Published in final edited form as:

Anesth Analg. 2018 January ; 126(1): 217–222. doi:10.1213/ANE.0000000000001907.

Study Design Rigor in Animal-Experimental Research Published in Anesthesia Journals

Janine M Hoerauf, B.A.¹ [Fourth Year Medical Student], Angela F Moss, M.S.² [Staff Biostatistician], Ana Fernandez-Bustamante, M.D., Ph.D.³ [Associate Professor of Anesthesiology], and Karsten Bartels, M.D.⁴ [Assistant Professor of Anesthesiology, Medicine, and Surgery]

¹University of Colorado Denver, School of Medicine. JANINE.HOERAUF@UCDENVER.EDU. Contribution: Janine M Hoerauf contributed to the study design, conduct of the study, data collection, data analysis, and wrote the manuscript. Attestation: Janine M Hoerauf approved the final manuscript. Janine M Hoerauf attests to the integrity of the original data and the analysis reported in this manuscript. Conflicts of Interest: None

²Adult and Child Consortium for Health Outcomes Research and Delivery Science, University of Colorado Denver. ANGELA.MOSS@UCDENVER.EDU. Contribution: Angela F Moss contributed to the study design, data analysis, and helped critically revise the manuscript. Attestation: Angela F Moss approved the final manuscript. Angela F Moss attests to the integrity of the data analysis reported in this manuscript. Conflicts of Interest: None

³Department of Anesthesiology, University of Colorado Denver. ANA.FERNANDEZ-BUSTAMANTE@UCDENVER.EDU. Contribution: Ana Fernandez-Bustamante contributed to interpretation of the data and helped critically revise the manuscript. Attestation: Ana Fernandez-Bustamante approved the final manuscript. Conflicts of Interest: None

⁴Department of Anesthesiology, University of Colorado Denver. KARSTEN.BARTELS@UC.DENVER.EDU. Contribution: Karsten Bartels secured funding for the study, contributed to the study design, conduct of the study, data collection, data analysis, and manuscript preparation. Attestation: Karsten Bartels approved the final manuscript. Karsten Bartels attests to the integrity of the original data and the analysis reported in this manuscript. Karsten Bartels is the archival author. Conflicts of Interest: None.

Abstract

Background—Lack of reproducibility of preclinical studies has been identified as an impediment for translation of basic mechanistic research into effective clinical therapies. Indeed,

Corresponding author: Karsten Bartels, M.D., Assistant Professor of Anesthesiology, Medicine, and Surgery, University of Colorado Denver, 12401 E. 17th Avenue, 7th Floor, B113, Aurora, CO 80045, Phone: 720-848-6845, Fax: 720-848-7375, karsten.bartels@ucdenver.edu.

Clinical trial number: Not applicable.

Preliminary results of this work were presented at the Young Investigator session at the ANESTHESIOLOGY 2016 meeting in Chicago, IL.

Disclosure of Funding: Karsten Bartels is supported by the National Institutes of Health / National Institute On Drug Abuse, Award Number K23DA040923. The content of this report is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

the National Institutes of Health have revised their grant application process to require more rigorous study design, including sample size calculations, blinding procedures, and randomization steps. We hypothesized that the reporting of such metrics of study design rigor has increased over time for animal-experimental research published in anesthesia journals.

Methods—PubMed was searched for animal-experimental studies published in 2005, 2010, and 2015 in primarily English-language anesthesia journals. A total of 1466 publications were graded on the performance of sample size estimation, randomization, and blinding. Cochran-Armitage test was used to assess linear trends over time for the primary outcome of whether or not a metric was reported. Inter-rater agreement for each of the three metrics (power, randomization, and blinding) was assessed using the weighted kappa coefficient in a 10% random sample of articles re-rated by a second investigator blinded to the ratings of the first investigator.

Results—A total of 1466 manuscripts were analyzed. Reporting for all three metrics of experimental design rigor increased over time (2005–2010–2015): For power analysis from 5% (27/516), to 12% (59/485), and to 17% (77/465), for randomization from 41% (213/516), to 50% (243/485), and to 54% (253/465), and for blinding from 26% (135/516), to 38% (186/485), and to 47% (217/465). The weighted kappa coefficients and 98.3% CI indicate almost perfect agreement between the two raters beyond that which occurs by chance alone [Power: 0.93 (0.85,1.0), Randomization: 0.91 (0.85,0.98), Blinding: 0.90 (0.84,0.96)].

Conclusions—Our hypothesis that reported metrics of rigor in animal-experimental studies in anesthesia journals have increased over the past decade was confirmed. More consistent reporting - or explicit justification for absence - of sample size calculations, blinding techniques, and randomization procedures could better enable readers to evaluate potential sources of bias in animal-experimental research manuscripts. Future studies should assess if such steps lead to improved translation of animal-experimental anesthesia research to successful clinical trials.

Introduction

Over the past decade, there have been growing concerns regarding a lack of reproducibility in scientific research in general and animal-experimental studies in particular.^{1–4} Accordingly, sample size calculations, proper use of blinding techniques, and implementation of randomization procedures are performed by scientists to increase reproducibility of their results.⁵ Yet, such measures are not routinely included in many experimental research protocols. For example, a 2009 survey of 271 bio-medical animal experimental studies found that 87% did not perform randomization, and 86% did not perform blinding.⁶ Robust experimental design is required to maximize the quality of results while minimizing resources as well as animal suffering.^{7–9} Most importantly, reproducible animal experimental research results are at the core of our ability to translate basic science into tangible benefits for patients. Only the most promising novel concepts should proceed towards clinical trials in humans.¹⁰ Yet, while more than 1,000 substances demonstrated neuroprotective effects in the laboratory, they failed to translate into effective clinical therapies.¹¹ Indeed, an evaluation of excess significance bias in animal studies of neurological disease came to the conclusion that only eight of 160 evaluated treatments should have even gone on to clinical trials in humans.¹²

The European Association of Science Editors^{*} and the International Committee of Medical Journal Editors[†] are continuously updating their guidelines to promote publications of complete, concise and standardized manuscripts. In 2010, the Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines were developed to improve transparency and reproducibility of animal studies.¹³ Accordingly, ARRIVE guidelines recommend proper sample size calculations, blinding, and randomization procedures.

Simple experimental design steps to reduce bias are common in clinical studies and have also long been recommended for animal experimental research.^{13,14} In 2014, the directors of the National Institutes of Health (NIH), Collins and Tabek, released an influential publication which outlined the NIH's proposals to enhance reproducibility.¹⁵ Subsequently, NIH's updated reviewer guidelines for federally funded research have emphasized the requirement for more scientific rigor: "The strict application of the scientific method to ensure robust and unbiased experimental design, methodology, analysis, interpretation and reporting of results."[‡]

The need to implement steps that enhance rigorous scientific methods to advance perioperative medicine is crucial for the field of anesthesiology. Here, we sought to examine trends in reporting of sample size estimates, blinding, and randomization in animal-experimental studies published in anesthesia journals in 2005, 2010, and 2015. We hypothesized that reporting of such study design metrics of rigor was more common in more recent publications.

Materials and Methods

Journal Selection

Journals were selected within the discipline of anesthesiology listed with impact factor in the Journal Citation Report from 2015 (2015 Journal Citation Reports® [Thomson Reuters, 2015]). Journals that had a primary reporting language other than English were excluded. This manuscript reports results of an observational bibliometric study and therefore follows the applicable Enhancing the QUALity and Transparency Of health Research EQUATOR guidelines (PROMIS and STROBE).[§]

Database Search for Studies

The sequence of search strategies and analyses of the articles is summarized in Figure 1. PubMed was searched for animal-experimental research in the selected anesthesia journals in 2005, 2010, and 2015. The search was limited by the "Date - Publication" field descriptor between January 1 and December 31 and the respective year. Each journal was selected using the field descriptor "Journal". In order to narrow the search results to primary research studies, the following terms were excluded from the search under the field descriptor "Study type": historical article, comment, letter, review, editorial, case report, and meta-analysis.

^{*} Available at: <http://www.ease.org.uk/publications/author-guidelines-authors-and-translators/>. Accessed September 1, 2016.

[†] Available at: <http://www.ICMJE.org>. Accessed September 1, 2016.

[‡] Available at: http://grants.nih.gov/grants/peer/guidelines_general/Reviewer_Guidance_on_Rigor_and_Transparency.pdf. Accessed September 1, 2016.

[§] Available at: <http://www.equator-network.org/>. Accessed 9/1/2016.

This search yielded a preliminary list of articles. An investigator then manually pruned this list to include only studies using animal models, which included primary animal studies and studies with both animal and human components. The National Library of Medicine Medical Subject Heading (MeSH) definition of “animal” was used for inclusion criteria, which is defined as follows: “Unicellular or multicellular, heterotrophic organisms, that have sensation and the power of voluntary movement. Under the older five-kingdom paradigm, Animalia was one of the kingdoms. Under the modern three domain model, Animalia represents one of the many groups in the domain Eukaryota.”** Experiments included, but were not limited to, animal behavior studies, isolated animal organs (i.e. spinal cord, aortic root, lung, liver), and primary cell cultures. Each list was then cross-referenced for validity with a list generated using the same date, journal, and publication type as described above with an added “animal” filter. The final list included 1466 publications of animal-experimental research in anesthesia journals with 516 from 2005, 485 from 2010, and 465 from 2015. A comprehensive list of journals, search terms, domains, and articles can be found in Supplemental Digital Content 1. In order to limit cognitive bias from the rater, studies were randomized prior to analysis using a computer-generated random number generator [“=rand()” function, Microsoft® Excel®, Microsoft Co., Seattle, WA].

A second investigator blinded to the animal-experimental study search from the first investigator performed a PubMed search for two randomly selected journals (*Journal of Anesthesia* and *European Journal of Pain*) for 2005, 2010, and 2015) using only the journal name and year as search-limiting terms. Following manual pruning, the second reviewer identified 192 of the 193 manuscripts that had been classified as an animal-experimental study by the first investigator. No additional manuscripts were identified. Un-blinded re-review of the manuscript not identified as an animal-experimental study during manual pruning by the second reviewer revealed that this manuscript, in fact, described an animal-experimental study.

Search term criteria

Three categories were chosen to investigate study design rigor: sample size estimate, blinding, and randomization. The word particles “power”, “sample size”, “sample” and “group” were used to evaluate sample size estimate. The word particle “rando” was used to evaluate randomization. The word particle “blind” was used to evaluate blinding. These metrics were chosen both because they are commonly recommended to ensure robust and reliable results, and they can be practically used in an electronic search function.^{16,17}

Collection of reporting

Adobe® Acrobat Reader’s (Adobe® Systems Incorporated, San Jose, CA) *find* function was used to search PDF versions of each article for the word particles above. Word particles were then assessed to ensure context within the experimental design description and rated on an ordinal scale from 0–3 to assess the quality of reporting. A score of 0 was assigned if the term was not found by the search function, 1 if the term was mentioned but metric was not performed, 2 if the metric was performed but the method was not detailed and 3 if the metric

** Available at: https://www.nlm.nih.gov/cgi/mesh/2016/MB_cgi?mode=&term=Animals&field=entry. Accessed September 1, 2016.

was performed and the method was described. A detailed description of evaluation criteria can be found in the Supplemental Digital Content 2.

Inter-rater Agreement

A second investigator blinded to the initial rating scored a random selection of 10% of the articles (n=147) analyzed in each year for sample size estimate, blinding, and randomization. The random selection was performed using the random number generator described above. These full-text articles were then scored using the same evaluation criteria and methods as were used for the initial analysis.

Statistical Analysis

To address the primary hypothesis ordinal scale rating scores (0-not mentioned, 1-mentioned but not performed, 2-performed but no details, 3-performed and details given) were collapsed into binary (performed/not performed) variables. To examine overall trends in reporting of quality metrics as binary outcomes across the three time points, the Cochran-Armitage Test for linear trend was used.^{18,19} Simple logistic regression with time as a continuous covariate was used to estimate the effect of time on the likelihood of each metric being performed in published articles. P-values and confidence intervals were corrected using the Bonferroni method for multiple comparisons.

To examine overall trends in reporting of quality metrics as ordinal outcomes across the three time points, the Cochran-Mantel Haenszel Test for linear trend was used.²⁰ P-values were adjusted for multiple comparisons using the Bonferroni correction.

Inter-rater agreement for each of the three metrics (power, randomization, and blinding) scored using the four-point ordinal scale was assessed using the weighted kappa coefficient.^{21,22} Kappa represents the difference in agreement between that which is actually observed (observed agreement) and that which occurs by chance alone. Weights were calculated with the Cicchetti-Allison method.

Analysis was done in SAS 9.4 SAS Institute Inc., Cary, North Carolina USA. All statistical tests were performed adjusting for multiple comparisons to maintain an overall 0.05 level of significance.

For the power analysis, we assumed a 25% absolute increase in reporting incidences for each of the three metrics over a ten-year interval in two independent proportions. Therefore, we anticipated a baseline reporting level of 25% in 2005 and a reporting level of 50% in 2015. In order to maintain a 0.05 significance level across the three outcome metrics, the Bonferroni method for multiple comparisons was used to adjust the alpha to 0.017. A total of 77 studies in each year (154 total) would yield 80% power to detect an absolute difference in the proportion of metrics identified of at least 25% as significant (alpha adjusted to 0.017 for multiple comparisons, overall alpha 0.05). This power calculation was done using G*Power Version 3.1.9.2.²³

Results

Study selection

This study included 1466 publications from 23 journals with 516 from 2005, 485 from 2010, and 465 from 2015 (Table 1).

Reporting of Quality Metrics

Reporting for all three metrics of experimental design rigor increased over time (2005–2010–2015): For power analysis from 5% (27/516), to 12% (59/485), and to 17% (77/465), for randomization from 41% (213/516), to 50% (243/485), and to 54% (253/465), and for blinding from 26% (135/516), to 38% (186/485), and to 47% (217/465) (Table 2). The odds of a power analysis being performed and mentioned in a published article increased 81% (40%, 134%) within a 5-year interval. The odds of a randomization being performed and mentioned in a published article increased 30% (12%, 52%) within a 5-year interval. Finally, the odds of blinding being performed and mentioned in a published article increases 57% (33%, 84%) within a 5-year interval.

Further, the quality of the reporting for power analysis, randomization, and blinding assessed on a four-point ordinal scale also significantly increased over time (all $p < 0.0001$), (Table 3). From 2005 to 2010 and to 2015, the percentage of articles that received the highest rating for reporting (defined as: “metric reported, performed, and details described”) increased. For power analysis: from 4% (22/516) to 11% (53/485) and to 14% (67/465). For randomization: from 4% (23/516) to 7% (36/485) and to 10% (46/465). For blinding: from 18% (93/516) to 26% (124/485) and to 31% (142/465). The percentage of articles that received the lowest rating for reporting (defined as: “metric word particle not found”) correspondingly decreased. For power analysis: from 90% (465/516) to 80% (389/485) and to 60% (280/465). For randomization: from 58% (300/516) to 49% (239/485) and to 43% (199/465). For blinding: from 72% (371/516) to 61% (294/485) and to 49% (226/465).

Inter-rater Agreement

Observed agreement between the two investigators for the 147 article assessed was 0.95, 0.93, 0.89 for power, randomization, and blinding respectively. The Cicchetti-Allison weighted Kappa coefficients and lower limits of the 98.3% CI indicate almost perfect agreement beyond that which occurs by chance alone [Power: 0.93 (0.85,1.0), Randomization: 0.91 (0.85,0.98), Blinding: 0.90 (0.84,0.96)].

Discussion

When comparing the years 2005, 2010, and 2015, animal experimental research manuscripts published in anesthesia journals showed increased rates of reporting for power analysis (from 5% to 12% to 17%), randomization (from 41% to 50% to 54%), and blinding (from 26% to 38% to 47%). Our hypothesis that reported metrics of rigor in animal experimental studies have increased over the past decade was confirmed. Yet, even in 2015, only a minority of published research manuscripts included information on all of the analyzed experimental design procedures to improve rigor.

Fields such as oncology and neuroscience have been amongst the first to embrace approaches to demand new standards for pre-clinical research. In 2010, Hess and colleagues reviewed 100 publications in the journal *Cancer Research* and found that 28% of the articles reported random allocation of animals, only 2% reported blinding, and 14% did not report sample size.²⁴ Concerns that novel mechanistic findings may lack reproducibility, and lead to the subsequent design of futile clinical trials, have prompted vigorous responses to establish more robust requirements for pre-clinical cancer research.² In the field of neuroscience, Button and colleagues reported an average statistical power of only 8% – 31%, indicating a need to address proper sample size estimation prior to pursuing experimental research.²⁵ Also within the field of neuroscience, Sena and colleagues found low reporting of randomization (2–12%), blinding (3–18%) and sample size calculation (0%) in animal experimental studies.²⁶ Increased attention to the basics of sound statistics and robust study design have since been demanded within the neuroscience community.²⁷ Our study is novel as it details the state of reporting on select experimental study design metrics in the field of anesthesia. While we did identify current shortcomings, we were also able to demonstrate significant improvements in the last ten years by authors and journals to include more detailed information on power analysis, randomization, and blinding in published manuscripts. Explicit editorial requirements for authors to use reporting checklists such as the ones recommended by the “EQUATOR Network”^{††} are likely to continue to enhance reporting for essential characteristics of experimental study design.^{28,29} Hence, the observed increases in reporting of recommended experimental study design features in our study may not come as a surprise, yet, the fact that the numbers remained as low as they were in 2015 is noteworthy.

Our study has several limitations: First, we cannot exclude that investigators actually performed power analysis, randomization or blinding - without reporting this in their manuscripts. Second, some studies may have used unusual wording that would not have been recognized by our search strategy. However, this concern was not confirmed during the manual pruning of relevant methods sections in the manuscripts analyzed. Third, robust, reproducible, and translatable basic science research requires more than *a priori* power analysis, randomization and blinding.³⁰ For example, in-bred strains of animals of the same age that are housed under identical conditions do not reflect the genetic heterogeneity and differing environments encountered in humans. In addition, gender-specific effects need to be accounted for in experimental protocols and efforts to translate basic science into clinical trials. However, for this study, we chose to narrow our focus to three important concepts as they are relatively easily implemented, recommended by current guidelines, and can be unambiguously reported in a manuscript.¹³ Lastly, a potential limitation of our study is observer bias because both the selection of articles and rating of metrics are subject to human error. We attempted to limit this source of bias through cross-referencing manually chosen articles with automated filters available within PubMed and comparative rating of the three metrics by a second, blinded rater.

^{††}Available at: <http://www.equator-network.org/> Accessed 9/1/2016.

In summary, we found an increase in the reporting of power analysis, randomization, and blinding in anesthesia journals from 2005 to 2010 and to 2015. Going forward, more scientific anesthesia journals could routinely require inclusion - or justification for the absence - of recommended study design features in published manuscripts. Enhanced reporting could enable readers to better assess potential sources of bias. Future studies should assess if such steps lead to improved translation of animal-experimental anesthesia research to successful clinical trials.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Individuals or organizations to be acknowledged: The authors thank Dr. William Henderson, Ph.D., M.P.H., Professor, Department of Biostatistics and Informatics, University of Colorado School of Public Health for assistance with the statistical analysis.

References

1. Prinz F, Schlange T, Asadullah K. Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets? *Nat Rev Drug Discov.* 2011; 10:712. [PubMed: 21892149]
2. Begley CG, Ellis LM. Drug Development: Raise Standards for Preclinical Cancer Research. *Nature.* 2012; 483:531–3. [PubMed: 22460880]
3. Ioannidis JP. Why Most Published Research Findings Are False. *PLoS Med.* 2005; 2:e124. [PubMed: 16060722]
4. Smith JA, Birke L, Sadler D. Reporting Animal Use in Scientific Papers. *Lab Anim.* 1997; 31:312–7. [PubMed: 9350702]
5. Baker M. 1,500 Scientists Lift the Lid on Reproducibility. *Nature.* 2016; 533:452–4. [PubMed: 27225100]
6. Kilkenny C, Parsons N, Kadyszewski E, et al. Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals. *PLoS One.* 2009; 4:e7824. [PubMed: 19956596]
7. Carbone L, Austin J. Pain and Laboratory Animals: Publication Practices for Better Data Reproducibility and Better Animal Welfare. *PLoS One.* 2016; 11:e0155001. [PubMed: 27171143]
8. Festing MF, Nevalainen T. The Design and Statistical Analysis of Animal Experiments: Introduction to This Issue. *ILAR J.* 2014; 55:379–82. [PubMed: 25541539]
9. Avey MT, Fenwick N, Griffin G. The Use of Systematic Reviews and Reporting Guidelines to Advance the Implementation of the 3rs. *J Am Assoc Lab Anim Sci.* 2015; 54:153–62. [PubMed: 25836961]
10. Begley CG, Ioannidis JP. Reproducibility in Science: Improving the Standard for Basic and Preclinical Research. *Circ Res.* 2015; 116:116–26. [PubMed: 25552691]
11. Bartels K, Karhausen J, Clambey ET, et al. Perioperative Organ Injury. *Anesthesiology.* 2013; 119:1474–89. [PubMed: 24126264]
12. Tsilidis KK, Panagiotou OA, Sena ES, et al. Evaluation of Excess Significance Bias in Animal Studies of Neurological Diseases. *PLoS Biol.* 2013; 11:e1001609. [PubMed: 23874156]
13. Kilkenny C, Browne WJ, Cuthill IC, et al. Improving Bioscience Research Reporting: The Arrive Guidelines for Reporting Animal Research. *J Pharmacol Pharmacother.* 2010; 1:94–9. [PubMed: 21350617]
14. Schulz KF, Altman DG, Moher D, et al. Consort 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials. *PLoS Med.* 2010; 7:e1000251. [PubMed: 20352064]

15. Collins FS, Tabak LA. Policy: Nih Plans to Enhance Reproducibility. *Nature*. 2014; 505:612–3. [PubMed: 24482835]
16. Festing MF, Altman DG. Guidelines for the Design and Statistical Analysis of Experiments Using Laboratory Animals. *ILAR J*. 2002; 43:244–58. [PubMed: 12391400]
17. Landis SC, Amara SG, Asadullah K, et al. A Call for Transparent Reporting to Optimize the Predictive Value of Preclinical Research. *Nature*. 2012; 490:187–91. [PubMed: 23060188]
18. Cochran WG. Some Methods for Strengthening the Common χ^2 Tests. *Biometrics*. 1954; 10:417–51.
19. Armitage P. Tests for Linear Trends in Proportions and Frequencies. *Biometrics*. 1955; 11:375–86.
20. Mantel N, Haenszel W. Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *J Natl Cancer Inst*. 1959; 22:719–48. [PubMed: 13655060]
21. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics*. 1977; 33:159–74. [PubMed: 843571]
22. Viera AJ, Garrett JM. Understanding Interobserver Agreement: The Kappa Statistic. *Fam Med*. 2005; 37:360–3. [PubMed: 15883903]
23. Faul F, Erdfelder E, Buchner A, et al. Statistical Power Analyses Using G*Power 3.1: Tests for Correlation and Regression Analyses. *Behav Res Methods*. 2009; 41:1149–60. [PubMed: 19897823]
24. Hess KR. Statistical Design Considerations in Animal Studies Published Recently in Cancer Research. *Cancer Res*. 2011; 71:625. [PubMed: 21239476]
25. Button KS, Ioannidis JP, Mokrysz C, et al. Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Nat Rev Neurosci*. 2013; 14:365–76. [PubMed: 23571845]
26. Sena E, van der Worp HB, Howells D, et al. How Can We Improve the Pre-Clinical Development of Drugs for Stroke? *Trends Neurosci*. 2007; 30:433–9. [PubMed: 17765332]
27. Couzin-Frankel J. When Mice Mislead. *Science*. 2013; 342:922–3. 5. [PubMed: 24264972]
28. Pittet JF, Vetter TR. Continuing the Terra Firma and Establishing a New Equator for Anesthesia & Analgesia. *Anesth Analg*. 2016; 123:8–9. [PubMed: 27183374]
29. Eisenach JC, Warner DS, Houle TT. Reporting of Preclinical Research in Anesthesiology: Transparency and Enforcement. *Anesthesiology*. 2016; 124:763–5. [PubMed: 26845144]
30. Traystman RJ, Herson PS. Misleading Results: Translational Challenges. *Science*. 2014; 343:369–70. [PubMed: 24458624]

Figure 1. Flow diagram for search methods and data analysis

Search strategy and analysis of articles reporting animal-experimental research in anesthesia journals.

Table 1**Journal Articles Included in the Study**

Journals were selected within the discipline of Anesthesiology listed with impact factor in the Journal Citation Report from 2015 (2015 Journal Citation Reports® [Thomson Reuters, 2015]). Percentages are given in brackets.

Journal	Impact Factor	Total (n= 1466)	2005 (n=516)	2010 (n=485)	2015 (n=465)
British Journal of Anaesthesia	5.616	97 (6.6)	42 (2.9)	33 (2.3)	22 (1.5)
Pain	5.557	330 (22.5)	127 (8.7)	111 (7.6)	92 (6.3)
Anesthesiology	5.555	278 (19.0)	104 (7.1)	88 (6.0)	86 (5.9)
Anesthesia & Analgesia	3.827	262 (17.9)	128 (8.7)	78 (5.3)	56 (3.8)
Anaesthesia	3.794	9 (0.6)	0	3 (0.2)	6 (0.4)
European Journal of Anaesthesiology	3.634	55 (3.8)	24 (1.6)	20 (1.4)	11 (0.8)
Regional Anesthesia and Pain Medicine	3.459	21 (1.4)	5 (0.3)	11 (0.8)	5 (0.3)
Pain Physician	3.407	10 (0.7)	1 (0.1)	1 (0.1)	8 (0.5)
European Journal of Pain	2.900	133 (9.1)	12 (0.8)	55 (3.8)	66 (4.5)
Journal of Neurosurgical Anesthesiology	2.828	19 (1.3)	7 (0.5)	7 (0.5)	5 (0.3)
Clinical Journal of Pain	2.712	1 (0.1)	0	0	1 (0.1)
Pain Practice	2.317	2 (0.1)	0	1 (0.1)	1 (0.1)
Canadian Journal of Anesthesia	2.139	18 (1.2)	9 (0.6)	9 (0.6)	0
Pediatric Anesthesia	2.082	11 (0.8)	0	4 (0.3)	7 (0.5)
Acta Anaesthesiologica Scandinavica	2.049	87 (5.9)	38 (2.6)	26 (1.8)	23 (1.6)
International Journal of Obstetric Anesthesia	2.040	2 (0.1)	1 (0.1)	1 (0.1)	0
Minerva Anesthesiologica	2.036	7 (0.5)	1 (0.1)	3 (0.2)	3 (0.2)
Current Opinion in Anesthesiology	1.916	0	0	0	0
Journal of Clinical Monitoring & Computing	1.819	15 (1.0)	0	4 (0.3)	11 (0.8)
Journal of Cardiothoracic and Vascular Anesthesia	1.519	16 (1.1)	2 (0.1)	9 (0.6)	5 (0.3)
Journal of Anesthesia	1.343	60 (4.1)	12 (0.8)	20 (1.4)	28 (1.9)
BMC Anesthesiology	1.320	29 (2.0)	1 (0.1)	0	28 (1.9)
Journal of Clinical Anesthesia	1.284	1 (0.1)	0	0	1 (0.1)
Anaesthesia and Intensive Care	1.283	3 (0.2)	2 (0.1)	1 (0.1)	0

Table 2
Performance of Power Analysis, Randomization, and Blinding in Animal-Experimental Research in Anesthesia

To examine overall trends in reporting of quality metrics (binary) across the three time points, the Cochran-Armitage Test for linear trend was used. P-values were adjusted for multiple comparisons using the Bonferroni correction. Simple logistic regression with time as a continuous covariate was used to estimate the effect of time on metric performed and mentioned in published articles. The reference group was “not performed” and odds ratios were calculated for 5-year increments in time. Percentages are given in brackets.

Variables	Total (n= 1466)	2005 (n=516)	2010 (n=485)	2015 (n=465)	P-value	Odds Ratio (98.3%CI)
Power						
Performed	163 (11)	27 (5)	59 (12)	77 (17)	<0.0001	1.81 (1.4, 2.34)
Not performed	1,303 (89)	489 (95)	426 (88)	388 (83)		
Randomization						
Performed	709 (48)	213 (41)	243 (50)	253 (54)	0.0001	1.30 (1.12,1.52)
Not performed	757 (52)	303 (59)	242 (50)	212 (46)		
Blinding						
Performed	538 (37)	135 (26)	186 (38)	217 (47)	<0.0001	1.57 (1.33,1.84)
Not performed	928 (63)	381 (74)	299 (62)	248 (53)		

Table 3
Qualitative Rating of Power Analysis, Randomization, and Blinding in Animal-Experimental Research in Anesthesia

To examine overall trends in reporting of quality metrics as ordinal outcomes across the three time points, the Cochran-Mantel Haenszel Test for linear trend was used. P-values were adjusted for multiple comparisons using the Bonferroni correction. Percentages are given in brackets.

Variables	Total (n=1466)	2005 (n=516)	2010 (n=485)	2015 (n=465)	P-value
Power					
Not mentioned	1,134 (77)	465 (90)	389 (80)	280 (60)	<0.0001
Mentioned but not performed	169 (12)	24 (5)	37 (8)	108 (23)	
Performed but no details	21 (1)	5 (1)	6 (1)	10 (2)	
Performed and details given	142 (10)	22 (4)	53 (11)	67 (14)	
Randomization					
Not mentioned	738 (50)	300 (58)	239 (49)	199 (43)	<0.0001
Mentioned but not performed	19 (1)	3 (1)	3 (1)	13 (3)	
Performed but no details	604 (41)	190 (37)	207 (43)	207 (45)	
Performed and details given	105 (7)	23 (4)	36 (7)	46 (10)	
Blinding					
Not mentioned	891 (61)	371 (72)	294 (61)	226 (49)	<0.0001
Mentioned but not performed	37 (3)	10 (2)	5 (1)	22 (5)	
Performed but no details	179 (12)	42 (8)	62 (13)	75 (16)	
Performed and details given	359 (25)	93 (18)	124 (26)	142 (31)	