COMMENTARY

A Science for the Art of Consensus

David L. Sackett*

Editor’s note: How is it that different people of good intent can look at the same data and draw opposing conclusions? The recent National Institutes of Health Consensus Development Conference on screening mammography for women ages 40 through 49 is a prime example of this dilemma: The conference was convened to weigh information on the effectiveness of screening this age group and to see if an across-the-board recommendation for mammography was warranted. The panel, composed of non-federal medical and public health experts and patient representatives, concluded that no recommendation should be made that would apply to all women in this age range. A minority report was produced by two panel members who thought otherwise. Several weeks later, on the basis of virtually the same data, the National Cancer Advisory Board, the principal advisory body to the National Cancer Institute, recommended in favor of screening women in their forties. Public reaction was polarized in the ensuing confusion.

In this issue, we publish the full text of the consensus statement, as well as the minority report (pages 1015-26). We have also included the following Commentary by Dr. Sackett to illustrate the nature of this kind of dilemma, one bound to arise many times at the interface of scientific inquiry and the formulation of public policy.

This Commentary presents two related theses: First, although consensus development conferences may be dismissed as art, there can be a science to them. Second, content experts and science bureaucrats should be the servants of such enterprises, not their masters. These theses were reinforced as I read the National Institutes of Health Consensus Development Conference Statement on breast cancer screening for women ages 40-49 (1), and the subsequent reactions to it. The two theses can be expressed in six propositions. For those with little time or short attention spans, these propositions are summarized in Table 1, and the “levels of evidence” that can provide science to the enterprise appear in Table 2.

The first three propositions have to do with screening in order to achieve early diagnosis, particularly of a potentially fatal or debilitating disease, and are included here because I haven’t often found them acknowledged by screeners and screenees alike:

1. Early diagnosis improves survival when treatment is worthless.

This is so for three reasons: First, like other individuals who accept and follow health advice and take their medicine [e.g., coronary patients who faithfully take their placebos are much less likely to die than those who don’t (2)], those who respond to invitations for screening often are destined for lower mortality at the outset. In a landmark trial of mammography and clinical examination of the breast, the two thirds of experimental women who accepted the invitation for screening had half the cardiovascular mortality of the one third who stayed away (3). Second, when the “lead time” of the earlier diagnosis achieved through screening is not corrected in short-term survival analyses, the spurious “leftward” shift in survival curves for screened patients guarantees an improvement in their short-term survival, even when treatment is ineffective (and individuals detected in this way are not given extra years forward, of life, but only extra years backward, of disease). Finally, slow-growing (better prognosis) cancers are “at risk” of detection for a longer time than fast-growing (poor prognosis) ones, and their overrepresentation in cohorts of patients whose cancers were detected through screening will, once again, suggest longer survival when therapy is worthless (4).

2. Even early diagnostic maneuvers that save lives will reduce health.

The screening maneuver that saves the life of one of three individuals diagnosed by the test can only shorten the healthy lives of the other two. The saved person is rewarded with an increased total life span; the other two are only harmed by premature knowledge of a “silent” morbidity. Thus, the introduction of even effective screening maneuvers lowers the average health of a town, and it is only when an abundance of saved lives accumulates that the net loss in the town’s health recovers even to zero.

3. Soliciting the symptomless must promise them better health.

When a patient comes to me with symptomatic incurable cancer, my pledge to them is restricted to doing my best, and I label as quack any charlatan who promises them cure or longer

*Correspondence to: David L. Sackett, M.D., Centre for Evidence-Based Medicine, University of Oxford, Level 5, John Radcliffe Hospital, Headington Way, Headington, Oxford OX3 9DU, U.K.

See “Notes” following References.”

© Oxford University Press
Propositions about screening for early diagnosis:
1. Early diagnosis improves survival when treatment is worthless.
2. Even early diagnostic maneuvers that save lives will reduce health.
3. Soliciting the symptomless must promise them better health.

Propositions about consensus conferences:
4. Conflicting conclusions might both be rational.
5. Experts aren’t.
6. Truth is not often best determined by scheduling a press conference.

<table>
<thead>
<tr>
<th>Level of evidence</th>
<th>Description</th>
<th>Grade of recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Meta-analyses or individual randomized trials in which the lower limit of the confidence interval for the treatment effect exceeds the minimal clinically important benefit</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>Meta-analyses or individual randomized trials in which the lower limit of the confidence interval for the treatment effect overlaps the minimal clinically important benefit</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>Nonrandomized concurrent cohort studies</td>
<td>C</td>
</tr>
<tr>
<td>4</td>
<td>Nonrandomized historic cohort studies</td>
<td>C</td>
</tr>
<tr>
<td>5</td>
<td>Case-series</td>
<td>C</td>
</tr>
</tbody>
</table>

3. Each views the conclusions of the other as not simply wrong, but irrational. In fact, both might be rational (even when wrong) and simply result from pursuing different, sometimes incompatible, goals. So, too, when disparate groups come together in efforts to achieve consensus about whether and how to intervene in the health of the public. Members who “must know” will clash with members who “must recommend.” The former, who regard achieving certainty about the health effects of the proposed maneuver as the highest objective, will insist on evidence from the top of the hierarchy and, in its absence, will advocate silence. They are certain to clash, in the absence of this certainty, with other members who, because they “must recommend,” will descend down the hierarchy until they can generate the clear-cut recommendation that constitutes their goal.

5. Experts aren’t.

By definition, experts seldom hesitate to tell us how to evaluate and manage our patients, and their opinion is sometimes placed on a par with Level 1 evidence by consensus committees. Thus, recent recommendations from the American College of Cardiology/American Heart Association defined Class 1 evidence as “conditions for which there is evidence [sic] and/or general agreement that a given procedure or treatment is beneficial, useful, and effective” (6). But surely the experts of former eras, whose absurd therapeutic recommendations we de- ride, were at least as smart as the contemporary experts whose therapeutic recommendations we implement with uncritical reverence. And a growing body of evidence shows that when experts put their advice on paper they practice second-rate science. Thus, Cynthia Mulrow showed (7) that expert review articles routinely failed to use scientific methods to identify, assess, and synthesize information, and Andrew Oxman and Gordon Guyatt documented (8) an inverse relationship between adherence to these methods and the experts’ self-professed expertise. (I have elsewhere described (9) the damage experts do to their areas of expertise and hope that readers of this note are indifferent to the news that I screen for some disorders but not for others.)

6. Truth is not often best determined by scheduling a press conference.

Three days is about enough time to carry out a clinical decision analysis for just one patient on my clinical service but too short a time for most of us to decide which automobile to buy, what color to paint the kitchen, or where to take our next holiday (despite prior experiences with all three). I was given five times as long simply to prepare this commentary, and the deliberation time of the gold standard assembly for systematically reviewing the effects of health care, a Review Group of the Cochrane Collaboration, is measured in years, not days. The imperatives imposed by scheduling a press release before a consensus conference brings us back to the fourth proposition: the goal is rational, but its relentless pursuit may do more harm than good to our patients.

How might these propositions be incorporated into a consensus conference? Such a deliberation could have five elements. First, it might begin by agreeing on a hierarchy of evidence, such as that described in Table 2 (or any other system judged suitable to its task). Second, it could use such a hierarchy to classify the evidence submitted by the experts who appear before it. Third, as shown in Table 2, it could append to each recommendation a notation indicating the level of evidence on which it was based (thereby labeling imperative recommendations based on shaky evidence with cautions about their validity). Fourth, the timing
of its report could be dictated only by the completion of its deliberations. Fifth, it could reconvene to update its recommendations when better evidence became available.

Impossible? Such an assembly has been held triennially since 1985 by the American College of Chest Physicians (10)—who commissioned the contents of Table 2. And to the five elements described above, it added a sixth. When participants at its first meeting recognized that their “strong” recommendation that patients with non-valvular atrial fibrillation receive anticoagulants was based on low-level evidence, they proceeded to design and carry out some of the randomized trials which, when powerfully positive, raised that recommendation to Grade A at a subsequent conference.

References


Notes

1 I didn’t attend the consensus conference on breast cancer screening and won’t comment on it other than to express my respect and admiration for many of those who took part in it.

Manuscript received June 24, 1997; accepted June 24, 1997.