Original Article

JCBFM

Harmonization of sensorimotor deficit assessment in a registered multicentre pre-clinical randomized controlled trial using two models of ischemic stroke

Alessia Valente[1,2,]* , Jacopo Mariani[2,]*, Serena Seminara[1],
Mauro Tettamanti[1], Giuseppe Pignataro[3] , Carlo Perego[1],
Luigi Sironi[4], Felicita Pedata[5], Diana Amantea[6] ,
Marco Bacigaluppi[7] , Antonio Vinciguerra[8],
Susanna Diamanti[2,9], Martina Viganò[2],
Francesco Santangelo[2] , Chiara Paola Zoia[2],
Virginia Rodriguez-Menendez[2], Laura Castiglioni[4],
Joanna Rzemieniec[4], Ilaria Dettori[5], Irene Bulli[5],
Elisabetta Coppi[5], Chiara Di Santo[6], Ornella Cuomo[3],
Giorgia Serena Gullotta[7], Erica Butti[7], Giacinto Bagetta[6],
Gianvito Martino[7], Maria-Grazia De Simoni[1],
Carlo Ferrarese[2,9], Stefano Fumagalli[1] and Simone Beretta[2,9];
for the TRICS study group

Abstract

Multicentre preclinical randomized controlled trials (pRCTs) are a valuable tool to improve experimental stroke research, but are challenging and therefore underused. A common challenge regards the standardization of procedures across centres. We here present the harmonization phase for the quantification of sensorimotor deficits by composite neuroscore, which was the primary outcome of two multicentre pRCTs assessing remote ischemic conditioning in rodent models of ischemic stroke. Ischemic stroke was induced by middle cerebral artery occlusion for 30, 45 or 60 min in mice and 50, 75 or 100 min in rats, allowing sufficient variability. Eleven animals per species were video recorded during neurobehavioural tasks and evaluated with neuroscore by eight independent raters, remotely and blindly. We aimed at reaching an intraclass correlation coefficient (ICC) ≥0.60 as satisfactory interrater agreement. After a first remote training we obtained ICC = 0.50 for mice and ICC = 0.49 for rats. Errors were identified in animal handling and test execution. After a second remote training, we reached the target interrater agreement for mice (ICC = 0.64) and rats (ICC = 0.69). In conclusion, a multi-step, online harmonization phase proved to be feasible, easy to implement and highly effective to align each centre's behavioral evaluations before project's interventional phase.

[1]Istituto Di Ricerche Farmacologiche Mario Negri IRCCS, Milan, Italy
[2]School of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy
[3]Department of Neuroscience, Reproductive Sciences and Dentistry, Federico II University of Naples, Napoli, Italy
[4]Department of Pharmaceutical Sciences, University of Milan, Milano, Italy
[5]Department of Neuroscience, Psychology, Drug Research and Child Health (NEUROFARBA), Division of Pharmacology and Toxicology, University of Florence, Firenze, Toscana, Italy
[6]Department of Pharmacy, Health and Nutritional Sciences, University of Calabria, Rende (Cosenza), Italy
[7]Neuroimmunology Unit, San Raffaele Hospital and Università Vita-Salute San Raffaele, Milano, Lombardia, Italy

[8]Department of Biomedical Science and Public Health, Marche Polytechnic University, Ancona, Italy
[9]Fondazione IRCCS San Gerardo dei Tintori, Monza, Italy

*These authors contributed equally to this work.

Corresponding authors:
Stefano Fumagalli, Laboratory of Stroke and Vascular Dysfunctions, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, via Mario Negri 2, 20156 Milan, Italy.
Email: stefano.fumagalli@marionegri.it

Simone Beretta, Laboratory of Experimental Stroke Research, Department of Medicine and Surgery, University of Milano Bicocca, Via Cadore 48, 20900 Monza (MI), Italy.
Email: simone.beretta@unimib.it

Introduction

Acute ischemic stroke is a leading cause of death and long-term disability worldwide.[1] Intravenous thrombolysis and endovascular mechanical thrombectomy are currently the best available therapies, with the aim of restoring cerebral blood flow in the hyperacute phase of acute ischemic stroke. The advent of recanalization therapies has widened the number of treated patients, which is now about 60% of total ischemic stroke patients. However even if successfully recanalized, patients may develop subsequent severe disability. As such ischemic stroke remains a medical emergency and there is an urgent need for adjunctive treatments limiting disease progression.

Among new therapeutic treatments, remote ischemic conditioning (RIC) after the initial event may be promising. Post-stroke RIC consists in inducing one or more transient periods of ischemia in a distant organ, far from the site of injury. Numerous studies reported that RIC can improve cerebral circulation, reduce infarct volume and promote both neurogenesis and angiogenesis.[2]

It should be mentioned that over the past decades, experimental studies have identified and tested different therapeutic targets for stroke in preclinical models.[3] However, none of the compounds or protective strategies identified preclinically have effectively translated into clinical trials.[4]

In view of fostering the transferability of preclinical stroke research, the Stroke Therapy Academy Industry Roundtable published reporting and operational recommendations to enhance the quality of preclinical studies.[5,6] These recommendations are now available in published guidelines like the ARRIVE[7] and the IMPROVE.[8] However, following these guidelines may not be enough to effectively enhance preclinical stroke research if most preclinical studies are single-centre trials.[9] Preclinical randomized controlled trials (pRCTs) conducted in a multicentric manner are a valuable tool to increase the reliability of experimental stroke research.[10] We therefore designed two pRCTs in mice and rats of both sexes, aimed at testing the efficacy of RIC following transient middle cerebral artery occlusion (tMCAo) to model acute ischemic stroke (TRICS BASIC project).[11] The strength of the TRICS BASIC is based on a pre-registered detailed protocol (see at https://preclinicaltrials.eu, ID: PCTE0000177) with a thorough implementation of the ARRIVE and IMPROVE guidelines. TRICS introduces a new step in multicentre study: a reproducible and valid method to assess the neurologic deficit following stroke, which is essential for multicentre trials.[7,8] As the predefined primary outcome of TRICS BASIC we selected the sensorimotor deficits measured at 48 hours after tMCAo by composite neuroscore (also defined to as the De Simoni neuroscore[11,12]). At variance with the clinical setting, the assessment of injury severity and outcome in experimental stroke models lacks a standard scoring system. We here chose a scoring system already used in the previous multicentric study by Llovera et al.,[13] proving to be 1) feasible across different centres, 2) well correlated to the histological measurement of the ischemic lesion (Pearson r 0.76 and 0.77 at 48 and 96 hours after tMCAo respectively[13]).

The present work reports the harmonization procedures for the evaluation of sensorimotor deficits by the neuroscore across the TRICS centres, performed before beginning the interventional phase of the project. The aim of this study was to verify whether the raters were able to assess the same sample of ischemic rats and mice with a substantial agreement. Different durations of MCA occlusion were used to allow sufficient variability in the neurologic outcome and raters were blinded to the experimental condition. We predefined our target for a satisfactory agreement at intraclass correlation coefficient (ICC) of 0.6, as described in the pre-registered study's protocol paper.[11]

Material and methods

Animal models and study setting

All experiments were carried out in animal facilities belonging to seven Italian academic or research institutions:

(I) Istituto di Ricerche Farmacologiche Mario Negri (IRFMN), the University of Calabria (UniCal) and San Raffaele Hospital (HSR) that used mice as animal model:

(II) The University of Firenze (UniFi), the University of Milano Bicocca (UniMib) and the University of Milano Statale (UniMi) that used rats as animal model.

(III) The University of Napoli (UniNa) used both species as animal models.

The experiments and the care of the animals were conducted in accordance with national (Decree-Law No. 26/2014) and international (EEC Council Directive 2010/63/UE; Dec. 12, 1987; Guide for the Care and Use of Laboratory Animals, US National Research Council Eighth Edition 2011) laws. All experiments on animals have been approved by the Ethics Committee of the University of Milano Bicocca (Organismo preposto al benessere animale: OPBA), the Coordinating body of the project and received authorization No. 1056/2020-PR, prot. FB7CC.43, by the Italian Ministry of Health. The experimental protocol of which the study is part of was registered with the following number: PCTE0000177 on https://preclinicaltrials.eu.

The protocols and details of this study are in accordance with the ARRIVE guidelines (see the list provided as a supplementary file).

Male C57BL/6J wild-type (WT) mice (24 g ± 10%, Charles Rivers Laboratories, Italy) and male Sprague-Dawley rats (250 g ± 5%) were housed in standard condition in an Specific Pathogen Free enclosure, in a single cage, exposed to 12 h controlled light/dark cycle and room temperature, food and water available *ad libitum*, for at least a week before any intervention. After surgery, the animals were housed under the same conditions for 48 hours.

Models of transient cerebral ischemia in rodents

Mice. Ischemia was performed by transient occlusion of the left or right middle cerebral artery (tMCAo).[14,15] Anaesthesia was administered by inhalation of 3% isoflurane in a gaseous mixture of oxygen and nitrous oxide (N_2O/O_2, 70%/30%) and maintained with 1.5% isoflurane in the same mixture. During the surgery, the animal was placed supine on a thermostatic bed equipped with a rectal probe to monitor and maintain the temperature at 37 ± 0.5°C. The surgical site was disinfected with clorexyderm 4% solution and a 1 cm incision was made in the midline of the neck. Using a dissecting microscope, the common carotid artery (CCA) was isolated and ligated upstream the bifurcation between the internal (ICA) and external (ECA) carotid artery. The occlusion of the middle cerebral artery (MCA) was achieved inserting a silicone rubber-coated monofilament (size 7–0, diameter 0.06–0.09 mm, diameter with coating 0.23 mm; coating length 6 mm, Doccol Corporation, Redlands, California, USA) into the ICA, which was pushed cranially to occlude the origin of the middle cerebral artery (MCA). Based on the surgical protocol available at each centre, the filament was inserted either from the CCA (IRFMN, HSR) or the ECA (UniCal, UniNa). To ensure better variability in the outcome, three different occlusion times were performed: 30, 45 or 60 minutes. During the occlusion, the animal was awakened from anaesthesia, kept in a warm box and tested for the presence of intra-ischemic deficits (for inclusion/exclusion criteria, see below). After the pre-established time of occlusion, the blood flow was restored by gently removing the filament, under anaesthesia. If the filament was inserted from the CCA, the artery was then permanently ligated. Otherwise the ECA was ligated and the CCA re-opened. Analgesia was achieved by local application of a local anaesthetic (EMLA, containing 2.5% lidocaine and prilocaine, Aspen Pharma). The established reperfusion period is 48 hours.

Rats. Anaesthesia was induced by 3% and maintained by 1.5% isoflurane inhalation in an N_2O/O_2 (70%/30%) mixture. The animal was subjected to occlusion of the origin of the MCA and to ensure a better variability in the outcome, three different times of ischemia[16] were performed: 50, 75 and 100 minutes, followed by a period of reperfusion of 48 hours. A silicone filament (size 5–0, diameter with coating 0.33 mm; length with coating 5–6 mm; Doccol Corporation, Redlands, California, USA), was introduced into the right external carotid artery and pushed through the internal carotid artery to occlude the origin of the right MCA. During the occlusion of the MCA, the rats were awakened from anaesthesia to assess the intra-ischemic clinical assessment which reveals the correct induction of ischemia. After the occlusion time (50, 75 and 100 minutes), blood flow was restored by carefully removing the filament under anaesthesia. During the surgery, the animal's body temperature was maintained at 37°C by a heating pad. After the surgery, all the rats were housed in single cages.

Inclusion and exclusion criteria

Rats and mice were included in the study if cerebral ischemia was successfully induced, that is, animals displayed the early focal deficits associated with the MCA occlusion. Namely, during the intra-ischemic period, we applied a clinical assessment score as described previously by centre IRFMN.[17] These inclusion criteria do not require specific tools to be applied and therefore could be easy to implement in a multicentric trial. Animals were judged ischemic, and included in the trial if presenting ≥3 of the following deficits during the intra-ischemic period:

1. The palpebral fissure has an ellipsoidal shape (not the normal circular one)
2. One or both ears extend laterally
3. Asymmetric body bending on the ischemic side

4. Limbs extend laterally and do not align with the body

   Animals would have been excluded in case of:

1. Death during MCA occlusion surgery
2. Major experimental protocol violations: errors or surgical complications (eg, major arterial or venous haemorrhage, section of the vagus nerve, carotid artery dissection, filament entrapment or displacement) during MCA occlusion procedure; errors in ischemia time.

Health monitoring

Animals were monitored at 24 and 48 hours after surgery, before the behavioural testing. A predefined Middle Cerebral Artery Occlusion (MCAo) health report (available at https://figshare.com, DOI: 10.6084/m9.figshare.13031861), prepared based on the Ischemia Models: Procedural Refinements Of in Vivo Experiments (IMPROVE) guidelines, was filled at baseline, at 24 hours and 48 hours with information on animal welfare. Animals showing signs of moderate distress, according to the MCAo health report, were treated subcutaneously with 0.05–0.1 mg/kg buprenorphine every 8–12 hours (this dose was used for both rats and mice).

Training for the execution of the neuroscore

We distributed tutorial videos to the involved centres, illustrating how to handle the animals during the execution of the behavioral test and how to evaluate the neuroscore. The videos showed the test execution both on an ischemic and on a healthy animal, to allow the detection of the difference in focal or general deficit and allow the centres to better understand the evaluation criteria. These video tutorials are available as Supplementary Information and present a clear description of the correct procedures to handle animals and assess the neuroscore.

Evaluation of neurological deficits

At 48 hours after the induction of the ischemia, each centre performed and recorded on video the neuroscore. The total amount of recorded videos were n = 11 for mice and n = 11 for rats from all the centres. The videos were sent to the coordinating centre for the blinding. A figure outside the study changed the number that identifies the name of the animal's video (numbers from T01 to T11 for mice and R01 to R11 for rats). The videos, thus blindly randomized, were uploaded to an online platform with free access to all the centres (Google Drive, shared folder TRICS Basic project, sub-folder Inter Rater Agreement). At each centre, an evaluator assigned a score to the 11 videos.

Evaluators were different from researchers doing surgery.[12] The score ranges from 0 (absence of deficits) to 56 (worst neurological result) and includes general and focal deficits. The general deficits describe the general well-being of the mouse with a score between 0 and 28. This score includes information on the physical appearance of the mouse, i.e.: fur (0–2), ears (0–2), eyes (0–4), posture (0–4), spontaneous activity (0–4) and presence of epileptic seizures (0–12). Focal deficits describe neurological damage with a score between 0 and 28 and were evaluated through observations on: body symmetry (0–4), gait (0–4), ability to climb a 45° inclined surface (0–4), circling behavior (0–4), forelimb symmetry (0–4), compulsory circling (0–4) and whisker response (0–4). All evaluations were entered on the REDCap online platform and retrieved by the coordinator for statistical analysis.

A detailed description of the neuroscore items can be found at https://figshare.com, DOI: 10.6084/m9.figshare.13031861, as presented in the protocol paper.[11] Deficits are registered regardless if seen on the left or the right side of the animal, so to allow the evaluation of either right or left-induced MCAo.

Measurement of the ischemic volume

After the neurobehavioral test, animals were sacrificed by deep narcosis with CO_2; brains were extracted and fixed in 10% formalin. Collected brains were sent to the TRICS coordinating unit UniMib and processed and evaluated by an operator blinded to the experimental condition (i.e. tMCAo duration and the centre executing the surgery). We collected 16 out of 22 total animals in trial 1, one of which could not be further processed for the histological analysis due to procedural errors. Coronal sections (100 μm of thickness) were obtained using Vibratome1000Plus (Leica) and stained using Cresyl Violet 0.1% (Bioptica, Milano, Italy). The ischemic volume was measured in 19 consecutive sections distanced by 200 μm (bregma +3.0 mm to −2.0 mm). Each section was mounted on a positively charged slide (SuperFrost Plus, Thermo Scientific) and rinsed in a saline solution (Dulbecco's Phosphate Solution w/Magnesium w/Calcium; Euroclone): only after 48 hours sections were stained with Cresyl Violet (Cresilvioletto Kluver Barrera 05–B16001; Bioptica) according to manufacturer's instructions. Sections were finally immersed in xylene (Sigma-Aldrich) to wash off the excess dye and dehydrate it, allowing assembly in dibutyl phthalate xylene (DPX non–aqueous mounting medium CL04.0401.0500; Chem_Lab NV). The ischemic volume was calculated using ImageJ image processing software (National Institute of Health, Bethesda, MD, USA), corrected for interhemispheric asymmetries due to cerebral edema with the

following equation: ischemic area = direct lesion volume − (ipsilateral hemisphere − contralateral hemisphere) and expressed in mm^3.

Intraclass correlation coefficient and definition of group size

The intraclass correlation (ICC) assesses the reliability of ratings by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects.

To limit the use of animals, the power analysis performed indicated that 11 animals for species were necessary. The sample size was calculated starting from the knowledge that 4 raters were available for each species. The expected intraclass correlation coefficient (ICC) was estimated to be approximately 0.80. The ICC ranges from −1 (perfect disagreement) to 0 (absence of agreement) to +1 (perfect agreement). When the sample size is 11, a two-sided 95% confidence interval computed using the large sample normal approximation for an intraclass correlation was calculated to extend about 0.17 from the observed intraclass correlation.

Fleiss's kappa. The interobserver agreement on the neuroscore comparing all raters was described using Fleiss' κ, ranging between 0 and 1.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Cohen's kappa. The evaluations that each of the 7 centres obtained from carrying out the neuroscore on the 11 videos of the mice and 11 of the rats were then compared in pairs. The interobserver agreement of the neuroscore comparing pairs of raters was described using Cohen's κ, ranging from $\kappa = 0$ (equivalent to chance) to $\kappa = 1$ (perfect agreement).

Correlation analysis

The analysis of correlation for paired centres' total neuroscore and for the neuroscore vs. ischemic volume was done using the Pearson correlation for normally distributed values or Spearman correlation for non-normally distributed values. Normality was assessed by the Kolmogorov-Smirnoff test.

Results

Training phase

All operated animals met the inclusion criteria and were thus included in the study. We did not observe any mortality for either species. After health monitoring, no animals showed signs of severe distress to require sacrifice before the experimental endpoint.

In the training phase of the project, we prepared video tutorials (available as Supplementary material) on sham and ischemic animals explaining the evaluation of sensorimotor deficits using the neuroscore. Tutorials were administered to each rater of the participating centres before starting the evaluation phase. The study was then conducted according to the plan depicted in Figure 1(a). Briefly, mice and rats were subjected to tMCAo with different durations (i.e. 30, 45 or 60 minutes for mice and 50, 75 or 100 minutes for rats) to increase variability of the observed deficits. After health monitoring at 24 and 48 hours post-surgery (according to the IMPROVE guidelines) animals underwent the neuroscore while being video recorded. The coordinating centre then collected all the videos and performed their blinding before redistributing them to each centre for the neuroscore assignment.

Interrater agreement showed a moderate consistency in the first trial

The interrater agreement on the total score range of the neuroscore (0–56) was described using the ICC. We reached a moderate agreement for mice ICC = 0.50 [0.22–0.77] (Figure 1(b)) and for rats ICC = 0.49 [0.21–0.77] (Figure 1(c)), which did not satisfy our cut-off of ICC ≥0.60.

We repeated the analysis after score dichotomisation, replacing with the parameter "good" if the total score was <21 and with "bad" if the total score was ≥21. This score cut-off was defined based on a previous work using the same neuroscore.[14] The Fleiss κ on the dichotomised score was $\kappa = 0.54$ for mice and $\kappa = 0.36$ for rats, meaning fair agreement for mice and slight for rats. As such, when score was dichotomized to discriminate between good and bad outcome the agreement was not satisfactory especially for rats.

In order to identify possible 'outlier centres', we calculated the interrater reliability on pairs of raters using the Cohen' κ coefficient. In mice, considering the dichotomised score, we obtained: fair agreement between HSR and UniCal ($\kappa = 0.30$), HSR and IRFMN ($\kappa = 0.30$); moderate agreement between HSR and UniNa ($\kappa = 0.42$); substantial agreement between UniCal and IRFMN ($\kappa = 0.61$), UniCal and UniNa ($\kappa = 0.79$), UniNa and IRFMN ($\kappa = 0.79$) (Figure 2(a)). In rats, we observed: poor agreement between UniFi and UniMi ($\kappa = 0$), UniFi and UniMiB ($\kappa = 0$); fair agreement between UniFi and UniNa ($\kappa = 0.39$); substantial agreement between UniMi and UniNa ($\kappa = 0.62$), UniMiB and UniNa ($\kappa = 0.62$); perfect agreement between UniMi and UniMiB ($\kappa = 1$) (Figure 2(b)). Thus HSR for mice and UniFi for rats
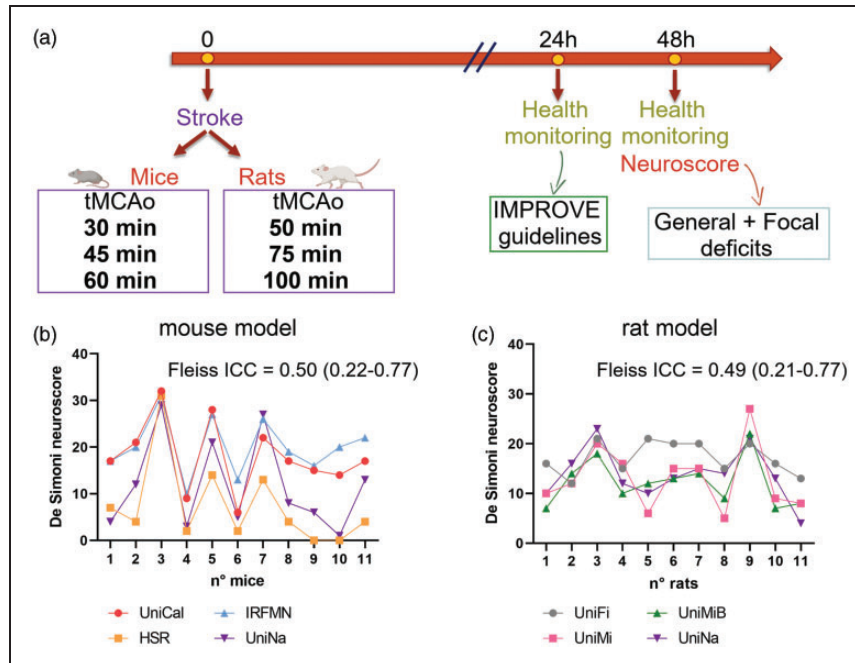
Figure 1. (a) Experimental plan. Eleven mice or rats were subjected to different durations of tMCAo. Animals were monitored at 24 and 48 hours post-surgery according to the IMPROVE guidelines. Sensorimotor deficits were assessed at 48 hours by the neuroscore and (b, c) Interrater agreement analysis on total score range of the neuroscore in the first trial. All scores given by centres are presented in the graphs. The interrater reliability was calculated by ICC and its values with 95% interval of confidence are indicated.

seemed to provide slightly different scores than other centres.

We then correlated the total score given by each rater using the Pearson or Spearman correlation coefficient depending on shape of data distribution. In mice we found Spearman $r = 0.88$ (UniCal-HSR, $p = 0.0006$), Pearson $r = 0.93$ (UniCal-IRFMN, $p < 0.0001$), Pearson $r = 0.84$ (UniCal-UniNa, $p = 0.0011$), Spearman $r = 0.74$ (HSR-IRFMN, $p = 0.0119$), Spearman $r = 0.78$ (HSR-UniNa, $p = 0.0059$), Pearson $r = 0.88$ (IRFMN-UniNa, $p = 0.0004$) (Figure 2(c)). In rats we found Pearson $r = 0.44$ (UniFi-UniMi, $p = 0.1711$), Pearson $r = 0.57$ (UniFi-UniMiB, $p = 0.0695$), Pearson $r = 0.47$ (UniFi-UniNa, $p = 0.1484$), Pearson $r = 0.84$ (UniMi-UniMiB, $p = 0.0012$), Pearson $r = 0.73$ (UniMi-UniNa, $p = 0.0115$), Pearson $r = 0.80$ (UniMiB-UniNa, $p = 0.0028$) (Figure 2(d)). The correlation analysis identified non significant correlations only when UniFi evaluations were paired with the other centres evaluating rats.

Overall scores correlated significantly with the ischemic volume measured at the same time (i.e. 48 hours), with a Spearman r of 0.61 and a $p = 0.018$ (Figure 2(e)).

Systematic errors during the execution of the neuroscore in the first trial

In order to identify the reasons for the poor agreement in the first trial - i.e. lower than our target of ICC

≥ 0.60 – we critically revised all videos to identify any experimental issues. We noticed errors during the evaluation of general and focal deficits, as reported in Figure 3. Typical errors regarded the observation of eyes (Figure 3(a) and (e)), the improper use of wool gloves (Figure 3(b)) and plastic sheets (Figure 3(f)) to assess animals' balance and the surface used to assess climbing (Figure 3(c) and (g)). We observed also errors in animal handling for evaluation of whisker response on the lesioned and contralateral side (Figure 3(d) and (h)), i.e. the use of pointed tweezers and the wrong position of the observer that was visible by the animals.

Interrater agreement showed a substantial consistency in the second trial

We replaced the videos with poor experimental execution with new correct ones. All videos were blinded to origin again and redistributed for evaluation according to the randomization plan depicted in Supplementary Table 1. In the second trial we reached a substantial agreement for mice, having an ICC $= 0.64$ [0.37–0.85] (Figure 4(a)) and for rats, ICC $= 0.69$ [0.44–0.88] (Figure 4(b)), both satisfactory according to our target (ICC ≥ 0.60). The Fleiss κ on the dichotomised score was $\kappa = 0.45$ for mice and $\kappa = 0.70$ for rats.

The interrater reliability calculated on pairs of raters was in mice: slight agreement between UniCal and
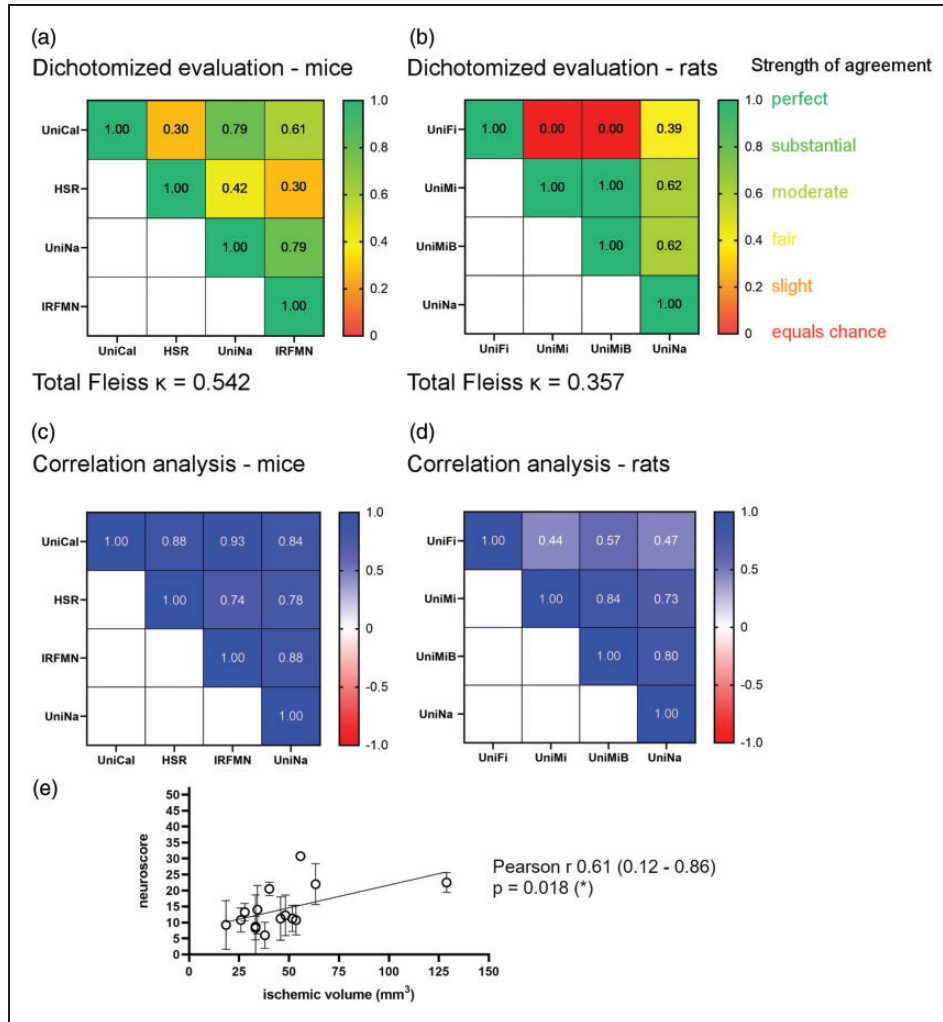
Figure 2. (a, b) Box presenting the interrater reliability calculated on pairs of raters using the Cohen' κ coefficient (indicated in each box). Red tones indicate poor while green tones strong agreement. (c, d) Box presenting the correlation between scores from pairs of raters. Red tones indicate poor while blue tones strong correlation. Pearson or Spearman correlation tests were performed for normal or non-normal distributed data per Kolmogorov-Smirnoff test and (e) Correlation between the neuroscore and the ischemic volume calculated at 48 hours after tMCAo. Data presented as mean of neuroscores attributed to each animal by the four centres $\pm$ SD and relative ischemic volume expressed in mm^3 (n $=$ 15 mice and rats, not all animals assessed for neuroscore could be analysed for the ischemic volume). Linear regression is shown. Spearman r 0.61, p $=$ 0.018.

UniNa ($\kappa = 0.23$), UniCal and IRFMN ($\kappa = 0.24$); moderate agreement between UniNa and IRFMN ($\kappa = 0.42$); substantial agreement between UniCal and HSR ($\kappa = 0.54$), HSR and IRFMN ($\kappa = 0.62$), HSR and UniNa ($\kappa = 0.74$) (Figure 4(c)). In rats, we observed moderate agreement between UniMiB and UniNa ($\kappa = 0.42$); substancial agreement between UniFi and UniNa ($\kappa = 0.62$), UniMi and UniNa ($\kappa = 0.62$), UniFi and UniMiB ($\kappa = 0.74$), UniMi and UniMiB ($\kappa = 0.74$); perfect agreement between UniFi and UniMi ($\kappa = 1$). With correlation analysis, we found in mice: Pearson r $= 0.69$ (UniCal-HSR, p $= 0.0184$), Pearson r $= 0.79$ (UniCal-IRFMN,

p $= 0.0037$), Pearson r $= 0.74$ (UniCal-UniNa, p $= 0.0092$), Pearson r $= 0.80$ (HSR-IRFMN, p $= 0.0033$), Pearson r $= 0.79$ (HSR-UniNa, p $= 0.0037$), Pearson r $= 0.95$ (IRFMN-UniNa, p < 0.0001) (Figure 4(e)). In rats we found Spearman r $= 0.54$ (UniFi-UniMi, p $= 0.0855$), Spearman r $= 0.63$ (UniFi-UniMiB, p $= 0.0414$), Spearman r $= 0.33$ (UniFi-UniNa, p $= 0.3185$), Spearman r $= 0.68$ (UniMi-UniMiB, p $= 0.0250$), Spearman r $= 0.38$ (UniMi-UniNa, p $= 0.2470$), Spearman r $= 0.44$ (UniMiB-UniNa, p $= 0.1735$) (Figure 4(f)).

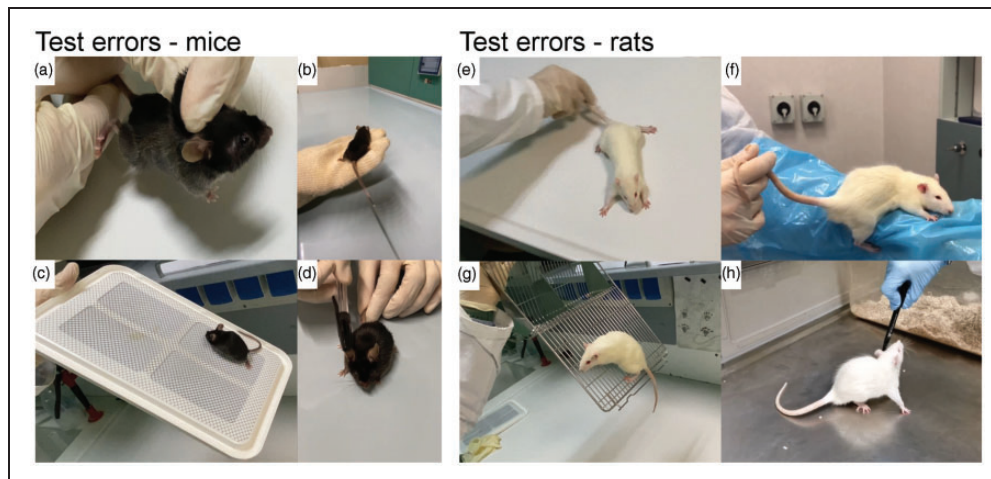Scores stratified by tMCAo duration did not differ in either trials (Supplementary Table 2).

Figure 3. Typical animal handling errors during the neuroscore first trial. In particular: (a, e) interference when observing the eyes; (b) improper use of wool gloves and (f) plastic sheet to assess animals' balance; (c, g) incorrect surface to assess climbing and (d, h) wrong handling during the evaluation of whisker response.

Intra-rater score correlations revealed good consistency between the two trials in mouse, but not rat evaluations

Exploiting the videos that were evaluated in both trials (7 for mice and 8 for rats) after blinding, we could calculate the intra-rater agreement, i.e. how the two blind evaluations on the same animal correlated for each rater (Figure 5). As reported in Table 1, raters evaluating mice were more consistent in the two trials compared to those evaluating rats, with an overall r of 0.83 (0.66–0.92 CI 95%), $p < 0.0001$, compared to 0.69 (0.45–0.84), $p > 0.0001$. In the second trial, the total score increased by +2.2 for mice and +1.2 for rats, indicating a better ability of raters to identify the deficits associated with the ischemic models.

Discussion

In the present study we harmonized the behavioral procedures for the evaluation of sensorimotor deficits across the centres involved in the TRICS project. This work originally presents an effective workflow to standardize the assessment of the pre-defined primary outcome in a multicentre preclinical study.

Preclinical randomized controlled trials (pRCTs) have been introduced to improve stroke preclinical research, with the final ambition of overcoming the lack of its clinical translability. A pRCT reported by Llovera et al.[13] anticipated the results of a clinical trial on Natalizumab efficacy, showing no improved outcomes of treated stroke patients compared to placebo,[18] thus proving pRCT as reliable predictive tools. However pRCTs are still uncommon and have shown some weaknesses, especially in how the studies were designed and performed. Specifically, not all the good practices for solid clinical trials have been implemented in pRCTs, including trial pre-registration and protocol standardization. Recently the SPAN pRCT was launched and its published stage 1 results could confirm that a large, multilaboratory, preclinical assessment effort to reduce known sources of bias is feasible and practical.[10]

Clinical trials in patients are conducted following detailed and pre-registered protocols, which clearly describe the study design, the randomization procedure, the primary outcome, secondary outcomes and sample size estimate. *In vivo* preclinical studies, particularly multi-centre confirmatory trials whose aim is to translate a promising therapy to clinical studies, should be as similar as possible to human clinical trials.[8,19] Preclinical researchers are expected to implement the available guidelines for the standardization and reporting data of *in vivo* animal research.[7,8] Nonetheless, a recent study revealed that over 85% of published animal studies did not describe any randomization or blinding strategies and over 95% lacked the estimation of sufficient sample size needed for detecting the true effects in the intervention studies.[20] Reported flaws of preclinical studies include the lack of *a priori* inclusion/exclusion criteria, randomization with intention-to-treat logic, *pre-hoc* power test study, replication of findings, reporting of full animal details and definition of a quality check strategy.[21] When we designed the TRICS BASIC study we planned to consider all the above points by implementing similar practices to those applied in phase III clinical trials,[22] i.e. 1) the experimental protocol has been pre-registered at https://preclinicaltrials.eu (ID: PCTE0000177), 2) a protocol paper has been published,[11] 3) all experimental
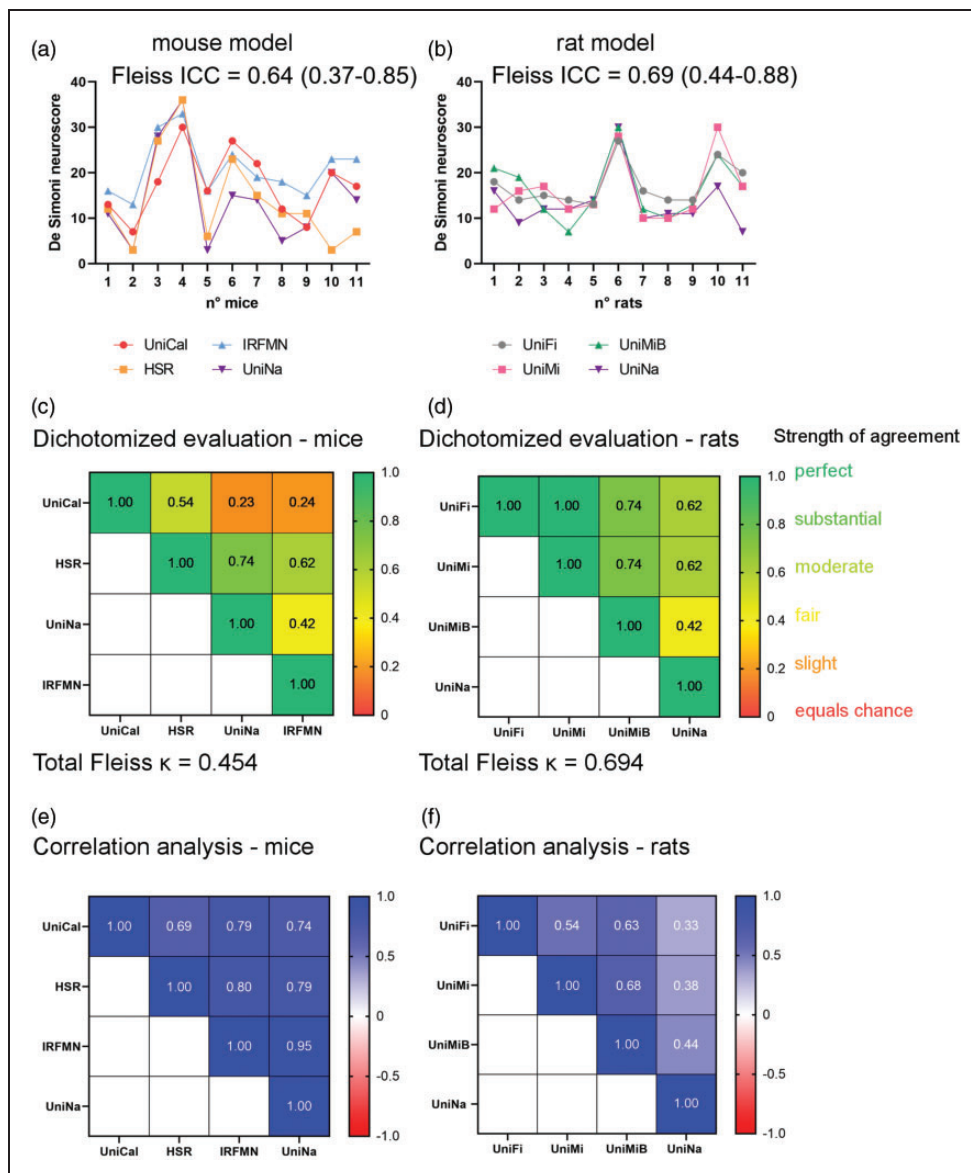
Figure 4. Improved interrater agreement after the second trial. (a, b) All scores given by centres are presented in the graphs. The interrater reliability was calculated by ICC and its values with 95% interval of confidence are indicated. (c, d) Box presenting the interrater reliability calculated on pairs of raters using the Cohen' κ coefficient (indicated in each box). Red tones indicate poor while green tones strong agreement and (e, f) Box presenting the correlation between scores from pairs of raters. Red tones indicate poor while blue tones strong correlation. Pearson or Spearman correlation tests were performed for normal or non-normal distributed data per Kolmogorov-Smirnoff test.

subjects will be recorded on a REDCap-based online database, 4) a 'preclinical monitor' will supervise centres' compliance to the experimental plan.

Key to standardization and quality check was to harmonize the evaluation of sensorimotor deficits, set as the primary outcome that was successfully obtained in the present work. As a multicentre trial, the agreement among the individuals collecting data – here referred to as inter-rater agreement – can be immediately observed due to the fluctuation among the raters.

Inter-rater agreement can vary on the individuals' different expertise with the specific assessments.[23–25] This is the reason why we decided to implement a training phase trial for data collectors (raters) before the start of the trial, in order to reduce the variability in the way raters assess and interpret the neurobehavioral data.[26] Although the perfect agreement is difficult to achieve, a substantial agreement was deemed to be required before starting animal randomization, considering the translational aim of multicentre preclinical trials.[27,28]
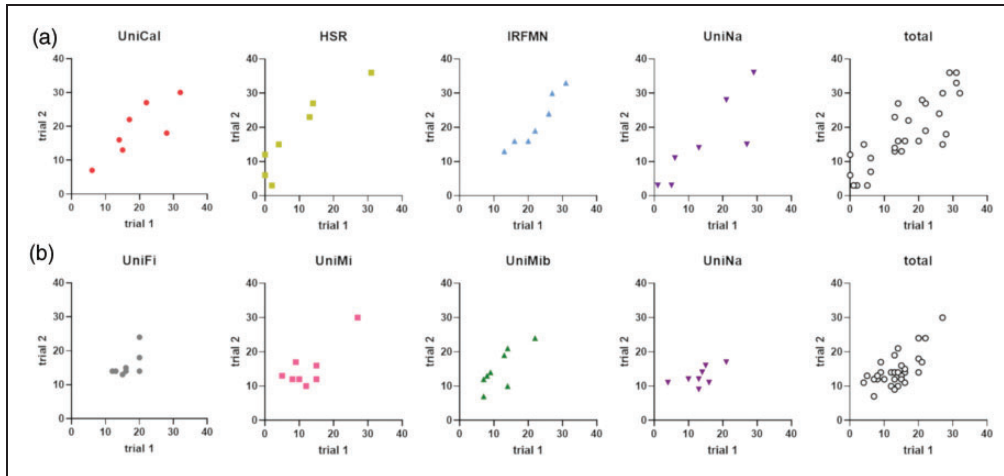
Figure 5. Intra-rater agreement, i.e. the correlation of trial 1 vs. trial 2 scores by the same rater on the same mouse (a) or rat (b). The relative Pearson r, 95%-CI and p values are depicted in Table 1.

Table 1. Intra-rater agreement's correlations and their p-values.

Species	Centre	Pearson r	95% CI	P value
Mouse				
	UniCal	0.81	0.16–0.97	0.0257 (*)
	HSR	0.93	0.57–0.99	0.0025 (**)
	IRFMN	0.95	0.68–0.99	0.0011 (**)
	UniNa	0.84	0.25–0.98	0.0170 (*)
	Total	0.83	0.66–0.92	<0.0001 (***)
Rat				
	UniFi	0.61	−0.17–0.92	0.1088
	UniMi	0.81	0.24–0.96	0.0150 (*)
	UniMib	0.78	0.16–0.96	0.0229 (*)
	UniNa	0.59	−0.21–0.91	0.1262
	Total	0.69	0.45–0.84	<0.0001 (***)

*p < 0.05; **p < 0.01; ***p < 0.001.

The interrater agreement on the total score range of the neuroscore (0–56) was described using the Intraclass Correlation Coefficient, setting a target of ICC $\geq$ 0.60 as a substantial agreement, as per protocol paper.[11] The cut-off was based on both methodological and translational reasons, since intraclass agreements greater than 0.60 are commonly considered good to excellent[29] and that previous studies reported that such ICC was achievable when assessing the NIH-stroke scale (NIHSS) or the modified Rankin score (mRS) in stroke patients.[30,31]

We could reach the satisfactory agreement with two rounds of training. After the second training, we improved the ICC from 0.50 to 0.64 for the evaluation of mice, and from 0.49 to 0.69 for that of rats. The fact that we could not obtain a substantial inter-rater agreement at the first trial largely depended on operators' mistakes that regarded animal handling or the use of unsuitable devices, then corrected before the second

trial. Also, the raters using for the first time this neuroscore tended to give low deficit scores, thus failing to identify deficits when not overtly present. In line with this, considering the seven randomized mouse videos analysed blindly in both trials, the overall increase of neuroscore compared to the first trial was +2.2 for mice and +1.2 for rats. Moreover the neuroscore was applied to the ischemic rats for the first time here, and required some protocol adjustments, especially in animal handling. When we analyzed the intra-rater agreement by correlating same rater's trial 1 vs. trial 2 score on the same animals, we obtained higher correlations for mice than rats. After the second training, our work proved the applicability of the neuroscore to the rat model of ischemic stroke, a key finding in view of the interventional phase of the TRICS project. It should be noted that, besides its under-cut-off inter-rater agreement, the first trial neuroscore correlated with the histological measure of the ischemic volume, thus confirming a previous observation over a larger cohort of animals.[13]

We believe that the neuroscore proposed here may be a standard neurobehavioral assessment in large multicentric preclinical stroke trials, due to its reliability and easy implementation not requiring complicated tools. In order to allow other scientists in the field of stroke research to implement the neuroscore, we provided in Supplementary Information the video tutorials presenting the assessment of an ischemic and a sham mouse.

Our scoring system would help align experimental stroke models to the clinical setting, where stroke patients are assessed by standard scoring systems, like the NIHSS for injury severity and the mRS for longer term outcome. We used here the neuroscore to assess

an early deficit, in line with the primary endpoint of the TRICS preclinical trial[11] and its parallel clinical trial,[32] i.e. observing an early neurological improvement after RIC. In view of using the neuroscore for long term outcome in rodents, it was previously shown to identify sensorimotor deficits over 5 weeks after the ischemic onset, when assessed longitudinally in a cohort of tMCAo mice.[33]

Our study is the first specifically designed to increase reliability of neurobehavioural scoring as a primary outcome in multicentre preclinical trials. A multi-step, online harmonization phase proved to be feasible, easy to implement and highly effective to improve the agreement between the raters of different centres and with different skills. A potential limitation of our study is generalizability since the harmonization phase performed for the TRICS preclinical trial might not be applied *tout-court* to other preclinical models or more complex neurobehavioral tests. A customized approach according to the study protocol is likely to be needed to maximize agreement under different experimental conditions.

To conclude, our findings strongly indicate that a harmonization phase reduces bias in the neurobehavioral assessment used as a primary outcome in multicentre preclinical stroke trials and could be considered as a basic requirement before starting animal randomization.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability

Raw data are available at Figshare, doi: 10.6084/m9.figshare.21346731

Authors' contributions

AV, JM performed the experiments, analyzed and interpreted the data and drafted the ms. MT conceived and coordinated the study statistical analysis, analyzed the data and revised the ms. SF, SB conceived the study, analyzed and interpreted the data, drafted the ms. MGDS, CF supervised the study and revised the ms. GP, LS, GB, GM, FP, MB supervised the work at participating centres, revised the ms. SS, CP, DA, CDS, OC, AV, SD, MV, FS, PCZ, VR-M, LC, JR, ID, IB, EC, SGG, EB performed the experiments and collected the data on REDCap. All authors have given their final approval of the current version.

ORCID iDs

Alessia Valente https://orcid.org/0000-0002-6917-9278
Giuseppe Pignataro https://orcid.org/0000-0002-7290-4397
Diana Amantea https://orcid.org/0000-0002-7513-3495
Marco Bacigaluppi https://orcid.org/0000-0002-0213-2328
Francesco Santangelo https://orcid.org/0000-0001-7724-6875

Supplementary material

Supplemental material for this article is available online.

References

1. Roth GA, Mensah GA, Johnson CO, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 study. *J Am Coll Cardiol* 2020; 76: 2982–3021.
2. Pignataro G, Scorziello A, Di Renzo G, et al. Post-ischemic brain damage: effect of ischemic preconditioning and postconditioning and identification of potential candidates for stroke therapy. *Febs J* 2009; 276: 46–57.
3. O'Collins VE, Macleod MR, Donnan GA, et al. 1,026 Experimental treatments in acute stroke. *Ann Neurol* 2006; 59: 467–477.
4. Schmidt-Pogoda A, Bonberg N, Koecke MHM, et al. Why most acute stroke studies are positive in animals but not in patients: a systematic comparison of preclinical, early phase, and phase 3 clinical trials of neuroprotective agents. *Ann Neurol* 2020; 87: 40–51.
5. Stroke Therapy Academic Industry Roundtable (STAIR). Recommendations for standards regarding preclinical neuroprotective and restorative drug development. *Stroke* 1999; 30: 2752–2758.
6. Fisher M, Feuerstein G, Howells DW, et al. Update of the stroke therapy academic industry roundtable preclinical recommendations. *Stroke* 2009; 40: 2244–2250.
7. Sert NP, Du Hurst V, Ahluwalia A, et al. The ARRIVE guidelines 2.0: updated guidelines for reporting animal research. *PLOS Biol* 2020; 18: e3000410.
8. Percie Du Sert N, Alfieri A, Allan SM, et al. The IMPROVE guidelines (ischaemia models: procedural refinements of in vivo experiments). *J Cereb Blood Flow Metab* 2017; 37: 3488–3517.
9. Dirnagl U. Thomas Willis lecture. *Stroke* 2016; 47: 2148–2153.
10. Lyden PD, Bosetti F, Diniz MA, et al. The stroke preclinical assessment network: rationale, design, feasibility, and stage 1 results. *Stroke* 2022; 53: 1802–1812.
11. Tettamanti M, Beretta S, Pignataro G, et al. Multicentre translational trial of remote ischaemic conditioning in acute ischaemic stroke (TRICS): protocol of multicentre, parallel group, randomised, preclinical trial in female and male rat and mouse from the Italian Stroke Organization (ISO) basic science network. *BMJ Open Sci* 2020; 4: e100063.
12. Orsini F, Villa P, Parrella S, et al. Targeting mannose-binding lectin confers long-lasting protection with

a surprisingly wide therapeutic window in cerebral ischemia. *Circulation* 2012; 126: 1484–1494.

13. Llovera G, Hofmann K, Roth S, et al. Results of a preclinical randomized controlled multicenter trial (pRCT): anti-CD49d treatment for acute brain ischemia. *Sci Transl Med* 2015; 7: 299ra121.

14. Neglia L, Oggioni M, Mercurio D, et al. Specific contribution of mannose-binding lectin murine isoforms to brain ischemia/reperfusion injury. *Cell Mol Immunol* 2020; 17: 218–226.

15. Bacigaluppi M, Pluchino S, Jametti LP, et al. Delayed post-ischaemic neuroprotection following systemic neural stem cell transplantation involves multiple mechanisms. *Brain* 2009; 132: 2239–2251.

16. Riva M, Pappadà GB, Papadakis M, et al. Hemodynamic monitoring of intracranial collateral flow predicts tissue and functional outcome in experimental ischemic stroke. *Exp Neurol* 2012; 233: 815–820.

17. Mercurio D, Piotti A, Valente A, et al. Plasma-derived and recombinant C1 esterase inhibitor: binding profiles and neuroprotective properties in brain ischemia/reperfusion injury. *Brain Behav Immun* 2021; 93: 299–311.

18. Elkind MSV, Veltkamp R, Montaner J, et al. Natalizumab in acute ischemic stroke (ACTION II): a randomized, placebo-controlled trial. *Neurology* 2020; 95: e1091–e1104.

19. Ioannidis JPA. Why most published research findings are false. *PLOS Med* 2005; 2: e124.

20. Laajala TD, Jumppanen M, Huhtaniemi R, et al. Optimized design and analysis of preclinical intervention studies in vivo. *Sci Rep* 2016; 6: 30723.

21. Dirnagl U. Bench to bedside: the quest for quality in experimental stroke research. *J Cereb Blood Flow Metab* 2006; 26: 1465–1478.

22. Nosek BA, Ebersole CR, DeHaven AC, et al. The pre-registration revolution. *Proc Natl Acad Sci U S A* 2018; 115: 2600–2606.

23. Gisev N, Bell JS and Chen TF. Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Res Soc Adm Pharm* 2013; 9: 330–338.

24. Goldstein LB, Bertels C and Davis JN. Interrater reliability of the NIH stroke scale. *Arch Neurol* 1989; 46: 660–662.

25. Benchoufi M, Matzner-Lober E, Molinari N, et al. Interobserver agreement issues in radiology. *Diagn Interv Imaging* 2020; 101: 639–641.

26. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Medica* 2012; 22: 276–282.

27. Tymianski M. Neuroprotective therapies: preclinical reproducibility is only part of the problem. *Sci Transl Med* 2015; 7: 299fs32.

28. Dewey HM, Donnan GA, Freeman EJ, et al. Interrater reliability of the national institutes of health stroke scale: rating by neurologists and nurses in a community-based stroke incidence study. *Cerebrovasc Dis* 1999; 9: 323–327.

29. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 1994; 6: 284–290.

30. Zhao H, Collier JM, Quah DM, et al. The modified Rankin scale in acute stroke has good inter-rater-reliability but questionable validity. *Cerebrovasc Dis* 2010; 29: 188–193.

31. Lyden P, Raman R, Liu L, et al. National institutes of health stroke scale certification is reliable across multiple venues. *Stroke* 2009; 40: 2507–2511.

32. Diamanti S, Beretta S, Tettamanti M, et al. Multi-center randomized phase II clinical trial on remote ischemic conditioning in acute ischemic stroke within 9 hours of onset in patients ineligible to recanalization therapies (TRICS-9): study design and protocol. *Front Neurol* 2021; 12: 724050.

33. Sammali E, Alia C, Vegliante G, et al. Intravenous infusion of human bone marrow mesenchymal stromal cells promotes functional recovery and neuroplasticity after ischemic stroke in mice. *Sci Rep* 2017; 7: 6962.