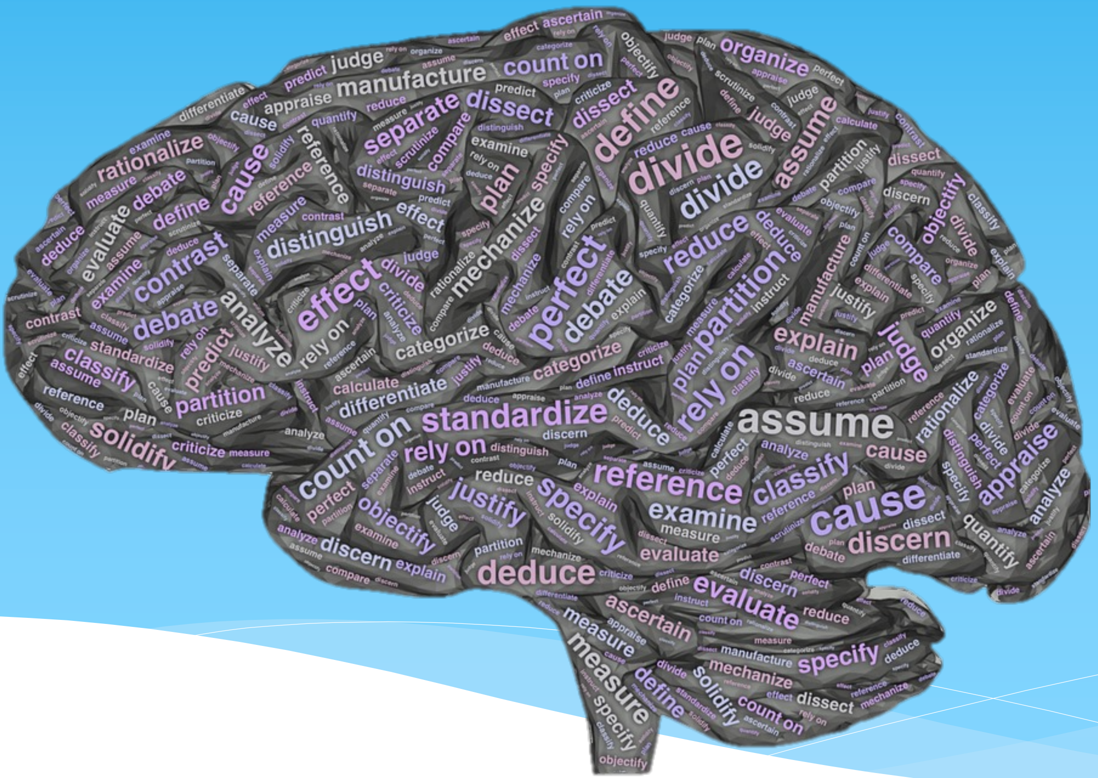
# Appraising the evidence



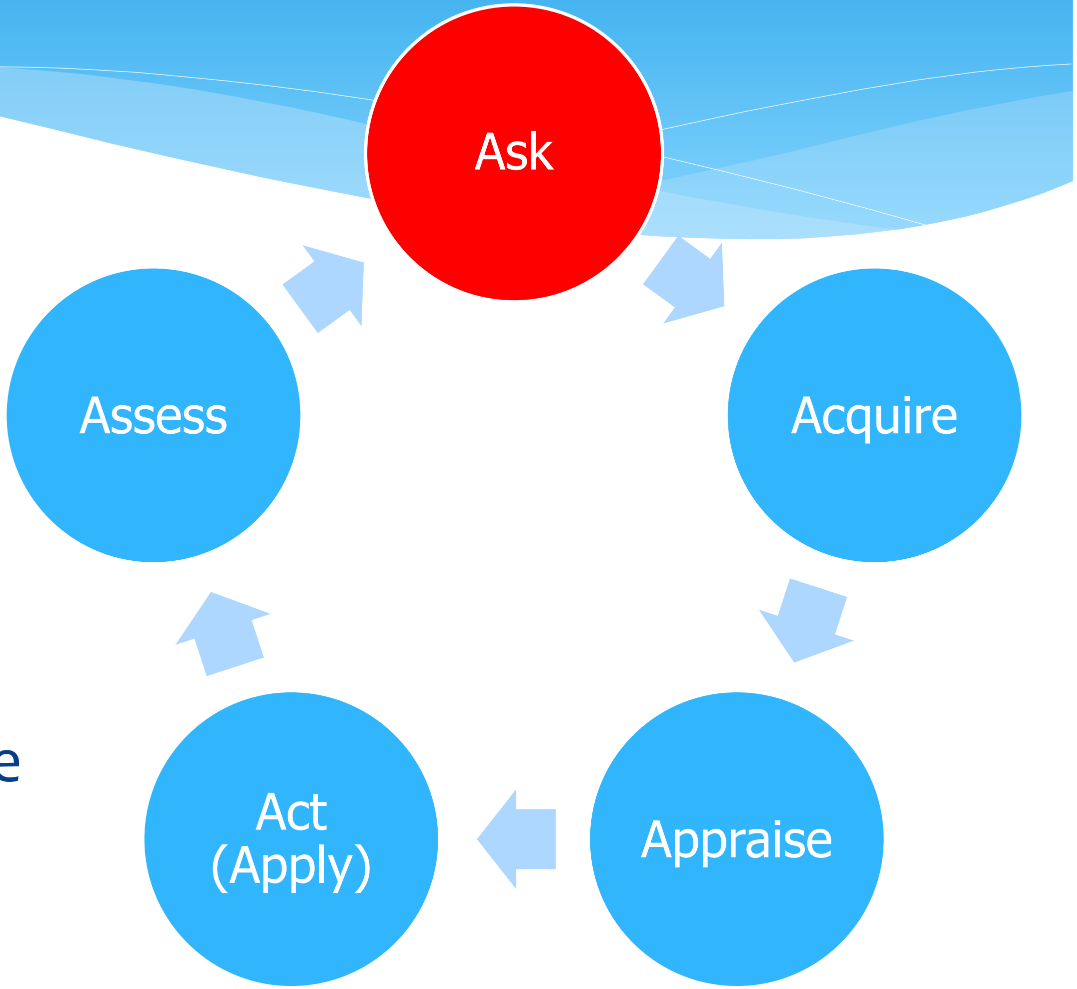
https://pixabay.com/illustrations/brain-dualistic-thought-rational-4014126/

# Objectives

- This presentation aims to present a general explanation of the methods used to appraise scientific evidence.
- At the end of this session, the participants are expected to;
  - Reiterate the 5As of evidence based medicine
  - Discuss internal and external validity of a study
  - Discuss ‘p’ values and significance
  - Discuss confidence interval and its interpretation
  - Present measures of effect sizes
  - Explain the use of the AGREE II instrument

# Steps of Evidence Based Practice: 5As

- Ask the right question
- Acquire the best evidence
- Appraise the evidence critically
- Act or Apply the evidence
- Assess



Users' Guides to Medical Literature. Essentials of Evidence-Based Clinical Practice 3rd Ed. McGraw Hill. 2015.

# Internal vs. External Validity

- Internal: Did the study measure what it said it would?
- External: Are the results applicable to your setting?

# Internal Validity
# Assessing for Risk of Bias

- A bias is a systematic error, or deviation from the truth, in results or inferences.
- Most common biases that need to be addressed are
- Selection Bias
- Allocation Bias
- Performance Bias
- Detection Bias
- Attrition Bias
- Publication Bias

http://methods.cochrane.org/bias/assessing-risk-bias-included-studies

# Significance

- p value
- Tests the null hypothesis
- Measures the probability that the result is due to chance
- A p value of 0.01 that the probability of the result occurring by chance is 1 in 100

# Statistical vs. Clinical Significance

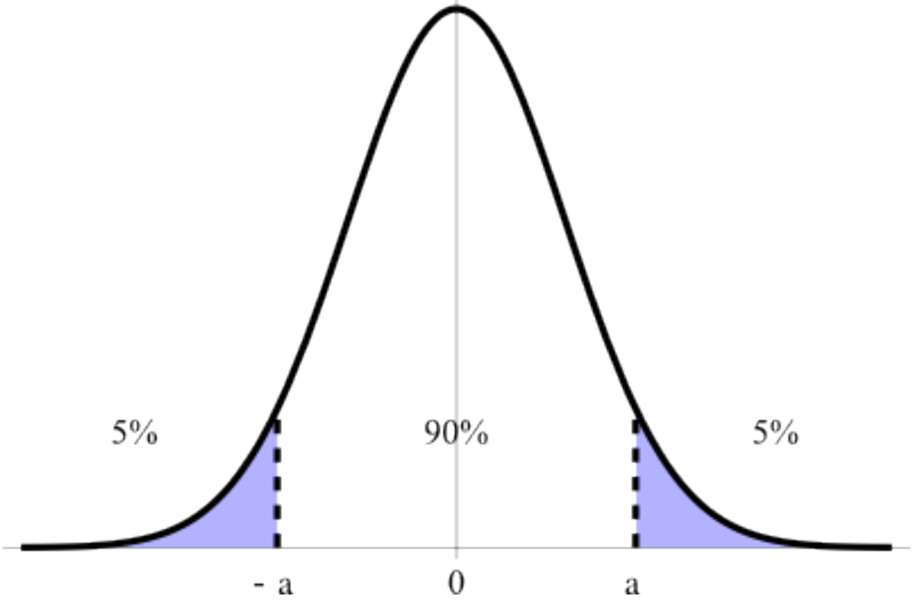
- Even though the results may be statistically significant, they may not be clinically significant.
- In studies with large sample sizes, despite statistical significance, the results may not be clinically significant (high costs, benefits may outweigh risks)
- On the other hand, a lack of statistical significance could be due to inadequate sample size rather the lack of a true effect

# Confidence Intervals

- Range of values that are likely to include the true result.
- 95% CI: If the study was repeated a 100 times, the measured statistical variable would fall within the interval 95 out of 100 times.
- Measure of the precision of the result.
- If the "No effect value" falls within the 95% CI, the result is not statistically significant.

# ‘p’ value and Confidence Intervals

- If your significance level is 0.05, the corresponding confidence level is 95%.
- If the confidence interval does not contain the null hypothesis value, the results are statistically significant.



https://commons.wikimedia.org/wiki/File:Confidence_Interval_90P.png

# 95%CI of difference

**A systematic review and dose-response meta-analysis on the efficacy of dapagliflozin in patients with type 1 diabetes mellitus**

- "The pooled results suggested a significant reduction in glycated hemoglobin A1C (HbA1C; WMD: -0.36%, 95% CI: -0.55, -0.18), body weight (WMD: -4.02 kg, 95% CI: -4.78, -3.25), and total daily insulin dose (TDID; WMD: -10.36%, 95% CI: -13.42, -7.29), as well as an increase in 24 -h urinary glucose excretion (24 -h UGE; WMD: 90.02 g/24 -h, 95% CI: 72.96, 107.09) in dapagliflozin group compared to control group. Dose of dapagliflozin had a significant effect on body weight reduction (Coef=-3.7, p = 0.01) and 24 -h UGE (coef = 0.85, p = 0.005)"

https://pubmed.ncbi.nlm.nih.gov/33515709/

# 95%CI of OR

## Risk Factors for Albuminuria in Normotensive Older Adults with Type 2 Diabetes Mellitus and Normal Renal Function: A Cross-Sectional Study

- **Results:** A total of 250 older adults were enrolled during the study period, including 124, 82, and 44 with normal albuminuria, microalbuminuria, and macroalbuminuria, respectively. We found that an extended duration of DM (odds ratio [OR] 1.085, 95% confidence interval [CI] 1.012-1.164, P = 0.022), elevated systolic blood pressure (OR 1.049, 95%CI 1.018-1.081, P < 0.01), elevated glycated hemoglobin (OR 1.734, 95% CI 1.332-2.258, P < 0.01), low insulin (OR 0.871, 95% CI 0.804-0.944, P < 0.01), and low C-peptide (OR 0.365, 95% CI 0.239-0.588, P < 0.01) were independent risk factors for albuminuria.

https://pubmed.ncbi.nlm.nih.gov/33515179/

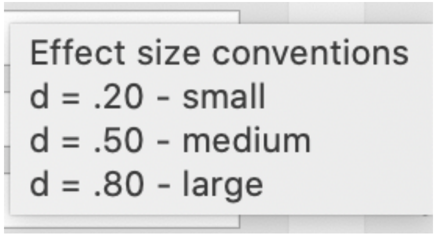
# Effect Size

- The simple definition of effect size is the magnitude, or size, of an effect
- ‘p’ value says if 2 groups are different or not
- Effect size helps us understand the magnitude of differences found, whereas statistical significance examines whether the findings are likely to be due to chance

Sullivan 2012. Journal of Graduate Medical Education

# Effect Size Indices

- Effect sizes between groups
- Odds ratio
- Relative risk
- Cohen’s d = mean difference/standard deviation
- Association indices between variables
- Correlation r
- $R^2$ co-efficient of determination



Effect size conventions
d = .20 - small
d = .50 - medium
d = .80 - large

# Treatment Effect: Odds Ratio

- Commonly used in cross-sectional and case-control studies
- Odds that an outcome will occur with an exposure compared to odds of the outcome occurring in the absence of the exposure/intervention

- OR=1
  - There is no difference in odds that an outcome will occur with an exposure compared to odds of the outcome occurring in the absence of the exposure/intervention
- OR<1
  - The odds are lower that an outcome will occur with an exposure compared to odds of the outcome occurring in the absence of the exposure/intervention
- OR>1
  - The odds are greater that an outcome will occur with an exposure compared to odds of the outcome occurring in the absence of the exposure/intervention

# Treatment Effect: Relative Risk

- Relative Risk= Experimental Event Rate/Control Event Rate
- Relative risk reduction: 1-RR
  - 0.46=RR, 1-RR=0.54, 0.54*100=54%
  - (How would you state this in a sentence?)

https://www.medcalc.org/manual/relativerisk_oddsratio.php

# Treatment Effect: Number Needed to Treat

- Number needed to treat (NNT) = 1/Absolute risk reduction
- NNT: The number of patients that need to be treated to achieve one therapeutic success
- Low NNT=Stronger therapeutic effect
- When an undesirable outcome is being evaluated the term used in number needed to harm (NNH).

# NNT

- ARF in Germany: 1/1 000 000 (0.0001%) per year.
- 30-80% of ARF is expected to cause rheumatic heart disease
- Absolute Risk (Expected RHD): 0.0001*0.80=0.00008%
- Penicillin treatment is expected to reduce the relative risk of ARF by 63%: 0.00008*0.63=0.0000504% decrease
- ARR=0.00008%-0.0000504%=0.0000296%
- NNT=1/ARR=1/0.0000296= **33 784** treatments to prevent 1 RHD
- Fatal anaphylaxis risk: 0.0015%
- Net benefit: 0.0015%*33784=0.5 deaths to prevent 1 RHD (**or 1 deaths to prevent 2 RHD**)

https://www.degam.de/files/Inhalte/Leitlinien-Inhalte/Dokumente/DEGAM-S3-Leitlinien/053-010_Halsschmerzen/053-010l_Halsschmerzen_211220.pdf p:60

# Null value

- The “no effect” (null) value for ratios (odds ratio and RR) is 1 and for risk difference is 0.
- 95%CI is often used as a proxy for the presence of statistical significance.
- If the 95%CI does not overlap the null value it implies statistical significance
- For e.g.: If the 95%CI excludes 1 for relative risk and 0 for absolute risk difference, then the result is considered statistically significant

# Guideline Evaluation Resources

- A Provisional Instrument for Assessing Clinical Practice Guidelines. (Institute of Medicine, 1992)
- Methodological Standards. (Shaneyfelt et al., 1999)
- Grading of Recommendations Assessment, Development and Evaluation (**GRADE**). (Brozek et al., 2009)

- Appraisal of Guidelines for Research & Evaluation (AGREE II). (Brouwers et al., 2010)
- Guidelines International Network Standards for Clinical Practice Guidelines. (Qaseem et al., 2012)
- CEP Trustworthy Guideline Appraisal Scale. (Center for Evidence-based Practice, University of Pennsylvania, 2016)

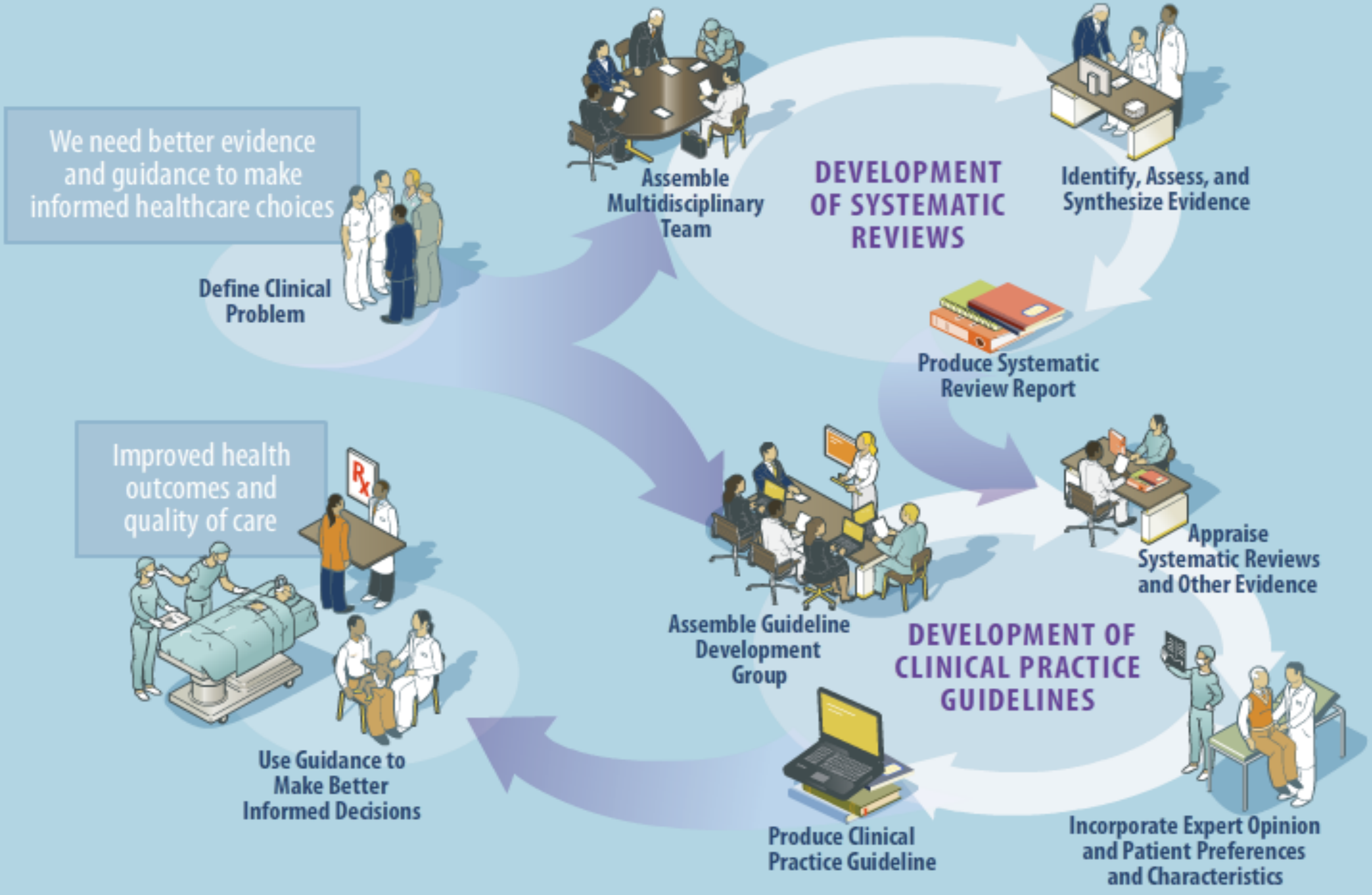
# Evaluating Guideline Development

- In 2008, the U.S. Congress asked the Institute of Medicine (IOM) to study best methods used in developing clinical practice guidelines.
- Expert committee established 8 standards for developing rigorous, trustworthy clinical practice guidelines.

https://www.ncbi.nlm.nih.gov/books/NBK209539/pdf/Bookshelf_NBK209539.pdf
"Clinical Practice Guidelines We Can Trust" (IOM, 2011).

# Systematic Reviews and Clinical Practice Guidelines Improve Healthcare Decision Making

Click on any text for more information



We need better evidence and guidance to make informed healthcare choices

Define Clinical Problem

Assemble Multidisciplinary Team

DEVELOPMENT OF SYSTEMATIC REVIEWS

Identify, Assess, and Synthesize Evidence

Produce Systematic Review Report

Appraise Systematic Reviews and Other Evidence

Incorporate Expert Opinion and Patient Preferences and Characteristics

DEVELOPMENT OF CLINICAL PRACTICE GUIDELINES

Assemble Guideline Development Group

Produce Clinical Practice Guideline

Use Guidance to Make Better Informed Decisions

Improved health outcomes and quality of care

SHARE:

INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

# Benefits of Guideline Development Standards

- Provide guideline developers with objective, standardized criteria to enhance quality of product.
- Trustworthy guidelines will enhance health care quality and outcomes.
- Potential users of guidelines can use IOM development standards to evaluate trustworthy practice guidelines.

# Standards for Developing Trustworthy Practice Guidelines

- Establishing transparency.
- Management of conflict of interest.
- Guideline development group composition.
- Clinical practice guideline-systematic review intersection.

(IOM, 2011)

# STANDARDS FOR DEVELOPING TRUSTWORTHY CLINICAL PRACTICE GUIDELINES (CPGS)

- “Clinical Practice Guidelines We Can Trust” (IOM, 2011).
- https://www.ncbi.nlm.nih.gov/books/NBK209539/pdf/Bookshelf_NBK209539.pdf

# 1. Establishing Transparency

- The processes by which a CPG is developed and funded should be detailed explicitly and publicly accessible.

# 2. Management of Conflict of Interest

- Prior to selection of the guideline development group (GDG), individuals being considered for membership should declare all interests and activities potentially resulting in COI with development group activity, by written disclosure to those convening the GDG

# 3. Guideline Development Group Composition

- The GDG should be multidisciplinary and balanced, comprising a variety of methodological experts and clinicians, and populations expected to be affected by the CPG.

# 4. Clinical Practice Guideline–Systematic Review Intersection

- Clinical practice guideline developers should use systematic reviews that meet standards set by the Institute of Medicine's Committee on Standards for Systematic Reviews of Comparative Effectiveness Research.

# 5. Establishing Evidence Foundations for and Rating Strength of Recommendations

- An explanation of the reasoning underlying the recommendation,
- A rating of the level of confidence in (certainty regarding) the evidence underpinning the recommendation
- A rating of the strength of the recommendation
- A description and explanation of any differences of opinion regarding the recommendation

# 6. Articulation of Recommendations

- 6.1 Recommendations should be articulated in a standardized form detailing precisely what the recommended action is, and under what circumstances it should be performed.
- 6.2 Strong recommendations should be worded so that com-pliance with the recommendation(s) can be evaluated.

# 7. External Review

- External reviewers should comprise a full spectrum of relevant stakeholders, including scientific and clinical ex-perts, organizations (e.g., health care, specialty societies), agencies (e.g., federal government), patients, and repre-sentatives of the public.

# 8. Updating

- The CPG publication date, date of pertinent systematic evidence review, and proposed date for future CPG review should be documented in the CPG.
- Literature should be monitored regularly following CPG publication to identify the emergence of new, potentially relevant evidence and to evaluate the continued validity of the CPG.

# The AGREE II instrument

- Assesses quality of guidelines.
- Assesses methodological rigor and transparency of development.
  - Does not assess validity of recommendations.
- Provides framework to understand what information should be reported and how information should be reported in guidelines.

https://www.agreetrust.org/wp-content/uploads/2013/10/AGREE-II-Users-Manual-and-23-item-Instrument_2009_UPDATE_2013.pdf

# AGREE II

- Six Domains (23 items):
  - Scope and Purpose
  - Stakeholder Involvement
  - Rigor of Development
  - Clarity of Presentation
  - Applicability
  - Editorial Independence

- Overall Assessment
  - 2 global rating items
    - Quality of guideline.
    - Recommend guideline for use in practice.

- Scope and Purpose (3 items)
  - Overall aim, specific health questions, target population.
- Stakeholder Involvement (3 items)
  - Inclusion of appropriate stakeholders, views of intended users.
- Rigor of Development (8 items)
  - Gathering, synthesis of evidence, methods used to develop recommendations and update them.

- Clarity of Presentation (3 items)
  - Format, structure, language used.
- Applicability (4 items)
  - Implications, barriers, facilitators to implementation.
- Editorial Independence (2 items)
  - Transparency of bias or conflict of interest.

- Tool provides detailed assessment criteria and specific considerations for each of the 23 items.
- Items rated on 7-point scale
  - Score of 1 (strongly disagree) = no information or concept very poorly reported.
  - Score of 7 (strongly agree) = excellent quality of reporting and full criteria for item are met.
- Authors emphasize that rating guidelines requires judgment.

# Scoring the Agree II Instrument

- Sum up all item scores within a domain.
  - Determine percentage of maximum and minimum possible scores for that domain.
    - Max score = 7 x 3 items x 4 appraisers = 84
    - Min score = 1 x 3 items x 4 appraisers = 12

$$Agree\ Score = \frac{(Obtained\ score - Min.possible\ score)}{(Max.possible\ score\ - Min.possible\ score)}$$

# Scoring the Agree II Instrument

| | Item 1 | Item 2 | Item 3 | Total |
|---|---|---|---|---|
| Appraiser 1 | 5 | 5 | 6 | 17 |
| Appraiser 2 | 6 | 6 | 7 | 19 |
| Appraiser 3 | 2 | 4 | 3 | 9 |
| Appraiser 4 | 3 | 3 | 2 | 8 |
| Total | 16 | 19 | 18 | 53 |

$$Agree\ Score = \frac{(53-12)}{(84-12)} = \frac{41}{72} = 0.569 \cong 57\%$$

Brouwers et al., 2013, p.12

# Interpreting AGREE II Domain Scores

- Domain scores enable comparisons across guidelines.
  - Scores help determine if guideline should be recommended for use in practice.
- The authors do not provide minimum domain scores or % across domains to indicate high versus poor quality.
  - Clinical judgment required.

# Interpreting AGREE II Domain Scores

- Overall assessment – 2 global items.
- Appraisers judge quality of guideline considering criteria met/not met in assessment process.
- Appraisers determine whether s/he recommends use of the guideline.

# What after having guidelines?

- Structures
  - Are evidence based policies/procedures in place – based on selected guidelines?
  - Do clinicians have the “tools” to provide recommended care?
- Processes
  - To what extent are providers complying with policies?
- Outcomes
  - Safety?
  - Cost?
  - Patient satisfaction?

# Legal Implications of Applying Practice Guidelines

- Concerns if selected guideline recommendations considered to be “standard of care” lead to adverse outcomes.
- Clinical practice guidelines enable evidence-based care.
- Providing care based on critically evaluated, evidence-based guidelines promotes a “new standard of practice” versus “this is how we do it and how we’ve always done it”.
- Ensure compliance!

# Summary

- Reiterate the 5As of evidence based medicine
- Discuss internal and external validity of a study
- Discuss ‘p’ values and significance
- Discuss confidence interval and its interpretation
- Present measures of effect sizes
- Explain the use of the AGREE II instrument