10

Subgroup Analyses

10.1 Bona Fide Gems or Fool’s Gold
Well-trained research investigators diligently work to identify every potentially valuable result. Having invested great time and effort in their studies, these scientists want and need to examine the data systematically and completely. They are well aware that interesting findings await them in non-prospectively declared analyses. Investigators believe that like real gems, these tantalizing surprises lie just below the surface, hidden from view, waiting to be unearthed. If the experiment demonstrated that an intervention reduces the incidence of heart attacks, then ethnicity or gender may affect the relationship between therapy and heart attacks.

Others can also raise these same intriguing questions. In the process of publication, reviewers and editors of the manuscript will sometimes ask that additional analyses be carried out. These analyses may warrant the consideration of the effect of the intervention in subsets of the data. Does the therapy work equally well in the young and old? Does it work equally in different regions of the world? What about in patients with a previous heart attack? These analyses are demanded by others but are also not prospectively stated. Similar inquiries can come from the manuscript’s readers.

The research program’s cost-effectiveness and the investigator’s desire for thoroughness require that all facets of a research effort’s data be thoroughly examined. After all, why collect the data if it will not be considered in an analysis? However, as we have seen earlier, the interpretation of non-prospective analyses is fraught with difficulty. The need to protect the community from the dissemination of mistaken results runs head-on into the need to make maximum use of the data that has been so carefully collected. These problems are exemplified in subgroup analyses.

As we will see, there may be no better maxim for guiding the interpretation of subgroup analyses in this setting than “Look, but don’t touch.”

10.2 What Are Subgroups?
A subgroup analysis is the evaluation of the exposure–disease relationship within a fraction of the recruited subjects. The analysis of subgroups is a popular, necessary,
Subgroup analysis as currently utilized in clinical research is tantalizing and controversial. As described in the beginning of Chapter Three, the results from subgroup assessments have traditionally been used to augment the persuasive power of a clinical trial’s overall results by demonstrating the uniform effect of the therapy in patients with different demographic and risk factor profiles. This uniformity leads to the development of easily understood and implemented rules to guide the use of therapy. Some clinical trials report these results in the manuscript announcing the trial’s overall results [1–4]. Others have separate manuscripts dealing exclusively with subgroup analyses [5–7]. Such subgroup analyses potentially provide new information about an unanticipated benefit (or hazard) of the clinical trial’s randomly allocated intervention.

However useful and provocative these results can be, it is well established that subgroup analyses are often misleading [8–11]. Assmann et al. [12] have demonstrated how commonly subgroup analyses are misused, while others point out the dangers of accepting subgroup analyses as confirmatory [13]. Consider the amlodipine controversy.

10.3 The Amlopidine Controversy
In the 1980s, the use of calcium channel blocking agents in patients with CHF was problematic. While initial studies suggested that patients with CHF experienced increased morbidity and mortality associated with these agents [14], additional developmental work on this class of medications proceeded. In the early 1990s, new calcium channel blocking agents appeared, and early events suggested they might be beneficial.

To evaluate this possibility formally, the Prospective Randomized Amlodipine Survival Evaluation (PRAISE) [15] trial was designed. PRAISE’s long-term objective was the assessment of the channel blocker amlodipine’s effect on morbidity and mortality in patients with advanced heart failure. The primary measurement in PRAISE was the composite endpoint of all-cause mortality and/or hospitalization. The protocol also stipulated that there would be analyses in the following subgroups of patients: sex, ejection fraction, NYHA class, serum sodium concentration, angina pectoris, and hypertension.

PRAISE began recruiting patients in March 1992. Patients with CHF (NYHA functional class IIIb/IV and LVEF < 30%) were randomized to receive either amlodipine or placebo therapy. Suspecting that the effect of amlodipine might depend on the cause of the patient’s CHF, the investigators stratified randomization into two groups, patients with ischemic cardiomyopathy and patients who

*The finding that a particular lipid lowering drug works better in women than in men can complicate the already complex decisions that practitioners must make as the number of available compounds increase.
had non-ischemic cardiomyopathy.* PRAISE followed 1153 patients for 33 months. At the conclusion of PRAISE, the investigators determined that amlodipine was not effective in the overall cohort for either the primary or secondary analyses.

The evaluation then turned to the etiology-specific CHF subgroups. PRAISE recruited 732 patients with an ischemic cause for their CHF and 421 patients with a non-ischemic cause. The analysis of the effect of therapy in these strata revealed that treatment with amlodipine reduced the frequency of primary and secondary endpoints in patients with non-ischemic dilated cardiomyopathy (58 fatal or nonfatal events in the amlodipine group and 78 in the placebo group, 31% risk reduction; 95% CI 2 to 51% reduction; \( p = 0.04 \)). Further evaluation of these events revealed that there were only 45 deaths in the amlodipine group while there were 74 deaths in the placebo group, representing a 46% reduction in the mortality risk in the amlodipine group (95% CI 21 – 63% reduction, \( p < 0.001 \)).

The therapy appeared to be effective in the non-ischemic cardiomyopathy stratum. Treatment with amlodipine did not affect the combined risk of morbidity and mortality in the ischemic cardiomyopathy group.

Thus, although amlodipine did not produce an overall beneficial effect, it reduced the combined incidence rate of all cause hospitalization/total mortality, and the total mortality rate in patients with non-ischemic dilated cardiomyopathy.

A second trial, PRAISE-2 [16], was conducted to verify the beneficial effect on mortality seen in the subgroup analysis of patients in PRAISE-1. PRAISE-2 was similar in design to PRAISE-1, the exception being its focus on patients with non-ischemic cardiomyopathy. The PRAISE-2 investigators randomized 1,650 patients to either amlodipine or placebo, following them for up to 4 years. However, the results of PRAISE-2 were quite different from PRAISE-1. Unlike the first study, there was no difference in mortality between the two groups (33.7% in the amlodipine arm and 31.7% in the placebo arm; odds ratio 1.09, \( p = 0.28 \)) in PRAISE-2. Thus, the marked mortality benefit seen in the subgroup analysis in PRAISE-1 for amlodipine was not confirmed in PRAISE-2. The positive subgroup analysis could not be confirmed.

Another example of the misleading results that subgroup analyses can provide are from MRFIT, discussed in Chapter Two. In that case, a misleading finding in a subgroup analysis suggested that the use of antihypertensive medications in a subgroup of hypertensive men who had ECG abnormalities at baseline would be harmful.

Nevertheless, the medical community continues to be tantalized by spectacular subgroup findings from clinical trials. A recent example is the subgroup analysis-based suggestion that medication efficacy is a function of race; this has appeared in both peer-reviewed journals [17–18] and the lay press [19].

* Stratified randomization is an adaptation of the random allocation to therapy process. In general, the random allocation of therapy ensures that each patient has the same chance of receiving active therapy as any other patient. Stratified randomization is randomization confined to a small subgroup of patients. Without it the small size of the subgroup may not permit the overall randomization process to be balanced.
In this chapter, we will review the definitions, concepts, and limitations of subgroup utilization in clinical trials.

10.4 Definitions
While the concept of subgroup analyses is straightforward, the terminology can sometimes be confusing.

A subgroup is the description of patient based characteristic, e.g., gender, that can be subdivided into categories. For example, if an investigator is interested in creating a gender subgroup, patients are classified into one of two groups—male or female. These groups are referred to as levels or strata. There is one stratum for each category.

The traditional subgroup analysis is an evaluation of the effect of therapy within each of the subgroup strata. In a gender-based subgroup, the subgroup analysis consists of an evaluation of the effect of therapy for males and an evaluation of the effect of therapy for females. Thus, each stratum analysis produces an effect size with standard error, a confidence interval, and \( p \)-value.

The definition of a subgroup can be complicated. Consider the categories for a subgroup entitled race. In the 1970s, US-based healthcare research commonly (although somewhat inappropriately) divided their participants into one of two strata: white or black. However, these strata were forced to expand, evolving with the changing US demographics. They now rarely include less than the four categories—white, black (or African-American), Hispanic, and Asian. However, commonly, contemporary research takes a more expansive view of this issue, expanding the consideration to not just race, but ethnicity as well. The nature of the research helps to guide the choice for the number of subgroup strata.

The most useful rules of thumb in developing subgroup strata are to (1) know the demography of the group from whom subjects will be recruited, and (2) understand how the subgroup analysis will be used in interpreting the research result.

10.5 Interpretation Difficulties
The illustrations of the previous sections have demonstrated that there are many possible ways to group patients. Investigators work to identify subgroup classifications that are meaningful. When examination of the therapy effect within a subgroup appears, it is only natural for the investigator to believe the rationale for the choice of the subgroup is justified. Furthermore, the scientist may think that the stratum-specific therapy effect is due to some effect-mediation ability produced by the subgroup trait. However, the very fact that patients are classified and divided can induce a subgroup effect of its own. Subgroup analyses commonly classify and reclassify patients making the attribution of effect unclear.

10.5.1 Re-evaluating the Same Patients
In addition, we must keep in mind that in a collection of subgroup analyses, patients may be stratified, then re-stratified in different ways. This observation can complicate the interpretation of a subgroup evaluations.
For example, consider a randomized, controlled clinical trial that is in its analysis phase. At this time, all of the patients are grouped into either having or not having diabetes mellitus. Once the stratum membership assignment is finished, the effect of the randomly allocated intervention is assessed in each of the two strata. The analyses reveal that the therapy has a greater effect in non-diabetic patients than in those with diabetes.

When completed, these same patients are then re-aggregated based not on diabetes but on age. Subjects are placed into one of the following three age strata: (1) less than 30 years of age, (2) between 30 and 60 years of age, and (3) greater than 60 years of age. When the subgroup analysis is carried out for the age strata, it appears that the effect of therapy is reduced in the older age group.

The investigators report that both age and diabetes modify the effect of therapy. Holding aside the complexities of generalizing from the sample to the population, focus just on the sample, and consider how we know which of these subgroups is responsible for the observed effects? All the investigator has done is to reclassify the same patients in two different ways. Perhaps the older patients are more likely to be diabetic. If this is the case, the diabetes subgroup effect may simply be another manifestation of the age effect. Thus the findings reported out as two different effects may simply be one effect, expressing itself in two different analyses.

It can be misleading to report both, because the investigator has simply reclassified the same patients. The results from these two subgroup analyses essentially demonstrate that the same patients when characterized one way (by diabetes) provide a different result than when characterized another way (by age). Was it really diabetes that modified the effect of therapy or was it the chance collection of patients that made it appear that diabetes gender was an influence? Since we can expect that the effect of treatment within a subgroup stratum depends on the patients within that stratum, then the value of the subgroup analysis must be tightly linked to the ability to demonstrate that the stratum characteristic (and not the random aggregation of patients) is producing the interesting effect.

### 10.5.2. Random Findings

It is difficult to separate a true stratum-specific “signal” from the random background “noise.”

Consider the following simple experiment. A classroom chosen at random has a capacity of seating 80 students. These 80 seats are divided by a central aisle, with 40 seats on each of the left-hand and right-hand side of the courtroom. Eighty students seat themselves as they choose, distributing themselves in an unrestricted manner among the seats on each side of the class. When all are seated, we measure the height of each person, finding that the average height is exactly 69 inches. Does that mean that the average height of those seated on the left-hand side of the classroom will be 69 inches? No, because the 69 inch measurement was produced from all eighty students in the room, not just the forty on the left-hand side. The sample of forty has a different mean simply because they are only part of and not the entire population of 80.
If the average height on the left-hand side of the classroom is greater than 69 inches, then those seated on the right-hand side will have an average height less than 69 inches. Thus, those sitting on the left-hand side have a greater height than those on the right-hand side. While that fact is undeniable in this one classroom during this one seating, would it be fair to generalize this conclusion to the population at large?

No. The random aggregation of observations has induced a subgroup effect that is based only on the play of chance here. Specifically, the “subgroup effect” was induced by selectively excluding individuals from the computation of the mean.

An illustration of this principle is how subgroups are induced by the process of sampling. Consider the result of a hypothetical clinical trial in which the investigators report that the randomly allocated intervention produced a 25% reduction in the prospectively defined endpoint of total mortality. The initial reaction to the demonstration of a beneficial effect of therapy in a well-designed, well-executed clinical trial is naturally to assume that the effect of therapy is homogenous. Our first response is therefore to believe that all collections of patients in the active group were beneficiaries of this 25% benefit, and that the beneficial effect of therapy provides protection that is broadly distributed (Figure 10.1, panel 1.) This treatment effect uniformity is the truth about the population at large. However, the examination of the same therapy effect within different subgroups of the clinical trial sample reveals a mosaic of treatment effect magnitudes (Figure 10.1, panel 2).

At first glance, it appears that the uniform mortality benefit has been replaced by a much more heterogeneous response. The uniform 25% reduction in the total mortality effect is still there; the population from which the research sample was derived still experienced a 25% reduction in the proportion of deaths. However, when that uniform effect is viewed through the prism of a small sample, the uniform effect is distorted. The same process that produced a difference in average height on either side of the room, i.e., the exclusion of individuals from the computation of the treatment effect, generates the differences in subgroup effect sizes in Figure 10.1, panel 2. Much as the random aggregation of students in the class from the previous example produced a subgroup effect, the random aggregation of patients in research efforts will induce a heterogeneity of responses.

An investigator, unaware of this effect-distorting phenomenon and having only the data from Figure 10.1, panel 2, believes that the response to treatment is heterogeneous since his data demonstrates variability between, for example, men and women, or African-Americans and Caucasians. However, this variability is not induced by a complex interrelationship among genetics, environment, and therapy response. It is induced solely by the sampling process.

We must also keep in mind that the reverse effect is possible. The sample may suggest that there is no differential effect of therapy between subgroup strata when in fact one is present, again, because of sampling error.

*If the average height of all in the classroom is 69 inches, and the average height on the left side is greater than 69 inches, then the average height on the right-hand side must be less than 69 inches in order for the average of all to be 69 inches.
The random selection of data from the population produces sampling error. Its presence will produce false findings and “red herrings”, just through the random aggregation of subjects in the population. However the data will also accurately reflect relationships that are embedded in the population. This random subgroup effect appears in all subgroup analyses, and we will have to integrate it into our interpretation of any subgroup effect that we see.

**Fig. 10.1.** Homogenous population effect (Panel 1) appears as a variable treatment effect in the sample (Panel 2).

It therefore is difficult correctly to classify a relationship that is observed in the data. Unfortunately, exploratory subgroup analysis is an imprecise means of classification. For example, in the MRFIT trial, the identification of a relationship that suggested that the treatment of hypertension may be harmful in men with abnormal hearts confused the medical community for several years. In the end, this finding was attributed to sampling error. These occurrences help to justify the admonition that the best descriptor of the effect in a subgroup is the finding that is observed in the overall cohort.

### 10.5.3 Clinical Trial-Mediated Subgroup “Effects”

As another example, consider the effect of therapy in a clinical trial designed to assess the role of LDL cholesterol level modification in reducing the incidence of fatal heart attack/nonfatal heart attack/revascularization. The analysis of the effect of therapy on this endpoint in the entire cohort was overwhelmingly positive. The effect of therapy in a variety of post hoc subgroups was also examined (Table 10.1).

Table 10.1 reveals the effect of therapy within each of five subgroups (A through E) in the clinical trial. Each subgroup contains two different strata. For

* See Chapter Two.
each stratum, the number of patients in the subgroup at risk for the disease, the number of patients in the subgroup who have the disease, and the cumulative incidence rate of the endpoint are provided. To assess the effect of therapy, both the relative risk of therapy and the $p$-value are included. For example, subgroup A contains two strata: I and II. In strata I, there are 576 patients; 126 of them experienced the endpoint. The cumulative incidence of the endpoint in the placebo group was 27.59 (i.e., 27.59% of these patients experienced a clinical endpoint), and 16.08 in the active group. The relative risk due to therapy was 0.54, i.e., the risk of an event in the active group was 54% of the risk of that event in the placebo group, suggesting that there was some benefit was associated with therapy. The $p$-value for this protective effect is 0.001. In this subgroup, the effect of therapy was notable in each of the two strata, (relative risk of 0.54 in stratum I versus 0.80 in strata II) with a further suggestion that patients in strata I received a greater benefit from therapy than patients in stratum II.

Examination of the data suggest that the therapy effect is not uniform across the subgroup strata, providing benefit for some and hazard for others. In subgroup C, for example, the relative risk for the event is 1.02 ($p$-value = 0.916) for strata I, suggesting no benefit. In stratum II of this subgroup, the relative risk is 0.71 ($p$-value < 0.001). Does this suggest that patients in subgroup C, strata I should not receive the therapy? If true, this would be an important message to disseminate, especially if the therapy was associated with significant cost or side effects.

There is a similar finding for subgroup D: again, stratum I demonstrates no effect of therapy while strata II demonstrates a possibly important effect. The analyses in subgroups C and D suggest that there may be a stratum dependent therapy, while subgroup E demonstrates that the effect of therapy was the same in each of its strata.

However, we have only the relative risk and the $p$-value to convey the significance of the therapy effect within subgroups, and our assessment of their joint message tempt us to draw conclusions about which of the subgroup strata demonstrated a therapy effect and which did not. In fact, it’s not hard to envision the explanations of imaginative investigators that would be offered to explain the different subgroup effects.

Table 10.2 reveals a more useful explanation. This table contains the same data as Table 10.1 and also the identities of the subgroups. We see that subgroups A, B, and C are clinical subgroups of direct relevance to the clinical investigation. However, subgroups D and E are completely random subgroups, i.e., a random number was used to generate the strata membership for these two subgroups. Nevertheless, the statistical findings of these random subgroups are equally as persuasive as the findings in the subgroups of clinical relevance. The results in the random subgroups mimic the findings in the “real subgroup.” From a comparison of Table 10.1 and 10.2, we see that one cannot differentiate a meaningful effect of therapy from just the random play of chance by examining the results.
### Table 10.1. Effect of therapy on endpoint incidence within unknown subgroups: Fatal and nonfatal MI plus revascularizations

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Patients</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Endpoint</td>
<td>80</td>
<td>46</td>
<td>126</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>290</td>
<td>286</td>
<td>576</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>Cum. Inc Rate *100</td>
<td>27.59</td>
<td>16.08</td>
<td>21.88</td>
<td>0.001</td>
</tr>
<tr>
<td>II</td>
<td>Endpoint</td>
<td>469</td>
<td>384</td>
<td>853</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1788</td>
<td>1795</td>
<td>3583</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cum. Inc Rate *100</td>
<td>26.23</td>
<td>21.39</td>
<td>23.81</td>
<td>0.001</td>
<td>0.80</td>
</tr>
<tr>
<td>I</td>
<td>Endpoint</td>
<td>258</td>
<td>217</td>
<td>475</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1003</td>
<td>1027</td>
<td>2030</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Cum. Inc Rate *100</td>
<td>25.72</td>
<td>21.13</td>
<td>23.40</td>
<td>0.018</td>
</tr>
<tr>
<td>II</td>
<td>Endpoint</td>
<td>291</td>
<td>213</td>
<td>504</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1075</td>
<td>1054</td>
<td>2129</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cum. Inc Rate *100</td>
<td>27.07</td>
<td>20.21</td>
<td>23.67</td>
<td>&lt;0.001</td>
<td>0.73</td>
</tr>
<tr>
<td>I</td>
<td>Endpoint</td>
<td>98</td>
<td>95</td>
<td>193</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>461</td>
<td>439</td>
<td>900</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>Cum. Inc Rate *100</td>
<td>21.26</td>
<td>21.64</td>
<td>21.44</td>
<td>0.916</td>
</tr>
<tr>
<td>II</td>
<td>Endpoint</td>
<td>451</td>
<td>335</td>
<td>786</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1617</td>
<td>1642</td>
<td>3259</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cum. Inc Rate *100</td>
<td>27.89</td>
<td>20.40</td>
<td>24.12</td>
<td>&lt;0.001</td>
<td>0.71</td>
</tr>
<tr>
<td>I</td>
<td>Endpoint</td>
<td>105</td>
<td>91</td>
<td>196</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>397</td>
<td>402</td>
<td>799</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Cum. Inc Rate *100</td>
<td>26.45</td>
<td>22.64</td>
<td>24.53</td>
<td>0.210</td>
</tr>
<tr>
<td>II</td>
<td>Endpoint</td>
<td>444</td>
<td>339</td>
<td>783</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1681</td>
<td>1679</td>
<td>3360</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cum. Inc Rate *100</td>
<td>26.41</td>
<td>20.19</td>
<td>23.30</td>
<td>&lt;0.001</td>
<td>0.75</td>
</tr>
<tr>
<td>I</td>
<td>Endpoint</td>
<td>186</td>
<td>137</td>
<td>323</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>653</td>
<td>639</td>
<td>1292</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Cum. Inc Rate *100</td>
<td>28.48</td>
<td>21.44</td>
<td>25.00</td>
<td>0.011</td>
</tr>
<tr>
<td>II</td>
<td>Endpoint</td>
<td>363</td>
<td>293</td>
<td>656</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1425</td>
<td>1442</td>
<td>2867</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cum. Inc Rate *100</td>
<td>25.47</td>
<td>20.32</td>
<td>22.88</td>
<td>0.001</td>
<td>0.77</td>
</tr>
</tbody>
</table>
### Table 10.2 Effect of therapy on endpoint incidence within known subgroups:
Fatal and nonfatal MI plus revascularizations

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Patients</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>Active</td>
<td>Total</td>
<td>P-value</td>
<td>Rel Risk</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>Endpoint</td>
<td>80</td>
<td>46</td>
<td>126</td>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>290</td>
<td>286</td>
<td>576</td>
<td></td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>Cum. Inc Rate *100</td>
<td>27.59</td>
<td>16.08</td>
<td>21.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>Endpoint</td>
<td>469</td>
<td>384</td>
<td>853</td>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1788</td>
<td>1795</td>
<td>3583</td>
<td></td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Cum. Inc Rate *100</td>
<td>26.23</td>
<td>21.39</td>
<td>23.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>Endpoint</td>
<td>258</td>
<td>217</td>
<td>475</td>
<td></td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1003</td>
<td>1027</td>
<td>2030</td>
<td></td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Cum. Inc Rate *100</td>
<td>25.72</td>
<td>21.13</td>
<td>23.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>Endpoint</td>
<td>291</td>
<td>213</td>
<td>504</td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1075</td>
<td>1054</td>
<td>2129</td>
<td></td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Cum. Inc Rate *100</td>
<td>27.07</td>
<td>20.21</td>
<td>23.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDL</td>
<td>Cholesterol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>Endpoint</td>
<td>98</td>
<td>95</td>
<td>193</td>
<td></td>
<td>0.916</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>461</td>
<td>439</td>
<td>900</td>
<td></td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>Cum. Inc Rate *100</td>
<td>21.26</td>
<td>21.64</td>
<td>21.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>Endpoint</td>
<td>451</td>
<td>335</td>
<td>786</td>
<td></td>
<td>0.210</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1617</td>
<td>1642</td>
<td>3259</td>
<td></td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>Cum. Inc Rate *100</td>
<td>27.89</td>
<td>20.40</td>
<td>24.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>Endpoint</td>
<td>105</td>
<td>91</td>
<td>196</td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>397</td>
<td>402</td>
<td>799</td>
<td></td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Cum. Inc Rate *100</td>
<td>26.45</td>
<td>22.64</td>
<td>24.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>Endpoint</td>
<td>444</td>
<td>339</td>
<td>783</td>
<td></td>
<td>0.210</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1681</td>
<td>1679</td>
<td>3360</td>
<td></td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>Cum. Inc Rate *100</td>
<td>26.41</td>
<td>20.19</td>
<td>23.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>Endpoint</td>
<td>186</td>
<td>137</td>
<td>323</td>
<td></td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>653</td>
<td>639</td>
<td>1292</td>
<td></td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Cum. Inc Rate *100</td>
<td>28.48</td>
<td>21.44</td>
<td>25.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>Endpoint</td>
<td>363</td>
<td>293</td>
<td>656</td>
<td></td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1425</td>
<td>1442</td>
<td>2867</td>
<td></td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Cum. Inc Rate *100</td>
<td>25.47</td>
<td>20.32</td>
<td>22.88</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Two points are to be made from this evaluation. The first is that neither the
\( p \)-value nor the relative risk discriminated between the clinical versus the random
subgroup. Relying on these measures can be quite misleading in subgroup analysis.
Secondly, it is not so much that the random subgroup looked like the clinical sub-
group. It is more to the point to say that the clinical subgroups are more likely to
produce random findings just through the aggregation of events.

Even if we concentrate on the first three subgroups whose subgroup defini-
tions are plausible and have scientific meaning, we still cannot be sure that the ef-
teffects demonstrated in Table 10.2 truly represent the findings in the population. The
gender examination suggests that women derive a greater benefit from cholesterol
reduction therapy than do men. There appears to be no real difference in effect of
therapy across the different age strata. Also, it appears that patients with baseline
LDL cholesterol greater than 125 mg/dl derive a profound effect from therapy
while those patients with a baseline LDL cholesterol \( \leq 125 \) mg/dl obtain no benefit
from therapy.

10.5.4 The Importance of Replication

These findings for women and patients with baseline LDL cholesterol \( \leq 125 \) mg/dl
were obtained from the CARE clinical trial and published by Sacks et al. [2], pro-
ducing much discussion among lipidologists. A follow-up manuscript elaborating
on the effect of cholesterol reduction therapy in women was published by Lewis et
al [6]. A manuscript examining the relationship between LDL cholesterol and clinici-
cal endpoints was also published in 1998 [7]. The subgroup findings from CARE
and the subsequent published manuscripts based on these findings were surprising
and useful, generating much debate. In neither case was the hypothesis stated pro-
spectively (with alpha allocation) in the protocol. Yet, in each case the subgroup
analysis was relevant and insisted upon by the scientific community. How can one
interpret the findings of these required but non-prospective evaluations?

One relevant tool for assessing subgroup results is the criterion of confir-
manation or replication of an experiment’s results. In its narrowest form, study repli-
cation involves the slavish reproduction of previous numeric results using the origi-
nal experimental protocol; its goal is to limit investigator error (either inadvertent or
intentional). More broadly, replication involves the systematic variation of research
conditions to determine whether an earlier result can be observed under new condi-
tions. This is particularly important when a number of different study designs are
used. The variety of study designs, each with its own unique strengths and weak-
nesses, minimizes the chance that all of the studies are making the same mistake.
Consistency is demonstrated when several of these studies giving the same results
[20].

It is this broader, more useful interpretation that would be most helpful in
interpreting the results from CARE. An independent clinical trial carried out in
Australia [4], completed after the end of CARE, examined the effect of the same
cholesterol reducing therapy. This second study differed in that (1) the Australian

* The protocol is the prospectively written plan which outlines the goal, design, execution,
and analysis of the experiment.
study (known by the acronym LIPID) was twice as large as CARE, (2) the entry criteria were somewhat different, and (3) the clinical endpoint LIPID was coronary heart disease death, a stronger endpoint than the well-accepted combined endpoint of fatal heart attack/nonfatal heart attack/revascularization used in the CARE sub-group analysis. The LIPID investigators carried out the subgroup analyses in women and in patients with low levels of baseline LDL cholesterol. They found no differential effect of therapy in women, a finding that contradicted the observation of CARE. However, the Australian study did replicate the baseline LDL cholesterol subgroup findings first identified in CARE. Thus, one would give more credence to the replicated finding involving baseline LDL cholesterol than to the finding involving women. The lesson is that subgroup analyses from one study are difficult to interpret on their own; they should be reproduced in an independent study before they are considered reliable.

10.6 Stratified Randomization

Another difficulty in the interpretation of subgroup analyses in clinical trials is that the patient classification process can undo one of the most important features of these research experiments — the ability to attribute differences in observed endpoints to the randomly allocated intervention. The absence of this key feature complicates the interpretation of the subgroup analysis.

Consider a clinical trial that has a control group and a treatment group. The random allocation of therapy can clarify a study’s results by requiring each of its patients have their therapy assignment based on factors separate from his/her own traits. Frequently, this means that each patient has the same probability of receiving the active intervention as the next patient. This feature distributes patients between the control and active group in such a way that the only difference between the two groups is that one group received active therapy and the other received control group therapy. There are not likely to be any important demographic, sociologic, or risk factor differences between the two groups. Therefore, differences between endpoint rates that occur at the end of the trial can be attributed to the only difference between the two groups — the randomly allocated therapy.

These finely balanced circumstances are perturbed when subgroups are analyzed. Unfortunately, membership in the subgroup stratum of interest may be very low, and randomization may not have had a real opportunity to balance patient distribution in this stratum. Thus, the effect of therapy within that particular subgroup stratum may be confounded, i.e., confused and intertwined with other characteristics that are different between the treatment groups. It can be virtually impossible to persuasively attribute any differences between the treatment groups to the therapy within this one stratum.

* Adaptations of this procedure can be executed to ensure that there is an exact balance between the number recruited to each of the control and treatment arms. However, therapy is not assigned based on a patient’s characteristic.
† The importance of prospectively declared scientific questions was addressed in Chapter Two.
Knowledgeable investigators anticipate the distribution of patients across the subgroup and the resultant small number of patients within some of the subgroup’s strata. These investigators sometimes actually force the random allocation therapy to work. That is, they adjust the randomization algorithm so that there are an equal number of patients allocated to each of the control and active groups of the trial within the subgroup stratum. This randomization within the strata, or stratified randomization, ensures that even though there is a small number of patients within the stratum, the allocation of therapy will be more effectively balanced. While not compensating for the different endpoint event rates stemming from the relatively small number of patients in the stratum, this adaptation does substantially improve the balance of baseline characteristics between the treatment groups. By doing this, investigators strengthen their ability to ascribe differences in the mortality rate to the randomly allocated therapy.

10.7 Proper Versus Improper Subgroups

A critical preliminary task in subgroup analysis is the proper classification of patients into each of the subgroup strata. Although membership determination may appear to be a trivial task, there are circumstances in which this classification is problematic. These concerns revolve around the timing of the subgroup membership determination.

There are two important possibilities for determining of the timing of subgroup membership. The first is the classification of patients into the correct subgroup stratum when the patients are randomized. The second choice is to classify patients into subgroup strata at some time during the execution of the trial. While each has advantages, the determination of subgroup membership at the beginning of the study is preferred.

Determining subgroup membership at the beginning of the trial requires not only that the subgroup must be defined at the beginning of the study, but also that the subgroup strata membership should be defined prospectively as well. This is a straightforward procedure to apply to the gender subgroup with its two strata. However, for other subgroups of clinical interest, the process can be complex. For example, while it can be relatively straightforward to evaluate the relationship between cholesterol levels subgroup strata 1) less than 175 mg/dl, and 2) greater than 175 mg/dl and the cumulative incidence of stroke, the evaluation of these strata when they are based in follow-up levels of cholesterol is problematic.

The problems arise for two reasons. The first is that patients can change subgroup strata as the study progresses and the their cholesterol levels fluctuate. This makes it difficult to determine definitively and convincingly subgroup membership, and the analysis can suffer from the observation that changing the subgroup membership of just a few patients can change the results of the subgroup analysis. Such brittle evaluations are unpersuasive.

Secondly, there are many influences that effect lipid measurements over time. If the exposure being evaluated reduces cholesterol levels, then patients with lower cholesterol levels are more likely to have received active therapy, and patients with the higher levels would have a greater chance of being in the control group. Thus the evaluation of lipid levels will be confounded with exposure to the
agents after the study was initiated, confounding (i.e., confusing) the attribution of
the observed effect on the endpoint.

In our first example in this section, we acknowledged that there were many
factors that influence baseline lipid levels. Race, gender, family history, and prior
treatment are but a few of them. However, the randomly assigned intervention did
not influence baseline LDL-cholesterol levels. It is the absence of any relationship
between the randomly allocated therapy and the baseline LDL-cholesterol level that
allows a clear examination of the effect of LDL-cholesterol level on the relationship
between the intervention and stroke. A subgroup whose strata membership criteria
are based on baseline characteristics of the patient is called a proper subgroup [21].
Improper subgroups are those whose strata membership can only be determined
after the patient has been randomized. Membership based on follow-up data is in-
fluenced by randomly allocated therapy and the interpretation is complicated.

There are circumstances in which this type of analysis is nevertheless car-
rried out. If the investigators are interested in an evaluation of the effect of lower
blood pressure on the incidence of stroke, regardless of how the blood pressure was
lowered, then analysis procedures are available.* However these evaluations are
exceedingly complicated and the results must be interpreted with great caution.
Similar evaluations have examined the relationship between lipid lowering medi-
cations and atherosclerotic morbidity and mortality [22–24].

Finally, we will hold aside the issue of the analysis of a proper subgroup
defined post hoc. In that circumstance, the subgroup criteria using baseline vari-
ables is defined at the end of the study. Since the subgroup analysis was planned
after the data were examined, the analysis is exploratory.

10.8 “Intention-to-Treat” Versus “As Treated”
Consider a clinical trial in which patients are randomized to receive an intervention
to reduce the total mortality rate from end-stage renal disease. At the inception of
the study, patients are randomized to receive either control group therapy or the
intervention. At the conclusion of the study, the investigators will compare the cu-
mulative mortality rates of patients in each of the two treatment groups. However,
how will the investigators decide which patients should be assigned to each group
in the final analysis? The commonly used approach is to assign treatment group
membership simply as the group to which the patient was randomized. This is the
“intention to treat” principle.

The “intention-to-treat” principle of analysis is the standard analysis pro-
cedure for the evaluation of clinical trial results. Undoubtedly, this analysis tends to
be a conservative one, since not every patient is treated as they were “intended.”
For example, some patients randomized to the active group may not take their med-
ication. These patients, although randomized to the active group, will have the con-
tral group experience and produce endpoints at rates similar to that of the control
group. However, they would be included in the active group since they were ran-
domized to and “intended to be treated” like active group patients. The inclusion of
these patients in the active group for analysis purposes tends to make the active

* Cox hazard analysis with time-dependent covariates has been one useful tool in this regard.
10.9 Effect Domination Principle

The examination of individual subgroup strata effects in healthcare research can be misleading for reasons that have been elaborated. If we cannot believe the event rates that are present in the stratum are the best measures of that stratum’s response to the exposure, what then is the best measure of the effect of an exposure on a subgroup stratum?

The classroom illustration provided in section 9.5.1 reveals the answer. In that circumstance where the average height was greater for those sitting on one side of the room than the other, the best estimate of the average height of those who sit on the left side of the room is the overall average height of all in the class.

We are not saying that the average height should replace the actual heights of those on each side of the room. Clearly, the best estimator of the average height of those sitting on the left side of the room is the average height of precisely those people. However, we are saying that if one wished to generalize results to the population at large, the best generalization of the average height of people who sit on the allowing the overall measure of effect in the entire cohort to supersede the subgroup stratum effects can be termed the effect domination principle and is attributable to Yusuf et al. [21]

* There are occasional complications in an “intention-to-treat” analysis. In some cases, a patient is tested and randomized, but then, subsequent to the randomization the test result reveals that the patient is not eligible for the trial for a prospectively stated reason. In this case, there was no “intent” to randomize this patient when the test result was known, and the patient is removed from the study.

† The effect of the magnitude of the treatment effect on the power of a study for fixed sample size is elaborated in Appendix B.
Thus, in the example from the PRAISE I clinical trial discussed in section 10.3, the generalizable effect of therapy in those with non-ischemic cardiomyopathy was the overall effect of therapy in general. This was borne-out in the PRAISE II findings. This principle of effect domination is not very provocative and contains none of the excitement of exploratory analyses. However, it is far more reliable, given the general non-confirmatory analyses that the majority of subgroup analyses constitute in healthcare results.

10.10 Confirmatory Subgroup Analyses

Since subgroup analyses have and will, in all likelihood, continue to engender the interest of the medical community, it is logical to ask why there aren’t more confirmatory analyses involving subgroup evaluations. This is an especially interesting question since there are clear circumstances in which subgroup evaluations can produce confirmatory results of a therapy effect within (or across) subgroup strata. When executed, these confirmatory results stand on their own, separate and apart from the result of the effect of therapy in the overall cohort. The criteria for these evaluations are clearly characterized by and are coincident with our development of confirmatory analyses in this text.

The first of these criteria for the development of confirmatory analyses in clinical trials is that the subgroup analysis must be prospectively designed and proper. This structure is required so that (1) the therapy effect size estimators that the subgroup analysis produces are trustworthy; and (2) that the effect of therapy to be evaluated in a subgroup is not confounded by (i.e., bound up with) post-randomization events as discussed in the previous chapter. In general, there has been no difficulty with meeting this requirement of confirmatory subgroup analyses. Many clinical trials make statements in their protocols describing the plans of investigators to evaluate the effect of therapy within their subgroups of interest. These subgroups are, proper subgroups, e.g., demographic traits, or the presence of risk characteristics at baseline.

However, the final requirement for a confirmatory subgroup analysis is the prospective allocation of type I and type II error rates. This last criterion has proved to be especially troublesome because of the severe sample size constraints this places on subgroup analyses. As we have pointed out earlier, the allocation of type I error rates for confirmatory testing must be such that the FWER, $\xi$, is conserved. This requires that statistical testing at the level of subgroup analyses be governed by test-specific $\alpha$ error rates that are generally less than 0.05.

The difficulty of executing subgroup analyses in the presence of FWER control and adequate statistical power is not difficult to understand. In fact, resources are generally strained to the breaking point for the analysis of the effect of therapy in the overall cohort. This overall analysis is typically carried out with the minimum acceptable power (80%) because of either financial constraints or patient recruitment difficulties. By definition, subgroup analyses (and certainly within-stratum subgroup analyses) will involve a smaller number of patients; it is a daunting task to allocate prospectively type I and type II error rates at acceptable levels in a smaller number of patients, although the methodology for the accurate computation of sample size is available [26]. Thus, the growth of the use of subgroups as
confirmatory tools has, to some extent, been stunted by the difficulty of construct-
ing a prospective clinical trial with an embedded, prospectively defined proper sub-
group for which tight statistical control is provided for type I and type II statistical
errors.

10.11 Assessment of Subgroup Effects

The evaluation of subgroup effects in clinical trials focuses on the effect of the ran-
donally allocated therapy on the subgroup of interest. However this assessment can
be carried out in two complementary manners. The first is the determination of a
differential effect of therapy across subgroup strata. The second is the evaluation of
the effect of therapy within a single subgroup stratum. Each approach, when pros-
pectively planned and concordantly executed, can supplement the information
provided by the evaluation of the main effect of a clinical trial.

10.11.1 Effect Modification and Interactions

We commonly think of the effect of the randomly allocated intervention in a clini-
cal trial as an effect across the entire research cohort. The examination of a dataset
for this effect, while complicated, has become a routine part of the evaluation of the
randomly allocated therapy’s influence in a clinical trial. The finding of both clini-
cal and statistical significance for this analysis suggests that the effect of therapy is
different for one subgroup stratum than for another.

This type of subgroup effect is commonly referred to as a treatment by
subgroup interaction; a notable product of this analysis is the $p$-value for interac-
tion. Typically, the analysis result is described as identifying how the subgroup
strata interacts with the therapy to alter the occurrence of the endpoint, and the
evaluation is called an interaction analysis. Alternatively, this approach is de-
scribed as effect modification, i.e., it examines the degree to which the subgroup
stratum modifies the effect of treatment on the endpoint.

We should not be surprised by the observation that statistically significant
effect modification analyses in research are uncommon. The subgroup analyses
involve an evaluation of an effect difference between smaller subsets of patients
within the research cohort. Everything else being equal, the smaller sample sizes
reduce the statistical power of the hypothesis tests. The presence of a test statistic
that does not fall in the critical region in a low-power environment is not a null
finding, but instead is merely uninformative; many of these subgroup analyses are
unhelpful.

Alternatively, the occurrence of a statistically significant effect size can be
particularly noteworthy. An example of such a finding occurred in the Cholesterol
and Recurrent Events (CARE) clinical trial. CARE was designed to demonstrate
that the reduction in LDL cholesterol in patients with moderate lipid levels would
reduce the incidence of cardiovascular disease in patients with a history of MI. In
that study, an examination of the effect of the HMG-CoA reductase inhibitor
pravastatin was assessed in the gender subgroup. The relevant analysis was the ef-
ect of pravastatin on the cumulative incidence rate of the post hoc composite end-
point of CAD death + nonfatal MI + coronary revascularization [7]. There were
4,159 patients recruited in the CARE study; of these, 576 (13.8%) were women and
3,583 (86.2%) were men. During the course of the trial the effect of the randomly allocated intervention pravastatin on lipids appeared to be the same in women and men, producing equivalent reductions in total cholesterol (20% in women, 19% in men), low density lipid (LDL) cholesterol (28% in women, 28% in men), and triglycerides (13% in women, 14% in men). There were also equivalent elevations in high-density lipoprotein (HDL) cholesterol (4% in women, 5% in men).

However, the subgroup analysis revealed an apparent difference in the effect of pravastatin therapy on the expanded endpoint in men and women (Table 10.3).

<table>
<thead>
<tr>
<th>Gender</th>
<th>Relative Risk</th>
<th>Interaction P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>0.761</td>
<td></td>
</tr>
<tr>
<td>Females</td>
<td>0.545</td>
<td>0.050</td>
</tr>
</tbody>
</table>

*CHD disease death + nonfatal MI + coronary revascularization.

Men in CARE experienced a relative risk of 0.761 on pravastatin therapy, while women who were randomly chosen to receive pravastatin therapy experienced a 0.545 relative risk. The p-value that assesses the difference in the effect for men and women was 0.05. Within CARE, the effect of therapy appeared to be modified by gender, women being the greater beneficiary of this effect than men.

However, care must be exercised in the interpretation of the analysis. This evaluation was but one of a number of secondary analyses executed in the study, and as such, was not subjected to a correction in the type I error rate for multiplicity. Thus its findings serve as merely supportive of the overall finding of cardioprotection from statin-based LDL reduction from CARE.

10.11.2 Within-Stratum Effects

The evaluation of a subgroup mediated effect modification may not directly address the question the investigators have raised about the subgroup. This is because the investigators’ interest may not be in the entire subgroup, but only in selected subgroup strata. Specifically, the investigators may not ask whether the effect of therapy is the same across subgroups, but instead ask whether there is an explicit effect of the intervention in the prospectively defined subgroup stratum of interest. This is a different question than that addressed by an interaction analysis.

One such situation would be when the stratum is composed of patients who have a very different prognosis from that of patients in other strata of the subgroup. While investigators may be most interested in the effect of a new intervention on thyroid cancer, they may be particularly interested in the effect of the therapy in patients with an advanced stage of the disease. This interest does not require
the investigators to ask whether the effect of therapy in patients with less advanced thyroid cancer is different from that of patients with advanced thyroid cancer; they wish to know only whether the therapy has been shown to have explicit efficacy in patients with advanced thyroid cancer.

Similarly, a new therapy for the treatment of CHF may hold promise for reducing mortality in all patients with CHF, but the investigator is motivated to demonstrate the effect of this therapy in patients with CHF whose etiology is non-ischemic. She is not interested in comparing or contrasting the efficacy of the intervention between ischemic versus non-ischemic etiologies of CHF. She is instead focused on two questions: (1) Is the therapy effective in the entire cohort and (2) Can the effect of this therapy be confirmed in the subcohort with CHF-non-ischemic etiology?

Is it possible that the therapy could be effective in the entire cohort but not the subcohort of interest? Yes. Consider the possibility that the therapy in fact is effective for patients with CHF-ischemic etiology but ineffective for patients with a non-ischemic etiology for their CHF. Let the research sample primarily contain patients with CHF-ischemic etiology with only a small number of patients who have a non-ischemic etiology for their heart failure. Since the research sample contains primarily those patients who will respond to the therapy, the result of the concordantly executed clinical trial will be positive (barring an effect that is driven by sampling error). The investigator will then argue that, since the trial is positive, this positive finding will apply to the CHF-non-ischemic subgroup as well. Essentially, the conclusion about the non-ischemic subcohort is based primarily on the findings of patients who are not in that subcohort at all.

This is the consequence of the effect domination principle, in which the findings in the overall cohort devolve on each of the subgroup strata. In this example, the principle produces the wrong conclusion; nevertheless, it is the best conclusion available in the absence of a confirmatory subgroup analysis. In order to avoid this possibility, the investigator is interested in reaching a confirmatory conclusion.

As another illustration of a circumstance in which prospectively specified, stratum-specific subgroup analyses can make an important contribution, consider the situation in which the adverse event profile of a therapy that is being studied in a controlled clinical trial is known to be different between women and men. As an illustration, consider a cholesterol-reducing drug that produces benign breast disease in women. In this circumstance, the risk–benefit profile of this drug is different for women than it is for men. Since women will be exposed to a greater risk with this therapy, it is reasonable to require investigators to produce a statistically valid demonstration of efficacy in women. The investigators are not disinterested in an effect in men; however, the relatively low risk of the drug in men allows the investigators to be satisfied with deducing the effect of the therapy in men from the effect of therapy in the overall cohort. It is the greater adverse event risk in women that requires an explicit demonstration of efficacy in them.

There are different questions that can be asked of subgroups. Some of these questions can be addressed by a heterogeneity of effect evaluation with its accompanying interaction analysis; however, there are others that are addressed by the direct demonstration of efficacy in a single subgroup stratum. Numerous exam-
ples and scenarios of the execution of subgroup stratum-specific analyses are available [27, 28].

10.12 Data Dredging — Caveat Emptor
Data dredging is the methodologic examination of a database for all significant relationships. These database evaluations are thorough, and the analysis procedures are wide ranging, spanning the gamut from simple $t$-testing to more complex time to event evaluations, repeated measures assessments, and structure equation modeling. Typically, little thought is given to endpoint triage as discussed in Chapter Nine.

The notion that if they look hard enough, work long enough, and dig deep enough they will dig up something “significant” in the database drives the data dredgers. Indeed, the investigators may identify a relationship that will ultimately be of great value to the medical community. However, while it is possible to discover a jewel in this “strip-mining” operation, for every rare jewel identified, there will be many false finds, fakes, and shams. As Miles pointed out, datasets that are tortured long enough will provide the answers that the investigators seek, whether the answers are helpful, truthful, or not [29].

Unfortunately, many of the important principles of good experimental methodology are missing in the direct examination of interesting subgroups. Inadequate sample size, poorly performing estimators,* low power, and the generation of multiple $p$-values combine to create an environment in which the findings of the data dredging operation are commonly not generalizable. Accepting the “significant” results of these data-dredging activates can misdirect researchers into expending critical research resources in fruitless pursuits, a phenomenon described by Johnson [30]. In his 1849 text Experimenta Agriculture, Johnson stated that a badly conceived experiment was not only wasted time and money, but led to both the adoption of incorrect results and the neglect of further research along more productive lines. It can therefore take tremendous effort for the medical and research community to sort out the wheat from the data-dredged chaff, often at great expense.

10.13 Conclusions
Subgroup analyses are most likely here to stay. As long as physicians focus on the treatment of individual patients, factoring in those patients’ unique and distinguishing characteristics when fashioning therapy, they will be interested in the results of subgroup analyses. This is an honest attempt to reduce the number of unknowns in the prediction of an individual patient’s response to treatment, and will continue to stoke the subgroup analysis fire.

This chapter demonstrates some of the contemporary difficulties that subgroup analyses create. Many subgroups evaluations can be misinterpreted because subgroup membership may merely be a surrogate for another, less obvious factor.

* The observation that the accuracy and precision of commonly used estimators is distorted in exploratory research is discussed in Chapter Two.
that determines efficacy. The investigator must consider this possible explanation for her subgroup specific effect in her interpretation of the analysis.

Yet another force that exerts traction on both investigators and regulators for subgroup interpretation is the lay press. Consider, as an example, an editorial appearing in the Wall Street Journal that purported to have identified an unnecessary obstruction put in place by the FDA’s drug approval process [31]. The editorial, complained openly, asking, “Why not allow companies to cull the relevant data from existing studies when a certain subgroup is clearly of help?”

Unfortunately, these calls for directed action that are based on the misdirected and misleading belief that subgroup analyses are not harmless. The genuine, heartfelt desire to come to the aid of ailing people must be tempered with a disciplined research strategy and execution. In the absence of this control, research efforts produce interventions that harm patients.

It is also more than likely that the majority of subgroup analyses will continue to be carried out as either prospectively declared secondary evaluations or as exploratory analyses. Thus the guidelines put forward by Yusuf will continue to be predominant for these evaluations—the best estimate of the effect of a therapy within a subgroup in such an analysis is the effect of the therapy that was seen in the overall cohort. These subgroup evaluations can suggest but not confirm the answers to questions about the risks and benefits of an exposure in clinical research. Fishing expeditions for significance commonly catch only the junk of sampling error.

References


levels in the average range Results of the Cholesterol and Recurrent Events (CARE) trial. *Annals of Internal Medicine* 129:681–689.


26. Peterson, B, George, SL (1993) Sample size requirements and length of study for testing interaction in a 1 x k factorial design when time-to-failure is the outcome *Controlled Clinical Trials* **14**:511–522.


