A New and Rapid Scoring System to Assess the Scientific Evidence from Clinical Trials

SIGMUND SILBER, M.D., F.E.S.C., F.A.C.C.

From the Cardiology Practice and Hospital, Munich, Germany

Guidelines are based on a scientific analysis from existing data of randomized controlled trials (RCTs), registry trials, simple registries, case reports, and the personal experience of the task force members. Furthermore, meta-analyses and subgroup analyses are used to derive the strengths of recommendations. Fortunately, the major cardiac societies, i.e., the American College of Cardiology (ACC), the American Heart Association (AHA), and the European Society of Cardiology (ESC), are essentially using the same definitions for the levels of recommendations. In the expanding field of cardiology, however, the overwhelming and increasing number of clinical studies reveals the limitations of the traditional ranking of these studies: Applying the standard definitions of the ACC/AHA/ESC criteria for the levels of recommendation, almost every PCI procedure would easily reach the level of evidence A, even with two small, underpowered studies and a surrogate endpoint. Although meta-analyses are important tools for creating an overview of major diagnostic or treatment modalities, they are bound to severe limitations. The compilation of several underpowered, small trials can generate a statistically artificial "significant" result. This is especially important because only two meta-analyses containing almost identical studies could easily yield an evidence level A. RCTs are usually designed and conducted according to a power calculation, for which a primary clinical or surrogate endpoint can be chosen. Surrogate endpoints, however, do not necessarily correlate with the clinical outcome. The history of medicine is full of errors introduced by underpowered studies with surrogate endpoints. Many investigators and companies attempt to tease out treatment effects in specific subpopulations of patients. These subgroup analyses are usually underpowered. Another major limitation of the ACC/AHA/ESC scoring system is that neither the power of a study nor the choice of a primary clinical endpoint is included in their definitions. Yet another limitation of the ACC/AHA/ESC grading system is that two "simple" registries may already lead to a level of evidence B. A new scoring system is presented addressing most of these limitations: a primary clinical endpoint receives three points, whereas all of the following receive one point: double-blind design, evaluation interval of primary endpoint $\geq$ 6 months, multicenter (at least three centers), independent data and safety monitoring, power of $\geq$ 80% for primary endpoint achieved, and follow-up $\geq$ 80% for a surrogate primary endpoint or follow-up of $\geq$ 95% for a clinical primary endpoint. Thus, the maximum achievable points is 10. This scoring system can also be applied for high-quality registry controlled trials using a predefined control group and power calculation. For simple registry studies and subgroup analyses, a modified scoring system has been developed (maximum achievable points is 5). The advantage of the suggested new scoring system is its transparency, reproducibility, and ease of use by quickly answering the key quality questions for clinical trials. The new scoring system suggested here should help make decisions regarding which treatment to use and stimulate discussions.
(J Interven Cardiol 2006;19:485–492)

Introduction

One of the first formal methods for the evaluation of the quality of controlled randomized clinical trials (RCTs, RaCTs) was developed in 1981 by the late Thomas Chalmers.[1] As a pioneer in this field, T. Chalmers developed a system to evaluate the design, implementation, and analysis of RaCTs with emphasis on the quadruple blinding (the randomization process, the physicians and patients as to therapy, and the physicians as to ongoing results), which he considered to be the most important aspect of any trial. Numerous other

Address for reprints: Sigmund Silber, M.D., F.A.C.C., F.E.S.C., Professor of Medicine, Cardiology Practice and Hospital, Am Isarkanal 36, 81379 Munich, Germany. Fax: +49-89-74215131; e-mail: sigmund@silber.com

scales and checklists have since been suggested to evaluate RaCT quality.[2–4]

Guidelines for diagnostic and treatment modalities are increasingly created and used in most fields of medical care. Guidelines are not legally binding; they are recommendations of what one could or should do—but not what one must do. However, if a problem arises, one might justify not having followed the guidelines. Guidelines are the result of an analysis of existing data but refer to only those clinical situations that have been investigated in clinical trials. Therefore, guidelines do not replace medical experience.

Guidelines are based on a scientific analysis of existing data from RaCTs, registry controlled trials (ReCTs), simple registries, case reports, and the personal experience of the task force members. Furthermore, meta-analyses and subgroup analyses are used to derive the strengths of recommendations. Fortunately, the major cardiac societies, i.e., the American College of Cardiology (ACC), the American Heart Association (AHA), and the European Society of Cardiology (ESC), are essentially using the same definitions for the levels of recommendations. In the expanding field of cardiology, however, the overwhelming and increasing number of clinical studies reveals the limitations of the traditional ranking of these studies. Applying the standard definitions of the ACC/AHA/ESC criteria for the levels of recommendation, almost every PCI procedure would easily reach the level of evidence A, even with two small, underpowered studies and a surrogate endpoint.

The following article therefore discusses the limitations of the traditional ranking of recommendations and suggests a new scoring system to rapidly check each individual study and make transparent and reproducible its key parameters of scientific evidence.

The Definition of the ACC/AHA/ESC Levels of Recommendations

The U.S.A.[5] and European[6] guidelines are based on a combination of the defined classes of recommendations (I, II, III) and the levels of evidence (A, B, C) as listed in Tables 1–3. Whereas the definitions of the classes of recommendations are almost identical between the United States and Europe (Table 1), there are minor differences in the definitions of the levels of evidence: level of evidence B requires "large" nonrandomized studies in Europe, whereas any nonrandomized studies are included in the United States. Other minor differences for level of evidence C are listed in Table 2.

Table 1. Classification of Recommendations as Defined in the ACC/AHA[5] and in the ESC[6] Guidelines

Class I	*ACC/AHA:* Conditions for which there is evidence for and/or general agreement that a given procedure or treatment is beneficial, useful, and effective. *ESC:* Evidence and/or general agreement that a given treatment or procedure is beneficial, useful, and effective.
Class II	*ACC/AHA:* Conditions for which there is conflicting evidence and/or a divergence of opinion about the usefulness/efficacy of a procedure or treatment. *ESC:* Conflicting evidence and/or divergence of opinion about the usefulness/efficacy of the treatment or procedure.
Class IIa	*ACC/AHA/ESC:* Weight of evidence/opinion is in favor of usefulness/efficacy.
Class IIb	*ACC/AHA/ESC:* Usefulness/efficacy is less well established by evidence/opinion.
Class III	*ACC/AHA:* Conditions for which there is evidence and/or general agreement that a procedure/treatment is not useful/effective and in some cases may be harmful. *ESC:* Evidence or general agreement that the treatment or procedure is not useful or effective and in some cases may be harmful.

The Limitations of the ACC/AHA/ESC Levels of Recommendations

It may be somewhat confusing that the evidence derived from a study appears in both categories, i.e., in the class of recommendation as well as in the level of evidence (Tables 1 and 2). The strength of evidence derived from a study should be accounted for exclusively in the level of evidence. Thus, the classes of recommendations should be based predominantly on the agreement of the opinion of the task force members: a unanimous agreement between the task force members would lead to a class I recommendation; if the majority of the task force members is in favor of a treatment, the recommendation would be a class IIa and if only a minority of task force members is in favor a class IIb. This would make the final, combined levels of recommendations (Table 3) more transparent.[7]

Table 2. Levels of Evidence as Defined in the ACC/AHA[5] and in the ESC[6] Guidelines

Level of evidence A	*ACC/AHA/ESC:* Data derived from multiple randomized clinical trials or meta-analyses.
Level of evidence B	*ACC/AHA:* Data derived from a single randomized trial or nonrandomized studies. *ESC:* Data derived from a single randomized clinical trial or large nonrandomized studies.
Level of evidence C	*ACC/AHA:* Only consensus opinion of experts, case studies, or standard-of-care. *ESC:* Consensus of opinion of the experts and/or small studies, retrospective studies, registries.

The Problem with a "Class III" Recommendation. The class III recommendation contains two basically different situations: If a procedure is not effective, it at least does not harm the patient. On the other hand, if a procedure is harmful, it must be avoided. The "recommendation not to recommend" is contradictory. Instead, one should simply not use the term "class III" anymore.[7] If a treatment/procedure is not effective, it can be so described in the guidelines—but without a class of recommendation. If a treatment/procedure is harmful, it can be separately listed under "warnings and contraindications"—without a class of recommendation. This is especially important when a study was prematurely terminated for safety reasons.[8]

Reproducibility of the Classification System. It is obvious that the unclear ACC/AHA/ESC definitions of classes of recommendations and levels of evidence leave room for different gradings of the same studies and procedures by different task forces: Despite the almost identical definitions and analyzing the same studies, different levels of recommendation between the ACC/AHA and the ESC are not unusual. For example, the use of abciximab for PCI in patients with ST-segment elevation myocardial infarction is at level IIa B in the ACC/AHA guidelines[9] but at level IIa A in the ESC guidelines.[7] On the other hand, drug-eluting stents (DES) "should be considered as an alternative to the bare-metal stent in subsets of patients in whom trial data suggest efficacy" at level I A in the ACC/AHA guidelines,[5] whereas DES were graded level I B for the same indication in the ESC guidelines.[7]

Table 3. The 12 Possibilities of the Final Levels of Recommendations Are Based on the Combination of the Classes of Recommendations and the Corresponding Levels of Evidence

	Level of Evidence A	Level of Evidence B	Level of Evidence C
Class I	I A	I B	I C
Class IIa	IIa A	IIa B	IIa C
Class IIb	IIb A	IIb B	IIb C
Class III	III A	III B	III C

The Problem of Meta-analyses. Meta-analyses are important tools for creating an overview of major diagnostic or treatment modalities; however, their severe limitations should be kept in mind.

The compilation of several underpowered, small trials can generate a statistically artificial "significant" result—although still underpowered. This is especially important, because only two meta-analyses containing almost identical studies could easily yield a level of evidence A (Table 2). Furthermore, different studies use different definitions for the same parameter: e.g., major cardiac events (MACE) in SIRIUS[10] included any death and target lesion revascularization (TLR), whereas MACE in TAXUS-IV[11] included only death from cardiac causes (which is lower than any death) and target vessel revascularization (TVR, which is usually higher than TLR). Similar differences in definitions of "stent thrombosis" can also be found. In addition to the angiographic confirmation, the TAXUS studies classified any Q-wave or non-Q-wave myocardial infarction (related to the target vessel territory) as "stent thrombosis", whereas the SIRIUS studies included only Q-wave infarctions related to the target vessel territory. In addition, any unexplained sudden death within 30 days after the procedure was considered a "stent thrombosis" in the TAXUS and SIRIUS programs, whereas in the ENDEAVOR program, any cardiac death within 30 days was interpreted as a "stent thrombosis". This meta-analysis "mishmash" of different studies with different inclusion/exclusion criteria using different definitions and different standards of external reviewing boards must be interpreted with caution.[12] Studies of low methodological quality in which the estimate of quality is incorporated into the meta-analyses can alter the interpretation of the benefit of intervention.[13] Furthermore, meta-analyses should not contain studies which are not yet fully published, especially if the orally presented data are different from the data printed in the abstract.[14] Patients are better served by clinicians

waiting at least for the evaluation of potential scientific advances by rigorous peer review.[15]

Meta-analyses are important for creating hypotheses—which should be confirmed by a subsequent RaCT—but should not serve as a tool to automatically receive a level of evidence A.

The Importance of "Power" and the Problem of Subgroup Analyses. Although some RaCTs have been published without a power calculation,[16] RaCTs are usually designed and conducted according to a power calculation. This power calculation results in a required minimum number of patients to be enrolled (sample size) in order to prove or disprove the respective hypothesis. The hypothesis can be proven according to a superiority trial design (e.g., DES superior to bare stents[10,11,17]) or according to a noninferiority trial design (e.g., DES-1 is not inferior to DES-2[18–20]). Noninferiority trials (= equivalency trials) involve various problems: In contrast to superiority trials, equivalency trials define an "acceptable difference" to the reference treatment effect, the so-called delta. This "delta" can be freely chosen by the investigators. To save money, studies could be conducted with a low number of patients, showing "statistically significant equivalency" of a new DES as compared to a standard DES. This "outcome drift" may imply a nonexisting equivalency between the established standard DES and the newly developed DES. Furthermore, the question whether a negative noninferiority study has proven a superiority (DES-1 is "not noninferior" to DES-2) is a matter of controversy.[18,19]

One of the major limitations of the ACC/AHA/ESC scoring system is that the power of a study is not included in their definitions (Tables 1 and 2), especially since the power of a study is more important than just the number of enrolled patients: A "small" study can have more power than a "large" study. The power of a study may also get lost if the analysis switches from "intention to treat" to "per protocol." If a study was originally designed as an "intention to treat" trial, the "per protocol" analysis should be interpreted with caution.

Subgroup analyses are a frequently used tool to demonstrate all kinds of "effects." Many investigators and companies attempt to tease out treatment effects in specific subpopulations of patients. One must, however, always keep in mind that subgroup analyses are usually underpowered, because they contain a lower patient number than the total number of patients required to achieve the necessary power.[21] Reporting of subgroup analysis needs to be substantially improved because emphasis on these secondary results may mislead treatment decisions.[21]

Therefore, the results of subgroup analyses are often a matter of chance and should be taken with caution, particularly when not all subgroup analyses go in the same direction. Simple registries should not be used to "prove" a superiority of DES-1 over DES-2. Like meta-analyses, simple registries should be used to generate a hypothesis, which will be corroborated by subsequent randomized trials.

The Importance of the Primary Clinical Endpoint. The primary endpoint of an RaCT can be chosen as a clinical endpoint or a surrogate endpoint.[22]

Clinical endpoints are defined as clinical events, like MACE, major cardiac and cerebrovascular events (MACCE), target vessel failure (TVF), TVR, and target lesion revascularization (TLR). For MACE, MACCE, and TVF, the patients do not have to undergo another cardiac catheterization, whereas TVR and TLR require this. It is anticipated that TVR and TLR are clinically driven—but sometimes the differentiation between a clinically driven re-PCI and an angiographically driven ("oculostenotic reflex") re-PCI in studies with a per protocol routine follow-up angiography can be complicated: The impact of systematic repeat angiography on "clinically driven" re-PCI rates has been nicely shown with considerably lower TLR and TVR rates in the nonangiography subset as compared to the angiography subset.[17] It is easy to imagine that the difference between strict clinically driven and angiographically driven re-PCI will be even higher when physicians are not blinded to the treatment groups: While randomization is the accepted way to create treatment groups that are free from treatment-related selection bias, it alone does not protect RaCTs against other types of bias.[23]

Surrogate endpoints are usually quantitative coronary analyses (QCA) of the angiograms, like diameter stenosis (DS in %), restenosis rate (RR in %), minimum lumen diameter (MLD in mm) or late loss (LL in mm)/late loss index (LLI in %). Also intravascular ultrasound (IVUS[24]), TIMI-flow, and "myocardial blush" have served as surrogate parameters. A surrogate endpoint, however, is insufficient for documenting an improvement in clinical outcome: In DELIVER-I,[25] although late lumen loss was statistically significantly reduced, there was no significant benefit for the clinical outcome. A similar missing correlation between surrogate and clinical endpoints can be derived

from the SIRIUS,[10] TAXUS-IV,[11] and ENDEAVOR-II[17] studies. Although there were significant differences at 9 months in LL (0.17 mm/0.39 mm/0.62 mm), the TVF rates were almost identical (8.6%/7.6%/8.1%). Also, after 1 year there was no difference among these DES regarding the TVF (9.8%/10.0%/9.9%). Nevertheless, LL might be an easy-to-measure parameter, independent of vessel size, to initially evaluate newly developed DES by generating hypotheses to be corroborated in subsequent randomized trials.[26]

Furthermore, if a clinical parameter is chosen as a secondary endpoint and even if it is significantly improved at $P < 0.05$, this does not necessarily mean that the statistical proof is sufficient to document an improved clinical outcome: Since the power of such a study was calculated according to the primary (surrogate) endpoint, the power of the secondary (clinical) endpoint is insufficient ("underpowered").[27] Underpowered studies have been misleading medicine for a long time (ELITE-I vs ELITE-II; Vesnarinone vs VEST; PRAISE-I vs PRAISE-II). The major "advantage" of choosing a surrogate parameter as primary endpoint is that the required sample size is smaller, so the study is completed earlier and less expensively.

Independent of the choice of the primary endpoint, the monitoring of the events by an external, independent "clinical event committee" (CEC) and/or a data safety monitoring board (DSMB) is crucial for the quality of the study. The members of the CEC/DSMB should not participate as investigators and not be members of the steering committee. Full disclosure of possible financial conflicts of interest of all actively involved persons is required.[28]

Underpowered studies in combination with a primary surrogate endpoint have led to the controversy of "Cypher vs Taxus."[27] This controversy was fueled by studies which were not "real" multicenter studies or had a primary surrogate endpoint.[18,19,29,30] The only major multicenter "head-to-head" trial with an independent, external review committee was REALITY[31]—showing no difference between Cypher (TVF after 12 months: 12.0%) and Taxus (12.9%). Head-to-head trials proving the superiority of DES-1 over DES-2 should be designed as superiority trials with a primary clinical endpoint, an adequate power calculated and achieved, a "real" multicenter status and an external, independent DSMB. Such a trial does not exist. For the detection of differences in (late) stent thrombosis, more than 10,000 patients would probably be needed to detect statistically significant differences at a power of >80%. From the clinical point of view one could state that if such megatrials are necessary to show a difference, the difference is probably not clinically relevant.

Not All Registries Are Equal. As opposed to RaCTs, registries do not enroll their patients according to a randomized protocol and therefore do not have a parallel, simultaneous control group. Usually, they compare the results with a not-predefined historical "control" group of another study, eventually trying to post hoc match the patients. These "standard" simple registries intrinsically do not have a power calculation, no sample size calculation, and therefore do not have a primary endpoint. It thus makes more sense to call the goals of a simple registry a "primary objective" rather than a "primary endpoint."

One of the major advantages of registries is their usually much larger number of patients with no inclusion/exclusion criteria, so they might better reflect the "real world" of "all comers." On the other hand, monitoring and follow-up in simple registries is usually lower than in RaCTs. One of the limitations of the ACC/AHA/ESC grading system is that two "simple" registries may already lead to a level of evidence B (Table 2).

Recently, some more elaborate registries prespecified the "to be matched control group" using identical inclusion/exclusion criteria and a sample size based on an a priori power calculation. These "registry controlled trials" (ReCTs) with a power calculation and a primary endpoint have to be differentiated from the simple registries without a power calculation and therefore without a primary endpoint.

A New Scoring System to Evaluate Clinical Trials for Their Scientific Evidence

A new scoring system should address most of the above-discussed relevant limitations, being fully transparent and therefore reproducible, tailored to all different designs of clinical studies and not too complicated to apply. The concept of this new scoring system could also be applied in medical fields beyond PCI and even beyond cardiology.

Randomized Controlled Trials. For the reasons discussed above, the following eight questions have to be answered to evaluate the scientific evidence of an RaCT:

Table 4. A New Score for Randomized Controlled Trials (RaCT) or Registry Controlled Trials (ReCT)

Clinical primary endpoint (TLR, TVR, TVF, MACE)	Yes = 3 No = 0
Double-blind (including physicians)	Yes = 1 No = 0
Evaluation interval of primary endpoint ≥ 6 months	Yes = 1 No = 0
Multicenter (at least 3 centers)	Yes = 1 No = 0
Clinical events committee / data safety monitoring board independent from the steering committee	Yes = 1 No = 0
Primary endpoint reached	Yes = 1 No = 0
Power of ≥ 80% for primary endpoint achieved	Yes = 1 No = 0
Follow-up percentage ≥ 80% for surrogate primary endpoint or follow-up percentage of ≥ 95% for clinical primary endpoint	Yes = 1 No = 0
Maximum possible Silber score	10

An ReCT is a nonrandomized trial requiring a prespecified matched control with prespecified inclusion/exclusion criteria and a sample size based on a power calculation.

1. Was the primary endpoint a clinical or a surrogate parameter?
2. Were the physicians blinded to the modality of treatment?
3. Was the evaluation interval for the primary endpoint long enough?
4. Was it a "real" multicenter study?
5. Was the clinical events committee/DSMB external and independent from the steering committee?
6. Was the primary endpoint reached?
7. Was a power calculation performed and the desired power actually reached?
8. Was the follow-up rate of patients sufficient?

Table 4 depicts the suggested new scoring system answering these questions. The maximum achievable points is 10. The Cypher-stent (SIRIUS-trial[10]), the Taxus-stent (TAXUS-IV trial,[11]), and the Endeavor-stent (ENDEAVOR-II trial[17]) have received a score of 10 each, as published at the TCTMD Web site.[32]

Registry Controlled Trials. These high-standard registries (ReCTs) can be evaluated analogous to the RaCTs (Table 4). Examples for such ReCTs are ARTS-II[33] and Taxus-ATLAS, each one having received a score of 8.[32]

Table 5. A New Score for "Simple Registries" (as Opposed to Registry Controlled Trials) or Subgroup Analyses from Randomized Controlled Trials

Data prospectively collected	Yes = 1 No = 0
Subgroup analysis from a randomized, controlled trial	Yes = 1 No = 0
Multicenter (at least 3 centers)	Yes = 1 No = 0
Clinical events committee / data safety monitoring board independent from steering committee	Yes = 1 No = 0
Monitoring ≥ 10% and follow-up percentage ≥ 90%	Yes = 1 No = 0
Maximum possible Silber score	5

Simple Registry Studies and Subgroup Analyses. As opposed to RaCTs and ReCTs, "simple" registries do not have power/sample size calculation and therefore no predefined calculation of an endpoint. The missing power calculation is also true for subgroup analyses of RaCTs. Therefore, the scoring system as in Table 4 cannot be applied under these circumstances. Nevertheless, for "simple" registries and subgroup analyses, there are also criteria to evaluate their scientific robustness, as in the following questions:

1. Were the data prospectively or retrospectively collected?
2. Were the data derived from a subgroup analysis of a randomized controlled trial?
3. Was it a "real" multicenter study?
4. Was the clinical events committee/DSMB external and independent from the steering committee?
5. Was there an adequate monitoring and follow-up rate?

Table 5 depicts the suggested new scoring system answering these questions. The maximum achievable points is 5. An "adequate" monitoring for registries is considered if at least 10% of the patients have been randomly controlled. Since registries are usually clinically oriented, a follow-up rate of at least 90% should be possible. Subgroup analyses of RaCTs receive an extra point because of the strong monitoring and follow-up procedures of RaCTs. Since a tremendous number of simple registries and subgroup analyses with many updates are presented at many meetings, they should only be graded after their final publication in a peer-reviewed journal.

Limitations of the New Scoring System. The choice of the factor 3 for a primary clinical endpoint

is arbitrary. But as discussed above, clinical studies with a primary clinical endpoint have the highest evidence to prove or disprove a given procedure or treatment. Surrogate measurements may not necessarily reflect differences in clinical outcome. Therefore, studies with a primary clinical endpoint and their inherent higher number of needed patients should be rewarded accordingly.

Another criticism might be that the score is too simplistic. "Between measurements based on randomized controled trials and benefit. . .in the community there is a gulf which has been much underestimated" (A. L. Cochrane, 1971, quoted from Ref. 34). Many determinants are needed for external validity in the design and reporting of RCTs.[34] While the presented score could certainly be expanded to include more evaluation parameters, that would inhibit its easy application. Some of the questions related to the quality of a clinical study that were not included in the scoring system are as follows: Was the possible financial conflict of interest of the investigators fully disclosed? What are the economic consequences of the study? Will society be able to afford the new treatment options? Although these aspects are important, they were not included in the score because they go beyond the "pure" scientific evidence. Guidelines are usually based on the "pure" scientific evidence[7]—only in a few exceptions are they based on cost-effectiveness analyses.[35] Cost-effectiveness analyses are very dependent on regional circumstances and complexity of lesions as has been shown for the United States[36] and Switzerland.[37]

The advantages of the suggested new scoring system are their transparency, reproducibility, and ease of use by quickly answering the key quality questions for RaCTs, ReCTs, and simple registries. The score is intended to help practicing cardiologists rapidly evaluate the strength of evidence from the many DES: In Europe, 13 different DES are currently CE certified and commercially available. The CE mark, however, is not at all a proof of safety and efficacy.[38] Therefore, the new scoring system suggested here should help decide which DES to use and stimulate discussions.

Acknowledgment: The author fully acknowledges the help of Debra Lynn Beck, TCTMD editorial staff.

References

1. Chalmers TC, Smith H Jr, Blackburn B, et al. A method for assessing the quality of a randomized control trial. Control Clin Trials 1981;2:31–49.
2. Moher D, Jadad AR, Nichol G, et al. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. Control Clin Trials 1995;16:62–73.
3. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Control Clin Trials 1996;17:1–12.
4. Davidoff F, Haynes B, Sackett D, et al. Evidence-based medicine. BMJ 1995;310:1085–1086.
5. Smith SC Jr, Feldman TE, Hirshfeld JW Jr, et al. ACC/AHA/SCAI 2005 Guideline Update for Percutaneous Coronary Intervention-Summary Article. A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (ACC/AHA/SCAI Writing Committee to Update the 2001 Guidelines for Percutaneous Coronary Intervention). J Am Coll Cardiol 2006;47:216–235.
6. ESC. Recommendations for Guidelines Production. http://www.escardio.org 2006.
7. Silber S, Albertsson P, Aviles FF, et al. Guidelines for Percutaneous Coronary Interventions: The task force for percutaneous coronary interventions of the European Society of Cardiology. Eur Heart J 2005;26:804–847.
8. Ioannidis JP, Evans SJ, Gotzsche PC, et al. Better reporting of harms in randomized trials: An extension of the CONSORT statement. Ann Intern Med 2004;141:781–788.
9. Antman EM, Anbe DT, Armstrong PW, et al. ACC/AHA guidelines for the management of patients with ST-elevation myocardial infarction—executive summary: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Revise the 1999 Guidelines for the Management of Patients With Acute Myocardial Infarction). Circulation 2004;110:588–636.
10. Moses JW, Leon MB, Popma JJ, et al. N Engl J Med 2003;349:1315–1323.
11. Stone GW, Ellis SG, Cox DA, et al. A polymer-based, paclitaxel-eluting stent in patients with coronary artery disease. N Engl J Med 2004;350:221–231.
12. Jadad AR, Cook DJ, Jones A, et al. Methodology and reports of systematic reviews and meta-analyses: A comparison of Cochrane reviews with articles published in paper-based journals. JAMA 1998;280:278–280.
13. 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.
14. Kastrati A, Dibra A, Eberle S, et al. Sirolimus-eluting stents vs paclitaxel-eluting stents in patients with coronary artery disease: Meta-analysis of randomized trials. JAMA 2005;294:819–825.
15. Falagas ME, Rosmarakis ES. Clinical decision-making based on findings presented in conference abstracts: Is it safe for our patients? Eur Heart J 2006;27:2038–2039.
16. Goy JJ, Stauffer JC, Siegenthaler M, et al. A prospective randomized comparison between paclitaxel and sirolimus stents in the real world of interventional cardiology: The TAXi trial. J Am Coll Cardiol 2005;45:308–311.
17. Fajadet J, Wijns W, Laarman GJ, et al. Randomized, double-blind, multicenter study of the Endeavor zotarolimus-eluting phosphorylcholine-encapsulated stent for treatment of native coronary artery lesions: Clinical and angiographic results of the ENDEAVOR II trial. Circulation 2006;114:798–806.

18. Dibra A, Kastrati A, Mehilli J, et al. Paclitaxel-eluting or sirolimus-eluting stents to prevent restenosis in diabetic patients. N Engl J Med 2005;353:663–670.
19. Mehilli J, Dibra A, Kastrati A, et al. Randomized trial of paclitaxel- and sirolimus-eluting stents in small coronary vessels. Eur Heart J 2006;27:260–266.
20. Mehilli J, Kastrati A, Wessely R, et al. Randomized trial of a nonpolymer-based rapamycin-eluting stent versus a polymer-based paclitaxel-eluting stent for the reduction of late lumen loss. Circulation 2006;113:273–279.
21. Hernandez AV, Boersma E, Murray GD, et al. Subgroup analyses in therapeutic cardiovascular clinical trials: Are most of them misleading? Am Heart J 2006;151:257–264.
22. Silber S. Which parameter should be chosen as primary endpoint for randomized drug-eluting stent studies? J Interv Cardiol 2004;17:375–385.
23. Jadad AR, Rennie D. The randomized controlled trial gets a middle-aged checkup. JAMA 1998;279:319–320.
24. Colombo A, Drzewiecki J, Banning A, et al. Randomized study to assess the effectiveness of slow- and moderate-release polymer-based paclitaxel-eluting stents for coronary artery lesions. Circulation 2003;108:788–794.
25. Lansky AJ, Costa RA, Mintz GS, et al. Non-polymer-based paclitaxel-coated coronary stents for the treatment of patients with de novo coronary lesions: Angiographic follow-up of the DELIVER clinical trial. Circulation 2004;109:1948–1954.
26. Mauri L, Orav EJ, Candia SC, et al. Robustness of late lumen loss in discriminating drug-eluting stents across variable observational and randomized trials. Circulation 2005;112:2833–2839.
27. Silber S. Cypher versus taxus: Are there differences? J Interven Cardiol 2005;18:441–446.
28. Choudhry NK, Stelfox HT, Detsky AS. Relationships between authors of clinical practice guidelines and the pharmaceutical industry. JAMA 2002;287:612–617.
29. Windecker S, Remondino A, Eberli FR, et al. Sirolimus-eluting and paclitaxel-eluting stents for coronary revascularization. N Engl J Med 2005;353:653–662.
30. Silber S, Albertsson P, Aviles FF, et al. Reply: Guidelines for percutaneous coronary interventions. Letter to the Editor. Eur Heart J 2006;27:1757–1759.
31. Morice MC, Colombo A, Meier B, et al. Sirolimus- vs paclitaxel-eluting stents in de novo coronary artery lesions: The REALITY trial: A randomized controlled trial. JAMA 2006;295:895–904.
32. Silber S. The Evidence-Based Medicine Center (EBMC). http://www.tctmd.com 2006.
33. Serruys P, Ong AT, Morice MC, et al. Arterial Revascularisation Therapies Study Part II—Sirolimus-eluting stents for the treatment of patients with multivessel de novo coronary artery lesions. EuroInterv 2005;1:147–156.
34. Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply?" Lancet 2005;365:82–93.
35. NICE (National Institute for Clinical Excellence). Coronary artery stents (No. 71), (replacing Drug-eluting stents No. 4). http://www.nice.org.uk 2004.
36. Cohen DJ, Bakhai A, Shi C, et al. Cost-effectiveness of sirolimus-eluting stents for treatment of complex coronary stenoses: Results from the Sirolimus-Eluting Balloon Expandable Stent in the Treatment of Patients With De Novo Native Coronary Artery Lesions (SIRIUS) trial. Circulation 2004;110:508–514.
37. Kaiser C, Brunner-La Rocca HP, Buser PT, et al. Incremental cost-effectiveness of drug-eluting stents compared with a third-generation bare-metal stent in a real-world setting: Randomised Basel Stent Kosten Effektivitats Trial (BASKET). Lancet 2005;366:921–929.
38. Kastrati A, Schömig A, Dirschinger J, et al. Increased risk of restenosis after placement of gold-coated stents: Results of a randomized trial comparing gold-coated with uncoated steel stents in patients with coronary artery disease. Circulation 2000;101:2478–2483.