauma 58, 47-50.

LITTLE R.J.A. (1992). Regression with missing X's: a review. J. Am. Stat. Assoc. 87, 1227-1237.

LITTLE R., AN H. (2004). Robust likelihood-based analysis of multivariate data with missing values. Statistica Sinica 14, 949-968.

MAAS A.I.R., STEYERBERG E.W., MURRAY G.D., et al. (1999). Why have recent trials of neuroprotective agents in head injury failed to show convincing efficacy? A pragmatic analysis and theoretical considerations. Neurosurgery 44, 1286-1298.

MACHADO S.G., MURRAY G.D., TEASDALE G.M. (1999). Evaluation of designs for clinical trials of neuroprotective agents in head injury. J. Neurotrauma 16, 1131-1138.

MARSHALL L.F., MARSHALL S.B., KLAUBER M.R., et al. (1991). A new classification of head injury based on computarized tomography. J. Neurosurg. 75(Suppl), S14-S20.

MASSON F. (2000). Epidemiology of severe cranial injuries [in French]. Ann. Fr. Anesth. Reanim. 19, 261-269.

MURRAY G.D., BARER D., CHOI S., et al. (2005). Design and analysis of phase III trials with ordered outcome scales: the concept of the sliding dichotomy. J. Neurotrauma 22, 511-517.

MUSHKUDIANI N., STEYERBERG E.W., MAAS A.I.R. (2004). Prognostic models in traumatic brain injury based on admission data: A systematic review. Med. Decis. Making 24, 437.

NAGELKERKE N.J.D. (1991). A note on a general definition of the coefficient of determination. Biometrika 78, 691-692.

NARAYAN R.K., MICHEL M.E., AND THE CLINICAL TRIALS IN HEAD INJURY STUDY GROUP. (2002). Clinical trials in head injury. J. Neurotrauma 19, 503-557.

STEYERBERG E.W., BOSSUYT P.M., LEE K.L. (2000). Clinical trials in acute

myocardial infarction: should we adjust for baseline characteristics? Am. Heart J. 139, 745-751.

STEYERBERG E.W., HARRELL F.E. Jr., BORSBOOM G.J.J.M., EIJKEMANS M.J.C., VERGOUWE Y., HABBEMA J.D.F. (2001). Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. J. Clin. Epidemiol. 54, 774-781.

WILSON J.T.L. (2001). Assessing outcome in head injury trials. Curr. Pharm. Des. 7, 1537-1552.

Effects of platelet glycoprotein IIb/IIIa receptor blockers in non-ST-segment elevation acute coronary syndromes: Benefit and harm in different age subgroups.

4.2

ABSTRACT

Purpose
To investigate whether the effects of platelet glycoprotein (GP) IIb/IIIa receptor blockers in patients with non-ST-elevation acute coronary syndromes (NSTE-ACS) depend on patient age.

Methods
An individual data meta-analysis of 6 trials (PRISM, PRISM-PLUS, PARAGON-A, PURSUIT, PARAGON-B, GUSTO4-ACS; n=31,402) was performed. Patients were randomized to a GP IIb/IIIa receptor blocker or placebo/control. We applied multivariable logistic regression analyses to evaluate the drug effects on death or non-fatal MI at 30 days, as well as on major bleeding, by age subgroups (<60, 60-69, 70-79, ≥80 years). We quantified the reduction of death or MI as number needed to treat (NNT), and the increase of major bleeding as number needed to harm (NNH) across age subgroups.

Results
Overall, 11,155 (35%) patients were <60, 9,727 (31%) were 60-69, 8,468 (27%) were 70-79, and 2,049 (7%) were ≥80 years-old. The relative benefit of GP IIb/IIIa receptor blockers did not differ significantly (p=0.5) across age subgroups (odds ratio [95% CI] for death or MI: 0.86 [0.74-0.99], 0.90 [0.80-1.02], 0.97 [0.86-1.10], 0.90 [0.73-1.16]; overall 0.91 [0.86-0.99]). Odds ratios (95% CIs) for major bleeding were 1.9 (1.3-2.8), 1.9 (1.4-2.7), 1.6 (1.2-2.1), and 2.5 (1.5-4.1), respectively. The overall NNT was 105, and the overall NNH was 90. The oldest patients had larger absolute increases in major bleeding, but also had the largest absolute reductions of death or MI. Patients ≥80 years had a half of the NNT and a third of the NNH in comparison with patients <60 years.

Conclusions
In patients with NSTE-ACS, the relative reduction of death or non-fatal MI with GP IIb/IIIa receptor blockers was independent of patient age. Larger absolute outcome reductions were seen in the elderly, but with a higher risk of major bleeding. Thus, a close monitoring of these patients is warranted.

Introduction

Platelet glycoprotein (GP) IIb/IIIa receptor blockers decrease the risk of death or non-fatal myocardial infarction (MI) at 30 days in patients with non-ST elevation acute coronary syndromes (NSTE-ACS) who are not routinely scheduled for early revascularization [1-4]. Age is an important risk factor for these patients, and if the relative benefits of effective interventions are the same across age groups, physicians should treat the elderly even more aggressively than the younger, since the absolute benefit may be larger [5]. However, in clinical practice, the utilization of GP IIb/IIIa receptor blockers is lower among elderly patients [6].

Elderly patients may be undertreated because of several reasons: they were underrepresented or excluded from randomized clinical trials (RCTs), clinicians may believe that benefits in younger may not generalize to the elderly, or they may be worried about harmful effects in elderly patients [5]. Researchers have argued that the benefit of GP IIb/IIIa receptor blockers is greater in younger patients [7], similar in old and younger patients [8], or greater in older patients given their higher baseline risk [5, 9].

Yet, it is difficult to determine how the efficacy of GP IIb/IIIa receptor blockers varies among age subgroups because most trials are not large enough to provide a reliable answer. Individual ACS trials have been inconclusive or even conflicting regarding the presence or absence of relative differences in drug effects across ages [10-15]. Usually, the patient population was only split in two age groups (e.g. <65 years, ≥65 years) [11, 13-15], and different primary endpoints were considered. An evaluation of the drug effects across age groups in a meta-analysis using individual data can better define its relative and absolute efficacies in older vs. younger patients.

One more issue is relevant in the interpretation of the effects of GP IIb/IIIa receptor blockers by age groups. The incorporation of harmful major bleeding rates in the evaluation of effects should be considered to further understand the net drug effectiveness across age strata [5, 9, 16].

We investigated whether the relative effects of GP IIb/IIIa receptor blockers were consistent across age subgroups in non-ST-segment elevation ACS patients. Further, we evaluated whether the absolute benefits and harms differed across age subgroups.

Methods

Trial selection

A meta-analysis of individual patient data was performed, including trials reported since 1990 with the following characteristics: randomization of patients with NSTE-ACS, comparison of a GP IIb/IIIa receptor blocker with placebo or control therapy, no-recommendation for early (<48h) coronary revascularization during study-drug infusion, and enrolment of at least 1000 patients. Six trials met the inclusion criteria -PRISM, PRISM-PLUS, PARAGON-A, PURSUIT, PARAGON-B, and GUSTO IV-ACS- [10-15] with a total of 31,402 patients. Details of the trial designs are available elsewhere [3].

Patient baseline characteristics

An electronic database consisting of data from individual patients in all eligible trials was available [3]. These data were checked for completeness, for internal consistency of patients' records, and for consistency with the published reports. For this analysis, baseline characteristics regarded as important predictors of the outcome for which information was almost complete (i.e. less than 1% missing) were age, gender, diabetes, smoking, previous myocardial infarction [MI], previous heart failure [HF], previous coronary artery bypass surgery (CABG), previous percutaneous coronary intervention (PCI), and ST-segment depression. Other important predictors had more than 20% of missing data: blood pressure and heart rate were not recorded in the GUSTO IV-ACS trial (n=7800, 25%); and baseline creatine kinase MB (CK-MB) was missing in 7469 patients (24%) across different trials. Blood pressure, heart rate and CK-MB were used in addition to the other predictors in secondary analyses that yielded largely similar results.

Endpoints

For this analysis, the primary efficacy endpoint was defined a priori as the composite of death of any cause or non-fatal MI at 30 days. MI was part of the composite outcome of all trials. The MI definitions had subtle differences across trials regarding the CK-MB threshold [3]. However, all trials had pre-specified definitions of MI [17, 18]. Secondary endpoints were: death; non-fatal MI; coronary artery bypass graft (CABG); percutaneous coronary intervention (PCI); and CABG or PCI. The primary harm endpoint was major bleeding within 30 days. Individual trial definitions of major bleeding had also at most subtle differences, and trial-specific definitions were retained [3].

Efficacy analysis by age

We divided the patient data into four subgroups according to age: <60, 60-69, 70-79, and ≥80 years old. The decision to group patients in these intervals was made a priori, and was based on decade intervals of common clinical use. The choice of other cut-off points (e.g. quartiles) yielded similar results (not shown). Relative differences between GP IIb/IIIa receptor blockers and placebo/control on the primary endpoint by age subgroups were assessed, within each trial and across all trials. Logistic regression models were used, and odds ratios (OR) and corresponding 95% confidence intervals (95% CI) were calculated. To evaluate GP IIb/IIIa receptor blocker effect modification by age in each individual trial and in all trials, interaction tests were used [19]. These tests also evaluated heterogeneity of effects across trials. The effects of GP IIb/IIIa receptor blockers and the interactions were adjusted for the previously described predictors, for trial, and for potential differences in age-related trends between trials. These effects were combined using random effects calculations [20]. Heterogeneity of interactions across trials was evaluated with the random effects inverse variance model (with trial being the random effect) [21].

Benefit and harm of GP IIb/IIIa receptor blockers by age subgroups

We performed analyses that incorporated the relation among the baseline risk (eBR, proportion of patients in the placebo/control group with the primary efficacy endpoint), the efficacy Odds Ratio (eOR), and the respective number needed to treat [NNT]. The calculation of NNT was done using eBR and eOR, with the formula: [1-eBR(1-eOR)]/ [eBR(1-eBR)(1-eOR)] [22]. The NNT is the number of patients who need to be treated in order to prevent one additional death or non-fatal MI. It is the inverse of the absolute risk reduction (ARR). Further, we looked at the relation among the baseline proportion of the primary harm endpoint in the placebo/control group (hBR), the harm Odds Ratio (hOR), and the respective number needed to harm [NNH]. The NNH was calculated using hBR and hOR, with the formula: [hBR(hOR-1)+1]/[hBR(1-hBR)(hOR-1)] [22]. The NNH is the number of patients who need to be treated in order to cause one major bleeding. It is the inverse of the absolute risk increase (ARI). The NNT and NNH calculations were done overall and by age subgroups.

Role of the funding source

The trials included in this analysis were sponsored by several pharmaceutical companies, which are mentioned in the main trial reports [10-15], and in the acknowledgements. This study was designed, conducted, and interpreted independently of the sponsors. These had the right to review the manuscript, but not censor the findings. No separate industrial grant was obtained for this investigation.

Results

Age subgroups and predictors

Overall, 11,155 (35%) patients were < 60, 9,727 (31%) were 60-69, 8,468 (27%) were 70-79, and 2,049 (7%) were ≥ 80 years-old. Baseline characteristics across age subgroups are shown in Table 1. The proportion of women and of patients with a history of diabetes, MI or HF, and ST depression increased with age. Further, patients ≥80 years had lower proportions of previous revascularization procedures than younger patients. The proportion of patients older than 70 years ranged between 30% in the PURSUIT and PRISM trials and 40% in the GUSTO IV-ACS trial.

Endpoints at 30 days by age subgroups

The overall adjusted relative reduction in the odds of death or MI at 30 days was 9% (OR 0.91; 95% CI [0.85-0.99]). There was no difference in the relative benefit of GP IIb/IIIa receptor blockers across age subgroups (p for interaction = 0.5) and this was true also for secondary efficacy endpoints (Table 2). Interestingly, the ratio of non-fatal MI over death decreased with increasing age. The overall adjusted relative increase in the odds of major bleeding was 83% (OR 1.83 [1.5-2.2]). This was especially high for patients ≥80 years (OR 2.5 [1.5-4.1]), but there were no significant differences across ages (p for interaction=0.3) (Table 2).

Benefit of GP IIb/IIIa receptor blockers per trial by age subgroups

With regard to the incidence of death or non-fatal MI, two trials showed significantly different relative effects across age subgroups, but in opposite directions (Table 3). The PRISM trial patients had a clear gradient of GP IIb/IIIa receptor blocker effect across ages: older patients had larger odds reductions than younger ones (p for interaction=0.01). Conversely, younger PURSUIT patients had larger

	<60 years (n=11,155)		60-69 years (n=9,727)		70-79 years (n=8,468)		≥80 years (n=2,049)	
	N	%	N	%	N	%	N	%
Gender								
Male	8275	74	6274	65	4841	57	997	49
Diabetes								
Yes	1771	16	2360	24	2269	27	461	23
Smoking								
Never	3931	35	3439	36	3269	39	861	42
Former	3144	28	3537	37	3133	37	621	31
Current	4036	36	2709	28	2015	24	552	27
Previous MI								
Yes	3164	28	3445	36	3162	37	877	43
Previous HF								
Yes	578	5	962	10	1191	14	437	21
Previous CABG								
Yes	1088	10	1305	13	1194	14	185	9
Previous PCI								
Yes	1454	13	1251	13	956	11	162	8
ST depression								
Yes	5096	46	5475	57	5441	65	1403	69
Trial								
PRISM	1274	11	1005	10	781	9	172	8
PRISMPLUS	693	6	603	6	495	6	124	6
PARAGON-A	737	7	728	8	631	8	183	9
PURSUIT	4082	37	3553	37	2763	33	550	27
PARAGON-B	1976	18	1513	16	1374	16	362	18
GUSTO4-ACS	2393	21	2325	24	2424	29	658	32

MI denotes myocardial infarction, HF: heart failure, CABG: coronary artery bypass graft, PCI: percutaneous coronary intervention, y: years. Differences among age subgroups were highly significant ($p<0.001$).

Table 1. Patient characteristics by age subgroups.

odds reductions than the older ones (p for interaction=0.03). The interactions between GP IIb/IIIa receptor blockers and age subgroup were heterogeneous across trials (p=0.002).

Benefit and harm of GP IIb/IIIa receptor blocker across age subgroups

The absolute risk of death or MI at 30 days correlated with age, varying from 8% in the youngest (<60 years) to 21% in the oldest group (≥80 years). Major bleeding at 30 days also correlated with age, from 0.8% in the youngest to 2.3% in the oldest. For the overall relative reduction in the odds of death or MI of 9%,

	<60y (n=11,155)			60-69y (n=9,727)			70-79y (n=8,468)			≥80y (n=2,049)		
	Events	%	OR* (95% CI)	Events	%	OR (95% CI)	Events	%	OR (95% CI)	Events	%	OR (95% CI)
Death†												
GP IIb/IIIa	70	1.1	0.86	165	2.9	0.98	281	5.6	0.91	115	9.5	0.90
Placebo/Control	58	1.2	(0.61-1.23)	124	3.0	(0.77-1.24)	215	6.2	(0.75-1.09)	88	10.5	(0.67-1.21)
Nonfatal MI‡												
GP IIb/IIIa	372	5.7	0.83	428	7.6	0.85	437	8.8	1.02	112	9.3	0.91
Placebo/Control	316	6.8	(0.72-0.97)	365	8.8	(0.74-0.99)	299	8.6	(0.87-1.19)	85	10.1	(0.68-1.23)
Death or MI												
GP IIb/IIIa	442	6.8	0.86	593	10.6	0.90	718	14.4	0.97	227	18.8	0.90
Placebo/Control	374	8.0	(0.74-0.99)	489	11.9	(0.80-1.02)	514	14.8	(0.86-1.10)	173	20.5	(0.73-1.16)
CABG												
GP IIb/IIIa	828	12.7	1.00	931	16.6	0.92	860	17.2	0.99	102	8.5	1.07
Placebo/Control	590	12.7	(0.90-1.13)	732	17.7	(0.83-1.03)	603	17.3	(0.88-1.11)	67	8.0	(0.77-1.47)
PCI												
GP IIb/IIIa	1839	28.3	0.92	1369	24.4	1.02	894	17.9	0.89	171	14.2	0.90
Placebo/Control	1404	30.1	(0.84-0.99)	991	24.0	(0.93-1.12)	684	19.7	(0.80-1.00)	131	15.6	(0.70-1.15)
CABG or PCI												
GP IIb/IIIa	2618	40.3	0.93	2264	40.4	0.97	1721	34.5	0.93	268	22.2	0.93
Placebo/Control	1960	42.1	(0.86-1.00)	1699	40.8	(0.89-1.05)	1258	36.2	(0.85-1.02)	197	23.4	(0.76-1.15)
Major bleeding												
GP IIb/IIIa	90	1.5	1.90	118	2.3	1.94	174	3.8	1.58	63	5.7	2.46
Placebo/Control	35	0.8	(1.28-2.81)	46	1.1	(1.38-2.74)	80	2.3	(1.21-2.07)	19	2.3	(1.46-4.14)

* Odds ratio of treatment effect between GP IIb/IIIa and Placebo/Control, GP IIb/IIIa denotes platelet glycoprotein IIb/IIIa receptor blockers; †Death within 30 days; ‡ Non-fatal myocardial infarction in patients who survived at least 30 days. Number of patients per age group: <60 y: GP 6496, Placebo/control 4659; 60-69 y: GP 5602, Placebo/control 4125; 70-79 y: GP 4991, Placebo/control 3477; >80 y: GP: 1207, Placebo/control: 842.

Table 2. Treatment effect on various endpoints at 30 days according to age subgroups.

	PRISM N=3,232	PRISMPLUS N=1,915	PARAGON-A N=2,282	PURSUIT N=10,948	PARAGON-B N=5,225	GUSTO IV-ACS N=7,800	TOTAL* N=31,402
Age <60 years							
OR (95% CI)	1.13 (0.66-1.96)	0.98 (0.54-1.78)	1.65 (0.83-3.30)	0.72 (0.59-0.88)	0.90 (0.64-1.27)	1.01 (0.65-1.55)	0.86 (0.74-0.99)
Age 60-69 years							
OR (95% CI)	0.86 (0.53-1.38)	0.58 (0.35-0.96)	0.87 (0.55-1.39)	0.93 (0.77-1.20)	0.81 (0.59-1.12)	1.19 (0.85-1.67)	0.90 (0.80-1.02)
Age 70-79 years							
OR (95% CI)	0.63 (0.36-1.09)	1.02 (0.61-1.70)	0.83 (0.53-1.31)	0.91 (0.76-1.11)	1.11 (0.82-1.50)	1.15 (0.88-1.50)	0.97 (0.86-1.10)
Age ≥80 years							
OR (95% CI)	0.45 (0.19-1.07)	0.94 (0.39-2.27)	0.82 (0.37-1.81)	1.27 (0.87-1.86)	0.84 (0.48-1.47)	0.80 (0.52-1.22)	0.90 (0.73-1.16)
All subgroups, adjusted for predictors†	0.80 (0.60-1.06)	0.83 (0.62-1.11)	0.95 (0.72-1.25)	0.88 (0.79-0.98)	0.92 (0.78-1.10)	1.07 (0.90-1.27)	0.91 (0.86-0.99)
Age by GP IIb/IIIa Interaction (p)‡	0.01	0.77	0.15	0.03	0.92	0.52	0.15

* Odds ratios of each age subgroup adjusted for trial.
† Predictors included: age, gender, diabetes, smoking, previous MI, previous heart failure, previous CABG, previous PTCA, ST depression.
‡ Odds ratios adjusted for predictors, and age trend. The interactions age by GP IIb/IIIa are significantly different among trials. p: p-value.

Table 3. Treatment effects on death or MI at 30 days according to age subgroups, by trial and overall.

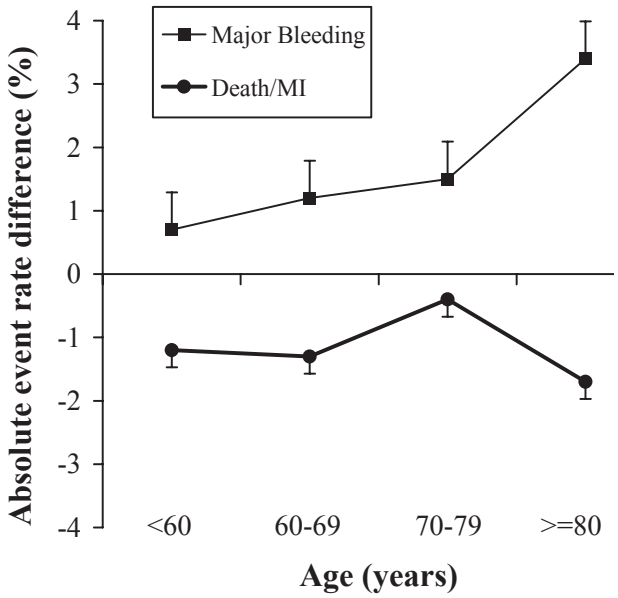

Figure. Absolute event rate differences between treatment arms (GP IIb/IIIa vs. placebo/control) by age subgroups in beneficial (death or myocardial infarction) and harmful (major bleeding) endpoints. GP IIb/IIIa denotes platelet glycoprotein IIb/IIIa receptor blockers.

the NNT was 105. For the overall relative increase in the odds of major bleeding of 83%, the NNH was 90.

The oldest patients had the largest absolute reductions of death or MI, but also had larger absolute increases in major bleeding. Patients younger than 70 years had higher NNTs and NNHs (149 and 163 for those younger than 60 years, and 105 and 110 for those between 60 and 69 years) than those older than 70 years (87 and 55 for those between 70 and 79 years, and 67 and 56 for those older than 80 years). The Figure shows the absolute event rate difference between GP IIb/IIIa receptor blocker and placebo/control arms across age subgroups. We noted a rather larger harm in patients ≥70 years and a somewhat variable benefit across all age subgroups.

Discussion

In patients with ACS without ST elevation, the relative reduction in the odds of death or MI at 30 days with GP IIb/IIIa receptor blockers was largely independent of age. The oldest patients had about 3-fold the baseline risk of the youngest ones, not only for death or MI, but also for major bleeding. In the oldest patients, the use of GP IIb/IIIa receptor blockers yielded larger absolute reductions of death/MI, but also larger absolute increases in major bleeding rates in comparison with the youngest patients.

This meta-analysis had more statistical power than individual trials to explore how the GP IIb/IIIa receptor blocker effects vary by age [7-9, 23]. Individual trials did not report these effects in detail across similar age subgroups [10, 11, 13-15], and they analyzed different endpoints. Previous analyses of the age effects in single trials have yielded inconclusive results [24]. Only the PURSUIT and GUSTO IV-ACS reported the same primary endpoint as we used in this paper. Also, these analyses did not adjust for important predictors of the primary endpoint. We found that the PRISM and the PURSUIT trials showed significant differential relative effects of GP IIb/IIIa receptor blockers across ages, but differences were in the opposite direction. We do not fully understand this phenomenon. We speculate that it could be related to the doses used as well as the duration of the study drug infusion. This might have resulted in different levels of platelet inhibition in the PRISM trial (where the dose was later shown to produce suboptimal platelet inhibition in young patients) as compared to the PURSUIT trial (where the dose was not adjusted for older age or modest renal impairment), which might have had different consequences in younger and older patients.

The effects of other antithrombotics have been evaluated in elderly patients with unstable angina or NSTE-ACS [9]. The low molecular weight heparin enoxaparin, compared with unfractionated heparin, appeared to have greater relative and absolute benefit in patients aged 65 years and older, as compared with younger patients. When comparing clopidogrel plus aspirin to aspirin alone, there was a consistent 20% relative reduction in cardiovascular death, myocardial infarction, or stroke in both elderly and younger patients. For GP IIb/IIIa receptor blockers, we found an equivalent relative benefit across age subgroups, which translated into a greater absolute benefit in older in comparison with younger patients.

Most trials, meta-analyses, and systematic reviews have neglected the contribution of major bleeding rates in the evaluation of the net GP IIb/IIIa receptor blocker effectiveness across age subgroups in NSTE-ACS patients [1-4, 10-15, 25-30]. Elderly patients have higher absolute risks of major bleeding [6, 31]. Therefore, the interpretation of the overall GP IIb/IIIa receptor blocker efficacy needs to

incorporate this harm. We should acknowledge that death or non-fatal MI and major bleeding do not have the same utility, and therefore are not comparable events. Determining the relative weights of these events is largely subjective. A recent review identified that the weight of a major bleeding related to a drug in the context of an acute coronary syndrome was 0.87, in comparison with the weight of death, which was equal to zero [32]. In order to describe the relative gain in primary efficacy end points by age subgroups, we defined a ratio of reduction of non-fatal MIs to reduction of deaths. For instance, a ratio higher than 1 shows a larger benefit in reduction of non-fatal MIs in comparison to reduction of deaths. Given that the ratio of non-fatal MI to death decreased with increased age, the use of GP IIb/IIIa receptor blockers in the oldest likely aborted more deaths than non-fatal MIs.

An appropriate dosing of GP IIb/IIIa receptor blockers is a requisite to obtain a higher benefit and a lower harm in elderly NSTE-ACS patients. The CRUSADE registry demonstrated that GP IIb/IIIa receptor blockers were underutilized and mis-dosed in elderly patients, who are at higher risk for adverse cardiac events [33]. An essential factor that increases the risk of major bleeding in elderly patients is low renal function, which is associated with higher serum levels of GP IIb/IIIa receptor blockers. Doses used in early trials were more aggressive than currently recommended doses, which are adjusted for renal dysfunction. Thus, elderly NSTE-ACS patients should receive adequate doses of GP IIb/IIIa receptor blockers to obtain the expected clinical benefit, and these doses should be adjusted for their level of renal function to avoid major bleeding events.

A recent decision analysis evaluated the efficacy of an unspecified potential drug on survival in patients with MI and unstable angina [5], and included serious adverse events (fatal complications) as an element of the evaluation of benefit-risk balance by age-related baseline risks. The authors used a registry database, and a hard primary endpoint (mortality at 1 year). The estimate of effectiveness was larger than in our randomized data (relative risk reduction 25%, absolute risk reduction 2%), and the registry population was more heterogeneous in risk (baseline risk of 2.3% in the youngest vs.27% in the oldest). They defined a threshold beyond which the treatment benefit would be outclassed by the treatment harm, and found that the fatal complication rate would have to be sevenfold greater in the oldest compared with the youngest age group to outweigh the survival benefits associated with treatment. These results need to be interpreted cautiously given that most major events in these patients do not lead to death. Moreover, retrospective observational data may sometimes inflate estimates of treatment efficacy [34].

Some limitations should be acknowledged. First, even with over 30,000 ran-

domized patients, subtle age interactions could have been missed, especially for rare events such as death. We did not see any age interactions for death based on the available data (not reported) and the clinical significance of subtle interactions is debatable. Second, a substantial amount of missing values for a few important predictors (blood pressure, heart rate, CK-MB) limited some possibilities of adjusted analysis. However, the results with imputed data yielded similar conclusions (not shown). Third, additional research into the appropriate weighting of events is needed, that can allow a more direct comparison between benefits and harms.

A series of nuances should be considered in interpreting these results. The trials included broad populations of patients with ACS. Through analysis of subgroups, it seems evident that higher risk patients, such as those with positive troponins, diabetes, and perhaps ST segment depression, achieve the greatest benefit. Further, it is likely that patients treated with the aggressive revascularization strategy achieve more benefit than those treated with the conservative strategy. The trials themselves were heterogeneous, as GUSTO IV-ACS showed no benefit and perhaps a detriment of abciximab, and PURSUIT used a very liberal definition of myocardial infarction that minimized the differences between eptifibatide and placebo. Finally, the category of major bleeding overestimates risk relative to the risk of blood transfusion, which is a more direct measure of risk and occurs less frequently (Mahaffey KW et al., Circulation, in press). The EARLY ACS trial is enrolling patients without age limits, it is testing whether the benefit of antithrombotic drugs is similar between elderly and young patients, and it is also addressing each of the above issues [35]. Allowing for these caveats, our analysis provides estimates for NNTs and NNHs by age subgroups that may be used in clinical decision making for the use of GP IIb/IIIa receptor blockers in NSTE-ACS patients.

In conclusion, the relative risk reduction of death or MI with GP IIb/IIIa receptor blocker is independent of age in patients with non-ST-elevation acute coronary syndromes. Larger absolute reductions of death or MI were observed in the oldest in comparison with the youngest patients, as well as larger absolute increases in major bleeding rates. Attention should be given to optimizing the benefit to elderly patients without increasing bleeding, by ensuring that doses adjusted for renal function are given. Moreover, elderly patients should be monitored more intensively.

Acknowledgements

The data included in this subgroup meta-analysis were provided by Merck Inc, White House Station, NJ, USA (sponsor of the PRISM and PRISM-PLUS trials); Hoffman La-Roche, Basel, Switzerland (sponsor of PARAGON-A and PARAGON-B trials); COR Therapeutics Inc, San Francisco, CA, USA, and Schering-Plough Inc, Kenilworth, NJ, USA (sponsors of the PURSUIT trial); and Centocor Inc, Malvern, PA, USA (sponsor of the GUSTO IV-ACS trial). Dr. Adrián V. Hernández received support from the Netherlands Organization for Scientific Research (ZON/MW 908-02-117).

References

1. Lincoff AM, Califf RM, Topol EJ. Platelet glycoprotein IIb/IIIa receptor blockade in coronary artery disease. J Am Coll Cardiol 2000; 35: 1103-15.
2. Bhatt DL, Topol EJ. Current role of platelet glycoprotein IIb/IIIa inhibitors in acute coronary syndromes. JAMA 2000; 284: 1549-58.
3. Boersma E, Harrington RA, Moliterno DJ, et al. Platelet glycoprotein IIb/IIIa inhibitors in acute coronary syndromes: a meta-analysis of all major randomised clinical trials. Lancet 2002; 359: 189-98.
4. Schulman SP. Antiplatelet therapy in non-ST-segment elevation acute coronary syndromes. JAMA 2004; 292: 1875-82.
5. Alter DA, Manuel DG, Gunraj N, Anderson G, Naylor CD, Laupacis A. Age, risk-benefit trade-offs, and the projected effects of evidence-based therapies. Am J Med 2004; 116: 540-5.
6. Avezum A, Makdisse M, Spencer F, et al. Impact of age on management and outcome of acute coronary syndrome: Observations from the Global Registry of Acute Coronary Events (GRACE). Am Heart J 2005; 149: 67-73.
7. Thompson SG, Higgins JP. Can meta-analysis help target interventions at individuals most likely to benefit?. Lancet 2005; 365: 341-6.
8. Mak KH, Effron MB, Moliterno DJ. Platelet glycoprotein IIb/IIIa receptor antagonists and their use in elderly patients. Drugs Aging 2000; 16: 179-87.
9. Cannon CP. Elderly patients with acute coronary syndromes: higher risk and greater benefit from antiplatelet therapy and/or interventional therapies. Am J Geriatr Cardiol 2003; 12: 259-62.
10. The PRISM Study Investigators. A comparison of aspirin plus tirofiban with aspirin plus heparin for unstable angina. N Engl J Med 1998; 338: 1498-505.
11. The PRISM-PLUS Study Investigators. Inhibition of the platelet glycoprotein IIb/IIIa receptor with tirofiban in unstable angina and non-Q-wave myocardial infarction. Platelet Receptor Inhibition in Ischemic Syndrome Management in Patients Limited by Unstable Signs and Symptoms. N Engl J Med 1998; 338: 1488-97.

12. The PARAGON Investigators. International, randomized, controlled trial of lamifiban (a platelet glycoprotein IIb/IIIa inhibitor), heparin, or both in unstable angina. Platelet IIb/IIIa Antagonism for the Reduction of Acute coronary syndrome events in a Global Organization Network. Circulation 1998; 97: 2386-95.
13. The PURSUIT Trial Investigators. Inhibition of platelet glycoprotein IIb/IIIa with eptifibatide in patients with acute coronary syndromes. N Engl J Med 1998; 339: 436-43.
14. The GUSTO IV-ACS Investigators. Effect of glycoprotein IIb/IIIa receptor blocker abciximab on outcome in patients with acute coronary syndromes without early coronary revascularisation: the GUSTO IV-ACS randomised trial. Lancet 2001; 357: 1915-24.
15. The PARAGON-B Investigators. Randomized, placebo-controlled trial of titrated intravenous lamifiban for acute coronary syndromes. Circulation 2002; 105: 316-21.
16. Patrono C, Coller B, FitzGerald GA, Hirsh J, Roth G. Platelet-active drugs: The relationships among dose, effectiveness, and side effects. The seventh ACCP conference on antithrombotic and thrombolytic therapy. Chest 2004; 126: 234S-264S.
17. Early Breast Cancer Trialists'Collaborative Group. Treatment of early breast cancer, vol 1: worldwide evidence 1985-1990. Oxford: Oxford University Press, 1990: 12-8.
18. Chan A-W, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials. Comparison of protocols to published articles. JAMA 2004; 291: 2457-65.
19. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355: 1064-9.
20. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986; 7: 177-88.
21. Clarke M, Oxman A, eds. Cochrane reviewers' handbook, version 4.2.0 (updated March 2003). In: The Cochrane Library, issue 4. Chichester: John Wiley & Sons, Ltd, 2003.
22. McQuay HJ, Moore RA. Using numerical results for systematic reviews in clinical practice. Ann Intern Med 1997; 126: 712-20.
23. Brookes ST, Whitley E, Egger M, Davey Smith G, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol 2004; 57: 229-36.
24. Hasdai D, Holmes DR Jr, Criger DA, Topol EJ, Califf RM, Harrington RA, for the PURSUIT trial investigators. Age and outcome after acute coronary syndromes without persistent ST-segment elevation. Am Heart J 2000; 139: 858-66.
25. Alexander JH, Harrington RA. Recent antiplatelet drug trials in acute coronary syndromes. Clinical interpretation of PRISM, PRISM-PLUS, PARAGON A and PURSUIT. Drugs 1998; 56: 965-76.
26. Vorchheimer DA, Badimon JJ, Fuster V. Platelet glycoprotein IIb/IIIa receptor

antagonists in cardiovascular disease. JAMA 1999; 281: 1407-14.
27. Casserly IP, Topol EJ. Glycoprotein IIb/IIIa antagonists – from the bench to practice. Cell Moll Life Sci 2002; 59: 478-500.
28. De Caterina R, Di Gioacchino L. Glycoprotein IIb-IIIa inhibitors in unstable coronary syndromes and percutaneous interventions – a conservative approach. Rev Port Cardiol 2003; 22: 995-1002.
29. Januzzi JL, Cannon CP, Theroux P, Boden WE. Optimizing glycoprotein IIb/IIIa receptor antagonist use for the non-ST-segment elevation acute coronary syndromes: risk stratification and therapeutic intervention. Am Heart J 2003; 146: 764-74.
30. Atwater BD, Roe MT, Mahaffey KW. Platelet glycoprotein IIb/IIIa receptor antagonists in non-ST segment elevation acute coronary syndromes. A review and guide to patient selection. Drugs 2005; 65: 313-24.
31. Ali Raza J, Movahed A. Use of cardiovascular medications in the elderly. Int J Cardiol 2002; 85: 203-15.
32. Tengs TO, Wallace A. One thousand health-related quality-of-life estimates. Med Care 2000; 38: 583-637.
33. Hoekstra JW, Roe MT, Peterson ED, et al. Early glycoprotein IIb/IIIa inhibitor use for non-ST-segment elevation acute coronary syndrome: patient selection and associated treatment patterns. Acad Emerg Med 2005; 12: 431-8.
34. Ioannidis JPA, Haidich A-B, Papa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 2001; 286: 821-30.
35. Giugliano RP, Newby LK, Harrington RA, et al. The Early Glycoprotein IIb/IIIa Inhibition in Non-ST-Segment Elevation Acute Coronary Syndrome (EARLY ACS) trial: A randomized placebo-controlled trial evaluating the clinical benefits of early front-loaded eptifibatide in the treatment of patients with non-ST-segment elevation acute coronary syndrome—Study design and rationale. Am Heart J 2005; 149: 994-1002.

Predictors of stroke within 30 days in patients with non-ST-segment elevation acute coronary syndromes

4.3

ABSTRACT

Purpose
Stroke is an uncommon but serious complication in patients with non-ST-segment elevation acute coronary syndromes (NSTE-ACS). We studied baseline patient characteristics that predict the development of stroke at 30 days.

Methods
We pooled data from 6 trials (n=31402) that randomized NSTE-ACS patients either to platelet glycoprotein (GP) IIb/IIIa receptor blockers or placebo/control. Potential predictors of stroke included treatment, and demographic and clinical characteristics. We quantified predictors using univariable and multivariable logistic regression models, and their performance was evaluated with calibration (Hosmer-Lemeshow test) and discrimination (c-statistic).

Results
We found 228 (0.7%) all-cause strokes: 155 (0.5%) non-hemorrhagic, 20 (0.06%) hemorrhagic, and 53 without CT confirmation. Patients with any type of stroke had a 30-day mortality of 25%. Randomization to GP IIb/IIIa receptor blockers was not significantly associated with all-cause stroke (OR [95% CI] 1.08 [0.83 – 1.41]. Older age (OR per increase of 10 years 1.5 [1.3-1.7]), prior stroke (2.1 [1.4-3.1]) and elevated heart rate (per increase of 10 beats 1.1 [1.0-1.2]) were the strongest predictors of 30-day all-cause stroke. Similar predictors were found for non-hemorrhagic and hemorrhagic strokes. Secondary predictors of all-cause stroke included smoking, previous myocardial infarction, diabetes and hypertension. The multivariable model to predict all-cause stroke was well calibrated, but the discriminative power was only moderate (c-statistic 0.69 [0.65 – 0.72]).

Conclusions
Stroke is a rare complication occurring early after NSTE-ACS, but is associated with high mortality. We found no evidence that GP IIb/IIIa receptor blockers increase stroke risks. There were few clinical characteristics that predicted a higher stroke risk. Thus, incident strokes in NSTE-ACS patients remain largely unexplained.

Introduction

The non-ST-segment elevation acute coronary syndrome (NSTE-ACS) is a heterogeneous disease. Risk stratification is essential for predicting prognosis, planning treatment strategy, and providing information to patients and relatives [1, 2]. Previous papers in patients with NSTE-ACS have evaluated the predictors associated with a range of clinical outcomes at 30 days or 6 months, such as death, cardiovascular death, and cardiovascular death or myocardial infarction (MI) [2-5].

Stroke is an uncommon but severe event in patients presenting with NSTE-ACS. Analyses with a few events in the PURSUIT trial found several clinical predictors of non-hemorrhagic stroke at 30 days [6]. These patients are also at increased risk for hemorrhagic strokes from polypharmacy anticoagulation. However, the confirmation of the importance of these predictors of stroke with a larger number of patients and events is desirable.

We aimed to identify baseline clinical and demographic patient characteristics that predict the development of all-cause, non-hemorrhagic and hemorrhagic strokes within 30 days. We analyzed 31,387 patients with NSTE-ACS from 6 large international trials. Moreover, we evaluated whether the use of GP IIb/IIIa receptor blockers was associated with an increased risk of stroke.

Methods

Clinical trials

We used individual patient data from 6 trials (PRISM, PRISM-PLUS, PARAGON-A, PURSUIT, PARAGON-B, and GUSTO IV-ACS) [7-12]. These trials were reported since 1990 with the following characteristics: randomization of patients with NSTE-ACS, comparison of platelet glycoprotein (GP) IIb/IIIa receptor blockers with placebo or control therapy, no-recommendation for early (<48h) coronary revascularization during study-drug infusion, and enrolment of at least 1000 patients. A total of 31,402 patients participated in these trials. Details of the trial designs are available elsewhere [13].

Potential predictors

An electronic database consisting of data from individual patients in all eligible trials was available [13]. These data were checked for completeness, for internal

consistency of patients' records, and for consistency with the published reports. For this analysis, we used available baseline demographic and clinical characteristics regarded as potential predictors of stroke [6]. Those with almost complete information (<1% of missing values) included age, gender, smoking, weight, and prior history all the following: hypertension, diabetes, stroke, MI, heart failure, angina pectoris, coronary artery bypass surgery (CABG), percutaneous coronary intervention (PCI), and use of aspirin. Two variables had 2% of missing values: history of hypercholesterolemia, and ST-depression at baseline.

Other variables had more than 20% of missing data: race, heart rate, blood pressure (systolic and diastolic), and prior use of beta blockers, ACE inhibitors, nitrates and calcium antagonists. Blood pressure and heart rate were not recorded in the GUSTO IV-ACS trial (n=7800); baseline creatine kinase MB (CK-MB) was missing in 7469 patients across different trials. Predictors with more than 20% of missing were imputed using the EM (estimated mean) procedure in SPSS (SPSS Inc., Chicago IL, USA, 1999). Atrial fibrillation and creatinine clearance were not available. The body mass index could not be calculated (i.e. no height was available) and it was not included in the analysis. The use of GP IIb/IIIa receptor blockers was also included as a potential predictor of stroke.

Outcomes

For this analysis, the primary outcomes defined a priori were all-cause stroke, non-hemorrhagic stroke, and hemorrhagic stroke within 30 days of the index ACS. Non-hemorrhagic and hemorrhagic strokes needed CT confirmation. All-cause stroke was missing for 12 patients. Non-hemorrhagic stroke was missing in 7434 patients, and hemorrhagic stroke was missing in 7474 patients. No formal attempt to impute these outcomes was done.

Statistical analysis

Univariable logistic regression models were used to evaluate the association between each potential predictor and the outcome. The predictive weight of each variable was expressed as a χ^2 statistic, which was calculated on the –2 log likelihood scale. The higher the number, the more important the predictor; a χ^2 exceeding 3.84 corresponds to $p<0.05$ for a predictor with 1 degree of freedom. All predictors were entered in a multivariable logistic regression model without further selection to properly evaluate their predictive effects while adjusting for the effects of each other predictor [14].

The performance of the multivariable models was studied with respect to discrimination and calibration. Discrimination refers to the ability to distinguish a stroke from no stroke. It was quantified by a measure of concordance, the c-statistic. For binary outcomes the c-statistic is identical to the area under the receiver operating characteristic (ROC) curve. The c-statistic lies between 0.5 and 1, and is better if closer to one [15]. Since the apparent c-statistic is optimistic with low numbers of events, we used a standard bootstrapping procedure to correct the estimates [14, 15]. Calibration refers to whether the predicted risks agree with the observed risk frequencies. Calibration was measured with the Hosmer-Lemeshow goodness-of-fit test [16]. Analyses were performed in SPSS 10.0 and S-PLUS 2000 (Insightful Inc, Seattle WA, USA).

Results

Patient characteristics

We found 228 (0.7%) all-cause strokes in the study population: 155 (0.5%) were non-hemorrhagic, 20 (0.06%) hemorrhagic, and 53 (0.2%) without CT confirmation. Older patients, those with a prior stroke, prior MI, diabetes, hypertension, and patients with elevated heart rate had higher risks of all-cause and non-hemorrhagic strokes (Table 1). Smoking was not clearly related with the stroke incidence. Patients with previous PTCA were at lower risk to develop any stroke. Less clear associations were seen in hemorrhagic strokes, probably due to small numbers.

A high proportion of patients who suffered a stroke died: 56 (25%) of those with all-cause stroke, 27 (17%) of those with non-hemorrhagic stroke and 13 (65%) of those with hemorrhagic stroke. The difference in mortality between non-hemorraghic and hemorrhagic strokes was highly statistically different ($p<0.001$). No clear relation was observed between predictors and death in patients who suffered any type of stroke (Table 1).

Predictors of stroke

The use of GP IIb/IIIa receptor blockers was not associated with a higher incidence of all-cause (OR [95% CI] 1.08 [0.83-1.41]), non-hemorrhagic (1.06 [0.77-1.47]), and hemorrhagic (1.70 [0.65-4.45]) strokes.

The strongest univariable predictors of all-cause stroke were older age

	All-cause strokes n=228			Non-hemorrhagic strokes n=155			Hemorrhagic strokes n=20		
Predictor	**N**	**n (%)**	**deaths (%)†**	**N**	**n (%)**	**deaths (%)†**	**N**	**n (%)**	**deaths (%)**
Age*									
<70 y	20874	98 (0.5)	20 (20)	15836	69 (0.4)	10 (15)	15822	6 (0.03)	4 (67)
≥70y	10513	130 (1.2)	36 (28)	8132	86 (1.1)	17 (20)	8108	14 (0.17)	9 (64)
Prior stroke*									
No	29890	201 (0.7)	50 (25)	22777	134 (0.6)	24 (18)	23744	16 (0.07)	10 (63)
Yes	1446	27 (1.9)	6 (22)	1141	21 (1.8)	3 (14)	1134	4 (0.35)	3 (75)
Heart rate									
<75	16807	104 (0.6)	18 (17)	12577	70 (0.6)	11 (16)	12564	10 (0.08)	5 (50)
≥75	14580	124 (0.9)	23 (19)	11391	85 (0.7)	10 (12)	11364	10 (0.09)	8 (80)
Smoking									
Never	11499	68 (0.6)	15 (22)	9516	55 (0.6)	8 (15)	9511	7 (0.07)	4 (57)
Former	10429	91 (0.9)	21 (23)	7577	55 (0.7)	10 (18)	7557	6 (0.08)	4 (67)
Current	9307	68 (0.7)	20 (29)	6768	44 (0.7)	9 (20)	6753	7 (0.10)	5 (71)
Prior MI									
No	20648	125 (0.6)	31 (25)	16345	90 (0.6)	14 (16)	16317	14 (0.09)	9 (64)
Yes	10646	103 (1.0)	25 (24)	7531	65 (0.9)	13 (20)	7519	6 (0.08)	4 (67)
Diabetes mellitus									
No	24488	159 (0.6)	43 (27)	18612	106 (0.6)	18 (17)	18590	16 (0.08)	11 (69)
Yes	6860	68 (1.0)	12 (18)	5317	49 (0.9)	9 (18)	5299	4 (0.08)	2 (50)
Hypertension									
No	14417	81 (0.6)	21 (26)	10908	60 (0.6)	11 (18)	10891	7 (0.06)	5 (71)
Yes	16935	147 (0.9)	35 (24)	13025	95 (0.7)	16 (17)	13002	13 (0.09)	8 (62)
GP IIb/IIIa RB ‡									
No	13097	91 (0.7)	21 (23)	9928	62 (0.6)	9 (15)	9908	6 (0.06)	2 (33)
Yes	18290	137 (0.8)	35 (26)	14040	93 (0.7)	18 (19)	14020	14 (0.09)	11 (79)

* P<0.001 for the comparison between categories
† Deaths within 30 days. The percentage refers to the number of deaths in patients who suffered a stroke
‡ Denotes platelet glycoprotein IIb/IIIa receptor blocker

Table 1: Distribution of patient baseline characteristics across stroke types (all-cause, non-hemorrhagic, and hemorrhagic).

(χ^2=69), prior stroke (χ^2=19), prior MI (χ^2=12), hypertension (χ^2=10), elevated heart rate (χ^2=9), lighter weight (χ^2=9), diabetes (χ^2=8), and smoking (χ^2=6). The associations are shown in Table 2. No interactions between predictors were statistically significant. The three most important predictors were older age (OR [95% CI] per 10 years: 1.5 [1.3-1.7]), prior stroke (2.1 [1.4-3.1]), and elevated heart rate (per 10 beats: 1.1 [1.0-1.2]). Multivariable predictors with relatively minor effects included smoking, prior MI, diabetes mellitus and hypertension.

The strongest univariable predictors of non-hemorrhagic stroke were older age (χ^2=38), prior stroke (χ^2=18), elevated heart rate (χ^2=9), prior MI (χ^2=7), and diabetes (χ^2=7). Lighter weight (χ^2=4), and hypertension (χ^2=3) had minor importance. The three most important predictors of non-hemorrhagic stroke had

Predictors	All-cause strokes		Non-hemorrhagic strokes		Hemorrhagic strokes	
	Univariable	Multivariable	Univariable	Multivariable	Univariable	Multivariable
Main						
Age (per 10 years)	1.68 (1.48-1.91)	1.51 (1.31-1.74)	1.59 (1.37-1.86)	1.45 (1.22-1.73)	2.26 (1.41-3.60)	2.17 (1.30-3.64)
Prior stroke	2.81 (1.87-4.21)	2.06 (1.36-3.12)	3.17 (1.99-5.04)	2.36 (1.46-3.80)	5.03 (1.68-15.06)	3.76 (1.21-11.70)
Heart rate (per 10 beats)	1.11 (1.05-1.19)	1.11 (1.04-1.18)	1.13 (1.05-1.20)	1.13 (1.05-1.21)	1.10 (0.95-1.28)	1.15 (0.96-1.37)
Secondary						
Smoking						
Former	1.48 (1.08-2.03)	1.45 (1.05-2.02)	1.26 (0.86-1.83)	1.25 (0.84-1.84)	1.08 (0.36-3.21)	1.75 (0.85-5.65)
Current	1.24 (0.88-1.73)	1.37 (0.98-1.95)	1.13 (0.76-1.67)	1.50 (0.99-2.27)	1.41 (0.49-4.02)	2.48 (0.83-7.40)
Prior MI	1.60 (1.23-2.08)	1.32 (0.99-1.77)	1.57 (1.14-2.16)	1.23 (0.86-1.75)	0.93 (0.36-2.42)	0.82 (0.29-2.38)
Diabetes mellitus	1.53 (1.15-2.04)	1.26 (0.93-1.69)	1.62 (1.16-2.28)	1.35 (0.94-1.93)	0.88 (0.29-2.62)	0.79 (0.26-2.43)
Hypertension	1.55 (1.18-2.03)	1.21 (0.91-1.61)	1.32 (0.96-1.84)	1.00 (0.70-1.40)	1.56 (0.62-3.90)	1.22 (0.46-3.22)

Table 2: Univariate and multivariate OR (95% CI) of predictors of stroke in NSTE-ACS patients.

comparable associations as those described for all-cause stroke. For hemorrhagic strokes, the strongest univariable predictors were older age (χ^2=12), prior stroke (χ^2=8), and lighter weight (χ^2=5). Similarly, the three most important predictors were those of the non-hemorrhagic strokes (Table 2).

Performance of predictive models

The calibration of the predictive model of all-cause stroke was good (Hosmer-Lemeshow test 10.4, p=0.24), but the discriminative power of this model was moderate (c-statistic [95% CI]: 0.69 [0.65-0.72]). Although the calibration of the predictive models of non-hemorrhagic and hemorrhagic strokes was good, the discriminative power was either moderate (c-statistic 0.67 [0.63-0.71]) or poor (c-statistic 0.58 [0.54-0.63]), respectively.

Discussion

Stroke occurred in 0.7% of patients within 30 days of presenting with NSTE-ACS. Two thirds of the strokes were non-hemorrhagic. Older age, prior stroke and elevated heart rate were the strongest predictors of all-cause, non-hemorrhagic, and hemorrhagic strokes. However, the discriminative power of these predictors was moderate, and especially poor for hemorrhagic strokes. Thus, it is difficult to accurately predict the incidence of stroke in this population.

The incidence of 30-day all-cause stroke in our patients is comparable to the incidence in similar populations: 0.8% in the GUSTO-IIb trial [17], and 0.5% in the OPUS-TIMI 16 trial [18]. However, in clinical practice the incidence of 30-day all-cause stroke may be larger, because the in-hospital incidence already reaches 0.7% [19]. In a Spanish nationwide registry (DESCARTES) the incidence of 30-day all-cause stroke was 0.9% (95% CI 0.4-1.3%) [20]. For comparison, the incidence of 30-day all-cause stroke in patients with ST-segment elevation-ACS (STE-ACS) treated with thrombolytics was 1.4% (between 1.2% and 1.6%) in the GUSTO-I trial [21], and 0.8% in nine trials from a meta-analysis [22]. The VALIANT registry, including both NSTE- and STE-ACS patients, had 1.5% in-hospital strokes [23]. The proportion of hemorrhagic strokes was around 50% of the total number of strokes in the GUSTO-I trial [21], and 13% in the meta-analysis [22]. Strokes in NSTE-ACS patients were associated with a high mortality rate (25%), which is lower than that observed in STE-ACS patients (41%) [21].

Importantly, the use of GP IIb/IIIa receptor blockers was not clearly associated with an increased incidence of all-cause stroke, non-hemorrhagic stroke or

hemorrhagic stroke. However, it should be recognized that the conclusion about the effect of GP IIb/IIIa receptor blockers on hemorrhagic strokes has substantial uncertainty, given the low numbers of events available and, hence, the limited power of the statistical analysis. The low frequency of hemorrhagic stroke in the overall population, coupled with lack of clear evidence of increased risk, provides reassurance that fear of intracranial hemorrhage should not be a reason to avoid these drugs. However, when patients receive these drugs on top of more aggressive antithrombotic therapy, the incidence of hemorrhagic strokes increases, as in patients with STE-ACS who received thrombolytics [24]. In our NSTE-ACS patients, predictors associated with the incidence of hemorrhagic stroke were similar to those associated with non-hemorrhagic stroke. In contrast, STE-ACS patients who take oral anticoagulation before admission, with less than 70 kg, and older than 65 years were at increased risk of hemorrhagic stroke.

Stroke has only been studied as an outcome in a secondary analysis of the PURSUIT trial [6]. Sixty-six non-hemorrhagic strokes in 9461 NSTE-ACS patients were studied. Hemorrhagic strokes were not studied. The strongest predictors were higher heart rate, older age, prior anterior MI, prior stroke or transient ischemic attack, and diabetes mellitus. Our analysis of 6 trials with a sample size of 31,387 patients increased the number of events and the power to find predictors of any type and all-cause stroke. However, the number of hemorrhagic strokes was still limited.

Age was an important predictor of non-hemorrhagic stroke in the PURSUIT trial [6], and the GUSTO-I trial [25]. In our analysis, age was the strongest predictor of all-cause, non-hemorrhagic, and hemorrhagic strokes, and its relative importance was slightly higher than the results of the PURSUIT trial. Elderly patients probably have a higher risk of stroke due to multiple co-morbidities associated with older age, such as atrial fibrillation, hypertension, physical inactivity, and asymptomatic carotid stenosis [26].

Prior stroke has been described as a predictor of stroke in the OPUS-TIMI 16 trial [27]. In this trial, the proportion of 10-month all-cause stroke was 2.9% in 1,173 patients with prior extra-cardiac vascular disease (peripheral + stroke + transient ischemic attack [TIA]) in comparison with 1.1% in 9108 patients without prior extra-cardiac vascular disease. In the PURSUIT trial and in the GUSTO-I trial [25] prior stroke was analyzed in conjunction of prior TIA, and this combined predictor was important. Prior stroke may be a marker of underlying cardiac, carotid or cerebral vascular disease in ACS patients.

Elevated heart rate was very important in the PURSUIT trial [6], and in the GUSTO-I trial [25]. An explanation for the association between elevated heart rate

and stroke is not clear [6]. The heart rate may correlate with larger infarctions that predispose patients to a higher likelihood of atrial arrhythmia and left ventricular thrombi. Heart rate is strongly associated with the presentation of atrial fibrillation in patients with NSTE-ACS [28]. Atrial fibrillation is a common complication of these patients, occurring in 6.4% of patients enrolled [29]. Moreover, an elevated baseline heart rate may simply be an expression of a prior atrial fibrillation. Unfortunately, our dataset did not provide information over prior or incident atrial fibrillation. Finally, a high heart rate may be an expression of a decompensated heart failure, related to the extent of the MI. Heart failure on admission has been described as an independent predictor of in-hospital all-cause stroke in the VALIANT registry [23].

Diabetes and prior MI were important predictors of stroke in the PURSUIT trial [6], but not in our analysis. Diabetes has a known association with a widespread atherosclerosis, and prior MI is associated with the formation of mural thrombus and emboli. Finally, lighter weight was weakly associated with hemorrhagic stroke. This was probably related to doses of GP IIb/IIIa receptor blockers and anticoagulants that were not reduced in lighter patients, and especially for the elderly.

Our study has some limitations. We had about 7500 patients with missing values for the non-hemorrhagic and hemorrhagic stroke outcomes. The number of non-hemorrhagic strokes was still larger (n=155) than the largest previously published (n=66) [6]. However, the number of hemorrhagic strokes was small (n=20), and this limited the conclusions about the predictors of hemorrhagic stroke. We imputed several patient characteristics. Of them, only heart rate remained as strong predictor, as was demonstrated previously [6].

In conclusion, stroke is an infrequent but serious early complication of patients with NSTE-ACS. Mortality is high, especially for hemorrhagic strokes. Platelet GP IIb/IIIa receptor blockers were not significantly associated with any type of stroke. Three main predictors of stroke were older age, prior stroke and elevated heart rate. Since the discriminative ability of these patient characteristics was at best moderate, it is difficult to predict which ACS patients will suffer a stroke.

References

1. Marschner IC, Colquhoun D, Simes RJ, et al. Long-term risk stratification for survivors of acute coronary syndromes. Results from the Long-Term Intervention With Pravastatin in Ischemic Disease (LIPID) Study. J Am Coll Cardiol 2001; 38: 56-63.
2. Kini AS, Lee PC, Mitre CA, et al. Prediction of outcome after percutaneous coronary intervention for the acute coronary syndrome. Am J Med 2003; 115: 708-15.
3. Cohen M, Stinnett SS, Weatherley DD, et al. Predictors of recurrent events and death in unstable coronary artery disease after treatment with combination antithrombotic therapy. Am Heart J 2000; 139: 962-70.
4. Boersma E, Pieper KS, Steyerberg EW, et al. Predictors of outcome in patients with acute coronary syndromes without persistent ST-segment elevation. Results from an international trial of 9461 patients. Circulation 2000; 101: 2557-67.
5. Sabatine MS, Januzzi JL, Snappin S, Theroux P, Jang I-K. A risk score system for predicting adverse outcomes and magnitude of benefit with glycoprotein IIb/IIIa inhibitor therapy in patients with unstable angina pectoris. Am J Cardiol 2001; 88: 488-92.
6. Mahaffey KW, Harrington RA, Simoons ML, et al. Stroke in patients with acute coronary syndromes. Incidence and outcomes in the Platelet glycoprotein IIb/IIIa inhibitors in Unstable angina: Receptor Supression Using Integrilin Therapy (PURSUIT) trial. Circulation 1999; 99: 2371-7.
7. The PRISM Study Investigators. A comparison of aspirin plus tirofiban with aspirin plus heparin for unstable angina. N Engl J Med 1998; 338: 1498-505.
8. The PRISM-PLUS Study Investigators. Inhibition of the platelet glycoprotein IIb/IIIa receptor with tirofiban in unstable angina and non-Q-wave myocardial infarction. Platelet Receptor Inhibition in Ischemic Syndrome Management in Patients Limited by Unstable Signs and Symptoms. N Engl J Med 1998; 338: 1488-97.
9. The PARAGON Investigators. International, randomised, controlled trial of lamifiban (a platelet glycoprotein IIb/IIIa inhibitor), heparin, or both in unstable angina. Platelet IIb/IIIa Antagonism for the Reduction of Acute coronary syndrome events in a Global Organization Network. Circulation 1998; 97: 2386-95.
10. The PURSUIT Trial Investigators. Inhibition of platelet glycoprotein IIb/IIIa with eptifibatide in patients with acute coronary syndromes. N Engl J Med 1998; 339: 436-43.
11. The GUSTO IV-ACS Investigators. Effect of glycoprotein IIb/IIIa receptor blocker abciximab on outcome in patients with acute coronary syndromes without early coronary revascularisation: the GUSTO IV-ACS randomised trial. Lancet 2001; 357: 1915-24.
12. The PARAGON-B Investigators. Randomized, placebo-controlled trial of titrated intravenous lamifiban for acute coronary syndromes. Circulation 2002; 105:

316-21.
13. Boersma E, Harrington RA, Moliterno DJ, et al. Platelet glycoprotein IIb/IIIa inhibitors in acute coronary syndromes: a meta-analysis of all major randomised clinical trials. Lancet 2002; 359: 189-98.
14. Harrell FE Jr. Regression Modeling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis. New York: Springer 2001.
15. Steyerberg EW, Harrell FE Jr., Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001; 54: 774-81.
16. Hosmer DW, Lemeshow S. Applied Logistic Regression. New York: John Wiley & Sons, Inc. 1989.
17. The GUSTO IIb Investigators. A comparison of recombinant hirudin with heparin for the treatment of acute coronary syndromes. N Engl J Med 1996; 335: 775-82.
18. Cannon CP, McCabe CH, Wilcox RG, et al. Oral glycoprotein IIb/IIIa inhibition with ortofiban in patients with unstable coronary syndromes (OPUS-TIMI 16) trial. Circulation 2000; 102: 149-56.
19. Hasdai D, Behar S, Wallentin L, et al. A prospective survey of the characteristics, treatments and outcomes of patients with acute coronary syndromes in Europe and the Mediterranean basin. The Euro Heart Survey of Acute Coronary Syndromes (Euro Heart Survey ACS). Eur Heart J 2002; 23: 1190-1201.
20. Bueno H, Bardaji A, Fernandez-Ortiz A, Marrugat J, Marti H, Heras M. Management of Non-ST-segment-elevation acute coronary syndromes in Spain. The DESCARTES (Descripción del Estado de los Sindromes Coronarios Agudos en un Registro Temporal ESpañol) Study. Rev Esp Cardiol 2005; 58: 244-52.
21. Gore JM, Granger CB, Simoons ML, et al. Stroke after thrombolysis: mortality and functional outcomes in the GUSTO-I trial. Circulation 1995; 92: 2811-8.
22. Fibrinolytic Therapy Trialists' (FTT) Collaborative Group. Indications for fibrinolytic therapy in suspected acute myocardial infarction: collaborative overview of early mortality and major morbidity results from all randomised trials of more than 1000 patients. Lancet 1994; 343: 311-22.
23. Szummer KE, Solomon SD, Velazquez EJ, et al. Heart failure on admission and the risk of stroke following acute myocardial infarction: the VALIANT registry. Eur Heart J 2005 (in press).
24. De Jaegere PP, Arnold AA, Balk AH, Simoons ML. Intracranial hemorrhage in association with thrombolytic therapy: Incidence and clinical predictive factors. J Am Coll Cardiol 1992; 19: 289-94.

25. Mahaffey KW, Granger CB, Sloan MA, et al. Risk factors for in-hospital nonhemorrhagic stroke in patients with acute myocardial infarction treated with thrombolysis: results from GUSTO-I. Circulation 1998; 97: 757-64.
26. Straus SE, Majumdar SR, McAlister FA. New evidence for stroke prevention.

Scientific Review. JAMA 2002; 288: 1388-95.
27. Cotter G, Cannon CP, McCabe CH, et al. Prior peripheral arterial disease and cerebrovascular disease are independent predictors of adverse outcomes in patients with acute coronary syndromes: Are we doing enough? Results from the Orbofiban in Patients with Unstable Coronary Syndromes-Thrombolysis in Myocardial Infarction (OPUS-TIMI) 16 Study. Am Heart J 2003; 145: 622-7.
28. Kovar D, Cannon CP, Bentley JH, Charlesworth A, Rogers WJ. Does initial and delayed heart rate predict mortality in patients with acute coronary syndromes?. Clin Cardiol 2004; 27: 80-6.
29. Al-Khatib SM, Pieper KS, Lee KL, et al. Atrial fibrillation and mortality among patients with acute coronary syndromes without ST-segment elevation: Results from the PURSUIT trial. Am J Cardiol 2001; 88: 76-9.

General Discussion

5

General discussion and conclusions

5.1

This thesis describes theoretical and practical aspects of covariate adjustment and subgroup analysis in randomized clinical trials (RCTs) with heterogeneous populations, with special interest in traumatic brain injury and acute coronary syndrome trials. In this chapter, we mention some theoretical aspects of covariate adjustment and subgroup analysis. Subsequently, use, reporting, and interpretation of covariate adjustment and subgroup analysis in recently published RCTs, and the practical application of these aspects in traumatic brain injury and acute coronary syndrome trials are discussed in the light of the theory. Finally, some general recommendations are formulated.

Summary answers to the research questions

Question 1: Which are the pros and the cons of adjustment of the treatment effect in RCTs for baseline covariates?.

Answer: Covariate adjustment increases the power to detect a significant treatment effect, and reduces the sample size requirements. This reduction in sample size is directly related to the strength of the adjusted predictors, and it is independent of the treatment effect and sample size of the trial. When time-to-event outcomes are considered, this reduction in sample size is independent of the censoring level. Adjustment for imbalance yields a very limited gain in the reduction of sample size.

Question 2: How well are covariate adjustment and subgroup analysis used, reported and interpreted in current internal medicine, oncology, cardiology, and neurosurgery trials?

Answer: Covariate adjustment was used as primary analysis in about 50% of trials, and few predictive covariates were utilized. Subgroup analysis was used in over 50% of trials, but the appropriate interactions tests were underutilized. Moreover, a minority of subgroup analyses were pre-specified, and an overemphasis on results of subgroups was common.

Question 3: How much can the adjustment for important predictors of unfavorable outcome decrease the requirements in sample size in TBI trials?

Answer: Adjustment for seven strong predictors of 6-month unfavourable Glasgow Outcome Scale reduced the sample size requirements by 25% in TBI trials. Adjustment for the three strongest predictors reduced the sample size requirements by 20%.

Question 4: Are the effects of GP IIb/IIIa receptor blockers similar in the young and the elderly in a meta-analysis of non-ST-segment elevation acute coronary syndrome trials?.

Answer: In an individual data subgroup meta-analysis of 6 large trials, GP IIb/IIIa receptor blockers had similar relative beneficial effects in old and young patients with NSTE-ACS. Absolute beneficial effects in old patients were larger than in young patients, given a higher baseline risk. However, these beneficial effects in the elderly were outweighed by a higher risk of major bleeding.

Question 5: Which subgroups of patients are at higher risk to develop a stroke after a non-ST-segment elevation acute coronary syndrome?

Answer: Patients who are old, with a prior stroke and with a high heart rate are at higher risk of stroke after sustaining a NSTE-ACS. GP IIb/IIIa receptor blockers did not increase the risk of any type of stroke.

Research Question 1: Simulations using covariate adjustment

Covariate adjustment provides a more individual-oriented treatment effect estimates, corrects for chance imbalance in baseline characteristics, and increases the power to detect an important treatment benefit [1, 2]. We performed simulation studies with dichotomous outcomes. In chapter 2.1, we showed that adjustment for a strong covariate (either pre-specified or tested) in logistic regression models led to a more beneficial treatment effect, with larger variability (i.e. broader confidence intervals) in direct relation to the strength of the covariate. However, the overall result was a gain in power. When we only adjusted when covariates were imbalanced, the gain in power was very limited. Moreover, we obtained a conservative type I error when the adjustment was performed due to imbalance, which is explained by the fact that adjusting for imbalances constrains the outcome variability between treatment groups [3].

In practical situations (Odds Ratio [OR] of the covariate between 2 and 5), a reduction in sample size between 3% and 14% was obtained with an outcome incidence of 50% and a covariate prevalence of 50%. The reduction was independent of the treatment effect and sample size, and makes it an attractive summary measure to express the benefit of covariate adjustment. We did not present calculations for covariate adjusted sample sizes in the RCT design phase. To quantify any anticipated sample gains, we would need to specify covariate effects and covariate distributions [4]. We would have to meet these assumptions to achieve

the calculated power. We advised to perform unadjusted sample calculations, which need fewer assumptions. However, reductions of sample size of about 10% can be expected when adjusting for one known strong covariate.

In chapter 2.2, we applied the same methodology in simulated RCTs with time-to-event outcomes, analyzed with the Cox proportional hazards model. The power to detect the treatment effect was higher when covariates were pre-specified and/or predictive of the outcome. The type I error was usually at the nominal level. The reduction in sample size for a covariate Hazard Ratio (HR) between 2 and 5 ranged between 15% and 44% (covariate prevalence of 50%) and between 4% and 12% (covariate prevalence 10%). The reduction was independent of the treatment effect and sample size. Importantly, this measure was also independent of the censoring level. We should notice that the RSS were not directly comparable between a RCT with dichotomous outcomes and a RCT with time-to-event outcomes. When a covariate prevalence of 50% and no censoring were settled, a given value of HR corresponded to a higher OR ($OR=HR*p_2/p_1$, where p_2 was the survival in the covariate group with the best prognosis, and p1 the survival with the covariate with the worst prognosis). Thus, a reduction of 16% was achieved with adjustment for a moderately predictive covariate (HR=2), and a reduction of 14% was obtained with adjustment for a highly predictive covariate (OR=5).

Research Question 2: Current reporting of covariate adjustment and subgroup analysis

When adequately reported, RCTs provide clinicians with valuable information that helps them to accept or reject treatments or interventions, and hence to improve their practice [5]. The revised recommendations of the Consolidated Standards of Reporting Trials (CONSORT) statement highlighted the appropriate use and interpretation of covariate adjustment and subgroup analysis [6, 7]. Covariate adjustment should be performed for a limited number of covariates, and the reasons for their choice should be clearly stated [1]. Likewise, only a limited number of pre-specified subgroup analyses should be performed, the appropriate interaction tests should be used, and subgroup analysis is to be considered as secondary analysis. Evaluation showed that the reporting of covariate adjustment and subgroup analysis in RCTs from the 80's and 90's often was inappropriate, such as inconsistencies in the use of covariate adjustment, underuse of tests of interactions, and overinterpretation of subgroup analyses [3, 8, 9].

We evaluated the reporting of covariate adjustment and subgroup analysis in RCTs from high impact factor medical journals, in general internal medicine (n=46), cardiology (n=21) and oncology (n=17) journals, published in an arbitrary

3-month period at the end of 2002. In chapter 3.1, it is described that covariate adjustment was used as primary analysis by only 50% of the RCTs, and predictive covariates were mainly used. Over 50% of RCTs used subgroup analysis, and about 20% used the appropriate interaction tests 8. Remarkably, 25% of the RCTs overemphasized the subgroup results. The reporting of covariate adjustment has improved in comparison to previous reports in the last two decades 8, 9. However, the reporting of subgroup analysis has not improved. Therefore, subgroup analysis had major shortcomings that may have resulted in incorrect medical decisions.

Then, we focused on the reporting of subgroup analysis in 63 therapeutic cardiovascular RCTs published in 2002 and 2004, in the same cardiology journals chosen for the previous evaluation, with a median of 500 patients (Chapter 3.2). The number of trials in this clinical area has increased dramatically in the last 20 years: from 5410 in the period 1980-1989 to 14845 in the period 1990-2000 [10]. The interest to explore treatment effects in subgroups has also increased, due to the heterogeneity of these patients respect to their clinical outcome [11]. Thus, many analyses have been done to look at differential treatment effects in groups such as elderly patients, female patients, diabetic patients, severe patients with acute coronary syndromes, heart failure, hypertension, hypercholesterolemia and stroke.

All 39 RCTs that reported subgroup analysis were not powered to detect subgroup effects, and 26 RCTs reported more than 5 subgroups. Fourteen trials reported fully pre-specified subgroups, only 11 RCTs used the appropriate tests of interaction, and 15 overemphasized the subgroup results. These results were worse than those from the assessment of RCTs as detailed in Chapter 3.1. A prior assessment of reporting of subgroup analysis was published [12]. Here, large cardiovascular trials (n>1000 patients) published between 1980 and 1997 had similar shortcomings as we found in smaller and more recent trials. However, subgroup analysis in larger trials have more power to detect true subgroups than the smaller trials. Many physicians and cardiologists may misinterpret the results, and can use a harmful treatment in a "wrong" subgroup of patients, or even worse, can withhold a beneficial treatment in the "right" subgroup of patients.

Finally, we studied the reporting of covariate adjustment and subgroup analysis in 18 traumatic brain injury trials (Chapter 3.3). This is another clinical area with substantial interest in these methods due to the heterogeneity of patients, and negative results of the therapeutic trials. Five of the RCTs reported covariate adjustment, and 4 of them used the adjusted effect as main efficacy parameter. They correctly used few covariates, which mainly were predictors. Eleven RCTs reported subgroup analyses, but only 3 were completely pre-specified, 1 used the

appropriate interaction testing, and 5 gave the subgroup results a similar emphasis as the overall results. These mistakes in the reporting of subgroup analysis may have been planned, or may have been driven by the results (post-hoc analyses) [13]. We had the opportunity to evaluate the correspondence between subgroup analysis planning and reporting for 6 protocols-trials combinations. The correspondence was poor, and this seems an indication of post-hoc analyses. The results of post-hoc subgroup analyses should be treated with skepticism, as they are data driven rather than stated a priori [14]. A better translation of the CONSORT recommendations on appropriate reporting of covariate adjustment and subgroup analysis can improve the understanding of clinicians about the findings in trials with heterogeneous populations.

Research Questions 3, 4, and 5: Clinical applications

We focused in two clinical areas: traumatic brain injury [15], and non-ST-segment elevation acute coronary syndromes [16].

Reduction in sample size with covariate adjustment for strong predictors in TBI trials

Most of the trials of patients with moderate or severe traumatic brain injury have not demonstrated the efficacy of treatments or interventions [17]. One of the explanations to this failure has been the reduced sample sizes of previously reported trials [18]. It seems unrealistic to increase the sample size of explanatory TBI trials due to financial and administrative constraints. One way to achieve a smaller sample size is to adjust for important predictors of the clinical outcome. RSSs were calculated in 8 trials of moderate to severe TBI, with sample sizes ranging from 126 to 1118 (Chapter 4.1). We sequentially adjusted for 1 to 9 strong predictors (clinical, radiological, biochemical) of a 6-month dichotomous Glasgow Outcome Scale.

We found that the adjustment for the 3 strongest predictors (age, motor score and pupils) yielded RSS between 16% and 23%, and adjustment for 9 strong predictors yielded RSS between 23% and 30%. Adjustment for 7 strong predictors gave similar results to adjustment for 9 predictors. These RSS were smaller than the ones in the context of TBI registries, which have less restrictive inclusion criteria and hence is expected to have more heterogeneous populations. These results highlighted the importance of the adjustment for strong predictors, and it is valuable tool in the analysis phase of new TBI trials. However, other more sophisticated methods (e.g. sliding dichotomy or proportional odds models) [19]

may reduce the sample size requirements even more. A reduction in sample size of 25% can be expected when adjusting for a group of known strong predictors.

Platelet glycoprotein IIb/IIIa receptor blocker effects in elderly patients with non-ST-segment elevation acute coronary syndromes

We performed a individual data subgroup meta-analysis of 6 large international trials, with a total of 31402 patients, to try to overcome the underrepresentation of the elderly in previous trials, and increase the power to find more reliable subgroup effects 20 (see Chapter 4.2). The appropriate interactions tests were used, and age subgroup effects were combined using random effects models [21]. The relative GP IIb/IIIa effects were similar across age subgroups, and, consequently, the absolute drug effects were larger in the elderly. This is related to the fact that elderly patients have a larger baseline risk than younger patients, and therefore receive a larger benefit from this drug [22]. However, elderly patients had larger major bleeding rates than younger patients. These results are more reliable than subgroup effects in each individual trial, and we studied a large number of elderly patients (n=2049). Likewise, harmful major bleeding complications have to be remembered when the decision to use these drugs is taken, and a close monitoring of these patients is warranted.

Subgroups of patients at higher risk of stroke after a non-ST-segment elevation acute coronary syndrome

We used the large meta-analysis database, with 155 non-hemorrhagic stroke and 228 all-cause strokes (Chapter 4.3). Elderly patients, with a prior stroke and elevated heart rate were at the highest risk to develop any type of stroke. The other previously reported predictors were not found in our analysis, and this is probably an indication of the selection of non-predictors when few events were available. Use of GP IIb/IIIa receptor blockers was not related to the presentation of any type of stroke. Although the model was well calibrated, the discriminative power of these 3 predictors was moderate (c-statistic 0.70). Thus, the prediction of stroke is difficult. We did not study long-term outcomes, and this can change the chosen predictors, and the performance of the models.

Recommendations for appropriate analysis of RCTs

Information about the appropriate analysis of RCTs with covariate adjustment and subgroup analysis in patients with heterogeneous populations is limited.

We summarize our recommendations, based on the literature and the papers presented in this thesis (Tables 1 and 2).

Subgroup analyses should be pre-specified [7]. Limitation to a small number of subgroups is desirable. This minimizes multiple testing and false-positive

Table 1: Recommendations for Covariate Adjustment

a. Specify the adjustment for few strong predictors in advance.
b. If no predictors are known, specify that you will test a limited number of them, and adjust if one or more are predictive [23].
c. Adjust for imbalances in covariates only when these covariates are predictors of the outcome.
d. Define in advance the type of model to be used.
e. Define the importance of the adjusted treatment effect in relation to the unadjusted effect (primary or secondary).
f. Perform usual unadjusted sample size calculations in the design phase. A 10% of reduction in sample size can be expected when adjusting for one strong covariate, and a 25% when adjusting for a group of strong covariates.

Table 2: Recommendations for Subgroup Analysis

Specification

a. Specify a limited number of subgroups in advance
b. Detail the rationale of the chosen subgroups.

Analysis

a. Use statistical interaction tests
b. Avoid performing analysis of the treatment effect within small subgroups of patients ('separate subgroup analysis').

Interpretation

a. Be sceptical if subgroups were not pre-specified, not biologically plausible, or no interaction tests were applied.
b. Evaluate previous reports for similar findings, and work on independent confirmation such as subgroup meta-analysis.
c. Subgroup analyses are mostly hypotheses-generating exercises to stimulate further research.
d. Keep emphasis on the overall results.

subgroup effects [8]. Subgroups should have a clear biological or pathophysiological rationale, and the appropriate interaction tests should be used. Subgroups must be confirmed independently to be more reliable. When possible, a subgroup meta-analysis increases the power to detect subgroup effects [17]. Finally, subgroup analysis only generates new hypotheses to be tested. Thus, the overall trial result is the best estimate of the treatment effect.

Conclusions

1. Covariate analysis is a valuable method to obtain more individual-oriented treatment effects. Covariate analysis allows a substantial gain in the power to demonstrate a significant treatment effect, and reduces the sample size requirements.
2. Covariate analysis is recommended in the analysis of RCTs when predictors are scientifically established. If predictors are not known, testing for them and adjustment for significant predictors are recommended.
3. A reduction of sample size of 10% can be expected when adjusting for one strong predictor, and a reduction of 25% can be expected when adjusting for a group of strong predictors. Adjustment for moderate predictors yields reductions in sample size about 5%.
4. Subgroup analysis generates hypotheses to test in further trials, and rarely provides definitive answers about subgroup effects. Its reporting is poor in current clinical trials, in different clinical and surgical medical fields.
5. A full pre-specification of subgroups, the use of interactions tests, and the consideration as secondary analysis may improve the sceptical vision of subgroup analysis.
6. Subgroup meta-analysis of trials improves the power to detect differential subgroup effects, and should be recommended when individual patient data of trials are available.

References

1. Hauck WW, Anderson S, Marcus SM: Should we adjust for covariates in non-linear regression analyses of randomized trials?. Control Clin Trials 1998; 19: 248-56.
2. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with non-linear regressions and omitted variables. Biometrika 1984; 71: 431-44.
3. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Sugroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. Stat Med 2002; 21: 2917-30.
4. Choi SC. Sample size in clinical trials with dichotmous endpoints: Use of covariables. J Biopharm Stat 1998; 8: 367-75.
5. DeMets DL. Clinical trials in the new millennium. Stat Med 2002; 21: 2779-87.
6. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001; 357: 1191-4.
7. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001; 134: 663-94.
8. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355: 1064-9.
9. Pocock SJ, Hughes MD, Lee RJ, Statistical problems in the reporting of clinical trials. A survey of three medical journals. N Engl J Med 1987; 317: 426-32.
10. Lader EW, Cannon CP, Ohman EM, et al. The clinician as investigator. Participating in clinical trials in the practice setting. Circulation 2004; 109; 2672-9.
11. DeMets DL, Califf RM. Lessons learned from recent cardiovascular clinical trials: Part I. Circulation 2002; 106: 746-51.
12. Parker AB, Naylor CD. Subgroups, treatment effects, and baseline risks: some lessons from major cardiovascular trials. Am Heart J 2000; 139: 952-61.
13. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991; 266: 93-8.
14. Cook DI, Gebski VJ, Keech AC. Subgroup analysis in clinical trials. Med J Aust 2004; 180: 289-91.
15. Jennett B. Epidemiology of head injury. J Neurol Neurosurg Psychiatry 1996; 60: 362-9.
16. Lincoff AM, Califf RM, Topol EJ. Platelet glycoprotein IIb/IIIa receptor blockade in coronary artery disease. J Am Coll Cardiol 2000; 35: 1103-15.
17. Maas AIR, Steyerberg EW, Murray GD, et al. Why have recent trials of neuroprotective agents in head injury failed to show convincing efficacy? A pragmatic analysis and theoretical considerations. Neurosurgery 1999; 44: 1286-98.

18. Dickinson K, Bunn F, Wentz R, Edwards P, Roberts I. Size and quality of randomised controlled trials in head injury: review of published studies. BMJ 2000; 320: 1308-11.
19. Murray GD, Barer D, Choi S, et al. Design and analysis of phase III trials with ordered outcome scales: the concept of sliding dichotomy. J Neurotrauma 2005; 22: 511-7.
20. Brookes ST, Whitley E, Egger M, Davey Smith G, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol 2004; 57: 229-36.
21. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986; 7: 177-88.
22. Cannon CP. Elderly patients with acute coronary syndromes: higher risk and greater benefit from antiplatelet therapy and/or interventional therapies. Am J Geriatr Cardiol 2003; 12: 259-62.
23. Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. Control Clin Trials 2000; 21: 330-42.

Summary/Samenvatting/ Resumen

5.2

Summary

Randomized clinical trials (RCTs) are essential to evaluate the usefulness of treatments and interventions, and clearly influence clinical practice. Trials are often performed in heterogeneous populations, such as patients with traumatic brain injury (TBI), acute coronary syndromes (ACS), stroke and cancer. Patients are heterogeneous regarding to their characteristics, such as age, gender, or disease severity. Heterogeneity may produce imbalance in randomized groups with respect to prognosis, and may dilute the beneficial effect of treatments in some subgroups of patients.

However, heterogeneity of patients offers some solutions to deal with these problems. Covariate adjustment and subgroup analysis are two methods used in the analysis phase of the trials. Covariate adjustment leads to adjusted estimates of treatment effects that relate to the "average" patient with a certain risk profile. It corrects for imbalance, and increases the statistical power to detect significant treatment effects. Subgroup analysis assesses differences in treatment effect across different subpopulations of patients.

We aimed to know which are the pros and cons of adjustment of the treatment effect for baseline characteristics in trials with dichotomous outcomes, to know how well are used, reported and interpreted both covariate adjustment and subgroup analysis in current clinical trials, and to apply these methods in trials performed in heterogeneous patient populations. We wanted to know how much can adjustment for important predictors decrease the sample size in TBI trials, whether the effects of glycoprotein IIb/IIIa receptor blockers were similar in the young and elderly patients with ACS, and which subgroups of patients were at higher risk to develop a stroke after an ACS.

Part 2 focuses on simulations studies to quantify the gain in power and the reduction in sample size requirements with the use of covariate adjustment in trials with dichotomous outcomes. In **Chapter 2.1** we showed that adjustment for a strong covariate (pre-specified or tested) in logistic regression models led to a more extreme estimate of the treatment effect, and a gain in power. With a strong covariate (OR between 2 and 5), adjustment yielded a reduction in sample size between 3% and 14%. This reduction was independent of the treatment effect and sample size, and it was an attractive summary measure to express the benefit of covariate adjustment. When we adjusted for covariates that were imbalanced, the gain in power and the reduction in sample size was very limited. In **Chapter 2.2** we applied the same methodology in simulation that considered time-to-event outcomes, analyzed with the Cox proportional hazards model. The statistical power to detect a significant treatment effect was higher than without

adjustment. Reductions in sample size were between 15% and 44% when adjustment was made for strong covariates (HR between 2 and 5), and lower when lower covariate prevalences were considered. The reduction in sample size was independent of treatment effect, sample size and censoring level.

Part 3 focuses on use, reporting and interpretation of covariate adjustment and subgroup analysis in internal medicine, oncology, cardiology and neurosurgery trials. In **Chapter 3.1** we evaluated internal medicine, cardiology and oncology trials published in 2002. Covariate adjustment was used in 50% of trials, and predictive covariates were used. Subgroup analysis was also used in 50% of trials. However, only 20% used the appropriate interaction tests, and 25% overemphasized the subgroup results. Subgroup analysis hence had major shortcomings that may be translated onto incorrect medical decisions. **Chapter 3.2** describes the situation of subgroup analysis in cardiovascular trials from 2002 and 2004. The interest in subgroup analysis in this field has increased dramatically in the last 20 years. Thirty-nine out of 63 trials reported subgroup analysis. Only 14 reported fully pre-specified subgroups, 11 used interaction tests, and 15 overemphasized the subgroup findings. The results were worse than those reported in chapter 3.1. Thus, many cardiologists may misinterpret the results of these subgroups, and can use a harmful treatment in "wrong" patients, and withhold a beneficial treatment in "right" patients. In **Chapter 3.3** we evaluated covariate adjustment and subgroup analysis in 18 TBI trials. Five trials used covariate adjustment, and correctly used few predictors. Eleven trials reported subgroup analysis, but only 3 fully pre-specified the subgroups, one used interactions testing, and 5 emphasized the results. We also evaluated whether these shortcomings in subgroup analysis were planned or were data-driven, by analyzing 6 available protocols of 6 trials. The correspondence was poor, and seems an indication of post-hoc analyses. Overall, the use, reporting and interpretation of covariate adjustment seems appropriate, but subgroup analysis have evident shortcomings.

Part 4 focuses on clinical applications of covariate adjustment and subgroup analysis in trials performed in two heterogeneous patient populations, such as TBI and ACS. In **Chapter 4.1** we evaluated the reduction in sample size that could be obtained when adjusting for strong predictors of unfavorable outcome in 8 TBI trials. We found that adjustment for the three strongest predictors yielded a reduction in sample size of 20%, and adjustment for the seven strongest predictors a reduction of 25%. For comparison, adjustment in TBI surveys, with more heterogeneous populations, yielded a larger reduction in sample size (reaching 40%). Adjustment for predictors is a valuable tool in the analysis phase of TBI trials. **Chapter 4.1** describes a subgroup meta-analysis of 6 large trials, which evaluated whether the effect of glycoprotein IIb/IIIa receptor blockers were similar between young and elderly patients with non-ST-segment-elevation

(NSTE) ACS. The relative beneficial effects of the drug were similar across age subgroups, and the absolute beneficial effects were larger in the elderly. This was due to a higher baseline risk of elderly patients. However, elderly patients had a higher absolute risk of major bleedings. Thus, these drugs are useful in the elderly as well as in the young, and the higher risk of bleeding warrants a careful close monitoring in the elderly. Finally, in **Chapter 4.3** we analyzed which subgroups of patients were at higher risk to develop a stroke after a NSTE-ACS. We found that patients with older age, with a prior stroke and with high heart rates were at increased risk of any type of stroke. Other previously described high-risk subgroups of patients (e.g. diabetics, hypertensive) were not at higher risk in our analysis. However, the model built with those 3 predictors had a moderate discrimination, making the prediction of stroke difficult.

Finally, in **Part 5** we put the main findings on perspective, and general recommendations for an appropriate use, reporting and interpretation of covariate adjustment and subgroup analysis are given. We conclude that covariate adjutment is a valuable method to obtain more individual-oriented treatment effects in trials, and substantial reductions in sample size requirements can be obtained when adjusting for predictors. Covariate adjustment should be recommended for future trials in heterogeneous populations. Subgroup analysis is commonly performed in current trials, but important shortcomings were found. Full pre-specification, use of interaction tests, and recognition of its secondary importance are essential to improve the sceptical vision on subgroup analysis. Subgroup meta-analysis substantially improves the power to detect subgroup effects across trials, and should be recommended as a standard procedure when individual data are available.

Samenvatting

Gerandomiseerde, klinische trials (RCT's) zijn zeer belangrijk voor onderzoek naar het nut van behandelingen en interventies en hebben veel invloed op de klinische praktijk. Trials worden vaak gedaan in heterogene populaties, zoals bijvoorbeeld patiënten met traumatisch hersenletsel (traumatic brain injury, TBI), acuut coronair syndroom (ACS), beroerte of kanker. De kenmerken van deze patiënten, zoals leeftijd, geslacht en ernst van de aandoening, zijn heterogeen. Deze heterogeniteit kan in gerandomiseerde groepen prognostische onbalans met zich meebrengen, waardoor het gunstige effect van een behandeling in sommige subgroepen van patiënten verwatert.

Er zijn verschillende oplossingen voor het probleem van de heterogeniteit van patiënten. De volgende twee methoden worden toegepast in de analysefase van trials: correctie voor covariabelen en subgroepanalyse. Correctie voor covariabelen leidt tot een aangepaste schatting van het behandeleffect voor een "gemiddelde" patiënt met een bepaald risicoprofiel. De onbalans wordt gecorrigeerd en het statistisch onderscheidingsvermogen (de power) voor het vaststellen van een significant behandeleffect wordt groter. Bij subgroepanalyse worden de verschillen in behandeleffect tussen verschillende subpopulaties van patiënten onderzocht.

We wilden onderzoeken welke voor- en nadelen correctie voor baseline kenmerken heeft op het behandeleffect in trials met dichotome uitkomstmaten, en in hoeverre correctie voor covariabelen en subgroepanalyse in de huidige klinische trials juist worden toegepast, beschreven en geïnterpreteerd. Tenslotte wilden we beide methoden toepassen in trials in heterogene patiëntenpopulaties: we wilden met name weten welke invloed correctie voor belangrijke voorspellers heeft op de steekproefomvang in trials naar TBI; of glycoproteïne-IIb/IIIa-receptorblockers in jonge en oudere patiënten met ACS een vergelijkbaar effect hebben; en welke subgroepen van patiënten na een ACS een hoger risico op een beroerte hebben.

In **deel 2** worden enkele simulatiestudies beschreven waarin voor trials met een dichotome uitkomstmaat de verandering in de power en de vereiste steekproefomvang door correctie voor covariabelen gemeten werd. In **hoofdstuk 2.1** laten we met behulp van logistische regressiemodellen zien dat correctie voor een sterke covariabele (vooraf bepaald of na toetsing) een extremere schatting van het behandeleffect en een grotere power met zich meebrengt. Correctie voor een sterke covariabele (OR tussen 2 en 5) resulteerde in een 3-14% kleinere steekproef. Deze reductie was onafhankelijk van het behandeleffect en de steekproefomvang, en bleek een geschikte samenvattende maat om de voordelen van correctie voor covariabelen in uit te drukken. Correctie voor ongebalanceerde

covariabelen had slechts een beperkte invloed op de power en de steekproefomvang. In **hoofdstuk 2.2** pasten we met behulp van Cox-regressieanalyse dezelfde methode toe in een simulatie met time-to-event uitkomstmaten. De statistische power om een significant behandeleffect aan te tonen was groter met correctie dan zonder correctie. Wanneer voor sterke covariabelen (HR tussen 2 en 5) werd gecorrigeerd, nam de steekproefomvang met 15-44% af; deze reductie was minder wanneer voor laagprevalente covariabelen werd gecorrigeerd. De mate van reductie van de steekproefomvang was onafhankelijk van het behandeleffect, de steekproefomvang en het niveau van censurering.

Deel 3 gaat in op de toepassing, de beschrijving en de interpretatie van correctie voor covariabelen en subgroepanalyse in interne, oncologische, cardiologische en neurochirurgische trials. In **hoofdstuk 3.1** evalueerden we de in 2002 gepubliceerde interne, oncologische en cardiologische trials. In 50% van de trials werd correctie voor covariabelen toegepast en gebruik gemaakt van voorspellende variabelen. In de andere helft van de trials werd subgroepanalyse toegepast. In slechts 20% echter werd een geschikte toets voor interactie gebruikt en in 25% werden de resultaten voor de subgroepen teveel benadrukt. Subgroepanalyse heeft dus belangrijke tekortkomingen die tot verkeerde medische beslissingen kunnen leiden. **Hoofdstuk 3.2** beschrijft de toepassing van subgroepanalyse in cardiovasculaire trials in de periode van 2002 tot 2004. In de laatste twintig jaar is de belangstelling voor subgroepanalyse in dit veld drastisch toegenomen. In 39 van de 63 trials bleek subgroepanalyse te zijn gebruikt. In slechts 14 trials was sprake van vooraf bepaalde subgroepen en in 11 trials vond toetsing voor interactie plaats; in 15 trials werden de resultaten voor de subgroepen teveel benadrukt. Deze bevindingen waren slechter dan die in hoofstuk 3.1. Cardiologen kunnen op deze manier de resultaten voor de subgroepen fout interpreteren en de "verkeerde" patiënten een schadelijke behandeling geven, terwijl de "goede" patiënten een gunstige behandeling onthouden wordt. In **hoofdstuk 3.3** onderzochten we correctie voor covariabelen en subgroepanalyse in 18 trials naar TBI. Vijf trials pasten correctie voor covariabelen toe en gebruikten - terecht - weinig voorspellers. Elf trials pasten subgroepanalyse toe, maar in slechts 3 trials waren de subgroepen vooraf bepaald en in 1 trial vond toetsing voor interactie plaats. In 5 trials kregen de resultaten voor de subgroepen teveel nadruk. We bekeken ook de onderzoeksprotocollen die voor 6 trials beschikbaar waren om na te gaan of de tekortkomingen in de subgroepanalyse gepland waren of door de data waren ingegeven. De inhoud was mager en kan een reden zijn geweest voor post hoc analysen. Correctie voor covariabelen wordt in het algemeen correct toegepast, beschreven en geïnterpreteerd, maar subgroepanalysen hebben duidelijke beperkingen.

Deel 4 gaat in op de klinische toepassing van correctie voor covariabelen

en subgroepanalyse in trials in twee heterogene patiëntenpopulaties met TBI en ACS. In **hoofdstuk 4.1** onderzochten we de reductie van de steekproefomvang in 8 trials naar TBI door te corrigeren voor sterke voorspellers van een ongunstige uitkomst. We vonden dat correctie voor de 3 sterkste voorspellers tot een 20% kleinere steekproef leidde; bij correctie voor de 7 sterkste voorspellers was deze reductie 25%. Ter vergelijking: correctie in onderzoeken naar TBI, in heterogenere populaties, leidde tot een grotere reductie van de steekproefomvang (tot 40%). Correctie voor voorspellers is een bruikbare methode in de analysefase van trials naar TBI. **Hoofdstuk 4.2** beschrijft een meta-analyse op subgroepniveau van 6 grote trials waarin het effect van glycoproteïne-IIb/IIIa-receptorblockers vergeleken werd bij jonge en bij oudere patiënten met ACS zonder elevatie van het ST-segment (NSTE-ACS). Relatief gezien waren de gunstige effecten van het geneesmiddel hetzelfde in de twee leeftijdsgroepen. Absoluut gezien waren de gunstige effecten het grootst bij oudere patiënten; dit had te maken met een hoger absoluut basisrisico. Oudere patiënten hadden echter ook een hoger absoluut risico op ernstige bloedingen. Dit houdt in dat dit middel bij zowel oudere als jongere patiënten nuttig is, maar dat oudere patiënten zorgvuldig in de gaten moeten worden gehouden vanwege het hogere risico op bloedingen. In **hoofdstuk 4.3** onderzochten we tenslotte subgroepen van patiënten met een hoger risico op een beroerte na NSTE-ACS. We vonden dat oudere patiënten, patiënten met een eerdere beroerte en patiënten met een snellere hartslag een verhoogd risico op een beroerte hadden. Andere eerder beschreven subgroepen van patiënten met een verhoogd risico (bijvoorbeeld patiënten met diabetes of hypertensie) hadden in onze analyse géén verhoogd risico. Het model met 3 voorspellers had echter een matig onderscheidend vermogen, waardoor het moeilijk was om een beroerte te voorspellen.

In **deel 5** plaatsen we de belangrijkste bevindingen in een breder perspectief en doen we algemene aanbevelingen voor een juiste toepassing, beschrijving en interpretatie van correctie voor covariabelen en subgroepanalyse. Wij concluderen dat correctie voor covariabelen in trials een bruikbare methode is voor het vaststellen van een meer op het individu gericht behandeleffect, en dat de steekproefomvang substantieel gereduceerd kan worden door te corrigeren voor voorspellers. Voor toekomstige trials in heterogene populaties is correctie voor covariabelen aan te bevelen. In huidige trials wordt vaak subgroepanalyse toegepast, maar deze schiet vaak tekort. Uitsluitend bepaling vooraf, toetsing voor interactie en onderkenning van het ondergeschikte belang zijn essentieel om de sceptische kijk op subgroepanalysen te verbeteren. Meta-analyse op subgroepniveau van verschillende trials vergroot de power om het effect in subgroepen te meten en is aan te bevelen als standaard procedure wanneer individuele gegevens beschikbaar zijn.

Resumen

Los estudios clínicos randomizados (RCTs) son esenciales para evaluar la utilidad de tratamientos e intervenciones, y claramente influencian la practica clínica. Estos estudios son frecuentemente hechos en poblaciones heterogéneas, tales como pacientes con daño cerebral traumático (TBI), sindromes coronarios agudos (ACS), enfermedad cerebrovascular, y cáncer. Estos pacientes son heterogeneos respecto a algunas de sus características, tales como edad, género o severidad de enfermedad. La hetorogeneidad puede producir imbalance en los grupos randomizados con respecto a pronóstico, y puede diluir el efecto beneficioso de los tratamientos en algunos subgrupos de pacientes.

Sin embargo, la heterogeneidad de pacientes ofrece algunas soluciones para lidiar con estos problemas. El ajuste de covariables y el análisis de subgrupos son dos métodos usados en la fase de análisis de los estudios clínicos. El ajuste de covariables produce estimaciones ajustadas de los efectos de tratamiento que se relacionan al "paciente promedio" con un cierto perfil de riesgo. Además corrige para imbalance e incrementa el poder estadístico para detectar efectos de tratamiento significativos. El análisis de subgrupos evalúa las diferencias en los efectos de tratamiento a través de diferentes subpoblaciones de pacientes.

Nos propusimos conocer cuáles son los pros y los contras del ajuste del efecto del tratamiento para características basales de los pacientes en estudios clínicos con resultados dicotómicos; conocer cuán bien son usados, reportados e interpretados el ajuste de covariables y el análisis de subgrupos en estudios clínicos actuales; y, aplicar estos métodos en estudios clínicos con poblaciones heterogéneas. Quisimos conocer cuánto puede el ajuste para predictors importantes disminuir el tamaño de muestra en estudios clínicos en TBI, si los efectos de los bloqueadores del receptor de glicoproteina IIb/IIIa son similares en jóvenes y ancianos con ACS, y cuáles subgrupos de pacientes están en mayor riesgo de desarrollar enfermedad cerebrovascular luego de un ACS.

En la **Parte 2** presentamos estudios de simulación para cuantificar el incremento del poder estadístico y la reducción del tamaño de muestra que se obtienen con el uso del ajuste de covariables en estudios clínicos con resultados dicótomos. En el **Capítulo 2.1** mostramos que el ajuste para una covariable fuerte (pre-especificada o probada) en modelos de regresión logística produjo un estimado del efecto del tratamiento más extremo, y un incremento en el poder estadístico. Con una covariable fuerte (OR entre 2 y 5) el ajuste produjo una reducción del tamaño de muestra entre 3% y 14%. Esta reducción fue independiente del la fuerza del efecto del tratamiento y del tamaño de muestra, y fue una medida resumen atractiva para expresar el beneficio obtenido con el adjuste de covariables. Cuando

ajustamos para covariables que tuvieron imbalance, el incremento del poder estadístico y la reducción del tamaño de muestra fue limitado. En el **Capítulo 2.2** aplicamos la misma metodología en un estudio de simulación que consideró resultados tiempo-a-evento, analizado con el modelo de hazards proporcionales de Cox. El poder estadístico para detectar un efecto de tratamiento significativo fue mayor que sin ajuste. Las reducciones en el tamaño de muestra fueron entre 15% y 44% cuando el ajuste fue hecho para covariables fuertes (HR entre 2 y 5), y menores cuando las prevalencias de las variables fueron mas pequeñas. La reducción del tamaño de muestra fue independiente del efecto del tratamiento, del tamaño de muestra y el nivel de censura.

La **Parte 3** se enfoca en el uso, reporte e interpretación del ajuste de covariables y del análisis de subgrupos en estudios clínicos de medicina interna, oncología, cardiología y neurocirugía. En el **Capítulo 3.1** evaluamos estudios clínicos de medicina interna, cardiología y oncología publicados en el 2002. El ajuste de covariables fue usado en el 50% de los estudios clínicos, y las covariables predictivas fueron usadas. El análisis de subgrupos fue usado en el 50% de los estudios. Sin embargo, solo 20% usó las apropiadas pruebas de interacción, y 25% sobreenfatizó los resultados obtenidos en los subgrupos. Por lo tanto, el análisis de subgrupos pudo importantes deficiencias que pueden ser traducidas en decisiones médicas incorrectas. El **Capítulo 3.2** describe la situación del análisis de subgrupos en estudios clínicos cardiovasculares publicados en 2002 y 2004. El interés en el análisis de subgrupos en este campo medico se ha incrementado dramáticamente en los últimos 20 años. Treinta y nueve de los 63 estudios reportaron análisis de subgrupos. Sólo 14 reportaron subgrupos completamente pre-especificados, 11 usaron las pruebas de interacción, y 15 sobreenfatizaron los hallazgos de los subgrupos. Estos resultados fueron peores que los reportados en el Capítulo 3.1. Así, muchos cardiólogos pueden malinterpretar los resultados de estos subgrupos, y pueden usar un tretamiento dañino en pacientes "equivocados", y pueden restringir su uso en pacientes "correctos". En el **Capítulo 3.3** evaluamos el ajuste de covariables y el análisis de subgrupos en 18 estudios clínicos en TBI. Cinco estudios usaron esl ajuste de covariables, y usaron correctamente pocos predictores. Once estudios reportaron análisis de subgrupos, pero solamente tres los pre-especificaron completamente, uno usó pruebas de interacción y cinco enfatizaron los resultados. También evaluamos si estas deficiencies en el análisis de subgrupos fueron planeadas or fueron basadas en los datos, al analizar 6 protocolos y 6 estudios disponibles. La correspondencia entre éstos fue pobre, y parece indicar la presencia de análisis post-hoc. En conjunto, el uso, reporte e interpretación del ajuste de covariables parece adecuado, pero el análisis de subgrupos tiene deficiencias evidentes.

La **Parte 4** se centra en las aplicaciones clínicas del ajuste de covariables y del

análisis de subgrupos en estudios clínicos realizados en poblaciones con pacientes heterogéneos, tales como TBI y ACS. En el **Capítulo 4.1** evaluamos la reducción del tamaño de muestra que podría ser obtenida cuando se ajusta para predictores importantes de resultado desfavorable en 8 estudios de TBI. Encontramos que el ajuste para los 3 predictores más fuertes produjo una reducción del tamaño de muestra del 20%, y que el ajuste para los 7 predictores más fuertes una reducción del 25%. En comparación, el ajuste de covariables en registros de TBI, con más poblaciones heterogeneas, proporcionaron una reducción mayor del tamaño de muestra (alrededor del 40%). El ajuste para predictores es una herramienta útil en la fase de análisis de estudios clínicos en TBI. El **Capítulo 4.2** describe un meta-análisis de subgrupos de 6 grandes estudios clínicos, los cuales evaluaron si los efectos de los bloqueadores del receptor de glicoproteina IIb/IIIa fueron similares entre pacientes jóvenes y ancianos con ACS sin elevación del segmento ST (NSTE-ACS). Los efectos benéficos relativos de estos fármacos fueron similares a través de los subgrupos de edad, y los efectos benéficos absolutos fueron mayores en los ancianos. Esto se debió a que los ancianos tuvieron un riesgo basal más alto que el de los jóvenes. Sin embargo, los pacientes ancianos tuvieron un riesgo absoluto mayor de sangrados significativos. Así, estos fármacos son útiles en los ancianos y en los jóvenes, y el mayor riesgo de sangrado significativo garantiza un monitoreo cuidadoso en los ancianos. Finalmente, en el **Capítulo 4.3** analizamos cuáles subgrupos de pacientes tuvieron un mayor riesgo de desarrollar enfermedad cerebrovascular luego de un NSTE-ACS. Encontramos que los pacientes ancianos, con una enfermedad cerebrovascular previa y con frecuencias cardiacas altas tuvieron mayor riesgo para desarrollar cualquier tipo de enfermedad cerebrovascular. Otros grupos de alto riesgo previamente descritos (por ejemplo, diabéticos, hipertensos) no estuvieron en alto riesgo en nuestro análisis. Sin embargo, el modelo construido con aquellos 3 predictores tuvo una moderada discriminación, haciendo difícil la predicción de la enfermedad cerebrovascular en estos pacientes.

Por ultimo, en la **Parte 5** pusimos los principales resultados en perspectiva, y presentamos recomendaciones generales para un uso, reporte e interpretación apropiadas del ajuste de covariables y del análisis de subgrupos. Concluimos que el ajuste de covariables es un método valioso para obtener efectos de tratamiento más individualizados en los estudios clínicos, y que reducciones substanciales en los requerimientos de tamaños de muestra pueden ser obtenidos cuando se ajusta para predictores. El ajuste de covariables debería ser recomendado para estudios clínicos futuros que incluyan poblaciones heterogéneas. El análisis de subgrupos es frecuentemente realizado en estudios clínicos actuales, pero con deficiencias importantes. La pre-especificación de los subgrupos, el uso de las pruebas de interacción, y el reconocimiento de su papel secundario son esenciales para mejoras la vision escéptica que se tiene de este tipo de análisis. El meta-análisis de

subgrupos mejora substancialmente el poder estadístico para detectar efectos en subgrupos a través de estudios clínicos, y debería ser recomendada como un procedimiento estándar cuando están disponibles datos individuales de pacientes.

Epilogue

Co-authors

Name	Affiliation
P.W. Armstrong, MD	Department of Cardiology, University of Alberta, Edmonton, Canada
E. Boersma, PhD	Department of Cardiology-Thoraxcenter, Erasmus MC Rotterdam
H. Bueno, MD, PhD	Department of Cardiology, Hospital General Universitario Gregorio Marañón, Madrid, Spain
I. Butcher, PhD	Division of Community Health Sciences, University of Edinburgh Medical School, Edinburgh, United Kingdom
R.M. Califf, MD	Department of Cardiology, Duke Clinical Research Institute, Durham, North Carolina, USA
S.C. Choi, PhD	Department of Biostatistics-Division of Neurosurgery, Virginia Commonwealth University, Richmond, USA
M.J.C. Eijkemans, PhD	Department of Public Health, Erasmus MC Rotterdam
J.D.F. Habbema, PhD	Department of Public Health, Erasmus MC Rotterdam
J.P.A. Ioannidis, MD	Department of Hygiene and Epidemiology, University of Ioannina, Ioannina, Greece
J. Lu, MD, MSc	Department of Biostatistics-Division of Neurosurgery, Virginia Commonwealth University, Richmond, USA
A.I.R. Maas, MD, PhD	Department of Neurological Surgery, Erasmus MC Rotterdam

A. Marmarou, PhD	Department of Biostatistics-Division of Neurosurgery, Virginia Commonwealth University, Richmond, USA
D.J. Moliterno, MD	Department of Cardiology, University of Kentucky, Lexington, Kentucky, USA
G.D. Murray, PhD	Division of Community Health Sciences, University of Edinburgh Medical School, Edinburgh, United Kingdom
N. Mushkudiani, PhD	Department of Public Health, Erasmus MC Rotterdam
M.L. Simoons, MD, PhD	Department of Cardiology-Thoraxcenter, Erasmus MC Rotterdam
E.W. Steyerberg, PhD	Department of Public Health, Erasmus MC Rotterdam
G.S. Taylor, MSc	Division of Community Health Sciences, University of Edinburgh Medical School, Edinburgh, United Kingdom
P. Theroux, MD	Department of Cardiology, Montreal Heart Institute, Montreal, Canada
L.C. Wallentin, MD	Department of Cardiology, University Hospital Uppsala, Uppsala, Sweden
C.M. Westerhout, MSc	Department of Cardiology, University of Alberta, Edmonton, Canada
H. White, MD	Department of Cardiology, Green Lane Hospital, Auckland, New Zealand

Acknowledgements

It has been a long journey since I arrived in Rotterdam in the summer of 2001. I passed through different kinds of experiences, and I had the honour to know and work with outstanding people. I hope to continue working with them. I would like to say thanks to some people who made this thesis possible:

My deepest gratitude to my promotor Prof. Dr. J. Dik F. Habbema, from whom I learned what means to be a good scientist. His insight in the research questions, the focus of the papers, and his brilliant suggestions improved my work tremendously. I am also very grateful to my co-promotor Dr. Ewout W. Steyerberg for allowing me to be part of his team, and for teaching me the basics and the applications of statistics and epidemiology in clinical research. His words of encouragement, his very good sense of humor, and his personal and professional advice have been, are, and will be unforgettable. Special thanks to Dr. Eric Boersma, whose huge experience in clinical research, detailed suggestions on the papers, and his chronic smile have improved the outcome of my work. The opportunity you gave me to know the Thoraxcenter and its work has been very important. Thanks to Prof. Dr. Theo Stijnen and Prof. Dr. Peter Koudstaal for agreeing to be part of the Small Evaluation Committee of this thesis.

I am especially grateful to some of my co-authors. To Dr. Andrew Maas, one of the most optimistic persons I have ever known, who gave the right clinical balance in some of our papers. Your leadership was vital for the success of the IMPACT Study Group. To Prof. Dr. Gordon Murray, Dr. Isabella Butcher and Dr. Gillian Taylor, our Scottish colleagues from Edinburgh. Gordon, thanks very much for your brightness, statistical experience, and fine humour. Izzy, I finally learned that Glasgow is not the capital city of Scotland. Gillian, hope you finish a marvelous 100,000-word PhD thesis very soon!. To Prof. Dr. John P. A. Ioannidis, one of the busiest scientists I know, and who always has time for discussing nice ideas. Your input in the meta-analysis was essential, and I was delighted to learn directly from you. To Dr. Héctor Bueno, my friend from Madrid and the supporter-number-one of Osasuna FC, for the opportunity to work with you and for the clinical balance you put to our papers. Thanks for presenting part of my work during the Cardiology Course at the Clínica San Pablo, Lima, in last August. I hope our future collaborative work be a success. To Dr. Cindy Westerhout, our collaborator from Edmonton. Although I do not know you personally, I appreciate very much your participation, and hope to see your PhD thesis soon.

In addition, I would like to thank my old and present colleagues from the Center for Clinical Decision Sciences (currently Center for Medical Decision Making), Erasmus MC Rotterdam. I am especially obliged to Dr. Yvonne

Vergouwe for her guidance during my MSc. period, and the continuous scientific discussions. Your move to Utrecht was a big loss to our Group, but probably a good decision for your scientific career. Hope to see you around after your time in Oxford!. I also thank Dr. Merel van Dijk whose insight, friendship and good sense of humor always helped me. Merel, hope to visit you in Amsterdam in the near future!. Thanks to Dr. Claudine Hunault, with whom I could talk during lunchtime about the difficulties of life during our stay in the Netherlands. I hope you are resting now after a few years of crazy work at home and at the University. Thanks also to Dr. René Eijkemans, who always had time to talk and to solve some non-easy-to-grasp statistical problems. I wish you a wonderful time in your new and difficult state of a married guy!. Finally, I would like to thank Dr. Cecile Janssens, Dr. Nino Mushkudiani, and Dr. Chantal Hukkelhoven for their input in my work.

I am thankful to some other colleagues from the Department of Public Health. To Dr. Hanny Groenewoud, my roommate, with whom I learned a lot of het Nederlands and who always had time to help me and solve questions of a broad range of topics. To Willemieke Kroeze, my former roommate, whose friendship and smile will not be forgotten. To Mateja Rebolj, my Slovenian good friend and paraninph, a lover of English music and long talks. I will visit Ljubljana in the near future, and hope to see you there!. To Dr. Katrina Giskes, for the good time during the "mgz dutch course". I am still waiting for your cake!. To Prof. Dr. Vicent Lorant, our Belgian visiting fellow, whose good energy and humor was infectious. Thanks for your hospitality during my visit to your house in Limar, Belgium. Finally, thanks to Dr. Esther de Vries, Marloes Berkers, Prof. Dr. Hans Brug, Prof. Dr. Jan Willem Coebergh, Peter Faas, Laetitia Verbeek, Dr. Elsbeth Voogt, Dr. Suzie Otto, Bruno Federico, Masoud Amiri, Sonja Deurloo-van Dam, Mirela Antonic, and Dr. Jan Hendrik Richardus for their support and help.

Some friends supported me during my time in Erasmus. Special words for the group of the 22nd floor (Genetic Epidemiology). To Alejandro Arias, my closest friend and paraninph, whose support have been invaluable. We spent a lot of time talking about our problems and expectations. I am delighted you will stay at Erasmus for a few years more, and hope to be in your wedding with Sandra!. To Dr. Omer Njajou, now in the USA, a world-class researcher and person, a fan of Roger Milla, the most famous football player from Cameroon. To Dr. Stefano Elefante, my Napolitan friend, a lover of pasta and pizza, a fanatic of Diego Armando Maradona, and current neighbor in Madrid. To Dr. Vincenzo Bonifati, another world-class researcher, usual writer in top medical journals and a very modest person. Thanks for giving me intensive lessons of humility and greatness. To Dr. Yuri Aoultchenko, my russian friend, whose vodka lessons were not well learned by this bad student. To Dr. Fernando Rivadeneira and

Dr. Carolina Pardo, for their hospitality, and important life-saving tips. To Dr. Fakhredin Sayed Tabatabaei for those uncountable moments of life and science discussions at lunch time. Finally, to Dr. Pascual Sánchez, usual at lunch talks and during basketball time (joder tio, ya nos veremos en España!), Dr. Luba Pardo, with whom I enjoyed nice conversations and who made me famous for one of my common questions ("y tú qué vendes?"), and to Gustavo González, el Oráculo de Rotterdam.

I would like also to thank some friends linked to the COEUR-Thoraxcenter courses, especially to Linda Smit and Dr. Jan Willem de Jong: thanks for your continuous help and advice, to Dr. Jiro Aoki and Dr. Keiichi Tsuchida, my Japanese friends, I am sure you will have excellent careers as cardiologists, and I hope to be in your PhD defenses!, to Dr. Thomas de Veert, a bright commentator of football matches, the best supporter of Ajax in Nederland, and future bright radiologist at Erasmus MC!, and to Dr. Vijay Kumar Thyagarajan, an excellent friend, a bright researcher in pharmacology and future medical resident in the USA.

Friends from other parts of the world have also been essential in this journey. Thanks in Lima to Hernán Sanabria for his eternal input and support; to Ricardo Illescas, José Hidalgo, Jaime Collins, José Claros, Rafael Chumpitaz, Juan Villena, Lourdes Rodríguez, Juan Vega, Luis Gutiérrez, Raúl Castillo, and Professor Raúl Salazar, and workers of the 1BO Service, HIV/AIDS Unit, G. Almenara General Hospital, Lima; to Andrés Villafuerte, to Eduardo Gotuzzo, Germán Valenzuela, Ernesto Yrigoyen and family; to Héctor Fuentes and Azucena, Fidel Gutiérrez and Regina, Jasson Zevallos, Belkyse and Gustavito, Manuel García Sandoval and Verónica, Kike Gal'lino, Javier and Miguel Lau and family, Walter Alcántara and family, Mario Paredes and Allyson, Luis Huarcaya, Moisés González, Willy Castillejos, and all the members of the Promoción 89 Salesiano de Lima; to Freddy Guzmán and Roxana, Percy Erazo and Lida, Luis Esteves and Julissa, Juan Marmolejo, Ivonne Guzmán and family. Thanks in Huancavelica to Andrés Herbozo and family. Thanks in Huancayo to Julio López. Thanks in São Paulo to José E. Vidal and Paula (hope to see you soon), Sergio Cimerman, and the collaborators from the Emilio Ribas Institute of Infectology. Thanks in Brasilia to César Carranza and Willeke. Thanks in Philadelphia to Gonzalo Huamán, Rocío and daugthers. Thanks in Nashville to Cecilia Chung. Thanks in Tucson to Wilder Díaz. Thanks in Los Angeles to Patricio Escalante. Thanks in Cleveland to Mike Kattan. Thanks in Madrid to Wenceslao Vásquez, Christian and Julia Domínguez, Maritza Molina, Quique Marzal, my nephew Quiquín, and family. Thanks in Jaén to Pío Zelaya, Elena and Rolito. Thanks in Vantaa to Maurizio Ruales and Mia. Thanks in México DF to my uncle José Hernández, aunt Meche, cousins Alfredo, Edita and Natalia, and Diego Flores

(half of the time in Wageningen). Thanks in Leiden to Yannina Sánchez and Michael. Thanks in Homa Bay, Kenya, to my dearest friend Walther Guerrero.

Thanks for the eternal inspiration to Enrique Bunbury, Juan Valdivia, Joaquín Cardiel, and Pedro Andreu of Héroes del Silencio; Gustavo Cerati, Charly Alberti, and Zeta Bossio of Soda Stereo; Saúl Hernández, Alfonso André, Alejandro Marcovich, Sabo Romo and César "Vampiro" López of Caifanes and Jaguares; Jorge González, Miguel Tapia and Claudio Narea of Los Prisioneros; Café Tacuba, Maldita Vecindad y los Hijos del Quinto Patio, Lucybell, La Ley, Aterciopelados, Molotov, Los Tres, Christina y los Subterráneos, Enanitos Verdes, GIT, Los Violadores, Ataque 77, Charly García, Jumbo, El Tri, Fobia, La Lupita, Babasónicos, Control Machete, El Gran Silencio, Julieta Venegas, Ely Guerra, Bohemia Suburbana, Alux Nahual, Fito Páez, Erica García, Andrés Calamaro, Los Rodríguez, Loquillo y los Trogloditas, El último de la Fila, Duncan Dhu, Amaral, Mar de Copas, Dolores Delirio, Líbido, Daniel F y Leuzemia, Micky González, Os Paralamas do Sucesso, Os Titãs, Legião Urbana, Capital Inicial, Skank, Engenheiros do Hawaii, Caetano Veloso, Cássia Eller, Daniela Mercury, Arnaldo Antunes, Juan Luis Guerra y la 4.40, Joe Arroyo, Sting, Phil Collins, Guns & Roses, Queen, Metallica, and the Beatles.

This thesis is dedicated to the loving memory of my mother Luz Angélica and my father Edgardo Alberto. Without them, nothing could have been possible. I remember the encouragement of my mother to make me a better person and professional. This thesis goes for you both. My brothers Julia and Edgardo are the living memory of them, and we enjoy very much our time together in Lima during my holidays. My close relatives have played a big role in these years, especially Esperanza, Jorge César, Julia, Marta, Tito, Maricucha, Nena, Eugenia, Elizabeth, Ezequiel, and Héctor.

Finally, and most important, I would like to thank Daniela for the wonderful years we have spent together. Extraordinary trips to Utrecht, Groningen and abroad, especially Madrid, Montreux, Bremen, Lima, Cuzco and Paracas, uncountable visits to Images Groningen --our headquarters--, uncountable songs, CDs, DVDs, and concerts, uncountable experiences. I learned from you the order, the charm, the warmness, the dedication to one's profession, and the love for one's family. I hope to share our lives forever very soon in Madrid. I keep you in the deepest part of my heart.

Curriculum Vitae

Adrián V. Hernández Díaz was born on March 9, 1973 in Lima, Peru. He graduated with honours from the high school (Colegio Salesiano de Lima) in 1989. He began his studies of medicine in 1990, and obtained his medical degree (M.D.) with honours in 1998 at the San Fernando Medical School, Universidad Nacional Mayor de San Marcos, Lima. During his time at the medical school, he was assistant teacher at the Department of Anatomy (supervisor: Dr. María Meza). He did a Residency Training in Internal Medicine between June 1999 and August 2001 at the HIV/AIDS Unit, First Service of Internal Medicine, Guillermo Almenara General Hospital, Lima (Head: Prof. Dr. Raúl Salazar Castro). He then received a scholarship from the Netherlands Organization for International Co-operation (NUFFIC) in 2001 to perform a Master of Sciences (MSc) in Clinical Epidemiology at the Netherlands Institute for Health Sciences [NIHES] (Tutor: Dr. Ewout W. Steyerberg). After his MSc. graduation in June 2002, he became research fellow at the Center for Clinical Decision Sciences, Department of Public Health, Erasmus MC Rotterdam, where he started the work described in this thesis (Promotor: Prof. Dr. Dik Habbema; Co-promotor: Dr. Ewout Steyerberg). This work was performed in close collaboration with the Thoraxcenter, Department of Cardiology (Dr. Eric Boersma), and the Department of Neurosurgery (Dr. Andrew Maas), both at Erasmus MC. He was finalist of the Young Investigators Award in Population Sciences during the 2004 European Society of Cardiology Conference, 31 August- 3 September, Munich, Germany, for his work on subgroup analysis in therapeutic cardiovascular trials. He was awarded with the best oral presentation during the 2005 Congress of the Brazilian Society of Tropical Medicine, 6-10 March, Florianópolis, Brazil, for his collaborative work on diagnosis of cerebral toxoplasmosis in AIDS patients. He is member of the Editorial Board of the PanAmerican Journal of Infectology since June 2005. In July 2005, he began his Residency Training in Pulmonary Medicine at the Department of Pneumology, Hospital General Universitario 12 de Octubre, Madrid, Spain (Head: Prof. Dr. Pedro Martín Escribano).

List of publications *by December 15, 2005*

Hernández AV, Steyerberg EW. Managing the resource demands of a large sample size in clinical trials: can you succeed with fewer subjects?. **Med J Aust** 2003 ; 178: 356-7.

Vidal JE, da Silva PR, Schiavon Nogueira R, Bonasser Filho F, **Hernández AV**. Liver abscess due to Salmonella enteritidis in a returned traveler with HIV infection: case report and review of the literature. **Rev Inst Med Trop Sao Paulo** 2003; 45: 115-7.

Hernández AV, Vergouwe Y, Steyerberg EW. Reporting of predictive logistic models should be based on evidence-based guidelines. **Chest** 2003; 124: 2034-5.

Hernández AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. **J Clin Epidemiol** 2004; 57: 454-60.

Vidal JE, **Hernández AV**, Oliveira AC, Souza A, Madalosso G, Silva PR, Dauar R. Cerebral tuberculomas in AIDS patients: a forgotten diagnosis?. **Arq Neuropsiquiatr** 2004; 62: 793-6.

Hernández AV, Steyerberg EW, Habbema JDF. Inappropriate use of baseline characteristics in current randomised controlled trials. **Med Decis Making** 2004; 24: 426.

Hernández AV, Eijkemans MJC, Steyerberg EW. Randomized controlled trials with time-to-event outcomes: how much does pre-specified covariate adjustment increase power?. **Ann Epidemiol** 2006; 16: 41-8.

Vidal JE, **Hernández AV**, Penalva de Oliveira AC, Dauar RF, Barbosa Jr. SP, Focaccia R. Cerebral toxoplasmosis in HIV+ patients in Brazil: Clinical features and predictors of treatment response in the HAART era. **AIDS Patient Care STDS** 2005; 19: 626-34.

Hernández AV, Boersma E, Murray GD, Habbema JDF, Steyerberg EW. Subgroup analysis in therapeutic cardiovascular clinical trials: Are most of them misleading?. **Am Heart J** 2006 (in press).

Collins JA, **Hernández AV**, Hidalgo JA, Villena J, Sumire J, Delgado V, Salazar

R. High proportion of T-cell systemic non-Hodgkin lymphoma in HIV-infected patients in Lima, Peru. **JAIDS** 2005; 40:558-64.

Hernández AV, Steyerberg EW, Taylor GS, Marmarou A, Habbema JDF, Maas AIR. Subgroup analysis and covariate adjustment in randomized clinical trials of traumatic brain injury: A systematic review. **Neurosurgery** 2005; 57: 1244-53.

Colombo FA, Vidal JE, Penalva de Oliveira AC, **Hernández AV**, Bonasser-Filho F, Nogueira RS, Focaccia R, Pereira-Chioccola VL. Diagnosis of cerebral toxoplasmosis in Brazilian AIDS patients: Importance of molecular and immunological methods in peripheral blood samples. **J Clin Microbiol** 2005; 43: 5044-7.

Hernández AV, Westerhout CM, Steyerberg EW, Ioannidis JPA, Bueno H, White H, Theroux P, Moliterno DJ, Armstrong PW, Califf RM, Wallentin LC, Simoons ML, Boersma E. Effects of platelet glycoprotein IIb/IIIa receptor blockers in non-ST-segment elevation acute coronary syndromes: Benefit and harm in different age subgroups. Submitted

Hernández AV, Steyerberg EW, Mushkudiani N, Murray GD, Taylor GS, Butcher I, Marmarou A, Choi SC, Lu J, Habbema JDF, Maas AIR. Adjustment for strong predictors of unfavorable outcome reduces sample size requirements by over 25% in traumatic brain injury trials. Submitted

Westerhout CM, **Hernández AV**, Steyerberg EW, Bueno H, White H, Theroux P, Moliterno DJ, Armstrong PW, Califf RM, Wallentin LC, Simoons ML, Boersma E. Predictors of stroke at 30 days in patients with non-ST-segment elevation acute coronary syndromes. Submitted

Hernández AV, Vergouwe Y, Rietveld E, Moll HA, Habbema JDF, Steyerberg EW. The risk of hospitalisation in young children due to respiratory syncytial virus (RSV) infection: Cross-seasonal validation of a predictive model. Submitted

Lederink T, **Hernández AV**, Boersma E, Simoons ML, Bueno H. Predictors of 30-day mortality in Spanish patients with first acute myocardial infarction and older than 75 years of age - The PPRIMM75 study. In preparation.

Hernández AV, Boersma E, Bueno H. Underuse of Cardiovascular Preventive Therapies in Type 2 Diabetes Mellitus Patients in Primary Care in Spain- The Trans-STAR study. In preparation

Butcher I, Taylor GS, Lu J, Steyerberg EW, **Hernández AV,** Mushkudiani N, Murray GD, Marmarou A, Maas AIR (the IMPACT study group). The prognostic

value of cause of injury in traumatic brain injury. In preparation.

Taylor GS, Engel DC, Butcher I, Steyerberg EW, Lu J, Mushkudiani N, **Hernández AV**, Marmarou A, Maas AIR, Murray GD (the IMPACT study group). The Prognostic Value of Secondary Insults in traumatic brain injury. In preparation.

Vidal JE, **Hernández AV**, Colombo F, Penalva de Oliveira AC, Bonasser Filho F, Schiavon Nogueira R, Focaccia R, Pereira-Chioccola VL. Polimerase Chain Reaction (PCR) in blood samples for diagnosis of cerebral toxoplasmosis in AIDS patients. In preparation.

Collins JA, Hidalgo JA, **Hernández AV**, Salazar R, for the Almenara Hospital AIDS Working Group. HTLV- I infection is not associated with a higher risk of death in Peruvian HIV-infected patients. In preparation.