Journal of
Cerebral Blood Flow
& Metabolism

Quality and validity of large animal experiments in stroke: a systematic review

Journal:	*Journal of Cerebral Blood Flow and Metabolism*
Manuscript ID	JCBFM-0015-20-REV.R1
Manuscript Type:	Review Article
Date Submitted by the Author:	16-Apr-2020
Complete List of Authors:	Kringe, Leona; University Hospital Freiburg, Neuroradiology - Neurocenter; University of Leipzig Faculty of Veterinary Medicine, Institute of Veterinary Anatomy, Histology and Embryology Sena, Emily; University of Edinburgh, Clinical Neurosciences Motschall, Edith; University of Freiburg, Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center Bahor, Zsanett; University of Edinburgh, Clinical Neurosciences Wang, Qianying; University of Edinburgh, Clinical Neurosciences Herrmann, Andrea; University Hospital Freiburg, Neuroradiology - Neurocenter; University of Leipzig Faculty of Veterinary Medicine, Institute of Veterinary Anatomy, Histology and Embryology Mülling, Christoph; University of Leipzig Faculty of Veterinary Medicine, Institute of Veterinary Anatomy, Histology and Embryology Meckel, Stephan; Universitatsklinikum Freiburg, Neuroradiology Boltze, Johannes; University of Warwick, School of Life Sciences
Research Topics:	Acute Stroke, Animal Models, Animal studies, Basic Science, Brain Imaging, Brain Ischemia, Cerebrovascular disease, Experimental, Focal Ischemia, Imaging, Interventional Neuroradiology, MRI, Neuroanatomy, Neurodegeneration, Neuroprotection, Neuroradiology, Stroke, Systematic reviews
Research Techniques:	Animal studies, Brain Imaging, Imaging, MRI

Quality and validity of large animal experiments in stroke: a systematic review

Leona Kringe, DVM[1,2]; Emily S. Sena, PhD[3]; Edith Motschall[4]; Zsanett Bahor, PhD[3]; Qianying Wang, MSc[3]; Andrea M. Herrmann, DVM[1,2]; Christoph Mülling, DVM, PhD[2]; Stephan Meckel, MD[1]; Johannes Boltze, MD, PhD[5]

[1]Department of Neuroradiology, Neurocenter,

Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

~~University Hospital Freiburg, Freiburg, Germany~~

[2]Faculty of Veterinary Medicine, Institute of Anatomy, Histology and Embryology, Leipzig University, Leipzig, Germany

[3]Centre for Clinical Brain Sciences, ~~The~~ University of Edinburgh, Edinburgh, United Kingdom

[4]Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

[5]School of Life Sciences, University of Warwick, Coventry, United Kingdom

Please address correspondence to:

Johannes Boltze, MD, PhD
School of Life Sciences, University of Warwick
Coventry CV4 7AL United Kingdom,
E-mail: johannes.Boltze@warwick.ac.uk

Running headline: Role and analysis of methodological quality in large animal experiments in stroke

Abstract

An important factor for successful translational stroke research is study quality. Low-quality studies are at risk of biased results and effect overestimation, as has been intensely discussed for small animal stroke research. However, little is known about the methodological rigor and quality in large animal stroke models, which are becoming~~e~~ more frequently used in the field.

Based on research in two databases, this systematic review surveys and analyses the methodological quality in large animal stroke research. Quality analysis was based on the Stroke Therapy Academic Industry Roundtable (STAIR) and the Animals in Research: Reporting In Vivo Experiments (ARRIVE) guidelines. Our analysis revealed that large animal models are utilized with similar shortcomings ~~than~~ as small animal models. Moreover, translational benefits of large animal models may be limited due to lacking implementation of important quality criteria such as randomization, allocation concealment, and blinded assessment of outcome. On the other hand, an increase of study quality over time and a positive correlation between study quality and journal impact factor were identified.

Based on the obtained findings, we derive recommendations for optimal study planning, conducting and data analysis/reporting when using large animal stroke models to fully benefit from the translational advantages offered by these models.

Key words: ~~large animal models, stroke, brain ischemia, translational research, preclinical research, quality guidelines, quality assurance, study validity, predictive value, bias, study reproducibility~~large animal, stroke, preclinical research, study quality, study validity

1. Introduction

Acute ischemic stroke management and care have profoundly improved with the introduction of intravenous thrombolysis and, recently, mechanical thrombectomy for large vessel occlusions.[1] However, by far not all patients can benefit from the therapeutic progress due to numerous contraindications, restricted availability and therapeutic time windows of these therapeutic approaches. This causes a tremendous need for novel treatment options, but the translation of preclinical findings into clinically applicable and efficient therapies has so far been mostly ineffective and prone to failure.[2]

Critical assessment of rodent studies revealed that one important reason for the translational failure is the lack of methodological quality in these preclinical studies, causing a higher risk for poor internal validity, overestimation of effect sizes, and biased conclusions thus affecting rationale and design of subsequent clinical trials.[3,4,5]

Large animal models become more frequently used in preclinical stroke research since they are believed to provide a number of significant advantages in the translational process.[6,7] On the other hand, large animal stroke models are both more laborious and more expensive to utilize than rodent models. Budgetary limitations often restrict sample sizes in large animal experiments, ~~what~~ which limits statistical power.[8] Hence, it is essential to conduct large animal experiments with highest methodological rigor and to predefine precise endpoints that can be assessed with sufficient statistical power ~~in order~~ to take full advantage of the translational value of large animal stroke models.

Little is known about the methodological rigor and quality of large animal stroke experiments. We performed a systematic review and quality assessment of studies using large animal stroke models. Our quality analysis was based on the Stroke Therapy Academic Industry Roundtable (STAIR)[9,10] and Animals in Research: Reporting In Vivo Experiments (ARRIVE) guidelines.[11] Based on the obtained results, we also provide suggestions for methodological improvements in large animal stroke research.

2. Material & Methods

2.1. Study selection

Literature research was performed by the first author (L.K.). L.K. was supported by E.M., a professional librarian with extensive experience in systematic literature research who helped with designing the search strategy. The two last authors (S.M. and J.B.) were consulted by L.K. in case of any doubts or questions when extracting information from the literature. Intra-assessor reproducibility was not assessed.

2.1.1. Search strategy

We conducted a systematic search for preclinical large animal experiments in stroke using the Medline via Ovid from Wolters Kluwer and Science Citation Index Expanded via Web of Science from Clarivate Analytics data bases.

The initial search was conducted on September 26th, 2017, and an update was performed on August 9th, 2019. Data base entries between January 1st, 1990 and August 8th, 2019 were covered.

Search terms were “large animal” (including any relevant species, e.g. dogs, cats, pigs, rabbits, non-human-primates, sheep, goats, etc.) and “ischemic stroke” (involving for instance “brain ischemia” OR “ischemic neuronal injury” OR “thrombembolic stroke” OR “cerebrovascular disorders”). In the search strategies we combined the aspects *large aninals* and *ischemic stroke* with AND. Within each aspect w~~W~~e generally combined keywords, their synonyms and – for indexed citations of MEDLINE – controlled for vocabulary terms (Medical Subject Headings) using the operator OR. Detailed search strategies are provided in Supplementary Tables 1 and 2. The search process was conducted and results were recorded according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Figure 1A).

2.1.2. Inclusion and exclusion criteria

We included preclinical large animal studies conducted and published between 1990 and 2019 that report investigations of therapeutic and/or diagnostic ~~interventions~~ procedures for ischemic stroke. The studies needed to compare at least two groups, i.e. one in which a new procedure (therapeutic or diagnostic) is tested by comparing it to a second group being subjected to a standard or reference procedure ("control group"). Only studies in English were included.~~report a control and an interventional arm. Only studies in English were included.~~

We excluded studies focusing on diseases other than ischemic stroke, using small animal (e.g., rodent) models, clinical trials, in vitro studies, reviews, and meta-analyses. Purely descriptive studies only reporting a method or procedure, or non-controlled experiments (e.g., cases series) were also excluded.

2.2. Data extraction

2.2.1. Basic study characteristics and impact factor

First, study meta-data were extracted. Those included information on species, type of intervention, year of publication and region of origin (North America, Europe, Asia & Oceania), aim of evaluation (e.g., safety, feasibility), the stroke model used, study duration and information on investigation of dose-response-relationship (if applicable), compliance with animal welfare regulations, subject health condition prior to enrolment, animal housing conditions, and additional veterinary care.

Second, we documented the ~~journal~~ impact factor (IF) of the journal in which the study results were published, measured in the year of publication. IFs were identified via the annual Thomson Reuters Journal Impact Factor report. ~~In case~~Where the IF could not be retrieved for the required year, we contacted the respective journal and asked to provide the IF for the particular year(s).

2.2.2. Group sizes

We further extracted the number of subjects in experimental groups for each species. Group sizes were obtained for control and the diagnostic or therapeutic procedure group(s).

2.3. Analysis

2.3.1. Assessment of Reporting Quality

We designated a scale that was applicable to both, diagnostic and therapeutic procedures, to assess study quality (Table 1).~~A score was designed to assess study quality (Table 1).~~ The quality score includes central STAIR and ARRIVE criteria, supplemented by additional quality items. The score comprised four categories, containing 6 items each. Category 1 addresses reporting of study subject details and welfare, category 2 covered the reporting of details on study design, category 3 addressed internal study validity, and category 4 assessed quality of outcome analysis and reporting. Each study was assigned a score from 0 (lowest quality) to 24 (highest quality), with each category having a quality value of 0 (lowest quality) to 6 (highest quality).

[Table 1 about here]

2.3.2. Additional aspects influencing study quality

We further investigated whether study quality improved after the implementation of the STAIR guidelines in 1999, and their update in 2009.[9,10] We also analyzed differences in quality with respect to species, region of study origin, and type of investigation (i.e., assessment of neuroprotectives, thrombolytics, cell therapies, diagnostics, and others). Furthermore, we evaluated possible associations between the quality score and ~~impact factor.~~IF.

2.3.3. Group sizes

Where a study reported more than one procedure group, they were all counted individually (maximum number was n=10). Average group sizes were calculated for control and procedure groups(s) for each species. We compared total group size (control plus procedure groups) across species as well as control and procedure groups separately.

2.4 Statistics

All statistical analyses were performed using GraphPad PRISM 5 Software. Statistical significance was determined as p<0.05. Statistical significance was indicated with a single asterisk (*) at p<0.05, or a double asterisk (**) at p<0.01, respectively. Median as well as IQR (interquartile range including 25% and 75% quartiles) were documented. Comparisons between two groups were performed using the Wilcoxon signed rank test for non-parametric data to conservatively account for relatively small sample sizes. In case more than two groups were compared, the Kruskal-Wallis test was used, followed by Dunn`s correction for multiple comparisons. Spearman’s correlation analysis was performed to evaluate associations between quality score and ~~impact factor.~~IF. Group sizes were analyzed by ANOVA on ranks (no normal distribution of data) followed by Dunn's multiple comparison test.

3. Results

3.1. Data set and year of publication

Initial and update searches identified a total of 10282 manuscripts being reduced to 8093 after elimination of duplicates. (Figure 1A; a list of all studies included can be found in the supplementary material). A total of 208~~9~~ studies were included in final analysis after screening abstracts and full text according to preset inclusion and exclusion criteria (Figure 1A). Results of basic study characteristics are shown in Table 2.

Analysis of publication output per year revealed that the number of large animal experiments published from 1990 to 2014 generally decreased from n=56 in 1990-1994 to n=21 in 2010-2014 (Figure 1B). However, there was a steep increase in published studies from 2015, reaching an all-time high (n=40) even though studies published in late 2019 are not yet included in our search strategy. This might be related to the milestone evidence for clinical benefit publication of mechanical thrombectomy in large vessel occlusion stroke by the publication of five randomized controlled trials in 2015 that may have sparked new interest in the field and an increased demand for large animal models to investigate related procedures.[12,13] ~~Analysis of publication output per year revealed that the number of large animal experiments published from 1990 to 2019 generally decreased after reaching a peak in the mid 1990s (Figure 1B). There were more publications in the 1990s (n=93) than in the last decade (n=62). However, publication output remarkably increased since 2014 with 47 studies published between 2014 and 2019.~~

[Figure 1 about here]

[Table 2 about here]

3.2. Study Quality

The overall median quality score was 11~~2~~ (range 3 to 22; IQR: 4 (9-13)) out of 24. The median quality score in the first category (reporting of study subject details and welfare) was 2 out of 6 (range 1 to 5; IQR: 1 (1-2)). The second category (study planning quality) also reached a median quality score of 2 (range 1 to 6; IQR: 1 (2-3)). The third category (study conductance quality) had a median score of 3 (range 0 to 6; IQR: 2 (2-4)). Category 4 (result reporting and analysis quality) had a median quality score of 4 (range 0 to 6; IQR: 1 (2~~3~~-4)). A significantly lower number of quality criteria were fulfilled in category 1 in comparison to the others ($p<0.05$).

3.2.1. Study subject details and welfare (category 1)

All studies reported the species used, but only 146 studies (70.2~~69.9~~%) reported that the study was approved by responsible animal welfare authorities. Sex and age were reported by 31 studies (15.0~~4.8~~%). Sex only was reported by 153 (73.6~~2~~%), while age was not reported solely. The pre-study health status was reported by only 12 studies (5.8~~7~~%). Medication details including the use of companion medication (e.g., analgetics, antibiotics) was reported in only 20 studies (9.6%). Comorbidities were not reported by any study.

3.2.2. Study planning (category 2)

Working hypotheses were reported in 207~~8~~ (99.5%) studies. However, primary study endpoints were nominally determined in only 10 studies (4.8%). 135 (64.6%) studies reported that the study rationale was based on earlier small animal (n=79; 38.0~~7.8~~%) or in vitro studies (n=25~~6~~; 12.1~~4~~%), or both (n=16; 7.7%). Effect size estimation and a priori sample size calculation can be performed based on such data. However, only 27 studies (13.0~~2.9~~%) actually reported an estimation of effect size and a priori sample size calculation. A specific primary working hypothesis explicitly referring to previous in vitro and/or in vivo studies was reported in 18 studies (8.7~~6~~%). Inclusion and exclusion criteria were reported in 104 studies (~~49.8~~50.0%), but only 2 studies (1.0%) determined these criteria a priori.

3.2.3. Study conductance (category 3)

Randomization was reported in 116 studies (55.8~~5~~%), and allocation concealment was reported in 59 cases (28.4~~2~~%). 104 studies (~~49~~50.0~~.8~~%) reported blinded outcome assessment. Measurement of physiological parameters was reported in 165~~6~~ cases (79.34%). The most frequently monitored parameters included mean arterial pressure (systemic), temperature, blood gases, blood pH, and exhalation gases. 186~~7~~ studies reported appropriate outcome analysis modalities (89.4~~5~~%; information on inappropriate analysis modalities are provided in Supplementary Table 3). These included survival rate (n=2; 1.0 %), functional outcome (n=67; 32.2%), infarct size (n=46; 22.1%, as determined by appropriate methods such as imaging or histology), other imaging (n=90; 43.3%) or histology (n=61; 29.3%) endpoints, clinical chemistry (n=52; 25.0%), general pathology (n=24; 11.5%) or both (n=18; 8.7%)~~These included survival rate (n=2; 1.0 %), functional outcome (n=67; 32.1%), imaging endpoints (n=91; 43.6%), clinical chemistry (n=52; 24.9%), general pathology (n=24; 11.5%) or histology (n=61; 29.2%) or both (n=18; 8.6%), as well as infarct size (n=46; 22%) as determined by appropriate methods such as imaging or histology~~. Only a fraction of studies that recorded physiological parameters finally analyzed those (n=52; ~~24.9~~25.0%). 100 studies (48.1~~7.8~~%) reported verification of infarct induction during intervention.

3.2.4. Result reporting and analysis (category 4)

168~~9~~ studies (80.8~~9~~%) adequately reported relevant data and findings in form of detailed tables or graphs. However, data were almost exclusively reported as means or medians. Individual data points were only provided by 16~~7~~ studies (7.7~~8.1~~%). Drop outs and excluded subjects were reported in 105 studies (50.5~~2~~%). Application of appropriate statistical tests was reported in 192~~3~~ studies (92.3%). 16 studies incompletely reported statistical analysis and, for ~~instance~~ example lacking information regarding statistical tests applied including post hoc tests. 91 studies (43.8~~5~~%) described potential sources of error and bias in the experiment, while 115 (55.3~~0~~%) reported limitations such as small sample size or ~~impossibility~~ that it was impossible to perform randomization. A conclusion fully justified by study findings was given in by most, but not all reports (n=190~~1~~; 91.34%).

3.3. Additional influences on study quality

3.3.1. Study quality versus origin, species and type of intervention

Total median quality score was highest in studies from North America (Median: 12~~1~~; IQR: ~~5.75 (8.25-14~~10-14~~)~~), ~~but not~~ statistically different from studies conducted in Asia & Oceania (Median: 10; IQR: ~~3.25 (~~8.75-12~~)~~) ~~as well as those from~~or Europe (Median: 10; IQR: ~~3.75 (~~8-11.75; ~~)) (~~p=0.~~1516~~ 0011 Figure 2A). Analysis of individual quality categories revealed no differences in category 1 (Figure 2B) but North American studies had statistically significantly higher scores in quality categories 2 (Median: 2.5; IQR: ~~2 (~~2-3)~~)~~ and 3 (Median: 4; IQR: ~~2 (~~3-5~~)~~) than their European counterparts (Median: 2; IQR: 1-2; p<0.01; Figure 2C). ~~In the second category, North American studies performed better than their European counterparts (Median: 2; IQR: 1 (1-2)) (p<0.01; Figure 2C).~~ Furthermore, North American studies were superior to Asian & Oceanian studies in category 3 (Median 3; IQR: ~~2 (~~2-4~~)~~; ~~) (~~p<0.01; Figure 2D). We did not find statistically significant differences regarding category 4 (Figure 2E). Quality scores were neither influenced by species used (Figure 2F) nor by the types of intervention (Figure 2G). Overall differences in median quality score in species varied significantly without any specific intergroup difference.

[Figure 2 about here]

3.3.2. Study quality in the post-STAIR era

M~~In general, m~~ethodological quality significantly improved after ~~publication~~ introduction of the ~~first~~ STAIR guidelines in 1999 (1990-1999 p~~P~~re-STAIR~~tair~~ m~~M~~edian: 10, IQR: ~~4 (~~8-12~~)~~; p~~P~~ost-STAIR~~Stair~~ m~~M~~edian: 12, IQR: ~~6 (~~9-15~~)~~; p<0.01; Figure 2H). We also compared quality scores of studies published prior to the first STAIR guidelines to quality scores of studies published~~(1990-1999; Median: 10, IQR: 4 (8-12)),~~ in the time between the first STAIR guideline publication and the 2009 update (2000-2009; m~~M~~edian: 11; IQR: ~~4 (~~9-13~~)~~), and to scores of studies published after the STAIR ~~preclinical guideline~~2009 update (2010-2019; m~~M~~edian: 13~~2~~.~~50~~; IQR: ~~5 (~~10-15~~)~~). Quality scores of studies published after the STAIR 2009 update were higher than those of studies published before the initial STAIR guideline publication (1990-1999; p<0.01). They were also higher than quality scores of studies published after the first publication of STAIR guidelines and prior to the 2009 update (2000-2009; p<0.05; Figure 2I).~~We found significantly higher quality in studies conducted between 2010-2019 compared to those performed prior STAIR guideline publication (1990-1999; p<0.01) and those performed after the first publication of STAIR guidelines (2000-2009; p<0.05; Figure 2I).~~

Improvements were particularly evident in categories 1 and 4. In category 1, quality scores were lower in pre-STAIR studies (1990-1999; median: 1, IQR: 1-2) as compared to studies published after the first publication of STAIR guidelines and prior to the 2009 update (2000-2009; median: 2; IQR: 1.25-2) and to studies published after the 2009 update (2010-2019; median: 2; IQR: 2-3; p<0.01). There was also a significant difference in category 1 quality scores of studies published after the 2009 update to studies published between 2000 and 2009 (p<0.01). In category 4, quality scores of studies published after the 2009 STAIR update (2010-2019; median: 4; IQR: 3-5) were higher than those of studies published before the STAIR guidelines introduction (1990-1999; median: 3; IQR: 2-4) and those of studies published between 2000 and 2009 (median 3; IQR: 2-4; p<0.01 each).

~~(1990-1999 (Median: 1, IQR: 1 (1-2)) vs. 2000-2009 (Median: 2; IQR: 0.75 (1.25-2)) and 1990-1999 vs. 2010-2019 (Median: 2; IQR: 1 (2-3)) and 2000-2009 vs. 2010-2019 (p<0.01)) and 4 (1990-1999 (Median: 3; IQR: 2 (2-4)) vs. 2010-2019 (4; IQR: 2 (3-5)) and 2000-2009 (Median 3; IQR: 2 (2-4)) vs. 2010-2019) (p<0.01).~~

3.3.3. Study quality versus impact factor

The IF was ~~documented~~ available for 172~~3~~ studies (82.7~~8~~%). We could not retrieve the IF for the remaining studies or no IF yet assigned on the particular journal in the year of publication (n=36; 17.3~~2~~%). These latter studies were therefore excluded from the following analyses. Median IF was 3.3 (range 0.1 to 41.6; IQR: ~~2.65 (~~2-4.6~~5)~~). Correlation analysis showed a statistically significant positive relationship between the total quality score and the ~~journal impact factor~~IF (r=0.~~2723~~2802; p<0.01, alpha=0.05; Figure 3). We also correlated each quality score category with the IF and found that quality scores in all individual categories positively correlated with the IF (category 1: r=0.~~1918~~1851; p<0.05; category 2: r=0.1653~~17~~; p<0.05; category 3: r=0.1858~~769~~; p<0.05; category 4: r=0.2297~~185~~; p<0.01; Supplementary Figure 1).

[Figure 3 about here]

3.3.4. Group sizes

Average group sizes across species are given in Table 3. Analysis of group sizes revealed that total (combined control and procedure) group size was largest in rabbits as compared to pigs (p<0.01), sheep and primates (p<0.05 each). Total group sizes in cats were larger than those in sheep (p<0.05; Figure 4A). Accordingly, control groups were largest in rabbits as compared to pigs (p<0.05) and primates (p<0.01; Figure 4B), while procedure groups were largest in rabbits as compared to pigs (p<0.01; Figure 4C).

[Table 3 about here]

[Figure 4 about here]

4. Discussion

Systematic bias may cause over- or underestimation of study results.[3] Quality items such as randomization, allocation concealment, and blinded assessment ~~help to~~ improve internal validity [14~~2~~], but are often neglected in small animal studies.[3, 15~~3~~, 16~~4~~]

Large animal models are believed to offer significant benefits for translational stroke research. ~~Those comprise a~~They have higher anatomical similarity to the human brain[17~~5~~] and to the cerebrovascular system~~cerebrovascular system anatomy~~.[6, 7, 18~~6~~] Another benefit is the potential to use these models in experiments closely mimicking a human clinical situation, and applying the same medical techniques and equipment for diagnostic and therapeutic interventions that would be used in human patients.[7, 19~~7~~] Moreover, physiological characteristics of large animal models including heart and respiratory frequency, blood pressure as well as pharmacodynamic and pharmacokinetic profiles are similar to humans.[20~~18~~, 21~~19~~] However, in view of these advantages, large animal studies require much greater efforts and resources. It is therefore important that quality in large animal studies is as high as possible to efficiently utilize the advantages large animal models offer for translational research.

Overall, we found that methodological quality in large animal stroke studies was mediocre. Although quality generally improved significantly over the last decades and potentially due to the 1999 publication and 2019 update of the STAIR criteria, our analysis revealed some important shortcomings. Improvements are needed in reporting study subject details and welfare (quality score category 1). Aspects such as sex and age, pre-study health conditions, and medications should be reported routinely for optimal study transparency and reproducibility, and transferability of study results.[9] The lack of comorbid large animal models is not surprising. Comorbidities are difficult to simulate in outbred large animal models as they occur due to age, distress, malnutrition and other factors according to the human situation~~Comorbidities may hardly be mimicked in outbred large animal models as they occur due to age, distress, malnutrition and other factors according to the human situation~~, and can take significant time in large animals to develop. Research on models exhibiting comorbidities may remain a domain of small animal research. Nevertheless, any spontaneously occurring comorbidities being diagnosed in large animals used for research should be reported.

Working hypotheses were reported in almost all studies (99.5%), but often without any obvious influence on study design. For instance, only 4.8% of the studies defined and reported primary endpoints, while analysis of expectable effect size and a priori sample size calculation were performed in few cases only (13~~2~~.0%). This may severely limit the translational benefits of large animal models since ~~neutral~~ study~~ies~~ results may be hard to interpret based on potentially poor statistical power. Given the significant resources required to perform large animal studies, considering these aspects is essential. On the other hand, determination of effect size can be challenging when previous research data is lacking or not entirely applicable. In these cases, we recommend to perform large animal pilot studies that may help to assess basic characteristics in the respective model, such as variability of ~~stroke~~ infarct size and its impact on the envisioned primary endpoint.

While ~~almost two thirds~~half of the studies reported inclusion and exclusion criteria (~~64.6~~50.0%), almost none (1.0~~1~~%) applied them a priori. Defining inclusion and exclusion criteria during or after the study is believed to be a major source of bias, particularly when a study is conducted in non-~~un~~blinded fashion. Hence, such bias can unfortunately not be excluded for most studies we analyzed.

Important quality aspects such as randomization (55.8~~5~~%), allocation concealment (28.4~~2~~%), and blinded assessment of outcome (~~49.7~~50.0%) were more frequently reported in large animal studies as compared to small animal stroke experiments (randomization: 33.3%; blinded assessment of outcome: 44.4%,[16~~4~~] allocation concealment: 25.9%; randomization, allocation concealment and blinded assessment of outcome: 24.1.%.[22~~0~~] Nevertheless, the number of studies not reporting those is still remarkably high in particular since blinding and randomization should~~all~~ be minimum standard quality assurance procedures in confirmative stroke research[23~~1~~] to which almost all large animal studies aim to contribute.

Imaging techniques such as magnetic resonance imaging, computed tomography, and angiography (43.5%) as well as physiological monitoring (80.4%) were utilized relatively frequently. This is a positive aspect since large animals are particularly suitable for clinical imaging techniques while thorough physiological monitoring creates meaningful information that may warrant subject in- or exclusion. However, verification of infarct induction (only reported in ~~47.8~~48.1%) as well as infarct size should be conducted thoroughly and routinely to avoid the risk of increasing inter-subject/-study/-group variability, further reducing statistical power of an experiment. Parameters such as ~~reduced~~ cerebral blood flow reduction for verification of infarct induction was documented by only 7.2% of studies. This is surprising since these parameters are relatively easy to determine in large animals, while clinical imaging techniques may be used to confirm the induced lesion directly.[21~~19~~]

Large animals are suitable for long-term studies including functional endpoint assessment. However, we only found a relatively low percentage (6.7%) of studies being conducted for more than one month, the minimum follow-up period recommended by the STAIR guidelines for functional endpoints. Next to costs, t~~T~~his may be due to the selection of other primary endpoints such as safety or efficacy of recanalization methods which can be assessed more rapidly. However, experimenters who wish to assess behavioral endpoints should take into consideration that functional consequences of stroke in large animals can be more heterogeneous than in rodent models, and may develop over longer time spans.[24~~2~~]

We recognized significant improvements in methodological quality since the publication of the first STAIR guidelines in 1999, and in particular after the STAIR guideline update in 2009. Comparable improvements were reported for small animal stroke studies from 2010 to 2013.[25~~3~~] These findings indicate the positive impact of specific good research practice guidelines, which should be advanced continuously as evidenced by the recent 2019 STAIR guideline updates.[26~~4~~] In contrast to previous findings in small animal studies,[27~~5~~] we also identified positive association (r=0.~~2723~~2802; p<0.01) between study quality and publication in high-impact journals. In particular, total quality score as well as quality scores in all single categories 1-4 significantly correlated with higher IF. This is an encouraging result since all these categories include items being important to prevent bias. These items are hence ~~essential~~ indispensable for a valid and transparent exchange of information between researchers.

Group sizes were significantly larger in rabbits as compared to other species. This is not surprising as rabbits are the smallest and cheapest of all large animal species what allows for larger group sizes. Importantly, group sizes in primates are generally not different to that of other species. This does not mean that group sizes were sufficient for each research question, but shows that costs related to primate experiments did not prevent the same group sizes as seen in other large animal species despite rabbits.

Our study has a number of limitations. We applied a predefined search strategy and protocol being developed together an expert in literature meta-analyses (E.M.) and experts in stroke research (J.B., S.M.). However, search strategy and protocol were not registered (ex ante protocol). Data extraction was not done in duplicate~~s~~, but senior experts were consulted in all doubtful cases. Intra-assessor reproducibility was not assessed. Moreover, we did not discriminate between studies focusing on therapeutic and diagnostic procedures. Large animal models provide a number of benefits over rodent models for diagnostic studies due to the larger brain size and in particular when clinical imaging is used.[33] However, those studies are often exploratory in nature. Since quality demands are different (and a bit lower) than in confirmative studies, those imaging-related studies would perform normally worse but still can contribute invaluably to their respective field.[34] Finally, we did not include a number of insightful imaging studies because they did not conduct a formal inter-group comparison.[35,36,37,38]

5. Conclusions and Recommendations

Although large animal models offer a offer a number of clear advantages for translational stroke~~clear benefit in many translational stroke studies~~, we found that they ~~are utilized~~have ~~with~~ similar shortcomings ~~than~~ to small animal models, limiting ~~this benefit~~this benefit. ~~Hence~~Therefore, we derived a number of recommendations ~~that may overcome~~to address these limitations but are, at the same time, relatively easy to implement.

5.1 Study planning and preparation

Large animal stroke studies are mostly confirmative studies. ~~Hence~~Therefore, study planning should be based on high quality standards applied for randomized controlled clinical trials (RCTs) when possible. Key elements of RCT planning and design such as a priori sample size calculation and endpoint definition should be conducted.[23~~1~~] We encourage to involve statisticians already in early planning steps~~Statisticians may be involved already in early planning steps~~ to optimize study design.[28~~6~~] Study planning can also be supported by specific software tools. For instance, the National Centre for the Replacement, Refinement and Reduction of Animals in Research provides a freeware called Experimental Design Assistance (https://eda.nc3rs.org.uk), which is free to use and was built to guide researchers through their study planning.[29~~7~~] Since optimal sample sizes may not be achieved for all endpoints, it is important to clearly define the most appropriate primary study endpoint, and to power the study properly. Collaboration between research teams in form of peer quality checks and validation of study design can highly increase objectivity and validity of a study.[30~~28~~] Inter-group collaboration and transfer of experience can also help to handle very complex models and/or experimental setups, helping to reduce inter-subject variability negatively affecting statistical power. Confirmative studies might be preregistered to maximize transparency.[39]

5.2 Effect size estimation and pilot trials

Collecting valid information from previous research is essential for reliable effect size estimation. If such data ~~is~~ are not available, pilot studies may be helpful for at least basically estimating variability of stroke impact and outcome in the model. In case previous experience with a particular model is low, variability is more likely to be ~~overestimated~~ higher and effect size is more likely to ~~be underestimated from~~lower in such pilot trials. ~~, contributing to~~ This will contribute to more conservative study planning since sample sizes calculated based on that information will be higher. ~~conservative study planning~~. An important side effect of pilot trials is experimenter training which limits experimenter-caused endpoint variability (see below) in the main experiment. In addition, meta-analyses can help to collect relevant information on effect size or regarding a specific research question from related fields.[31~~29~~]

5.3 Reducing the effect of sample size limitations and endpoint variability

Financial and logistical restrictions often impact sample and group sizes in large animal experiments. This is an understandable limitation which is ~~hard~~ difficult to overcome. Selection of a proper and relevant endpoint that can be adequately powered with respected to the addressed research question ~~(not necessarily functional outcome)~~ is therefore important to minimize the risk for low statistical power. Of note, some endpoints often used in studies assessing therapeutic interventions~~,~~ including infarct size and functional deficits, exhibit a higher variability in large animal models than in rodent~~ ones~~. ~~, making~~ This makes comparison of absolute data more difficult.[24~~2~~] Relative analysis of repeatedly assessed endpoints, i.e. in comparison to the individual initial infarct size and/or functional deficit can efficiently compensate for such variability~~, allows to efficiently compensate for such variability~~. Repeated assessments also allow calculating the area under the curve for particular endpoints. This may provide a benefit in statistical power to identify whether a real outcome benefit is present over time. However, this comes at the cost of temporal resolution: it cannot be concluded exactly when this benefit became evident. There is also preliminary evidence for fast and slow stroke progressors in large animals, indicating different collateral status and somewhat resembling the human situation, but further contributing to inter-subject variability. It is recommended to consider this fact when planning an acute stroke study.[32~~0~~]

In experiments of highly similar design, controls may be pooled. Of note, this counteracts randomization and therefore requires extremely thorough validation of comparability of control subjects from different experiments/sources. If comparability is thoroughly proven, this may help to increase statistical power, but the limitations of this approach and potentially resulting bias need to be discussed transparently and in detail when publishing results.

The possibility to repeatedly collect a broad spectrum of physiological data should be utilized ~~to the best~~where possible~~ extend~~, as deviation from normal parameter ranges may explain variability and warrant post-hoc exclusion of subjects in single cases.

5.4 Study duration and documentation

We recommend considering long-term experiments whenever meaningful and possible and meeting animal welfare requirements. Even though long-term experiments involve greater efforts, the amount of data collected for individual subjects may be much higher, providing a better overall picture on the assessed intervention. Documentation should be as transparent as possible ~~since ~~because transparency is not challenging or laborious, but contributes significantly to increased scientific rigor, reproducibility, and unbiased study result interpretation~~at all~~. Methodological limitations including lacking quality aspects due to good reason should be clearly stated as this allows better interpretation of positive, neutral and negative study results.

6. Acknowledgements

ES is supported by the Stroke Association (SA L-SNC 18\1003).~~**We thank Qianying Wang, Centre for Clinical Brain Sciences, University of Edinburgh for her expert guidance on data analysis.**~~

7. Disclosure

The authors do not report relevant disclosures.

8. Supplementary data

1. Supplementary Table 1: Search strategy in Medline
2. Supplementary Table 2: Search strategy in Web of Science
3. Supplementary Table 3: Type and frequency of inappropriate analysis methods
4. Supplementary Figure 1: Association between impact factor and quality score within individual categories
5. Supplementary reference list

Supplementary material for this paper can be found at the journal website: http://journals.sagepub.com/home/jcb

9. References

1. Bush CK, Kurimella D, Cross LJ, et al. Endovascular Treatment with Stent-Retriever Devices for Acute Ischemic Stroke: A Meta-Analysis of Randomized Controlled Trials. *PLoS One*. 2016; e0147287.
2. O'Collins VE, Macleod MR, Donnan GA, et al. 1,026 experimental treatments in acute stroke. *Ann Neurol* 2006; 59: 467-477.
3. Macleod MR, Fisher M, O`Collins V, et al. Reprint: Good laboratory practice: preventing introduction of bias at the bench. *Int J Stroke* 2004; 59: 3-5.
4. Macleod MR, Lawson Mc Lean A, Kyriakopoulou A, et al. Risk of Bias in Reports of In Vivo Research: A Focus for Improvement. *PLoS Biol* 2015; 13: e1002273.
5. Sena ES, van der Worp HB, Bath PM, et al. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biol* 2010; 8: e1000344.
6. Herrmann AM, Meckel S, Gounis MJ, et al. Large animal in neurointerventional research: A systematic review on models, techniques and their application in endovascular procedures for stroke, aneurysms and vascular malformations. *J Cereb Blood Flow Metab*. 2019; 39(3): 375-394.
7. Traystman RJ. Animal models of focal and global cerebral ischemia. *ILAR J* 2003; 44: 85-95.
8. Sena ES, Van der Worp HB, Howells D, et al. How can we improve the pre-clinical development of drugs for stroke? *Trends Neurosci* 2007; 30: 433-439.
9. STAIR. Recommendations for standards regarding preclinical neuroprotective and restorative drug development. *Stroke* 1999; 30: 2752-2758.
10. Fisher M, Feuerstein G, Howells D, et al. Update of the stroke therapy academic industry roundtable preclinical recommendations. *Stroke* 2009; 40: 2244-2250.
11. Kilkenny C, Browne WJ, Cuthill IC, et al. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol* 2010; 8: e1000412.

12. Jovin TG, Albers, GW, Liebeskind DS, et al. Stroke Treatment Academic Industry Roundtable: The Next Generation of Endovascular Trials. *Stroke* 2016; 47(19): 2656-2665.

13. Goyal M, Menon BK, van Zwam WH, et al. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomized trials. *Lancet* 2016; 387(10029): 1723-1731.

~~11.~~

~~12.~~14. Sena ES, Currie GL, McCann SK, et al. Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically. *J Cereb Blood Flow Metab* 2014; 34: 737-742.

~~13.~~15. Sena ES, Van der Worp HB, Bath PM, et al. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biol* 2010; 8: e1000344.

~~14.~~16. Macleod MR, Van der Worp HB, Sena ES, et al. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke* 2008; 39: 2824-2829.

~~15.~~17. Boltze J, Nitzsche F, Jolkkonen J, et al. Concise Review: Increasing the Validity of Cerebrovascular Disease Models and Experimental Methods for Translational Stem Cell Research. *Stem Cells* 2017; 35: 1141-1153.

~~16.~~18. Sommer CJ. Ischemic stroke: experimental models and reality. *Acta Neuropathol* 2017; 133: 245-261.

~~17.~~19. Mehra M, Henninger N, Hirsch JA et al. Preclinical acute ischemic stroke modeling. J *Neurointerv Surg* 2012; 4: pp. 307-3013.

~~18.~~20. Helke KL and Swindle MM. Animal models of toxicology testing: the role of pigs. *Expert Opin Drug Metab Toxicol* 2013; 9: 127-139.

~~19.~~21. Herrmann AM, Cattaneo GFM, Eiden, SA, et al. Development of a Routinely Applicable Imaging Protocol for Fast and Precise Middle Cerebral Artery Occlusion Assessment and Perfusion Deficit Measure in an Ovine Stroke Model: A Case Study. *Front Neurol.* 2019; 10: 1113.

~~20.~~22. Minnerup J, Zentsch V, Schmidt A, et al. Methodological Quality of Experimental Stroke Studies Published in the Stroke Journal: Time Trends and Effect of the Basic Science Checklist. *Stroke* 2016; 47: 267-272.

~~21.~~23. Dirnagl U. Bench to bedside: the quest for quality in experimental stroke research. *J Cereb Blood Flow Metab* 2006; 26: 1465-1478.

~~22.~~24. Boltze J, Modo MM, Mays RW, et al. Stem Cells as an Emerging Paradigm in Stroke 4: Advancing and Accelerating Preclinical Research. *Stroke* 2019; 50 (11): 3299-3306.

~~23.~~25. Minnerup J, Wersching H and Diederich K. Methodological quality of preclinical stroke studies is not required for publication in high-impact journals. *J Cereb Blood Flow Metab* 2010; 30: 1619-1624.

~~24.~~26. Savitz SI, Baron JC, Fisher M, et al. Stroke Treatment Academic Industry Roundtable X: Brain Cyoprotection Therapies in the Reperfusion Era. *Stroke* 2019. 50(4): 1026- 1031.

~~25.~~27. Minnerup J, Wersching H and Diederich K. Methodological quality of preclinical stroke studies is not required for publication in high-impact journals. *J Cereb Blood Flow Metab* 2010; 30: 1619-1624.

~~26.~~28. Würbel H. More than 3Rs: the importance of scientific validity for harm-benefit analysis of animal research. *Lab Anim (NY)* 2017; 46: 164-166.

~~27.~~29. Percie du Sert N, Bamsey I, Bate ST, et al. The Experimental Design Assistant. *PLoS Biol* 2017; 15: e2003779.

30. Sena ES, Currie GL, McCann SK, et al. Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically. *J Cereb Blood Flow Metab* 2014; 34: 737-742.

~~28.~~

~~29.~~31. Begley CG and Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res* 2015; 116: 116-126.

32. Shazeeb MS, King RM, Brooks OW, et al. Infarct Evolution in a Large Animal Model of Middle Cerebral Artery Occlusion. *Transl. Stroke Research* 2019. doi: 10.1007/s12975-019-00732-9.

33. Werner P, Saur D, Zeisig V, et al. Simultaneous PET/MRI in stroke: a case series. *J Cereb Blood Flow* 2015; 35(9): 1421-1425.

34. Dirnagl U, Hakim A, Macleod M, et al. A concerted appeal for international cooperation in preclinical stroke research. *Stroke* 2013; 44(6): 1754-1760.

35. Boltze J, Ferrara F, Hainsworth AH, et al. Lesional and perilesional tissue characterization by automated image processing in a novel gyrencephalic animal model of peracute intracerebral hemorrhage. *J Cereb Blood Flow Metab* 2019; 39(12): 2521-2535.

36. Haque ME, Gabr RE, Zhao X, et al. Serial quantitative neuroimaging of iron in the intracerebral hemorrhage pig model. *J Cereb Blood Flow Metab* 2018; 38(3): 375-381.

37. Kamimura HA, Flament J, Valette J, et al. Feedback control of microbubble cavitation for ultrasound-mediated blood-brain barrier disruption in non-human primates under magnetic resonance guidance. *J Cereb Blood Flow Metab* 2019; 39(7): 1191-1203.

38. Sander CY, Mandeville JB, Wey HY, et al. Effects of flow changes on radiotracer binding: Simultaneous measurement of neuroreceptor binding and cerebral blood flow modulation. *J Cereb Blood Flow Metab* 2019; 39(1): 131-146.

39. Kimmelmann J and Anderson JA. Should preclinical studies be registered? *Nat Biotechnol* 2012; 30(6): 488-489.

Figure legends

Figure 1. Overview on quantitative search results and frequency of large animal experiments in stroke research since 1990.

(A) Flow diagram of publication identification. N = Number of publications. Records were excluded after screening title and abstracts. Full-text articles were then screened and excluded for a priori determined reasons. (B) Timeline of publication in large animal stroke research (1990-2019): The increase of large animal stroke studies in the last years is potentially due to the breakthrough in recanalization therapies, prompting a number of follow-on translational studies utilizing large animal stroke models.

Figure 2. Influence of study origin and STAIR criteria publication on study quality.

(A) Total quality score, (B) Category 1: Reporting of study subject and animal welfare, (C) Category 2: Study planning quality (North America vs. Europe $p<0.01$), (D) Category 3: Study conductance quality (North America vs. Asia & Oceania $p<0.01$), (E) Category 4: Result reporting and analysis quality (North America vs. Europe $p<0.01$), (F) ~~Improvement in total methodological quality since the publication of the first STAIR criteria in 1999 ($p<0.01$),~~Influence of species, (G) Improvement in total methodological quality since the publication of the first STAIR criteria ~~comparing to their amendment~~ in 1999 (~~2010-2019 vs. 1990-1999~~ $p<0.01$), ~~and 2010-2019 vs. 2000-2009 $p<0.05$),~~ (H) Improvement in total methodological quality since the publication of the first STAIR criteria in 1999 comparing to their amendment in 2009 (2010-2019 vs. 1990-1999 $p<0.01$, and 2010-2019 vs. 2000-2009 $p<0.05$), ~~Influence of species,~~ (I) Influence of type of intervention. Horizontal lines and whiskers indicate the median~~an~~ with lower and upper 95% CI. *$p<0.05$; **$p<0.01$.

Figure 3. Association between total quality score versus impact factor.

Scatterplot shows correlation between quality score and ~~impact factor~~ IF ($p<0.01$).

Number of included studies is 172~~3~~, no IF could be retrieved for 36 studies. The latter studies were excluded from this analysis.

Figure 4. Group sizes across species.

(A) Total group sizes were largest in rabbits as compared to pigs (p<0.01), primates and sheep (p<0.05 each). (B) Control group sizes were larger in rabbits as compared to primates (p<0.01) and pigs (p<0.05). (C) Procedure group sizes were larger in rabbits as compared to pigs (p<0.01). Horizontal lines and whiskers indicate the median with lower and upper 95% CI. *p<0.05, **p<0.01.

Tables

Table 1. Quality score items.

Category 1: Reporting of study subject details and welfare		**Category 2: Study planning quality**	
Item	*Score point allocation*	*Item*	*Score point allocation*
1. Animal protocol approved	Reported yes=1/no=0	1. Study hypothesis	Reported yes=1/no=0
2. Species	Reported yes=1/no=0	2. A priori endpoint definition	Reported yes=1/no=0
3. Sex and Age	Reported yes=1/no=0	3. A priori sample size calculation	Reported yes=1/no=0
4. Pre-Study Health	Reported yes=1/no=0	4. Reference to previous studies	Reported yes=1/no=0

5. Comorbidities	Reported yes=1/no=0	5. Inclusion/Exclusion criteria	Reported yes/no=0
6. Adequate medication	Reported yes=1/no=0	6. Effect size/Treatment effect	Reported yes=1/no=0
Category 3: Internal study validity		**Category 4: Outcome analysis and reporting**	
Item	*Score point allocation*	*Item*	*Score point allocation*
1. Blinding	Reported yes=1/no=0	1. Individual data points	Reported yes=1/no=0
2. Randomization	Reported yes=1/no=0	2. Drop outs/Excluded subjects	Reported yes=1/no=0
3. Allocation concealment	Reported yes=1/no=0	3. Appropriate statistical tests	Used yes=1/no=0
4. Physiological parameters	Measuring reported yes=1/no=0	4. Potential error sources	Reported yes=1/no=0
5. Analysis modalities	Appropriate modalities reported[#] yes=1/no=0	5. Study/Methodological limits	Reported yes=1/no=0
6. Infarct induction confirmation	Reported yes=1/no=0	6. Justified conclusion given[##]	Provided yes=1/no=0

[#] analysis modalities were considered appropriate when being sufficient to assess the respective research question or endpoint (see Supplementary Table 3 for details).

[##] conclusion was considered justified when supported by correctly analyzed results.

Table 2. Basic Characteristics of included Animal Experimental Studies.

Item	Frequency (%)	Item	Frequency (%)	Item	Frequency (%)
Species		**Type of intervention**		**Study duration**	
Rabbit	n=96 (~~45.9~~46.1%)	Neuroprotectives	n=113 (54.3~~1~~%)	Acute phase (<24h)	n=139~~40~~ (~~67.0~~66.9%)
Cat	n=43 (20.7~~6~~%)	Thrombolytics	n=52 (~~24.9~~25.0%)	1-3 days	n=26 (12.5~~4~~%)
Dog	n=16 (7.7%)	Cell therapies	n=7~~8~~ (~~3.8~~3.4%)	<1 week	n=15 (7.2%)
Non-Human-Primate	n=32 (15.4~~3~~%)	Diagnostics	n=15 (7.2%)	<1month	n=14 (6.7%)
Pig	n=19~~20~~ (9.1~~6~~%)	Others[#]	n=21 (10.1~~0~~%)	>1 month	n=14 (6.7%)
Non-Human-Primate & Rabbit	n=1 (0.5%)				
Sheep	n=1 (0.5%)				
Region		**Primary endpoint**		**Stroke model**	
North America	n=134~~5~~ (64.4~~6~~%)	Efficacy	n=162 (77.9~~5~~%)	Transient	n=120 (57.7~~4~~%)
Europe	n=24 (11.5%)	Safety	n=12 (5.8~~7~~%)	Permanent	n=76~~7~~ (36.5~~8~~%)
Asia/Oceania	n=50 (24.1~~3.9~~%)	Feasibility	n=22~~3~~	Transient +Permanent	n=1 (0.5%)

			(~~11.0~~10.5%)		
		Safety + Feasibility	n=1 (0.5%)	Not reported	n=11 (5.3%)
		Safety + Efficacy	n=11 (5.3%)		
Further information					
Additional veterinary care reported	n=11 (5.3%)				
Dose-response relationship reported	n=30 (14.4%)				
Compliance with animal welfare regulations reported	n=128 (61.5~~2~~%)				
Pre-study quarantine reported	n=3 (1.4%)				
Animal housing conditions[##] reported	n=23 (11.1~~0~~%)				

#these included hypothermia (n=7), hemodilution (n=5), facial nerve stimulation (n=2), hyperglycemia, retrograde transvenous perfusion, crosslinked hemoglobin transfusion, alkalinization of systemic pH, omental transposition, induced hypertension, RIPC (short term remote ischemic postconditioning) (n=1 each)

##e.g., feeding, light/dark circle, single or grouped housing

Table 3. Median experimental group sizes across large animal species.

Non-human primate			Rabbit			Dog			Cat			Sheep			Pig		
C	P	T	C	P	T	C	P	T	C	P	T	C	P	T	C	P	T
7.4	6.3	6.6	12.4	10.0	11.0	7.1	9.0	8.3	8.7	8.6	8.6	6	4.25	4.2	5.8	6.4	6.2
(1-24)	(2-17)	(1-24)	(2-50)	(2-57)	(2-57)	(5-10)	(1-16)	(1-16)	(2-17)	(3-18)	(2-18)	(6)	(3-6)	(3-6)	(2-11)	(1-10)	(1-11)
n=35	n=64	n=99	n=108	n=267	n=375	n=15	n=25	n=40	n=45	n=77	n=122	n=1	n=4	n=5	n=16	n=45	n=60

C: control group; P: procedure group(s); T: total (combined) groups. Ranges (min.-max.) are given in brackets. n describes numbers of groups throughout the included literature.

689

Quality and validity of large animal experiments in stroke: a systematic review

Leona Kringe, DVM[1,2]; Emily S. Sena, PhD[3]; Edith Motschall[4]; Zsanett Bahor, PhD[3]; Qianying Wang, MSc[3]; Andrea M. Herrmann, DVM[1,2]; Christoph Mülling, DVM, PhD[2]; Stephan Meckel, MD[1]; Johannes Boltze, MD, PhD[5]

[1]Department of Neuroradiology, Neurocenter,

Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

[2]Faculty of Veterinary Medicine, Institute of Anatomy, Histology and Embryology, Leipzig University, Leipzig, Germany

[3]Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom

[4]Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

[5]School of Life Sciences, University of Warwick, Coventry, United Kingdom

Please address correspondence to:

Johannes Boltze, MD, PhD

School of Life Sciences, University of Warwick

Coventry CV4 7AL United Kingdom,

E-mail: johannes.Boltze@warwick.ac.uk

Running headline: Role and analysis of methodological quality in large animal experiments in stroke

Abstract

An important factor for successful translational stroke research is study quality. Low-quality studies are at risk of biased results and effect overestimation, as has been intensely discussed for small animal stroke research. However, little is known about the methodological rigor and quality in large animal stroke models, which are becoming more frequently used in the field.

Based on research in two databases, this systematic review surveys and analyses the methodological quality in large animal stroke research. Quality analysis was based on the Stroke Therapy Academic Industry Roundtable (STAIR) and the Animals in Research: Reporting In Vivo Experiments (ARRIVE) guidelines. Our analysis revealed that large animal models are utilized with similar shortcomings as small animal models. Moreover, translational benefits of large animal models may be limited due to lacking implementation of important quality criteria such as randomization, allocation concealment, and blinded assessment of outcome. On the other hand, an increase of study quality over time and a positive correlation between study quality and journal impact factor were identified.

Based on the obtained findings, we derive recommendations for optimal study planning, conducting and data analysis/reporting when using large animal stroke models to fully benefit from the translational advantages offered by these models.

Key words: large animal, stroke, preclinical research, study quality, study validity

1. Introduction

Acute ischemic stroke management and care have profoundly improved with the introduction of intravenous thrombolysis and, recently, mechanical thrombectomy for large vessel occlusions.[1] However, by far not all patients can benefit from the therapeutic progress due to numerous contraindications, restricted availability and therapeutic time windows of these therapeutic approaches. This causes a tremendous need for novel treatment options, but the translation of preclinical findings into clinically applicable and efficient therapies has so far been mostly ineffective and prone to failure.[2]

Critical assessment of rodent studies revealed that one important reason for the translational failure is the lack of methodological quality in these preclinical studies, causing a higher risk for poor internal validity, overestimation of effect sizes, and biased conclusions thus affecting rationale and design of subsequent clinical trials.[3,4,5]

Large animal models become more frequently used in preclinical stroke research since they are believed to provide a number of significant advantages in the translational process.[6,7] On the other hand, large animal stroke models are both more laborious and more expensive to utilize than rodent models. Budgetary limitations often restrict sample sizes in large animal experiments, which limits statistical power.[8] Hence, it is essential to conduct large animal experiments with highest methodological rigor and to predefine precise endpoints that can be assessed with sufficient statistical power to take full advantage of the translational value of large animal stroke models.

Little is known about the methodological rigor and quality of large animal stroke experiments. We performed a systematic review and quality assessment of studies using large animal stroke models. Our quality analysis was based on the Stroke Therapy Academic Industry Roundtable (STAIR)[9,10] and Animals in Research: Reporting In Vivo Experiments (ARRIVE) guidelines.[11] Based on the obtained results, we also provide suggestions for methodological improvements in large animal stroke research.

2. Material & Methods

2.1. Study selection

Literature research was performed by the first author (L.K.). L.K. was supported by E.M., a professional librarian with extensive experience in systematic literature research who helped with designing the search strategy. The two last authors (S.M. and J.B.) were consulted by L.K. in case of any doubts or questions when extracting information from the literature. Intra-assessor reproducibility was not assessed.

2.1.1. Search strategy

We conducted a systematic search for preclinical large animal experiments in stroke using the Medline via Ovid from Wolters Kluwer and Science Citation Index Expanded via Web of Science from Clarivate Analytics data bases.

The initial search was conducted on September 26th, 2017, and an update was performed on August 9th, 2019. Data base entries between January 1st, 1990 and August 8th, 2019 were covered.

Search terms were "large animal" (including any relevant species, e.g. dogs, cats, pigs, rabbits, non-human-primates, sheep, goats, etc.) and "ischemic stroke" (involving for instance "brain ischemia" OR "ischemic neuronal injury" OR "thrombembolic stroke" OR "cerebrovascular disorders"). In the search strategies we combined the aspects *large aninals* and *ischemic stroke* with AND. Within each aspect we generally combined keywords, their synonyms and – for indexed citations of MEDLINE – controlled for vocabulary terms (Medical Subject Headings) using the operator OR. Detailed search strategies are provided in Supplementary Tables 1 and 2. The search process was conducted and results were recorded according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Figure 1A).

2.1.2. Inclusion and exclusion criteria

We included preclinical large animal studies conducted and published between 1990 and 2019 that report investigations of therapeutic and/or diagnostic procedures for ischemic stroke. The studies needed to compare at least two groups, i.e. one in which a new procedure (therapeutic or diagnostic) is tested by comparing it to a second group being subjected to a standard or reference procedure ("control group"). Only studies in English were included.

We excluded studies focusing on diseases other than ischemic stroke, using small animal (e.g., rodent) models, clinical trials, in vitro studies, reviews, and meta-analyses. Purely descriptive studies only reporting a method or procedure, or non-controlled experiments (e.g., cases series) were also excluded.

2.2. Data extraction

2.2.1. Basic study characteristics and impact factor

First, study meta-data were extracted. Those included information on species, type of intervention, year of publication and region of origin (North America, Europe, Asia & Oceania), aim of evaluation (e.g., safety, feasibility), the stroke model used, study duration and information on investigation of dose-response-relationship (if applicable), compliance with animal welfare regulations, subject health condition prior to enrolment, animal housing conditions, and additional veterinary care.

Second, we documented the impact factor (IF) of the journal in which the study results were published, measured in the year of publication. IFs were identified via the annual Thomson Reuters Journal Impact Factor report. Where the IF could not be retrieved for the required year, we contacted the respective journal and asked to provide the IF for the particular year(s).

2.2.2. Group sizes

We further extracted the number of subjects in experimental groups for each species. Group sizes were obtained for control and the diagnostic or therapeutic procedure group(s).

2.3. Analysis

2.3.1. Assessment of Reporting Quality

We designated a scale that was applicable to both, diagnostic and therapeutic procedures, to assess study quality (Table 1). The quality score includes central STAIR and ARRIVE criteria, supplemented by additional quality items. The score comprised four categories, containing 6 items each. Category 1 addresses reporting of study subject details and welfare, category 2 covered the reporting of details on study design, category 3 addressed internal study validity, and category 4 assessed quality of outcome analysis and reporting. Each study was assigned a score from 0 (lowest quality) to 24 (highest quality), with each category having a quality value of 0 (lowest quality) to 6 (highest quality).

[Table 1 about here]

2.3.2. Additional aspects influencing study quality

We further investigated whether study quality improved after the implementation of the STAIR guidelines in 1999, and their update in 2009.[9,10] We also analyzed differences in quality with respect to species, region of study origin, and type of investigation (i.e., assessment of neuroprotectives, thrombolytics, cell therapies, diagnostics, and others). Furthermore, we evaluated possible associations between the quality score and IF.

2.3.3. Group sizes

Where a study reported more than one procedure group, they were all counted individually (maximum number was n=10). Average group sizes were calculated for control and procedure groups(s) for each species. We compared total group size (control plus procedure groups) across species as well as control and procedure groups separately.

2.4 Statistics

All statistical analyses were performed using GraphPad PRISM 5 Software. Statistical significance was determined as p<0.05. Statistical significance was indicated with a single asterisk (*) at p<0.05, or a double asterisk (**) at p<0.01, respectively. Median as well as IQR (interquartile range including 25% and 75% quartiles) were documented. Comparisons between two groups were performed using the Wilcoxon signed rank test for non-parametric data to conservatively account for relatively small sample sizes. In case more than two groups were compared, the Kruskal-Wallis test was used, followed by Dunn`s correction for multiple comparisons. Spearman's correlation analysis was performed to evaluate associations between quality score and IF. Group sizes were analyzed by ANOVA on ranks (no normal distribution of data) followed by Dunn's multiple comparison test.

3. Results

3.1. Data set and year of publication

Initial and update searches identified a total of 10282 manuscripts being reduced to 8093 after elimination of duplicates. (Figure 1A; a list of all studies included can be found in the supplementary material). A total of 208 studies were included in final analysis after screening abstracts and full text according to preset inclusion and exclusion criteria (Figure 1A). Results of basic study characteristics are shown in Table 2.

Analysis of publication output per year revealed that the number of large animal experiments published from 1990 to 2014 generally decreased from n=56 in 1990-1994 to n=21 in 2010-2014 (Figure 1B). However, there was a steep increase in published studies from 2015, reaching an all-time high (n=40) even though studies published in late 2019 are not yet included in our search strategy. This might be related to the milestone evidence for clinical benefit publication of mechanical thrombectomy in large vessel occlusion stroke by the publication of five randomized controlled trials in 2015 that may have sparked new interest in the field and an increased demand for large animal models to investigate related procedures.[12,13]

[Figure 1 about here]

[Table 2 about here]

3.2. Study Quality

The overall median quality score was 11 (range 3 to 22; IQR: 4 (9-13)) out of 24. The median quality score in the first category (reporting of study subject details and welfare) was 2 out of 6 (range 1 to 5; IQR: 1 (1-2)). The second category (study planning quality) also reached a median quality score of 2 (range 1 to 6; IQR: 1 (2-3)). The third category (study conductance quality) had a median score of 3 (range 0 to 6; IQR: 2 (2-4)). Category 4 (result reporting and analysis quality) had a median quality score of 4 (range 0 to 6; IQR: 1 (2-4)). A significantly lower number of quality criteria were fulfilled in category 1 in comparison to the others ($p<0.05$).

3.2.1. Study subject details and welfare (category 1)

All studies reported the species used, but only 146 studies (70.2%) reported that the study was approved by responsible animal welfare authorities. Sex and age were reported by 31 studies (15.0%). Sex only was reported by 153 (73.6%), while age was not reported solely. The pre-study health status was reported by only 12 studies (5.8%). Medication details including the use of companion medication (e.g., analgetics, antibiotics) was reported in only 20 studies (9.6%). Comorbidities were not reported by any study.

3.2.2. Study planning (category 2)

Working hypotheses were reported in 207 (99.5%) studies. However, primary study endpoints were nominally determined in only 10 studies (4.8%). 135 (64.6%) studies reported that the study rationale was based on earlier small animal (n=79; 38.0%) or in vitro studies (n=25; 12.1%), or both (n=16; 7.7%). Effect size estimation and a priori sample size calculation can be performed based on such data. However, only 27 studies (13.0%) actually reported an estimation of effect size and a priori sample size calculation. A specific primary working hypothesis explicitly referring to previous in vitro and/or in vivo studies was reported in 18 studies (8.7%). Inclusion and exclusion criteria were reported in 104 studies (50.0%), but only 2 studies (1.0%) determined these criteria a priori.

3.2.3. Study conductance (category 3)

Randomization was reported in 116 studies (55.8%), and allocation concealment was reported in 59 cases (28.4%). 104 studies (50.0%) reported blinded outcome assessment. Measurement of physiological parameters was reported in 165 cases (79.3%). The most frequently monitored parameters included mean arterial pressure (systemic), temperature, blood gases, blood pH, and exhalation gases. 186 studies reported appropriate outcome analysis modalities (89.4%; information on inappropriate analysis modalities are provided in Supplementary Table 3). These included survival rate (n=2; 1.0 %), functional outcome (n=67; 32.2%), infarct size (n=46; 22.1%, as determined by appropriate methods such as imaging or histology), other imaging (n=90; 43.3%) or histology (n=61; 29.3%) endpoints, clinical chemistry (n=52; 25.0%), general pathology (n=24; 11.5%) or both (n=18; 8.7%). Only a fraction of studies that recorded physiological parameters finally analyzed those (n=52; 25.0%). 100 studies (48.1%) reported verification of infarct induction during intervention.

3.2.4. Result reporting and analysis (category 4)

168 studies (80.8%) adequately reported relevant data and findings in form of detailed tables or graphs. However, data were almost exclusively reported as means or medians. Individual data points were only provided by 16 studies (7.7%). Drop outs and excluded subjects were reported in 105 studies (50.5%). Application of appropriate statistical tests was reported in 192 studies (92.3%). 16 studies incompletely reported statistical analysis and, for example lacking information regarding statistical tests applied including post hoc tests. 91 studies (43.8%) described potential sources of error and bias in the experiment, while 115 (55.3%) reported limitations such as small sample size or that it was impossible to perform randomization. A conclusion fully justified by study findings was given in by most, but not all reports (n=190; 91.3%).

3.3. Additional influences on study quality

3.3.1. Study quality versus origin, species and type of intervention

Total median quality score was highest in studies from North America (Median: 12; IQR: 10-14), statistically different from studies conducted in Asia & Oceania (Median: 10; IQR: (8.75-12) or Europe (Median: 10; IQR: 8-11.75; p=0.0011 Figure 2A). Analysis of individual quality categories revealed no differences in category 1 (Figure 2B) but North American studies had statistically significantly higher scores in quality categories 2 (Median: 2.5; IQR: 2-3) and 3 (Median: 4; IQR: 3-5) than their European counterparts (Median: 2; IQR: 1-2; p<0.01; Figure 2C). Furthermore, North American studies were superior to Asian & Oceanian studies in category 3 (Median 3; IQR: 2-4; p<0.01; Figure 2D). We did not find statistically significant differences regarding category 4 (Figure 2E). Quality scores were neither influenced by species used (Figure 2F) nor by the types of intervention (Figure 2G). Overall differences in median quality score in species varied significantly without any specific intergroup difference.

[Figure 2 about here]

3.3.2. Study quality in the post-STAIR era

Methodological quality significantly improved after introduction of the STAIR guidelines in 1999 (1990-1999 pre-STAIR median: 10, IQR: 8-12; post-STAIR median: 12, IQR: 9-15; p<0.01; Figure 2H). We also compared quality scores of studies published prior to the first STAIR guidelines to quality scores of studies published in the time between the first STAIR guideline publication and the 2009 update (2000-2009; median: 11; IQR: 9-13), and to scores of studies published after the STAIR 2009 update (2010-2019; median: 13.; IQR: 10-15). Quality scores of studies published after the STAIR 2009 update were higher than those of studies published before the initial STAIR guideline publication (1990-1999; p<0.01). They were also higher than quality scores of studies published after the first publication of STAIR guidelines and prior to the 2009 update (2000-2009; p<0.05; Figure 2I).

Improvements were particularly evident in categories 1 and 4. In category 1, quality scores were lower in pre-STAIR studies (1990-1999; median: 1, IQR: 1-2) as compared to studies published after the first publication of STAIR guidelines and prior to the 2009 update (2000-2009; median: 2; IQR: 1.25-2) and to studies published after the 2009 update (2010-2019; median: 2; IQR: 2-3; p<0.01). There was also a significant difference in category 1 quality scores of studies published after the 2009 update to studies published between 2000 and 2009 (p<0.01). In category 4, quality scores of studies published after the 2009 STAIR update (2010-2019; median: 4; IQR: 3-5) were higher than those of studies published before the STAIR guidelines introduction (1990-1999; median: 3; IQR: 2-4) and those of studies published between 2000 and 2009 (median 3; IQR: 2-4; p<0.01 each).

3.3.3. Study quality versus impact factor

The IF was available for 172 studies (82.7%). We could not retrieve the IF for the remaining studies or no IF yet assigned on the particular journal in the year of publication (n=36; 17.3%). These latter studies were therefore excluded from the following analyses. Median IF was 3.3 (range 0.1 to 41.6; IQR: 2-4.6). Correlation analysis showed a statistically significant positive relationship between the total quality score and the IF (r=0.2802; p<0.01, alpha=0.05; Figure 3). We also correlated each quality score category with the IF and found that quality scores in all individual categories positively correlated with the IF (category 1: r=0.1851; p<0.05; category 2: r=0.1653; p<0.05; category 3: r=0.1858; p<0.05; category 4: r=0.2297; p<0.01; Supplementary Figure 1).

[Figure 3 about here]

3.3.4. Group sizes

Average group sizes across species are given in Table 3. Analysis of group sizes revealed that total (combined control and procedure) group size was largest in rabbits as compared to pigs (p<0.01), sheep and primates (p<0.05 each). Total group sizes in cats were larger than those in sheep (p<0.05; Figure 4A). Accordingly, control groups were largest in rabbits as compared to pigs (p<0.05) and primates (p<0.01; Figure 4B), while procedure groups were largest in rabbits as compared to pigs (p<0.01; Figure 4C).

[Table 3 about here]

[Figure 4 about here]

4. Discussion

Systematic bias may cause over- or underestimation of study results.[3] Quality items such as randomization, allocation concealment, and blinded assessment improve internal validity [14], but are often neglected in small animal studies.[3, 15, 16]

Large animal models are believed to offer significant benefits for translational stroke research. They have higher anatomical similarity to the human brain[17] and to the cerebrovascular system.[6, 7, 18] Another benefit is the potential to use these models in experiments closely mimicking a human clinical situation, and applying the same medical techniques and equipment for diagnostic and therapeutic interventions that would be used in human patients.[7, 19] Moreover, physiological characteristics of large animal models including heart and respiratory frequency, blood pressure as well as pharmacodynamic and pharmacokinetic profiles are similar to humans.[20, 21] However, in view of these advantages, large animal studies require much greater efforts and resources. It is therefore important that quality in large animal studies is as high as possible to efficiently utilize the advantages large animal models offer for translational research.

Overall, we found that methodological quality in large animal stroke studies was mediocre. Although quality generally improved significantly over the last decades and potentially due to the 1999 publication and 2019 update of the STAIR criteria, our analysis revealed some important shortcomings. Improvements are needed in reporting study subject details and welfare (quality score category 1). Aspects such as sex and age, pre-study health conditions, and medications should be reported routinely for optimal study transparency and reproducibility, and transferability of study results.[9] The lack of comorbid large animal models is not surprising. Comorbidities are difficult to simulate in outbred large animal models as they occur due to age, distress, malnutrition and other factors according to the human situation, and can take significant time in large animals to develop. Research on models exhibiting comorbidities may remain a domain of small animal research. Nevertheless, any spontaneously occurring comorbidities being diagnosed in large animals used for research should be reported.

Working hypotheses were reported in almost all studies (99.5%), but often without any obvious influence on study design. For instance, only 4.8% of the studies defined and reported primary endpoints, while analysis of expectable effect size and a priori sample size calculation were performed in few cases only (13.0%). This may severely limit the translational benefits of large animal models since study results may be hard to interpret based on potentially poor statistical power. Given the significant resources required to perform large animal studies, considering these aspects is essential. On the other hand, determination of effect size can be challenging when previous research data is lacking or not entirely applicable. In these cases, we recommend to perform large animal pilot studies that may help to assess basic characteristics in the respective model, such as variability of infarct size and its impact on the envisioned primary endpoint.

While half of the studies reported inclusion and exclusion criteria (50.0%), almost none (1.0%) applied them a priori. Defining inclusion and exclusion criteria during or after the study is believed to be a major source of bias, particularly when a study is conducted in non-blinded fashion. Hence, such bias can unfortunately not be excluded for most studies we analyzed.

Important quality aspects such as randomization (55.8%), allocation concealment (28.4%), and blinded assessment of outcome (50.0%) were more frequently reported in large animal studies as compared to small animal stroke experiments (randomization: 33.3%; blinded assessment of outcome: 44.4%,[16] allocation concealment: 25.9%; randomization, allocation concealment and blinded assessment of outcome: 24.1.%.[22] Nevertheless, the number of studies not reporting those is still remarkably high in particular since blinding and randomization should be minimum standard quality assurance procedures in confirmative stroke research[23] to which almost all large animal studies aim to contribute.

Imaging techniques such as magnetic resonance imaging, computed tomography, and angiography (43.5%) as well as physiological monitoring (80.4%) were utilized relatively frequently. This is a positive aspect since large animals are particularly suitable for clinical imaging techniques while thorough physiological monitoring creates meaningful information that may warrant subject in- or exclusion. However, verification of infarct induction (only reported in 48.1%) as well as infarct size should be conducted thoroughly and routinely to avoid the risk of increasing inter-subject/-study/-group variability, further reducing statistical power of an experiment. Parameters such as cerebral blood flow reduction for verification of infarct induction was documented by only 7.2% of studies. This is surprising since these parameters are relatively easy to determine in large animals, while clinical imaging techniques may be used to confirm the induced lesion directly.[21]

Large animals are suitable for long-term studies including functional endpoint assessment. However, we only found a relatively low percentage (6.7%) of studies being conducted for more than one month, the minimum follow-up period recommended by the STAIR guidelines for functional endpoints. Next to costs, this may be due to the selection of other primary endpoints such as safety or efficacy of recanalization methods which can be assessed more rapidly. However, experimenters who wish to assess behavioral endpoints should take into consideration that functional consequences of stroke in large animals can be more heterogeneous than in rodent models, and may develop over longer time spans.[24]

We recognized significant improvements in methodological quality since the publication of the first STAIR guidelines in 1999, and in particular after the STAIR guideline update in 2009. Comparable improvements were reported for small animal stroke studies from 2010 to 2013.[25] These findings indicate the positive impact of specific good research practice guidelines, which should be advanced continuously as evidenced by the recent 2019 STAIR guideline updates.[26] In contrast to previous findings in small animal studies,[27] we also identified positive association (r=0.2802; p<0.01) between study quality and publication in high-impact journals. In particular, total quality score as well as quality scores in all single categories 1-4 significantly correlated with higher IF. This is an encouraging result since all these categories include items being important to prevent bias. These items are hence indispensable for a valid and transparent exchange of information between researchers.

Group sizes were significantly larger in rabbits as compared to other species. This is not surprising as rabbits are the smallest and cheapest of all large animal species what allows for larger group sizes. Importantly, group sizes in primates are generally not different to that of other species. This does not mean that group sizes were sufficient for each research question, but shows that costs related to primate experiments did not prevent the same group sizes as seen in other large animal species despite rabbits.

Our study has a number of limitations. We applied a predefined search strategy and protocol being developed together an expert in literature meta-analyses (E.M.) and experts in stroke research (J.B., S.M.). However, search strategy and protocol were not registered (ex ante protocol). Data extraction was not done in duplicate, but senior experts were consulted in all doubtful cases. Intra-assessor reproducibility was not assessed. Moreover, we did not discriminate between studies focusing on therapeutic and diagnostic procedures. Large animal models provide a number of benefits over rodent models for diagnostic studies due to the larger brain size and in particular when clinical imaging is used.[33] However, those studies are often exploratory in nature. Since quality demands are different (and a bit lower) than in confirmative studies, those imaging-related studies would perform normally worse but still can contribute invaluably to their respective field.[34] Finally, we did not include a number of insightful imaging studies because they did not conduct a formal inter-group comparison.[35,36,37,38]

5. Conclusions and Recommendations

Although large animal models offer a offer a number of clear advantages for translational stroke, we found that they have similar shortcomings to small animal models, limiting this benefit. Therefore, we derived a number of recommendations to address these limitations but are, at the same time, relatively easy to implement.

5.1 Study planning and preparation

Large animal stroke studies are mostly confirmative studies. Therefore, study planning should be based on high quality standards applied for randomized controlled clinical trials (RCTs) when possible. Key elements of RCT planning and design such as a priori sample size calculation and endpoint definition should be conducted.[23] We encourage to involve statisticians already in early planning steps to optimize study design.[28] Study planning can also be supported by specific software tools. For instance, the National Centre for the Replacement, Refinement and Reduction of Animals in Research provides a freeware called Experimental Design Assistance (https://eda.nc3rs.org.uk), which is free to use and was built to guide researchers through their study planning.[29] Since optimal sample sizes may not be achieved for all endpoints, it is important to clearly define the most appropriate primary study endpoint, and to power the study properly. Collaboration between research teams in form of peer quality checks and validation of study design can highly increase objectivity and validity of a study.[30] Inter-group collaboration and transfer of experience can also help to handle very complex models and/or experimental setups, helping to reduce inter-subject variability negatively affecting statistical power. Confirmative studies might be preregistered to maximize transparency.[39]

5.2 Effect size estimation and pilot trials

Collecting valid information from previous research is essential for reliable effect size estimation. If such data are not available, pilot studies may be helpful for at least basically estimating variability of stroke impact and outcome in the model. In case previous experience with a particular model is low, variability is more likely to be higher and effect size is more likely to lower in such pilot trials. This will contribute to more conservative study planning since sample sizes calculated based on that information will be higher.. An important side effect of pilot trials is experimenter training which limits experimenter-caused endpoint variability (see below) in the main experiment. In addition, meta-analyses can help to collect relevant information on effect size or regarding a specific research question from related fields.[31]

5.3 Reducing the effect of sample size limitations and endpoint variability

Financial and logistical restrictions often impact sample and group sizes in large animal experiments. This is an understandable limitation which is difficult to overcome. Selection of a proper and relevant endpoint that can be adequately powered with respected to the addressed research question is therefore important to minimize the risk for low statistical power. Of note, some endpoints often used in studies assessing therapeutic interventions including infarct size and functional deficits, exhibit a higher variability in large animal models than in rodent. This makes comparison of absolute data more difficult.[24] Relative analysis of repeatedly assessed endpoints, i.e. in comparison to the individual initial infarct size and/or functional deficit can efficiently compensate for such variability. Repeated assessments also allow calculating the area under the curve for particular endpoints. This may provide a benefit in statistical power to identify whether a real outcome benefit is present over time. However, this comes at the cost of temporal resolution: it cannot be concluded exactly when this benefit became evident. There is also preliminary evidence for fast and slow stroke progressors in large animals, indicating different collateral status and somewhat resembling the human situation, but further contributing to inter-subject variability. It is recommended to consider this fact when planning an acute stroke study.[32]

In experiments of highly similar design, controls may be pooled. Of note, this counteracts randomization and therefore requires extremely thorough validation of comparability of control subjects from different experiments/sources. If comparability is thoroughly proven, this may help to increase statistical power, but the limitations of this approach and potentially resulting bias need to be discussed transparently and in detail when publishing results.

The possibility to repeatedly collect a broad spectrum of physiological data should be utilized where possible, as deviation from normal parameter ranges may explain variability and warrant post-hoc exclusion of subjects in single cases.

5.4 Study duration and documentation

We recommend considering long-term experiments whenever meaningful and possible and meeting animal welfare requirements. Even though long-term experiments involve greater efforts, the amount of data collected for individual subjects may be much higher, providing a better overall picture on the assessed intervention. Documentation should be as transparent as possible because transparency is not challenging or laborious, but contributes significantly to increased scientific rigor, reproducibility, and unbiased study result interpretation. Methodological limitations including lacking quality aspects due to good reason should be clearly stated as this allows better interpretation of positive, neutral and negative study results.

6. Acknowledgements

ES is supported by the Stroke Association (SA L-SNC 18\1003).

7. Disclosure

The authors do not report relevant disclosures.

8. Supplementary data

1. Supplementary Table 1: Search strategy in Medline

2. Supplementary Table 2: Search strategy in Web of Science

3. Supplementary Table 3: Type and frequency of inappropriate analysis methods

4. Supplementary Figure 1: Association between impact factor and quality score within individual categories

5. Supplementary reference list

Supplementary material for this paper can be found at the journal website: http://journals.sagepub.com/home/jcb

9. References

1. Bush CK, Kurimella D, Cross LJ, et al. Endovascular Treatment with Stent-Retriever Devices for Acute Ischemic Stroke: A Meta-Analysis of Randomized Controlled Trials. *PLoS One*. 2016; e0147287.
2. O'Collins VE, Macleod MR, Donnan GA, et al. 1,026 experimental treatments in acute stroke. *Ann Neurol* 2006; 59: 467-477.
3. Macleod MR, Fisher M, O`Collins V, et al. Reprint: Good laboratory practice: preventing introduction of bias at the bench. *Int J Stroke* 2004; 59: 3-5.
4. Macleod MR, Lawson Mc Lean A, Kyriakopoulou A, et al. Risk of Bias in Reports of In Vivo Research: A Focus for Improvement. *PLoS Biol* 2015; 13: e1002273.
5. Sena ES, van der Worp HB, Bath PM, et al. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biol* 2010; 8: e1000344.
6. Herrmann AM, Meckel S, Gounis MJ, et al. Large animal in neurointerventional research: A systematic review on models, techniques and their application in endovascular procedures for stroke, aneurysms and vascular malformations. *J Cereb Blood Flow Metab.* 2019; 39(3): 375-394.
7. Traystman RJ. Animal models of focal and global cerebral ischemia. *ILAR J* 2003; 44: 85-95.
8. Sena ES, Van der Worp HB, Howells D, et al. How can we improve the pre-clinical development of drugs for stroke? *Trends Neurosci* 2007; 30: 433-439.
9. STAIR. Recommendations for standards regarding preclinical neuroprotective and restorative drug development. *Stroke* 1999; 30: 2752-2758.
10. Fisher M, Feuerstein G, Howells D, et al. Update of the stroke therapy academic industry roundtable preclinical recommendations. *Stroke* 2009; 40: 2244-2250.
11. Kilkenny C, Browne WJ, Cuthill IC, et al. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol* 2010; 8: e1000412.

12. Jovin TG, Albers, GW, Liebeskind DS, et al. Stroke Treatment Academic Industry Roundtable: The Next Generation of Endovascular Trials. *Stroke* 2016; 47(19): 2656-2665.

13. Goyal M, Menon BK, van Zwam WH, et al. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomized trials. *Lancet* 2016; 387(10029): 1723-1731.

14. Sena ES, Currie GL, McCann SK, et al. Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically. *J Cereb Blood Flow Metab* 2014; 34: 737-742.

15. Sena ES, Van der Worp HB, Bath PM, et al. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biol* 2010; 8: e1000344.

16. Macleod MR, Van der Worp HB, Sena ES, et al. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke* 2008; 39: 2824-2829.

17. Boltze J, Nitzsche F, Jolkkonen J, et al. Concise Review: Increasing the Validity of Cerebrovascular Disease Models and Experimental Methods for Translational Stem Cell Research. *Stem Cells* 2017; 35: 1141-1153.

18. Sommer CJ. Ischemic stroke: experimental models and reality. *Acta Neuropathol* 2017; 133: 245-261.

19. Mehra M, Henninger N, Hirsch JA et al. Preclinical acute ischemic stroke modeling. J *Neurointerv Surg* 2012; 4: pp. 307-3013.

20. Helke KL and Swindle MM. Animal models of toxicology testing: the role of pigs. *Expert Opin Drug Metab Toxicol* 2013; 9: 127-139.

21. Herrmann AM, Cattaneo GFM, Eiden, SA, et al. Development of a Routinely Applicable Imaging Protocol for Fast and Precise Middle Cerebral Artery Occlusion

Assessment and Perfusion Deficit Measure in an Ovine Stroke Model: A Case Study. *Front Neurol.* 2019; 10: 1113.

22. Minnerup J, Zentsch V, Schmidt A, et al. Methodological Quality of Experimental Stroke Studies Published in the Stroke Journal: Time Trends and Effect of the Basic Science Checklist. *Stroke* 2016; 47: 267-272.

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

24. Boltze J, Modo MM, Mays RW, et al. Stem Cells as an Emerging Paradigm in Stroke 4: Advancing and Accelerating Preclinical Research. *Stroke* 2019; 50 (11): 3299-3306.

25. Minnerup J, Wersching H and Diederich K. Methodological quality of preclinical stroke studies is not required for publication in high-impact journals. *J Cereb Blood Flow Metab* 2010; 30: 1619-1624.

26. Savitz SI, Baron JC, Fisher M, et al. Stroke Treatment Academic Industry Roundtable X: Brain Cyoprotection Therapies in the Reperfusion Era. *Stroke* 2019. 50(4): 1026-1031.

27. Minnerup J, Wersching H and Diederich K. Methodological quality of preclinical stroke studies is not required for publication in high-impact journals. *J Cereb Blood Flow Metab* 2010; 30: 1619-1624.

28. Würbel H. More than 3Rs: the importance of scientific validity for harm-benefit analysis of animal research. *Lab Anim (NY)* 2017; 46: 164-166.

29. Percie du Sert N, Bamsey I, Bate ST, et al. The Experimental Design Assistant. *PLoS Biol* 2017; 15: e2003779.

30. Sena ES, Currie GL, McCann SK, et al. Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically. *J Cereb Blood Flow Metab* 2014; 34: 737-742.

31. Begley CG and Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res* 2015; 116: 116-126.

32. Shazeeb MS, King RM, Brooks OW, et al. Infarct Evolution in a Large Animal Model of Middle Cerebral Artery Occlusion. *Transl. Stroke Research* 2019. doi: 10.1007/s12975-019-00732-9.

33. Werner P, Saur D, Zeisig V, et al. Simultaneous PET/MRI in stroke: a case series. *J Cereb Blood Flow* 2015; 35(9): 1421-1425.

34. Dirnagl U, Hakim A, Macleod M, et al. A concerted appeal for international cooperation in preclinical stroke research. *Stroke* 2013; 44(6): 1754-1760.

35. Boltze J, Ferrara F, Hainsworth AH, et al. Lesional and perilesional tissue characterization by automated image processing in a novel gyrencephalic animal model of peracute intracerebral hemorrhage. *J Cereb Blood Flow Metab* 2019; 39(12): 2521-2535.

36. Haque ME, Gabr RE, Zhao X, et al. Serial quantitative neuroimaging of iron in the intracerebral hemorrhage pig model. *J Cereb Blood Flow Metab* 2018; 38(3): 375-381.

37. Kamimura HA, Flament J, Valette J, et al. Feedback control of microbubble cavitation for ultrasound-mediated blood-brain barrier disruption in non-human primates under magnetic resonance guidance. *J Cereb Blood Flow Metab* 2019; 39(7): 1191-1203.

38. Sander CY, Mandeville JB, Wey HY, et al. Effects of flow changes on radiotracer binding: Simultaneous measurement of neuroreceptor binding and cerebral blood flow modulation. *J Cereb Blood Flow Metab* 2019; 39(1): 131-146.

39. Kimmelmann J and Anderson JA. Should preclinical studies be registered? *Nat Biotechnol* 2012; 30(6): 488-489.

Figure legends

Figure 1. Overview on quantitative search results and frequency of large animal experiments in stroke research since 1990.

(A) Flow diagram of publication identification. N = Number of publications. Records were excluded after screening title and abstracts. Full-text articles were then screened and excluded for a priori determined reasons. (B) Timeline of publication in large animal stroke research (1990-2019): The increase of large animal stroke studies in the last years is potentially due to the breakthrough in recanalization therapies, prompting a number of follow-on translational studies utilizing large animal stroke models.

Figure 2. Influence of study origin and STAIR criteria publication on study quality.

(A) Total quality score, (B) Category 1: Reporting of study subject and animal welfare, (C) Category 2: Study planning quality (North America vs. Europe p<0.01), (D) Category 3: Study conductance quality (North America vs. Asia & Oceania p<0.01), (E) Category 4: Result reporting and analysis quality (North America vs. Europe p<0.01), (F) Influence of species, (G) Improvement in total methodological quality since the publication of the first STAIR criteria in 1999 (p<0.01), (H) Improvement in total methodological quality since the publication of the first STAIR criteria in 1999 comparing to their amendment in 2009 (2010-2019 vs. 1990-1999 p<0.01, and 2010-2019 vs. 2000-2009 p<0.05), (I) Influence of type of intervention. Horizontal lines and whiskers indicate the median with lower and upper 95% CI. *p<0.05; **p<0.01.

Figure 3. Association between total quality score versus impact factor.

Scatterplot shows correlation between quality score and IF (p<0.01). Number of included studies is 172, no IF could be retrieved for 36 studies. The latter studies were excluded from this analysis.

Figure 4. Group sizes across species.

(A) Total group sizes were largest in rabbits as compared to pigs ($p<0.01$), primates and sheep ($p<0.05$ each). (B) Control group sizes were larger in rabbits as compared to primates ($p<0.01$) and pigs ($p<0.05$). (C) Procedure group sizes were larger in rabbits as compared to pigs ($p<0.01$). Horizontal lines and whiskers indicate the median with lower and upper 95% CI. *$p<0.05$, **$p<0.01$.

Tables

Table 1. Quality score items.

Category 1: Reporting of study subject details and welfare		**Category 2: Study planning quality**	
Item	*Score point allocation*	*Item*	*Score point allocation*
1. Animal protocol approved	Reported yes=1/no=0	1. Study hypothesis	Reported yes=1/no=0
2. Species	Reported yes=1/no=0	2. A priori endpoint definition	Reported yes=1/no=0
3. Sex and Age	Reported yes=1/no=0	3. A priori sample size calculation	Reported yes=1/no=0
4. Pre-Study Health	Reported yes=1/no=0	4. Reference to previous studies	Reported yes=1/no=0
5. Comorbidities	Reported yes=1/no=0	5. Inclusion/Exclusion criteria	Reported yes/no=0
6. Adequate medication	Reported yes=1/no=0	6. Effect size/Treatment effect	Reported yes=1/no=0

Category 3: Internal study validity		**Category 4: Outcome analysis and reporting**	
Item	*Score point allocation*	*Item*	*Score point allocation*
1. Blinding	Reported yes=1/no=0	1. Individual data points	Reported yes=1/no=0
2. Randomization	Reported yes=1/no=0	2. Drop outs/Excluded subjects	Reported yes=1/no=0
3. Allocation concealment	Reported yes=1/no=0	3. Appropriate statistical tests	Used yes=1/no=0
4. Physiological parameters	Measuring reported yes=1/no=0	4. Potential error sources	Reported yes=1/no=0
5. Analysis modalities	Appropriate modalities reported# yes=1/no=0	5. Study/Methodological limits	Reported yes=1/no=0
6. Infarct induction confirmation	Reported yes=1/no=0	6. Justified conclusion given##	Provided yes=1/no=0

#analysis modalities were considered appropriate when being sufficient to assess the respective research question or endpoint (see Supplementary Table 3 for details).

##conclusion was considered justified when supported by correctly analyzed results.

Table 2. Basic Characteristics of included Animal Experimental Studies.

Item	Frequency (%)	Item	Frequency (%)	Item	Frequency (%)
Species		**Type of intervention**		**Study duration**	
Rabbit	n=96 (46.1%)	Neuroprotectives	n=113 (54.3%)	Acute phase (<24h)	n=139 (66.9%)
Cat	n=43 (20.7%)	Thrombolytics	n=52 (25.0%)	1-3 days	n=26 (12.5%)
Dog	n=16 (7.7%)	Cell therapies	n=7 (3.4%)	<1 week	n=15 (7.2%)
Non-Human-Primate	n=32 (15.4%)	Diagnostics	n=15 (7.2%)	<1month	n=14 (6.7%)
Pig	n=19 (9.1%)	Others[#]	n=21 (10.1%)	>1 month	n=14 (6.7%)
Non-Human-Primate & Rabbit	n=1 (0.5%)				
Sheep	n=1 (0.5%)				
Region		**Primary endpoint**		**Stroke model**	
North America	n=134 (64.4%)	Efficacy	n=162 (77.9%)	Transient	n=120 (57.7%)
Europe	n=24 (11.5%)	Safety	n=12 (5.8%)	Permanent	n=76 (36.5%)
Asia/Oceania	n=50 (24.1%)	Feasibility	n=22 (10.5%)	Transient +Permanent	n=1 (0.5%)
		Safety +	n=1 (0.5%)	Not reported	n=11 (5.3%)

		Feasibility	
		Safety + Efficacy	n=11 (5.3%)
Further information			
Additional veterinary care reported	n=11 (5.3%)		
Dose-response relationship reported	n=30 (14.4%)		
Compliance with animal welfare regulations reported	n=128 (61.5%)		
Pre-study quarantine reported	n=3 (1.4%)		
Animal housing conditions## reported	n=23 (11.1%)		

#these included hypothermia (n=7), hemodilution (n=5), facial nerve stimulation (n=2), hyperglycemia, retrograde transvenous perfusion, crosslinked hemoglobin transfusion, alkalinization of systemic pH, omental transposition, induced hypertension, RIPC (short term remote ischemic postconditioning) (n=1 each)

##e.g., feeding, light/dark circle, single or grouped housing

Table 3. Median experimental group sizes across large animal species.

Non-human primate			Rabbit			Dog			Cat			Sheep			Pig		
C	**P**	**T**	**C**	**P**	**T**	**C**	**P**	**T**	**C**	**P**	**T**	**C**	**P**	**T**	**C**	**P**	**T**
7.4	6.3	6.6	12.4	10.0	11.0	7.1	9.0	8.3	8.7	8.6	8.6	6	4.25	4.2	5.8	6.4	6.2
(1-24)	(2-17)	(1-24)	(2-50)	(2-57)	(2-57)	(5-10)	(1-16)	(1-16)	(2-17)	(3-18)	(2-18)	(6)	(3-6)	(3-6)	(2-11)	(1-10)	(1-11)
n=35	n=64	n=99	n=108	n=267	n=375	n=15	n=25	n=40	n=45	n=77	n=122	n=1	n=4	n=5	n=16	n=45	n=60

C: control group; P: procedure group(s); T: total (combined) groups. Ranges (min.-max.) are given in brackets. n describes numbers of groups throughout the included literature.

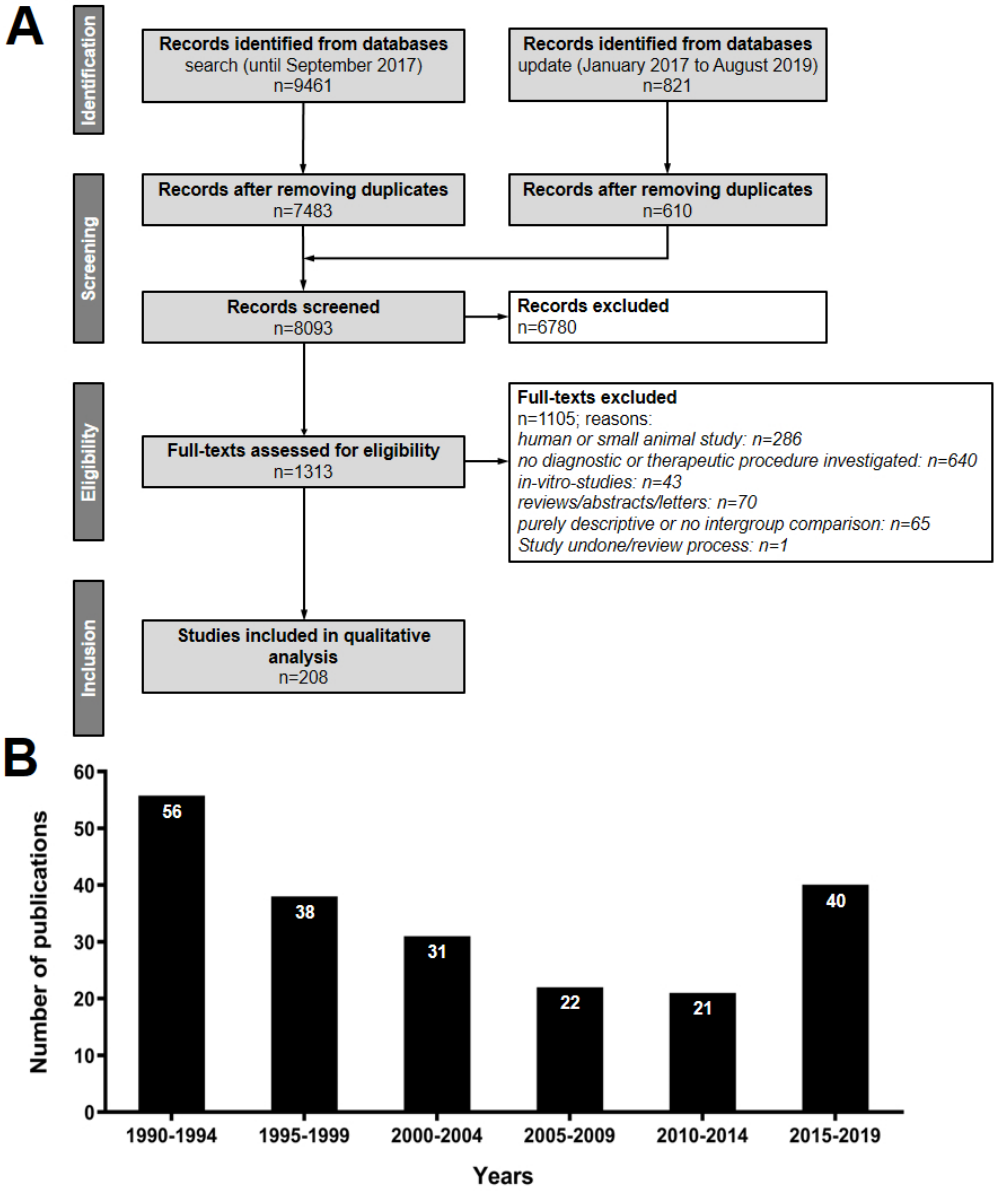

Figure 1. Overview on quantitative search results and frequency of large animal experiments in stroke research since 1990.

(A) Flow diagram of publication identification. N = Number of publications. Records were excluded after screening title and abstracts. Full-text articles were then screened and excluded for a priori determined reasons. (B) Timeline of publication in large animal stroke research (1990-2019): The increase of large animal stroke studies in the last years is potentially due to the breakthrough in recanalization therapies, prompting a number of follow-on translational studies utilizing large animal stroke models.

233x281mm (96 x 96 DPI)

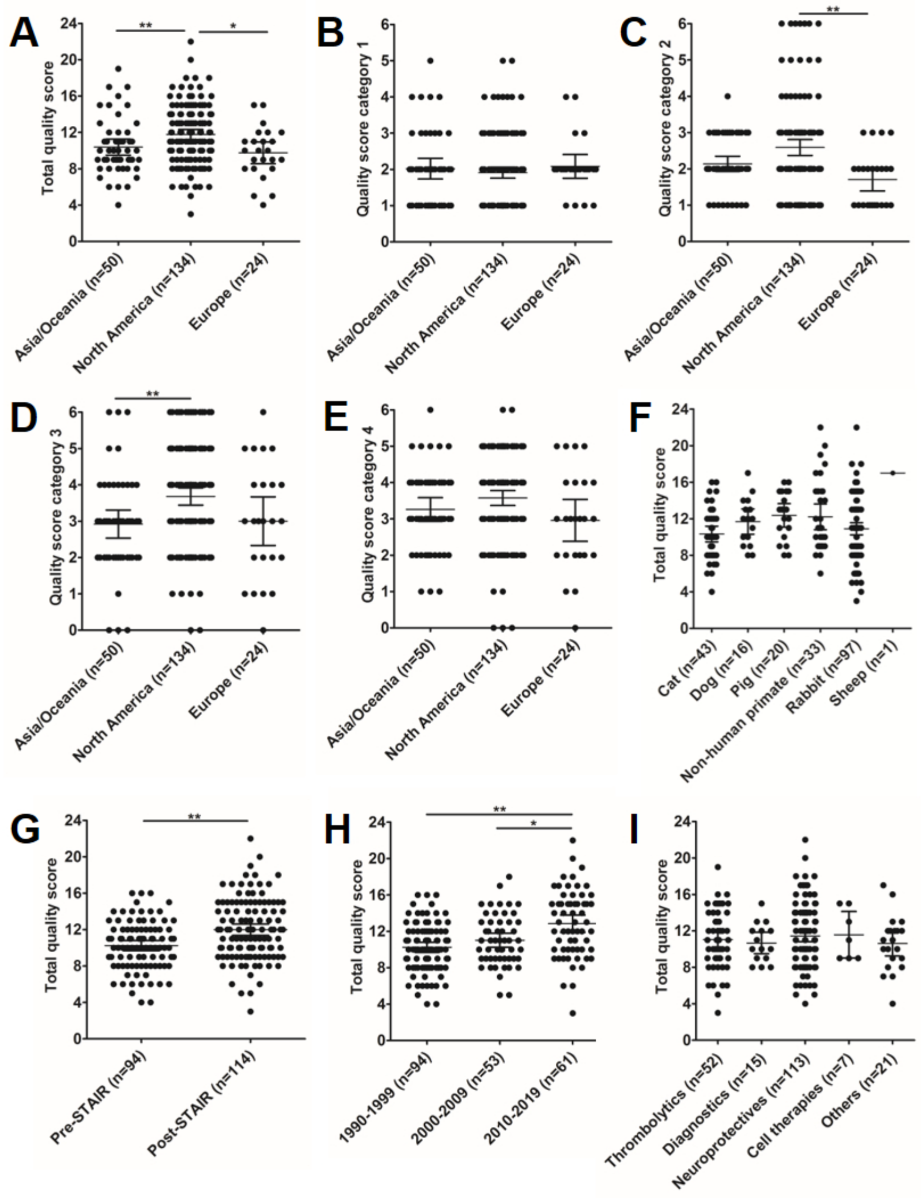

Figure 2. Influence of study origin and STAIR criteria publication on study quality. (A) Total quality score, (B) Category 1: Reporting of study subject and animal welfare, (C) Category 2: Study planning quality (North America vs. Europe p<0.01), (D) Category 3: Study conductance quality (North America vs. Asia & Oceania p<0.01), (E) Category 4: Result reporting and analysis quality (North America vs. Europe p<0.01), (F) Influence of species, (G) Improvement in total methodological quality since the publication of the first STAIR criteria in 1999 (p<0.01), (H) Improvement in total methodological quality since the publication of the first STAIR criteria in 1999 comparing to their amendment in 2009 (2010-2019 vs. 1990-1999 p<0.01, and 2010-2019 vs. 2000-2009 p<0.05), (I) Influence of type of intervention. Horizontal lines and whiskers indicate the median with lower and upper 95% CI. *p<0.05; **p<0.01.

216x278mm (96 x 96 DPI)

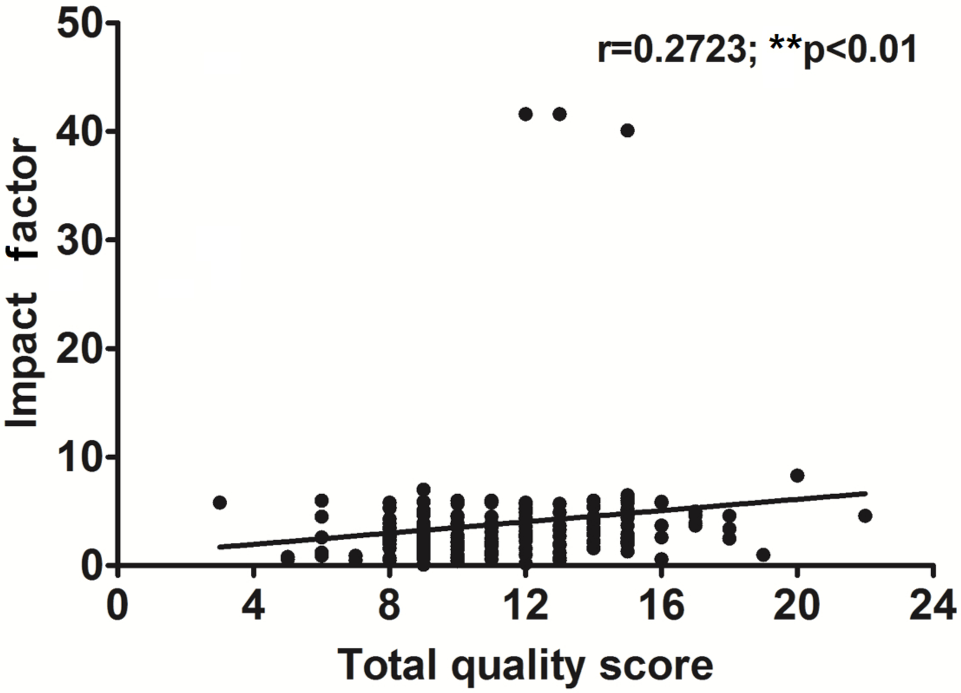

Figure 3. Association between total quality score versus impact factor.
Scatterplot shows correlation between quality score and IF (p<0.01). Number of included studies is 172, no IF could be retrieved for 36 studies. The latter studies were excluded from this analysis.

304x222mm (96 x 96 DPI)

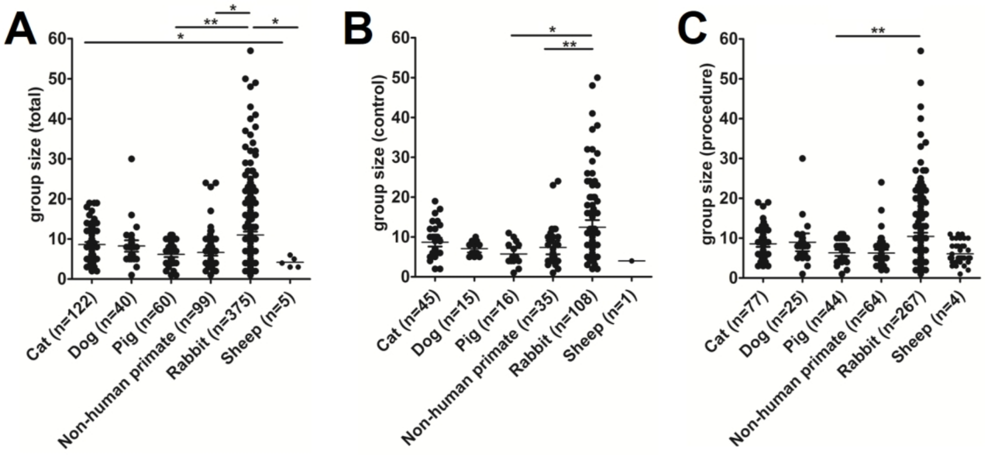

Figure 4. Group sizes across species.

(A) Total group sizes were largest in rabbits as compared to pigs (p<0.01), primates and sheep (p<0.05 each). (B) Control group sizes were larger in rabbits as compared to primates (p<0.01) and pigs (p<0.05). (C) Procedure group sizes were larger in rabbits as compared to pigs (p<0.01). Horizontal lines and whiskers indicate the median with lower and upper 95% CI. *p<0.05, **p<0.01.

377x175mm (120 x 120 DPI)