How Evidence-Based Are the Recommendations in Evidence-Based Guidelines?

Finlay A. McAlister[1*], Sean van Diepen[1], Rajdeep S. Padwal[1], Jeffrey A. Johnson[2,3], Sumit R. Majumdar[1]

1 The Division of General Internal Medicine, University of Alberta, Edmonton, Canada, **2** The Institute of Health Economics, University of Alberta, Edmonton, Canada, **3** The School of Public Health, University of Alberta, Edmonton, Canada

Funding: This study was completed without external funding support, and none of the salary-support funders for any of the authors had input into the design or conduct of the study; collection, management, analysis, or interpretation of the data; nor preparation, review, approval, or decision to submit the manuscript for publication.

Competing Interests: FAM is co-chair of the Central Review Committee for the Canadian Hypertension Education Program and RSP is a member of the Canadian Hypertension Society Education Program Central Review Commitee.

Academic Editor: Alessandro Liberati, Italian Cochrane Centre, Italy

Citation: McAlister FA, van Diepen S, Padwal RS, Johnson JA, Majumdar SR (2007) How evidence-based are the recommendations in evidence-based guidelines? PLoS Med 4(8): e250. doi:10.1371/journal.pmed.0040250

Received: February 27, 2007
Accepted: June 20, 2007
Published: August 7, 2007

Abbreviations: AGREE, Appraisal of Guidelines Research and Evaluation; CHEP, Canadian Hypertension Education Program; GRADE, Grading of Recommendations Assessment, Development and Evaluation; RCT, randomized controlled trial

* To whom correspondence should be addressed. E-mail: Finlay.McAlister@ualberta.ca

ABSTRACT

Background

Treatment recommendations for the same condition from different guideline bodies often disagree, even when the same randomized controlled trial (RCT) evidence is cited. Guideline appraisal tools focus on methodology and quality of reporting, but not on the nature of the supporting evidence. This study was done to evaluate the quality of the evidence (based on consideration of its internal validity, clinical relevance, and applicability) underlying therapy recommendations in evidence-based clinical practice guidelines.

Methods and Findings

A cross-sectional analysis of cardiovascular risk management recommendations was performed for three different conditions (diabetes mellitus, dyslipidemia, and hypertension) from three pan-national guideline panels (from the United States, Canada, and Europe). Of the 338 treatment recommendations in these nine guidelines, 231 (68%) cited RCT evidence but only 105 (45%) of these RCT-based recommendations were based on high-quality evidence. RCT-based evidence was downgraded most often because of reservations about the applicability of the RCT to the populations specified in the guideline recommendation (64/126 cases, 51%) or because the RCT reported surrogate outcomes (59/126 cases, 47%).

Conclusions

The results of internally valid RCTs may not be applicable to the populations, interventions, or outcomes specified in a guideline recommendation and therefore should not always be assumed to provide high-quality evidence for therapy recommendations.

The Editors' Summary of this article follows the references.

Introduction

There has been a rapid expansion in the number of clinical practice guidelines over the past decade and, as a result, clinicians are frequently faced with several guidelines for treatment of the same condition. Unfortunately, recommendations may differ between guidelines [1,2], leaving the clinician with a decision to make about which guideline to follow. While it is easy to say that one should follow only those guidelines that are "evidence based," very few guideline developers declare their documents to be non–evidence based, and there is ambiguity about what "evidence based" really means in the context of guidelines. The term may be interpreted differently depending on who is referring to the guideline—the developer, who creates the guidelines, or the clinician, who uses them. To their developers, "evidence-based guidelines" are defined as those that incorporate a systematic search for evidence, explicitly evaluate the quality of that evidence, and then espouse recommendations based on the best available evidence, even when that evidence is not high quality [3]. However, to clinicians, "evidence based" is frequently misinterpreted as meaning that the recommendations are based solely on high-quality evidence (i.e., randomized clinical trials [RCTs]) [4]. Previous studies of guidelines have focused almost exclusively on the elements embodied in the first definition of an evidence-based guideline. For example, guideline appraisal tools assess the methodology used in developing the guideline and the clarity with which recommendations and the type of underlying evidence are communicated in that guideline [5,6].

However, few studies have addressed the issue raised in the second interpretation of an evidence-based guideline—that is, the quality of the evidence underpinning "evidence-based" guidelines. Given the widespread availability of electronic databases to search the literature, one would expect that evidence-based guidelines would usually cite the same evidence. However, an analysis of 15 guidelines for type 2 diabetes mellitus revealed little overlap—only ten studies (less than 1% of all citations) were cited in at least six of these guidelines, and the most frequently cited study in these guidelines (the Diabetes Complication Control Trial, referenced in 11 of 15 guidelines) was conducted exclusively in patients without type 2 diabetes mellitus [7].

How then should the quality of evidence underlying recommendations be evaluated? In those guidelines that use explicit scales, virtually all are based solely on considerations of study design and internal validity [8]. For example, RCTs are graded higher than observational studies irrespective of sample size, study conduct, endpoints evaluated, or the applicability (i.e., generalizability) of the RCT to the populations, interventions, and outcomes specified in the guideline recommendation. In order to incorporate external validity, applicability, and clinical relevance into evidence appraisal, we used an evidence grading scheme that has been used (and refined) for almost a decade by the Canadian Hypertension Education Program (CHEP) [9]. The CHEP scheme evaluates three (type of study, internal validity, and directness) of the four domains recommended by the Grading of Recommendations Assessment, Development and Evaluation [GRADE] working group (http://www.gradeworkinggroup.org/, last accessed 29 January 2007)—consistency between studies is not explicitly evaluated in the CHEP scheme (except in the assessment of meta-analyses) and, instead, CHEP places primacy on the study achieving the highest evidence grading for a particular recommendation [3].

We designed this study to evaluate the quality of the evidence cited for cardiovascular risk management recommendations in evidence-based clinical practice guidelines.

Methods

Selection of Guidelines

Based on the prevalence of these conditions in the outpatient primary care setting and our collective areas of clinical expertise, we restricted ourselves to national guidelines for the management of diabetes mellitus, dyslipidemia, and hypertension. We a priori chose the most recent guidelines from the United States, Canada, and Europe for each disorder, as these are the national-level guidelines Canadian clinicians are most commonly exposed to [10–19]. We defined recommendations as any statements that advocated a specific intervention for application in clinical care. We focused on evaluating the evidence base for cardiovascular risk management interventions in these guidelines, and did not examine the evidence base underlying recommendations on diagnosis, monitoring, or prevention. Moreover, as our interest was on the chronic treatment of these conditions, we a priori excluded recommendations for pregnant, hospitalized, or peri-operative individuals.

Quality Appraisal of the Guidelines

We appraised the quality of each guideline using the standardized Appraisal of Guidelines Research and Evaluation (AGREE) instrument (http://www.agreetrust.org/, last accessed 29 January 2007), which consists of 23 items that rate the various dimensions of quality for the overall guideline using four-point Likert scales. The AGREE instrument is organized into six independent domains, and the score for each domain is calculated as a percentage of the maximum possible score for that domain [20]. Three investigators independently completed the AGREE instrument for each guideline, and discrepancies were resolved by consensus.

It should be noted that the AGREE appraisal is based on consideration of the whole guideline (i.e., not for specific recommendations within each guideline) and includes any recommendations on diagnosis, monitoring, or prevention. In contrast, our detailed analysis of the evidence underlying recommendations focused solely on therapy recommendations.

Extraction of Recommendations and Underlying Evidence

After completing a training set to calibrate evidence ratings, two investigators with expertise in both critical appraisal and the topic areas of the chosen guidelines reviewed each guideline, independently extracted recommendations and the references cited by each guideline in support of each cardiovascular therapy recommendation, and graded the quality of the evidence as described in detail below. We did not conduct literature searches and used only the studies cited directly in the references list of each guideline to evaluate the evidence base for that recommendation. All evidence ratings for each cardiovascular therapy recommendation were independently checked by a third

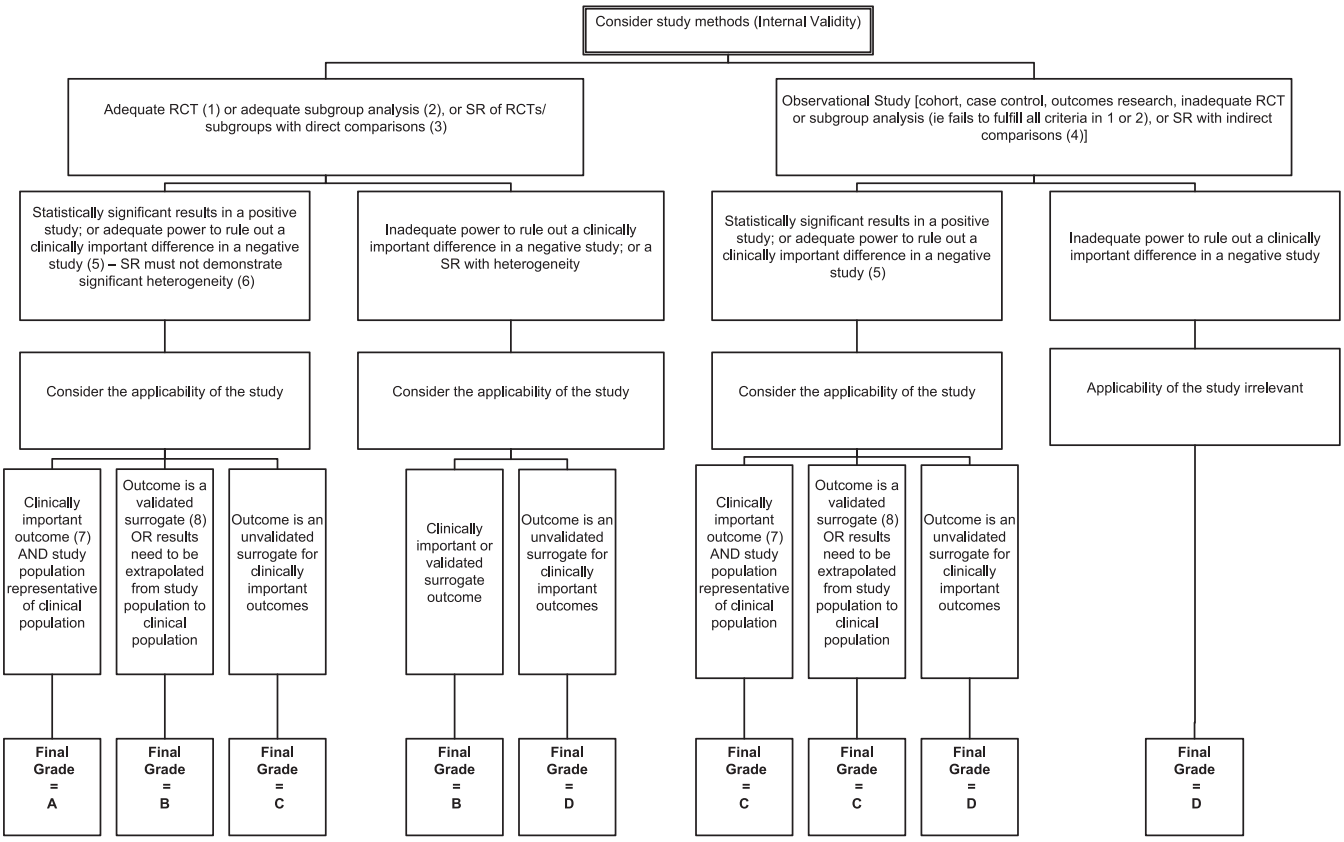

Figure 1. The Evidence-Grading Scheme Employed in this Study

Adapted from CHEP [9].

An "adequate" RCT is one with allocation concealment, blinded assessment of outcomes (if subjective), intention-to-treat analysis, adequate follow-up (i.e., at least 90%, or losses to follow-up are too few to materially affect the results), and sufficient sample size to detect a clinically important difference with power $>$ 80% (1). Subgroup analysis was a priori, done within an adequate RCT, one of only a few tested, and there was sufficient sample size within the examined subgroup to detect a clinically important difference with power $>$ 80% (2). A sytematic review (SR) with direct comparisons is one in which the comparison arms are derived from head-to-head comparisons within the same RCT (3). A systematic review with indirect comparisons is one in which the comparison arms are derived from different RCTs, then extrapolations are made across RCTs (4). Adequate power in a negative study implies that the 95% confidence interval excludes a clinically important difference (5). Effect estimates in each study included in the systematic review are qualitatively similar (i.e., in the same direction) (6). Clinically important outcomes are "hard" endpoints such as death, stroke, or myocardial infarction (7). End points have been consistently shown to be associated with the clinical end point in multiple studies (observational or RCT), and RCTs have consistently demonstrated that improvement in the surrogate translates into a consistent and predictable improvement in the clinical end point (8).

doi:10.1371/journal.pmed.0040250.g001

reviewer (FAM), and any discrepancies were resolved by consensus.

Grading Quality of the Underlying Evidence

We appraised the quality of evidence cited to support each cardiovascular risk management recommendation using the CHEP evidence-grading scheme (Figure 1). The CHEP grading scheme was developed in 1999, has been used and refined in the seven annual updates of the Canadian hypertension guidelines since then, and explicitly operationalizes many elements of the GRADE scheme using a priori standardized rules of evidence (based on reference [21]) in a fashion that has been shown to be reproducible within and between reviewers trained in critical appraisal [9]. As we were interested in exploring how many cardiovascular therapy recommendations considered to be high quality in current guidelines would have their evidence grading reduced when factors beyond study design were taken into account, we focused specifically on those recommendations citing RCTs or systematic reviews of RCTs. In situations where several RCTs or systematic reviews were cited in support of a particular recommendation, we based our evidence grade on the RCT or systematic review achieving the highest ranking for that recommendation.

Pooled across all nine guidelines, inter-rater agreement on whether a recommendation was based on RCT evidence or not was 91% (kappa 0.80) and inter-rater agreement on whether RCT evidence was high quality or not (grade A versus B, C, or D on the CHEP scheme) was 89% (kappa 0.78).

Results

Characteristics of the Guidelines and Recommendations

The nine guidelines in our study (Table 1) [10–19] ranged in size from ten pages to 284 pages and cited between 44 and 1,121 references. They provided a total of 1,005 recommendations for diagnosis and prevention ($n = 362$) or treatment ($n = 643$) for patients with diabetes mellitus, dyslipidemia, or hypertension; 369 of the treatment recommendations in these guidelines advocated cardiovascular risk management therapies. Although only four of the guidelines [10,11,14,17] provided grades for their underlying evidence, seven guide-

Table 1. Description of Guidelines and Quality of Evidence

Medical Condition	Organization	Year	AGREE Instrument Rigor of Development Domain Score, %	Cardiovascular Risk Management Therapy Recommendations, *n*	Cardiovascular Therapy Recommendations Based on RCT Evidence, *n* (%)	Cardiovascular Therapy Recommendations Based on High-Quality (Grade A) Evidence[a], *n* (%)
Diabetes mellitus	Canadian Diabetes Association [10]	2003	52%	44	29 (66%)	13 (30%)
	American Diabetes Association[b] [11]	2005	43%	47	27 (57%)	10 (21%)
	International Diabetes Federation [12]	2005	19%	26	No direct link between recommendations and evidence	Not applicable
Lipids	The Canadian Working Group on Hypercholesterolemia and Other Dyslipidemias [13]	2003	10%	14	8 (57%)	4 (29%)
	National Cholesterol Education Panel [14]	2002	57%	42	34 (81%)	14 (33%)
	European Society of Cardiology[c] [15]	2003	24%	5	No direct link between recommendations and evidence	Not applicable
Hypertension	Canadian Hypertension Education Program [16,17]	2006	86%	95	67 (71%)	32 (34%)
	Joint National Committee [18]	2003	57%	52	31 (60%)	14 (27%)
	European Society of Hypertension [19]	2003	24%	44	35 (80%)	18 (41%)

[a]High-quality evidence is defined as Grade A evidence in the CHEP evidence grading scheme (Figure 1).
[b]For the American Diabetes Association guidelines we included only recommendations from the Standards of Medical Care Section.
[c]For the European Society of Cardiology guidelines on cardiovascular disease prevention, we extracted only recommendations dealing with diagnosis and management of dyslipidemia.
doi:10.1371/journal.pmed.0040250.t001

Figure 2. Summary of AGREE Domain Scores for Guidelines, Averaged over Each Condition
doi:10.1371/journal.pmed.0040250.g002

lines explicitly linked recommendations to evidence (either within the text of the recommendation itself or at the end of relevant sections of the guideline text); these 338 recommendations formed the sample for this study.

Quality of the Guidelines

The standardized scores for each domain of the AGREE instrument (averaged across guidelines within the same topic areas) are outlined in Figure 2. While these guidelines were scored highly for the clarity of their presentations and their descriptions of "scope and purpose" and were judged to have reasonable levels of editorial independence, they scored poorly on the "stakeholder involvement" (due particularly to lack of involvement of patient groups and prepiloting of the guidelines) and "applicability" domains of the AGREE instrument. Within the applicability domain, these guidelines infrequently discussed resource implications, audit criteria, or potential organizational changes necessary to implement recommended therapies.

Quality of the Underlying Evidence

Of the 338 recommendations that endorsed particular cardiovascular risk management therapies and cited evidence, 231 (68%) cited an RCT or systematic review of RCTs in support of that recommendation (Table 1). Of these 231 RCT-based recommendations, 105 (45%) received a grade A using the CHEP grading scheme (corresponding to "high-quality evidence" in the GRADE scheme), with the proportion ranging between 37% and 51% in different guidelines (Figure 3). Thus, only 28% (range 21% to 41% between guidelines) of the 369 cardiovascular risk management recommendations in these nine prominent national evidence-based guidelines were directly supported by high-quality evidence (Table 1).

The most frequent reason for downgrading RCT-based therapy recommendations (64 [51%] of the 126 cases) were concerns about the need to extrapolate from a highly selected RCT population to the scenario and/or the target population specified in the guideline. As an illustration, consider the differences in quality of evidence underlying two antihypertensive recommendations that were both based on RCT

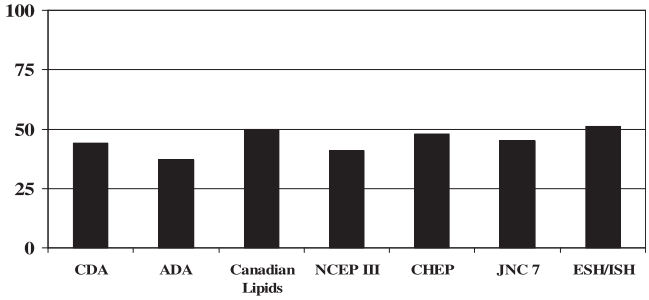

Figure 3. Percentage of Recommendations Citing RCTs That Were Based on High-Quality Evidence in Each Guideline

Legend and references: ADA, American Diabetes Association [11]; Canadian Lipids, the Canadian Working Group on Hypercholesterolemia and Other Dyslipidemias [13]; CDA, Canadian Diabetes Association [10]; CHEP, the Canadian Hypertension Education Program [16,17]; ESH/ISH, European Society of Hypertension/International Society of Hypertension [19]; JNC 7, Joint National Committee seventh report [18]; NCEP III, Third Report of the National Cholesterol Education Program [14].
doi:10.1371/journal.pmed.0040250.g003

evidence. The first recommendation was for the use of thiazides in individuals with hypertension. It cited a meta-analysis [22] of 30 RCTs that proved conclusively the efficacy of thiazides in reducing cardiovascular events and mortality in over 70,000 individuals with hypertension—this recommendation was judged to be supported by high-quality (grade A) evidence. In contrast, the second recommendation was for the use of spironolactone for individuals with hypertension and cited an RCT [23] of 1,663 patients with advanced heart failure in which baseline blood pressures were normotensive (mean 122/75 mm Hg), and results for the hypertensive subgroup were not reported separately. Although this spironolactone recommendation was based on RCT evidence, its relevance and applicability to the general hypertensive population is sufficiently uncertain that we classified it as grade B (moderate quality) evidence at best. As an aside, it should be noted that RCT evidence was not downgraded if the RCT was conducted primarily in men but the recommendation referred to people in general, or if the RCT was conducted in a country other than that from which the guideline arose. The second most common reason for downgrading RCT-based recommendations (59 cases, 47%) were concerns about the clinical relevance of the RCT—for example, the RCT reported the effect of the recommended therapy on surrogate outcomes only (e.g., levels of glucose, low-density lipoprotein cholesterol, or blood pressure) rather than patient-centered outcomes such as death, myocardial infarction, or stroke. Illustrative examples of RCT-based recommendations that were downgraded are provided in Table 2.

The quality of the evidence base cited for cardiovascular risk management therapies was not related to guideline length (Pearson correlation coefficient $[r] = 0.13$, $p = 0.78$), number of references ($r = 0.20$, $p = 0.67$), or score on the "rigor of development" domain of the AGREE instrument ($r = -0.03$, $p = 0.96$).

Discussion

In summary, we found that while two-thirds of cardiovascular risk management therapy recommendations made in the nine different guidelines we examined were based on RCT evidence, less than half of these RCT-based recommendations were deemed "high quality" using an evidence-grading scheme that went beyond considerations of internal validity alone to take into account clinical relevance and direct applicability of the RCT to that recommendation. As a result, less than one-third of recommendations that advocated specific cardiovascular risk management therapies in these evidence-based guidelines were actually based on high-quality evidence.

The most frequent reason for RCT-based recommendations to be down-graded was that the RCT was conducted to answer a particular question in a restricted study population but was then extrapolated in the guideline to justify using the tested intervention in a related, but different, clinical scenario and/or in a more general population. In a similar vein, other investigators have recently questioned whether the evidence cited in the Third Report of the National Cholesterol Education Program, or NCEP III, as support for recommendations to use statins for primary prevention of cardiovascular disease is directly applicable, since one-tenth of the patients in the 16 "primary prevention" trials cited in that guideline had cerebrovascular or peripheral vascular disease at baseline [24].

As a corollary, it is evident that while a particular RCT may be used as the basis for multiple recommendations, RCTs will not provide the same quality of evidence for each recommendation (and in some cases guideline developers may extrapolate beyond the limits of the evidence in making particular recommendations). For example, the 2003 Kidney Disease Outcome Quality Initiative guidelines [25] recommended statins for all patients with chronic kidney disease and LDL $>$ 2.59 mmol/l, including those with end-stage renal disease, on the basis of RCTs such as the Heart Protection Study which were positive, but excluded patients with end-stage renal disease [26]. However, a recently published RCT conducted in 1255 hemodialysis patients with type 2 diabetes mellitus found no reduction in the primary outcome of cardiovascular events or death but instead an unexpected increase in the risk of stroke with statin therapy [27].

We do not mean to imply that recommendations should not be made in the absence of high quality evidence or that RCT evidence should not be extrapolated beyond the limits of trial eligibility criteria. Indeed, we recognize that trialists design RCTs with relatively homogenous populations in order to maximize internal validity (at the expense of external validity), and there are published guides on how and when to extrapolate RCT evidence to individual patient situations [28]. However, we do believe that transparency about any extrapolation of RCT evidence is critical, particularly in light of studies demonstrating that the composition and interpersonal dynamics of a guideline panel influence the extent to which their consensus recommendations diverge from the available evidence base [29–31].

Our findings that only some guidelines linked their recommendations to citations and that only some used explicit grading systems to communicate the quality of the evidence echo earlier reviews [7,32–34]. Similarly, our finding that many treatment recommendations are not based on RCT evidence has been reported before in other fields [35]. However, our unique finding is that even those recommendations in evidence-based guidelines that cite internally valid RCTs as support may not be underpinned by high-quality

Table 2. Examples of Guideline Recommendations That Cited an Internally Valid Randomized Trial Which Did Not Provide High-Quality Evidence for That Recommendation

Reason for Downgrade	Recommendation	Evidence Cited	CHEP Grade	Rationale for Downgrade
RCT evidence was downgraded because of concerns about applicability of the RCT results to the population specified in the guideline recommendation	Aldosterone antagonists for the treatment of patients with hypertension and concomitant heart failure	RALES [23]	B	Recommendation required extrapolation from an RCT of 1,663 patients with advanced heart failure in which baseline blood pressures were normotensive (mean 122/75 mm Hg) and results for the hypertensive subgroup were not reported separately
	Fibrates for primary prevention in patients with diabetes mellitus and low-density lipoprotein cholesterol below target but elevated fasting triglycerides	VA-HIT [40]	B	Recommendation required extrapolation from a secondary prevention RCT in 2,531 men with coronary heart disease (all 627 individuals with diabetes mellitus in the trial had coronary heart disease at baseline)
RCT evidence was downgraded because of concerns about clinical relevance (i.e., primary outcome was a surrogate outcome)	DASH diet for all hypertensive individuals	Sacks, et al. [41]	B	Statistically significant results, but primary outcome in this 412-patient RCT was change in blood pressure (a validated surrogate outcome)
	ACE inhibitor plus angiotensin receptor blocker as dual therapy for patients with hypertension and advanced renal disease	COOPERATE [42]	D	Statistically significant results, but primary outcome in this 263 patient RCT was "doubling of serum creatinine or development of end stage renal failure" (an unvalidated surrogate outcome)
RCT evidence was downgraded because recommendation was based on post hoc subgroup results	Aspirin for primary prevention of cardiovascular events in patients with diabetes mellitus	Antithrombotic Trialists Collaboration [43]	B	Although the pooled estimate for 144,051 patients in 195 RCTs demonstrated a statistically significant 22% reduction in vascular events, the data in the 5,116 patients with diabetes mellitus and no vascular disease at baseline was inconclusive (OR 0.93, 95% CI 0.80–1.08)
	Angiotensin receptor blockers for the treatment of patients with isolated systolic hypertension	LIFE secondary publication [44]	B	Randomization was not stratified within the subgroup of 1,326 patients with isolated systolic hypertension; all patients in this trial also had left ventricular hypertrophy on ECG; losartan reduced the composite cardiovascular endpoint compared to atenolol, but difference was statistically significant only in adjusted analysis
RCT evidence was downgraded because the RCT had inadequate power to rule out a clinically important difference in a negative study	Beta-blockers for the treatment of hypertension in patients with diabetes mellitus unable to tolerate ACE inhibitor or angiotensin receptor blocker	UKPDS-39 [45]	B	Although there was no statistically significant difference for the primary outcome with atenolol versus captropril in this 758-patient trial, this comparison was underpowered (134 deaths [RR 1.27, 95% CI 0.82–1.97], 107 myocardial infarctions [RR 1.20, 95% CI 0.82–1.76], and 259 "any diabetes related endpoints" [RR 1.10, 95% CI 0.86–1.41])

ACE, angiotensin converting enzyme; CI, confidence interval; DASH, Dietary Approaches to Stop Hypertension; ECG, electrocardiogram; OR, odds ratio; RR, relative risk.
doi:10.1371/journal.pmed.0040250.t002

evidence. We did not compare the countries of origin for cited studies, because previous studies have already established that local evidence tends to be over-represented in guidelines [7,36,37]. While the guidelines we studied were published at different times, the range was narrow (2003–2006) and our study focused on the type of evidence cited by each guideline rather than by the specific recommendations made.

Despite a number of strengths, our study has some limitations. First, we did not systematically search for different guidelines, but instead examined only a small sample of guidelines; future studies should expand our work to explore the quality of evidence underlying guidelines in other topic areas. For example, we believe a systematic examination of all guidelines produced by a particular organization (or, alternatively, all published guidelines in a particular topic area) would provide useful additional insights. To do so, we advocate the use of explicit grading schemes such as those of CHEP or GRADE (as in the recently reported framework for World Health Organization Rapid Advice Guidelines [38]). However, while we found a high degree of inter-rater reliability for assessing whether RCT evidence was high quality or not (kappa 0.78 after all investigators completed a training set), future studies should also assess the inter-rater reliability scores for the GRADE or CHEP schemes if used by investigators less familiar with the schemes. Second, because the guidelines were inconsistent in how they cited studies in support of therapy recommendations (with some providing the citation directly with the recommendation and others providing numerous citations at the end of supporting text associated with recommendations), there is a potential risk we may have misattributed citations to particular recommendations. We attempted to minimize this risk by having two investigators extract recommendations and citations independently for each guideline and by always biasing in favor of the guideline (i.e., if several citations were attached to a

recommendation, we assigned the highest evidence rating achieved by any of the studies to that recommendation). However, future researchers may want to consider prospectively surveying guideline developers to determine exactly which pieces of evidence are considered for each recommendation; also, documentation of the debate around the evidence for particular recommendations in different overlapping guidelines would provide potentially interesting insights. Finally, although our choice to restrict our analysis to cardiovascular therapy recommendations may be perceived as a limitation, it in fact strengthens our conclusions since therapy recommendations are those most likely to be based on RCT evidence. Thus, our findings represent a "best-case" scenario, insofar as very few preventive or diagnostic guideline recommendations are based on RCT evidence.

In conclusion, our finding that less than one-third of treatment recommendations (and less than half of those citing RCTs in support of the advocated treatment) were based on high-quality evidence in national evidence-based guidelines for common conditions should sound a note of caution to consumers of clinical practice guidelines who assume that the sobriquet "evidence based" means that all recommendations contained therein are derived from high-quality evidence. In particular, we have documented that even evidence arising from internally valid RCTs may not be directly applicable to the populations, interventions, and outcomes specified in a guideline recommendation. As a recent editorial noted, "external validity is the neglected dimension in evidence ranking" [39]. Indeed, in order to make the evidence base underlying therapy recommendations more transparent in future guidelines, we advocate wider adoption of evidence-rating schemes (such as the CHEP system or the GRADE system) that go beyond just judging the internal validity of supporting evidence but also incorporate considerations of the clinical relevance and applicability of that evidence to the clinical scenario the recommendation is being made for. A clearer understanding of the strengths and limitations of the underlying evidence base will then permit clinicians to individualize the application of practice guideline recommendations to their patients.

Acknowledgments

FAM, JAJ, and SRM are supported by the Alberta Heritage Foundation for Medical Research and the Canadian Institutes of Health Research (CIHR). FAM is also supported by the Merck Frosst/Aventis Chair in Patient Health Management at the University of Alberta. JAJ is supported by the CIHR through the Canada Research Chairs Program.

Author contributions. FAM had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. FAM conceived the study. FAM and SvD designed the study. FAM, SvD, RSP, JAJ, and SRM were involved in the acquisition of data and analysis/interpretation of the data. FAM wrote the first draft of the manuscript; FAM, SvD, RSP, JAJ, and SRM participated in revising the manuscript. FAM, SvD, RSP, JAJ, and SRM have approved submission of the final draft.

References

1. McMurray J, Swedberg K (2006) Treatment of chronic heart failure: A comparison between the major guidelines. Eur Heart J 27: 1773–1777.
2. McAlister FA, Campbell NRC, Zarnke K, Levine M, Graham I (2001) The management of hypertension in Canada: A review of current guidelines, their shortcomings, and implications for the future. CMAJ 164: 517–522.
3. GRADE Working Group (2004) Grading quality of evidence and strength of recommendations. BMJ 328: 1490.
4. Grol R, Dalhuijsen J, Thomas S, in't Veld C, Rutten G, et al. (1998) Attributes of clinical guidelines that influence use of guidelines in general practice: Observational study. BMJ 317: 858–861.
5. Burgers JS (2006) Guideline quality and guideline content: Are they related? Clin Chem 52: 3–4.
6. Vlayen J, Aertgeerts B, Hannes K, Sermeus W, Ramaekers D (2005) A systematic review of appraisal tools for clinical practice guidelines: Multiple similarities and one common deficit. Int J Qual Health Care 17: 235–242.
7. Burgers JS, Vailey JV, Klazinga NS, van der Bij AK, Grol R, et al. (2002) Inside guidelines. Comparative analysis of recommendations and evidence in diabetes guidelines from 13 countries. Diabetes Care 25: 1933–1939.
8. Atkins D, Eccles M, Flottorp S, Guyatt GH, Henry D, et al. (2004) Systems for grading the quality of evidence and the strength of recommendations I: Critical appraisal of existing approaches. BMC Health Services Research 4: 38.
9. McAlister FA (2006) The Canadian Hypertension Education Program (CHEP)—A Unique Canadian Initiative. Can J Cardiol 22: 559–564.
10. Harris SB, Capes SE, Lillie D, Lank CN, Mahon J, et al. (2003) Canadian Diabetes Association 2003 clinical practice guidelines for the prevention and management of diabetes in Canada. Can J Diabetes 27: S1–S152.
11. American Diabetes Association (2005) Standards of medical care in diabetes. Diabetes Care 28: S4–36.
12. International Diabetes Federation (2005) Clinical guidelines task force: Global guidelines for type 2 diabetes. International Diabetes Federation website. Available: http://www.idf.org/home/index.cfm?node=1457. Accessed 11 December 2006.
13. Genest J, Frohlich J, Fodor G, McPherson R (2003) Recommendations for the management of dyslipidemia and the prevention of cardiovascular disease. CMAJ 169: 921–924.
14. Grundy SM, Becker D, Clark LT, Cooper RS, Denke MA, et al. (2001) Executive summary of the third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults. National Heart Lung and Blood Institute Website. Available: http://www.nhlbi.nih.gov. Accessed 11 December 2006.
15. De Backer G, Ambrosioni E, Borch-Johnsen K, Brotons C, Cifkova R, et al. (2003) European guidelines on cardiovascular disease prevention in clinical practice: Third joint task force of European and other societies on cardiovascular disease prevention in clinical practice. Eur Heart J 24: 1601–1610.
16. Hemmelgarn B, McAlister FA, Grover S, Myers MG, McKay DW, et al. (2006) The 2006 Canadian Hypertension Education Program (CHEP) recommendations for the management of hypertension: Part I, Blood pressure measurement, diagnosis, and assessment of risk. Can J Cardiol 22: 573–581.
17. Khan NA, McAlister FA, Rabkin SW, Padwal R, Feldman RD, et al. (2006) The 2006 Canadian Hypertension Education Program (CHEP) recommendations for the management of hypertension: Part 2, Therapy. Can J Cardiol 22: 583–593.
18. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, et al. (2003) The seventh report of the Joint National Committee on prevention, detection, evaluation, and treatment of high blood pressure. The JNC 7 Report. JAMA 289: 2560–2572.
19. Guidelines Committee (2003) 2003 European Society of Hypertension–European Society of Cardiology guidelines for the management of arterial hypertension. J Hypertens 21: 1011–1053.
20. Cluzeau FA, Littlejohns P, Grimshaw JM, Feder G, Moran SE (1999) Development and application of a generic methodology to assess the quality of clinical guidelines. Int J Qual Health Care 11: 21–28.
21. Guyatt GH, Rennie Deditors (2002) Users' guides to the medical literature. A manual for evidence-based clinical practice. Chicago: American Medical Association. 705 p.
22. Psaty BM, Lumley T, Furberg CD, Schellenbaum G, Pahor M, et al. (2003) Health outcomes associated with various antihypertensive therapies used as first-line agents: A network meta-analysis. JAMA 289: 2534–2544.
23. Pitt B, Zannad F, Remme WJ, Cody R, Castaigne A, et al. (1999) The effect of spironolactone on morbidity and mortality in patients with severe heart failure. N Engl J Med 341: 709–717.
24. Abramson J, Wright JM (2007) Are lipid-lowering guidelines evidence-based? Lancet 369: 168–169.
25. Eknoyan G, Levin A, Levin N, Kasiske B, Cosio FG, et al. for the National Kidney Foundation (2003) K/DOQI clinical practice guidelines for managing dyslipidemias in chronic kidney disease. Am J Kidney Dis 41: S1–S91.
26. Heart Protection Study Collaborative Group (2002) MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20 536 high-risk individuals: A randomised placebo-controlled trial. Lancet 360: 7–22.
27. Wanner C, Krane V, Marz W, Olschewski M, Mann JF, et al. (2005) Atorvastatin in patients with type 2 diabetes mellitus undergoing hemodialysis. N Engl J Med 353: 238–248.
28. Rothwell PM (2005) External validity of randomised controlled trials: "To whom do the results of this trial apply?" Lancet 365: 82–93.
29. Burgers JS, van Everdingen JJE (2004) Beyond the evidence in clinical guidelines. Lancet 364: 392–393.
30. Raine R, Sanderson C, Hutchings A, Carter S, Larkin K, et al. (2004) An

experimental study of determinants of group judgements in clinical guideline development. Lancet 364: 429–437.
31. Pagliari C, Grimshaw J, Eccles M (2001) The potential influence of small group processes on guideline development. J Eval Clin Pract 7: 165–173.
32. Shaneyfelt TM, Mayo-Smith MF, Rothwangl J (1999) Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature. JAMA 281: 1900–1905.
33. Graham ID, Beardall S, Carter AO, Glennie J, Hebert P, et al. (2001) What is the quality of drug therapy clinical practice guidelines in Canada? CMAJ 165: 157–163.
34. Grilli R, Magrini N, Penna A, Mura G, Liberati A (2000) Practice guidelines developed by specialty societies: the need for a critical appraisal. Lancet 355: 103–106.
35. Vincent JL (2006) Is the current management of severe sepsis and septic shock really evidence based? PLoS Med 3: e346. doi:10.1371/journal.pmed.0030346
36. Grant J, Cotttrell R, Cluzeau F, Fawcett G (2000) Evaluating "payback" on biomedical research from papers cited in clinical guidelines: Applied bibliometric study. BMJ 320: 1107–1111.
37. Ramsay LE, Wallis EJ, Yeo WW, Jackson PR (1998) The rationale for differing national recommendations for the treatment of hypertension. Am J Hypertens 11: 79S–88S.
38. Schünemann HJ, Hill SR, Kakad M, Vist GE, Bellamy R, et al. (2007) Transparent development of the WHO rapid advice guidelines. PLoS Med 4: e119 doi:10.1371/journal.pmed.0040119
39. Persaud N, Mamdani MM (2006) External validity: The neglected dimension in evidence ranking. J Eval Clin Practice 12: 450–453.
40. Rubins HB, Robins SJ, Collins D, Fye CL, Anderson JW, et al. (1999) Gemfibrozil for the secondary prevention of coronary heart disease in men with low levels of high-density lipoprotein cholesterol. N Engl J Med 341: 410–418.
41. Sacks FM, Svetkey LP, Vollmer WM, Appel LJ, Bray GA, et al. (2001) Effects on blood pressure of reduced dietary sodium and the Dietary Approaches to Stop Hypertension (DASH) diet. N Engl J Med 344: 3–10.
42. Nakao N, Yoshimura A, Morita H, Takada M, Kayano T, et al. (2003) Combination treatment of angiotensin-II receptor blocker and angiotensin-converting-enzyme inhibitor in non-diabetic renal disease (COOPERATE): A randomized controlled trial. Lancet 361: 117–124.
43. Antithrombotic Trialists' Collaboration (2002) Collaborative meta-analysis of randomized trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients. BMJ 324: 71–86.
44. Kjeldsen SE, Dahlöf B, Devereux RB, Julius S, Aurup P, et al. (2002) Effects of losartan on cardiovascular morbidity and mortality in patients with isolated systolic hypertension and left ventricular hypertrophy: A Losartan Intervention for Endpoint Reduction (LIFE) substudy. JAMA 288: 1491–1498.
45. UK Prospective Diabetes Study Group (1998) Efficacy of atenolol and captopril in reducing risk of macrovascular and microvascular complications in type 2 diabetes (UKPDS 39). BMJ 317: 713–720.

Editors' Summary

Background. Until recently, doctors largely relied on their own experience to choose the best treatment for their patients. Faced with a patient with high blood pressure (hypertension), for example, the doctor had to decide whether to recommend lifestyle changes or to prescribe drugs to reduce the blood pressure. If he or she chose the latter, he or she then had to decide which drug to prescribe, set a target blood pressure, and decide how long to wait before changing the prescription if this target was not reached. But, over the past decade, numerous clinical practice guidelines have been produced by governmental bodies and medical associations to help doctors make treatment decisions like these. For each guideline, experts have searched the medical literature for the current evidence about the diagnosis and treatment of a disease, evaluated the quality of that evidence, and then made recommendations based on the best evidence available.

Why Was This Study Done? The recommendations made in different clinical practice guidelines vary, in part because they are based on evidence of varying quality. To help clinicians decide which recommendations to follow, some guidelines indicate the strength of their recommendations by grading them, based on the methods used to collect the underlying evidence. Thus, a randomized clinical trial (RCT)—one in which patients are randomly allocated to different treatments without the patient or clinician knowing the allocation—provides higher-quality evidence than a nonrandomized trial. Similarly, internally valid trials—in which the differences between patient groups are solely due to their different treatments and not to other aspects of the trial—provide high-quality evidence. However, grading schemes rarely consider the size of studies and whether they have focused on clinical or so-called "surrogate" measures. (For example, an RCT of a treatment to reduce heart or circulation ["cardiovascular"] problems caused by high blood pressure might have death rate as a clinical measure; a surrogate endpoint would be blood pressure reduction.) Most guidelines also do not consider how generalizable (applicable) the results of a trial are to the populations, interventions, and outcomes specified in the guideline recommendation. In this study, the researchers have investigated the quality of the evidence underlying recommendations for cardiovascular risk management in nine evidence-based clinical practice guides using these additional criteria.

What Did the Researchers Do and Find? The researchers extracted the recommendations for managing cardiovascular risk from the current US, Canadian, and European guidelines for the management of diabetes, abnormal blood lipid levels (dyslipidemia), and hypertension. They graded the quality of evidence for each recommendation using the Canadian Hypertension Education Program (CHEP) grading scheme, which considers the type of study, its internal validity, its clinical relevance, and how generally applicable the evidence is considered to be. Of 338 evidence-based recommendations, two-thirds were based on evidence collected in internally valid RCTs, but only half of these RCT-based recommendations were based on high-quality evidence. The evidence underlying 64 of the guideline recommendations failed to achieve a high CHEP grade because the RCT data were collected in a population of people with different characteristics to those covered by the guideline. For example, a recommendation to use spironolactone to reduce blood pressure in people with hypertension was based on an RCT in which the participants initially had congestive heart failure with normal blood pressure. Another 59 recommendations were downgraded because they were based on evidence from RCTs that had not focused on clinical measures of effectiveness.

What Do These Findings Mean? These findings indicate that although most of the recommendations for cardiovascular risk management therapies in the selected guidelines were based on evidence collected in internally valid RCTs, less than one-third were based on high-quality evidence applicable to the populations, treatments, and outcomes specified in guideline recommendations. A limitation of this study is that it analyzed a subset of recommendations in only a few guidelines. Nevertheless, the findings serve to warn clinicians that evidence-based guidelines are not necessarily based on high-quality evidence. In addition, they emphasize the need to make the evidence base underlying guideline recommendations more transparent by using an extended grading system like the CHEP scheme. If this were done, the researchers suggest, it would help clinicians apply guideline recommendations appropriately to their individual patients.

Additional Information. Please access these Web sites via the online version of this summary at http://dx.doi.org/10.1371/journal.pmed.0040250.

- Wikipedia contains pages on evidence-based medicine and on clinical practice guidelines (note: Wikipedia is a free online encyclopedia that anyone can edit; available in several languages)
- The National Guideline Clearinghouse provides information on US national guidelines
- The Guidelines International Network promotes the systematic development and application of clinical practice guidelines
- Information is available on the Canadian Hypertension Education Program (CHEP) (in French and English)
- See information on the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group, an organization that has developed an grading scheme similar to the CHEP scheme (in English, Spanish, French, German, and Italian)