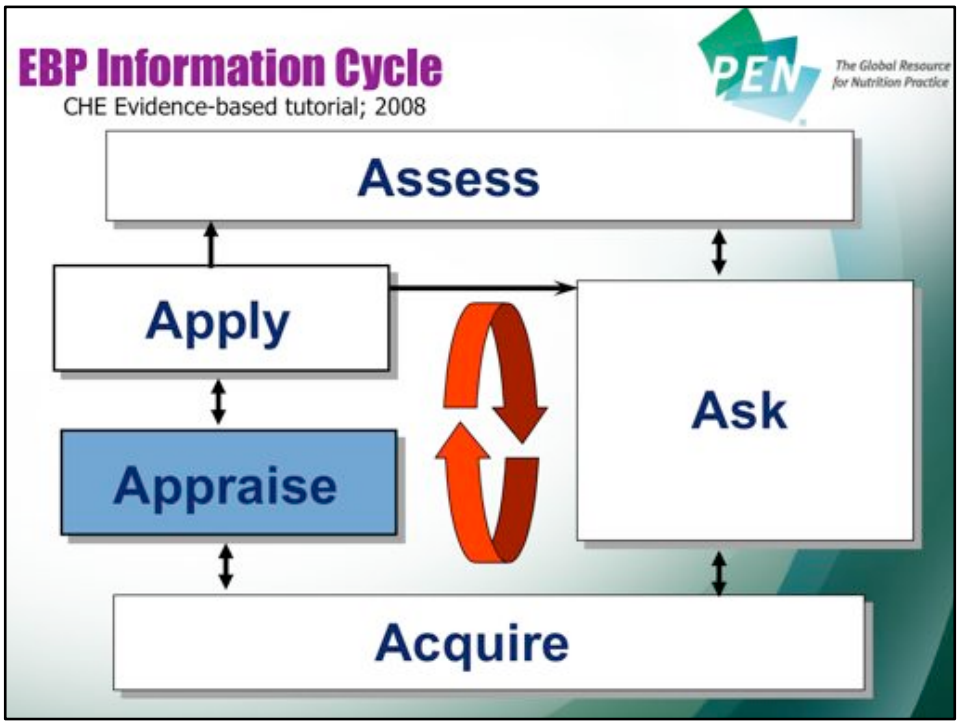
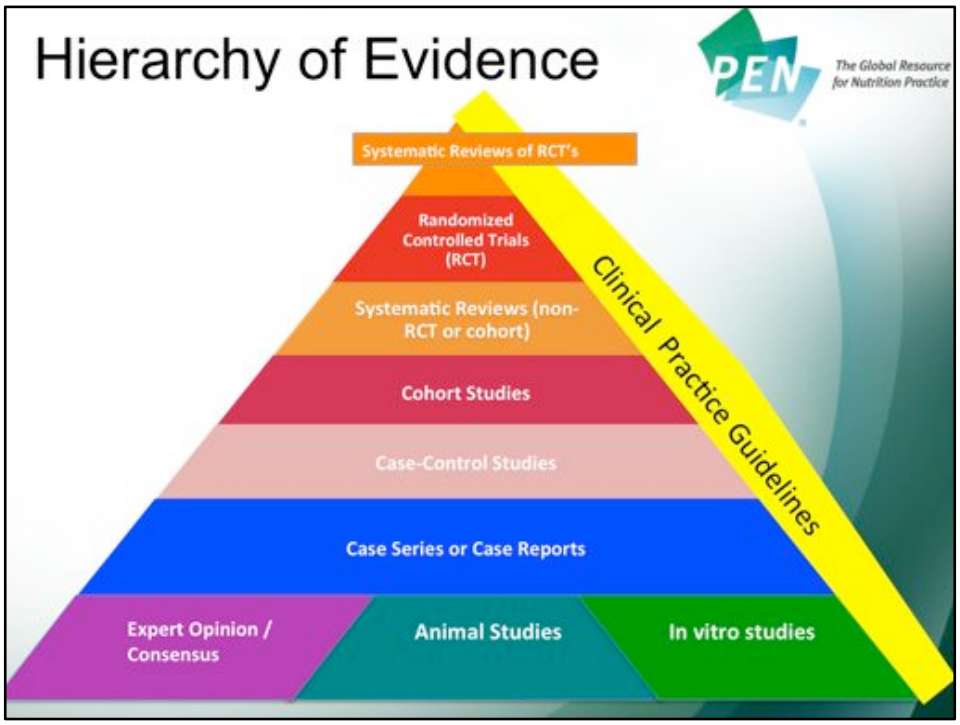
Practice-based Evidence in Nutrition (PEN®, www.pennutrition.com) is the global resource for nutrition practice.

This is a PEN® training module specific to critically appraising the literature. This module contains links to additional learning modules on appraisal as well as an additional guideline for appraising clinical practice guidelines (CPG's).

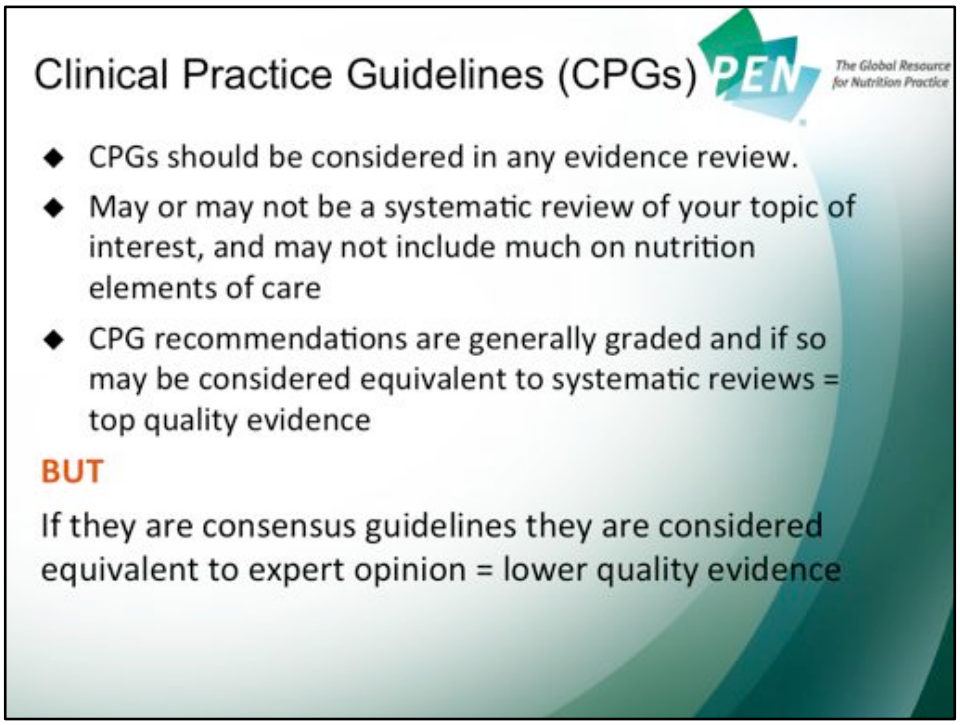


Critically appraising the evidence comes after acquiring the highest quality evidence.

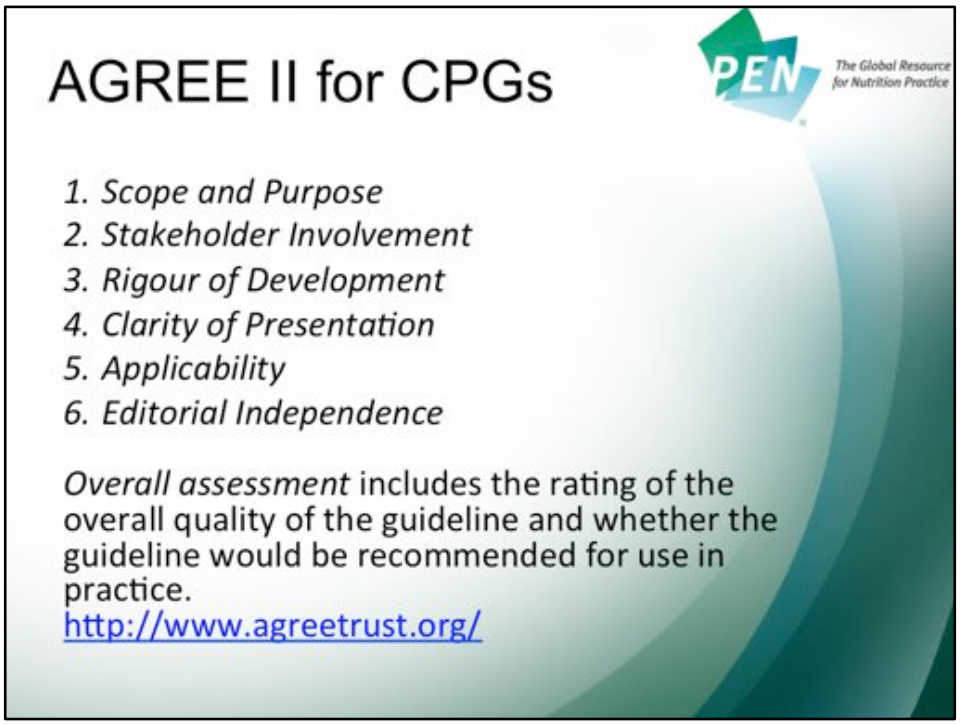


Hierarchy of Evidence: to help us understand that every study is not created equal and some are better quality evidence than others. Some types of studies are more appropriate depending on the question being asked.
In addition to the type or design of the study, the quality of the study (e.g. well designed, size, risk of bias…) is a key component in determining how it contributes to the graded evidence used in PEN®.
Other slides in this training module discuss the PEN® grading system and the PEN® Evidence Grading Checklist in some detail.
Depending on your quality assessment of the document then generally the higher quality evidence in terms of design is from:
Systematic reviews – to be highest quality must be of randomized controlled trials (RCTs)
RCTs
THEN systematic reviews of nonrandomized or cohort studies
Cohort studies
Case-control studies
Lower quality:
Case series or Case reports – descriptions of a single or a series of cases of some illness or disease
Expert opinion – not actually evidence, but opinion
Animal research – studies conducted on animals cannot be applied directly to humans as the biology may be different
In vitro - a biological process conducted in a laboratory container such as a test tube or petri dish
It should be noted that there are some issues/questions where animal and in vitro studies are the most appropriate research design but the quality of the evidence is still graded low.

Note: (October 1, 2013) This is the current hierarchy of evidence but we are aware of continuing discussions taking place within groups such as Cochrane and GRADE. The PEN® team continues to monitor these discussions.



Read this slide… in the next slide, we share some tools to evaluate CPGs…



AGREE stands for "Appraisal of Guidelines for Research and Evaluation" (AGREE)" and is intended to advance the science of practice guidelines. AGREE II is a valid and reliable tool for use in evaluating or developing clinical practice guidelines.

The AGREE II tool consists of 23 key items to evaluate practice guideline quality organized within 6 domains followed by 2 global rating items ("Overall Assessment"). Each domain captures a unique dimension of guideline quality.

Scope and Purpose is concerned with the overall aim of the guideline, the specific health questions, and the target population.

Stakeholder Involvement focuses on the extent to which the guideline was developed by the appropriate stakeholders and represents the views of its intended users.

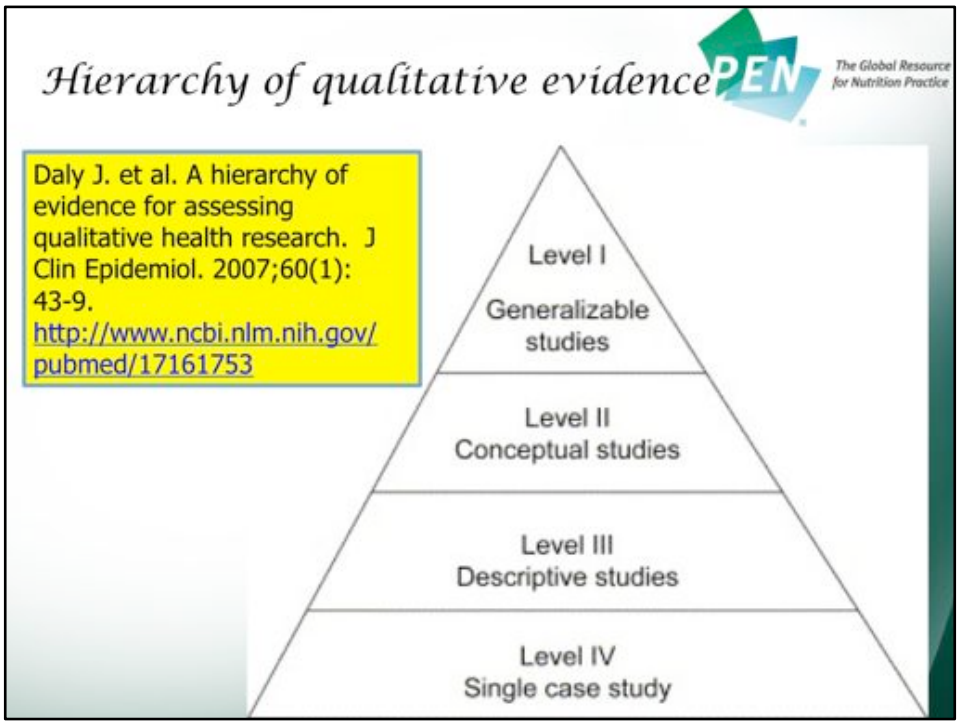
Rigour of Development relates to the process used to gather and synthesize the evidence, the methods to formulate the recommendations, and to update them.

Clarity of Presentation deals with the language, structure, and format of the guideline.

Applicability pertains to the likely barriers and facilitators to implementation, strategies to improve uptake, and resource implications of applying the guideline.

Editorial Independence is concerned with the formulation of recommendations not being unduly biased with competing interests.

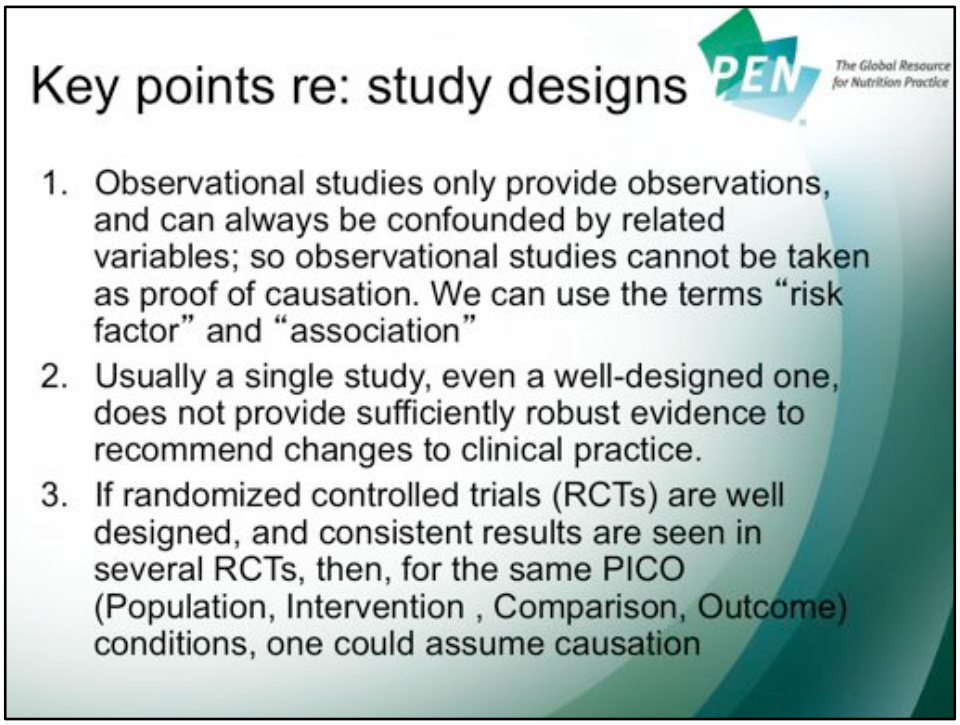
Overall assessment includes the rating of the overall quality of the guideline and whether the guideline would be recommended for use in practice.



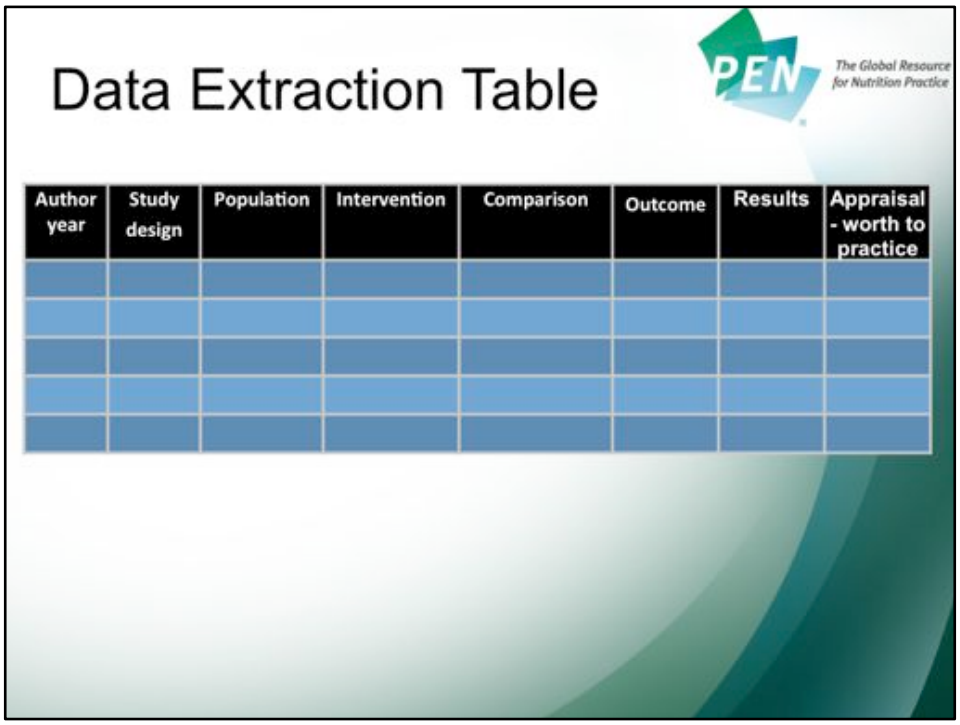
Here is a hierarchy for appraising qualitative evidence.

Qualitative studies use non-numerical information, and can provide valuable insight into how people experience states of health and illness or how things came about.

Generally qualitative studies are beyond the scope of this presentation. However, in case you are assessing qualitative evidence, here is a reference to a tool to assess their rigor.
Daly J. et al. A hierarchy of evidence for assessing qualitative health research. J Clin Epidemiol. 2007;60(1):43-9. The absract is available at:
http://www.ncbi.nlm.nih.gov/pubmed/17161753

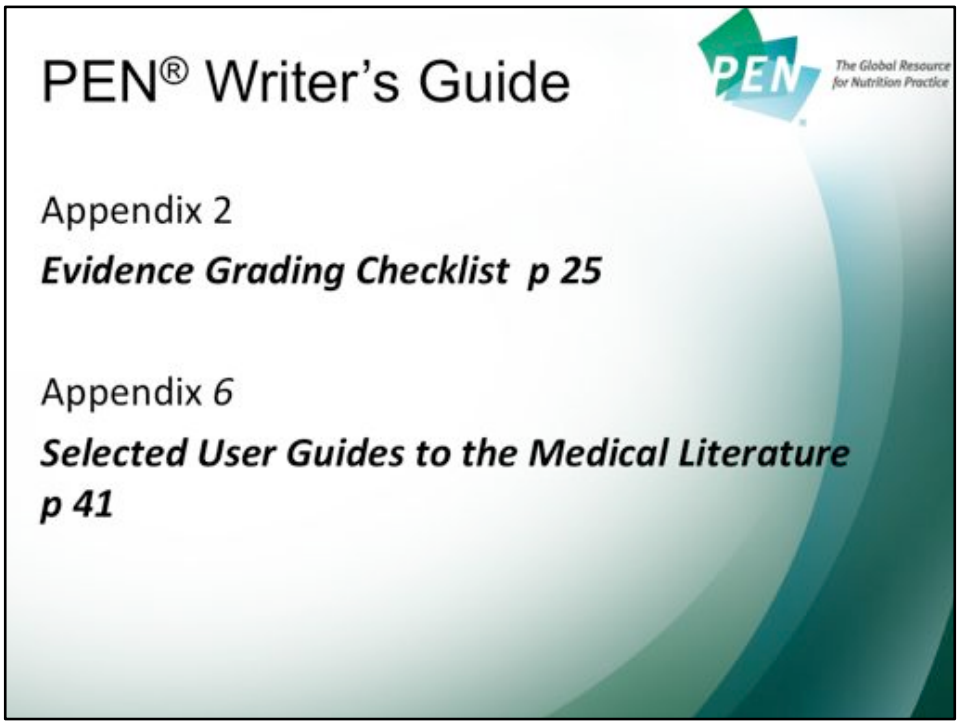


Read slide. If study design is an area that you would like to look at in more detail complete the "PEN® Quick Review of Study Designs" training module available at: http://www.pennutrition.com/WriterGuide.aspx



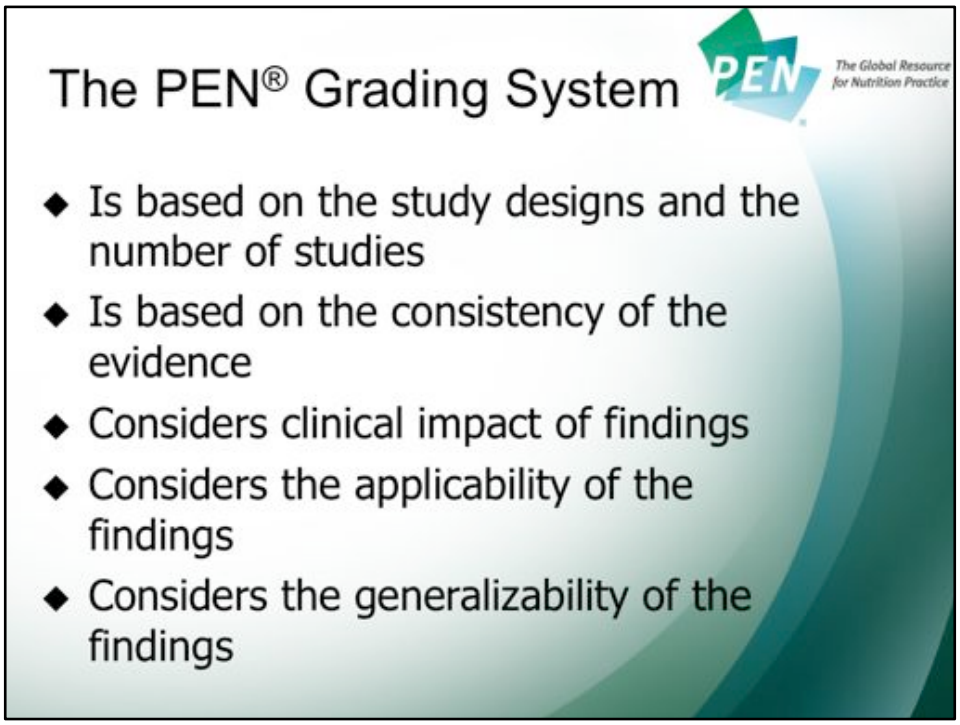
It helps to summarize your findings in a table like this one, in order to look over the study designs and results, to help you understand the data you found, and to see similarities and differences across the studies. Note, however, that when writing for PEN® we primarily present our evidence in paragraphs. If you feel it is best to present your evidence in PEN® as a table, discuss this with the editor.

If you found several studies, it can help to prepare separate tables with PICO in mind. For example, if your topic was calcium and bone health - a separate table by Population may help (e.g. one for adolescent females, one for post-menopausal women etc), by Intervention (e.g. one for calcium from food, one for calcium from supplements), by Outcome (e.g. one for bone mineral density, one for fracture incidence). This will also help you organize your key practice points so that they are clear, concise and organized.

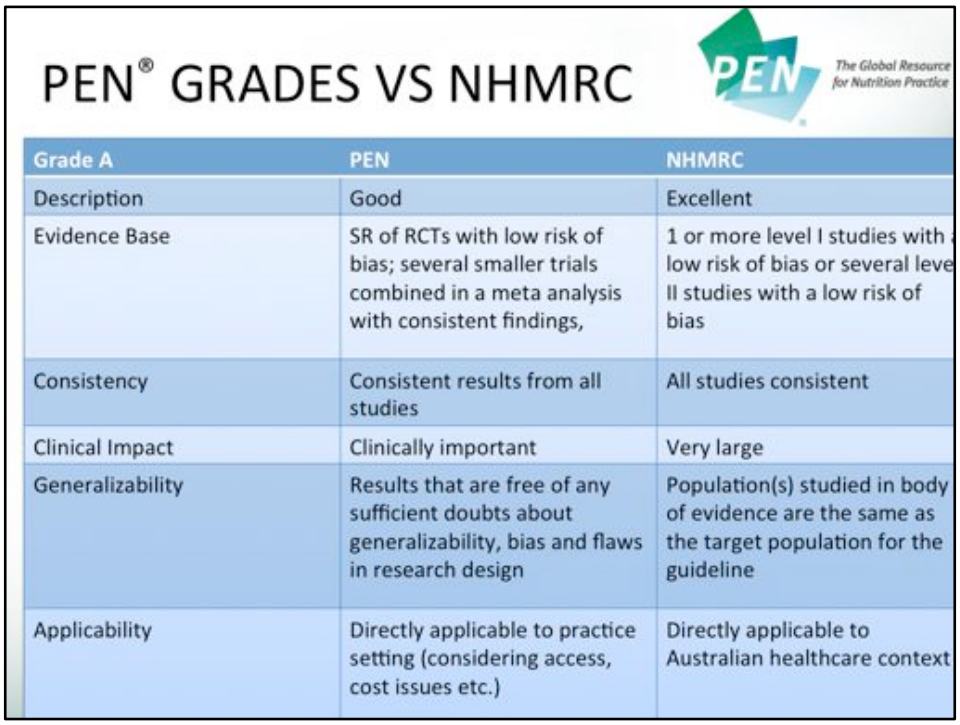


The PEN® Evidence Grading Checklist is in the PEN® Writer's Guide.
User guides are also in PEN® Writer's Guide and PDF documents of both are linked just below the link to the "PEN® Evidence-based Process" training module in PEN® .
All of these can be accessed from: http://www.pennutrition.com/WriterGuide.aspx

The Critical Appraisal Skills Programme (CASP) also has critical appraisal checklists to assess the applicability, reliability and validity of research. These tools can be downloaded from http://www.caspinternational.org/?o=1012

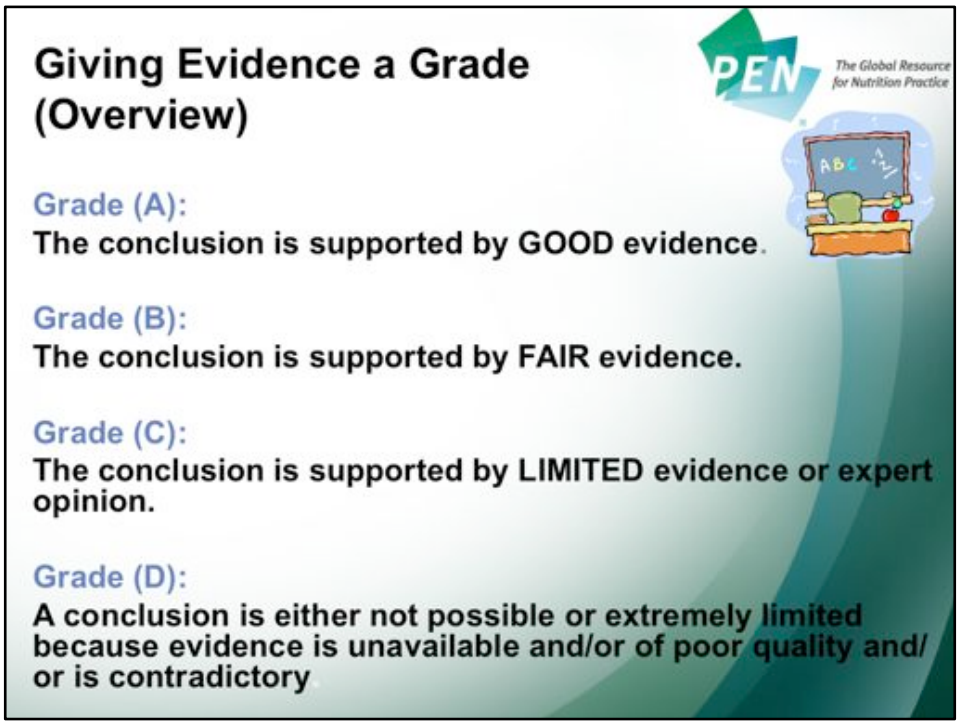


Similar to other credible grading frameworks, the PEN® Grading system…read slide



Grade A	PEN	NHMRC
Description	Good	Excellent
Evidence Base	SR of RCTs with low risk of bias; several smaller trials combined in a meta analysis with consistent findings,	1 or more level I studies with a low risk of bias or several level II studies with a low risk of bias
Consistency	Consistent results from all studies	All studies consistent
Clinical Impact	Clinically important	Very large
Generalizability	Results that are free of any sufficient doubts about generalizability, bias and flaws in research design	Population(s) studied in body of evidence are the same as the target population for the guideline
Applicability	Directly applicable to practice setting (considering access, cost issues etc.)	Directly applicable to Australian healthcare context

IN 2010 as part of the DC DAA partnership discussions we compared the PEN® grading system with the National Health and Medical Research Council (NHMRC, Australia) and found tremendous congruency. We actually tweaked the PEN® grading system to be closer in line with NHMRC



A number of information sources were used in developing a grading system for PEN®.

Gray GE and Gray LK. Evidence-based medicine: Applications in dietetic practice. J Am Diet Assoc 2002; 102: 1263-1272. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=12792624&query_hl=3; accessed 25 Aug 2005.

Grandage KK; Slawson DC; Shaughnessy AF. When less is more: a practical approach to searching for evidence-based answers. J Med Libr Assoc July 2002; 90(3): 298-2002. Available from http://www.pubmedcentral.gov/picrender.fcgi?artid=116402&blobtype=pdf; accessed 25 Aug 2005.

Personal communication. R Hayward. Centre for Health Evidence. 2004.

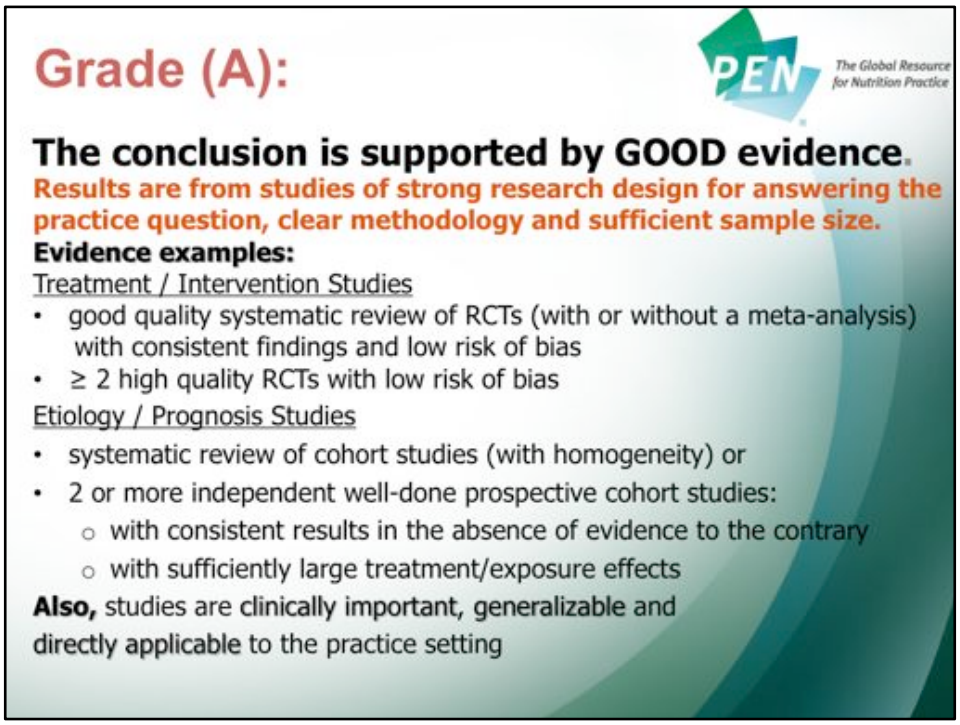
Canadian Task Force on Preventive Health Care. New grades for recommendations from the Canadian Task Force of Preventive Health Care. CMAJ. 2003; 169(3): 207-208. Available from: CMAJ accessed 25 Aug 2005.

Upshur, REG. Are all evidence-based practices alike? Problems in the ranking of evidence. CMAJ; 2003; 169(7): 672-673. Available from: http://www.cmaj.ca/cgi/content/full/169/7/672; accessed 25 August 2005.

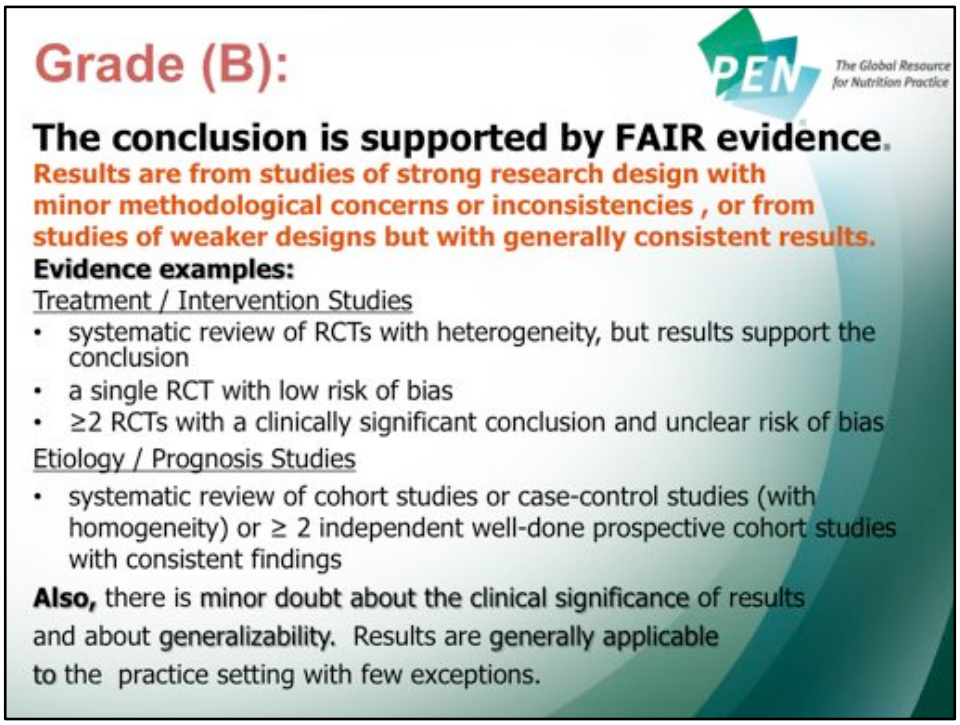
Schüremann HJ, Best D, Vist G, Oxmann AD On behalf of the GRADE Working Group.Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations CMAJ. 2003; 169(7):677. Available from: http://www.cmaj.ca/cgi/content/full/169/7/677; accessed 25 August 2005.

Ebell MH, Siwek J, Weiss BD, Woolf SH, Susman J, Ewigan B, Bowman M. SORT: A Patient Centered Approach to Grading the Literature. Am Fam Physician. 2004; 69(3):548-56. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=14971837&query_hl=6; accessed 25 August 2005.

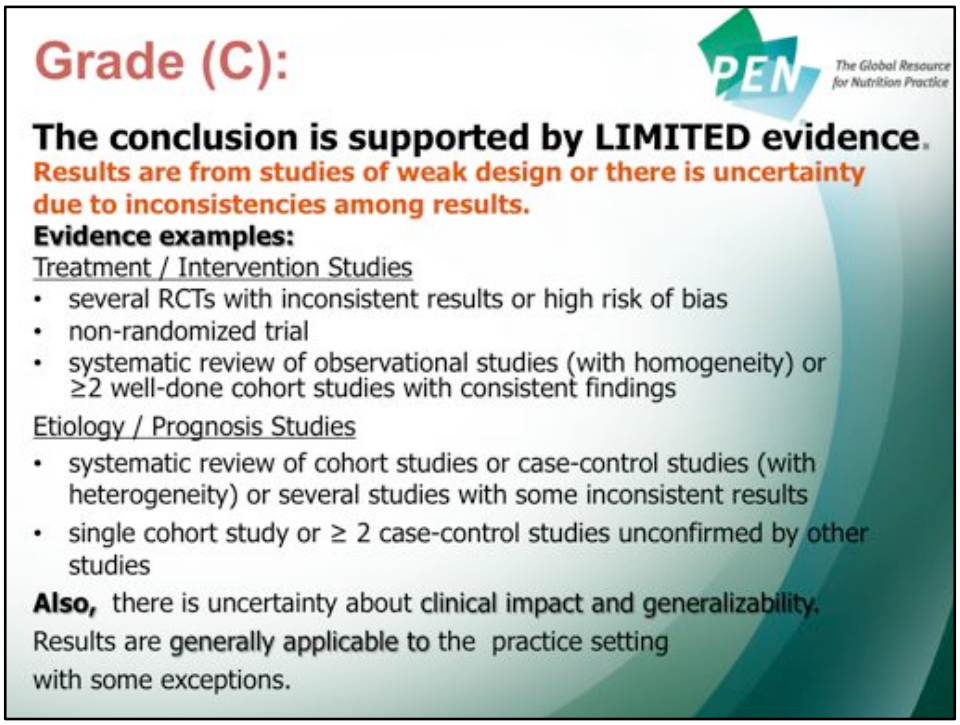
GRADE Working Group. Grading Quality of Evidence and Strength of Recommendations. BMJ. 2004; 328:1490. Available from: http://bmj.bmjjournals.com/cgi/content/full/328/7454/1490; accessed 25 August 2005.



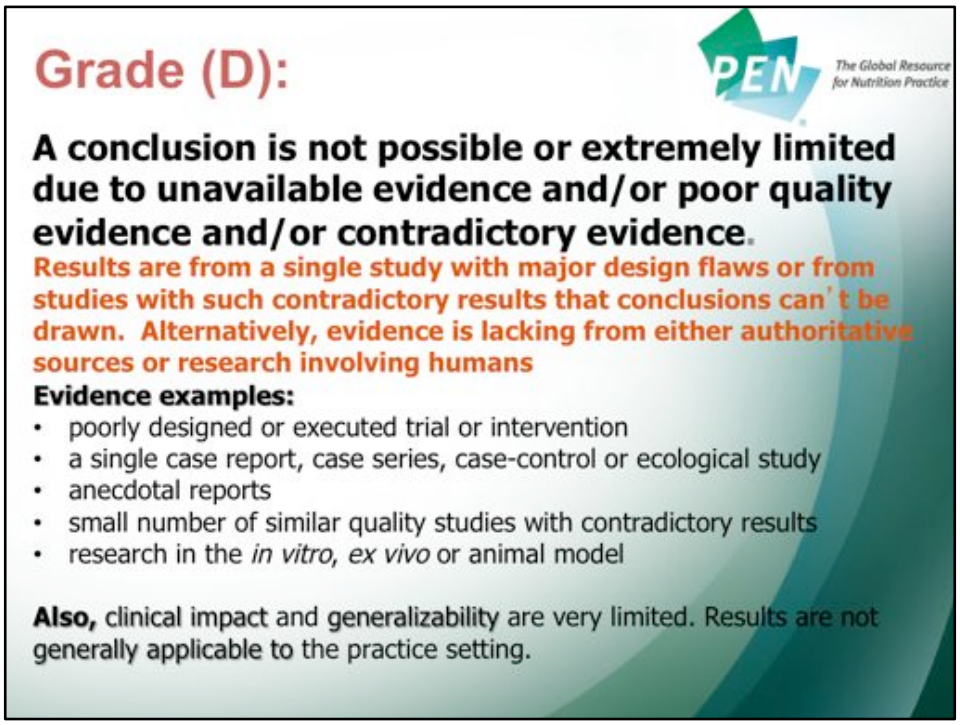
This slide and the next 3 slides provide examples of evidence and other factors to consider in grading the evidence (this information is drawn from the PEN® Evidence Grading Checklist in Appendix 2, p. 25 of the PEN® Writer's Guide). A direct link to the PEN® Evidence Grading checklist can be found just below the link to the "PEN® Evidence-based Process" training module in PEN® , which can be accessed from: http://www.pennutrition.com/WriterGuide.aspx


Grade (B):
PEN
The Global Resource for Nutrition Practice
The conclusion is supported by FAIR evidence.
Results are from studies of strong research design with minor methodological concerns or inconsistencies , or from studies of weaker designs but with generally consistent results.
Evidence examples:
Treatment / Intervention Studies
• systematic review of RCTs with heterogeneity, but results support the conclusion
• a single RCT with low risk of bias
• ≥2 RCTs with a clinically significant conclusion and unclear risk of bias
Etiology / Prognosis Studies
• systematic review of cohort studies or case-control studies (with homogeneity) or ≥ 2 independent well-done prospective cohort studies with consistent findings
Also, there is minor doubt about the clinical significance of results and about generalizability. Results are generally applicable to the practice setting with few exceptions.

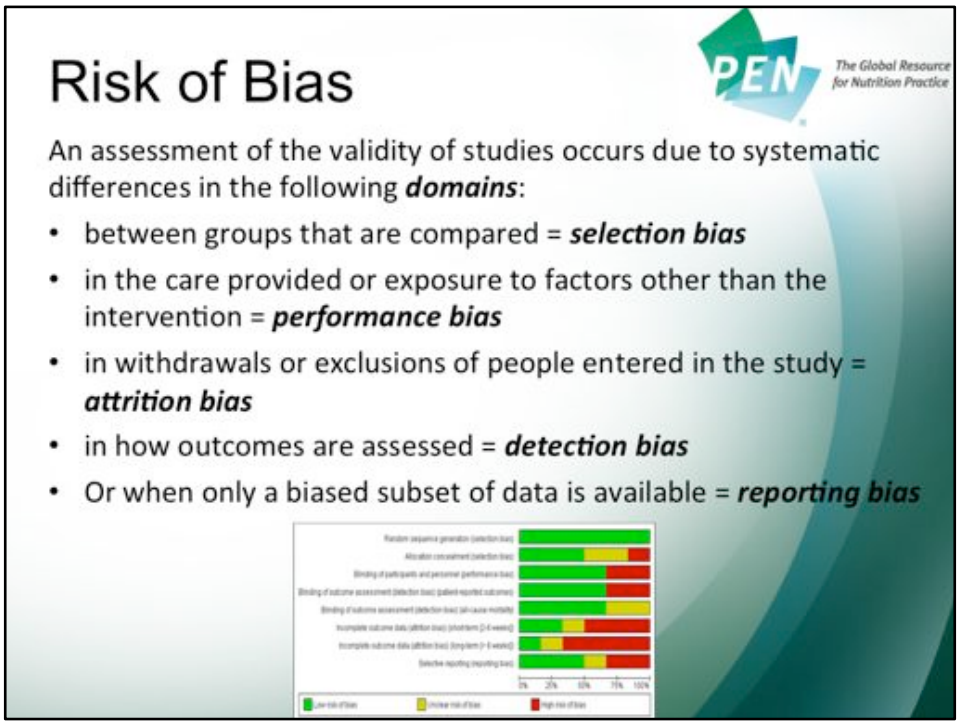
Read slide….


Grade (C):
PEN
The Global Resource for Nutrition Practice
The conclusion is supported by LIMITED evidence.
Results are from studies of weak design or there is uncertainty due to inconsistencies among results.
Evidence examples:
Treatment / Intervention Studies
• several RCTs with inconsistent results or high risk of bias
• non-randomized trial
• systematic review of observational studies (with homogeneity) or ≥2 well-done cohort studies with consistent findings
Etiology / Prognosis Studies
• systematic review of cohort studies or case-control studies (with heterogeneity) or several studies with some inconsistent results
• single cohort study or ≥ 2 case-control studies unconfirmed by other studies
Also, there is uncertainty about clinical impact and generalizability.
Results are generally applicable to the practice setting
with some exceptions.

Read slide….

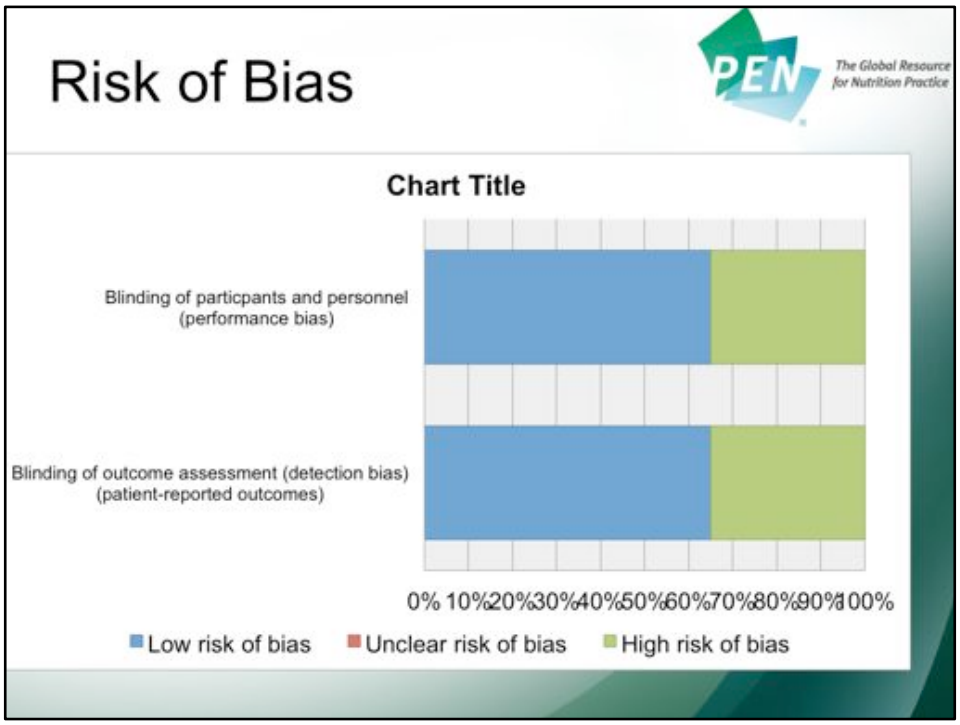

Grade (D):
PEN
The Global Resource for Nutrition Practice
A conclusion is not possible or extremely limited due to unavailable evidence and/or poor quality evidence and/or contradictory evidence.
Results are from a single study with major design flaws or from studies with such contradictory results that conclusions can't be drawn. Alternatively, evidence is lacking from either authoritative sources or research involving humans
Evidence examples:
• poorly designed or executed trial or intervention
• a single case report, case series, case-control or ecological study
• anecdotal reports
• small number of similar quality studies with contradictory results
• research in the in vitro, ex vivo or animal model
Also, clinical impact and generalizability are very limited. Results are not generally applicable to the practice setting.

Read slide….

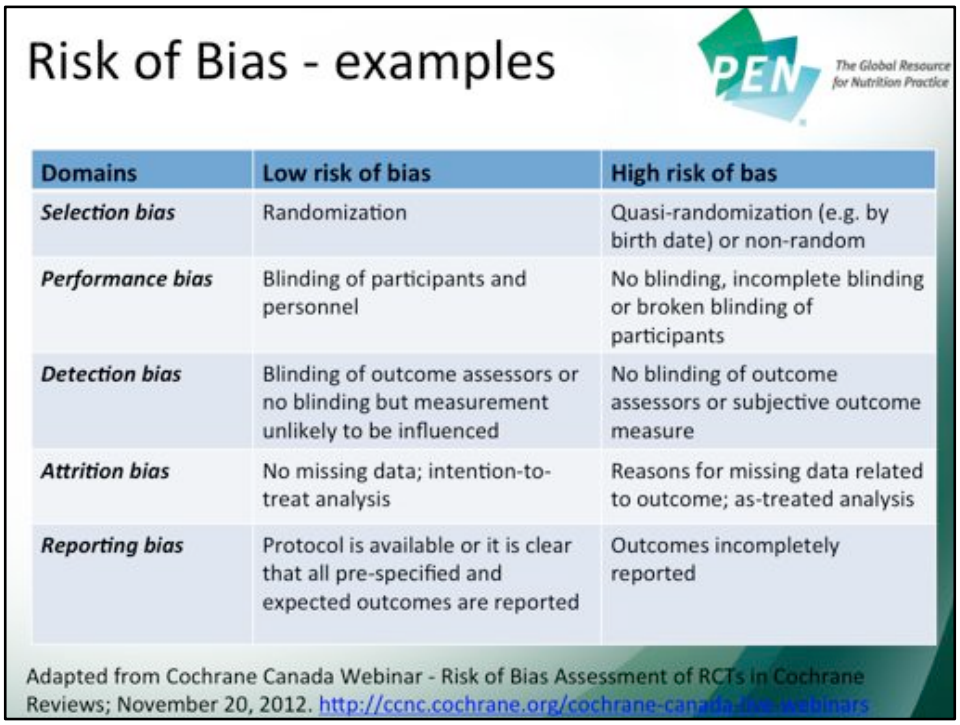


Study Quality or Risk of Bias?

Bias can occur in either direction, resulting in an underestimation or overestimation of the true effect, and can range from trivial to substantial. The term, *risk of bias* is used in Cochrane reviews and in PEN®'s Evidence Grading Checklist to help explain variation in the results of studies, including those reported in systematic reviews (e.g. helps to explain heterogeneity of results). Assessing risk of bias establishes the degree to which results of studies should be believed. Risk of bias is distinguished from study quality, which refers to the critical appraisal of a study. Bias can occur in well conducted studies, and not all methodological flaws introduce bias.



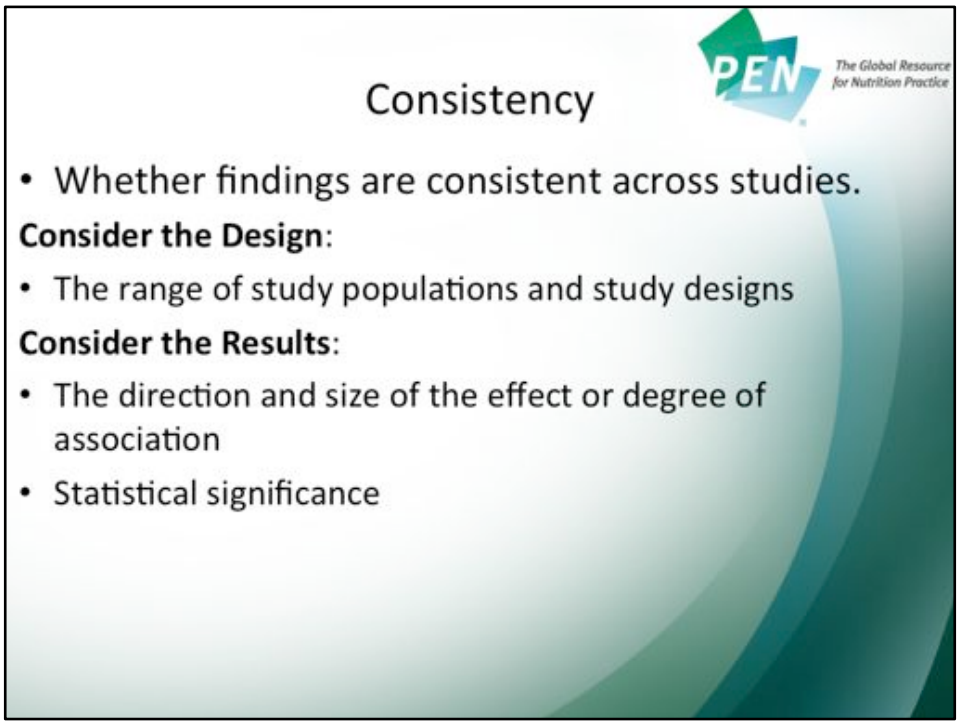
Overall risk of bias – Drawing conclusions about the overall risk of bias within trials or across trials, means summarizing assessments for each outcome, and deciding which domains are most important in the context of the research question being asked. Cochrane often reports this as a graph so you can see the proportion of studies that have low, unclear, or high risk of bias for each domain. Within a trial, a low risk of bias generally means that a low risk of bias was identified for all of the key domains; a high risk of bias within a trial generally means that a high risk of bias was identified for one or more key domains. For systematic reviews, a low risk of bias means that most information is from trials at low risk of bias; a high risk of bias means that a proportion of information from trials at high risk of bias is sufficient to affect the interpretation of results.



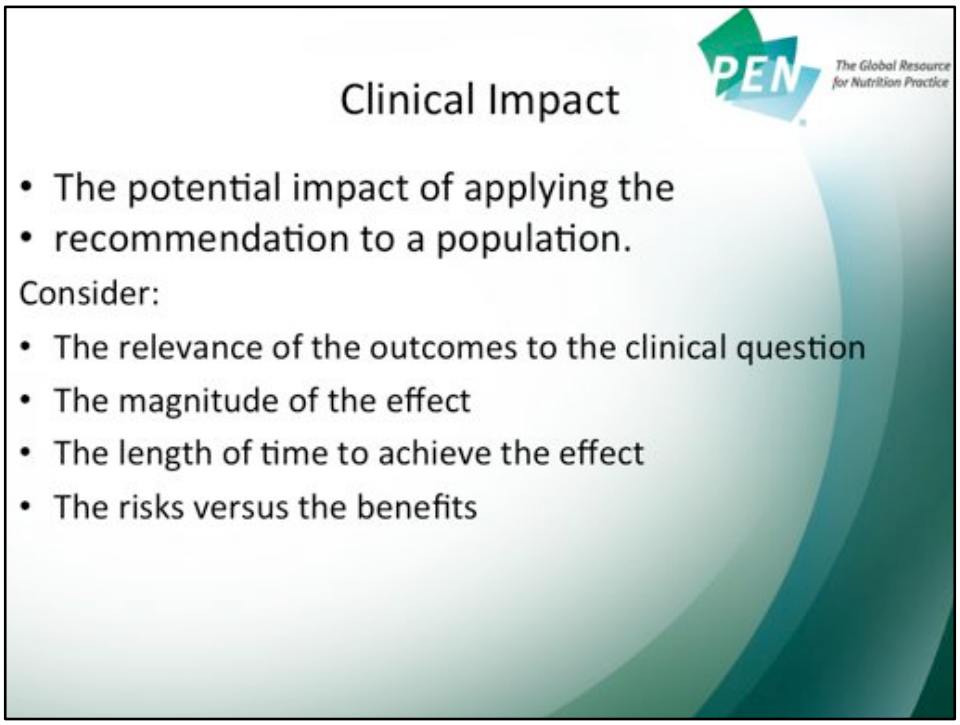
Cochrane uses several domains to evaluate bias and encourages categorizing each domain as a low or high risk of bias (unclear risk of bias is also an option) The table shows examples that demonstrate low and high risk of bias for each of these domains. This information was adapted from a Cochrane webinar available at: http://ccnc.cochrane.org/cochrane-canada-live-webinars
A good reference paper is: Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al.; Cochrane Bias Methods Group; Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011 Oct 18 [cited 2012 Nov 21];343:d5928. Available from: http://www.bmj.com/content/343/bmj.d5928?view=long&pmid=22008217

Additional information on risk of bias can be found in PEN®eNews, '*HOW MUCH SHOULD I BELIEVE THE RESULT?*' available at: http://www.pennutrition.com/enews.aspx?id=10#119

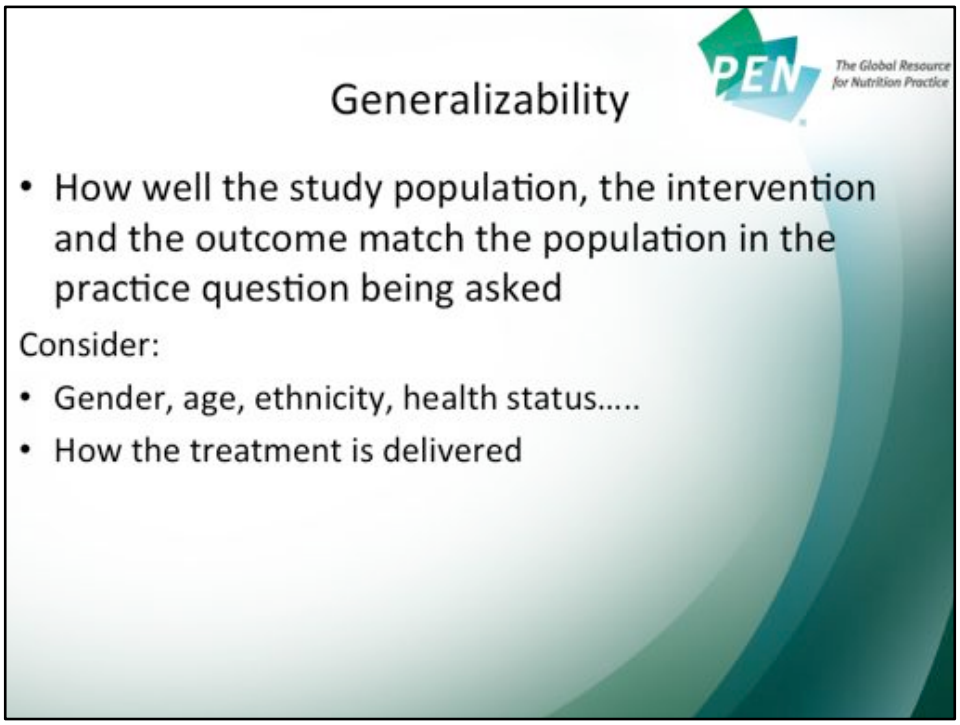
PEN® eNews also covers other relevant topics on bias, including:
'SEARCHING FOR ANSWERS: COMBATING PUBLICATION BIAS' available from: http://www.pennutrition.com/enews.aspx?id=8#79
'AND ANOTHER THING…ABOUT BIAS' available from: http://www.pennutrition.com/enews.aspx?id=12#137



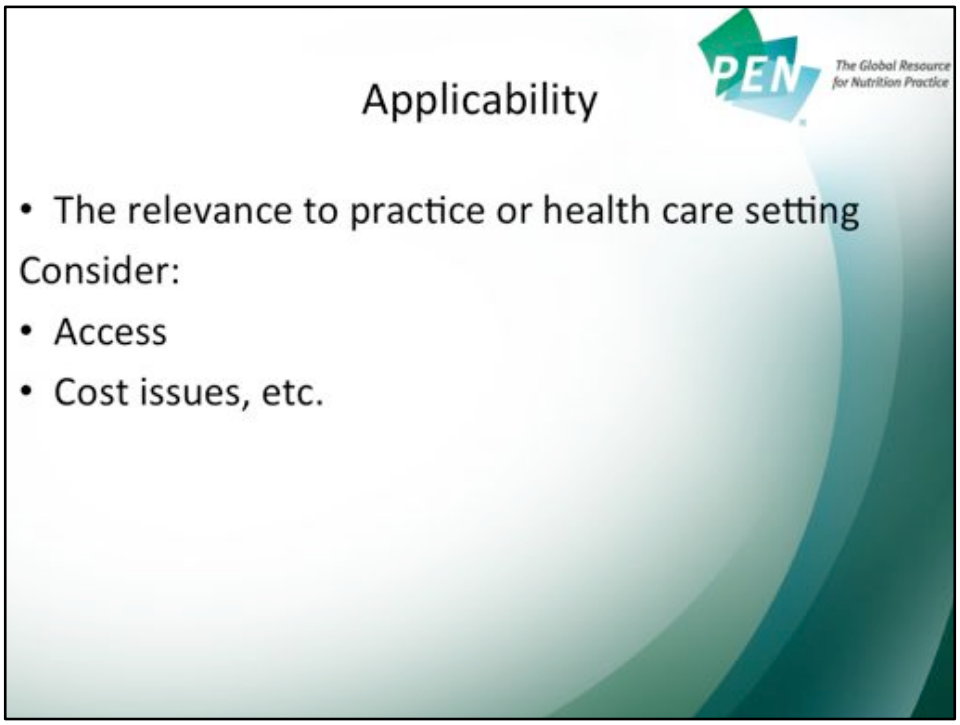
Other factors to consider are consistency….. Read slide



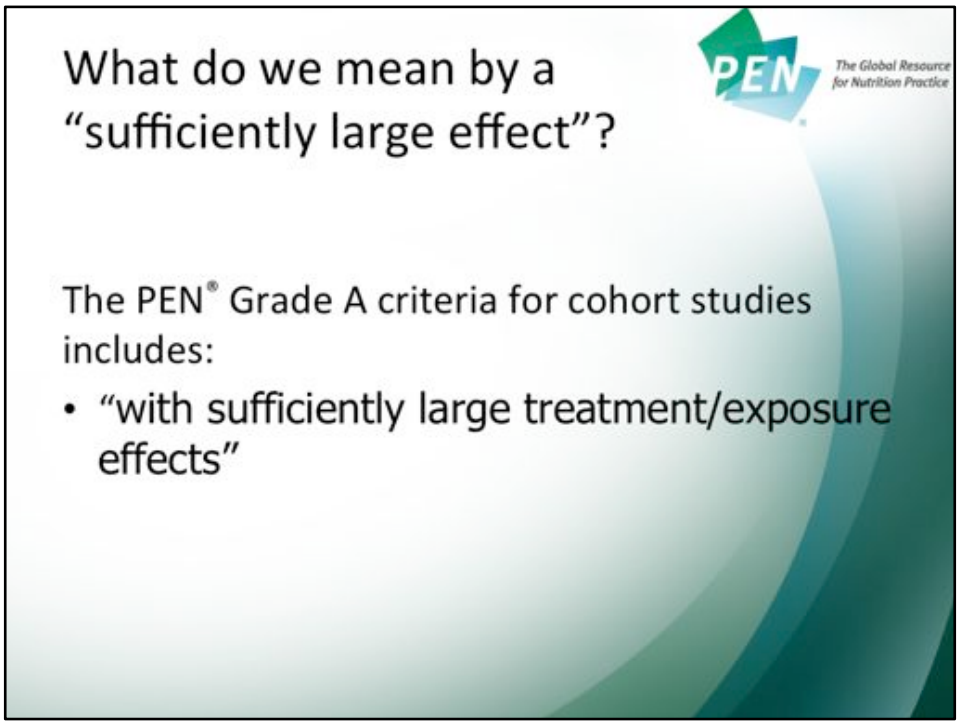
clinical impact….. Read slide



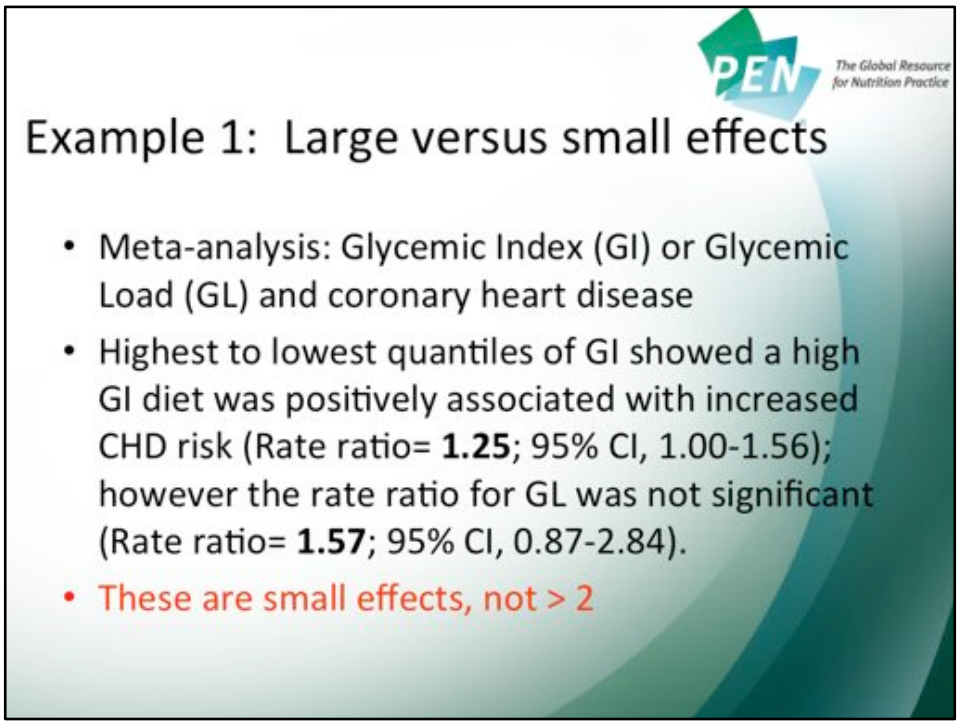
generalizability….Read slide



and applicability….Read slide



What is meant by a large "effect" is referring to the results, that is the size of the estimate of the association, which in observational studies is often estimated as an odds ratio (OR) or a risk ratio (RR, sometimes referred to as the relative risk). You see ORs or RRs in studies that report the proportions of research participants who have or do not have an outcome (two possible outcomes, a dichotomous outcome). Other studies may have a continuous outcome, such as changes in weight or lab values. All of these outcomes are generally referred to as the "effect" or "effect size" of the study.

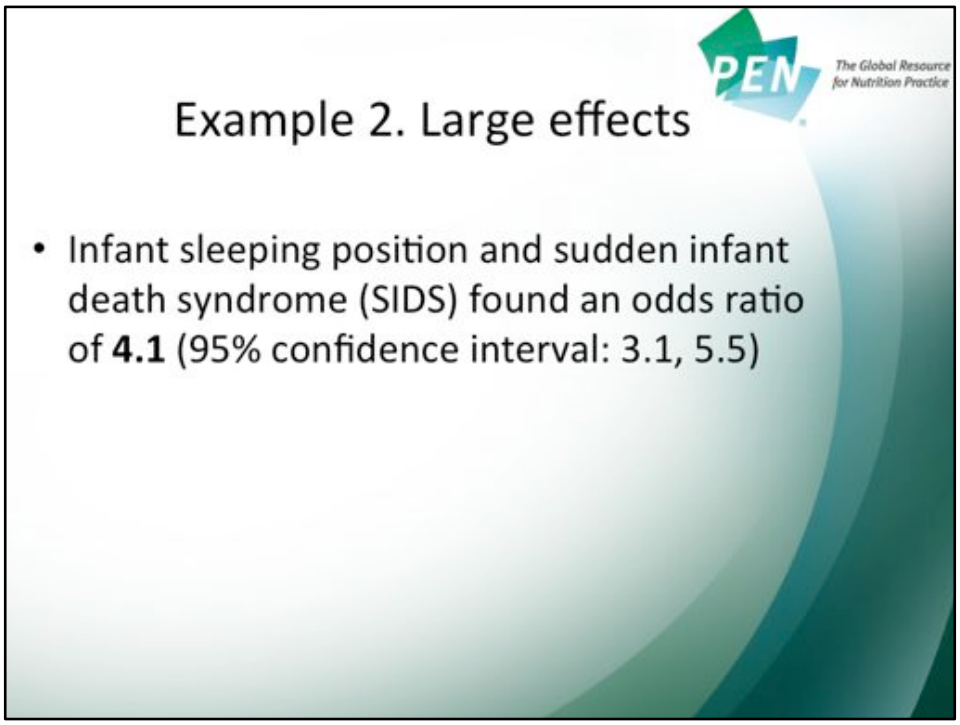


The Grades of Recommendation, Assessment, Development and Evaluation (GRADE) Working Group (http://www.gradeworkinggroup.org/) suggests that observational evidence is supported if the results show: **A large effect when RR or OR > 2 or very large effect when RR or OR > 5; (when protective, a large effect is when RR or OR <0.5 and a very large effect is when RR or OR < 0.2)**

This example shows: Both estimates are small effects, with Rate Ratios of 1.25 and 1.57 – so this is not strong evidence.

(Interesting that the smaller value was statistically significant, while the larger value of 1.57 was not. To be statistically significant, as well as the value of the rate ratio, sample size and the frequency of occurrence in the sample are important determinants. Since this data came from a large sample size study, even the small rate ratio of 1.25 was statistically significant. The main point GRADE is making is that these are not important rate ratios, even though one of them was statistically significant.)

Study reference: Barclay AW, Petocz P, McMillan-Price J, Flood VM, Prvan T, Mitchell P, et al. Glycemic index, glycemic load, and chronic disease risk--a meta-analysis of observational studies. Am J Clin Nutr. 2008 Mar [cited 2013 Nov 5];87(3): 627-37. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18326601

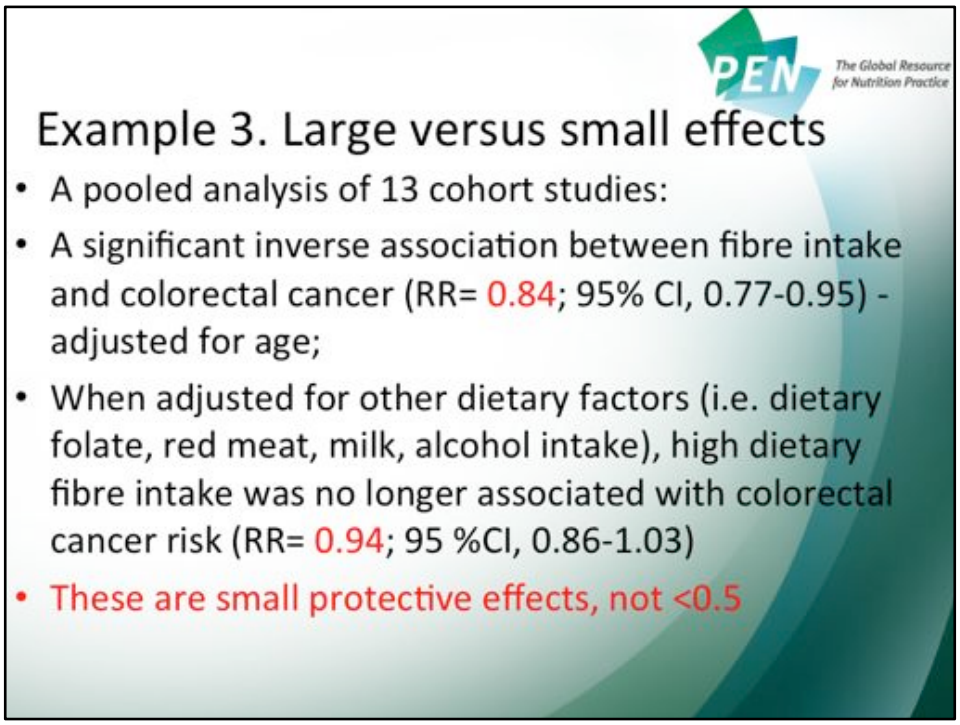


This example shows:
A strong effect with an Odds Ratio of 4.1
This study was quoted in the Guyatt et al. GRADE guidelines: 9. Rating up the quality of evidence. Am J Epidemiol. 2011;64:1311-16. Available from: http://www.ceb-institute.org/fileadmin/upload/refman/J_Clin_Epidemiol_2011_64_1311_Guyatt.pdf

We don't usually see effect sizes as large as 4 in nutrition observational studies. This result was statistically significant, as you can see that the confidence interval does not overlap 1.

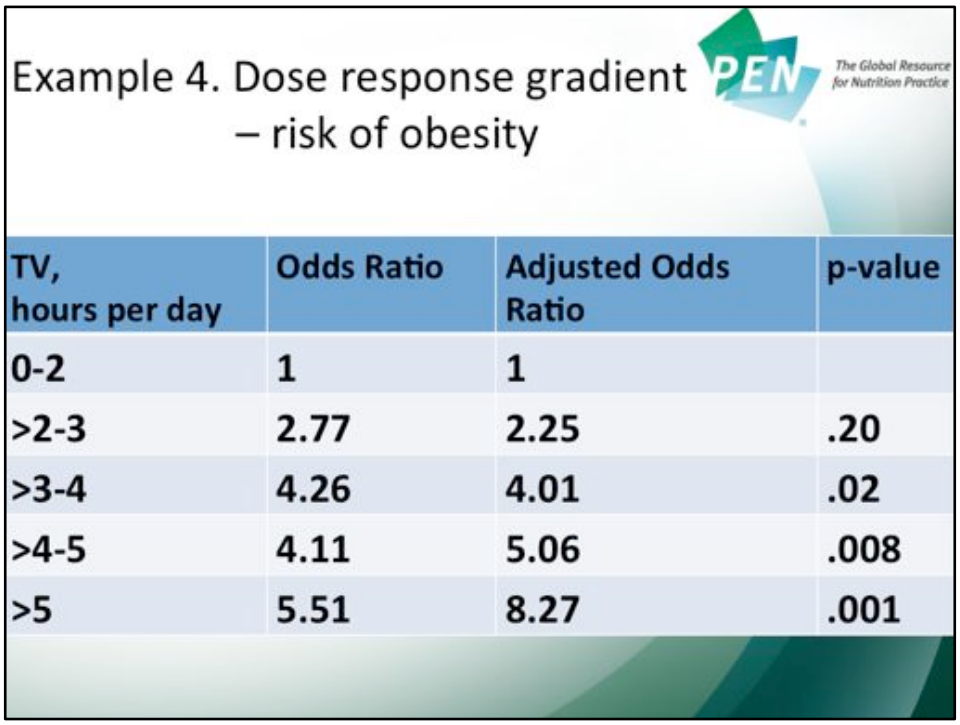
Dr. Spock encouraged front sleeping in case the baby regurgitated, to prevent aspiration. This was reasonable thinking at the time, based on knowledge of babies and biologic plausibility. Dr. Ben Goldacre estimated that Spock contributed to the SIDS death of "tens of thousands of avoidable crib deaths" through this well-meaning advice, since he was considered the source of advice for a generation of parents beginning in 1946. Ref- Bad Science, 2008

(Interesting that Dr. Spock, Pediatrician and Author, encouraged babies put to bed on their tummies, based on biologic knowledge, but has been found wrong through further study)



GRADE suggests observational evidence is supported (i.e. a large protective effect) if the results show a RR<0.5 (or <0.2 for a very large protective effect).
This example shows:
Although the first RR was statistically significant, the "Effect size" or value for the Relative Risk is close to 1, not <0.5 or <0.2, thus it is a very weak estimate of the relationship between fibre and colorectal cancer. Not a large enough "effect size" to be considered strong evidence.
The adjusted Relative Risk is the better estimate of the true relationship than the crude (unadjusted estimate), since it takes into account the other variables that were mixing of effects or "confounding" the relationship.
These are both not strong effects – the correct conclusion for this systematic review is that the evidence does not support a protective effect of fibre on colorecatal cancer. This is the finding even though we might think that fibre is helpful in terms of preventing prolonged contact between intestinal contents and the intestine. This is evidence, and our thinking about the mechanism is not evidence.

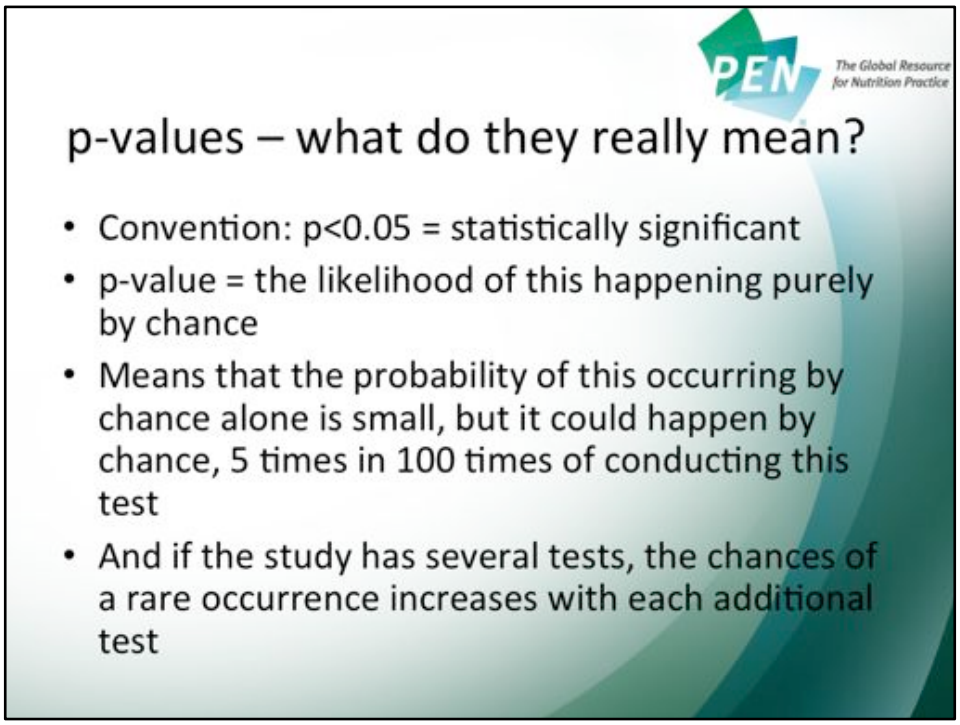
Study reference: Park Y, Hunter DJ, Spiegelman D, Bergkvist L, Berrino F, van den Brandt PA, et al. Dietary fiber intake and risk of colorectal cancer: a pooled analysis of prospective cohort studies. JAMA. 2005 Dec 14 [cited 2013 Nov 5]; 294(22):2849-57. Abstract available from: http://www.ncbi.nlm.nih.gov/pubmed/16352792



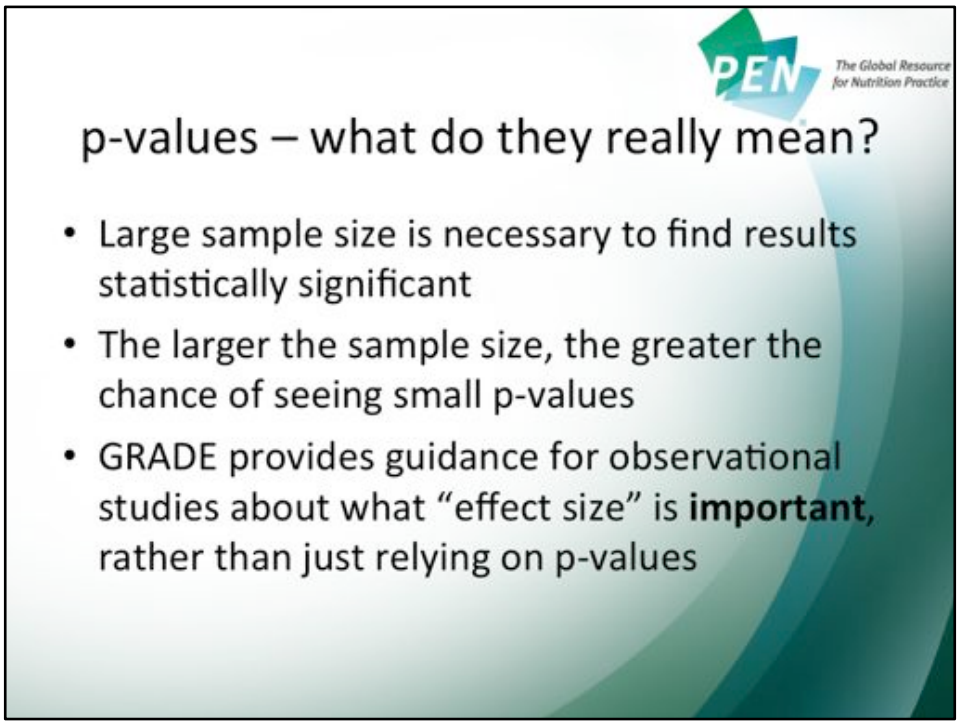
GRADE suggests observational evidence is supported if the results show a large effect when RR or OR >2 or a very large effect when RR or OR >5.
This study shows a strong effect, as the Adjusted Odds ratio was over 2 for all 4 categories, and up to 8 times for more than 5 hours of TV watching per day. This means that children who watch TV 5 or more hours per day had eight times the risk of obesity compared to children who watched less than 2 hours of TV a day.

This study also shows: A Dose-response gradient – another indicator of strength of the association
This dose-response relationship for obesity can be seen in both the hours-per-day of TV, as well as the p-value, as they get stronger with increased TV watching. They both suggest a stronger relationship between obesity and more TV watching. This is a good example of a dose response gradient. It may not be TV watching that caused the obesity, but rather the lifestyle that includes many hours per day of TV watching. This study was published in 1996, and now extended time using computers could also be a risk factor for obesity.

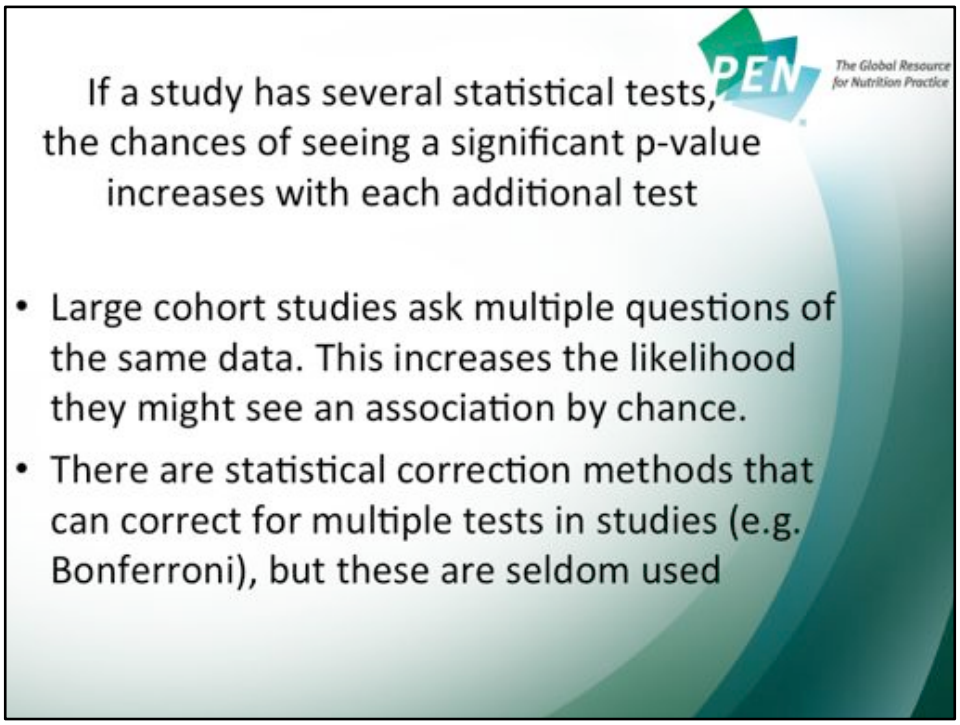
The adjusted Odds Ratio is the better estimate of the true relationship, since it takes into account the other variables that were adjusted for. The reason that the first row has "1"s is that it was the comparison group that the others were compared to. They adjusted for baseline characteristics – exactly what they did was not specified, but baseline characteristics measured in the study included (looks like a good representation): age, gender, maternal education, poverty, maternal working, marital status, # children in family, maternal and child weight status.
Study reference: Gortmaker SL, Must A, Sobol AM, Peterson K, Colditz GA, Dietz WH. Television viewing as a cause of increasing obesity among children in the United States, 1986-1990. Arch Pediatr Adolesc Med. 1996 Apr [cited 2010 Aug 20];150(4):356-62. Abstract available from: http://www.ncbi.nlm.nih.gov/pubmed/8634729


PEN
The Global Resource for Nutrition Practice
p-values – what do they really mean?
• Convention: p<0.05 = statistically significant
• p-value = the likelihood of this happening purely by chance
• Means that the probability of this occurring by chance alone is small, but it could happen by chance, 5 times in 100 times of conducting this test
• And if the study has several tests, the chances of a rare occurrence increases with each additional test

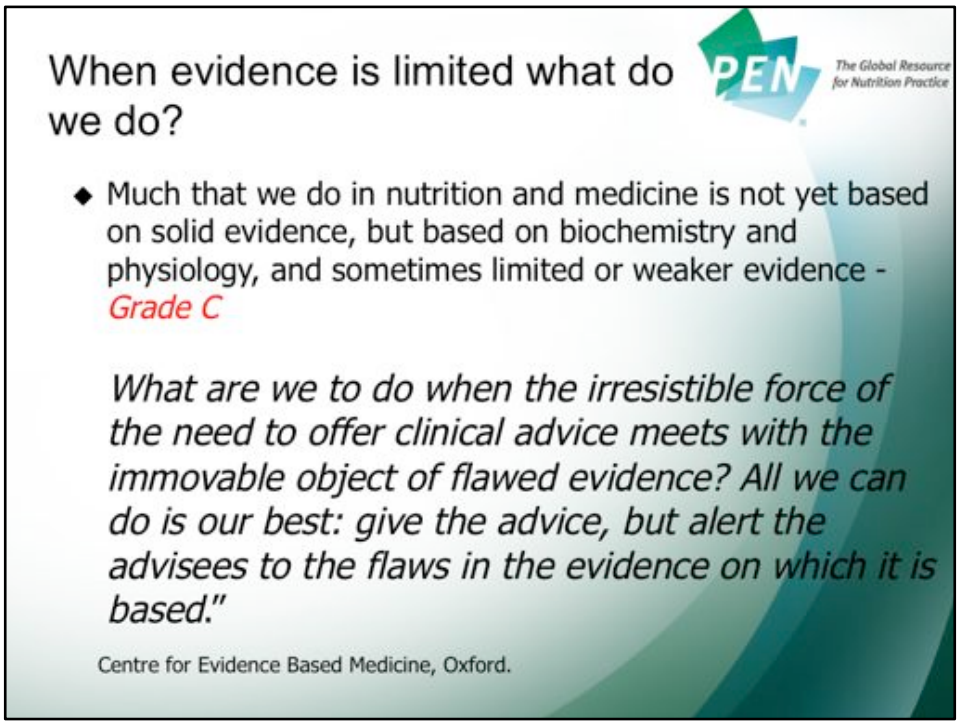
Read slide.



As mentioned in the previous examples, the GRADE Working Group provides guidance on large-effects for observational studies.


If a study has several statistical tests, the chances of seeing a significant p-value increases with each additional test
• Large cohort studies ask multiple questions of the same data. This increases the likelihood they might see an association by chance.
• There are statistical correction methods that can correct for multiple tests in studies (e.g. Bonferroni), but these are seldom used

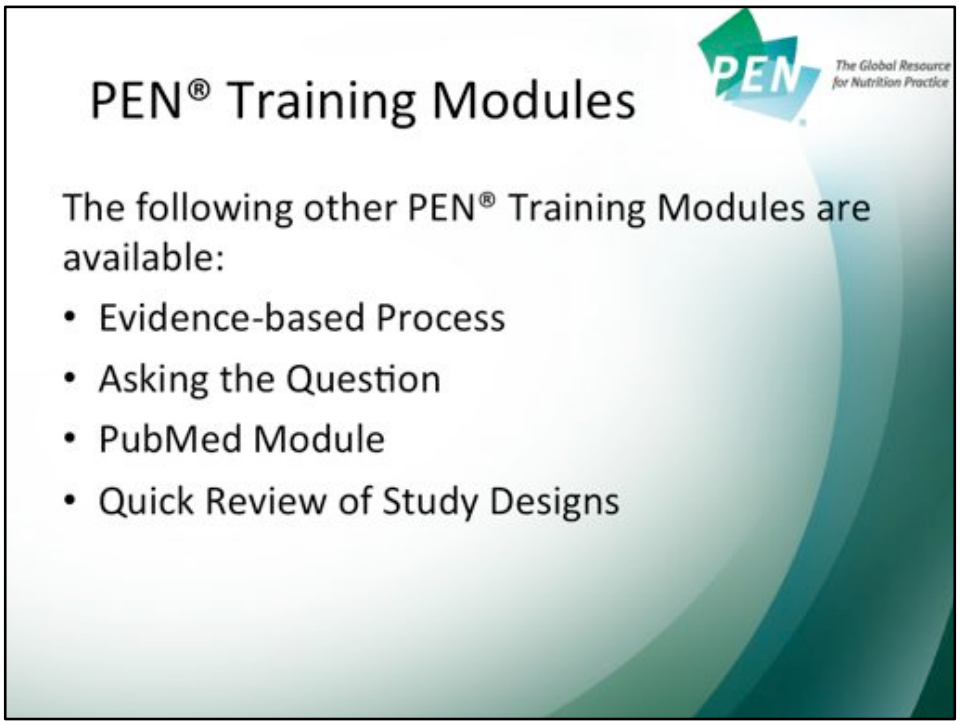
Read slide



Read the slide. In other words, our job is to identify the level of evidence, whether it is high or low.

Although RCTs provide the strongest evidence, since the randomization process, if well conducted, evenly distributes confounding variables or potential prognostic variables between the groups, it is not always possible or ethical to conduct an RCT. One example is to prove whether parachutes improve safety in sky diving. It is unlikely that an Ethics committee would allow one to randomize subjects to a no-parachute group, and unlikely that subjects would agree to be randomized to this group :). When RCTs are not possible or not available, we are left with observational data or research conducted on animals or in test-tubes at best.

Nutrition examples where RCTs could not be conducted include:
Randomization to breastfeeding versus formula feeding
Randomization to a lifetime of vegetarian versus meat eating



The other PEN® Writer training modules can be accessed at: http://www.pennutrition.com/WriterGuide.aspx


PEN
The Global Resource for Nutrition Practice
Thank you for your time in reviewing this module.


PEN
The Global Resource for Nutrition Practice
Come be social with PEN
pennutrition
PENNutrition
PEN Nutrition