The GATE Notes: a Generic Appraisal Tool for Epidemiology

THE GATE NOTES:

- These guides incorporate most of the questions from the JAMA series of "Users' Guides to the Medical Literature" (1-9), but they have been rearranged to more systematically link design and appraisal. It is recommended that you use the JAMA guides or an updated version of the key JAMA guides such as the EBM handbook by Sackett et al (10), as reading to complement the GATE guides.
- Each section of GATE starts with a brief explanation of the study type, then a checklist (2 pages) and a User guide for each checklist. It is useful to fill out as you go through a full set of appraisal questions.
- When you have completed each Section of the checklist, it is important to weigh up the overall quality of that aspect of the study.

AN OVERVIEW OF THE SUBTYPES OF EPIDEMIOLOGICAL STUDIES:

Epidemiological studies can be differentiated into major subtypes based on how the exposure and comparison subgroups in the study population are assigned (i.e. experimentally or non-experimentally) and based on the types of occurrence measures used (i.e. prevalence in cross-sectional studies and incidence in longitudinal studies). Some study designs are modifications of these major subtypes. Each study subtype can be derived using the Generic Appraisal Tool for Epidemiology (GATE) approach based on the 5 part PECOT diagram. A brief overview of each study subtype is given below.

There are 5 types of studies that are used in guideline development.

1. Randomised controlled trials. (RCTs): This is an experimental study where participants are randomly allocated to exposure(s) or comparison intervention (sometimes a placebo). Outcomes are typically measured over a period of time in RCTs, therefore most RCTs are longitudinal studies measuring incidence, however outcomes can also be measured cross-sectionally (i.e. prevalence measures) in RCTs . *Screening studies* investigate the effect of a screening test on a health-related outcome and should ideally be RCTs in which the test allocation is randomly allocated) but are sometimes cohort studies (see below) if the use of the test is ascertained rather than allocated by the investigator).

2. Cohort studies: If participants are assigned to exposure(s) and comparison groups based on the MEASUREMENT of these factors (rather than being randomly allocated), the study is non-experimental. These studies are often called observational studies, although outcomes are observed in all studies, both experimental and non-experimental, so the

term “non-experimental” is more appropriate than “observational”. Cohort studies can be considered as non-experimental versions of RCTs in which the exposure and comparison groups assignment is determined by measurement of these factors in the study participants and outcomes are measured over a follow-up period. Cohort studies are non-experimental longitudinal studies .

3. Case-control studies (non experimental) are “nested” inside cohorts and can be considered as efficient versions of cohort studies (not included in these notes)

4. Prognostic studies (non experimental) are cohort studies in which the objective is to investigate how well an exposure(s) predicts the occurrence of outcomes rather than whether or not the association is causal.

5. Cross-sectional studies (non experimental) are similar in design to cohort studies, except that outcomes are measured at one point in time; usually at the same time as the study population exposure and comparison groups are defined . *Diagnostic test studies* are cross-sectional studies that compare the accuracy of a diagnostic test with a reference standard.

1. RANDOMISED CONTROLLED TRIALS (Treatment studies)
(Relevant JAMA User's Guides, Numbers IIA & B: references (3,4)
Introduction:

The most valid study design for assessing the effectiveness (both the benefits and harms) of therapeutic or preventive interventions is the randomised controlled trial (RCT). This is an experiment in which the investigator controls the random allocation of participants or study communities in the study population to the interventions of interest (i.e. exposure or intervention subgroup/s) or a comparison subgroup (i.e. the control group).

Trials are considered the "purest" type of epidemiological study because the investigator has control over exposure allocation. If the investigator randomises individual participants or communities to intervention and comparison subgroups, it is possible to minimise differences in baseline characteristics between the groups that might influence the outcome of interest (i.e. it minimises confounding).

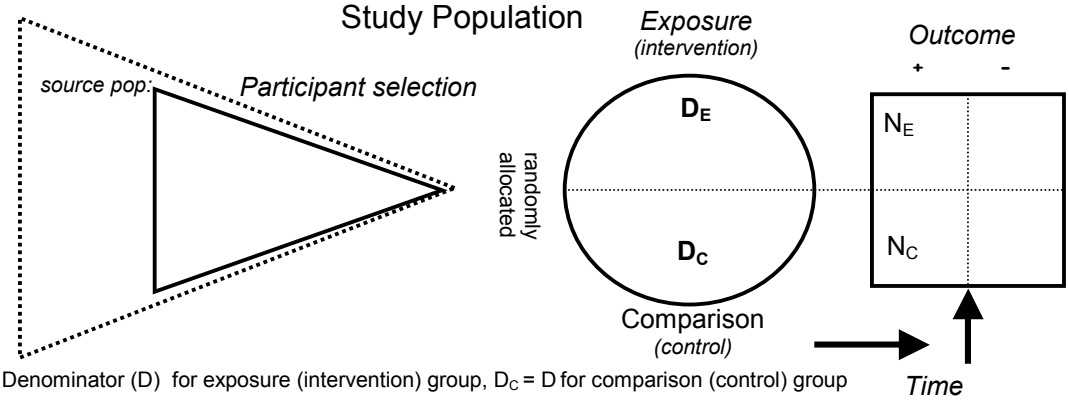
The comparison or control group may be allocated a placebo intervention, an alternative real intervention or no intervention at all.

If randomisation is successful and the groups are similar at baseline, the investigator can be more confident that observed differences in outcomes between the groups are related to the intervention rather than confounding factors.

Trials have a number of potential limitations compared with other designs. For practical and ethical reasons some important questions cannot be investigated using an experimental design. Moreover when trials are possible, they are often conducted in artificial environments and with highly motivated volunteers. This may limit their generalisability to populations of interest.

GATE Checklist for Randomised Controlled Trials (Intervention: benefit or harm)

Study author, title, publication reference	*Key 5 part study question (PECOT). Was it focussed?*



D_E = Denominator (D) for exposure (intervention) group, D_C = D for comparison (control) group

N_E = Numerator (N) for exposure group, N_C = N for comparison group

SECTION 1: STUDY VALIDITY		Appraised by:	
Evaluation criterion		*How well was this criterion addressed?*	Quality ✓ ? x
Participants	What were the key selection (inclusion & exclusion) criteria? Were they well defined? Were they replicable?		
	Were inclusion & exclusion criteria appropriate given study question?		
Exposures & Comparison	What were the exposures (interventions) & comparison? Well defined? Replicable?		
	Was assignment to exposure & comparison groups randomised?		
	Was randomisation concealed?		
	Was randomisation successful: were exposure & comparison groups similar at start of study?		
	Were all participants analysed in groups to which randomised?		
	Were participants, health workers, researchers blind to interventions?		
	Apart from study interventions, were groups treated equally?		
	Was compliance with interventions measured? Was it sufficient?		
Outcomes	What outcome measures were used? Well defined? Replicable?		
	How complete was follow up? Was it sufficient? How many dropouts?		
	Was outcome assessment blind?		
Time	Was follow up time sufficiently long to detect important effects on outcomes of interest?		
QUALITY OF STUDY DESIGN: *How successfully do you think the study minimised bias? Very well = +, okay = ∅, poorly = -*			

SECTION 2: STUDY RESULTS: MAGNITUDE & PRECISION OF EFFECTS	
What measures of occurrence (incidence / prevalence) & intervention effects (RR /RD /NNTs) were reported?	
What measures of precision of effects were reported (CIs, p-values)?	

THE NUMBERS TABLE: OCCURRENCE, EFFECT ESTIMATES & PRECISION

Outcomes* & Time (T)	**Exposure event rate ($EER=N_E/D_E/T$) or mean***	**Comparison event rate ($CER=N_C/D_C/T$) or mean***	**Relative Risk* (RR = EER/CER) ± (95% CI)**	**Risk difference or mean difference (RD = CER-EER) ± (95% CI)**	**Number Needed to Treat* (NNT = 1/RD) ± (95% CI)**

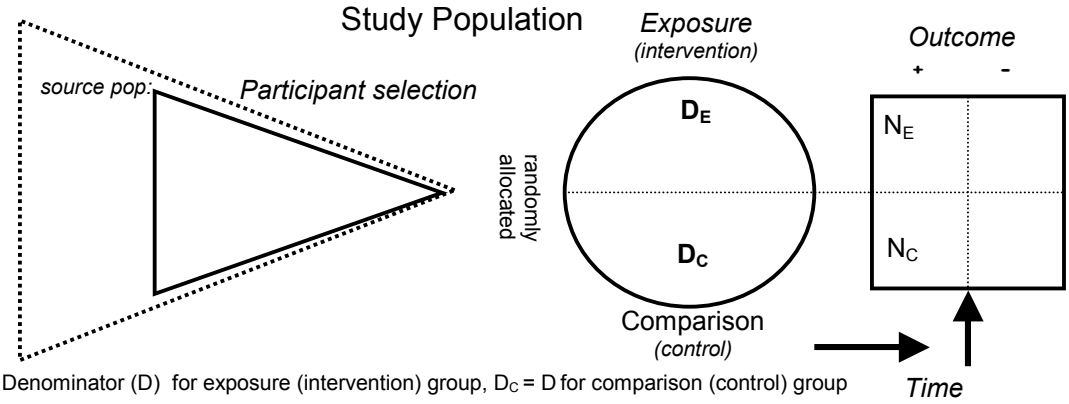
** if outcomes continuous, can calculate means, mean differences, but not NNTs (don't usually calculate relative means)*
D_E = Denominator (D) for exposure (intervention) group(s), D_C = D for comparison (control) group
N_E = Numerator (N) for exposure group(s), N_C = N for comparison group

	Quality ✓ ? x
Could useful effect estimates (e.g. RR, RDs or mean differences, NNTs) be calculated? For benefits & harm?	
What was the magnitude and direction of the effect estimates?	
Was the precision of the effect estimates sufficient?	
If no statistically significant effects detected, was there sufficient power?	
If multi-centred RCT - were the results homogeneous between sites?	
QUALITY OF STUDY RESULTS: *Useful, precise +/or sufficient power? Very good = +, okay = ∅, poor = -*	

SECTION 3: STUDY APPLICABILITY			
Participants	Was the source population for participants well described?		
	Were participants representative of source population?		
	Can the relevance / similarity of the participants to a specific target group(s) be determined?		
Exposures & Comparison	Were the characteristics of the study setting well described? *e.g. rural, urban, inpatient, primary care*		
	Can the applicability of interventions be determined?		
	Can the relevance of the comparison group management be determined?		
Outcomes	Were all important outcomes considered: benefits? harms? costs?		
	Are likely benefits greater than potential harms & costs (or vice versa)? In what target group(s)?		
QUALITY OF STUDY APPLICABILITY: (a) Was it possible to determine applicability? *Very well = +, okay = ∅, poorly = - (b) Are findings applicable in your practice/setting? Very well = +, okay = ∅, poorly = -*			

USERS GUIDE for GATE Checklist for Randomised Controlled Trials

Study author, title, publication reference	*Key 5 part study question (PECOT). Was it focussed?*



D_E = Denominator (D) for exposure (intervention) group, D_C = D for comparison (control) group

N_E = Numerator (N) for exposure group, N_C = N for comparison group

SECTION 1: STUDY VALIDITY		Appraised by:	
Evaluation criterion		*How well was this criterion addressed?*	Quality ✓ ? x
Participants	What were the key selection (inclusion & exclusion) criteria? Were they well defined? Were they replicable?	List important selection criteria; e.g. age group, gender, risk profile, medical history. Usually in Methods section. There should be sufficient information in the paper (or referenced) to allow the reader to theoretically select a similar population	
	Were inclusion & exclusion criteria appropriate given study question?	Are the participants a relevant group to apply the study intervention to? (e.g. diagnostic tests are not very helpful in people with a very high probability of disease).	
Exposures & Comparison	What were the exposures (interventions) & comparison? Well defined? Replicable?	Examples include: dosage of drugs, description of surgical procedure, number of faecal occult blood tests in a screening study, management in comparison group (e.g. check what care the comparison (placebo group) receive).	
	Was assignment to exposure & comparison groups randomised?	Random allocation of interventions is fundamental to this type of study. The method of randomisation should be described (e.g. using a computer algorithm). If the description of the randomisation method is poor, or the process used is not truly random (e.g. allocation by date, alternating between one group and another) or can otherwise be seen as flawed, the study should be given a lower quality rating.	
	Was randomisation concealed?	This is an important quality issue as it has been shown that RCTs that failure to conceal allocation typically lead to an overestimate of the treatment effect. The ideal concealment process would involve an independent group that registered each new participant, determined the allocation and then informed the caregivers of the allocation. This can be done relatively simply by phone or fax or by using an automatic web-based system. This reduces the chance of the care giver influencing allocation.	

	Was randomisation successful: were exposure & comparison groups similar at start of study?	This can be judged by examining the similarity between baseline characteristics of the groups in each arm of the study. Successful randomisation will produce similar groups. The study should report significant differences in the composition of the study groups in relation to gender mix, age, stage of disease (if appropriate), social background, ethnic origin, or comorbid conditions. Of note, statistically significant differences are not always important in practice; consider if the differences are likely to be meaningful given the study questions.	
	Were all participants analysed in groups to which randomised?	It is rarely that all participants allocated to the intervention group receive the intervention throughout the trial, or that all those in the comparison group do not. Participants may refuse treatment, or contra-indications may arise. If the comparability of groups through randomisation is to be maintained, however, patient outcomes *should* be analysed according to the group to which they were originally allocated irrespective of the treatment they actually received. (This is known as *intention to treat* analysis.) If analysis were not on an intention to treat basis, the study validity could be compromised. Bias can be quantified by attempting both "on-treatment" and "intention-to-treat" analyses.	
	Were participants, health workers, researchers blind to interventions?	Blinding can be carried out on up to three levels. Single blinding is where participants are unaware of which intervention they are receiving; in double blind studies neither the care giver nor the patient know which intervention is being given; in triple blind studies neither patients, care givers, nor those conducting the analysis are aware of which participants received which intervention. Most studies described as double blind studies are usually triple blind. The higher the level of blinding, the lower the risk of bias in the study. Although the nature of the intervention may unblind the allocation in some cases (eg. Surgical trials) the allocation to intervention & comparison groups should be blind.	
	Apart from study interventions, were groups treated equally?	If some participants received additional interventions, even if of a minor nature or consisting of advice and counselling rather than a physical intervention, this can introduce confounding and may invalidate the results. If there is unequal intervention (apart from the study intervention) the study results should be interpreted with caution and given a low quality rating.	
	Was compliance with interventions measured? Was it sufficient?	Compliance is often a problem in studies involving ongoing interventions such as daily medication or behaviour change. Pill counts and blood levels of drugs are examples of objective methods of measuring compliance, although self-reports are more common but less reliable.	
Outcomes	What outcome measures were used? Well defined? Replicable?	Criteria for assessing outcomes such as diagnostic algorithms should be well described or referenced. It should be theoretically possible for the reader to replicate the process.	
	How complete was the follow up? How many dropouts were there?	The number of participants who drop out of a study is a concern if the number is very high. Conventionally, a 20% drop out rate is regarded as acceptable, but this depends on the study question. Some regard should be paid to *why* participants dropped out, as well as how many. It should be noted that the drop out rate may be expected to be higher in studies conducted over a long period of time. A higher drop out rate will normally lead to downgrading, rather than rejection of a study.	
	Was outcome assessment blind?	Ideally the assessors who measure & record outcomes should be blind to participant allocation. This is more important for assessing outcomes that are not clear cut & where knowledge of the intervention may influence the diagnostic assessment.	
Time	Was follow up time sufficiently long to detect important effects on outcomes of interest?	This is specific to the study intervention and outcomes assessed	
QUALITY OF STUDY DESIGN: How successfully do you think the study minimised bias? Very well = +, okay = ∅, poorly = -			

SECTION 2: STUDY RESULTS: MAGNITUDE & PRECISION OF EFFECTS	
What measures of occurrence (incidence / prevalence) & intervention effects (RR /RD /NNTs) were reported?	Some studies do not provide the relevant number of participants (D) in the exposure and comparison groups, the number of outcomes (N), the event rates / proportions with outcomes (N/D) in each study group, or the relevant measures of effect (RR, etc). If they are not reported or cannot be calculated, it is not possible to ascertain the accuracy of the effect estimates such as relative risk (RR), risk difference (RD) or mean differences (if continuous measures of outcome are given) and numbers needed to treat (NNT) – see definitions below in the Numbers Table below.
What measures of precision of effects were reported (CIs, p-values)?	Either confidence intervals or p values for the estimates of effect should be reported or be possible to calculate.

THE NUMBERS TABLE: OCCURRENCE, EFFECT ESTIMATES & PRECISION

Outcomes* & Time (T)	**Exposure event rate ($EER=N_E/D_E/T$) or mean***	**Comparison event rate ($CER=N_C/D_C/T$) or mean***	**Relative Risk* (RR = EER/CER) ± (95% CI)**	**Risk difference or mean difference (RD = CER-EER) ± (95% CI)**	**Number Needed to Treat* (NNT = 1/RD) ± (95% CI)**
complete	complete	complete	complete	complete	complete

* *if outcomes continuous, can calculate means, mean differences, but not NNTs (don't usually calculate relative means)*
D_E = Denominator (D) for exposure (intervention) group(s), D_C = D for comparison (control) group
N_E = Numerator (N) for exposure group(s), N_C = N for comparison group

		Quality ✓ ? x
Could useful effect estimates (e.g. RR, RDs or mean differences, NNTs) be calculated? For benefits & harm?	These numbers should be reported or able to be calculated in the Numbers Table (above). To be useful, they need to have some meaning in practice. For example a change of one point on a visual analogue scale of symptoms may have little meaning unless clearly linked to a symptom description.	
What was the magnitude and direction of the effect estimates?(RR, RD, mean differences, NNTs)	These numbers are the bottom line of every study. All other appraisal questions relate to the validity, precision and applicability of these numbers. The importance of these numbers in practice depends on the group to which they are applied (see Applicability - next section).	
Was the precision of the effect estimates sufficient?	If 95% confidence intervals are wide and include the no effect point (e.g. RR=1, RD=0) or p-values are >> 0.05, then the precision of the estimates is likely to be poor & insufficient	
If no statistically significant effects detected, was there sufficient power?	If an effect estimate is not significantly different from no effect and the confidence interval is wide, the study is probably not large enough to detect a real difference between treatment and comparison groups (i.e. a low power study). A non significant effect associated with a tight CI suggests there is no effect and that the study has adequate power. Look for a power calculation in the methods section.	
If multi-centred RCT - were effects homogeneous between sites?	In multi-site studies, confidence in the results should be increased if it can be shown that similar results were obtained at the different participating centres.	
QUALITY OF STUDY RESULTS: Useful, precise +/or sufficient power? Very good = +, okay = ∅, poor = -		

SECTION 3: STUDY APPLICABILITY			
Participants	Was the source population for participants well described?	If the source population is not well described it is not easy to assess the generalisability of the study findings to a target group or whether the study participants are a typical or atypical subset of the source population.	
	Were participants representative of source population?	As above	
	Can the relevance / similarity of the participants to a specific target group(s) be determined?	As above	
Exposures & Comparison	Were the characteristics of the study setting well described? *e.g. rural, urban, inpatient, primary care*	This helps determine the applicability of the interventions	
	Can the applicability of interventions be determined?	These should be described in some detail in the paper or referenced. It should be theoretically possible for the reader to replicate the process.	
	Can the relevance of the comparison group management be determined?	It is important to determine whether the comparison group receive no interventions (e.g .placebo only) or whether they receive "usual care." As usual care may differ in different settings, it is important to determine what usual care involves	
Outcomes	Were all important outcomes considered: benefits? harms? costs?	Many studies only report data on benefits of interventions. Decisions to intervene need to balance benefits, harms and costs.	
	Are likely benefits greater than potential harms & costs (or vice versa)? In what target group(s)?	The benefits, harms and costs of interventions may differ between different groups of people due to severity, co-morbidities etc. Ideally studies should describe the overall balance of risks, benefits and costs in different subgroups.	
QUALITY OF STUDY APPLICABILITY: (a) Was it possible to determine applicability? *Very well = +, okay = ∅, poorly = - (b) Are findings applicable in your practice/setting? Very well = +, okay = ∅, poorly = -*			

STUDIES OF THE ACCURACY OF DIAGNOSTIC TESTS:
(Relevant JAMA Users' Guide Numbers IIIA & B: references (5,6))

Introduction:

The most valid study design for assessing the accuracy of diagnostic tests is a non-experimental cross-sectional study that compares a test's classification of a diagnosis with a reference standard's classification, in a relevant study population.

The conceptual starting point of a diagnostic test study is to apply the reference (or gold) standard to determine which study participants have the disease or condition (D_E) - equivalent to exposed subgroup in other studies described in this module - and which participants don't have it (D_C) - equivalent to the comparison subgroup. In many diagnostic test studies information on test results rather than the reference standard are collected first, however applying the reference standard remains the conceptual starting point.

The outcome of interest in a diagnostic test study is the test result (N). This may initially appear counter-intuitive as the outcome of interest in most studies is the disease. In the simplest example illustrated in the PECOT diagram (page 12), the test result is either positive (N+) or negative (N-). If the test is positive in someone with the condition (i.e. reference standard positive) then we use the symbol $N+_E$; if the test is positive in someone without the condition (i.e. reference standard negative) then we use the symbol $N+_C$. Similarly we can derive test negative categories $N-_E$ and $N-_C$.

The "Outcomes" square in the PECOT diagram (page 12) is equivalent to the 2x2 table often described in texts and studies about diagnostic tests, however we have turned it on its side. For some reason most 2x2 tables have the reference standard results across the top of the table and the test results down the side of the table. We suggest you use our table format because when you draw the PECOT diagram, it is more obvious where the 2x2 table comes from.

The most useful single measure of accuracy of a diagnostic test is the likelihood ratio (LR). The LR is equivalent to a relative risk in other epidemiological studies and is calculated in the same way. However it is possible to calculate LRs for different test result (e.g. for a positive or a negative test result) – see boxes below for definitions.

These numbers can also be used to calculate sensitivity and specificity, which are the more traditionally described characteristics of a diagnostic test study. While they provide useful information (see definitions in boxes below), the LR has the advantage of combining sensitivity and specificity in one number. Moreover, as long as you remember that it is equivalent to a relative risk, it is easy to derive the LR from the PECOT diagram.

If you know the LRs for a test and you have an idea of the average disease prevalence in the group of patients you would apply the test to (known as the pre-test probability), you can also use a simple tool, called a likelihood ratio nomogram (reference 6, page 705 or reference 11, page 79), to estimate the probability that the patient has the disease once you have received the test result (known as the post-test probability of disease).

For those readers who feel more comfortable with sensitivity and specificity, the LR for a positive test is the sensitivity/(1 – specificity) and the LR for a negative test is (1-sensitivity/specificity.

The likelihood ratio for a positive test (LR+ve) is the ratio of: i.) the likelihood of a positive test in people with disease to: ii) the likelihood of a positive test in people without disease.

$$\text{Likelihood Ratio for positive test (LR+ve)} = \frac{\text{number of } N+_E \text{ outcomes / number in } D_E}{\text{number of } N+_C \text{ outcomes / number in } D_C}$$

The likelihood ratio for a negative test (LR-ve) is the ratio of: i.) the likelihood of a negative test in people with disease to: ii) the likelihood of a negative test in people without disease.

$$\text{Likelihood Ratio for negative test (LR-ve)} = \frac{\text{number of } N-_E \text{ outcomes / number in } D_E}{\text{number of } N-_C \text{ outcomes / number in } D_C}$$

The sensitivity of a test is its ability to detect people who have disease; it is the proportion of all people with disease who are identified as positive by the test.

Sensitivity = number of $N+_E$ outcomes / number in D_E

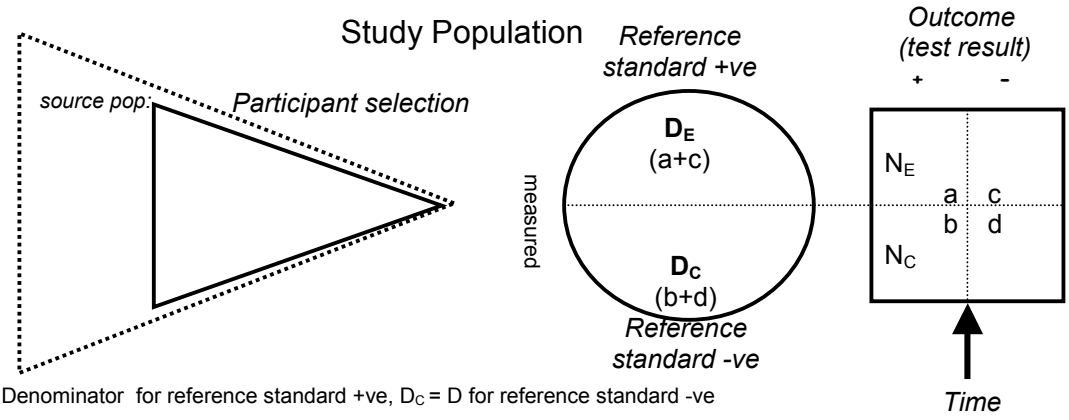
The specificity of a test is its ability to detect people who do not have disease; it is the proportion of all people without disease who are identified as negative by the test.

Specificity = number of $N-_C$ outcomes / number in D_C

The effectiveness of a diagnostic test in reducing the occurrence of a health problem (i.e. the effectiveness of screening with a diagnostic test) is best evaluated in a randomised controlled trial (see appraisal guide for experimental studies).

GATE Checklist for Diagnostic Test Studies (cross-sectional)

Study author, title, publication reference	*Key 5 part study question (PECOT). Was it focussed?*



D_E = Denominator for reference standard +ve, D_C = D for reference standard -ve

N_E = Numerator for reference standard +ve, N_C = N for reference standard -ve

SECTION 1: STUDY VALIDITY		Appraised by:	
Evaluation criterion		*How well was this criterion addressed?*	Quality ✓ ? x
Participants	What were the key selection (inclusion & exclusion) criteria? Were they well defined? Were they replicable?		
	Were selection criteria appropriate given study question?		
	Did selection lead to an appropriate spectrum of participants (like those assessed in practice)		
Exposure/ Comparison	What was the reference standard of diagnosis? Was it clearly defined, independent & valid?		
	Was the reference standard applied regardless of test result?		
	Was the reference standard assessed blind to test result?		
Outcomes	What tests were used? Were they well defined? Replicable?		
	Was the test applied regardless of the reference standard result?		
	Was test assessment blind to reference standard result?		
	Was the test validated in a second, independent group?		
QUALITY OF STUDY DESIGN: *How successfully do you think the study minimised bias? Very well = +, okay = ∅, poorly = -*			

SECTION 2: STUDY RESULTS: ACCURACY & PRECISION	
What measures of test accuracy were reported (sensitivity, specificity, LRs)?	
What measures of precision were reported (CIs, p-values)?	

THE NUMBERS TABLE: LIKELIHOODS, LIKELIHOOD RATIO ESTIMATES & PRECISION

TEST RESULT (N[O])	**IF REFERENCE STANDARD + VE: likelihood of a specific test result (N[O]) = L+ve = $(N[O]_E / D_E)$***	**IF REFERENCE STANDARD - VE: likelihood of a specific test result N[O]) = L–ve = $(N[O]_C / D_C)$***	**LIKELIHOOD RATIO LR = L+ve / L-ve (similar to RR)**	**± 95% CI**
+ve	**= sensitivity (a/a+c)**	**= 1 - specificty (b/b+d)**		
-ve	**= 1 - sensitivity (c/a+c)**	**= specificty (d/b+d)**		
etc				

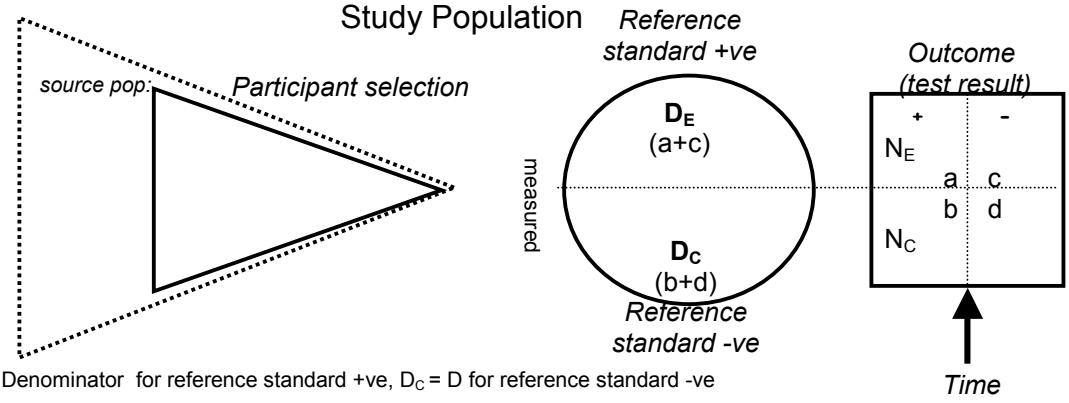
** N[O] represents the generic test result (e.g. +ve, -ve, or a level of a test)*

		Quality ✓ ? x
Could useful measures of test accuracy (i.e.likelihood ratios [LR]) be calculated?		
What was the magnitude of the LR estimates?		
Was the precision of the LR estimates sufficient?		
If no statistically significant associations detected, was there sufficient power?		
QUALITY OF STUDY RESULTS: *Useful, precise +/or sufficient power? Very good = +, okay = ∅, poor = -*		

SECTION 3: STUDY APPLICABILITY			
Participants	Was the source population for participants well described?		
	Were participants representative of source population?		
	Can the relevance of the participants to a specific target group(s) be determined?		
Exposures & Comparison	Were the characteristics of the study setting well described? *e.g. rural, urban, inpatient, primary care*		
	Can sensible estimates of individual patient's pre-test probabilities be determined from the study? (or from elsewhere?)		
Outcomes	Is the test available, affordable and reproducible in the target settings?		
	Will resulting post-test probabilities affect management and help patients? For which target group(s)?		
QUALITY OF STUDY APPLICABILITY: (a) Was it possible to determine applicability? *Very well = +, okay = ∅, poorly = - (b) Are findings applicable in your practice/setting? Very well = +, okay = ∅, poorly = -*			

USERS GUIDE for GATE Checklist for Diagnostic Test Studies

Study author, title, publication reference	*Key 5 part study question (PECOT). Was it focussed?*



D_E = Denominator for reference standard +ve, D_C = D for reference standard -ve

N_E = Numerator for reference standard +ve, N_C = N for reference standard -ve

SECTION 1: STUDY VALIDITY		Appraised by:	
Evaluation criterion		*How well was this criterion addressed?*	Quality ✓ ? x
Participants	What were the key selection (inclusion & exclusion) criteria? Were they well defined? Were they replicable?	List important selection criteria; e.g. age group, gender, risk profile, medical history. Usually in Methods section. There should be sufficient information in the paper (or referenced) to allow the reader to theoretically select a similar population	
	Were selection criteria appropriate given study question?	Are the participants a relevant group to apply the study intervention to? (e.g. diagnostic tests are not very helpful in people with a very high probability of disease).	
	Did selection lead to an appropriate spectrum of participants (like those assessed in practice)	Studies including participants with the range of common presentations of the target disorder and with commonly confused diagnoses are far more informative than studies that only include the extreme ends of the spectrum (florid cases & asymptomatic volunteers only	
Exposure / Comparison	What was the reference standard of diagnosis? Was it clearly defined, independent & valid?	The validity of the study requires that there is an accepted, valid and replicable reference (gold) standard of diagnosis. Readers should give careful and critical consideration to the authors' choice of a reference standard. In addition, those applying and interpreting the reference standard should ideally be unaware of the result of the test to avoid conscious or unconscious bias. This is not always possible, and can lead to over or under-interpretation of the reference standard results.	
	Was the reference standard applied regardless of test result?	Reference standards are often not applied to participants with negative tests, particularly if invasive. An alternative is to follow these participants for an extended period to identify any false negative cases.	
	Was the reference standard assessed blind to test result?	see above, reduces under and over-interpretation of reference standard	
Outcomes	What tests were used? Were they well defined? Replicable?	The methods for undertaking tests should be well described or referenced. It should be theoretically possible for the reader to replicate the process.	
	Was the test applied regardless of the reference standard result?	All participants who are assessed with the reference standard should be tested. Untested participants are equivalent to cases "lost to follow-up"	

	Was test assessment blind to reference standard result?	see above, reduces under and over-interpretation of test	
	Was the test validated in a second, independent group?	As diagnostic tests are predictors, not explainers,of diagnoses, it is possible that the findings in a participant group are related to the characteristics of those selected. Demonstration of test accuracy in a second participant group increases confidence in the findings.	
QUALITY OF STUDY DESIGN: *How successfully do you think the study minimised bias? Very well = +, okay = ∅, poorly = -*			

SECTION 2: STUDY RESULTS: ACCURACY & PRECISION	
What measures of test accuracy were reported (sensitivity, specificity, LRs)?	Some studies do not provide the relevant number of participants (D) in the study population who were assessed using the reference standard, the numbers who were tested (N), the proportions with various test results (N/D) in each reference stand group, or the relevant measures of test accuracy. If they are not reported or cannot be calculated, it is not possible to ascertain the accuracy of the test(s) - see definitions below in the Numbers Table below.
What measures of precision were reported (CIs, p-values)?	Either confidence intervals or p values for sensitivity, specificity & LRs should be reported or be possible to calculate

THE NUMBERS TABLE: LIKELIHOODS, LIKELIHOOD RATIO ESTIMATES & PRECISION

TEST RESULT **(N[O])**	**IF REFERENCE STANDARD + VE: likelihood of a specific test result (N[O]) = L+ve = $(N[O]_E / D_E)$***	**IF REFERENCE STANDARD - VE: likelihood of a specific test result N[O]) = L–ve = $(N[O]_C / D_C)$***	**LIKELIHOOD RATIO LR = L+ve / L-ve** **(similar to RR)**	**± 95% CI**
+ve	**= sensitivity (a/a+c)**	**= 1-specificty (b/b+d)**		
-ve	**=1-sensitivity (c/a+c)**	**= specificity (d(b+d)**		
etc				

** N[O] represents the generic test result (e.g. +ve, -ve, or a level of a test)*

		Quality ✓ ? x
Could useful measures of test accuracy (i.e.likelihood ratios [LR]) be calculated?	LRs should be reported or able to be calculated in the Numbers Table (above). If sensitivity & specificity are reported, it is possible to calculate LRs	
What was the magnitude of the LR estimates?	These numbers are the bottom line of every study. All other appraisal questions relate to the validity, precision and applicability of these numbers. The importance of these numbers in practice depends on the group to which they are applied (see Applicability - next section).	
Was the precision of the LR estimates sufficient?	If 95% confidence intervals are wide and include the no effect point (LR=1) or p-values are >> 0.05, then the precision of the estimates is likely to be poor & insufficient	
If no statistically significant associations detected, was there sufficient power?	If an LR estimate is not significantly different from 1 and the confidence interval is wide, the study is probably not large enough to determine if the test is accurate (i.e. a low power study). A non significant LR associated with a tight CI suggests the test is not useful and that the study has adequate power. Look for a power calculation in the methods section.	
QUALITY OF STUDY RESULTS: *Useful, precise +/or sufficient power? Very good = +, okay = ∅, poor = -*		

SECTION 3: STUDY APPLICABILITY			
Participants	Was the source population for participants well described?	If the source population is not well described it is not easy to assess the generalisability of the study findings to a target group or whether the study participants are a typical or atypical subset of the source population.	
	Were participants representative of source population?	As above	
	Can the relevance of the participants to a specific target group(s) be determined?	As above	
Exposures & Comparison	Were the characteristics of the study setting well described? *e.g. rural, urban, inpatient, primary care*	This helps determine the applicability of the test	
	Can sensible estimates of individual patient's pre-test probabilities be determined from the study? (or from elsewhere?)	The importance of a test depends to a large extent on the pre-test probability of the target condition (i.e. the prevalence of the condition) in the people to whom the test is applied in practice. This information is often difficult to find and readers often depend on the study to determine this.	
Outcomes	Is the test available, affordable and reproducible in the target settings?	The reproducibility of a test may depend on the expertise of those performing and evaluating the test. Information on reproducibility and training in the study setting can help determine reproducibility in other settings.	
	Will resulting post-test probabilities affect management and help patients? For which target group(s)?	The post-test probabilities of the target condition (i.e. the probability of having the target condition if the test is positive or if the test is negative) depends on both the pre-test probability in the whole group tested and the test accuracy (LR). As pre-test probabilities are likely to differ between groups, the usefulness of a test will vary from group to group.	
QUALITY OF STUDY APPLICABILITY: (a) Was it possible to determine applicability? *Very well = +, okay = ∅, poorly = - (b) Are findings applicable in your practice/setting? Very well = +, okay = ∅, poorly = -*			

NON-EXPERIMENTAL STUDIES ABOUT BENEFIT, HARM OR CAUSATION – cohort and cross-sectional studies examining the benefits and harm of exposures, including therapy and other interventions. (Relevant JAMA Users' Guide Number IV: references (7))

Introduction:

- The structure of a non-experimental study investigating benefit, harm or causation is very similar to a controlled trial. The occurrence of health outcomes (prevalence or incidence) is measured in subgroups defined by specific exposures and comparisons, which may be interventions in some instances.
- The major difference between experimental and non-experimental studies is that the investigator controls the allocation of the exposures (or treatment) to participants in an experiment, whereas the investigator in a non-experimental study categorises participants into exposure and comparison subgroups after measuring factors (i.e. the exposures) that the study participants are exposed to. In other words, the non-experimental study investigator observes a "natural" experiment rather than conducting one.
- The main weakness of non-experimental studies is the potential for confounding. As exposure allocation is not controlled by the investigator, it is common to find differences between exposure and comparison subgroups other than the main exposures of interest, that also influence health outcomes. These differences are known as confounding factors causing a mixing of effects. For questions about the benefits or harm of therapy, experimental studies (particularly randomised controlled trials) are usually superior to non-experimental studies because of the large potential for confounding in the latter.

Longitudinal (or cohort) studies:

- Cohort studies are basically non-experimental versions of controlled trials and are undertaken to investigate the effects (both benefits and harms) of exposures. In a cohort study, participants are recruited into the study population, exposures are measured, and then participants are followed up over time to measure outcomes.
- As mentioned above, cohort studies are not the most appropriate study design for examining the effects of interventions, because the potential for confounding is typically greatest when people are selected or self-select the exposures of interest (particularly therapies or exposures requiring a conscious decision by the participant, such as taking leisure time physical activity). Nevertheless, cohort studies are often used to

investigate the effects of therapy because of other the shortcomings of experimental studies.

- Cohort studies can often be conducted in situations where controlled trials are not possible. In some situations a trial would be unethical (e.g. investigating the adverse effects of a dangerous exposure such as electromagnetic radiation or cigarette smoking). In addition, trials are often not feasible when the effect of exposure (e.g. cigarette smoking) takes many years to cause an outcome (e.g. lung cancer) or when the outcomes of interest are uncommon (e.g. asthma death) and very large numbers of study participants are required to identify sufficient outcomes.
- Non-experimental longitudinal studies are also the most appropriate design for investigating prognosis.

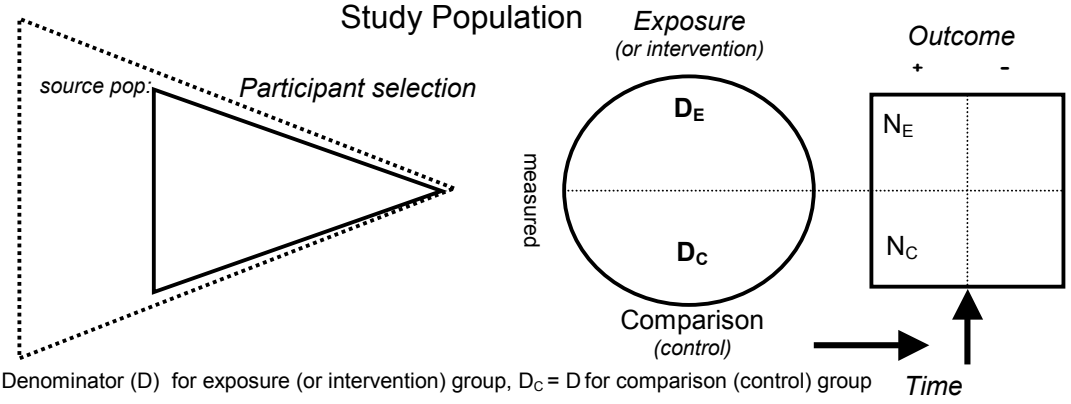
Cross-sectional studies:

- The cross-sectional study has an identical structure to the cohort study except that the exposures and outcomes are measured at the same time (i.e. cross-sectionally), whereas in a cohort study outcomes are typically measured after the exposure/s has been measured (i.e. longitudinally).
- Cross-sectional studies are the design of choice for assessing the prevalence of health-related outcomes in a target population. In such studies it is very important that the study population is representative of the target or source population of interest (i.e. the findings in the study population must be generalisable to the target population).
- Cross-sectional studies are also the design of choice for comparing diagnostic tests with a reference standard.
- Cross-sectional studies may be undertaken to investigate causal associations between exposures and outcomes, although they are not ideally suited for this purpose; especially if the outcomes of interest are acute events. As outcome measurements are made at one point in time in cross-sectional studies, many acute outcomes would be missed, particularly if they are either fatal (e.g. coronary death) or recovery occurs quickly and there are no lasting signs or symptoms of the event (e.g. asthma attacks).
- If the outcome of interest can affect the exposure of interest (e.g. a myocardial infarction may lower blood pressure), then it is not possible to validly investigate the association in a cross-sectional study, because the outcome (myocardial infarction) may be measured before the exposure (blood pressure) has been measured.
- It is therefore important to document whether the exposure was measured before the outcome occurred (i.e. check if the association is temporally correct).

- As cohort studies and most cross-sectional studies are simply longitudinal and cross-sectional versions of the same study design, they are considered together in one appraisal guide.

GATE Checklist for Cohort & Cross-sectional Studies (causation or intervention, benefit or harm)

Study author, title, publication reference	*Key 5 part study question (PECOT). Was it focussed?*



D_E = Denominator (D) for exposure (or intervention) group, D_C = D for comparison (control) group

N_E = Numerator (N) for exposure group, N_C = N for comparison group

SECTION 1: STUDY VALIDITY		Appraised by:	
Evaluation criterion (NAXS = not applicable for cross-sectional studies)		*How well was this criterion addressed?*	Quality ✓ ? x
Participants	What were the key selection (inclusion & exclusion) criteria? Were they well defined? Were they replicable?		
	Were inclusion & exclusion criteria appropriate given study question?		
Exposures & Comparison	What were the exposures (or interventions) & comparison? Well defined? Replicable?		
	Was measurement of variables similar & valid in all groups?		
	Were exposure & comparison groups similar at start of study except for study exposures?		
	If not, were differences stratified / adjusted for in analyses?		
	Were all participants analysed in groups to which initially assigned?		
	Were participants, health workers, researchers blind to exposures?		
	Apart from study exposures, were groups treated equally?		
	Were exposures remeasured during follow-up & were there important changes? (NAXS)		
Outcomes	What outcome measures were used? Well defined? Replicable?		
	How complete was follow up? Was it sufficient? How many dropouts? (NAXS)		
	Was outcome assessment blind?		
Time	Was follow up time sufficiently long to detect important effects on outcomes of interest? (NAXS)		
QUALITY OF STUDY DESIGN: *How successfully do you think the study minimised bias? Very well = +, okay = ∅, poorly = -*			

SECTION 2: STUDY RESULTS: MAGNITUDE & PRECISION OF EFFECTS	
What measures of occurrence (incidence / prevalence) & exposure effects (RR /RD /NNTs) were reported?	
What measures of precision of effects were reported (CIs, p-values)?	

THE NUMBERS TABLE: OCCURRENCE, EFFECT ESTIMATES & PRECISION

Outcomes* & Time (T)	**Exposure event rate ($EER=N_E/D_E/T$) or mean***	**Comparison event rate ($CER=N_C/D_C/T$) or mean***	**Relative Risk* (RR = EER/CER) ± (95% CI)**	**Risk difference or mean difference (RD = CER-EER) ± (95% CI)**	**Number Needed to Treat* (NNT = 1/RD) ± (95% CI)**

* *if outcomes continuous, can calculate means, mean differences, but not NNTs (don't usually calculate relative means)*
D_E = Denominator (D) for exposure (intervention) group(s), D_C = D for comparison (control) group
N_E = Numerator (N) for exposure group(s), N_C = N for comparison group

	Quality ✓ ? x
Could useful effect estimates (e.g. RR, RDs or mean differences, NNTs) be calculated? For benefits & harm?	
What was the magnitude and direction of the effect estimates?	
Was the precision of the effect estimates sufficient?	
If no statistically significant effects detected, was there sufficient power?	
QUALITY OF STUDY RESULTS: *Useful, precise +/or sufficient power? Very good = +, okay = ∅, poor = -*	

SECTION 3: STUDY APPLICABILITY			
Participants	Was the source population for participants well described?		
	Were participants representative of source population?		
	Can the relevance / similarity of the participants to a specific target group(s) be determined?		
Exposures & Comparison	Were the characteristics of the study setting well described? *e.g. rural, urban, inpatient, primary care*		
	Can the applicability/relevance of exposures be determined?		
	Can the relevance of the comparison group be determined?		
Outcomes	Were all important outcomes considered: benefits? harms? costs?		
	Are likely benefits greater than potential harms & costs (or vice versa)? In what target group(s)?		
QUALITY OF STUDY APPLICABILITY: (a) Was it possible to determine applicability? *Very well = +, okay = ∅, poorly = - (b) Are findings applicable in your practice/setting? Very well = +, okay = ∅, poorly = -*			

STUDIES ABOUT PROGNOSIS

(Relevant JAMA Users' Guide Number V: references (8))

Introduction:

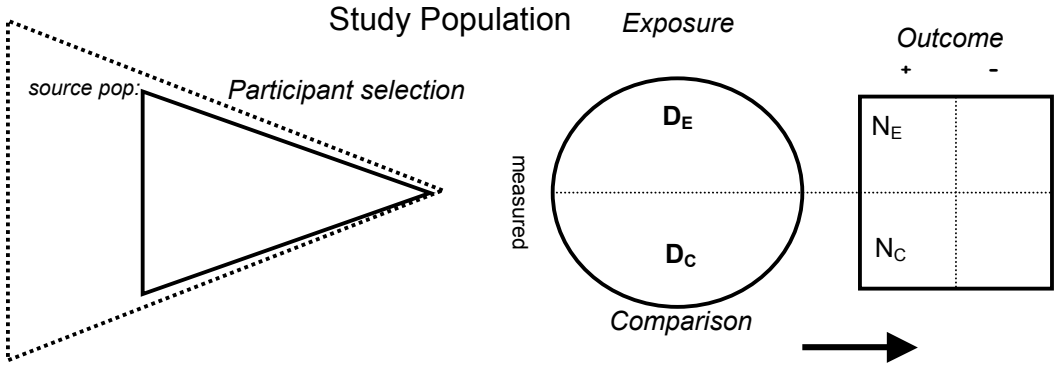
- Prognosis describes the expected occurrence, or probability, of an outcome (either good or bad) in a person with a specified condition or set of characteristics. The standard prognostic study is a cohort study in which a group of people with a particular condition or set of characteristics is followed over a period of time. At the start of the period a range of factors that may influence outcomes are measured and outcomes are measured over the period.

- Factors demonstrated to predict outcomes in a prognostic study are known as prognostic factors. These factors are equivalent to the exposures and confounding factors in a cohort study. As prognostic factors do not have to be causal, confounders can be prognostic factors. Any factors that may have an important effect on the occurrence of outcomes should be measured and classified as potential prognostic factors.

- In prognostic studies it is particularly important that the study population is a well-described and representative sample from a relevant and recognisable group of people who have a specified condition or set of characteristics and are at a similar stage in the development of a disease or other health-related outcome. Sometimes the control group in a randomised trial is used to assess prognosis, however this may be quite inappropriate if the controls are a highly selected, unrepresentative subset of usual patients.

GATE Checklist for Prognostic Studies

Study author, title, publication reference	*Key 5 part study question (PECOT). Was it focussed?*



D_E = Denominator (D) for exposure (or intervention) group, D_C = D for comparison (control) group

N_E = Numerator (N) for exposure group, N_C = N for comparison group

SECTION 1: STUDY VALIDITY		Appraised by:	
Evaluation criterion		*How well was this criterion addressed?*	Quality ✓ ? x
Participants	What were the key selection (inclusion & exclusion) criteria? Well defined? Replicable?		
	Were inclusion & exclusion criteria appropriate given study question?		
	Were participants at a common point in the course of their disease?		
Exposures & Comparison	What were the prognostic groups? Well defined? Replicable?		
	Was measurement of variables similar & valid in all groups?		
	Were different prognostic groups similar at start of study except for study prognostic factors?		
	If not, were differences stratified / adjusted for in analyses?		
	Were all participants analysed in groups to which initially assigned?		
	Were participants, health workers, researchers blind to prognostic factors?		
	Were groups treated equally?		
	Were prognostic factors remeasured during follow-up & were there important changes?		
Outcomes	What outcome measures were used? Well defined? Replicable?		
	How complete was follow up? Was it sufficient? How many dropouts?		
	Was outcome assessment blind?		
Time	Was follow up time sufficiently long to detect important prognostic factors?		
QUALITY OF STUDY DESIGN: *How successfully do you think the study minimised bias? Very well = +, okay = ∅, poorly = -*			

SECTION 2: STUDY RESULTS: MAGNITUDE & PRECISION OF PROGNOSIS	
What measures of prognosis (over what time) & differences between groups (e.g. RR /RD) were reported?	
What measures of precision were reported (CIs, p-values)?	

THE NUMBERS TABLE: PROGNOSIS & PRECISION

Outcomes	**Time of follow-up**	**Prognostic group E event rate ($EER=N_E/D_E/T$) or mean* ± (95% CI)**	**Prognostic group C event rate ($CER=N_C/D_C/T$) or mean* ± (95% CI)**	**Relative Risk* (RR = EER/CER) ± (95% CI)**

* *if outcomes continuous, can calculate means & mean differences*
D_E = Denominator (D) for exposure group(s), D_C = D for comparison group
N_E = Numerator (N) for exposure group(s), N_C = N for comparison group

		Quality ✓ ? x
Could useful estimates of prognosis be calculated? (i.e. events/person/time)		
What was the magnitude and direction of the prognostic estimates?		
Was the precision of the prognostic estimates sufficient?		
If no statistically significant estimates detected, was there sufficient power?		
QUALITY OF STUDY RESULTS: *Useful, precise +/or sufficient power? Very good = +, okay = ∅, poor = -*		
SECTION 3: STUDY APPLICABILITY		
Participants	Was the source population for participants well described?	
	Were participants representative of source population?	
	Can the relevance / similarity of the participants to a specific target group(s) be determined?	
Exposures & Comparison	Were the characteristics of the study setting well described? *e.g. rural, urban, inpatient, primary care*	
	Can the applicability/relevance of prognostic factors be determined?	
Outcomes	Were all important outcomes considered: benefits? harms?	
	Would the prognostic information have an important impact on patient or practitioner decisions? In what target group(s)?	
QUALITY OF STUDY APPLICABILITY: (a) Was it possible to determine applicability? *Very well = +, okay = ∅, poorly = - (b) Are findings applicable in your practice/setting? Very well = +, okay = ∅, poorly = -*		

SYSTEMATIC REVIEWS OF STUDIES OF THERAPY OR OTHER INTERVENTIONS:

(Relevant JAMA Users' Guide Number VI: references (9))

Introduction:

- Few individual trials are large enough or include a wide enough range of relevant groups of people to definitively answer questions about therapy or prevention.

- Systematic reviews of trials attempt to summarise all relevant trials that have addressed a particular question, by following a set of systematic rules to ensure both the completeness of the review and the validity of the findings.

- By bringing together all relevant trials it is possible to determine more precise estimates of effect than that available in any single trial. Moreover if all trials of good quality show similar effect estimates, the reviewer can be more confident in the findings.

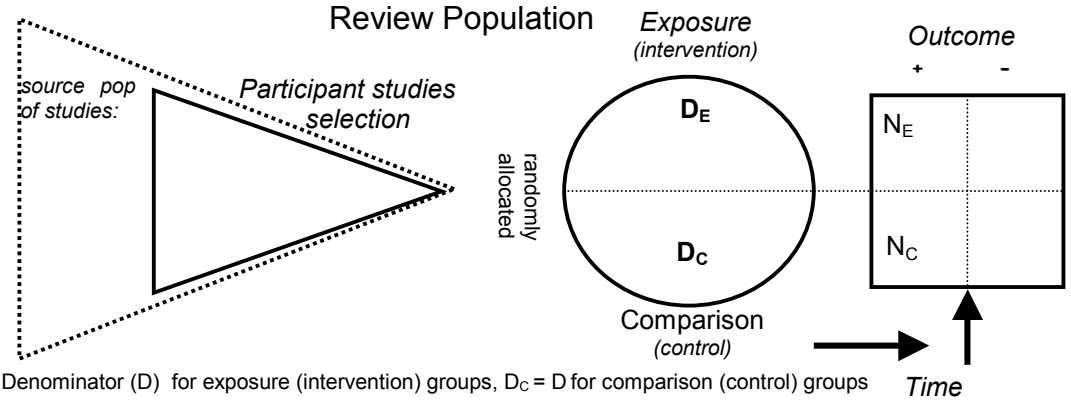
- All trials included in a systematic review should first be assessed using the critical appraisal guide for experimental studies. Ideally evidence-based decisions should be based on systematic reviews of evidence rather than individual studies.

- It is also possible to undertake systematic reviews of non-experimental studies.

- It is beyond the scope of this guide to describe the mathematical approach to synthesising data from different studies. This is covered in many papers and texts.

GATE Checklist for Systematic Reviews of Randomised Controlled Trials

Review author, title, reference	*Key 5 part Review question (PECOT). Was it focused?*



D_E = Denominator (D) for exposure (intervention) groups, D_C = D for comparison (control) groups

N_E = Numerator (N) for exposure groups, N_C = N for comparison groups

SECTION 1: REVIEW VALIDITY		Appraised by:	
Evaluation criterion		*How well was this criterion addressed?*	Quality ✓ ? x
Participant studies	What were the key (inclusion & exclusion) criteria for selecting studies? Were they well defined? Were they replicable?		
	Were inclusion & exclusion criteria appropriate given study question?		
	Was the search for studies comprehensive? Complete?		
	How was the validity of individual studies assessed?		
Exposures & Comparison	What were the exposures (interventions) & comparison? Well defined? Replicable? Similar from study to study?		
	Was assignment to groups randomised in all studies? Was randomisation concealed?		
	Was randomisation successful in all studies? If not, how were potential confounders dealt with?		
	Was intention-to-treat analyses used in all studies?		
Outcomes	What outcome measures were used? Well defined? Replicable? Similar in all studies?		
	Was follow-up sufficient in all studies?		
Time	Was follow up time sufficiently long to detect important effects on outcomes of interest in all studies?		
Was it reasonable to combine the studies (i.e. were the participants, interventions, comparisons, outcomes and times of follow-up similar enough from study to study)?			
QUALITY OF REVIEW DESIGN: *How successfully do you think the Review minimised bias? Very well = +, okay = ∅, poorly = -*			

SECTION 2: REVIEW RESULTS: MAGNITUDE & PRECISION OF EFFECTS	
What measures of occurrence (incidence / prevalence) & exposure (intervention) effects (e.g. RR or ORs) were reported for each study?	
What measures of precision of effects were reported (CIs, p-values)?	

THE NUMBERS TABLE: OCCURRENCE, EFFECT ESTIMATES & PRECISION

Author & year	Quality rating	Number of participants, brief description of selection critieria	Exposures & Comparison	Outcomes & time	Exposure event rate ($EER=N_E/D_E/T$) or mean*	Comparison event rate ($CER=N_C/D_C/T$) or mean*	Relative Risk* RR = EER/CER) or Odds Ratio (±95% CI)

* *If outcomes continuous, can calculate means, mean differences.*

- *D_E = Denominator (D) for exposure (intervention) group(s), D_C = D for comparison (control) group.*
- *N_E = Numerator (N) for exposure group(s), N_C = N for comparison group*

			Quality ✓ ? x
Could useful summary effect estimates (e.g. RR, ORs or mean differences) be calculated? For benefits & harm?			
What was the magnitude and direction of the summary effect estimates?			
Was the precision of the summary effect estimates sufficient?			
If no statistically significant effects detected, was there sufficient power?			
Were the effect estimates consistent from study to study?			
QUALITY OF REVIEW RESULTS: *Useful, precise +/or sufficient power? Very good = +, okay = ∅, poor = -*			
SECTION 3: REVIEW APPLICABILITY			
Participants	Were source populations for participants described?		
	Can the relevance / similarity of the participants in the Review to a specific target group(s) be determined?		
Exposures & Comparison	Were the characteristics of the study settings well described? *e.g. rural, urban, inpatient, primary care*		
	Can the applicability of exposures (interventions) be determined?		
	Can the relevance of the comparison groups management be determined?		
Outcomes	Were all important outcomes considered:benefits? harms? costs?		
	Are likely benefits greater than potential harms & costs (or vice versa)? In what target group(s)?		
QUALITY OF REVIEW APPLICABILITY: (a) Was it possible to determine applicability? *Very well = +, okay = ∅, poorly = - (b) Are findings applicable in your practice/setting? Very well = +, okay = ∅, poorly = -*			

REFERENCES:

1. Evidence-Based Medicine Working Group. Evidence-Based Medicine. A new approach to teaching the practice of medicine. JAMA 1992;268:2421-5.

2. Oxman A, Sackett DL, Guyatt GH. Users' guides to the medical literature. I. How to get started. JAMA 1993;270:2093-2095.

3. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? JAMA 1993;270:2598-2601.

4. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? JAMA 1994;271:59-63.

5. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? JAMA 1994;271:389-391.

6. Jaeschke R, Gordon H, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. what are the results and will they help me in caring for my patients? JAMA 1994;271:703-707.

7. Levine M, Walter S, Lee H, Haines T, Holbrook A, Moyer V. Users' guides to the medical literature. IV. How to use an article about harm. JAMA 1994;271:1615-1619.

8. Laupacis A, Wells G, Richardson S, Tugwell P. Users' guides to the medical literature. V. How to use an article about prognosis. JAMA 1994;272:234-237.

9. Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. JAMA 1994;272:1367-1371.

10. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-Based Medicine (second edition). Churchill Livingstone: London 2000.

EVIDENCE TABLE TEMPATE
for therapy

Clinical question: Treatment 1 vs Treatment 2 or placebo or no treatment

Study authors and year	Study Design	Participants	Exposure/ Comparison	Outcomes	Results					Quality Scores
					EER	CER	RR	RD	NNT	