



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

Dr Richard Wakefield & Dr Liz Hensor

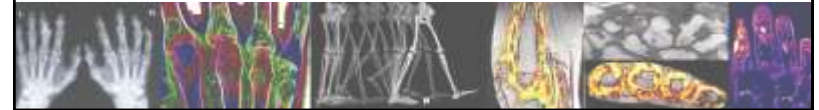
NIHR Leeds Musculoskeletal Biomedical Research Unit and
Leeds Institute of Rheumatic and Musculoskeletal Medicine

The Leeds Teaching Hospitals NHS Trust

NHS
National Institute for
Health Research

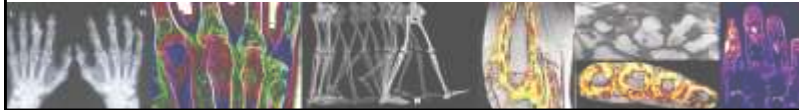
Outline

- What is inflammatory arthritis and why is it important
- Rheumatoid arthritis – synovitis as the target
- The role of US in detecting synovitis and the challenges of measurement
- Description of scoring methods
- The statistical challenges presented by the data
- The rationale for the planned reliability study (IACON)
- The selection of patients to be included
- The creation of the image bank



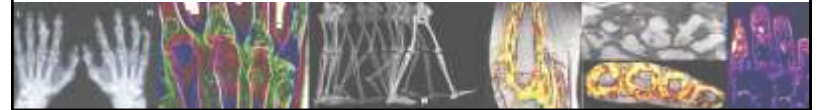
What is inflammatory arthritis (IA) ?

- Arthritis characterized by signs of joint inflammation – stiffness, pain, warmth and swelling
- Common examples include rheumatoid arthritis, psoriatic arthritis and gout
- Each disease has its own target for inflammation e.g. synovial membrane +/- tendons +/- ligaments



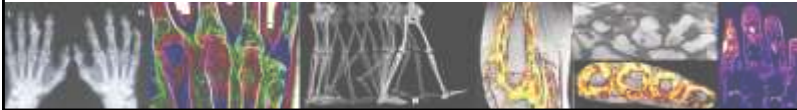
Why is it important ?

- If unrecognized, IA leads to increased risk of structural damage (soft tissue and bone), poorer functional outcome and disability
- Good evidence that early aggressive therapy improves outcome with there being a 'window of opportunity'
- Concept of 'Treat to Target' where aim for maximal suppression of disease

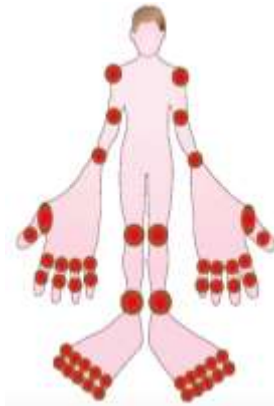


Rheumatoid disease

- Common cause of disability
- Chronic deforming arthritis + systemic features
- Polyarticular – multiple joints
- Autoimmune – antibodies
- Synovium
 - Site of initiation
 - Membrane that lines joint spaces and tendon sheaths
- If left untreated leads to tendon and bone damage



Polyarticular disease; synovial disease

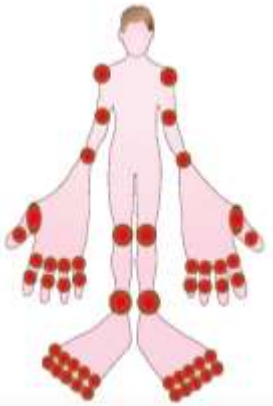


Predominantly a disease of wrists and 'small joints' of fingers and toes – 85% present this way

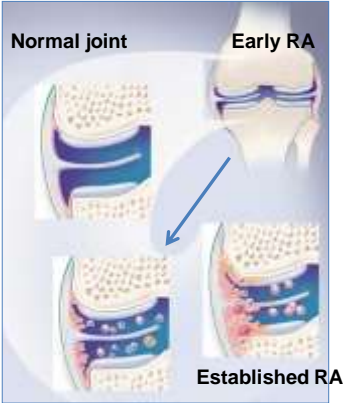
Also affects larger joints

Choy NEJM 2001

Polyarticular disease; synovial disease



Choy NEJM 2001



INFLAMMATION - SYNOVITIS



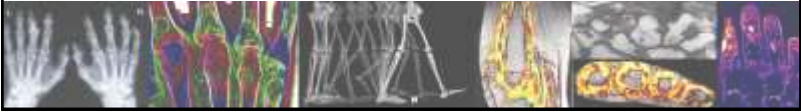
DAMAGE



BONE EROSION

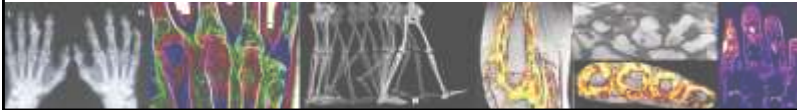


TENDON RUPTURE



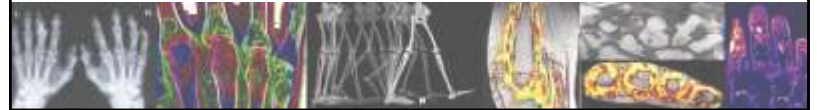
Limitations of clinical assessment

- Clinical examination (CE) insensitive and non specific
- Inflammatory markers (ESR, CRP) do not always correlate with CE
- Xray – insensitive to detect mild bone and cartilage changes



Need for new methods of assessment

- MRI – often described as gold standard – tomographic but lacks feasibility esp for multiple assessments
- US – widely available, immediate decision making, multi –joint assessment at multi-time points



The ultrasound equipment



Computer



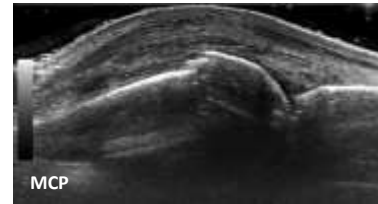
Probe
6-20 MHz



Gel

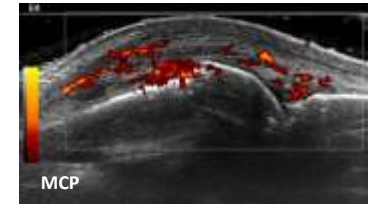
The US images....

Gray scale



qualitative
structural changes

Doppler (usually PD)



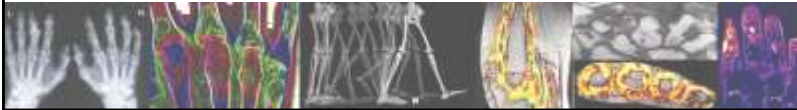
functional assessment
(vascularity)



Conventional scanning views

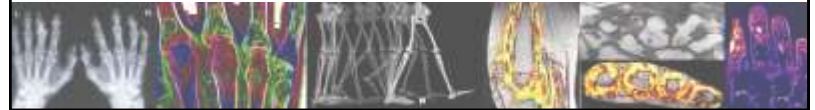


Different views taken / joint



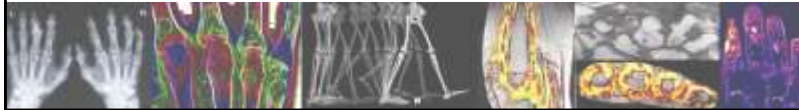
Conventional scanning views

- Shoulder – posterior GHJ, axillary GHJ (2)
- Elbow – anterior, radio-humeral, posterior (3)
- Wrist – midline, medial and lateral (3)
- MCPJ – dorsal and volar (2)
- PIPJ – dorsal and volar (2)
- Knees – midline, medial and lateral (3)
- MTPJ – dorsal only (1)



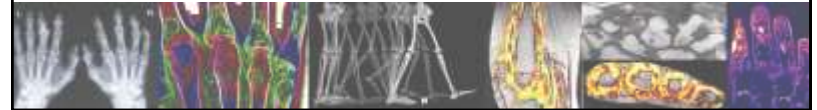
Scoring systems

- **Joint level (per individual joint)**
 - Binary (present/absent)
 - Semi-Quantitative
 - Commonest 0-3 (OMERACT-EULAR) – for GS and PD (or combined); pragmatic
 - Quantitative
 - Pixel counting
 - Resistive index of vessels (best of 3) – score 0-1
 - High RI (> 0.7) - normal
 - Low RI (< 0.7) - inflammation
 - Contrast agents – rate of uptake



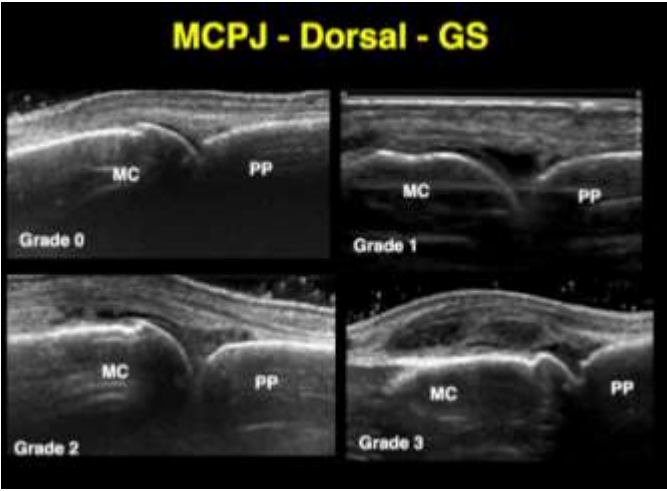
Scoring systems

- **Patient level (multi-joint)**
 - Joints chosen might depend on whether early (i.e. for diagnosis) or established disease (for monitoring)
 - Total scores for GS, PD, combined
 - Counts of joints



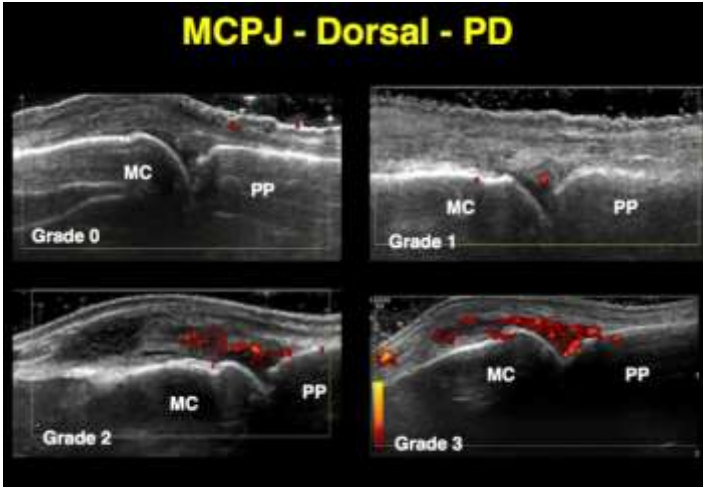
OMERACT-EULAR

MCPJ - Dorsal - GS

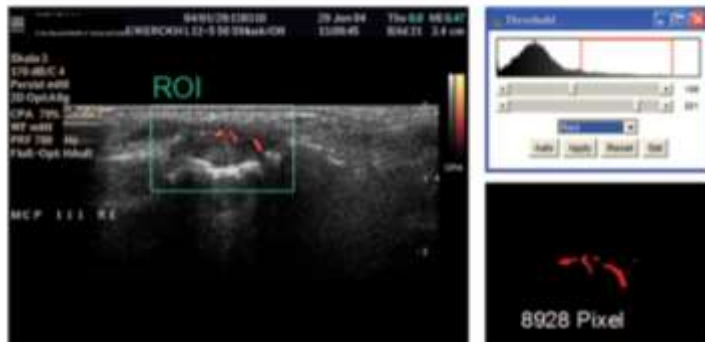


OMERACT-EULAR

MCPJ - Dorsal - PD



Pixel counting

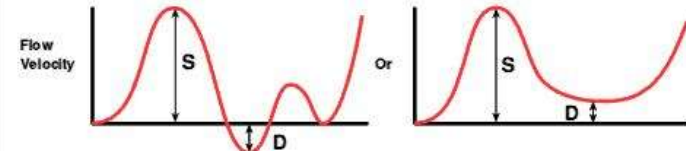


Albrecht K et al. Clin Exp Rheum 2007;25:630-38

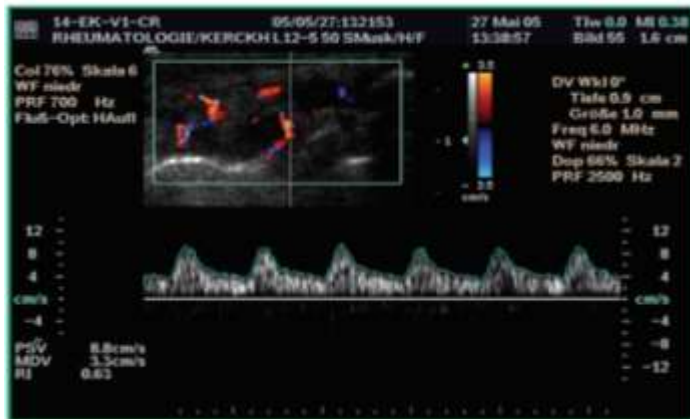
Resistive index

Resistive Index (RI) = $\frac{\text{Peak Systolic Velocity} - \text{Lowest Diastolic Velocity}}{\text{Peak Systolic Velocity}}$

$$(RI) = \frac{S - D}{S}$$



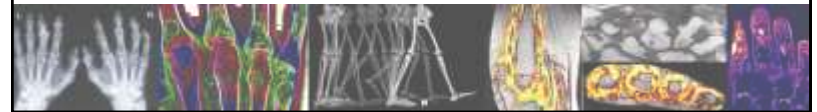
Resistive index



Albrecht K et al. Clin Exp Rheum 2007;25:630-38

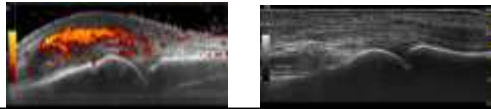
Challenges of US scoring

- **Physical limitations of ultrasound**
 - Unable to visualize whole joint (cf MRI- tomographic)
 - Sensitivity of GS and Doppler differs between machines
 - Torp-Pederson S et al. Arthritis Rheum 2015



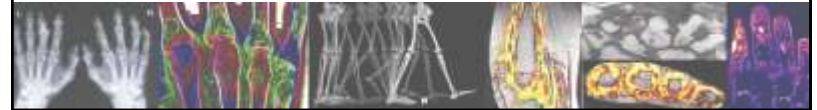
Challenges of US scoring

- **Standardization of exam**
 - Environment
 - Ambient temperature, (Ellegaard K et al Rheumatol 2009)
 - level of pre scan physical activity, (Ellegaard K et al, Rheum Int 2013))
 - pre scan use of medications eg steroids/ NSAIDS (Zayat A et al, ARD, 2011)
 - Position of joint (Zayat A et al. Rheum 2012)
 - Pressure of probe (Joshua F et al. Australasia Radiol 2005)
 - Position of probe (Vlad et al. BMC Musc Disorders 2011)



Knowing what is normal

- Small amounts of fluid and synovial hypertrophy are common in healthy controls
- Identifying which vessels are normal intra- and extra-articular vessels

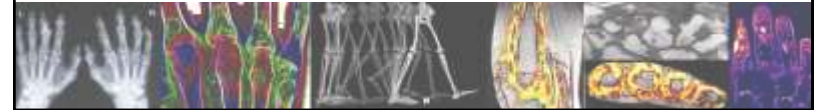
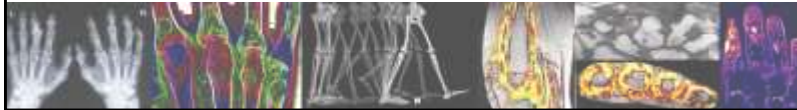


Methods for testing reliability

	Pros	Cons
Static	<ul style="list-style-type: none"> • Easy to acquire • Test multiple times 	<ul style="list-style-type: none"> • Only best images selected • Does not reflect acquisition
Video	<ul style="list-style-type: none"> • Captures whole joint • Test multiple times 	<ul style="list-style-type: none"> • Difficult to acquire in standardised way • Video might be biased to reader i.e. might concentrate on certain areas
Real-time (patient)	<ul style="list-style-type: none"> • Real life: tests reading and acquisition 	<ul style="list-style-type: none"> • Difficult to organise • Less suitable for multiple observers

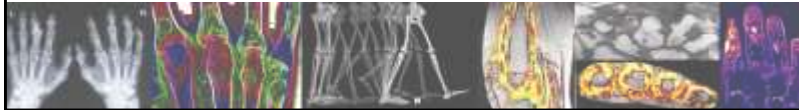
Outline

- What is inflammatory arthritis and why is it important
- Rheumatoid arthritis – synovitis as the target
- The role of US in detecting synovitis and the challenges of measurement
- Description of scoring methods
- The statistical challenges presented by the data
- The rationale for the planned reliability study (IACON)
- The selection of patients to be included
- The creation of the image bank



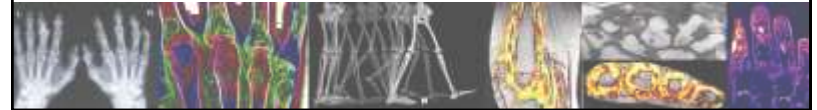
Statistical challenges

- How to deal with clustered data at the joint level
 - compartments within joints
 - joints within patients
- How to properly assess agreement in joints where inflammation is less prevalent



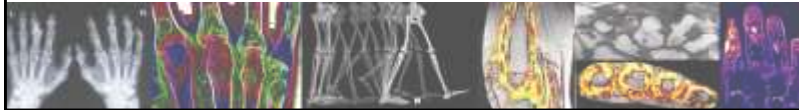
Statistical challenges

- How to summarise at the patient level
 - Two inter-related elements (GS and PD)
 - Ordinal scaling of total scores
 - Accounting for joint size



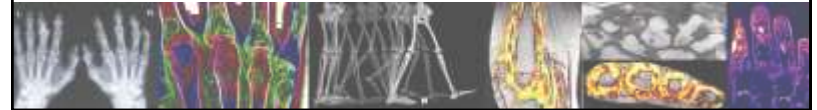
Clustered data

- How to combine GS/PD scores from different joint compartments into one score
 - Small joint eg MCPJ – volar and dorsal
 - Large joint eg knee – SPP, MJS and LJS
- Necessary to compare against CE
- Typically maximum score is used
 - Treatment is given at the joint level



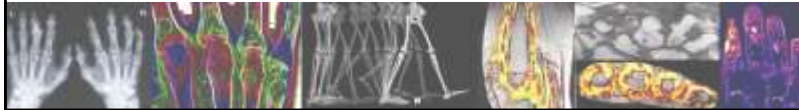
Clustered data

- How to deal with clustering of joints within patients when assessing agreement at joint level
- Stratified Kappa is possible
 - Weighted by inverse of variance (Fleiss 2003)
 - Common correlation model (Donner & Klar 1996)
 - Weighting by stratum size (Barlow 1991)



Low prevalence in some joints

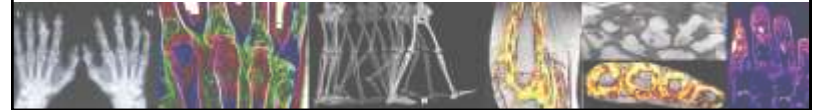
- How to assess operator agreement in joints that rarely affected
 - Agreement may vary by joint type
 - Prevalence of inflammation varies by joint type
 - Hard to measure agreement in less commonly affected joints; inflammation may be absent in sample
 - May require careful selection of individuals



Patient-level data

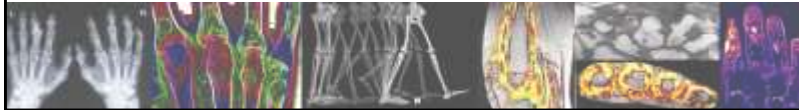
- Total GS / total PD (summated 0-3 scores)
- Counts of joints with GS present / PD present
- Combined GS and PD

GS	PD			
	0	1	2	3
0	0			
1	1	1	2	3
2	2	2	2	3
3	3	3	3	3



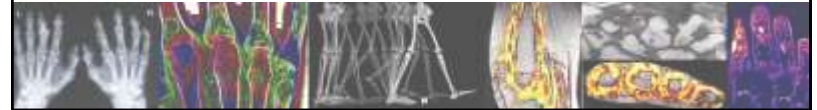
Ordinal scaling

- Although described as semi-quantitative at joint level, scores cannot be considered interval-scaled
 - GS: Absent; mild; moderate; marked hypertrophy
 - PD:
 - **Grade 0** = no flow in the synovium (gray scale area)
 - **Grade 1** = up to 3 single spots signals or up to 2 confluent spots or 1 confluent spot + up to 2 single spots
 - **Grade 2** = vessel signals in less than half of the area of the synovium (< 50%)
 - **Grade 3** = vessel signals in more than half of the area of the synovium (> 50%)



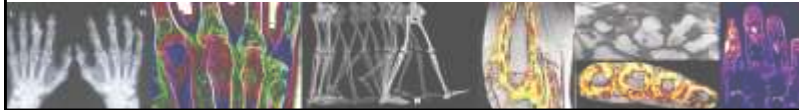
Ordinal scaling

- Ordinal scales not valid for longitudinal changes
- Limits usefulness of US scores as clinical trial outcomes



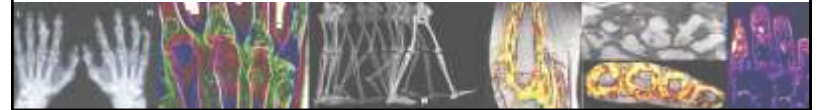
Ordinal scaling

The practice of misusing ordinal scales as though they were interval measures was re-emphasized by Merbitz and colleagues (2) in their seminal paper “Ordinal scales and foundations of misinference” [...]. They went on to state that if ordinal scales are manipulated mathematically, the results are not logically valid, and conclusions may therefore be misleading. They concluded that readers should not permit the lack of a complete interval or ratio level functional outcome scale to make the practice of misinference socially acceptable.



Accounting for joint size

- Should joints be weighted in total scores and counts?
- Lansbury & Haut 1956
 - Used component bone ends of skeleton joints
 - Carefully covered cartilage areas with Al foil
 - Weighed several times
 - Converted to surface area

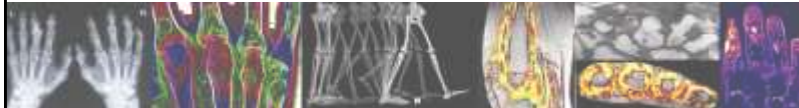


Accounting for joint size

TABLE 2.—VALUES FOR INDIVIDUAL JOINTS EXPRESSED IN WHOLE NUMBERS FOR CALCULATING TOTAL AMOUNT OF JOINT INVOLVEMENT IN RHEUMATOID ARTHRITIS.

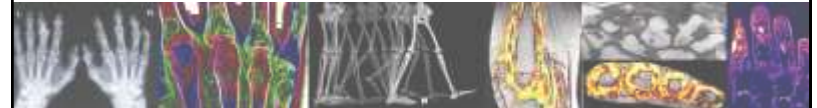
<i>Upper Extremity</i>		<i>Lower Extremity</i>	
Each terminal interphalangeal joint.....	1	Each terminal interphalangeal joint.....	0.5
Each proximal interphalangeal joint.....	2	First proximal interphalangeal joint.....	3
Each metacarpophalangeal joint.....	5	Remaining proximal interphalangeal joints.....	1
Each carpometacarpal joint.....	4	First metatarsophalangeal joint.....	8
Transverse intercarpal joint area.....	15	Remaining metatarsophalangeal joints.....	5
Wrist.....	15	Tarsometatarsal joint area.....	25
Elbow.....	52	Transverse intertarsal joint area.....	12
Shoulder.....	45	Talonavicular-calcaneocuboid.....	91
Acromioclavicular.....	4	Talocalcaneal (subtalar).....	18
Sternoclavicular.....	12	Ankle.....	35
Temporomandibular.....	4	Knee with patella.....	104
		Hip.....	82

To determine percentage of total joint involvement, add up the values for each affected joint, place a decimal point before the last digit of the total figure.

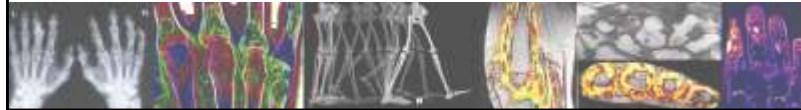
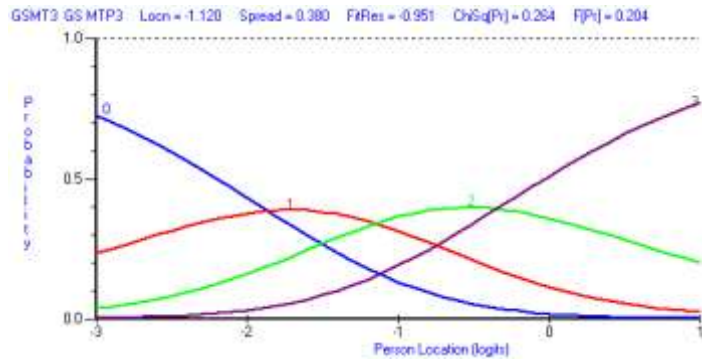


Item response theory

- Rasch model (single parameter model)
 - Probabilistic form of Guttman scaling
- Model tests data for measurement axioms:
 - Unidimensionality (required for valid total score)
 - Invariance of item ordering
 - Appropriate category ordering
 - Absence of differential item functioning
 - Absence of residual correlation

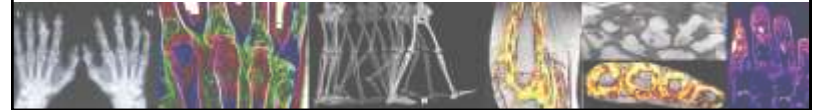


Item response category ordering

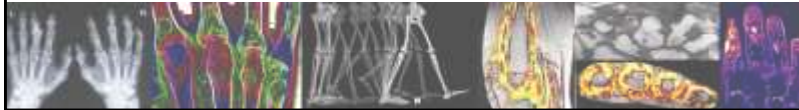
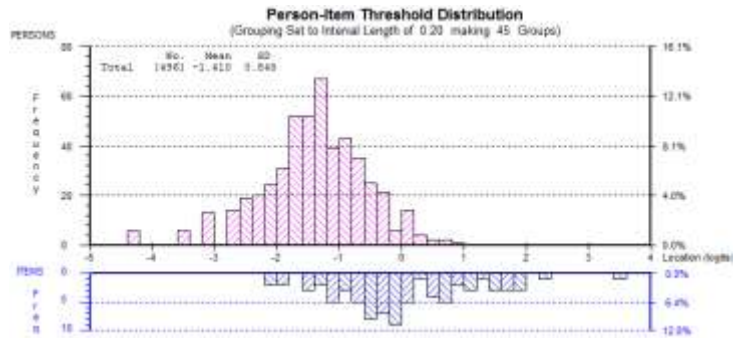


Item response theory

- Targeting of persons and items
- Reliability
 - Extent to which scale can reliably distinguish between people with different levels of the latent trait
- Sample size (n=200 ideally)
- Software: RUMM, WINSTEPS, Stata, SAS

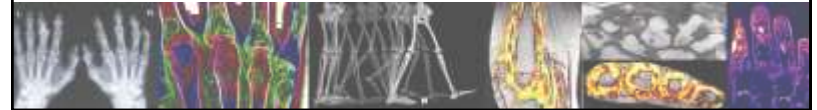


Example of poorly targeted scale



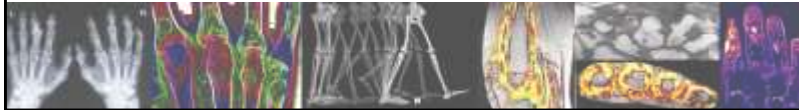
Rationale for the Leeds study

- Small scale reliability studies common
 - Often added onto an existing study
 - Rarely powered
 - Inclusion criteria often at odds with requirements for reliability
- Potentially misleading & wasteful of resources



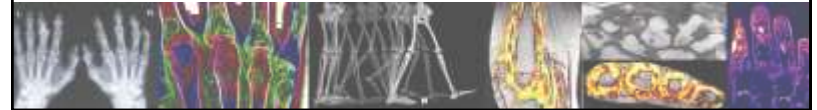
The IACON cohort

- Leeds Inflammatory Arthritis CONTinuum
- Cohort study of early IA
- >1200 patients since 2010
- US at baseline, 6m, 12m then annually
- Joints scored by sonographers for GS and PD
- View selected and stored



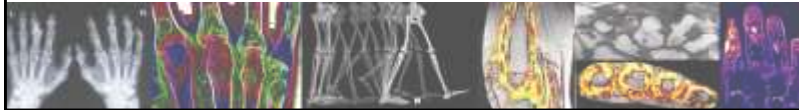
The IACON cohort

- The following joints are captured bilaterally:
 - Elbow
 - Wrist
 - Metacarpophalangeal (MCP) joints 2 & 3
 - Proximal interphalangeal (PIP) joints 2 & 3
 - Knee
 - Ankle
 - Metatarsophalangeal (MTP) joints 1 - 5



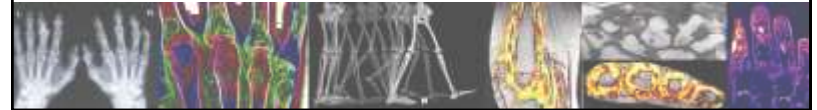
Study design

- Initially designed to assess reliability of the Leeds US team
- At least 5 different operators
- Each to score all joints twice at an interval of at least 2 weeks
- Intra-operator repeatability to be assessed
- Inter-operator reliability to be assessed overall (all operators) and relative to single reference score from expert operator



Study design

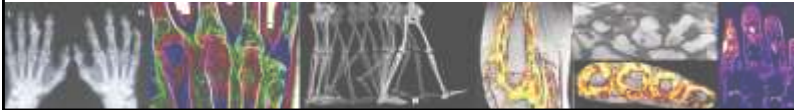
- Analysis of joint-level data
 - Quadratic-weighted Kappa by joint type
 - Maximum attainable Kappa
 - Proportions of positive agreement per category
- Analysis of patient-level data
 - Bland-Altman plots (each operator vs expert)
 - Kendall's coefficient of concordance
 - ICCs (potentially using rank-based versions)



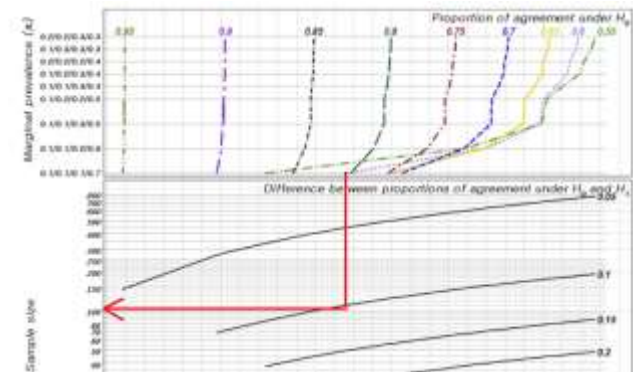
Study design

- Sample size: Kw for joint-level data
 - Minimum required $n = 2k^2 = 32$
- Sample size: ICC for patient-level data
 - Methods of Shoukri et al. 2004
 - Stata module sampicc
 - $\rho_0 = 0.6, \rho_1 = 0.7, \text{reps} = 5, \alpha = 0.05, \beta = 0.20: n = 99$
 - 95% CI width 0.15

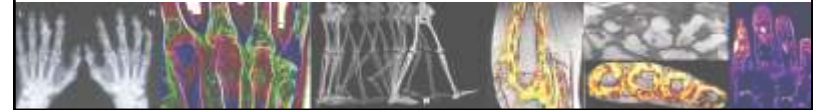
Hong et al 2014



Sample size for K: 4 nominal categories

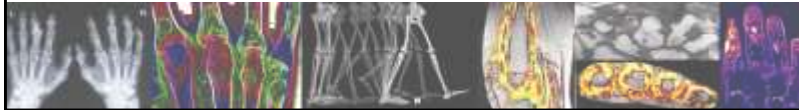


Hong et al 2014

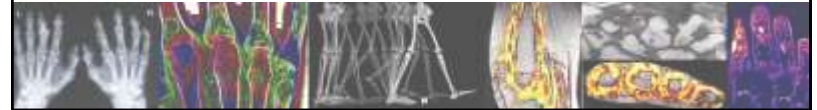
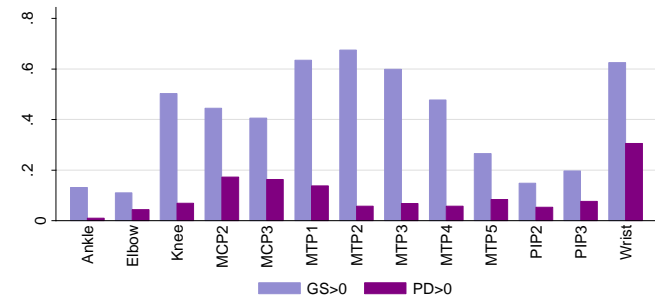


Study design

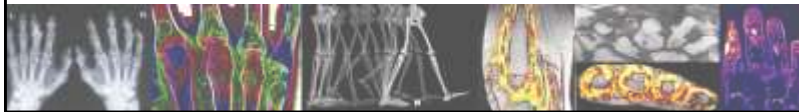
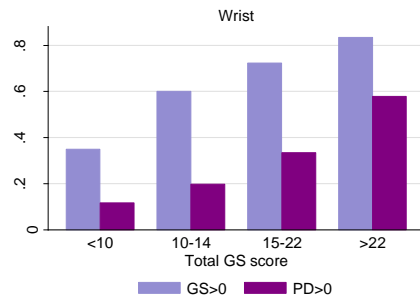
- Sample size: Proportion of positive agreement
 - Could use rules of thumb
 - to obtain stable estimate of a proportion: $n=60$
- Calculated per category, per joint
- Four score categories (0, 1, 2, 3)
 - 240 scores needed (= 120 joints)
 - Total number of patients required 60 if joints on left and right sides pooled
 - Note that this is 'best case' score prevalence



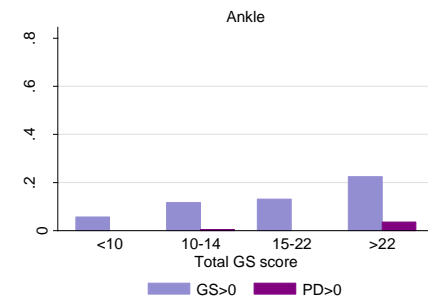
Variation in prevalence



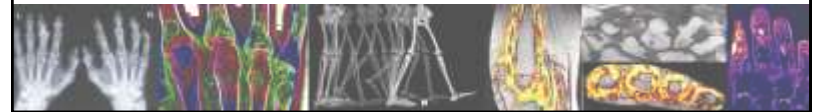
Is there evidence of Guttman scaling



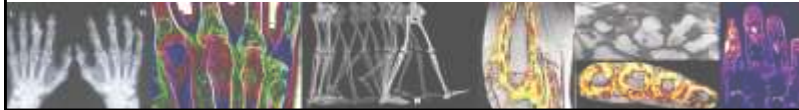
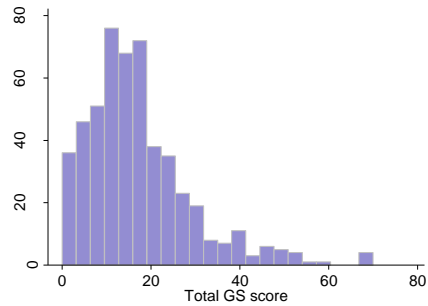
Is there evidence of Guttman scaling



We might expect higher proportion of ankle joints with PD>0 in a cohort with more severe inflammation (ankle = 'difficult item')

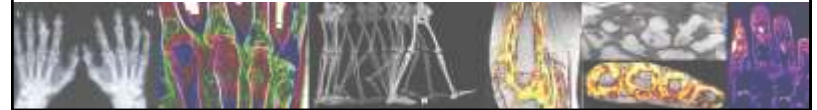


Total GS scores low in our sample



Effect of sample size

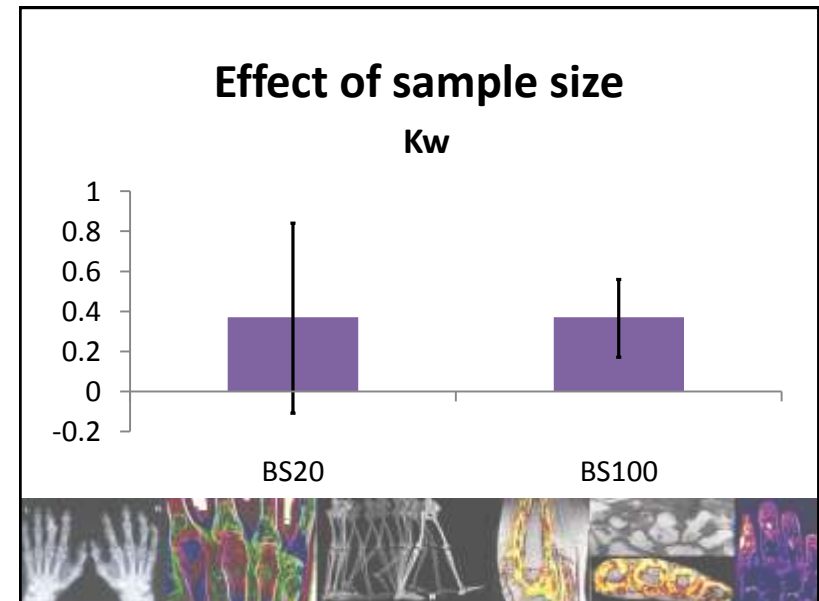
- PD in MCP2 as example (L and R as 'raters')
- Data from 514 joints available
- Bootstrapped using 1000 reps, size 20 or 100
- In full sample (n=514):
 - PEA = 80%
 - Kw = 0.37
 - 33 out of 1028 'ratings' score PD=3

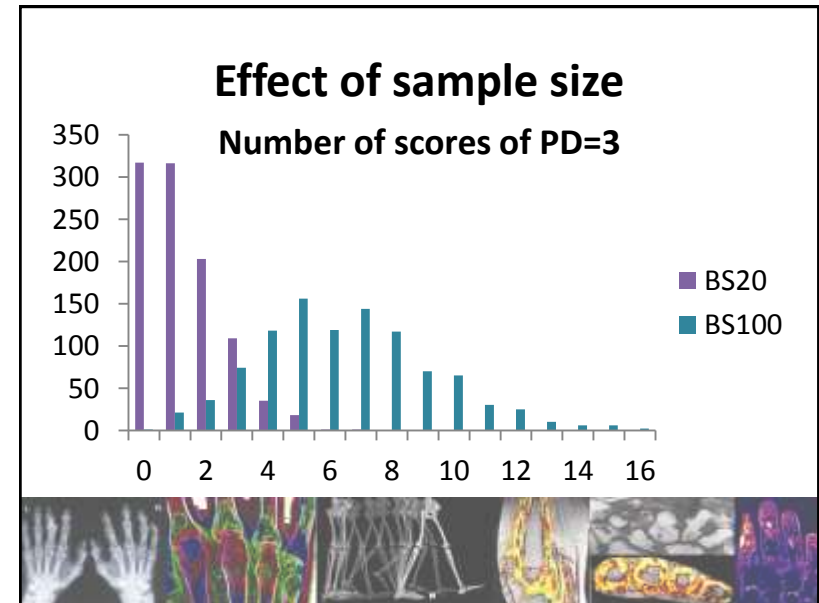
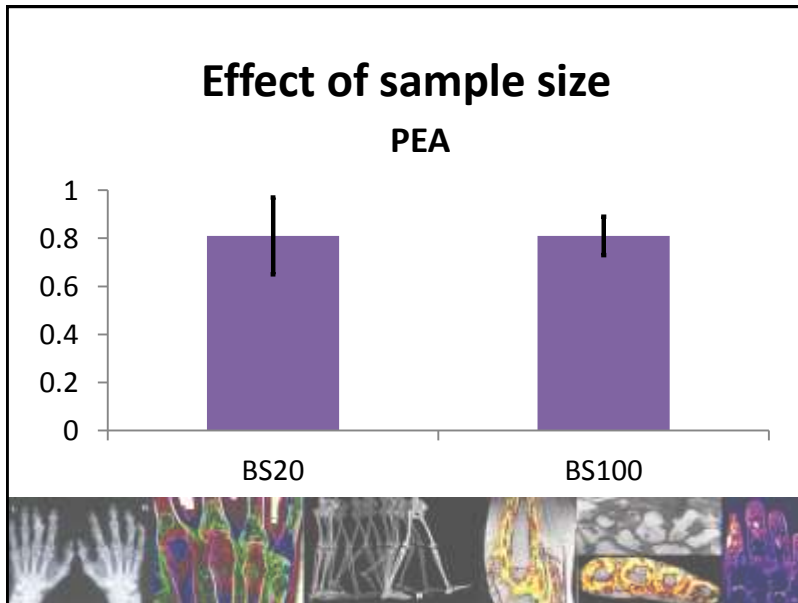


Effect of sample size

PD score 'rater 1'	'rater 2'				Total
	0	1	2	3	
0	387	15	12	8	422
1	22	8	5	0	35
2	13	6	17	3	39
3	7	4	3	4	18
Total	429	33	37	15	514

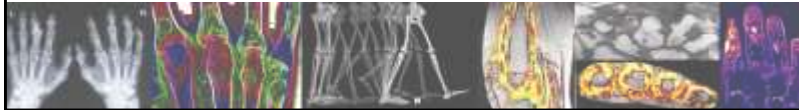
Ppos0 = 91%; Ppos1=24%; Ppos2=45%; Ppos3=24%





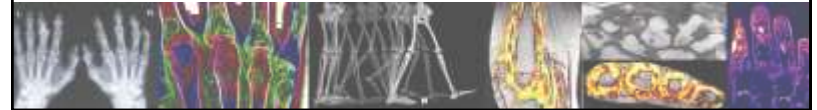
Selection of patients

- Improve distribution by oversampling PD>0
 - Calculate maximum PD per joint (right or left)
 - Rank joint types according to prevalence of PD>0
 - Starting with least prevalent joint and category, sample iteratively according to whether ‘ideal’ joint sample size attained, given current selection, until required n



Selection of patients

- With 100 of each joint and 4 categories, ideal n is 25 per score category
- Start with least prevalent joint and category (here PD=3 in ankle); if ≤ 25 patients with a score of 3 available, select all of them
- Move to second least prevalent joint and repeat; at each stage query how many more patients are required to reach n=25 for that joint (if possible)
- If more than enough patients available, choose enough at random to reach n=25
- Repeat for PD=3 in each joint type, then start with PD=2 in least prevalent joint again until required N reached



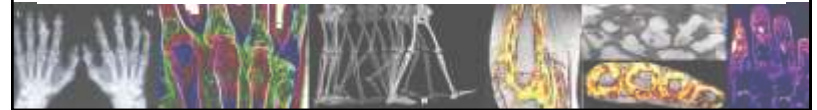
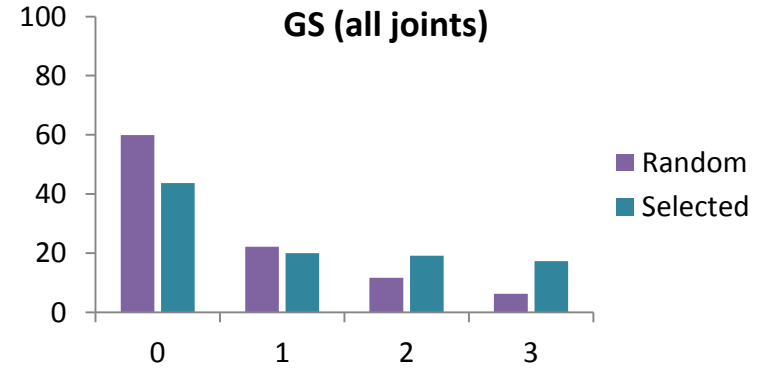
Selection of patients

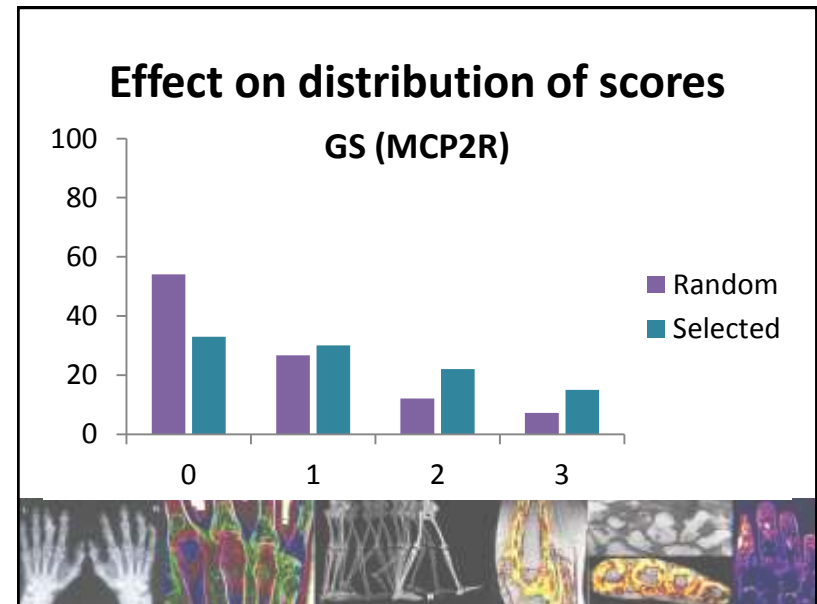
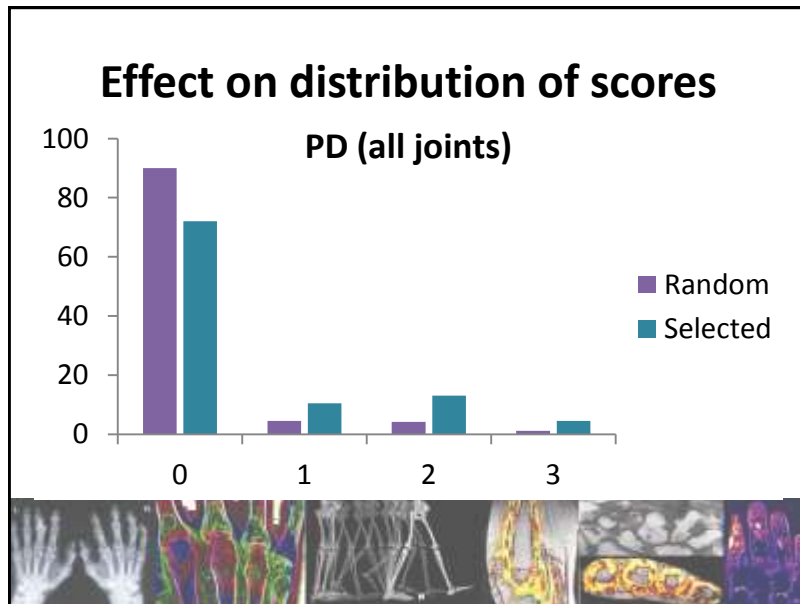
Joint	PD=0	PD=1	PD=2	PD=3
Ankle	505	5	3	1
Elbow	479	15	18	2
PIP2	471	17	22	4
MTP2	467	27	14	6
MTP4	467	22	22	3
Knee	465	31	14	4
MTP3	461	23	24	6
PIP3	458	18	22	16
MTP5	453	29	23	9
MTP1	410	54	42	8
MCP3	388	52	52	22
MCP2	387	45	53	29
Wrist	312	72	104	26

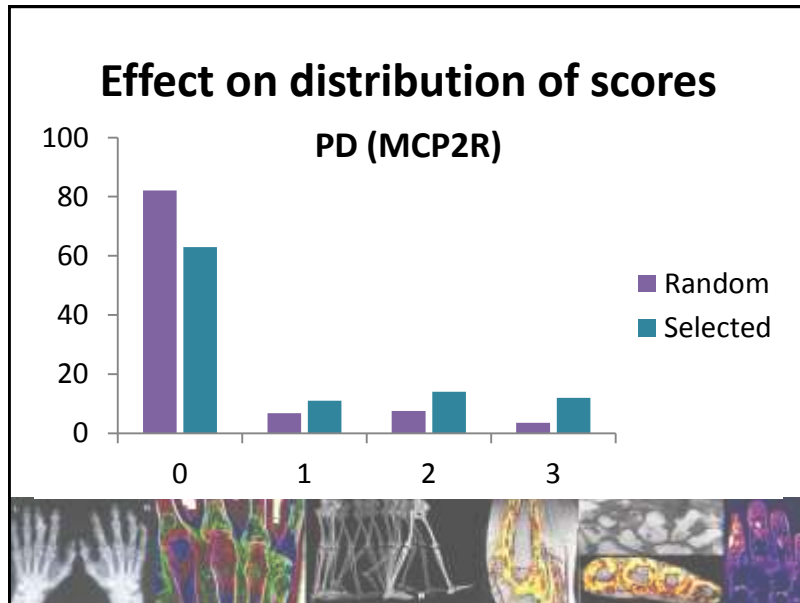
Start here



Effect on distribution of scores

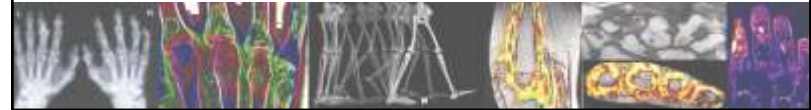






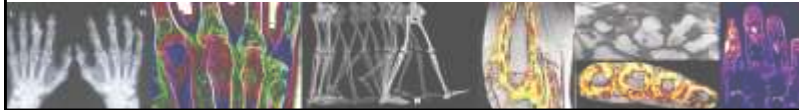
Selection of images

- Image quality as an outcome
 - Important to assess operator ability to grade quality
- Best available image will be selected; some poor quality images will be included
 - May be possible to collate pool of images of varying quality for separate assessment of agreement over quality



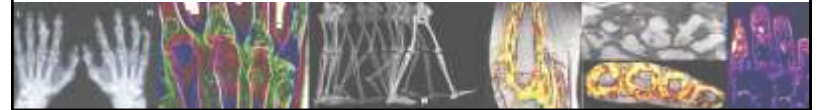
Creation of image bank

- Images from 2600 joints in 100 patients
 - 6057 DICOM files = 15.65GB
 - Reduces to 1.47GB when converted to JPEGs
- Anonymisation and cataloguing
- Learning management system
- Hosting costs



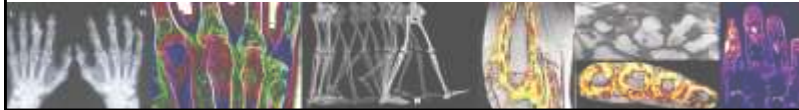
Creation of image bank

- Presentation of images in storybook
 - Per patient, in order
 - Per patient, random order
 - By joint type
 - Completely at random
- Facility to bookmark progress
- Potential training and assessment tool across different centres




Future work

- Comparison of semi-quantitative scores with quantitative
- Comparison of reliability in early and late IA
- Assessment of in vivo scoring performance



Acknowledgements

- The Leeds ultrasound team
Jane Freeston, Laura Horton, Alwyn Jackson,
Jacqueline Nam, Ai Lyn Tan, Ahmed Zayat
- Our colleagues at LIRMM and LMBRU

The Leeds Teaching Hospitals 
NHS Trust


UNIVERSITY OF LEEDS

