Evidence-based Medicine: An Overview of Key Concepts

Marissa J. Carter, PhD

Abstract
Since Sackett et al established the defining principles almost 20 years ago, the concept evidence-based medicine (EBM) has grown and is reflected in the current US healthcare reform debate. To determine the relevancy of EBM elements to wound care, a review of the literature was conducted using several electronic databases. English-language literature was searched for articles published between 1985 and 2009 with regard to EBM principles, study design, systematic reviews, meta-analysis, safety, and clinical practice guidelines to find insightful and balanced studies and commentary. Current literature shows general agreement on the hierarchy of clinical trial design and the value of systematic reviews and meta-analysis but a divergence of opinions about optimal EBM rating schemes remains. Randomized controlled trials will continue to be the gold standard of trial design to evaluate treatment efficacy, but they often are poorly executed. The value of comparative observational trials is frequently underrated and although systematic reviews and meta-analysis are valuable for summarizing the evidence, care needs to be taken to avoid pitfalls. Research also suggests that clinical practice guideline implementation studies are needed if their content is to be fully appreciated by practicing clinicians. Thus, while EBM has advanced considerably in the last 20 years, tradition-based care is still evident in wound care. The challenge is to ensure EBM continues to integrate individual clinical expertise with the best available external clinical evidence from high quality research.

Key Words: review, evidence-based medicine, assessment schemes, randomized controlled trials, meta-analysis

Potential Conflicts of Interest: none disclosed

Although it is generally agreed that the US health system needs reform, debate continues on what that reform should be. Approximately 47 million individuals are not insured in the US and that number will likely increase another 7 million by the end of 2010. Uninsured adults have less access to recommended care, receive poorer quality of care, and experience worse health outcomes than insured adults. Moreover, individuals with existing medical insurance spend more on premiums and out-of-pocket costs and avoid needed healthcare because of costs. The Obama Family Foundation estimates that in the US these costs have increased by a factor of 2.33 from 1992 to 2007 — ie, in healthcare, what cost $100,000 in 1992 now costs at least $233,000. The Obama administration’s healthcare proposals cover a variety of approaches, including the creation of a plan that would ensure coverage of all individuals through a government plan; other aspects focus on costs savings and creating more efficient systems. On February 17, 2009, President Obama signed into law the American Recovery and Reinvestment Act (ARRA), which includes $19.2 billion for healthcare information technology and emphasizes the importance of comparative effectiveness under the National Institutes of Health (NIH) Challenge Grants in Health and Science Research. Aside from the potential cost savings associated with healthcare administration and better infrastructure, how can significant cost savings be achieved? The answer boils down to two concepts: 1) what works and how well does it work and 2) is it cost-effective.

The first concept falls under the umbrella of evidence-based medicine (EBM). The term evidence-based medicine first appeared in the literature in 1992 and although a variety of definitions have been published since, Sackett et al’s definition is still useful: The practice of evidence-based medicine means integrating individuals’ clinical expertise with the best available external evidence from systematic research.
Although both EBM and cost-effectiveness have been periodically under fire when they impinge on healthcare policy and autonomous clinical decision-making, a fundamental tenet remains — without evidence in medicine, patients are diagnosed and treated in an arbitrary fashion. Much of medicine has been founded on tradition-based care that dates back centuries. The application of EBM has been steadily analyzing those traditions, in many cases finding that such practices have no scientific basis. In wound care, a good example is the continuing use of gauze dressings. Consequently, when evidence compels, tradition-based medicine should yield to EBM.

The first of a two-part series examining treatment in healthcare, this article describes EBM with emphasis on what has been learned in the last 20 years and EBM’s strengths, weaknesses, and current limitations. The purpose of this review is to discuss the principles of EBM and critically examine the tools and resources commonly used in EBM practice — ie, the trial designs that collect the evidence to help formulate clinical practice guidelines (CPGs).

Methods

The EBM literature is vast, requiring judicious selection of articles. PubMed, Scopus, and Google served as the primary databases searched to obtain EBM definitions, details of websites that focus on EBM, design facets of randomized controlled trials and observation studies, strengths and weaknesses of systematic reviews and meta-analysis, and safety and adverse events issues. Further searches concentrated on techniques used in the synthesis of evidence, rating of the evidence, development of recommendations, and formulation/posting of CPGs. Several hundred combinations of key words were employed to ensure that coverage was comprehensive. Ultimately, the author searched English-language literature published between 1985 and 2009 in print and online.

Website pages, studies, and commentary were selected as sources on the basis of insightfulness, accurate content, and balance to ensure a lack of bias. However, several sources with distinct bias also were selected in the context of controversial subjects so all points of view were covered. Where possible, examples were drawn from the wound care literature for illustrative purposes.

Principles of Evidence-based Medicine

Hierarchy. The concept hierarchy is intimately associated with the level of evidence (see Table 1).

Expert opinion. If 100 wound care experts were asked question, “What is the best treatment for noninfected diabetic foot ulcers?” likely, several answers would be offered. As such, expert opinions are given the lowest evidence ranking.

Case series. Case series, the next highest level of evidence, report on a series of patients with an outcome of interest but which involves no control group. Because no matched control groups are involved, no conclusions regarding efficacy are possible. Efficacy is the capacity or power to produce a desired effect; whereas, effectiveness is the quality of being able to bring about an effect, specifically in real world practice. RCTs are typically used to assess efficacy; while observational studies are used to address effectiveness. Efficacy is a measure of the benefit resulting from an intervention for a given health problem under the ideal conditions of an investigation and can be defined in terms of how well a modality or procedure works relative to some standard control modality or procedure. For example, in a recent wound care case series that addressed the application of equine pericardium collagen wound dressing in the treatment of the neuropathic diabetic foot wounds, it was reported that when the equine material was removed (mean, 2.9 weeks), 30 of the wounds (94%) had improved, with an average reduction in wound size of 44.3% (P < .0001). Although the authors concluded the treatment was beneficial, it is not known how efficacious the treatment is compared to the gold standard of treatment (which is offloading using total contact casting [TCC]). Consequently, the case series design is considered a step above expert opinion.

Observational studies. Observational studies examine events in a clinical setting and do not involve a deliberate intervention on the part of investigators. They also do not have matched control groups; thus, they are not controlled trials. Common designs in this category include cohort and case-control studies.
In addition, observational studies include case series, as well as cohort and case control designs; thus, an observational study may or may not have a comparison group. However, the control group may sometimes be a historical cohort when a control group cannot be followed prospectively. Case-control designs are retrospective and involve identifying patients with the investigated outcome (the cases) and control patients without the same outcome and examining whether they had the exposure of interest. Such study designs are ranked above case series because they have comparative groups. An example of this study design is the cohort study of Faglia et al, who followed the long-term prognosis of critical limb ischemia in patients with diabetes over several years. Results show that major amputations were performed in 8.2% of patients who received peripheral angioplasty and 21.1% of patients who received a bypass graft, but 59.2% of patients in the group received no revascularization. Clearly, it seems that without some kind of revascularization, the risk of amputation is much higher. However, the exact risk of amputation is debatable because neither the treatment nor the no-treatment condition was randomly assigned to patients; thus, many different kinds of bias might exist that were not controlled in the study.

**Randomized controlled trials.** The randomized controlled trial (RCT) is considered the highest level of study because a well-designed RCT can control for most of the potential biases that can occur. This kind of study is defined as a group of patients who are randomized into an experimental group and a control group with the groups followed up for the variables or outcomes of interest. The RCT conducted by Rullan et al is a good example. The study involved the assessment of bemiparin in the treatment of chronic diabetic foot ulcers. The study was triple-blind — ie, the patients, the clinical assessors, and the analysts did not know which group received bemiparin or a placebo plus standard care. Ulcer improvement rates (objective decrease in ulcer area of ≥50% and/or any decrease in Wagner’s ulcer grade at 3 months) were 70.3% (26 of 37 patients) in the bemiparin group and 45.5% (15 of 33 patients) in the placebo group; 95% confidence interval ([CI] 2.3–47.3; \( P = 0.035 \)). This case provides a good idea of the magnitude of the effect, which is small, and in this size of trial is marginally statistically significant. An open-label (no blinding) RCT conducted by Blume et al in which patients with diabetic foot ulcers were randomized to negative wound pressure therapy (NPWT) or advanced moist wound therapy (AMWT) showed that a greater proportion of foot ulcers achieved complete ulcer closure with NPWT (73 of 169, 43.2%) than with AMWT (48 of 166, 28.9%) within the 112-day active treatment phase (\( P = 0.007 \)). The 95% CI for these results was not reported but can be calculated using Revman 5.0 Information Management software (Mantel-Haenszel, fixed effects, risk-difference): RD = 0.14, 95% CI = 0.4–0.24. Using the same procedure, the results of Rullan et al also were calculated: RD = 0.25, 95% CI = 0.02–0.47. The larger trial has a smaller effect size than the smaller trial yet the CIs are far narrower and the result has a higher statistical significance, despite being an open-label study. Which is the better trial and why? Would both be considered equal in terms of the level of evidence? To answer these questions, assessment methods used in EBM need consideration.

**Assessment methods: level of the evidence and recommendations.** Evidence assessment requires a systematic method for evidence collection and methodological processes for analyzing the quality and strength of the evidence and formulating recommendations. Evidence quality and strength of the evidence can be assessed using expert consensus based on a committee or Delphi process, a subjective review, or weighting according to a rating scheme.

Two key concepts to understand throughout this process are the **strength of the evidence** and **strength of the recommendations**. The strength of the evidence refers to the quality, quantity, and consistency of the evidence in any body of studies. Strength of the recommendations communicates the importance of adherence to a recommendation and is based on both the quality of the evidence and the magnitude of anticipated benefit or harm. Study ratings typically follow a I, II, III format in which level I is higher than II; whereas, most recommendation schemes follow an A, B, C evidence-level format in which A is higher than B or where recommendations are provided in terms of strong/weak.

The problem is that there are dozens of rating schemes. Some provide only general frameworks while others are specific, structured as checklists or scoring questionnaires suitable for assessing RCTs or observational studies but not necessarily both. In addition, some schemes are all-encompassing, linking the strength of the recommendation(s) to the level of the evidence and others providing more leeway. Although most rating schemes use the hierarchy outlined in Table 1, some place systematic reviews and meta-analyses at the same level as RCTs. Furthermore, some specific rating schemes only address the quality of studies and others just assess bias.

**Examples of assessment schemes.** The Oxford Centre for Evidence-based Medicine has a relatively simple set of schemes in which evidence levels are linked to recommendation grades in five categories: therapy/prevention, etiology/harm; prognosis; diagnosis; differential diagnosis/symptom prevalence study; and economic and decision analyses (see Table 2). The most important feature of the Oxford scheme is that it introduces the concept of a poor quality study, although it does not specifically define poor quality in all instances. Another issue is the operational definition of a narrow CI. The Rullan study could be easily downgraded to level II because of a relatively wide CI and the Blume study accorded a narrow CI but these would be subjective judgments to some extent.

Many schemes simply assess bias in relevant studies. Bias is prejudice, preconception, or a partiality that prevents objective consideration of an issue or situation; applied to studies, it implies a distortion of some kind. For many years, the Cochrane group has used a risk-of-bias scheme that guides its authors in systematic reviews and defines bias as “a systematic error, or deviation from the truth, in results or inferences.” The Cochrane Collaboration’s recommended tool for assessing risk of bias is neither a
scale nor a checklist but a “domain-based evaluation, in which critical assessments are made separately for different domains.” The tool asks yes/no questions in six domains, with opportunity for appropriate illustrative description. The domains include adequate sequence generation (ie, method of randomization of patients to an intervention); allocation concealment (assurance that clinicians who recruit prospective patients to a trial do not know what treatment a patient might receive once he or she is enrolled); blinding (ie, masking of the treatment to patients, investigators, and analysts); incomplete outcome data; selective outcome reporting (not reporting a protocol’s pre-specified outcomes); and other potential threats to validity. The Cochrane Handbook also contains a great deal of information useful to potential reviewers. The Cochrane approach assesses study quality through a risk-of-bias tool (sometimes referred to as internal validity). A recent study of risk bias cautions that substantial variation in agreement can occur across the domains used in the tool, so this approach may require considerable user training. On the other hand, the simplest and most widely used quality assessment scheme (known as the Jadad scale), employs only five questions and is scored as 1 for yes and 0 for no, with interpretation of scores of 0 to 2 as low quality and 3 to 5 as high quality (see Table 3). This scale has been criticized as overly simplistic and yet is still capable of generating poor inter-rater scores.

Two relatively recent schemes use more comprehensive approaches to assessing the evidence and developing recommendations. The Scottish Intercollegiate Guidelines Network (SIGN) approach uses separate checklists for different categories of studies in section 1 (as well as systematic reviews and meta-analyses) — ie, internal validity — in which one of six possible attributes is assigned: well covered, adequately addressed, poorly addressed, not addressed, not reported, and not applicable. Based on the responses in the first section (overall assessment of the study) the methodological quality of the study is rated in the second section according to the following system:

- All or most of the criteria have been fulfilled. Where they have not been fulfilled, the conclusions of the study (or systematic review/meta-analysis) are thought very unlikely to alter (++);
- Some of the criteria have been fulfilled. Where they have not been fulfilled or adequately described, the conclusions of the study (or systematic review/meta-analysis) are thought unlikely to alter (+);
- Few or no criteria are fulfilled. The conclusions of the study are thought likely or very likely to alter (−).

The levels of evidence in SIGN then are described from a combination of the hierarchy of studies in combination with the methodological quality rating.
with the methodological quality of the study with grades of recommendation following A, B, C, D (see Table 4). In developing the recommendations, the SIGN group emphasized that a considered judgment must be made about 1) the relevance and applicability of the evidence to the target group of patients for the guideline, 2) the consistency of the evidence base, and 3) the likely clinical impact of the intervention.30

The Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach32,33 involves making sequential judgments about:

- The quality of evidence across studies for each important outcome;
- Which outcomes are critical to a decision;
- The overall evidence across these critical outcomes;
- The balance between benefits and harms;
- The strength of recommendations.

In assessing quality of the evidence, GRADE indicates reviewers should consider study design (hierarchy); study quality (no explicit scheme suggested); consistency (a measure of the similarity of estimates of effect across studies); and directness (the extent to which people, interventions, and outcome measures are similar to those of interest — i.e., applicability of the studies to the population of interest). In assigning the grade of evidence, grades may be decreased based on limitations involving study quality, important inconsistencies, uncertainty about directness, imprecise or sparse data, or a high probability of reporting bias.30 Likewise, grades can be increased based on strong or very strong evidence of association, evidence of a dose-response relationship, or that plausible confounders would have reduced the effect. The quality of evidence is assigned the following grades:

- High — further research is very unlikely to change confidence in the estimate of the effect;
- Moderate — further research is likely to have an important impact on confidence in the estimate of the effect and may change the estimate;
- Low — further research is very likely to have an important impact on confidence in the estimate of the effect and is likely change the estimate;
- Very low — any estimate of effect is very uncertain.

GRADE recommendations consider the fundamental question, “Does the intervention do more good than harm?” and suggest using two recommendation categories (strong and weak) to help clinicians make informed decisions.32 Strong recommendations imply that patients should receive the recommended course of action; weak recommendations involve more contextual evaluation of the recommendation by a clinician for a given patient.

It has been evident for some time that all systems have some limitations in regard to effectiveness, harm, diagnosis, and prognosis, as well as usability by professionals, patients, and policymakers.34 The GRADE approach has been endorsed by many journals and medical societies, including the United Kingdom National Institute for Clinical Excellence (NICE), the Agency for Health Care Research and Quality (AHRQ), the World Health
Organization (WHO), and the Cochrane Collaboration. SIGN believes its approach is based on multidisciplinary, group-based judgments and suggests that GRADE is based on methodology-based judgments, which are becoming formally more complex and time-consuming and require more training. Nevertheless, the vigorous debate between the SIGN and GRADE approaches as exemplified at a recent symposium in Edinburgh suggests that EBM continues to evolve as a science and hopefully in the not-too-distant future more uniform frameworks will be reached through consensus.

Study Design

RCTs. The RCT has always been considered the gold standard of study designs because it randomizes the treatment or diagnosis of interest to groups of patients. A well-conducted RCT in which the treatment is blinded to patients, assessors, and analysts (an example of triple-blinding) with allocation concealment of treatment to patient and recruiters is likely to rate highly in EBM schemes, provided the RCT is adequately powered for the effect it is measuring and no other systematic bias is present.

One aim of a therapeutic RCT is to demonstrate a clinical difference between the control and experimental group(s) as a result of treatment. In study design, the concept of the minimal clinical important difference is defined as the smallest treatment effect that would result in a change in patient management, given its side effects, costs, and inconveniences. Yet, a random sample of 27 RCTs published in five major medical journals showed that authors do not consistently provide an interpretation of the clinical importance of their results nor provide sufficient information to allow readers to make their own interpretation. The best example of this problem is one in which the difference in primary outcome (ie, the delta value) is smaller than the minimally clinical important difference. In this instance, even if a statistically significant difference is found, it has no clinical value. Thus, it would behoove study authors to define the minimally clinical important difference for various outcomes in their methodology and demonstrate that the study was powered to detect these differences before discussing the findings in context of these differences.

In underpowered trials in which the sample size is inadequate to detect an effect, a difference between two groups may be observed that is not statistically significant. This is the classical type II error (a type I error is a situation in which an investigator rejects the null hypothesis when it is, in fact, true; and a type II error

<table>
<thead>
<tr>
<th>Levels of evidence</th>
<th>Grades of recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1++</strong> High quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias</td>
<td><strong>A</strong> At least one meta-analysis, systematic review, or RCT rated as 1 ++ and directly applicable to the target population; or a systematic review of RCTs or a body of evidence consisting principally of studies rated as 1 + directly applicable to the target population and demonstrating overall consistency of results</td>
</tr>
<tr>
<td><strong>1+</strong> Well conducted meta-analyses, systematic reviews of RCTs, or RCTs with a low risk of bias</td>
<td><strong>B</strong> Body of evidence including studies rated as 2 ++ directly applicable to the target population and demonstrating overall consistency of results; or extrapolated evidence from studies rated as 1 ++ or 1 +</td>
</tr>
<tr>
<td><strong>1-</strong> Meta-analyses, systematic reviews or RCTs, or RCTs with a high risk of bias</td>
<td><strong>C</strong> A body of evidence including studies rated as 2 + directly applicable to the target population and demonstrating overall consistency of results; or extrapolated evidence from studies rated as 2 ++</td>
</tr>
<tr>
<td><strong>2++</strong> High-quality, systematic reviews of case-control or cohort studies or high-quality case-control or cohort studies with a very low risk of confounding, bias, or chance and a high probability that the relationship is causal</td>
<td><strong>D</strong> Evidence level 3 or 4 or extrapolated evidence from studies rated as 2 +</td>
</tr>
<tr>
<td><strong>2+</strong> Well-conducted case-control or cohort studies with a low risk of confounding, bias, or chance and a moderate probability that the relationship is causal</td>
<td></td>
</tr>
<tr>
<td><strong>2-</strong> Case-control or cohort studies with a high risk of confounding, bias, or chance and a significant risk that the relationship is not causal</td>
<td></td>
</tr>
<tr>
<td><strong>3</strong> Non-analytic studies — eg, case reports, case series</td>
<td></td>
</tr>
<tr>
<td><strong>4</strong> Expert opinion</td>
<td></td>
</tr>
</tbody>
</table>

*aRCT = randomized controlled trial*

---

**Table 4. The SIGN** revised grading system for recommendations in evidence-based guidelines

<table>
<thead>
<tr>
<th>Levels of evidence</th>
<th>Grades of recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1++</strong> High quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias</td>
<td><strong>A</strong> At least one meta-analysis, systematic review, or RCT rated as 1 ++ and directly applicable to the target population; or a systematic review of RCTs or a body of evidence consisting principally of studies rated as 1 + directly applicable to the target population and demonstrating overall consistency of results</td>
</tr>
<tr>
<td><strong>1+</strong> Well conducted meta-analyses, systematic reviews of RCTs, or RCTs with a low risk of bias</td>
<td><strong>B</strong> Body of evidence including studies rated as 2 ++ directly applicable to the target population and demonstrating overall consistency of results; or extrapolated evidence from studies rated as 1 ++ or 1 +</td>
</tr>
<tr>
<td><strong>1-</strong> Meta-analyses, systematic reviews or RCTs, or RCTs with a high risk of bias</td>
<td><strong>C</strong> A body of evidence including studies rated as 2 + directly applicable to the target population and demonstrating overall consistency of results; or extrapolated evidence from studies rated as 2 ++</td>
</tr>
<tr>
<td><strong>2++</strong> High-quality, systematic reviews of case-control or cohort studies or high-quality case-control or cohort studies with a very low risk of confounding, bias, or chance and a high probability that the relationship is causal</td>
<td><strong>D</strong> Evidence level 3 or 4 or extrapolated evidence from studies rated as 2 +</td>
</tr>
<tr>
<td><strong>2+</strong> Well-conducted case-control or cohort studies with a low risk of confounding, bias, or chance and a moderate probability that the relationship is causal</td>
<td></td>
</tr>
<tr>
<td><strong>2-</strong> Case-control or cohort studies with a high risk of confounding, bias, or chance and a significant risk that the relationship is not causal</td>
<td></td>
</tr>
<tr>
<td><strong>3</strong> Non-analytic studies — eg, case reports, case series</td>
<td></td>
</tr>
<tr>
<td><strong>4</strong> Expert opinion</td>
<td></td>
</tr>
</tbody>
</table>

*aRCT = randomized controlled trial*
is a situation in which the investigator fails to reject the null hypothesis when it is false.) Similarly, when an RCT compares two treatments side by side — often called the noninferiority trial — and the power of the trial is not set to discriminate a reasonable difference between outcomes, the two treatments may be considered equivalent when they are not. Such a trial demands that, at a minimum, other previous RCTs already have evaluated the treatments separately against placebos.39

On the other hand, statistically significant results do not necessarily equate to clinically important differences. Many large drug trials use thousands of patients in each arm; statistically significant differences often are apparent, but clinically important differences are absent. One way to independently assess this situation is to calculate the number needed to treat (NNT) obtained by examining the risk differences between groups for the effect under consideration.40 Although no absolute benchmarks exist with regard to NNTs, values of 2 to 4 can be considered beneficial (values of <2 are relatively rare); whereas, an NNT of 50 (ie, 50 patients would have to be treated in order for one patient to benefit), would indicate that the treatment is ineffective and that a clinically important difference is probably absent regardless of the significance of the result.41 Unfortunately, the NNT and its companion NNH (the number needed to harm) are not explicitly required in the Consolidated Standards of Reporting Trials (CONSORT) reporting criteria.42 Thus, huge studies in which relatively small effects are observed can be misinterpreted as having important clinical significance when they are not, even though the differences between groups are statistically significant.

A situation that can increase type I error is the significance testing of multiple outcomes that may or may not be correlated. The classicists base their argument on the fact that if one tests often enough, one statistically significant outcome is inevitable.43 For example, if it is assumed that alpha, the a priori probability of incorrectly rejecting the null hypothesis is set at 0.05 by the researcher, the probability of obtaining a statistically significant result by chance will be 1 in 20. However, if 20 independent statistical tests are conducted, $P$ (the probability of finding the observed result by chance alone) will be 0.642 ($1 – (1 – 0.05)^{20}$).44 Clearly, on average one would find a statistically significant result. In other words, $P$ values should be adjusted downward to remain statistically significant to reflect the possibility of incorrectly declaring statistical significance. Although the Bonferroni43 correction is often used to adjust $P$ values in this regard, it is often too conservative; other methods have been developed to correct for these problems.44 Another school of thought argues that such adjustments depend on whether the outcomes are related and that if $P$ value adjustments are incorrectly made, a type II instead of a type I error can result.45 To avoid this minefield, researchers should present both unadjusted and adjusted $P$ values where possible and explain why a particular method was used. Discussing the magnitude of the effect in regard to multiple testing is also

| Table 5. Important considerations that affect bias in randomized controlled trials, effects in regard to quality assessment, and possible mitigations |
|---|---|---|
| Factor | Effect on results | Mitigation |
| Small block designs in multisite trials | Can lead to inadequate allocation concealment | Use mixed randomization or other techniques$^{59}$ |
| No blinding | Overestimates treatment effects by 15% in non-double-blind trials$^{50}$ | Study creating ways of achieving blinding |
| No allocation concealment | Overestimates treatment effects by 30%$^{50}$ | Use foolproof methods of assignment$^{61}$; use urn or mixed randomization methods for multisite trials$^{49}$ |
| Extensive losses to follow-up | Considerably underestimates or overestimates effects, depending on make-up of lost patients$^{48}$ | Use run-in periods or identify patients intolerant of treatment ahead of time; analyze lost and surviving patients with regard to baseline parameters$^{46}$ |
| Excessive or restrictive inclusion/exclusion criteria | Poor generalizability of results (external validity)$^{32}$ | Calculated number of patients likely to be excluded in more general populations$^{53}$; minimize exclusion criteria where possible |
| Imbalances between groups regarding baseline parameters (also differences in patient parameters between enrolment sites) | | Perform unadjusted and adjusted analyses and compare$^{56}$; consider minimization, covariate adaptive randomization, or covariate-adjusted response-adaptive randomization methods if concern arises during clinical trial design phase$^{55-58}$ |
important. Last, using one primary outcome measure or a
global assessment measure rather than many measures may
be considered.45

Missing data in trials also can have important effects. The
standard method of analysis is intention to treat (ITT) — ie, all
randomized subjects are analyzed according to their assigned
treatment.46 Use of a per-protocol analysis in which the effect of
treatment is analyzed in persons who follow the study protocol
(sometimes known as an efficacy analysis) often can lead to bias
because excluded subjects may differ in characteristics from an-
alyzed subjects. For example, Lachin47 estimated that in large
studies, the inflation in type I error probability can be severe (0.50
or higher) even when the null hypothesis is true. Many re-
searchers typically use the last observation carried forward
(LOCF) principle to fill in the missing data, although other op-
tions are possible.46 However, when treatment is effective but
nonadherence to protocol is substantial, ITT will underestimate
the magnitude of the treatment effect that occurs in adherent pa-
tients.48 When losses are severe (perhaps ≥20%, as mentioned in
the Oxford assessment scheme), particularly when losses are quite
different between groups, the bias could be large enough to con-
sider downgrading the quality of the trial on bias grounds.

Although a detailed discussion of the many considerations
that can affect assessment of RCTs is beyond the scope of this re-
view, a list of some of the more important considerations, possi-
ble mitigations, and effects in regard to quality assessment are
listed in Table 5. Not all of wound care’s problems with regard to
RCTs can be ameliorated59 but with careful forethought and plan-
ning, many issues can be defused.

Observational trials and RCTs: hierarchy revisited. Traditionally,
observational studies have been relegated to lower evi-
dence levels; clinicians taught that evidence was not as good as
that derived from RCTs. For example, a 1982 study60 that assessed
six therapies using 50 RCTs and 56 trials with historical controls
concluded that biases in patient selection may irretrievably weight
the outcome of historical controls in favor of new therapies. Al-
though no one suggests the absence of poor quality observations,
not every RCT is high quality.

A more recent evaluation61 of five medical topics based on as-
essment of meta-analyses of RCTs and observational studies also
challenged the notion that evidence from observational studies
is second best. The conclusion of the study was that well-designed
observational trials (cohort or case-control) did not systematically
overestimate the magnitude of treatment effects compared to
RCTs.

Why is so much emphasis placed on the RCT? Does a large,
well-designed, comparative observational study constitute a good
alternative in some instances? One way to assess this issue is to
contrast some of the points between observational trials and
RCTs raised by Bagshaw and Bellomo62 (see Table 6). These au-
thors contend that many of the issues raised are not considered
in the GRADE system, which rates evidence from RCTs and ob-
servational studies as high and low, respectively, and notes prob-
lems in each type of study. In essence, the notion of a rigid
hierarchy of study design needs more refinement. Better assessment criteria might permit some observational studies to be deemed level 1 and some RCTs downgraded to level 2 to better appreciate the evidence. For example, in the field of digestive surgery, Shikata et al found sample size and quality sufficiently affected RCTs to forestall proclaiming superiority based solely on study design; these authors also contend that the higher heterogeneity observed in primary outcomes between observational studies reflects the higher heterogeneity of the study subjects (in RCTs, populations studied tend to be more homogeneous). Thus, while respecting that RCTs are more rigorously controlled studies, the information provided by observational studies may be complementary.

In general, observational studies offer several advantages over RCTs: they have been found to be less expensive to conduct, offer greater flexibility with regard to length of study period and can accommodate a broader range of patients. Moreover, they can be useful in clinical settings in which random allocation is not easily accepted by patients or clinicians. In some areas of complex treatment, it might be better to start with an observational trial rather than attempting an ambitious RCT in which confounding variables may end up plaguing interpretation of the results.

Finally, observational trials may be the only practical way to obtain certain kinds of data. In wound care, rare conditions would not justify the expense of an RCT and controlled clinical study designs are not always the best method to provide evidence about aspects of wound and patient assessment (including risk assessment). In the field of hyperbaric medicine, most RCTs are relatively small, precluding realistic estimates of adverse events. The only way researchers have been able to learn about the effects of oxygen toxicity in relation to hyperbaric pressure and estimate the NNH is through observational trials; the most informative were all large case series.

RCTs also have been criticized in the context of comparative effectiveness research as ill-suited to meet evidentiary needs — specifically, the comparison of effective interventions among patients in typical patient care settings, with decisions tailored to patient needs. However, several proposals regarding design principles, operational procedures, and statistical methods have been offered to remedy these deficiencies. In wound care, some RCTs may never reach the highest levels of evidence because it is impossible to blind patients and investigators from the treatment; negative pressure wound therapy is a good example. Moreover, when exclusion criteria are too narrow, the generalizability of RCT outcomes may be difficult when transferred to real world wound-care populations. However, none of these issues should be used as arguments to specifically negate RCTs; rather, they should serve as stimuli for obtaining the highest level of evidence using the best-designed studies, being creative and more adaptive where necessary.

### Table 6. Contrast of some issues raised by Bagshaw and Bellomo in regard to observational and randomized controlled trials (RCTs)

<table>
<thead>
<tr>
<th>Issue</th>
<th>Observational trial</th>
<th>RCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inconsistent current therapy given to patients in control groups</td>
<td>What is observed, but still a problem for a comparison group</td>
<td>Potential problem</td>
</tr>
<tr>
<td>Trials of interventions in which blinding is impossible</td>
<td>Not applicable</td>
<td>Problem</td>
</tr>
<tr>
<td>Type I and II errors</td>
<td>Susceptible when comparison groups are used</td>
<td>Susceptible</td>
</tr>
<tr>
<td>Denial of critical or important treatment or diagnosis for condition</td>
<td>Ethical (observes what happens and collect data)</td>
<td>Unethical (control group)</td>
</tr>
<tr>
<td>Biological implausibility/coherence (eg, results look too good to be true or results contradict what is known)</td>
<td>Potential problem</td>
<td>Potential problem</td>
</tr>
<tr>
<td>Temporality (intervention precedes effect)</td>
<td>Problem if inconsistency in care occurs or intervention is complex</td>
<td>Problem if inconsistency in care occurs or intervention is complex</td>
</tr>
<tr>
<td>Consistency</td>
<td>Potential problem</td>
<td>Potential problem</td>
</tr>
<tr>
<td>Generalizability (applicability of findings to general populations)</td>
<td>Less of a problem</td>
<td>More of a problem</td>
</tr>
</tbody>
</table>

www.o-wm.com

OSTOMY WOUND MANAGEMENT APRIL 2010
Assessing a study. The sheer number of issues involved with clinical trial design can seem overwhelming to the EBM novice, but a surprising amount of information can be quickly gleaned by asking a few simple questions. First, what is the level of evidence to which the study belongs? RCTs often provide the most convincing evidence of treatment effects, but as deMaria,70 Editor-in-Chief of the Journal of the American College of Cardiology, has pointed out, “Data from RCTs represent the beginning of the decision-making process, not the end.” Also, it may not be possible or ethical to test certain treatments with an RCT, in which case a well-conducted observational study can contribute to the evidence. Second, was the sample size small or large? In the context of wound care, 100 subjects or more could be considered large and larger studies usually provide more precise results. Third, did the study truly demonstrate any clinical benefit? The stated definition of clinical benefit in the study may be the most relevant here and the NNT parameter may be helpful. Fourth, is harm versus benefit clearly demonstrated? Studies that do not report harm or do not adequately address adverse events should raise questions. Last, are there any significant study limitations? Issues of blinding and allocation concealment tend to have the most effect on outcomes but other issues can be critical.

Systematic reviews. Currently, many organizations include systematic reviews as class I evidence — notably, the Oxford Centre for Evidence-Based Medicine, SIGN, the National Health and Medical Research Council (Australia) when the review only includes RCTs,71 and NICE, when the systematic review includes meta-analysis. In comparison to narrative reviews, systematic reviews are an attempt to judge the evidence in a field with impartiality and can are defined by Cook et al72 as “overviews developed with the application of scientific strategies that limit bias by the systematic assembly, critical appraisal, and synthesis of all relevant studies on a particular topic.” In practice, because of eligibility criteria for systematic reviews, different methods of assessment, and the kind of evidence actually available, no systematic review can ever be completely bias-free but bias can be minimized and when carefully performed systematic reviews constitute valuable sources of information for the clinician.

Sometimes, readers of systematic reviews are frustrated because conclusions seem to be unable to provide specific guidance as a consequence of lacking or conflicting data.73 However, an inconclusive finding is important and indicates that further data are needed to distinguish between absence of evidence and evidence of absence of an effect, which are different concepts.74 Thus, a good clinician always will examine the summary graphs and tables and find studies on clinically relevant patients and use them to inform relevant practice decisions. In the final analysis, EBM is about informing patient-oriented clinical decisions with the best available data, not waiting for perfect evidence before making the decision.

Subject scope varies considerably in systematic reviews; generally, the broader the review, the less scrutiny individual trials are likely to receive. Systematic reviews that cover 50 studies cannot cover as much ground as ones that describe 10 studies. In addition, review quality is also extremely variable; thus, reporting can be as important as it is for RCTs. The Quality of Reporting Meta-analyses (QUOROM) protocol75 was originally applied to meta-analyses but its successor, Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)76,77 applies to systematic reviews as well. Wide adoption of this protocol is hoped to improve the standard of systematic reviews.

The most critical aspect of a systematic review is what it covers. In practice, systematic reviews commonly cover only RCTs

---

Figure 1. Forest plot of a fictitious meta-analysis illustrating a significant effect with no statistical heterogeneity.

Figure 2. Forest plot of a fictitious meta-analysis illustrating a significant effect and significant statistical heterogeneity.
unless trial data at that level are sparse or a large, high-quality body of observational data is available. For example, the evidence for a subject might include three poor-quality RCTs, two cohort studies (one very large and well-conducted), two published abstracts of RCTs, and one large but unpublished RCT that did not show any benefit. It can be argued that the systematic review ought to cover all of the evidence; thus, investigators need to track down all published and unpublished trials, regardless of format. The reason for this approach is that it is common in EBM to find publication bias — ie, the observation that interesting, large, well-funded studies with significant, positive outcomes are more likely to be published than studies that lack these characteristics. Tracking down abstracts and unpublished studies and obtaining sufficient information upon which to judge can be problematic, especially if a study was completed many years in the past.

In the absence of complete information, how should the evidence and systematic review be judged? The simplest way to understand in practice what a systematic review covers is to ask the question “What class of evidence does the systematic review include?” In the fictitious example presented here, if the systematic review only covered the published RCTs, it would most likely be class II, depending on the rating system used. Clinical practice guidelines posted on National Guideline Clearinghouse (NGC) must utilize the guideline template and state what system was used to rate the evidence to communicate and the level of evidence presented and, by implication the level of the systematic review, if that approach was utilized in constructing the guideline.

Similar to clinical trials, systematic reviews can be done well or badly. It is incumbent on the reader to not blindly accept what the authors of a systematic review state in terms of conclusions without first judging whether the methodology of the review was appropriate. For example, in wound care, the Food and Drug Administration considers complete wound healing to be the only satisfactory endpoint in clinical trials testing the treatment of a product or process designed to assist in wound healing. This is not an appropriate parameter in all instances — systematic reviews that discuss only complete wound healing may not be telling the whole story.

Upshur noted that the concept of EBM was founded, in part, to reduce unnecessary inconsistencies and encourage standardized practice, resulting in different hierarchies that have been proposed over the years. Systematic reviews seemingly contradict these principles. Because this may be true, the merits of considering systematic reviews as part of an EBM hierarchy no doubt will be debated for many years.

**Meta-analysis.** Meta-analysis is the pooling of similar types of summary outcome data from studies to obtain an estimate of an effect, not a simple averaging of the data at that level. A well-conducted meta-analysis can provide a more objective appraisal of the evidence than traditional narrative reviews, more precise estimates of treatment effects, and may explain heterogeneity between studies.81

The process begins by developing a protocol similar to that used for systematic reviews. The two approaches frequently are combined for convenience, although often the results from many studies cannot be pooled due to differing designs. The first step is to ask an appropriate question — eg, do daily healing rates increase when venous ulcers are treated with healing factor X in addition to usual care? This question might have been tested in several trials of different design, so data first will be sorted by hierarchy — ie, RCTs, cohort, case-control, and case series. Next, all outcome data should be converted to the same units where feasible. Finally, the researcher must decide whether the trials are broadly similar in design, quality, and other characteristics such that their outcome data can be pooled — the “apples and oranges” dilemma.82 The problem is that meta-analysis works best when all trials are virtually identical and their data combined provide a perfectly homogeneous result, which suggests no need to do a meta-analysis in the first place. Conversely, when disparate trial data are combined, the result can be statistical heterogeneity (the measure of between-study variability), a result that can potentially invalidate the meta-analysis. The solution is that perhaps meta-analysis works best with sufficient “like” data from relatively small trials where combining data makes sense.

In the broadest sense, statistical heterogeneity can be seen as the similarity or difference between studies with regard to specific outcome parameters and can be easily visualized in a forest plot. In Figure 1, which depicts a forest plot of some fictitious data with regard to complete wound healing, a fixed effects model has been chosen with computed relative risk (RR) of having a wound completely healed based on four studies that followed a 4-week treatment for 10 weeks. The weighting in this instance, represented by the size of the blue boxes in the diagram on the right and described numerically under the Weight column, is a function of variance using the Mantel-Haenszel method. The 95% CIs of the effect for each individual study show that the difference between studies is not great; thus, heterogeneity is not expected. The Cochran Q statistic, shown in the bottom left of the diagram (chi square 0.46) is not statistically significant ($P = 0.93$). The $I^2$ (inconsistency) statistic, which is another measure of heterogeneity less susceptible to number of trials in the meta-analysis and their size,83 is 0%. Thus, this meta-analysis demonstrates an unequivocal significant effect ($P < 0.0001$) with the summary estimate (RR = 1.70, 95% CI 1.32–2.19) shown graphically as the black diamond that does not cross the vertical line of no effect. This can be contrasted to the result in Figure 2 in which high heterogeneity is present ($I^2 = 75\%$) even though the overall result is still significant. One of the advantages of the $I^2$ statistic is that it quantifies the amount of heterogeneity (0% to 100%), which in this case is high (<25% is low, 25% to 50% is moderate, and >50% is high).83

When high heterogeneity is encountered, a conservative approach might suggest that the meta-analysis as conducted is inappropriate. Hence, it either should not be performed or the protocol should be modified by re-examining the clinical heterogeneity of outcomes between studies. Another approach would
be to explore the source of heterogeneity using a pre-specified protocol. For example, one could remove studies one at a time, or clusters of studies, either empirically based on a given factor or more formally using algorithms, to determine what effects these changes have on heterogeneity. Other alternatives might employ subgroup analyses or meta-regression.

Deciding on whether a fixed effects or random effects model should be used can be challenging. Where fixed effects models assume that variability is exclusively due to random variation, random effects models assume different underlying effects for each study and take this into consideration, thus providing larger CIs. In practice, the results from these two types of models will tend to be similar unless substantial heterogeneity is present. Under these circumstances, it should be pointed out that random effects models penalize high-precision studies and thus favor low-precision studies; used in this context, precision is the inverse of variance — high precision would be inversely proportional to low variance and low precision inversely proportional to high variance. Moreover, random effects models may not be conservative because point estimates are not invariably closer to the null value nor are their $P$ values necessarily larger than those of fixed-effect summaries. The best that can be said is that neither approach represents the "truth" and that both approaches are viable when used appropriately. The specific use of random effects models to account for significant heterogeneity is not appropriate; in such cases, summary point estimates should be relegated to a minor role alongside the attempt to probe the origins of heterogeneity.

As discussed previously, meta-analysis has a number of potential pitfalls and the advent of programs such as the Cochrane group’s RevMan makes it easy to perform meta-analysis without due consideration. Interpreting even simple meta-analyses can be a problem. A good illustration is the meta-analysis of Schierhout and Roberts that examined mortality rates in crystalloid versus colloid resuscitation in critically ill patients. Its subgroup analysis for patients (four studies) with burns showed a pooled RR of 1.21 (95% CI: 0.88–1.66) with regard to mortality with a chi-square value of 4.63. Alarming as the findings may have looked at the time, despite the absence of statistical heterogeneity, the result was not statistically significant. Yet, despite that fact, many burns units in the...
UK immediately discontinued the use of colloids, fearing the worst without truly understanding the situation.

This example acutely underlines the role meta-analyses should play in EBM. Although many assessment schemes incorporate high-quality meta-analyses as level 1 evidence, it can be quite difficult for experts to agree whether a particular meta-analysis adds value. Certainly meta-analytical evidence should be considered because it can augment existing trial data, but it can never replace or be used to completely summarize trial data. Most importantly, badly conducted meta-analyses or meta-analyses with significant heterogeneity can be wrongly interpreted, leading to erroneous conclusions.

Safety

The safety and adverse events of treatments and diagnostic tests are important parts of the evidence required to assess benefit versus harm, one of the steps considered in the formulation of any clinical practice guideline. One common approach utilizes the NNH, which can be derived from adverse event data in any comparative clinical trial. However, to calculate appropriate NNHs, the data have to be reported; there is evidence that reporting of adverse events and safety concerns in RCTs of drugs is poor due to lack of journal space for articles describing drug treatment studies and many other reasons.

First is the “the more you search, the more you find” paradigm, which depends on what symptoms subjects in clinical trials are asked to report and how patients are asked relevant questions. Second, definitions of harm are critical in demonstrating what events constitute an adverse event to be reported; frequently, the procedures by which the presence of such events are reported changes over time, making comparisons between trials difficult. Last, administrative issues (e.g., number and timing of follow-up visits, how forms are administered to subjects, whether expected and unexpected events are given the same weight, and whether adverse events are judged to be attributable to the treatment or procedure under test) are all important and must be considered.

Which trial type provides the best safety data? In theory, data from an RCT should be the best because of the way the control and experimental groups are selected and treated. However, in practice, data may be limited either by the sample size or the length of the trial, both of which are issues in wound care. Thus, any comprehensive evaluation of adverse events also should include evidence from large observational studies and case reports for late-appearing and rare but serious events, unexpected and difficult-to-measure harm, and identification of patient populations at especially high risk for adverse events. This point may be particularly important for systematic reviews in which the analysis only depends on RCTs. Finally, most of the reported safety data are in the form of averages. Assessing the harm versus benefit for an individual patient is much trickier because patient characteristics may differ substantially from average characteristics reported in clinical trials.

Clinical Practice Guidelines

EBM reports on subjects of interest are produced in many ways. Best practice guidelines and CPGs, as well as technical reports, represent the most popular forms of data dissemination, whether formally published in a peer-reviewed journal, posted on a website, or added to the repository of CPGs on the NGC website. The NGC was the brainchild of the AHRQ and was established to provide a resource for healthcare professionals to access CPGs. The work of Shekelle et al suggests that on average the usefulness of CPGs is about 6 years without updating and that each CPG should be reviewed every 3 years. As part of the NGC Annual Verification, the NGC now removes guidelines from their website at the end of the year for not meeting their inclusion criteria, which requires that guidelines represented in their database have been developed, reviewed, or revised within the last 5 years. This procedure represents a forward-thinking policy.

According to the Institute of Medicine, CPGs are “systematically developed statements to assist practitioner and patient decisions about appropriate healthcare for specific clinical circumstances.” Yet, despite the fact that the NGC database contains more than 2,500 CPGs, how many of these have been tested or validated is not known. The last large survey of CPGs (n = 431) was published in 2000; 67% did not report any description of the type of stakeholders, 88% provided no information on searches for published studies, and 82% did not provide any explicit grading of the strength of recommendations. Although CPGs seem to be improving over time (the study was conducted from 1988 to 1998), only 5% of the guidelines met the three quality criteria. More recent but smaller surveys suggest major validation issues with regard to CPGs.

Twenty years ago, CPGs were relatively rare; now in many areas of medicine, including wound care, their numbers have proliferated and nonexpert clinicians are tasked with distinguishing the best, most current, and the most relevant for his or her patients. The UK’s NHS Evidence website promises an accreditation system to mark the “best and most trusted sources” and provides a means of highlighting important new evidence on a monthly and annual basis. It will be interesting to see how this resource develops.

Many researchers have noted the gap between evidence and clinical practice. In a pilot study of 29 individuals and three focus group (n = 10) interviews, physicians reported they relied on clinical experience, the opinions of colleagues, and EBM summarizing electronic clinical resources more often than referring directly to the EBM literature when making clinical decisions. One reason may be that CPGs often are not well-integrated within practice routines because of issues with the guidelines, the target group of professionals, and the patients, as well as the cultural and social setting and the organizational and economic environment in which the guideline is being implemented.

In wound care, the gap can be demonstrated through two examples. A survey of general surgeons and specialist nurses in the northeast (US) showed wet-to-dry dressings were commonly preferred over traditional dressings and the north east (US) showed wet-to-dry dressings were commonly preferred over traditional dressings and the
prescribed for situations with little evidence to support their use (eg, in open surgical wounds healing by secondary intention). Although 75% of respondents had access, more sophisticated dressings were not utilized because of conflict with traditional approaches to wound care, lack of education, and cost issues (note: not cost-effectiveness, which apparently was not understood by the respondents). The traditional approach to wound care, which may be interpreted here as “I’ve always used it and I know it works,” figured prominently in the findings. This study illustrates how clinicians may not keep up with the evidence.

TCC for insensitive foot ulcers is instructive for different reasons. TCC is the gold standard for offloading pressure on an insensitive foot ulcer to facilitate wound healing and has strong evidence for its efficacy. However, two studies found that <6% of wound care providers used TCC to treat such ulcers. The reasons for such high neglect of an evidence-based treatment appear to be complex, with lengthy and messy procedures, staff training, lack of reimbursement, and patient tolerance the most commonly cited factors.

These examples and the result of other studies suggest that as much time and effort should be spent optimizing the implementation of CPGs as is currently spent developing and publishing them.

Conclusion

In the last 20 years, EBM principles have advanced considerably and the tools and resources available to practitioners have enabled clinicians to learn much more about what and how well particular products and approaches work. At the same time, the number and types of EBM assessment schemes and methods of evaluating clinical trial data have exploded, challenging the average clinician to understand the nuances involved. Although EBM has not been achieved regarding some EBM techniques regarding evidence assessment, wound care is moving toward a broader appreciation of the issues involved and how to correct deficiencies. The challenge is for EBM to present data in a format clinicians can readily apply to their patients, who really are an N of 1, while ensuring that EBM never supplants good clinical judgment. The practice of EBM is about integrating individual clinical expertise with the best available external clinical evidence from systematic research. Most importantly, even when contentious issues arise in medicine as a result of EBM, the temptation to regress back to the concept of tradition-based medical care must be resisted. Traditions may be honorable but their place in modern healthcare should be limited, lest we run the risk of “the blind leading the blind.”

EBM practitioners should be encouraged to review and adhere to EBM definitions as originally stated by Sackett and others. For the sake of wound care patients, clinicians should learn and teach accurate definitions, recognize why and how they work in other medical fields, and provide compelling examples and objective criteria for the integration of EBM in their wound care practice.

References
