

**Manuscript version: Author's Accepted Manuscript**

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

**Persistent WRAP URL:**

<http://wrap.warwick.ac.uk/136529>

**How to cite:**

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

**Copyright and reuse:**

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

**Publisher's statement:**

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk).

**Journal of  
Cerebral Blood Flow  
& Metabolism**

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

Journal:	<i>Journal of Cerebral Blood Flow and Metabolism</i>
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; Universitätsklinikum 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

1  
2  
3 **1 Quality and validity of large animal experiments in stroke: a systematic review**

4  
5 2  
6  
7  
8 3 Leona Kringe, DVM<sup>1,2</sup>; Emily S. Sena, PhD<sup>3</sup>; Edith Motschall<sup>4</sup>; Zsanett Bahor, PhD<sup>3</sup>;

9  
10 4 [Qianying Wang, MSc<sup>3</sup>](#); Andrea M. Herrmann, DVM<sup>1,2</sup>; Christoph Mülling, DVM,  
11  
12 5 PhD<sup>2</sup>; Stephan Meckel, MD<sup>1</sup>; Johannes Boltze, MD, PhD<sup>5</sup>

13  
14 6  
15  
16  
17 7 <sup>1</sup>Department of Neuroradiology, Neurocenter,

18  
19 8 [Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany](#)

20  
21 9 [University Hospital Freiburg, Freiburg, Germany](#)

22  
23  
24 10 <sup>2</sup>Faculty of Veterinary Medicine, Institute of Anatomy, Histology and Embryology,  
25  
26 11 Leipzig University, Leipzig, Germany

27  
28 12 <sup>3</sup>Centre for Clinical Brain Sciences, ~~The~~ University of Edinburgh, Edinburgh, United  
29  
30 13 Kingdom

31  
32  
33 14 <sup>4</sup>Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center,  
34  
35 15 University of Freiburg, Freiburg, Germany

36  
37 16 <sup>5</sup>School of Life Sciences, University of Warwick, Coventry, United Kingdom

38  
39  
40 17  
41  
42 18 Please address correspondence to:

43  
44 19 Johannes Boltze, MD, PhD

45  
46 20 School of Life Sciences, University of Warwick

47  
48 21 Coventry CV4 7AL United Kingdom,

49  
50 22 E-mail: [johannes.Boltze@warwick.ac.uk](mailto:johannes.Boltze@warwick.ac.uk)

51  
52  
53 23  
54  
55  
56 24  
57  
58 25 Running headline: Role and analysis of methodological quality in large animal  
59  
60 26 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 ~~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

53

## 54 1. Introduction

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

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

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

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

79 improvements in large animal stroke research.

## 80 **2. Material & Methods**

### 81 **2.1. Study selection**

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

87

#### 88 **2.1.1. Search strategy**

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

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

95 Search terms were “large animal” (including any relevant species, e.g. dogs, cats, pigs,  
96 rabbits, non-human-primates, sheep, goats, etc.) and “ischemic stroke” (involving for instance  
97 “brain ischemia” OR “ischemic neuronal injury” OR “thrombembolic stroke” OR  
98 “cerebrovascular disorders”). In the search strategies we combined the aspects *large animals*

99 and *ischemic stroke* with AND. Within each aspect wWe generally combined keywords, their  
100 synonyms and – for indexed citations of MEDLINE – controlled for vocabulary terms (Medical  
101 Subject Headings) using the operator OR. Detailed search strategies are provided in  
102 Supplementary Tables 1 and 2. The search process was conducted and results were recorded  
103 according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses  
104 (PRISMA) guidelines (Figure 1A).

105

### 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~~ 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).

131

### 132 **2.2.2. Group sizes**

133 We further extracted the number of subjects in experimental groups for each species.

134 Group sizes were obtained for control and the diagnostic or therapeutic procedure group(s).

135

136

## 137 **2.3. Analysis**

### 138 **2.3.1. Assessment of Reporting Quality**

139 We designated a scale that was applicable to both, diagnostic and therapeutic procedures,

140 to assess study quality (Table 1).~~A score was designed to assess study quality (Table 1).~~ The

141 quality score includes central STAIR and ARRIVE criteria, supplemented by additional quality

142 items. The score comprised four categories, containing 6 items each. Category 1 addresses

143 reporting of study subject details and welfare, category 2 covered the reporting of details on

144 study design, category 3 addressed internal study validity, and category 4 assessed quality of

145 outcome analysis and reporting. Each study was assigned a score from 0 (lowest quality) to 24

146 (highest quality), with each category having a quality value of 0 (lowest quality) to 6 (highest

147 quality).

148

149 [Table 1 about here]

150

### 151 **2.3.2. Additional aspects influencing study quality**

152 We further investigated whether study quality improved after the implementation of the

153 STAIR guidelines in 1999, and their update in 2009.<sup>9,10</sup> We also analyzed differences in quality

154 with respect to species, region of study origin, and type of investigation (i.e., assessment of

155 neuroprotectives, thrombolytics, cell therapies, diagnostics, and others). Furthermore, we

156 evaluated possible associations between the quality score and impact factor.IF.



157

### 158 **2.3.3. Group sizes**

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

163

## 164 **2.4 Statistics**

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

175

## 176 **3. Results**

### 177 **3.1. Data set and year of publication**

178 Initial and update searches identified a total of 10282 manuscripts being reduced to 8093  
179 after elimination of duplicates. (Figure 1A; a list of all studies included can be found in the  
180 supplementary material). A total of 2089 studies were included in final analysis after screening  
181 abstracts and full text according to preset inclusion and exclusion criteria (Figure 1A). Results  
182 of basic study characteristics are shown in Table 2.

1  
2  
3 183 Analysis of publication output per year revealed that the number of large animal  
4  
5 184 experiments published from 1990 to 2014 generally decreased from n=56 in 1990-1994 to n=21  
6  
7 185 in 2010-2014 (Figure 1B). However, there was a steep increase in published studies from 2015,  
8  
9 186 reaching an all-time high (n=40) even though studies published in late 2019 are not yet included  
10  
11 187 in our search strategy. This might be related to the milestone evidence for clinical benefit  
12  
13 188 publication of mechanical thrombectomy in large vessel occlusion stroke by the publication of  
14  
15 189 five randomized controlled trials in 2015 that may have sparked new interest in the field and an  
16  
17 190 increased demand for large animal models to investigate related procedures.<sup>12,13</sup>~~Analysis of~~  
18  
19 191 ~~publication output per year revealed that the number of large animal experiments published~~  
20  
21 192 ~~from 1990 to 2019 generally decreased after reaching a peak in the mid 1990s (Figure 1B).~~  
22  
23 193 ~~There were more publications in the 1990s (n=93) than in the last decade (n=62). However,~~  
24  
25 194 ~~publication output remarkably increased since 2014 with 47 studies published between 2014~~  
26  
27 195 ~~and 2019.~~  
28  
29  
30  
31  
32

196

197 [Figure 1 about here]

198 [Table 2 about here]

199

### 200 3.2. Study Quality

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

209

### 3.2.1. Study subject details and welfare (category 1)

All studies reported the species used, but only 146 studies (70.269.9%) reported that the study was approved by responsible animal welfare authorities. Sex and age were reported by 31 studies (15.04.8%). Sex only was reported by 153 (73.62%), while age was not reported solely. The pre-study health status was reported by only 12 studies (5.87%). 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.

217

### 3.2.2. Study planning (category 2)

Working hypotheses were reported in 2078 (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; 38.07.8%) or in vitro studies (n=256; 12.14%), 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 (13.02.9%) 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 (8.76%). Inclusion and exclusion criteria were reported in 104 studies (49.850.0%), but only 2 studies (1.0%) determined these criteria a priori.

228

### 3.2.3. Study conductance (category 3)

Randomization was reported in 116 studies (55.85%), and allocation concealment was reported in 59 cases (28.42%). 104 studies (4950.0.8%) reported blinded outcome assessment. Measurement of physiological parameters was reported in 1656 cases (79.34%). The most frequently monitored parameters included mean arterial pressure (systemic), temperature, blood gases, blood pH, and exhalation gases. 1867 studies reported appropriate outcome

analysis modalities (89.45%; 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 (48.17.8%) reported verification of infarct induction during intervention.

#### 3.2.4. Result reporting and analysis (category 4)

1689 studies (80.89%) 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 167 studies (7.78.1%). Drop outs and excluded subjects were reported in 105 studies (50.52%). Application of appropriate statistical tests was reported in 1923 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 (43.85%) described potential sources of error and bias in the experiment, while 115 (55.30%) 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=1901; 91.34%).

### 3.3. Additional influences on study quality

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

261 Total median quality score was highest in studies from North America (Median: ~~12~~; IQR: ~~5.75 (8.25-14)~~), ~~but not~~ statistically different from studies conducted in Asia & Oceania (Median: 10; IQR: ~~3.25 (8.75-12)~~) ~~as well as those from~~ 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

279 ~~In general,~~ methodological quality significantly improved after ~~publication-introduction~~ of the ~~first~~ STAIR guidelines in 1999 (~~1990-1999 p~~Pre-STAIR ~~m~~Median: 10, IQR: ~~4 (8-12)~~; ~~p~~Post-STAIR ~~m~~Median: 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~~Median: 11; IQR: ~~4 (9-13)~~), and to scores of studies published after the STAIR ~~preclinical-guideline~~2009 update (2010-2019; mMedian: 13.50; IQR: 5 (10-15)). Quality scores of studies published after the STAIR 2009

1  
2  
3 287 update were higher than those of studies published before the initial STAIR guideline  
4  
5 288 publication (1990-1999; p<0.01). They were also higher than quality scores of studies published  
6  
7 289 after the first publication of STAIR guidelines and prior to the 2009 update (2000-2009; p<0.05;  
8  
9  
10 290 Figure 2I). We found significantly higher quality in studies conducted between 2010-2019  
11  
12 291 compared to those performed prior STAIR guideline publication (1990-1999; p<0.01) and those  
13  
14 292 performed after the first publication of STAIR guidelines (2000-2009; p<0.05; Figure 2I).

15  
16  
17 293 Improvements were particularly evident in categories 1 and 4. In category 1, quality  
18  
19 294 scores were lower in pre-STAIR studies (1990-1999; median: 1, IQR: 1-2) as compared to  
20  
21 295 studies published after the first publication of STAIR guidelines and prior to the 2009 update  
22  
23 296 (2000-2009; median: 2; IQR: 1.25-2) and to studies published after the 2009 update (2010-  
24  
25 297 2019; median: 2; IQR: 2-3; p<0.01). There was also a significant difference in category 1  
26  
27 298 quality scores of studies published after the 2009 update to studies published between 2000 and  
28  
29 299 2009 (p<0.01). In category 4, quality scores of studies published after the 2009 STAIR update  
30  
31 300 (2010-2019; median: 4; IQR: 3-5) were higher than those of studies published before the STAIR  
32  
33 301 guidelines introduction (1990-1999; median: 3; IQR: 2-4) and those of studies published  
34  
35 302 between 2000 and 2009 (median 3; IQR: 2-4; p<0.01 each).  
36  
37 303 (1990-1999 (Median: 1, IQR: 1 (1-2)) vs. 2000-2009 (Median: 2; IQR: 0.75 (1.25-2)) and 1990-  
38  
39 304 1999 vs. 2010-2019 (Median: 2; IQR: 1 (2-3)) and 2000-2009 vs. 2010-2019 (p<0.01)) and 4  
40  
41 305 (1990-1999 (Median: 3; IQR: 2 (2-4)) vs. 2010-2019 (4; IQR: 2 (3-5)) and 2000-2009 (Median  
42  
43 306 3; IQR: 2 (2-4)) vs. 2010-2019) (p<0.01).

44  
45  
46  
47  
48  
49 307

### 308 3.3.3. Study quality versus impact factor

50  
51  
52 309 The IF was ~~documented~~ available for 1723 studies (82.78%). We could not retrieve the  
53  
54 310 IF for the remaining studies or no IF yet assigned on the particular journal in the year of  
55  
56 311 publication (n=36; 17.32%). These latter studies were therefore excluded from the following  
57  
58 312 analyses. Median IF was 3.3 (range 0.1 to 41.6; IQR: ~~2.65~~ (2-4.65)). Correlation analysis  
59  
60

1  
2  
3 313 showed a statistically significant positive relationship between the total quality score and the  
4  
5 314 ~~journal impact factor~~IF (r=0.27232802; p<0.01, alpha=0.05; Figure 3). We also correlated each  
6  
7  
8 315 quality score category with the IF and found that quality scores in all individual categories  
9  
10 316 positively correlated with the IF (category 1: r=0.19181851; p<0.05; category 2: r=0.165317;  
11  
12 317 p<0.05; category 3: r=0.1858769; p<0.05; category 4: r=0.2297485; p<0.01; Supplementary  
13  
14 318 Figure 1).

15  
16  
17 319  
18  
19 320 [Figure 3 about here]

### 21 321 **3.3.4. Group sizes**

22  
23  
24 322 Average group sizes across species are given in Table 3. Analysis of group sizes  
25  
26 323 revealed that total (combined control and procedure) group size was largest in rabbits as  
27  
28 324 compared to pigs (p<0.01), sheep and primates (p<0.05 each). Total group sizes in cats were  
29  
30 325 larger than those in sheep (p<0.05; Figure 4A). Accordingly, control groups were largest in  
31  
32 326 rabbits as compared to pigs (p<0.05) and primates (p<0.01; Figure 4B), while procedure groups  
33  
34 327 were largest in rabbits as compared to pigs (p<0.01; Figure 4C).

35  
36  
37 328  
38  
39 329 [Table 3 about here]

40  
41  
42 330 [Figure 4 about here]

## 43 331 44 332 **4. Discussion**

45  
46  
47 333 Systematic bias may cause over- or underestimation of study results.<sup>3</sup> Quality items such  
48  
49 334 as randomization, allocation concealment, and blinded assessment ~~help to~~ improve internal  
50  
51 335 validity<sup>142</sup>, but are often neglected in small animal studies.<sup>3, 153, 164</sup>

52  
53  
54 336 Large animal models are believed to offer significant benefits for translational stroke  
55  
56 337 research. ~~Those comprise a~~ They have higher anatomical similarity to the human brain<sup>175</sup> and  
57  
58 338 to the cerebrovascular system ~~cerebrovascular system anatomy~~.<sup>6, 7, 186</sup> Another benefit is the

1  
2  
3 339 potential to use these models in experiments closely mimicking a human clinical situation, and  
4  
5 340 applying the same medical techniques and equipment for diagnostic and therapeutic  
6  
7 341 interventions that would be used in human patients.<sup>7, 197</sup> Moreover, physiological characteristics  
8  
9 342 of large animal models including heart and respiratory frequency, blood pressure as well as  
10  
11 343 pharmacodynamic and pharmacokinetic profiles are similar to humans.<sup>2018, 2119</sup> However, in  
12  
13 344 view of these advantages, large animal studies require much greater efforts and resources. It is  
14  
15 345 therefore important that quality in large animal studies is as high as possible to efficiently utilize  
16  
17 346 the advantages large animal models offer for translational research.

18  
19 347 Overall, we found that methodological quality in large animal stroke studies was  
20  
21 348 mediocre. Although quality generally improved significantly over the last decades and  
22  
23 349 potentially due to the 1999 publication and 2019 update of the STAIR criteria, our analysis  
24  
25 350 revealed some important shortcomings. Improvements are needed in reporting study subject  
26  
27 351 details and welfare (quality score category 1). Aspects such as sex and age, pre-study health  
28  
29 352 conditions, and medications should be reported routinely for optimal study transparency and  
30  
31 353 reproducibility, and transferability of study results.<sup>9</sup> The lack of comorbid large animal models  
32  
33 354 is not surprising. Comorbidities are difficult to simulate in outbred large animal models as they  
34  
35 355 occur due to age, distress, malnutrition and other factors according to the human  
36  
37 356 situation~~Comorbidities may hardly be mimicked in outbred large animal models as they occur~~  
38  
39 357 ~~due to age, distress, malnutrition and other factors according to the human situation~~, and can  
40  
41 358 take significant time in large animals to develop. Research on models exhibiting comorbidities  
42  
43 359 may remain a domain of small animal research. Nevertheless, any spontaneously occurring  
44  
45 360 comorbidities being diagnosed in large animals used for research should be reported.

46  
47 361 Working hypotheses were reported in almost all studies (99.5%), but often without any  
48  
49 362 obvious influence on study design. For instance, only 4.8% of the studies defined and reported  
50  
51 363 primary endpoints, while analysis of expectable effect size and a priori sample size calculation  
52  
53 364 were performed in few cases only (132.0%). This may severely limit the translational benefits



365 of large animal models since ~~neutral~~ studies results may be hard to interpret based on  
366 potentially poor statistical power. Given the significant resources required to perform large  
367 animal studies, considering these aspects is essential. On the other hand, determination of effect  
368 size can be challenging when previous research data is lacking or not entirely applicable. In  
369 these cases, we recommend to perform large animal pilot studies that may help to assess basic  
370 characteristics in the respective model, such as variability of ~~stroke infarct size~~ and its impact  
371 on the envisioned primary endpoint.

372 While ~~almost two third~~ half of the studies reported inclusion and exclusion criteria  
373 (~~64.650.0%~~), almost none (1.01%) applied them a priori. Defining inclusion and exclusion  
374 criteria during or after the study is believed to be a major source of bias, particularly when a  
375 study is conducted in ~~non-un~~blinded fashion. Hence, such bias can unfortunately not be  
376 excluded for most studies we analyzed.

377 Important quality aspects such as randomization (55.85%), allocation concealment  
378 (28.42%), and blinded assessment of outcome (~~49.750.0%~~) were more frequently reported in  
379 large animal studies as compared to small animal stroke experiments (randomization: 33.3%;  
380 blinded assessment of outcome: 44.4%,<sup>164</sup> allocation concealment: 25.9%; randomization,  
381 allocation concealment and blinded assessment of outcome: 24.1%.<sup>220</sup> Nevertheless, the  
382 number of studies not reporting those is still remarkably high in particular since blinding and  
383 randomization ~~should~~ be minimum standard quality assurance procedures in confirmative  
384 stroke research<sup>231</sup> to which almost all large animal studies aim to contribute.

385 Imaging techniques such as magnetic resonance imaging, computed tomography, and  
386 angiography (43.5%) as well as physiological monitoring (80.4%) were utilized relatively  
387 frequently. This is a positive aspect since large animals are particularly suitable for clinical  
388 imaging techniques while thorough physiological monitoring creates meaningful information  
389 that may warrant subject in- or exclusion. However, verification of infarct induction (only  
390 reported in ~~47.848.1%~~) as well as infarct size should be conducted thoroughly and routinely to

1  
2  
3 391 avoid the risk of increasing inter-subject/-study/-group variability, further reducing statistical  
4  
5 392 power of an experiment. Parameters such as ~~reduced~~ cerebral blood flow reduction for  
6  
7  
8 393 verification of infarct induction was documented by only 7.2% of studies. This is surprising  
9  
10 394 since these parameters are relatively easy to determine in large animals, while clinical imaging  
11  
12 395 techniques may be used to confirm the induced lesion directly.<sup>2119</sup>

13  
14 396 Large animals are suitable for long-term studies including functional endpoint  
15  
16 397 assessment. However, we only found a relatively low percentage (6.7%) of studies being  
17  
18 398 conducted for more than one month, the minimum follow-up period recommended by the  
19  
20 399 STAIR guidelines for functional endpoints. Next to costs, tThis may be due to the selection of  
21  
22 400 other primary endpoints such as safety or efficacy of recanalization methods which can be  
23  
24 401 assessed more rapidly. However, experimenters who wish to assess behavioral endpoints  
25  
26 402 should take into consideration that functional consequences of stroke in large animals can be  
27  
28 403 more heterogeneous than in rodent models, and may develop over longer time spans.<sup>242</sup>

29  
30 404 We recognized significant improvements in methodological quality since the publication  
31  
32 405 of the first STAIR guidelines in 1999, and in particular after the STAIR guideline update in  
33  
34 406 2009. Comparable improvements were reported for small animal stroke studies from 2010 to  
35  
36 407 2013.<sup>253</sup> These findings indicate the positive impact of specific good research practice  
37  
38 408 guidelines, which should be advanced continuously as evidenced by the recent 2019 STAIR  
39  
40 409 guideline updates.<sup>264</sup> In contrast to previous findings in small animal studies,<sup>275</sup> we also  
41  
42 410 identified positive association ( $r=0.27232802$ ;  $p<0.01$ ) between study quality and publication  
43  
44 411 in high-impact journals. In particular, total quality score as well as quality scores in all single  
45  
46 412 categories 1-4 significantly correlated with higher IF. This is an encouraging result since all  
47  
48 413 these categories include items being important to prevent bias. These items are hence **essential**  
49  
50 414 **indispensable** for a valid and transparent exchange of information between researchers.

51  
52 415 Group sizes were significantly larger in rabbits as compared to other species. This is not  
53  
54 416 surprising as rabbits are the smallest and cheapest of all large animal species what allows for

1  
2  
3 417 larger group sizes. Importantly, group sizes in primates are generally not different to that of  
4  
5 418 other species. This does not mean that group sizes were sufficient for each research question,  
6  
7 419 but shows that costs related to primate experiments did not prevent the same group sizes as seen  
8  
9  
10 420 in other large animal species despite rabbits.

11  
12 421 Our study has a number of limitations. We applied a predefined search strategy and  
13  
14 422 protocol being developed together an expert in literature meta-analyses (E.M.) and experts in  
15  
16 423 stroke research (J.B., S.M.). However, search strategy and protocol were not registered (ex ante  
17  
18  
19 424 protocol). Data extraction was not done in duplicates, but senior experts were consulted in all  
20  
21 425 doubtful cases. Intra-assessor reproducibility was not assessed. Moreover, we did not  
22  
23 426 discriminate between studies focusing on therapeutic and diagnostic procedures. Large animal  
24  
25 427 models provide a number of benefits over rodent models for diagnostic studies due to the larger  
26  
27 428 brain size and in particular when clinical imaging is used.<sup>33</sup> However, those studies are often  
28  
29 429 exploratory in nature. Since quality demands are different (and a bit lower) than in confirmative  
30  
31 430 studies, those imaging-related studies would perform normally worse but still can contribute  
32  
33 431 invaluably to their respective field.<sup>34</sup> Finally, we did not include a number of insightful imaging  
34  
35 432 studies because they did not conduct a formal inter-group comparison.<sup>35,36,37,38</sup>  
36  
37  
38  
39  
40 433

## 41 434 **5. Conclusions and Recommendations**

42  
43  
44 435 Although large animal models offer a offer a number of clear advantages for translational  
45  
46 436 strokeclear benefit in many translational stroke studies, we found that they are utilizedhave with  
47  
48 437 similar shortcomings than—to small animal models, limiting this—benefitthis benefit.  
49  
50 438 HenceTherefore, we derived a number of recommendations that may overcome to address these  
51  
52 439 limitations but are, at the same time, relatively easy to implement.  
53  
54  
55  
56 440

### 57 441 **5.1 Study planning and preparation**

58  
59 442 Large animal stroke studies are mostly confirmative studies. HenceTherefore, study  
60

1  
2  
3 443 planning should be based on high quality standards applied for randomized controlled clinical  
4  
5 444 trials (RCTs) when possible. Key elements of RCT planning and design such as a priori sample  
6  
7  
8 445 size calculation and endpoint definition should be conducted.<sup>234</sup> We encourage to involve  
9  
10 446 statisticians already in early planning steps~~Statisticians may be involved already in early~~  
11  
12 447 ~~planning steps~~ to optimize study design.<sup>286</sup> Study planning can also be supported by specific  
13  
14 448 software tools. For instance, the National Centre for the Replacement, Refinement and  
15  
16 449 Reduction of Animals in Research provides a freeware called Experimental Design Assistance  
17  
18 450 (<https://eda.nc3rs.org.uk>), which is free to use and was built to guide researchers through their  
19  
20 451 study planning.<sup>297</sup> Since optimal sample sizes may not be achieved for all endpoints, it is  
21  
22 452 important to clearly define the most appropriate primary study endpoint, and to power the study  
23  
24 453 properly. Collaboration between research teams in form of peer quality checks and validation  
25  
26 454 of study design can highly increase objectivity and validity of a study.<sup>3028</sup> Inter-group  
27  
28 455 collaboration and transfer of experience can also help to handle very complex models and/or  
29  
30 456 experimental setups, helping to reduce inter-subject variability negatively affecting statistical  
31  
32 457 power. Confirmative studies might be preregistered to maximize transparency.<sup>39</sup>  
33  
34  
35  
36  
37  
38  
39

458

## 459 5.2 Effect size estimation and pilot trials

460 Collecting valid information from previous research is essential for reliable effect size  
461  
462 estimation. If such data ~~is-are~~ not available, pilot studies may be helpful for at least basically  
463  
464 estimating variability of stroke impact and outcome in the model. In case previous experience  
465  
466 with a particular model is low, variability is more likely to be ~~overestimated-higher~~ and effect  
467  
468 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

1  
2  
3 469 effect size or regarding a specific research question from related fields.<sup>3129</sup>

4  
5 470

### 6 7 8 471 **5.3 Reducing the effect of sample size limitations and endpoint variability**

9  
10 472 Financial and logistical restrictions often impact sample and group sizes in large animal  
11  
12 473 experiments. This is an understandable limitation which is ~~hard~~ difficult to overcome. Selection  
13  
14 474 of a proper and relevant endpoint that can be adequately powered with respected to the  
15  
16 475 addressed research question (not necessarily functional outcome) is therefore important to  
17  
18 476 minimize the risk for low statistical power. Of note, some endpoints often used in studies  
19  
20 477 assessing therapeutic interventions, including infarct size and functional deficits, exhibit a  
21  
22 478 higher variability in large animal models than in rodent ~~ones~~, ~~making~~ This makes comparison  
23  
24 479 of absolute data more difficult.<sup>242</sup> Relative analysis of repeatedly assessed endpoints, i.e. in  
25  
26 480 comparison to the individual initial infarct size and/or functional deficit can efficiently  
27  
28 481 compensate for such variability, ~~allows to efficiently compensate for such variability~~. Repeated  
29  
30  
31  
32 482 assessments also allow calculating the area under the curve for particular endpoints. This may  
33  
34 483 provide a benefit in statistical power to identify whether a real outcome benefit is present over  
35  
36 484 time. However, this comes at the cost of temporal resolution: it cannot be concluded exactly  
37  
38 485 when this benefit became evident. There is also preliminary evidence for fast and slow stroke  
39  
40 486 progressors in large animals, indicating different collateral status and somewhat resembling the  
41  
42 487 human situation, but further contributing to inter-subject variability. It is recommended to  
43  
44  
45 488 consider this fact when planning an acute stroke study.<sup>320</sup>

46  
47  
48  
49 489 In experiments of highly similar design, controls may be pooled. Of note, this counteracts  
50  
51 490 randomization and therefore requires extremely thorough validation of comparability of control  
52  
53 491 subjects from different experiments/sources. If comparability is thoroughly proven, this may  
54  
55 492 help to increase statistical power, but the limitations of this approach and potentially resulting  
56  
57 493 bias need to be discussed transparently and in detail when publishing results.

58  
59  
60 494 The possibility to repeatedly collect a broad spectrum of physiological data should be

1  
2  
3 495 utilized ~~to the best~~where possible ~~extend~~, as deviation from normal parameter ranges may  
4  
5 496 explain variability and warrant post-hoc exclusion of subjects in single cases.  
6  
7

8 497

#### 10 498 **5.4 Study duration and documentation**

12 499 We recommend considering long-term experiments whenever meaningful and possible  
13  
14 500 and meeting animal welfare requirements. Even though long-term experiments involve greater  
15  
16 501 efforts, the amount of data collected for individual subjects may be much higher, providing a  
17  
18 502 better overall picture on the assessed intervention. Documentation should be as transparent as  
19  
20  
21 503 possible ~~since~~ because transparency is not challenging or laborious, but contributes  
22  
23 504 significantly to increased scientific rigor, reproducibility, and unbiased study result  
24  
25 505 interpretation ~~at all~~. Methodological limitations including lacking quality aspects due to good  
26  
27 506 reason should be clearly stated as this allows better interpretation of positive, neutral and  
28  
29 507 negative study results.  
30  
31

32  
33 508  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Review Only

1  
2  
3 509 **6. Acknowledgements**  
4

5 510 ~~ES is supported by the Stroke Association (SA L-SNC 18\1003). We thank Qianying Wang,~~

6  
7  
8 511 ~~Centre for Clinical Brain Sciences, University of Edinburgh for her expert guidance on~~

9  
10 512 ~~data analysis.~~  
11

12 513

13  
14 514 **7. Disclosure**

15  
16  
17 515 The authors do not report relevant disclosures.  
18

19 516

20  
21 517 **8. Supplementary data**

22  
23  
24 518 1. Supplementary Table 1: Search strategy in Medline

25  
26 519 2. Supplementary Table 2: Search strategy in Web of Science

27  
28 520 3. Supplementary Table 3: Type and frequency of inappropriate analysis methods

29  
30 521 4. Supplementary Figure 1: Association between impact factor and quality score within

31  
32 522 individual categories

33  
34  
35 523 5. Supplementary reference list

36  
37 524

38  
39  
40 525 Supplementary material for this paper can be found at the journal website:

41  
42 526 <http://journals.sagepub.com/home/jcb>  
43

44 527

45  
46  
47 528  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



1  
2  
3 529 **9. References**  
4

- 5 530 1. Bush CK, Kurimella D, Cross LJ, et al. Endovascular Treatment with Stent-Retriever  
6  
7 531 Devices for Acute Ischemic Stroke: A Meta-Analysis of Randomized Controlled  
8  
9 532 Trials. *PLoS One*. 2016; e0147287.  
10  
11  
12 533 2. O'Collins VE, Macleod MR, Donnan GA, et al. 1,026 experimental treatments in acute  
13  
14 534 stroke. *Ann Neurol* 2006; 59: 467-477.  
15  
16  
17 535 3. Macleod MR, Fisher M, O'Collins V, et al. Reprint: Good laboratory practice:  
18  
19 536 preventing introduction of bias at the bench. *Int J Stroke* 2004; 59: 3-5.  
20  
21  
22 537 4. Macleod MR, Lawson Mc Lean A, Kyriakopoulou A, et al. Risk of Bias in Reports of  
23  
24 538 In Vivo Research: A Focus for Improvement. *PLoS Biol* 2015; 13: e1002273.  
25  
26 539 5. Sena ES, van der Worp HB, Bath PM, et al. Publication bias in reports of animal stroke  
27  
28 540 studies leads to major overstatement of efficacy. *PLoS Biol* 2010; 8: e1000344.  
29  
30  
31 541 6. Herrmann AM, Meckel S, Gounis MJ, et al. Large animal in neurointerventional  
32  
33 542 research: A systematic review on models, techniques and their application in  
34  
35 543 endovascular procedures for stroke, aneurysms and vascular malformations. *J Cereb*  
36  
37 544 *Blood Flow Metab*. 2019; 39(3): 375-394.  
38  
39  
40 545 7. Traystman RJ. Animal models of focal and global cerebral ischemia. *ILAR J* 2003; 44:  
41  
42 546 85-95.  
43  
44  
45 547 8. Sena ES, Van der Worp HB, Howells D, et al. How can we improve the pre-clinical  
46  
47 548 development of drugs for stroke? *Trends Neurosci* 2007; 30: 433-439.  
48  
49 549 9. STAIR. Recommendations for standards regarding preclinical neuroprotective and  
50  
51 550 restorative drug development. *Stroke* 1999; 30: 2752-2758.  
52  
53  
54 551 10. Fisher M, Feuerstein G, Howells D, et al. Update of the stroke therapy academic  
55  
56 552 industry roundtable preclinical recommendations. *Stroke* 2009; 40: 2244-2250.  
57  
58 553 11. Kilkenney C, Browne WJ, Cuthill IC, et al. Improving bioscience research reporting:  
59  
60 554 the ARRIVE guidelines for reporting animal research. *PLoS Biol* 2010; 8: e1000412.



- 1  
2  
3 555 [12. Jovin TG, Albers, GW, Liebeskind DS, et al. Stroke Treatment Academic Industry](#)  
4  
5 556 [Roundtable: The Next Generation of Endovascular Trials. \*Stroke\* 2016; 47\(19\): 2656-](#)  
6  
7 [2665.](#)  
8 557  
9  
10 558 [13. Goyal M, Menon BK, van Zwam WH, et al. Endovascular thrombectomy after large-](#)  
11  
12 559 [vessel ischaemic stroke: a meta-analysis of individual patient data from five](#)  
13  
14 560 [randomized trials. \*Lancet\* 2016; 387\(10029\): 1723-1731.](#)  
15  
16  
17 561 ~~11.~~  
18  
19 562 ~~12.~~14. Sena ES, Currie GL, McCann SK, et al. Systematic reviews and meta-analysis  
20  
21 563 of preclinical studies: why perform them and how to appraise them critically. *J Cereb*  
22  
23 564 *Blood Flow Metab* 2014; 34: 737-742.  
24  
25 565 ~~13.~~15. Sena ES, Van der Worp HB, Bath PM, et al. Publication bias in reports of  
26  
27 566 animal stroke studies leads to major overstatement of efficacy. *PLoS Biol* 2010; 8:  
28  
29 567 e1000344.  
30  
31  
32 568 ~~14.~~16. Macleod MR, Van der Worp HB, Sena ES, et al. Evidence for the efficacy of  
33  
34 569 NXY-059 in experimental focal cerebral ischaemia is confounded by study quality.  
35  
36 570 *Stroke* 2008; 39: 2824-2829.  
37  
38  
39 571 ~~15.~~17. Boltze J, Nitzsche F, Jolkkonen J, et al. Concise Review: Increasing the  
40  
41 572 Validity of Cerebrovascular Disease Models and Experimental Methods for  
42  
43 573 Translational Stem Cell Research. *Stem Cells* 2017; 35: 1141-1153.  
44  
45 574 ~~16.~~18. Sommer CJ. Ischemic stroke: experimental models and reality. *Acta*  
46  
47 575 *Neuropathol* 2017; 133: 245-261.  
48  
49 576 ~~17.~~19. Mehra M, Henninger N, Hirsch JA et al. Preclinical acute ischemic stroke  
50  
51 577 modeling. *J Neurointerv Surg* 2012; 4: pp. 307-3013.  
52  
53 578 ~~18.~~20. Helke KL and Swindle MM. Animal models of toxicology testing: the role of  
54  
55 579 pigs. *Expert Opin Drug Metab Toxicol* 2013; 9: 127-139.  
56  
57  
58  
59  
60

- 1  
2  
3 580 [19-21.](#) Herrmann AM, Cattaneo GFM, Eiden, SA, et al. Development of a Routinely  
4  
5 581 Applicable Imaging Protocol for Fast and Precise Middle Cerebral Artery Occlusion  
6  
7 582 Assessment and Perfusion Deficit Measure in an Ovine Stroke Model: A Case Study.  
8  
9 *Front Neurol.* 2019; 10: 1113.  
10 583  
11  
12 584 [20-22.](#) Minnerup J, Zentsch V, Schmidt A, et al. Methodological Quality of  
13  
14 585 Experimental Stroke Studies Published in the Stroke Journal: Time Trends and Effect  
15  
16 586 of the Basic Science Checklist. *Stroke* 2016; 47: 267-272.  
17  
18  
19 587 [21-23.](#) Dirnagl U. Bench to bedside: the quest for quality in experimental stroke  
20  
21 588 research. *J Cereb Blood Flow Metab* 2006; 26: 1465-1478.  
22  
23  
24 589 [22-24.](#) Boltze J, Modo MM, Mays RW, et al. Stem Cells as an Emerging Paradigm in  
25  
26 590 Stroke 4: Advancing and Accelerating Preclinical Research. *Stroke* 2019; 50 (11):  
27  
28 591 3299-3306.  
29  
30 592 [23-25.](#) Minnerup J, Wersching H and Diederich K. Methodological quality of  
31  
32 593 preclinical stroke studies is not required for publication in high-impact journals. *J*  
33  
34 594 *Cereb Blood Flow Metab* 2010; 30: 1619-1624.  
35  
36  
37 595 [24-26.](#) Savitz SI, Baron JC, Fisher M, et al. Stroke Treatment Academic Industry  
38  
39 596 Roundtable X: Brain Cyoprotection Therapies in the Reperfusion Era. *Stroke* 2019.  
40  
41 597 50(4): 1026- 1031.  
42  
43  
44 598 [25-27.](#) Minnerup J, Wersching H and Diederich K. Methodological quality of  
45  
46 599 preclinical stroke studies is not required for publication in high-impact journals. *J*  
47  
48 600 *Cereb Blood Flow Metab* 2010; 30: 1619-1624.  
49  
50  
51 601 [26-28.](#) Würbel H. More than 3Rs: the importance of scientific validity for harm-  
52  
53 602 benefit analysis of animal research. *Lab Anim (NY)* 2017; 46: 164-166.  
54  
55  
56 603 [27-29.](#) Percie du Sert N, Bamsey I, Bate ST, et al. The Experimental Design Assistant.  
57  
58 604 *PLoS Biol* 2017; 15: e2003779.  
59  
60

- 1  
2  
3 605 [30.](#) Sena ES, Currie GL, McCann SK, et al. Systematic reviews and meta-analysis of  
4  
5 606 preclinical studies: why perform them and how to appraise them critically. *J Cereb*  
6  
7 607 *Blood Flow Metab* 2014; 34: 737-742.  
8  
9  
10 608 ~~28.~~  
11  
12 609 ~~29.~~ [31.](#) Begley CG and Ioannidis JP. Reproducibility in science: improving the standard  
13  
14 610 for basic and preclinical research. *Circ Res* 2015; 116: 116-126.  
15  
16  
17 611 [32.](#) Shazeeb MS, King RM, Brooks OW, et al. Infarct Evolution in a Large Animal Model  
18  
19 612 of Middle Cerebral Artery Occlusion. *Transl. Stroke Research* 2019. doi:  
20  
21 613 10.1007/s12975-019-00732-9.  
22  
23  
24 614 [33.](#) Werner P, Saur D, Zeisig V, et al. Simultaneous PET/MRI in stroke: a case series. *J*  
25  
26 615 *Cereb Blood Flow* 2015; 35(9): 1421-1425.  
27  
28  
29 616 [34.](#) Dirnagl U, Hakim A, Macleod M, et al. A concerted appeal for international  
30  
31 617 [cooperation in preclinical stroke research.](#) *Stroke* 2013; 44(6): 1754-1760.  
32  
33 618 [35.](#) Boltze J, Ferrara F, Hainsworth AH, et al. Lesional and perilesional tissue  
34  
35 619 [characterization by automated image processing in a novel gyrencephalic animal](#)  
36  
37 620 [model of peracute intracerebral hemorrhage.](#) *J Cereb Blood Flow Metab* 2019; 39(12):  
38  
39 621 [2521-2535.](#)  
40  
41  
42 622 [36.](#) Haque ME, Gabr RE, Zhao X, et al. Serial quantitative neuroimaging of iron in the  
43  
44 623 [intracerebral hemorrhage pig model.](#) *J Cereb Blood Flow Metab* 2018; 38(3): 375-  
45  
46 624 [381.](#)  
47  
48  
49 625 [37.](#) Kamimura HA, Flament J, Valette J, et al. Feedback control of microbubble cavitation  
50  
51 626 [for ultrasound-mediated blood-brain barrier disruption in non-human primates under](#)  
52  
53 627 [magnetic resonance guidance.](#) *J Cereb Blood Flow Metab* 2019; 39(7): 1191-1203.  
54  
55  
56 628 [38.](#) Sander CY, Mandeville JB, Wey HY, et al. Effects of flow changes on radiotracer  
57  
58 629 [binding: Simultaneous measurement of neuroreceptor binding and cerebral blood flow](#)  
59  
60 630 [modulation.](#) *J Cereb Blood Flow Metab* 2019; 39(1): 131-146.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

631 39. Kimmelman J and Anderson JA. Should preclinical studies be registered? *Nat*

632 *Biotechnol* 2012; 30(6): 488-489.

633

634

635

Confidential: For Review Only

1  
2  
3 **636 Figure legends**  
4

5 **637 Figure 1. Overview on quantitative search results and frequency of large animal**  
6  
7 **638 experiments in stroke research since 1990.**

8  
9  
10 639 (A) Flow diagram of publication identification. N = Number of publications. Records  
11  
12 640 were excluded after screening title and abstracts. Full-text articles were then screened and  
13  
14 641 excluded for a priori determined reasons. (B) Timeline of publication in large animal stroke  
15  
16 642 research (1990-2019): The increase of large animal stroke studies in the last years is potentially  
17  
18 643 due to the breakthrough in recanalization therapies, prompting a number of follow-on  
19  
20 644 translational studies utilizing large animal stroke models.  
21  
22  
23  
24  
25

26 **646 Figure 2. Influence of study origin and STAIR criteria publication on study quality.**  
27

28 647 (A) Total quality score, (B) Category 1: Reporting of study subject and animal welfare,  
29  
30 648 (C) Category 2: Study planning quality (North America vs. Europe  $p < 0.01$ ), (D) Category 3:  
31  
32 649 Study conductance quality (North America vs. Asia & Oceania  $p < 0.01$ ), (E) Category 4: Result  
33  
34 650 reporting and analysis quality (North America vs. Europe  $p < 0.01$ ), (F) ~~Improvement in total~~  
35  
36 651 ~~methodological quality since the publication of the first STAIR criteria in 1999~~  
37  
38 ~~( $p < 0.01$ ), Influence of species,~~ (G) Improvement in total methodological quality since the  
39  
40 652 ~~publication of the first STAIR criteria comparing to their amendment in 1999 (2010-2019 vs.~~  
41  
42 653 ~~1990-1999  $p < 0.01$ ), and 2010-2019 vs. 2000-2009  $p < 0.05$ ), (H) ~~Improvement in total~~  
43  
44 654 ~~methodological quality since the publication of the first STAIR criteria in 1999 comparing to~~  
45  
46 655 ~~their amendment in 2009 (2010-2019 vs. 1990-1999  $p < 0.01$ , and 2010-2019 vs. 2000-2009~~  
47  
48 656  ~~$p < 0.05$ ), Influence of species,~~ (I) Influence of type of intervention. Horizontal lines and  
49  
50 657 whiskers indicate the ~~median~~ with lower and upper 95% CI. \* $p < 0.05$ ; \*\* $p < 0.01$ .  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60~~

61 **660 Figure 3. Association between total quality score versus impact factor.**

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

662 Number of included studies is 17~~23~~, no IF could be retrieved for 36 studies. The latter studies  
 663 were excluded from this analysis.

664

665 **Figure 4. Group sizes across species.**

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

671

672

673 **Tables**

674 **Table 1. Quality score items.**

Category 1: Reporting of study subject details and welfare		Category 2: Study planning quality	
<i>Item</i>	<i>Score point</i>	<i>Item</i>	<i>Score point</i>
	<i>allocation</i>		<i>allocation</i>
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
<b>Category 3: Internal study validity</b>		<b>Category 4: Outcome analysis and reporting</b>	
<i>Item</i>	<i>Score point</i>	<i>Item</i>	<i>Score point</i>
	<i>allocation</i>		<i>allocation</i>
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

675 #analysis modalities were considered appropriate when being sufficient to assess the

676 respective research question or endpoint (see Supplementary Table 3 for details).

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

678

679

680 **Table 2. Basic Characteristics of included Animal Experimental Studies.**

Item	Frequency (%)	Item	Frequency (%)	Item	Frequency (%)
<b>Species</b>		<b>Type of intervention</b>		<b>Study duration</b>	
Rabbit	n=96 (45.946.1%)	Neuroprotectives	n=113 (54.31%)	Acute phase (<24h)	n=13940 (67.066.9%)
Cat	n=43 (20.76%)	Thrombolytics	n=52 (24.925.0%)	1-3 days	n=26 (12.54%)
Dog	n=16 (7.7%)	Cell therapies	n=78 (3.83.4%)	<1 week	n=15 (7.2%)
Non-Human-Primate	n=32 (15.43%)	Diagnostics	n=15 (7.2%)	<1month	n=14 (6.7%)
Pig	n=1920 (9.16%)	Others#	n=21 (10.10%)	>1 month	n=14 (6.7%)
Non-Human-Primate & Rabbit	n=1 (0.5%)				
Sheep	n=1 (0.5%)				
<b>Region</b>		<b>Primary endpoint</b>		<b>Stroke model</b>	
North America	n=1345 (64.46%)	Efficacy	n=162 (77.95%)	Transient	n=120 (57.74%)
Europe	n=24 (11.5%)	Safety	n=12 (5.87%)	Permanent	n=767 (36.58%)
Asia/Oceania	n=50 (24.13.9%)	Feasibility	n=223	Transient +Permanent	n=1 (0.5%)



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46

		( <del>11.0</del> 10.5%)		
	Safety +	n=1 (0.5%)	Not reported	n=11 (5.3%)
	Feasibility			
	Safety + Efficacy	n=11 (5.3%)		
<b>Further information</b>				
Additional veterinary care	n=11 (5.3%)			
reported				
Dose-response	n=30 (14.4%)			
relationship reported				
Compliance with animal	n=128 (61. <del>5</del> 2%)			
welfare regulations				
reported				
Pre-study quarantine	n=3 (1.4%)			
reported				
Animal housing	n=23 (11. <del>1</del> 0%)			
conditions## reported				

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

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

685

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

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

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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

689

Confidential: For Review Only

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

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

<sup>1</sup>Department of Neuroradiology, Neurocenter,

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

<sup>2</sup>Faculty of Veterinary Medicine, Institute of Anatomy, Histology and Embryology, Leipzig University, Leipzig, Germany

<sup>3</sup>Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom

<sup>4</sup>Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

<sup>5</sup>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.<sup>1</sup> 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.<sup>2</sup>

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.<sup>3,4,5</sup>

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.<sup>6,7</sup> 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.<sup>8</sup> 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)<sup>9,10</sup> and Animals in Research: Reporting In Vivo Experiments (ARRIVE) guidelines.<sup>11</sup> 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 26<sup>th</sup>, 2017, and an update was performed on August 9<sup>th</sup>, 2019. Data base entries between January 1<sup>st</sup>, 1990 and August 8<sup>th</sup>, 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 animals* 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.<sup>9,10</sup> 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

1  
2  
3 experiments published from 1990 to 2014 generally decreased from n=56 in 1990-1994 to n=21  
4 in 2010-2014 (Figure 1B). However, there was a steep increase in published studies from 2015,  
5 reaching an all-time high (n=40) even though studies published in late 2019 are not yet included  
6 in our search strategy. This might be related to the milestone evidence for clinical benefit  
7 publication of mechanical thrombectomy in large vessel occlusion stroke by the publication of  
8 five randomized controlled trials in 2015 that may have sparked new interest in the field and an  
9 increased demand for large animal models to investigate related procedures.<sup>12,13</sup>  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20

21 [Figure 1 about here]

22 [Table 2 about here]

### 23 24 25 26 27 28 **3.2. Study Quality**

29  
30  
31 The overall median quality score was 11 (range 3 to 22; IQR: 4 (9-13)) out of 24. The  
32 median quality score in the first category (reporting of study subject details and welfare) was 2  
33 out of 6 (range 1 to 5; IQR: 1 (1-2)). The second category (study planning quality) also reached  
34 a median quality score of 2 (range 1 to 6; IQR: 1 (2-3)). The third category (study conductance  
35 quality) had a median score of 3 (range 0 to 6; IQR: 2 (2-4)). Category 4 (result reporting and  
36 analysis quality) had a median quality score of 4 (range 0 to 6; IQR: 1 (2-4)). A significantly  
37 lower number of quality criteria were fulfilled in category 1 in comparison to the others  
38 (p<0.05).  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50

#### 51 **3.2.1. Study subject details and welfare (category 1)**

52  
53 All studies reported the species used, but only 146 studies (70.2%) reported that the study  
54 was approved by responsible animal welfare authorities. Sex and age were reported by 31  
55 studies (15.0%). Sex only was reported by 153 (73.6%), while age was not reported solely. The  
56 pre-study health status was reported by only 12 studies (5.8%). Medication details including  
57  
58  
59  
60

1  
2  
3 the use of companion medication (e.g., analgetics, antibiotics) was reported in only 20 studies  
4  
5 (9.6%). Comorbidities were not reported by any study.  
6  
7  
8  
9

### 10 **3.2.2. Study planning (category 2)**

11  
12 Working hypotheses were reported in 207 (99.5%) studies. However, primary study  
13  
14 endpoints were nominally determined in only 10 studies (4.8%). 135 (64.6%) studies reported  
15  
16 that the study rationale was based on earlier small animal (n=79; 38.0%) or in vitro studies  
17  
18 (n=25; 12.1%), or both (n=16; 7.7%). Effect size estimation and a priori sample size calculation  
19  
20 can be performed based on such data. However, only 27 studies (13.0%) actually reported an  
21  
22 estimation of effect size and a priori sample size calculation. A specific primary working  
23  
24 hypothesis explicitly referring to previous in vitro and/or in vivo studies was reported in 18  
25  
26 studies (8.7%). Inclusion and exclusion criteria were reported in 104 studies (50.0%), but only  
27  
28 2 studies (1.0%) determined these criteria a priori.  
29  
30  
31  
32  
33  
34

### 35 **3.2.3. Study conductance (category 3)**

36  
37 Randomization was reported in 116 studies (55.8%), and allocation concealment was  
38  
39 reported in 59 cases (28.4%). 104 studies (50.0%) reported blinded outcome assessment.  
40  
41 Measurement of physiological parameters was reported in 165 cases (79.3%). The most  
42  
43 frequently monitored parameters included mean arterial pressure (systemic), temperature,  
44  
45 blood gases, blood pH, and exhalation gases. 186 studies reported appropriate outcome analysis  
46  
47 modalities (89.4%; information on inappropriate analysis modalities are provided in  
48  
49 Supplementary Table 3). These included survival rate (n=2; 1.0%), functional outcome (n=67;  
50  
51 32.2%), infarct size (n=46; 22.1%, as determined by appropriate methods such as imaging or  
52  
53 histology), other imaging (n=90; 43.3%) or histology (n=61; 29.3%) endpoints, clinical  
54  
55 chemistry (n=52; 25.0%), general pathology (n=24; 11.5%) or both (n=18; 8.7%). Only a  
56  
57  
58  
59  
60

1  
2  
3 fraction of studies that recorded physiological parameters finally analyzed those (n=52; 25.0%).  
4  
5 100 studies (48.1%) reported verification of infarct induction during intervention.  
6  
7  
8  
9

### 10 **3.2.4. Result reporting and analysis (category 4)**

11  
12 168 studies (80.8%) adequately reported relevant data and findings in form of detailed  
13 tables or graphs. However, data were almost exclusively reported as means or medians.  
14  
15 Individual data points were only provided by 16 studies (7.7%). Drop outs and excluded  
16  
17 subjects were reported in 105 studies (50.5%). Application of appropriate statistical tests was  
18  
19 reported in 192 studies (92.3%). 16 studies incompletely reported statistical analysis and, for  
20  
21 example lacking information regarding statistical tests applied including post hoc tests. 91  
22  
23 studies (43.8%) described potential sources of error and bias in the experiment, while 115  
24  
25 (55.3%) reported limitations such as small sample size or that it was impossible to perform  
26  
27 randomization. A conclusion fully justified by study findings was given in by most, but not all  
28  
29 reports (n=190; 91.3%).  
30  
31  
32  
33  
34  
35  
36  
37

## 38 **3.3. Additional influences on study quality**

### 39 **3.3.1. Study quality versus origin, species and type of intervention**

40  
41 Total median quality score was highest in studies from North America (Median: 12; IQR:  
42 10-14), statistically different from studies conducted in Asia & Oceania (Median: 10; IQR:  
43 (8.75-12) or Europe (Median: 10; IQR: 8-11.75;  $p=0.0011$  Figure 2A). Analysis of individual  
44  
45 quality categories revealed no differences in category 1 (Figure 2B) but North American studies  
46  
47 had statistically significantly higher scores in quality categories 2 (Median: 2.5; IQR: 2-3) and  
48  
49 3 (Median: 4; IQR: 3-5) than their European counterparts (Median: 2; IQR: 1-2;  $p<0.01$ ; Figure  
50  
51 2C). Furthermore, North American studies were superior to Asian & Oceanian studies in  
52  
53 category 3 (Median 3; IQR: 2-4;  $p<0.01$ ; Figure 2D). We did not find statistically significant  
54  
55 differences regarding category 4 (Figure 2E). Quality scores were neither influenced by species  
56  
57  
58  
59  
60

1  
2  
3 used (Figure 2F) nor by the types of intervention (Figure 2G). Overall differences in median  
4  
5 quality score in species varied significantly without any specific intergroup difference.  
6  
7  
8  
9

10 [Figure 2 about here]  
11

### 12 13 14 **3.3.2. Study quality in the post-STAIR era**

15  
16 Methodological quality significantly improved after introduction of the STAIR guidelines in  
17  
18 1999 (1990-1999 pre-STAIR median: 10, IQR: 8-12; post-STAIR median: 12, IQR: 9-15;  
19  
20  $p < 0.01$ ; Figure 2H). We also compared quality scores of studies published prior to the first  
21  
22 STAIR guidelines to quality scores of studies published in the time between the first STAIR  
23  
24 STAIR guidelines to quality scores of studies published in the time between the first STAIR  
25  
26 guideline publication and the 2009 update (2000-2009; median: 11; IQR: 9-13), and to scores  
27  
28 of studies published after the STAIR 2009 update (2010-2019; median: 13.; IQR: 10-15).  
29  
30 Quality scores of studies published after the STAIR 2009 update were higher than those of  
31  
32 studies published before the initial STAIR guideline publication (1990-1999;  $p < 0.01$ ). They  
33  
34 were also higher than quality scores of studies published after the first publication of STAIR  
35  
36 guidelines and prior to the 2009 update (2000-2009;  $p < 0.05$ ; Figure 2I).  
37  
38  
39

40 Improvements were particularly evident in categories 1 and 4. In category 1, quality  
41  
42 scores were lower in pre-STAIR studies (1990-1999; median: 1, IQR: 1-2) as compared to  
43  
44 studies published after the first publication of STAIR guidelines and prior to the 2009 update  
45  
46 (2000-2009; median: 2; IQR: 1.25-2) and to studies published after the 2009 update (2010-  
47  
48 2019; median: 2; IQR: 2-3;  $p < 0.01$ ). There was also a significant difference in category 1  
49  
50 quality scores of studies published after the 2009 update to studies published between 2000 and  
51  
52 2009 ( $p < 0.01$ ). In category 4, quality scores of studies published after the 2009 STAIR update  
53  
54 (2010-2019; median: 4; IQR: 3-5) were higher than those of studies published before the STAIR  
55  
56 guidelines introduction (1990-1999; median: 3; IQR: 2-4) and those of studies published  
57  
58 between 2000 and 2009 (median 3; IQR: 2-4;  $p < 0.01$  each).  
59  
60

### 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.<sup>3</sup> Quality items such

1  
2  
3 as randomization, allocation concealment, and blinded assessment improve internal validity<sup>14</sup>,  
4  
5 but are often neglected in small animal studies.<sup>3, 15, 16</sup>  
6  
7

8 Large animal models are believed to offer significant benefits for translational stroke  
9  
10 research. They have higher anatomical similarity to the human brain<sup>17</sup> and to the  
11  
12 cerebrovascular system.<sup>6, 7, 18</sup> Another benefit is the potential to use these models in experiments  
13  
14 closely mimicking a human clinical situation, and applying the same medical techniques and  
15  
16 equipment for diagnostic and therapeutic interventions that would be used in human patients.<sup>7,</sup>  
17  
18 <sup>19</sup> Moreover, physiological characteristics of large animal models including heart and  
19  
20 respiratory frequency, blood pressure as well as pharmacodynamic and pharmacokinetic  
21  
22 profiles are similar to humans.<sup>20, 21</sup> However, in view of these advantages, large animal studies  
23  
24 require much greater efforts and resources. It is therefore important that quality in large animal  
25  
26 studies is as high as possible to efficiently utilize the advantages large animal models offer for  
27  
28 translational research.  
29  
30  
31

32  
33 Overall, we found that methodological quality in large animal stroke studies was  
34  
35 mediocre. Although quality generally improved significantly over the last decades and  
36  
37 potentially due to the 1999 publication and 2019 update of the STAIR criteria, our analysis  
38  
39 revealed some important shortcomings. Improvements are needed in reporting study subject  
40  
41 details and welfare (quality score category 1). Aspects such as sex and age, pre-study health  
42  
43 conditions, and medications should be reported routinely for optimal study transparency and  
44  
45 reproducibility, and transferability of study results.<sup>9</sup> The lack of comorbid large animal models  
46  
47 is not surprising. Comorbidities are difficult to simulate in outbred large animal models as they  
48  
49 occur due to age, distress, malnutrition and other factors according to the human situation, and  
50  
51 can take significant time in large animals to develop. Research on models exhibiting  
52  
53 comorbidities may remain a domain of small animal research. Nevertheless, any spontaneously  
54  
55 occurring comorbidities being diagnosed in large animals used for research should be reported.  
56  
57  
58  
59  
60

Working hypotheses were reported in almost all studies (99.5%), but often without any



1  
2  
3 obvious influence on study design. For instance, only 4.8% of the studies defined and reported  
4 primary endpoints, while analysis of expectable effect size and a priori sample size calculation  
5 were performed in few cases only (13.0%). This may severely limit the translational benefits  
6 of large animal models since study results may be hard to interpret based on potentially poor  
7 statistical power. Given the significant resources required to perform large animal studies,  
8 considering these aspects is essential. On the other hand, determination of effect size can be  
9 challenging when previous research data is lacking or not entirely applicable. In these cases,  
10 we recommend to perform large animal pilot studies that may help to assess basic  
11 characteristics in the respective model, such as variability of infarct size and its impact on the  
12 envisioned primary endpoint.

13  
14  
15 While half of the studies reported inclusion and exclusion criteria (50.0%), almost none  
16 (1.0%) applied them a priori. Defining inclusion and exclusion criteria during or after the study  
17 is believed to be a major source of bias, particularly when a study is conducted in non-blinded  
18 fashion. Hence, such bias can unfortunately not be excluded for most studies we analyzed.

19  
20  
21 Important quality aspects such as randomization (55.8%), allocation concealment  
22 (28.4%), and blinded assessment of outcome (50.0%) were more frequently reported in large  
23 animal studies as compared to small animal stroke experiments (randomization: 33.3%; blinded  
24 assessment of outcome: 44.4%,<sup>16</sup> allocation concealment: 25.9%; randomization, allocation  
25 concealment and blinded assessment of outcome: 24.1%.<sup>22</sup> Nevertheless, the number of studies  
26 not reporting those is still remarkably high in particular since blinding and randomization  
27 should be minimum standard quality assurance procedures in confirmative stroke research<sup>23</sup> to  
28 which almost all large animal studies aim to contribute.

29  
30  
31 Imaging techniques such as magnetic resonance imaging, computed tomography, and  
32 angiography (43.5%) as well as physiological monitoring (80.4%) were utilized relatively  
33 frequently. This is a positive aspect since large animals are particularly suitable for clinical  
34 imaging techniques while thorough physiological monitoring creates meaningful information  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 that may warrant subject in- or exclusion. However, verification of infarct induction (only  
4 reported in 48.1%) as well as infarct size should be conducted thoroughly and routinely to avoid  
5 the risk of increasing inter-subject/-study/-group variability, further reducing statistical power  
6 of an experiment. Parameters such as cerebral blood flow reduction for verification of infarct  
7 induction was documented by only 7.2% of studies. This is surprising since these parameters  
8 are relatively easy to determine in large animals, while clinical imaging techniques may be used  
9 to confirm the induced lesion directly.<sup>21</sup>

10  
11  
12  
13  
14  
15  
16  
17  
18  
19 Large animals are suitable for long-term studies including functional endpoint  
20 assessment. However, we only found a relatively low percentage (6.7%) of studies being  
21 conducted for more than one month, the minimum follow-up period recommended by the  
22 STAIR guidelines for functional endpoints. Next to costs, this may be due to the selection of  
23 other primary endpoints such as safety or efficacy of recanalization methods which can be  
24 assessed more rapidly. However, experimenters who wish to assess behavioral endpoints  
25 should take into consideration that functional consequences of stroke in large animals can be  
26 more heterogeneous than in rodent models, and may develop over longer time spans.<sup>24</sup>

27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37 We recognized significant improvements in methodological quality since the publication  
38 of the first STAIR guidelines in 1999, and in particular after the STAIR guideline update in  
39 2009. Comparable improvements were reported for small animal stroke studies from 2010 to  
40 2013.<sup>25</sup> These findings indicate the positive impact of specific good research practice  
41 guidelines, which should be advanced continuously as evidenced by the recent 2019 STAIR  
42 guideline updates.<sup>26</sup> In contrast to previous findings in small animal studies,<sup>27</sup> we also identified  
43 positive association ( $r=0.2802$ ;  $p<0.01$ ) between study quality and publication in high-impact  
44 journals. In particular, total quality score as well as quality scores in all single categories 1-4  
45 significantly correlated with higher IF. This is an encouraging result since all these categories  
46 include items being important to prevent bias. These items are hence indispensable for a valid  
47 and transparent exchange of information between researchers.

1  
2  
3 Group sizes were significantly larger in rabbits as compared to other species. This is not  
4 surprising as rabbits are the smallest and cheapest of all large animal species what allows for  
5 larger group sizes. Importantly, group sizes in primates are generally not different to that of  
6  
7 larger group sizes. Importantly, group sizes in primates are generally not different to that of  
8 other species. This does not mean that group sizes were sufficient for each research question,  
9 but shows that costs related to primate experiments did not prevent the same group sizes as seen  
10 in other large animal species despite rabbits.  
11  
12  
13  
14  
15

16  
17 Our study has a number of limitations. We applied a predefined search strategy and  
18 protocol being developed together an expert in literature meta-analyses (E.M.) and experts in  
19 stroke research (J.B., S.M.). However, search strategy and protocol were not registered (ex ante  
20 protocol). Data extraction was not done in duplicate, but senior experts were consulted in all  
21 doubtful cases. Intra-assessor reproducibility was not assessed. Moreover, we did not  
22 discriminate between studies focusing on therapeutic and diagnostic procedures. Large animal  
23 models provide a number of benefits over rodent models for diagnostic studies due to the larger  
24 brain size and in particular when clinical imaging is used.<sup>33</sup> However, those studies are often  
25 exploratory in nature. Since quality demands are different (and a bit lower) than in confirmative  
26 studies, those imaging-related studies would perform normally worse but still can contribute  
27  
28 invaluably to their respective field.<sup>34</sup> Finally, we did not include a number of insightful imaging  
29 studies because they did not conduct a formal inter-group comparison.<sup>35,36,37,38</sup>  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46

## 47 **5. Conclusions and Recommendations**

48  
49 Although large animal models offer a offer a number of clear advantages for translational  
50 stroke, we found that they have similar shortcomings to small animal models, limiting this  
51 benefit. Therefore, we derived a number of recommendations to address these limitations but  
52 are, at the same time, relatively easy to implement.  
53  
54  
55  
56  
57  
58  
59  
60

### 5.1 Study planning and preparation

1  
2  
3 Large animal stroke studies are mostly confirmative studies. Therefore, study planning  
4 should be based on high quality standards applied for randomized controlled clinical trials  
5 (RCTs) when possible. Key elements of RCT planning and design such as a priori sample size  
6 calculation and endpoint definition should be conducted.<sup>23</sup> We encourage to involve  
7 statisticians already in early planning steps to optimize study design.<sup>28</sup> Study planning can also  
8 be supported by specific software tools. For instance, the National Centre for the Replacement,  
9 Refinement and Reduction of Animals in Research provides a freeware called Experimental  
10 Design Assistance (<https://eda.nc3rs.org.uk>), which is free to use and was built to guide  
11 researchers through their study planning.<sup>29</sup> Since optimal sample sizes may not be achieved for  
12 all endpoints, it is important to clearly define the most appropriate primary study endpoint, and  
13 to power the study properly. Collaboration between research teams in form of peer quality  
14 checks and validation of study design can highly increase objectivity and validity of a study.<sup>30</sup>  
15 Inter-group collaboration and transfer of experience can also help to handle very complex  
16 models and/or experimental setups, helping to reduce inter-subject variability negatively  
17 affecting statistical power. Confirmative studies might be preregistered to maximize  
18 transparency.<sup>39</sup>  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41

## 42 **5.2 Effect size estimation and pilot trials**

43  
44 Collecting valid information from previous research is essential for reliable effect size  
45 estimation. If such data are not available, pilot studies may be helpful for at least basically  
46 estimating variability of stroke impact and outcome in the model. In case previous experience  
47 with a particular model is low, variability is more likely to be higher and effect size is more  
48 likely to lower in such pilot trials. This will contribute to more conservative study planning  
49 since sample sizes calculated based on that information will be higher.. An important side effect  
50 of pilot trials is experimenter training which limits experimenter-caused endpoint variability  
51 (see below) in the main experiment. In addition, meta-analyses can help to collect relevant  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 information on effect size or regarding a specific research question from related fields.<sup>31</sup>  
4

### 5 **5.3 Reducing the effect of sample size limitations and endpoint variability**

6

7  
8 Financial and logistical restrictions often impact sample and group sizes in large animal  
9  
10 experiments. This is an understandable limitation which is difficult to overcome. Selection of  
11  
12 a proper and relevant endpoint that can be adequately powered with respected to the addressed  
13  
14 research question is therefore important to minimize the risk for low statistical power. Of note,  
15  
16 some endpoints often used in studies assessing therapeutic interventions including infarct size  
17  
18 and functional deficits, exhibit a higher variability in large animal models than in rodent. This  
19  
20 makes comparison of absolute data more difficult.<sup>24</sup> Relative analysis of repeatedly assessed  
21  
22 endpoints, i.e. in comparison to the individual initial infarct size and/or functional deficit can  
23  
24 efficiently compensate for such variability. Repeated assessments also allow calculating the  
25  
26 area under the curve for particular endpoints. This may provide a benefit in statistical power to  
27  
28 identify whether a real outcome benefit is present over time. However, this comes at the cost  
29  
30 of temporal resolution: it cannot be concluded exactly when this benefit became evident. There  
31  
32 is also preliminary evidence for fast and slow stroke progressors in large animals, indicating  
33  
34 different collateral status and somewhat resembling the human situation, but further  
35  
36 contributing to inter-subject variability. It is recommended to consider this fact when planning  
37  
38 an acute stroke study.<sup>32</sup>  
39  
40  
41  
42  
43

44  
45 In experiments of highly similar design, controls may be pooled. Of note, this counteracts  
46  
47 randomization and therefore requires extremely thorough validation of comparability of control  
48  
49 subjects from different experiments/sources. If comparability is thoroughly proven, this may  
50  
51 help to increase statistical power, but the limitations of this approach and potentially resulting  
52  
53 bias need to be discussed transparently and in detail when publishing results.  
54

55  
56 The possibility to repeatedly collect a broad spectrum of physiological data should be  
57  
58 utilized where possible, as deviation from normal parameter ranges may explain variability and  
59  
60 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>

Only

## 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. Kilkenney C, Browne WJ, Cuthill IC, et al. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol* 2010; 8: e1000412.



- 1  
2  
3 12. Jovin TG, Albers, GW, Liebeskind DS, et al. Stroke Treatment Academic Industry  
4  
5 Roundtable: The Next Generation of Endovascular Trials. *Stroke* 2016; 47(19): 2656-  
6  
7 2665.  
8  
9  
10 13. Goyal M, Menon BK, van Zwam WH, et al. Endovascular thrombectomy after large-  
11  
12 vessel ischaemic stroke: a meta-analysis of individual patient data from five  
13  
14 randomized trials. *Lancet* 2016; 387(10029): 1723-1731.  
15  
16  
17 14. Sena ES, Currie GL, McCann SK, et al. Systematic reviews and meta-analysis of  
18  
19 preclinical studies: why perform them and how to appraise them critically. *J Cereb*  
20  
21 *Blood Flow Metab* 2014; 34: 737-742.  
22  
23  
24 15. Sena ES, Van der Worp HB, Bath PM, et al. Publication bias in reports of animal  
25  
26 stroke studies leads to major overstatement of efficacy. *PLoS Biol* 2010; 8: e1000344.  
27  
28  
29 16. Macleod MR, Van der Worp HB, Sena ES, et al. Evidence for the efficacy of NXY-  
30  
31 059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke*  
32  
33 2008; 39: 2824-2829.  
34  
35  
36 17. Boltze J, Nitzsche F, Jolkkonen J, et al. Concise Review: Increasing the Validity of  
37  
38 Cerebrovascular Disease Models and Experimental Methods for Translational Stem  
39  
40 Cell Research. *Stem Cells* 2017; 35: 1141-1153.  
41  
42  
43 18. Sommer CJ. Ischemic stroke: experimental models and reality. *Acta Neuropathol*  
44  
45 2017; 133: 245-261.  
46  
47  
48 19. Mehra M, Henninger N, Hirsch JA et al. Preclinical acute ischemic stroke modeling. *J*  
49  
50 *Neurointerv Surg* 2012; 4: pp. 307-3013.  
51  
52  
53 20. Helke KL and Swindle MM. Animal models of toxicology testing: the role of pigs.  
54  
55 *Expert Opin Drug Metab Toxicol* 2013; 9: 127-139.  
56  
57  
58 21. Herrmann AM, Cattaneo GFM, Eiden, SA, et al. Development of a Routinely  
59  
60 Applicable Imaging Protocol for Fast and Precise Middle Cerebral Artery Occlusion

- 1  
2  
3 Assessment and Perfusion Deficit Measure in an Ovine Stroke Model: A Case Study.  
4  
5 *Front Neurol.* 2019; 10: 1113.  
6  
7  
8 22. Minnerup J, Zentsch V, Schmidt A, et al. Methodological Quality of Experimental  
9  
10 Stroke Studies Published in the Stroke Journal: Time Trends and Effect of the Basic  
11  
12 Science Checklist. *Stroke* 2016; 47: 267-272.  
13  
14 23. Dirnagl U. Bench to bedside: the quest for quality in experimental stroke research. *J*  
15  
16 *Cereb Blood Flow Metab* 2006; 26: 1465-1478.  
17  
18  
19 24. Boltze J, Modo MM, Mays RW, et al. Stem Cells as an Emerging Paradigm in Stroke  
20  
21 4: Advancing and Accelerating Preclinical Research. *Stroke* 2019; 50 (11): 3299-  
22  
23 3306.  
24  
25  
26 25. Minnerup J, Wersching H and Diederich K. Methodological quality of preclinical  
27  
28 stroke studies is not required for publication in high-impact journals. *J Cereb Blood*  
29  
30 *Flow Metab* 2010; 30: 1619-1624.  
31  
32  
33 26. Savitz SI, Baron JC, Fisher M, et al. Stroke Treatment Academic Industry Roundtable  
34  
35 X: Brain Cyoprotection Therapies in the Reperfusion Era. *Stroke* 2019. 50(4): 1026-  
36  
37 1031.  
38  
39  
40 27. Minnerup J, Wersching H and Diederich K. Methodological quality of preclinical  
41  
42 stroke studies is not required for publication in high-impact journals. *J Cereb Blood*  
43  
44 *Flow Metab* 2010; 30: 1619-1624.  
45  
46  
47 28. Würbel H. More than 3Rs: the importance of scientific validity for harm-benefit  
48  
49 analysis of animal research. *Lab Anim (NY)* 2017; 46: 164-166.  
50  
51  
52 29. Percie du Sert N, Bamsey I, Bate ST, et al. The Experimental Design Assistant. *PLoS*  
53  
54 *Biol* 2017; 15: e2003779.  
55  
56  
57 30. Sena ES, Currie GL, McCann SK, et al. Systematic reviews and meta-analysis of  
58  
59 preclinical studies: why perform them and how to appraise them critically. *J Cereb*  
60  
*Blood Flow Metab* 2014; 34: 737-742.

- 1  
2  
3 31. Begley CG and Ioannidis JP. Reproducibility in science: improving the standard for  
4  
5 basic and preclinical research. *Circ Res* 2015; 116: 116-126.  
6  
7  
8 32. Shazeeb MS, King RM, Brooks OW, et al. Infarct Evolution in a Large Animal Model  
9  
10 of Middle Cerebral Artery Occlusion. *Transl. Stroke Research* 2019. doi:  
11  
12 10.1007/s12975-019-00732-9.  
13  
14 33. Werner P, Saur D, Zeisig V, et al. Simultaneous PET/MRI in stroke: a case series. *J*  
15  
16 *Cereb Blood Flow* 2015; 35(9): 1421-1425.  
17  
18  
19 34. Dirnagl U, Hakim A, Macleod M, et al. A concerted appeal for international  
20  
21 cooperation in preclinical stroke research. *Stroke* 2013; 44(6): 1754-1760.  
22  
23  
24 35. Boltze J, Ferrara F, Hainsworth AH, et al. Lesional and perilesional tissue  
25  
26 characterization by automated image processing in a novel gyrencephalic animal  
27  
28 model of peracute intracerebral hemorrhage. *J Cereb Blood Flow Metab* 2019; 39(12):  
29  
30 2521-2535.  
31  
32  
33 36. Haque ME, Gabr RE, Zhao X, et al. Serial quantitative neuroimaging of iron in the  
34  
35 intracerebral hemorrhage pig model. *J Cereb Blood Flow Metab* 2018; 38(3): 375-  
36  
37 381.  
38  
39  
40 37. Kamimura HA, Flament J, Valette J, et al. Feedback control of microbubble cavitation  
41  
42 for ultrasound-mediated blood-brain barrier disruption in non-human primates under  
43  
44 magnetic resonance guidance. *J Cereb Blood Flow Metab* 2019; 39(7): 1191-1203.  
45  
46  
47 38. Sander CY, Mandeville JB, Wey HY, et al. Effects of flow changes on radiotracer  
48  
49 binding: Simultaneous measurement of neuroreceptor binding and cerebral blood flow  
50  
51 modulation. *J Cereb Blood Flow Metab* 2019; 39(1): 131-146.  
52  
53  
54 39. Kimmelman J and Anderson JA. Should preclinical studies be registered? *Nat*  
55  
56 *Biotechnol* 2012; 30(6): 488-489.  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Confidential: For Review Only

## 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.**

<b>Category 1: Reporting of study subject details and welfare</b>		<b>Category 2: Study planning quality</b>	
<i>Item</i>	<i>Score point</i>	<i>Item</i>	<i>Score point</i>
	<i>allocation</i>		<i>allocation</i>
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

<b>Category 3: Internal study validity</b>		<b>Category 4: Outcome analysis and reporting</b>	
<i>Item</i>	<i>Score point</i>	<i>Item</i>	<i>Score point</i>
	<i>allocation</i>		<i>allocation</i>
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 <sup>#</sup> 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 <sup>##</sup>	Provided yes=1/no=0

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

<sup>##</sup>conclusion was considered justified when supported by correctly analyzed results.

**Table 2. Basic Characteristics of included Animal Experimental Studies.**

<b>Item</b>	<b>Frequency (%)</b>	<b>Item</b>	<b>Frequency (%)</b>	<b>Item</b>	<b>Frequency (%)</b>
<b>Species</b>		<b>Type of intervention</b>		<b>Study duration</b>	
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%)	<1 month	n=14 (6.7%)
Pig	n=19 (9.1%)	Others <sup>#</sup>	n=21 (10.1%)	>1 month	n=14 (6.7%)
Non-Human-Primate & Rabbit	n=1 (0.5%)				
Sheep	n=1 (0.5%)				
<b>Region</b>		<b>Primary endpoint</b>		<b>Stroke model</b>	
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%)
<b>Further information</b>	
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 <sup>##</sup> reported	n=23 (11.1%)

<sup>#</sup>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.

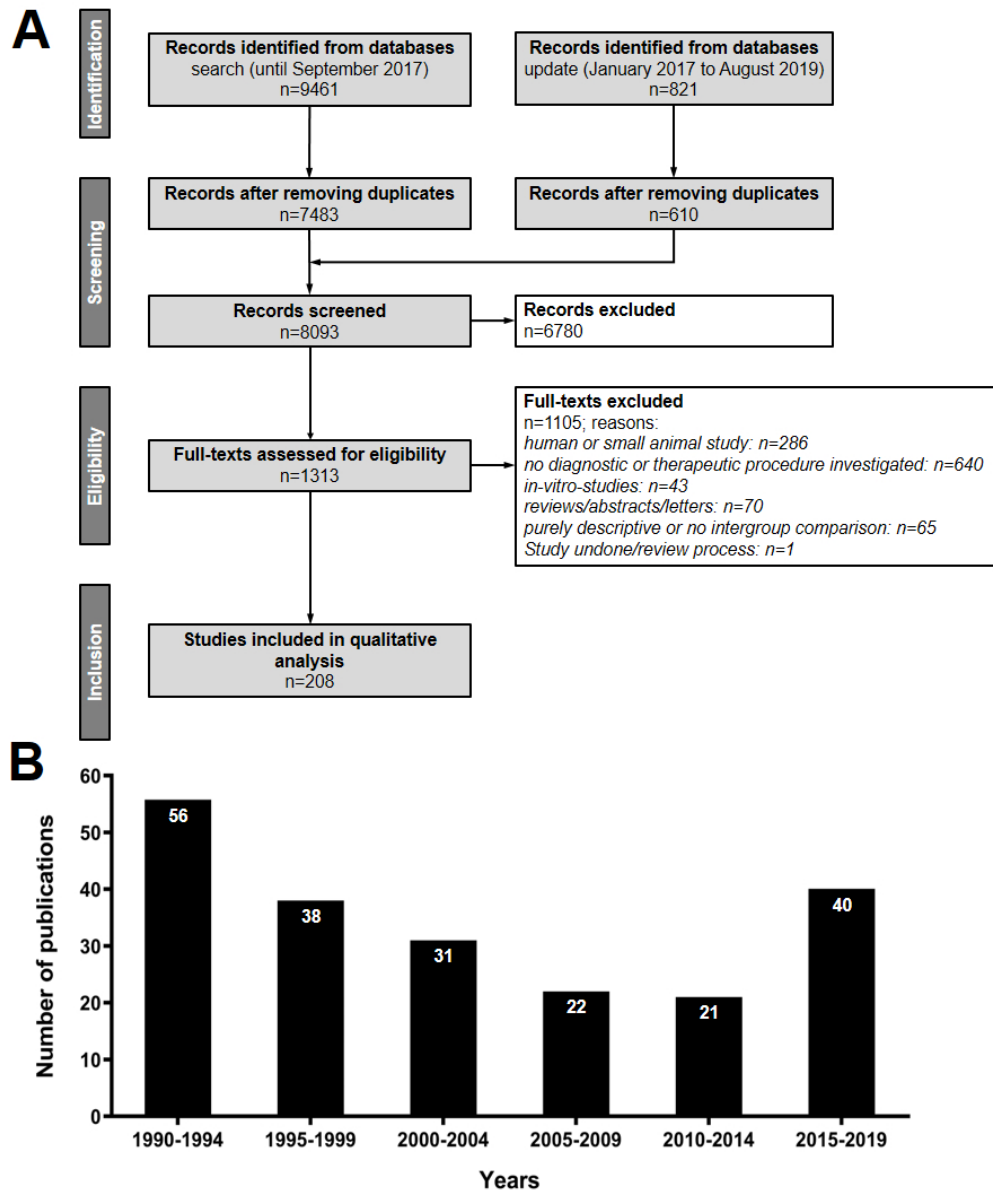


Figure 1. Overview on quantitative search results and frequency of large animal experiments in stroke research since 1990.

(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)

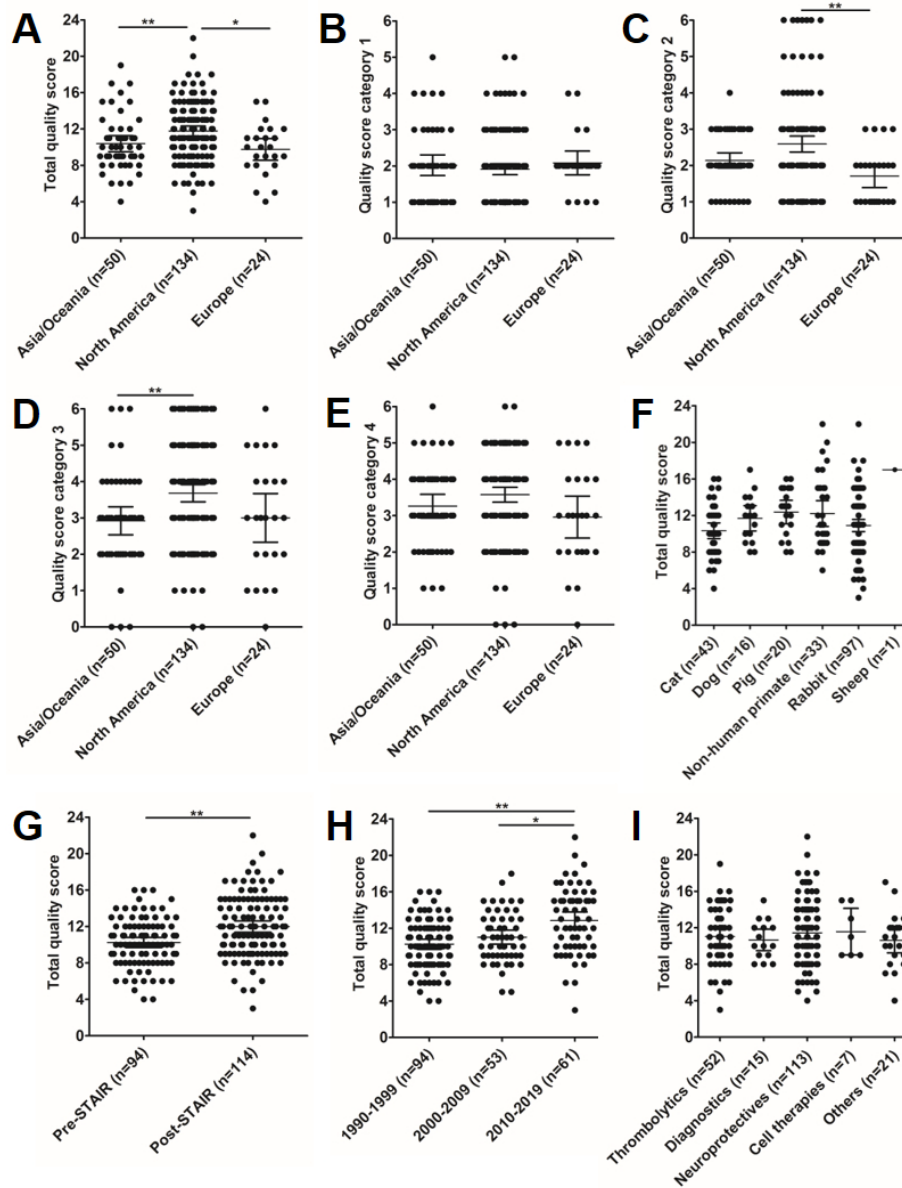


Figure 2. Influence of study origin and STAIR criteria publication on study quality.

(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. \* $p<0.05$ ; \*\* $p<0.01$ .

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

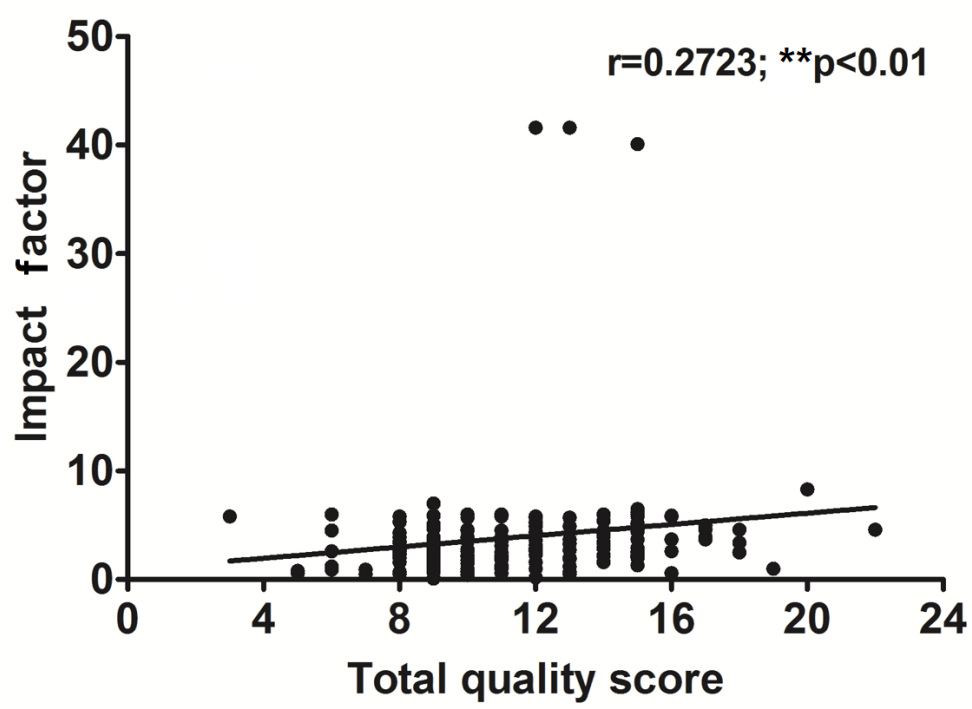


Figure 3. Association between total quality score versus impact factor. Scatterplot shows correlation between quality score and IF ( $p<0.01$ ). Number of included studies is 172, no IF could be retrieved for 36 studies. The latter studies were excluded from this analysis.

304x222mm (96 x 96 DPI)

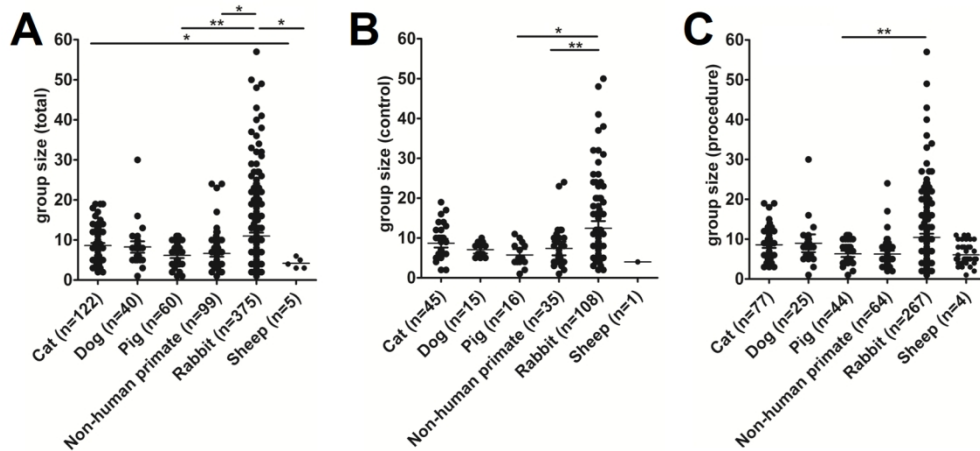


Figure 4. Group sizes across species.

(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)