

Quantification of Uncertainty in the Mammography Controversy

or

The Dangers of Dichotomy

Carl V. Phillips¹
Constance Wang²
Cindy Leong-Wu¹

¹University of Texas SPH, Management and Policy Sciences

²University of Texas SPH, Epidemiology

Background

Last year Olsen and Gotzsche (O&G) published their Cochrane review on screening for breast cancer (BC) with mammography (Cochrane review on screening for breast cancer with mammography, Lancet 2001).

Their conclusion that there was no reliable evidence that screening for breast cancer reduces mortality set off a firestorm (interestingly, they were merely confirming their earlier findings which did not generate much attention).

The popular press picked up the message, creating much alarm and anxiety. Women and their doctors no longer had a clear recommendation to follow for breast cancer screening.

The dilemma seems to have been "solved" by everyone but academics quietly agreeing to forget the whole embarrassing incident. Several organizations declared that their recommendations for screening have not changed -- a reasonable conclusion, but seemingly based on inertia rather than direct responses to the serious problems the O&G controversy demonstrated.

The Public, Confused by Dichotomy

Public perception of the value of mammography changed,

from virtually an obligation like seatbelts or vegetables

to apparently no value, almost overnight.

It is not surprising popular opinion changed like this, given that the people's experts (physicians, talking heads, etc.) opinions seem to have changed in the same manner.

Quite reasonably, people asked how such a sea change could have occurred.

Aside: one immediate result was an lot of useless advice

Common advice to women wondering whether to continue screening was "keep doing what you were doing until we know more." This is logically absurd. Perhaps the women who previously ignored the constant advice to screen are now doing the sensible thing. Perhaps screening is still a good idea. But whichever is the case, it cannot be contingent on what someone used to do.

Even worse was "we cannot be sure, so you will have to decide for yourself," which is similar to civil engineers saying "*we are not sure if this bridge is structurally sound, so you will have to decide yourself whether to drive over it.*" The effectiveness of screening is not a matter for personal preference or individual belief; it is a matter experts must decide.

Dichotomy Damages Health Research

The confusion in the reactions of clinicians, policy makers, and researchers seems largely to result from the false dichotomies that characterize how research results are often presented:

A study is **good** (and thus objective scientific truth) or **bad** (and can be ignored).

A result is **significant** (considering random sampling and allocation errors alone) or **not significant** (and so tells us nothing, or worse, is taken as supporting a null effect even when it is too imprecise to be conclusive)

A causal association is either **proven** (because a good study had a significant result) or we are **completely ignorant** (and so should do nothing, awaiting proof).

Ok, perhaps we exaggerate a bit, but it is not far from the truth...

Dichotomy versus Mammography

In June 2000, the benefit of screening mammography was taken as given. Most research on the topic focused on how to get more women to do it.

In January 2002, many neutral observers changed their beliefs in response to O&G's analysis and adopted the view that there was no evidence that mammography was beneficial, with some even believing that there was proof of no benefit.

By June 2002, new evidence seems like it should make us believers again (e.g., Nystrom et al. Long-term effects of mammography screening: updated overview of the Swedish randomized trials. Lancet 2002.).

Or perhaps we are somewhere between certain and ignorant.

But that is really where we always were.

Information, pre-O&G

For the 9 randomized trials O&G identified as the basis for our opinion, point estimates for the risk ratio from mammography on BC mortality ranged from ~1/3 reduction in BC mortality to zero reduction.

Information from non-RCT sources:

Observational studies tend to suggest a benefit.

The simple logic of the situation tells us that screening must reduce BC mortality some (though not necessarily enough to be worth doing).

But, the large reduction in BC mortality that should have accompanied the increase in screening mammography has not appeared (improvements in treatment seem to account for the entire reduction, or even seem like they should have caused more reduction than actually observed).

Information from Olsen and Gotzsche

All of the trials left something to be desired (in terms of study methods, at least as reported).

Three (or two, depending on how you divide them) trials were the only ones that met O&G's inclusion criterion, which was based on a dichotomous, good-study vs. bad-study scale.

The included studies showed no mortality benefit from screening, resulting in the conclusion of "no reliable supporting evidence."

The excluded trials, with one exception, were not dismissed because of identifiable biases that favored a beneficial effect of screening, but on what might be called "technicalities," particularly unexplained irregularities in the randomization process.

No doubt the noted irregularities should make us less certain of the protective effect of screening...

But what did we really know (and when did we know it)?

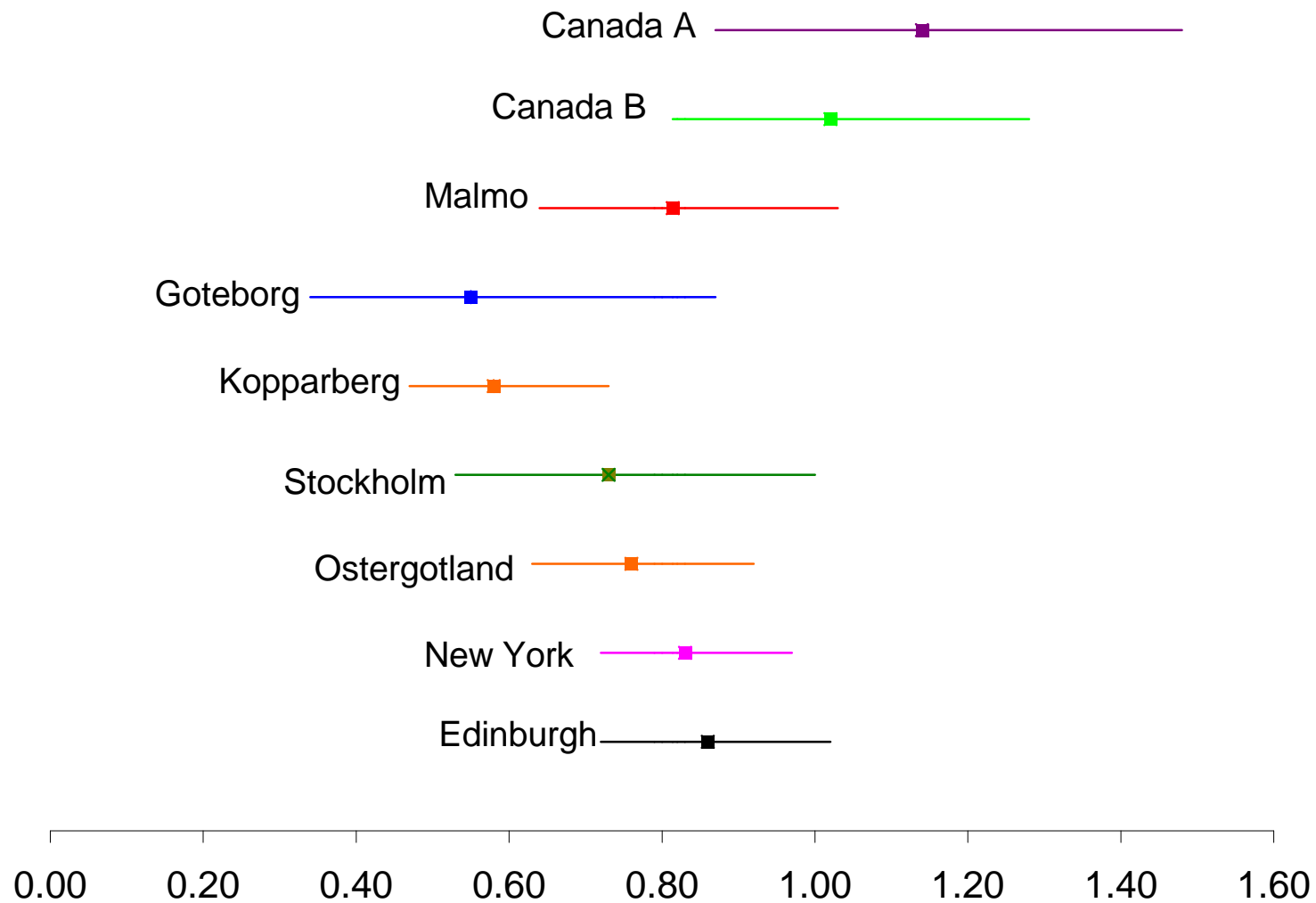
What do the nine trials tell us if we just accept their results?

(i.e., what did we know before O&G?)

To begin to answer this question, we will start with the premise that underlies the current confusion: the trial data tell us everything we know about this association, trumping all other sources of information. (We would hope that few SER attendees would believe this, but it is a convenient starting point.)

The point estimates and confidence intervals for the trials cover a wide range.

Point Estimates and 90% CIs or risk ratios for "the" nine trials



****WHISKER PLOT****

risk ratio for breast cancer mortality, screening mammography vs. not

(Data from: O&G. Systematic review of screening for breast cancer with mammography. <http://image.thelancet.com/lancet/extra/fullreport.pdf>)

The distribution of results should be recognized, rather than summarized away

This distribution of results would pass typical tests of homogeneity, but the range of results still suggests a fair amount of uncertainty. Based on these results, it does not seem justifiable to claim that we are certain of a substantial protective effect.

Synthetic meta-analysis (as O&G set out to do and others have done for some measures of mammography's effectiveness) implicitly assumes that the point estimates differ only due to random errors, and that the variation does not result from systematic errors, effect modification, or simply measuring different things.

In this way, synthetic meta-analysis obliterates the uncertainty about the different methods, measures, and errors reflected in the distribution of results, and moves us to a false dichotomy.

A proposed method for realistically assessing the uncertainty

Instead of treating the trials as flawless measures of the same causal effect, differing only due to random allocation error, consider the possibility that each trial might correctly measure the causal effect (but for random error), while the others deviate because of systematic errors or study design flaws, or because they do not actually measure what we want to know.

As a rough estimate of plausible values for each study, consider a triangular distribution, with a mode at the point estimate and a range covering the 90% c.i.

A rough, quick-cut measure of uncertainty can thus be generated by combining these distributions for the nine studies:

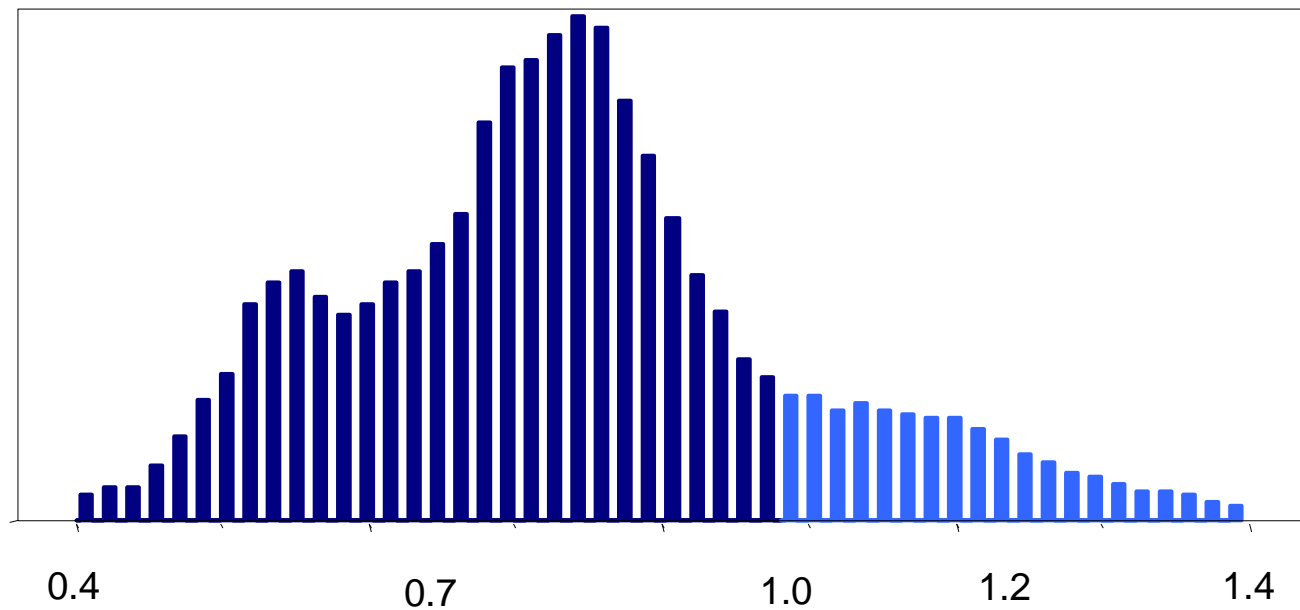
One study is picked at random and a point estimate is drawn from the triangular distribution that describes its plausible results.

This random generation is iterated (50,000 times for the results presented) and the point estimates are collected into a histogram that approximates a probability density function.

Result 1

Figure 1 shows the overall distribution of plausible risk ratios using the proposed specification. The distribution suggests that our level of certainty that screening mammography reduces BC deaths should be high, but definitely short of 100%. (For this model specification, it is about 5/6 likely that the risk ratio is less than 1.0.)

Plausible values of the risk ratio
for the effect of screening mammography on BC mortality
(not corrected for possible study biases)



What did the trials tell us after considering O&G's concerns?

The above proposed method addresses the fact that studies should not be assumed to result in valid estimates of effect, and thus should not be aggregated in a way that obscures uncertainty.

In order to move away from the other side of the dichotomy -- to recognize that a study still contributes knowledge even when flaws are identified -- we depart from O&G's strategy of eliminating studies in favor of incorporating plausible levels of systematic error into the above quantification.

Specifically, we adjust each study result according to O&G's assessment of its validity as follows:

Malmo, Canada A & B: O&G cite some problems with the stated sample sizes, some uneven randomization, and loss of subjects. They conclude that these are not sufficiently troublesome to exclude these studies. Following their relatively positive assessment (and our inability to identify clear sources of potential bias) we use the same distribution as before.

Goteborg: O&G note a possible publication bias because results have only been published for women aged 39-49, though data apparently existed for those 50-59. To incorporate this possibility, we simulate the missing population such that the risk ratio for them ranges uniformly from 0.75 to

1.5 (skewed to reflect O&G's speculation about the direction of bias). We then pool the age groups and sample from the resulting 90% CI as before.

Stockholm, Ostergotland, Kopparberg: O&G determined that there was plausible evidence of inadequate randomization in each of these based on poorly described methodology, the numbers in study arms, and differing values for covariates. They state that, "Trials with inadequate or unclearly described randomization methods exaggerate the estimated treatment effect by 30-40% on average." Imperfect randomization does not necessarily cause a problem and we are not sure of the basis for O&G's quantification of bias, but to defer to their criticism, we attenuated the risk ratios for those studies by a factor of .3 to .4 (uniformly distributed).

New York: O&G find plausible evidence of selection bias: 853 women were excluded from the screened group due to previous breast cancer while only 336 were excluded from the same-sized control group based on what should have been the same criterion. This is potentially very important because the risk ratio would have been 1.0 if the extra exclusions had not happened and less than 10% of them were BC fatalities. We correct for the uncertain level of bias by adding back in the excess exclusions, randomly selecting a mortality rate for them, between the average for the treatment group (which would cause no bias) and 10%. O&G also note evidence of possible differential misclassification resulting in 44 controls who should have been counted as BC deaths. To reflect the possibility, we add back between 0 and 44 BC fatalities to the treatment group

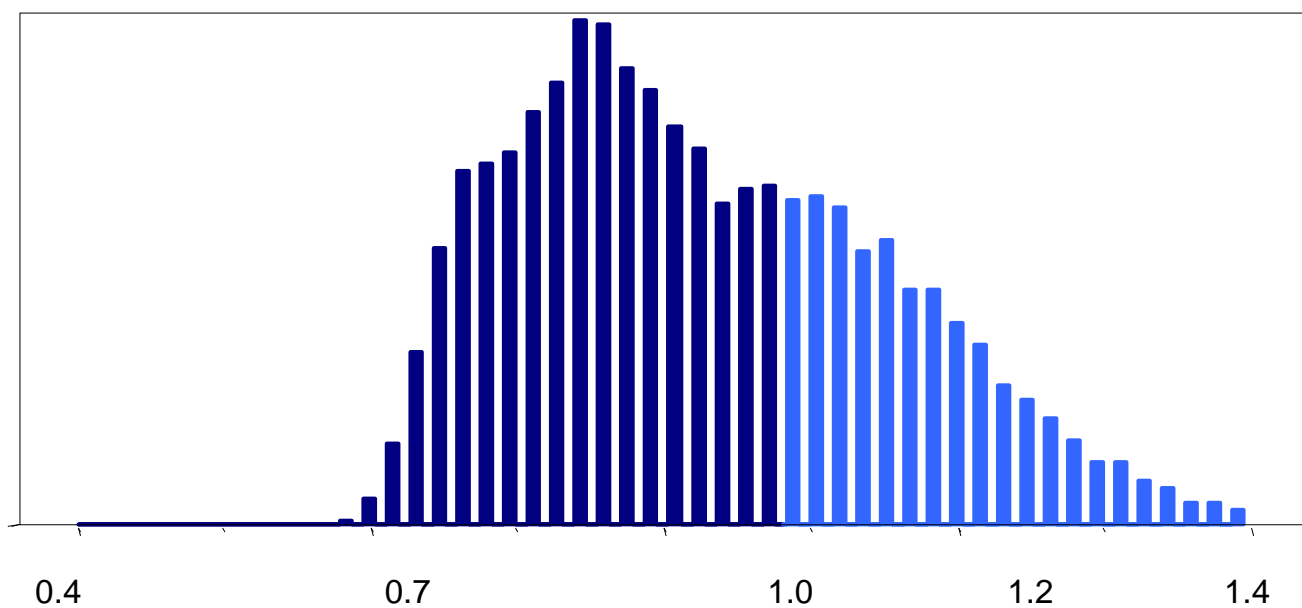
Edinburgh: O&G identify sufficiently large problems with randomization and other study procedures that it is difficult to figure exactly what this study is measuring or how to correct for errors. Serious unexplained problems make this study seem much less likely to be the "right" one, so for simplicity we accept O&G's strategy and eliminate it from consideration.

O&G provide more details about several potential problems in the studies, but we believe the above-mentioned ones are the most quantifiable and roughly capture the uncertainty created by the plausible sources of bias.

Result 2

Repeating the quantification with distributions of corrections for identified errors yields a second distribution. It suggests that we should still be fairly confident that screening mammography reduces BC mortality. (Assuming the model is right, about 2/3 probability.) It is a reduction in our confidence, but far from a sea-change.

Plausible values of the risk ratio for the effect
of screening mammography on BC mortality
(eight trials, corrected for O&G's hypothesized biases)



What do the trials tell us if we accept the responses to O&G's criticisms

This analysis can be updated as we acquire new information.

There have been many comments on O&G's analysis. The one that seems to most fundamentally address O&G's criticisms of the studies is Nystrom et al. [Lancet 2002], which offered explanations for most of the apparent problems with the Swedish studies.

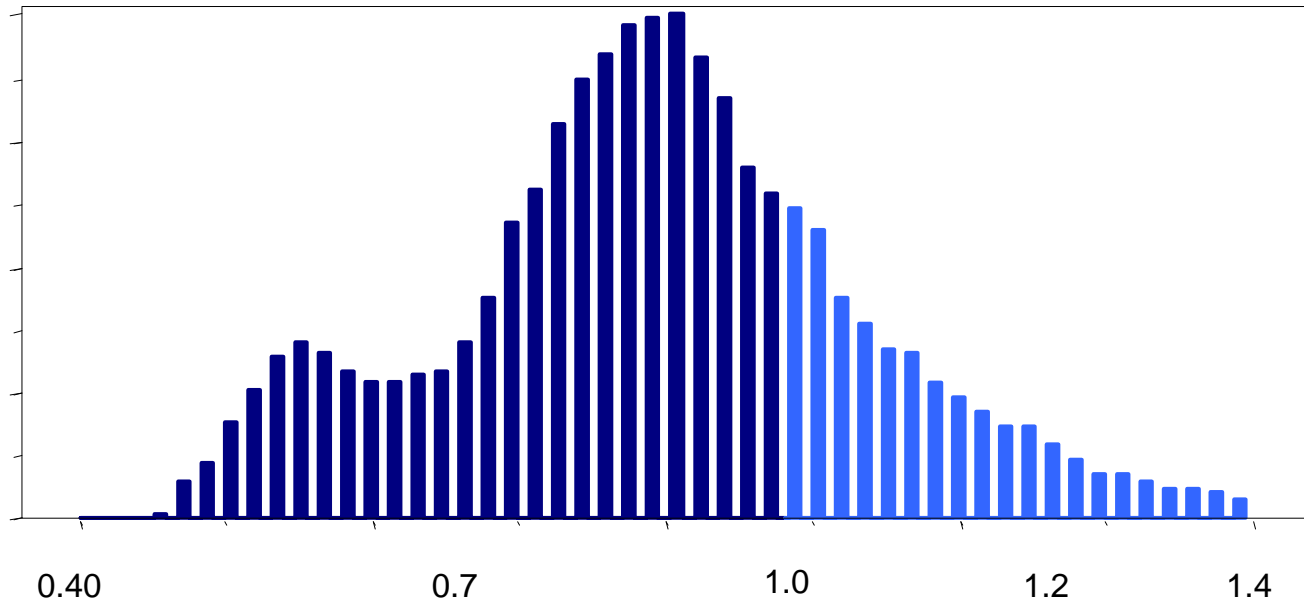
Nystrom et al.'s explanations of randomization discrepancies seem adequate, providing full descriptions of methods and explaining the effects of cluster randomization. Additionally, Nystrom et al. provide updated data, which we include in the final model run below. In particular, they provide data for the missing age range for Goteborg, eliminating the speculation of publication bias.

A second convincing response to O&G was a letter by Miller [Lancet, Dec 2001], accepted as sufficient by O&G, that effectively countered the speculation of differential misclassification for New York. O&G did not accept this claim, but we accept it for purposes of calculating the third result. Miller also tried to explain the post-randomization pre-existing BC exclusion imbalance, but we found this less than convincing.

Result 3

Removing the rebutted hypothesized biases and including the new adjustments yields the following distribution of plausible true risk ratios shown in the following figure. This shows a reduction of uncertainty toward the original distribution, though it remains wider and further to the right. (Assuming the model parameters are correct, we should be about 3/4 certain of a beneficial effect.)

Plausible values the risk ratio for the effect
of screening mammography on BC mortality
(eight trials, corrected for responses to O&G's hypothesized biases)



Discussion

This is clearly a very rough method for estimating uncertainty, but it strikes us as basically plausible and practical, and much better than no attempt to estimate the total uncertainty in the estimated effects.

We would not put much faith in the exact magnitude of uncertainty, let alone a particular probability from the distributions, but this does not change the main lessons:

- ? We should generally consider ourselves much less certain than the usual dichotomy implies.

- ? Almost all new information from individual sources of data changes our total uncertainty only modestly.

The evidence we had at any time could never justify any claim of certainty, but neither did it leave us in total ignorance. Assessed realistically, we were fairly sure that mammography saved lives at a substantial (though not necessarily cost-effective) rate before O&G's analysis. After O&G's analysis, we were merely fairly sure.

Thus, if the evidence really was enough to warrant heavily marketing mammography before O&G, it was almost certainly enough after -- not because O&G should have been ignored or dismissed, but *even if they were right* in their concerns (though not in their solutions).

General Lessons

Dichotomy is always the wrong way to consider our results -- it merely becomes more evident for a situation like this one.

Good study procedure is not sufficient to make a study error-free and identifiable bad procedure does not render it vacuous.

We need to become better at quantifying uncertainty. Quantified uncertainty is what is needed to make optimal policy decisions. Expert panels that are charged with or adopt the responsibility of advising policy should endeavor to find better ways to report uncertainty, rather than forcing consensus and synthesis that hides it.

More uncertainty is necessary.