CPH 931 — Fall 2008 — Dr. Charnigo

Lecture 11

Lecture 11A: Subgroup Analyses

Preface. Lecture 11A is based on pages 553-565 in the *Users' Guides to the Medical Literature* (2005) edited by Gordon Guyatt and Drummond Rennie. I also refer to a 1983 *Circulation* paper by Furberg and Byington titled "What Do Subgroup Analyses Reveal About Differential Response to Beta-Blocker Therapy?"

Introduction. The Beta-blocker Heart Attack Trial (BHAT), discussed in the paper by Furberg and Byington, investigated whether propranolol would lower mortality in patients with acute myocardial infarctions. Besides attempting to answer the question for the overall patient population to which the study was generalizable, BHAT entailed analyses of 146 subgroups. The idea was to determine whether some subgroups of patients might benefit more from propranolol therapy than others. After all, as the authors of the *Users' Guides* state, "Clinicians faced with a treatment decision in a particular patient are interested in the evidence that pertains most directly to that individual" (page 554).

For instance, noting the antiarrhythmic action of propranolol, Furberg and Byington argue for biological plausibility of the notion that propranolol therapy might be especially beneficial to patients in the subgroup with ventricular tachycardia after hospital admission. On the other hand, Furberg and Byington argue for biological plausibility of the notion that propranolol therapy might not be so beneficial to patients in the subgroup defined by moderate to heavy leisure physical activity before infarct.

Yet, Furberg and Byington, along with the authors of the *Users' Guides*, caution readers that not all subgroup analyses should be taken at face value. The goals of Lecture 11A are to explain the concept of interaction implicit in subgroup analyses, why the caution called for by Furberg and Byington is warranted, and how readers can assess the credibility of subgroup analyses.

The concept of interaction. Implicit in the performance of subgroup analyses is a belief that there may be some "interactions" between treatment and risk factors such as age, gender, and comorbidities. What are interactions? Simply put, they are variations in treatment effects that correspond to variations in risk factors.

For instance, if the relative risk of mortality (risk on active treatment divided by risk on placebo) is 0.60 for patients aged less than 65 and 0.80 for patients aged greater than 65, then there is an interaction between treatment and age because the treatment effects are stronger for younger patients. On the other hand, if the relative risk is 0.70 for males and 0.70 for females, then there is no interaction between treatment and gender.

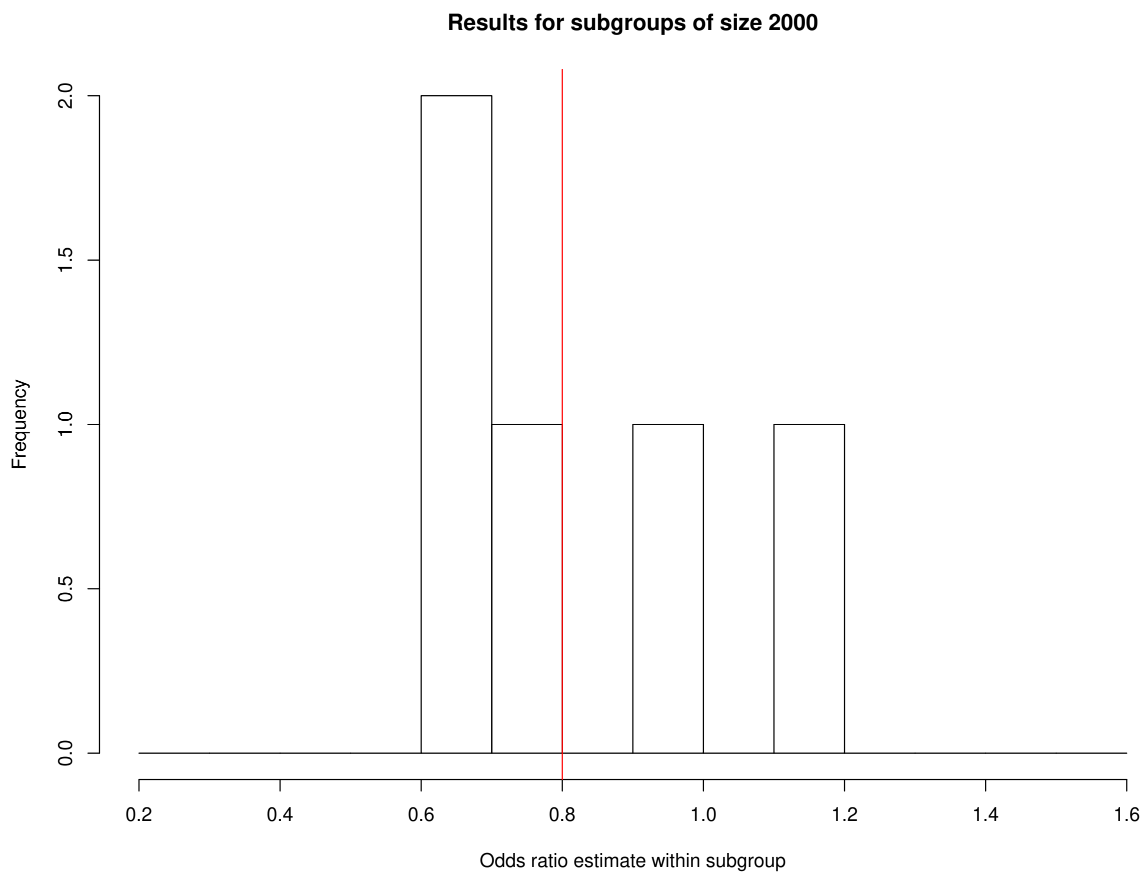
Note that a risk factor can be important and yet not interact with treatment. For example, suppose that mortality risk is 15.0% for diabetics on placebo, 10.5% for diabetics on active treatment, 10.0% for non-diabetics on placebo, and 7.0% for non-diabetics on active treatment. Clearly, diabetes is an important risk factor in this example. Even so, the relative risks for non-diabetics (7.0%/10.0% = 0.70) and diabetics (10.5%/15.0% = 0.70) are identical, meaning that there is no interaction between treatment and diabetes.

Why caution is warranted in evaluating subgroup analyses. Although subgroup analyses have the potential to identify patient subgroups in which a treatment is more (or less) effective than in the overall patient population to which the study is generalizable, we must be concerned with the great multiplicity of hypothesis tests performed in subgroup analyses. In fact, we are confronted by this issue with two different kinds of hypothesis tests in subgroup analyses.

First, a null hypothesis of no treatment effects is usually tested within each subgroup. This would look like "The odds ratio for males is 1.00" or "The odds ratio for females is 1.00". If we declare statistical significance every time we encounter a p-value less than 0.05, then even if there truly are no treatment effects whatsoever we can expect to find statistically significant results for 5% of the subgroups. Thus, in BHAT we would expect to flag $146 \times 0.05 \approx 7$ subgroups even if propanolol conferred no benefit (or harm) whatsoever. In practice, the small p-values that we encounter are usually a mix: some of them are "real" (i.e., represent genuine, important treatment effects), some of them are not real, and we do not know which are which. Moreover, because the sample sizes in the subgroups are typically much smaller than the overall sample size, we may have very low power to detect genuine treatment effects in the subgroups.

Second, a null hypothesis of no interaction can be tested for each risk factor that defines subgroups. This would look like "The odds ratio for males is the same as the odds ratio for females". If we declare statistical significance every time we encounter a p-value less than 0.05, then even if there truly are no interactions whatsoever we can expect to find statistically significant results for 5% of the risk factors. As before, the difficulty is that we do not know which of the small p-values we encounter are real (i.e., represent genuine interactions) and which are not.

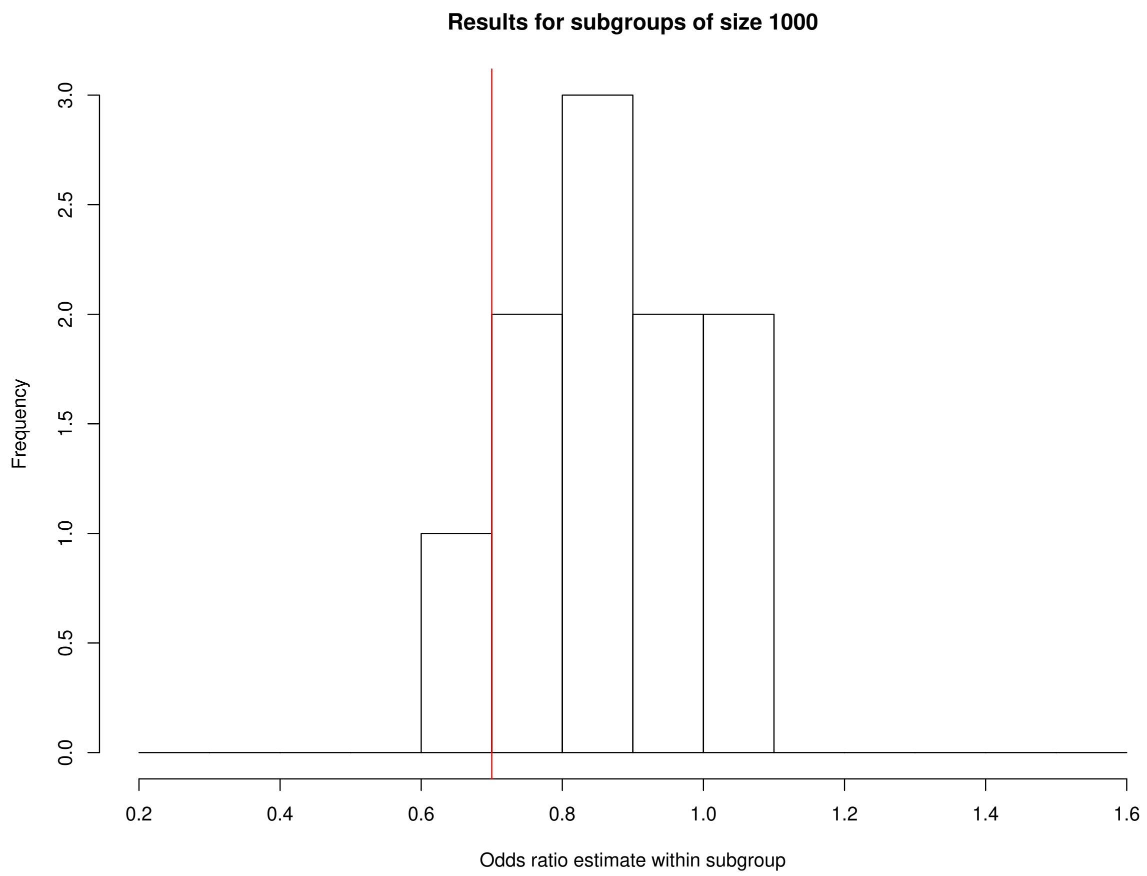
Figure 1:



An illustrative simulation. I conducted a computer simulation to assess how much estimated odds ratios (EOR) could vary across subgroups in a clinical trial without there being any genuine interactions. I created 40 subgroups, ranging in size from 400 to 2000. The true placebo event rate was fixed at 10.0% for all subgroups, and the true treatment event rate was fixed at 8.0% for all subgroups.

The results are summarized in Figures 1 through 4. Figure 1 shows the EORs obtained for the five subgroups of size 2000. Those EORs to the left of the red vertical line would have been identified as statistically different from 1 at a significance level of 0.05. Figure 2 shows the EORs obtained for the 10 subgroups of size 1000, Figure 3 shows the EORs obtained for the 25 subgroups of size 400, and Figure 4 shows all 40 EORs.

Figure 2:



Several observations can be made. One, there is more variation in the EORs among smaller subgroups than among larger subgroups. Two, for smaller subgroups only EORs considerably less than the true odds ratio are declared statistically significant, affirming that there may be very little power in subgroup analyses. Three, the EORs of 0.30 and 1.52 may appear strikingly different, but the difference is entirely due to chance; the true odds ratio common to both subgroups is 0.78.

Figure 3:

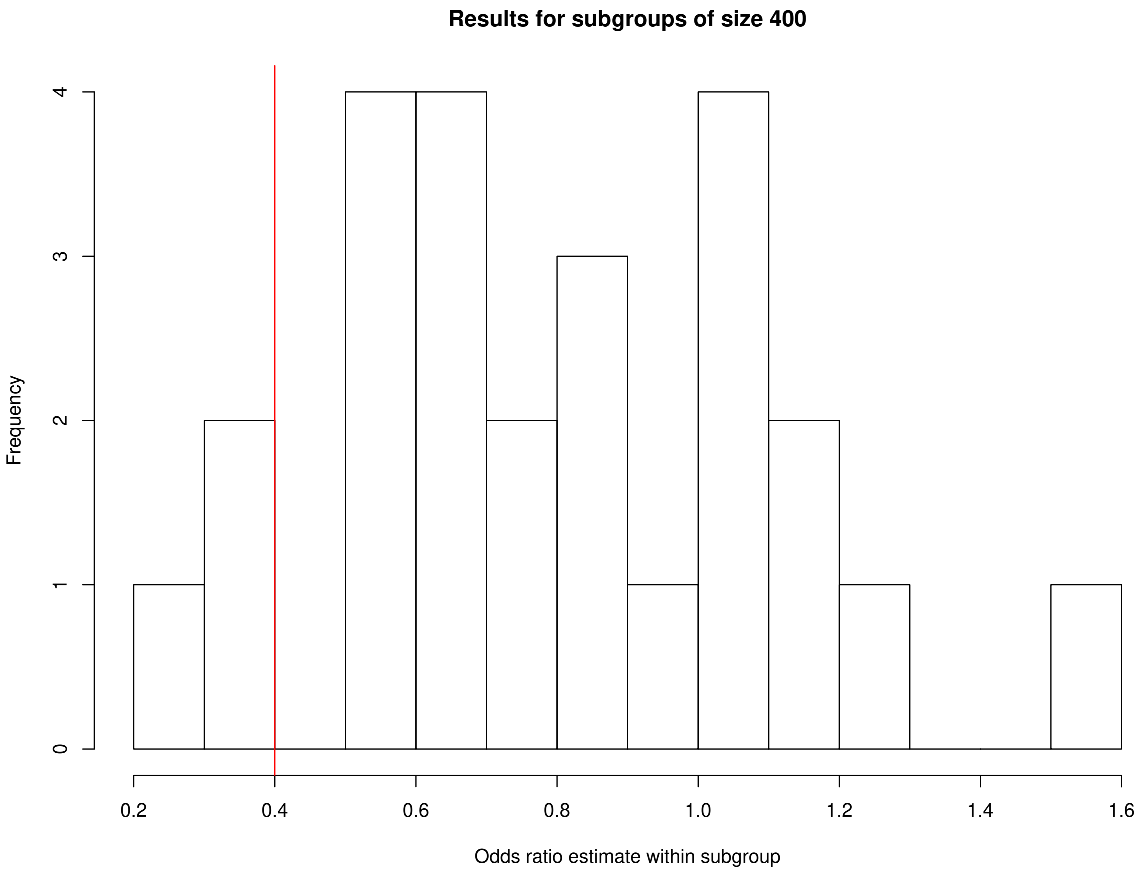
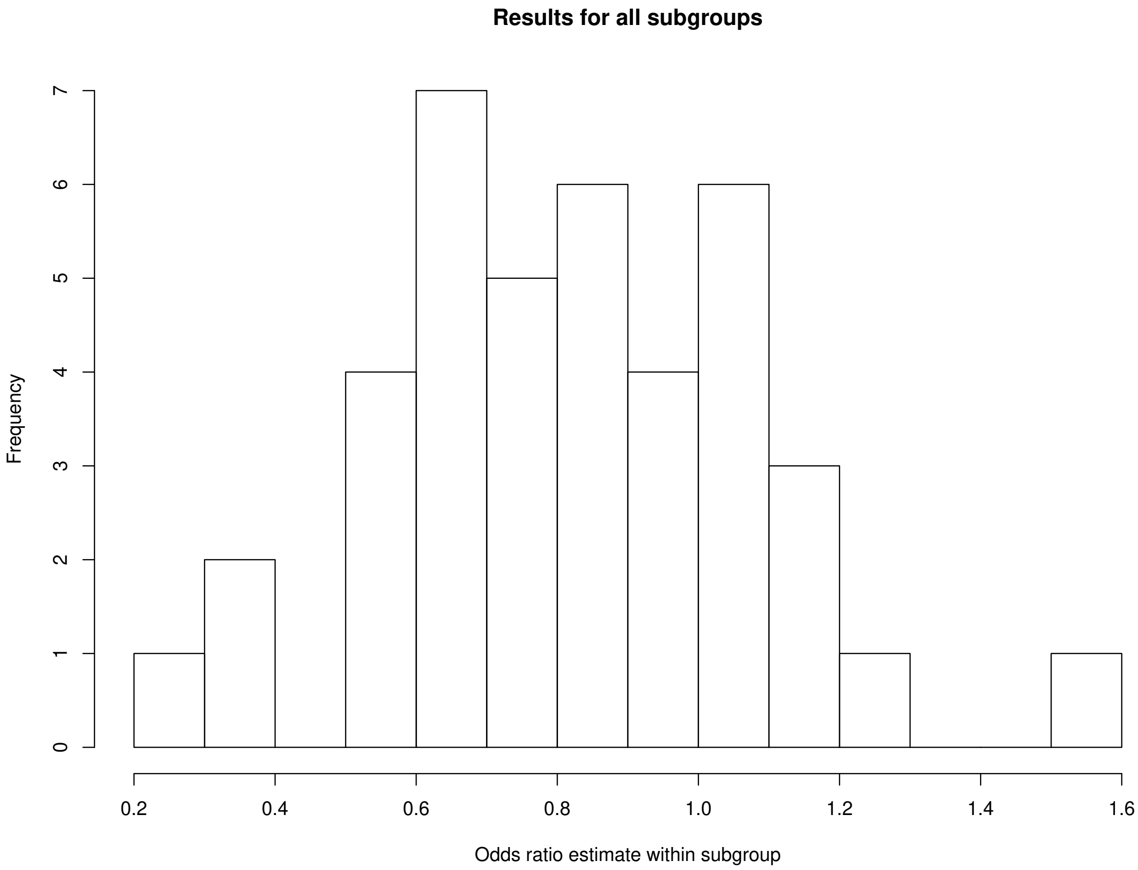


Figure 4:



Assessing the credibility of subgroup analyses. The illustrative simulation and the discussion preceding it have revealed the potential problems with taking subgroup analyses at face value. But exactly how does one exercise "caution" in evaluating subgroup analyses?

To answer this question, the authors of the *Users' Guides* provide several guidelines for deciding how much trust to place in subgroup analyses.

Within or between studies. Suppose that one trial, in which all of the participants are males, yields an EOR of 0.80 while a separate trial, in which all of the participants are females, yields an EOR of 1.20. Even if the authors could produce a small p-value supporting the claim of a real difference underlying the 0.80 and the 1.20, readers should be skeptical because there could have been "hidden" interactions, unnoticed or ignored by the authors, due to other differences in the study populations. For instance, one trial might have enrolled patients that were generally much younger and healthier. Thus, more trust can be placed in a claim of interaction if it involves a within-study comparison than if it involves a between-study comparison.

Planned in advance. More trust can be placed in subgroup analyses that have been planned in advance than in subgroup analyses that have not. The reason is that, in the former case, we are assured that the authors have not merely sifted through their data, creating subgroups after seeing how to define them in a way to obtain statistically significant results.

To drive home the point, consider the absurd extreme where one subgroup is defined to consist of all placebo patients who experienced the event and all treatment patients who did not, while the other subgroup is defined to consist of all treatment patients who experienced the event and all placebo patients who did not. Clearly, nothing useful for clinical practice is revealed. While researchers are never quite this flagrant, skepticism is warranted when the authors define subgroups after the data are in hand.

Number of subgroups. A p-value less than 0.05 becomes less impressive as the number of subgroup analyses increases, because we know that the authors can get one or more p-values less than 0.05 just by chance if they do enough subgroup analyses. We are less inclined toward skepticism when, for example, there are six subgroups than when there are 146.

Magnitude of apparent difference. If the EOR for males is 0.80 and the EOR for females is 1.20, then we are far more concerned with the apparent difference than if the EOR for males is 0.80 and the EOR for females is 0.85. This is because the former apparent difference is large enough that, if it were real, clinical practice should differ by gender. In contrast, the latter apparent difference does not seem large enough to motivate differential clinical practice.

Statistical significance. An important conceptual mistake, noted by the authors of the *Users' Guides* (page 560), is that some people will declare the treatment effects to differ between Subgroup 1 and Subgroup 2 when the estimated relative risk or EOR is statistically significant in Subgroup 2 but not in Subgroup 1. In essence, the mistake is trying to combine the results from two hypothesis tests of the first kind rather than explicitly performing a hypothesis test of the second kind. This would entail saying, for example, "The EOR for males is 0.80 with confidence interval 0.65 to 0.98, the EOR for females is 0.85 with confidence interval 0.70 to 1.04, and so I conclude that the treatment works differently by gender because only the male confidence interval excludes 1.00" when one should say "I need to determine whether the 0.80 and 0.85 differ significantly from each other".

Consistent across studies. We are more convinced that an interaction is real if it is suggested by several studies than if it is suggested by only one study. Indeed, Furberg and Byington state that the "strongest support for a subgroup finding in one trial is a replication from another trial".

Note that the authors of the *Users' Guides* are not advocating between-study comparisons in the sense described earlier. Rather, they are looking for similarities in the within-study comparisons from multiple trials.

Indirect evidence. We are more convinced that an interaction is real if there is indirect evidence in its favor. The authors of the *Users' Guides* identify three kinds of indirect evidence: studies involving different populations, studies involving different but related outcomes, and studies involving different but related treatments.

Lecture 11B: Meta-analysis, fixed and random effects models, publication bias

Preface. Lecture 11B is based on pages 529-552 in the *Users' Guides to the Medical Literature* (2005) edited by Gordon Guyatt and Drummond Rennie. I also refer to a 2008 JACC paper by Fauchier et al titled "Antiarrhythmic Effect of Statin Therapy and Atrial Fibrillation".

Overview of meta-analysis. The purpose of meta-analysis is to rigorously combine the results from several studies to yield an overall answer to some scientific question that has not been resolved definitively by any of the studies individually.

As an example, Fauchier et al (2008) state that "statins had been suggested to protect against atrial fibrillation in some clinical and experimental studies but [their use] remained inadequately explored". For this reason, Fauchier et al (2008) performed a meta-analysis of six randomized controlled trials involving statins in which data on atrial fibrillation outcomes were collected. The results from these six trials are summarized below.

Table 1: Results from six trials

Study	Incidence in statin arm	Incidence in control arm	Est OR	95% CI
MIRACL	93/1539	96/1548	0.97	[0.72,1.31]
Tveit	18/51	17/51	1.09	[0.48,2.47]
Dernellis	14/40	36/40	0.06	[0.02,0.20]
ARMYDA 3	35/101	56/99	0.41	[0.23,0.72]
Chello	2/20	5/20	0.33	[0.06,1.97]
Ozaydin	3/24	11/24	0.17	[0.04,0.72]

The odds ratio estimates of 0.97, 1.09, 0.06, 0.41, 0.33, and 0.17 are not in qualitative agreement. Three of the 95% confidence intervals include the neutral value of 1, but one of those three confidence intervals is so wide that it also includes the decidedly nonneutral value of 0.10. The other three confidence intervals lie entirely to the left of 1. The 95% confidence interval from MIRACL has no overlap with three of the other confidence intervals. Likewise, the 95% confidence interval from Dernellis has no overlap with three of the other confidence intervals.

By performing a meta-analysis, Fauchier et al (2008) distilled the results from the six trials into an overall odds ratio estimate of 0.39 with an accompanying overall 95% confidence interval of 0.18 to 0.85. Thus, their overall answer to the scientific question was that statins decreased "incidence or recurrence of atrial fibrillation in patients in sinus rhythm with a history of previous atrial fibrillation or undergoing cardiac surgery or after acute coronary syndrome".

Fixed effects and random effects models. One issue discussed extensively in the *Users' Guides* is whether to adopt a fixed effects or random effects statistical model to describe the estimated odds ratios from different studies. A fixed effects model treats within-study variation as random but between-study variation as nonrandom, while a random effects model treats both kinds of variation as random.

A fixed effects model is conceptually appealing if the studies whose results we are combining constitute the totality of all relevant studies, while a random effects model is conceptually appealing if the studies whose results we are combining constitute only a sample of all relevant studies that were conducted or that hypothetically could have been conducted. The authors of the *Users' Guides* clearly lean toward the latter perspective (and I agree with them), as they are "interested in extrapolating results beyond the [studies in the meta-analysis] to patients in [their] own practice" (page 544). Indeed, if we are extrapolating to patients in our own practice, then the studies in the meta-analysis do not constitute the totality of circumstances in which the question of scientific interest needs to be addressed.

Practical considerations. Despite leaning toward the random effects model conceptually, the authors of the *Users' Guides* express practical reservations about relying on it exclusively. They provide an example (page 543, Figure 2E-4) in which a random effects meta-analysis leads to what they regard as a counterintuitively and unreasonably wide overall 95% confidence interval. On the other hand, they provide another example (page 542, Figure 2E-3) in which a fixed effects meta-analysis leads to what they regard as an unreasonably narrow overall 95% confidence interval. Ultimately, the authors of the *Users' Guides* favor being open to either statistical model if the other produces an unreasonable overall 95% confidence interval.

In fact, what the authors of the *Users' Guides* observe in their examples on pages 542 and 543 reflects a general phenomenon: the overall 95% confidence interval obtained from a random effects meta-analysis tends to be wider than the overall 95% confidence interval obtained from a fixed effects meta-analysis involving the same studies, especially when between-study variation is large.

Another point raised by the authors of the *Users' Guides* is that a random effects meta-analysis tends to weight the individual study results more equally than a fixed effects meta-analysis, especially when between-study variation is large.

Illustration. That random effects meta-analysis tends to weight the individual study results nearly equally is apparent in Fauchier et al (2008). While the largest study (MIRACL, 3087 patients) receives the greatest weight (22.48%) and the smallest study (Chello, 40 patients) receives the least weight (10.57%) in their random effects meta-analysis, these weights are closer to a scheme of equal weighting for all studies than to a scheme of weighting in proportion to the sample size (MIRACL contributed 86.79% of the patients, while Chello contributed only 1.12% of the patients).

Heterogeneity between studies. If the results from individual studies are too different, one can make the argument that meta-analysis is inappropriate: taking a weighted average of strongly conflicting results may not constitute a reasonable extrapolation to patients in our own practice.

The authors of the *Users' Guides* provide an example on page 549 (Figure 2E-5) in which two studies favoring treatment have 95% confidence intervals that do not even come close to overlapping with the 95% confidence intervals for two other studies not favoring treatment. In this instance, the authors of the *Users' Guides* admit that they are uncomfortable with meta-analysis.

However, the authors of the *Users' Guides* stop short of saying that meta-analysis should not be performed in such situations. Rather, they encourage investigators to examine possible explanations for the heterogeneity such as "differences in study participants, interventions, outcomes, and study methodology" (page 552) and to "maintain extra caution in recommending treatments on the basis of pooled estimates associated with unexplained heterogeneity" (page 552).

Illustration. The meta-analysis by Fauchier et al (2008) exhibits considerable heterogeneity between trials. A formal statistical test for heterogeneity yielded a chi-square statistic of 29.47 and an accompanying p-value less than 0.0001. The I^2 inconsistency statistic was 83.0%. Briefly, I^2 is a somewhat ad-hoc though useful measure of heterogeneity defined to range from 0 to 1. Larger values of I^2 indicate greater heterogeneity. The main rationale for reporting I^2 is that, for example, we can always regard an I^2 greater than 75.0% as being quite large; whether a chi-square statistic of 25 is large depends on the number of trials in the meta-analysis.

To their credit, Fauchier et al (2008) documented a number of differences between the six trials: one trial used pravastatin while five used atorvastatin, two trials used standard therapy in the control arm while four used placebo, different doses were used, study populations varied from patients with acute coronary syndrome to patients with scheduled coronary bypass surgery, and patients were monitored for different lengths of time.

Publication bias. The authors of the *Users' Guides* define publication bias as the "selective publication of manuscripts based on the magnitude, direction, and statistical significance of the study results" (page 530). In particular, studies that do not find statistically significant differences favoring treatment — especially small studies — are less likely to be written up and submitted for publication, less likely to be accepted for publication when they are submitted, and less likely to appear in MEDLINE-indexed journals when they are published. Thus, a researcher performing a meta-analysis may not be aware of all relevant studies, and the studies of which the researcher is aware may disproportionately favor treatment.

A researcher performing a meta-analysis can try to reduce the impact of publication bias by asking experts to help identify references, reviewing conference proceedings, considering that some relevant studies may be published in languages other than English, and searching a wide variety of databases. The authors of the *Users' Guides* also express hope that, one day, prospective study registration will further reduce the impact of publication bias.

In any event, once a meta-analysis has been performed, a researcher can employ various visual and quantitative tools to assess the impact of publication bias. A commonly used visual tool is the so-called funnel plot, illustrated on pages 534 and 535 (Figure 2E-2A, Figure 2E-2B). The standard error associated with the estimated odds ratio for each study is plotted against the estimated odds ratio. If the impact of publication bias is negligible, the funnel plot is anticipated to be nearly symmetric. If the impact of publication bias is considerable, the funnel plot is anticipated to be asymmetric, the lack of symmetry arising from the absence of small studies not finding statistically significant differences favoring treatment. However, whether a given funnel plot warrants concern may be difficult to assess when the meta-analysis involves only a few studies.

Illustration. The meta-analysis by Fauchier et al (2008) appears vulnerable to publication bias because the authors considered only studies from peer-reviewed journals indexed in MEDLINE and disregarded studies published in languages other than English. The funnel plot in Figure 2 appears quite asymmetric to me, about as asymmetric as it can get with only five points (where is the sixth?), despite the authors' claim that it is relatively symmetric.

Final remarks about meta-analysis. While meta-analysis is performed to yield an overall answer to some scientific question, there is no guarantee that the answer will be definitive. For instance, even in Fauchier et al (2008) we see that the overall 95% confidence interval for primary prevention of atrial fibrillation is 0.27 to 1.37. While the inclusion of 1 in this confidence interval does not permit the authors to assert that statins are useful for primary prevention, the inclusion of 0.4 does not permit them (or anyone else) to rule out that possibility.

As Fauchier et al (2008) acknowledge explicitly for their own meta-analysis, a meta-analysis generally does not use patient-level data from the individual studies. Rather, only summary data (numbers of events and patients in each arm) are used. If patient-level data are available for all studies of interest, and if the patient-level data are compatible across different studies, then a researcher can go beyond meta-analysis to a statistical technique (generalized linear mixed modeling or generalized estimating equations) that allows for the analysis of patient-level data across multiple studies when the observations within a study are assumed to be correlated.

Discussion Questions

1. In Lecture 11A how would you go about testing a null hypothesis like "The odds ratio for males is different from the odds ratio for females"?

2. From Section 13.8 of Rosner, the formula for a confidence interval from a fixed effects model is

$$\exp\left[\frac{\sum_{i=1}^{k} w_{i,fixed}\ y_i}{\sum_{i=1}^{k} w_{i,fixed}} \pm 1.96\sqrt{\frac{1}{\sum_{i=1}^{k} w_{i,fixed}}}\right]$$

and the formula for a confidence interval from a random effects model is

$$\exp\left[\frac{\sum_{i=1}^{k} w_{i,random}\ y_i}{\sum_{i=1}^{k} w_{i,random}} \pm 1.96\sqrt{\frac{1}{\sum_{i=1}^{k} w_{i,random}}}\right],$$

where y_i is the logarithm of the estimated odds ratio from study i, $w_{i,fixed}$ is the weight for study i in a fixed effects model, and $w_{i,random}$ is the weight for study i in a random effects model. Moreover, $w_{i,fixed}$ and $w_{i,random}$ are related by the formula

$$\frac{1}{w_{i,random}} = \frac{1}{w_{i,fixed}} + \hat{\Delta}^2,$$

where $\hat{\Delta}^2$ is a nonnegative quantity. Using these facts, can you substantiate my claim in Lecture 11B that confidence intervals from random effects models tend to be wider than those from fixed effects models?