Since its inception in 1984, the U.S. Preventive Services Task Force (USPSTF) has acted as a highly credible opinion-rendering body. It has sometimes carried the day and sometimes been drowned out. It has never recommended prostate-specific antigen screening for prostate cancer (1), but physicians keep ordering the test; when it proposed reducing the frequency and increasing the age of mammography screening in 2009, it was vilified, then ignored.

A 2008 law gave Medicare the power (but not the requirement) to add new preventive services if the Task Force gave them an “A” or a “B” (2). In the 2010 Patient Protection and Affordable Care Act, those services that the Task Force recommended with an “A” or a “B” earned waivers of copayments and deductibles in Medicare and private insurance. New “A” or “B” recommendations will be included in the annually updated standards from the U.S. Department of Health and Human Services for private health plans. Medicare will not face the same mandate.

The USPSTF’s expanded role since passage of the Affordable Care Act and this latest recommendation on lung cancer screening (3) provide an opportunity to take stock of the Task Force’s processes. Many things are to be commended. The committee membership is broadly representative, and the evidence reviews that underlie the recommendations are comprehensive and unbiased. However, the Task Force could break out its recommendations and the grades that accompany them to the level of granularity that the available evidence enables. It could also be more cautious about relying on modeling data to fill in gaps in the evidence, particularly when the models do not match the empirical data that are available.

Today’s grading system considers both the magnitude of the net benefit delivered by the service and the certainty of that estimate. An “A” only goes to those services for which there is high certainty that the net benefit is large. A “B” is earned when there is high certainty of a moderate net benefit, moderate certainty of a high net benefit, or even only moderate certainty of a moderate net benefit. Lung cancer screening fell into the last category.

However, the expected degree of net benefit or level of certainty about the evidence is rarely uniform, even for selected populations. In lung cancer screening, even among persons who are deemed to be “high-risk” and were eligible for the NLST (National Lung Screening Trial), there is a predictable and broad spectrum of both anticipated benefit and anticipated benefit–harm tradeoff (what the Task Force would call “net benefit”) (4). Across quintiles of lung cancer risk within the NLST, the number of participants who needed to be screened to prevent a lung cancer death, which is a measure of the probability of benefit for a person, varied by 33-fold from the lowest- to highest-risk group (5276 vs. 161 needed to screen). The number of false-positive results per prevented lung cancer death, which is a measure of the expected benefit–harm tradeoff for a person, varied 25-fold from 1648 false-positive results per prevented death to only 65 (5). Perhaps the high-risk group should have qualified for an “A”; perhaps the latter should get only a “C,” a service that should be only selectively offered.

Then there is the matter of the Task Force relying heavily on disease state models to extrapolate beyond the empirical data (6). On the basis of models, the Task Force chose to lengthen the duration of screening to a maximum of 26 years and increase the upper age of eligibility for screening to 80 years, even though NLST participants were screened for only 3 years and were ineligible to enroll if they were older than 74 years (only 8.8% of participants were aged 70 years or older at enrollment) (7). This may be appropriate, but here, too, the grading of this extrapolation should match the low level of evidence supporting it. The American College of Chest Physicians grades extrapolations outside of studied populations as a “C” (8). Most hierarchies of evidence would place modeling studies, even those with great rigor, in the category of expert opinion, the lowest level of evidence.

In this specific case, I found the Task Force’s reliance on the modeling dismaying, particularly now that its “B” rating will be converted into insurance mandates. Lung cancer is a poorly understood and highly heterogeneous condition. Even the highly accomplished Cancer Intervention and Surveillance Modeling Network (CISNET) researchers who generated the models do not seem to have been able to generate models of lung cancer that parallel its natural history or simulate the empirical pattern of benefit seen from computed tomography screening in the NLST (6). A cumulative plot of the lung cancer mortality ratio between computed tomography and chest radiography screening (see Appendix Figure 2 in the article by de Koning and colleagues [6]) seen in the NLST is flat at an approximate 20% benefit. Some CISNET models predict increasing benefits, and others predict decreasing benefits. The models only match the data approximately at the 6-year time point, and this is because they were calibrated to do so.

Seeing this, the Task Force might have stopped short of relying on these models for extrapolation well beyond the empirical data. In addition, it might have considered how little is known about the net benefit of screening annually over many years. Benefits may increase, plateau, or decrease; the harms from false-positive results may decrease per year of screening, but overdiagnosis would be expected to compound (9, 10).

Likewise, using the models to estimate the magnitude of the benefit or benefit–harm tradeoff under different conditions can give readers a sense of the uncertainty inherent in the process.
screening scenarios seems problematic. Between the models, an important estimate of benefit of computed tomography screening, the number of life-years gained per 100,000 persons, ranged from 2020 to 10,153 (6). An important measure of harm, the number of persons overdiagnosed with lung cancer, varied almost 6-fold (from 72 to 426) (6). The Task Force seems to have looked for findings where there was “consensus” between the models as a way of overcoming the heterogeneity between them. However, because they are starkly different on so many fronts, looking only for the overlap is reminiscent of the Texas sharpshooter and the fallacy that accompanies him. The sharpshooter shoots first at the barn and then draws the target around the greatest cluster of hits.

Peter B. Bach, MD, MAPP
Memorial Sloan-Kettering Cancer Center
New York, New York

Acknowledgment: The author thanks Geoffrey Schnorr, BS, from the Memorial Sloan-Kettering Cancer Center, who provided research, editorial, and administrative assistance.

Potential Conflicts of Interest: None disclosed. The form can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.0do?msNum=M13-2926.

Requests for Single Reprints: Peter B. Bach, MD, MAPP, 1275 York Avenue, New York, NY 10065.

This article was published online first at www.annals.org on 31 December 2013.

References