

Evaluating Evidence and Grading Recommendations: The SIS/IDSA Guidelines for the Treatment of Complicated Intra-Abdominal Infections

Joseph S. Solomkin

Abstract

Background: Guidelines for the management and treatment of complicated intra-abdominal infections have been generated by a joint effort of the Surgical Infection Society and the Infectious Diseases Society of America. Continued review is needed of the process of these guideline development efforts, the evidence collected, and the recommendations developed by this collaboration.

Methods: The literature employed in the development of these guidelines and the process for the development of the recommendations was reviewed.

Results: The process for the development of the guidelines required providing answers to questions surrounding the type of evidence, the quality of that evidence, and whether the data used were generated from clinical trials with investigators blinded to the randomization scheme. The recommendations most commonly follow randomized comparative trials, with observational reports and expert opinions assuming a lesser role. Sources of bias and conflict of interest considerations also must be important in guideline development, with full disclosure from participants.

Conclusions: Standards for the evaluation of data and recommendations in the guidelines for complicated intra-abdominal infection provided a rigorous evaluation of available clinical information. This process serves as a model for the development of additional guidelines for patient care.

THE RECENTLY PUBLISHED Surgical Infection Society (SIS)/ Infectious Diseases Society of America (IDSA) Guidelines on Complicated Intra-Abdominal Infections were generated from deliberations of a panel co-chaired by John Mazuski and this writer [1,2]. This guideline is a revision of two previous documents, one produced by the SIS and one by the IDSA [3,4]. The panel for this update and merger of the two societies' activities was composed of individuals identified by each society, but generally re-created the panel that developed the 2003 guidelines. There were equivalent numbers of IDSA and SIS panelists, although each group had non-overlapping areas of expertise and knowledge.

The purpose of this article is not to restate the guideline content, but rather to discuss the more difficult process of evaluating evidence and grading recommendations. Grading has two components: The quality of the information (blinded and randomized trials vs. expert opinion) and the strength of the review group's recommendations (strong-moderate-

weak). To see the value of this general approach, it is necessary to understand the benefits of properly done prospective trials. It also is important to note that the field of surgery has been slow to turn away from "eminence-based" practice guidelines. In many situations, randomized trials are difficult if not impossible. However, the case is made that understanding the key design elements in well-done prospective trials provides insight into evaluating non-randomized trials.

Evidence-Based Review Scheme

The organization of these guidelines followed a rigid format developed by the IDSA for the large number of guidelines generated by that group (www.idsociety.org/Content.aspx?id=9088). The guidelines are formed around key questions that the panel accepted a priori (Table 1).

How recommendations are generated to answer these key questions in an evidence-based manner is a crucial issue for

TABLE 1. KEY QUESTIONS USED TO DEVELOP SPECIFIC RECOMMENDATIONS

1.	What are the appropriate procedures for initial evaluation of patients with suspected intra-abdominal infections?
2.	When should fluid resuscitation be started for patients with suspected intra-abdominal infections?
3.	When should antimicrobial therapy be initiated for patients with suspected or confirmed intra-abdominal infections?
4.	What are the proper procedures for obtaining adequate source control?
5.	When and how should microbiological specimens be obtained and processed?
6.	What are appropriate antimicrobial regimens for patients with community-acquired intra-abdominal infections of mild-to-moderate severity?
7.	What are appropriate antimicrobial regimens for patients with community-acquired intra-abdominal infections of high severity?
8.	What antimicrobial regimens should be used in patients with healthcare-associated intra-abdominal infections, particularly with regard to <i>Enterococcus</i> , methicillin-resistant <i>Staphylococcus aureus</i> , and <i>Candida</i> ?
9.	What are appropriate diagnostic and antimicrobial therapeutic strategies for acute cholecystitis and cholangitis?
10.	What are appropriate antimicrobial regimens for pediatric patients with community-acquired intra-abdominal infections?
11.	What constitutes appropriate antibiotic dosing?
12.	How should microbiological culture results be used to adjust antimicrobial therapy?
13.	What is the appropriate duration of therapy for patients with complicated intra-abdominal infections?
14.	What patients should be considered for oral or outpatient antimicrobial therapy and what regimens should be used?
15.	How should suspected treatment failure be managed?
16.	What are the key elements that should be considered in developing a local appendicitis pathway?

all guideline developers, and the processes available to do this all have serious flaws. The system used for these guidelines, required by IDSA rules, is based on an older set developed for evaluating elements in a periodic health examination in the Canadian health system [5]. Because of the enormous number of such examinations performed and the health and economic consequences of these studies, this is an important healthcare practice to scrutinize.

The published studies were first categorized according to study design and quality. Then, the recommendations developed from these studies were graded according to the strength of the evidence behind them (Table 2). It is important to note that IDSA is now moving to the GRADE system (vide infra).

Why Are High Quality Randomized Trials Considered Best Evidence?

The central problem in any clinical trial is to prevent bias from confounding the results. A simple example would be an open-label study where the treating physician believes one agent to be superior to another. It then is far more likely that any fever would be considered a treatment failure in a patient receiving the “inferior” therapy than in a patient receiving the “superior” treatment, where fever might be attributed to the slow resolution of disease or non-infectious complications such as allergy or thrombophlebitis. This form of bias is often taken advantage of in prospective commercial trials where a meeting is held of all investigators and lectures are provided detailing the safety or efficacy problems with the control agent to be used, and how these are avoided with the investigational agent. During the conduct of the study, even though assignment is blinded, the individual’s treatment is known to the treatment team.

In a proper prospective analysis, this potential source of bias is avoided by blinding all parties to the treatment administered: The treating team of nurses, doctors, and pharmacists; the patient; and the individuals tasked with collecting data and assessing outcome. Blinding is recognized less frequently as improving compliance and retention of trial participants and reducing biased supplemental care or treat-

ment (sometimes called “co-intervention”) [6]. An additional reality of clinical practice is that there is a real placebo effect that applies even to effective therapy. By blinding a trial, that extra benefit—which the writer surmises reflects a transfer of optimism—is removed. The treatment effect observed is then the *minimum* that would be seen in clinical practice.

The bias of investigators for or against the interventions can be transferred directly to participants by investigator attitudes. Their inclinations also may lead to differential use of supplemental care or treatment (co-interventions). Investigator bias also could encourage or discourage continuation in the trial on the basis of knowledge of the intervention group assignment [6].

More subjective outcomes present greater opportunities for bias. Pain scores assessed by participants are a good

TABLE 2. STRENGTH OF RECOMMENDATION AND QUALITY OF EVIDENCE

Category/Grade	Definition
Strength of recommendation	
A	Good evidence to support a recommendation for <i>or against</i> use
B	Moderate evidence to support a recommendation for <i>or against</i> use
C	Poor evidence to support a recommendation
Quality of evidence	
I	From ≥ 1 properly randomized, controlled trial
II	From ≥ 1 well-designed clinical trial, without randomization; from cohort or case-controlled analytic studies (preferably from >1 center); from multiple time series; or from dramatic results from uncontrolled experiments
III	From opinions of respected authorities based on clinical experience, descriptive studies, or reports of expert committees

example of a subjective outcome. Even some outcomes considered objective can be fraught with subjectivity—for example, soft tissue infection with cellulitis [6]. For this reason, objectifying such parameters, for example, by on-site photography, is recommended.

Blinding becomes less important to reduce observer bias as the outcomes become less subjective. “Hard” outcomes leave little room for bias. For example, knowledge of the intervention would have little effect on measuring a “hard” outcome, such as death (but still could influence the attributed cause of death). Of importance, even when participants and investigators have not been blinded, blinding of outcome assessors is often possible and advisable [6].

Inadequate Blinding Leads to Bigger Estimates of Effect

There are multiple steps encompassed by the term “blinding.” The first is allocation concealment; no one knows what the next eligible patient will receive. Allocation concealment prevents those who admit patients to a trial from knowing the upcoming assignments. Although not double-blinding appears to introduce bias, its average effect—exaggerating estimates by about 19% [7]—appears weaker than that of allocation concealment. Trials with inadequate or unclear allocation concealment yield larger estimates of treatment effects than those that use adequate concealment (on average, 41% and 33%, respectively), with commensurate results when the inadequate and unclear categories are lumped [7,8]. Double-blinding appears to be important in preventing bias, but not as important as allocation concealment.

Blinding also refers to masking the nature of the treatment being given during the trial. How big are the statistical effects of inadequate blinding? Schulz et al. assessed the methodological quality of 250 controlled trials from 33 meta-analyses and then analyzed the associations between those assessments and estimated treatment effects [9]. Compared with trials in which authors reported adequately concealed treatment allocation, trials in which concealment was either inadequate or unclear yielded larger estimates of treatment effects ($p < 0.001$). Adjusted for other aspects of quality, odds ratios were exaggerated by 41% for inadequately concealed trials and by 30% for unclearly concealed trials. Trials that were not double-blind also yielded larger estimates of effect ($p = 0.01$), with odds ratios being exaggerated by 17%.

“GRADEing” the Quality of a Trial: The Current Rage

Various tools have been developed to assist in the methodologic qualitative assessment of clinical trials. The GRADE system provides a framework for evaluating the utility of a specific study (Table 3) [10]. This system states that the quality of evidence for each main outcome can be determined after considering each of the elements: Study design, study quality, consistency, and directness.

Study design has to do with patient selection criteria, the statistical background of the study, dosing regimens, and other factors. This can be assessed simply by reading the protocol. Study quality has to do with how well the investigators complied with the study design. Were the specified patients enrolled? Were all patients eligible for the trial enrolled? Directness refers to the extent to which the people,

interventions, and outcome measures are similar to those of interest. For example, there may be uncertainty about the directness of the evidence if the people of interest are older, sicker, or have more co-morbidity than those in the studies.

The GRADE was developed to assist in meta-analysis by allowing an estimate of the quality of the study (Table 3). This and other evaluation systems can, nonetheless, be used to assess the quality of individual studies.

A more direct system, which provides a clear numeric score for a single clinical trial, is that devised by van Nieuwenhoven et al. [11]. They developed a clear analytic tool, which is reproduced in modified form in Table 4. They examined the role of methodologic quality in trials to determine the role of selective digestive decontamination in preventing pneumonia in patients in the intensive care unit. Their analysis demonstrated an inverse relation between the methodological quality of selective digestive decontamination (SDD) studies and the observed effects on the incidence of nosocomial pneumonia: The higher the quality score, the smaller the relative risk benefit of SDD in preventing pneumonia. No association was found for death. The trial quality characteristics associated with the outcome of pneumonia were patient selection, allocation of the intervention, and blinding.

To determine whether important uncertainty exists, it can be asked whether there is a compelling reason to expect important differences in the size of the effect. In the specific area of antimicrobial therapy for intra-abdominal infection, there are believed to be important differences in effect size in different patient groups. Patients with perforated or abscessed appendicitis benefit greatly from antibiotic therapy, with an odds ratio of 3 vs. 6 (mortality rates with antibiotics/without) [12]. However, patients with severe illness (in intensive care units) with postoperative infection are believed to

TABLE 3. CRITERIA FOR ASSIGNING GRADE OF EVIDENCE UTILIZING THE GRADE SYSTEM

Type of evidence
Randomized trial = high
Observational study = low
Any other evidence = very low
Decrease grade if:
• Serious (⊥1) or very serious (⊥2) limitation to study quality
• Important inconsistency (⊥1)
• Some (⊥1) or major (⊥2) uncertainty about directness
• Imprecise or sparse data (⊥1)
• High probability of reporting bias (⊥1)
Increase grade if:
• Strong evidence of association—significant relative risk of >2 (<0.5) based on consistent evidence from two or more observational studies, with no plausible confounders (+1) ₄₆
• Very strong evidence of association—significant relative risk of 5 (<0.2) based on direct evidence with no major threats to validity (+2) ₄₆
• Evidence of a dose response gradient (+1)
• All plausible confounders would have reduced the effect (+1)

TABLE 4. CRITERIA FOR ASSESSMENT OF METHODOLOGICAL QUALITY

Score	Criterion
	Population
	<i>Patient selection</i>
2	Consecutive eligible consenting patients
1	Attempt made to enroll as such, with failure outlined explicitly
0	Selected patients or not described
	<i>Patient characteristics</i>
2, Groups comparable on ≥ 6 characteristics	Age (means differ $<10\%$)
1, Groups comparable on 3 to 5 characteristics	APACHE, TISS, or other validated rating scale score (means differ by $<10\%$)
0, Groups comparable on ≤ 2 characteristics	<i>Diagnoses (proportion with the following differing by $<10\%$)</i>
	Respiratory failure (mechanical ventilation)
	Sepsis syndrome
	Renal failure
	Central nervous system disease
	Trauma
	Coagulopathy
	Hepatic failure
	Major operation
	Peptic ulcer disease, gastrointestinal tract surgery
	Diabetes mellitus or alcoholism
	Intervention
	<i>Allocation sequence</i>
2	Random allocation sequence
1	No information
0	Quasi-randomization (hospital number, date)
	<i>Concealment of allocation</i>
2	Non-manipulable
1	Potentially manipulable (sealed envelope) or randomization stated with no further information
0	Open-label
	<i>Blinding</i>
2	Blinding of patients, health care team, data collectors, and assessors to treatment group
1	Blinding of less than all of these groups
0	Potentially unblinded, unblinded, or cannot tell
	Outcome definitions
3	Gram stain or cultures of postoperative intra-abdominal fluid collection or surgical site infection diagnosed by CDC criteria
1	Need for additional antimicrobial therapy based on clinical evidence of fever or leukocytosis
0	Provision of additional antimicrobial therapy or prolonged LOS

APACHE = Acute Physiology and Chronic Health Evaluation; CDC = U.S. Centers for Disease Control and Prevention; LOS = length of stay; TISS = Therapeutic Intervention Scoring System.

have a lesser effect size because of confounding acute disease (e.g., acute lung injury, renal failure) and because adequate source control may be difficult to obtain. This is of great importance because the increasing incidence of appendicitis in more recent clinical trials is at the point where these trials provide little examination of antimicrobial agents in critical illness. Indeed, assessment of the effect and recommendations for therapy in severely ill patients are extrapolations of microbiologic activity and pharmacokinetics.

Evaluation of Data from Observational Trials

Non-randomized studies include experimental studies (such as quasi-randomized trials) and observational studies with controls (such as controlled before–after studies, concurrent cohort studies, and case–control studies) or without concurrent controls (such as before–after studies, cross-

sectional studies, and case series) [13]. Most published articles (68% to 87% of the feature articles and brief communications in *Annals of Internal Medicine*, *BMJ*, and *The New England Journal of Medicine*) are non-randomized [14]. Historically, there has been little randomized-trial evidence in the areas of devices and procedures, and it is estimated that randomized trials account for less than 10% of the evidence base for surgical interventions [15].

These reports suffer greatly from lack of patient matching, other unnoted or undocumented changes in care practices that affect outcome, and considerable risk of bias in outcome assessment, particularly if subjective or indirect criteria are used. The central problem of such studies, particularly if done as a time series (“we compared patients treated in one time interval to those in another”), is the effect of changes in other care items unnoted in the analysis. A simple example would be the apparent effect of adding screening for

methicillin-resistant *Staphylococcus aureus* and mupirocin decontamination for preoperative patients. If done as a time series, it is highly probable that the screening and decontamination were begun as part of a program or bundle of activities that likely would have included measures such as vancomycin prophylaxis. Such studies therefore establish only temporal associations, not cause-effect relations. The concept of “bundling” several interventions and examining before-and-after data at least recognizes these problems, but cannot account for the beneficial effect of simply studying a problem.

A major stated concern regards the ethics of randomizing patients to therapies not believed to be highly effective. The counter response has to do with sacrificing large numbers of future patients to inadequate or unsafe treatment, absent information that could have been obtained in a small group of patients. Ethics is a moving target, and current Institutional Review Board practices effectively prevent trials that pose any risk of systemic harm.

Recently, an extensive comparison of the results of randomized vs. observational trials (often referred to as “quasi-experimental”—enough said) was reported by MacLehose et al. [13]. They found that most quasi-experiments they reviewed were of poor quality. They believed that a high priority should be given to the development of standards for reporting quasi-experimental and observational studies. Enforcement of such standards, in the long term, might be expected to improve the standard of the research as well as the reporting.

There is a need to develop methods for identifying studies that provide a direct comparison of estimates from randomized and non-randomized data. A register should be established and studies entered in the register as they are identified. There also is a need for innovative search strategies.

I believe it is possible to use properly done observational trials as a basis for decision-making if these quasi-experimental studies adhere to certain basic precepts. These are rather well captured in the system presented in Table 4 (please note the outcome variables will depend on the disease under study). A hidden benefit of focusing on the methodologic quality of observational trials in guidelines is to increase the gain individual researchers get from performing higher quality studies.

What Was Learned from the SIS/IDSA Process?

The cIAI guidelines are the result of a laborious consensus development process, and, if consensus could not be reached, a painful search for wording that briefly outlined the minority position. The panel met on four occasions, three times via teleconference and once face to face, to complete the guidelines. The meetings were held to discuss the questions to be addressed, make writing assignments, and discuss recommendations. There was a large volume of e-mail comments as drafts were regularly circulated electronically. All panel members participated in the preparation and review of the draft guideline.

Feedback from external peer reviewers was obtained. The guideline was reviewed and endorsed by the Pediatric Infectious Diseases Society, American Society for Microbiology, the American Society of HealthSystem Pharmacists, and the Society of Infectious Disease Pharmacists. The guideline was finally reviewed and approved by the IDSA Standards and Practice Guidelines Committee (SPGC), the IDSA Board of

Directors, the SIS Therapeutic Agents Committee, and the SIS Executive Council prior to publication.

So what does this elaborate review process add to the document the authors created? Probably believability. What emerged through the review process, at both the individual and the organizational level, was simply a look at how committees function. The process was, at the end of the day, impacted heavily by individuals with specific issues. It was noteworthy that few reviewers disagreed with the grading given specific recommendations.

The panel was keenly aware through this process that recommendations against specific common current practices could well cause difficulties for the practitioners involved. These are particular problems with what could be termed “legacy practices,” such as aminoglycoside usage in community-acquired infections. This becomes particularly problematic when the quality of the evidence is not strong, or when it is based on surrogate markers such as microbiologic activity of alternative agents or efficacy in other patient groups. Can we extrapolate appendicitis efficacy identified in a non-inferiority study design to septic shock from colon-derived diffuse peritonitis? Can non-inferiority data ever result in a A-I recommendation?

Conflict of Interest

One of the more common concerns raised since these guidelines were published is the possibility that recommendations might have been biased by conflicts of interests. The main focus of this concern has been a table describing agents useful for community-acquired agents of mild-to-moderate severity infection (appendicitis primarily).

My own response to these comments has been that the guidelines are very clear on this: “The efficacy and cost advantages of generic agents are noted . . .” The table detailing available agents is not entitled, “Recommended Agents,” and it lists agents available for use and approved by the U.S. Food and Drug Administration for this indication. Further concerns are expressed about the potentially negative effects of ertapenem, tigecycline, and moxifloxacin on stewardship efforts.

I am strongly in favor of disclosures detailing the dollar amount of payments to individuals by source. This information is becoming more available through the publication of such payments on pharmaceutical websites.

But ultimately, there is no scoring system for bias. Further, bias can be subtle. Claims that multiple conflicts mean there are none is humorous but misleading. The conflict may not be favoring a specific agent but rather being exposed to commercial relationships masquerading as personal relationships, and human responses to the latter. This is only one of the biasing pressures encountered in advisory boards and other interactions. Another problem is that being immersed in antimicrobial issues may lead to an altered view of the role of other factors, particularly source control, in taking care of these patients. We worked hard to expand the realm of these guidelines to encompass these items.

Review by another set of individuals focused on appearance of bias is not the answer. The review process now is too complex and consumes too much time. Another layer, with the requirements for recycling of the new “final” draft, would stretch the process out so much that guidelines would never be current.

Having individuals writing guidelines without industrial contacts is unwise. Because of the complex interactions of clinical research, education, and practice, it is unlikely that such individuals would be sufficiently versed in the science of this therapeutic area without interacting with pharmaceutical companies. This is perhaps one of the most important areas of further thinking and even research in the area of therapeutic guidelines development.

Author Disclosure Statement

No conflicting financial interests exist.

References

- Solomkin JS, Mazuski JE, Bradley JS, et al. Diagnosis and management of complicated intra-abdominal infection in adults and children: guidelines by the Surgical Infection Society and the Infectious Diseases Society of America. *Surg Infect (Larchmt)* 2010;11:79–109.
- Solomkin JS, Mazuski JE, Bradley JS, et al. Diagnosis and management of complicated intra-abdominal infection in adults and children: guidelines by the Surgical Infection Society and the Infectious Diseases Society of America. *Clin Infect Dis* 2010;50:133–164.
- Mazuski JE, Sawyer RG, Nathens AB, et al. The Surgical Infection Society guidelines on antimicrobial therapy for intra-abdominal infections: An executive summary. *Surg Infect* 2002;3:161–174.
- Solomkin JS, Mazuski JE, Baron EJ, et al. Guidelines for the selection of anti-infective agents for complicated intra-abdominal infections. *Clin Infect Dis* 2003;37:997–1005.
- Canadian Task Force on the Periodic Health Examination. The periodic health examination. *Can Med Assoc J* 1979;121:1193–1254.
- Schulz KF, Chalmers I, Altman DG. The landscape and lexicon of blinding in randomized trials. *Ann Intern Med* 2002;136:254–259.
- Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ* 2001;323:42–46.
- Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998;352:609–613.
- Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408–412.
- Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. *BMJ* 2004;328:1490.
- van Nieuwenhoven CA, Buskens E, van Tiel FH, Bonten MJ. Relationship between methodological trial quality and the effects of selective digestive decontamination on pneumonia and mortality in critically ill patients. *JAMA* 2001;286:335–340.
- Barnes BA, Behringer GE, Wheelock FC, Wilkins EW. Treatment of appendicitis at the Massachusetts General Hospital (1937–1959). *JAMA* 1962;180:122–126.
- MacLehose RR, Reeves BC, Harvey IM, et al. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technol Assess* 2000;4:1–154.
- Ray JG. Evidence in upheaval: Incorporating observational data into clinical practice. *Arch Intern Med* 2002;162:249–254.
- McCulloch P, Taylor I, Sasako M, et al. Randomised trials in surgery: Problems and possible solutions. *BMJ* 2002;324:1448–1451.

Address correspondence to:

Dr. Joseph S. Solomkin
Department of Surgery (ML 0558)
University of Cincinnati College of Medicine
231 Albert Sabin Way
Cincinnati, OH 45267-0558

E-mail: joseph.solomkin@uc.edu