NIHR Statistics Group – Imaging Studies section

11th November 2015

Public Health Building, University of Birmingham

Statistical Issues in designing a large-scale reliability exercise in ultrasonography of the joint synovium

Clinical Background

(Presented by Dr Richard Wakefield, senior lecturer and honorary consultant in rheumatology, NIHR Leeds Musculoskeletal Biomedical Research Unit and Leeds Institute of Rheumatic and Musculoskeletal Medicine.)

The purpose of the meeting was to discuss statistical issues in designing a reliability exercise for ultrasonography of the joint synovium, the site of initiation in rheumatoid disease.

Rheumatoid disease is a condition affecting approximately 1% of the population in the UK, typically affecting multiple joints. Clinical examination is non-specific, and often shows poor correlation with markers of inflammation such as ESR (erythrocyte sedimentation rate) and CRP (C-reactive protein).

Available imaging modalities include:

- X-ray: insensitive to mild changes in bone and cartilage
- Magnetic resonance imaging (MRI): regarded as the gold standard, and capable of detecting changes in soft tissue, but infeasible for multiple assessments
- Ultrasound (US): quicker than MRI and more practical for multi-joint assessment

US provides two types of image: 'Grey scale' (GS), which is used to indicate the presence of synovitis (inflammation of the joint synovium) and structural abnormalities, and 'Power Doppler', which is used to indicate vascularity (blood flow) and for functional assessment.

For a given joint, these can be scored using different views (below and above), which may give rise to different results, with a differential effect seen for different joints. Each patient presenting with possible rheumatoid disease is scored at multiple joints, possibly including repeat measurements of the same joint. Scoring systems can be binary, ordinal or continuous, but GS and PD are typically scored as Grade 0 (healthy) to Grade 3 for each joint. An alternative outcome variable is the 'Resistive Index', which requires several assessments of the joint(s) to be made.

Practical issues in the scoring of US images include the difficulty of standardising the examination between patients, and between readers; reliability considerations when assessing US results from a static image as opposed to a video or in real time; and difficulties in interpreting scores above Grade 0 as some individuals exhibit mild synovitis (typically, scoring Grade 1) even when no rheumatoid disease is present.

Statistical considerations

(Presented by Dr Liz Hensor, statistician, NIHR Leeds Musculoskeletal Biomedical Research Unit and Leeds Institute of Rheumatic and Musculoskeletal Medicine.)

Reliability exercises are commonly performed within larger studies of treatment efficacy or effectiveness. As a result, they are often small-scale and rarely adequately powered.

The Leeds Inflammatory Arthritis Continuum (IACON) is a cohort study for individuals with early Inflammatory Arthritis (IA), and has recruited more than 1200 patients since 2010. GS and PD measurements are scored per joint, at time of recruitment, after 6 months and 12 months, and annually subsequently. Plan is to select images from 100 patients (26 joints each); five operators independently score each joint twice, at an interval of two weeks.

General statistical considerations relevant to the design and analysis of IA reliability exercises such as IACON include:

- Clustering: scoring is carried out for different 'compartments' of the joint, leading to a hierarchy of compartment within joint within patient. This needs to be accounted for in calculating measures of agreement.
- Low prevalence of severe outcomes makes reliability difficult to assess – for example, if the majority of patients have measurements of Grade 0 or Grade 1.
- Can GS and PD data be combined to form a summary score? A contingency table of GS versus PD has been proposed.
- Grades 0 to 3 cannot be considered interval scaled, as intervals between adjacent categories cannot be interpreted as equal; implications for reliability analysis (ICCs)
- It may be necessary to account for joint size (e.g. knee versus finger).
- Methods used for Item Response Theory, such as the Rasch model, have been proposed for analysing GS and PD data as a latent continuous measurement. However, these methods typically require a large sample size.

Other considerations specific to the design of the IACON study include:

- Can high PD scores be 'oversampled' when selecting images for scoring, in order to increase prevalence?
- Images vary in quality – is there a relationship between the image quality and the level of agreement?
- Should images be presented to raters per patient (in random order); per joint type; or completely at random?
- Can the IACON image bank be used as a training/assessment tool in different centres?
- Might knowledge of the oversampling method encourage raters to overestimate scores?

Discussion

The remainder of the meeting focused on the five discussion points below. A brief summary of the conclusions reached is presented. Relevant background information given to discussants during the meeting appears in the Appendix.

Discussion point 1

How would you approach the assessment of inter- and intra-reader reliability of scores for individual joints (0-3 for both GS and PD)? Would you provide statistics at the joint level for all joints, irrespective of joint type? What issues might you need to consider before calculating an overall Kappa value for all joints?

- *Decision as to joint-level/patient level depends on what is useful clinically/for research purposes*
- *Analyse different joints individually*
    - *Supplement with analysis that accounts for correlated joints*
    - *GS & PD likely to be correlated, could account for this, but start separately*
- *Possibility of using multilevel modelling to look at right and left sides; joint level/patient level/reader level*
- *Is Kappa the best measure to use? Descriptive statistics are potentially much more informative*
- *Stratified Kappa values can be calculated which account for clustering by combining Kappa values from several different strata together. Assuming patients vary in their level of disease activity (and thus in the prevalence of scores >0) it may not be valid to combine per-patient Kappa into an overall Kappa, as there is likely to be considerable heterogeneity.*

Discussion point 2

Considering the pertinent information, can you identify ways in which a total combined score for GS and PD might fail to satisfy the following axioms of measurement?

- Unidimensionality
    - *Perhaps GS and PD represent two different dimensions (unlikely but possible)*
- appropriate score category ordering
    - *If there is often disagreement over scores of 1 in some joints we might expect to see inconsistent category ordering*
- absence of differential item functioning (different score for same level of inflammation)

- *There might be DIF by age (due to scarring for example), rater (representing disagreement over scoring definitions), time (operator learning curve), possibly sex (if anatomical differences exist)*

- absence of residual correlation (dependence of score for one joint on the score of another)
    - *Within a joint, GS and PD might be residually correlated as it is not possible to score PD>0 if GS=0; joints of the same type on right and left side might show residual correlation; joints of the same type (MCPs) on the same side might show residual correlation (which could be a consequence of 'true' biological spread of condition or may reflect rater bias)*

Discussion point 3

If grey scale scores of 1 are not uncommon in healthy controls, and agreement is found to be poor over scores of 1 in some joints, how might this affect the use of grey scale as a screening tool, or total grey scale as an outcome? If ultrasound is to be useful in early disease, how might the measurement systems need to be adapted?

- *Scores of 1 could lead to false positives/overtreatment*
- *Impact depends on context*
- *Could raise threshold for treatment to deal with this*
- *Could incorporate some other elements like inflammatory markers*
- *Might want to weight different joints or drop some from scanning schedule*
- *Might need to consider moving towards quantitative methods of measurement*

Discussion point 4

The proposed Leeds study will use stored static images to assess reliability. Are there elements of ultrasound scanning that might contribute to measurement error which this design will not detect? How could these be investigated in a future study? What recommendations might you make about sample size and inclusion criteria?

- *May need to assess machine and patient factors prior to capturing data*
- *Operator expertise*
    - *Ability to identify important images*
    - *Ability to capture good quality images*
    - *Protocols at hospital level (training?)*
    - *Different scanners*

- *Patient cooperation*
- *Amount of gel*
- *Joint position etc*
- *Multiple scans within a patient, reflecting variability*
- *How much noise is upstream of scan visual vs downstream?*

Discussion point 5

In rheumatology it is common for small reliability substudies to be performed as part of the main study ("the reader will rescore 10% of the films to check reliability").

1. Is this common to other areas of research in your experience?
2. Are these ad hoc small scale reliability studies worthwhile?
3. Could / should we encourage researchers and reviewers to reconsider this approach?
4. Could / should one-off, larger scale reliability studies take their place?

- *The attendees agreed there is a general tendency to tack reliability onto existing studies*
- *There is a value in doing stand-alone studies, but would they be difficult to fund?*
    - *Perhaps not if direct benefit to future trials can be demonstrated (better training/identification of potential biases = more responsive outcomes)*
- *More emphasis should be placed on implications of reliability results than reporting them in passing*
- *Something is better than nothing and individual small studies could address different aspects of reliability, but…*
    - *How to combine different elements together?*
- *Changes in devices – will results of larger studies eventually be outdated, limiting the extent to which it will be valid to refer back to the results in future?*
- *For rare conditions / expensive modalities one solution might be sharing of images from different centres to create a sufficiently large image bank (with caveats surrounding differences in equipment etc)*

Acknowledgements

Thanks to Sue Mallett and Yemisi Takwoingi for arranging the venue, and to Liz Hensor, Richard Wakefield and the Imaging Studies section steering group for developing and advising on the scientific content.

Future meetings

The NIHR Statistics Group meeting planned for May 2016 is likely to include a session on the design of Imaging Studies. The next meeting of the Imaging Studies section is planned for October/November 2016. Details of both events will be announced through the NIHR-STATS-IMAGING mailing list and on the website (http://www.statistics-group.nihr.ac.uk).

References

Abramson J. WINPEPI updated: computer programs for epidemiologists, and their teaching potential. Epidemiol Perspect Innov. 2011;8:1.

Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: time to end malpractice? J Rehabil Med. 2012 Feb;44(2):97-8.

Hong H, Choi Y, Hahn S, Park SK, Park BJ. Nomogram for sample size calculation on a straightforward basis for the kappa statistic. Ann Epidemiol. 2014 Sep;24(9):673-80.

Iagnocco A, Naredo E, Wakefield R, Bruyn GA *et al*. Responsiveness in rheumatoid arthritis. A report from the OMERACT 11 ultrasound workshop. J Rheumatol. 2014 Feb;41(2):379-82.

Kottner J, Audigé L, Brorson S, Donner A *et al*. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011 Jan;64(1):96-106.

Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study: review and new results. Stat Methods Med Res. 2004;13:251-271.

Siemons L, ten Klooster PM, Taal E, Kuper IH, van Riel PL, van de Laar MA, Glas CA. Validating the 28-tender joint count using item response theory. J Rheumatol. 2011 Dec;38(12):2557-64.

Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005 Mar;85(3):257-68.

Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum. 2007 Dec 15;57(8):1358-62.

Appendix

Background information provided for discussion points.

Typically multiple joints are scanned per patient and scored 0-3 for hypertrophy (grey scale; GS) and 0-3 for vascularisation (power Doppler; PD). For example, the following joints might be scanned bilaterally:

Shoulder, elbow, wrist [global score, plus scores for individual radio-carpal, ulna-carpal, inter-carpal joints], metacarpophalangeal joints 1-5, proximal interphalangeal joints 1-5, knees, ankles, metatarsophalageal joints 1-5.

The different joint types vary in size and ease of access. Some are located close together.

In clinical practice, joints are typically scanned in a consistent order.

A characteristic of rheumatoid arthritis and some other arthritides is that there is symmetry in the joint involvement, particularly in the small joints of the hands.

It is not uncommon for healthy controls to have GS=1 in some joints. Previous studies suggest that agreement is poor over scores of GS=1 in some joints.

To an extent, PD score depends on GS score, as PD scores represent the proportion of the synovium that is affected, and GS scores represent the extent to which the synovium is enlarged. Joints cannot score PD>0 unless they also score GS>0.