Exploratory Subgroup Analyses in Clinical Research,
First Edition. Gerd Rosenkranz.

Companion website: www.wiley.com/go/rosenkranz/exploratory

1

Some History of Subgroup Analysis

The essence of tragedy has been described as the destructive collision of two sets of protagonists, both of whom are correct. The statisticians are right in denouncing subgroups that are formed post hoc from exercises in pure data dredging. The clinicians are also right, however, in insisting that a subgroup is respectable and worthwhile when established a priori from pathophysiological principles. (*Feinstein, 1998*)

Alvan R Feinstein (1925–2001)

The Problem of Cogent Subgroups: A Clinicostatistical Tragedy.

1.1 INTRODUCTION

The history of subgroup analysis is characterized by a strong difference in opinions about its value.

One group of scientists has a skeptical attitude towards the topic warning of the risks of subgroup analysis and other attempts to target treatments. For example, Yusuf et al. (1984) stated that " … it would be unfortunate if desire for the perfect (i.e. knowledge of exactly who will benefit from treatment) were to become the enemy of the possible (i.e. knowledge of the direction and approximate size of the effects of treatment of wide categories of patients)." Many clinicians are afraid of applying the overall results of large trials to individual patients without consideration of determinants of individual responses (Rothwell, 2005) while most prominently statisticians have raised concerns (Assmann et al., 2000, Sleight, 2000, Lagakos, 2006, Guillemin, 2007, Lonergan et al., 2017) and requested that:

- Investigators should be cautious when undertaking subgroup analyses.
- Subgroup findings should be exploratory, and only exceptionally should they affect the conclusions from trials.
- Editors and reviewers of journals need to correct any inappropriate, over-enthusiastic uses of subgroup analyses.

The statement "subgroups kill people" was attributed – rightly or wrongly – to statistician Sir Richard Peto in van Gijn and Algra (1994). In fact, Peto commented on subgroup analyses

undertaken on the GISSI[1] study (GISSI Study Group, 1986): "The GISSI study … is one of the most important randomized trials ever conducted and when it was published provided the best evidence then available that thrombolytic therapy reduced mortality. But the ability of the GISSI report to save lives could be substantially compromised by misinterpretation by clinicians of some of the data-dependent subset analyses that it contained." (Peto, 1990)

A second camp of scientists and pharmaceutical executives is more attracted by the opportunities than by the risks of subgroup analysis driven by the vision of "personalized" medicine. In 1977, Sir Richard Sykes, at the time chief executive officer of Glaxo-Wellcome, later chairman of GlaxoSmith-Kline and rector of Imperial College London, wrote:

"It will soon be possible for patients in clinical trials to undergo genetic tests to identify those individuals who will respond favorably to the drug candidate, based on their genotype, and therefore the underlying mechanism of their disease. This will translate into smaller, more effective clinical trials with corresponding cost savings and ultimately better treatment in general practice. In addition, clinical trials will be capable of screening for genes involved in the absorption, metabolism and clearance of drugs and the genes that are

[1]GISSI = Gruppo Italiano per lo Studio della Streptochinasi nell'Infarcto Miocardico

likely to predispose a patient to drug-induced side-effects. In this way, individual patients will be targeted with specific treatment and personalized dosing regimens to maximize efficacy and minimize pharmacokinetic problems and other side-effects." (Sykes, 1977), quoted from Senn (2001). It took another 20+ years until the first targeted medicine in oncology, trastuzumab for HER2 positive breast cancer, was approved by the US Food and Drug Administration (FDA) in 1998.

More and more drugs were approved for targeted patient populations during the following years. A selective list is displayed in Table 1.1. In 2013, the FDA issued a report "Paving the way to personalized medicine" (FDA, 2013) describing how the agency was planning to support the development of new drugs with companion

Table 1.1 Approved targeted therapies

Indication	Marker	Compound
Breast cancer	HER2+	trastuzumab pertuzumab
	HER2-/ER+	everolimus
Colorectal cancer	KRAS	cetuximab panatumumab
Cystic fibrosis	G551D	ivacaftor
Melanoma	BRAF V600E	vemurafenib dabrafenib
	BRAF V600E or V600K	trametinib
NSCLC	ALK	crizotinib

diagnostics to guide their use. In 2017, the agency approved 16 targeted medicines (FDA, 2017).

1.2 QUESTIONABLE SUBGROUP ANALYSES

1.2.1 Star Signs May Matter

A trial that Peto mentioned in his critique on the GISSI study to justify his concerns on subgroup analyses was ISIS–2[2] (ISIS–2 Collaborative Group, 1988). This study enrolled 17 187 patients in 417 hospitals up to 24 h after the onset of suspected myocardial infarction. Patients were randomized to (i) a one hour iv infusion of streptokinase; (ii) one month of 160 mg/day aspirin; (iii) both active treatments; or (iv) neither. In the end, streptokinase reduced five week vascular mortality by $25 \pm 4\%$ as compared to placebo ($2p < 0.00001$). Aspirin reduced five week vascular mortality by $23 \pm 4\%$ as compared to placebo ($2p < 0.00001$). The combination of both aspirin and streptokinase reduced five week vascular mortality by $45 \pm 5\%$ as compared to placebo ($2p < 0.00001$).

The study authors concluded on subgroup analyses: "Even in a trial as large as ISIS–2, reliable identification of subgroups of patients among whom treatment is particularly advantageous (or among whom it is ineffective) is unlikely to be

[2]ISIS = International Study of Infarct Survival

possible. When in a trial with a clearly positive overall result many subgroup analyses are considered, false negative results in some particular subgroups must be expected." They underlined their opinion with "the most entertaining example of an inappropriate subgroup analysis" (Horton, 2000): "For example, subdivision of the patients in ISIS–2 with respect to their astrological birth signs appears to indicate that for patients born under Gemini or Libra there was a slightly adverse effect of aspirin on mortality (9% ± 13% odds increase; NS), while for patients born under all other astrological signs there was a strikingly beneficial effect (28% ± 5%) odds reduction; $2p < 0.00001$)." The results for the aspirin–placebo comparison are shown in Table 1.2

The reason for this odd item appearing in the paper originated in negotiations between authors and editors. The Lancet was keen to include clinically relevant subgroup findings. The authors agreed under the proviso that the journal allowed

Table 1.2 Vascular deaths in the ISIS–2 study

Star sign	Aspirin			Placebo			Odds ratio
	N	deaths	(%)	*N*	deaths	(%)	
Gemini/ Libra	1357	150	11.1	1442	147	10.2	1.09
Other	7228	654	9.0	7157	868	12.1	0.72

Source: ISIS–2 Collaborative Group (1988)

the star sign groups to appear first to underline for readers the reliance they might put (or not) on the validity of these analyses (Horton, 2000).

1.2.2 Unjustified Under-treatment

The artificial subgroup analysis just described has at least one real world counterpart that caused serious under-treatment of a subgroup of patients for at least a decade because of a subgroup analysis: a Canadian Cooperative Study Group trial came to the conclusion that aspirin was effective in preventing stroke and death in men but not in women. The gender by treatment interaction turned out to be significant ($p = 0.003$) and aspirin was effective in preventing stroke and death in men (RR$= 0.52$, $p < 0.005$) but not in women (1.42, $p = 0.35$) (The Canadian Cooperative Study Group, 1978).

As part of a major meta-analysis of studies of high risk subjects in which individual patient data were obtained it was concluded that antiplatelet therapy for high risk patients appeared to reduce the odds of vascular events by a roughly similar proportion regardless of age or gender of the subjects (Antiplatelet Trialists' Collaboration, 1994). Thus the notion that women might not benefit from antiplatelet therapy (which arose from data dependent subgroup analyses of a few trials) is contradicted by much more reliable, prospectively planned overview analyses.

1.2.3 Misinterpretation of Center Effects

Most randomized controlled clinical trials are conducted in multiple centers since a single center is unable to provide all of the necessary patients for a definitive study, for example, when the condition under study is rare or the anticipated treatment effect is small. A controversy that has emerged concerning the analysis of multi-center trials is the interpretation of divergent center results. In a post hoc analysis of the Beta-Blocker Heart Attack Trial (BHAT), Horwitz et al. (1996) illustrated the occurrence of substantial variation in results among the participating centers.

BHAT enrolled 3837 subjects in 31 clinical centers in the United States and Canada. Eligible subjects included men and women between the ages of 30 and 39 years who had been hospitalized with an acute myocardial infarction. Patients were randomized to receive propanolol or a matching placebo after their condition had stabilized. The minimum length of follow-up was 12 months and the average time on trial was 25 months. Overall, 1916 subjects were randomized to propanolol and 1921 to placebo. At 25 months the estimated mortality rates were 7.2% for propanolol and 9.8% for placebo, for a relative risk of 0.73 and a 95% confidence interval (CI)= (0.59, 0.90).

In the course of a post hoc investigation, Horwitz et al. (1996) divided the centers into two groups: 21 dominant centers in which mortality

rates were higher for patients on placebo and 10 divergent centers in which higher mortality rates occurred for patients on propanolol. The relative risk in the dominant centers favoring propanolol was 0.5, 95% CI= (0.38, 0.67). In the divergent centers, they obtained a relative risk of 1.33 and a 95% CI= (0.95, 1.88). These numbers are summarized in Table 1.3.

The test for qualitative interactions by Gail and Simon (1985) was significant on the 5% level supporting the view of the authors that propanolol is potentially helpful for patients in the dominant centers and potentially harmful for the diverging centers.

This view was criticized in a dissent by Senn and Harrell (1997) who were not at all surprised that Horwitz et al. (1996) found a significant difference between the two groups of centers. In fact they argued that "a very similar analysis applied to *any multicenter trial whatsoever* will always be significant at the 5% level provided only that the

Table 1.3 Results from BHAT by center subgroups and overall

Subgroup	Number of subjects	Relative risk (95% CI)
21 dominant centers	2480	0.50 (0.38, 0.67)
10 divergent centers	1357	1.33 (0.95, 1.88)
All centers	3837	0.73 (0.59, 0.90)

Source: Horwitz et al. (1996)

number of centers is at least equal to 8." The analysis they proposed was to rank the centers according to the treatment effect, divide them into two groups above and below the median and carry out a rank test on the groups so defined.

They argued further that given the overall mortality rates of 0.072 and 0.098 for the treatment and control group, no heterogeneity between centers and an average center size of 62, one obtains a probability of an effect reversal favoring control of 0.25 per center. The expected number of effect reversal for a trial of 31 centers of equal size would then amount to $0.25 \times 31 = 7.7$ and the probability of 10 or more effect reversals would be 0.22. They also complain about the improper use of the Gail–Simon test (Gail and Simon, 1985), which requires subsets to be specified in advance and not based on observed event rate differences.

In a rejoinder, Horwitz et al. (1997) re-iterated that the results of clinical trials providing average effects on intention to treat (ITT) patient populations that ignore post randomization interventions do not support the needs of treating physicians to prescribe the best treatment for an individual patient.

1.2.4 The End of a Career

A study that had a material impact on the players involved was the Actimmune trial in idiopathic pulmonary fibrosis (IPF) (Raghu et al., 2004). The study enrolled 330 IPF patients who were

randomized between placebo and Actimmune (Interferon gamma-1b). At the study conclusion, no significant difference in the primary endpoint (progression free survival) nor in any of nine secondary endpoints could be found. However, the mortality rate was 40% lower under the test treatment compared to placebo ($p = 0.084$). In addition, in a post hoc subgroup of 254 patients with mild to moderate disease, mortality was reduced by 70% ($p = 0.004$).

The company issued a press release[3] entitled "InterMune announces Phase III Data Demonstrating Survival Benefit of Actimmune in IPF: Reduces Mortality by 70% in Patients With Mild to Moderate Disease" with the following conclusions:

- "Preliminary data … demonstrate a significant survival benefit in patients with mild to moderate disease randomly assigned to Actimmune versus the control treatment ($p = 0.004$)."
- "There was also approximately a 10% relative reduction in the rate of progression-free survival associated with Actimmune versus placebo, the trial's primary endpoint, but this was not a statistically significant difference."

InterMune then promoted the drug off-label in IPF, while the FDA never approved it in this indication. In 2003, the company initiated the INSPIRE trial, a study of Actimmune in patients

[3]http://www.sec.gov/Archives/edgar/data/1087432/000091205702033878/a2088367zex-99_1.htm

with mild to moderate IPF to confirm the results of the post hoc subgroup analysis. Unfortunately, the study was terminated in 2007 after an interim analysis showed no survival benefit. InterMune sold the drug in 2012.

The CEO of InterMune was prosecuted by the US Department of Justice for " … fraudulently promoting the drug Actimmune", by issuing " … false and misleading information about the drugs effectiveness in treating idiopathic pulmonary fibrosis." In 2009 the CEO was found guilty of wire fraud by a jury. The conviction was affirmed by the Ninth Circuit of the United States Court of Appeals in March 2013 and a petition for writ of certiorari was denied by the US Supreme Court in December 2013.

1.3 ENCOURAGING SUBGROUP ANALYSES

1.3.1 Higher Efficacy

The growth factor receptor HER2 is over-expressed in 25 to 30% of breast cancers, increasing the aggressiveness of the tumor. To evaluate the efficacy and safety of trastuzumab, a recombinant monoclonal antibody against HER2, in women with metastatic breast cancer over-expressing HER2, 469 patients were enrolled in a prospective clinical study, 234 of whom were randomized

to receive standard chemotherapy and 235 to receive standard chemotherapy plus trastuzumab. Patients who had not previously received adjuvant therapy with an anthracycline were treated with doxorubicin or epirubicine in combination with cyclophosphamide with (143 women) or without (138 women) trastuzumab. Patients who had previously been treated with anthracyclines were treated with paclitaxel alone (96 woman) or paclitaxel plus trastuzumab (92 woman). The addition of trastuzumab was associated with a longer time to disease progression, a higher rate of objective response and a higher one year survival rate (Slamon et al., 2001). The results for time to progression (primary endpoint) and mortality are summarized in Table 1.4.

The interesting feature of this study is that it included only patients with HER2 positive cancer, i.e. exclusively the predefined subgroup, without a direct comparison in HER2 negative tumors. The evidence that trastuzumab would be primarily efficacious in HER2 positive disease was obtained from earlier studies.

1.3.2 Harm Prevention

Idiosyncratic drug-induced liver injury (DILI) is a major safety concern and has been a common cause for the marketing withdrawal of a range of drugs. Due to the unpredictable and rare nature of these events it is often not until the post-marketing phase that a drug's propensity for DILI is revealed.

Table 1.4 Results on time to progression and survival from the trastuzumab study in HER2 positive breast cancer

End point	C+T	C	A+T	A	P+T	P
Median time to disease progression (months)	7.4	4.6	7.8	6.1	6.9	3.0
—Relative risk of progression (95% CI)	0.51 (0.41–0.63)		0.62 (0.47–0.81)		0.38 (0.27–0.53)	
—*p*-value	< 0.001		< 0.001		< 0.001	
Median survival time (months)	25.1	20.3	26.8	21.4	22.1	18.4
—Relative risk of death (95% CI)	0.80 (0.64–1.00)		0.82 (0.61–1.09)		0.80 (0.56–1.11)	
—*p*-value	0.046		0.16		0.17	

C+T = chemotherapy plus trastuzumab , C = chemotherapy alone, A+T = anthracycline plus trastuzumab, A = anthracycline alone, P+T = paclitaxel plus trastuzumab, P = paclitaxel alone. Time to progression and survival were analyzed nine months and 31 months after enrollment of the last patient, respectively. Source: Slamon et al. (2001)

The discovery of genetic markers able to identify individuals at risk could make otherwise safe and efficacious drugs available for use.

Concerns over hepatoxicity have contributed to the withdrawal or non-approval of the selective COX-2 inhibitor lumiracoxib, which proved to be efficacious in osteoarthritis and acute pain (Bannwarth and Berenbaum, 2007). To identify genetic markers able to select individuals at risk for developing drug induced liver injury a case-control genome-wide association study was conducted in 41 lumiracoxib treated patients with liver enzyme elevations above five times the upper limit of normal (ULN) and 176 patients without liver injury (Singer et al., 2010) using DNA samples collected from the TARGET[4] study (Farkouh et al., 2004). Endpoints were time to liver enzyme elevations above five times ULN. Fine mapping identified a strong association with a common HLA haplotype. HLA-DQA1*0102 had the best results in terms of negative predictive value (99%) and sensitivity (73.6%).

To further examine the performance characteristics of the marker, all remaining 4518 lumiracoxib treated patients from the TARGET study with DNA available who had given informed consent were genotyped for the presence or absence of HLA-DQA1*0102. Kaplan–Meier (KM) estimates of the cumulative incidence

[4]TARGET = Therapeutic Arthritis Research and Gastrointestinal Event Trial

of liver enzyme elevations were obtained for HLA-DQA1*0102 carriers and non-carriers and compared to estimates for all patients treated with lumiracoxib, ibuprofen or naproxen. As it turns out, the KM curve for lumiracoxib treated subjects who are DQA*0102 carriers is increasing much faster over time than the KMs for patients treated with the comparator drugs. The risk of non-carriers under lumiracoxib is similar to the risk in the overall population under the comparator treatments (Figure 2 in Singer et al. (2010)). The paper concludes: "The results presented here provide strong evidence that the HLA-DQA1*0102 allele would have clinical utility as a screening marker to exclude carriers from lumiracoxib treatment." In any case, the study moved the field of personalized medicine into the safety area with one of the first DILI safety markers.

1.3.3 Avoiding Unnecessary Treatment

Hormone-receptor-positive, axillary node-negative disease accounts for approximately half of all cases of breast cancer in the United States. Adjuvant chemotherapy reduces the risk of recurrence with effects greater in younger women. These findings let a National Institute of Health (NIH) panel recommend adjuvant chemotherapy for most patients, leading to a declining breast cancer mortality. However, the majority of patients may receive chemotherapy unnecessarily.

A recurrence score ranging from 0 to 100 based on the 21-gene breast cancer assay (OncotypeDX) predicts chemotherapy benefit if it is high and a low risk of recurrence without chemotherapy if it is low (Sparano and Paik, 2008, Sparano et al., 2015). However, there was uncertainty about benefit of chemotherapy for most patients who have a midrange score. To close this knowledge gap, the National Cancer Institute (NCI) sponsored the Trial Assigning Individualized Options for Treatment (TAILORx) (Sparano et al., 2018).

TAILORx was a prospective trial involving 10273 women with hormone-receptor-positive, human epidermal growth factor receptor 2 (HER2)-negative, axillary node-negative breast cancer. Of the 9719 eligible patients, 6711 had a midrange recurrence score of 11 to 25. These subjects were randomly assigned to either chemoendocrine therapy or endocrine therapy alone. The trial was designed to show non-inferiority of endocrine therapy alone for invasive disease-free survival, defined as freedom from invasive disease recurrence, second primary cancer or death. A five year rate of invasive disease-free survival rate of 90% with chemoendocrine therapy and of 87% or less with endocrine therapy alone, which corresponds to a hazard ration of 1.322, was specified as unacceptable.

After conclusion of the trial, the hazard ratio for invasive disease-free survival of endocrine relative to chemoendocrine therapy was 1.08 with a 95% confidence interval of (0.94, 1.24).

Non-inferiority could also be concluded in other endpoints. A significant interaction between age and chemotherapy treatment was found that confirmed previous findings. However, it is unclear whether the string of covariates underlying the exploratory analyses were pre-specified. The study confirmed the generally good results for scores below 10 on endocrine therapy alone and the slightly worse results under chemoendocrine treatment for scores of 26 and above.

1.4 SUBGROUPS AND DRUG APPROVALS

1.4.1 A Convincing Subgroup

In 2008, the Food and Drug Administration (FDA) cleared the first transcranial magnetic stimulation (TMS) device to treat depression in patients who failed on one antidepressant. In an editorial (Hines et al., 2009), the authors raised concerns that the decision was based on a post hoc subgroup analysis of a published negative randomized controlled trial (O'Reardon et al., 2007), re-analyzed by Lisanby et al. (2009).

In the double-blind study, 301 patients with major depression who had not benefited from prior treatment were randomized; 155 to active and 146 to sham TMS. The primary outcome was a change from baseline to week 4 on the

Montgomery–Asberg depression rating scale (MADRS).

The difference between treatment groups was 1.7 points on the 60 points MARDS at four weeks with a p-value of 0.057, "both statistically and clinically non-significant." (Hines et al., 2009). The finding became significant ($p = 0.038$) after the post hoc exclusion of six patients with baseline MADRS below 20. In the result section of the publication, the authors claimed that "active TMS was significantly superior to sham TMS on the MADRS (with a post hoc correction for inequality in symptom severity between groups at baseline)."

When presented with these data, the FDA Advisory Committee concluded that TMS' "clinical effect was perhaps marginal, borderline, questionable, and perhaps a reasonable person could ask whether there was an effect at all" and rejected the device. FDA overruled its committee and granted approval. In the agency's favor it may be argued that the original result was close to significance and that subjects with low MARDS at baseline do not have a good chance to improve. This may impact the results if these subjects are not evenly distributed amongst treatment groups. However, an analysis of covariance may have been an alternative option.

1.4.2 Inconsistencies Across Regions

Another subgroup analysis that gave the FDA and its Advisory Committee a hard time was presented

as part of the New Drug Application (NDA) of ticagrelor. The platelet inhibition and clinical outcomes (PLATO) trial revealed a significant benefit of ticagrelor over clopidogrel in terms of the composite endpoint of cardiovascular death, myocardial infarction or stroke at one year in patients with acute coronary symptoms [hazard ratio 0.84, 95% CI (0.77, 0.92), $p < 0.001$] (Wallentin et al., 2009). Much discussion was generated by the finding that ticagrelor offered no benefit over clopidogrel in the United States (Table 1.5).

A similar situation occurred earlier in 2001 for the MERIT-HF[5] where a post hoc subgroup analysis showed a mortality hazard ratio of 1.05 (95% CI (0.71, 1.56)) for the United States and 0.55 (0.43, 0.70) for all other countries combined and a significant quantitative interaction ($p = 0.003$) (Wedel et al., 2001).

In search for an explanation of the PLATO results, the differing aspirin (ASA) dose between the regions was identified as a potential cause by the sponsor. Specifically, a higher maintenance dose of aspirin was associated with relatively unfavorable outcomes with tigacrelor in both US and non-US patients while the hazard ratios for low-dose aspirin look similar, however with a much lower sample size in the United States as compared to the non-US region (see Table 1.5).

[5] MERIT-HF = metoprolol controlled release randomized intervention trial in heart failure

Table 1.5 PLATO results overall and by region and ASA dose (mg)

Subgroup	Tigacrelor		Clopidogrel		HR (95% CI)
	N	Events	*N*	Events	
All	9333	864	0291	1014	0.84 (0.77, 0.92)
US	707	84	706	67	1.27 (0.92, 1.75)
Non-US	8626	780	8585	947	0.81 (0.74, 0.90)
ASA $\geq$ 300	464	68	492	50	1.45 (1.01, 2.09)
100 < ASA < 300	525	64	527	65	0.99 (0.70, 1.40)
ASA $\leq$ 100	7733	565	7706	723	0.77 (0.69, 0.86)

https://www.accessdata.fda.govdrugsatfda_docs/label/2015/022433s015lbl.pdf

The other potential explanation was a difference in monitoring among regions. It seems hard, however, to imagine how this should have affected mortality. At the end of the day, ticagrelor was approved in the United States; however, the lack of efficacy in the US and the potential impact of the aspirin dose had to be included in the label (Gaglia and Waksman, 2011). Buyse and Marschner (2011) provided further support of no variation in treatment effect beyond chance.

It would have been interesting to see how the PMDA (Pharmaceuticals and Medical Devices Agency), the Japanese health authority, would have reacted if the results had occurred in Japan instead of the United States. Following their guidelines (MHLW, 2007), they only accept the results of an international trial for drug approval if the observed treatment effect in Japan is at least half of the overall study effect estimate or if all regional effects are less than 0 (if a reduction of an endpoint constitutes a success). Applying either criterion, tigacrelor may have failed in Japan.

1.4.3 Detecting Non-responders

An example where post hoc subgroup analyses led to product label changes resulting in restrictions of the patient population is cetuximab. The drug got approval for epidermal growth factor receptor (EGFR) expressing metastatic colorectal cancer in 2004. However, retrospective subgroup analyses suggested a lack of efficacy in patients with

Table 1.6 Results for the PLATO trial by ASA dose within region

Region	ASA dose (mg)	Tigacrelor		Clopidogrel		HR (95% CI)
		N	Events	*N*	Events	
US	≥ 300	324	40	352	27	1.62 (0.99, 2.64)
	(100, 300)	22	2	16	2	—
	≤ 100	284	19	263	24	0.73 (0.40, 1.33)
Non-US	≥ 300	140	28	140	23	1.23 (0.71, 2.14)
	(100, 300)	503	62	511	63	1.00 (0.71, 1.42)
	≤ 100	7449	546	7443	699	0.78 (0.69, 0.87)

https://www.accessdata.fda.govdrugsatfda_docs/label/2015/022433s015lbl.pdf

metastatic disease whose tumors have KRAS[6] mutations. KRAS, an essential component of the EGFR signaling cascade, can acquire mutations that render EGFR inhibitors ineffective.

We summarize the results of one of the studies (Jonker et al., 2007) that was analyzed further in Karapetis et al. (2008) after the observation of resistance to cetuximab with a 50% progression rate at the first assessment of disease progression and a median progression-free survival that did not differ between the groups (1.8 months in the supportive-care group versus 1.9 months in the cetuximab group). The disease was stable or responded to therapy in only 39.4% of the patients in the cetuximab group, a result indicating a need for predictive biomarkers to identify patients who could benefit from such treatment (Jonker et al., 2007).

Tumor samples obtained from 394 of 572 patients were analyzed to look for activating mutations of the K-ras gene. Of the tumors evaluated, 42.3% in the supportive care group and 40.9% in the cetuximab group had at least one mutation. The interaction of K-ras mutation status with overall survival and progression-free survival was significant with $p = 0.01$ and $p < 0.001$, respectively, indicating an association between these variables and genetic status. The main results of the analysis are shown in Table 1.7. The subgroup analyses eventually led to a labeling

[6] v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog

Table 1.7 Median (progression-free) survival and hazard ratios

	Survival (months)			PFS (months)		
K-ras	Cet	SC	HR (95% CI)	Cet	SC	HR (95% CI)
Mutation	4.5	4.6	0.98 (0.70, 1.37)	1.8	1.8	0.99 (0.73, 1.35)
Wild type	9.5	4.8	0.55 (0.41, 0.74)	3.7	1.9	0.40 (0.30, 0.54)

PFS = progression-free survival, Cet = cetuximab, SC = supportive care, HR = hazard ratio, CI = confidence interval. Source: (Karapetis et al., 2008)

revision by FDA in 2009: "Retrospective subset analyses of metastatic or advanced colorectal cancer trials have not shown a treatment benefit for Erbitux in patients whose tumors had KRAS mutations in codon 12 or 13. Use of Erbitux is not recommended for the treatment of colorectal cancer with these mutations."[7]

1.4.4 In Search for Benefit

Another interesting debate on the interpretation of the overall study result and subgroup analyses took place in the Journal of Clinical Epidemiology in 2010. The authors of the CAPRIE[8] study (CAPRIE Steering Committee, 1996) felt annoyed by the, in their opinion, inconsistent assessments of two agencies, the British National Institute for Health and Clinical Excellence (NICE) and the German Institute for Quality and Effiency in Healthcare (IQWiQ) (Hasford et al., 2010).

CAPRIE compared the use of clopidogrel versus aspirin in the secondary prevention of vascular events (myocardial infarction (MI), ischemic stroke (IS), or vascular death) in patients with atherothrombosis as diagnosed by recent MI, recent stroke, or symptomatic peripheral arterial disease (PAD). The intention-to-treat (ITT) analysis of 19 185 patients showed an annual 5.32%

[7] https://www.accessdata.fda.gov/drugsatfda_docs/label/2009/125084s167lbl.pdf

[8] CAPRIE = Clopidogrel versus aspirin in patients at risk of ischemic events

risk of IS, MI, or vascular death under clopidogrel compared with 5.83% under aspirin, yielding a statistically significant ($p = 0.043$) relative risk reduction of 8.7% in favor of clopidogrel (95% confidence interval 0.3 to 16.5).

The study protocol considered a stratified randomization and sample size planning accounting for different event rates for MI, stroke or PAD. The publication contained an analysis for each of these subgroups summarized as in Table 1.8 and using a forest plot as shown in Figure1.1. A treatment by subgroup interaction test was statistically significant ($p = 0.042$). Despite that, FDA approved clopidogrel for the entire study population in 1997 mentioning in their report that it is not clear whether the difference is real or a chance occurrence. In a meta-analysis, the agency discovered the strongest results of aspirin in patients with recent MI and lower efficacy in other subgroups (FDA, 1997). IQWiQ, however, considered these subgroups as preplanned and

Table 1.8 Relative risk reduction with 95% confidence interval (CI) by subgroup and overall in CAPRIE study

Subgroup	RRR(%)	95% CI	*p*-value
Stroke	7.3	(−5.7, 18.7)	0.26
MI	−3.7	(−22.1, 12.0)	0.66
PAD	23.8	(8.9, 36.2)	0.0028
All	8.7	(0.3, 16.5)	0.043

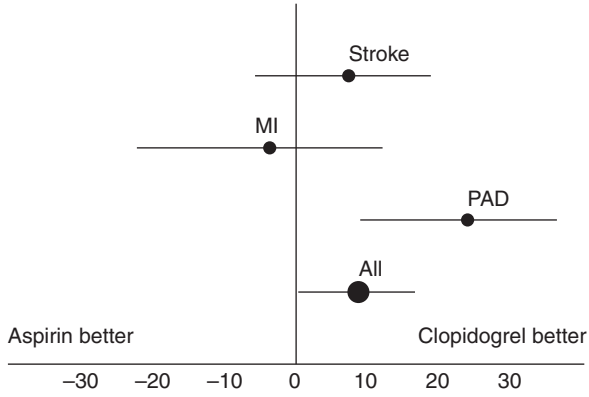

Figure 1.1 Relative risk reduction with 95% confidence intervals by subgroup and overall in CAPRIE study.

acknowledged an additional benefit of clopidogrel exclusively for the PAD subgroup because of heterogeneity of results (Bender et al., 2010).

In the view of Hasford et al. (2010), NICE concluded with respect to efficacy in accordance with the primary analysis of the overall population. However, it considered the balance between clinical effectiveness and cost-effectiveness not to justify a replacement of aspirin by clopidogrel to prevent vascular events. Therefore, NICE's and IQWiQ's conclusions were not that far apart since both agencies stated that the data do not show consistent benefit of clopidogrel but rather non-inferiority. In the last editorial on the topic (Hasford et al., 2011) the opponents concluded that "Standards for subgroup analyses are needed?—we could not agree more."

1.5 CONCLUDING REMARKS

This chapter has presented a selection of the good, the bad, and the ugly subgroup analyses from the past. In those cases that were part of submissions, regulatory agencies had to make a decision on how the results from subgroups affect the label of the compound. It becomes obvious from this account that there is at times a fine line between the reasonable and the ridiculous. Unfortunately, the ridiculous is not always as obvious as the star-sign determined subgroups in the ISIS–2 study.

In many cases, a biological rationale is helpful to "separate the wheat from the chaff." However, preferably the rationale should be available at the planning stage and not post hoc, otherwise it may become triggered by suggestive results. In any case, some form of replication of results should occur to assess credibility of subgroup findings. Comparing results with other research and searching for corroboration should be undertaken. For results with a big impact replication may be the action of choice or, as Rothwell (2005) has put it: the best test of the validity of subgroup analyses is not significance but replication.