

RECOMMENDATIONS TO THE PHARMACEUTICAL ADVERTISING ADVISORY BOARD REGARDING THE USE OF SCIENTIFIC INFORMATION IN ADVERTISING

January 14,
2012

Don Husereau


Adjunct Professor,

Department of Epidemiology and Community Medicine,
University of Ottawa

Senior Scientist

Institute for Public Health, Medical Decision Making and Health Technology Assessment

UMIT - Private Universität für Gesundheitswissenschaften, Medizinische Informatik und Technik GmbH



“In every line involving scientific details a censor is appointed. The ad-writer, however well informed, may draw wrong inferences from facts. ... The ad seems so simple, and it must be simple to appeal to simple people. But back of that ad may lie reams of data, volumes of information, months of research. So this is no lazy mans field”

-Claude C. Hopkins



KEY MESSAGES

- The growth and evolution of scientific methods in health research and the complexity in understanding them prompts questions as to whether these new methods are appropriate for substantiating claims of effectiveness, comparative effectiveness, and cost-effectiveness.
- Recommendations regarding changes to the PAAB Code of Advertising Acceptance ('Code') were developed. These recommendations focused on the use and reporting of statistical analysis, review articles, meta-analysis, unpublished studies, post-hoc analyses, observational studies, mathematical, non-inferiority trials, economic evaluations, and patient-reported outcomes/health-related quality of life studies.
- All of the recommendations were developed using a standard approach and thorough review of existing evidence and best practice. Each recommendation also considered the current regulatory environment and feasibility of adoption. Recommendations also considered the views of key scientific opinion leaders.
- Some areas were identified in which no change to the PAAB Code is required; this includes disallowing the use of *post hoc* analyses, meta-analysis, and network-meta-analysis to make claims about effectiveness or comparative effectiveness
- Some areas will require changes that should be straightforward to implement; this includes using 95% confidence intervals instead of *P* values to convey statistical information, and re-wording requirement 4.2
- Some areas will require further scientific input and consensus but may be equally straightforward to implement; these include claims based on analysis from Bayesian statistics, trial subgroups, secondary outcomes, observational studies, economic evaluations and health-related quality of life/patient-reported outcome studies.
- Very important but potentially controversial areas of change include adopting biomedical journal standards for supporting and providing access to registered trial protocols, the use of systematic reviews in advertising claims, unpublished research findings, and mathematical modeling



EXECUTIVE SUMMARY

INTRODUCTION

Advertising is an activity designed to influence individual choice. The Pharmaceutical Advertising Advisory Board of Canada (PAAB) defines advertising and its associated promotional activities as “any paid message communicated by Canadian media with the intent to influence the choice, opinion or behavior of those addressed by commercial messages.”

The use of scientific analysis in advertising and promoting pharmaceuticals is commonplace given the role of science in substantiating claims for regulators and clinical decision makers. The population of health care providers that advertising is attempting to influence *are* scientists and the methods taught for making therapeutic decisions are generally science and evidence-based.


The growth and evolution of scientific methods in health research and the complexity in understanding them prompts questions as to whether these new methods are appropriate for substantiating claims of effectiveness, comparative effectiveness, and cost-effectiveness. As part of maintaining its Code of Advertising Acceptance (‘Code’), PAAB solicited an independent analysis of several evolving methodological areas including the use and reporting of statistical analysis, review articles, meta-analysis, unpublished studies, post-hoc analyses, observational studies, mathematical, non-inferiority trials, economic evaluations, and patient-reported outcomes/health-related quality of life studies.

METHODS

Recommendations were developed using a multi-stage approach. First, an analytic framework was developed as a basis for establishing the goals of advertising and creating consistency across recommendations.

Secondly, a literature search was undertaken to identify current best practice and evidence of validity of each method. After a thorough examination of the evidence, options for changing the code were developed and a draft of the evidence synthesis, framework and options were circulated to national and international leaders in the fields of consumer policy, observational and outcomes research, biomedical journal editing, economic evaluation and modeling, systematic review, meta-analysis and network meta-analysis, epidemiology, biostatistics, and health-related quality of life measurements (see Appendix).

Lastly, final recommendations were developed based on the comments and suggestions of the expert panel, the evidence identified the analytic framework



as well as a consideration of their feasibility in implementation. The recommendations were developed independently from PAAB.

FINDINGS

This analytic framework describes the ultimate goal of advertising as improving health and well-being of Canadians. This is done through informing the belief of patient providers and influencing therapeutic choice. Within this framework, it is assumed that an unreliable method will mislead providers into making decisions that lead to suboptimal levels of health. That is, the use of scientific methods that intentionally or unintentionally mislead health care providers will harm, rather than benefit Canadians.


Seventeen recommendations describing opportunities to change and revisit the Code were developed. A rationale for each method was also developed that describes the evidence and best practice information supporting the recommendation, the feasibility of the recommendation and how it is consistent with achieving better health for Canadians through improved clinical decision making. Some of the recommendations are relatively straightforward to implement and uncontroversial, whereas others will require some further discussion, consensus and development. The recommendations are listed below:

Recommendations consistent with the code and requiring no or little change to the code

1. Post hoc analysis should continue to be discouraged
2. The use of meta-analysis for making claims of effectiveness should be discouraged.
3. The use of network meta-analysis for making claims of relative effectiveness should be discouraged

Recommendations relatively easy to adopt, requiring little additional work

1. *P* values should be discouraged wherever possible except under exceptional circumstances and consistent with current guidance from biomedical journals
2. Confidence intervals should be encouraged instead of *P* values wherever possible and consistent with current guidance from biomedical journals. PAAB should suggest only 95% Confidence Intervals (CI) are appropriate for the presentation of findings rather than *P* values.


- 
3. The wording of PAAB Code requirement 4.2 needs correction and revisiting

Recommendations that can be adopted after some moderate additional work

1. PAAB should revisit Code requirement 4.2 and make additional provisions in the Code Explanatory Notes that Bayesian statistical testing is acceptable.
2. PAAB can allow the use of subgroup analysis, but with specific conditions.
3. PAAB can allow the use of claims from secondary outcomes, but with specific conditions.
4. PAAB can allow the use of claims from observational studies, but with specific conditions.
5. PAAB can allow the use of claims of comparative effectiveness from non-inferiority trials, but with specific conditions
6. PAAB should allow claims based on economic evaluation when adequate qualifying language is provided and other regulations are consistently applied.
7. PAAB should allow claims based on HRQoL and PRO measures, but with specific conditions.

Recommendations that require fundamental change and may be viewed as controversial

1. Publication of information from clinical trials should be discouraged if research protocols and outcomes have not been registered and are readily accessible by PAAB and the health care providers that they serve. PAAB should additionally mandate manufacturers provide a link to the registered information in advertisements AND endorse the Ottawa statement
2. If claims from individual studies are used, information regarding the total number of similar studies conducted (in terms of patients, interventions, design) from a *systematic* review of available evidence should be made available to reduce selection bias or claims based on exaggerated study findings.
3. The use of unpublished research findings should not be discouraged.



4. PAAB should allow claims based on mathematical modelling when adequate qualifying language is provided and consumers are given an opportunity to interact with the model.

LIMITATIONS

The recommendations were based on a review of specific aspects of the use of science in advertising and should not be interpreted as a complete review of the PAAB Code. There are other important aspects outside of the scope of these recommendations that may also require exploration currently or in the future. This includes not only evolving scientific methods but how scientific content is depicted and framed for consumers.

CONCLUSIONS

These 17 recommendations are feasible to implement and consistent with the perceived goal of advertising health products – namely, improving the health and well-being of Canadians. They are also consistent with the current regulatory framework for health products and best practices in using evidence to inform decision making. A serious consideration of each recommendation will allow PAAB to achieve its Vision of “trusted healthcare product communication that promotes optimal health” while maintaining and upholding its corporate values of integrity, competency, credibility, independence, excellence, and transparency.



TABLE OF CONTENTS

TABLE OF CONTENTS

Key Messages.....	3
Executive Summary.....	4
Introduction.....	4
Methods.....	4
Findings.....	5
Limitations.....	7
Conclusions.....	7
Table of Contents.....	8
Introduction.....	11
The Role of Advertising.....	11
Defining consumer demand in healthcare.....	11
The Role of Science in Advertising.....	12
Purpose of the Study.....	12
Key Questions.....	13
Analytic Framework.....	14
Goal of Advertising.....	15
Achieving Optimal Health.....	16
The Economics of Optimal Health.....	17
Choice and Misleading Information.....	19
Can Science Mislead?.....	20
General Approach.....	21
Findings for Section 4.2.....	23
How should statistical information be presented so the reader can assess validity, reliability and level of significance?.....	23
Introduction.....	23
Evidence.....	26
Summary of Evidence.....	29
Options.....	29



Recommendations	30
Illustrative Example	33
Findings for Section 3.1 – Claims, Quotations and References	35
Should review articles, pooled data and meta-analysis be used to support clinical/therapeutic claims of effectiveness?.....	35
Introduction	35
Terminology	35
Evidence.....	37
Should unpublished studies be used to support clinical/ therapeutic claims of effectiveness?	42
Introduction	42
Evidence.....	45
Should secondary outcomes, subgroup analysis, and post-hoc analysis be used to support clinical/ therapeutic claims of effectiveness?.....	46
Introduction	46
Evidence.....	47
Should observational (i.e., non-experimental) studies be used to support clinical/therapeutic claims of effectiveness?.....	51
Introduction	51
Evidence.....	53
Summary of Evidence	56
Options	57
Recommendations	59
Illustrative Example	62
Findings for Sections 5.7 - 5.12	64
Should Mathematical Modeling Be Used To Support Comparative Claims Of Effectiveness?	64
Introduction	64
Evidence.....	65
Should Indirect Comparisons Be Used To Support Comparative Claims Of Effectiveness?	67
Introduction	67
Evidence.....	69



Should Non-Inferiority Studies Be Used To Support Comparative Claims Of Effectiveness?	71
Introduction	71
Evidence.....	75
Summary of Evidence.....	78
Options	79
Recommendations	80
How Should Health Economic Claims Be Made?	81
Introduction	81
Evidence.....	84
How Should Claims Of Improvements In Patient-Reported Outcomes/ Health-Related Quality Of Life Be Made?	91
Introduction	91
Evidence.....	95
Summary of Evidence.....	97
Options	98
Recommendations	98
Summary of Recommendations	100
References.....	103
Appendixes	134
APPENDIX A Expert Reviewers.....	134
Experts (alphabetically, by surname), <i>Affiliations</i>	134



INTRODUCTION

THE ROLE OF ADVERTISING

Advertising is an activity designed to influence individual choice. The Pharmaceutical Advertising Advisory Board of Canada defines advertising and its associated promotional activities as “any paid message communicated by Canadian media with the intent to influence the choice, opinion or behavior of those addressed by commercial messages.” Advertising’s role in the context of promoting pharmaceuticals is a commercial one, and implicitly intended to create revenue from increased consumer demand and product sales. This role is more directly stated by one of advertising’s greatest pioneers of advertising, Claude C. Hopkins, who in 1923 reminded the world of advertising’s real purpose: “The only purpose of advertising is to make sales. It is profitable or unprofitable according to its actual sales.”


DEFINING CONSUMER DEMAND IN HEALTHCARE

The world of medicine and delivery of health care is unique in that it is most often paid for through pooled risk-sharing schemes, called health care insurance. Those who receive medical products and services may not directly pay for them and those who pay for health care goods and services may never receive them. Although in Canada, drugs outside of a hospital setting do not fall under universal health insurance, public and private sector insurance plans account for 83% of expenditure on prescribed drugs (1).

Health care providers, often physicians, also do not pay for health products and services, and unlike most consumers, may not have information about its price. Nonetheless, providers act as agents and make choices for patients. Patients may have some say in the choice of medical goods and services they receive, but substantial empirical evidence suggests providers are the ultimate agent of demand for health products and services (2).

Providers, then, play the role of both *suppliers*¹ of health care products and services and *consumers* at the same time, because they are predominantly responsible for the therapeutic choices of their patients. Additionally, the current regulatory framework in Canada limits advertising directly to patients. So for the purpose of the recommendations given in this document, we will focus on the information that exists in advertising to care providers. In this sense,

¹ There are manufacturers (i.e., producers), wholesalers, and retailers of health products who are the actual suppliers. However, health care providers act as agents for these suppliers and are the interface between supply and demand.



health care providers act as the *consumers* that advertising is directed to. Different types of health care providers play the role of agent and *consumer* for different types of health care products.


THE ROLE OF SCIENCE IN ADVERTISING

The use of science information in advertising has a long and evolving history. It has been argued that scientific advances can be important to improving social welfare and that the timely and effective dissemination of information to consumers will be beneficial to society.(3) Additionally, the use of science has great appeal to advertisers, who are motivated to make compelling arguments to their audience, using claims that will promote their product, and that can be *substantiated*. In the 21st century, *science* is a well accepted vehicle for substantiation and gauging the accuracy and truthfulness of claims in advertising.(4) In advertising to care providers, science becomes more salient since the consumers receiving information *are* scientists and the methods taught for making therapeutic decisions are science and evidence-based.

Despite the potential benefits of using science in medical advertising, there has been considerable tension between science and advertising in medicine. (5,6) One medical commentator noted that “Science may be defined as a critical analysis of data from well-designed studies. Advertising, on the other hand, is a self-serving and biased promulgation of data.” Others have cautioned about the nature of science and scientific information itself. How we arrive at “truth” depends on how we think about science (i.e., the philosophy of science) and how it informs our beliefs. Despite some who might believe the contrary, science cannot tell us what is “true”. It is generally accepted that what is true today in medicine may not be tomorrow (7–9), due to the evolving nature of scientific evidence and how we measure and interpret it.(10,11) This highlights the need of adopting standards of “truthfulness” and rules for how science information in advertising can be applied.

PURPOSE OF THE STUDY

The Pharmaceutical Advertising Advisory Board (PAAB) is an independent review agency whose primary role is to ensure that healthcare product communications are accurate, balanced, evidence-based, and reflect current and best practice. In granting the PAAB approval and with it the authorization to use the PAAB logo on advertising materials, the PAAB adopts the standards specified in its Code of Advertising Acceptance (the ‘Code’) to all categories of health care products including prescription drugs, non-prescription drugs, and natural health products.



One of the principle activities of the PAAB is maintaining its Code. In doing this, from time to time, the PAAB must consider re-assessing the standards specified in its Code. As stated in PAAB's mandate, "The PAAB also monitors trends in health product advertising and promotion and adjusts its code and practices as required to fulfill its mandate."

Decisions to revisit the code can be made by the PAAB Board of Directors. Such a decision may be based, in part, on research that informs an assertion that the code requires revisiting.

In response to a request by PAAB, the information that follows attempts to synthesize the current state of knowledge and identify best practices in regards to the use of scientific methods and the reporting of scientific information in specific areas of concern to PAAB. It focuses on three key areas in the Code: 1) Claims, Quotations and References (specifically, interpretation of item 3.1); 2) Data Presentations (specifically, item 4.2); and 3) and Comparative Claims (specifically sections 5.7-5.12). The report provides options and some preliminary direction to PAAB regarding how the Code should be changed.

The key questions subjects and research questions developed in consultation with PAAB can be seen below. Each section of the report will attempt to answer these questions. Each question is a normative question, i.e., a question whose answers will result in recommendation or advice for changes to the code. For each question, a comprehensive search for key pieces of information required to properly address policy questions was conducted.

KEY QUESTIONS

Questions addressing specific sections of PAAB Code of Advertising Acceptance

Section 4: Data Presentations


Section 4.2 - Statistics must be presented so as to accurately reflect their validity, reliability and level of significance

- 1) How should statistical information be presented so the reader can assess validity, reliability and level of significance?

Questions addressing specific sections of PAAB Code of Advertising Acceptance

Section 3: Claims, Quotations and References

Section 3.1 - Claims and/or quotations in Advertising/ Promotion Systems (APS) must be consistent with, and within the limitations of, the Health Canada Terms of Market Authorization, or prescribing information for products with no product monograph. Any APS containing direct or indirect product claims [11.7] and/or quotations from the scientific literature must include a complete listing of the scientific references. Labeling must be authorized by Health Canada.

- 
- 1) Should review articles, pooled data and meta-analysis be used to support clinical/therapeutic claims of effectiveness?
 - 2) Should unpublished studies be used to support clinical/therapeutic claims of effectiveness?
 - 3) Should post-hoc analyses be used to support clinical/therapeutic claims of effectiveness?
 - 4) Should observational (i.e., non-experimental) studies be used to support clinical/therapeutic claims of effectiveness?

Questions addressing specific sections of PAAB Code of Advertising Acceptance
Section 5: Comparisons

Section 5.7 - Comparative claims of efficacy and safety require support of evidence from head-to-head well-designed, adequately controlled, blinded, randomized clinical studies. Open-label studies are not considered to be a high level of evidence and are not acceptable if subjective end-points are included in the study. Comparative claims should be relevant to current medical opinion and practice.


- 1) Should mathematical modeling be used to support comparative claims of effectiveness?
- 2) Should non-inferiority designs be allowed to support comparative claims of comparative effectiveness?

Section 5.10 of the code states, “All direct and indirect comparisons must not mislead, and be supported by reliable current data”. In the explanatory section, it is stated “Pharmacoeconomic and quality of life claims must be supported by high-quality studies. Disclosure of study parameters, Section 5.11, is important for interpretation of results.” There is no specific guidance for study parameters that apply to pharmacoeconomics studies.

- 1) How should health economic claims be made?
- 2) How should claims of improvements in patient-reported outcomes/health-related quality of life be made?

ANALYTIC FRAMEWORK

This section describes how the analysis was carried out – specifically how information was considered to create individual options and recommendations. An analytic framework is intended to improve the transparency and consistency in judgment across recommendations and can also assist others with future



improvements to the code. If the logic and assumptions of the framework are questioned, then each of the individual recommendations can be changed accordingly.

GOAL OF ADVERTISING

The first question we need to consider in the analytic framework is what the ultimate goal of advertising is. We have already stated that advertising is intended to influence individual choice. And from a commercial standpoint, influencing choice increases sales. But from this standpoint, we would not have to regulate the type of information contained in advertising since the positive (i.e., increased sales) or negative (i.e., no increase in sales) consequences of advertising would be the direct responsibility of the advertiser.

Instead, the need for regulation stems from concerns about the goals of the *consumer* and whether consumers would make choices differently with better knowledge. Federal statutes to protect consumers are outlined in Canada's Competition Act (12). According to Section 52 of the act, contravention of the act applies to advertising that is "misleading in a material respect"(12). "Material" in this context refers to the degree of influence the deceptive act or practice has and whether the consumer would have chosen differently without its existence(13). It is assumed what is material to *consumers* (and their patients) in the advertising of health products is the goal of obtaining health. That is, consumers make choices to use health products in order to obtain health. Consistent with this, one economic theory, called agency theory, suggests that providers acting as agents on behalf of their patients will make better choices (i.e., more closely aligned and with the goal of health) when more complete information is available to them.(14–17)

The World Health Organization (WHO) has defined health as defined health as "a state of complete physical, mental and social well being and not merely the absence of disease or infirmity." Although this definition has met with some criticism(18–21), health system decision makers generally measure health and the performance of health systems in terms of *goodness* and *fairness*(22). Goodness is achieving the best attainable (i.e., optimal) level of health for the population; fairness is the smallest feasible difference among individuals and groups. It is assumed that advertising, for the most part, will not promote greater differences in opportunities for health across populations. That is, it will not lead to unfair gains for certain populations. It is assumed that the goal of advertising is focused more on goodness, similar to the implied goal of PAAB itself as stated in the PAAB Vision Statement, namely the promotion of "optimal health" (23).

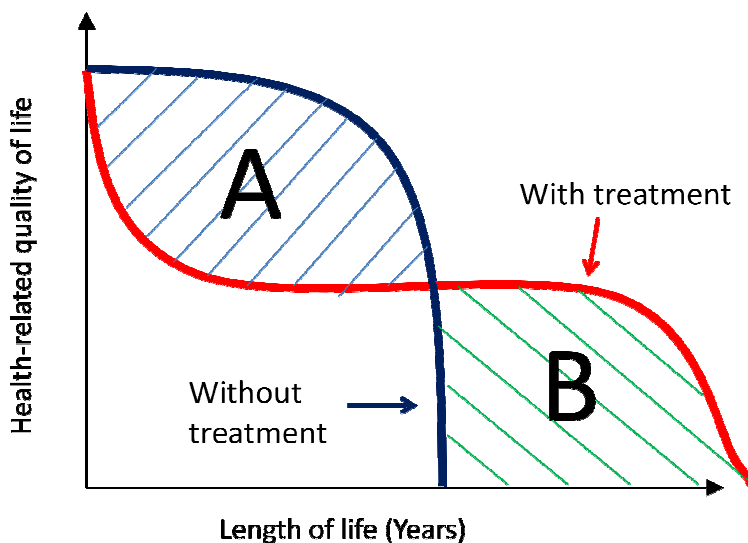



ACHIEVING OPTIMAL HEALTH

It is generally accepted by patients and health care providers that improving longevity and reducing disease are important clinical outcomes (24). Measures have been developed that combine both quality and quantity of life to arrive at a single measure (25). For example, the WHO uses a measure incorporating life expectancy and health-related quality of life to compare average levels of population health across countries(22). These types of measurements are also used in health care decision making. A popular unit used for reimbursement decisions and health economic evaluation is the quality-adjusted life-year (QALY)(26).

To illustrate the QALY concept, we can imagine a person who is told she has one year to live in perfect health after which time, she is likely to die. She is given the choice to take a new health product (drug) that can extend life by an additional year. (See Figure 1) The medicine is likely to produce severe side effects that reduce the quality of life in half; the two years of life multiplied by a 50% quality of life can be thought of as equivalent to a single year in perfect health. By putting length of life on a horizontal axis and health-related quality of life on a vertical axis, we can depict the relationship of quality and quantity of life experienced with the life-extending therapy (depicted by region B). We can see the area under the blue line (without treatment) is roughly equivalent to the area under the red line (with treatment). We can also see that the difference between no treatment and treatment (region A) is roughly equivalent to the difference between treatment and no treatment (region B).

FIGURE 1: ILLUSTRATION OF THE QALY





It has been argued that there are important dimensions beyond simply length and (health-related) quality of life (24). These include the quality of life of family and caregivers, and convenience to patients. Also the measurements do not consider unmet need or distinguish between additional health gained by the very old (for which there is precious little health remaining) or for the very sick (who may value small health gains to a greater extent). Nonetheless, QALYs and similar metrics are standard and conventional international approaches to considering gains in health. For the purpose of our framework, we, too, will assume that length and health-related quality of life are the most important considerations for patients and providers when therapeutic decisions are made with the goal of improving or protecting health. We will assume that patients and providers wish to achieve net health benefit from therapeutic choices— that is, there will be an overall measurable increase in the value of quality and quantity of life achieved through choosing one therapy versus another.

THE ECONOMICS OF OPTIMAL HEALTH

One additional important consideration for the achievement of optimal health is the consideration of scarce health care resources. On an individual level, patients and providers must consider health gains achieved in terms of the benefits versus harms from therapy. That is, they need to ask themselves if the potential gains in terms of length and quality of life from therapy outweigh potential harms in terms of reduced length or quality of life from disease or unintended consequences of the drug.

On a third-party payer or societal level, we need to consider gains achievable within a budget constraint (27). We have to ask ourselves if the societal cost of providing a new therapy is prohibitive, even one with the potential to increase net health benefit to an individual patient relative to an existing therapy. That is, if the costs to the health system of providing the new therapy could have produced more health if invested elsewhere, the new therapy will actually *decrease* overall population health, even if the new therapy *increases* individual health benefits. This can be thought of as the opportunity cost of making a therapeutic choice.

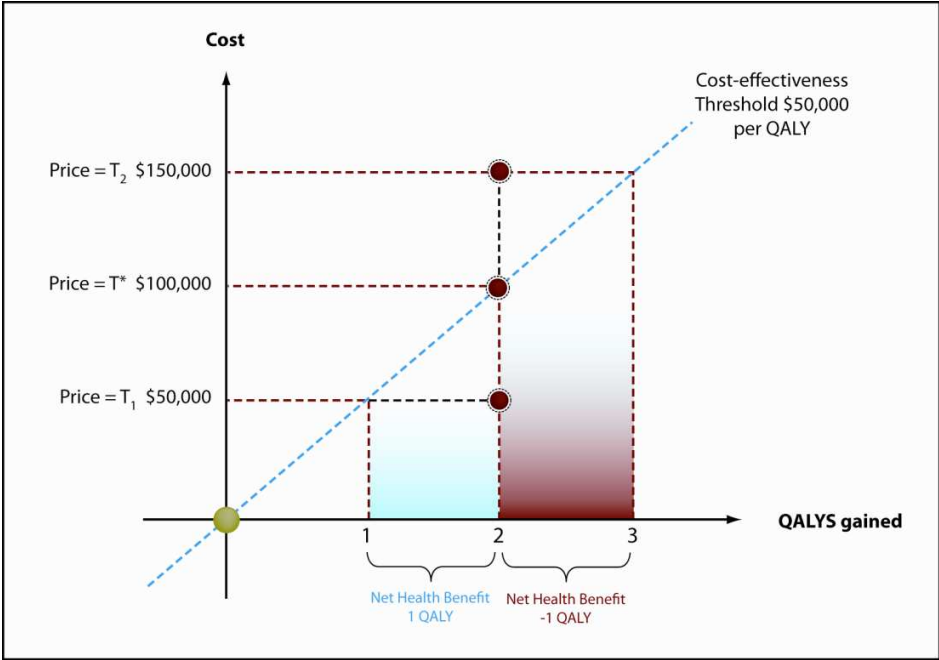
Within this economic framework, promoting therapies of either reasonable or high value is beneficial to both patients and society. Promoting the use of low-value (potentially from being over-priced or because there is too much uncertainty) therapies can still be good for an individual, but has the unintended consequences of removing the opportunity for health gains for other patients who access the health system – essentially harming patients indirectly by wasting health care resources.(28) It is important not to overlook this important




aspect of achieving optimal health, especially in the context of advertising and promoting therapy using health economic information.

To illustrate this, we can consider a therapy which produces net health gains for individuals, but is available for sale to the health system at three different prices. (Figure 2) At the lowest price (T_1), we are able to produce two quality-adjusted life-years (QALYs) for the resources normally needed to generate one QALY, so by investing in the new therapy, we can actually produce an extra QALY and society benefits (net health benefit of 1 QALY). At a higher price, (T^*), it takes the same resources to produce a QALY with a new therapy that we would need to generate the same amount of health with existing therapies. So investing in the new therapy does not produce any health gains for society. With the highest price (T_2), we could actually produce 3 QALYs in the current system with the resources required to just generate 2 additional QALYs with the new therapy. Despite the fact the therapy can benefit an individual, investing in the therapy has led to an overall societal loss (of 1 QALY) since resources could have been put to better use elsewhere.

FIGURE 2: OPPORTUNITY COSTS IN HEALTHCARE



In summary, we will assume for the purpose of the framework and consistent with the delivery of health care and PAAB's vision that decisions regarding therapeutic choices should be made with the intended *goal* of achieving the highest possible net health gains, in terms of length and quality of life, across all
Page | 18



individuals in society. This framework must consider both the potential net health gains for individuals as well as the net health gains for a society given available health system resources.

CHOICE AND MISLEADING INFORMATION

The first statement of the General Requirements of the PAAB Code suggests “Statements or illustrations must not mislead” (23). This is consistent with provisions in the Competition Act (12) and the PAAB Mission and Mandate of being trustworthy, accurate, balanced, evidence-based and reflecting best practice for clinical/therapeutic claims of effectiveness (23). The following section will illustrate how the definition of what is “misleading” in advertising should be thought of in terms of the choices we make and the consequences of our choices.

As already stated the *goal* of advertising for consumers and of making therapeutic decisions is to provide patients with the highest possible net health gains, in terms of length and quality of life. Good advertising will therefore influence consumer choice and *lead* the consumer to this worthy goal. Misleading advertising will allow the consumer to believe they will be led to this goal, but will actually lead the consumer somewhere else, and without their foreknowledge. It is the consequence of the advertising that makes it misleading, not its intent. Misleading advertising is similar to a poorly printed or hard-to-decipher map. A consumer who must rely on this information may never wind up at their chosen destination.

We can imagine this as a choice diagram as in Figure 3. *Consumers* (i.e., providers) are exposed to advertising and this influences their choices to use a new therapy. In a world of no advertising, consumers will have certain beliefs and may or may not discover and use the new therapy. With advertising, more use occurs and this increased level of use then translates into some measure of improved health for society. The net health gains are the difference between the level of health achieved in a world with advertising (B) and the level of health achieved in a world where consumers are not exposed to advertising (and presumably choose the new therapy less often, A). These net health gains = *net health benefit B minus net health benefit C*.

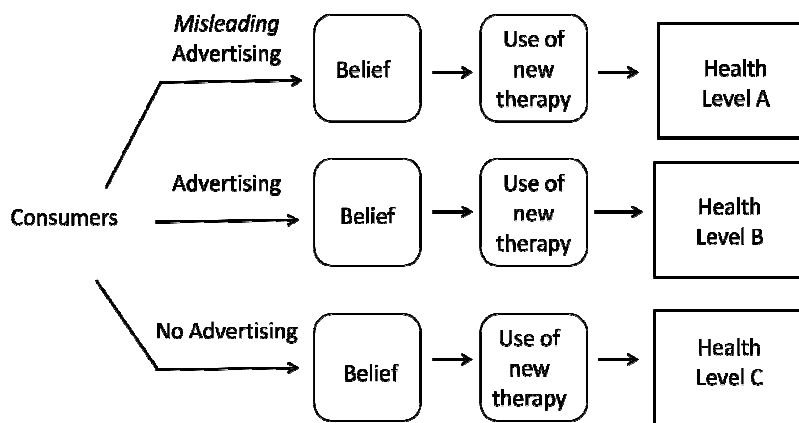
Exposing consumers to misleading, or deceptive, advertising will also influence individual choice and change levels of use of the new therapy. But misleading advertising will change beliefs in consumers and lead some consumers to use a new therapy that they normally would not have used *had they know better*.

We assume consumers will choose with the goal of health in mind. Misleading advertising will lead to suboptimal levels of health. The difference between the



health gains from misleading advertising (A) and the health gains under advertising (B) that is not misleading is the health lost from deception. Health lost could mean harm from inappropriate use, or it could mean harm from using a less effective therapy. It could also mean harm from using a similarly effective therapy using scarce resources that could have been used elsewhere. Misleading advertising, then, has the potential to be harmful to both patients and society. It is the difference between net health benefit A minus net health benefit B.


FIGURE 3: HOW MISLEADING ADVERTISING IS HARMFUL



In summary, misleading advertising and promotion is defined as advertising that does not lead to levels of belief, therapeutic choice associated with the highest possible net health gains, in terms of length and quality of life, across all individuals in society, when compared to the use of advertising that is not misleading. Misleading advertising will harm patients, by leading reasonable consumers concerned about health to make therapeutic choices that they otherwise would not have made, had they been given information that was consistent with PAAB’s Mission – trustworthy, accurate, balanced, evidence-based and reflecting best practice.

CAN SCIENCE MISLEAD?

There are many ways in which advertising and promotion can mislead or be deceiving to consumers. It is important to bear in mind that deceptive advertising does not imply deliberate deception on behalf of the advertiser. As suggested in The Competition Act and the Canadian Code of Advertising Standards, what is important is the subjective impression and interpretation by the consumer (12,29). Similarly, the US Federal Trade Commission (FTC) drafted a policy statement on deception in an attempt to guide public understanding on the subject. It suggests examining the likelihood that a reasonable consumer



would be misled from a deceptive act or practice (30). A deceptive act is further defined as either a misrepresentation or omission of information (30).

The research and recommendations in this report are limited to the use of science information in advertising to health care providers— both what types of scientific methods and claims are reasonable and how they should be reported. It does not focus on other aspects of advertising of which there are numerous (e.g., confirmatory bias effects and framing with misleading graphics and text, non-scientific misleading verbal or written communication), other audiences (e.g., children, patients) or non-therapeutic (i.e., health-related) claims.


Since science and its methods are often promoted as objective measures, it might seem unlikely at first that the use of science information can mislead consumers. However, science can and has misled consumers (even consumer-scientists) in numerous ways:

- **Framing** – Questions can be framed so that even proper methods cannot reveal important results.
- **Misrepresentation** – Using appropriate underlying scientific methods, scientific results can be misreported, falsely reported or misrepresented graphically or verbally.
- **Misapplication** - Scientific methods can be incorrectly applied, or correctly applied but inappropriate within a specific context. Results from properly applied and appropriate methods can be misinterpreted.
- **Omission** - Scientific methods or results can be not reported or underreported.

Properly applying scientific method is tricky business, even for scientists. As suggested previously, misleading advertising with scientific advertising does not equate with deliberate deception; scientific methods continue to evolve and both consumers and advertisers may be unaware of the appropriate application in certain contexts. However, it is because consumers (or in the case of health care, patients) can be harmed by unfair or deceptive practices that regulation and law enforcement are required. It is hoped that the recommendations that follow embody a current state of thinking about the proper application and use of scientific method to better inform current regulation.

GENERAL APPROACH

For each research question, a rigorous and transparent review and analysis of relevant literature from biomedical and social science databases along with



unpublished literature from other relevant sources (such as government publications) was conducted. Conceptual and empirical studies addressing each research question were identified and synthesized for each question under the section “Evidence”.

Studies were sought that described how using different scientific methods can alter estimates of the underlying true result. In scientific terms, if a method or approach consistently leads to different estimates from an underlying true estimate, this is known as a *bias*. The reviews conducted for each research question attempt to quantify the potential bias from using one scientific method or approach versus another. It is assumed that if a significant bias has been demonstrated, then a reasonable consumer might be materially misled by a particular choice (i.e., if they are misled, this could have consequences on patient health). If there is a potential for bias that can be regulated, options for detecting and dealing with bias are reported. We may have information that suggests different approaches arrive at different answers, but no single approach will consistently arrive at an answer which is closest to the truth (i.e., that minimizes bias). Studies that measure the frequency of current usage of methods and how a consumer might respond to information were also sought, in an attempt to better understand whether the method represents current *best practice* and what the *subjective impression* that would be left on a “reasonable consumer” of information, namely a health care provider.

Options for changing the code were made in light of the information identified and the potential positive and negative consequences of each of the options are described under the heading “Options”. These options were made within the current context of regulated advertising – that is, options to improve the use of scientific methods and reporting must still be entirely compatible with the information contained with the manufacturer’s Health Canada-authorized product monograph. Options were not considered where information not consistent with the indications listed in the authorized product monograph would need to be used. (Consistent with PAAB Code Requirement 3.1) The evidence and options for each section were then reviewed by national and international experts (see Appendix to inform the development of a final recommendation and accompanying rationale.

In this phase, options could be revisited if relevant evidence was identified as missing by the expert, if flaws in the analysis were discovered, or if new options were identified. Experts were also asked for commentary and this is reported where given.

FINDINGS FOR SECTION 4.2

The requirements of the code as stated in Section 4.2 state, “Statistics must be presented so as to accurately reflect their validity, reliability and level of significance”. In the explanatory notes to this section, it is suggested that “Statistical information should include dosage and the level of significance e.g. p-value, in the presentation. Information such as patient numbers, time span, dosage, etc. that are needed to assess the data may appear in the product summary box in the prescribing information.”

HOW SHOULD STATISTICAL INFORMATION BE PRESENTED SO THE READER CAN ASSESS VALIDITY, RELIABILITY AND LEVEL OF SIGNIFICANCE?


INTRODUCTION

P-VALUES AND HYPOTHESIS TESTING

Statistics has been defined as “...the department of study that has for its object the collection and arrangement of numerical facts or data, whether relating to human affairs or to natural phenomena.”(31). With the increased prominence of clinical trials as the basis for substantiating knowledge claims of effectiveness, conventions for arranging and interpreting data have arisen in the past decades (32). These approaches have in turn led to standards for statistical reporting in leading biomedical journals.(33)

Claims of clinical effectiveness typically rely on results from randomized controlled trials. The results from these experiments allow a comparison between a group exposed to an intervention and an unexposed group. The biomedical community has adopted a combination of two statistical approaches for creating these comparisons: 1) The *P* value, developed by Ronald Fisher in the 1920s as an index to measure discrepancy between observation and a null hypothesis (i.e., no clinical benefit) (34) and; 2) Hypothesis testing, developed Jerzy Neyman and Egon Pearson as a method of making decisions about which hypotheses to reject (35). Although originally intended as alternative and incompatible methods, they are now often and mistakenly regarded as a single approach, sometimes called the ‘null hypothesis significance test’.(36)

The *P* value is defined as the probability, under the assumption of no effect or no difference (the null hypothesis), of obtaining a result equal to or more extreme than what was actually observed (34). To calculate a *P* value, the investigator



must have information about the observed difference in means of the outcome values between population samples, and the standard error of the difference in means. Standard error is calculated from a measure of standard deviation and sample size (37). The investigator must then make some assumption about how the range of possible observed differences, given there is really no difference, could be distributed (typically normal, or Gaussian). As a measure of how severely the null hypothesis is contradicted by observation, Fisher proposed a P value of 0.05 as a “convenient . . . limit in judging whether a deviation is to be considered **significant** or not”(34,38) (Emphasis mine) This is the value most often used in clinical trials today. To properly interpret the results of a significance test, the reader must know the values, distributional assumptions and level of significance tested.

With hypothesis testing, the investigator creates two hypotheses about nature: typically one that states there is no observed difference between a therapy and no treatment (i.e., the null hypothesis); the other is typically the opposite hypothesis – i.e., that there is some kind of effect. Given the results of an experiment, the investigator has two options – to accept one hypothesis and reject the other. In doing so, the investigator stands a chance of making an error. If the investigator rejects the null hypothesis (i.e., no difference) and accepts the hypothesis that there is an effect based on the experiment, but the opposite is actually true, they have committed a Type I error (also known as a false-positive). If the null hypothesis is accepted when it should have been rejected, they have committed a Type II error (also known as a false negative). Typically, clinical trialists design trials so that Type I errors are committed in 5% (often referred to as an α of 0.05) of trial observations and that Type II errors (β) only occur in 20%. The probability of not committing a Type II error is often referred to as the statistical power of the experiment – that is, the chance of not declaring no observed difference by accident (i.e., equivalent to $1 - \beta$).

As stated by Goodman, “Hypothesis tests are equivalent to a system of justice that is not concerned with which individual defendant is found guilty or innocent (that is, ‘whether each separate hypothesis is true or false’) but tries instead to control the overall number of incorrect verdicts (that is, ‘in the long run of experience, we shall not often be wrong’).”(36) Hypothesis testing was proposed as a way of limiting the number of wrong conclusions, but there is no way for an individual investigator to know if their individual conclusion is a mistaken one. To properly interpret the results of a hypothesis test, the reader must have the same information as for P value/significance testing (difference in means, standard error of difference) but they must also know what hypotheses are being tested and the error rates (α , β) assumed.

ALTERNATIVE STATISTICAL MEASURES

P values depend on both size of the observed differences *and* the precision of the estimate (based on standard deviation and sample size). As a consequence, a small effect in a study with large sample size can have the same *P* value as a large effect in a study with a small sample size. As such, they have been called “confounded information” (39). Confidence intervals allow the reader to see each of the “confounded” pieces of information separately, the size of effect and precision of the estimate. One approach to reporting both size and range of effects that are compatible with the observed data is to use a confidence interval.

P values and hypothesis testing are part of prominent school of medical statistics called *frequentist* statistics. Frequentist statistics allow clinicians to make deductions about observations – namely, they provide us with a sense of the probability of the data given a reality (i.e., a hypothesis). Neither significance testing with *P* values nor hypothesis testing with Type I and II error thresholds can give us any measure of the degree to which a claim of effectiveness is true. Hence, conclusions based on the data must consider a number of important pieces of information in addition to the results of the statistical tests: these include the magnitude of the effect; its clinical significance; its consistency with other measured endpoints; its consistency with evidence from previous studies; and (controversially) its consistence with biological theories and other forms of indirect knowledge.(36,40)

Although less prevalent, there exists an entirely different branch of statistics called Bayesian statistics. Bayesian statistics allow us to make inferences of a different quality – they provide us with a sense of the probability of a reality given the data. Bayesian statistics use different information values: Bayes factors instead of *P* values; and credible intervals instead of confidence intervals. Instead of using assumptions regarding the distribution of outcomes given no difference, Bayesian statistics requires assumptions about the distribution of outcomes prior to new information (called a Bayesian prior distribution). Randomized controlled trials for regulatory purposes have until now been based on frequentist statistics; the first trials based on Bayesian statistics are forthcoming.(41)

Significance testing, hypothesis testing, and Bayesian approaches are mainstream approaches to the analysis of clinical data. The question is not whether or under what circumstances each approach should be applied. Rather, assuming that each brings useful information to those who make clinical decisions, the question is how properly reporting each can contribute to an understanding of the “validity, reliability and level of significance”.

EVIDENCE


There is a preponderance of evidence to suggest the application and reporting of statistics in medicine is less than ideal or even poor (42–44), even in highly regarded medical journals, such as *Nature* and *BMJ* (45). An entire website has been dedicated to the extensive body of literature that describes the misuse and abuse of null hypothesis significance testing (46). Some common themes for improvement and best practices identified from the literature search and adapted from (45) are shown below.

BOX 1: COMMON PROBLEMS WITH STATISTICAL REPORTING IN BIOMEDICAL RESEARCH (ADAPTED FROM GARCIA-BERTHOU (45))

- 1) Confidence intervals are often more appropriate than hypothesis testing. If hypothesis testing is used, it is desirable to report not only the *P* values but also the observed values of test statistics and the degrees of freedom.
- 2) When hypothesis testing occurs, it should also describe clinical significance, power, sample size, and significant deviations from research protocols.
- 3) Bayesian statistics can be useful for medical decision making and complement current frequentist approaches.
- 4) If *P* values are required, exact values (to no more than two significant figures) should be given rather than reporting $P > 0.05$ or $P < 0.01$. It is unnecessary to specify levels of *P* lower than 0.0001.
- 5) Spurious precision adds no value to a paper and even detracts from its readability and credibility. Results need to be rounded.
- 6) Numerical results should be correctly rounded.
- 7) The preparation and editing of manuscripts should be more carefully checked.
- 8) Authors should make raw data freely available and journals should implement and stimulate this practice.
- 9) The software version or code used should also be stated, since this provides additional information regarding the methods used.

CONFIDENCE INTERVALS

Although *P* values have not been entirely discouraged, confidence intervals are increasingly promoted by leading biomedical journals and reporting guidelines over the use of *P* values (47). As already stated, confidence intervals are perceived as useful because they disentangle effect size from the precision of



the estimate. Confidence Intervals still harbor some problems, however. Firstly, some observers have cautioned that they must always be accompanied by the interval they are attempting to capture (e.g., 90%, 95%, 99%) to avoid misinterpretation (48,49). Secondly, and similar to *P* values, confidence intervals may be mistakenly misinterpreted as a measure of confidence in a hypothesis rather than a plausible range of values explained by the data (36,38,48). Thirdly, confidence intervals have been shown to overestimate precision in single trials, when heterogeneity across previous results is not factored in.(50) An alternative approach to confidence intervals, called the *P* value function, was previously proposed but not widely adopted (51).


Confidence intervals may also be used incorrectly for hypothesis testing. They have a potential to mislead consumers who assume that if two 95% confidence intervals overlap, then the null hypothesis must be accepted (49). Although this is generally the case, it is not always the case. In an essay titled “*P*-Values and Confidence Intervals: Two Sides of the Same Unsatisfactory Coin”, Dr. Alvan Feinstein summarized confidence intervals this way: “Despite the advantage of demonstrating how large [the therapeutic effect] might really be, the confidence-interval approach can produce a new problem while preserving two old ones. The new problem, according to Dr. Feinstein was using an arbitrary level of significance (such as 90%) without telling the reader. The preserved problems were omitting standard error values and improperly applying/calculating values without the readers’ knowledge.

HYPOTHESIS TESTING

Despite hypothesis testing being used extensively, the statistical test, which is one part of hypothesis testing, is often overemphasized compared to other equally important factors. There is substantial evidence to suggest underreporting of clinical significance or importance(52,53), power and sample size(53,54), and the error rates chosen a priori along with protocol deviations (55,56) to aid readers in understanding the validity of the conclusions drawn. Even when this information is available, a preponderance of small, underpowered trials with an inadequate application of hypothesis testing occurs frequently, even in premier biomedical journals (57). This phenomenon, coupled with a growth in molecular and genome-wide association studies (which have additional challenges for hypothesis testing) led one investigator to speculate that most currently published research findings were likely to be “false” (58).

P-VALUES

As already mentioned, the vast majority of literature describing how consumers can be misled from *P* values describes their overuse in the context of hypothesis testing (36,39,59,60). Although some have suggested they should almost always



not be used (39), others have argued that they can be useful in some situations (61) but that care must be taken to adjust the results of the statistical tests when an experiment involves multiple endpoints, subgroups, interim analyses, sequential testing, or protocol deviations (62,63) One report cautioned readers against the interpretation of extreme P values, suggesting very low values (e.g., $P=0.0001$) may not be significant at all, but instead be a symptom of a corrupt experiment (64). Other research has suggested P values may often be incorrectly rounded or be too precise (i.e., too many digits) (45).


BAYESIAN STATISTICS

A vast literature describes the proper conduct and interpretation of Bayesian statistics in the context of clinical trials. The consensus opinion is that Bayesian statistics are feasible (65,66) and useful (67,68), particularly as a means of encouraging flexibility in trial design. Bayesian analyses can be applied to clinical trials that have already undergone frequentist analyses – some have suggested this usually leads to less extreme results (69) while others have suggested that results can be the same or divergent, but show no reliable direction in difference (70). The consensus opinion currently is that Bayesian and frequentist approaches complement each other by providing unique pieces of information for decision making (69,71).

PROTOCOL INFORMATION AND RAW DATA

Because statistical information, particularly from frequentist statistics, cannot be interpreted without understanding the context and design of the experiment, the registration of trial protocol information has been advocated as a means of ensuring the validity of reported analyses. Principles for the registration of protocol information from clinical trials have been published (called “The Ottawa Statement”) (72) and regulatory and academic incentives have been put in place. For example, the International Committee of Medical Journal Editors (ICMJE) established a policy requiring trial registration prior to conduct as a condition of publication. This policy has been adopted by leading biomedical journals. Other legislative changes promoting the disclosure of analysis plans and data are ongoing, particularly in the US where the US Food and Drug Administration (FDA) Amendments Act of 2007 led to the legal requirement to disclose trial results.(73)

As a means to ensure the reproducibility of published research, biomedical journals are also increasingly adopting policies and standards for sharing raw data and software code for the purpose of ensuring reproducibility in research (74–76,76). Similarly, and in response to observed bias in industry-sponsored analyses of clinical trials, the Journal of the American Medical Association established a policy that the entire raw data set should be given to an



independent biostatistician, along with the study protocol and the pre-specified plan for data analysis (77) Other related initiatives have been a strengthening of declarations of funding and competing interests by report authors.


SUMMARY OF EVIDENCE

- There is considerable evidence to suggest that statistics are often misapplied or misreported in clinical trial research, and that this has the potential to mislead clinical decision makers.²
- Practices that have been promoted to reduce the potential for bias include
 - De-emphasizing or discouraging use of *P* values or the results of statistical tests
 - Encouraging the use of confidence intervals
 - Ensuring access to or describing the context, analysis plans, protocols, raw data and other factors (clinical significance, power, sample size, and significant deviations from research protocols.) to aid in interpretation of null hypothesis testing
- Bayesian statistics is an acceptable complement to current approaches but requires different reporting metrics (e.g., credible intervals) and language.

OPTIONS

1. *P* values should be discouraged wherever possible except under exceptional circumstances and consistent with current guidance from biomedical journals
 2. Confidence intervals should be encouraged instead of *P* values wherever possible and consistent with current guidance e from biomedical journals.
- [Option1] – PAAB should suggest only 95% CI are appropriate for the presentation of findings rather than *P* values. Since 95%CI can always and quite easily be calculated from *P* values and estimated differences, there is no reason for this information to not be available
 - -[Option 2]- PAAB should suggest 95%CI with optional *P* values are appropriate – this may be less confusing to clinicians used to *P* values (although they are unlikely to properly apply them) but it undermines many

² There are also issues related to the application and reporting of statistics in relevant biomedical research beyond randomized controlled trials including meta-analysis, cost-effectiveness trials, non-inferiority trials with factorial design (78), cluster randomized trials, and stepped-wedge trials (79). Meta-analysis, cost-effectiveness trials and non-inferiority trials are covered elsewhere in the report.



of the problems posed by the use and abuse of *P* values. It is, however, consistent with many published reports in biomedical journals

3. Publication of information from clinical trials should be discouraged if research protocols and outcomes have not been registered and are readily accessible by PAAB and the health care providers that they serve

- [Option 1] – PAAB can insist that null hypothesis testing information only be allowed from clinical trials with fully registered clinical trials. In this way, valid interpretation of statistical tests can be made.
- [Option 2] - PAAB can insist that null hypothesis testing information only be allowed from clinical trials with fully registered clinical trials AND provide a link to the information on advertisements. This is more consistent with the PAAB Value of transparency
- [Option 3] - PAAB can insist that null hypothesis testing information only be allowed from clinical trials with fully registered clinical trials AND provide a link to the information on advertisements AND endorse the Ottawa statement(72)


4. Bayesian statistical analysis of trials should be allowed

- [Option 1] – PAAB should make provision in the Code that Bayesian statistical testing is acceptable but therapeutic benefit must be expressed in 95% credible intervals and use proper language

RECOMMENDATIONS

1. *P* values should be discouraged wherever possible except under exceptional circumstances and consistent with current guidance from biomedical journals

Rationale: Although it is tempting to allow reporting of *P* values, since they have been an historical feature of biomedical reporting in claims of effectiveness, there is abundant evidence that *P* values are often misunderstood by consumers, are uninformative for clinical decision making and have the potential to mislead even those consumers who understand them, because they do not provide information about wither the size or precision of the effect. Leading epidemiologists have discouraged their use and biomedical journals in which drug advertisements appear have adopted policies to discourage their use. If confidence intervals are adopted, eliminating *P* values has no downside and may actually provide better information for decision making to consumers.



2. Confidence intervals should be encouraged instead of *P* values wherever possible and consistent with current guidance from biomedical journals. PAAB should suggest only 95% Confidence Intervals (CI) are appropriate for the presentation of findings rather than *P* values.


Rationale: Confidence intervals provide better information to consumers since they separate the size and precision of effect. 95%CI can always and quite easily be calculated from *P* values and estimated differences, so there is no reason for this information to not be available. Although confidence intervals can overestimate precision and may not adequately capture systematic bias, neither do *P* values. As more sophisticated techniques to adjust for bias become available, confidence intervals can be adjusted accordingly.

3. Publication of information from clinical trials should be discouraged if research protocols and outcomes have not been registered and are readily accessible by PAAB and the health care providers that they serve. PAAB should additionally mandate manufacturers provide a link to the registered information in advertisements AND endorse the Ottawa statement

Rationale: Endorsing the Ottawa Statement is recognition that public availability of information about all trials in healthcare is essential to ensuring ethical and scientific integrity in medical research. It is consistent with PAAB Values. There is no manner of validating a study's findings from reporting information related to statistical tests. To properly understand and critically interpret the validity and reliability of a study, providers must also be given information about how a trial was planned and executed. Although providing this level of detail in an ad is not feasible, some information directing the consumer to trial registration information would require fewer than 10 words and would ensure providers are not misled. Trial registration and its endorsement have become increasingly popular. As the vast majority of large pharmaceutical companies already register their clinical trials, and registration is required by leading biomedical journals, this additional requirement should not be viewed as onerous.

4. The wording of PAAB Code requirement 4.2 needs correction and revisiting.

Rationale: In its current format, the wording in Code requirement 4.2 "Statistics must be presented so as to accurately reflect their validity, reliability and level of significance." does not accurately reflect the intent of statistical null hypothesis testing. Specifically, the terms "validity" and "reliability" require re-visiting. Validity is the degree to which a study is actually able to make the claim, something that requires information about its design and execution. Statistics cannot be presented to reflect validity, but rather statistics can be presented to



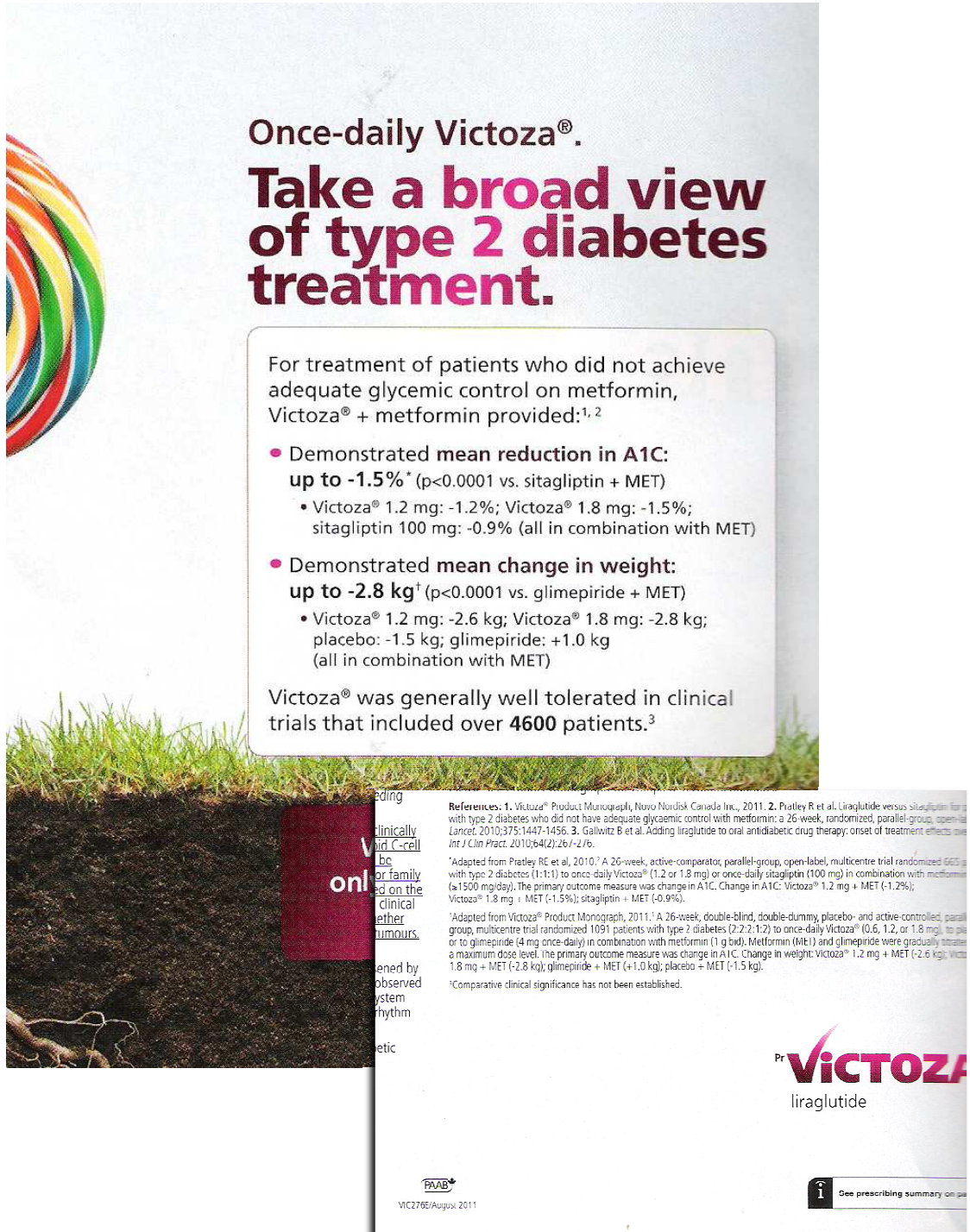
aid in interpreting validity. Reliability generally implies repeatability or consistency – something which cannot be addressed when findings from an individual trial are reported. In addition to being specific to frequentist statistics (see below), the level of significance and associated statistical tests (i.e., *P* Values) have a large potential for misleading, as already stated above. Changing the wording can re-align the intent of this requirement with what was its original intent.

5. PAAB should revisit Code requirement 4.2 and make additional provisions in the Code Explanatory Notes that Bayesian statistical testing is acceptable.

Rationale: In its current format, the wording in Code requirement 4.2 “Statistics must be presented so as to accurately reflect their validity, reliability and level of significance.” does not accurately reflect the future use of Bayesian statistics. References to frequentist statistics (i.e., level of significance) should be removed and an explanatory note should be developed to suggest how therapeutic benefit based on Bayesian statistics should be expressed (i.e., in 95% credible intervals). By doing this, PAAB is preparing itself for future claims which may be based on Bayesian statistical testing. The downside to allowing this is that consumers may not adequately understand what the result from Bayesian statistical tests mean.

ILLUSTRATIVE EXAMPLE

Advertisement appearing November, 2011 issue of Canadian Family Physician Journal



Once-daily Victoza®.
Take a broad view of type 2 diabetes treatment.

For treatment of patients who did not achieve adequate glycemic control on metformin, Victoza® + metformin provided:^{1, 2}

- **Demonstrated mean reduction in A1C: up to -1.5%*** (p<0.0001 vs. sitagliptin + MET)
 - Victoza® 1.2 mg: -1.2%; Victoza® 1.8 mg: -1.5%; sitagliptin 100 mg: -0.9% (all in combination with MET)
- **Demonstrated mean change in weight: up to -2.8 kg†** (p<0.0001 vs. glimepiride + MET)
 - Victoza® 1.2 mg: -2.6 kg; Victoza® 1.8 mg: -2.8 kg; placebo: -1.5 kg; glimepiride: +1.0 kg (all in combination with MET)

Victoza® was generally well tolerated in clinical trials that included over **4600 patients.**³

References: 1. Victoza® Product Monograph, Novo Nordisk Canada Inc., 2011. 2. Prallely R et al. Liraglutide versus sitagliptin for patients with type 2 diabetes who did not have adequate glycaemic control with metformin: a 26-week, randomized, parallel-group, open-label trial. *Lancet*. 2010;375:1447-1456. 3. Galkwitz B et al. Adding liraglutide to oral antidiabetic drug therapy: onset of treatment effects over time. *J Clin Pract*. 2010;64(2):267-270.

*Adapted from Prallely RC et al, 2010.² A 26-week, active-comparator, parallel-group, open-label, multicentre trial randomized 660 patients with type 2 diabetes (1:1:1) to once-daily Victoza® (1.2 or 1.8 mg) or once-daily sitagliptin (100 mg) in combination with metformin (≤1500 mg/day). The primary outcome measure was change in A1C. Change in A1C: Victoza® 1.2 mg + MET (-1.2%); Victoza® 1.8 mg + MET (-1.5%); sitagliptin + MET (-0.9%).

†Adapted from Victoza® Product Monograph, 2011.¹ A 26-week, double-blind, double-dummy, placebo- and active-controlled, parallel-group, multicentre trial randomized 1091 patients with type 2 diabetes (2:2:2:1:2) to once-daily Victoza® (0.6, 1.2, or 1.8 mg), to placebo or to glimepiride (4 mg once-daily) in combination with metformin (1 g bid). Metformin (MET) and glimepiride were gradually titrated to a maximum dose level. The primary outcome measure was change in A1C. Change in weight: Victoza® 1.2 mg + MET (-2.6 kg); Victoza® 1.8 mg + MET (-2.8 kg); glimepiride + MET (+1.0 kg); placebo + MET (-1.5 kg).

¹Comparative clinical significance has not been established.

Pr VICTOZA
liraglutide

PAAB
VIC2766/August 2011

1 See prescribing summary on page 1



Applying the recommendations 1-3 would change an advertisement as follows

Changes in Statistical reporting (recommendations 1 and 2)

For treatment of patients who did not achieve adequate glycemic control on metformin, Victoza® + metformin provided:^{1, 2}

- **Demonstrated mean reduction in A1C: up to -1.5% (95%CI: -1.63 to -1.37, n=218)**
 - Victoza® 1.2 mg: -1.2% (95%CI: -1.37 to -1.11, n=221)
 - Victoza® 1.8 mg: -1.5% (95% CI -1.63 to -1.37, n=218)
 - sitagliptin 100 mg: -0.9% (95%CI: -1.03 to -0.77, n=219)
 - (all in combination with MET)

References and PAAB Logo (Recommendation 3)

...ing
...linically
...bid C-cell
...be
...or family
...ed on the
...clinical
...ether
...tumours.
...ened by
...observed
...ystem
...hythm
...etic

... References: 1. Victoza® Product Monograph, Novo Nordisk Canada Inc., 2011. 2. Pratley R et al. Liraglutide versus sitagliptin for...
... with type 2 diabetes who did not have adequate glycaemic control with metformin: a 26-week, randomized, parallel-group, open-label...
... Lancet. 2010;375:1447-1456. 3. Gallwitz B et al. Adding liraglutide to oral antidiabetic drug therapy: onset of treatment effects...
... Int J Clin Pract. 2010;64(2):267-276.

*Adapted from Pratley RE et al, 2010.² A 26-week, active-comparator, parallel-group, open-label, multicentre trial randomized 665...
... with type 2 diabetes (1:1:1) to once-daily Victoza® (1.2 or 1.8 mg) or once-daily sitagliptin (100 mg) in combination with metformin...
... (≥1500 mg/day). The primary outcome measure was change in A1C. Change in A1C: Victoza® 1.2 mg + MET (-1.2%);
... Victoza® 1.8 mg + MET (-1.5%); sitagliptin + MET (-0.9%).

*Adapted from Victoza® Product Monograph, 2011.¹ A 26-week, double-blind, double-dummy, placebo- and active-controlled, paral...
... group, multicentre trial randomized 1091 patients with type 2 diabetes (2:2:2:1:2) to once-daily Victoza® (0.6, 1.2, or 1.8 mg), to pi...
... or to glimepiride (4 mg once-daily) in combination with metformin (1 g bid). Metformin (MET) and glimepiride were gradually titrate...
... a maximum dose level. The primary outcome measure was change in A1C. Change in weight: Victoza® 1.2 mg + MET (-2.6 kg); Victo...
... 1.8 mg + MET (-2.8 kg); glimepiride + MET (-1.0 kg); placebo + MET (-1.5 kg).

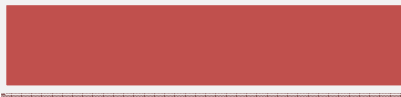
¹Comparative clinical significance has not been established.

This trial is registered with ClinicalTrials.gov, number [NCT00700817](https://clinicaltrials.gov/ct2/show/study/NCT00700817)

VICTOZA
liraglutide

PAAB endorses the Ottawa Statement on trial registration
VIC276E/August 2011

See prescribing summary on p...



FINDINGS FOR SECTION 3.1 – CLAIMS, QUOTATIONS AND REFERENCES

The explanatory notes to Section 3.1 of the PAAB Code suggest that “Clinical/therapeutic claims must be based on published, well-controlled and/or well-designed studies with clinical and statistical significance clearly indicated. Publication in peer-reviewed journals is usually a good criterion for establishing scientific rigor. Review articles, pooled data, meta-analysis and post-hoc analysis are generally regarded as not being high-level evidence to support claims in drug advertising.”

In the next section, we will attempt to answer the following questions related to the application of the Code. Specifically, we will focus on the potential bias that could result from reliance on review articles/ meta-analysis, unpublished studies, post-hoc analysis/secondary outcomes/subgroup analysis, and observational studies to establish claims of effectiveness.

SHOULD REVIEW ARTICLES, POOLED DATA AND META-ANALYSIS BE USED TO SUPPORT CLINICAL/THERAPEUTIC CLAIMS OF EFFECTIVENESS?

INTRODUCTION

Review articles of clinical effectiveness are conducted in an attempt to use information beyond a single study to understand the degree and likelihood of effectiveness of a health product. Review articles can be conducted with varying degrees of transparency and rigour. The various approaches to clinical review and terminology are outline here for the sake of clarity.

TERMINOLOGY

The term ‘systematic’ review refers to a review where studies are identified using selection criteria that are systematically applied to identify relevant studies. The studies considered relevant are used to draw conclusions about effectiveness. These identified studies can be thought of as admissible evidence (80). Methods for systematic reviews have evolved in an attempt to improve replicability of reviews and reduce the introduction of a selection bias, where an author can deliberately (or accidentally) pick studies for a review to be consistent with his beliefs or to influence a conclusion (81,82). Publication of systematic reviews has grown since the late 1970s, following their introduction in the social sciences (83). It is estimated that the number of systematic reviews published daily averages 11 and continues to grow (84). (Figure 4)

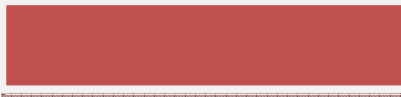
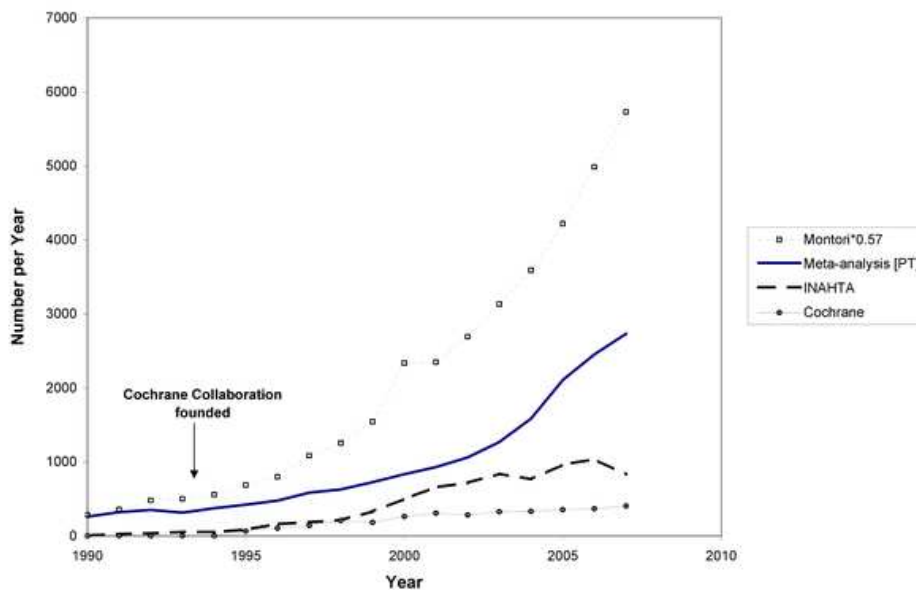



FIGURE 4: NUMBER OF SYSTEMATIC REVIEWS AND META-ANALYSIS CONDUCTED YEARLY (FROM (31))



The terminology surrounding these types of studies can be confusing and has evolved along with the methods used to conduct them. The term ‘narrative’ review is now used to distinguish between a literature review where evidence was not identified systematically versus one where it is (i.e., a systematic review). This has also been labeled an ‘informal’ review, a ‘traditional’ review, a ‘literature’ review or a commentary.(85) The term ‘scoping’ review has been used to identify a study where evidence is identified in a rigorous and transparent way, but not necessarily synthesized to form a single conclusion.(86)

At its origin, the term ‘meta-analysis’ was defined as an “analysis of analyses” and the “statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings.”(83,87) The terms ‘meta-analysis’ and ‘research overview’ have often been used synonymously with the term ‘systematic review’ in the biomedical literature(47,88). The term ‘meta-analysis’ is now reserved to describe the combination of quantitative information through statistical methods, consistent with its original definition. However, meta-analysis has also expanded to the concept of pooling individual (or primary) patient data across studies, rather than only combining data that has already been analyzed (87)³ The term ‘overview’ is now much less commonly used.

³ Meta-analysis was originally proposed as an alternative to pooling individual patient data; not because this approach lacked merit, but because it was becoming increasingly difficult to locate these data in the field of educational research. In his seminal publication, Glass acknowledges the merits of pooling data from individual patients.



Data ‘pooling’ is also used synonymously with the term ‘meta-analysis’; it refers to combining quantitative data using statistical methods. Pooling is sometimes implied to mean a less careful combination of data than meta-analysis. This is because data can be pooled using approaches not appropriately grounded in statistical theory. The term network meta-analysis refers to a type of meta-analysis used to make comparative claims and will be covered in a later section (Comparative Claims).


Review and meta-analysis are distinct concepts, so the following section will focus on them separately. A body of evidence can always be reviewed, whether or not it is feasible or appropriate to conduct meta-analysis of the studies identified is a separate question.⁽⁸⁹⁾ Firstly, we will look at the potential impact from the use of review articles versus the use of individual studies. Secondly, we will look at the potential impact of using meta-analysis.

EVIDENCE

REVIEW ARTICLES

As already mentioned, there are various types of review articles, ranging from narrative reviews that lack explicit selection criteria and rely heavily on anecdote and clinical experience to well-conducted systematic reviews that are replicable, transparent and rigorous. There is evidence to suggest that compared to systematic reviews, narrative reviews have the potential to mislead, either because they carry the real potential for bias (from selection), or the findings reflect the subject experts background (i.e., interpretation bias).^(81,82,90) Narrative reviews have also been shown to carry the very real threat of unintentionally misleading clinicians through omitting important information.^(91–95) In the 1990s, Harvard researchers provided evidence from the treatment of myocardial infarction that traditional narrative reviews failed to capture benefit from therapy, where a systematic review revealed compelling evidence of a benefit.^(94,95) Because the intentional or unintentional omission of relevant information can lead to dramatically different conclusions regarding clinical effectiveness, we can conclude that review articles not conducted systematically have the real potential to mislead.

Once a systematic review has been conducted, there remains the problem of integrating and synthesizing findings from different studies to draw overarching conclusions. There are two types of approaches and various hybrid approaches to doing this.⁽⁸⁵⁾ First, the authors may scrutinize each study individually using a defined framework, then, draw conclusions from one or more of the studies identified that appear to be the most salient, usually based on validity and transferability to a clinical context – this has been termed a ‘methodologic’ or ‘critical appraisal’ approach to review.⁽⁸⁵⁾ A second approach is to statistically



combine the findings from one or more studies from the review, using accepted statistical methods (meta-analysis).⁴ Systematic review coupled with meta-analysis will be discussed in the next section.


The relevant question is whether the findings of a methodologic systematic review can be misleading compared with arbitrarily selected “published, well-controlled and/or well-designed studies? Single studies can harbor several limitations.(50) Small studies are associated with larger random error from sampling variability so are more likely to produce a false negative result – problematic when claims of similarity or comparability are being made.(50) This can also happen because the number of patients in a trial is inadequate,(96) and achieving large sample sizes is difficult or prohibitively expensive. Although a single trial may be adequately powered (and well-conducted and designed) to detect a more immediate, or surrogate outcome, it may also not have an adequate period of follow-up to detect the effect of a therapy on less-frequently occurring outcomes such as death and serious morbidity. There may also be a lack of clear evidence to support the surrogate outcome. Because of this, larger, more adequately powered studies of clinical effectiveness, called mega-trials, overcome many of these shortcomings.(97)

It is still possible for two well-conducted systematic reviews, like two well-conducted trials, to arrive at different conclusions. This is because they may be addressing different clinical questions, have been conducted at different times, have applied different selection criteria, or have searched for studies in a different fashion.(98) Nonetheless, unlike discrepant trial results, the reasons for the differences should be plainly obvious to a reader of the reviews. Still, it highlights the fact that a systematic review has the power to mislead if the criteria used to identify studies or determine what studies are most relevant to drawing conclusions are arbitrary or unjustified.⁵ This same limitation applies to the use of single studies, since the single study may have been chosen for equally suspect reasons. The difference is the systematic reviewer must reveal their preferences to the reader.

What distinguishes systematic reviews from the arbitrary choice of single studies is that the consumer will be aware of what other studies might be helpful in answering a specific clinical question. This will help the consumer understand whether the results have been repeated successfully or are in line with the findings from other studies. The consumer is now provided with additional

⁴ There are also unacceptable statistical approaches to integrating study findings, including combining based on *P*-values or *vote counting*.

⁵ It is assumed that the systematic review only identifies those studies that are consistent with indications, dosage regimens, efficacy and safety information contained in the Health Canada Terms of Market Authorization and that these criteria are explicit.




information with regards to how and why a single study was selected to establish clinical effectiveness. It can help them understand whether focusing on a single study is appropriate.

Systematic reviews are more consistent with the PAAB Values of “transparency” (23) and the conduct of regulatory bodies (in terms of considering the totality of evidence) in general. They have the additional advantage of mitigating the introduction of a selection bias, compared to the use of single studies. Systematic reviews have become best practice for the assessment of clinical effectiveness in health care. Taken together, we have to conclude that the use of even a well-conducted single trial has the capacity to mislead more than a well-conducted systematic review addressing the same question of clinical effectiveness.

META-ANALYSIS

Meta-analysis refers to the combination of data from separate studies or experiments. Methods to combine information from different datasets were developed in fields outside of medicine in the early 19th century and the first example in medical research was a reported analysis of the effectiveness of military inoculation for typhoid (enteric) fever in 1904.(99) However, growth in the use and publication of meta-analysis began in the 1980s. A search for published meta-analysis in 1987 found only 13 meta-analyses published in the 1970s and 69 meta-analyses between 1980 and 1986.(100) In the 1990s, considerably more meta-analyses appeared. A survey of 8 leading journals found 272 meta-analyses published between 1993 and 2002.(101)


By the 1990s, the topic of meta-analysis and its potential deficiencies received considerable attention in leading biomedical journals.(90,100,102–108) The first and most important practical issue for meta-analysis is deciding what studies should be combined. To avoid drawing misleading conclusions from an arbitrary selection of studies (and introducing a selection bias), many authors have suggested that meta-analysis should only be conducted in the context of a systematic review.(89,109) Even adopting this approach, does not necessarily lead to a consistent population of studies because what is included depends on how the problem is framed. For example, a systematic review intended to meta-analyze the effect of *adalimumab* on *patients with rheumatoid arthritis* should lead to a larger collection of studies than a more narrowly focused study meta-analyzing the effect of a specific dosing regimen of *adalimumab* in *females with rheumatoid arthritis who have failed at least two drugs*. The latter may be much more relevant for clinical decisions (and reimbursement) while the former may be academically interesting or relevant to understand opportunities for future research.



But even if a systematic review is conducted to identify studies for meta-analysis, there are many more opportunities to introduce a bias that would lead to misleading conclusions from additional steps in the analysis.(110) Meta-analysis is not as straightforward as entering the results from all identified studies into a calculator to arrive at an overall measure of effectiveness. There is a risk of introducing bias from combining studies that vary in terms of how patients were selected, important population characteristics, how interventions were delivered, in what context, how studies were designed and executed, and importantly, the assumptions underlying the statistical model used to combine information.(111) A central issue in meta-analysis is how we to draw conclusions from studies that have multiple sources of variation and lead to different results. Although there have been some attempts to do this, including the use of statistical models that assume elements of randomness can explain variation, they may not be adequate (106) Because of this, one prominent epidemiologist has suggested that meta-analysis should not be viewed as a means of estimating a single effect, but rather, “a method for identifying the sources of disparity and conflict among studies”(112) Meta-analysis may also exaggerate findings since unpublished studies or outcomes are more likely to be negative(113) Taken together, all of these factors can contribute to misleading results, specifically when attempts are made to detect small effects on health.(114) A partial list of problems identified in the last two decades appears in Box 2.

BOX 2: POTENTIAL PROBLEMS WITH META-ANALYSIS.

- Difference in inclusion and exclusion of meta-analysis
- Identical selection criteria - differences in baseline characteristics across studies
- Certain design features in trials are associated with both increased random noise and a tendency to exaggerate treatment effects
- Variability in intervention/control (e.g., dose, timing, formulation)
- Variability in management , response to intermediate outcomes, patient care settings
- Variation in quality and design of execution of studies(115)
 - Use of RCTs only does not address all clinical questions(116)
 - How to incorporate quasi-experimental design(117)
 - How to incorporate observational studies(118)
 - Use of abstracts, unpublished and inadequately reported information(119,120)
 - Use of small trials(121)
- Variation in analysis from trials (how dropouts were handled etc.)
 - Interpretation bias(122)
 - Type II errors (123)
 - Use of standardized effects(124)


- 
- Addressing heterogeneity(125,126)
 - Industry funding leads to exaggerated results(127,128)
 - Harm is not analyzed consistently (129)
 - Longitudinal Data(130)
 - Quality of Life data(131)
 - Unpublished data not incorporated(132)

Beyond theoretical concerns, important empirical evidence began to emerge in the 1990s that compared the results of meta-analysis with the results of large randomized controlled trials. Concerns were raised when discrepancies occurred between findings of meta-analysis and large trials.(89,133–138) Critics of meta-analysis argued meta-analysis had the real potential to mislead as discrepancies beyond chance were reported up to 23% of the time(138), while defenders of meta-analysis suggested these analyses had flaws - there is actually little discrepancy if you compare appropriately(139). Further defense came from another study demonstrated that large randomized trials disagree with each other at a similar proportion to large trials disagreeing with meta-analyses(140). Debate continues as to whether a single trial or multiple trials is more appropriate for estimating clinical effectiveness.(50,141)

To respond to questions about the reliability of meta-analysis, new methods and processes for their conduct have been proposed. Many have argued that findings from meta-analysis are reliable if the meta-analysis is conducted well.(85,138,142–146) A considerable effort has been spent on creating quality measures and checklists so that only studies judged to be “high quality” would be used for meta-analysis. The argument was that the quality of the meta-analysis depended on the quality of studies used. This was shown to be problematic as different checklists produced different results.(147)

Statistical methods were proposed by some to overcome the problems in reliability introduced by study design(148), trial heterogeneity(125,149,150), underpowered analyses(151) and use of small trials(152). Others have commented that statistics will not overcome fundamental problems with meta-analysis (153). Others have suggested novel meta-analytic approaches to aiding clinicians and researchers to understand what is known(154,155) and problems with these new approaches have been identified.(156) Many have emphasized that meta-analysis of individual patient data, which removes many of the statistical assumptions that can lead to misleading results, is a superior form of analysis and should be encouraged.(157,158)

Despite these criticisms of meta-analysis, the use and conduct of systematic review-based meta-analysis continues to be widely promoted for the purpose of supporting clinical decision making and clinical practice guidelines. Graphs and



displays of meta-analysis have been devised with the intent of giving the reader a complete picture of information, including important factors that could produce misleading results, such as precision, consistency, potential for publication bias, and direction. Even with visual aid, studies have shown that physicians are likely unable to properly interpret meta-analytic plots(159), and meta-analysts need to be highly selective in deciding how to depict data because of poor reproducibility of graphic displays (160). The interpretation and conclusions drawn from meta-analysis appears to be highly subjective, even among experienced researchers (97), even if the methodology is rigorous.

Given some of the pitfalls of meta-analysis, proponents of the use of evidence in medical decision making have de-emphasized the need for meta-analysis, emphasizing more qualitative examination of the totality of available evidence (from a systematic review) in terms of its risk of bias (including publication bias), consistency, directness, and precision.(89,161,162) In total, it appears concerns about the conduct and interpretation of meta-analysis suggest it has the real potential to mislead consumers, even if conducted well. If anything, meta-analysis performed appropriately using all relevant patient-level data appears to represent the least opportunity for biased estimates of effectiveness.

SHOULD UNPUBLISHED STUDIES BE USED TO SUPPORT CLINICAL/ THERAPEUTIC CLAIMS OF EFFECTIVENESS?

INTRODUCTION

Peer review is a widely accepted system of audit both within and outside the biomedical science and has been widely adopted by biomedical journals. It is intended to promote consistency and reliability of research findings.(163) However, it lacks a strict operational definition or universally accepted standards of conduct.(164) At its most basic, an editor of a peer-reviewed journal will accept a manuscript from authors hopeful for submission. The editor will assign the manuscript to individuals who are assumed to be knowledgeable about the subject matter of the article. The reviewers will provide comments to authors to help improve the report and pass judgment as to whether the article is acceptable for publication. With rare exception do peer reviewers have access to the original data sets or study protocols.(165) As veteran editor for the British Medical Journal observed, this system, “is little better than tossing a coin, because the level of agreement between reviewers on whether a paper should be published is little better than you'd expect by chance”.(166)

These sentiments are shared by experts in biomedical peer review. One long-time editor, Drummond Rennie, Editor-in-Chief of the Journal of the American Medical Association, noted that peer review, like democracy, is more faith than science and despite its flaws, the best system there is.(167) He observed:

" the healthy result that more people feel involved, more are educated about the professional values that undergird the system, and the widespread paranoia that exists when authors offer up their precious reports to the tender mercies of journal editors may be lessened."(167)


Because peer review is intended to promote scientific rigour, it may seem intuitive to use only literature which has undergone peer review when drawing conclusions about the clinical effectiveness of a therapeutic product. However, the negative consequences of restricting an analysis to only published information were first identified in the late 1970s. Called the 'file drawer' problem(168), one observer noted that authors and journal editors are generally more interested in original contribution to knowledge in the form of positive results, and studies with inconclusive or negative findings would be less likely to be published, since investigators were less likely to submit them. At its extreme, a reviewer might most only have access the most positive findings (some of which would be false positives) and negative studies would forever remain in the 'file drawer' of investigators.

Since that time, considerable attention has been paid to the potential merits and failings of using only published information. Proponents of using published information have suggested that unpublished studies are more likely to be small and prone to error, and will not reasonably affect conclusions drawn about therapeutic effectiveness. Critics have suggested that in addition to studies being left in the file drawers of investigators, there are other, equally serious problems that can threaten the validity of research. Bias may result from a number of aspects related to publishing – for example, study reports may be published but harder to find (dissemination bias), or they may be easy to find but omit important information (outcome bias). They may also be in the process of being published, but currently unavailable (time lag bias). A complete list of potential bias related to publication appears below. (Box 3)

BOX 3: POTENTIAL BIAS RELATED TO DRAWING CONCLUSIONS FROM ONLY PUBLISHED INFORMATION (ADAPTED FROM SONG(169))

Dissemination bias: Occurs when the dissemination profile of a study's results depends on the direction or strength of its findings. The dissemination profile is defined as the accessibility of research results or the possibility of research findings being identified by potential users. The spectrum of the dissemination profile ranges from completely inaccessible to easily accessible, according to whether, when, where and how research is published.

Publication bias: Occurs when the publication of research results depends on the nature and direction of the results. Because of publication bias, the results of



published studies may be systematically different from those of unpublished studies.

Outcome reporting bias: Occurs when a study in which multiple outcomes were measured reports only those outcomes that were significant.

Time lag bias: Occurs when the speed of publication depends on the direction and strength of the trial results. For example, studies with significant results may be published earlier than those with non-significant results.

Grey literature bias: Occurs when the results reported in journal articles are systematically different from those presented in reports, working papers, dissertations or conference abstracts.

Full publication bias: Occurs when the full publication of studies that have been initially presented at conferences or in other informal formats is dependent on the direction and/or strength of their findings.

Language bias: Occurs when languages of publication depend on the direction and strength of the study results.

Multiple publication bias (duplicate publication bias): Occurs when studies with significant or supportive results are more likely to generate multiple publications than studies with non-significant or unsupportive results. Duplicate publication can be classified as 'overt' or 'covert'. Multiple publication bias is particularly difficult to detect if it is covert, when the same data are published in different places or at different times without providing sufficient information about previous or simultaneous publication.


Place of publication bias: Defined as occurring when the place of publication is associated with the direction or strength of the study findings. For example, studies with positive results may be more likely to be published in widely circulated journals than studies with negative results. The term was originally used to describe the tendency for a journal to be more enthusiastic towards publishing articles about a given hypothesis than other journals, for reasons of editorial policy or readers' preference.

Citation bias: Occurs when the chance of a study being cited by others is associated with its result. For example, authors of published articles may tend to cite studies that support their position. Thus, retrieving literature by scanning reference lists may produce a biased sample of articles and reference bias may also render the conclusions of an article less reliable.

Database bias (indexing bias): Occurs when there is biased indexing of published studies in literature databases. A literature database, such as MEDLINE or EMBASE, may not include and index all published studies on a topic. The literature search will be biased when it is based on a database in which the results of indexed studies are systematically different from those of non-indexed studies.

Media attention bias: Occurs when studies with striking results are more likely to be covered by the media (newspapers, radio and television news).

Omission of information that is misleading to consumers is also central to questions of deception in advertising.(30) In the next section, a description of the evidence addressing the relevant question of interest will be presented –



namely, what is the potential impact of allowing only published evidence versus allowing published and unpublished evidence on clinical/therapeutic claims of effectiveness?

EVIDENCE

EFFECTIVENESS OF PEER REVIEW


The most rigorous attempt to evaluate the effectiveness of peer review in the biomedical sciences was first reported in the Journal of the American Medical Association in 2002 (170) and more recently updated in 2007.(171) Despite an exhaustive search, the authors were not able to find compelling evidence to refute or confirm that peer review improved the validity of research findings. Instead, a single study (172) was identified that suggested very little real improvement and two additional studies were identified that suggested peer review resulted in better reporting.(173,174)

Studies have also been conducted to test the ability of peer review to discover major design and conduct flaws. Major errors were intentionally inserted into submitted manuscripts by editors to see if peer reviewers would detect them.(175,176) The results revealed reviewers seldom spotted errors with most reviewers identifying only 25% of major errors and no reviewer identifying all errors. Improvement in detection could not be altered by better training or improvements to the review system.(175,177) A recent report by the UK Science and Technology Committee acknowledged the lack of “solid evidence on the efficacy of pre-publication editorial peer review”, but identified its many other important facets, including career development and training of future researchers.(178)

BIAS FROM USING ONLY PUBLISHED FINDINGS

The theory that negative studies exist in the file drawers of investigators was confirmed in studies conducted in the late 1980s. In a landmark study, 156 investigators responded to a survey and confirmed that over 25% (271/1041) of the randomized controlled trials they had been involved with were unpublished. Of 178 of these that had been completed, 14% (26/178) had been positive versus 55% (423/767) of their published counterparts.(179) Subsequent studies confirmed this phenomenon (169) and revealed that this happens even when studies are published in very high proportion.(180) Other research revealed even well-designed and well-conducted studies are not published almost half of the time (181)

Further investigation revealed the problem did not depend on study design and stemmed from investigators not submitting trial reports, rather than the fault of editors who would not accept them. (169)(182) There is a growing body of evidence suggesting investigators may be less motivated to submit due to



pressure from research sponsors, instruction from journal editors, and requirements of the research award system(169).

Several studies have attempted to quantify the effect of these unpublished trials on estimates of effectiveness (183). A landmark study showed exaggerated estimates of effectiveness (i.e., a bias) when information outside of the realm of biomedical journal reports was ignored.(132) In a comprehensive review of bias stemming from publication, Song and colleagues confirmed that despite some caveats about the studies used to assess bias(184,185), “There is consistent empirical evidence that the publication of a study that exhibits statistically significant or ‘important’ results is more likely to occur than the publication of a study that does not show such results”.

More recently, the effects of other important potential sources of bias related to publication have also been empirically demonstrated. A large body of evidence has demonstrated the prevalence and effect of outcome reporting bias (186,187) including Canadian studies.(188) Harm outcomes tend to be underreported although evidence of significant bias from this phenomenon is lacking.(189)

Suggestions to combat this phenomenon include the setting up of trial registries and the advance publication of detailed protocols with an explicit description of outcomes and analysis plans and the use of more comprehensive search strategies.(169,186) Methods have also been developed to detect and adjust for exaggerated estimates of effectiveness from publication and outcome bias.(190–201)

SHOULD SECONDARY OUTCOMES, SUBGROUP ANALYSIS, AND POST-HOC ANALYSIS BE USED TO SUPPORT CLINICAL/ THERAPEUTIC CLAIMS OF EFFECTIVENESS?

INTRODUCTION

Clinical and therapeutic claims of effectiveness must be contextualized according to several important variables. Firstly, treatments must be tailored to the type of patient requiring therapy. This can be done by taking into consideration demographic, anatomic, biologic, genetic, prognostic, or pathophysiologic characteristics. Secondly, the intervention needs to be defined according to formulation, setting, frequency, and dose. A counterfactual, namely, care without the new health product must also be defined similarly. Finally, an outcome relevant to the patient must be satisfied, whether this is a measure of the disease burden, harm from the intervention, satisfaction with care or quality of life.



To substantiate these claims, there should be compelling scientific evidence demonstrating that a desired outcome occurred more often in these types of patients from this type of intervention compared to similar patients without the intervention. Human experiments, rigorously conducted, such as large, double-blinded randomized controlled trials with broad eligibility criteria, have been developed as a reliable method of providing compelling evidence of such clinical claims of effectiveness.

The findings from these experiments, however, may offer additional information that is of interest to clinicians. Firstly, patients may be observed to greatly benefit according to measured outcomes not of primary interest (i.e., a secondary outcome) to trial investigators or for which the trial was designed to measure. Secondly, some identifiable subgroups of patients may be shown to benefit to a much larger or smaller degree than the average effect seen across the trial. Thirdly, the secondary outcome or subgroup may be identified through an analysis completed during or after the completion of the trial.

Relevant to the overarching question of scientific information in advertising is if these types of analyses (subgroup analyses, secondary outcomes, post-hoc analyses) can be made reliable for clinical decision making and the degree to which they can mislead consumers.

EVIDENCE

SUBGROUP ANALYSES

Subgroup analyses in randomized controlled trials are commonly conducted(202–208). International standards for the conduct of subgroup analysis for clinical trials have been published and suggest that unless these analyses are appropriately planned and conducted, they should be considered exploratory (209). The guidance also states that conclusions based solely on exploratory analyses are “unlikely to be accepted.”(209).

Original guidance to aid clinicians in the interpretation of subgroup analyses for making clinical decisions based on existing evidence have been published by individual authors and widely adopted (210,211). Limitations in the applicability of these guidelines have led to a recent update (202). Criteria for examining the credibility of a claim based on an analysis of subgroups based on these previous suggestions are listed in Box 4.

BOX 4: CRITERIA TO ASSESS THE CREDIBILITY OF SUBGROUP ANALYSES (202)

DESIGN

- Is the subgroup variable a characteristic measured at baseline or after randomisation?*

- Is the effect suggested by comparisons within rather than between studies?
- Was the hypothesis specified a priori?
- Was the direction of the subgroup effect specified a priori*
- Was the subgroup effect one of a small number of hypothesised effects tested?

ANALYSIS

- Does the interaction test suggest a low likelihood that chance explains the apparent subgroup effect?
- Is the significant subgroup effect independent?*


CONTEXT

- Is the size of the subgroup effect large?
- Is the interaction consistent across studies?
- Is the interaction consistent across closely related outcomes within the study?*
- Is there indirect evidence that supports the hypothesised interaction (biological rationale)?

*New criteria.

There is considerable evidence to suggest subgroup analyses are frequently improperly conducted, reported and interpreted(203,204,206,207,212) and the extent of the problem is likely underestimated. (213) For example, in a review of 63 randomized controlled trials in cardiovascular medicine with a median of 496 patients, one group of investigators found tests for interaction (an appropriate method to analyze subgroups) were only performed in 30% of trials. (206)In this same study , 40% of trialists gave equal emphasis to overall results and findings from subgroup analysis.(206) Real effects can be missed because the original studies were not designed to detect them (i.e., false negatives)(204), and identified effects can be false because of multiple testing and natural within-trial variability (207,214–216) Industry–funded trials have been very recently demonstrated to harbor subgroup analyses that are even more prone to bias than their publicly-funded trial counterparts.(217)

An additional problem with subgroup analyses is that they are susceptible to bias. That is, the comparability of patients provided by randomizing patients into two separate groups is lost when subgroups are analyzed since these subgroups are not randomized in the same way. One solution to overcome this problem have been to use a stratified randomization procedure, but this increases the number of patients and costs of conducting trials and is not commonly employed. Taken together, despite some benefits from conducting these analyses (40,62), it is likely that the care and nuance required to conduct them




means the routine use of subgroup analyses in advertising harbors a great potential to mislead consumers, and lead to suboptimal care.

SECONDARY OUTCOMES

It is common for trials to have multiple outcomes. Although the prevalence of this has not been studied, one observer noted trials are “major undertakings for sponsors and investigators... It would be odd for a single outcome to encompass all that interests investigators.”(218) If a trial has a single endpoint, interpretation of statistical significance from a test is less complicated – “the significance level truly represents the probability of rejecting the null hypothesis of no treatment difference when it is in fact true (the Type I error).”(204) The chance of a Type I error increases with the introduction of secondary outcomes and is directly proportional to the number of outcomes in the trial and their relationship to each other. For example, if 10 unrelated outcomes were tested at a 0.05 significance level, the chance of having one false positive result would be almost 50%.(208) Type II errors, where no treatment difference is observed when in fact one exists, can also occur if the trial is not sufficiently powered to measure both the primary and secondary outcomes. An additional problem with secondary outcomes is that they may be less valid since their findings are not intended to create substantive arguments for effectiveness. Those conducting the trial may be less prepared to measure and record secondary outcomes, or the outcome may be an exploratory patient reported outcome or surrogate that has is yet to be fully validated.

Subgroups and secondary outcomes are two sides of the same statistical coin. The chance of introducing an error is similar whether we are interested in multiple outcomes in the same population or the same outcome in different subpopulations. Like subgroups, positive effects observed in a secondary outcome are more plausible if the direction and measurement are specified *a priori*, there are fewer outcomes tested, the effect is large and consistent with other similar outcomes (or subgroups) in the trial.

Many observers have cautioned that the results from testing multiple outcomes must make sense (204) For example, in a trial intended to reduce cardiovascular events, we would not expect a significant reduction in fatal heart attacks (a rare outcome) if the incidence of all heart attacks (fatal and non-fatal) were not reduced. For example, one group of investigators conducted subgroup analysis in a large RCT of a calcium antagonist in chronic heart failure. They were able to showed no reduction in mortality in patients expected to benefit (i.e., with ischemic cardiomyopathy, RR=1.04, 95%CI=0.83-1.29) but a major benefit in participants expected not to benefit (i.e., with non-ischemic cardiomyopathy, RR=0.64, 95%CI=0.37-0.79), (interaction $p=0.004$)(219). However, a subsequent



investigation through another trial failed to confirm the benefit in non-ischemic cardiomyopathy.(220)

In a similar example, a trial that randomized patients with heart failure to combination therapy with an angiotensin II receptor blocker (losartan) or the ACE inhibitor captopril showed a survival benefit in a subgroup of elderly patients.(221) This promising finding was not replicated in a subsequent, larger randomized trial with similar design. (222)

There are many proposed solutions for reducing the potential for bias from multiple testing ranging from trial conduct (223) to trial analysis (208,212,224,225) to reduce bias from multiple testing from secondary outcomes or subgroups. Because of the compelling evidence that the number of outcomes and pre-specified analyses are often obscured in trial reports (169,186,188), unless original protocols are available to help interpret findings, analyses of secondary outcomes (and subgroup analyses) have the great potential to be misleading.


POST HOC ANALYSES

Findings from *post hoc* (after the fact) analysis have been criticized for being similar to declaring the winner of a horse race after the fact. However, they may not be that bad - it is possible to further reduce susceptibility to bias from *post hoc* analyses through appropriate design and structuring of an experiment. For this reason, international guidelines for clinical trials have placed strong emphasis on the use of pre-specified protocols with statistical plans and blinded statistical reviewers when examining primary outcomes (209).

The problem is greatly exaggerated when multiple outcomes or subgroups are part of the trial design, which is almost certainly the case in every major clinical trial.(204,208) To demonstrate the problem of multiple testing of non-specified hypotheses, a Canadian study (226) was able to demonstrate that residents of Ontario born under the astrologic sign Leo had a significantly higher probability of being hospitalized for gastrointestinal hemorrhage compared to all other signs combined ($P = 0.0447$).⁶ In a similar study conducted on a large multi-centre trial (ISIS-2)(227), in which aspirin showed a large benefit over placebo for reducing cardiovascular events ($P < 0.00001$), an analysis of performance in patients born under 12 astrologic signs revealed that Geminis and Libras experienced higher rates of adverse effect ($9\% \pm 13\%$)(228)

As one observer noted “Post hoc observations are not automatically invalid (many medical discoveries have been fortuitous), but they should be regarded as

⁶ The investigators were subsequently able to eliminate many of the false positives through the use of further, appropriate, analytic methods.



unreliable unless they can be replicated.” (62) Taken together, it seems the perils and real potential for the introduction of bias from *post hoc* analysis can be very misleading.

SHOULD OBSERVATIONAL (I.E., NON-EXPERIMENTAL) STUDIES BE USED TO SUPPORT CLINICAL/THERAPEUTIC CLAIMS OF EFFECTIVENESS?

INTRODUCTION

In an ideal world, the best way to assess therapeutic effectiveness would be to accurately measure differences in health outcomes in two different realities where everything else is the same— in one reality people received a technologic intervention; in the other they don’t. Experimental studies attempt to approximate this theoretical, counterfactual world, by allowing the investigator to fully control and observe who is exposed to treatment versus who is not.


Experimental study designs include n-of-1 trials, controlled parallel group trials, quasi-randomized controlled trials, and randomized trials. Experimental study designs minimize the introduction and influence of important variables that can affect the final measured outcome, specifically variables related to why a person might choose a therapy to begin with. They also allow more reliable estimations of whether the observed effect is a chance effect, since (in theory), all variables that can influence the effect are known to the observer.

In a non-experimental or observational study, a researcher has no control over who receives treatment, but attempts to make a comparison with the observations he has. These studies are also known as quasi-experimental studies or natural experiments. Some more common non-experimental studies are listed in Box 5.

BOX 5: NON-EXPERIMENTAL STUDY DESIGNS

1. Controlled before-and-after study - A follow-up study of participants who have received an intervention and those who have not, measuring the outcome variable both at baseline and after the intervention period, comparing either final values if the groups are comparable at baseline, or change scores. It can also be considered an experimental design if the investigator has control over, or can deliberately manipulate, the introduction of the intervention.

2. Concurrent cohort study - A follow-up study that compares outcomes between participants who have received an intervention and those who have not. Participants are studied during the same (concurrent) period either prospectively or, more commonly, retrospectively.



3. Historical cohort study - A variation on the traditional cohort study where the outcome from a new intervention is established for participants studied in one period and compared with those who did not receive the intervention in a previous period, i.e. participants are not studied concurrently.

4. Case-control study - Participants with and without a given outcome are identified (cases and controls respectively) and exposure to a given intervention(s) between the two groups compared.

5. Before-and-after study- Comparison of outcomes from study participants before and after an intervention is introduced. The before and after measurements may be made in the same participants, or in different samples. It can also be considered an experimental design if the investigator has control over, or can deliberately manipulate, the introduction of the intervention.

6. Cross-sectional study - Examination of the relationship between disease and other variables of interest as they exist in a defined population at one particular time point.

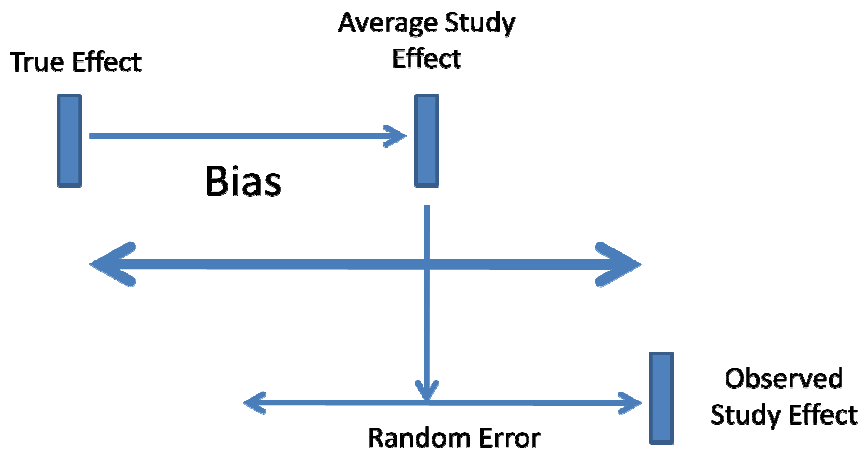
7. Case series - Description of a number of cases of an intervention and outcome (no explicit comparison with a control group).

There are theoretically limitless types of designs of both experimental and non-experimental studies depending on the design features they choose to incorporate. There are even hybrid designs that can incorporate elements of both (229,230). Importantly, what a study calls itself (e.g., RCT or case-control) does not automatically allow us to understand its validity. For example, random error is always present when we measure health outcomes across interventions regardless of trial design. Random error only decreases as we increase the number of participants (observations) in a study. (See Figure 5 below) Other systematic and measurement error that can contribute to biased effect estimates are related to trial design and analysis features.

As one epidemiologist suggested, “RCTs eliminate systematic confounding but at the expense of creating a very artificial situation, with the odd sort of people who would volunteer to have an exposure assigned to them, exposures which may not represent a realistic range of what people actually experience, and forces people to do something they might never have chosen. Figuring out whether the upside or downside matters more requires some scientific common sense” (231) The important question that needs to be addressed is whether the design feature of assignment to treatment in experimental studies can reduce error when compared to similar studies where treatment is not assigned by the investigator.



FIGURE 5 BIAS (SYSTEMATIC AND MEASUREMENT ERROR) AND RANDOM ERROR




(Source: Adapted from Goodman, Presentation to the Institute of Medicine, 2009)

EVIDENCE

The formal use of large non-experimental data to understand relationships between interventions (or exposures) and health outcomes was developed in the 1960s within the realm of the social sciences.(232) It was generally accepted that these non-experimental data sets could be adjusted statistically for selection biases introduced by self-assignment and other unknown factors, and produce results similar to experimental studies.(233) Outcomes research using non-experimental data similarly grew with the advent of information technology in the medical sciences(234), where large experimental studies such as randomized controlled trials had been commonplace. Observational studies as an extension to proof-of-concept clinical trials became commonplace, and similar statistical adjustments were promoted (235). In the 1980s, questions were raised as to whether non-experimental data could truly approximate experimental data in both the social (236) and medical sciences(237).

In an attempt to observe differences between experimental and non-experimental designs, researchers from McGill University conducted an observational study and cluster randomized trial in the same setting to compare the results (237). In an attempt to determine whether in-hospital infant formula supplementation led to changes in breastfeeding rates, the studies revealed different findings. Mothers offered formula in the experimental study had similar rates of breast feeding 9 weeks post partum to similar mothers not given formula. Yet the observational study, which clearly showed mothers who used in-hospital formula breastfed to a much smaller extent than mothers who did




not choose formula (237). The only difference in the two study designs was who controlled the formation of the two comparison groups.

This experiment highlights the differences between non-experimental and experimental research. Firstly, the studies ask two entirely different (but clinically relevant) questions. The experimental study asks how rates of breastfeeding would be affected if *everyone* was given infant-formula. The second asks what happens to *people who choose infant formula*. A number of confounders, or variables that have an effect on the final outcome other than infant formula were identified by the authors. For example, mothers unable to breastfeed from cracked or sore nipples may chosen or been advised to feed using infant formula. The observational study may also not adequately capture cause and effect: does a decision to breast feed lead to reduced rates of infant formula use or is reduced breast feeding a natural result of a decision to use formula?(237)

Since this time, many additional studies have been conducted to explore whether there are predictable differences between observational and experimental research. Initially, there was compelling evidence that poorly conducted randomized controlled trials led to exaggerated estimates of effectiveness compared to their well-conducted counterparts. For example, a bias can be introduced if assignment to treatment is conducted improperly(238), if participants or providers become unmasked, if outcomes cannot be reliably measured, or if important confounding from concomitant therapy or placebo are not appropriately managed(239,240). This evidence of exaggerated effects from poorly conducted randomized controlled trials has been used to suggest randomized controlled trials will always be more reliable than observational studies. The argument, paraphrased, is that if an experimental study with poor assignment to treatment exaggerates clinical effectiveness than surely an observational study, where no assignment occurs, must be more unreliable. This, in turn, has led to evidence hierarchies with clinical decisions needing to consider first and foremost randomized controlled trials (at the top of the hierarchy) and then observational studies.

Despite this, there is a large and compelling body of literature that suggests on average, observational studies do not systematically overestimate results as compared to randomized controlled trials(241–248). Well conducted non-experimental studies produce similar results to randomized controlled trials when asking similar questions (249). There has been some limited evidence that measurements which are more objective (e.g., blood pressure, myocardial infarction) may also be more reliable (250) .

Observational research has been more recently acknowledged as central to comparative effectiveness research and its role in evidence-based decision making. (251) Factors such as patient compliance and adherence may have a



dramatic effect on the real-world effectiveness of a drug but rates of compliance are most often unusually high in the context of a randomized control trial. For example For example, despite the weak effect of oral asthma medications relative to inhaled medications demonstrated in RCTs, real-world observational data provided compelling evidence that these agents have a similar effect in children, largely due to improved adherence rates.(252,253) Currently, observational research is already used in advertising to identify harms from drug therapy, as the much larger numbers of patient observations more easily allow producers and regulators to identify rare but serious effects.

This has led to more reflection on the role of observational research in making clinical decisions (254,255). It is now largely acknowledged that evidence hierarchies for making clinical decisions are inconsistent with current knowledge about the reliability of observational research (245,256–258). It has been suggested, even by the most ardent supporters of randomized controlled trials, that observational research can be used more decision making and is more reliable when the observed effect is large, when there is a dose–response gradient, and when all plausible confounders or other biases increase our confidence in the estimated effect (259). It has also been suggested clinicians must be careful not to interpret *P* values from observational studies in a similar fashion to experimental studies (260). This means observational research can reliably answer clinically relevant questions in a complimentary way (261–265) and it can sometimes provide important information for clinical decisions in an easier fashion than experimental research (266,267). For example, Hayward and colleagues effectively demonstrated that evidence of the effectiveness from treating to cholesterol targets cannot logically be answered using a randomized trial design; they proposed an observational approach to deal with the lack of evidence for this question.(268)


Ultimately, claims of effectiveness from observational studies still rest on more elaborate techniques for measuring and analyzing data. There is no real consensus on what statistical adjustment techniques are most reliable and the empirical evaluation of these methods against each other and against randomized data on a large and systematic scale remains a vital area for future research. Because the findings from observational research may be largely influenced by the scientific judgments of the researchers analyzing the data, one excellent and large RCT may provide adequate evidence, whereas the reliability of a single observational study is much more suspect. Causal claims from observational data will always require closer scrutiny, as well as evidence that is reliable – i.e., consistent with different studies from different researchers using different methods of observation and adjustment.



SUMMARY OF EVIDENCE


The key findings from the previous section can be summarized as follows:

- Systematic reviews have become best practice for assessing clinical effectiveness and are less prone to bias than the arbitrary selection of single studies.
- Meta-analysis as a means to providing an estimate of effectiveness is becoming less well-accepted as it has known pitfalls.
- Well-conducted meta-analysis of individual patient-level data based on a systematic review of relevant studies provides the best opportunity for reliably estimating outcomes, but can still be misleading as the results can be influenced by assumptions not easy to detect.
- Meta-analysis is more appropriate for exploring between-study differences.
- The identification and use of unpublished research has become a best practice.
- There is no compelling evidence that peer review improves the validity of research findings - since negative or equivocal studies are less likely to be published, the use of published-only information has the potential to mislead clinical decision makers into believing therapies are more beneficial than they actually are.
- There is considerable evidence to suggest subgroup analyses are frequently improperly conducted, reported and interpreted and the extent of the problem is likely underestimated. Industry-funded trials may be more prone to bias than their publicly-funded counterparts.
- Subgroup analyses are more plausible if the direction and measurement are specified a priori, there are fewer outcomes tested, the effect is large and consistent with other subgroups in the trial
- Issues for secondary outcomes are similar to those for subgroup analyses, although subgroups uniquely suffer from potential bias from defeating randomization and secondary outcomes uniquely suffer from problems with measurement and extrapolation.

- 
- Post hoc analysis should be considered unreliable unless they have been replicated
 - There is no compelling evidence to suggest observational studies are more or less reliable than randomized controlled trials for making therapeutic claims of effectiveness although there are additional factors (e.g., selection bias, measurement error) that must be considered to assess their validity.
 - Some claims required for clinical decision making (not necessarily effectiveness claims) may be better substantiated by observational evidence.
 - Clinicians must be careful not to interpret *P* values (or confidence intervals) from observational studies in a similar fashion to experimental studies

OPTIONS

1. The use of up-to-date systematic reviews should be encouraged as a means to provide decision makers with an understanding of what evidence is currently available.
2. The use of meta-analysis should be discouraged
 - [Option 1] – PAAB can insist that these clinical effectiveness claims only be based on an individual patient data meta-analysis based on an updated systematic review of available data in a particular patient population. This is problematic since there may be a lack of an updated systematic review and the applicant will be forced to conduct one. On the other hand, during a marketing period, the advertiser should have complete knowledge of all trials that have been conducted with the drug.
 - [Option 2] – PAAB can insist that clinical effectiveness claims can be made from individual trials but that all other similar trials conducted in the same population with the same drug (and dosage), population and effect sizes must be reported to allow the clinician to understand the plausible range of effect sizes measured from study data.

- 
- [Option 3] - – PAAB can insist that clinical effectiveness claims can be made from individual trials but that all other similar trials conducted in the same population with the same drug (and dosage), population and effect sizes must be depicted in a forest plot (based on a systematic review). In this way, a clinician can readily see if the trial being promoted is an outlier compared to other available information. The forest plot can easily depict what trial is part of the advertisement.

3. The use of unpublished research findings should be encouraged


- [Option 1] – PAAB should remove its restriction on the use of unpublished data; bearing in mind that the proper interpretation of statistical null hypothesis testing requires trial protocol and outcomes information accessible to clinicians.

4. Subgroup analysis should be carefully managed

- [Option 1] – PAAB should not allow the use of subgroup analysis
- [Option 2] – PAAB can allow the use of subgroup analysis, but only if a forest plot or equivalent relevant information that states the number of groups tested, identifying characteristics of the groups, the direction and magnitude of the effect from each group are reported and trial protocol information is available
- [Option 3] - PAAB can allow the use of subgroup analysis, but only if it meets some pre-specified conditions for validity, AND if a forest plot or equivalent relevant information that states the number of groups tested, identifying characteristics of the groups, the direction and magnitude of the effect from each group are reported and trial protocol information is available

5. Secondary outcome analysis should be managed carefully

- [Option 1] – PAAB should not allow the use of secondary outcome analysis
- [Option 2] – PAAB can allow the use of secondary outcome analysis, but only if a forest plot or equivalent relevant information that states the number of outcomes tested, the identifying characteristics of each outcome, the direction and magnitude of the effect from each outcome are reported and trial protocol information is available
- [Option 3] - PAAB can allow the use of secondary outcome analysis, but only if it meets some pre-specified conditions for validity AND a forest plot or equivalent relevant information that states the number of outcomes tested, the identifying characteristics of each outcome, the



direction and magnitude of the effect from each outcome are reported and trial protocol information is available

6. Post hoc analysis should continue to be discouraged
7. Observational study claims should be encouraged
 - [Option 1] – PAAB should allow claims based on the use of observational studies if sufficient information is made available to PAAB to assess the validity of the claim

RECOMMENDATIONS

1. If claims from individual studies are used, information regarding the total number of similar studies conducted (in terms of patients, interventions, design) from a *systematic* review of available evidence should be made available to reduce selection bias or claims based on exaggerated study findings.

Rationale: A consumer provided with a single estimate of effectiveness will be left with the impression (i.e., misled to believe) that this estimate represents an achievable level of effectiveness in her patient. Yet, even large randomized trials may report findings that are not entirely explained by the drug. Although, we cannot always know what factors may have contributed to an estimate of effectiveness that is exaggerated or inconsistent with other studies, it is important to be aware that a range of effects have been observed. Only a systematic of available evidence can reliably and consistently demonstrate what similar studies exist. This should be very feasible for the advertiser - in the marketing phase of a drug, and if all trials have been registered and identified to drug regulators, the process of identifying all available similar studies should be quite rapid and not place a large burden on manufacturers.

2. The use of meta-analysis for making claims of effectiveness should be discouraged.

Rationale: The results of meta-analysis may lead to an estimate of effectiveness which is misleading for reasons which are not obvious to a consumer, even if provided with the full details of analysis, or even obvious to the analyst who conducted the meta-analysis, specifically when studies are combined that have very different findings. Although meta-analysis is helpful for understanding why studies may differ, it is not a currently accepted best practice for guiding clinical decision making.



3. The use of unpublished research findings should not be discouraged.


Rationale: The identification and use of unpublished research has become a best practice due to substantive and compelling evidence that ignoring unpublished research can lead to biased estimates of effectiveness. There is also growing evidence that suggests peer review does not improve the validity of findings. Although a single unpublished study may not provide the best information for creating a claim of clinical effectiveness, unpublished research findings should be identified when the results of systematic reviews are presented. This will allow the consumer to understand whether published results are providing more optimistic estimates compared to their unpublished counterparts.

4. PAAB can allow the use of subgroup analysis, but with specific conditions.

Rationale: Subgroups may be important for clinical decision making but have the potential to mislead if they are improperly conducted. PAAB should consider allowing the use of subgroup analysis, but only if it meets some pre-specified conditions for validity; at minimum these should include a consideration of the criteria in Box 4, such as whether the subgroup is biologically plausible, identified *a priori*, is large and consistent etc. Similar to presenting the results from a single trial, those wishing to show results from a subgroup analysis should be willing to explicitly report the primary outcome findings of the trial and show other subgroups using a forest plot or equivalent relevant information that states the number of groups tested, identifying characteristics of the groups, and the direction and magnitude of the effect from each group. Because subgroup analyses are susceptible to *post hoc* biases, they should only be allowed if trial protocol information is available

5. PAAB can allow the use of claims from secondary outcomes, but with specific conditions.

Rationale: Claims based on secondary outcomes may be important for clinical decision making but have the potential to mislead if they are improperly conducted. PAAB should consider allowing the use of secondary outcomes if they meet certain pre-specified conditions for validity; at minimum the outcome should have been 1) replicated in more than one independent study and should be the strongest claim available for that outcome. For example, a claim of mortality benefit from a secondary outcome analysis in a trial should not be made if another similar trial (similar patients, interventions and design) showed no estimable difference for that same outcome in a different trial. The secondary outcome must be plausible – for example, a claim of reduction in mortality in a healthier population does not make sense if a sicker population expected to further benefit does not demonstrate the same (or better) effect. The secondary



outcome must also be identified *a priori* in a trial statistical analysis plan. There should also be sufficient evidence that the outcome itself is valid (i.e., for surrogate outcomes) and has adequate properties for reliable measurement (i.e., for patient-reported outcomes). Similar to reporting results from individual trials and individual subgroups, secondary outcomes should be reported in a way that allows the consumer to see findings from all other subgroups tested (e.g., using a forest plot) or equivalent relevant information that states the number of groups tested, identifying characteristics of the groups, and the direction and magnitude of the effect from each group. Because secondary outcome analyses are susceptible to *post hoc* biases, they should only be allowed if trial protocol information is available.

6. Post hoc analysis should continue to be discouraged

Rationale: Post hoc analyses continue to be unaccepted practice for establishing claims of effectiveness. Unless the claim of effectiveness is strikingly obvious, they will require validation through subsequent *a priori* testing. However, if the effect is *strikingly* obvious, advertising as a means to influence consumer behaviour should be unwarranted.

7. PAAB can allow the use of claims from observational studies, but with specific conditions.

Rationale: Claims of effectiveness based on observational studies may be more important for clinical decision making than randomized trials but have the potential to mislead if improperly conducted. Unlike a large and well-conducted randomized trial, claims of effectiveness are also theoretically more susceptible to bias from judgments used by investigators in conducting analyses and unidentified effect modifiers. PAAB should consider allowing the use of claims based on observational studies if they meet certain pre-specified conditions for validity and , and have been replicated by independent research groups using different patients and approaches and the findings are consistent with other evidence or have a strong underlying theoretical basis. Specific claims related to effectiveness (but not claims of the causal impact of a drug on a clinical outcome) may be particularly warranted. For example, claims of patient adherence may be poorly based on randomized trial evidence (although very pragmatic designs, unlike current phase II designs, or hybrid randomized designs may be sufficient to demonstrate this). Moreover, some claims, like the benefits of reaching a laboratory marker target can never be made based on randomized trial evidence and should require observational data.

ILLUSTRATIVE EXAMPLE

Advertisement appearing October 4, 2011 issue of Canadian Medical Association Journal

LIPITOR
Our indications are supported by our evidence

STROKE
relative risk reduction shown*
95% CI (-69% to -11%)
ARR[†] = 1.3% = stroke
p = 0.016

MYOCARDIAL INFARCTION
relative risk reduction shown*
95% CI (-50% to -17%)
ARR[†] = 1.1% in non-fatal MI and fatal CHD^{††}
p = 0.0005

HYPERCHOLESTEROLEMIA
LDL-C reductions shown^{**} across the dose range

ORIGINAL LIPITOR STILL REIMBURSED BY RAMQ AND MANY PRIVATE INSURANCE PLANS

LIPITOR is indicated to reduce the risk of MI and stroke in adult patients with type 2 diabetes mellitus and hypertension without clinically evident CHD but with other risk factors.*

LIPITOR is indicated to reduce the risk of MI in adult hypertensive patients without clinically evident CHD but with at least three additional risk factors for CHD.[‡]

LIPITOR In patients with hypercholesterolemia

Up to **60%**
(**39% - 60%**)

48%

36%

* p < 0.05 compared to placebo
† See text for details of events, results for women were non-significant



Changes from adopting recommendations 1-3 (the statistics are not factually correct and are for illustrative purposes)

LIPITOR®

Our indications are supported by our evidence

LIPITOR
is indicated to reduce the risk of MI and stroke in adult patients with type 2 diabetes mellitus and hypertension without clinically evident CHD but with other risk factors.*

STROKE *

48%

relative risk reduction shown
95% CI (-69% to -11%)
ARR[†] = 1.3% in stroke
p=0.016

**ORIGINAL LIPITOR
STILL REIMBURSED BY RAMQ AND
MANY PRIVATE INSURANCE PLANS**

LIPITOR
is indicated to reduce the risk of MI in adult hypertensive patients without clinically evident CHD but with at least three additional risk factors for CHD.³

MYOCARDIAL INFARCTION

36%

relative risk reduction shown[†]
95% CI (-50% to -17%)
ARR[†] = 1.1% in non-fatal MI and fatal CHD^{†††}
p=0.0005

†† LIPITOR is not indicated to prevent fatal CHD.
††† Due to a small number of events, results for women were inconclusive.

LIPITOR
In patients with hypercholesterolemia

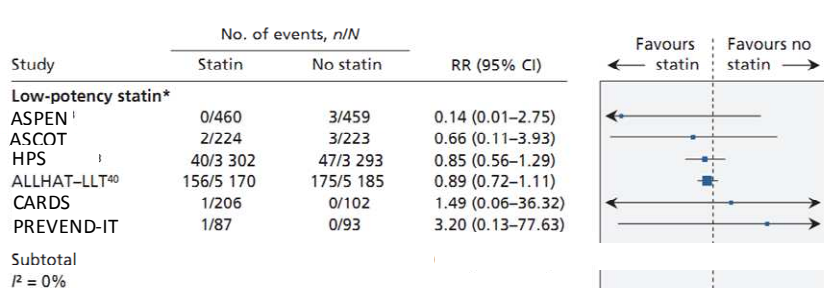
HYPERCHOLESTEROLEMIA

Up to 60%

(39% - 60%)

LDL-C reductions shown^{††} across the dose range

*This outcome has been studied in six trials enrolling over 10,000 patients. The direction and size of the effect is seen in the diagram below





FINDINGS FOR SECTIONS 5.7 - 5.12

Section 5.7 of the PAAB Code states “Comparative claims of efficacy and safety require support of evidence from head-to-head well-designed, adequately controlled, blinded, randomized clinical studies. Open-label studies are not considered to be a high level of evidence and are not acceptable if subjective end-points are included in the study. Comparative claims should be relevant to current medical opinion and practice.”

SHOULD MATHEMATICAL MODELING BE USED TO SUPPORT COMPARATIVE CLAIMS OF EFFECTIVENESS?

INTRODUCTION

A model can be defined as a “simplified representation of reality that captures some of that reality’s essential properties and relationships (e.g. logical, quantitative, cause/effect).”(269) In a simulation model, an “actual or proposed system is replaced by a functioning or interactive representation of the system” (269) Today’s simulation models generally use mathematics and require computation that is facilitated by the use of computers. Models of clinical effectiveness often attempt to predict health outcomes experienced by people with specific characteristics and then compare outcomes experienced to those exposed to a treatment. These types of models have also been labeled disease outcome models. These models provide a framework for assembling knowledge and values from different sources, allowing a “formal quantified comparison of health technologies” (270) and can aid clinical decision making.

Mathematical modeling of effectiveness often includes the use of statistical (i.e., extrapolation based on current data) models. Meta-analysis and network meta-analysis are types of statistical models and will be dealt with in the next section. This section will focus strictly on mathematical models.


Although models can be used to predict outcomes, such as those associated with comparative claims of effectiveness, their additional value is in allowing a decision maker to understand the relationship between a model input (e.g., a patient characteristic, such as blood pressure) and output (e.g., risk of heart attack with therapy). This led to a now-famous quote by a pioneer in the field of mathematical modeling research, George Box, to declare “Essentially, all models are wrong, but some are useful” (271). This is not to say models are inherently misleading. As Sculpher and colleagues responded, “... some are useful, and it is the usefulness of models that is the appropriate test of validity not the accuracy of predictions. Indeed, this case has been made about scientific activities in general.”(272). The question is whether better decisions are made with the results of a modeling study versus without it.



EVIDENCE

The role of modeling to inform policy decision making in regards to the impact of new technology has a long history with applications across population growth, agricultural, environmental, energy, and fiscal policy sectors (273). Disease modeling in health care has seen more widespread uptake in recent years, particularly with the growth and acceptance of cost-effectiveness analysis for reimbursement decisions in health care (274,275). In the UK, for example, the National Institute for Health and Clinical Effectiveness program commissions systematic literature reviews and meta-analysis and uses these as inputs to disease models for the purpose of providing guidance on reimbursement to the UK National Health Service (276). Disease outcome models to inform clinical guidance are also emerging. Recently a leader in the development of clinical practice guidelines, the Agency for Healthcare Research and Quality-sponsored US Preventive Services Task Force changed the name of their technology assessments from systematic evidence reviews to *evidence syntheses*; in part to reflect the use of statistical and mathematical models to estimate the effectiveness of preventive interventions (277).

Despite a lack of direct evidence, the systematic and structured organization and analysis of data, in theory, is an improvement over the qualitative use and subjective (or informal) combination of data. Credible, high quality disease outcome models use statistical models (such as meta-analysis or indirect comparisons, below) to provide inputs for the short term effectiveness of treatment and use mathematical models to extrapolate these estimates over years, based on local population data. For example, a recent analysis undertaken by a UK government-commissioned research group compared clopidogrel and modified-release dipyridamole for the prevention of occlusive vascular events (278). The analysts used information from a systematic review that identified a total of four trials (279–282) to describe how likely patients on either treatment would experience fatal and non-fatal cardiovascular events. A modeling study conducted by the manufacturer of clopidogrel did not use data from all available trials which led to more optimistic projections about the number of quality-adjusted life-years gained with clopidogrel by patients with multi-vascular disease (an additional 0.6 versus an additional 0.154). The model was primarily based on results from a single trial (279). Many other differences in approach explain the differences in results, including the manufacturers model making more optimistic assumptions about what happens to patients in the unobserved years beyond the clinical trial time period, and the government-funded analysts using more conservative statistical assumptions about how patients disease progress (based on more sophisticated statistical analysis). Some models with



especially detailed statistical assumptions have demonstrated an ability to accurately predict health outcomes in different populations (283).

There is little direct evidence to suggest mathematical modeling leads to better decisions about health. A systematic review of the value of computer simulation modeling in population and health care interventions found very little in terms of evidence, and concluded “Despite the increasing numbers of quality papers published in medical or health services research journals we were unable to reach any conclusions on the value of modelling in health care because the evidence of implementation was so scant.”(284) Another study intending to examine clinical judgment from expert panels in cardiovascular disease with decision-analytic modeling found disease outcome models and expert panels performed similarly when faced with the same decision problems(285). The authors concluded “neither approach is superior to the other” (285).

There is a large body of literature that provides compelling indirect evidence that predictions based on modelling are at least as (if not more) accurate than informal methods of synthesis. In a systematic review of 136 studies comparing human (clinical) with actuarial (statistical or mathematical) predictions of diagnoses or human behaviour, 64 studies demonstrated the superiority of actuarial methods, 64 showed similar accuracy and only 8 favored clinical judgments (286). The authors highlight the fact that these findings were entirely consistent with a previous review conducted 50 years earlier.

The real advantage of mathematical models, and beyond the scope of their use in advertising, is their ability to make decisions more transparent, and provide an opportunity for legitimacy and accountability among for decision makers (287,288). Methodological and reporting guidance has been published to facilitate the transparency and usefulness of decision models (272,289). Mathematical models allow decisions to be revisited in light of new information or poor outcomes and provide a coherent framework across multiple decisions. While mathematical modeling may or may not more accurately predict future health outcomes, it does provide a broader and more legitimate approach to creating public and clinical policy decisions (290,291) This means for models to be meaningful and lead to better clinical decision making, they should not be simply a static analysis. Translating their real advantage to decision making means allowing decision makers to interact with the models. This means making the models available to decision makers, rather than simply conveying the results under a certain set of conditions that may not apply to a specific situation.

SHOULD INDIRECT COMPARISONS BE USED TO SUPPORT COMPARATIVE CLAIMS OF EFFECTIVENESS?

INTRODUCTION

The question of clinical effectiveness or whether a drug might do something versus doing nothing has become much less relevant given the number of new therapies emerging yearly. More relevant for the consumer is judging whether the new drug therapy works better than an existing one. If there is a lack of direct evidence of relative effectiveness from studies comparing two separate treatments, the consumer is forced to make an indirect comparison. They could, for example, compare reductions in myocardial infarction from two drugs by examining estimates of effectiveness from randomized trials (preferably based on systematic reviews). However, this method of comparing introduces the opportunity for error and a biased estimate of relative effectiveness – the reason for improved reductions in myocardial infarction could be due to the fact that one trial enrolled patients better able to benefit or a myriad of other factors (See Box 6 below)(292,293). This method of indirect comparison has been labeled unadjusted or naïve indirect comparison (294).

Box 6: Factors that might introduce differences in relative treatment effects from indirect comparisons (adapted from (292))

A. Different quality or methods of randomized trials

- i. Adequate concealment of randomization
- ii. Blinding
- iii. Duration of follow-up
- iv. Loss to follow-up
- v. Cross-over

B. Confounding factors in relation to participant populations

- i. Age
- ii. Sex
- iii. Genetic variation
- iv. Diagnostic workup
- v. Intensity of surveillance
- vi. Severity of pathology
- vii. Physiological reserve
- viii. Stage or duration of disease
- ix. Prior therapy
- x. Co-existing disease
- xi. Background therapy of concomitant treatments/advances in standard of care

C. Confounding factors in relation to circumstances

- i. Health systems
- ii. Geography
- iii. Setting in hospital or ambulatory care
- iv. Date of trials

D. Different treatment (common reference and interventions)

- i. Dose
- ii. Duration
- iii. Timing


E. Different outcome measures and methods of statistical analysis

- i. Definition
- ii. Rating instrument
- iii. Frequency of measurement
- iv. Start point of measurement against duration or progression of disease or treatment, especially in time-to-event analyses

A more reliable approach is to create comparisons adjusting for these factors. Statistical approaches have been developed that allow comparisons across results from two or more randomized controlled trials (295–300). These approaches are an extension of the same concepts underlying meta-analysis, and rely on the analysis of a network of randomized trials, connected by common underlying properties. Meta-analytic approaches to indirectly compare two or more studies with at least one treatment in common (i.e., in a network) have been called adjusted indirect treatment comparison, anchored indirect treatment comparison, cross-study comparison, connected comparative experiment, network meta-analysis, mixed comparison, and virtual comparison (293). The blanket term “network meta-analysis” can be used to describe all of these techniques.(301)

Similar to meta-analysis, these methods are compromised when they are based on corrupted or missing information. Hence, the starting point for valid statistical methods of indirect comparison is a systematic review of the available evidence and careful selection of comparable studies for combination (302). Like meta-analyses, they can provide a single estimate of effectiveness but even more importantly allow the analyst to examine the consistency (or coherency) of the direct and indirect evidence available (112). They allow exploration of between-trial differences and extreme results (294). These comparisons can also be conducted using evidence from both direct and indirect sources. There are concerns that techniques for creating indirect comparisons, like techniques used to adjust for bias in non-experimental studies, may be inadequate and lead to potential bias themselves. One working group has suggested the term “Common reference-based indirect comparison” instead of “Adjusted indirect comparison”

Page | 68



as to not lure readers into a false sense of complacency that the estimates have been entirely adjusted (292).⁷ On the other hand, in the absence of formal comparisons or data from head-to-head trials, consumers may be tempted to create their own unadjusted comparisons, even unconsciously. Additionally, head-to-head trials themselves may have significant shortcomings that may mislead consumers (303). The question of whether network meta-analytic methods for creating indirect comparison are sufficiently reliable to not mislead consumers will be addressed in the next section.

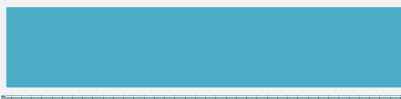
EVIDENCE

Indirect comparison approaches using network meta-analysis are receiving significant attention. As one recent article suggests, given “...their enormous value for health intervention decision-making, clinicians, drug manufacturers, regulatory agencies and the public are now requiring meta-analysis to identify the most effective intervention among a range of alternatives”. The Canadian Institutes of Health Research and the Agency for Healthcare Research and Quality have funded research (304) exploring optimal methods for indirect comparison. Guidance for appraisal and reimbursement in the UK has changed to reflect these methods (276) while guidance in Australia is currently undergoing consultation (292). Both the Cochrane Collaboration and the International Society for Pharmacoeconomics and Outcomes Research have established special working groups to identify best practices in these areas (305,306)

The validity of approaches to network meta-analysis has been explored in several studies (294,307–309). Song and colleagues identified 44 meta-analyses where competing interventions could be compared both directly and indirectly. They found that “adjusted indirect comparisons usually but not always agree with the results of head-to-head randomized trials” (307). They suggested that the validity of the trials being analyzed must be examined as it can compromise the overall results. To illustrate they provide an example of a meta-analysis of acetaminophen plus codeine versus acetaminophen alone for postsurgical pain.

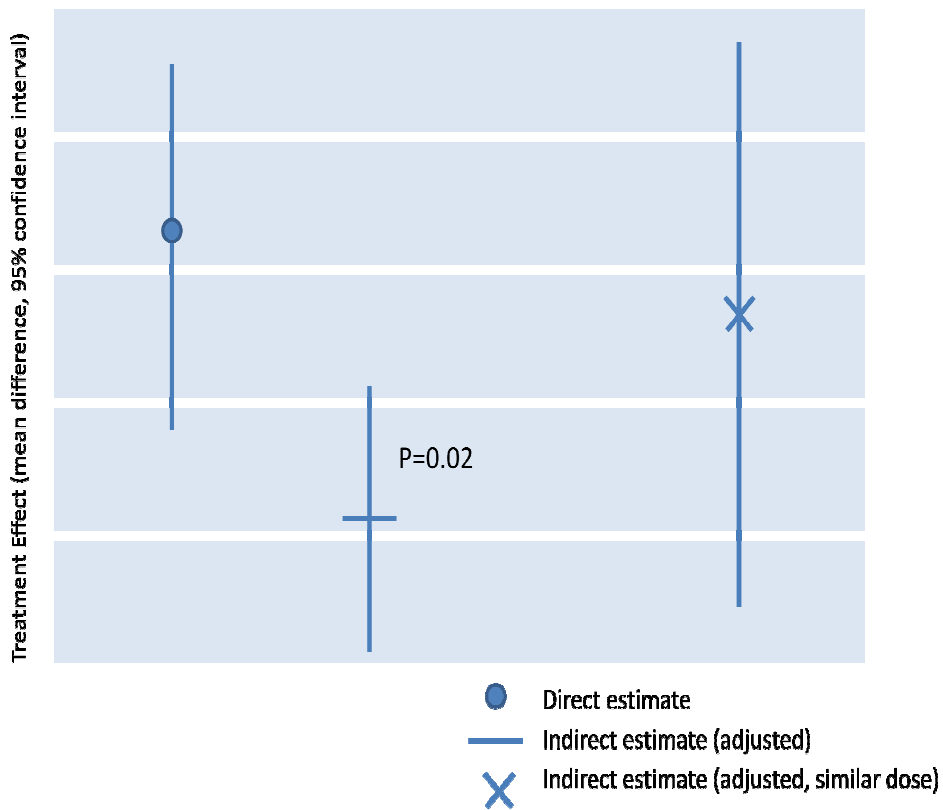
“Based on the results of 13 trials, the direct estimate indicated a significant difference in treatment effect (mean difference 6.97, 95% confidence interval 3.56 to 10.37). The adjusted indirect comparison that used a total of 43 placebo controlled trials suggested there was no difference between the interventions (– 1.16, – 6.95 to 4.64). The discrepancy between the direct and the adjusted indirect estimate was significant (P=0.02). (See Figure 6) However, most of the

⁷ The term “anchored indirect treatment comparison” has also been used
Page | 69




trials (n=10) in the direct comparison used 600-650 mg [acetaminophen] and 60 mg codeine daily, while many placebo controlled trials (n=29) used 300 mg [acetaminophen] and 30 mg codeine daily. When the analysis included only trials that used 600-650 mg [acetaminophen] and 60 mg codeine, the adjusted indirect estimate (5.72, - 5.37 to 16.81) was no longer significantly different from the direct estimate (7.28, 3.69 to 10.87).

FIGURE 6: IMPORTANCE OF SIMILARITY BETWEEN TRIALS IN ADJUSTED COMPARISON



Song and colleagues also suggested adjusted indirect comparisons may be less biased than direct comparisons (310) based on limited empirical evidence and simulation. However their simulation study which demonstrated placebo-controlled studies of new drugs have less potential for introducing a bias than that of old drugs did not specifically examine differences between direct and indirect estimate. In an update and further exploration on this topic, Song and colleagues were able to demonstrate empirically that significant inconsistency



existed when direct and indirect comparisons were compared. (309) However, in a response to the article, one group of investigators noted that these findings may be more reflective of the weaknesses of the effect of underlying data on meta-analysis in general, rather than the methods used.

Wells and colleagues also reviewed various methods for adjusted indirect comparison. They highlighted that not all methods are suitable under all circumstance. When comparing their performance through simulation, they concluded that there is a potential for bias when event rates are small, but suggested conclusions about the degree of bias associated with direct versus indirect effect estimates are not consistent with results of Song's study. However, it is entirely possible that the approach to examining small event rates (and zero values), a difficult issue, may have been more responsible for this finding. The study provides some evidence that adjusted indirect comparisons using a method by Bucher (300) are a robust approach.

Like meta-analysis, the outcomes of this type of analysis can be accurate if care is taken in its conduct. This is reflected in recommendations made in a recent report, "Validity of the adjusted indirect comparison methods depends on the consistency of treatment effects across studies, and the appropriateness of an indirect comparison needs to be assessed on a case-by-case basis." (304) However, changes in underlying assumptions not readily obvious to a consumer can change the outcomes of these analyses.

SHOULD NON-INFERIORITY STUDIES BE USED TO SUPPORT COMPARATIVE CLAIMS OF EFFECTIVENESS?

INTRODUCTION

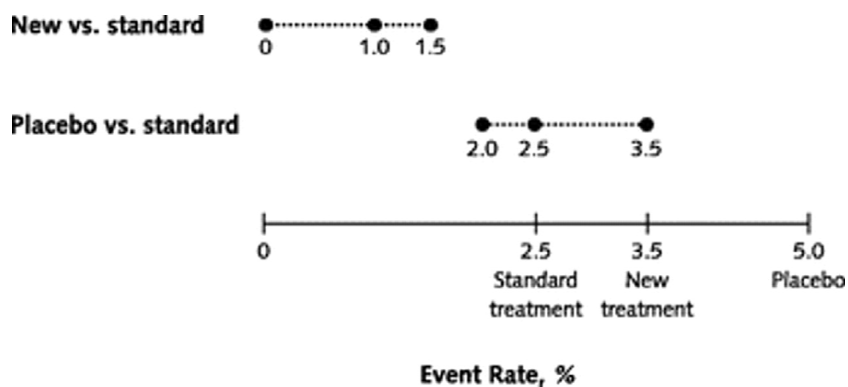
The *superiority* of being exposed to a treatment versus no treatment has traditionally been investigated through placebo-controlled experiments, which in turn have been used to substantiate claims for regulatory purposes. In some cases, a placebo-controlled trial may be unethical when multiple therapeutic options are available; additionally, a new therapy may or may not necessarily offer a therapeutic advantage but rather anticipate a market share by being perceived as having a better safety profile, being more convenient or reducing healthcare costs. Under these circumstances an experiment using existing therapy as a comparator is required. If the therapy is not anticipated to be superior in terms of health gains, the trial can be designed to detect whether the therapy produces similar health benefits (i.e., produces an effect that is similar to the active comparator within a margin of error) or that it is non-inferior (i.e., produces an effect that is not worse than the comparator within a margin of error). Trial designs that attempt to answer questions about whether the new therapy is not worse than existing therapy are called *non-inferiority* trials.



An example of a non-inferiority trial is provided by Kaul (311). In this example two therapies are to be assessed for a disease known to be associated with serious irreversible clinical outcomes and death. We can imagine that an existing standard therapy lowers the mortality rate from 5% (based on placebo response) to 2.5%. In this case the absolute benefit is 1 additional death avoided for every 40 treated, or a reduction in death of 2.5%.

If a new drug with a similar mechanism of action and predicted benefit is available, we will want to test it against the existing therapy. We design a trial to compare the standard treatment with the new treatment, as a placebo arm is judged to be inferior and unethical. What criteria for success of the new treatment should be used to inform clinical decision making?

Basic concepts of noninferiority assessment. dotted line.




Kaul S, Diamond G A Ann Intern Med 2006;145:82-89

©2006 by American College of Physicians

Annals of Internal Medicine

“To be fairly certain that the new treatment is better than placebo, we want to be convinced that its mortality rate is not more than 2.5 percentage points greater than the standard. One such criterion could be that the 95% CI around the trial estimate should not include a mortality rate increase of 2.5 percentage points. This would be fulfilled by observing a mortality rate increase of 1 percentage point (CI, 0 to 2.0 percentage points).

The upper limit statistically excludes the mortality rate difference of 2.5 percentage points in persons taking placebo, but there are 2 problems with such a criterion as the basis for a conclusion of non-inferiority. First, although the absolute benefit of the new treatment might show that it is likely to be better



than placebo, it might be lower than the standard (for example, 2%) and thus still not preferable. Second, in this example, we assume that we know the standard treatment benefit with certainty. In fact, the treatment benefit of 2.5 percentage points would always have a range of uncertainty, for example, a CI of 2 to 3.5 percentage points. For a new treatment to be better than placebo, it would have to be shown that its mortality rate was less than 2.0% (the smallest expected standard treatment effect), not 2.5% (the point estimate of standard treatment effect) and thus no more than 2.0 percentage points (or less) higher than the standard. In summary, the degree of tolerable inferiority, that is, the non-inferiority margin [i.e., irrelevance margin], must take into account the uncertainty in the estimated difference over placebo, and it must be outweighed by the superiority of the new treatment in other respects.”(311)


Non-inferiority trials create challenges for those who conduct trials and those who need to interpret their findings to make decisions. First, they require assumptions about comparator performance which rely on methods of quantitative synthesis, like meta-analysis. They must also rely on assumptions about what it is to be worse, and what margin of error may be achieved. They also use different approaches to statistical testing and interpretation. For example, if the results of a well-conducted non-inferiority trial lead to a positive (statistically significant) result, the consumer may be misled into believing the new therapy is similar or equivalent to an existing therapy. It may be especially difficult to reconcile the meaning of a “negative” trial that shows a new therapy is not non-inferior than established therapy. And unlike superiority trials, an underpowered non-inferiority trial may be more likely to produce an untrue positive result. Concerns and issues highlighted in international guidance documents for trialists (312–314) have been synthesized by Wangge(315) (Figure 7).



FIGURE 7: SYNTHESIS OF CONCERNS REGARDING NON-INFERIORITY TRIAL DESIGN FROM FDA (DRAFT) AND ICH GUIDANCE DOCUMENTS

Issues in NI trials	Requirements in the guidelines
Blinding method	<ul style="list-style-type: none"> - Blinding is necessary to minimize bias (<i>ICH E9 and E10</i>)It is critical to provide reassurance and procedures that ensure maintenance of blinding (<i>draft FDA guideline on NI trial 2010</i>)
NI margin	<ul style="list-style-type: none"> - An acceptable non-inferiority margin should be defined (<i>ICH E10, CPMP/EMEA 2000</i>) - Should be pre-specified, and can be no larger than the presumed entire effect of the active control in the NI trial (<i>draft FDA guideline on NI trial 2010</i>) - Should be specified in publication (<i>CONSORT statement extension, 2006</i>)
Method to determine NI margin	<ul style="list-style-type: none"> - The determination of the margin in a non-inferiority trial is based on both statistical reasoning and clinical judgment (<i>ICH E10</i>) - Margin is chosen by defining the largest difference that is clinically acceptable, so that a difference bigger than this would matter in practice (<i>CPMP/EMEA 2000</i>) - The NI margin should be generally identified based on previous experience in placebo-controlled trials of adequate design under conditions similar to those planned for the current trial, but could also be supported by dose response or active control superiority studies.(<i>ICH E10, CHMP/EMEA 2005</i>) - Fixed margin method (two CIs method) is recommended. It is referred to as fixed because the past studies comparing the drug with placebo are used to derive a single fixed value for statistical margin, even though this value is based on results of placebo-controlled trials (one or multiple trials versus placebo) that have a point estimate and confidence interval for the comparison with placebo. This approach is relatively conservative, as it keeps separate the variability of estimates of the treatment effect in the historical studies and the variability observed in the NI trial, and uses a fixed value for the estimate of the control effect based on historical data (the 90% or 95% CI lower bound), a relatively conservative estimate of the control drug effect. (<i>draft FDA guideline on NI trial 2010</i>) - should be specified in publication (<i>CONSORT statement extension, 2006</i>)
Type of statistical analysis	<ul style="list-style-type: none"> - Use of the full analysis set is generally not conservative and its role should be considered very careful (<i>ICH E9</i>) - Both ITT and PP have equal importance (<i>CPMP/EMEA 2000</i>) - Important to conduct both ITT and as-treated analyses. Differences in results using the two analyses will need close examination. (<i>draft FDA guideline on NI trial 2010</i>)
Assay sensitivity	<ul style="list-style-type: none"> - A trial should have the ability to distinguish an effective from an ineffective drug (<i>ICH 10, draft FDA guideline on NI trial 2010</i>) - A three-armed trial with test, reference and placebo allows some within-trial validation of the choice of non-inferiority margin and should be used wherever possible.(<i>CHMP/EMEA 2005, draft FDA guideline on NI trial 2010</i>)
Constancy assumption	<ul style="list-style-type: none"> - The similarity of the new trial to the historical trial should be sufficient (<i>CHMP/EMEA 2005, draft FDA guideline on NI trial 2010</i>)
Similarity with trial of reference treatment	<ul style="list-style-type: none"> - The report should contain whether the eligibility criteria, interventions and outcomes are identical (or very similar) to that of any trial that established efficacy of the reference treatment (<i>CONSORT statement extension, 2006</i>)

Note: The draft FDA guideline is not in effect yet and still open for changes (as per 18th March 2010).
doi:10.1371/journal.pone.0013550.t001



Despite differences in the design, approach and interpretation of non-inferiority trials, what is relevant is whether these can be readily reported and interpreted in a way that is meaningful to consumers, without having a significant potential to mislead. As one observer commented, “drug and device manufacturers may not be willing to state in an advertisement that ‘our product was not inferior to the standard product with regard to our predefined margin of the smallest clinically meaningful difference.’”(316)

EVIDENCE

Several comprehensive reviews have been conducted on the conduct and reporting of non-inferiority trials (317–321). In a systematic review of the use and choice of equivalence and non-inferiority (i.e., irrelevance) margins, Lange and Freitag identified 332 publications of 327 unique trials (317). Although they did not report results for non-inferiority and equivalence trials separately, they discovered a rationale for the irrelevance margin was given in approximately half of trials but substantiated in only 30% (86/314) of trials. They also observed that in nearly half the trials, the margins appear to be too large, corresponding to odds ratios of at least 2.2. They specifically highlighted that in trials with mortality as an endpoint, at least half the trials had margins corresponding to odds ratios of 1.5, much too large to reliably inform decisions.

In a similar review of how non-inferiority and equivalence trials are reported, Le Henanff and colleagues observed similar rates of a lack of rationale for the irrelevance margin (319). The results of a recent review of reporting discovered similar rates (321). They noted, “Putting merely a statement that the margin was determined based on clinically acceptable difference is not sufficient for any subsequent trial replications.”(321) The sensitivity of drawing conclusions from the adequate development and substantiation of an appropriate irrelevance margin, has been highlighted by several authors (311,317,320,322–324).

Additional problems with non-inferiority trials have also been identified, and include lack of information on patient flow (319,321), lack of appropriate sample size calculations(319), inappropriate use of open-label design, and inappropriate use of intention-to-treat analysis (321). A comprehensive review of cardiovascular non-inferiority trials published over 8 years in the *New England Journal of Medicine* and *Journal of the American Medical Association* (318) identified considerable problems. The authors observed that even in these high impact journals, substantial problems in conduct, analysis, reporting and interpretation led them to conclude that “potentially suboptimal treatments might be introduced into routine clinical practice” (318) In this study and a

follow-up study by the same authors with more in-depth analysis of cardiovascular trials, they have provided a checklist for the critical appraisal of non-inferiority trials.

Box 7: Essentials of Non-inferiority Assessment (318,325)

1. Ethical imperative:

- (a) Placebo control cannot be used because effective standard treatment is available.
- (b) New treatment should offer substantial benefits in safety, cost, or convenience over the standard treatment.

2. Choice of active control: best available comparator with large, reliable, and consistent treatment effect in placebo-controlled trials.

3. Non-inferiority margin:

- (a) Defined a priori based on clinical judgment and statistical reasoning.
- (b) Relative risk difference scale (risk, odds, or hazard ratio) preferred over absolute risk difference.

4. Adequate power and sample size to minimize type II error (false negative).

5. Proper trial design and high quality of conduct:

- (a) Identical patient population and protocols in historical placebo-controlled trials
- (b) Maximize protocol adherence.

6. Critical assumptions:

- (a) Assay sensitivity (internal validity), assured if optimal choice for active control used in the current trial
- (b) Constancy—active control effect is similar in current trial as in historical trials, assured by proper trial design and high quality of conduct.

7. Statistical analysis:

i. Fixed margin analysis

- (a) Indirect CI comparison: upper limit of 2-sided 95% CI of treatment difference < margin
- (b) Hypothesis testing: $P \leq 0.025$ to reject the null hypothesis of inequality (risk difference \geq margin)
- (c) Bayesian analysis: posterior probability of non-inferiority

≥0.975

ii. Putative placebo analysis

- (a) Superiority over imputed placebo: OR of new vs. standard treatment <1.0
- (b) Fraction preservation of active control: at least 50% for non-inferiority claim.
- (c) Bayesian analysis: posterior probability of superiority over imputed placebo >1.0 and 50% fraction preservation ≥0.975.


8. Robust interpretive criteria for non-inferiority

- (a) Stringent marginal and fractional threshold and confidence interval (2-sided 95% over 1-sided 95%)
- (b) Stability of non-inferiority inference for relative vs. absolute outcomes, and for ITT vs. per-protocol analysis
- (c) Non-inferiority claim for efficacy and superiority claim for safety/tolerability established in the same trial.

An additional problem identified by several observers and discussed in depth in a recent guidance document for comparative effectiveness reviews conducted by the Agency for Healthcare Research and Quality (326) is the problematic use of language when interpreting findings. Even when trials appear to be reported correctly, results could still be misinterpreted as to mislead. Of 33 articles (20.3%) that were adequately reported in one review (319), 4 (12.1%) had “misleading conclusions.”

Firstly, clinicians and investigators may be tempted to interpret the findings of non-inferiority trials, like superiority trials, solely on the basis of the results of the statistical test of significance. This ignores other equally important factors in interpretation, including the clinical significance, power, sample size, and significant deviations from research protocols. With non-inferiority trials, positive results more easily follow poorly conducted or underpowered trials.

Secondly, even with adequate methods and a positive test result, investigators may erroneously conclude that the comparator is non-inferior to the test drug (rather than the other way around), or that the test drug is *at least as effective* as the comparator, or that the test drug is *similar* to the comparator. All of these wording are incorrect. A conclusion that a test drug is significantly non-inferior to a comparator, although accurate, is difficult to interpret. More appropriate wording might be that the test drug is *no worse than* or *not inferior to* the comparator. Bayesian statistics must use different wording to interpret findings. From a Bayesian analysis it may be entirely appropriate to claim the probability



is 98% that the test drug is not inferior to the comparator in reducing mortality (326).

A non-inferiority trial may also allow for even *post hoc* conclusions of superiority and this is allowable if the trial is designed appropriately (324) . For example a trial may have both a placebo arm and active control arm (318). However, a non-inferiority claim is entirely inappropriate in a trial with a superiority design. Hence appropriate phraseology and conclusions are entirely dependent on the design (including sample size, power, and test of significance) and conduct of the trial.

SUMMARY OF EVIDENCE

- Mathematical models are more than likely to be better than expert opinion, but may produce varied results when compared to single head-to-head trials. They can provide information unavailable from clinical trials but are sensitive to assumptions that may not be obvious to readers. This includes assumptions about the comparative effectiveness of therapy that are based on statistical methods, such as adjusted indirect comparisons using randomized controlled trial data
- Mathematical models provide an opportunity for combining all available evidence into a coherent and transparent framework for decision making. Their value is better decision making when decision makers are able to interact with them.
- Network meta-analysis approaches for creating adjusted indirect comparisons have similar issues to meta-analysis in general
- Taken together, there is no evidence to refute or support evidence of a difference between well-conducted adjusted indirect comparisons and head-to-head trials. Although theoretically, there are some instances where one would expect direct evidence to provide better estimates.
- Non-inferiority trials are becoming more prominent. They are often inadequately reported in biomedical journals to facilitate interpretation of findings.
- Non-inferiority trials require the same level of transparency as superiority trials to facilitate their interpretation and some additional factors and assumptions unique to their design

- The language used to draw conclusions from non-inferiority trials requires special attention as it can be misleading.

OPTIONS

1. The use of mathematical models to draw conclusions about comparative effectiveness should be carefully managed.
 - [Option 1] – PAAB should discourage any claim of comparative effectiveness based on modeling. Because the underlying assumptions may not be obvious even to a reader of the original report, models harbour a potential to be misleading
 - [Option 2] – PAAB should allow claims based on mathematical modelling but only when adequate qualifying language is provided and consumers are given an opportunity to interact with the model. Qualifying language includes a disclosure of what assumptions the results of the model were most sensitive to.
2. Indirect comparisons should be discouraged or carefully managed
 - [Option 1] – PAAB can continue to insist that comparative effectiveness claims only be based on head to head trials of available data.
 - [Option 2] – PAAB can allow claims based on adjusted indirect comparisons but these should be based on a systematic review of available evidence and explanations as to why trials excluded from meta-analysis (but identified in the systematic review) were excluded and how sensitive the results are to these exclusions.
3. Non-inferiority trials should be encouraged
 - [Option 1] – In addition to measures already described under statistical reporting, PAAB should encourage forest plots depicting minimal important differences and trial results to facilitate interpretation of the results of these trials. PAAB should also insist on standard wording (e.g., no worse than)



RECOMMENDATIONS

1. PAAB should allow claims based on mathematical modelling when adequate qualifying language is provided and consumers are given an opportunity to interact with the model.


Rationale: Mathematical models can provide important information for decision making unavailable from individual drug studies. However, their real value is in allowing consumers to understand how sensitive findings of effectiveness are to variables within the care setting through careful interaction with the model. User friendly models are becoming more prevalent for clinical decision making. Mathematical models are most suitable in an interactive forum, such as promotional activities like detailing. For print advertising, a link to a model would need to be provided. Uninterrogable claims of effectiveness or relative effectiveness based on the output of a model should be strictly avoided.

2. The use of network meta-analysis for making claims of relative effectiveness should be discouraged.

Rationale: The results of a network meta-analysis may lead to an estimate of effectiveness which is misleading for the same reasons that meta-analysis in general can be misleading. These factors may not be readily obvious to a consumer, even if provided with the full details of analysis, or even obvious to the analyst who conducted the network meta-analysis, specifically when studies are combined that have very different findings. Although network meta-analysis can be helpful for understanding why studies may differ, it should be avoided until issues surrounding its reliability are resolved.

3. PAAB can allow the use of claims of comparative effectiveness from non-inferiority trials, but with specific conditions.

Rationale: Non-inferiority trials are becoming increasingly predominant and will provide an evidentiary basis for decision making when no other evidence is available. Because the clinical community currently poorly understands the implications of their findings, PAAB regulations will need to serve to both educate, through the use of qualifying language, and to ensure results are portrayed in a standard fashion. This includes appropriately stating what the active comparator was, the non-inferiority margin associated with its use, its variance, and how this was derived. Results should use consistent language (either “not worse than” or “not inferior to”). All conditions related to appropriately conveying statistical validity (section 4.2) should also apply here,



including availability of trial protocol through registration, so that an analysis plan can be appropriately appraised.

Section 5.10 of the code states, “All direct and indirect comparisons must not mislead, and be supported by reliable current data”. In the explanatory section, it is stated “Pharmacoeconomic and quality of life claims must be supported by high-quality studies. Disclosure of study parameters, Section 5.11, is important for interpretation of results.” There is no specific guidance for study parameters that apply to pharmacoeconomics studies.

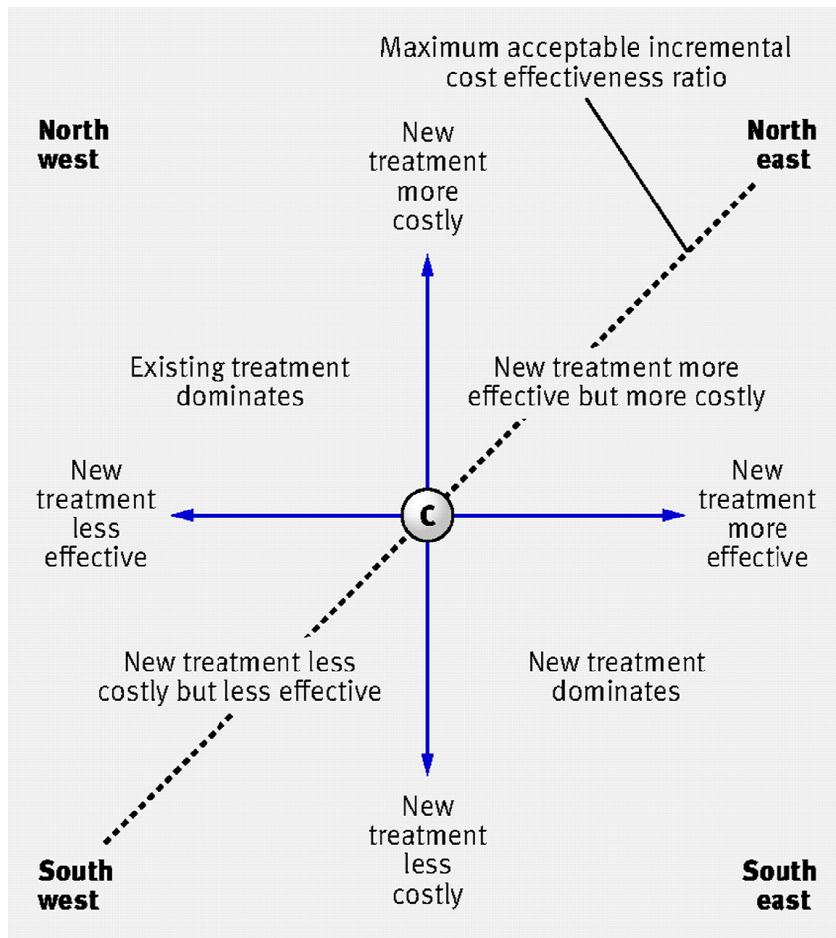
HOW SHOULD HEALTH ECONOMIC CLAIMS BE MADE?

INTRODUCTION

Economic evaluations attempt to capture the cost and consequences of choices in health care (327). Economic evaluations aim to provide decision makers with one or more measures of value compared with one or more measures of resource use so that decision makers can decide whether or where resources are best spent. If a single metric of effectiveness is used, the results of the study can be depicted on a cost-effectiveness plane. (See Figure) The terms economic evaluation, economic analysis and economic study mean the same thing, although in recent years *economic evaluation* has been identified as a preferred term (327). The term pharmacoeconomic analysis refers specifically to analysis of different drugs but has become less widely used. Cost-effectiveness analysis has also been used synonymously with the term economic evaluation. Cost-effectiveness literally implies a comparison of costs and effectiveness (hence the hyphen), however cost-effectiveness has also been used to describe a specific type of economic evaluation where costs are compared with a unit of health. Hence the term “cost-effectiveness” can lead to confusion.



FIGURE 8: THE COST-EFFECTIVENESS PLANE




Petrou S , Gray A *BMJ* 2011;342:bmj.d1548

©2011 by British Medical Journal Publishing Group



Types of economic evaluation have been distinguished according to what *metric of value* healthcare resource costs are being compared to. Palmer describes two types of economic evaluations: cost-benefit and cost-effectiveness analysis (328). Cost-benefit analysis value all resources used or saved as a result of the intervention in monetary terms. They compare net benefits (benefits minus costs measured in monetary units) associated with one drug to the net benefits associated with another. They are less frequently conducted, as they force analysts to put a value on human life. Cost-effectiveness analysis, in contrast, compares the increased costs of the resources associated with the use of a new treatment compared to a direct measure of health, such as a myocardial infarction avoided, or years of life gained.




A third form of economic evaluation is cost-minimization analysis (CMA). CMA treats the outcomes of the interventions or technologies being evaluated as identical, shifting the focus to which has the lowest costs. For example, two treatments for hypertension may both reduce systolic blood pressure by an average of 10 mm of mercury; a CMA would then focus on the costs of each, with the goal of identifying which is the least costly approach to treating hypertension. Cost-minimization analyses are problematic in that it is rare for any two treatments to be *identical* in reality, and the assumption of equivalence can ignore small and uncertain effects with meaningful economic consequences (329).

A fourth type of economic evaluation is a cost-consequences analysis. This type of study does not try to present tradeoffs of costs for a single measure of effectiveness or benefit, but rather, presents multiple measures that might be of interest to patients and policymakers. Often, these are health outcomes of interest. For example, an examination of screening for colorectal cancer could present the incremental costs compared with unnecessary referrals avoided, cancer cases avoided, and deaths averted.(330)

A special type of cost-effectiveness analysis is the cost-utility analysis. This type of analysis, widely promoted for informing reimbursement decisions, is a type of cost-effectiveness analysis where the unit of effectiveness incorporates some measure of patient preferences for health. That is, an objective unit of health (like additional years of life) is adjusted for a patient's health-related quality of life, functional status or other subjective endpoint. A popular unit of health in cost-utility analyses is the quality-adjusted life-year. It should be recognized that the use of these terms may not be entirely consistent among economists – for example some economists would not use the term “cost-utility” analysis altogether. Challenges with these types of analyses stem from adequately capturing individual preferences for health while still providing a measure that can be used for appropriately allocating scarce resources.(331)

Economic evaluations must always incorporate the various methods of clinical and comparative analysis already discussed in other sections of this report to estimate the clinical effectiveness of an intervention. This includes the use of mathematical modeling (e.g., disease outcome modeling), observational data, trial analysis (including non-inferiority trials), systematic review, patient-reported outcomes, meta-analysis, and adjusted indirect comparisons. Interestingly, economic evaluations may eschew the use of standard clinical null hypothesis testing, since the underlying uncertainty in any healthcare decision also represents value (in monetary or health terms) for the decision maker and must also be taken into account (332). Rather than being straightforward, judgments must also be made to estimate resource use and their associated



costs which can in turn affect the findings, such as top-down or bottom-up approaches to (333). Because an economic evaluation may require “stitching together” many disparate pieces of information from multiple sources, one observer poignantly labeled them “Frankenstein’s Monster” (334).

The first relevant question is what elements require reporting to allow proper interpretation of an economic evaluation? A second and related question is what study parameters are of most importance in assessing the validity and reliability of an economic evaluation and determining which studies are of sufficiently high quality to allow comparative claims of cost-effectiveness?

Considerable effort has been taken to produce checklists and guidance to aid authors and readers in the correct reporting and interpretation of economic evaluations. Although related concepts, there is much confusion about checklists to promote the interpretation, quality and conduct of evaluations versus checklists to improve the reporting of evaluations. For example, one paper suggests in its title guidance for reporting but actually gives guidance for conduct of studies (335). Although elements for reporting and appraisal are related, they are not the same: Firstly, lack of reporting does not necessarily imply lack of conduct. For example, information may be omitted for the purpose of satisfying editorial space requirements. Secondly, reporting guidance provides authors with instructions on what aspects of a study must be reported, but not how they should be analyzed. Thirdly, elements needed for reporting may not have equal weight in terms of their contribution to the validity of an economic evaluation – an appraisal checklist should allow readers to discriminate between high and low-quality studies.

To explore the first question, what is required to allow proper interpretation of a study, the next section will examine the reporting checklists that have been and are currently under development. Then, the following section will explore checklists and published guidance for interpreting economic evaluations to better understand what elements are most important to properly assess the validity and reliability of an economic evaluation.

EVIDENCE

REPORTING

Reporting guidance is intended to make reports of economic evaluations interpretable by promoting consistency and transparency. Figure 9 provides a list of all currently existing published guidance documents and checklists for reporting economic evaluation.



FIGURE 9: PUBLISHED REPORTING CHECKLISTS FOR ECONOMIC EVALUATION

First Author	Year	Description	Items	Ref
Drummond	1996	Consensus panel - Instructions for authors to BMJ – wide uptake	40	(327,336)
Gold	1996	Consensus panel - US Public Health Service Appointed – wide uptake	37	(337,338)
Vintzileos	2004	Intended for economic evaluation in obstetrics	33	(339)
Drummond	2005	Suggestions for improving generalizability and uptake of studies	10	(340)
Ramsey	2005	ISPOR Task Force guidance for economic evaluation alongside clinical trials	14	(341)
Goetghebeur	2008	Suggestions for structured reporting to improve decision making	11	(342)
Petrou	2011	General guidance for economic evaluation alongside modelling and clinical trials	N/A	(343,344)

There are additionally many country-specific guidance documents for reporting economic evaluation specific to confidential or less-widely circulated reports related to reimbursement processes (345). For example, a Dutch Task Force recommended guidelines for reporting pharmacoeconomic analyses for applications to the Dutch health system (346,347) Similarly, guidelines produced by the Academy of Managed care Pharmacy in the US enable pharmaceutical companies to respond to requests for economic data from health plans, within the guidelines laid down by the Food and Drug Administration and have enjoyed wide uptake (348). Three iterations of country-specific guidance for reporting have also been developed in Canada by the Canadian Agency for Drugs and Technologies in Health (349).

The most well known and rigorously-developed reporting checklists are arguably two independent checklists produced for the BMJ medical journal and US Public Health Service (336,337). These guidelines used a consensus approach across varying perspectives to determine the elements reporting elements essential to the interpretability of economic evaluations. A synthesis of elements from these guidance documents as well as others is shown below. Although an International standard is not yet available, the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) has convened an International Task Force to develop Reporting Standards for Health Economic Evaluations which are still in development (350).

FIGURE 10: ELEMENTS FOR REPORTING ECONOMIC EVALUATIONS IDENTIFIED IN PREVIOUS GUIDANCE DOCUMENTS

Reporting Element	Description/Reporting Guidance
TITLE	
Title	Identification the report as an economic evaluation
ABSTRACT	
Structured Abstract	Provide a structured abstract including background, objectives, perspective, form of analysis, population, whether trial- or model-based or both, comparators, measure of benefit, measure of costs, discount rate, and analysis of uncertainty as well base case results and key results from the uncertainty analysis and conclusions that are supported by the results
INTRODUCTION	
Question and rationale	Provide an explicit statement of the broader context for the study as well as the decision problem, a question(s) in answerable form that is relevant to the decision, and the importance of the study question for health policy or practice. The question should contain the target population, comparators of interest, key outcome and the setting.
METHODS - General	
Target population and subgroups with rationale	Describe characteristics of the target population that is being considered for the base case analysis, and if there evidence of heterogeneity of costs or effects in different subgroups, justify the performance or not of any subgroup analysis and the characteristics that are important to the decision problem.
Setting and location	State the country(ies) in which the economic evaluation is set and the clinical setting/level in which the intervention is provided and any other relevant aspects of the health system/s in which the decision/s needs to be made.
Perspective and rationale	Describe the perspective (health system, payer, society) in terms of costs included, the associated components (direct costs, indirect costs, and to whom), and how this fits the needs of the target audience.
Comparators and rationale	Describe and justify what interventions are being compared and relate these to the decision context. An intervention can be to “do nothing”.
Time horizon and rationale	State the length of time in which health outcomes and resource use are being evaluated and why this is appropriate given the clinical or policy decision.
Type of analysis and rationale	Describe and justify the form(s) of analysis of the evaluation (cost-benefit, outcome, effectiveness, utility, or minimization study).
Data sources and rationale	In a table, describe source of data for costs and outcomes of interventions and associated uncertainty. Describe if it is study-based, model-based, or it has both components. If a modeling-based analysis, describe source of data for each parameter.
Data synthesis methods	Describe approach to transforming data from source data to use in economic evaluation in sufficient detail to allow replication.
Year of Costing	Present the year used for the presentation of the cost estimates
Discount rate and rationale	An annual rate, described in percentage, used to calculate current values of resources or health outcomes

Availability and cost of data	Describe whether data are available to others, how it can be obtained and whether there are associated fees for use.
METHODS - Measurement and valuation of outcomes	
Outcomes and rationale	Describe what outcomes were used in the evaluation and their relevance to the type of analysis. These might include, but are not limited to, outcomes expressed in natural units, e.g. life years gained or lives saved, for the purposes of cost-effectiveness analysis; outcomes expressed in terms of quality-adjusted life years (QALYs) for the purposes of cost-utility analysis; or outcomes expressed in monetary terms for the purposes of cost-benefit analysis.
Measurement of clinical effectiveness	<p><i>Single study</i>-Report design features of single effectiveness study, including the methods of selection of study population; methods of allocation of study subjects; whether the intention to treat analysis was used, time of follow-up, and methods for handling potential biases in the study design, for example, selection biases. It is important to justify why the single study was a sufficient source of clinical effectiveness data.</p> <p><i>Synthesis of evidence</i>-Describe methods of synthesis of clinical effectiveness data, including the methods of systematic review and search strategy; potential biases arising from the inclusion of non-randomized studies; methods of study selection and data extraction; methods of meta analysis; and the use of indirect and mixed treatment comparisons where appropriate.</p>
Measurement of preference-based outcomes	Describe the method of measurement of preference-based outcomes, for example, the use of a multi-attribute utility measure (e.g. EQ-5D, SF-6D), a direct scaling technique (e.g. standard gamble approach, time trade-off approach), contingent valuation, discrete choice experiment, etc. The format and timing(s) of these measurements should also be described
Valuation of preference-based outcomes	The population from whom valuations of preference-based outcomes were obtained should be described in terms of size and characteristics (i.e., patients, general public, carers). This population may differ from the study population for the economic evaluation. Statistical modeling techniques used to derive valuations should also be outlined.
METHODS – Measurement, valuation and analysis of costs	
Methods of estimation of resource quantities	The methods of estimation of resource quantities should be described. For trial-based economic evaluations, describe approaches to estimating resource use associated with care of patients, for example, trial data collection forms, separately designed economic questionnaires, data extracted from routine data collection systems, etc. For modeling-based economic evaluations, describe approaches to estimating resource use associated with health states or prognoses.

Methods of valuation of resource quantities (unit costs)	Describe primary research methods for valuing each resource item in terms of its unit cost. If secondary sources are used for unit costs, describe the underpinning accounting procedures. Describe any adjustments made to approximate to opportunity costs.
Reporting of resource quantities, unit costs and total costs	Report mean resource use and variability for the main resource items of interest. Separately report unit costs associated with each resource item and the source of this information. Report mean cost and variability for the main cost categories of interest, and for total cost, as well as mean differences between the comparator groups and their associated confidence intervals.
Currency and price date	Report the dates for the estimation of resource quantities and unit costs. Identify the currency in which costs are reported,
Price adjustments and currency conversion	Describe methods for adjusting costs to a recent price level (e.g. health care specific pay and prices index). For economic evaluations performed on a multinational basis, describe methods for converting costs into a common currency base (e.g. purchasing power parities).
Analysis of costs	Describe methods for dealing with skewed, missing or censored cost data where these arise.
METHODS – Modeling (for modeling studies)	
Model type and rationale	If modeling, describe the type of model or simulation model method used (decision tree, Markov model, system dynamic) and specific type of simulation model employed (e.g., decision tree, semi-Markov, Markov decision process, discrete event simulation, agent-based simulation)
Detailed model structure	If modeling, describe and/or illustrate the complete structure of the model in a way that allows replication.
Model input parameter values and values for sensitivity analysis	Present a tabulated listing of each of the parameters required to run the model and their associated values including distribution or other relevant values related to uncertainty and variability. This includes ALL clinical and economic parameters that would be needed by a reader wishing to replicate the model.
Sources for input parameters	Describe the rationale for selection of the data sources used.
Valuation of parameters from selected data sources	Describe the methods used to generate the input parameter values, ranges, and if used, probability distributions from the selected data sources for ALL clinical and economic parameters.
Valuation of parameters for variability analysis	Describe methods used to quantify parameter values and distributions for variability analysis. E.g., Patient subgroups
Model validation/calibration	Describe if and how the model was validated and/or calibrated using real-world data.
Model assumptions	Describe ALL underlying structural or other assumptions in the model that would be needed by a reader wishing to replicate the model.
RESULTS	

Incremental costs and consequences	Report incremental costs and outcomes separately both as total costs and outcomes as well as individual cost categories and clinical outcomes from which they were derived.
Sensitivity analysis of model structure or key assumptions	Describe the sensitivity of the results to structural or other key assumptions. If no sensitivity analysis is conducted, justify this. (e.g., model validity)
Sensitivity analysis of model inputs	Describe the sensitivity of the results to parameter value uncertainty.
Variability analysis of model inputs	Describe the impact of variability (e.g., alternative patient or practice or market characteristics) on the findings.
DISCUSSION	
Study findings	Summarize key study findings and conclusions.
Study strengths	Describe the strengths of the approach taken and how this strengthens the conclusions drawn.
Study limitations	Describe the weaknesses of the approach taken and key changes that would affect the study's conclusions.
Generalizability of results	Describe the applicability of the analysis to the participants in the setting and what settings the findings will not apply to.
Ethical issues	Identify any issues of ethics that are not addressed by the economic evaluation
Equity issues	Identify any relevant equity issues. Specifically identify any impacts relating to geographical equity, equity by socio-economic status and impacts on minorities
Implications for health care system policy	Identify other issues that might be relevant to decision makers. For example, budgetary impact and affordability, need for training, changes in skill-mix or other organizational impacts
Implications for practice	Identify any issues for health care providers.
Implications for research, including economic evaluation research	Identify needs for further research revealed by the findings.
OTHER	
Source of funding	Describe how the study was funded and the role of the funder in the identification, design, conduct and reporting of the analysis.
Contributor conflicts of interest	Describe any potential for conflict of interest across study contributors that have occurred within the last 5 years.
Original model / code	Provide a guarantor who can be contacted for access to the original model. Additionally provide a link or code to original model.
Protocol	Provide a description of a guarantor who can be contacted to obtain the original study protocol and subsequent amendments for 5 years after study publication. Additionally provide a link if available.

CONDUCT/QUALITY


Checklists and guidance, with the intent of providing a list of key items to guide or judge the conduct of the study (rather than guide or rate what has been

reported) have also been developed (272,335,351–355). Published guidance for clinicians on the critical appraisal of economic evaluations was loosely based on Drummond’s BMJ guidance for reporting, with more prescriptive quality measures enforced (351). Similarly, one group of researchers developed and assessed scoring methods based on items from Drummond’s BMJ Checklist for Authors (352).

In another more robust attempt at creating a weighted grading system, all available checklists for conduct and reporting were pooled and the weight of each item graded. The authors developed a 16-item checklist (the Quality of Health Economic Studies Instrument, or QHES) (353) with both face and construct validity subsequently shown to be capable of discriminating high and low-quality studies (356). The QHES is shown below. A similar effort based on 25 International guidelines led to a checklist of 20 items (354) called the Consensus on Health Economic Criteria (CHEC) guidance document. There are also many country-specific guidance documents for economic evaluation, that may or may not have a checklist associated with them, but have been used by authors to interpret the validity and reliability of economic evaluations.(357,358)

Box 7: The Quality Of Health Economic Studies Instrument (QHES)

1. Was the study objective presented in a clear, specific, and measurable manner?
2. Were the perspective of the analysis (societal, third-party payer, etc.) and reasons for its selection stated?
3. Were variable estimates used in the analysis from the best available source (i.e. Randomized Control Trial—Best, Expert Opinion—Worst)?
4. If estimates came from a subgroup analysis, were the groups pre-specified at the beginning of the study?
5. Was uncertainty handled by: 1) statistical analysis to address random events; 2) sensitivity analysis to cover a range of assumptions?
6. Was incremental analysis performed between alternatives for resources and costs?
7. Was the methodology for data abstraction (including value health states and other benefits) stated?
8. Did the analytic horizon allow time for all relevant and important outcomes? Were benefits and costs that went beyond 1 year discounted (3–5%) and justification given for the discount rate?
9. Was the measurement of costs appropriate and the methodology for the estimation of quantities and unit costs clearly described?
10. Were the primary outcome measure(s) for the economic evaluation clearly stated and were the major short term, long term and negative outcomes included?


- 
11. Were the health outcomes measures/scales valid and reliable? If previously tested valid and reliable measures were not available, was justification given for the measures/scales used?
 12. Were the economic model (including structure), study methods and analysis, and the components of the numerator and denominator displayed in a clear transparent manner?
 13. Were the choice of economic model, main assumptions and limitations of the study stated and justified?
 14. Did the author(s) explicitly discuss direction and magnitude of potential biases?
 15. Were the conclusions/recommendations of the study justified and based on the study results?
 16. Was there a statement disclosing the source of funding for the study?

A more recent assessment of both QHES and CHEC Checklists suggests they perform similarly and can be reliable, but scores are often more dependent on the rater rather than the study being rated (359). This suggests the need for caution when a single or untrained reviewer applies the checklist for the purpose of appraising a study. In an attempt to move away from checklists, in part because of evidence suggesting reliance on quality scores of RCTs can be misleading (147), the use of a case-by-case multi-criteria decision approach (GRADE) to appraising the quality is being promoted (162). A GRADE approach to economic evaluation has been developed (355). The GRADE approach more strongly emphasizes examining evidence in context and across multiple dimensions and raters rather than relying on single scores.

HOW SHOULD CLAIMS OF IMPROVEMENTS IN PATIENT-REPORTED OUTCOMES/ HEALTH-RELATED QUALITY OF LIFE BE MADE?

INTRODUCTION

Measures of clinical effectiveness typically reflect morbidity related to the disease or patient longevity. These outcomes that can be measured without asking patients insofar as the presence of a death, myocardial infarction or malignant growth can be identified and measured by someone other than the patient and using a clinical definition. What these measures do not tell us is how a patient is feeling. The obvious need to use measures of wellbeing as a goal becomes even more important in clinical situations where the primary goal of treatment is wellbeing rather than prolongation of life or amelioration of disease. More importantly, individuals with the same health status or disease may perceive their health-related quality of life (HRQoL) quite differently, as



their ability to cope with limitations and disability and other factors can alter perception about satisfaction with life (360). Health-related quality of life measures have also seen application in policy and administration of health care including screening and monitoring for psychosocial problems, population surveys of perceived health problems, medical audits, health services or evaluation research, and economic evaluation (361).


Health-related quality of life measures are intended to capture patient experiences and have been promoted as a means of understanding how a patient is feeling about their own health. The generic term “quality of life” has been used interchangeably with *self-reported health*, *patient-assessed outcomes*, *patient-reported outcomes*, *person-reported outcomes*, *patient outcomes* and *outcomes*. Although there is no overarching consensus on the proper use of terms, the term *health-related quality of life* has been widely adopted and promoted, as the term *quality of life* implies an evaluation of the effect of all aspects of life, rather than health-specific aspects on general well-being.

The term *patient-reported outcome* (PROs) has grown with changes in FDA regulations to allow the use of patient-reported outcome measures to support labeling claims (362). The FDA defines a patient-reported outcome as “a measurement based on a report that comes directly from the patient (i.e., study subject) about the status of a patient’s health condition without amendment or interpretation of the patient’s response by a clinician or anyone else. “. It then states, “A PRO can be measured by self-report or by interview provided that the interviewer records only the patient’s response. “ (362) According to FDA terminology, the term patient-reported outcome is more specific than HRQoL as a patient-reported outcome may only capture one aspect of health (such as psychological well-being) whereas as HRQoL captures all aspects of health – physical, mental, and social wellbeing – reflecting the current WHO definition.

Patient *health status* and *functional status* have also been used synonymously with the term HRQoL despite the fact that these measures do not necessarily require information from the patient’s perspective. In some cases input from the patient and others, like the provider, are combined while in other cases the measure reflects input from the patient or provider alone. Similarly, there exist proxy-reported outcomes, which are derived from information from parents, providers or caregivers about their perceptions of how a patient is feeling. For the purpose of this section of the report, we will focus strictly on PROs and HRQoL.

MEASUREMENT OF QUALITY OF LIFE - TERMINOLOGY

To measure PROs or HRQoL we must have systems to extract information from patients. The term *instrument (or index)* is used to describe a method to capture



data plus the setting and other relevant information required to support its use. For example an *instrument* could be a questionnaire along with the instructions for administration or responding to the questionnaire, a standard format for data collection, and methods for scoring, analysis, and interpretation of results in the patient population (362).

An *item* refers to an “An individual question, statement, or task (and its standardized response options) that is evaluated by the patient to address a particular concept”. An *item* could be a question like “How alert do you feel?” A *scale* is used as a means of responding to an item. The patient may be given an opportunity to answer the previous question using an open-ended response, or be provided with categorical answers (e.g., ‘Very alert’, ‘Somewhat alert’, ‘Not alert’, ‘Difficulty staying awake’) or with a visual analogue response (e.g., a scale with ‘Very alert’ at one end and ‘Difficulty staying awake’ at the other) where the patient is asked to describe feelings using a continuous, rather than categorical system.


Instruments composed of many items are developed with the specific goal of measuring one or more *concepts*. A concept is the thing to be measured (e.g., pain intensity improvement, symptoms associated with a condition, or HRQoL). Depending on its complexity, a concept may or may not require multiple items. It also may or may not be further divided into specific *domains*. A *domain* refers to a particular focus of attention or subconcept and may be addressed by one or more *items*. For example, performance in the domain of “cognitive function” may be measured by responses to one or more *items* related to this domain whereas “emotional function” may be measured by one or more items that may or may not overlap with these items.

Various types of instruments have been developed to capture PROs and can be classified according to the type of concepts they are trying to measure and the outputs of the information collected. *Generic* instruments can be used in a wide variety of populations, conditions, or treatment settings whereas *specific* instruments are suitable only in specific patient groups, conditions or areas of function (360,363). Modular instruments and batteries of scales combine generic and the disease-specific approaches by adding disease- and therapy-specific questions to a core module of questions if needed. A taxonomy of measures is shown in Box 8

BOX 8: TYPES OF PATIENT-REPORTED OUTCOME MEASURES (ADAPTED FROM (364))

Dimension-specific measures focus on particular aspects of health such as psychological wellbeing and usually produce a single score—for example, Beck depression inventory.

Disease- or population-specific measures include aspects of health that are



relevant to particular health problems and may measure several health domains—for example, asthma quality of life questionnaire.

Generic measures can be used across different patient populations; they usually measure several health domains—for example, SF-36.

Individualized measures allow respondents to include and weight the importance of aspects of their own life; they usually sum to produce a single score—for example, patient generated index.


Utility measures have been developed to measure individual preferences for health states, and produce a single index—for example, EuroQol EQ-5D.

The development and correct interpretation of measures of PRO concepts first requires an understanding of inferences that can be drawn given the study design. PRO instruments may be administered in either an experimental or non-experimental study design. For example, an instrument may be administered as a population survey according to a cross-sectional design. Alternatively, it could be measured in the context of a large randomized controlled trial. The validity of estimates must first take into account the potential for systematic and random error inherent in these research designs. They may also be the primary variable of interest for which a study was designed, or more often a secondary variable of interest.

To minimize the introduction of measurement error, analysts must consider the reliability, validity, and ability to detect change of the instrument employed (360,363,365). Reliability refers to the “the ability of a PRO instrument to yield consistent, reproducible estimates of true treatment effect” (362). This can be inter- or intra-interviewer reliability or the reliability of the test itself.

Validity can be broken down into several dimensions: firstly, *face validity* refers to whether the instrument covered the relevant range of topics (361). *Content validity* is an attempt to see if the instrument actually measures what it is supposed to. *Construct validity* is an attempt to ensure measures are consistent with other measures and knowledge about the relationships between domains and concepts. Construct validity can be established by looking at the degree to which the instrument can distinguish between groups known to have different perceptions of illness.

An ability to detect change can be characterized by both the responsiveness of an instrument to change and the sensitivity of the instrument. Responsiveness refers to changes in the outcomes reflecting changes in the patients. For example, increasing doses of a pain-reliever should result in decreasing pain (360). Sensitivity refers to the ability for changes in the patient to cause measurable changes in the scale of the instrument (360). An additional concern



that has been raised is whether an instrument is appropriate, particularly when used in different cultural or linguistic settings (361,366).

Finally, it is important to recognize that the instruments and measurements in and of themselves provide us with a descriptive measure but not a measure of *value*. Assigning a value, such as a score or utility to the descriptive findings of these instruments requires identifying a means to score them. This could be through using preference-based, non-preference based or other approaches with a sample of the general population or a specific population.(367)

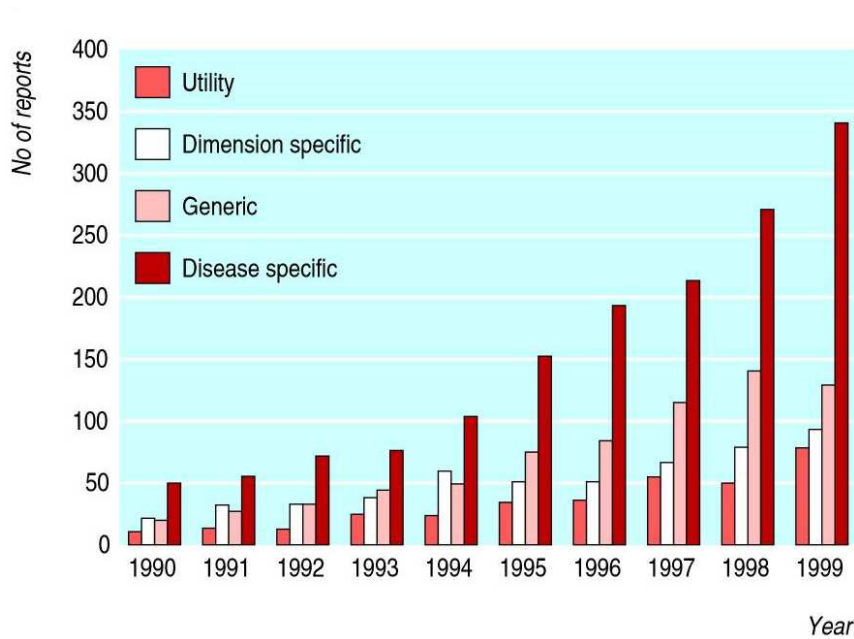
Given the analytic complexity and terminology associated with the use of PRO and HRQoL instruments, it is no wonder that their usefulness for informing clinical decisions has been questioned (365). The question to be addressed is what information is most useful for the correct interpretation of studies of these outcomes, so as not to mislead consumers.

EVIDENCE

The use of patient reported outcome measures has grown. In a comprehensive search for published studies describing the development and evaluation of measures, one group of investigators discovered published reports rose from 144 to 650 per year during period of 1990-9 (364). During that same time period, they identified 3921 reports describing development and evaluation of PROs; of those they could classify, 1819 (46%) were disease or population specific, 865 (22%) were generic, 690 (18%) were dimension specific, 409 (10%) were utility, and 62 (1%) were individualized measures (364). (See Figure 6) From 1980-97 reporting on quality of life increased from 0.63% to 4.2% for trials from all disciplines, from 1.5% to 8.2% for cancer trials, and from 0.34% to 3.6% for cardiovascular trials (368).




FIGURE 11: INCIDENCE OF PUBLISHED STUDIES OF PROS



With the advent of FDA guidance, specific attention has been paid to the proper validation and application of PROs (369–371) in the context of clinical research. Tools have been identified to help researchers with the validation and development of new instruments. Beyond specific issues related to trial design and conduct (e.g., sample size, hypotheses being tested, statistical analyses) particular issues related to the evaluation of instruments in the context of clinical trials have been identified (362,365,366,372–374) A summary of concerns from the published literature (365,366,375–377) is shown in Box 9.

Box 9: Issues specific to studies measuring PROs and HRQoL

- Are patients or a proxy being asked for information?
- How were the instruments chosen?
- How were the instruments scored?
- How was the instrument validated in this population?
- What was the time frame and timing of assessment?
- Was there a need for cultural adaptation?
- What is the minimal important difference?
- What was the magnitude of change in the scale?
- What were both the number of assessments completed and items completed versus the number expected?
- How were missing data handled?
- Were interviewers trained?
- What quality assurance mechanisms were in place?
- Has multiple testing been addressed?
- How do the findings compare with other studies?



There is considerable evidence that reports of quality of life assessments from clinical trials are poorly reported and may be prone to bias (365,378,379) Although clinical trials reporting quality of life information appear to be similarly reported to those that don't, specific items related to the choice and application of concept-specific instruments were lacking (379). A guideline for reporting quality of life measures in clinical trials comprised of 76 items with 8 main section headings has been proposed by an expert panel (378)

Even if the conduct of a trial is transparent, the meaning of the findings may be uninterpretable to clinicians for making individual decisions. Guyatt and colleagues have created a series of proposals for making information regarding quality of life less misleading to clinical decision makers (376,380–383). They emphasize adequate reporting of the minimal important difference (MID) on PRO instruments, and focusing away from a mean difference and examining the proportions of patients achieving and not achieving a benefit and the magnitude of benefit according relative to the MID. They also advocate for the use of interpretation aids for clinicians and patients (376).

SUMMARY OF EVIDENCE

- There are many currently available reporting standards for economic evaluation, although no one, widely recognized international standard
- Appraisal checklists and guidance have been developed in a robust fashion for assessing the validity of economic evaluations
- Reporting of health-related quality of life from clinical trials tends to be poor and prone to bias. Some reporting standards have been developed although there is no one, widely recognized international standard
- Appraisal checklists and guidance have been developed for assessing the validity of health-related quality of life assessments
- Focusing away from mean differences and focusing on the MID for PRO instruments is a feasible and reasonable approach to reporting PRO data in a less misleading way for consumers.




OPTIONS

1. Comparative effectiveness claims from economic evaluations should be scrutinized
 - [Option 1] – PAAB can use either the CHEC or QHES checklists to determine whether a study is suitable. PAAB could only allow CEA from clinical trials. Costs and effects should be disaggregated and claims of “cost-effectiveness” should be accompanied by a stated assumption about willingness to pay.
 - [Option 2] – PAAB can use either the CHEC or QHES checklists to determine whether a study is suitable. PAAB could allow CEA from either clinical trials or mathematical modeling of effectiveness. It can then insist that cost-effectiveness studies only be based on effectiveness analysis based on systematic reviews and with modeling sensitivity assumptions stated. Costs and effects should be disaggregated and claims of “cost-effectiveness” should be accompanied by a stated assumption about willingness to pay.
2. Comparative effectiveness claims from HRQoL and PRO studies should be scrutinized
 - [Option 1] – PAAB can use existing appraisal instruments (for example, (377) to determine whether a study is suitable.
 - [Option 2] – PAAB can use existing appraisal instruments (for example, (377) to determine whether a study is suitable AND it should then endorse the approach advocated by Guyatt (376) for reporting results.

RECOMMENDATIONS

1. PAAB should allow claims based on economic evaluation when adequate qualifying language is provided and other regulations are consistently applied.

Rationale: Economic evaluations can provide important information for decision making unavailable from individual drug studies. PAAB could pre-qualify a study using CADTH economic evaluation guidelines and applying a QHES or CHEC checklist with a PAAB rater who is appropriately trained in appraisal and use of the checklist. Qualifying language about this type of study should be present in the advertisement. Standard information and language will need to be adopted when reporting the findings of an economic evaluation. At a minimum, costs and



measures of effectiveness should be reported separately and use of the term cost-effective should only apply to those drugs which reduce costs and improve effectiveness. Because many evaluations are based on mathematical modeling and extrapolation, recommendations for modeling will apply as they, too, are most suitable in an interactive forum, such as promotional activities like detailing. For print advertising, a link to an economic model would need to be provided. In rare cases where an economic evaluation is based on the results of a single study, PAAB regulations governing reporting of that type of study (e.g., superiority RCT, non-inferiority RCT, or observational) will need to be applied.

2. PAAB should allow claims based on HRQoL and PRO measures, but with specific conditions.

Rationale: HrQoL and PRO measures provide additional information that may be helpful to consumers when making therapeutic choices. However, studies using these measures are susceptible to bias and should be appraised before being approved for use. Because the clinical community currently poorly understands the implications of their findings, PAAB regulations will need to serve to both educate, through the use of qualifying language, and to ensure results are portrayed in a standard fashion. The approach most recently proposed by Guyatt (376) is very feasible and an excellent starting point for reporting results in a fashion that are useful to consumers.



SUMMARY OF RECOMMENDATIONS

The following 17 recommendations were developed through consideration of current best practice, a thorough examination of the evidence of the susceptibility to bias of the methods employed and consultation with national and international leaders in the fields of consumer policy, observational and outcomes research, biomedical journal editing, economic evaluation and modeling, systematic review, meta-analysis and network meta-analysis, epidemiology, biostatistics, and health-related quality of life measurements (see Appendix). Each recommendation was developed using the analytic framework presented in the introduction and through a consideration of their feasibility within the PAAB context.

A comment accompanies recommendations that might be perceived as controversial, where there are questions about feasibility, and where further work based on knowledge gaps is required.

RECOMMENDATIONS TO THE PHARMACEUTICAL ADVERTISING ADVISORY BOARD REGARDING THE USE OF SCIENTIFIC INFORMATION IN ADVERTISING

1. *P* values should be discouraged wherever possible except under exceptional circumstances and consistent with current guidance from biomedical journals

Comment: This relatively straightforward recommendation must only be adopted if an alternative (recommendation 2) is adopted. It may be seen as controversial by some (those used to seeing P values) but is not scientifically controversial.


2. Confidence intervals should be encouraged instead of *P* values wherever possible and consistent with current guidance from biomedical journals. PAAB should suggest only 95% Confidence Intervals (CI) are appropriate for the presentation of findings rather than *P* values.

3. Publication of information from clinical trials should be discouraged if research protocols and outcomes have not been registered and are readily accessible by PAAB and the health care providers that they serve. PAAB should additionally mandate manufacturers provide a link to the registered information in advertisements AND endorse the Ottawa statement

Comment: This relatively straightforward recommendation may be seen as controversial by some of those who have previously raised concerns about overexposure of commercial in confidence information.

4. The wording of PAAB Code requirement 4.2 needs correction and revisiting

5. PAAB should revisit Code requirement 4.2 and make additional provisions in the Code Explanatory Notes that Bayesian statistical testing is acceptable.



Comment: This is straightforward but some further scientific consensus as to what Bayesian information is required.

6. If claims from individual studies are used, information regarding the total number of similar studies conducted (in terms of patients, interventions, design) from a *systematic* review of available evidence should be made available to reduce selection bias or claims based on exaggerated study findings.

Comment: This may be confusing to some and controversial to others. A meta-analysis is not being asked for, nor is a reference to a published systematic review. The advertiser must simply identify all similar studies based on their current knowledge and provide information on the across-study variance for the effectiveness claim.

7. The use of meta-analysis for making claims of effectiveness should be discouraged.

Comment: Some may see this as too restrictive since the conduct of meta-analysis is so widespread.

8. The use of unpublished research findings should not be discouraged.

Comment: Some may see this as surprising or controversial. Adopting this recommendation requires an understanding of how reliance on published-only information can be misleading. Similar to allowing scrutiny of trial protocols, some may be concerned about knowledge management aspects associated with this approach and exposing commercial in confidence information.

9. PAAB can allow the use of subgroup analysis, but with specific conditions.

Comment: This is straightforward but will need further scientific consensus as to how these claims are scrutinized and what information is required


10. PAAB can allow the use of claims from secondary outcomes, but with specific conditions.

Comment: This is straightforward but will need further scientific consensus as to how these claims are scrutinized and what information is required

11. Post hoc analysis should continue to be discouraged

12. PAAB can allow the use of claims from observational studies, but with specific conditions.

Comment: This is straightforward but will need further scientific consensus as to how these claims are scrutinized and what information is required



13. PAAB should allow claims based on mathematical modelling when adequate qualifying language is provided and consumers are given an opportunity to interact with the model.

Comment: This may seem controversial since meta-analysis, often used to inform mathematical modeling is being discouraged. However, promoting interactivity with models is consistent with the added value that models bring. Consumers can be provided an opportunity to change meta-analytic inputs as well as other important variables to allow them to see how long-term outcomes, not empirically observed, may change.

14. The use of network meta-analysis for making claims of relative effectiveness should be discouraged

Comment: Some may see this as too restrictive since the conduct of network meta-analysis is increasing in prominence.

15. PAAB can allow the use of claims of comparative effectiveness from non-inferiority trials, but with specific conditions

Comment: This is straightforward but will need further scientific consensus as to how these claims are scrutinized and what information is required

16. PAAB should allow claims based on economic evaluation when adequate qualifying language is provided and other regulations are consistently applied.


Comment: This is straightforward but will need further scientific consensus as to how these claims are scrutinized and what information is required


17. PAAB should allow claims based on HRQoL and PRO measures, but with specific conditions.


Comment: This is straightforward but will need further scientific consensus as to how these claims are scrutinized and what information is required


REFERENCES


1. Canadian Institute for Health Information. Drug expenditure in Canada, 1985 to 2010. Ottawa, Ont. : Canadian Institute for Health Information,; 2011.
2. Mulley AG. Inconvenient truths about supplier induced demand and unwarranted variation in medical practice. *BMJ*. 2009 Oct 20;339:b4073–b4073.
3. Ippolito PM, Mathios AD. The regulation of science-based claims in advertising. *Journal of Consumer Policy*. 1990 Dec;13:413–45.
4. FTC Policy Statement on Advertising Substantiation, 49 Fed. Reg. 30,999 (1984), REPRINTED IN THOMPSON MEDICAL CO., 104 F.T.C. 648, 839 (1984), AFF'D, 791 F. 2d 189 (D.C. Cir. 1986), CERT. DENIED, 479 U.S. 1086 (1987) (“Substantiation Statement”). 1984;
5. Cohn J. Science and advertising [editorials]. *Circulation*. 1982;65(5):839.
6. Science in Advertising. *BMJ*. 1938 Oct 22;2(4059):842.
7. Blau J. Half-life of truth in medicine. *The Lancet*. 1998 Jan;351:376.
8. Hall JC, Platell C. Half-life of truth in surgical literature. *The Lancet*. 1997 Dec;350:1752.
9. Poynard T, Munteanu M, Ratziu V, Benhamou Y, Di Martino V, Taieb J, et al. Truth survival in clinical research: an evidence-based requiem? *Ann. Intern. Med.* 2002 Jun 18;136(12):888–95.
10. Polanyi M. *Science, faith and society*. University of Chicago Press; 1966. 96 p.
11. Popper K. *The myth of the framework: in defence of science and rationality*. Reprint. London [u.a.]: Routledge; 1997.
12. Competition Act [Internet]. [cited 2011 Nov 7];Available from: <http://laws-lois.justice.gc.ca/eng/acts/C-34/>
13. Competition Bureau - Misleading Advertising Guidelines [Internet]. [cited 2011 Oct 11];Available from: <http://www.competitionbureau.gc.ca/eic/site/cb-bc.nsf/eng/01222.html#a>
14. Jensen MC, Meckling WH. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of financial economics*. 1976;3(4):305–60.


- 
15. Eddy DM. Variations in physician practice: the role of uncertainty. *Health Aff (Millwood)*. 1984;3(2):74–89.
 16. Mott DA, Schommer JC, Doucette WR, Kreling DH. Agency theory, drug formularies, and drug product selection: implications for public policy. *Journal of Public Policy & Marketing*. 1998;:287–95.
 17. Eisenhardt KM. Agency theory: An assessment and review. *Academy of management review*. 1989;:57–74.
 18. Callahan D. The WHO definition of 'health'. *Hastings Center Studies*. 1973;:77–87.
 19. Saracci R. The World Health Organisation needs to reconsider its definition of health. *Bmj*. 1997;314(7091):1409.
 20. Bircher J. Towards a dynamic definition of health and disease. *Medicine, Health Care and Philosophy*. 2005;8(3):335–41.
 21. Jadad AR, O'Grady L. How should health be defined? *BMJ*. 2008;337.
 22. World Health Organization. WHO | The world health report 2000 - Health systems: improving performance [Internet]. [cited 2011 Jun 26]; Available from: <http://www.who.int/whr/2000/en/>
 23. PAAB. Pharmaceutical Advertising Advisory Board Code of Advertising Acceptance.
 24. Hausman DM. Valuing health: a new proposal. *Health Econ*. 2010 Mar;19(3):280–96.
 25. Maynard A. Logic in medicine: an economic perspective. *British medical journal (Clinical research ed.)*. 1987;295(6612):1537.
 26. Buxton MJ. How much are health-care systems prepared to pay to produce a QALY? *Eur J Health Econ*. 2005 Dec;6(4):285–7.
 27. Epstein DM, Chalabi Z, Claxton K, Sculpher M. Efficiency, Equity, and Budgetary Policies: Informing Decisions Using Mathematical Programming. *Medical Decision Making*. 2007 Mar;27(2):128–37.
 28. McKenna C, Chalabi Z, Epstein D, Claxton K. Budgetary policies and available actions: a generalisation of decision rules for allocation and research decisions. *J Health Econ*. 2010 Jan;29(1):170–81.
 29. The Canadian Code of Advertising Standards [Internet]. [cited 2011 Nov 7]; Available from: <http://www.adstandards.com/en/Standards/canCodeOfAdStandards.aspx>


- 
30. FTC. See *Cliffdale Assocs., Inc.* 103 F.T.C. 110, 175 (1984), reprinted as appendix letter dated Oct. 14, 1983, from the Commission to The Honorable John D. Dingell, Chairman, Committee on Energy and Commerce, U. S. House of Representatives (“Deception Statement”). 1984;
 31. statistics, n. [Internet]. OED Online. [cited 2011 Oct 29]; Available from: <http://www.oed.com/view/Entry/189322?redirectedFrom=statistics>
 32. Matthews JR. *Quantification and the quest for medical certainty.* Princeton University Press; 1995. 195 p.
 33. Cobo E, Selva-O’Callaghan A, Ribera J-M, Cardellach F, Dominguez R, Vilardell M. Statistical Reviewers Improve Reporting in Biomedical Articles: A Randomized Trial. *PLoS ONE.* 2007 Mar 28;2:e332.
 34. Fisher SRA. *Statistical methods for research workers.* Oliver and Boyd; 1970. 378 p.
 35. Neyman J, Pearson ES. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character.* 1933 Jan 1;231:289–337.
 36. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann. Intern. Med.* 1999 Jun 15;130(12):995–1004.
 37. Altman DG, Bland JM. Standard deviations and standard errors. *BMJ.* 2005 Oct 15;331(7521):903.
 38. Feinstein AR. P-Values and Confidence Intervals: Two Sides of the Same Unsatisfactory Coin. *Journal of Clinical Epidemiology.* 1998 Apr;51(4):355–60.
 39. Lang JM, Rothman KJ, Cann CI. That confounded P-value. *Epidemiology.* 1998 Jan;9(1):7–8.
 40. Feinstein AR. The problem of cogent subgroups: a clinicostatistical tragedy. *J Clin Epidemiol.* 1998 Apr;51(4):297–9.
 41. Health C for D and R. *Guidance Documents (Medical Devices and Radiation-Emitting Products) - Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials* [Internet]. [cited 2011 Oct 30]; Available from: <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm>
 42. Altman DG. Statistics in medical journals: some recent trends. *Stat Med.* 2000 Dec 15;19(23):3275–89.


- 
43. Berle D, Starcevic V. Inconsistencies between reported test statistics and p-values in two psychiatry journals. *Int J Methods Psychiatr Res.* 2007;16(4):202–7.
 44. Faulkner C, Fidler F, Cumming G. The value of RCT evidence depends on the quality of statistical analysis. *Behav Res Ther.* 2008 Feb;46(2):270–81.
 45. García-Berthou E, Alcaraz C. Incongruence between test statistics and P values in medical papers. *BMC Med Res Methodol.* 2004 May 28;4:13.
 46. thompson5 [Internet]. [cited 2011 Oct 30];Available from: <http://warnercnr.colostate.edu/~anderson/thompson1.html>
 47. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ.* 2010 Mar 23;340:c869–c869.
 48. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed).* 1986 Mar 15;292(6522):746–50.
 49. Knol MJ, Pestman WR, Grobbee DE. The (mis)use of overlap of confidence intervals to assess effect modification. *Eur J Epidemiol.* 2011 Apr;26(4):253–4.
 50. Borm GF, Lemmers O, Fransen J, Donders R. The evidence provided by a single trial is less reliable than its statistical analysis suggests. *J Clin Epidemiol.* 2009 Jul;62(7):711–5.e1.
 51. Kulldorff M, Graubard B, Velie E. The P-value and P-value function. *Epidemiology.* 1999 May;10(3):345–7.
 52. Barrett B, Brown D, Mundt M, Brown R. Sufficiently important difference: expanding the framework of clinical significance. *Med Decis Making.* 2005 Jun;25(3):250–61.
 53. Lindgren BR, Wielinski CL, Finkelstein SM, Warwick WJ. Contrasting clinical and statistical significance within the research setting. *Pediatr. Pulmonol.* 1993 Dec;16(6):336–40.
 54. Bezeau S, Graves R. Statistical power and effect sizes of clinical neuropsychology research. *J Clin Exp Neuropsychol.* 2001 Jun;23(3):399–406.
 55. Chan A-W, Hróbjartsson A, Jørgensen KJ, Gøtzsche PC, Altman DG. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. *BMJ.* 2008;337:a2299.


- 
56. Moyé LA. P-value interpretation and alpha allocation in clinical trials. *Ann Epidemiol*. 1998 Aug;8(6):351–7.
 57. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA*. 1994 Jul 13;272(2):122–4.
 58. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005 Aug;2(8):e124.
 59. Freeman PR. The role of p-values in analysing trial results. *Stat Med*. 1993 Aug;12(15-16):1443–52; discussion 1453–8.
 60. Ebramzadeh E, McKellop H, Dorey F, Sarmiento A. Challenging the validity of conclusions based on P-values alone: a critique of contemporary clinical research design and methods. *Instr Course Lect*. 1994;43:587–600.
 61. Compton S. Do not discourage the use of P values. *Ann Emerg Med*. 2003 Apr;41(4):584.
 62. Rothwell P. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*. 2005 Jan 8;365(9454):176–86.
 63. Gillen DL, Emerson SS. A note on P-values under group sequential testing and nonproportional hazards. *Biometrics*. 2005 Jun;61(2):546–51.
 64. Cleophas TJ. Clinical trials and p-values, beware of the extremes. *Clin. Chem. Lab. Med*. 2004 Mar;42(3):300–4.
 65. Berry DA. A case for Bayesianism in clinical trials. *Stat Med*. 1993 Aug;12(15-16):1377–93; discussion 1395–404.
 66. O’Hagan A, Stevens JW. Bayesian methods for design and analysis of cost-effectiveness trials in the evaluation of health care technologies. *Stat Methods Med Res*. 2002 Dec;11(6):469–90.
 67. Burton PR. Helping doctors to draw appropriate inferences from the analysis of medical studies. *Stat Med*. 1994 Sep 15;13(17):1699–713.
 68. Bauer P, Brannath W. The advantages and disadvantages of adaptive designs for clinical trials. *Drug Discov. Today*. 2004 Apr 15;9(8):351–7.
 69. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of internal medicine*. 1999;130(12):1005.
 70. Bloom BS, de Pouvourville N, Libert S. Classic or Bayesian research design and analysis. Does it make a difference? *Int J Technol Assess Health Care*. 2002;18(1):120–6.


- 
71. Wijeyesundera DN, Austin PC, Hux JE, Beattie WS, Laupacis A. Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials. *J Clin Epidemiol*. 2009 Jan;62(1):13–21.e5.
 72. Krleža-Jerić K, Chan A-W, Dickersin K, Sim I, Grimshaw J, Gluud C. Principles for international registration of protocol information and results from human trials of health related interventions: Ottawa statement (part 1). *BMJ*. 2005 Apr 23;330(7497):956–8.
 73. [cited 2011 Oct 31];Available from: http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=110_cong_public_laws&docid=f:publ085.110
 74. Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible Research: Moving toward Research the Public Can Really Trust. *Annals of Internal Medicine*. 2007 Mar 20;146(6):450–3.
 75. Groves T. The wider concept of data sharing: view from the BMJ. *Biostatistics*. 2010 Jun 10;11:391–2.
 76. Peng RD, Dominici F, Zeger SL. Reproducible Epidemiologic Research. *American Journal of Epidemiology*. 2006 May 1;163(9):783–9.
 77. Fontanarosa PB, Flanagin A, DeAngelis CD. Reporting Conflicts of Interest, Financial Aspects of Research, and Role of Sponsors in Funded Studies. *JAMA: The Journal of the American Medical Association*. 2005 Jul 6;294(1):110–1.
 78. McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. *JAMA*. 2003 May 21;289(19):2545–53.
 79. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol*. 2006;6:54.
 80. Hadorn DC. Natural Kinds, Evidence, and Randomness in Health Outcomes Research. Thesis, University of Wellington. 2001 Apr 15;
 81. Mulrow CD. The medical review article: state of the science. *Ann. Intern. Med*. 1987 Mar;106(3):485–8.
 82. Teagarden JR. Meta-analysis: whither narrative review? *Pharmacotherapy*. 1989;9(5):274–81; discussion 281–4.
 83. Glass GV. Primary, secondary, and meta-analysis of research. *Educational researcher*. 1976;5(10):3–8.
 84. Bastian H, Glasziou P, Chalmers I. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? *PLoS Med*. 2010;7(9):e1000326.


- 
85. Gerbarg ZB, Horwitz RI. Resolving conflicting clinical trials: guidelines for meta-analysis. *J Clin Epidemiol*. 1988;41(5):503–9.
 86. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*. 2005 Feb;8:19–32.
 87. Light RJ, Pillemer DB. *Summing up: the science of reviewing research*. Harvard University Press; 1984. 212 p.
 88. Yusuf S, Simon R, Ellenberg S. *Proceedings of the Workshop on methodologic issues in overviews of randomized clinical trials, held in Bethesda on 15-16 May, 1986*. John Wiley and Sons; 1987. 193 p.
 89. Egger M, Smith GD, Sterne JA. Uses and abuses of meta-analysis. *Clin Med*. 2001 Dec;1(6):478–84.
 90. Spector TD, Thompson SG. The potential and limitations of meta-analysis. *J Epidemiol Community Health*. 1991 Jun;45(2):89–92.
 91. Fergusson D, Glass KC, Hutton B, Shapiro S. Randomized controlled trials of aprotinin in cardiac surgery: could clinical equipoise have stopped the bleeding? *Clin Trials*. 2005;2(3):218–29; discussion 229–32.
 92. Sutton AJ, Cooper NJ, Jones DR. Evidence synthesis as the key to more coherent and efficient research. *BMC Med Res Methodol*. 2009;9:29.
 93. Williams JW Jr, Ranney L, Morgan LC, Whitener L. How reviews covered the unfolding scientific story of gabapentin for bipolar disorder. *Gen Hosp Psychiatry*. 2009 Jun;31(3):279–87.
 94. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. *Treatments for myocardial infarction*. *JAMA*. 1992 Jul 8;268(2):240–8.
 95. Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N. Engl. J. Med*. 1992 Jul 23;327(4):248–54.
 96. Bailar JC, Hoaglin DC. *Medical uses of statistics*. John Wiley & Sons; 2009. 541 p.
 97. Shrier I, Boivin J-F, Platt RW, Steele RJ, Brophy JM, Carnevale F, et al. The interpretation of systematic reviews with meta-analyses: an objective or subjective process? *BMC Med Inform Decis Mak*. 2008;8:19.
 98. Jadad AR, Cook DJ, Browman GP. A guide to interpreting discordant systematic reviews. *CMAJ*. 1997 May 15;156(10):1411–6.

- 
99. Report on Certain Enteric Fever Inoculation Statistics. *Br Med J.* 1904 Nov 5;2(2288):1243–6.
 100. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *N. Engl. J. Med.* 1987 Feb 19;316(8):450–5.
 101. Gerber S, Tallon D, Trelle S, Schneider M, Jüni P, Egger M. Bibliographic study showed improving methodology of meta-analyses published in leading journals 1993-2002. *J Clin Epidemiol.* 2007 Aug;60(8):773–80.
 102. Thacker SB. Meta-analysis. A quantitative approach to research integration. *JAMA.* 1988 Mar 18;259(11):1685–9.
 103. Naylor CD. Two cheers for meta-analysis: problems and opportunities in aggregating results of clinical trials. *CMAJ.* 1988 May 15;138(10):891–5.
 104. Jenicek M. Meta-analysis in medicine. Where we are and where we want to go. *J Clin Epidemiol.* 1989;42(1):35–44.
 105. Chalmers TC. Problems induced by meta-analyses. *Stat Med.* 1991 Jun;10(6):971–9; discussion 979–80.
 106. Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. *Am. J. Epidemiol.* 1994 Aug 1;140(3):290–6.
 107. Jones DR. Meta-analysis: weighing the evidence. *Stat Med.* 1995 Jan 30;14(2):137–49.
 108. Egger M, Smith GD. Misleading meta-analysis. *BMJ.* 1995 Mar 25;310(6982):752–4.
 109. Yuan Y, Hunt RH. Systematic reviews: the good, the bad, and the ugly. *Am. J. Gastroenterol.* 2009 May;104(5):1086–92.
 110. Vavken P, Dorotka R. A systematic review of conflicting meta-analyses in orthopaedic surgery. *Clin. Orthop. Relat. Res.* 2009 Oct;467(10):2723–35.
 111. Naylor CD. The case for failed meta-analyses. *J Eval Clin Pract.* 1995 Nov;1(2):127–30.
 112. Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. *American Journal of Epidemiology.* 1994;140(3):290.
 113. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan A-W, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE.* 2008;3(8):e3081.

- 
114. Pereira TV, Ioannidis JPA. Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. *J Clin Epidemiol* [Internet]. 2011 Mar 29 [cited 2011 Jun 8]; Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21454050>
 115. Altman DG. Commentary: Systematic reviewers face challenges from varied study designs. *BMJ*. 2002 Aug 31;325(7362):461.
 116. Atkins D. Creating and synthesizing evidence with decision makers in mind: integrating evidence from clinical trials and other study designs. *Med Care*. 2007 Oct;45(10 Supl 2):S16–22.
 117. Lopez-Lee D. Indiscriminate data aggregations in meta-analysis. A cause for concern among policy makers and social scientists. *Eval Rev*. 2002 Oct;26(5):520–44.
 118. Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. *BMJ*. 1998 Jan 10;316(7125):140–4.
 119. Piedbois P, Buyse M. Meta-analyses based on abstracted data: a step in the right direction, but only a first step. *J. Clin. Oncol*. 2004 Oct 1;22(19):3839–41.
 120. Robertson C, Idris NRN, Boyle P. Beyond classical meta-analysis: can inadequately reported studies be included? *Drug Discov. Today*. 2004 Nov 1;9(21):924–31.
 121. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann. Intern. Med*. 2001 Dec 4;135(11):982–9.
 122. Lindbaek M. Seeing what you want to see in randomised controlled trials. Meta-analyses may suffer from interpretation bias too. *BMJ*. 2000 Oct 28;321(7268):1079.
 123. Koretz RL. Methods of meta-analysis: an analysis. *Curr Opin Clin Nutr Metab Care*. 2002 Sep;5(5):467–74.
 124. Cummings P. Meta-analysis based on standardized effects is unreliable. *Arch Pediatr Adolesc Med*. 2004 Jun;158(6):595–7.
 125. Ades AE, Lu G, Higgins JPT. The interpretation of random-effects meta-analysis in decision models. *Med Decis Making*. 2005 Dec;25(6):646–54.
 126. Schroll JB, Moustgaard R, Gøtzsche PC. Dealing with substantial heterogeneity in Cochrane reviews. Cross-sectional study. *BMC Med Res Methodol*. 2011;11:22.


- 
127. Gøtzsche PC, Johansen HK. Misleading statements in industry-sponsored meta-analysis of itraconazole. *J. Clin. Oncol.* 2005 Dec 20;23(36):9428–9; author reply 9429–32.
 128. Jørgensen AW, Maric KL, Tendal B, Faurschou A, Gøtzsche PC. Industry-supported meta-analyses compared with meta-analyses with non-profit or no support: differences in methodological quality and conclusions. *BMC Med Res Methodol.* 2008;8:60.
 129. Golder S, Loke Y, McIntosh HM. Room for improvement? A survey of the methods used in systematic reviews of adverse effects. *BMC Med Res Methodol.* 2006;6:3.
 130. Connis RT, Evans RL, Hendricks RD. Meta-analysis with longitudinal studies: controlling for analytical bias. *Psychol Rep.* 1996 Dec;79(3 Pt 2):1383–6.
 131. Sloan JA. Quality-of-life meta-analysis: why treat it differently? *J. Clin. Oncol.* 2003 Apr 1;21(7):1422.
 132. McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *Lancet.* 2000 Oct 7;356(9237):1228–31.
 133. Villar J, Carroli G, Belizán JM. Predictive ability of meta-analyses of randomised controlled trials. *Lancet.* 1995 Mar 25;345(8952):772–6.
 134. LeLorier J, Grégoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *N. Engl. J. Med.* 1997 Aug 21;337(8):536–42.
 135. Cappelleri JC, Ioannidis JP, Schmid CH, de Ferranti SD, Aubert M, Chalmers TC, et al. Large trials vs meta-analysis of smaller trials: how do their results compare? *JAMA.* 1996 Oct 23;276(16):1332–8.
 136. Brazzi L, Bertolini G, Minelli C. Meta-analyses versus randomised controlled trials in intensive care medicine. *Intensive Care Med.* 2000 Feb;26(2):239–41.
 137. Hennekens CH, DeMets D. The Need for Large-Scale Randomized Evidence Without Undue Emphasis on Small Trials, Meta-analyses, or Subgroup Analyses. *JAMA: The Journal of the American Medical Association.* 2009 Dec 2;302(21):2361–2.
 138. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet.* 1998 Jan 10;351(9096):123–7.
 139. Ioannidis JP, Cappelleri JC, Lau J. Issues in comparisons between meta-analyses and large trials. *JAMA.* 1998 Apr 8;279(14):1089–93.


- 
140. Furukawa TA, Streiner DL, Hori S. Discrepancies among megatrials. *J Clin Epidemiol*. 2000 Dec;53(12):1193–9.
 141. Shun Z, Chi E, Durrleman S, Fisher L. Statistical consideration of the strategy for demonstrating clinical evidence of effectiveness—one larger vs two smaller pivotal studies. *Stat Med*. 2005 Jun 15;24(11):1619–37; discussion 1639–56.
 142. Boissel JP, Blanchard J, Panak E, Peyrieux JC, Sacks H. Considerations for the meta-analysis of randomized clinical trials. Summary of a panel discussion. *Control Clin Trials*. 1989 Sep;10(3):254–81.
 143. Laird NM, Mosteller F. Some statistical methods for combining experimental results. *Int J Technol Assess Health Care*. 1990;6(1):5–30.
 144. Greenland S, Salvan A. Bias in the one-step method for pooling study results. *Stat Med*. 1990 Mar;9(3):247–52.
 145. Pogue J, Yusuf S. Overcoming the limitations of current meta-analysis of randomised controlled trials. *Lancet*. 1998 Jan 3;351(9095):47–52.
 146. Davey Smith G, Egger M. Meta-analyses of randomised controlled trials. *Lancet*. 1997 Oct 18;350(9085):1182.
 147. Jüni P, Witschi A, Bloch R, Egger M. The Hazards of Scoring the Quality of Clinical Trials for Meta-analysis. *JAMA: The Journal of the American Medical Association*. 1999;282(11):1054–60.
 148. Salanti G, Ioannidis JPA. Synthesis of observational studies should consider credibility ceilings. *J Clin Epidemiol*. 2009 Feb;62(2):115–22.
 149. Eddy DM, Hasselblad V, Shachter R. An introduction to a Bayesian method for meta-analysis: The confidence profile method. *Med Decis Making*. 1990 Mar;10(1):15–23.
 150. Su XY, Li Wan Po A. Combining event rates from clinical trials: comparison of Bayesian and classical methods. *Ann Pharmacother*. 1996 May;30(5):460–5.
 151. Li J, Mehrotra DV. An efficient method for accommodating potentially underpowered primary endpoints. *Stat Med*. 2008 Nov 20;27(26):5377–91.
 152. Rücker G, Carpenter JR, Schwarzer G. Detecting and adjusting for small-study effects in meta-analysis. *Biom J*. 2011 Mar;53(2):351–68.
 153. van Amelsvoort LGPM, Viechtbauer W, Spigt MG. Spuriously precise results from meta-analysis. Is better statistical correction or a more





critical methodological assessment warranted? *J Clin Epidemiol*. 2009 Feb;62(2):123–5; discussion 126–7.


154. Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. *J Clin Epidemiol*. 2008 Aug;61(8):763–9.
155. van der Tweel I, Bollen C. Sequential meta-analysis: an efficient decision-making tool. *Clin Trials*. 2010 Apr;7(2):136–46.
156. Janket S-J, Moles DR, Lau J, Needleman I, Niederman R. Caveat for a cumulative meta-analysis. *J. Dent. Res*. 2005 Jun;84(6):487; author reply 487.
157. Clarke MJ, Stewart LA. Systematic reviews of randomized controlled trials: the need for complete data. *J Eval Clin Pract*. 1995 Nov;1(2):119–26.
158. Carl van W. Individual patient meta-analysis—rewards and challenges. *Journal of Clinical Epidemiology*. 2010 Mar;63(3):235–7.
159. Raina PS, Brehaut JC, Platt RW, Klassen TP, Moher D, St John P, et al. The influence of display and statistical factors on the interpretation of metaanalysis results by physicians. *Med Care*. 2005 Dec;43(12):1242–9.
160. Bax L, Ikeda N, Fukui N, Yaju Y, Tsuruta H, Moons KGM. More than numbers: the power of graphs in meta-analysis. *Am. J. Epidemiol*. 2009 Jan 15;169(2):249–55.
161. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology*. 2011 Apr;64(4):401–6.
162. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology*. 2011 Apr;64(4):383–94.
163. Godlee F, Jefferson T. *Peer review in health sciences*. BMJ Books; 2003. 367 p.
164. Laine C, Mulrow C, for the Editors. *Peer Review: Integral to Science and Indispensable to Annals*. *Annals of Internal Medicine*. 2003 Dec 16;139(12):1038–40.
165. Morgan WK. On evidence, embellishment and efficacy. *J Eval Clin Pract*. 1997 Apr;3(2):117–22.
166. Smith R. Peer review: a flawed process at the heart of science and journals. *JRSM*. 2006 Apr 1;99(4):178–82.


- 
167. Rennie D. Integrity in Scientific Publishing. *Health Services Research*. 2010 Mar 10;45:885–96.
 168. Rosenthal R. The file drawer problem and tolerance for null results. *Psychological bulletin*. 1979;86(3):638.
 169. Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al. Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess*. 2010 Feb;14(8):iii, ix-xi, 1–193.
 170. Jefferson T, Alderson P, Wager E, Davidoff F. Effects of Editorial Peer Review. *JAMA: The Journal of the American Medical Association*. 2002 Jun 5;287(21):2784–6.
 171. Jefferson T, Rudin M, Brodney Folse S, Davidoff F. Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database Syst Rev*. 2007;2.
 172. Rune E. Are road safety evaluation studies published in peer reviewed journals more valid than similar studies not published in peer reviewed journals? *Accident Analysis & Prevention*. 1998 Jan;30(1):101–18.
 173. Goodman SN, Berlin J, Fletcher SW, Fletcher RH. Manuscript Quality before and after Peer Review and Editing at *Annals of Internal Medicine*. *Annals of Internal Medicine*. 1994 Jul 1;121(1):11–21.
 174. Pierie J-PE, Walvoort HC, Overbeke AJP. Readers' evaluation of effect of peer review and editing on quality of articles in the *Nederlands Tijdschrift voor Geneeskunde*. *The Lancet*. 1996 Nov 30;348(9040):1480–3.
 175. Godlee F, Gale CR, Martyn CN. Effect on the Quality of Peer Review of Blinding Reviewers and Asking Them to Sign Their Reports. *JAMA: The Journal of the American Medical Association*. 1998 Jul 15;280(3):237–40.
 176. Schroter S. Response to Scientific journals are “faith based”: is there a science behind peer review? *Journal of the Royal Society of Medicine*. 2007 Mar 1;100:117–8.
 177. Schroter S, Black N, Evans S, Carpenter J, Godlee F, Smith R. Effects of training on quality of peer review: randomised controlled trial. *BMJ*. 2004 Mar 20;328(7441):673.
 178. House of Commons - Peer review in scientific publications - Science and Technology Committee [Internet]. [cited 2011 Dec 2]; Available from: <http://www.publications.parliament.uk/pa/cm201012/cmselect/cmsctech/856/85602.htm>
 179. Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith H Jr. Publication bias and clinical trials. *Control Clin Trials*. 1987 Dec;8(4):343–53.


- 
180. Dickersin K, Min YI. NIH clinical trials and publication bias. *Online J Curr Clin Trials*. 1993 Apr 28;Doc No 50:[4967 words; 53 paragraphs].
 181. von Elm E, Röllin A, Blümle A, Huwiler K, Witschi M, Egger M. Publication and non-publication of clinical trials: longitudinal study of applications submitted to a research ethics committee. *Swiss Med Wkly*. 2008 Apr 5;138(13-14):197–203.
 182. Dickersin K, Min YI. Publication bias: the problem that won't go away. *Ann. N. Y. Acad. Sci*. 1993 Dec 31;703:135–46; discussion 146–8.
 183. Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. *Cochrane Database Syst Rev*. 2007;(2):MR000010.
 184. Sterne JAC, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in “meta-epidemiological” research. *Stat Med*. 2002 Jun 15;21(11):1513–24.
 185. van Driel ML, De Sutter A, De Maeseneer J, Christiaens T. Searching for unpublished trials in Cochrane reviews may not be worth the effort. *J Clin Epidemiol*. 2009 Aug;62(8):838–44.e3.
 186. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan A-W, Cronin E, et al. Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias. *PLoS ONE*. 2008;3(8):e3081.
 187. Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ*. 2010;340:c365.
 188. Chan A-W, Krleža-Jerić K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *Canadian Medical Association Journal*. 2004;171(7):735–40.
 189. Golder S, Loke YK, Bland M. Unpublished data can be of value in systematic reviews of adverse effects: methodological overview. *J Clin Epidemiol*. 2010 Oct;63(10):1071–81.
 190. Copas J, Jackson D. A bound for publication bias based on the fraction of unpublished studies. *Biometrics*. 2004 Mar;60(1):146–53.
 191. Formann AK. Estimating the proportion of studies missing for meta-analysis due to publication bias. *Contemp Clin Trials*. 2008 Sep;29(5):732–9.
 192. Riley RD, Sutton AJ, Abrams KR, Lambert PC. Sensitivity analyses allowed more appropriate and reliable meta-analysis conclusions for multiple


- 
- outcomes when missing data was present. *J Clin Epidemiol*. 2004 Sep;57(9):911–24.
193. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics*. 1994 Dec;50(4):1088–101.
 194. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997 Sep 13;315(7109):629–34.
 195. Tang JL, Liu JL. Misleading funnel plot for detection of bias in meta-analysis. *J Clin Epidemiol*. 2000 May;53(5):477–84.
 196. Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Stat Med*. 2001 Feb 28;20(4):641–54.
 197. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol*. 2005 Sep;58(9):882–93.
 198. Schwarzer G, Antes G, Schumacher M. A test for publication bias in meta-analysis with sparse binary data. *Stat Med*. 2007 Feb 20;26(4):721–33.
 199. Harbord RM, Egger M, Sterne JAC. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med*. 2006 Oct 30;25(20):3443–57.
 200. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. *JAMA*. 2006 Feb 8;295(6):676–80.
 201. Rücker G, Schwarzer G, Carpenter J. Arcsine test for publication bias in meta-analyses with binary outcomes. *Stat Med*. 2008 Feb 28;27(5):746–63.
 202. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ*. 2010 Mar 30;340:c117–c117.
 203. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*. 2000 Mar 25;355(9209):1064–9.
 204. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. *N. Engl. J. Med*. 1987 Aug 13;317(7):426–32.
 205. Bhandari M, Devereaux PJ, Li P, Mah D, Lim K, Schünemann HJ, et al. Misuse of baseline comparison tests and subgroup analyses in surgical trials. *Clin. Orthop. Relat. Res*. 2006 Jun;447:247–51.


- 
206. Hernández AV, Boersma E, Murray GD, Habbema JDF, Steyerberg EW. Subgroup analyses in therapeutic cardiovascular clinical trials: Are most of them misleading? *American Heart Journal*. 2006 Feb;151(2):257–64.
 207. Gabler NB, Duan N, Liao D, Elmore JG, Ganiats TG, Kravitz RL. Dealing with heterogeneity of treatment effects: is the literature up to the challenge? *Trials*. 2009;10:43.
 208. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*. 2007;357(21):2189–94.
 209. International conference on harmonisation; guidance on statistical principles for clinical trials; availability--FDA. Notice. *Fed Regist*. 1998 Sep 16;63(179):49583–98.
 210. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA*. 1991 Jul 3;266(1):93–8.
 211. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med*. 1992 Jan 1;116(1):78–84.
 212. Lagakos SW. The challenge of subgroup analyses--reporting without distorting. *N. Engl. J. Med*. 2006 Apr 20;354(16):1667–9.
 213. Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technol Assess*. 2001;5(33):1–56.
 214. Altman DG. Within trial variation--a false trail? *J Clin Epidemiol*. 1998 Apr;51(4):301–3.
 215. Senn S. Individual response to treatment: is it a valid assumption? *BMJ*. 2004 Oct 23;329(7472):966–8.
 216. Senn S. Controversies concerning randomization and additivity in clinical trials. *Stat Med*. 2004 Dec 30;23(24):3729–53.
 217. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, et al. The influence of study characteristics on reporting of subgroup analyses in randomised controlled trials: systematic review. *BMJ*. 2011;342:d1569.
 218. Freemantle N. Interpreting the results of secondary end points and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic? *BMJ*. 2001 Apr 21;322(7292):989–91.


- 
219. Packer M, O'Connor CM, Ghali JK, Pressler ML, Carson PE, Belkin RN, et al. Effect of amlodipine on morbidity and mortality in severe chronic heart failure. Prospective Randomized Amlodipine Survival Evaluation Study Group. *N. Engl. J. Med.* 1996 Oct 10;335(15):1107–14.
 220. Wijeyesundera HC, Hansen MS, Stanton E, Cropp AS, Hall C, Dhalla NS, et al. Neurohormones and oxidative stress in nonischemic cardiomyopathy: relationship to survival and the effect of treatment with amlodipine. *Am. Heart J.* 2003 Aug;146(2):291–7.
 221. Pitt B, Segal R, Martinez FA, Meurers G, Cowley AJ, Thomas I, et al. Randomised trial of losartan versus captopril in patients over 65 with heart failure (Evaluation of Losartan in the Elderly Study, ELITE)*. *The Lancet.* 1997;349(9054):747–52.
 222. Pitt B, Poole-Wilson PA, Segal R, Martinez FA, Dickstein K, Camm AJ, et al. Effect of losartan compared with captopril on mortality in patients with symptomatic heart failure: randomised trial—the Losartan Heart Failure Survival Study ELITE II. *The Lancet.* 2000;355(9215):1582–7.
 223. Hayward RA, Kent DM, Vijan S, Hofer TP. Reporting Clinical Trial Results To Inform Providers, Payers, And Consumers. *Health Affairs.* 2005 Nov 1;24(6):1571–81.
 224. Bauer P. Multiple testing in clinical trials. *Stat Med.* 1991 Jun;10(6):871–89; discussion 889–90.
 225. Gordi T, Khamis H. Simple solution to a common statistical problem: interpreting multiple tests. *Clin Ther.* 2004 May;26(5):780–6.
 226. Austin PC, Mamdani MM, Juurlink DN, Hux JE. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *Journal of clinical epidemiology.* 2006;59(9):964–9.
 227. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. *Lancet.* 1988 Aug 13;2(8607):349–60.
 228. Sleight P. Debate: Subgroup analyses in clinical trials: fun to look at - but don't believe them! *Curr Control Trials Cardiovasc Med.* 2000;1(1):25–7.
 229. Zelen M. A new design for randomized clinical trials. *N. Engl. J. Med.* 1979 May 31;300(22):1242–5.
 230. Brown CH, Ten Have TR, Jo B, Dagne G, Wyman PA, Muthén B, et al. Adaptive designs for randomized trials in public health. *Annu Rev Public Health.* 2009 Apr 29;30:1–25.


- 
231. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology*. 2011 Apr;64(4):401–6.
 232. Riecken HW & B, Robert F. [Eds]. *Social Experimentation: A Method for Planning and Evaluating Social Intervention*. Academic Press; 1974.
 233. Heckman JJ. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*. 1979;:153–61.
 234. Karakiewicz PI, Briganti A, Chun FK-H, Valiquette L. Outcomes research: a methodologic review. *Eur. Urol*. 2006 Aug;50(2):218–24.
 235. Hemming K, Hutton JL, Maguire MJ, Marson AG. Open label extension studies and patient selection biases. *J Eval Clin Pract*. 2008 Feb;14(1):141–4.
 236. Fraker T, Maynard R. The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources*. 1987;:194–227.
 237. Gray-Donald K, Kramer MS. Causality inference in observational vs. experimental studies. *American journal of epidemiology*. 1988;127(5):885.
 238. 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 Feb 1;273(5):408–12.
 239. Bajard A, Chabaud S, Pérol D, Boissel J-P, Nony P. Revisiting the level of evidence in randomized controlled clinical trials: A simulation approach. *Contemp Clin Trials*. 2009 Sep;30(5):400–10.
 240. Jane-wit D, Horwitz RI, Concato J. Variation in results from randomized, controlled trials: stochastic or systematic? *J Clin Epidemiol*. 2010 Jan;63(1):56–63.
 241. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N. Engl. J. Med*. 2000 Jun 22;342(25):1887–92.
 242. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N. Engl. J. Med*. 2000 Jun 22;342(25):1878–86.
 243. Deeks JJ, Dinnes J, D’Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. *Health Technol Assess*. 2003;7(27):iii-x, 1–173.


- 
244. Kunz R, Khan KS, Neumayer HH. Observational studies and randomized trials. *N. Engl. J. Med.* 2000 Oct 19;343(16):1194–5; author reply 1196–7.
 245. Concato J. Observational versus experimental studies: what’s the evidence for a hierarchy? *NeuroRx.* 2004 Jul;1(3):341–7.
 246. Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials. *Cochrane Database Syst Rev.* 2007;(2):MR000012.
 247. Oliver S, Bagnall AM, Thomas J, Shepherd J, Sowden A, White I, et al. Randomised controlled trials for policy interventions: a review of reviews and meta-regression. *Health Technol Assess.* 2010 Mar;14(16):1–165, iii.
 248. Odgaard-Jensen J, Vist GE, Timmer A, Kunz R, Akl EA, Schünemann H, et al. Randomisation to protect against selection bias in healthcare trials. *Cochrane Database Syst Rev.* 2011;(4):MR000012.
 249. Kunz R. Randomized trials and observational studies: still mostly similar results, still crucial differences. *J Clin Epidemiol.* 2008 Mar;61(3):207–8.
 250. Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ.* 2008 Mar 15;336(7644):601–5.
 251. Dreyer NA, Tunis SR, Berger M, Ollendorf D, Mattox P, Gliklich R. Why Observational Studies Should Be Among The Tools Used In Comparative Effectiveness Research. *Health Affairs.* 2010 Oct 1;29(10):1818–25.
 252. Bukstein DA, Luskin AT, Bernstein A. “Real-world” effectiveness of daily controller medicine in children with mild persistent asthma. *Ann. Allergy Asthma Immunol.* 2003 May;90(5):543–9.
 253. Macie C, Wooldrage K, Manfreda J, Anthonisen NR. Introduction of leukotriene receptor antagonists in Manitoba. *Can. Respir. J.* 2006 Mar;13(2):94–8.
 254. Berger M, Mamdani M, Atkins D, Johnson M. Good research practices for comparative effectiveness research: defining, reporting and interpreting non-randomized studies of treatment effects using secondary data sources. *ISPOR TF Report 2009—Part I. Value Health.* 2009;12:1044–52.
 255. Garrison Jr LP, Neumann PJ, Erickson P, Marshall D, Mullins CD. Using Real-World Data for Coverage and Payment Decisions: The ISPOR Real-World Data Task Force Report. *Value in Health.* 2007;10(5):326–35.

- 
256. Edwards AG, Russell IT, Stott NC. Signal versus noise in the evidence base for medicine: an alternative to hierarchies of evidence? *Fam Pract*. 1998 Aug;15(4):319–22.
 257. Borgerson K. Valuing evidence: bias and the evidence hierarchy of evidence-based medicine. *Perspect. Biol. Med*. 2009;52(2):218–33.
 258. Gugiu PC, Gugiu MR. A critical appraisal of standard guidelines for grading levels of evidence. *Eval Health Prof*. 2010 Sep;33(3):233–55.
 259. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. *Journal of Clinical Epidemiology*. 2011 Dec;64:1311–6.
 260. Ludwig DA. Use and misuse of p-values in designed and observational studies: guide for researchers and reviewers. *Aviat Space Environ Med*. 2005 Jul;76(7):675–80.
 261. Lecky FE, Driscoll PA. The clinical relevance of observational research. *J Accid Emerg Med*. 1998 May;15(3):142–6.
 262. Black N. What observational studies can offer decision makers. *Horm. Res*. 1999;51 Suppl 1:44–9.
 263. Ligthelm RJ, Borzi V, Gumprecht J, Kawamori R, Wenying Y, Valensi P. Importance of observational studies in clinical practice. *Clin Ther*. 2007 Jun;29(6 Pt 1):1284–92.
 264. Hoppe DJ, Schemitsch EH, Morshed S, Tornetta P 3rd, Bhandari M. Hierarchy of evidence: where observational studies fit in and why we need them. *J Bone Joint Surg Am*. 2009 May;91 Suppl 3:2–9.
 265. Foody JM, Mendys PM, Liu LZ, Simpson RJ Jr. The utility of observational studies in clinical decision making: lessons learned from statin trials. *Postgrad Med*. 2010 May;122(3):222–9.
 266. Rawlins M. De testimonio: on the evidence for decisions about the use of therapeutic interventions. *Lancet*. 2008 Dec 20;372(9656):2152–61.
 267. Worrall J. Evidence and ethics in medicine. *Perspect. Biol. Med*. 2008;51(3):418–31.
 268. Hayward RA, Hofer TP, Vijan S. Narrative review: lack of evidence for recommended low-density lipoprotein treatment targets: a solvable problem. *Annals of internal medicine*. 2006;145(7):520.
 269. Stahl JE. Modelling methods for pharmacoeconomics and health technology assessment: an overview and guide. *Pharmacoeconomics*. 2008;26(2):131–48.

- 
270. Brennan A, Chick SE, Davies R. A taxonomy of model structures for economic evaluation of health technologies. *Health Economics*. 2006 Dec;15:1295–310.
 271. Box G. *Empirical model-building and response surfaces*. New York: Wiley; 1987.
 272. Sculpher M, Fenwick E, Claxton K. Assessing quality in decision analytic cost-effectiveness models: a suggested framework and example of application. *Pharmacoeconomics*. 2000;17(5):461–77.
 273. Coates JF. The role of formal models in technology assessment. *Technological Forecasting and Social Change*. 1976;9(1-2):139–90.
 274. Neumann PJ, Greenberg D, Olchanski NV, Stone PW, Rosen AB. Growth and quality of the cost-utility literature, 1976-2001. *Value Health*. 2005 Feb;8(1):3–9.
 275. Neumann PJ. The arrival of economic evidence in managed care formulary decisions: the unsolicited request process. *Med Care*. 2005 Jul;43(7 Suppl):27–32.
 276. NICE. Guide to the methods of technology appraisal [Internet]. [cited 2011 Nov 1]; Available from: [http://www.nice.org.uk/aboutnice/howwework/devnicetech/technologyappraisalprocessguides/guidetothemethodsoftechnologyappraisal.jsp?do media=1&mid=B52851A3-19B9-E0B5-D48284D172BD8459](http://www.nice.org.uk/aboutnice/howwework/devnicetech/technologyappraisalprocessguides/guidetothemethodsoftechnologyappraisal.jsp?do%20media=1&mid=B52851A3-19B9-E0B5-D48284D172BD8459)
 277. Screening for Colorectal Cancer: An Updated Systematic Review - NCBI Bookshelf [Internet]. [cited 2011 Nov 1]; Available from: <http://www.ncbi.nlm.nih.gov.proxy.bib.uottawa.ca/books/NBK35179/>
 278. Greenhalgh J, Bagust A, Boland A, Martin Saborido C, Oyee J, Blundell M, et al. Clopidogrel and modified-release dipyridamole for the prevention of occlusive vascular events (review of Technology Appraisal No. 90): a systematic review and economic analysis. *Health Technol Assess*. 2011 Sep;15(31):1–178.
 279. Sacco RL, Diener H-C, Yusuf S, Cotton D, Ounpuu S, Lawton WA, et al. Aspirin and extended-release dipyridamole versus clopidogrel for recurrent stroke. *N. Engl. J. Med*. 2008 Sep 18;359(12):1238–51.
 280. Halkes PHA, van Gijn J, Kappelle LJ, Koudstaal PJ, Algra A. Aspirin plus dipyridamole versus aspirin alone after cerebral ischaemia of arterial origin (ESPRIT): randomised controlled trial. *Lancet*. 2006 May 20;367(9523):1665–73.


- 
281. Diener HC, Cunha L, Forbes C, Sivenius J, Smets P, Lowenthal A. European Stroke Prevention Study. 2. Dipyridamole and acetylsalicylic acid in the secondary prevention of stroke. *J. Neurol. Sci.* 1996 Nov;143(1-2):1–13.
 282. A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). CAPRIE Steering Committee. *Lancet.* 1996 Nov 16;348(9038):1329–39.
 283. Stern M, Williams K, Eddy D, Kahn R. Validation of Prediction of Diabetes by the Archimedes Model and Comparison With Other Predicting Models. *Diabetes Care.* 2008 Aug;31(8):1670–1.
 284. Fone D, Hollinghurst S, Temple M, Round A, Lester N, Weightman A, et al. Systematic review of the use and value of computer simulation modelling in population health and health care delivery. *J Public Health Med.* 2003 Dec;25(4):325–35.
 285. Kuntz KM, Tsevat J, Weinstein MC, Goldman L. Expert panel vs decision-analysis recommendations for postdischarge coronary angiography after myocardial infarction. *JAMA.* 1999 Dec 15;282(23):2246–51.
 286. Grove WM, Meehl PE. Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law.* 1996;2:293–323.
 287. Dowie J. Evidence based medicine. Needs to be within framework of decision making based on decision analysis. *BMJ.* 1996 Jul 20;313(7050):170–1.
 288. Dowie J. “Evidence-based”, “cost-effective” and “preference-driven” medicine: decision analysis based medical decision making is the prerequisite. *J Health Serv Res Policy.* 1996 Apr;1(2):104–13.
 289. Weinstein MC, O’Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C, et al. Principles of Good Practice for Decision Analytic Modeling in Health-Care Evaluation: Report of the ISPOR Task Force on Good Research Practices—Modeling Studies. *Value in Health.* 2003 Jan;6:9–17.
 290. Daniels N, Sabin JE. Accountability for reasonableness: an update. *BMJ.* 2008;337:a1850.
 291. Culyer A. Deliberative processes in decisions about health care technologies: combining different types of evidence, values, algorithms and people. [Internet]. 2009 Jun; Available from: <http://www.ohe.org/publications/recent-publications/list-by-title-20/detail/date////deliberative-processes-in-decisions-about-health-care-technologies.html>

- 
292. Report of the Indirect Comparisons Working Group to the Pharmaceutical Benefits Advisory Committee: assessing indirect comparisons [Internet]. 2009 Jan; Available from: http://www.pbs.gov.au/industry/useful-resources/PBAC_feedback_files/ICWG%20Report%20FINAL2.pdf
 293. Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, et al. Indirect comparisons of competing interventions. *Health Technol Assess*. 2005 Jul;9(26):1–134, iii-iv.
 294. Song F, Loke YK, Walsh T, Glenny A-M, Eastwood AJ, Altman DG. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. *BMJ*. 2009;338:b1147.
 295. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med*. 2004 Oct 30;23(20):3105–24.
 296. Ades AE. A chain of evidence with mixed comparisons: models for multi-parameter synthesis and consistency of evidence. *Stat Med*. 2003 Oct 15;22(19):2995–3016.
 297. A chain of evidence with mixed comparisons: models ... [Stat Med. 2003] - PubMed - NCBI [Internet]. [cited 2011 Nov 2]; Available from: <http://www.ncbi.nlm.nih.gov.proxy.bib.uottawa.ca/pubmed?term=A%20chain%20of%20evidence%20with%20mixed%20comparisons%3A%20models%20for%20multi-parameter%20>
 298. Eddy DM, Hasselblad V, Shachter RD. *Meta-analysis by the confidence profile method: the statistical synthesis of evidence*. Academic Press; 1992. 444 p.
 299. Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat Med*. 2002 Aug 30;21(16):2313–24.
 300. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol*. 1997 Jun;50(6):683–91.
 301. Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, et al. Interpreting Indirect Treatment Comparisons and Network Meta-Analysis for Health-Care Decision Making: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 1. *Value in Health*. 2011 Jun;14(4):417–28.
 302. Salanti G, Higgins JPT, Ades AE, Ioannidis JPA. Evaluation of networks of randomized trials. *Stat Methods Med Res*. 2008 Jun;17(3):279–301.
 303. Heres S, Davis J, Maino K, Jetzinger E, Kissling W, Leucht S. Why Olanzapine Beats Risperidone, Risperidone Beats Quetiapine, and




Quetiapine Beats Olanzapine: An Exploratory Analysis of Head-to-Head Comparison Studies of Second-Generation Antipsychotics. *American Journal of Psychiatry*. 163(2):185–94.


304. Fu R, Gartlehner G, Grant M, Shamliyan T, Sedrakyan A, Wilt TJ, et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. *J Clin Epidemiol*. 2011 Nov;64(11):1187–97.
305. About Us | Comparing Multiple Interventions Methods Group [Internet]. [cited 2011 Nov 2]; Available from: <http://cmimg.cochrane.org/about-us>
306. Good Research Practices: Interpreting Indirect Treatment Comparison Studies for Decision-making – Part 1 [Internet]. [cited 2011 Nov 2]; Available from: <http://www.ispor.org/taskforces/ITC.asp>
307. Song F. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ*. 2003 Mar 1;326:472–472.
308. Canadian Agency for Drugs and Technologies in Health, Wells G. Indirect evidence indirect treatment comparisons in meta-analysis. Ottawa, ON :: Canadian Agency for Drugs and Technologies in Health=Agence canadienne des médicaments et des technologies de la santé,; 2009.
309. Song F, Xiong T, Parekh-Bhurke S, Loke YK, Sutton AJ, Eastwood AJ, et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. *BMJ*. 2011 Aug 16;343(aug16 2):d4909–d4909.
310. Song F, Harvey I, Lilford R. Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. *Journal of clinical epidemiology*. 2008;61(5):455–63.
311. Kaul S, Diamond GA. Good Enough: A Primer on the Analysis and Interpretation of Noninferiority Trials. *Annals of Internal Medicine*. 2006 Jul 4;145(1):62–9.
312. Lewis JA. Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline. *Statistics in Medicine*. 1999;18(15):1903–42.
313. Brown D, Volkert P, Day S. An introductory note to CHMP guidelines: choice of the non-inferiority margin and data monitoring committees. *Statistics in Medicine*. 2006 May 30;25:1623–7.
314. Guideline ICHHT. CHOICE OF CONTROL GROUP AND RELATED ISSUES IN CLINICAL TRIALS E10. CPMP/ICH/364/96; 2000.


- 
315. Wangge G, Klungel OH, Roes KCB, de Boer A, Hoes AW, Knol MJ. Room for Improvement in Conducting and Reporting Non-Inferiority Randomized Controlled Trials on Drugs: A Systematic Review. *PLoS ONE*. 2010 Oct 27;5(10):e13550.
 316. Gotzsche PC. Lessons from and cautions about noninferiority and equivalence randomized trials. *JAMA: the journal of the American Medical Association*. 2006;295(10):1172.
 317. Lange S, Freitag G. Choice of delta: requirements and reality--results of a systematic review. *Biom J*. 2005 Feb;47(1):12–27; discussion 99–107.
 318. Kaul S, Diamond GA. Good enough: a primer on the analysis and interpretation of noninferiority trials. *Annals of internal medicine*. 2006;145(1):62.
 319. Le Henanff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. *JAMA: the journal of the American Medical Association*. 2006;295(10):1147.
 320. McCaffrey N, Merlin T, Hiller J, Health Technology Assessment International. Meeting (3rd : 2006 : Adelaide SA). LACK OF CLINICAL RATIONALE PROVIDED FOR NON-INFERIORITY MARGINS. [Internet]. 2006 [cited 2011 Nov 6]. p. 28. Available from: <http://gateway.nlm.nih.gov/MeetingAbstracts/ma?f=103724794.html>
 321. Wangge G, Klungel OH, Roes KCB, de Boer A, Hoes AW, Knol MJ. Room for improvement in conducting and reporting non-inferiority randomized controlled trials on drugs: a systematic review. *PLoS ONE*. 2010;5(10):e13550.
 322. Ellenberg SS, Temple R. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 2: practical issues and specific cases. *Annals of internal medicine*. 2000;133(6):464.
 323. D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues-the encounters of academic consultants in statistics. *Statistics in Medicine*. 2002;22(2):169–86.
 324. Hung H, Wang SJ, O'Neill R. A Regulatory Perspective on Choice of Margin and Statistical Inference Issue in Non-inferiority Trials. *Biometrical Journal*. 2005;47(1):28–36.
 325. Kaul S, Diamond GA. Making sense of noninferiority: a clinical and statistical perspective on its application to cardiovascular clinical trials. *Prog Cardiovasc Dis*. 2007 Feb;49(4):284–99.
 326. Assessing Equivalence and Non-Inferiority - Research Report - Draft | AHRQ Effective Health Care Program [Internet]. [cited 2011 Nov


6];Available from:
<http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=802>

327. Drummond MF, Sculpher MJ, Torrance G, O'Brien B, Stoddart G. *Methods for the Economic Evaluation of Health Care Programmes*. Third Edition. Oxford University Press; 2005.
328. Palmer S, Byford S, Raftery J. Types of economic evaluation. *BMJ*. 1999;318(7194):1349.
329. Briggs AH, O'Brien BJ. The death of cost-minimization analysis? *Health Economics*. 2001;10(2):179–84.
330. Husereau D, Morrison A, Battista R, Goeree R. Health Technology Assessment: A Review of International Activity and Examples of Approaches With Computed Tomographic Colonography. *Journal of the American College of Radiology*. 2009 May;6(5):343–52.
331. Pliskin JS, Shepard DS, Weinstein MC. Utility functions for life years and health status. *Operations research*. 1980;28:206–24.
332. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *Journal of Health Economics*. 1999;18(3):341–64.
333. Mihaylova B, Briggs A, O'Hagan A, Thompson SG. Review of statistical methods for analysing healthcare resources and costs. *Health Econ*. 2011 Aug;20(8):897–916.
334. O'Brien B. Economic Evaluation of Pharmaceuticals: Frankenstein's Monster or Vampire of Trials? *Medical Care*. 1996 Dec 1;34(12):DS99–108.
335. Nicholson A, Berger K, Bohn R, Carcao M, Fischer K, Gringeri A, et al. Recommendations for reporting economic evaluations of haemophilia prophylaxis: a nominal groups consensus statement on behalf of the Economics Expert Working Group of The International Prophylaxis Study Group. *Haemophilia*. 2008 Jan;14(1):127–32.
336. Drummond M, Jefferson T. Guidelines for authors and peer reviewers of economic submissions to the *BMJ*. *Bmj*. 1996;313(7052):275.
337. Gold MR. *Cost-effectiveness in health and medicine*. Oxford University Press; 1996. 462 p.
338. Siegel JE, Weinstein MC, Russell LB, Gold MR. Recommendations for reporting cost-effectiveness analyses. Panel on Cost-Effectiveness in Health and Medicine. *JAMA*. 1996 Oct 23;276(16):1339–41.

- 
339. Vintzileos AM, Beazoglou T. Design, execution, interpretation, and reporting of economic evaluation studies in obstetrics. *American journal of obstetrics and gynecology*. 2004;191(4):1070–6.
 340. Drummond M, Manca A, Sculpher M. Increasing the generalizability of economic evaluations: recommendations for the design, analysis, and reporting of studies. *International journal of technology assessment in health care*. 2005;21(2):165–71.
 341. Ramsey S, Willke R, Briggs A, Brown R, Buxton M, Chawla A, et al. Good Research Practices for Cost-Effectiveness Analysis Alongside Clinical Trials: The ISPOR RCT-CEA Task Force Report. *Value in Health*. 2005;8(5):521–33.
 342. Goetghebeur M, Wagner M, Khoury H, Levitt R, Erickson L, Rindress D. Evidence and Value: Impact on DEcisionMaking–the EVIDEM framework and potential applications. *BMC health services research*. 2008;8(1):270.
 343. Petrou S, Gray A. Economic evaluation alongside randomised controlled trials: design, conduct, analysis, and reporting. *BMJ*. 2011;342.
 344. Petrou S, Gray A. Economic evaluation using decision analytical modelling: design, conduct, analysis, and reporting. *BMJ*. 2011 Apr 11;342:d1766–d1766.
 345. Pharmacoeconomic Guidelines Around The World [Internet]. [cited 2011 Nov 6]; Available from: <http://www.ispor.org/peguidelines/index.asp>
 346. Nuijten MJC, Brorens MJA, Hekster YA, Van Der Kuy A, Lockefeer JHM, De Smet PAGM, et al. Reporting format for economic evaluation: part I: application to the Dutch healthcare system. *Pharmacoeconomics*. 1998;14(2):159–63.
 347. Nuijten C, Pronk MH, Brorens MJA, Hekster YA, Lockefeer JHM, De Smet PAGM, et al. Reporting format for economic evaluation: Part II: Focus on modelling studies. *Pharmacoeconomics*. 1998;14(3):259–68.
 348. The AMCP format for formulary submissions version 3.0. *J Manag Care Pharm*. 2010 Jan;16(1 Suppl A):1–30.
 349. Health CA for D and T in. Guidelines for the Economic Evaluation of Health Technologies: Canada. Canadian Agency for Drugs and Technologies in Health; 2006. 46 p.
 350. HEALTH ECONOMIC EVALUATION PUBLICATION GUIDELINES GOOD RESEARCH PRACTICES TASK FORCE [Internet]. [cited 2011 Nov 6]; Available from: <http://www.ispor.org/TaskForces/EconomicPubGuidelines.asp>

- 
351. Drummond MF, Richardson WS, O'Brien BJ, Levine M, Heyland D. Users' Guides to the Medical Literature. *JAMA: The Journal of the American Medical Association*. 1997 May 21;277(19):1552–7.
 352. Gonzalez-Perez JG. Developing a scoring system to quality assess economic evaluations. *Eur J Health Econ*. 2002;3(2):131–6.
 353. Chiou C-F, Hay JW, Wallace JF, Bloom BS, Neumann PJ, Sullivan SD, et al. Development and validation of a grading system for the quality of cost-effectiveness studies. *Med Care*. 2003 Jan;41(1):32–44.
 354. Evers S, Goossens M, de Vet H, van Tulder M, Ament A. Criteria list for assessment of methodological quality of economic evaluations: Consensus on Health Economic Criteria. *Int J Technol Assess Health Care*. 2005;21(2):240–5.
 355. Brunetti M, Ruiz F, Lord J, Pregno S, Oxman AD. Grading Economic Evidence [Internet]. In: Shemilt I, Mugford M, Vale L, Marsh K, Donaldson C, editors. *Evidence-Based Decisions and Economics*. Oxford, UK: Wiley-Blackwell; 2010 [cited 2011 Oct 4]. p. 114–33. Available from: <http://doi.wiley.com/10.1002/9781444320398.ch10>
 356. Ofman JJ, Sullivan SD, Neumann PJ, Chiou C-F, Henning JM, Wade SW, et al. Examining the value and quality of health economic analyses: implications of utilizing the QHES. *J Manag Care Pharm*. 2003 Feb;9(1):53–61.
 357. Tarn T, Smith MD. Pharmacoeconomic guidelines around the world. *ISPOR connections*. 2004;10(4):5–15.
 358. Mauskopf J, Walter J, Birt J, Bowman L, Copley-Merriman C, Drummond M. Differences among formulary submission guidelines: Implications for health technology assessment. *International Journal of Technology Assessment in Health Care*. 2011;27(03):261–70.
 359. Gerkens S, Crott R, Cleemput I, Thissen J-P, Closon M-C, Horsmans Y, et al. Comparison of three instruments assessing the quality of economic evaluations: a practical exercise on economic evaluations of the surgical treatment of obesity. *Int J Technol Assess Health Care*. 2008;24(3):318–25.
 360. Testa MA, Simonson DC. Assessment of quality-of-life outcomes. *New England Journal of Medicine*. 1996;334(13):835–40.
 361. Fitzpatrick R, Fletcher A, Gore S, Jones D, Spiegelhalter D, Cox D. Quality of life measures in health care. I: Applications and issues in assessment. *British Medical Journal*. 1992;305(6861):1074.

- 
362. Food U. Drug Administration: Guidance for Industry. Patient-reported outcome measures: Use in medical product development to support labeling claims. *Health Qual Life Outcomes*. 2006;4:79.
 363. Guyatt GH, Van Zanten SJV, Feeny DH, Patrick DL. Measuring quality of life in clinical trials: a taxonomy and review. *CMAJ: Canadian Medical Association Journal*. 1989;140(12):1441.
 364. Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. *BMJ : British Medical Journal*. 324(7351):1417.
 365. Gill TM, Feinstein AR. A critical appraisal of the quality of quality-of-life measurements. *JAMA*. 1994 Aug 24;272(8):619–26.
 366. Calvert MJ, Freemantle N. Use of health-related quality of life in prescribing research. Part 2: methodological considerations for the assessment of health-related quality of life in clinical trials. *J Clin Pharm Ther*. 2004 Feb;29(1):85–94.
 367. Brazier J, Papaioannou D, Cantrell A, Paisley S, Herrmann K. Identifying and Reviewing Health State Utility Values for Populating Decision Models [Internet]. In: Shemilt I, Mugford M, Vale L, Marsh K, Donaldson C, editors. *Evidence-Based Decisions and Economics*. Oxford, UK: Wiley-Blackwell; 2010 [cited 2011 Nov 5]. p. 93–105. Available from: <http://doi.wiley.com/10.1002/9781444320398.ch8>
 368. Sanders C, Egger M, Donovan J, Tallon D, Frankel S. Reporting on quality of life in randomised controlled trials: bibliographic study. *BMJ*. 1998 Oct 31;317(7167):1191–4.
 369. Patrick DL, Burke LB, Powers JH, Scott JA, Rock EP, Dawisha S, et al. Patient-reported outcomes to support medical product labeling claims: FDA perspective. *Value in Health*. 2007;10:S125–37.
 370. Rothman ML, Beltran P, Cappelleri JC, Lipscomb J, Teschendorf B. Patient-reported outcomes: conceptual issues. *Value Health*. 2007 Dec;10 Suppl 2:S66–75.
 371. Townshend AP, Chen C-M, Williams HC. How prominent are patient-reported outcomes in clinical trials of dermatological treatments? *Br. J. Dermatol*. 2008 Nov;159(5):1152–9.
 372. Groenvold M. Methodological issues in the assessment of health-related quality of life in palliative care trials. *Acta Anaesthesiol Scand*. 1999 Oct;43(9):948–53.

- 
373. Guyatt GH, Naylor CD, Juniper E, Heyland DK, Jaeschke R, Cook DJ. Users' Guides to the Medical Literature. JAMA: The Journal of the American Medical Association. 1997 Apr 16;277(15):1232–7.
 374. Symonds T, Berzon R, Marquis P, Rummans TA. The clinical significance of quality-of-life results: practical considerations for specific audiences. Mayo Clin. Proc. 2002 Jun;77(6):572–83.
 375. Diamond GA, Denton TA. Alternative perspectives on the biased foundations of medical technology assessment. Ann. Intern. Med. 1993 Mar 15;118(6):455–64.
 376. Guyatt G, Schunemann H. How can quality of life researchers make their work more useful to health workers and their patients? Qual Life Res. 2007 Sep;16(7):1097–105.
 377. Fletcher A. Quality-of-life measurements in the evaluation of treatment: proposed guidelines. British journal of clinical pharmacology. 1995;39(3):217.
 378. Staquet M, Berzon R, Osoba D, Machin D. Guidelines for reporting results of quality of life assessments in clinical trials. Qual Life Res. 1996 Oct;5(5):496–502.
 379. Lee CW, Chi KN. The standard of reporting of health-related quality of life in clinical cancer trials. J Clin Epidemiol. 2000 May;53(5):451–8.
 380. Brożek JL, Guyatt GH, Schünemann HJ. How a well-grounded minimal important difference can enhance transparency of labelling claims and improve interpretation of a patient reported outcome measure. Health and quality of life outcomes. 2006;4(1):69.
 381. Schünemann HJ, Akl EA, Guyatt GH. Interpreting the results of patient reported outcome measures in clinical trials: The clinician's perspective. Health and quality of life outcomes. 2006;4(1):62.
 382. Symonds T, Berzon R, Marquis P, Rummans TA. The clinical significance of quality-of-life results: practical considerations for specific audiences. In: Mayo Clinic Proceedings. 2002. p. 572.
 383. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. In: Mayo Clinic Proceedings. 2002. p. 371.





APPENDIXES

APPENDIX A EXPERT REVIEWERS

Each of these reviewers was asked to provide comments on one or more particular sections of the report. The report was revised in light of their comments and recommendations were developed once the comments were assimilated. The final report should not be interpreted as an endorsement by any individual expert. Also, none of the experts were asked to review the report in its entirety.

EXPERTS (ALPHABETICALLY, BY SURNAME), *AFFILIATIONS*
[REPORT SECTION(S)]

Tony Ades BSc PhD MSc Professor of Public Health Science, School of Social and Community Medicine, University of Bristol


[Two Sections: 1) Should review articles, pooled data and meta-analysis be used to support clinical/therapeutic claims of effectiveness? and 2) Should indirect comparisons be used to support comparative claims of effectiveness?]

Tony Ades leads a programme of research on formal methods for multi-parameter evidence synthesis in epidemiology and medical decision making, which has been mainly supported by grants from the Medical Research Council. Among the main areas of investigation have been network meta-Analysis, bias models, synthesis of multiple outcomes, synthesis for Markov models, and synthesis applications in infectious disease epidemiology. Previously he was Reader in Epidemiology and Biostatistics at the Institute of Child Health, London, working especially on infectious diseases in the mother, fetus and newborn, on national surveillance systems for infectious diseases, and on a range of issues in antenatal and neonatal screening.

Marc Berger, MD, Executive Vice-President and Senior Scientist at OptumInsight

[Should observational (i.e., non-experimental) studies be used to support clinical/therapeutic claims of effectiveness?]

Marc L. Berger, MD is Executive Vice-President and Senior Scientist at OptumInsight (formerly Ingenix). Marc was formerly Vice President, Global Health Outcomes at Eli Lilly and Company. In this role, he has consolidated health outcomes functions across Lilly into a single organization whose mission is to provide expertise and scientific information that enables Lilly to develop and provide products that deliver better patient outcomes and are valued by payers and providers. A native of New York, he joined Lilly in April 2007 after retiring from Merck & Co., Inc. where he held the position of Vice President, Outcomes Research and Management.



He holds an M.D. degree from Johns Hopkins University School of Medicine and has adjunct appointments as Senior Fellow at the Leonard Davis Institute at the Wharton School of the University of Pennsylvania and Professor in the Department of Health Policy and Administration at the University of North Carolina at Chapel Hill School of Public Health. He also serves on the Medicare Evidence Development & Coverage Advisory Committee (MedCAC) for the Center for Medicare & Medicaid Services (CMS), the evidence-based medicine advisory committee for the National Pharmaceutical Council (NPC), and the editorial advisory board of Value in Health. In addition, he recently completed his term on the steering committee for the Agency for Health Care Research and Quality (AHRQ) Centers for Research and Education on Therapeutics (CERTs). He has published widely in peer-reviewed journals in health services research, outcomes research, health economics, and health policy. An active ISPOR member, Dr. Berger co-edited "Health Care, Cost, Quality, and Outcomes: ISPOR Book of Terms" and co-chaired the ISPOR 2010 Vision Committee.

Trish Groves. Deputy Editor, BMJ and Editor in chief, BMJ Open (qualifications MBBS, MRCPsych)


[Should unpublished studies be used to support clinical/ therapeutic claims of effectiveness?]

I have worked at the BMJ (British Medical Journal, bmj.com) for more than 20 years. I am one of three deputy editors and am also senior research editor. I lead the BMJ team that peer reviews and publishes original research articles, and also lead our international outreach programme, with key responsibility for helping researchers to maximise their chances of publication and for encouraging authors to send the BMJ their research. I write and maintain the BMJ's editorial policies and instructions to authors, and have co-developed the BMJ's regular workshops on peer review training.

On behalf of the BMJ I have been a member of several research-related organisations and groups: the council of the Committee on Publication Ethics (2008-10), the CONSORT 2010 group on reporting randomised controlled trials, and the SPIRIT group on reporting trial protocols. I am also participating in strategic efforts to encourage the sharing of raw research data, to develop prognosis research methods, to revise the EU clinical trials directive, and to improve the practice of grant review.

I helped to develop BMJ Open - the online-only open access general medical journal launched by BMJ Group in early 2011 (bmjopen.bmj.com) – and am its Editor in chief. BMJ Open is dedicated to publishing medical research from all disciplines and therapeutic areas and considers all research study types, from study protocols to phase I trials to meta-analyses, including small or potentially low-impact studies.

Before joining the BMJ I trained in medicine at London's Royal Free Hospital School of Medicine and then specialised in psychiatry, gaining MRCPsych in 1989. In 1998 I was an honorary research fellow at the School for Public Policy, University College London.



I have presented programmes and series for BBC World Service radio, presented TVam's Doc Spot, co-authored the HarperCollins Consumer Guide to Mental Health (winner of the Medical Journalists' Association best book of 1995), and edited the BMJ book Countdown to Community Care (1993).

Paul Kind, MSc.


[How should claims of improvements in patient-reported outcomes/ health-related quality of life be made?]

Paul Kind is an Honorary Professor in the York Centre for Health Economics. His background includes several relevant academic disciplines including economics and psychology, but his work prior to the mid 1970s had been in engineering research and the computer industry. For most of the past 30 years he has been concerned with the development of methods for use in measuring health outcomes, often but not exclusively for application in economic evaluation. He is a founder member and a past-President of the EuroQoL Group which celebrated its 20th Anniversary in 2007, and currently chairs its Scientific Executive. Paul has been a Visiting Scientist at McGill University, Montreal and a Visiting Professor at the University of Wisconsin, Madison and has acted as a coordinator of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Quality of Life Special Interest Group. He is an Honorary Fellow at the Multinational Quality of Life Centre, St Petersburg and is currently Visiting Professor in the Department of Pharmacy, University of Uppsala. He has served as an elected Board Member of the International Society for Quality of Life Research (ISOQOL). Paul provides expert opinion on all aspects of the measurement of health outcomes for a range of organisations in the UK and internationally.

Malcolm Maclure, ScD

[How should statistical information be presented so the reader can assess validity, reliability and level of significance?]

Malcolm Maclure, ScD, is a health services epidemiologist specializing in methodology. He is "British Columbia Chair in Patient Safety" and Professor in the Department of Anesthesiology, Pharmacology and Therapeutics at the University of British Columbia, Vancouver, Canada. He is currently employed as manager of research in the Pharmaceutical Services Division of the British Columbia Ministry of Health in Victoria. From 2002-2006, he was Professor in the School of Health Information Science at the University of Victoria, BC, funded by a Michael Smith Foundation Distinguished Scholar Award. Trained in the Department of Epidemiology at Harvard School of Public Health, he continues his affiliation there as Adjunct Professor. He is current president of the Society for Epidemiologic Research. Among epidemiologists, he is best known for inventing the case-crossover design to study triggers of acute events. He is interested in the causation and measurement of bias in epidemiologic studies. His recent investigations concern drug policy impact evaluations using administrative databases and randomized pragmatic trials. He is exploring the potential to combine these methods to evaluate the effectiveness of drugs in the real world and health system safety improvements.



Muhammad Mamdani, PharmD, MA, MPH Director of the Applied Health Research Centre (AHRC) of the Li Ka Shing Knowledge Institute (LKSKI) at St. Michael's Hospital.


[Should secondary outcomes, subgroup analysis, and post-hoc analysis be used to support clinical/ therapeutic claims of effectiveness?]

The recipient of a Caldwell Partners International Top 40 under 40, Dr. Mamdani provides overall leadership of the AHRC and is also actively involved in clinical research as the lead of the Ontario Drug Policy Research Network. Dr. Mamdani completed his Doctor of Pharmacy (PharmD) degree from the University of Michigan, a fellowship in pharmacoeconomics and a Master of Arts (MA) degree in economics from Wayne State University and a Master of Public Health (MPH) in quantitative methods from Harvard University. He is an internationally renowned pharmacoepidemiologist who has published over 200 articles in peer-reviewed medical journals such as the New England Journal of Medicine, the Lancet, and the Journal of the American Medical Association, the British Medical Journal, and the Canadian Medical Association Journal. In addition to his leadership role with the AHRC, Dr. Mamdani is an Associate Professor of Medicine and Pharmacy in the respective faculties at the University of Toronto, an Adjunct Scientist at the Institute for Clinical Evaluative Sciences (ICES), and a member of the Human Drug Advisory Panel of the Patented Medicine Prices Review Board (PMPRB) of Canada.

Alan Mathios is currently a Professor at Cornell University and the Rebecca Q. and James C. Morgan Dean for the College of Human Ecology.

[Introduction]

He is a member of the Department of Policy Analysis and Management and served as Associate Chair and Director of Undergraduate Studies for the Department. He is the North American Editor of the Journal of Consumer Policy and on the Editorial Boards of the Journal of Consumer Affairs and the Journal of Public Policy and Marketing. He was also the Project Leader on the Merck Foundation Co. Program Consumers, Pharmaceutical Policy and Health. He came to Cornell following six years of employment at the Federal Trade Commission (FTC), where he served as a staff economist in the Division of Economic Policy Analysis and as an econometrics consultant to the Bureau of Economics. While serving at the FTC her received numerous awards including the Outstanding Scholarship award, the Excellence in Economics Award, and the Award for Superior Service. A major focus of his research is on the effect of Food and Drug Administration regulatory policies on consumer and firm behaviour. His research



also focuses on government tax policy and its impact on smoking onset and cessation. His research has been funded by a variety of sources including the National Cancer Institute, the Robert Wood Johnson Foundation and the Merck Foundation Co. He has been the recipient of a number of teaching and advising awards including the SUNY Chancellor's Award for Excellence in Teaching and the Cornell University Kendal S. Carpenter Advising Award.

Josephine A. Mauskopf, PhD, MHA, Vice President of Health Economics at RTI Health Solutions


[Two sections: 1) How Should Health Economic Claims Be Made? And 2) Should mathematical modeling be used to support comparative claims of effectiveness?]

Josephine Mauskopf has extensive experience both as a consultant and within the pharmaceutical industry designing and implementing pharmacoeconomic research strategies. She has designed pharmacoeconomic research programs for drugs for bacterial infections, viral infections, psychiatric illness, and neurologic diseases. Dr. Mauskopf has estimated budget impacts for new products for schizophrenia, bipolar disease, breast cancer, and HIV infection. She has estimated the cost-effectiveness of antiretroviral drugs, as well as drugs for treating herpes zoster, epilepsy, neonatal respiratory distress syndrome, digoxin toxicity, community-acquired pneumonia, intra-abdominal infections, and primary pulmonary hypertension. Dr. Mauskopf was previously vice president at MEDTAP International, department head of Economics Research at Burroughs Wellcome, and director of Pharmacoeconomics Research for anti-virals and anti-infectives at Glaxo Wellcome. She recently completed 8 years as Editor-in-Chief of Value in Health. She has presented her research at numerous national and international symposia, and has published extensively in journals.

Mark Sculpher PhD

[Two sections: 1) How Should Health Economic Claims Be Made? And 2) Should mathematical modeling be used to support comparative claims of effectiveness?]

Mark Sculpher PhD is Professor of Health Economics at the Centre for Health Economics, University of York, UK, and is Director of the Programme on Economic Evaluation and Health Technology Assessment. He is also Director of Oxford Outcomes Ltd. Mark has worked in the field of economic evaluation and health technology assessment for over 20 years. He has researched in a range of clinical areas including heart disease and cancer. He has also contributed to methods in the field, in particular relating to decision analytic modelling and



techniques to handle uncertainty. He has over 160 peer-reviewed publications and is a co-author of two major text books in the area: *Methods for the Economic Evaluation of Health Care Programmes* (OUP, 2005 with Drummond, Torrance, O'Brien and Stoddart) and *Decision Modelling for Health Economic Evaluation* (OUP, 2006 with Briggs and Claxton).

Mark is a member of the UK National Institute of Health Research College of Senior Investigators. He has also been a member of the National Institute for Health and Clinical Excellence (NICE) Technology Appraisal Committee and the NICE Public Health Interventions Advisory Committee. He chaired NICE's 2004 Task Group on methods guidance for economic evaluation and advised the Methods Working Party for the 2008 update of this guidance; he has also advised health systems internationally on HTA methods including those in Ireland, Portugal and New Zealand. He has been a member of the Commissioning Board for the UK NHS Health Technology Assessment programme and currently sits on the UK Medical Research Council's Methodology Research Panel.

Mark was Issue Panel Committee Co-Chair for the ISPOR 10th Annual European Congress, teaches the ISPOR Short Course, "Advanced Decision Modeling for Health Economic Evaluations", and participated on the Leadership Group of the Transferability of Economic Evaluations Across Jurisdictions: ISPOR Good Research Practices Task Force.