[Special Features]

CHEST

Extending the Reach of Evidence-Based Medicine

A Proposed Categorization of Lower-Level Evidence

Frank C. Detterbeck, MD; Michael K. Gould, MD; Sandra Z. Lewis, PhD; and Sheena Patel, MPH

Clinical practice involves making many treatment decisions for which only limited formal evidence exists. While the methodology of evidence-based medicine (EBM) has evolved tremendously, there is a need to better characterize lower-level evidence. This should enhance the ability to appropriately weigh the evidence against other considerations, and counter the temptation to think it is more robust than it actually is. A framework to categorize lower-level evidence is proposed, consisting of *nonrandomized comparisons*, *extrapolation using indirect evidence*, *rationale*, and *clinical experience* (ie, an accumulated general impression). Subtypes are recognized within these categories, based on the degree of confounding in nonrandomized comparisons, the uncertainty involved in extrapolation from indirect evidence, and the plausibility of a rationale. Categorizing the available evidence in this way can promote a better understanding of the strengths and limitations of using such evidence as the basis for treatment decisions in clinically relevant areas that are devoid of higher-level evidence.

CHEST 2017; ■(■):■-■

KEY WORDS: evidence-based medicine; GRADE; lower-level evidence

The practice of medicine has become increasingly complex, due to the explosion of knowledge and advancements in clinical science. Practicing on a virtual island, using an individual physician's judgment, is no longer acceptable. Judgment must be enhanced by clinical science—that is, evidence-based medicine.

Paralleling advances in clinical science are advances in methodologic science. Standards have evolved for a good-quality systematic review and grading levels of evidence.[1-4] Clinical guidelines distill the expanding body of evidence down to succinct clinical treatment recommendations.[5-7] Standards for high-quality clinical guidelines continue to be refined.[7-10]

The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) system, which has been widely adopted among guideline developers,[1,11-14] classifies guideline recommendations as strong (when there is little uncertainty that benefits outweigh harms for most patients, typically based on high- or moderate-quality evidence) or weak (when there is greater uncertainty, typically involving low- or very low-quality evidence).

ABBREVIATIONS: EBM = evidence-based medicine; GRADE = Grading of Recommendations Assessment, Development, and Evaluation; RCT = randomized controlled trial; SBRT = stereotactic body radiotherapy

AFFILIATIONS: From the Section of Thoracic Surgery, Department of Surgery (Dr Detterbeck), Yale University, New Haven, CT; the Department of Research and Evaluation (Dr Gould), Kaiser Permanente Research, Pasadena, CA; Doctor Evidence (Dr Lewis), Santa Monica, CA; and American College of Chest Physicians (Ms Patel), Chicago, IL.

CORRESPONDENCE TO: Frank Detterbeck, MD, Section of Thoracic Surgery, Department of Surgery, Yale University, BB205, 330 Cedar St, New Haven, CT 06520; e-mail: frank.detterbeck@yale.edu

DOI: https://doi.org/10.1016/j.chest.2017.09.006

REV 5.5.0 DTD ■ CHEST1330_proof ■ 26 September 2017 ■ 10:38 pm ■ EO: CHEST-17-0837

Well-designed randomized controlled trials (RCTs) without serious limitations are typically considered high-quality evidence and observational studies as low-quality evidence; however, RCTs with flaws, inconsistencies, indirectness, or imprecision may be downgraded to moderate (or low) in quality, and exceptional observational studies may be upgraded to moderate (or rarely high) in quality.[1,12] The primary focus of GRADE is on guideline recommendations and the level of evidence acceptable as a basis for these.

In actual practice, however, there are differences between the nature of clinical guidelines and that of clinical care. One aspect is that there is only lower-level evidence for most of the decisions that clinicians must make. When developing guideline recommendations, there is a natural tendency to focus on areas with higher quantity and quality of evidence, and to avoid making statements when the evidence is meager (although this is not the proper guideline development/evidence-based medicine approach). In clinical practice, however, one cannot avoid making decisions, regardless of how much or little evidence there is.

Another aspect is that clinical guidelines outline what should be done in an ideal setting for an average patient (is there such a thing?). In clinical practice, one makes decisions about an individual and the setting at hand. One integrates what is known in general with an assessment of what will happen to an individual patient. This is clinical judgment, and involves weighing the strength of various arguments and considerations and our confidence in our knowledge.

In general, physicians' abilities to make such complex decisions are good. Nevertheless, we want to enhance simple "gut feelings" with available evidence as much as possible. However, clinicians are often confused and struggle to weigh the evidence appropriately. Tools describing the strength and limitations of lower-level evidence are needed to provide a structure to appropriately weigh (ie, not overplay) the evidence base in clinical decision-making. A better understanding can lead to clearer thinking, recognition of the limitations of the evidence, and enhanced ability to incorporate clinical science into the full spectrum of patient care. This article proposes a structure to categorize lower-level evidence in a way that is useful for application by clinicians in the course of clinical care.

Proposed Categories of Lower-Level Evidence

We distinguish four categories of lower-level evidence to address how to use it in clinical decision-making (Table 1). This paper focuses on nonrandomized evidence; while randomized studies with flaws also constitute lower-level evidence, discussion of this is beyond the scope of this paper.

Nonrandomized Comparison

The category of nonrandomized comparison requires that data be available from two groups (eg, receiving different treatment). Groups can be chosen in many ways, such as an intervention group vs a historical cohort, a contemporary (untreated) cohort, a matched case-control study, and so on. We must assume that there are (recognized or unrecognized) differences between the groups. Frequently the treatment selection for a patient is based on patient and disease characteristics that are inherently also associated with prognosis, for example, sicker patients are less (or more) often treated but already at greater risk of poor survival. This is known as *confounding*—technically defined as the presence of a variable that is associated with both the intervention and the outcome of interest. If not controlled for, confounding can lead to spurious associations or spuriously absent associations between the intervention and the outcome. We distinguish four subcategories, reflecting the likelihood of spurious results.

Probably Not Confounded Comparison: Rarely, one can be reasonably confident that there is little unaccounted confounding in a nonrandomized comparison. The key requirement is assurance that the cohorts are similar with respect to *all potential* confounding factors (eg, demographics, selection, disease state, health care structural aspects, time periods) or that appropriate statistical methods have adjusted for differences. The bar for this subcategory must be set very high. The topic should be well studied (meaning major confounding factors have been identified) and the comparison must address *all known or suspected* confounding factors. Nevertheless, one must be wary of unknown confounders. Ideally, there should be several comparative studies with consistent results. Lack of clarity regarding all potential confounding factors in each group or how matching or adjustment was accomplished makes it inappropriate to categorize a comparison as probably not confounded.

Furthermore, even when an association between an intervention and an outcome is quite certain, we must

REV 5.5.0 DTD ■ CHEST1330_proof ■ 26 September 2017 ■ 10:38 pm ■ EO: CHEST-17-0837