Statistics & outcome measures

Kenneth Rankin

Academic Clinical Lecturer in Orthopaedics

20th September 2010

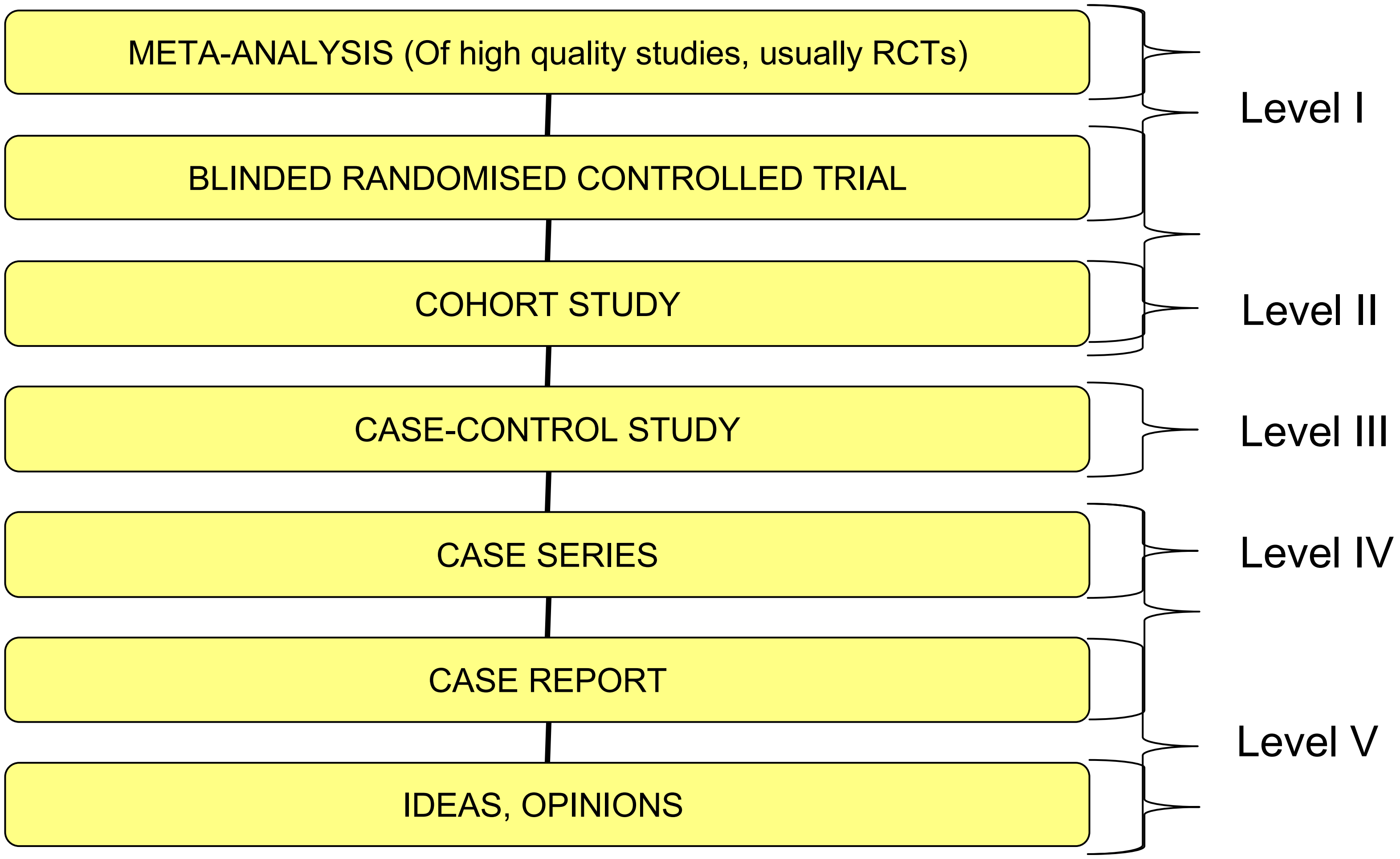
Study types

- Observational or experimental
- Observational
  - Epidemiological
  - Cross-sectional
  - Longitudinal
    - Prospective
    - Retrospective
  - Cohort
    - Prospective study of population group to see who develops a condition of interest
    - Data is presented with calculation of relative risk
  - Case-control
    - Retrospective study of patients with a condition and a control group without
    - Data presented with calculation of odds ratio

Study types (cont)

- Experimental
  - Investigator intervenes to effect the outcome
  - Longitudinal and prospective
    - Case series: non-comparative
    - Clinical trial: comparative i.e. controlled
      - May be randomised +/- blinded

Hierarchy of evidence



META-ANALYSIS (Of high quality studies, usually RCTs)

BLINDED RANDOMISED CONTROLLED TRIAL

COHORT STUDY

CASE-CONTROL STUDY

CASE SERIES

CASE REPORT

IDEAS, OPINIONS

Level I

Level II

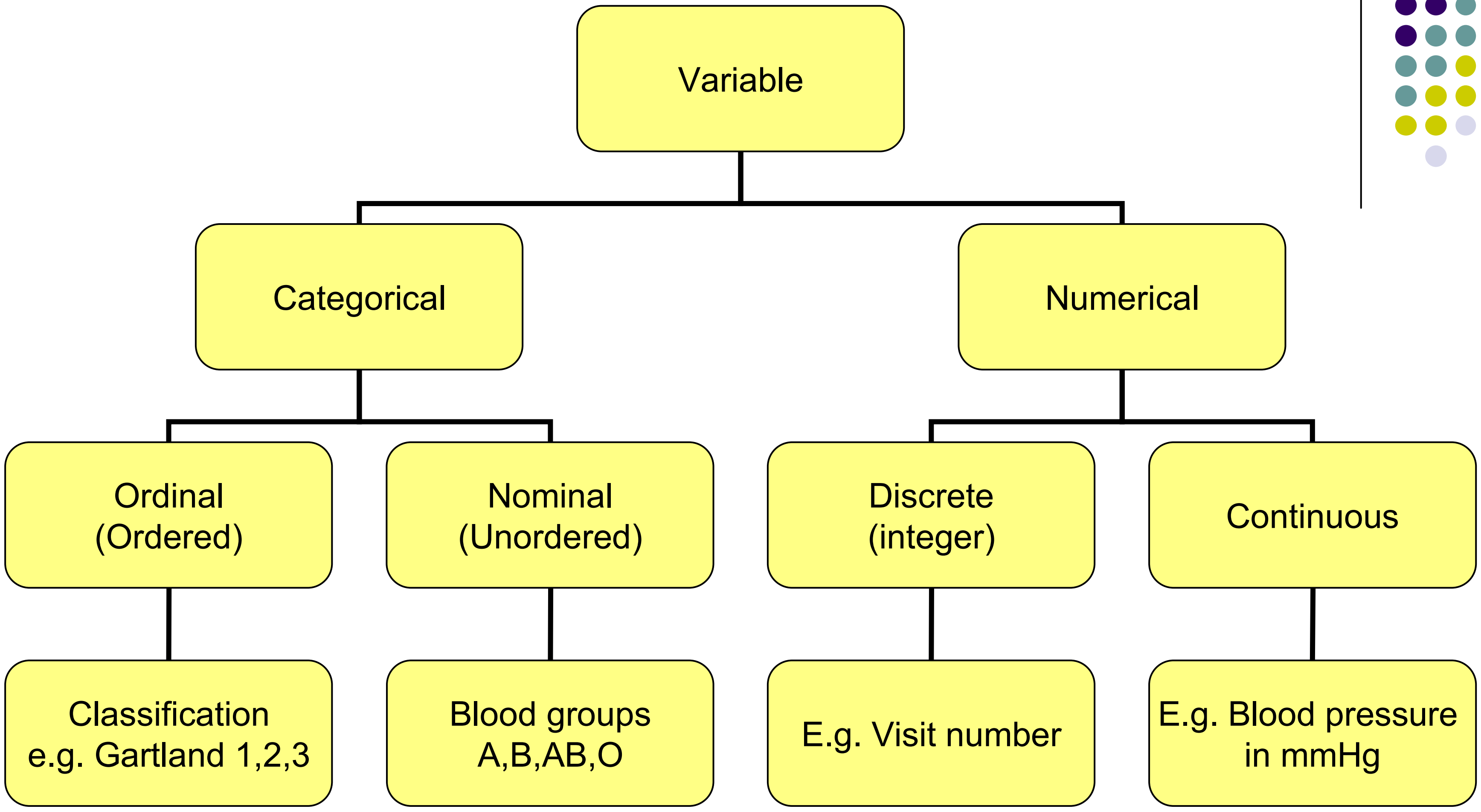
Level III

Level IV

Level V

Statistics definitions

- Data
  - The observations made on one or more variables of interest
- Statistics
  - The methods of collecting, summarising, presenting, analysing, and drawing conclusions from data
- Descriptive statistics
  - Summary of a dataset e.g. table or diagram
- Inferential statistics
  - Sample of the population which we hope is representative
  - Involves estimation of the population parameters e.g. normal distribution or not
  - Testing of hypotheses related to the population
- Values
  - Categorical: value is assigned to a particular category e.g. a sample of hip replacements: cemented v uncemented
  - Numerical: value is purely a number e.g. number of millilitres of blood lost during a procedure
- Diagrams
  - Table or diagram to illustrate the frequency distribution of a variable
  - Categorical values: a bar chart or pie chart
  - Numerical values: histogram


Variable

Categorical

Numerical

Ordinal
(Ordered)

Nominal
(Unordered)

Discrete
(integer)

Continuous

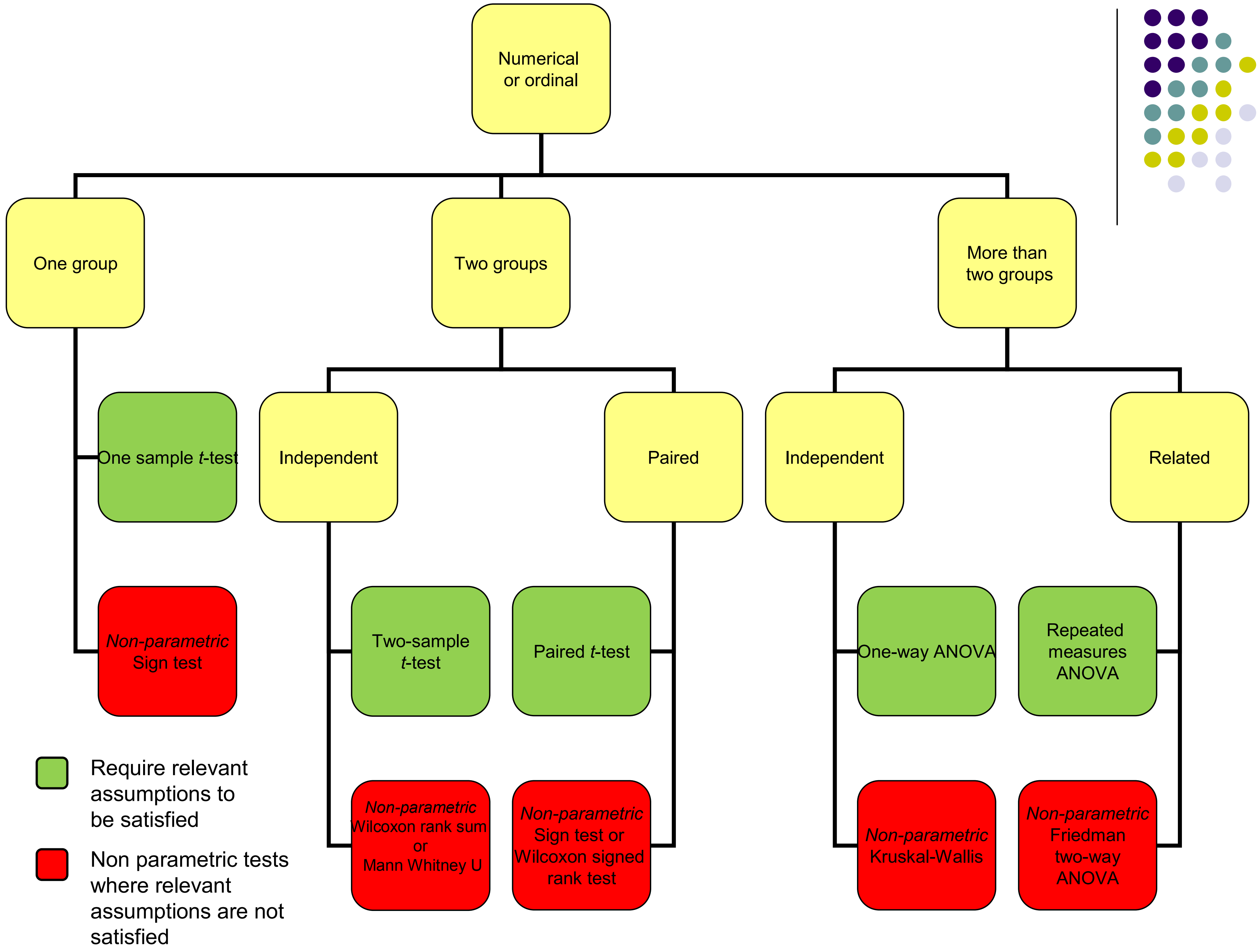
Classification
e.g. Gartland 1,2,3

Blood groups
A,B,AB,O

E.g. Visit number

E.g. Blood pressure
in mmHg

Binary variable: present or not
e.g. union or non union



Numerical or ordinal

One group

One sample *t*-test

Non-parametric Sign test

Two groups

Independent

Two-sample *t*-test

Non-parametric Wilcoxon rank sum or Mann Whitney U

Paired

Paired *t*-test

Non-parametric Sign test or Wilcoxon signed rank test

More than two groups

Independent

One-way ANOVA

Non-parametric Kruskal-Wallis

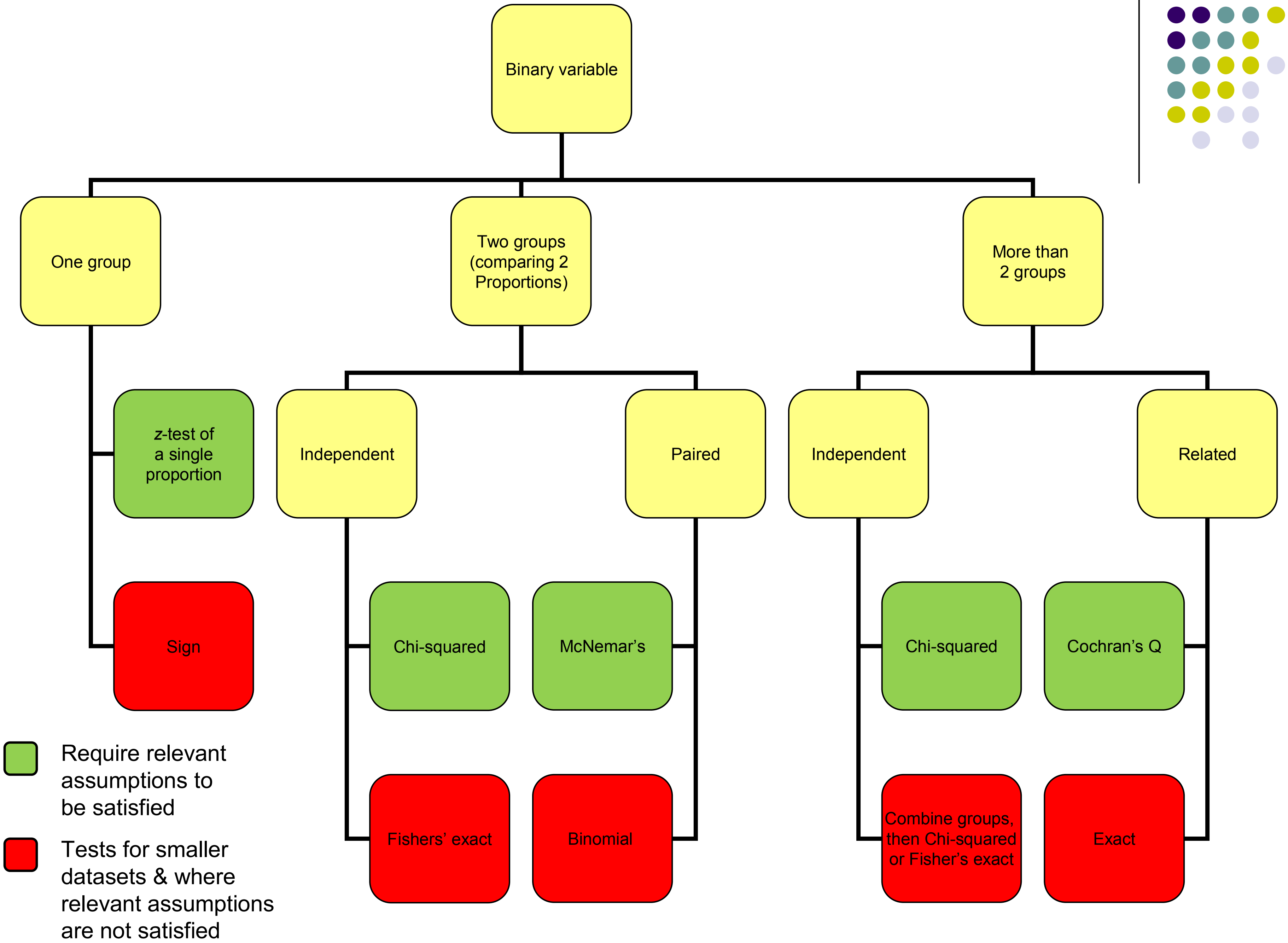
Related

Repeated measures ANOVA

Non-parametric Friedman two-way ANOVA

🟩 Require relevant assumptions to be satisfied

🟥 Non parametric tests where relevant assumptions are not satisfied



Binary variable

One group

z-test of a single proportion

Sign

Two groups (comparing 2 Proportions)

Independent

Chi-squared

Fishers' exact

Paired

McNemar's

Binomial

More than 2 groups

Independent

Chi-squared

Combine groups, then Chi-squared or Fisher's exact

Related

Cochran's Q

Exact

🟩 Require relevant assumptions to be satisfied

🟥 Tests for smaller datasets & where relevant assumptions are not satisfied

Summary measures

- Mean
  - Sum of values divided by the number of values
  - Useful for statistical tests
  - Affected by outliers which may produce inappropriate results
- Median
  - Middle value in a series
  - Not affected by outliers
- Mode
  - Most common value in a series
- Range
  - Simplest measure of spread
  - Heavily influenced by outliers
  - Interquartile range often used to avoid this i.e. takes the central 50% of a series of values

Summary measures (cont)

- Variance
  - Alternative measure of spread using all the data
  - Allows a calculation of the standard deviation (SD)
  - SD is essentially an average of the deviations from the mean

Estimating parameters

- Standard error of the mean (SEM)
  - A calculation to estimate how close our sample mean is to our population mean
  - SEM=SD/√n
  - Often put on graphs to make the ‘error bars’ look smaller!
- 95% confidence intervals
  - Range of values within which the true mean would lie 95% of the time if the project was performed repeatedly
  - Not quoted in many papers

Testing hypotheses

- Research idea
- Build a simple hypothesis
- Background
  - Blood transfusions are often necessary following TKR and various drugs e.g. Tranexamic acid administered perioperatively may reduce the transfusion requirement
- Hypothesis
  - Tranexamic acid given perioperatively reduces the transfusion requirement
  - The null hypothesis is that the Tranexamic acid is not effective
  - If there is a statistically significant reduction in transfusion requirement then we reject the null hypothesis
- Error types
  - Type I: we incorrectly reject the null hypothesis i.e. we display our results as significant, but they are not e.g. use of a parametric test to assess skewed data resulting in false significance
  - Type II: we incorrectly accept the null hypothesis e.g. We find the Tranexamic acid to have no effect on transfusion requirement but have not removed outliers who have bled a large amount for other reasons such as occult coagulopathy

Which test to use?

- Fundamental part of the study design- should not be thought about after data is collected
- Is the variable categorical or numerical?
- How many groups are being compared
- Are the assumptions underlying the proposed test satisfied? E.g. normal v skewed
  - Test for whether the data is normally distributed or not, e.g. SPSS to add a curve to the histogram
  - Kolmogorov-smirnov test: http://www.physics.csbsju.edu/stats/KS-test.n.plot_form.html
- Normal distribution: use a parametric test
- Skewed data: use a non-parametric test

Sample size estimation

- How large
  - Enough to significantly show an important treatment effect
  - Not so large as to waste resources and delay potential benefits to all patients
- Consider
  - Significance level, usually $p<0.05$
  - Power of the test, usually >80% to detect a significant effect
- Methods
  - Computer e.g. nQuery Advisor
  - Books of tables e.g. Machin et al
  - Diagram e.g. Altman’s nomogram
  - Formulae e.g. Lehr

Relationships between variables

- 2 or more variables are frequent in orthopaedic studies
- Regression models
- 2 variables: univariable linear regression
- More than 2 variables: multivariable linear regression

Univariable

- Data must be checked for distribution
- Linear correlation measurement is provided by the correlation coefficient (*r*) which ranges from -1 to +1
- Pearson correlation coefficient is parametric for data of normal distribution & Spearman correlation coefficient is non-parametric
- -1 one variable decreases as the other increases
- +1 one variable increases as the other increases
- A value of 0 indicates no correlation
- Significance is attached to it and is highly dependent on the number of observations in the sample

Multivariable

- Essentially an extension of univariable
- Usefulness expires if number of variables exceeds 1/10th of the number of observations e.g. 10 variables assessed in a study of 80 patients
- For binary variables: linear logistic regression analysis performed giving an odds ratio
  - Defined end point

Survival analysis

- Binary end point e.g. revision rate over varying length of time
- Kaplan-Meier curve
  - Non-parametric i.e. takes into account the data may not be normally distributed
- A time point to discuss survival can be taken, but the curve is not calculated up to an end point
- Non-parametric log rank test for comparing different curves on the plot

Diagnostic tests

- Indicates whether a patient has a particular disease/condition or not
- May be used for screening an apparently healthy person for a condition/disease
- Important features of a test
  - Sensitivity
    - If a person has the condition how often the test is positive
  - Specificity
    - If a person does not have the condition how often the test is negative
  - Confidence intervals should be quoted for further detail as to the accuracy of the test
  - Positive predictive value
    - The percentage of people with a positive test result who actually have the condition
  - Negative predictive value
    - The percentage of people with a negative test who do not have the condition
- Bayesian approach
  - Use of clinical features other than the test to assess of that test

Reliability studies

- To assess the accuracy of a measurement or classification
- Intra-observer i.e. the same person making repeated assessments
- Inter-observer i.e. different people making the assessments
- Categorical data: Kappa statistic
- Numerical data: Bland-Altman plot

Common Errors

- Design
  - Inappropriate or no control group
  - No randomisation (experimental study)
  - No blinding
  - Inadequate response rate
- Analysis
  - No checking of underlying assumptions e.g. Normality of data
  - Inappropriate use of arithmetic mean to summarise skewed data
  - Failure to recognise dependencies in data e.g. using the measurements from 2 limbs of the same patient and treating the data as independent as if it came from different patients
  - Failure to use the correct unit of analysis
  - Inappropriate analysis of variance

Common errors (cont)

- Presentation
  - Not specifying the primary aim of the study
  - Not providing an adequate description of the randomisation process in an experimental study
  - Not reporting exact p-values e.g. p<0.05
  - Not providing measures of precision e.g. no CI
  - Poor diagrams with inadequate labelling, using bar charts for continuous data
  - Describing data as parametric or non-parametric
    - The data is in a normal distribution or not
    - The test to assess the data is parametric or non-parametric

Common errors (cont)

- Interpretation
  - Conclusions go beyond what the data warrants
  - Conclusions are not a reasonable reflection of the data presented

Source material

- Journal article
  - Statistics in orthopaedic papers

    Petrie A

    J Bone Joint Surg Br. 2006 Sep;88(9):1121-36
- Book
  - Medical statistics made easy

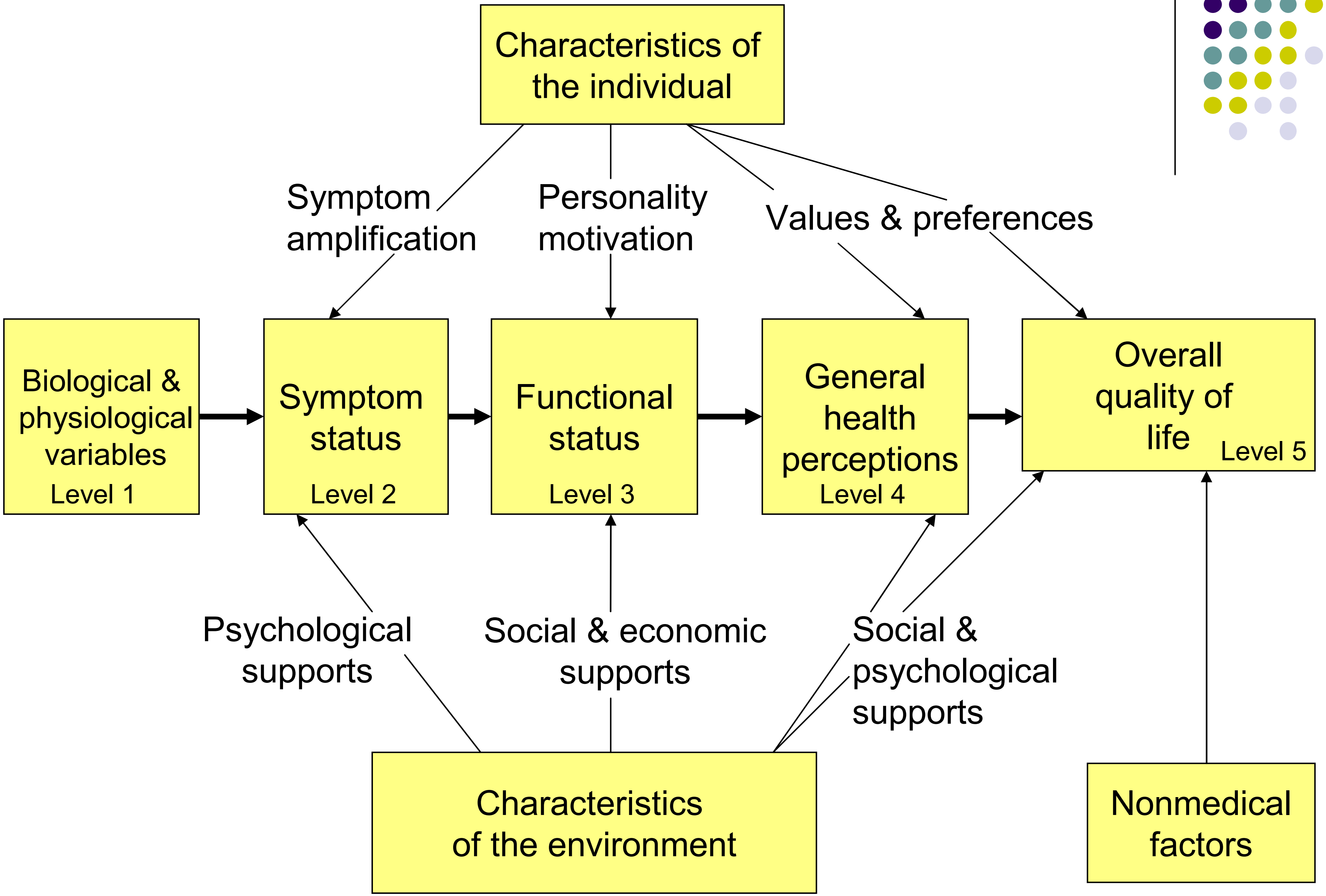
    M. Harris and G.Taylor

QUESTIONS?

Outcome measures

- Outcome is a visible or practical result
- Provide the basis for clinical research and audit
- Increasing use to provide detailed assessment of patient outcome following treatment, non-operative or operative
- Objective (hard)
  - Radiology e.g. alignment/ time to union
  - ROM measurements
- Subjective (soft)
  - Pain score
  - Patient satisfaction questionnaire
- Objective measurements are not necessarily superior to subjective
  - The most important feature is the ability of a measure to indicate improvement in patient function or symptoms **from the patient's point of view**
  - A TKR may look perfect on the radiograph and ROM may be from 0-130°, but if the patient has significant residual pain an objective outcome measure may be misleading

Wilson & Cleary 1995



Modes of administration of outcome instruments

Mode of administration	Advantages	Disadvantages
Interviewer	Maximal response rate Can clarify questions Higher completion rate Control over who is the respondent Control over order of questions	Costly Interviewer bias Reporting bias Characteristics of interviewer may influence bias
Telephone	Good response rate Relatively inexpensive Quick data collection Probe for complete answers Clarification of ambiguous answers	Excludes those without a telephone Voice inflection of interviewer may introduce bias
Mail	Relatively inexpensive No bias from interviewer May reach more respondents Respondents can take time to locate information for their answers	Low response rate Possibility of bias from non-response No control of who is the respondent May misunderstand question May miss questions
Computer based	Consistent presentation Prompts for omissions Can be web based Reliable scoring & transfer to database	Demands subject sits/stands in front of computer Demands some computer skills
Self	Maximal response rate Inexpensive	May misunderstand question May miss questions
Proxy	Can collect info on patients who would otherwise not be represented	Response may differ from that of target respondent

Types of outcome measures

- Mixed clinician based and functional outcomes
  - Questioning the patient and performing a clinical examination to document scores
  - Not recommended due to variability in clinical examination
- System specific
  - Related to one body system, usually one joint e.g. Oxford knee score, DASH
- Disease specific
  - Measuring a patient’s well being e.g. quality of life assessment of patient with OA
- General health related quality of life measures
  - Detailed assessment of a person’s functional abilities without focusing on a disease e.g. Short Form-36
  - ↑Use since mid 1990s to complement traditional clinical examination & radiographic scores

Selecting an outcome measure

- Define the research question
- Consult experienced musculoskeletal clinical researchers
  - Nurse practitioners
  - Rheumatologists
- Identify measures to ideally cover all 5 levels
- Literature search to assess for system specific & overall quality measures
  - Literature available summarising the best tools for upper limb, lower limb and general quality of life
  - *Outcome instruments: rationale for their use*

    Poolman et al JBJS (Am) 2009 May;91 Suppl 3:41-9

Commonly used measures

- SF-36
  - Generic questionnaire for any disease
  - Measures overall quality of life
- WOMAC (Western Ontario & McMaster Universities)
  - Hip or knee OA
  - 24 item questionnaire
  - Useful for assessing outcomes pre & post treatment
- www.orthopaedicscore.com

REGION	Clinician completed	Patient completed
Hip	Harris Hip Score	Oxford Hip Score
		HOOS (Hip disability and Osteoarthritis Outcome)
		WOMAC Score
Knee (Osteoarthritis)	Knee Society Score (KSS)	Oxford Knee Score
		KOOS (Knee Injury and Osteoarthritis Outcome)
		WOMAC Score
		IKDC
Knee (Anterior Cruciate Ligament)	Modified Cincinatti Rating System	KOOS (Knee Injury and Osteoarthritis Outcome)
	Tegner Lysholm Knee Scoring Scale	Modified Cincinatti Rating System
		Tegner Lysholm Knee Scoring Scale
Foot/Ankle	American Foot & Ankle Score	Foot & Ankle disability Index
Shoulder	Constant Shoulder Score	Oxford Shoulder Score
	UCLA Shoulder rating scale	DASH (Disabilities of arm, shoulder and hand) Score
		Quick-DASH Score
Shoulder (Instability)	ROWE Score for instability	Oxford Instability Score
Elbow	MAYO Elbow Score	Oxford Elbow Score
		DASH (Disabilities of arm, shoulder and hand) Score
		Quick-DASH Score
Wrist	MAYO Wrist Score	DASH (Disabilities of arm, shoulder and hand) Score
		Quick-DASH Score
Hand		DASH (Disabilities of arm, shoulder and hand) Score
		Quick-DASH Score
Lumbar Spine		Oswestry Low Back Pain Score
		Modified Oswestry Low Back Pain Score
		Back pain Index
Cervical Spine		Vernon & Mior Cervical Spine Score

Trauma outcome measures

- More than 50 scores for evaluation at the scene to A&E to theatre & to ITU
- Useful for tracking acute patient progress and for auditing outcomes
- Three main groups
    - Anatomical
        - Abbreviated injury scale (AIS)
        - Injury severity score (ISS)
        - New injury severity score (NISS)
        - Anatomic profile
    - Physiological
        - Revised trauma score (RTS)
        - Glasgow coma scale (GCS)
        - Acute physiology and chronic health evaluation (APACHE)
    - Combined
        - Trauma and injury severity score (TRISS)
        - International classification of diseases-based ISS (ICISS)
- Mangled extremity severity score (MESS)

Quality criteria for outcome measures

- Content validity
  - Avoidance of deviation e.g. a knee instability measure including questions on OA may have little relevance to an athlete out of training due to knee instability
- Internal consistency
  - Different subscales in a tool may measure similar features
  - Conversely, the tool may have very divergent subscales
  - Cronbach alpha should be measured: a low result indicates poor correlation of the measures & a high result indicates good correlation but redundancy
- Criterion validity
  - Compare the tool to a gold standard (if available)
- Construct validity
  - When no gold standard is available, attempts should be made to validate the tool with reference to existing data

Quality criteria for outcome measures

- Reproducibility
  - Agreement: the extent to which repeated scores are close to each other (absolute measurement error)
  - Reliability: the extent to which patients can be
- Responsiveness
  - Ability of the tool to measure clinically important changes over time
- Floor & ceiling effects
  - The tool should not produce too many results with near perfect scores
- Interpretability
  - The degree to which qualitative meaning can be assigned to quantitative scores i.e. a tool may pick up statistically significant small changes which make no clinical difference e.g. a large sample of 2 groups of TKRs with small significant differences in alignment & ROM scores that have no difference in patient satisfaction or overall function scores

Methodological considerations

- Use one tool for each outcome level
- For multiple outcomes adjust for this when applying statistical tests
  - 5 levels: increase significance to $p<0.01$ from 0.05 to account for the extra levels
  - Run Bonferroni post hoc
- Report all the results from the tool, not just the interesting/significant ones

Minimally important differences

Defined as

“the smallest difference in a score of a domain of interest that patients perceive to be beneficial and that would mandate, in the absence of troublesome side effects and excessive costs, a change in the patient’s management”

Mathematically expressed as ½ of a SD (continuous measure only)

Categorical v continuous

- Categorical
  - Usually dichotomous i.e. One of two categories
  - Requires larger sample size
- Continuous outcomes
  - Numerical value e.g. Blood pressure, time to fracture union
- Statistical analysis differs

Sample size calculation for dichotomous outcomes

- Define the 2 outcomes
- Determine the level of clinically relevant difference (5% improvement)
- Set the power of the study (80%)
- Results in over 1000 patients per group in most calculations due to the necessary formula

Sample size calculation for continuous variables

- Define the primary outcome variable e.g. SF-36 physical functioning score
- Determine the effect size (0.5 of SD)
- Set the power of the study (80%)
- Results in much more reasonable groups of <100 (again due to appropriate formula)

Composite outcomes

- Additional outcomes added to the dichotomous
- Increases statistical precision
- Reduces sample size
- Increased care needed for interpretation

Future of outcome measures

- Increased cohesion of orthopaedic outcome measures
- Follow the principles of OMERACT to produce standardised well validated tools to be used for all areas of research

Source material

- Journal articles
  - Outcome instruments: rationale for their use

    Poolman RW, Swiontkowski MF, Fairbank JC, Schemitsch EH, Sprague S, de Vet HC.

    J Bone Joint Surg Am. 2009 May;91 Suppl 3:41-9
  - Outcome measures and implications for sample size calculations

    Zlowodzki M, Bhandari M.

    J Bone Joint Surg Am. 2009 May;91 Suppl 3:35-40
- Book
  - Outcome measures in trauma

    P.B. Pynsent, J.C.T. Fairbank, A.J. Carr

QUESTIONS?