CME

COMMENTARY

Critical Appraisal of Clinical Studies: An Example from Computed Tomography Screening for Lung Cancer

By Nicholas P Emptage, MAE
Marguerite A Koster, MA, MFT
Joanne E Schottinger, MD
Diana B Petitti, MD, MPH

Introduction

Every physician is familiar with the impact that findings from studies published in scientific journals can have on medical practice, especially when the findings are amplified by popular press coverage and direct-to-consumer advertising. New studies are continually published in prominent journals, often proposing significant and costly changes in clinical practice. This situation has the potential to adversely affect the quality, delivery, and cost of care, especially if the proposed changes are not supported by the study's data. Reports about the results of a single study do not portray the many considerations inherent in a decision to recommend or not recommend an intervention in the context of a large health care organization like Kaiser Permanente (KP).

Moreover, in many cases, published articles do not discuss or acknowledge the weaknesses of the research, and the reader must devote a considerable amount of time to identifying them. This creates a problem for the busy physician, who often lacks the time for systematic evaluation of the methodologic rigor and reliability of a study's findings. The Southern California Permanente Medical Group's Technology Assessment and Guidelines (TAG) Unit critically appraises studies published in peer-reviewed medical journals and provides evidence summaries to assist senior leaders and physicians in applying study findings to clinical practice. In the following sections, we provide a recent example of the TAG Unit's critical appraisal of a highly publicized study, highlighting key steps involved in the critical appraisal process.

Critical Appraisal: The I-ELCAP Study

In its October 26, 2006, issue, the *New England Journal of Medicine* published the results of the International Early Lung Cancer Action Program (I-ELCAP) study, a large clinical research study examining annual computed tomography (CT) screening for lung cancer in asymptomatic persons. Though the authors concluded that the screening program could save lives, and suggested that this justified screening asymptomatic populations, they offered no discussion of the shortcomings of the study. This report was accompanied by a favorable commentary containing no critique of the study's limitations,[1] and it garnered positive popular media coverage in outlets including the *New York Times*, CNN, and the *CBS Evening News*. Nevertheless, closer examination shows that the I-ELCAP study had significant limitations. Important harms of the study intervention were ignored. A careful review did not support the contention that screening for lung cancer with helical CT is clinically beneficial or that the benefits outweigh its potential harms and costs.

… in many cases, published articles do not discuss or acknowledge the weaknesses of the research …

Critical appraisals of published studies address three questions:

1. Are the study's results valid?
2. What are the results?
3. Will the results help in caring for my patient?

We discuss here the steps of critical appraisal in more detail and use the I-ELCAP study as an example of the way in which this process can identify important flaws in a given report.

Are the Study's Results Valid?

Assessing the validity of a study's results involves addressing three issues. First, *does the study ask a*

Nicholas P Emptage, MAE, (top, left) is a Senior Analyst, Technology Assessment and Guidelines Unit for the Southern California Permanente Medical Group. E-mail: nicholas.p.emptage@kp.org.
Marguerite A Koster, MA, MFT, (top, right) is a Practice Leader, Technology Assessment and Guidelines Unit for the Southern California Permanente Medical Group. E-mail: marguerite.a.koster@kp.org.
Joanne E Schottinger, MD, (bottom, left) is an Oncologist and Hematologist at the Panorama City Medical Center and Regional Assistant Medical Director for Quality and Clinical Analysis for the Southern California Permanente Medical Group. E-mail: joanne.e.schottinger@kp.org.
Diana B Petitti, MD, MPH, (bottom, right) is Adjunct Professor, Department of Preventive Medicine, Keck School of Medicine, University of Southern California. E-mail: dbpetitti@verizon.net.

clearly focused clinical question? That is, does the paper clearly define the population of interest, the nature of the intervention, the standard of care to which the intervention is being compared, and the clinical outcomes of interest? If these are not obvious, it can be difficult to determine which patients the results apply to, the nature of the change in practice that the article proposes, and whether the intervention produces effects that both physician and patient consider important.

… does the paper clearly define the population of interest, the nature of the intervention, the standard of care to which the intervention is being compared, and the clinical outcomes of interest?

The clinical question researched in the I-ELCAP study[2] of CT screening for lung cancer is only partly defined. Although the outcomes of interest—early detection of lung carcinomas and lung cancer mortality—are obvious and the intervention is clearly described, the article is less clear with regard to the population of interest and the standard of care. The study population was not recruited through a standardized protocol. Rather, it included anyone deemed by physicians at the participating sites to be at above-average risk for lung cancer. Nearly 12% of the sample were individuals who had never smoked nor been exposed to lung carcinogens in the workplace; these persons were included on the basis of an unspecified level of secondhand smoke exposure. It is impossible to know whether they were subjected to enough secondhand smoke to give them a lung cancer risk profile similar to that of a smoker. It is also not obvious what was considered the standard of care in the I-ELCAP study. Although it is common for screening studies to compare intervention programs with "no screening," the lack of a comparison group in this study leaves the standard entirely implicit.

Second, *is the study's design appropriate to the clinical question*? Depending on the nature of the treatment or test, some study designs may be more appropriate to the question than others. The randomized controlled trial, in which a study subject sample is randomly divided into treatment and control groups and the clinical outcomes for each group are evaluated prospectively, is the gold standard for studies of screening programs and medical therapies.[3,4] Cohort studies, in which a single group of study subjects is studied either prospectively or at a single point in time, are better suited to assessments of diagnostic or prognostic tools[3] and are less valid when applied to screening or treatment interventions.[5] Screening evaluations conducted without a control group may overestimate the effectiveness of the program relative to standard care by ignoring the benefits of standard care. Other designs, such as nonrandomized comparative studies, retrospective studies, case series, or case reports, are rarely appropriate for studying any clinical question.[5] However, a detailed discussion of threats to validity arising within particular study designs is beyond the scope of this article.

The I-ELCAP study illustrates the importance of this point. The nature of the intervention (a population screening program) called for a randomized controlled trial design, but the study was in fact a case series. Study subjects were recruited over time; however, because the intervention was an ongoing annual screening program, the number of CT examinations they received clearly varied, and it is impossible to tell from the data presented how the number of examinations per study subject is distributed within the sample. With different study subjects receiving different "doses" of the intervention, it thus becomes impossible to interpret the average effect of screening in the study. In particular, it is unclear how to interpret the ten-year survival curves the report presents; if the proportion of study subjects with ten years of data was relatively small, the survival rates would be very sensitive to the statistical model chosen to estimate them.

The lack of a control group also poses problems. Without a comparison group drawn from the same population, it is impossible to determine whether early detection through CT screening is superior to any other practice, including no screening. Survival data in a control group of unscreened persons would allow us to determine the lead time, or the interval of time between early detection of the disease and its clinical presentation. If individuals in whom stage I lung cancer was diagnosed would have survived for any length of time in the absence of screening, the mortality benefit of CT screening would have been overstated. Interpreting this interval as life saved because of screening is known as lead-time bias. The lack of a comparable control group also raises the question of overdiagnosis; without survival data from control subjects, it cannot be known how many of the lung cancers detected in I-ELCAP would have progressed to an advanced stage.

The types of cancers detected in the baseline and annual screening components of the I-ELCAP study only underscore this concern. Of the cancers diagnosed at baseline, only 9 cancers (3%) were small cell cancer, 263 (70%) were adenocarcinoma, and 45 (22%) were squamous cell cancer. Small cell and squamous cell cancers are almost always due to smoking. Data from nationally representative samples of lung cancer cases generally show that 20% of lung cancers are small cell, 40% are adenocarcinoma, and 30% are squamous cell. The prognosis for adenocarcinoma is better even at stage I than the prognoses for other cell types, especially small cell.[6] The I-ELCAP study data suggest that baseline screening might have detected the slow-growing tumors that would have presented much later.

A third question is *whether the study was conducted in a methodologically sound way*. This point concerns the conduct of the study and whether additional biases apart from those introduced by the design might have emerged. A discussion of the numerous sources of bias, including sample selection and measurement biases, is beyond the scope of this article. In randomized controlled trials of screening programs or therapies, it is important to know whether the randomization was done properly, whether the study groups were comparable at baseline, whether investigators were blinded to group assignments, whether contamination occurred (ie, intervention or control subjects not complying with study assignment), and whether intent-to-treat analyses were performed. In any prospective study, it is important to check whether significant attrition occurred, as a high dropout rate can greatly skew results.

In the case of the I-ELCAP study,[2] these concerns are somewhat overshadowed by those raised by the lack of a randomized design. It does not appear that the study suffered from substantial attrition over time. Diagnostic workups in the study were not defined by a strict protocol (protocols were recommended to participating physicians, but the decisions were left to the physician and the patient). This might have led to variation in how a true-positive case was determined.

What Are the Results?

Apart from simply describing the study's findings, the results component of critical appraisal requires the reader to address the *size of the treatment effect* and the *precision of the treatment-effect estimate* in the case of screening or therapy evaluations. The treatment effect is often expressed as the average difference between groups on some objective outcome measure (eg, SF-36 Health Survey score) or as a relative risk or odds ratio when the outcome is dichotomous (eg, mortality). In cohort studies without a comparison group, the treatment effect is frequently estimated by the difference between baseline and follow-up measures of the outcome, though such estimates are vulnerable to bias. The standard errors or confidence intervals around these estimates are the most common measures of precision.

The results of the I-ELCAP study[2] were as follows. At the baseline screening, 4186 of 31,567 study subjects (13%) were found by CT to have nodules qualifying as positive test results; of these, 405 (10%) were found to have lung cancer. An additional five study subjects (0.015%) with negative results at the baseline CT were given a diagnosis of lung cancer at the first annual CT screening, diagnoses that were thus classified as "interim." At the subsequent annual CT screenings (delivered 27,456 times), 1460 study subjects showed new noncalcified nodules that qualified as significant results; of these, 74 study subjects (5%) were given a diagnosis of lung cancer. Of the 484 diagnoses of lung cancer, 412 involved clinical stage I disease. Among all patients with lung cancer, the estimated ten-year survival rate was 88%; among those who underwent resection within one month of diagnosis, estimated ten-year survival was 92%. Implied by these figures (but not stated by the study authors) is that the false-positive rate at the baseline screening was 90%—and 95% during the annual screens. Most importantly, without a control group, it is impossible to estimate the size or precision of the effect of screening for lung cancer. The design of the I-ELCAP study makes it impossible to estimate lead time in the sample, which was likely substantial, and again, the different "doses" of CT screening received by different study subjects make it impossible to determine how much screening actually produces the estimated benefit.

Will the Results Help in Caring for My Patient?

Answering the question of whether study results help in caring for one's patients requires careful consideration of three points. First, *were the study's patients similar to my patient*? That is, would my patient have met the study's inclusion criteria, and if not, is the treatment likely to be similarly effective in my patient? This question is especially salient when we are contemplating new indications for a medical therapy. In the I-ELCAP study,[2] it is unclear whether the sample was representative of high-risk patients generally; inso-

… would my patient have met the study's inclusion criteria, and if not, is the treatment likely to be similarly effective in my patient?

far as nonsmokers exposed to secondhand smoke were recruited into the trial, it is likely that the risk profiles of the study's subjects were heterogeneous. The I-ELCAP study found a lower proportion of noncalcified nodules (13%) than did four other chest CT studies evaluated by our group (range, 23% to 51%), suggesting that it recruited a lower-risk population than these similar studies did. Thus, the progression of disease in the presence of CT screening in the I-ELCAP study might not be comparable to disease progression in any other at-risk population, including a population of smokers.

The second point for consideration is *whether all clinically important outcomes were considered*. That is, did the study evaluate all outcomes that both the physician and the patient are likely to view as important? Although the I-ELCAP study did provide data on rates of early lung cancers detected and lung cancer mortality, it did not address the question of morbidity or mortality related to diagnostic workup or cancer treatment, which are of interest in this population.

… did the study evaluate all outcomes that both the physician and the patient are likely to view as important?

Finally, physicians should consider *whether the likely treatment benefits are worth the potential harms and costs*. Frequently, these considerations are blunted by the enthusiasm that new technologies engender. Investigators in studies such as I-ELCAP are often reluctant to acknowledge or discuss these concerns in the context of interventions that they strongly believe to be beneficial. The I-ELCAP investigators did not report any data on or discuss morbidity related to diagnostic procedures or treatment, and they explicitly considered treatment-related deaths to have been caused by lung cancer. Insofar as prior research has demonstrated that few pulmonary nodules prove to be cancerous, and because few positive test results in the trial led to diagnoses of lung cancer, it is reasonable to wonder whether the expected benefit to patients is offset by the difficulties and risks of procedures such as thoracotomy. The study report also did not discuss the carcinogenic risk associated with diagnostic imaging procedures. Data from the National Academy of Sciences' Seventh report on health risks from exposure to low levels of ionizing radiation[7] suggest that radiation would cause 11 to 22 cases of cancer in 10,000 persons undergoing one spiral CT. This risk would be greatly increased by a strategy of annual screening via CT, which would include many additional CT and positron-emission tomography examinations performed in diagnostic follow-ups of positive screening results. Were patients given annual CT screening for all 13 years of the I-ELCAP study, they would have absorbed an estimated total effective dose of 130 to 260 mSv, which would be associated with approximately 150 to 300 cases of cancer for every 10,000 persons screened. This is particularly critical for the nonsmoking study subjects in the I-ELCAP sample, who might have been at minimal risk for lung cancer; for them, radiation from screening CTs might have posed a significant and unnecessary health hazard.

In addition to direct harms, Eddy[5] and other advocates of evidence-based critical appraisal have argued that there are indirect harms to patients when resources are spent on unnecessary or ineffective forms of care at the expense of other services. In light of such indirect harms, the balance of benefits to costs is an important consideration. The authors of I-ELCAP[2] argued that the utility and cost-effectiveness of population mammography supported lung cancer screening in asymptomatic persons. A more appropriate comparison would involve other health care interventions aimed at reducing lung cancer mortality, including patient counseling and behavioral or pharmacologic interventions aimed at smoking cessation. Moreover, the authors cite an upper-bound cost of $200 for low-dose CT as suggestive of the intervention's cost-effectiveness. Although the I-ELCAP study data do not provide enough information for a valid cost-effectiveness analysis, the data imply that the study spent nearly $13 million on screening and diagnostic CTs. The costs of biopsies, positron-emission tomography scans, surgeries, and early-stage treatments were also not considered.

Summary

Using the example of a recent, high-profile study of population CT screening for lung cancer, we discussed the various considerations that constitute a critical appraisal of a clinical trial. These steps include assessments of the study's validity, the magnitude and implications of its results, and its relevance for patient care. The appraisal process may appear long or tedious, but it is important to remember that the interpretation of emerging research can have enormous clinical and operational implications. In other words, in light of the stakes, we need to be sure that we understand what a given piece of research is telling us. As our critique of the I-ELCAP study report makes clear, even high-profile studies reported in prominent journals can have important weaknesses that may not be obvious on a cursory read of an article. Clearly, few physicians have time to critically evaluate all the research coming out in their field. The Technology Assessment and Guidelines Unit located in Southern California is available to assist KP physicians in reviewing the evidence for existing and emerging medical technologies. ❖

Acknowledgment

Katharine O'Moore-Klopf of KOK Edit provided editorial assistance.

References

1. Unger M. A pause, progress, and reassessment in lung cancer screening. N Engl J Med 2006 Oct 26;355(17):1822–4.
2. The International Early Lung Cancer Action Program Investigators. Survival of patients with stage I lung cancer detected on CT screening. N Engl J Med 2006 Oct 26;355(17):1763–71.
3. Campbell DT, Stanley JC. Experimental and quasi-experimental designs for research. Chicago: Rand McNally; 1963.
4. Holland P. Statistics and causal inference. J Am Stat Assoc 1986;81:945–60.
5. Eddy DM. A manual for assessing health practices and designing practice policies: the explicit approach. Philadelphia: American College of Physicians; 1992.
6. Kufe DW, Pollock RE, Weichselbaum RR, et al (editors). Cancer Medicine (6th ed). Hamilton, Ontario, Canada: BC Decker; 2003.
7. National Academy of Sciences. Health risks from exposure to low levels of ionizing radiation: BEIR VII. Washington, DC: National Academies Press; 2005.

Perfection

You know you've achieved perfection in design
not when you have nothing more to add,
but when you have nothing more to take away.

— Antoine de Saint-Exupèry, 1900-1944, pioneer aviator, poet and novelist