What Constitutes Reasonable Evidence of Efficacy and Effectiveness to Guide Oncology Treatment Decisions?

DANIEL SARGENT

Mayo Clinic, Rochester, Minnesota, USA

Key Words. Evidence-based medicine • Pragmatic trials • Cluster randomization • Comparative effectiveness

ABSTRACT

The need to practice evidence-based medicine is the current prevailing paradigm within the medical community. Evidence to guide practice can and should come from a variety of sources, including clinical trials, observational studies, and meta-analyses of both or either. This paper discusses the relative strengths and weaknesses of data that arise from these various sources. The different types of evidence required to demonstrate “efficacy” versus “effectiveness,” a critical and often overlooked distinction, are discussed. In the genomic age, in which targeted therapies with or without specific biomarkers are emerging in cancer care, new approaches are necessary to generate the evidence required for decision making.

INTRODUCTION

The need to practice evidence-based medicine is the current prevailing paradigm within the medical community. It has been estimated that only 50% of current medical practice is evidence based [1], clearly demonstrating a compelling need to collect and analyze additional data to better inform practice. These data can and should come from a variety of sources, including clinical trials, observational studies, and meta-analyses of both or either. This paper discusses the relative strengths and weaknesses of data that arise from these various sources, with an emphasis on methods to obtain the required data in the most rapid, cost-effective, yet reliable manner. Several specific examples are provided. This is followed by a discussion of the different types of evidence required to demonstrate “efficacy” versus “effectiveness,” a critical and often overlooked distinction. The manuscript closes with a discussion of approaches that are necessary to generate the evidence required for decision making in the genomic age, in which targeted therapies with or without specific biomarkers are emerging in cancer care.

SOURCES AND STRENGTHS OF EVIDENCE

The U.S. Preventive Services Task Force has established a well-recognized scale to evaluate the level of evidence arising from various types of studies [2] (Table 1). This scale is based on the premise that the strongest evidence comes from randomized clinical trials (RCTs), which are appropriately the “gold standard”; under these criteria, RCTs are necessary to generate level I evidence. Other types of stud-
ies generate level II or III evidence, depending on the relative strengths of their design.

RCTs

The cornerstone nature of the RCT in generating evidence is based on the premise that the act of randomization balances both the known and, more importantly, the unknown characteristics of the patients treated with the two or more interventions under study. This act of randomization allows causal inference—it assures that the only systematic difference in the patients on the treatments in question is the intervention, and thus allows the ability to conclude that any difference in outcome between study arms is caused by the intervention. The strength of the study is heightened when the randomization is stratified to guarantee balance between the study arms for factors of known prognostic importance. Other critical elements of quality in an RCT, or any study, include prespecified endpoints and specific hypotheses regarding the effect of the treatment under study on those endpoints. Clearly defined eligibility criteria, a consistent method for treatment delivery, unbiased endpoint ascertainment (optimally through placebo control), and fully defined statistical considerations are also required for a high-quality RCT. Clinical trials that meet these standards are generally required to allow for full U.S. Food and Drug Administration (FDA) approval of new therapies.

The appropriate endpoint for an RCT is a critical decision. Per FDA standards, the most appropriate endpoint is one that demonstrates “clinical benefit” to the patient [3]. In oncology, a clinical benefit is generally accepted as an improvement in survival or, in some cases, a clear improvement in time to disease progression together with an improvement in symptoms or quality of life. Because demonstrating an improvement in a clinical benefit endpoint is challenging in many cases, and may require lengthy and costly studies, trials often seek to use a “surrogate endpoint,” which is an endpoint that reliably predicts the eventual outcome of the associated clinical benefit endpoint. An example of a validated surrogate endpoint is disease-free survival as a surrogate for overall survival in the adjuvant setting for colon cancer [4]. The validation of surrogate endpoints typically requires a meta-analysis of data from multiple RCTs [5].

As currently implemented, most RCTs are lengthy and expensive, which greatly limits both the number of trials that can be completed and the number of physicians who are willing and/or able to enter their patients into such trials. These limitations reduce the ability of the RCT to inform treatment decisions, and imply that most RCTs can only inform on issues of treatment efficacy, as opposed to effectiveness, as described more fully in the next section. There are, however, existing methods that are currently greatly underused that would expand the ability of RCTs to generate reliable evidence in a much broader setting. The first of these is the concept of “pragmatic trials” [6, 7], alternately referred to as “large-simple trials” [8]. In these trials, the goal is to make patient eligibility criteria and treatment as close to standard medical practice as necessary, to keep data collection to a minimum, and to let randomization as well as the large number of patients overwhelm the variability that is introduced into the trial by the lack of complete standardization. Although a number of such trials have been successfully completed [9], such trials remain relatively rare in the U.S.

A second method that preserves the benefit of randomization while reducing the complexities involved in individual patient randomization is the cluster randomized trial. In these trials, the unit of randomization is not the individual patient but rather a predefined group of patients, such as all patients treated by a single physician or institution, or all patients treated within a certain geographic region (such as a city or county). Such trials are ideal for circumstances in which individual patient randomization is challenging or impractical, such as for a public health intervention. Such trials require specialized analysis methodology, because of intracluster correlation in individual patient outcomes; however, this is a small price to pay to preserve the benefits of randomization in cases in which individual patient randomization is infeasible or impossible.

Nonrandomized Studies

A multitude of study designs are present to provide evidence regarding treatment efficacy in the nonrandomized setting. These include cohort studies, in which a group of patients is identified and then followed; case–control studies, in which patients who have the condition of interest are identified and then attempted to be matched with similar in-

<p>| Table 1. U.S. Preventive Services Task Force levels of evidence |
|-------------------------|-----------------------------|</p>
<table>
<thead>
<tr>
<th>Level</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Evidence from at least one properly randomized, controlled trial</td>
</tr>
<tr>
<td>II-1</td>
<td>Evidence from well-designed, controlled trials without randomization</td>
</tr>
<tr>
<td>II-2</td>
<td>Evidence from well-designed, cohort or case–control analytic studies</td>
</tr>
<tr>
<td>II-3</td>
<td>Evidence from multiple time series with or without the intervention</td>
</tr>
<tr>
<td>III</td>
<td>Opinions of authorities, clinical experience; descriptive studies and case reports; reports of expert committees</td>
</tr>
</tbody>
</table>
individuals who do not have the condition of interest; and case series, in which a series of patients uniformly treated is followed. Nonrandomized trials are critical to provide information on conditions for which RCTs have not been conducted, and to generate hypotheses that can be tested in RCTs. However, the lack of randomization in such trials implies that findings are always susceptible to potential bias.

The literature is conflicting regarding the reliability of evidence generated by large nonrandomized studies, with some reports of similar conclusions from RCTs and nonrandomized trials [10, 11] and other reports reaching the opposite conclusion. The very prominent example of the conflicting findings from observational trials versus the RCT evidence generated from the Women’s Health Initiative for reducing the risk for heart disease with the use of hormone replacement therapy (HRT) in menopausal women clearly demonstrates the limitations of the use of nonrandomized data to inform treatment decisions [12]. In the case of the Women’s Health Initiative, the RCT also identified a previously unrecognized association between HRT and a higher breast cancer risk [13].

Several analysis methods are in development and use to try to improve the ability of nonrandomized trials to provide more reliable evidence of treatment efficacy. These approaches include the use of propensity scores, which is a method that attempts to develop a model to predict which patients are more or less likely to receive treatment and then use this model to estimate the effect of a treatment if the patients who received and did not receive the treatments were directly comparable [14]. Although these methods do provide benefit over an unadjusted analysis, they are limited in that they can only adjust for factors that are known and measured, when in fact, frequently, the patient characteristics that dictate treatment are poorly measurable or unknown, such as comorbidity, patient preference, or the influence of socioeconomic factors.

**Efficacy Versus Effectiveness**

RCTs are typically designed to provide evidence of “efficacy” of a new agent, where efficacy is defined as proof that the agent has a therapeutic effect. Efficacy is typically demonstrated in carefully defined patient populations, often using surrogate endpoints, and in a protocol in which treatment is given in a highly standardized manner, often compared with a placebo. It is becoming widely recognized that such efficacy studies, although providing critical evidence, do not provide adequate information to judge the impact of a new treatment when used in the real-world setting. In part, this is a result of the fact that, in oncology, only approximately 3%–5% of patients enter clinical trials [15], and the fact that clinical trials often have restrictive eligibility criteria [16]. Other factors that contribute to this discrepancy include the fact that RCTs typically exclude patients with comorbid conditions, are often over-represented with patients treated in academic medical centers (as opposed to the community setting), and frequently under-accrue racial minorities and the elderly, and the fact that once treatments are introduced into a clinical setting the regimens, dosages, and other factors deviate from the specified FDA label. The benefit of treatments when used in daily practice is referred to as “effectiveness.”

The recognition of the gap in the available evidence of effectiveness of many agents in common use has led to an increasing emphasis on (and funding for) comparative effectiveness research (CER). CER seeks to define specifically the relative benefit of treatments in daily practice. Such studies generally require an active control (as opposed to a placebo), focus on simple, pragmatic endpoints, have a greater focus on patient-reported outcomes such as satisfaction and quality of life, and frequently include a cost-effectiveness component. Ideally, such trials should be performed in clinical situations in which a direct mechanism exists to implement the findings into general clinical practice [17]. Pragmatic and/or cluster RCTs are valuable methodologies to generate CER data.

**Generating Evidence for Targeted Therapies and Biomarker-Based Treatment Decisions**

The recent successful demonstration of the clinical efficacy of targeted cancer therapies, such as imatinib [18] and trastuzumab [19] as well as the identification of biomarkers to predict the efficacy (or lack of efficacy) of agents (e.g., KRAS status for cetuximab [20] and panitumumab [21]) highlight the potential for our increasing knowledge of cancer biology to translate into more effective and less toxic therapies. The demonstration of both efficacy and effectiveness for targeted agents, or diagnostics to guide therapy, requires new approaches that are still in development.

Focusing first on establishing the efficacy of such agents or tests, standard RCT designs are likely suboptimal. For therapies that are effective in only a subset of patients, the ability to limit the population tested to that most likely to benefit may accelerate the development process and reduce the sample size necessary for definitive trials [22]. These “targeted trials,” however, do not allow the ability to assess the activity of the agent in a larger population, because only a subset of patients based on a specific characteristic are enrolled. For example, some early trials with the epidermal growth factor receptor (EGFR) inhibitor cetuximab restricted the eligibility to patients with positive immunohis-
tochemical staining for EGFR expression, which later was shown to have no relationship with the efficacy of that therapy [23]. Alternative designs for clinical trials, in which all patients are enrolled but the primary hypothesis is specified only for a specific subgroup, allow testing of the treatment efficacy in both the subgroup and the general population [24].

Effectiveness studies of these newer targeted agents at this time are virtually nonexistent [25]. A key issue in such studies involves the reliability and reproducibility of the assay: in efficacy studies, the assay is often performed in a central lab, but in community practice, that assay may be performed locally. The targeting of therapy based on a biomarker also complicates the treatment algorithm at the local site: tissue must be obtained and tested, which may delay treatment, some patients may have invaluable assay results (such patients typically would not have been included in the efficacy trials), and for agents that were tested only in a specific subpopulation, usage outside that population may be considered in cases for which no other options exist.

Critical to both efficacy and effectiveness research in the current era of biomarkers and targeted therapies is biospecimen collection. Carefully annotated biospecimens with accompanying long-term patient outcomes will greatly facilitate the identification of biomarkers for both new and existing therapies, possibly through “prospective/retrospective” studies, in which characteristics of existing biospecimens are tested for association with biomarkers based on a prospectively defined protocol [24].

**CONCLUSIONS**

The practice of evidence-based medicine requires the use of careful and disciplined methodologies to provide reliable information to guide clinical practice. Whenever possible, randomized data are clearly preferred to results based on nonrandomized trials. In the presence of bias, large sample sizes are of little benefit, and in fact may only make us more confident in the “wrong” answer. Pragmatic and/or cluster randomized trials may provide a critical link between the highly regulated “efficacy studies” that dominate current oncology research and the “effectiveness” studies that are critical to clinical practice. The establishment and validation of personalized medicine require further innovation in these areas. The need for innovative methods in oncology research is readily apparent; successful development and implementation of such methods will allow the generation of reliable evidence to guide the future of cancer care.

**REFERENCES**


19. Romond EH, Perez EA, Bryant J et al. Trastuzumab plus adjuvant chemo-


