

Original Research Article

Video Rating in Neurodegenerative Disease Clinical Trials: The Experience of PRION-1

Christopher Carswell^a Michael Rañopa^b Suvankar Pal^a
Rebecca MacFarlane^a Durre Siddique^a Dafydd Thomas^a
Tom Webb^a Steve Wroe^a Sarah Walker^b Janet Darbyshire^b
John Collinge^a Simon Mead^a Peter Rudge^a

^aMRC Prion Unit and Department of Neurodegenerative Disease, University College London Institute of Neurology, and National Prion Clinic, National Hospital for Neurology and Neurosurgery, University College London Hospitals NHS Trust, and ^bMRC Clinical Trials Unit, London, UK

Key Words

Creutzfeldt-Jakob disease · Rating scale · Clinical trial · Video rating · PRION-1

Abstract

Background/Aims: Large clinical trials including patients with uncommon diseases involve assessors in different geographical locations, resulting in considerable inter-rater variability in assessment scores. As video recordings of examinations, which can be individually rated, may eliminate such variability, we measured the agreement between a single video rater and multiple examining physicians in the context of PRION-1, a clinical trial of the antimalarial drug quinacrine in human prion diseases. **Methods:** We analysed a 43-component neurocognitive assessment battery, on 101 patients with Creutzfeldt-Jakob disease, focusing on the correlation and agreement between examining physicians and a single video rater. **Results:** In total, 335 videos of examinations of 101 patients who were video-recorded over the 4-year trial period were assessed. For neurocognitive examination, inter-observer concordance was generally excellent. Highly visual neurological examination domains (e.g. finger-nose-finger assessment of ataxia) had good inter-rater correlation, whereas those dependent on non-visual clues (e.g. power or reflexes) correlated poorly. Some non-visual neurological domains were surprisingly concordant, such as limb muscle tone. **Conclusion:** Cognitive assessments and selected neuro-

logical domains can be practically and accurately recorded in a clinical trial using video rating. Video recording of examinations is a valuable addition to any trial provided appropriate selection of assessment instruments is used and rigorous training of assessors is undertaken.

Copyright © 2012 S. Karger AG, Basel

Introduction

The lack of disease-modifying therapies in neurodegeneration is one of the most significant failings of modern neurology. The slow rate of discovery and progress of investigational therapeutics through trials is an important factor, but the performance of outcome measures and trial methodologies is also being questioned [1–6]. Unnecessary variability and blunt measurement tools mean larger and longer trials. Very large trials are expensive but are at least feasible in common neurodegenerative diseases. The problems are compounded in rare conditions because large trials may be impossible without international collaboration across multiple centres or long recruitment periods. As a result, problems of unnecessary variability may then be exacerbated by geographical, inter-rater and time-dependent factors [7, 8]. Video rating is a potential solution to these problems.

Prion diseases are a group of fatal, transmissible neurodegenerative conditions caused by misfolding of the prion protein [9]. Typically, prion diseases involve multiple cognitive and neurological domains, requiring very broad assessment tools to capture changes in the neurocognitive examination over time. Currently, there is no effective treatment for human prion disease [10, 11]. Quinacrine was known to be a relatively safe oral drug which had been shown to be efficacious against prion infection in cell culture [12]. After extensive consultation, PRION-1 was started, an open-label patient-preference trial of the anti-malarial compound quinacrine in human prion disease. This clinical trial remains the largest conducted in prion disease.

A total of 107 sporadic, variant or inherited Creutzfeldt-Jakob disease (CJD) patients were enrolled and prescribed quinacrine before repeat neurological and cognitive assessment during a follow-up period totalling 77 patient-years. There were three reasons for video recording these assessments. First, it was originally hoped to conduct a double-blind placebo-controlled trial. One particular problem with quinacrine is that the drug causes a yellowish discoloration of the skin in many subjects, which unblinds the patient and assessing physicians. Partially to ameliorate this problem, the video assessment was done using a monochrome filter ensuring that the video rating could be done blind to treatment received. Secondly, a permanent record of the assessment was obtained. Finally, given the length of the trial from inception to termination (4 years), it was likely that the clinical assessors would not be the same individuals throughout this period for those patients with the longest duration in the trial, emphasising the need for a consistent independent assessment. Thus, to improve overall agreement and objectivity in clinical assessment and reduce the reliance upon inter-observer agreement, we performed video assessments rated by a common senior neurologist assessor throughout the trial [11].

Video consultation has been validated in other diseases and, although some outcomes agree favourably in inter- or intra-observer assessment, the results are variable and would appear specific to the clinical signs being assessed as well as the training of the individual assessors being compared [13–18]. The opportunity in PRION-1, provided by the breadth of the neurocognitive assessment domains, is for general conclusions to be made of relevance to many other neurological disorders. We therefore analysed the agreement between a bedside and video-rated neurocognitive assessment of patients involved in the PRION-1 trial to understand the potential benefit of this medium in future therapeutic studies in neurodegenerative diseases.

Table 1. Details of scale type and vocal prompt for cognitive examination

Cognitive assessment	Scale type	No. of levels	Example of prompt/action
Fragmented letters	Ordinal	4	Can you see any letters here?
Calculation	Ordinal	5	What is 5 + 4?
Spelling	Ordinal	7	Can you spell the word build?
Fragmented objects	Ordinal	5	Can you see any objects here?
Words beginning with letter	Ordinal	31	Can you give me as many words beginning with the letter 'F' as possible?
Copying gestures	Ordinal	4	Can you copy these shapes with your hand?
Digit span	Ordinal	6	Can you repeat these numbers after me?
Memory	Ordinal	4	What is your name?
Line drawings	Ordinal	11	Can you tell me what these drawings are?
Recall	Ordinal	4	Can you tell me everything you can remember about the passage you read?
Frontal lobe sequencing	Ordinal	4	Can you copy this?
Proverbs	Categorical	3	Can you tell me what 'Too many cooks spoil the broth' means?
Letter cancelling	Ordinal	13	Can you show me all the letter As?
Miming	Ordinal	4	Show me how you brush your hair?
Reading passage	Ordinal	4	Can you read this passage out for me?

Methods

PRION-1 Study Design

Details of the PRION-1 methodology and the lack of benefit from quinacrine have been published [11]. In brief, patients were offered a choice between taking quinacrine, not taking quinacrine, or being randomised to immediate quinacrine or quinacrine deferred for 24 weeks. Whether chosen or randomised, quinacrine was given orally with a loading dose of 1 g over 24 h (200 mg every 6 h), followed by 100 mg three times daily. In total, 107 individuals were enrolled (45 sporadic, 19 acquired (17/19 variant) and 43 inherited prion disease). The objective was to obtain data on the effect of quinacrine in human prion disease, from a randomised comparison where acceptable and otherwise from observational comparisons. Patients were seen at enrolment and subsequently either at the National Prion Clinic, National Hospital for Neurology and Neurosurgery, London, or at their homes.

Cognitive Examination

Fifteen separate tests were applied to test cognitive ability (table 1). Attention was assessed using a cancellation task in which 12 identical letters randomly embedded in a matrix of 68 other letters had to be identified within 1 min. The number of letters cancelled and errors were determined. Visual perception was assessed using 3 incomplete figures and 3 letters in which 30% of the pixels were randomly deleted. In the miming task, being unable to initiate the action, poor performance or using a body part as the implement were all classified as abnormal. Frontal lobe sequencing was determined by the ability to open and close the hand alternating between the left and right upper limb eight consecutive times. Concrete interpretation of proverbs was scored as abnormal. Reading was assessed using a passage of approximately 100 regular and irregular words determining fluency, dysphasia and dyslexia. Recall of this passage was done 15–30 min later, subjectively scoring the number of salient points recalled.

Table 2. Details of scale type and vocal prompt for cognitive examination

Overall assessment	Scale type	No. of levels	Example of prompt/action
Cooperation	Categorical	3	Is overall cooperation satisfactory? (yes/no/fluctuates)
Cerebellar	Ordinal	4	Overall impression of cerebellar impairment was none/mild/moderate/severe
Cognitive impairment	Ordinal	4	Overall impression of cognitive impairment was...
Cope	Categorical	3	Is the patient able to cope with the test demands?
Attention	Categorical	3	Is overall attention satisfactory?
Extrapyramidal impairment	Ordinal	4	Overall impression of extrapyramidal impairment was...
Pyramidal impairment	Ordinal	4	Overall impression of pyramidal impairment was...

Table 3. Details of scale type and vocal prompt for neurological examination

Neurological assessment	Scale type	No. of levels	Example of prompt/action
Walking scale	Ordinal	7	Can you walk for me? (0–6)
Sequential opposition	Ordinal	4	normal/mild/moderate/severe
Rapid alternating hand movements	Ordinal	4	normal/mild/moderate/severe
Romberg	Binary	2	normal/abnormal
Tone	Binary	2	normal/abnormal
Walking Overall	Ordinal	6	mild/moderate/severe/wheelchair/bedbound
Finger-nose testing	Ordinal	4	normal/mild/moderate/severe
Heel toe walking	Ordinal	6	mild/moderate/severe/wheelchair/bedbound
Observation – myoclonus	Binary	2	normal/abnormal
Primitive reflexes – grasp	Binary	2	normal/abnormal
Sequential finger tapping	Ordinal	4	normal/mild/moderate/severe
Primitive reflexes – glabellar tap	Binary	2	normal/abnormal
Primitive reflexes – pout	Binary	2	normal/abnormal
Eye movements	Categorical	4	Normal/failure of upgaze/nystagmus/other
Power	Binary	2	normal/abnormal
Observation – chorea	Binary	2	normal/abnormal
Walking gait	Categorical	4	normal/apraxic/ataxic/cerebellar
Reflexes	Binary	2	normal/abnormal
Observation – tremor	Binary	2	normal/abnormal
Observation – other	Binary	2	normal/abnormal

Overall Impression

This was the subjective assessment of the observing physician regarding overall attention, ability to cope with the task, cooperation, and an overall measure of cognitive, extrapyramidal, pyramidal and cerebellar impairment using a qualitative 4-point scale ranging from normal to severely impaired (table 2).

Neurological Assessment

A standardised neurological examination was performed on each assessment (table 3). Tone was designated normal or increased and was classified as pyramidal or extrapyramidal. Reflexes were classified as absent, normal or increased. The MRC grading of power was used. Finger-nose-finger assessment, rapid alternating hand movements, sequential index finger

tapping and sequential opposition were all qualitatively graded on a 4-point scale from normal to severely impaired. Gait was described categorically, and impairment was rated on both a 7-point scale and functionally.

Video Protocol

Large amounts of data on a variety of sequential clinical scores were collected throughout this trial using the NV-GS10 digital video camera (Panasonic) and Premiere Pro 2.0 editing software (Adobe). The assessments were made wherever the patient resided, i.e. in the hospital, at the patient's home or in care homes throughout the UK. These data included assessments of neurological and cognitive examinations, producing a score for each domain and an overall score for cooperation, attention and ability to cope with the test as well as a summary score of motor and cognitive function, the frequency of testing depended on the type of prion disease. Visits were initially scheduled at baseline, at 1, 2, 4 and 6 months and thereafter at 3-month intervals. The duration of the trial was 2 years, and patients were recruited over a period of 5 years. The assessments were conducted by research registrars; all examinations in a particular patient were done, as far as possible, by the same registrar.

Subsequently, colour videos were copied and edited by a registrar to give sequences of each examination module. These videos were scored blindly, using a yellow monochrome screen by 1 senior neurologist in a randomised order.

Statistical Analysis

The κ statistic, a chance-corrected measure of agreement, with 95% confidence intervals, was used to assess the agreement between the registrar and blind assessor scoring for neurological and cognitive examinations. κ statistics were interpreted based on the convention by Landis and Koch [19]: <0, no agreement; 0–0.20, slight agreement; >0.20–0.40, fair agreement; >0.40–0.60, moderate agreement; >0.60–0.80, substantial agreement, and >0.80–1.0, almost perfect agreement.

The standard κ statistic does not take into account the degree of disagreement between the categories selected by the raters, and all disagreement is treated equally as total disagreement. For tests with 4 or more levels in an ordinal scale, a weighted κ statistic (with linear weights) was used to allow for the fact that disagreement by 1 ordinal category is less severe than by 2, 3 or 4, etc.

Agreement between the registrar and blinded assessor was also displayed for ordinal tests with 4 or more values using Bland-Altman plots, which scatter the difference between two measurements versus the mean. Good agreement is indicated by points scattered evenly around the horizontal line of no difference, with no trend for increasing or decreasing differences with larger or smaller mean values.

All statistical analyses were performed with STATA version 11.0.

Results

A total of 513 assessments involving 104 patients were performed over the 4-year trial, of which 445 (86%) were recorded on video. After baseline video quality control, 425 (83%) of the videoed assessments were available for paired scoring assessment on 101 patients. Between 64 and 335 (median 210) assessments for each individual domain across all patient visits were rated by both the registrar and the blinded assessor. The most common cause for an absent score was inability to perform the task due to advanced patient disability, e.g. akinetic mutism. One domain (the palmomental reflex) was not available for final comparison because too few assessments were rated by both assessors.

Table 4. The inter-assessor agreement in cognitive examination measured by the kappa statistic when the Creutzfeldt-Jakob disease patients were examined at the bedside and remotely by video recording

Cognitive assessment	n	κ statistic	SE
Fragmented letters	190	0.940	0.062
Calculation	212	0.937	0.050
Spelling	210	0.921	0.052
Fragmented objects	214	0.890	0.052
Words beginning with letter	206	0.860	0.041
Copying gestures	172	0.850	0.062
Digit span	163	0.819	0.052
Memory	314	0.817	0.048
Line drawings	212	0.812	0.050
Recall	147	0.763	0.064
Frontal lobe sequencing	177	0.757	0.063
Proverbs	163	0.755	0.060
Letter cancelling	173	0.610	0.052
Miming	220	0.471	0.050
Reading passage	164	0.468	0.050

Table 5. The inter-assessor agreement in overall examination measured by the kappa statistic when the Creutzfeldt-Jakob disease patients were examined at the bedside and remotely by video recording

Overall assessment	n	κ statistic	SE
Cooperation	333	0.652	0.043
Cerebellar	300	0.625	0.041
Cognitive impairment	311	0.589	0.037
Cope	335	0.583	0.043
Attention	332	0.529	0.042
Extrapyramidal impairment	309	0.410	0.039
Pyramidal impairment	310	0.155	0.036

Comparison of the degree of agreement for the cognitive and neurological scores between the blinded and unblinded assessors is shown in figure 1, table 4 and table 6.

The inter-observer agreement was generally excellent for the neurocognitive examination (median $\kappa = 0.817$, range 0.468–0.940) as the agreement in 9 of the 15 cognitive domains was excellent, substantial in 4 and moderate in only 2 (fig. 1a; table 4). Agreement was best in those measures with a clearly defined outcome such as fragmented letters test ($\kappa = 0.940 \pm 0.05$) and calculation ($\kappa = 0.940 \pm 0.06$), and worst in miming ($\kappa = 0.470 \pm 0.05$) and reading a passage ($\kappa = 0.471 \pm 0.050$), which do not have such a clearly defined outcome. There were, however, some surprising exceptions to this pattern; proverb interpretation had substantial agreement ($\kappa = 0.755 \pm 0.060$), whereas letter cancellation ($\kappa = 0.610 \pm 0.052$), which involved striking through 12 letters in a set random display, only had substantial/moderate agreement (fig. 1a; table 1). Two factors probably accounted for this reduction in expected agreement in the cancellation task: we did not define a time limit to the test and we could have defined better rules for failure of the task due to perseveration causing repeated cancellation of the same letter. The relatively poor agreement on reading a passage was partly due to dysarthria and technical inadequacies such as poor microphone placement and high background noise. Of note, paired completion of the cognitive tests across the 425 assessments was relatively poor, ranging from 65–74%.

Overall, neurological observations were the most consistently measured assessments (71–79%) and had slightly better agreement than the neurological examination itself (me-

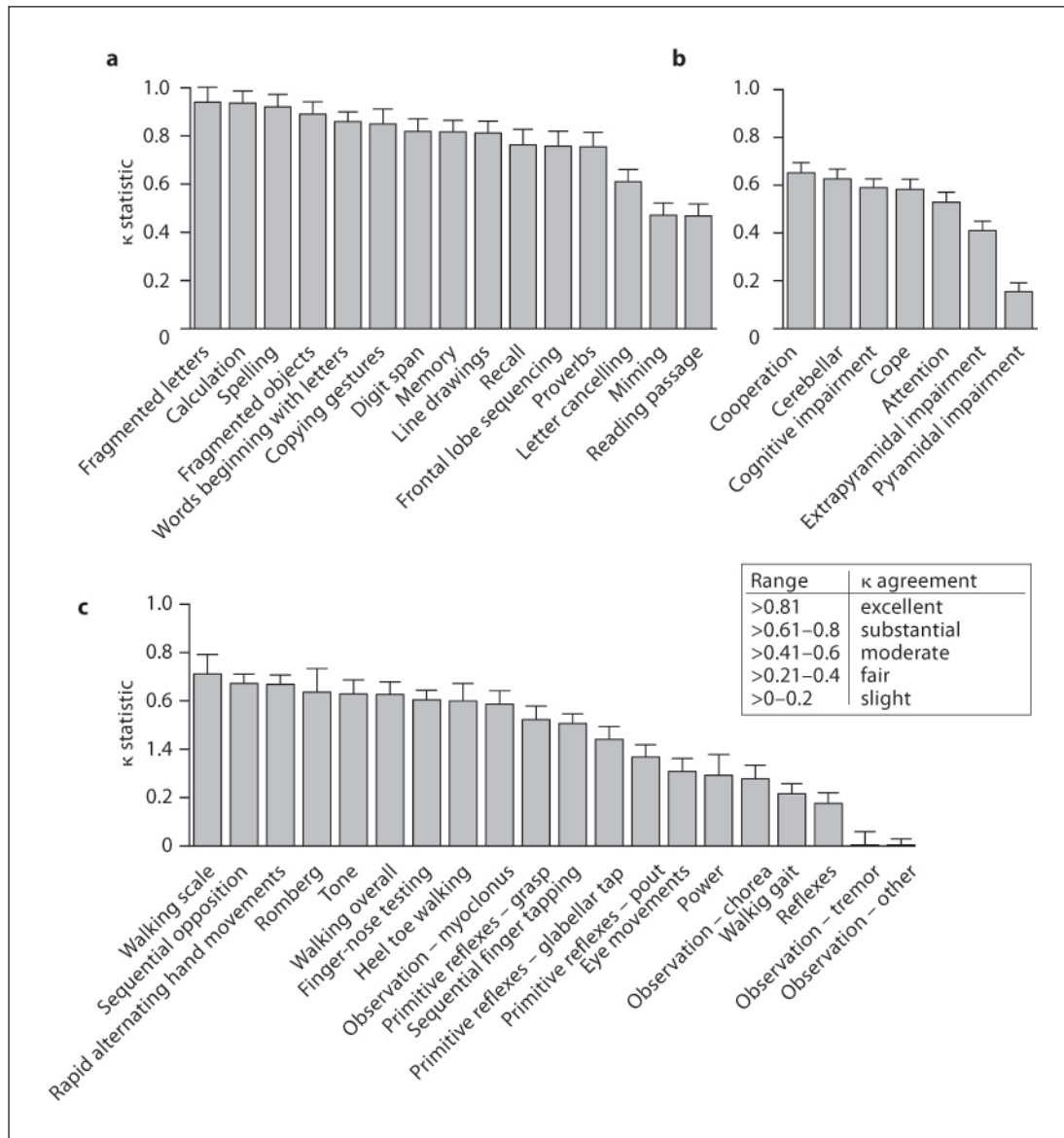


Fig. 1. Creutzfeldt-Jakob disease patients were assessed by the bedside and remotely by video recording. The inter-assessor variability between the assessments is shown using the kappa statistic for cognitive examination (a), overall impression (b), and neurological examination (c).

dian $\kappa = 0.583$, range 0.155–0.652 vs. median $\kappa = 0.523$, range –0.064 to 0.712). Interestingly, the attention of the patient and the ability to cope and cooperate with the examination all had moderate/substantial agreement and were scored on almost every visit ($n = 332/335$), but no domain had excellent agreement (fig. 1b). The agreement between the blinded observer and clinician for the overall impression of cognitive and cerebellar deficit was moderate or substantial, but that for pyramidal or extrapyramidal deficit was poor (fig. 1b; table 5).

The agreement in the neurological examination was very variable between tests. The agreement was substantial in 6 of the 27 domains, moderate also in 6 of the 27 domains and fair/slight in 7 of the 27 domains; a negative κ ($\kappa = -0.064 \pm 0.056$) for observation of other movements was observed, indicating that agreement occurred less often than predicted by

Table 6. The inter-assessor agreement in neurological examination measured by the kappa statistic when the Creutzfeldt-Jakob disease patients were examined at the bedside and remotely by video recording

Neurological assessment	n	κ statistic	SE
Walking scale	91	0.712	0.080
Sequential opposition	159	0.672	0.038
Rapid alternating hand movements	169	0.668	0.039
Romberg	68	0.636	0.097
Tone	283	0.628	0.058
Walking overall	166	0.627	0.052
Finger-nose testing	178	0.603	0.040
Heel toe walking	64	0.600	0.072
Observation – myoclonus	313	0.586	0.056
Primitive reflexes – grasp	320	0.523	0.056
Sequential finger tapping	164	0.507	0.039
Primitive reflexes – glabellar tap	320	0.440	0.053
Primitive reflexes – pout	320	0.368	0.051
Eye movements	235	0.309	0.053
Power	81	0.292	0.086
Observation – chorea	313	0.278	0.057
Walking gait	184	0.215	0.042
Reflexes	290	0.176	0.045
Observation – tremor	313	0.004	0.055
Observation – other	313	–0.064	0.056

chance alone (fig. 1c; table 6). Unsurprisingly, neurological examinations relying on simple observations, such as sequential finger opposition (κ (right) = 0.672 ± 0.038) or rapid alternating hand movements (κ (right) = 0.668 ± 0.039), had substantial agreement, whereas domains which relied upon the examiner's on-the-spot interpretation of the stimulus and response, such as reflexes (κ (right) = 0.176 ± 0.045), usually had poor agreement (fig 1c; table 6). Despite this generality, there were some neurological tests which were unexpectedly concordant, for example tone, which we had thought would be hard to assess by video (κ (right) = 0.628 ± 0.058). The domains which agreed poorly included pout reflex, eye movements, observation of all unintentional movements (except myoclonus), gait assessment and power (fig. 1c). Of these domains with poor agreement, it is noticeable that the character of gait did not agree well despite the fact that both the functional impression and scale of impairment did (fig. 1c; table 6). The agreement of gait abnormality, which can be complex in prion diseases, may have been confounded by the clinical experience of the visiting registrars. Similarly to the cognitive examinations, paired completion of neurological assessments varied from 15–75%.

The κ statistic gives an overall impression of agreement but does not give a measure of the agreement throughout the range of scores, and so Bland-Altman plots were constructed to illustrate the level of agreement across the range of scores observed [20]. For tests in which there was almost perfect agreement, such as calculation, it can be seen that the Bland-Altman plot has nearly all observations on the horizontal line through zero, indicating no difference at every observed score (fig. 2a). Other tests with a moderate/substantial agreement on κ agree well only at data extremes, e.g. letter cancellation (fig. 2b). Such assessment instruments are poorly suited to the measurement of clinical deterioration despite the 'substantial' agreement indicated by the κ statistic because they will not perform equally well at different levels of impairment. Conversely, there are domains in which there is poor Bland-Altman correlation at any score, e.g. overall extrapyramidal impairment (fig. 2c), and in these κ is also small.

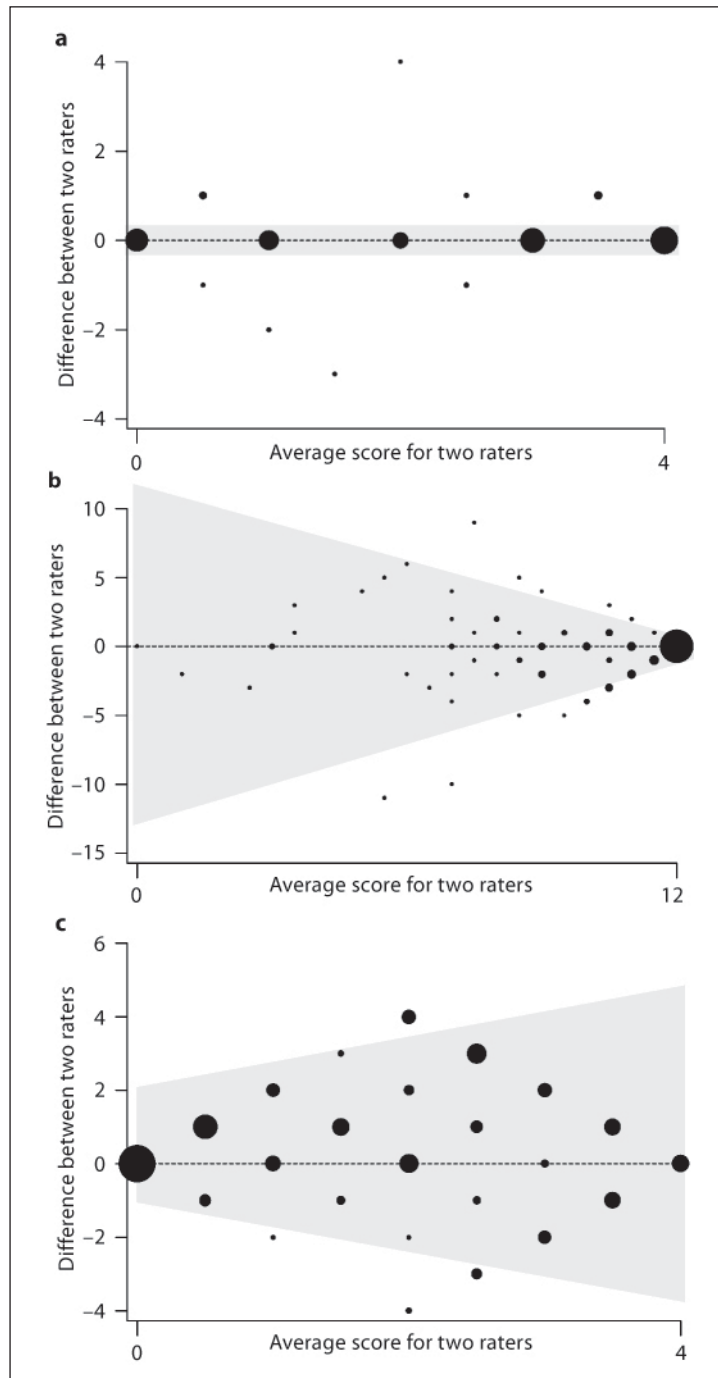


Fig. 2. Bland Altman plots describing the variability of agreement throughout the range of each examination scale between bedside and remote assessments of Creutzfeldt-Jakob disease patients in the PRION-1 trial for the domains: **a** Calculation, **b** Letter Cancellation, and **c** Overall extrapyramidal assessment.

Discussion

The PRION-1 trial involved the longitudinal follow-up of 107 prion disease patients throughout the whole of the UK in an array of different clinical and domestic settings. Despite such logistical challenges and the high disability level of the patients involved, we have succeeded in video recording a formal cognitive and neurological examination on 335 visits over 4 years, without the direct involvement of audiovisual specialists.

We found in general that for assessments which depend on visual or auditory input, such as cognitive examinations, agreement was good. This was in contrast to the variable agreement on neurological assessment where some tests were concordant but others showed considerable inter-observer variation; this will inform the development of assessment instruments in future therapeutic trials as those examinations which lend themselves to video assessment can be prioritised to detect early clinical findings. There are three broad reasons, which are not mutually exclusive, for why the neurological examination provided such widely varying results.

Firstly, there are a number of technical factors related to audiovisual recording. The videos were obtained by a nurse with neurological training in most cases. Although this person had some training to obtain videos, most were taken in far from ideal conditions. Lighting was often suboptimal relying on ambient illumination in the patient's house/hospital, and commonly it was not possible to obtain a light source behind the camera, resulting in dark images. The sites where the recordings were obtained were often extremely cramped so that it was not possible to assess certain parameters because of lack of space, for example, the patient being too near to the camera during a walking assessment. The resolution of the images is of great importance in certain domains. This was particularly the case for recording eye movements. Although the camera had a zoom facility, this was not used in most cases so that it was impossible for the video assessor to determine whether the pursuit movement was saccadic, or if nystagmus was present. Clamping the camera while filming the patient is necessary so that the movement of the camera operator does not confound observed movements in the patient. As examination teams were travelling around the UK, they did not always carry lighting and tripod during the PRION-1 trial. This issue is especially important in assessing dynamic movements, such as limb tremor, where there was virtually no agreement between the two examiners. Further, it is necessary to have the patient at right angles to get a consistent vector measure; in our study, the direction of recording varied greatly.

Secondly, audiovisual assessment is only suitable for certain domains, i.e. those that do not require patient contact. For example, determination of power requires actual examination. For similar reasons, the agreement of attempts to classify the neurological systems involved, i.e. pyramidal and extrapyramidal, was poor. Counter-intuitively tone assessment agreement was substantially better, suggesting that some unexpected signs can be inferred by observation of an examination.

Thirdly, agreement requires two assessors to use the exact same scoring criteria. For example, although the Romberg test is binary, the agreement was still only substantial; some raters in the PRION-1 trial classified this test as positive if there was swaying rather than a fall. Similarly, classification of movement disorders and primitive reflexes is largely subjective and, as a result, some interpretations were unsurprisingly fair/slight, e.g. chorea and pout. Likewise, assessment of gait where the overall impression was normal or abnormal resulted in reasonable agreement, but when describing its nature as apraxic or ataxic, which is more subjective, the agreement was only fair. Clearly defined criteria and extensive training of assessors should overcome this problem.

This study demonstrates that video assessment is feasible and useful in clinical trials requiring a range of neurocognitive tests and when patients are dispersed throughout a country. However, we also highlight that video rater disagreement may be substantial for several specific components of the neurocognitive examination. We recommend a rigorous training programme for all assessors with regard to technical aspects of video recording and well-defined scoring criteria for certain neurological tests. Video rating provides an objective backup and permanent record to trial assessment and offers an opportunity to blind assessments. A clinical trial should be tailored so that those modalities which are not suited to video assessment are marginalised in the video rating.

Acknowledgements

Trial Steering Committee: D. Armstrong, I. Chalmers, L. Firkins, F. Certo, J. Collinge, J. Darbyshire, C. Kennard, A. Kennedy (until 2004), J. Ironside, A. MacKay, H. Millar, J. Newsom-Davies, J. Nicholl, J. Stephenson, M. Wile and S. Wroe.

Trial Observers: R. Knight, M. Rossor, A.S. Walker (trial statistician) and G. Keogh.

Data and Safety Monitoring Committee: M. Ferguson Smith, V. Farewell, I. McDonal and R. Collins.

We thank all the individuals, their caregivers and their families, who took part in the PRION-1 study, and UK neurologists and the National CJD Surveillance Unit for referring patients. We thank the former Chief Medical Officer Sir Liam Donaldson, officials at the Department of Health, Medical Research Council Research Management Group staff, co-chairs of the PRION-1 trial steering committee, and our colleagues at the National CJD Surveillance Unit for establishing the National CJD referral arrangements, without which PRION-1 would not have been possible. The pilot phase, consumer group consultation, establishment of the National Referral Agreement, and the design and implementation of PRION-1 spanned several years, and many individuals assisted in many ways to enable this study. We thank all past and present colleagues at the National Prion Clinic (formerly at St. Mary's Hospital, London, and now at the National Hospital for Neurology and Neurosurgery, Queen Square, London), in particular Kathryn Prout, Nora Heard, Clare Morris, Rita Wilkinson, Chris Rhymes, Suzanne Hampson, Claire Petersen, Ekaterina Kassidi and Colm Treacy, and at the MRC Clinical Trials Unit, in particular Geraldine Keogh, Moira Spyer, Debbie Johnson, Liz Brodnicki and Patrick Kelleher. We particularly acknowledge the contributions of our distinguished neurological colleagues John Newsom-Davies and Ian McDonald who died during the study.

Some of this work was undertaken at UCLH/UCL that received a proportion of funding from the Department of Health's NIHR Biomedical Research Centres funding scheme. Ray Young assisted with figure design.

Disclosure Statement

J.C. is a director and shareholder of D-Gen Ltd, an academic spin-out company working in the field of prion disease diagnosis, decontamination and therapeutics. No other author has any other financial relationship to disclose or conflict of interest to declare.

References

- 1 Salloway S, Sperling R, Gilman S, et al: A phase 2 multiple ascending dose trial of bapineuzumab in mild to moderate Alzheimer disease. *Neurology* 2009;73:2061–2070.
- 2 Kiebertz K, McDermott MP, Voss TS, et al: A randomized, placebo-controlled trial of latrepirdine in Huntington disease. *Arch Neurol* 2010;67:154–160.
- 3 Green RC, Schneider LS, Amato DA, et al: Effect of tarenflurbil on cognitive decline and activities of daily living in patients with mild Alzheimer disease: a randomized controlled trial. *JAMA* 2009;302:2557–2564.
- 4 Doody RS, Gavrillova SI, Sano M, et al: Effect of dimebon on cognition, activities of daily living, behaviour, and global function in patients with mild-to-moderate Alzheimer's disease: a randomised, double-blind, placebo-controlled study. *Lancet* 2008;372:207–215.
- 5 Feldman HH, Doody RS, Kivipelto M, et al: Randomized controlled trial of atorvastatin in mild to moderate Alzheimer disease: LEADe. *Neurology* 2010;74:956–964.

- 6 Courtney C, Farrell D, Gray R, et al: Long-term donepezil treatment in 565 patients with Alzheimer's disease (AD2000): randomised double-blind trial. *Lancet* 2004;363:2105–2115.
- 7 Quinn TJ, Dawson J, Walters MR, Lees KR: Variability in modified Rankin scoring across a large cohort of international observers. *Stroke* 2008;39:2975–2979.
- 8 Brown CR, Hillman SJ, Richardson AM, Herman JL, Robb JE: Reliability and validity of the Visual Gait Assessment Scale for children with hemiplegic cerebral palsy when used by experienced and inexperienced observers. *Gait Posture* 2008;27:648–652.
- 9 Collinge J: Prion diseases of humans and animals: their causes and molecular basis. *Annu Rev Neurosci* 2001;24:519–550.
- 10 Stewart LA, Rydzewska LH, Keogh GF, Knight RS: Systematic review of therapeutic interventions in human prion disease. *Neurology* 2008;70:1272–1281.
- 11 Collinge J, Gorham M, Hudson F, et al: Safety and efficacy of quinacrine in human prion disease (PRION-1 study): a patient-preference trial. *Lancet Neurology* 2009;2009:334–344.
- 12 Korth C, May BC, Cohen FE, Prusiner SB: Acridine and phenothiazine derivatives as pharmacotherapeutics for prion disease. *Proc Natl Acad Sci USA* 2001;98:9836–9841.
- 13 Adler CH, Bansberg SF, Hentz JG, et al: Botulinum toxin type A for treating voice tremor. *Arch Neurol* 2004;61:1416–1420.
- 14 Wang S, Lee SB, Pardue C, et al: Remote evaluation of acute ischemic stroke – reliability of National Institutes of Health Stroke Scale via telestroke. *Stroke* 2003;34:E188–E191.
- 15 Goetz CG, Leurgans S, Hinson VK, et al: Evaluating Parkinson's disease patients at home: utility of self-videotaping for objective motor, dyskinesia, and ON-OFF assessments. *Mov Disord* 2008;23:1479–1482.
- 16 Quinn J, Moore M, Benson DF, et al: A videotaped CIBIC for dementia patients – validity and reliability in a simulated clinical trial. *Neurology* 2002;58:433–437.
- 17 Handforth A, Ondo WG, Tatter S, et al: Vagus nerve stimulation for essential tremor – a pilot efficacy and safety trial. *Neurology* 2003;61:1401–1405.
- 18 Homma A, Nakamura Y, Kobune S, et al: Reliability study on the Japanese version of the Clinician's Interview-Based Impression of Change. *Dement Geriatr Cogn Disord* 2006;21:97–103.
- 19 Landis JR, Koch GG: The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174.
- 20 Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–310.