


# Why Most Acute Stroke Studies Are Positive in Animals but Not in Patients: A Systematic Comparison of Preclinical, Early Phase, and Phase 3 Clinical Trials of Neuroprotective Agents

Antje Schmidt-Pogoda, MD <sup>1,†</sup> Nadine Bonberg, MSc,<sup>2,†</sup> Mailin Hannah Marie Koecke,<sup>1</sup> Jan-Kolja Strecker, PhD,<sup>1</sup> Jürgen Wellmann, PhD,<sup>2</sup> Nils-Martin Bruckmann, MD,<sup>1</sup> Carolin Beuker, MD,<sup>1</sup> Wolf-Rüdiger Schäbitz, MD,<sup>3</sup> Sven G. Meuth, MD, PhD,<sup>1</sup> Heinz Wiendl, MD,<sup>1</sup> Heike Minnerup, MD, MSc,<sup>2</sup> and Jens Minnerup, MD<sup>1</sup>

**Objective:** To analyze why numerous acute stroke treatments were successful in the laboratory but failed in large clinical trials.

**Methods:** We searched all phase 3 trials of medical treatments for acute ischemic stroke and corresponding early clinical and experimental studies. We compared the overall efficacy and assessed the impact of publication bias and study design on the efficacy. Furthermore, we estimated power and true report probability of experimental studies.

**Results:** We identified 50 phase 3 trials with 46,008 subjects, 75 early clinical trials with 12,391 subjects, and 209 experimental studies with >7,141 subjects. Three (6%) phase 3, 24 (32%) early clinical, and 143 (69.08%) experimental studies were positive. The mean treatment effect was 0.76 (95% confidence interval [CI] = 0.70–0.83) in experimental studies, 0.87 (95% CI = 0.71–1.06) in early clinical trials, and 1.00 (95% CI = 0.95–1.06) in phase 3 trials. Funnel plot asymmetry and trim-and-fill revealed a clear publication bias in experimental studies and early clinical trials. Study design and adherence to quality criteria had a considerable impact on estimated effect sizes. The mean power of experimental studies was 17%. Assuming a bias of 30% and pre-study odds of 0.5 to 0.7, this leads to a true report probability of <50%.

**Interpretation:** Pivotal study design differences between experimental studies and clinical trials, including different primary end points and time to treatment, publication bias, neglected quality criteria and low power, contribute to the stepwise efficacy decline of stroke treatments from experimental studies to phase 3 clinical trials. Even under conservative estimates, less than half of published positive experimental stroke studies are truly positive.

ANN NEUROL 2020;87:40–51

Stroke is a leading cause of death and disability.<sup>1</sup> Nonetheless, intravenous thrombolysis with tissue plasminogen activator (tPA) and thrombectomy are still the only approved treatments for acute ischemic stroke,<sup>2</sup> and stroke

researchers are desperate to identify novel treatment strategies beyond recanalization, i.e. neuroprotective treatments. In the last few decades, more than 1,000 acute stroke treatments have been tested in the laboratory.<sup>3</sup> Strikingly, the

View this article online at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/ana.25643). DOI: 10.1002/ana.25643

Received Mar 20, 2019, and in revised form Nov 11, 2019. Accepted for publication Nov 11, 2019.

Address correspondence to Dr Schmidt-Pogoda, Albert-Schweitzer-Campus 1, Gebäude A1, 48149 Münster, Germany.

E-mail: [antje.schmidt-pogoda@ukmuenster.de](mailto:antje.schmidt-pogoda@ukmuenster.de)

<sup>†</sup>A.S.-P. and N.B. contributed equally.

From the <sup>1</sup>Department of Neurology, Institute of Translational Neurology, University of Münster, Münster; <sup>2</sup>Institute of Epidemiology and Social Medicine, University of Münster, Münster; and <sup>3</sup>Department of Neurology, Evangelical Hospital Bethel, Bielefeld, Germany

Additional supporting information can be found in the online version of this article.

majority of acute stroke treatments achieved great success in animal studies but eventually failed in large clinical trials.<sup>3</sup> As a consequence, the perception that “everything works in animals but nothing works in people” dominates stroke research, and numerous commentaries and review articles on the so called “translational roadblock” between animal studies and clinical trials have been published.<sup>3–5</sup> Strictly speaking, we do not observe a single roadblock between animal experiments and human stroke patients but rather a stepwise efficacy decline from highly successful animal studies to partly successful early clinical trials to neutral phase 3 clinical trials. Here, we systematically analyzed differences in the study design of experimental studies, early phase clinical trials, and phase 3 clinical trials, and we evaluated in how far these differences led to overestimated efficacy of acute stroke treatments in experimental stroke studies and early clinical trials.

## Methods

### Identification of Studies and Data Extraction

We systematically searched the databases Clinicaltrials.gov (searched in August 2017) and the Internet Stroke Center (searched in August 2017) to identify all phase 3 clinical trials of medical treatments for acute ischemic stroke, i.e. candidate neuroprotectants. Our search of the database Clinicaltrials.gov included the terms “closed studies” for “recruitment,” “all studies” for “study result,” “interventional studies” for “study type,” and “phase III” for “phase.” Our search of the Internet Stroke Center involved the terms “completed” and “terminated” for “status,” and “phase III” for “phase.” Because we aimed to focus on acute treatments, trials were only included if the intervention started within the first 72 hours from symptom onset. To determine corresponding experimental studies and early phase clinical trials of identified phase 3 trials, we searched Pubmed from the beginning until August 2017 using the terms “stroke” or “ischemia” or “infarct” and “drug name” or “abbreviated drug name.” We included only studies reporting functional outcomes and/or infarct volumes. Studies using hemorrhagic stroke models, electrolytic brain lesions and pretreatment, and treatment arms involving comedication other than recombinant tPA (rtPA) were excluded. An overview of inclusion and exclusion criteria is provided in the supplementary material (Supplementary Table). We extracted data on the number of subjects in each treatment arm, functional outcomes, infarct volumes, study quality, time to treatment, and statistical methods. When data were presented graphically only, values were read off the graphics using ImageJ (public domain software). Study identification and data extraction were performed by 2 independent investigators (A.S.-P. and M.K. identified all studies, A.S.-P. and

M.K. extracted data from experimental studies, A.S.-P. and N.-M.B. extracted data from early clinical trials and phase 3 clinical trials). Disagreements were resolved in discussion with a third investigator (C.B.).

### Statistical Methods

We performed meta-analyses to illustrate the gradual efficacy decline of acute stroke treatments from experimental studies to early clinical trials to phase 3 clinical trials. For continuous outcomes like infarct volumes, we extracted mean values and their standard errors (SEs) for each treatment group and control group. If not provided directly, the SE was computed by dividing the standard deviation by the square root of the number of animals per group. If only median and quantiles were reported, we used the median as location parameter and estimated the SE of the mean with a scaled interquartile range due to normality assumption. For infarct volumes and functional outcomes in experimental studies, we quantified treatment effects by the ratios of the outcomes of treatment groups and corresponding control groups, and we approximated the corresponding SEs by the delta method.<sup>6</sup> We used the ratio of means instead of the difference of means for the following reasons: First, animal experiments were performed in rats, mice, and rabbits, which all have different brain sizes. Second, we included studies with different infarct models, for example, middle cerebral artery occlusion and photothrombosis, which produce different infarct sizes. Third, infarct sizes were provided in different units (mm<sup>3</sup> and percent). Using the ratio of means, units and multiplication factors can be cancelled so that these studies can be compared, independent of the animal strain, infarct model, and the reported unit for infarct size. Another advantage is that response ratios are multiplicative values indicating the proportion of brain that can be saved by a certain treatment. For studies reporting binary outcomes—for example, the proportion of patients achieving a modified Rankin scale of 0 to 1 after 3 months—we quantified treatment effects by odds ratios and their confidence intervals (CIs). If more than one verum group was available, verum groups were combined. To this end, means of continuous outcomes were weighted with inverse variance, and means of proportions were weighted with group size.

Our meta-analyses included all experimental studies in which mean infarct volumes or means of functional outcomes and corresponding SEs were either denoted directly or could be determined as described previously and all early clinical trials and phase 3 clinical trials that provided binary functional outcomes. Calculations were performed with R version 3.5.2. Meta-analyses were performed with the metafor package using random-effects models, as previously described.<sup>7</sup> To analyze how much time to treatment and adherence to quality criteria such as

TABLE Overview of Included Studies

Treatment	Phase 3 Trials, n	Subjects in Phase 3 Trials, n	Early Clinical Trials, n	Subjects in Early Clinical Trials, n	Experimental Studies, n	Subjects in Experimental Studies, n
Acetaminophen	1	1,400	5	458	0	0
Albumin	1	841	2	346	17	462
Aptiganel	1	628	2	659	7	165
Atenolol	1	201	0	0	0	0
BMS-204352	1	1,977	0	0	1	108
Candesartan	1	2,029	1	342	18	500
Cervene	1	368	2	366	0	0
Chlormethiazole	2	2,558	2	1,550	4	60
Citicoline	3	3,591	6	934	18	512
Diazepam	1	880	0	0	1	37
DP-b99	1	446	1	150	0	0
Ebselen	1	302	1	105	5	225
Edaravone	1	814	9	1,429	20	253
Eliprodil	1	483	0	0	3	179
Enlimomab	1	635	0	32	7	262
Epoetin alfa	1	522	3	262	33	>1,372
FGF	1	286	0	0	16	655
Flunarizine	1	331	2	357	1	74
Fosphenytoin	1	462	0	0	0	0
Gavestinel	2	3,171	4	474	2	84
Ginsenoside-Rd	1	386	1	199	0	0
GM1 ganglioside	3	1,581	4	727	4	93
Isradipine	1	357	0	0	3	61
Lubeluzole	2	2,507	4	1,029	3	87
Magnesium	2	4,086	5	296	3	112
Naftidrofuryl	1	620	1	100	2	>16
NeuroAiD	1	1,099	1	150	0	0
Nimodipine	4	3,028	12	1,769	10	200
NXY-059	2	5,028	1	134	10	540
ONO-2506	1	841	1	92	1	108
Piracetam	1	927	0	0	2	50
Repinotan	1	681	1	240	1	316
Selfotel	1	567	1	32	2	37
Tirilazad	2	682	1	111	8	137
Uric acid	1	411	2	48	0	0
UK-279,276	1	966	0	0	0	0
YM872	1	312	0	0	7	436
Total	50	46,008	75	12,391	209	>7,141

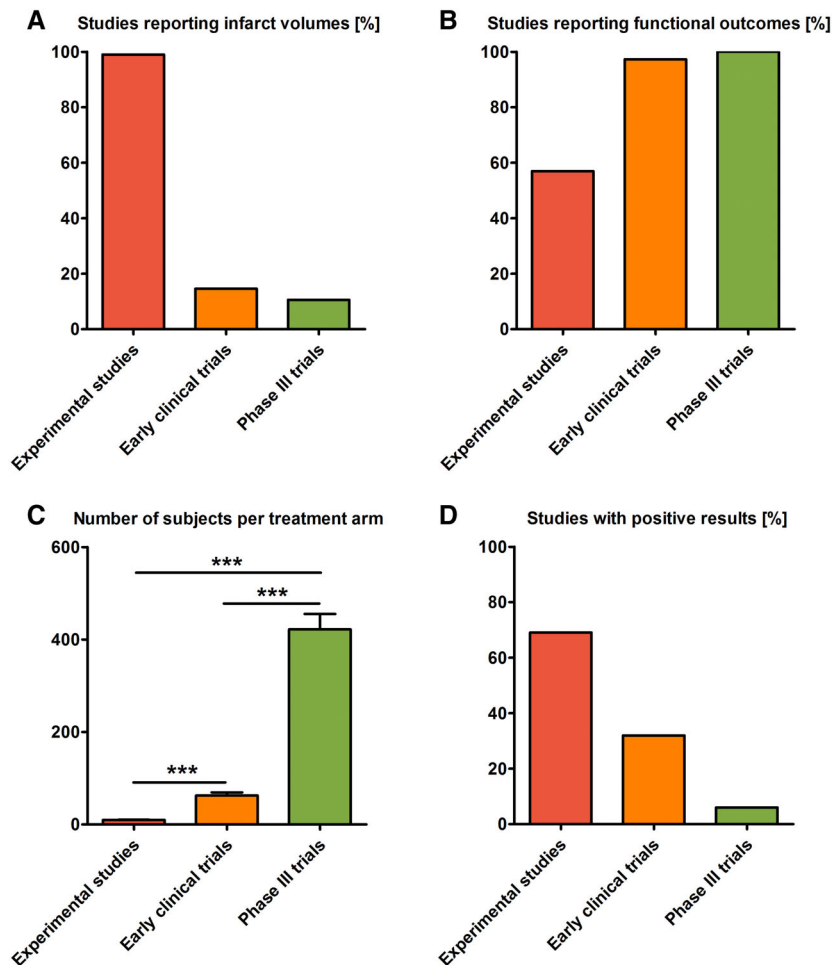
FGF = fibroblast growth factor.

randomization, blinded outcome assessment, and the inclusion of subjects with comorbidities influence the treatment effect, we added these variables as influential variables in random-effects models.<sup>7</sup>

The publication bias in experimental studies, early clinical trials, and phase 3 clinical trials was estimated by funnel plot asymmetry and a trim-and-fill approach, as previously described.<sup>8</sup> To prepare funnel plots, we plotted the effect estimates from individual studies against each study's precision, that is, the reciprocal value of the SE. We chose a reversed vertical scale that places larger, more powerful studies toward the top. In consequence, effect estimates from smaller studies should scatter more widely at the bottom, whereas the scatter of larger studies narrows toward the top.<sup>9</sup> The concept of funnel plotting is based on the assumption that imprecise studies are as likely to overstate efficacy as to understate efficacy, thus resulting in a plot that

resembles an inverted funnel. If publication bias is present, there is an imbalance in favor of imprecise studies which overstate efficacy, whereas the counterpart of imprecise studies that understate efficacy is missing.<sup>8</sup> The trim-and-fill method is a nonparametric data augmentation technique proposed by Duval and Tweedie.<sup>10</sup> In the trim-and-fill approach, studies with the most extreme effect sizes are suppressed, and the overall estimate of efficacy is recalculated. This process is repeated until there are no more studies to exclude. Thereafter, suppressed studies are replaced together with imputed counterparts, which are determined by reflection around the recalculated overall estimate of efficacy.<sup>10</sup> Meta-analysis of this data set gives an estimate of the actual efficacy, and the number of imputed counterparts gives an idea of the number of unpublished studies.<sup>10</sup>

Power analyses for experimental studies were performed with the pwr package.<sup>11</sup> Power analyses were



**FIGURE 1:** General study characteristics. (A) Almost all experimental studies (99.03%) presented infarct volumes as main end point, whereas only 14.66% of early clinical and 10.64% of phase 3 clinical trials reported infarct volumes. (B) All phase 3 trials and almost all early clinical trials (97.33%) presented functional outcomes as main end point, whereas only 57% of experimental studies provided functional outcomes. (C) There were highly significant differences (triple asterisks) in the mean number of subjects per treatment arm ( $p < 0.001$ , one-way analysis of variance). (D) There were 69.08% of experimental studies, 32% of early clinical, and 6% of phase 3 clinical trials that reported at least 1 significant positive major end point, that is, either improved functional outcome or infarct volume reduction.

based on a 2-sample (2-sided)  $t$  test, and we assumed that the logarithms of infarct volumes in placebo groups  $\ln(V_p)$  and verum groups  $\ln(V_v)$  were distributed normally. As expectable effect, we chose the estimator of the meta-analysis  $\ln(\text{response ratio [RR]}) = \ln(0.76) = \ln(V_v) - \ln(V_p)$  from experimental studies and assumed homogeneity of variance. As an estimator for variance, we calculated a with the sample size weighted mean of variances in experimental studies and determined Cohen's  $d = -0.47$ .

The true report probability (trp) of experimental studies was estimated by the previously defined formula  $\text{trp} = ([1 - \beta]R + u\beta R)/(R + \alpha - \beta R + u - u\alpha + u\beta R)$ .<sup>12</sup> In this formula,  $R$  is the prestudy odds of true  $H_0$  to  $H_1$  hypotheses,  $U$  is the proportion of probed analyses that would not have been “research findings” but ended up being reported as such because of bias,  $(1 - \beta)$  is the study power, and  $\alpha$  is the type 1 error rate.<sup>10</sup>

## Results

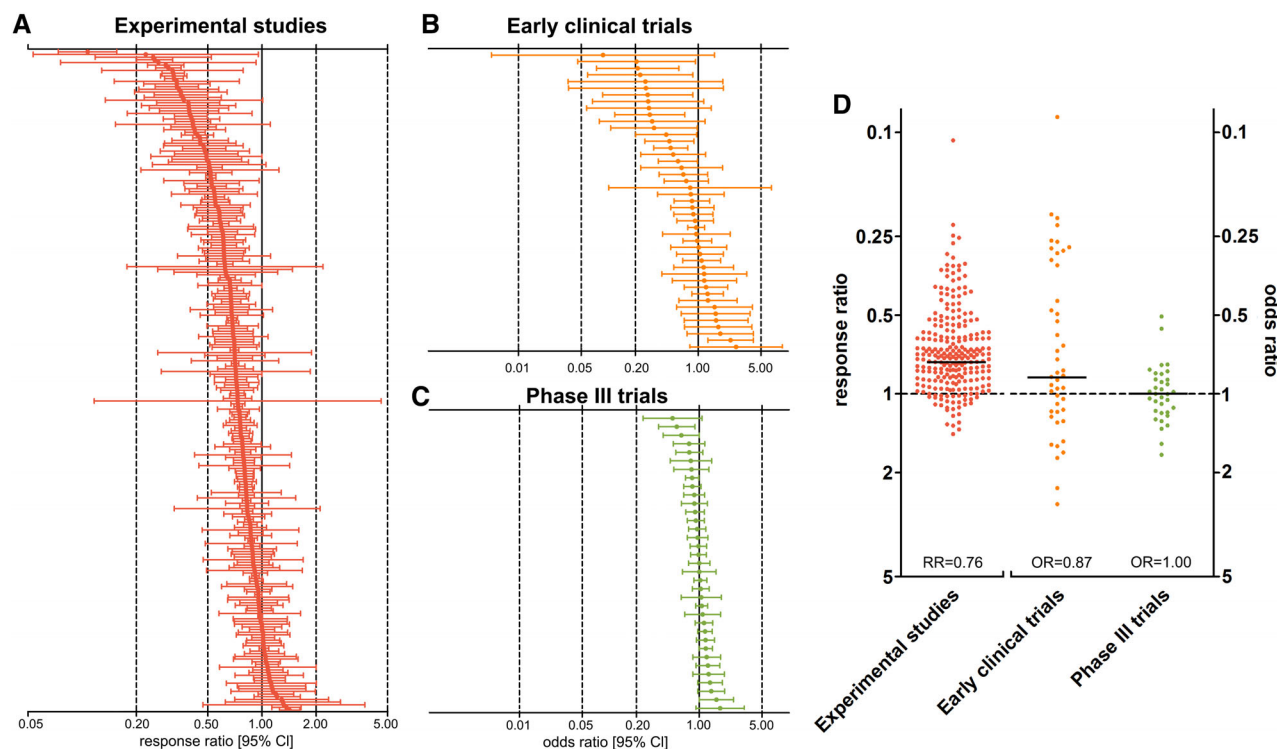
### Included Studies and General Study Characteristics

Our search identified 50 phase 3 clinical trials on 37 treatments with 46,008 patients; 209 corresponding animal

experimental studies with more than 7,141 subjects; and 75 corresponding early clinical trials with 12,391 patients (Table). Vote-count meta-analyses revealed intriguing differences in general study characteristics between experimental studies, early clinical trials, and phase 3 clinical trials: 99.03% of experimental studies provided infarct volumes as main end points, whereas most clinical studies reported only neurological outcomes (Fig 1). Mean numbers of subjects per treatment arm differed significantly between animal studies ( $9.96 \pm 0.21$ ), early clinical trials ( $62.79 \pm 6.65$ ), and phase 3 clinical trials ( $422.10 \pm 33.30$ ;  $p < 0.0001$ , one-way analysis of variance). Most important, 143 (69.08%) experimental studies reported at least 1 significant positive major outcome, that is, either infarct volume reduction or improved functional outcome, whereas only 24 (32.0%) early clinical studies and 3 (6%) phase 3 clinical trials presented a significant positive outcome.

### Efficacy of Acute Stroke Treatments

For infarct volumes, the overall response ratio of acute stroke treatments in experimental studies was 0.76 (95% CI = 0.70–0.83; Fig 2A), indicating that acute stroke



**FIGURE 2:** Treatment efficacy. (A–C) Comparisons are ranked according to the effect on infarct volumes in experimental studies and according to the effect on functional outcomes in early clinical trials and phase 3 clinical trials. Horizontal error bars indicate 95% confidence interval (CI). (D) Scatter plots show the treatment effect of acute stroke studies for individual studies, and black lines indicate the overall treatment effect. The overall treatment effect was 0.76 (95% CI = 0.70–0.83) in experimental studies, 0.87 (95% CI = 0.71–1.06) in early clinical, and 1.00 (95% CI = 0.95–1.06) in phase 3 clinical trials. Note that response ratios are used to compare treatment effects on infarct volumes in experimental studies, whereas odds ratios are used to compare treatment effects on functional outcomes in clinical trials.

treatments reduced stroke severity by 24% on average. For functional outcomes, the overall response ratio of acute stroke treatments in experimental studies was 0.78 (95% CI = 0.67–0.91; Supplementary Fig 2).

In early clinical trials, the overall odds ratio was 0.87 (95% CI = 0.71–1.06; see Fig 2B), meaning that acute stroke treatments improved outcome by 13% on average. In phase 3 clinical trials, the overall odds ratio was 1.00 (95% CI = 0.95–1.06; see Fig 2C).

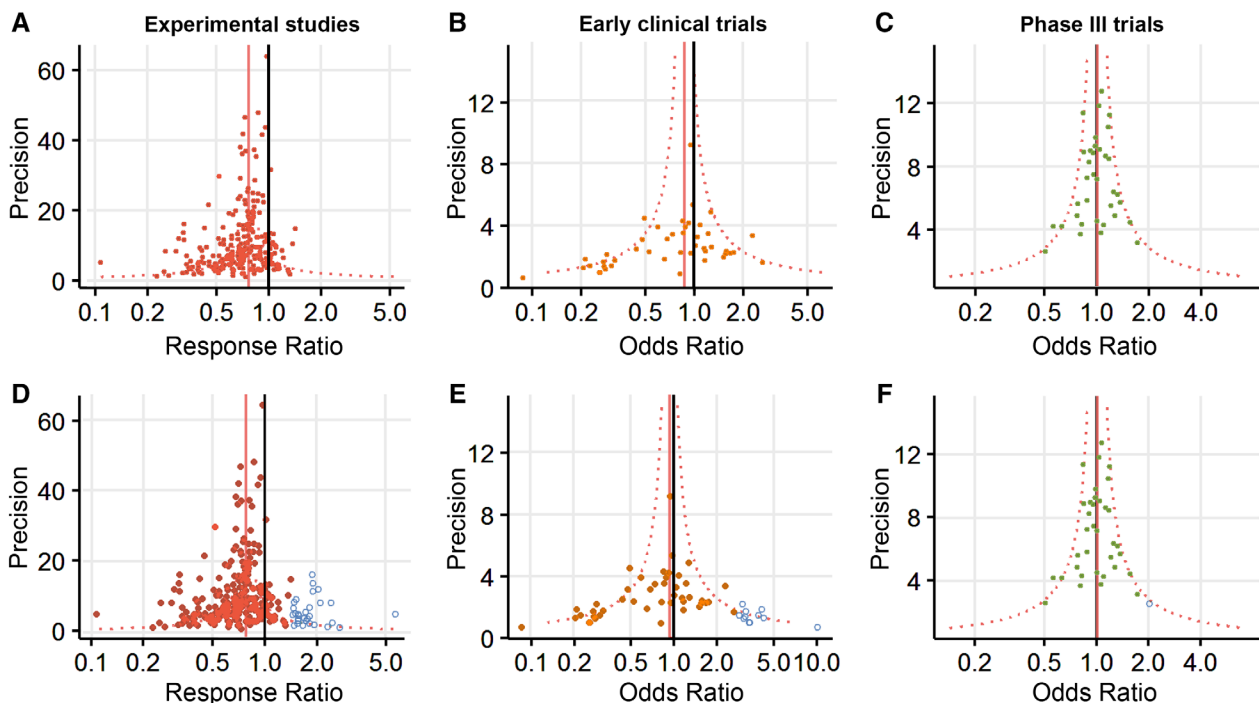
Altogether, our findings demonstrate a stepwise efficacy decline of acute stroke treatments from experimental to early clinical trials to phase 3 clinical trials (see Fig 2D). We next aimed to identify contributing factors explaining this observed stepwise efficacy decline. We used infarct volume reduction in experimental studies and reduction of neurological deficits in clinical trials to determine treatment efficacy because the majority of experimental studies provided infarct volumes as main outcomes, whereas most clinical studies reported only neurological outcomes.

### Publication Bias

The presence of publication bias is likely to contribute to an overestimation of treatment efficacy in meta-analyses and systematic reviews.<sup>8</sup> Publication bias describes the phenomenon that positive results are more likely to be published, whereas negative data may either remain unpublished

or be presented in such a way that they become positive.<sup>9</sup> We analyzed the prevalence of publication bias by funnel plot asymmetry and a trim-and-fill approach.<sup>8</sup> In brief, the concepts of funnel plotting and trim and fill are based on the assumption that imprecise studies are as likely to overstate efficacy as to understate efficacy. If publication bias is present, there is an imbalance in favor of imprecise studies that overstate efficacy, whereas the counterpart of imprecise studies that understate efficacy is missing. Graphically, publication bias results in funnel plot asymmetry.<sup>8</sup>

Visual inspection of funnel plots showed that only the plot of phase 3 clinical trials had an inverted funnel shape (Fig 3A–C), indicating an absence of publication bias in phase 3 clinical trials. In experimental studies, there was a clear imbalance, with abundant imprecise studies with large effect sizes, suggesting a high prevalence of publication bias. In early clinical trials, there was a moderate imbalance, with a modest number of imprecise studies with large effect sizes, thus indicating a moderate prevalence of publication bias. In the trim-and-fill approach, we imputed missing counterparts of imprecise studies to get an idea of the number of unpublished experiments (see Fig 3D–F). Trim-and-fill method suggested 30 missing experimental studies, 11 missing early clinical trials, and 1 missing phase 3 clinical trial. In addition, we recalculated the overall efficacy of the new data set to get



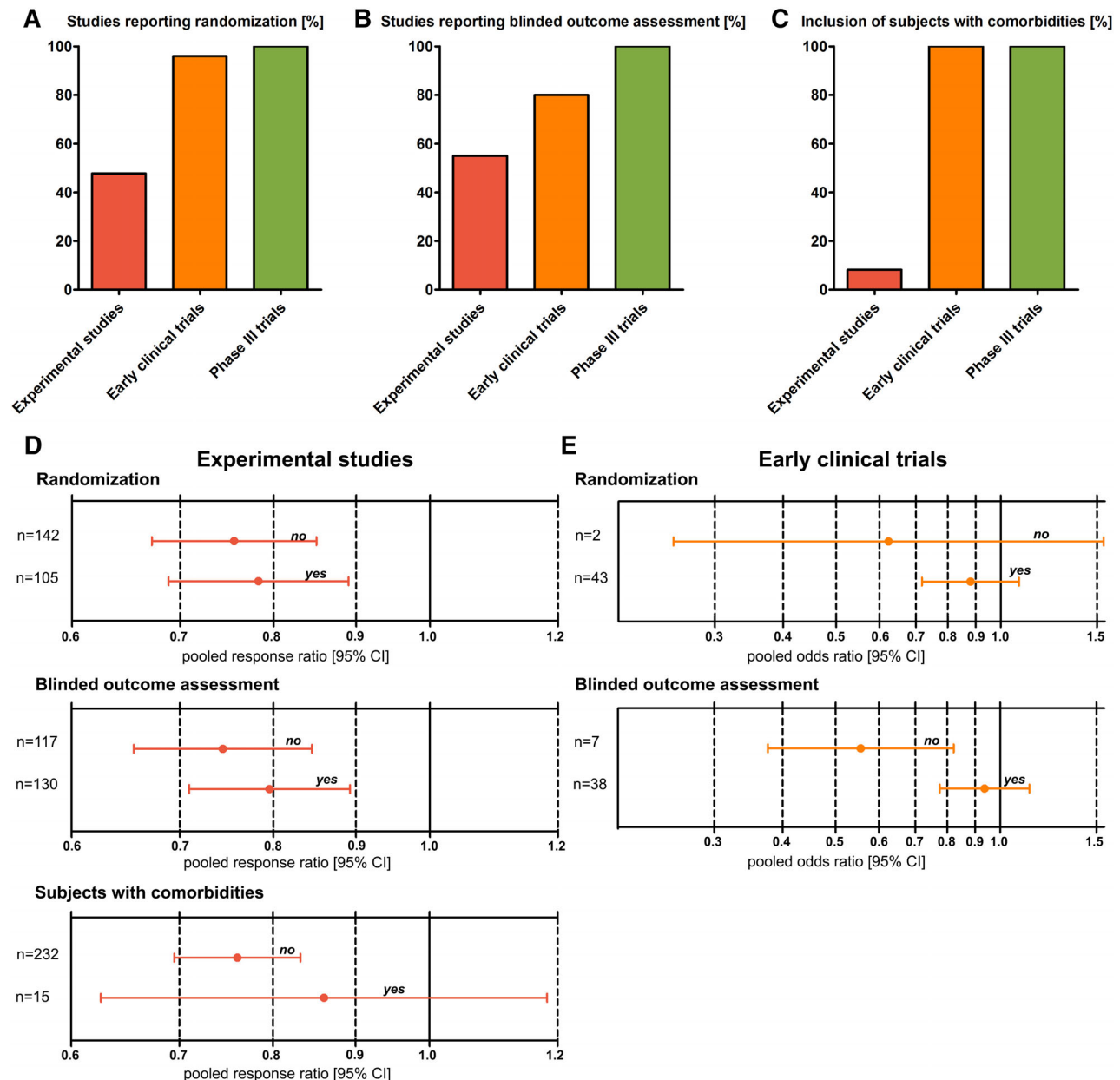
**FIGURE 3:** Publication bias. (A–C) Funnel plots with precision, that is, the reciprocal value of the standard error, plotted against the effect size. In the absence of publication bias, the plots should resemble an inverted funnel. Visual inspection of funnel plot asymmetry indicates the presence of publication bias in experimental studies and early clinical trials. (D–F) In the trim-and-fill approach, missing counterparts of imprecise studies were added in green to the data set of (A–C) to get an estimate of the number of unpublished studies.



an estimate of the actual efficacy in the absence of publication bias. The recalculated overall response ratio in experimental studies was 0.78 (95% CI = 0.70–0.87), the recalculated overall odds ratio in early clinical trials was 0.93 (95% CI = 0.7–1.19), and the recalculated overall odds ratio in phase 3 clinical trials was 1.01 (95% CI = 0.95–1.07). To verify our estimation of publication bias, we additionally determined publication bias in clinical trials by comparing trials that are registered with publications within 3 years following completion, similar to a

previously reported method.<sup>13</sup> Following this approach, we identified 12 missing early clinical trials and no missing phase 3 trials. These results are in accordance with our estimation of publication bias by funnel plot asymmetry and trim and fill, which suggested 11 missing early clinical trials and 1 missing phase 3 trial.

Altogether, our findings identify publication bias as an important contributing factor that leads to overstated efficacy of stroke treatments in experimental studies and early clinical trials.

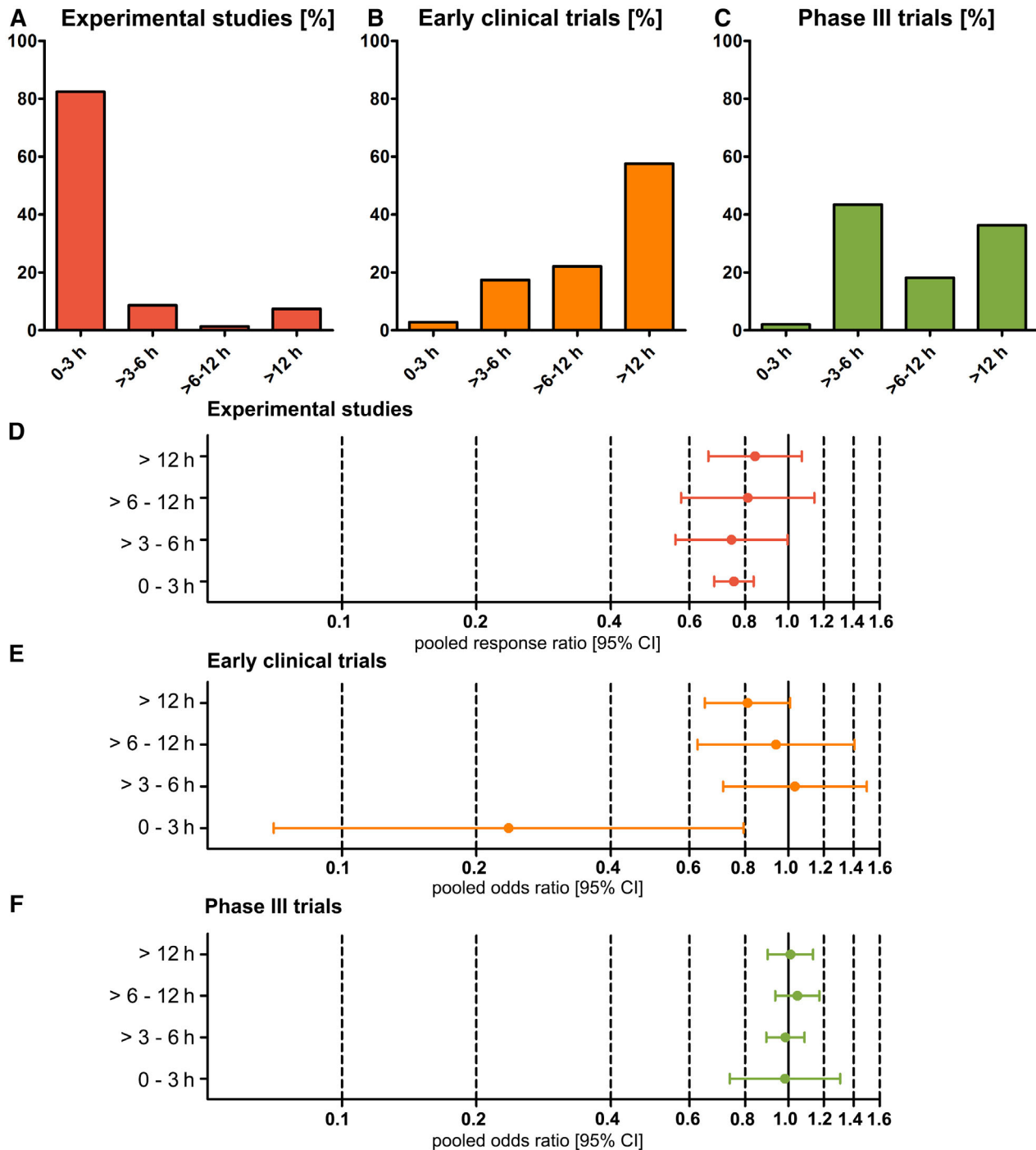


**FIGURE 4: Impact of study quality on treatment efficacy.** (A–C) Bar graphs showing remarkable differences in the adherence to quality criteria between experimental studies, early clinical trials, and phase 3 trials. (D, E) Meta-analyses illustrating the impact of randomization, blinded outcome assessment, and inclusion of subjects with comorbidities on treatment efficacy. Most interestingly, acute stroke therapies did not show a significant treatment effect in animals with comorbidities. Another important observation is that unblinded early clinical trials demonstrated efficacy of acute stroke treatments, whereas blinded studies did not show significant treatment effects. CI = confidence interval.

### Impact of Study Quality on Treatment Efficacy

Low methodological quality of experimental studies is commonly regarded as a major culprit for the translational failure of animal research.<sup>14,15</sup> We systematically reviewed the adherence to quality criteria and conducted meta-analyses to assess the impact of randomization, blinding, and the

use of animals with comorbidities on the estimated efficacy of acute stroke treatments in experimental studies and clinical trials. Only about half of the experimental studies reported randomization and blinded outcome assessment, whereas a vast majority of early clinical trials and all phase 3 trials obeyed these quality criteria (Fig 4). Also, only



**FIGURE 5:** Impact of time to treatment on efficacy. (A–C) Bar graphs illustrating considerable differences in the mean maximum time to treatment between experimental studies and clinical trials. (D–F) Meta-analyses illustrating the impact of time to treatment on treatment efficacy. Our findings demonstrate that the efficacy of acute stroke treatments is highly time dependent. CI = confidence interval.



8.2% of experimental studies included subjects with comorbidities, whereas all early and phase 3 clinical trials included patients with comorbidities. Notably, acute stroke therapies did not have a significant treatment effect in animals with comorbidities (response ratio = 0.86, 95% CI = 0.63–1.18) versus animals without comorbidities (response ratio = 0.76, 95% CI = 0.69–0.83). Another important finding of our meta-analyses is that unblinded early clinical trials demonstrate efficacy of acute stroke treatments, whereas blinded studies do not show significant treatment effects (odds ratio = 0.56, 95% CI = 0.38–0.82; vs odds ratio = 0.94, 95% CI = 0.77–1.13).

In summary, these data show that neglecting quality criteria contributes to an overestimation of treatment efficacy in experimental studies and early clinical trials.

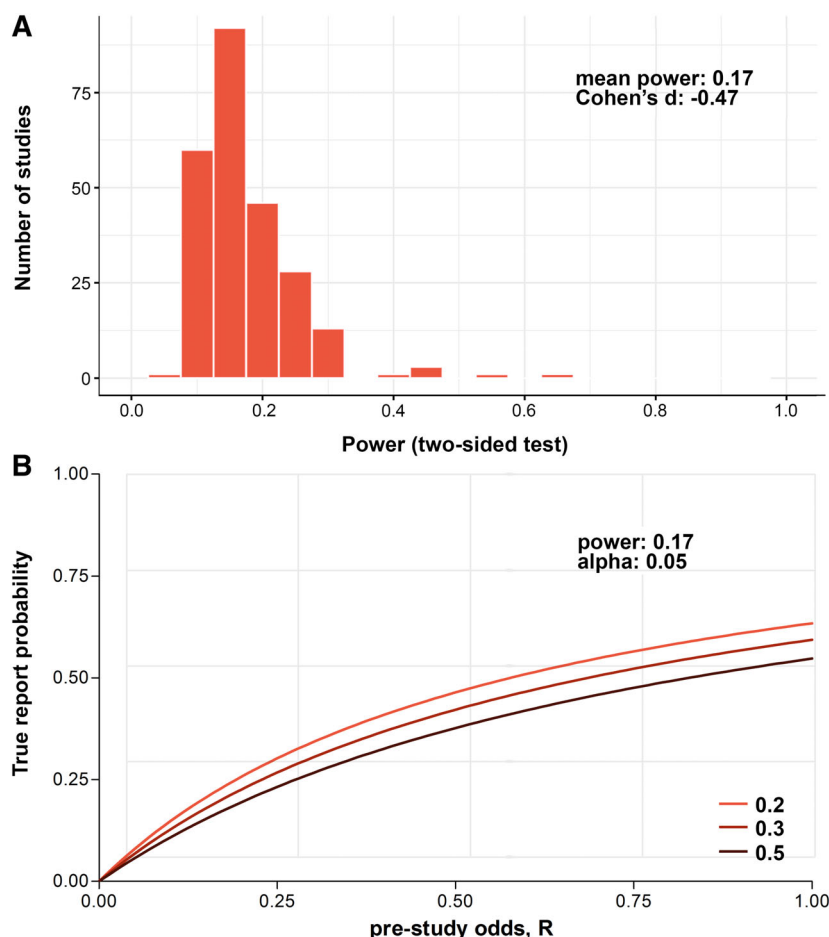
### Impact of Time to Treatment on Efficacy

Treatment delays have a major impact on the efficacy of acute stroke treatments. We analyzed time to treatment in the

different study stages and assessed the influence of time to treatment on observed treatment efficacy. Of note, we found a remarkable difference in time to treatment between experimental studies and clinical trials. In the majority (82.38%) of experimental studies, first drug administration occurred within the first 3 hours after ischemia onset, whereas in most clinical trials treatment was initiated later (Fig 5A–C).

Meta-analyses suggested that acute stroke treatments in experimental studies achieved the greatest treatment effects when initiated within the first 3 hours after ischemia onset (see Fig 5D). With a treatment delay of >6 hours, which is mostly seen in clinical trials, acute stroke treatments did not yield overall significant positive results in experimental studies.

Similarly, acute stroke treatments achieved positive effects in early clinical trials with a time to treatment of up to 3 hours after ischemia onset, whereas studies with later treatment initiation did not achieve overall significant positive results (see Fig 5E).



**FIGURE 6: Power and true report probability.** (A) The mean power in experimental studies was 17%. (B) True report probability depends on power, prestudy odds of true  $H_0$  to  $H_1$  hypotheses, and bias. Because bias and prestudy odds cannot be assessed precisely, we assumed different levels of bias and estimated the true report probability as a function of the prestudy odds. Even under conservative assumptions, with a low bias of 0.3 and prestudy odds between 0.5 and 0.7, the estimated true report probability of experimental stroke studies was lower than 50%.

Altogether, these data demonstrate that the efficacy of acute stroke treatments is highly time dependent (Supplementary Fig 3). Better developed collateral systems in human stroke patients compared to rodents may increase the ischemic tolerance and potentially extend the time window for neuroprotective treatments in human stroke patients.<sup>16,17</sup> However, delayed treatment initiation in clinical trials still remains a major factor contributing to the efficacy decline of acute stroke treatments from experimental studies to clinical trials.

### Power and True Report Probability

The reliability of a research finding may be defined by its true report probability, that is, the number of statistically significant true positive findings divided by the total number of reported statistically positive findings.<sup>12,18</sup> True report probability depends on power, prestudy odds of true  $H_0$  to  $H_1$  hypotheses, and bias, which in this context refers to any manipulation that leads to the presentation of intrinsically nonsignificant results as statistically significant.<sup>12,18</sup> For instance, a bias of 0.3 means that 30% of findings that would be statistically nonsignificant in the absence of bias will be reported as significant.<sup>18</sup>

To estimate the true report probability of experimental stroke studies, we performed power calculations and estimated a mean power of 17% (Fig 6). Because bias and prestudy odds cannot be assessed precisely, we assumed different levels of bias and estimated the true report probability as a function of prestudy odds. Even under the optimistic assumption of a low bias of 0.3 and prestudy odds between 0.5 and 0.7, the estimated true report probability of experimental stroke studies was lower than 50%.

These findings indicate that even under conservative estimates, less than half of the published positive experimental stroke studies are truly positive.

### Discussion

Our study shows a stepwise efficacy decline of acute stroke treatments from experimental studies to early clinical trials to phase 3 clinical trials. This stepwise efficacy decline is explained by publication bias, noncompliance with quality criteria, differences in study design (eg, treatment time window), and low power of experimental studies and early clinical trials.

We found a clear publication bias in experimental studies and early clinical trials, with a number of unpublished studies. Nonpublication of negative data obviously contributes to overestimated efficacy of experimental studies and early clinical trials. In addition, nonpublication of animal experiments raises major ethical concerns. First, animals have been sacrificed without advancing the knowledge of medical research. Second,

patients partaking in clinical trials are deceived by overestimated treatment efficacy in preceding animal studies.<sup>8</sup> To avoid publication bias in future studies, the implementation of a database has been suggested, registering all planned animal stroke studies in advance and ensuring studies are published even if study results are neutral or negative.<sup>19</sup> If primary end points and statistical methods were defined a priori, this would also prevent selective outcome reporting and selective analysis reporting.

Another important finding is that noncompliance with quality criteria contributed to an overestimation of treatment efficacy in experimental studies and early clinical trials. Randomization and blinding have been proposed as means to maximize the internal validity of animal experiments, meaning that the difference observed between animals of different treatment arms is in fact attributable to the intervention under investigation.<sup>20</sup> Randomization prevents the experimenter from allocating animals thought to perform either particularly well or badly to a certain treatment arm.<sup>20</sup> Blinding inhibits a systematic distortion of study results, which may occur when the investigator who assesses outcomes knows the treatment allocation. Moreover, blinding excludes differences in care between the treatment groups.<sup>20</sup>

The use of animals with comorbidities has been proposed as a means to increase the external validity of experimental stroke studies.<sup>20</sup> In this context, high external validity means that animal models mirror the clinical situation, where most stroke patients are old and morbid. Animal studies performed in young and healthy animals, as is the case in most experimental stroke studies, have low external validity. Our meta-analysis highlights that acute stroke treatments had no overall significant treatment effects in animals with comorbidities, whereas there was a large treatment effect in young and healthy animals. These findings demonstrate that poor external validity contributed considerably to overestimated efficacy of stroke treatments in experimental studies.

Our systematic analysis highlights intriguing differences in the study designs of experimental studies and clinical trials. For instance, almost all experimental studies provided infarct volumes as main end points, whereas most clinical studies reported neurological outcomes only. In addition, the treatment initiation was substantially delayed in clinical trials compared to experimental studies. Because early initiated treatment achieved best effects in experimental studies, it is likely that the treatment delay observed in clinical trials reduced the efficacy of acute stroke treatments, thus adding to the observed efficacy decline from experimental studies to clinical trials. To prevent translational failures due to study design differences between experimental studies and clinical trials, we suggest

conceptualizing future experimental studies in a study design that is 1:1 transferable into clinical trials.

Power has a major impact on the reliability of research findings.<sup>21</sup> According to our power estimation, the mean power of experimental studies was 17%. Technically, low power means that a study can only detect large treatment effects.<sup>21</sup> However, low-powered studies that coincidentally overestimate the dimension of a treatment effect can also detect treatment effects that are in truth small.<sup>21</sup> By inflation of effect sizes, low-powered experimental studies may thus have contributed to an overestimated efficacy of acute stroke treatments. In addition, low power is a main culprit for poor true report probability in experimental studies. As detailed previously, true report probability depends on power, prestudy odds of true  $H_0$  to  $H_1$  data, and bias. The prestudy odds cannot be influenced by the experimenter. Bias cannot be completely excluded either, although major types of bias like publication bias and reporting bias could be eliminated by preregistering studies and a priori defining primary end points. Power, by contrast, could be increased easily by implementing a priori sample size calculations.

Our study has strengths and limitations. A limitation of our study is that we cannot assess selective outcome reporting and selective analysis reporting in experimental studies. This leads to an underestimation of bias in experimental studies. For early clinical trials and phase 3 clinical trials, we systematically compared primary and secondary end points listed in ClinicalTrials.gov to reported end points in published studies and found no evidence for selective reporting in registered early clinical trials and phase 3 trials. Another limitation is that power estimations were based on results that were influenced by publication bias. This leads to an overestimation of power, with the true power being even lower than our estimated value. We addressed heterogeneity by comparing effects sizes between different stroke models and stroke subtypes. Our results show that experimental studies and early phase clinical trials overestimate treatment effects of neuroprotectants independent of stroke models in experimental studies and stroke subtypes in clinical trials. These findings strongly support the existence of an efficacy decline independent of differences in heterogeneity. Nonetheless, we cannot rule out heterogeneity as a contributing factor to the failure of trials at phase 3.

The comprehensive and systematic approach is a strength of our study. We provide the first systematic review and meta-analysis of virtually all medical treatments for acute ischemic stroke that have been under investigation in experimental studies, early clinical trials, and phase 3 clinical trials. Our large data set allows meaningful conclusions about translational difficulties in stroke research,

and some of our methodological conclusions can be transferred to other fields of medical research.

## Conclusion

In addition to the well-characterized translational roadblock between preclinical and clinical studies, our study unveils further obstacles on the path from early phase to phase 3 clinical trials. Most importantly, we identified major study design differences between experimental studies and clinical trials. For instance, most experimental studies use rodent models of middle cerebral artery occlusion with large hemispheric infarcts and examine infarct volumes in the acute phase as primary outcomes, whereas most clinical trials include various stroke etiologies and determine neurological outcomes in the chronic phase as primary outcomes. Delayed treatment initiation in clinical trials is another major study design difference explaining failed translation from experimental studies to clinical trials. Beyond these differences in study design, noncompliance with quality criteria and low power are pivotal methodological weaknesses contributing to the gradual efficacy decline of acute stroke treatments from experimental studies to early phase clinical trials and from early phase clinical trials to phase 3 clinical trials. Surprisingly, our estimations of power and true report probability suggest that less than half of the published positive experimental stroke studies are truly positive.

## Acknowledgment

This study was supported by a research grant of the Interdisciplinary Center for Clinical Research of the Medical Faculty of the University of Münster (SEED07, A.S.-P.), by the Deutsche Forschungsgemeinschaft (MI 1547/3-1), by the Else Kröner-Fresenius-Stiftung (J.M., 2014\_EKES.16), and by the DAAD grant 57403633.

## Author Contributions

A.S.-P., N.B., H.M., and J.M. developed the study concept and design; A.S.-P., N.B., M.H.M.K., J.W., N.-M. B., and C.B. were involved in data acquisition and analysis; and A.S.-P., J.-K.S., W.-R.S., S.G.M., H.W., and J.M. drafted the manuscript and figures.

## Potential Conflicts of Interest

Nothing to report.

## References

1. Benjamin EJ, Blaha MJ, Chiuve SE, et al. Heart disease and stroke statistics-2017 update: a report from the American Heart Association. *Circulation* 2017;135:e146–e603.

2. Powers WJ, Rabinstein AA, Ackerson T, et al. 2018 guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 2018;49:e46–e110.
3. O'Collins VE, Macleod MR, Donnan GA, et al. 1,026 experimental treatments in acute stroke. *Ann Neurol* 2006;59:467–477.
4. Hackam DG. Translating animal research into clinical benefit. *BMJ* 2007;334:163–164.
5. Hackam DG, Redelmeier DA. Translation of research evidence from animals to humans. *JAMA* 2006;296:1731–1732.
6. Hedges LV, Gurevitch J, Curtis PS. The meta-analysis of response ratios in experimental ecology. *Ecology* 1999;80:1150–1156.
7. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw* 2010;36:1–48.
8. 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.
9. Sterne JA, Sutton AJ, Ioannidis JP, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011;343:d4002.
10. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 2000;56:455–463.
11. Champely S. Pwr: basic functions for power analysis. R package version 1.2-2. Available at: <https://CRAN.R-project.org/package=pwr>. Last assessed May 2019.
12. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2:e124.
13. J Gill A, A Barletta J. Annual review issue: endocrine pathology: a pathological, clinical and molecular integration. *Histopathology* 2018;72:3.
14. Sena E, van der Worp HB, Howells D, Macleod M. How can we improve the pre-clinical development of drugs for stroke? *Trends Neurosci* 2007;30:433–439.
15. Macleod MR, Lawson McLean A, Kyriakopoulou A, et al. Risk of bias in reports of in vivo research: a focus for improvement. *PLoS Biol* 2015;13:e1002273.
16. McColl BW, Carswell HV, McCulloch J, Horsburgh K. Extension of cerebral hypoperfusion and ischaemic pathology beyond MCA territory after intraluminal filament occlusion in C57Bl/6J mice. *Brain Res* 2004;997:15–23.
17. Nogueira RG, Jadhav AP, Haussen DC, et al. Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. *N Engl J Med* 2018;378:11–21.
18. Szucs D, Ioannidis JP. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol* 2017;15:e2000797.
19. Liebeskind DS, Kidwell CS, Sayre JW, Saver JL. Evidence of publication bias in reporting acute stroke clinical trials. *Neurology* 2006;67: 973–979.
20. van der Worp HB, Howells DW, Sena ES, et al. Can animal models of disease reliably inform human studies? *PLoS Med* 2010;7: e1000245.
21. Button KS, Ioannidis JP, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 2013;14:365–376.