Oxford Biomedical Research Centre
Enabling translational research through partnership

National Institute for Health Research
Office for Clinical Research Infrastructure

NIHR Statistical Group
Imaging in Translation Research Meeting

Challenges in the design and analysis of studies evaluating imaging modalities

POSTER BOOKLET

Pembroke College, Oxford, OX1 1DW
Tuesday 1st October 2013

nihr-stats@kcl.ac.uk

Oxford Biomedical Research Centre
Enabling translational research through partnership

NHS National Institute for Health Research
Office for Clinical Research Infrastructure

Contents

Analysis of MRI measurements in a longitudinal observational study of Huntington's Disease: Track-HD and Track-ON

Amy Mulick Cassidy, Ruth Keogh, Chris Frost.
Department of Medical Statistics, London School of Hygiene and Tropical Medicine (LSHTM)

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

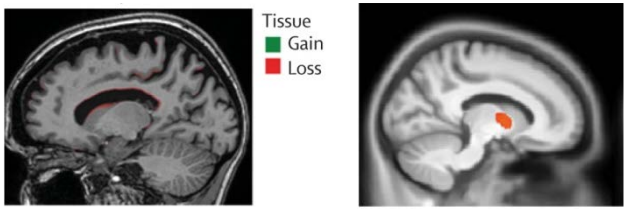
STUDY DESCRIPTION: TRACK-HD

Track-HD[1-4] is a multi-centre, multi-national observational study of individuals who have inherited the Huntington's Disease (HD) genetic mutation, together with a control group. The HD group was recruited to comprise individuals who have already had the clinical onset of disease symptoms and individuals who are pre-manifest, some of whom are predicted to be many years from onset.

The primary goal of Track-HD was to identify biomarkers for use in future clinical trials in HD. It is of particular interest to identify biomarkers for pre-manifest individuals who do not yet show outward signs of disease.

Participants have attended 4 annual study visits, at which a range of clinical measurements were made (motor function, cognition, neuropsychiatry). At annual visits participants also underwent **structural Magnetic Resonance Imaging (MRI)** (T1- T2- and diffusion weighted) on 3T scanners. Main imaging outcomes are:

- Whole brain, ventricular and caudate **volumes**
- Whole brain, ventricular and caudate **atrophy**, calculated using the **boundary shift integral (BSI)**[5]
- Grey and white matter **atrophy**, calculated using **voxel-based morphometry (VBM)**[6]



Visualisation of 'direct' measures of 12-month change via the BSI (left), which integrates a series of 'difference' images (red areas) obtained by accurate positioning of identical scans at different time points, and VBM (right), which integrates a series of voxel-compression maps from images matched using fluid registration. [2]

STATISTICAL CHALLENGES

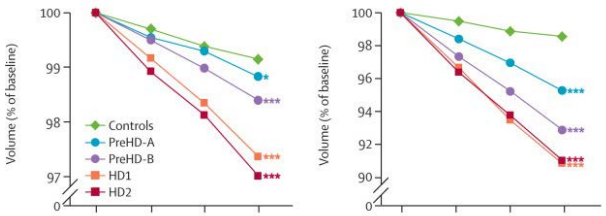
Measures of atrophy calculated with the BSI or VBM are 'direct' measures of change from scans taken at two points in time. This is in contrast to measuring change by subtracting measurements made at two time points. Measures of direct change are less prone to error and hence it may be possible to increase the statistical power of randomized clinical trials with atrophy as the outcome by utilizing study designs involving direct measures of *change*. Linear mixed models are used to estimate changes in brain region volumes over time. However to **apply linear mixed models to repeated measures of direct change** requires modification because there is an expected negative correlation between consecutive differences within the same individual.

For a change C in person *i* between the *j*th and *k*th visit, with random slope b and subject effect u, we specify the following mixed effects model:[7]

$$C_{ijk} = (\beta + b_i)(t_{ik} - t_{ij}) + u_{ik} - u_{ij} + d_{ijk},$$

$$u_{i.} \sim N(0, \sigma^2_u),\ d_{ijk} \sim N(0, \sigma^2_d).$$

Here we make explicit that random error in a measure of 'direct' change comprises variability introduced at the time of the first measurement (u_{ij}), variability introduced at the time of the second measurement (u_{ik}) and error in making the actual measurement (d_{ijk}). The sharing of error terms between the 'direct' change from, for example, the first to the second time points and that from the second to the third time points allows for anticipated negative correlation between consecutive changes.



Adjusted mean whole brain (left) and caudate (right) atrophy at 12, 24 and 36 months for controls, premanifest subjects ≥10.8 or <10.8 years from predicted disease onset (preHD-A and -B, respectively) and diagnosed subjects at Stage 1 and 2. [4]

CURRENT AND FUTURE CHALLENGES: TRACK-ON

Track-On is a follow-on study from Track-HD, with the aim of making further advancements in understanding the pre-manifest HD population. Pre-manifest individuals and controls attended two Track-On study visits (2012-13). In addition to structural MRI, **functional Magnetic Resonance Imaging (fMRI) was performed in resting state and while performing tasks**.

Pre-manifest HD individuals were found in Track-HD to have brain atrophy, but they continue to perform well in clinical tasks, suggesting existence of compensatory mechanisms in the brain. **A primary goal of Track-On is to make use of fMRI data to identify compensatory mechanisms.** Challenges we face include:

- High-dimensional data on connectivity between brain regions, including repeated measures.
- Using repeated measures of 'activation' data (relating onset of movement to changes in blood oxygenation) from task-based fMRI, where we anticipate error and potentially large variability over time.
- Integrating structural MRI and fMRI measures.
- Making use of resting state fMRI data when there is no corresponding clinical task.

1 Tabrizi SJ et al. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. *Lancet Neurol* 2009; 8: 791-801.
2 Tabrizi SJ et al. Biological and clinical changes in premanifest and early stage Huntington's disease in the TRACK-HD study: the 12-month longitudinal analysis. *Lancet Neurol* 2011; 10: 31-42.
3 Tabrizi SJ et al. Potential endpoints for clinical trials in premanifest and early Huntington's disease in the TRACK-HD study: analysis of 24 month observational data. *Lancet Neurol* 2012; 11: 42-53.
4 Tabrizi SJ et al. Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. *Lancet Neurol* 2013; 12: 637-649.
5 Freeborough P & Fox N. The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. *IEEE Trans Med Imaging* 1997; 16 (5), 623-9.
6 Ashburner J & Friston K. Voxel-based morphometry--the methods. *Neuroimage* 2000; 11 (6 Pt 1), 805-21.
7 Frost C et al. The analysis of repeated 'direct' measures of change illustrated with an application in longitudinal imaging. *Stat Med* 2004; 23:3275-3286.

Measuring low back inter-vertebral motion patterns with quantitative fluoroscopy

Fiona Mellor BSc(Hons), Alan Breen DC PhD, MIPEM
Institute for Musculoskeletal Research and Clinical Implementation
AECC, Bournemouth BH5 2DF
fmellor@aecc.ac.uk

Introduction

Back pain represents the 6th largest cause of global disability and is assumed to be mainly mechanical [1]. The functional stability of the spine depends on its ability to maintain patterns of displacement under normal loads [2], but is difficult to measure at an inter-vertebral level in vivo. Quantitative fluoroscopy (video x-rays) can extract and analyse such kinematic information with a similar radiation dose to static radiographs.

Methods

Recording

Forty patients with suspected mechanical back pain and 40 healthy controls were imaged from L2-5 using a Siemens Avantic portable C-arm fluoroscope recording at 15fps while being moved passively in the supine (left and right) and lateral decubitus positions (flexion extension) over 40 degrees global motion.

Processing

DICOM sequences were transferred to a desktop computer and 800 images per direction underwent frame to frame registration using bespoke software written in Matlab (v R2007b, the Mathworks Ltd.). Inter-vertebral rotational patterns were generated and then normalised to control for variations in overall motion.

Analysis

Inter and intra-observer repeatability were calculated from maximum rotational ranges in 10 participants. Reference regions (+/-2Sd) for normalised ranges were derived from controls and applied across the motion . Motion pattern variation was quantified by summing the normalised ranges between the 3 levels and calculating their variances. Variances of the 4 directions were then summed to produce a combined normalised range variance (CNRV) value.

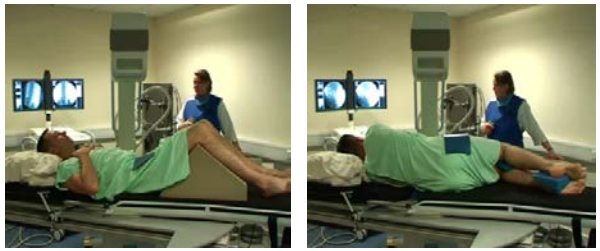


Figure 1: Image Acquisition

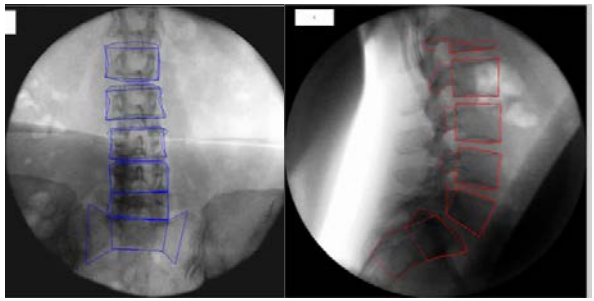


Figure 2: Image Tracking

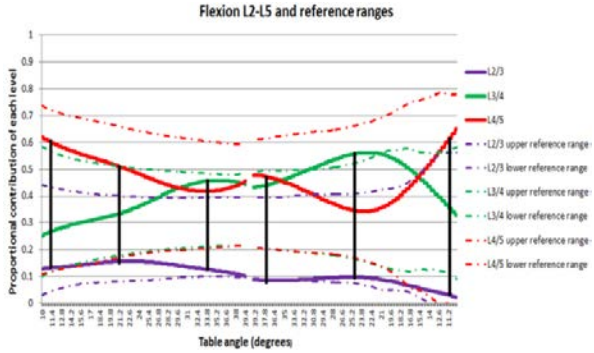
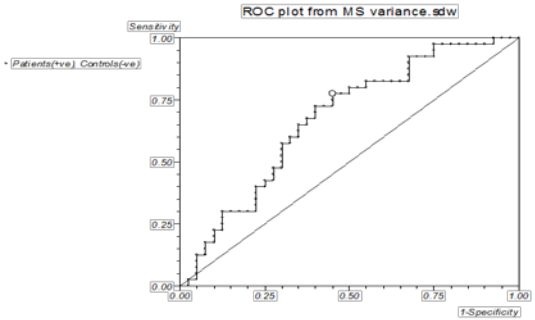


Figure 3: Patterns of normalised rotational motion from 3 levels in one direction with reference regions (+/-2SD) and examples of ranges between levels (black bars)

Results

For repeatability, inter and intra-observer ICCs (2,1) were 0.956 and 0.990 respectively and SEMs (consistency) 0.15° and 0.07°. 36/40 patients had at least one level that exceeded the reference intervals for normalised range (X^2=4.781, p=0.028). CNRV values were significantly higher in patients (0.011 vs 0.008, Mann-Whitney 2-sided p=0.008). ROC analysis found sensitivity 78% (60-95), specificity 55% (37-73) and AUC=0.672 for back pain.



Conclusion

Motion pattern variation was significantly higher in patients with persistent non-specific back pain implying a mechanical subgroup.

Acknowledgements

This research was funded by the NIHR-Clinical Academic Training scheme. Trial registration number UKCRN11478
Ethical approval was obtained from the UK NRES (09/HOSO2/99)

References

1. Murray, C.J.L., et al., Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. The Lancet, 2012. 380(9859): p. 2197-2223.
2. White, A.A., Panjabi, M.M., Clinical Biomechanics of the Spine 1990, Philadelphia: JB Lippincott.

WATER CHEMICAL SHIFT IN CHILDHOOD BRAIN TUMOURS: WHAT DOES IT MEASURE?

B. Babourina-Brooks[1], M. Wilson[1,2], T.N. Arvanitis[2,3], A.C. Peet[1,2], and N.P. Davies[1,4]

[1]School of Cancer Sciences, University of Birmingham, Birmingham, West Midlands, United Kingdom, [2]Birmingham Children's Hospital NHS Foundation Trust, Birmingham, West Midlands, United Kingdom, [3]School of Electronic, Electrical & Computer Engineering, University of Birmingham, Birmingham, West Midlands, United Kingdom, [4]Imaging & Medical Physics, University Hospitals Birmingham NHS Foundation Trust, Birmingham, West Midlands, United Kingdom

Introduction

- New prognostic biomarkers needed for childhood brain tumours
- Tumour micro-environment (ME) has been shown to be relevant to treatment response & prognosis [1,2]
- MRS water proton resonance frequency values are related to ME factors including temperature, chemical exchange, ionic strength. [2]
- MRS measures of brain tumour ME relatively unexplored

Aim: Investigate water PRF (ME) measure through known metabolite biomarkers

Group	Tumour Type	Before QC	After QC
Gliomas	Pilocytic astrocytoma (grade 1)	18	16
	Diffuse astrocytoma (grade 2)	5	4
	Gliometosis cerebri (grade 3)	1	1
	Glioblastoma (grade 4)	2	1
PNETs	Medulloblastoma (grade 4)	21	19

Analysis

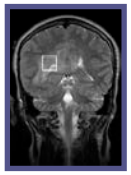
- Spectra were processed and metabolite concentrations calculated in TARQUIN [3]
- Ref water fit at ~4.68 ppm , Cr ~3.03ppm & Cho ~3.22ppm
- NAA not often present, Δ(H2O-Cr), and Δ(H_2O-Cho), calculated for each spectrum and compared
- PRF's were standardised to relative water PRF
  - $PRF_{rw} = \Delta\ (H_2O_{act} - Met_{act}) + Met_{ref}$
- PRF_{rw}(Cho) & PRF_{rw}(Cr), A^2 weighted averages were calculated [4,5] & compared
- Performed correlation test between Met Conc & PRF
- All significance determined by Student T tests

Method

- 1.5 T, single-voxel PRESS, TR 1500ms, TE 30ms
- Patient cohort information Table 1
  - Data from 20 apparently normal brains, BG & WM

Results & Discussion

- Sig PRF shift differences between PNETS & gliomas (Fig 2)
- Amp W increased the sig for healthy vs. Tumour, PNETs vs. gliomas
- PRF shift is dependent on temperature, ionic strength and fast proton exchange effects. [6]
  - varies in tissue types
- Met conc vs PRF within groups ($p<0.05$) (Fig 3)
  - PNET's = Lip, Gly -> necrosis/apoptosis & malignancy
  - Glioma = -NAA, -PCH -> malignancy markers
  - PRF changes most likely ME rather than temperature
- Met conc across group trends agreed with Davies et al [7]
  - PRF changes due to ME and temperature
- PRF correlates to malignancy markers within a tumour group



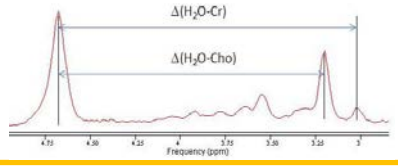


Figure 1. Spectral localisation image and analysed spectrum for a patient with Gliomatosis Cerebri (grade 3) showing a well behaved partially suppressed water peak suitable for water-PRF shift measurement relative to reference metabolite peaks

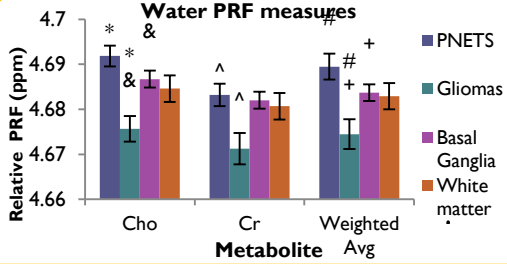
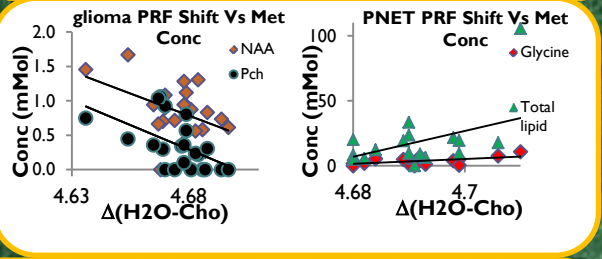


Figure 2: Relative water PRF against metabolite reference averaging graphs for PNETS, Gliomas, Basal Ganglia and White Matter cohorts. Error bars are standard error and significances indicated.

Figure 3 (below): PRF shift vs met Conc within the Glioma(left) and PNET group(right).



Conclusions

- Study suggests that water PRF varies between tumour types and for each type compared with normal brain in children
- Differences within tumour types are likely ME rather than temp, however across groups temp contribution increases
- PRF correlates with malignancy MRS markers
- Further work into the unique information the PRF can provide is required

Abbreviations: A^2, Amplitude squared; BG, Basal Ganglia; Cho, choline; Cr, creatine; ME, Micro-Environment; Met, Metabolite; NAA, N-acetyl-aspartate; PNET, Primitive Neuroectodermal Tumour; PRESS, point-resolved spectroscopy; PRF, proton resonant frequency; SE, standard error; WM, White Matter;

Acknowledgements: Funded by Birmingham Children's Hospital Research Foundation. The views expressed in this poster are those of the authors and not necessarily those of the NHS, the NIHR or the Dept. of Health. The MRUI software package was kindly provided by the participants of the EU Network programmes: Human Capital and Mobility, CHRX-CT94-0432 and Training and Mobility of Researchers, ERB-FMRX-CT970160.

References: [1] Cady EB, et al. Magn Reson Med 1995; 33: 862-867. [2] Davies et al. Abstract 1769 ISMRM 2011 [3] Wilson M. et al. Mag Reson Med 2011;65:1-12. [4] Cavassila S et al. J Magn Reson. 2000;143:311-320. [5] Cady et al. NMR Biomed. 2011;24:865-872. [6] Vescovo et al. NMR Biomed. 2013; 26: 213–223. [7] Davies NP, et al. NMR Biomed, 2008. 21: p. 908-918.

b.brooks@bham.ac.uk a.peet@bham.ac.uk

nigel.davies@nhs.net

Statistical models for repeated measures of imaging outcomes and related work: theory and applications to longitudinal studies and clinical trials in neurology

Chris Frost, Department of Medical Statistics, London School of Hygiene and Tropical Medicine (LSHTM)

Frost C, **Kenward MG**, Fox NC. Optimizing the design of clinical trials where the outcome is a rate. Can estimating a baseline rate in a run-in period increase efficiency? *Statistics in Medicine,* 27:3717-3731, 2008.

The statistical power of randomised controlled trials with a continuous outcome can be increased by adjusting for a pre-randomisation baseline measure of the outcome. For a trial where the outcome measure is a rate, for example in a therapeutic trial in Alzheimer's disease, the relevant covariate is a pre-randomisation measure of that rate. Obtaining this requires separating the total follow-up period into two periods, with all patients 'off-treatment' in the first 'run-in' period to facilitate calculation of baseline rates. In this paper we use linear mixed models to establish a methodological framework that is then used to assess the extent to which such designs can increase statistical power.

Kapoor R, Furby J, Hayton T, Smith KJ, **Altmann DR**, Brenner R, Chataway J, Hughes RAC, Miller DH. Lamotrigine for neuroprotection in secondary progressive multiple sclerosis: a randomised, double-blind, placebo-controlled, parallel-group trial. *Lancet Neurology*, 9:681 - 688, 2010.

The primary outcome for this randomised clinical trial was partial cerebral volume, obtained six-monthly over five time points, with the primary analysis being a comparison of the rate of change over 24 months. There was no evidence of a treatment effect.

Results from another randomised controlled trial in secondary progressive multiple sclerosis, this one investigating the effectiveness of Simvastatin, have been presented at the European Committee for Treatment and Research in Multiple Sclerosis. A publication is currently under consideration by a medical journal.

The papers summarised here are some that have arisen from longstanding research collaborations between statisticians at LSHTM (shown in **bold**) and various groups at the Institute of Neurology, UCL.

Also see poster by **Cassidy, Keogh** and **Frost** describing some of our collaborative work on TRACK-HD, a major international longitudinal study in Huntington's disease.

Frost C, **Kenward MG**, Fox NC. The analysis of repeated 'direct' measures of change illustrated with an application in longitudinal imaging. *Statistics in Medicine,* 23:3275-3286, 2004.

This paper introduces a statistical framework for the analysis of a novel type of repeated measures data that often arises in longitudinal imaging studies. The data in question consists of a series of measures of *change* in brain volume (rather than a series of volumes). Such repeated measures of change have a dependency pattern that is non-standard. Here we introduce a novel family of linear mixed models that make appropriate allowance for the correlation structure of such data, hence facilitating their correct analysis.

Henderson AP, **Altmann DR**, Trip AS, **Kallis C**, Jones SJ, Schlottmann PG, Garway-Heath DF, Plant GT, Miller DH. A serial study of retinal changes following optic neuritis with sample size estimates for acute neuroprotection trials. *Brain*,133:2592-602, 2010.

In unilateral optic neuritis the affected optic nerve area swells during the acute phase, making it unsuitable as a baseline covariate for comparison of follow-up affected nerve area between trial arms. This study calculated sample sizes for future clinical trials based on a follow-up comparison adjusted for baseline *unaffected* optic nerve area. Results were used to power an ongoing trial of phenytoin in unilateral optic neuritis.

Bartlett JW, **Frost C**, Mattsson N, Skillback T, Blennow K, Zetterberg H, Schott JM.
Determining cut-points for Alzheimer's disease biomarkers: statistical issues, methods and challenges. *Biomarkers in Medicine*, 6:391-400, 2012.

New proposed criteria for the clinical diagnosis of Alzheimer's disease increasingly incorporate biomarkers, often measured on a continuous scale. Operationalizing such criteria usually requires selection of a cut-point at which to dichotomize. In this paper, we review the statistical principles underlying the choice of cut-points, describe some of the most commonly adopted statistical approaches, highlight potential pitfalls with some of the approaches and characterize in what sense the estimated cut-point from each approach is optimal. We also emphasize that how a cut-point is selected must be made with reference to how the resulting dichotomized biomarker is to be used, specifically the actions will follow from a positive or negative test result.

Schott JM, **Frost C**, Whitwell JL, MacManus DG, Boyes RG, Rossor MN, Fox NC. Combining short interval MRI in Alzheimer's disease: Implications for therapeutic trials. *Journal of Neurology,* 253:1147-1153, 2006.

Schott JM, **Frost C**, MacManus DG, **Ibrahim F**, Rossor MN, Waldman AD, Fox NC. Short echo time proton MR spectroscopy in Alzheimer's disease: a longitudinal multiple time point study. *Brain,* 133:3315-3322, 2010.

These two papers describe results from longitudinal analysis of the MIRIAD cohort study in Alzheimer's disease. Conclusions are drawn regarding the design (ideal length of follow-up, advantages of introducing interim visits *etc.*) of future clinical trials that plan to utilise magnetic resonance imaging and MR spectroscopy respectively.

NIHR
Leeds Musculoskeletal
Biomedical Research Unit

Scoring of joint synovitis measured by ultrasound: A proposed study design for assessing reliability

Elizabeth Hensor & the LMBRU ultrasound team*

UNIVERSITY OF LEEDS

Introduction:

Ultrasound measurement of synovitis is increasingly important in the assessment of patients with arthritis. Inter- and intra-reader reliability of image scoring needs to be measured in a sufficiently large number of individuals, with adequate score distributions; for some joints, scores >0 are relatively rare.

Aim:

To design an optimal study to comprehensively assess the inter- and intra-reader reliability of a team of rheumatologists when scoring joint synovitis measured by ultrasound.

Methods:

The Leeds Inflammatory Arthritis CONtinuum (IACON) is a cohort study of early IA which has recruited 1200 patients since 2010. Ultrasound scans are performed at baseline, 6, 12 months then annually. Joints are scored for grey scale synovial hypertrophy (0-3) and power Doppler signal (0-3) by the sonographer during 'live' scanning (see *Score Definitions*); they then select a view of each joint which is captured and stored. The following images are stored: bilateral elbow, wrist, PIPs 2&3, MCPs 2&3, knee, ankle, MTPs 1-5. All of the joint images stored for a patient at a given visit are defined as a 'set'. Based on the existing sonographer's scores, sets of images from 100 patients will be selected for the reliability exercise (see *Rules for Image Selection*). Each member of the imaging team will then score the images twice, at least 2 weeks apart, blind to patient ID.

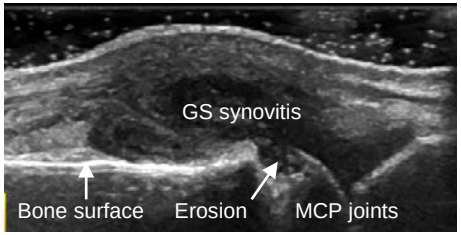
Sample Size:

If at least 5 measurements are made on 100 patients this will give 80% power to demonstrate that an anticipated ICC of 0.7 is significantly higher than 0.6 (=minimum for substantial agreement[1]). The 95% CI will be ±0.15.

Score Definitions:

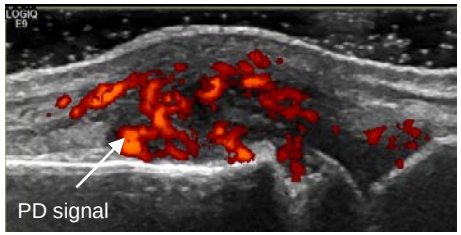
Grey Scale (GS) – hypoechoic synovial thickening

- Grade 0: None
- Grade 1: Mild [filling the angle between the peri-articular bones without bulging over the line linking tops of the bones]
- Grade 2: Moderate [bulging over the line with predominantly flat or concave upper surface]
- Grade 3: Severe [bulging over the line with a convex upper surface]



Power Doppler (PD) – vascularisation of synovium

- Grade 0: No flow in the synovium
- Grade 1: Isolated spots of signal (up to 3 single or 2 confluent)
- Grade 2: Vessel signals in <50% of the area of the synovium
- Grade 3: Vessel signals in ≥50% of the area of the synovium



Rules for Image Selection:

Image sets will be selected according to PD score, which tends to be lower than GS. Based on PD score frequencies in the dataset, elbows and ankles are the least commonly affected; all sets with scores >0 will be selected. For PIP2, PIP3, Knee, MTP2, MTP3, MTP4, MTP5, starting with the least common, sets scoring >1 will be selected. For the remaining joints, sets scoring >2 will be selected. This process will continue until 100 sets have been selected; each patient will contribute just 1 set.

Proposed analyses:

To assess reliability of scoring individual joint types, just one score per patient will be included in the analysis; the right side score will be selected unless the left side image was given a higher score by the sonographer who stored the images. This will eliminate within-patient clustering of scores and will achieve a more balanced score distribution. Total GS and PD scores will also be produced, summating image scores for all joints bilaterally.

Reliability of individual joint scoring will be assessed using paired and multi-reader quadratic-weighted Kappa, which will be supplemented with proportions of exact, close (±1 category) and category-specific agreement.

Reliability of total joint scores will be assessed using intraclass correlation coefficients and Kendall's coefficient of concordance. Bland-Altman plots will be used to help identify any consistent bias or association between the magnitude of measurement and the extent of disagreement; each reader's scores will be compared to those of the sonographer who initially scanned the joints.

Potential Limitation:

Scoring of static images does not exactly replicate the clinical situation as during 'live' scanning the US probe can be moved around the joint.

*LMBRU Ultrasound Team:

Jane Freeston, Laura Horton, Alwyn Jackson, Jacqueline Nam, Ai Lyn Tan, Richard Wakefield

Images courtesy of Ahmed Zayat

1. Landis JR & Koch GG (1977). The measurement of observer agreement for categorical data. *Biometrics* **33** (1): 159–174.

A trilogy of studies: Diagnostic accuracy & use of SeHCAT test in England

Jennifer Summers[1,2], Ana Pascoal[1], Stephen Keevil[1], Cornelius Lewis[1], Victoria Cornelius[1,2], Janet Peacock[1,2]

[1] KITEC - King's Imaging Technology Assessment Centre, King's Health Partners AHSC, King's College Hospital NHS Foundation Trust, Guy's and St Thomas' NHS Foundation Trust

[2]Division of Health and Social Care Research, King's College London

BACKGROUND: Chronic diarrhoea is a common problem often resulting in protracted negative and frequently repeated investigations and unsatisfactory outcome. It is a significant burden on the NHS as well as on affected patients. Bile Acid Malabsorption (BAM) is common in many diarrhoea conditions, such as diarrhoea-predominant irritable bowel syndrome (IBS-D). However, robust data on the prevalence of primary and secondary BAM does not exist and the proportion of patients for whom BAM is secondary to another condition, but who would benefit from treatment with with bile acid sequestrates (BAS), is unknown.

Confirmation of a clinical diagnosis of BAM can be made using SeHCAT (tauroselcholic [^{75}selenium] acid), a radiolabelled synthetic bile acid. The SeHCAT test is a measure of the retention of radioactivity in the patient following administration of a capsule containing SeHCAT. The patient is scanned with a gamma camera one to three hours after taking the capsule, and the scanning is repeated after seven days to measure retention of the radiolabelled bile acid. Low SeHCAT retention levels at day seven represent an abnormal result for the test, indicating a positive diagnosis of BAM. Patients with a confirmed clinical diagnosis of BAM from a positive SeHCAT test may be offered treatment with BAS.

NICE published Diagnostic Guidance (DG7 [NICE, 2012]) on SeHCAT for the investigation of diarrhoea due to BAM in people with IBS-D or Crohn's disease without ileal resection. DG7 concluded that there was potential patient and health system benefits from using SeHCAT for diagnosis of BAM. However, insufficient evidence currently exists to determine whether SeHCAT is a cost-effective diagnostic option, and therefore, a programme of research was recommended.

Following on from the research recommendations in DG7, King's Imaging Technology Evaluation Centre (KITEC) was commissioned to undertake research to address the DG7 research recommendations.

RESEARCH DIFFICULTIES: KITEC's preliminary work has identified various issues in conducting research on SeHCAT:

- Patchy adoption of SeHCAT testing across NHS centres.
- Poor patient adherence to BAS treatment.
- Little information on outcome of patients that are diagnosed as BAM negative.
- Current lack of evidence on value of SeHCAT test.
- Various types of expertise are required in different stages of the studies, with participation from various personnel in a range of NHS centre departments (eg. Gastroenterology and Nuclear Medicine departments).

STUDY OVERVIEW: KITEC have proposed a trilogy of sequential studies. The primary aim of these studies is to gather relevant evidence (clinical, technical, organisational and economic) on the use of SeHCAT in NHS centres in England and inform the design of the study to address the research questions in DG7. KITEC is able to provide a multidisciplinary team of various expertise: nuclear medicine physicists & physicians, gastroenterologists, statisticians and database developers.

1. **RETROSPECTIVE SURVEY** will be based on existing British Nuclear Medical Society (BNMS) data as described in Smith et al (2013). This will provide information on use of the SeHCAT test in NHS centres to inform overall use of SeHCAT test and variability of protocol.
    - Data collected includes: type of hospital, total/frequency of tests, organisational aspects, current SeHCAT use/protocol and economic data.
    - Identification of centres will be an opportunity to initiate collaboration with centres that are essential for implementing the next two stages of research.

2. **PROSPECTIVE SURVEY** will inform the prevalence and severity of BAM and the continuum of results for SeHCAT tests on individual patients.
    - Data will be collected on clinical history, SeHCAT results, post-SeHCAT care, patient follow-up, adherence to treatment and economic data.
    - Data will be collected through web based application.
    - Survey will provide the necessary background information to help optimize the research protocol/design of the observational study.

3. **MULTI-CENTRE PROSPECTIVE OBSERVATIONAL STUDY** will compare the diagnostic accuracy of SeHCAT and biomarkers in predicting a positive indication of BAM and will also investigate the efficacy of treatment with BAS in patients with BAM.
    - Investigate feasibility of using SeHCAT test to provide indication of BAM severity.
    - Will help determine appropriateness of developing standard threshold values for BAM diagnosis.
    - Provide information of BAM negative patients care pathways.

REFERENCES: NICE, 2012. *Diagnostics Guidance 7: SeHCAT (tauroselcholic [75 selenium] acid) for the investigation of diarrhoea due to bile acid malabsorption in people with diarrhoea-predominant irritable bowel syndrome (IBS-D) or Crohn's disease without ileal reseaction.* s.l.: National Institute for Health and Clinical Excellence.

Smith M & Perkins A, 2013. A survey of the clinical use of SeHCAT in the UK. *Nuclear Medicine Communications,* 34(4), pp. 306-313.

The promise of image derived biomarkers: are we fishing in the desert?

A Chalkidou, MJ O'Doherty, PK Marsden
Corresponding author: anastasia.chalkidou@kcl.ac.uk
Comprehensive Cancer Imaging Centre, Division of Imaging Sciences and Biomedical Engineering, Kings College London and Guys and St Thomas NHS Foundation Trust

Introduction

The use of imaging biomarkers is evolving from qualitative interpretation to quantitative analysis with the use of various image-based metrics.

A number of exploratory studies have proposed that a family of multiple image derived indices, generated by an algorithm and called texture features, can predict clinical outcome in patients with cancer.

The study of multiple image derived indices on a data set can lead to multiple hypotheses testing and great inflation of the type I error.

We conducted a systematic review in order to investigate the extent of the inflation of the type I error in studies analysing patient outcome with texture features.

Materials & Methods

Systematic review inclusion criteria:

1) Inclusion of patients with any cancer type

2) Investigation of the relationship between different texture features extracted from PET or CT images and patient outcome

3) Publication as a full paper

1) Various information were extracted for each study (table 1)

2) The true type I error and the corrected significance level for the published results, was recalculated based on:

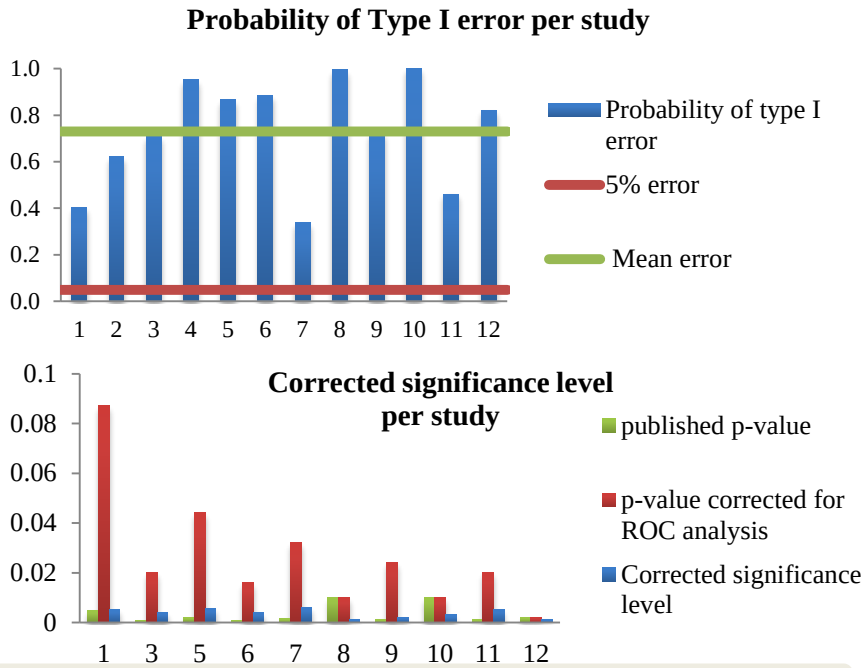
a) the formula proposed by Altman et al [1] for the minimum p-value approach and

b) the Benjamini/Hochberg method [2].

Table 1

No of studies	Mean sample size	Mean No of hypotheses tested/ study	Mean ratio of sample size/hypotheses	No of studies with optimum cutpoint	No of studies with type I error adjustment	No of studies with validation set
12	38	47	2.17	7	0	1

Results



Conclusions

- An average probability for the Type I error of 73% (range: 34-99%) was estimated for the included studies.
- None of the published p-values remained statistically significant after applying the corrections for the optimum cutpoint and multiple hypotheses testing.
- In an era where the lack of reproducibility in research findings has become the most significant problem, emerging trends in the field of imaging biomarkers should be carefully scrutinised for the validity of their results.
- This is very crucial for studies with image derived indices generated by algorithms without corresponding biological function.

References:
1) Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst. 1994 Jun 1;86(11):829-35.
2) Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289-300.