More than 20 years ago in this journal, Furberg and Byington \(^1\) showed how dicing a population from a randomized trial could unearth apparent subgroup findings. They used the Beta-Blocker in Heart Attack Trial, a double-blind randomized trial that showed conclusive evidence of benefit of \(\beta\)-blockers on mortality in patients who had experienced a myocardial infarction. In spite of this overall benefit, the authors showed that some subgroups exhibited no effect and some subgroups exhibited harm; they argued cogently that these “findings” were likely spurious. A few years later came the well-known example of spurious interaction from the International Study of Infarct Survival 2 (ISIS-2),\(^2\) a trial that showed a medically important and statistically significant benefit of aspirin in preventing fatal heart attacks. To demonstrate the pitfalls of subgroup analysis, the ISIS investigators classified all study participants by their astrological signs; subgroup analysis showed that whereas aspirin significantly benefited the population at large, it appeared actually to harm people born under Gemini or Libra. Most nonastrologers would believe the findings due to chance.

Most randomized clinical trials aim to show an effect of an experimental treatment in a given study population. Because most illnesses are variable, it seems quite natural to look to clinical trials for information about specific subgroups of patients in the hope that subgroup-specific estimates of treatment effects, rather than the overall effect, will predict more accurately the effect on specific patients. Unfortunately, few trials have sufficient statistical power to measure reliably the effect of a treatment within subgroups. Trials are even less likely to provide reliable information on interactions, that is, whether the effect of a treatment differs among subgroups.

The persistence of subgroup analyses, in spite of nearly all statisticians’ and many clinicians’ oft-stated warnings about their unreliability, must be due in part to the high probability that if one looks at enough subgroups, something interesting will emerge and the conviction that what emerges may in fact be true. Although looking at subgroups sounds fairly straightforward, the actual identification of subgroup-specific effects is treacherous. To some extent, the problem stems from our lack of knowledge about the diseases in question, so that defining clinically relevant subgroups may not be straightforward. To an even greater extent, however, the problems are statistical, relating to the need for large samples to insure reliable inference.\(^3\)–\(^7\) In an effort to deal with statistical traps, sometimes the articles present tests of interaction, a practice that should be used more often. Some analyses are accompanied by corrections for multiplicity, a task made daunting because the subgroups are not independent and available statistical methods are most useful for independent groups. If we are lucky, all the reported subgroups will show essentially the same thing, and the summary of results will include a statement that “the findings were consistent across subgroups of interest.”

Confronted with a surprising finding in a subgroup within a randomized controlled trial—by surprising I mean a finding that differs in important ways from the overall finding in the trial—one has 3 philosophical choices. The researcher who believes that data are unlikely to play tricks will accept, perhaps with some trepidation, the finding as indicating that this subgroup in fact differs from the others. The skeptic will almost automatically dismiss the finding as most likely due to the willingness of chance. And the agnostic, although perhaps intellectually siding with the skeptic, will have an uneasy feeling that the data may in fact be showing something real.

Those who tend to believe that a finding in a subgroup represents an underlying truth can argue from empiricism: What we see is what we get. Subgroup skeptics take a different tack. They (lest I be too coy, we) argue that the best estimate of the effect in a subgroup is not what the subgroup shows but either the estimate in the population at large\(^8\) or a weighted average of what the subgroup and the remainder of the population show, where the weights are a function of the variances within and between subgroups\(^9,10\) or a function not only of the data but also of one’s prior belief in subgroup differences.\(^9,10\) The following argument is a simple version of the reason the subgroup skeptic shies away from assuming that observations reflect reality. Imagine a trial that compares an experimental treatment to placebo. Further imagine that the treatment is totally inert so that the trial is really comparing 2 placebos. Split the study population into 2 mutually exclusive subgroups. The probability is 0.64 that at least one subgroup will show a \(P\) value <0.05. (For 50 subgroups, the probability is 0.92; for 100 subgroups, the probability is >0.99). Those probabilities are not in themselves surprising: the more shots at a target, the more likely even someone with poor aim is likely to hit it. Perhaps more surprising is that the smaller the subgroup, the larger the observed effect size must be to show statistical significance and because each subgroup, no matter what its size, is equally likely to show statistical significance when absolutely no effect of treatment exists, a false-positive effect in a small subgroup will appear larger than a false-positive effect in a

The opinions expressed in this article are not necessarily those of the editors or of the American Heart Association.

From Statistics Collaborative, Washington DC.

Correspondence to Janet Wittes, PhD, Statistics Collaborative, 1625 Massachusetts Ave NW, Suite 600, Washington, DC 20036. E-mail janet@statcollab.com

(Circulation. 2009;119:912-915.)

© 2009 American Heart Association, Inc.

Circulation is available at http://circ.ahajournals.org
DOI: 10.1161/CIRCULATIONAHA.108.836601
large subgroup. Consider, for example, a 10 000-patient trial comparing the effect of 2 placebos on the probability of experiencing an event E of interest. Suppose the probability of the event is 0.10 in all people. Now split the population into subgroups of sizes 1000 and 9000. Imagine that the data dealt us a type 1 error: a statistically significant result in 1 of the 2 subgroups. Statistical significance would require the 9000-patient subgroup to show a relative risk of about 0.88; however, for statistical significance, the relative risk in the smaller subgroup need be 0.67. Thus, if a statistically significant result is a false positive (and in any particular case we don’t know whether an observation reflects a true effect or a false-positive effect), on average its magnitude will be greater in small subgroups than in large ones, reflecting the greater variability in small subgroups. Thus, an unusually large treatment effect in a small subgroup raises the suspicion of play of chance. To not be “fooled by randomness,”11 skeptics assert their eyes, seeing the large effect as yet another version of the warning, “if it’s too good to be true, don’t buy.”

All of these caveats lead to the question of how the agnostic reader, or even more important, the practitioner, should react to surprising findings in a subgroup. Yusuf et al6 provide 5 rules of thumb for evaluating the likelihood that an effect seen in a subgroup is real.

1. Use statistical methods that capture the framework of the prior hypotheses.
2. Place greater emphasis on the overall result than on what may be apparent within a particular subgroup.
3. Distinguish between prior and data-derived hypotheses. Do not calculate P values for data-derived hypotheses because such P values usually bear little resemblance to what could occur if the hypothesis were tested independently in another study.
4. Use tests of “interactions,” and/or correct for multiplicity of statistical comparisons. (Nominal P values are usually misleading.)
5. Interpret the results in the context of similar data from other trials, from the architecture of the entire set of data on all patients, and from principles of biological coherence.

I started with 2 examples of trials where the overall effect showed benefit but subgroups did not. Conversely, a trial that shows no effect of treatment overall will likely yield a subgroup that appears to benefit. Such is the case in the Collins et al12 article appearing in the current issue of this journal. The article shows 51 subgroups in theRaloxifene Use for The Heart (RUTH) trial; 50 show no evidence that raloxifene prevents coronary events; however, the 51st, women aged 60 years, shows a 40% reduction in event rate. Consider this subgroup in light of the Yusuf rules of thumb. As the authors state, they did not specify the 60-or-younger age group a priori. Among the 3 prespecified age subgroups (≤65, 65 to <70, and ≥70), the interaction P value was 0.29, hardly suggesting a difference by age. None of the 3 age groups had a hazard ratio significantly different from 1. Under usual situations, the subgroup agnostics would stop here, thinking that the authors must have rummaged around for an age group that would show statistical significance. But the case is more subtle here: The authors changed their age breakdown “to match those in the Women’s Health Initiative randomized trials.”13 This choice makes sense, for it allows comparison across similar trials. What is troublesome is that the data do not appear similar (the smoothness of the curve of hazard ratio as a function of age reflects the fact that the model forces smoothness). The Collins article does not show neighboring subgroups, but it does provide some flavor for results in nearby subgroups. As seen in Table 1, the estimated effect size is considerably greater in the subgroup <60 (hazard ratio=0.59) than it is in the subgroup <65 (hazard ratio=0.84) and very much greater than in the subgroup of women <10 years postmenopause (hazard ratio=0.94). In order to believe that all these numbers mean what they say, one must create a complicated hypothesis relating to age <60, age 60 through 65, and time since menopause. The subgroup agnostic, applying Occam’s razor, would say that the low hazard ratio in the <60-year-old group more likely reflects chance than a true effect, or, if women aged <60 years really experience cardiovascular benefit from raloxifene, the benefit is very unlikely to be as large as that observed.

If reporting on subgroups is tempting but treacherous, failing to report on them seems unscientific and incurious. When surprising results from subgroups occur, they should be accompanied with other analyses. I suggest 3 in particular. Two are simple, but the third is more complicated. First, present graphs of ratios, including hazard ratios, on a semilog scale; ratios are by their nature multiplicative; displaying them on a linear scale distorts both the magnitude of the effect and the length of the confidence intervals. In Figure 1, reproduced from Figure 2 in Collins et al, the eye interprets 1.4 and 0.6 as equally distant from 1.0 (ie, 0.4 linear units), but in fact 1.4 and 0.71=1/1.4 are equally distant from 1.0 because ratios form a multiplicative scale. In this particular graph, the similarity of the widths of the 3 confidence intervals for age is deceptive. In RUTH, 1086 women

### Table 1. Effect of Raloxifene on the Coronary Outcomes in RUTH by Age Group and Year Since Menopause

<table>
<thead>
<tr>
<th>Age at Randomization (RUTH Prespecified)</th>
<th>Age at Randomization (Will Defined Groups)</th>
<th>Years Since Menopause</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgroup</td>
<td>Subgroup</td>
<td>Subgroup</td>
</tr>
<tr>
<td>≤65</td>
<td>&lt;60</td>
<td>&lt;10</td>
</tr>
<tr>
<td>&gt;65 and &lt;70</td>
<td>≥60 and ≥70</td>
<td>≥10 and ≥20</td>
</tr>
<tr>
<td>≥70</td>
<td>&gt;70</td>
<td>&gt;20</td>
</tr>
<tr>
<td>P for interaction</td>
<td>0.29</td>
<td>0.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimated Hazard Ratio</th>
<th>Estimated Hazard Ratio</th>
<th>Estimated Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤65 0.84</td>
<td>&lt;60 0.59</td>
<td>&lt;10 0.94</td>
</tr>
<tr>
<td>&gt;65 and &lt;70 1.09</td>
<td>≥60 and ≥70 1.06</td>
<td>≥10 and ≥20 1.02</td>
</tr>
<tr>
<td>≥70 0.97</td>
<td>&gt;70 0.98</td>
<td>&gt;20 1.02</td>
</tr>
</tbody>
</table>

Wittes On Looking at Subgroups 913
experienced a primary coronary event, but only 134 of them were younger than 60 years. The confidence intervals are of roughly equal width because the linear scale distorts widths of confidence intervals of ratios. See Figure 2 for a more accurate depiction. Adjusting for multiplicity would widen the confidence intervals still further.

Aside from adequacy of sample size and displaying uncertainty more accurately than is typically done, few fixes to the subgroup problem are available. As we have seen, subgroups that are too small will yield highly variable estimates. One way to prevent spurious findings is to limit exploration of subgroups to those with an a priori power of at least 40% or 50%. As shown in Table 2, for studies originally powered at 90%, this rule would preclude looking at subgroups with sizes <30% of the original study size; for a study with 80% power overall, only subgroups with sizes of at least 40% would be fair game for exploration. This simple caution would prevent many, but not all, spurious findings.

Finally, if one does identify a subgroup of interest, the nearby subgroups matter. If Collins et al had shown the 8 age subgroups with 130 events each and had the data shown a reasonably smooth relationship between age and effect size, the surprising result in the under-60 age group would be more credible. In the absence of knowing what happened in the equally powered groups, the results appear at best overstatements and at worst the consequence of randomness. For the case at hand, raloxifene, the choice of treatment for those under 60 years of age is unlikely to be affected by one’s interpretation of the 40% reduction in coronary events for this subgroup. As the authors point out, the decision to treat will more likely be influenced by a woman’s risk of breast cancer and fracture, both of which raloxifene reduces. For other drugs, where subgroup analyses are more likely to affect choice of treatment, this subgroup skeptic pleads for more caution.

Disclosures
Dr Wittes has served as chair of the data safety monitoring committees (DSMB) of both the RUTH and the Women’s Health Initiative (WHI) trials. As a member of the DSMB, she received honoraria from Eli Lilly.

References

Key Words: Editorials cardiovascular diseases prevention statistics
On Looking at Subgroups
Janet Wittes

Circulation. 2009;119:912-915
doi: 10.1161/CIRCULATIONAHA.108.836601
Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2009 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the
World Wide Web at:
http://circ.ahajournals.org/content/119/7/912

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published
in Circulation can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial
Office. Once the online version of the published article for which permission is being requested is located,
click Request Permissions in the middle column of the Web page under Services. Further information about
this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation is online at:
http://circ.ahajournals.org//subscriptions/