METHODOLOGICAL RIGOUR IN PRECLINICAL RESEARCH: IMPLICATIONS FOR ITS SCIENTIFIC VALIDITY AND BIOMEDICAL PROGRESS

FRANCISCO DANIEL RAMIREZ

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies in partial fulfillment of the requirements for the M.Sc. degree in Epidemiology

Epidemiology and Public Health

Faculty of Medicine

University of Ottawa

ABSTRACT

Preclinical research using animals often precedes and informs clinical trials; however, most attempts to translate findings from “bench-to-bedside” fail. There is growing concern that an important cause of failed translations is that much of preclinical research is not reproducible, with poor experimental methodology believed to be a major contributor. Four studies were conducted: (1) an assessment of reported study designs of preclinical experiments published in leading cardiovascular journals; (2) an examination of sex bias in preclinical cardiovascular research; (3) a comparison of experimental practices between male and female preclinical cardiovascular researchers; and (4) an analysis of the influence of journal initiatives on preclinical research quality. These studies suggest that (1) methodological shortcomings are prevalent and persistent in preclinical cardiovascular research; (2) women’s involvement in preclinical cardiovascular research is positively associated with considering sex as a biological variable; and (3) journals can exert considerable influence on the quality of published data.

EXECUTIVE SUMMARY

Research using animals is often used to inform and justify the testing of therapeutics and diagnostics in humans. This preclinical research therefore meaningfully influences research funding decisions, drug development, and ultimately patient care.[1] Given its importance and influence, ensuring that results from preclinical experiments are scientifically valid, precise, and reproducible through rigorous study design and reporting is essential to biomedical progress.[1-3] In clinical trials, key study design elements (SDEs) are widely accepted and expected to minimize the risk of bias and imprecision and potentially improve external validity.[4] For instance, randomization protects against selection bias; blinding protects against performance, detection, and attrition biases; sample size estimation ensures adequate statistical power, improving effect estimate precision and protecting against spurious conclusions; and including male and female animals and considering sex as a potentially important variable increase the likelihood that the results will be relevant to both men and women.[5-8] Although preclinical experiments are susceptible to similar risks of bias and threats to study validity, these SDEs are not routinely implemented in this phase of research. Influential scientific organizations, journals, and academics have focused considerable effort on improving the rigour of preclinical science given growing concerns that prevalent methodological shortcomings in animal research underlie problems with the reproducibility of results,[5, 8, 9] that poorly designed experiments result in animals being unjustifiably subjected to harm,[3, 6, 8] and that biased or spurious results from animal studies can inappropriately prompt or deter clinical trials in humans.[10-12]

Despite the above concerns, there are limited empirical data on the quality of published preclinical research (i.e. the extent of the problem) or on potential strategies to improve it. This

thesis sought to begin to address these gaps. Its main findings and recommendations can be summarized as follows:

1. Methodological shortcomings are prevalent in preclinical research. Despite increasing recognition among influential stakeholders within the research community of the importance and rationale for implementing key SDEs, research practices have largely remained unchanged, even in high-impact journals.
2. Preclinical cardiovascular researchers preferentially use male animals in experiments – a practice that is *increasing*. Given that data from experiments in animals are often used to inform clinical trials, excluding females in preclinical stages of research has the potential to undermine progress made in clinical research by biasing influential results toward male predominant patterns.
3. It has been proposed that the makeup of research teams can influence the quality of the research they produce, arguing for the value of diversity within teams. Researcher sex has been proposed as one such influential variable. Our data suggest that researcher sex is not associated with most measures of methodological rigour or with research impact. However, female researchers are more likely to study female animals and to consider sex as a biological variable when conducting and reporting experiments, even after adjusting for potential confounders.
4. It can be reasoned that journals are well positioned to influence the quality of preclinical research by acting as the gatekeepers of what is published. If so, the pressure to publish among academics, although generally viewed as a negative force, could be used to motivate researchers to accept and adhere to evolving experimental and reporting standards. Our data suggest that journal interventions that require preclinical researchers

to disclose key SDEs can effectively improve the reporting and potentially the methodological rigour of published research. Journals and their editors are therefore important stakeholders to engage in efforts to improve the transparency and rigour of preclinical research.

References

1. Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet. 2014;383(9912):166-75.
2. Henderson VC, Kimmelman J, Fergusson D, Grimshaw JM, Hackam DG. Threats to validity in the design and conduct of preclinical efficacy studies: A systematic review of guidelines for in vivo animal experiments. PLoS Med. 2013;10(7):e1001489.
3. van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, et al. Can animal models of disease reliably inform human studies? PLoS Med. 2010;7(3):e1000245.
4. Muhlhausler BS, Bloomfield FH, Gillman MW. Whole animal experiments should be more like human randomized controlled trials. PLoS Biol. 2013;11(2):e1001481.
5. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. Nature. 2014;505(7485):612-3.
6. Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, Fry D, et al. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PLoS One. 2009;4(11):e7824.
7. Clayton JA, Collins FS. Policy: NIH to balance sex in cell and animal studies. Nature. 2014;509(7500):282-3.

8. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature. 2012;490(7419):187-91.
9. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533(7604):452-4.
10. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012;483(7391):531-3.
11. Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. Stroke. 2008;39(10):2824-9.
12. Horn J, de Haan RJ, Vermeulen M, Luiten PG, Limburg M. Nimodipine in animal model experiments of focal cerebral ischemia: A systematic review. Stroke. 2001;32(10):2433-8.

CONTRIBUTION OF AUTHORS

Six manuscripts or letters were prepared and submitted for publication as part of this thesis. The student (FDR) was first or senior author on five submissions, for which he was responsible for study design, data collection, data analysis and interpretation, and manuscript preparation and revision. Given the size and scope of the datasets required, the student supervised and mentored several junior trainees who assisted predominantly with data collection but also with other aspects of study conduct (e.g. data formatting and cleaning, figure preparation, manuscript revision). Similarly, several manuscripts were completed through collaboration with senior researchers. These colleagues were offered authorship on relevant manuscripts commensurate with their contributions. Drs. Wells, Hibbert, and Birnie provided guidance and feedback on all projects.

ACKNOWLEDGEMENTS

I am deeply grateful to the following friends and colleagues without whom this work would not have been possible:

George A. Wells, PhD*†

Benjamin Hibbert, MD PhD†

David H. Birnie, MBChB MD†

Richard G. Jung, BSc (MD/PhD candidate)

Pouya Motazedian, MD (resident)

Pietro Di Santo, MD (resident, MSc candidate)

Robert Moreland, MD (fellow)

Zachary MacDonald, MD (resident)

Dylan Perry-Nguyen, BMSc (MD candidate)

Aisling A. Clancy, MD MSc (fellow, MPH candidate)

Alisha Labinaz (BSc candidate)

Steven Promislow, MD

Trevor Simard, MD (resident, PhD candidate)

Steeven Provencher, MD MSc

Sébastien Bonnet, PhD

Vivian A. Welch, PhD

Ian D. Graham, PhD

*Thesis supervisor

†Thesis advisory committee (TAC) members

LIST OF RELEVANT MANUSCRIPTS/LETTERS

1. **Ramirez FD**, Motazedian P, Jung RG, Di Santo P, MacDonald ZD, Moreland R, Simard T, Clancy AA, Russo JJ, Welch VA, Wells GA, Hibbert B. Methodological rigor in preclinical cardiovascular studies: targets to enhance reproducibility and promote research translation. *Circulation Research* 2017;120(12):1916-1926
2. **Ramirez FD**, Motazedian P, Jung RG, Di Santo P, MacDonald Z, Simard T, Clancy AA, Russo JJ, Welch V, Wells GA, Hibbert B. Sex bias is increasingly prevalent in preclinical cardiovascular research: implications for translational medicine and health equity for women. A systematic assessment of leading cardiovascular journals over a 10-year period. *Circulation* 2017;135(6):625-626
3. Provencher S, Archer SL, **Ramirez FD**, Hibbert B, Paulin R, Boucherat O, Lacasse Y, Bonnet S. Standards and methodological rigor in pulmonary arterial hypertension preclinical and translational research. *Circulation Research* 2018;122(7):1021-1032
4. **Ramirez FD**, Hibbert B. Letter by Ramirez and Hibbert regarding article, "Consideration of sex differences in design and reporting of experimental arterial pathology studies: a statement from the Arteriosclerosis, Thrombosis, and Vascular Biology Council". *Arteriosclerosis, Thrombosis, and Vascular Biology* 2018;38(6):e99-100
5. Labinaz A, Marbach JA, Jung R, Moreland R, Motazedian P, Di Santo P, Clancy AA, MacDonald Z, Simard T, Hibbert B, **Ramirez FD**. Female authorship in preclinical cardiovascular research: temporal trends and influence on experimental design. *JACC Basic to Translational Science* (Accepted)
6. **Ramirez FD**, Jung R, Motazedian P, Perry-Nguyen D, Di Santo P, MacDonald Z, Clancy AA, Labinaz A, Promislow S, Simard T, Provencher S, Bonnet S, Graham ID, Wells GA,

Hibbert B. Journal initiatives to enhance preclinical research: analyses of Stroke, Nature Medicine, Science Translational Medicine. (Submitted)

This work is dedicated to Evelyn and Caleb

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

ABBREVIATIONS

AHA: American Heart Association

ARRIVE: Animal Research: Reporting of *In Vivo* Experiments

ATVB: *Arteriosclerosis, Thrombosis, and Vascular Biology*

CI: confidence interval

CIHR: Canadian Institutes of Health Research

Circ Res: *Circulation Research*

CM/HF: cardiomyopathy/heart failure

CVD: cardiovascular disorder/disease

IICARus: Intervention to Improve Compliance with the ARRIVE guidelines

IQR: interquartile range

MI: myocardial infarction

NIH: National Institutes of Health

NPQIP: Nature Publication Quality Improvement Project

OR: odds ratio

PH: pulmonary hypertension

R&D: research and development

RR: relative risk

SDE: study design element

Science TM: *Science Translational Medicine*

CHAPTER 1: INTRODUCTION

1.1 Outline

This thesis has been prepared in a manuscript-based format. Five manuscripts or letters are included, four of which form chapters (Chapters 3-6), in addition to a Background chapter (Chapter 2) and a Discussion and Recommendations chapter (Chapter 7). The remaining letter is included as Appendix I.[1]

1.2 Objectives

Our objectives were as follows:

1. To determine the prevalence of important methodological shortcomings in preclinical cardiovascular research, including sex bias in animal models.
2. To identify proven or promising interventions to improve the reporting and methodological quality of preclinical research.

1.3 Chapter summaries and hypotheses

Chapter 1: Introductory chapter providing an overview of the organization of the thesis and of chapter objectives.

Chapter 2: Background information on the relevance of preclinical experiments to clinical research and patient care, the reproducibility "crisis" in preclinical research, the rationale for including animals of both sexes in preclinical experiments, factors associated with methodological rigour in preclinical studies, and relevant interventions that have been attempted.

Chapter 3: Manuscript of an analysis of the extent of methodological shortcomings and temporal patterns in experimental practices in preclinical cardiovascular studies.[2] We hypothesized that methodological shortcomings were prevalent and not improving. *Post hoc* analyses motivated the study presented in Chapter 6.

Chapter 4: Manuscript of an analysis of sex bias in preclinical cardiovascular experiments and a discussion on its potential impact on women's health.[3] We hypothesized that preclinical cardiovascular researchers preferentially studied male animals.

Chapter 5: Manuscript of a study of the association of female authorship with measures of methodological rigour and research impact in preclinical cardiovascular research. We hypothesized that male and female preclinical cardiovascular researchers would not differ in any measure of experimental rigour or research impact.

Chapter 6: Submitted manuscript of longitudinal analyses of the potential impact of journal initiatives designed to improve reporting and methodological quality in preclinical research. We hypothesized that journal initiatives were effective in improving the reporting and methodological quality of published reports.

Chapter 7: Summary of main study findings and discussion of their implications.

References

1. Ramirez FD, Hibbert B. Letter by Ramirez and Hibbert regarding article, "Consideration of sex differences in design and reporting of experimental arterial pathology studies: A statement

from the Arteriosclerosis, Thrombosis, and Vascular Biology Council". Arterioscler Thromb Vasc Biol. 2018;38(6):e99-e100.

2. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald ZD, Moreland R, et al. Methodological rigor in preclinical cardiovascular studies: Targets to enhance reproducibility and promote research translation. Circ Res. 2017;120(12):1916-26.
3. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald Z, Simard T, et al. Sex bias is increasingly prevalent in preclinical cardiovascular research: Implications for translational medicine and health equity for women: A systematic assessment of leading cardiovascular journals over a 10-year period. Circulation. 2017;135(6):625-6.

CHAPTER 2: BACKGROUND

2.1 Relevance of preclinical experiments to clinical research and patient care

Preclinical research and clinical therapy development exist in a symbiotic continuum but are influenced by discordant reporting standards, regulatory forces, and reward systems.[1, 2] In clinical trials, key study design elements are routinely incorporated to reduce the risk of bias and to ensure effect estimate precision, such as randomization, blinding, and power calculations. However, these study design elements (SDEs) are rarely implemented or reported in preclinical research involving non-human animals, even though it often precedes and informs clinical trials. Discrepant levels of methodological rigour between these phases of research could compromise the scientific validity of preclinical research,[3] contribute to failed attempts to translate findings from 'bench-to-bedside', and ultimately hinder progress in patient care.[4, 5]

2.2 Reproducibility "crisis"

Low methodological rigour in preclinical experiments has been associated with less precise and more extreme estimates of therapeutic effect than those reported in more rigorously performed experiments, suggesting that poorly conducted studies are at greater risk of bias and therefore less easily reproduced.[6-12] Preclinical methodological rigour and adequate reporting have therefore been advocated by prominent scientific bodies and leading journals as essential to improving reproducibility, including by the National Institutes of Health (NIH),[2] the National Institute of Neurological Disorders and Stroke (NINDS),[13] the Stroke Therapy Academic Industry Roundtable (STAIR),[14, 15] *Nature*,[16] *Science*,[17] *Cell*, and all American Heart Association journals (including *Circulation* and *Circulation Research*),[18] among many others (the Canadian

Institutes of Health Research [CIHR] have not clearly expressed a position on this issue). These concerns and calls to action have been echoed by leading experts in the field[1, 3, 19-21] and by the biopharmaceutical industry,[12, 22] the latter of which undeniably plays an important role in translating preclinical research into clinical therapies. Both Bayer HealthCare and Amgen have reported high levels of drug development failure directly attributed to an inability to reproduce experimental findings.[12, 22] Their message that the reliability of preclinical research is poor and hindering drug development is echoed by aggregate data on attrition rates throughout the drug development process from prominent pharmaceutical companies, which show that attrition at phase 2 of clinical trials is where most drugs fail once clinical testing is pursued. Only 1 in 4 compounds that enter phase 2 proceed to phase 3 clinical trials.[5] This is relevant because phase 2 trials are considered clinical 'proof-of-concept' studies that aim to demonstrate that a drug has the expected biological effect in humans. High failure rates at this stage therefore suggest poor biological target/therapeutic selection[5] and have fostered a growing perception that published preclinical research poorly predicts what is seen in clinical trials, thereby damaging investigators' confidence in animal model data to identify promising therapies for testing in humans.[5, 12, 22, 23] Furthermore, 90% of scientists believe that there exists a reproducibility "crisis" in research with >90% believing that irreproducibility is attributable to low statistical power or poor analysis and nearly 90% to poor experimental design (based on the results of a *Nature* survey of more than 1,500 researchers).[24] As well, nearly 90% of researchers in the survey selected "more robust experimental design" as an approach to improving reproducibility. Thus, despite limited direct data, the preclinical research community links experimental methods with reproducibility based on an understanding of the fundamental biases that study design

elements such as randomization and blinding are designed to prevent[13, 25, 26] and on robust albeit indirect evidence of the impact of fundamental aspects of experimental design.

2.3 Inclusion of female animals in preclinical research

Since the 1980's, there has been an increasing recognition that fundamental yet poorly understood differences exist between men and women at biological, psychological, and sociological levels and that these differences could engender disparities in health if ignored. The landmark 1985 Report of the Public Health Service Task Force on Women's Health Issues specifically identified sex and gender bias in research as a barrier to advancing women's health, noting that data obtained from studies using only male subjects may not be applicable to the female population.[27] The NIH Revitalization Act of 1993 accordingly mandated that women be included in NIH-funded clinical trials with accompanying guidelines cautioning that "increased attention, therefore, must be given to gender… in earlier stages of research to allow for informed decisions at the Phase III clinical trial stage" (https://grants.nih.gov/grants/guide/notice-files/not94-100.html). Considerable effort has since been focused on ensuring that women are adequately represented in clinical trials with undeniable results: women now account for more than half of NIH-funded clinical trial participants and our understanding of the impact of sex and gender on the natural history of many diseases and on the effectiveness of therapies has markedly improved.[28-30] In contrast, this "revolution" has conspicuously not permeated preclinical stages of research.[31-39] In response, the NIH announced in 2014 that it would require that sex be considered as a biological variable in applications for preclinical research funding.[31, 40] Similarly, CIHR now expects that research applications integrate sex and gender in research designs, when appropriate, and refers applicants to several relevant NIH guidance documents and

reporting guidelines (http://www.cihr-irsc.gc.ca/e/50836.html). Yet, there continues to be a perceived reluctance to include female animal models in preclinical experiments[41] and little evidence that reporting guidelines are effecting change.[42] Inadequate female representation in preclinical research has the potential to disadvantage women by skewing our understanding of disease processes toward male-predominant patterns and by reducing the likelihood of female-specific therapeutics advancing to the clinical realm.[43]

2.4 Factors associated with methodological rigour

Limited data regarding factors that influence experimental rigour in preclinical research exist. However, several factors are potentially important, including field of research,[44] journal of publication,[45, 46] and characteristics of the research team.[47] Journals act as ultimate gatekeepers of the published literature and are therefore particularly well-positioned to influence its quality through editorial culture and policies. Few studies have examined this possibility and most have had important methodological limitations.[45, 48, 49] Research team composition, including diversity with regards to sex, have also been proposed. The notion that men and women approach research differently and that these differences can impact the quality of the research produced is held by many academics and scientific organizations. In fact, the NIH has explicitly identified assessing the impact of diversity within research teams on the quality and outputs of their research as the first of four "major diversity challenges facing the biomedical ecosystem" [50]. The Director of the NIH has argued that diversity in the "biomedical research workforce" (including better representation of women) is "vital to harnessing the complete intellectual capital of the nation" [50]. There is a dedicated Scientific Workforce Diversity Office at the NIH whose strategic plan is titled "Great Minds Think Differently" and which highlights their commitment to diversity as a

means of increasing innovation and broadening scopes of research inquiry (https://diversity.nih.gov/sites/coswd/files/images/2018-06/SWD_StrategicPlan_layout_final_links-508c.pdf). It is argued that researcher sex is associated with different viewpoints, research questions, and research areas of interest [47]. Meaningfully different practice patterns and patient outcomes have been reported between the sexes in clinical medicine, for instance [51, 52], in part attributed to these potential differences.

2.5 Interventions

As alluded to above, in part because of concerns that there exists a reproducibility "crisis" in research, numerous influential scientists, organizations, and journals have advocated for change or enacted initiatives to promote methodological rigour and transparency through the development of guidelines, funding requirements, and submission checklists. For instance, the ARRIVE (Animals in Research: Reporting *In Vivo* Experiments) guidelines were introduced in 2010 and are widely endorsed;[39] however, enforcement has been suboptimal among journals.[42] The NIH and CIHR additionally endorse the SAGER (Sex and Gender Equity in Research) reporting guidelines,[53] but few journals have formally endorsed them (http://www.ease.org.uk/communities/gender-policy-committee/gpc-endorse-adopt-collab/). Based on reporting guidelines developed by NINDS, the NIH has proposed a set of guidelines and funding policies aimed at improving the reproducibility of preclinical research.[54] These guidelines have been endorsed by numerous influential journals, but their impact has not been adequately assessed. Several journals, including influential publishing groups such as *Nature*[16] and *Science*[17, 55] have instituted editorial policies designed to improve the quality of their published findings, but their effectiveness is similarly not well established. Prospective

registration and public deposition of preclinical studies and their results have been proposed[56] (analogous to systems such as clinicaltrials.gov for clinical trials) but have not gained widespread acceptance. Therefore, although various interventions have been proposed to address what are believed to be widespread methodological and reporting shortcomings in preclinical research, there remain little data on the extent of the problem or on how to address it.

References

1. Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet. 2014;383(9912):166-75.
2. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. Nature. 2014;505(7485):612-3.
3. Henderson VC, Kimmelman J, Fergusson D, Grimshaw JM, Hackam DG. Threats to validity in the design and conduct of preclinical efficacy studies: A systematic review of guidelines for in vivo animal experiments. PLoS Med. 2013;10(7):e1001489.
4. Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical R&D. Nat Rev Drug Discov. 2011;10(6):428-38.
5. Bunnage ME. Getting pharmaceutical R&D back on target. Nat Chem Biol. 2011;7(6):335-9.
6. Macleod MR, O'Collins T, Howells DW, Donnan GA. Pooling of animal experimental data reveals influence of study design and publication bias. Stroke. 2004;35(5):1203-8.
7. O'Collins VE, Macleod MR, Donnan GA, Horky LL, van der Worp BH, Howells DW. 1,026 experimental treatments in acute stroke. Ann Neurol. 2006;59(3):467-77.

8. Macleod MR, O'Collins T, Horky LL, Howells DW, Donnan GA. Systematic review and metaanalysis of the efficacy of FK506 in experimental stroke. J Cereb Blood Flow Metab. 2005;25(6):713-21.
9. Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. Stroke. 2008;39(10):2824-9.
10. Hirst JA, Howick J, Aronson JK, Roberts N, Perera R, Koshiaris C, et al. The need for randomization in animal trials: An overview of systematic reviews. PLoS One. 2014;9(6):e98856.
11. Bebarta V, Luyten D, Heard K. Emergency medicine animal research: Does use of randomization and blinding affect the results? Acad Emerg Med. 2003;10(6):684-7.
12. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012;483(7391):531-3.
13. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature. 2012;490(7419):187-91.
14. Stroke Therapy Academic Industry R. Recommendations for standards regarding preclinical neuroprotective and restorative drug development. Stroke. 1999;30(12):2752-8.
15. Fisher M, Feuerstein G, Howells DW, Hurn PD, Kent TA, Savitz SI, et al. Update of the Stroke Therapy Academic Industry Roundtable preclinical recommendations. Stroke. 2009;40(6):2244-50.
16. Journals unite for reproducibility. Nature. 2014;515(7525):7.
17. McNutt M. Journals unite for reproducibility. Science. 2014;346(6210):679.

18. National Institutes of Health. Principles and guidelines for reporting preclinical research. [Available from: https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research].
19. Begley CG, Ioannidis JP. Reproducibility in science: Improving the standard for basic and preclinical research. Circ Res. 2015;116(1):116-26.
20. Begley CG. Six red flags for suspect work. Nature. 2013;497(7450):433-4.
21. Perel P, Roberts I, Sena E, Wheble P, Briscoe C, Sandercock P, et al. Comparison of treatment effects between animal experiments and clinical trials: Systematic review. BMJ. 2007;334(7586):197.
22. 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(9):712.
23. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124.
24. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533(7604):452-4.
25. van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, et al. Can animal models of disease reliably inform human studies? PLoS Med. 2010;7(3):e1000245.
26. Muhlhausler BS, Bloomfield FH, Gillman MW. Whole animal experiments should be more like human randomized controlled trials. PLoS Biol. 2013;11(2):e1001481.
27. Women's health. Report of the public health service task force on women's health issues. Public Health Rep. 1985;100(1):73-106.
28. United States General Accounting Office. Women's health. NIH has increased its efforts to include women in research. Washington, DC. 2000.
29. United States General Accounting Office. Women's health. Women sufficiently represented in new drug testing, but FDA oversight needs improvement. Washington, DC. 2001.

30. United States Government Accountability Office. National Institutes of Health. Better oversight needed to help ensure continued progress including women in health research. Washington, DC. 2015.
31. Clayton JA, Collins FS. Policy: NIH to balance sex in cell and animal studies. Nature. 2014;509(7500):282-3.
32. Putting gender on the agenda. Nature. 2010;465(7299):665.
33. Check Hayden E. Sex bias blights drug studies. Nature. 2010;464(7287):332-3.
34. Wald C, Wu C. Biomedical research. Of mice and women: The bias in animal models. Science. 2010;327(5973):1571-2.
35. Zucker I, Beery AK. Males still dominate animal studies. Nature. 2010;465(7299):690.
36. Miller VM. In pursuit of scientific excellence: Sex matters. Am J Physiol Endocrinol Metab. 2012;302(9):E1025-6.
37. Shah K, McCormack CE, Bradbury NA. Do you know the sex of your cells? Am J Physiol Cell Physiol. 2014;306(1):C3-18.
38. Yoon DY, Mansukhani NA, Stubbs VC, Helenowski IB, Woodruff TK, Kibbe MR. Sex bias exists in basic science and translational surgical research. Surgery. 2014;156(3):508-16.
39. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. PLoS Biol. 2010;8(6):e1000412.
40. Clayton JA. Studying both sexes: A guiding principle for biomedicine. FASEB J. 2016;30(2):519-24.
41. Mogil JS. Perspective: Equality need not be painful. Nature. 2016;535(7611):S7.

42. Baker D, Lidster K, Sottomayor A, Amor S. Two years later: Journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. PLoS Biol. 2014;12(1):e1001756.
43. Klein SL, Schiebinger L, Stefanick ML, Cahill L, Danska J, de Vries GJ, et al. Opinion: Sex inclusion in basic research drives discovery. Proc Natl Acad Sci U S A. 2015;112(17):5257-8.
44. Sugimoto CR, Ahn YY, Smith E, Macaluso B, Lariviere V. Factors affecting sex-related reporting in medical research: A cross-disciplinary bibliometric analysis. Lancet. 2019;393(10171):550-9.
45. The NPQIP Collaborative group. Did a change in Nature journals' editorial policy for life sciences research improve reporting? BMJ Open Science. 2019;3:e000035.
46. Minnerup J, Zentsch V, Schmidt A, Fisher M, Schabitz WR. Methodological quality of experimental stroke studies published in the Stroke journal: Time trends and effect of the basic science checklist. Stroke. 2016;47(1):267-72.
47. Nielsen MW, Alegria S, Borjeson L, Etzkowitz H, Falk-Krzesinski HJ, Joshi A, et al. Opinion: Gender diversity leads to better science. Proc Natl Acad Sci U S A. 2017;114(8):1740-2.
48. Han S, Olonisakin TF, Pribis JP, Zupetic J, Yoon JH, Holleran KM, et al. A checklist is associated with increased quality of reporting preclinical biomedical research: A systematic review. PLoS One. 2017;12(9):e0183591.
49. Hair K, Macleod M, Sena E, Collaboration I. A randomised controlled trial of an intervention to improve compliance with the arrive guidelines (IICARus). BioRxiv. 2018.

50. Valantine HA, Collins FS. National Institutes of Health addresses the science of diversity. Proc Natl Acad Sci U S A. 2015;112(40):12240-2.
51. Tsugawa Y, Jena AB, Figueroa JF, Orav EJ, Blumenthal DM, Jha AK. Comparison of hospital mortality and readmission rates for medicare patients treated by male vs female physicians. JAMA Intern Med. 2017;177(2):206-13.
52. Wallis CJ, Ravi B, Coburn N, Nam RK, Detsky AS, Satkunasivam R. Comparison of postoperative outcomes among patients treated by male and female surgeons: A population based matched cohort study. BMJ. 2017;359:j4366.
53. Heidari S, Babor TF, De Castro P, Tort S, Curno M. Sex and gender equity in research: Rationale for the SAGER guidelines and recommended use. Res Integr Peer Rev. 2016;1:2.
54. National Institutes of Health. Principles and guidelines for reporting preclinical research [Available from: https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research].
55. Kelner KL. Playing our part. Sci Transl Med. 2013;5(190):190ed7.
56. Kimmelman J, Anderson JA. Should preclinical studies be registered? Nat Biotechnol. 2012;30(6):488-9.

CHAPTER 3: METHODOLOGICAL RIGOUR IN PRECLINICAL CARDIOVASCULAR STUDIES

The following is a manuscript published in *Circulation Research* examining reported experimental practices in preclinical cardiovascular research. The objective was to determine the extent of methodological rigour in this field and to examine its association with citation counts as a measure of research impact. These results were cited as a driver of new initiatives at *Circulation Research* intended to improve the rigour and reproducibility of its preclinical research publications.

Additional documentation related to this manuscript is provided in the following appendices:

Appendix A: Published article in *Circulation Research*.

Appendix B: Editorial by Schäbitz *et al* in *Stroke*.

Appendix C: Editorial by Jones in *Circulation Research*.

Appendix D: Basic Cardiovascular Sciences Scientific Sessions 2017 article describing "the next best thing in cardiovascular research" in *Journal of the American Heart Association*.

Appendix E: Article in *The Scientist* reporting on the study's findings.

Appendix F: Editorial by the former editor-in-chief of *Circulation Research* highlighting the study's findings as a driver of new initiatives being adopted by the journal to improve the quality of its published research.

Appendix G: Published article on standards and methodological rigour in preclinical pulmonary arterial hypertension research in *Circulation Research* (lead author: Dr. Steeve Provencher)

Methodological rigor in preclinical cardiovascular studies: targets to enhance reproducibility and promote research translation

F. Daniel Ramirez MD,[a,b,c] Pouya Motazedian BSc,[b] Richard G. Jung BSc,[b,d,e] Pietro Di Santo MD,[a,b] Zachary MacDonald BSc,[b] Robert Moreland MD,[b,f] Trevor Simard MD,[a,b,d,e] Aisling A. Clancy MD MSc,[g] Juan J. Russo MD,[a,b] Vivian Welch PhD,[c,h,i] George A. Wells PhD,[c,j] Benjamin Hibbert MD PhD[a,b,d,e]

[a] Division of Cardiology, University of Ottawa Heart Institute, Ottawa, Ontario, Canada

[b] CAPITAL Research Group, University of Ottawa Heart Institute, Ottawa, Ontario, Canada

[c] School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Ontario, Canada

[d] Vascular Biology and Experimental Medicine Laboratory, University of Ottawa Heart Institute, Ottawa, Ontario, Canada

[e] Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, Ontario, Canada

[f] Department of Radiology, University of Ottawa, Ottawa, Ontario, Canada

[g] Department of Obstetrics and Gynecology, University of Ottawa, Ottawa, Ontario, Canada

[h] Bruyère Research Institute, University of Ottawa, Ottawa, Ontario, Canada

[i] Centre for Global Health, University of Ottawa, Ottawa, Ontario, Canada

[j] Cardiovascular Research Methods Centre, University of Ottawa Heart Institute, Ottawa, Ontario, Canada

Short title: Methodological rigor in preclinical research

Total manuscript word count: 7,174

Subject codes: animal models of human disease, translational studies, basic science research

Corresponding author:

Dr. Benjamin Hibbert

University of Ottawa Heart Institute

40 Ruskin Street, H-4238

Ottawa, Ontario, Canada K1Y 4W7

Tel.: (613) 761-3059

Fax: (613) 761-4874

Email: bhibbert@ottawaheart.ca

Abstract

Rationale: Methodological sources of bias and suboptimal reporting contribute to irreproducibility in preclinical science and may negatively impact research translation. Randomization, blinding, sample size estimation, and considering sex as a biological variable are deemed crucial study design elements to maximize the quality and predictive value of preclinical experiments.

Objective: To examine the prevalence and temporal patterns of recommended study design element implementation in preclinical cardiovascular research.

Methods and Results: All articles published over a 10-year period in 5 leading cardiovascular journals were reviewed. Reports of *in vivo* experiments in non-human mammals describing pathophysiology, genetics, and/or therapeutic interventions relevant to specific cardiovascular disorders were identified. Data on study design and animal model use were collected. Citations at 60 months were additionally examined as a surrogate measure of research impact in a pre-specified subset of studies, stratified by individual and cumulative study design elements. Of 28,636 articles screened, 3,396 met inclusion criteria. Randomization was reported in 21.8%, blinding in 32.7%, and sample size estimation in 2.3%. Temporal and disease-specific analyses show that the implementation of these study design elements has overall not appreciably increased over the past decade, except in preclinical stroke research, which has uniquely demonstrated significant improvements in methodological rigor. In a subset of 1,681 preclinical studies, randomization, blinding, sample size estimation, and inclusion of both sexes were not associated with increased citations at 60 months.

Conclusions: Methodological shortcomings are prevalent in preclinical cardiovascular research, have not substantially improved over the past 10 years, and may be overlooked when basing subsequent studies. Resultant risks of bias and threats to study validity have the potential to hinder progress in cardiovascular medicine as preclinical research often precedes and informs clinical trials. Stroke research quality has uniquely improved in recent years, warranting a closer examination for interventions to model in other cardiovascular fields.

Keywords: experimental models, research methodology, reproducibility, bias, cardiovascular research

Novelty and significance

What is known?

- Preclinical research often precedes and informs clinical trials
- Randomization, blinding, sample size estimation, and considering sex as a biological variable are considered crucial study design elements to maximize the predictive value of preclinical experiments

What new information does this article contribute?

- Key study design elements are rarely implemented in preclinical cardiovascular research
- The implementation of these elements has not appreciably improved over the past decade, except in stroke research
- Methodological shortcomings in preclinical experiments may be overlooked when researchers design subsequent studies

Preclinical and clinical research are guided by different methodological standards. Resultant discrepancies in study designs can compromise scientific validity, give rise to experimental irreproducibility, and may contribute to failures in research translation. Randomization, blinding, sample size estimation, and considering sex as a biological variable have been identified as key study design elements to promote rigor and reproducibility of animal experiments and research relevance to both men and women. However, there exist limited data on the extent of suboptimal methodological rigor in preclinical research. We identified and analyzed 3,396 preclinical studies in leading cardiovascular journals over a 10-year period, including citation counts at 60 months as a surrogate measure of research impact for those published in the first 5 years. We show that

these design elements were rarely implemented – a situation that has not substantially improved over the past decade, except in stroke research. The implementation of these design elements was also not associated with increased citations (individually or cumulatively). Together, these data indicate that methodological shortcomings are prevalent in preclinical cardiovascular research and suggest that they may be overlooked when researchers design subsequent studies. Stroke research warrants a closer examination for interventions to model in other cardiovascular fields.

Non-standard Abbreviations and Acronyms

AHA: American Heart Association

ATVB: *Arteriosclerosis, Thrombosis and Vascular Biology*

Circ Res: *Circulation Research*

CM/HF: cardiomyopathy/heart failure

CVD: cardiovascular disorder

MI: myocardial infarction

NIH: National Institutes of Health

R&D: research and development

Introduction

Preclinical research and clinical therapy development exist in a symbiotic continuum yet are guided by different reporting standards, regulatory forces, and reward systems.[1, 2] Consequent discrepancies in study designs and analytical methods can compromise scientific validity,[3] result in irreproducible results (particularly from preclinical experiments using animal models[2]), and may contribute to the high rate of attrition seen in early stages of clinical development.[4, 5] Improved methodological rigor and transparent reporting practices have therefore been advocated to improve the predictive value of animal model data.[2, 6-10]

The issue of irreproducibility in preclinical research and its impact are well-recognized in academic circles[2, 11-14] and increasingly so among the lay public[15-17] and the biopharmaceutical industry.[13, 18] In 2011, Bayer HealthCare published an analysis of 67 in-house target identification and validation projects in the fields of oncology, women's health, and cardiovascular disease over a 4-year period, reporting that the published data were reproducible in less than one third of cases and that in nearly two thirds inconsistencies either prolonged the validation process or led to project termination.[11] In 2012, Amgen similarly published the results of their attempt to reproduce 53 "landmark" studies in the fields of hematology and oncology, succeeding in only 6 cases even after contacting the original authors for guidance, exchanging reagents, and in certain instances repeating experiments in the laboratory of the original investigators.[12] Some findings could not be reproduced even by the original investigators in their own laboratories when the experiments were repeated in a blinded fashion.[19] Disconcertingly, some of the irreproducible research was deemed to have prompted clinical studies. Others have likewise described candidate therapies that have advanced to testing in humans despite irreproducible or inconsistent results in animals.[8, 20, 21]

Since the mid-1990's, pharmaceutical research and development (R&D) productivity has declined with decreasing numbers of medications approved, increasing attrition rates at all stages of research, and lengthening development times despite rising expenditures.[4] This decline has been attributed in part to flawed preclinical research methodology, highlighting the importance of preclinical research for advancing clinical care, but equally portending a waning confidence in animal model data to identify promising therapeutic targets.[5, 7, 11, 12, 22, 23] Cardiovascular R&D, in particular, has fared poorly in recent years relative to other disease categories, exhibiting the greatest decline as a percentage of total projects[4] and among the lowest success rates and likelihood of approval at all phases of clinical trials.[24] Moreover, R&D investment patterns in general are increasingly deterring incremental innovation (i.e. improving upon available effective therapies), concentrating instead on novel disease targets with higher revenue potential but at higher risk of failure (e.g. specific cancers).[4] Poor reproducibility in preclinical research coupled with this increasingly unfavorable landscape for successful cardiovascular therapy development could therefore substantially hinder progress in cardiovascular care.

In response to the above systematic issues and in collaboration with editors from major journals, funding agencies, and scientific leaders, the National Institutes of Health (NIH) proposed a set of reporting guidelines and funding policies to improve the reproducibility and rigor of preclinical research.[25] A core set of standards first proposed by the NIH's National Institute of Neurological Disorders and Stroke were adopted, which identified randomization, blinding, sample size estimation, and data handling as minimum reporting requirements to promote transparency in animal studies.[7] In addition, the NIH announced that it would require that sex be considered as a biological variable in applications for preclinical research funding.[26] These principles and guidelines have been endorsed by prominent academic societies, associations, and journals,

including all American Heart Association (AHA) journals,[25] with evidence of editorial commitment to complying.[27, 28] However, there are limited data on the extent of suboptimal methodological rigor and/or incomplete reporting in preclinical cardiovascular science and therefore no baseline from which to gauge progress. A commitment to improving the quality and impact of preclinical research and to maintaining the trust of public and private stakeholders requires transparency on current and future states of scientific practice.[13] We therefore reviewed all preclinical cardiovascular studies published in leading AHA journals over the last decade to determine the prevalence of randomization, blinding, sample size estimation, and inclusion of both sexes as well as trends in these practices over time.

Methods

As previously described,[29] all preclinical cardiovascular studies published in AHA journals with archives spanning at least 10 years were reviewed. Five journals met these criteria: *Circulation*; *Circulation Research*; *Hypertension*; *Stroke*; and *Arteriosclerosis, Thrombosis, and Vascular Biology* (*ATVB*). All reports published over a 10-year period (July 2006 to June 2016) were screened. Studies were included if they represented original research published as full manuscripts; reported results of *in vivo* experiments in non-human mammals; and described pathophysiology, genetics, and/or therapeutic interventions that was/were stated to be directly relevant to a specific cardiovascular disorder (CVD) in humans (see pre-specified list of CVDs below). Studies on physiologic and/or genetic characteristics were included if potential therapeutic applications or implications of the study findings were proposed in the article. These criteria are consistent with previously proposed definitions of preclinical ('confirmatory' or 'proof-of-concept') experiments.[7, 30, 31] CVDs of interest included atherosclerosis and/or vascular

homeostasis, arterial aneurysms and/or dissections, myocardial infarction (MI), valvular disease, cardiomyopathy and/or heart failure, cardiac transplantation, pulmonary hypertension, cardiac arrhythmia, stroke, resuscitative medicine, hypertension, metabolic or endocrine diseases, and hematological disorders (including thrombosis). Studies deemed to report on clinically relevant cardiovascular conditions but not falling into one of the pre-specified categories could be included at the journal reviewers' discretion (category 'other'). Each journal article was independently reviewed and data extracted using standardized case report forms by 2 reviewers, allowing for assessments of interrater agreement. To permit our team to review a large volume of articles, we allowed for studies to be excluded as soon as it was clear that a manuscript violated *any* of our inclusion criteria, which was most often because of the model used (e.g. zebrafish or humans) or because they were published in formats other than full manuscripts (e.g. conference abstracts). The specific inclusion criterion/criteria that was/were deemed to have been violated for each excluded article therefore varied according to journal reviewer and were often multiple (e.g. a non-mammalian animal model was used and no specific reference to therapeutic implications/applications was made). Discrepancies were resolved by consensus or by an independent adjudicator before building a final locked database for analysis.

Pre-specified data, including the date of publication, CVD investigated, animal model(s) used and their sex, whether animals were randomized to treatment groups, whether any blinding was implemented (concealed allocation or blinded outcome assessment), and whether *a priori* sample size/power estimations were performed were collected. Subgroup analyses restricted to studies of therapeutic interventions and by CVD studied were performed as were *post hoc* comparisons of these practices before and after the publication of NIH guidelines and policies for reporting preclinical research and the implementation of a 'Basic Science checklist' by *Stroke*,[28] which is

purported to have improved the quality and designs of preclinical studies published in that journal.[32]

Finally, the number of citations has been used as a surrogate measure of research impact and influence.[33-37] Therefore, Scopus (Elsevier) was queried to identify original research articles that cited preclinical cardiovascular studies published between July 2006 and June 2011. This 5-year period was selected as it ensured that each article had ≥5 years of citation data available in June 2016. Reviews, conference papers, editorials, letters, books/book chapters, and errata were excluded. Scopus was selected as it is the largest database available for citation analysis and retrieves a greater number of citations when compared to others.[38] Pre-specified analyses of citations at 60 months following the index publication were performed, stratified by individual and cumulative study design elements and adjusted for journal of publication, year of publication, and CVD studied. Given the possibility that certain publications may have garnered attention in the short-term but could have ultimately been disproven or deemed irreproducible, a sensitivity analysis of citation counts at 36 months was also undertaken as this timepoint approximates contemporary mean time-to-retraction.[39]

Categorical variables are reported as number (%) and were compared via chi-square tests. Continuous variables are reported as median (IQR) and were compared using Wilcoxon signed-rank or Kruskal-Wallis tests. Interrater agreement was calculated using Cohen's kappa statistic and percent agreement. Temporal patterns in the proportions of studies reporting randomization, blinding, and sample size estimations were evaluated via Cochrane-Armitage trend tests and journal-specific logistic regression models adjusting for CVD studied and animal model used when the number of events per predictor variable was adequate.[40-42] For the latter adjusted analyses, backward elimination was used for model building using a criterion of $P<0.20$ for

specific CVD and animal model predictor variable inclusion. Associations between study design elements and citation counts were examined via stratification and multivariable linear regression. Non-normally distributed variables were log-transformed, when required. All analyses were performed using SAS 9.4 (SAS Institute Inc., Cary, NC) using a two-tailed alpha level of 0.05 to define statistical significance.

Results

Study selection and characteristics

As previously described,[29] of 28,636 articles screened, 3,396 met inclusion criteria and were analyzed (**Figure 1**). Interrater agreement for study inclusion prior to resolution was 94.5% (κ=0.72, 95% CI 0.70-0.73). Atherosclerosis, hypertension, stroke, and cardiomyopathy/heart failure were the most commonly studied CVDs (range 14.2-19.5%). Most other CVDs were the focus of <5.0% of studies. Ten studies on resuscitative medicine were identified and were included in the 'other' category, which comprised 3.0% of studies. Nearly one third of studies examined a therapeutic intervention. Mice or rats were most often used by researchers (used in 89.8% of studies) whereas guinea pigs, gerbils, or hamsters were used in <0.2%. Multiple animal models were used in 2.8% of studies (**Table**).

Randomization

Randomization of animals was reported in 740 studies (21.8%) overall, but was more frequently noted in the subset examining therapeutic interventions (38.3% vs. 15.0%, P<0.0001). When

studies were stratified by CVD studied, significant differences in the proportions reporting randomization were noted (P<0.0001, range 5.6-46.5%) with a lack of randomization predominating in all cases except stroke (**Figure 2A**). Significant differences in the proportions of studies reporting randomization were also observed when stratified by animal model used (P<0.0001, range 14.6-45.0%) with a lack of randomization predominating in all cases except when pigs or a combination of animal models were employed (**Figure 2B**).

Blinding

Blinding of treatment allocation or outcome assessment was reported in 1,110 studies (32.7%) overall, but was also more frequently noted in the subset examining therapeutic interventions (41.9% vs. 28.9%, P<0.0001). When stratified by CVD studied, significant differences in the proportions of studies reporting any blinding were observed (P<0.0001, range 10.9-62.6%). Studies on stroke ranked highest among CVDs though arterial aneurysms/dissections and transplantation medicine had comparable numbers of blinded and non-blinded studies. For all other CVDs, studies with blinding formed the minority (**Figure 3A**). No significant difference in the proportions of studies reporting blinding were seen when stratified by animal model used (P=0.369, range 20.9-36.8%) (**Figure 3B**).

Sample size estimation

A priori sample size estimations and/or power calculations were reported in 79 studies (2.3%). This aspect of study design was also more frequently reported in the subset examining therapeutic interventions (4.2% vs. 1.6%, P<0.0001). Significant differences in the proportion of

studies reporting sample size calculations were observed when stratified by CVD studied (P<0.0001, range 0-9.5%) with stroke ranking highest (**Figure 4A**), but not when stratified by animal model used (P=0.270, range 0-6.1%) (**Figure 4B**).

Inclusion of both sexes

Sex bias prevalence and detailed temporal trends have been described previously.[29] After excluding studies that did not report the sex of the animals used, significant differences were noted in the proportions including animals of both sexes when stratified by CVD of interest (P<0.0001, range 2.4-34.7%) and by animal model used (P<0.0001, range 6.7-26.9%) (**Supplemental Figure**).

Temporal patterns in study design element implementation

Over the past 10 years, there have been no significant changes in the proportions of studies reporting blinding or randomization (inclusion of both sexes has been reported previously[29]). There has been a significant increase in the proportion reporting *a priori* sample size estimations (P_{trend}<0.0001) though it remains below 7% (**Figure 5**). There was no substantial difference in the prevalence of any of these study design elements before and after the NIH principles and guidelines for reporting preclinical research were published in 2014 (range of differences 0.7-3.6%). *Post hoc* calculations suggest that our sample sizes had ≥80% power to detect a difference of 5.0% for all study design elements.

CVD-specific temporal analyses suggest that preclinical studies on stroke are increasingly incorporating randomization, blinding, sample size estimations, and inclusion of both sexes – a pattern that is not seen in studies on atherosclerosis, hypertension, or cardiomyopathy/heart failure (the most commonly studied CVDs). Studies on atherosclerosis have shown a modest increase in the proportion reporting sample size estimation in recent years, but an overall significant decrease in the proportion reporting blinding, whereas the proportion of studies on cardiomyopathy/heart failure that include both sexes has significantly decreased. No other significant changes were observed over the 10-year period (**Figure 6**).

Given the disproportionate amount of stroke research published in *Stroke*, which has reported improvements in the quality and design of published preclinical research following the introduction of their 'Basic Science Checklist' in 2011,[32] CVD- and animal model-adjusted comparisons of study design implementation before and after this time point were performed for each journal. *Stroke* uniquely exhibited significant improvements in all measures of methodological quality (range of adjusted odds ratios 2.4-8.2, $P<0.0001$ for all study design elements) (**Supplemental Table**). These analyses also identified stroke as the CVD studied as an independent positive predictor of one or more study design element in every journal.

Citations according to index study methodology

At 60 months, 41,441 articles citing the 1,681 preclinical studies that were published between July 2006 and June 2011 were identified. The median citation count per preclinical study was 20 (IQR 13-31, range 0-131). Studies that implemented randomization, blinding, or sample size estimation had similar numbers of citations as those that did not; however, studies that included

both males and females were cited less frequently (median 18 vs. 20, *P*=0.023). The cumulative number of these study design elements was not associated with citation counts (**Figure 7**). No study implemented all 4 of the above design elements and only 20 studies (1.2%) incorporated 3, therefore studies with ≥2 were grouped together for this analysis. The above associations (or lack thereof) between study design element(s) and number of citations persisted after adjusting for journal of publication, year of publication, and/or CVD studied in multivariable regression models and all findings were comparable in sensitivity analyses of citation counts at 36 months.

Discussion

Preclinical cardiovascular research using animal models plays an integral role in advancing the care of patients afflicted with CVDs; however, its impact is contingent on its scientific validity, reproducibility, and relevance to human physiology and disease. It is understood that inherent limitations of using animals to model human diseases can undermine the predictive value of preclinical findings, rendering it difficult for even the most skilled scientists to make impactful discoveries.[12] However, this difficulty is compounded when methodological sources of bias are introduced. We systematically examined a continuous and large body of preclinical cardiovascular research to determine how often randomization, blinding, and *a priori* sample size estimation are incorporated and did so over a sufficiently long period to draw meaningful conclusions on the temporal patterns of these practices. We report that, overall, these design elements are rarely implemented in studies published in leading peer-reviewed cardiovascular journals, paralleling the previously reported prevalence of sex bias.[29] Furthermore, apart from a modest increase in the proportion of studies reporting sample size calculations, there has been no improvement in these practices over the last decade, though stroke research may be a notable

exception. Lastly, analyses of citation counts suggest that crucial methodological aspects of preclinical studies may be overlooked by cardiovascular researchers, affording potentially biased research comparable influence on scientific research efforts as methodologically more robust studies.

Poorly designed preclinical studies not only contribute to experimental irreproducibility[2, 7] and wasted resources,[43, 44] but may result in erroneous conclusions regarding treatment effects,[9, 21, 23, 45, 46] which can ultimately spur or deter clinical trials in humans with consequent risk of harm.[12, 20, 47] The experimental design elements evaluated in our study are deemed crucial to improving the quality of preclinical research by addressing selection bias (randomization); minimizing performance, detection, and attrition bias (blinding); ensuring adequate statistical power and the ethical use of animals (sample size estimation); and promoting research relevance for both men and women (inclusion of both sexes).[1, 6, 7, 10, 13, 26, 48, 49] Though these elements are routinely implemented in clinical trials as they have been shown to protect against bias and imprecision,[50-53] they have conspicuously not permeated preclinical experiments.[1, 23, 54-56] The high degree of experimental control that is possible in preclinical research (e.g. via genetic and environmental homogeneity) can reduce variation and the sample sizes required relative to clinical trials; however, variations in injury or disease induction as well as the potential for persistent and unrecognized confounders represent important sources of bias.[6, 31, 57]

Furthermore, despite a pervasive belief among scientists that there is a reproducibility 'crisis' in research that is attributable to methodological and/or reporting shortcomings,[14] our citation analyses suggest that greater methodological rigor in preclinical cardiovascular research does not translate into greater scientific influence. A similar observation was noted by Amgen in a study of 53 preclinical cancer research publications: studies that they could *not* adequately reproduce

were cited as often, if not more often, than those that they could successfully reproduce, irrespective of the journal of publication's impact factor.[12] The suggestion that methodologically robust studies are given equal consideration as studies at greater risk of bias by cardiovascular researchers is troubling as it undermines the reputed "self-correcting" tenet of science and increases the risk of pursuing unfruitful avenues of research.[2, 5, 44, 48]

Cogent analyses and arguably practical solutions to improve preclinical research quality (and secondarily to enhance research translation) have been proposed; however, evidence of changes in research practice and/or data on the impact of corrective actions are scarce.[13] Several guidelines and checklists have been developed to improve preclinical methodology[45, 58, 59] and/or reporting[25, 45, 60-62] yet there is little indication that they are effecting change despite being widely endorsed.[43, 58, 63] Systematic reviews and meta-analyses have been advocated as safeguards to expose bias in preclinical research prior to embarking on clinical trials,[9, 47, 64, 65] but they are limited by the internal validity of included studies[66] and are not routinely performed. Changes in research funding such as those undertaken by the NIH[2, 26] may bolster these efforts; however, our analyses of animal studies in the cardiovascular sciences have not shown signs of substantial changes in research practices or reporting since they were announced.[29]

Journal editors and reviewers directly influence what is published thereby often serving as ultimate gatekeepers of research findings; however, few journals effectively encourage the use of reporting guidelines.[67] At the time of writing, there exists considerable variation in author guidelines for reporting animal studies among leading cardiovascular journals with few requiring authors to report randomization, blinding, sample size estimation, or the sex of the animals used – requirements that could increase transparency and reinforce the importance of these study design elements.[13] Uniquely, however, *Stroke* implemented a 'Basic Science Checklist' in 2011

(updated in 2016),[28] which includes all of these relevant elements, forms part of the manuscript submission process, and is evaluated by editors and reviewers. Over the 2.5 years following its introduction, Minnerup *et al* noted improvements in randomization, blinding, and allocation concealment compared to the preceding 18 months.[32] We noted significant improvements in the quality of preclinical stroke research, 90% of which was published in *Stroke*, raising the question of whether these improvements were due to journal editorial policies/culture and/or driven by changes in research practices among stroke researchers in general. Our *post hoc* analyses suggest that it is likely a combination of both: *Stroke* uniquely showed improvements in all study design elements even after adjusting for CVD studied and animal model used, but stroke as the CVD studied was identified as an independent predictor for at least one study design element for every journal examined. Our findings therefore corroborate those of Minnerup *et al* and expand upon them by demonstrating that these improvements have continued into 2016, extend to sample size estimation (and to a lesser extent inclusion of both sexes), contrast with an extended preceding period of suboptimal reporting, and have not been paralleled in other prominent cardiovascular journals. Therefore, though not conclusive, our data suggest that journal editors and reviewers can exert considerable influence on preclinical research practices and reporting. Yet, our findings also suggest that stroke research has independently improved in quality – a finding that may be attributable to its community's early appreciation of the importance of methodological rigor in preclinical science and extensive involvement in efforts to improve research translation.[7, 20, 21, 46, 58, 68-71]

Our study is not without limitations. The study sample comprised articles from 5 cardiovascular journals over 10 years, which may not fully represent preclinical cardiovascular research publications or practices. However, we deliberately selected AHA journals given their

prominence in the field of cardiovascular medicine, established reputation, collective focus on a broad range of CVDs, and unanimous endorsement of the NIH guidelines on rigor and reproducibility. Given that none of the American College of Cardiology or European Society of Cardiology journals have endorsed the guidelines thus far,[25] our data may actually overestimate the methodological quality of most preclinical cardiovascular research. There is no accepted definition for preclinical studies therefore criteria were developed for this study. It is therefore possible that not all relevant studies were included in our analysis. However, the inclusion criteria used, the number of journals reviewed, and the substantial interrater agreement for study inclusion support the validity of our results and render our data the best available on methodological rigor in preclinical cardiovascular research. Furthermore, our inclusion criteria are in line with previously proposed distinctions between 'exploratory' and 'confirmatory' preclinical research.[7, 30, 31] Our analysis considered blinding as a single and dichotomous variable, which limits detailed assessments of this design element, and did not examine all relevant potential sources of bias. Data handling and publication bias, for instance, are important factors that may contribute to irreproducibility in science.[2, 6, 7, 9-11, 13, 43] As well, factors other than experimental methodology can influence the predictive value of preclinical research, including how well an animal model reflects human physiology and disease,[6, 9, 45, 72] the appropriateness of statistical analyses,[73] and a lack of standardization of definitions and surrogate markers,[74, 75] which were not examined. Though article citation is the most commonly used measure of research impact, it is an imperfect indicator.[33, 34] Lastly, study design elements may have been implemented but not reported, which could result in underestimates of their prevalence. However, underreporting of measures of methodological rigor is believed to be low[49, 65] and therefore unlikely to have significantly influenced our results.

Our analysis of preclinical research published in leading cardiovascular journals over the last 10 years demonstrates that methodological sources of potential bias and imprecision are prevalent, have not appreciably improved over time, and may be overlooked by researchers when basing subsequent studies. Concerted efforts to address this problem are urgently needed. Stroke research has uniquely shown substantial improvement in several measures of quality in recent years, warranting a closer examination to identify drivers of its success.

Acknowledgements: None.

Sources of funding: None.

Disclosures: None.

References

1. Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, Schulz KF, Tibshirani R. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*. 2014;383:166-175
2. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. *Nature*. 2014;505:612-613
3. Henderson VC, Kimmelman J, Fergusson D, Grimshaw JM, Hackam DG. Threats to validity in the design and conduct of preclinical efficacy studies: A systematic review of guidelines for in vivo animal experiments. *PLoS Med*. 2013;10:e1001489

4. Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical R&D. *Nat Rev Drug Discov*. 2011;10:428-438
5. Bunnage ME. Getting pharmaceutical R&D back on target. *Nat Chem Biol*. 2011;7:335-339
6. van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, Macleod MR. Can animal models of disease reliably inform human studies? *PLoS Med*. 2010;7:e1000245
7. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, Crystal RG, Darnell RB, Ferrante RJ, Fillit H, Finkelstein R, Fisher M, Gendelman HE, Golub RM, Goudreau JL, Gross RA, Gubitz AK, Hesterlee SE, Howells DW, Huguenard J, Kelner K, Koroshetz W, Krainc D, Lazic SE, Levine MS, Macleod MR, McCall JM, Moxley RT, 3rd, Narasimhan K, Noble LJ, Perrin S, Porter JD, Steward O, Unger E, Utz U, Silberberg SD. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*. 2012;490:187-191
8. Scott S, Kranz JE, Cole J, Lincecum JM, Thompson K, Kelly N, Bostrom A, Theodoss J, Al-Nakhala BM, Vieira FG, Ramasubbu J, Heywood JA. Design, power, and interpretation of studies in the standard murine model of ALS. *Amyotroph Lateral Scler*. 2008;9:4-15
9. Perel P, Roberts I, Sena E, Wheble P, Briscoe C, Sandercock P, Macleod M, Mignini LE, Jayaram P, Khan KS. Comparison of treatment effects between animal experiments and clinical trials: Systematic review. *BMJ*. 2007;334:197
10. Muhlhausler BS, Bloomfield FH, Gillman MW. Whole animal experiments should be more like human randomized controlled trials. *PLoS Biol*. 2013;11:e1001481
11. 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

12. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. *Nature*. 2012;483:531-533
13. Begley CG, Ioannidis JP. Reproducibility in science: Improving the standard for basic and preclinical research. *Circ Res*. 2015;116:116-126
14. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature*. 2016;533:452-454
15. Unreliable research: Trouble at the lab. *The Economist*. 2013
16. Lehrer J. The truth wears off: Is there something wrong with the scientific method? *The New Yorker*. 2010
17. Freedman DH. Lies, damned lies, and medical science. *The Atlantic*. 2010
18. Osherovich L. Hedging against academic risk. *Science-Business eXchange*. 2011;4:doi:10.1038/scibx.2011.1416
19. Begley CG. Six red flags for suspect work. *Nature*. 2013;497:433-434
20. Horn J, de Haan RJ, Vermeulen M, Luiten PG, Limburg M. Nimodipine in animal model experiments of focal cerebral ischemia: A systematic review. *Stroke*. 2001;32:2433-2438
21. Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke*. 2008;39:2824-2829
22. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2:e124
23. Hackam DG, Redelmeier DA. Translation of research evidence from animals to humans. *JAMA*. 2006;296:1731-1732
24. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. *Nat Biotechnol*. 2014;32:40-51

25. National Institutes of Health. Principles and guidelines for reporting preclinical research. https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research. Accessed December 2, 2016.
26. Clayton JA, Collins FS. Policy: NIH to balance sex in cell and animal studies. *Nature*. 2014;509:282-283
27. Daugherty A, Hegele RA, Mackman N, Rader DJ, Schmidt AM, Weber C. Complying with the national institutes of health guidelines and principles for rigor and reproducibility: Refutations. *Arterioscler Thromb Vasc Biol*. 2016;36:1303-1304
28. Vahidy F, Schabitz WR, Fisher M, Aronowski J. Reporting standards for preclinical studies of stroke therapy. *Stroke*. 2016;47:2435-2438
29. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald Z, Simard T, Clancy AA, Russo JJ, Welch V, Wells GA, Hibbert B. Sex bias is increasingly prevalent in preclinical cardiovascular research: Implications for translational medicine and health equity for women: A systematic assessment of leading cardiovascular journals over a 10-year period. *Circulation*. 2017;135:625-626
30. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biol*. 2014;12:e1001863
31. Unger EF. All is not well in the world of translational research. *J Am Coll Cardiol*. 2007;50:738-740
32. Minnerup J, Zentsch V, Schmidt A, Fisher M, Schabitz WR. Methodological quality of experimental stroke studies published in the Stroke journal: Time trends and effect of the basic science checklist. *Stroke*. 2016;47:267-272

33. Ranasinghe I, Shojaee A, Bikdeli B, Gupta A, Chen R, Ross JS, Masoudi FA, Spertus JA, Nallamothu BK, Krumholz HM. Poorly cited articles in peer-reviewed cardiovascular journals from 1997 to 2007: Analysis of 5-year citation rates. *Circulation*. 2015;131:1755-1762
34. Danthi N, Wu CO, Shi P, Lauer M. Percentile ranking and citation impact of a large cohort of National Heart, Lung, and Blood Institute-funded cardiovascular R01 grants. *Circ Res*. 2014;114:600-606
35. Kaltman JR, Evans FJ, Danthi NS, Wu CO, DiMichele DM, Lauer MS. Prior publication productivity, grant percentile ranking, and topic-normalized citation impact of NHLBI cardiovascular R01 grants. *Circ Res*. 2014;115:617-624
36. Lauer MS, Danthi NS, Kaltman J, Wu C. Predicting productivity returns on investment: Thirty years of peer review, grant funding, and publication of highly cited papers at the National Heart, Lung, and Blood Institute. *Circ Res*. 2015;117:239-243
37. Fabsitz RR, Papanicolaou GJ, Sholinsky P, Coady SA, Jaquish CE, Nelson CR, Olson JL, Puggal MA, Purkiser KL, Srinivas PR, Wei GS, Wolz M, Sorlie PD. Impact of National Heart, Lung, and Blood Institute-supported cardiovascular epidemiology research, 1998 to 2012. *Circulation*. 2015;132:2028-2033
38. Kulkarni AV, Aziz B, Shams I, Busse JW. Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. *JAMA*. 2009;302:1092-1096
39. Steen RG, Casadevall A, Fang FC. Why has the number of scientific retractions increased? *PLoS One*. 2013;8:e68397

40. Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. *J Clin Epidemiol*. 1995;48:1495-1501
41. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol*. 1995;48:1503-1510
42. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49:1373-1379
43. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, Michie S, Moher D, Wager E. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet*. 2014;383:267-276
44. Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. *PLoS Biol*. 2015;13:e1002165
45. Macleod MR, O'Collins T, Howells DW, Donnan GA. Pooling of animal experimental data reveals influence of study design and publication bias. *Stroke*. 2004;35:1203-1208
46. Macleod MR, O'Collins T, Horky LL, Howells DW, Donnan GA. Systematic review and metaanalysis of the efficacy of FK506 in experimental stroke. *J Cereb Blood Flow Metab*. 2005;25:713-721
47. Pound P, Ebrahim S, Sandercock P, Bracken MB, Roberts I. Where is the evidence that animal research benefits humans? *BMJ*. 2004;328:514-517
48. Clayton JA. Studying both sexes: A guiding principle for biomedicine. *FASEB J*. 2016;30:519-524

49. Macleod MR, Lawson McLean A, Kyriakopoulou A, Serghiou S, de Wilde A, Sherratt N, Hirst T, Hemblade R, Bahor Z, Nunes-Fonseca C, Potluru A, Thomson A, Baginskaite J, Egan K, Vesterinen H, Currie GL, Churilov L, Howells DW, Sena ES. Risk of bias in reports of in vivo research: A focus for improvement. *PLoS Biol*. 2015;13:e1002273
50. Savovic J, Jones HE, Altman DG, Harris RJ, Juni P, Pildal J, Als-Nielsen B, Balk EM, Gluud C, Gluud LL, Ioannidis JP, Schulz KF, Beynon R, Welton NJ, Wood L, Moher D, Deeks JJ, Sterne JA. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Ann Intern Med*. 2012;157:429-438
51. Tzoulaki I, Siontis KC, Ioannidis JP. Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: Meta-epidemiology study. *BMJ*. 2011;343:d6829
52. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282:1061-1066
53. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ*. 2006;174:469-476
54. Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, Fry D, Hutton J, Altman DG. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One*. 2009;4:e7824
55. Beery AK, Zucker I. Sex bias in neuroscience and biomedical research. *Neurosci Biobehav Rev*. 2011;35:565-572
56. Yoon DY, Mansukhani NA, Stubbs VC, Helenowski IB, Woodruff TK, Kibbe MR. Sex bias exists in basic science and translational surgical research. *Surgery*. 2014;156:508-516

57. Reardon S. A mouse's house may ruin experiments. *Nature*. 2016;530:264
58. Fisher M, Feuerstein G, Howells DW, Hurn PD, Kent TA, Savitz SI, Lo EH, Group S. Update of the Stroke Therapy Academic Industry Roundtable preclinical recommendations. *Stroke*. 2009;40:2244-2250
59. Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW. SYRCLE's risk of bias tool for animal studies. *BMC Med Res Methodol*. 2014;14:43
60. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. *PLoS Biol*. 2010;8:e1000412
61. Announcement: Reducing our irreproducibility. *Nature*. 2013;496
62. Hooijmans CR, Leenaars M, Ritskes-Hoitinga M. A gold standard publication checklist to improve the quality of animal studies, to fully integrate the three Rs, and to make systematic reviews more feasible. *Altern Lab Anim*. 2010;38:167-182
63. Baker D, Lidster K, Sottomayor A, Amor S. Two years later: Journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS Biol*. 2014;12:e1001756
64. Sandercock P, Roberts I. Systematic reviews of animal experiments. *Lancet*. 2002;360:586
65. Hirst JA, Howick J, Aronson JK, Roberts N, Perera R, Koshiaris C, Heneghan C. The need for randomization in animal trials: An overview of systematic reviews. *PLoS One*. 2014;9:e98856
66. van Luijk J, Bakker B, Rovers MM, Ritskes-Hoitinga M, de Vries RB, Leenaars M. Systematic reviews of animal studies; missing link in translational research? *PLoS One*. 2014;9:e89981

67. Hirst A, Altman DG. Are peer reviewers encouraged to use reporting guidelines? A survey of 116 health research journals. *PLoS One*. 2012;7:e35621
68. Stroke Therapy Academic Industry R. Recommendations for standards regarding preclinical neuroprotective and restorative drug development. *Stroke*. 1999;30:2752-2758
69. O'Collins VE, Macleod MR, Donnan GA, Horky LL, van der Worp BH, Howells DW. 1,026 experimental treatments in acute stroke. *Ann Neurol*. 2006;59:467-477
70. Crossley NA, Sena E, Goehler J, Horn J, van der Worp B, Bath PM, Macleod M, Dirnagl U. Empirical evidence of bias in the design of experimental stroke studies: A metaepidemiologic approach. *Stroke*. 2008;39:929-934
71. 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
72. Perrin S. Preclinical research: Make mouse studies work. *Nature*. 2014;507:423-425
73. Garcia-Berthou E, Alcaraz C. Incongruence between test statistics and p values in medical papers. *BMC Med Res Methodol*. 2004;4:13
74. Johnson VE. Revised standards for statistical evidence. *Proc Natl Acad Sci U S A*. 2013;110:19313-19317
75. Ioannidis JP. How to make more published research true. *PLoS Med*. 2014;11:e1001747

Table. Study characteristics

	Circulation (*n*=672)	*Circ Res* (*n*=486)	*Hypertension* (*n*=860)	*Stroke* (*n*=501)	*ATVB* (*n*=877)	Total (*n*=3396)
Disease studied						
Atherosclerosis	119 (17.7)	113 (23.3)	14 (1.6)	2 (0.4)	385 (43.9)	633 (18.6)
Hypertension	34 (5.1)	35 (7.2)	577 (67.1)	1 (0.2)	14 (1.6)	661 (19.5)
Arterial aneurysm	16 (2.4)	21 (4.3)	15 (1.7)	24 (4.8)	62 (7.1)	138 (4.1)
Myocardial infarction	95 (14.1)	56 (11.5)	19 (2.2)	0 (0)	115 (13.1)	285 (8.4)
Cardiomyopathy/heart failure	173 (25.7)	154 (31.7)	135 (15.7)	0 (0)	21 (2.4)	483 (14.2)
Pulmonary hypertension	57 (8.5)	29 (6.0)	34 (4.0)	0 (0)	15 (1.7)	135 (4.0)
Arrhythmia	65 (9.7)	35 (7.2)	4 (0.5)	0 (0)	0 (0)	104 (3.1)
Valvular disease	9 (1.3)	4 (0.8)	0 (0)	0 (0)	14 (1.6)	27 (0.8)
Stroke	20 (3.0)	4 (0.8)	9 (1.1)	435 (86.8)	18 (2.1)	486 (14.3)
Transplantation medicine	28 (4.2)	15 (3.1)	0 (0)	31 (6.2)	32 (3.7)	106 (3.1)
Metabolic/endocrine disorders	19 (2.8)	10 (2.1)	23 (2.7)	2 (0.4)	75 (8.6)	129 (3.8)
Hematologic disorders	19 (2.8)	8 (1.7)	1 (0.1)	0 (0)	79 (9.0)	107 (3.2)
Other	18 (2.7)	2 (0.4)	4 (0.5)	6 (1.2)	47 (5.4)	102 (3.0)
Animal model(s) used						
Mouse	463 (68.9)	391 (80.5)	357 (41.5)	223 (44.5)	744 (84.8)	2178 (64.1)
Rat	84 (12.5)	44 (9.1)	443 (51.5)	237 (47.3)	63 (7.3)	872 (25.7)
Rabbit	10 (1.5)	15 (3.1)	16 (1.9)	18 (3.6)	23 (2.6)	82 (2.4)
Dog	37 (5.5)	12 (2.5)	13 (1.5)	5 (1.0)	0 (0)	67 (2.0)
Pig	84 (12.5)	4 (0.8)	9 (1.1)	3 (0.6)	19 (2.2)	60 (1.8)
Primate	3 (0.5)	1 (0.2)	2 (0.2)	3 (0.6)	8 (0.9)	17 (0.5)
Other	10 (1.5)	4 (0.8)	9 (1.1)	1 (0.2)	1 (0.1)	25 (0.7)
Combination	40 (6.0)	15 (3.1)	11 (1.3)	11 (2.2)	18 (2.1)	95 (2.8)
Therapeutic study	162 (24.1)	85 (17.5)	221 (25.7)	264 (52.7)	256 (29.2)	988 (29.1)

Values reported as number (%). *ATVB*: *Arteriosclerosis, Thrombosis and Vascular Biology*; *Circ Res*: *Circulation Research*.

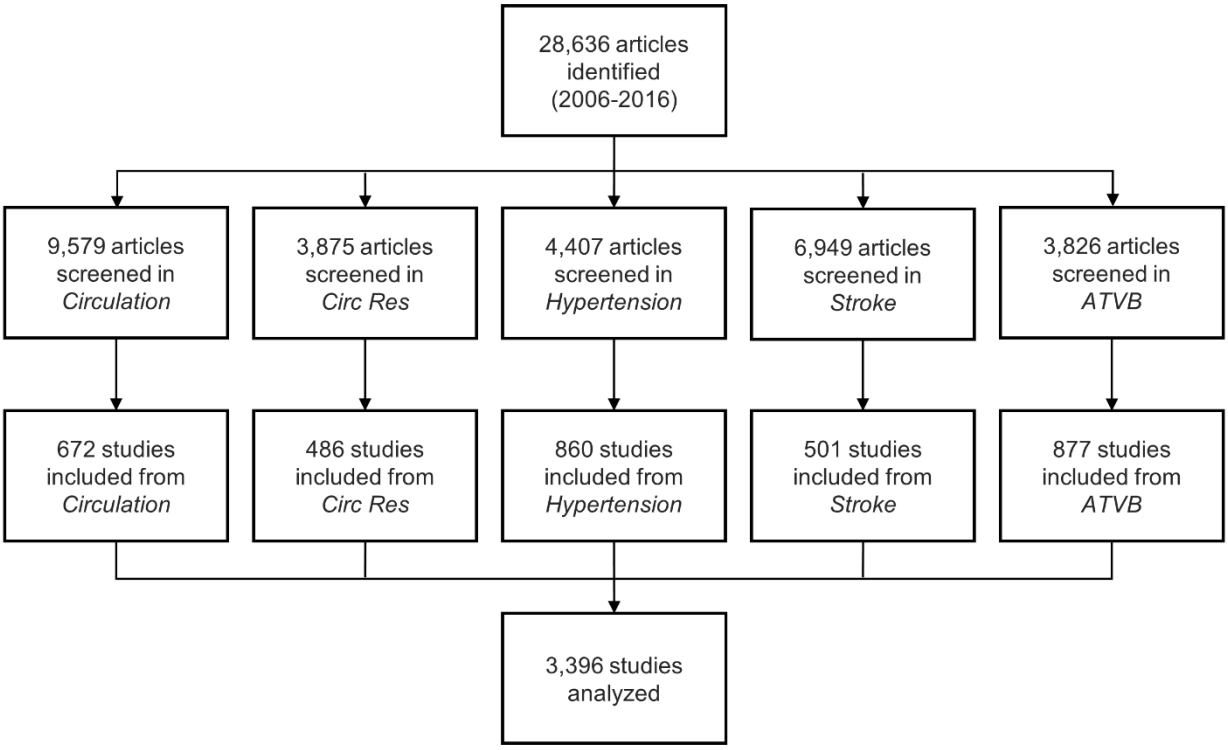


Figure 1. Literature search and results. *ATVB*: *Arteriosclerosis, Thrombosis and Vascular Biology*; *Circ Res*: *Circulation Research*.

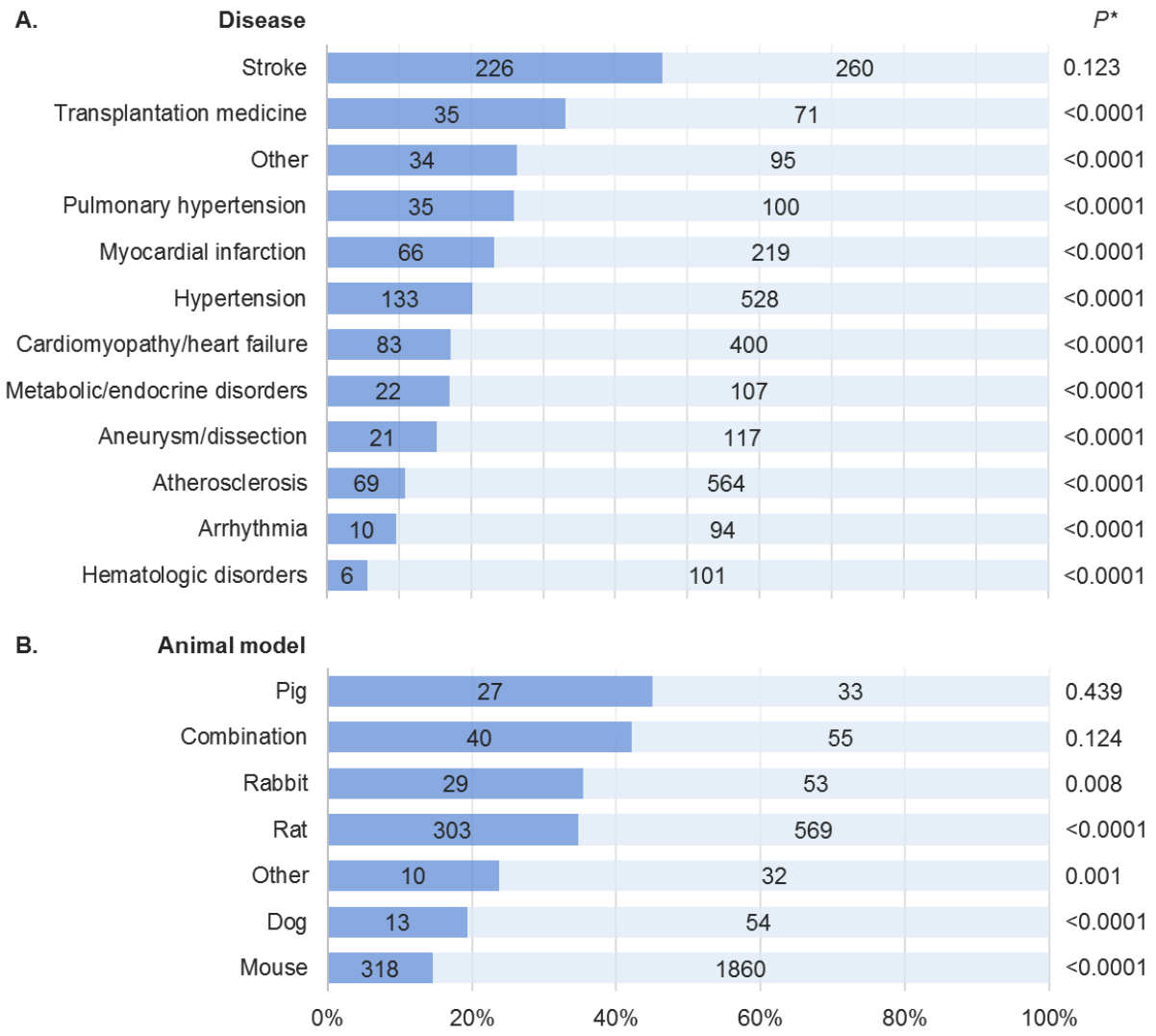


Figure 2. Randomization in preclinical cardiovascular research studies published over a 10-year period stratified by (*A*) disease studied and (*B*) species used. Dark blue corresponds to the proportion of studies implementing randomization; numbers in bars correspond to absolute numbers of studies. Valvular disease and resuscitative medicine studies were included in the 'Other' category due to the small number of relevant publications. 'Combination' refers to more than one animal species used within the same publication; animal models in the 'Other' category included guinea pig, gerbil, and hamster. *For comparison of studies incorporating randomization vs. not.

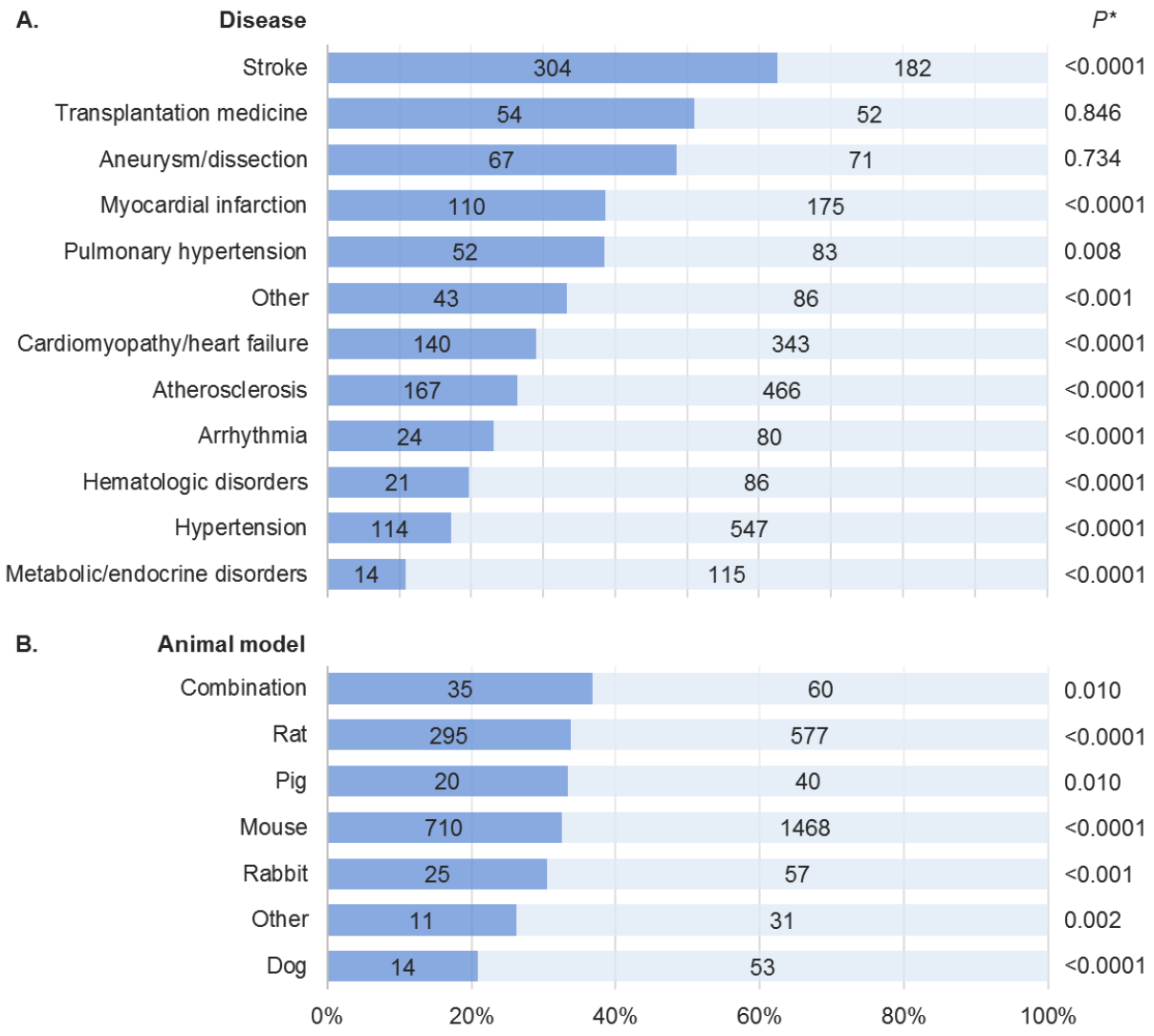


Figure 3. Blinding in preclinical cardiovascular research studies published over a 10-year period stratified by (*A*) disease studied and (*B*) species used. Dark blue corresponds to the proportion of studies implementing blinding; numbers in bars correspond to absolute numbers of studies. Valvular disease and resuscitative medicine studies were included in the 'Other' category due to the small number of relevant publications. 'Combination' refers to more than one animal species used within the same publication; animal models in the 'Other' category included guinea pig, gerbil, and hamster. *For comparison of studies incorporating blinding vs. not.

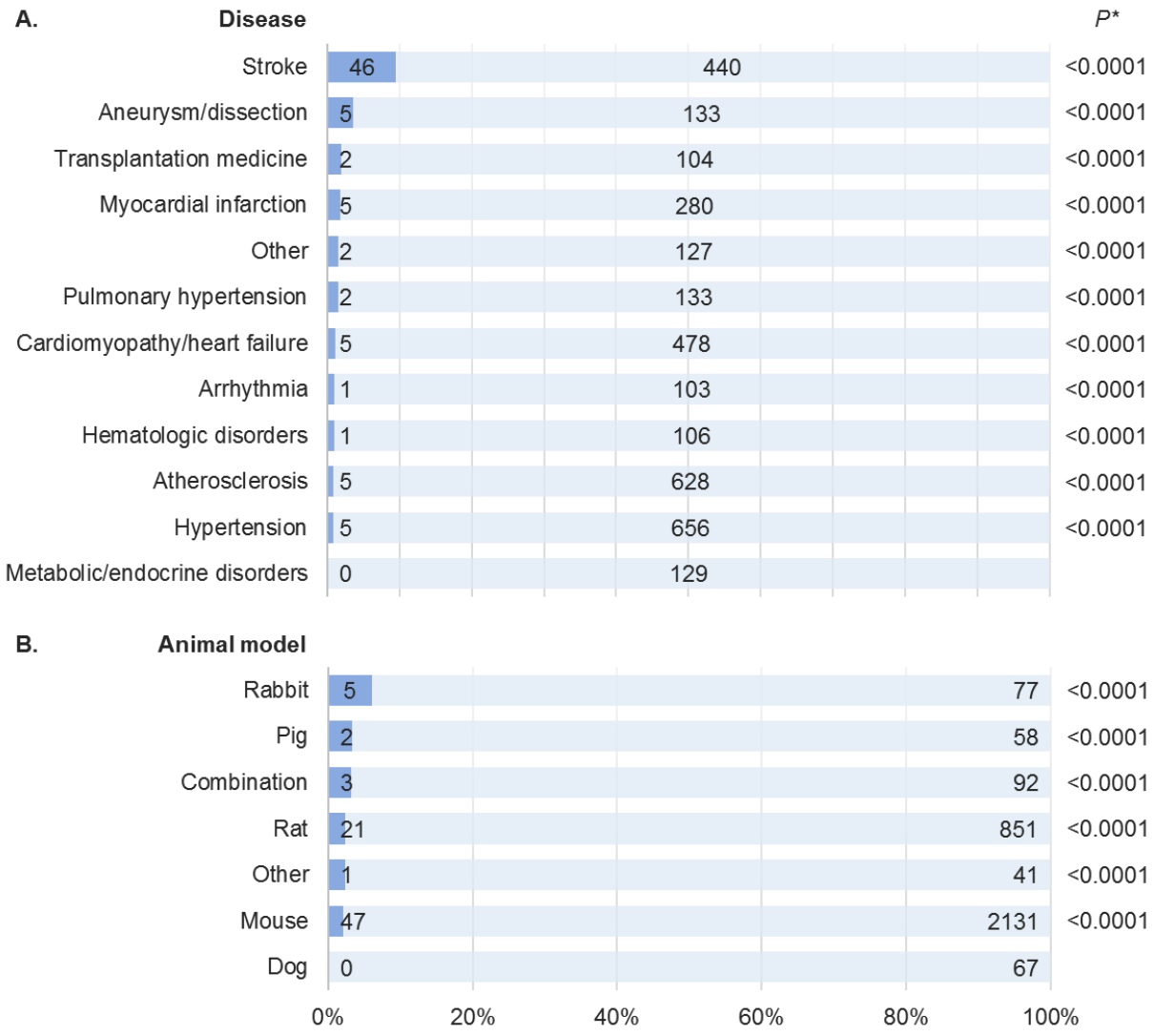


Figure 4. Sample size estimation in preclinical cardiovascular research studies published over a 10-year period stratified by (*A*) disease studied and (*B*) species used. Dark blue corresponds to the proportion of studies reporting sample size estimations/power calculations; numbers in bars correspond to absolute numbers of studies. Valvular disease and resuscitative medicine studies were included in the 'Other' category due to the small number of relevant publications. 'Combination' refers to more than one animal species used within the same publication; animal models in the 'Other' category included guinea pig, gerbil, and hamster. *For comparison of studies incorporating sample size estimation vs. not.

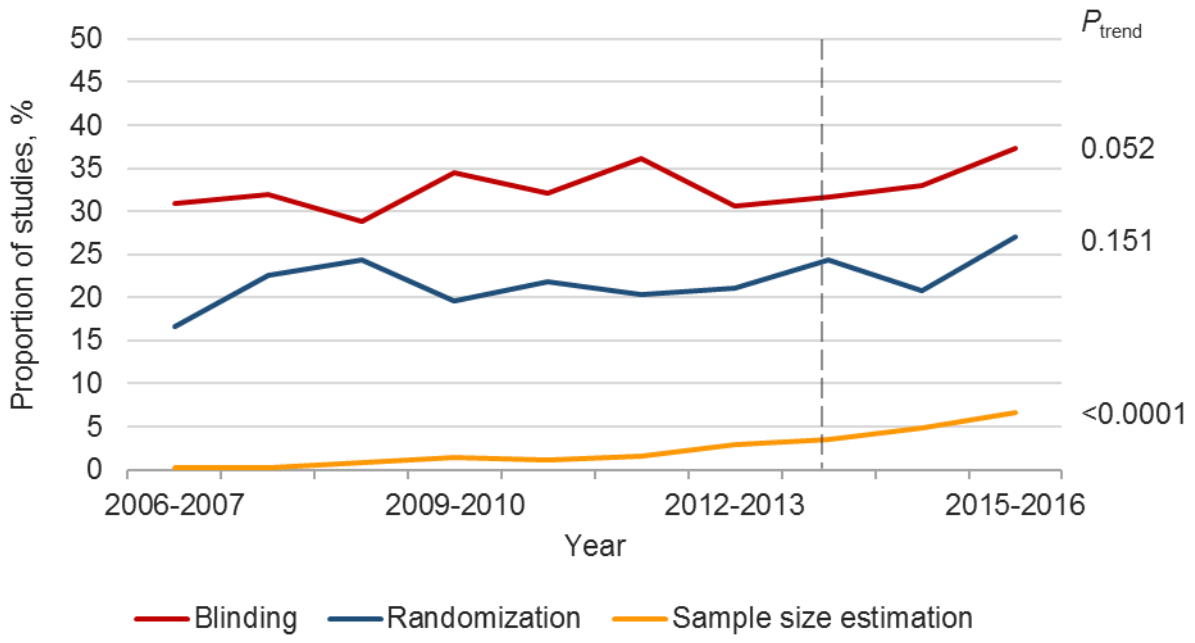


Figure 5. Temporal patterns in randomization, blinding, and sample size estimation in preclinical cardiovascular studies. Dashed line indicates the publication of the National Institutes of Health guidelines and policies for reporting preclinical research. Data for inclusion of both sexes have been reported previously.[29]

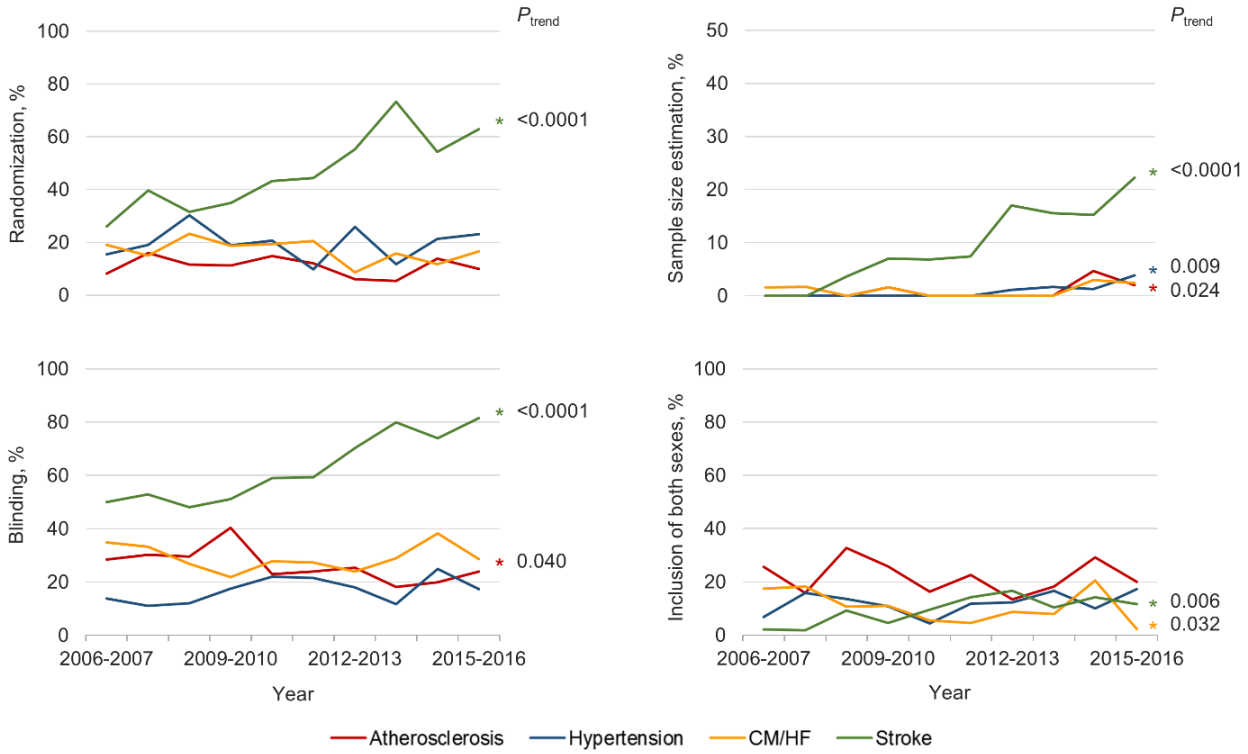


Figure 6. Temporal patterns in randomization, blinding, sample size estimation, and inclusion of both sexes in preclinical research for the most commonly studied cardiovascular diseases. Note the different scale of y-axis for sample size estimation. CM/HF: cardiomyopathy/heart failure.

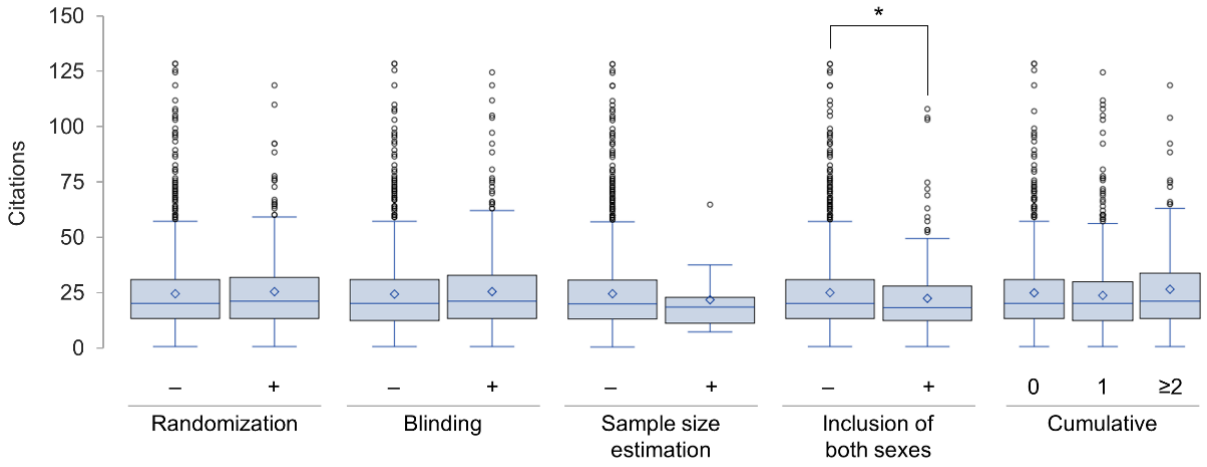
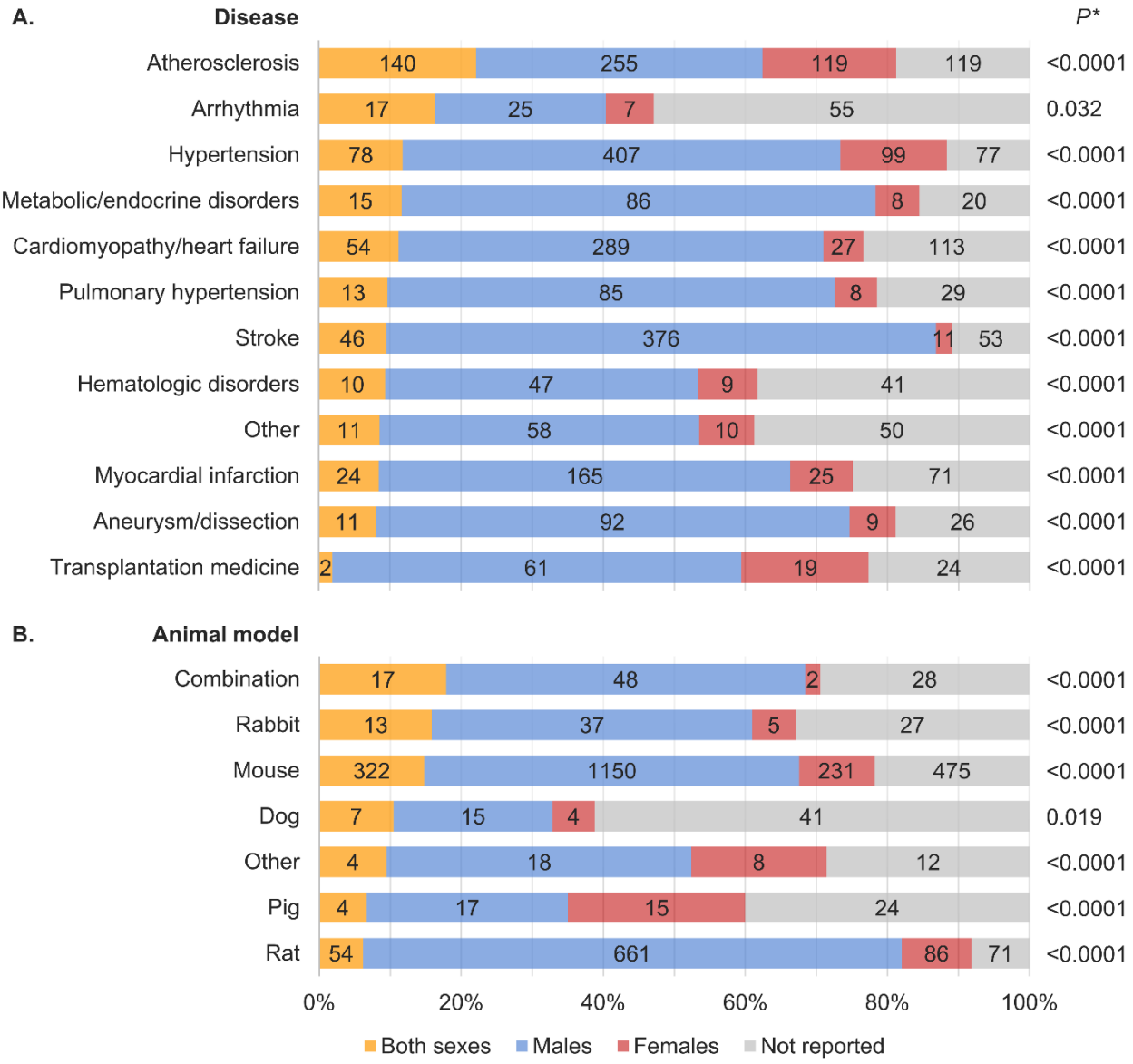


Figure 7. Box-whisker plot of the number of citations by methodological rigor of index preclinical study (*n*=1,681). Diamond symbol plotted at mean. Outliers identified as beyond 1.5 IQR. Cumulative refers to sum of randomization, blinding, sample size estimation, and inclusion of both sexes (each contributing 1 point). *P<0.05.

SUPPLEMENTAL MATERIAL



Online Supplemental Figure. Inclusion of both sexes in preclinical cardiovascular research studies published over a 10-year period stratified by (*A*) disease studied and (*B*) species used. Valvular disease and resuscitative medicine studies were included in the 'Other' category due to the small number of relevant publications. Numbers in bars correspond to absolute numbers of studies. 'Combination' refers to more than one animal species used within the same publication. *For comparison of proportion of studies including both sexes vs. males or females only.

Supplemental Table: Comparison of study design element implementation in preclinical studies before and after the implementation of the *Stroke* Basic Science Checklist, stratified by journal of publication

	Period 1* *n* (%)	Period 2* *n* (%)	Crude OR (95% CI)	*P*	Adjusted OR (95% CI)†	*P*†
Circulation	*n*=464	*n*=208				
Randomization	107 (23.1)	36 (17.3)	0.7 (0.5-1.1)	0.093	0.7 (0.4-1.1)	0.119
Blinding	169 (36.4)	59 (28.4)	0.7 (0.5-1.0)	0.042	0.7 (0.5-1.0)	0.043
Sample size estimation	7 (1.5)	5 (2.4)	1.6 (0.5-5.1)	0.422	NR	
Inclusion of both sexes	64 (13.8)	29 (13.9)	1.0 (0.6-1.6)	0.959	1.0 (0.6-1.6)	0.967
Circulation Research	*n*=303	*n*=183				
Randomization	35 (11.6)	29 (15.8)	1.4 (0.8-2.5)	0.176	1.4 (0.8-2.5)	0.261
Blinding	93 (30.7)	60 (32.8)	1.1 (0.7-1.6)	0.630	0.9 (0.6-1.4)	0.788
Sample size estimation	1 (0.3)	1 (0.3)	1.7 (0.1-26.7)	0.721	NR	
Inclusion of both sexes	57 (18.8)	33 (18.0)	0.9 (0.6-1.5)	0.830	1.0 (0.6-1.6)	0.937
Hypertension	*n*=485	*n*=375				
Randomization	104 (21.4)	81 (21.6)	1.0 (0.7-1.4)	0.956	1.2 (0.9-1.7)	0.298
Blinding	101 (20.8)	86 (22.9)	1.1 (0.8-1.6)	0.457	1.1 (0.8-1.5)	0.617
Sample size estimation	0 (0)	1 (0.3)	→∞ (0.0-∞)	0.946	NR	
Inclusion of both sexes	43 (8.9)	36 (9.6)	1.1 (0.7-1.7)	0.712	1.1 (0.7-1.7)	0.798
Stroke	*n*=316	*n*=185				
Randomization	120 (38.0)	119 (64.3)	2.9 (2.0-4.3)	<0.0001	3.2 (2.1-4.7)	<0.0001
Blinding	171 (54.1)	144 (77.8)	3.0 (2.0-4.5)	<0.0001	3.0 (2.0-4.5)	<0.0001
Sample size estimation	10 (3.2)	35 (18.9)	7.1 (3.4-14.8)	<0.0001	8.2 (3.7-18.4)	<0.0001
Inclusion of both sexes	15 (4.7)	20 (10.8)	2.4 (1.2-4.9)	0.012	2.4 (1.2-4.9)	<0.0001
ATVB	*n*=476	*n*=401				
Randomization	61 (12.8)	48 (12.0)	0.9 (0.6-1.4)	0.706	0.9 (0.6-1.4)	0.668
Blinding	130 (27.3)	97 (24.2)	0.8 (0.6-1.2)	0.293	0.7 (0.5-1.0)	0.026
Sample size estimation	2 (0.4)	10 (2.5)	6.1 (1.3-27.8)	0.021	NR	
Inclusion of both sexes	72 (15.1)	52 (13.0)	0.8 (0.6-1.2)	0.361	0.8 (0.6-1.3)	0.411

NR: not reported due to small number of events per predictor variable; OR: odds ratio

*Periods 1 and 2 correspond to before and after the date of implementation of the 'Basic Science Checklist' by *Stroke*, respectively

†Adjusted for cardiovascular disease studied and animal model used

CHAPTER 4: SEX BIAS IN PRECLINICAL CARDIOVASCULAR RESEARCH

The following is a manuscript published in *Circulation* examining preclinical cardiovascular studies for evidence of sex bias. Specifically, it sought to identify the sex of the animals used, the frequency of sex-matching, the frequency of sex-based reporting of results, and whether the use of animals of a single sex is acknowledged as a limitation. A manuscript reporting on a subgroup analysis of studies on atherosclerosis and arterial aneurysms/dissection was published in *Arteriosclerosis, Thrombosis, and Vascular Biology* and is included in the Appendices.

Additional documentation related to this research letter is provided in the following appendices:

Appendix H: Published research letter in *Circulation*.

Appendix I: Published letter to the editor in *Arteriosclerosis, Thrombosis, and Vascular Biology*.

Sex bias is increasingly prevalent in preclinical cardiovascular research: implications for translational medicine and health equity for women

A systematic assessment of leading cardiovascular journals over a 10-year period

F. Daniel Ramirez MD,[a,b,c]* Pouya Motazedian BSc,[b]* Richard G. Jung BSc,[b,d,e] Pietro Di Santo MD,[a,b] Zachary MacDonald BSc,[b] Trevor Simard MD,[a,b,d,e] Aisling A. Clancy MD MSc,[f] Juan J. Russo MD,[a,b] Vivian Welch PhD,[c,g,h] George A. Wells PhD;[c,i] Benjamin Hibbert MD PhD[a,b,d,e]

[a] Division of Cardiology, University of Ottawa Heart Institute, Ottawa, Ontario, Canada

[b] CAPITAL Research Group, University of Ottawa Heart Institute, Ottawa, Ontario, Canada

[c] School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Ontario, Canada

[d] Vascular Biology and Experimental Medicine Laboratory, University of Ottawa Heart Institute, Ottawa, Ontario, Canada

[e] Department of Cellular and Molecular Medicine, University of Ottawa, Ontario, Canada

[f] Department of Obstetrics and Gynecology, University of Ottawa, Ottawa, Ontario, Canada

[g] Bruyère Research Institute, University of Ottawa, Ottawa, Ontario, Canada

[h] Centre for Global Health, University of Ottawa, Ottawa, Ontario, Canada

[i] Cardiovascular Research Methods Centre, University of Ottawa Heart Institute, Ottawa, Ontario, Canada

* Equally contributing authors

Word count: 795

Short title: Sex bias in preclinical cardiovascular research

Corresponding author:

Dr. Benjamin Hibbert

University of Ottawa Heart Institute

40 Ruskin Street, H-4238

Ottawa, Ontario, Canada K1Y 4W7

Tel.: (613) 761-4409

Fax: (613) 761-4874

Email: bhibbert@ottawaheart.ca

Ensuring that women are adequately represented in clinical trials is recognized as essential for gender equity in health. However, the use of female animal models and sex-based reporting have not been equally enforced in preclinical stages of research, which often precede and inform clinical trials. In 2014, the National Institutes of Health (NIH) announced that it would require that sex be considered as a biological variable in applications for preclinical research funding[1] yet a reluctance to include female animal models in preclinical experiments persists. Inappropriately inferring experimental findings to both sexes when a single sex is studied or when sex is not specified has the potential to disadvantage women by skewing our understanding of disease processes toward male-predominant patterns and by reducing the likelihood of female-specific therapeutics advancing to the clinical realm. We therefore systematically examined all preclinical cardiovascular studies published in American Heart Association (AHA) journals with archives spanning at least 10 years (*Circulation*, *Circulation Research*, *Hypertension*, *Stroke*, and *Arteriosclerosis, Thrombosis, and Vascular Biology* (*ATVB*)) for evidence of sex bias.

Full manuscripts published between July 2006 to June 2016 and reporting original data from *in vivo* experiments in non-human mammals on pathophysiology, genetics, and/or therapeutic interventions directly relevant to a specific cardiovascular disorder in humans were included. Studies on physiologic and/or genetic characteristics were included if they proposed potential therapeutic applications or implications of the study findings. Each journal article was independently reviewed and pre-specified data extracted using standardized case report forms by two authors, including the cardiovascular disease (CVD) investigated, animal model(s) used and their sex, whether study samples were sex-matched (in studies using both sexes), whether at least one study result was reported by sex, and whether the use of a single sex was reported as a

limitation (in single-sex studies). Interrater agreement was calculated using Cohen's kappa statistic and percent agreement. Discrepancies were resolved by consensus or independent adjudication. Categorical variables were compared via chi-square tests. Temporal patterns were evaluated via Pearson correlation or Cochrane-Armitage trend tests. All analyses were performed using SAS 9.4 (SAS Institute Inc., Cary, NC) using an alpha level of 0.05 to define statistical significance.

Of 28,636 articles screened, 3,396 met inclusion criteria and were analyzed. Interrater agreement for study inclusion prior to resolution was 94.5% (κ=0.72, 95% CI 0.70-0.73). The sex of the animals used was not reported in 20.0% of studies. Males were exclusively used in 71.6% of studies in which sex was reported whereas females were exclusively used in 12.9% and both sexes in 15.5%. Sex-matching of animals was reported in 17.1% of studies that included both sexes. Restricting this analysis to the 988 studies of therapeutic interventions did not appreciably change these distributions. When stratified by CVD studied, males were exclusively used significantly more often than females or both sexes in all cases except atherosclerosis and arrhythmia. When stratified by animal model used, males were exclusively used significantly more often when mice, rats, rabbits, or a combination of animal models were employed (accounting for 95.0% of studies).

The number of preclinical studies published yearly over the past decade has remained stable; however, there has been a significant increase in the proportion using exclusively males and a significant decrease in the proportion using exclusively females. There have not been significant changes in the proportions of studies using both sexes or reporting the sex of the animals used (**Figure**). These trends were unchanged when studies in which sex was not specified were excluded.

Sex-based reporting of results occurred in 35.1% of studies in which sex was reported using the liberal criterion of *any* results being reported in reference to the sex of the animals used at least once. There was no appreciable difference in the prevalence of sex-based reporting before and after the publication of planned NIH policies to improve sex and gender inclusion in preclinical research[1] (34.9% vs. 35.7%, respectively). Of the 2,297 studies in which only one sex was reportedly used, 1.6% mentioned this aspect of the study design as a potential limitation.

Though the need to include women in clinical trials is now a well-established requirement to enhance external validity, analogous standards have not gained a foothold in preclinical stages. Reasons put forward for the preferential use of males include increased variability due to fluctuating gonadal hormone levels throughout the estrous cycle and sample size implications of planning sex-based analyses.[2, 3] Evidence to support these concerns is limited for most experimental settings, however.[3-5]

Our review of preclinical research published in leading cardiovascular journals over the last 10 years demonstrates that sex bias is prevalent and increasing, contrasting with advances made in clinical research designs and reporting. Concerted efforts on the part of researchers, journal editors, peer reviewers, universities, industry, professional organizations, and patient advocacy groups are urgently needed to address this discrepancy.

Acknowledgements: None.

Sources of funding: None.

Disclosures: None.

References

1. Clayton JA and Collins FS. Policy: NIH to balance sex in cell and animal studies. *Nature*. 2014;509:282-3.

2. Zucker I and Beery AK. Males still dominate animal studies. *Nature*. 2010;465:690.

3. Clayton JA. Studying both sexes: a guiding principle for biomedicine. *FASEB J*. 2016;30:519-24.

4. Prendergast BJ, Onishi KG and Zucker I. Female mice liberated for inclusion in neuroscience and biomedical research. *Neurosci Biobehav Rev*. 2014;40:1-5.

5. Itoh Y and Arnold AP. Are females more variable than males in gene expression? Meta-analysis of microarray datasets. *Biol Sex Differ*. 2015;6:18.

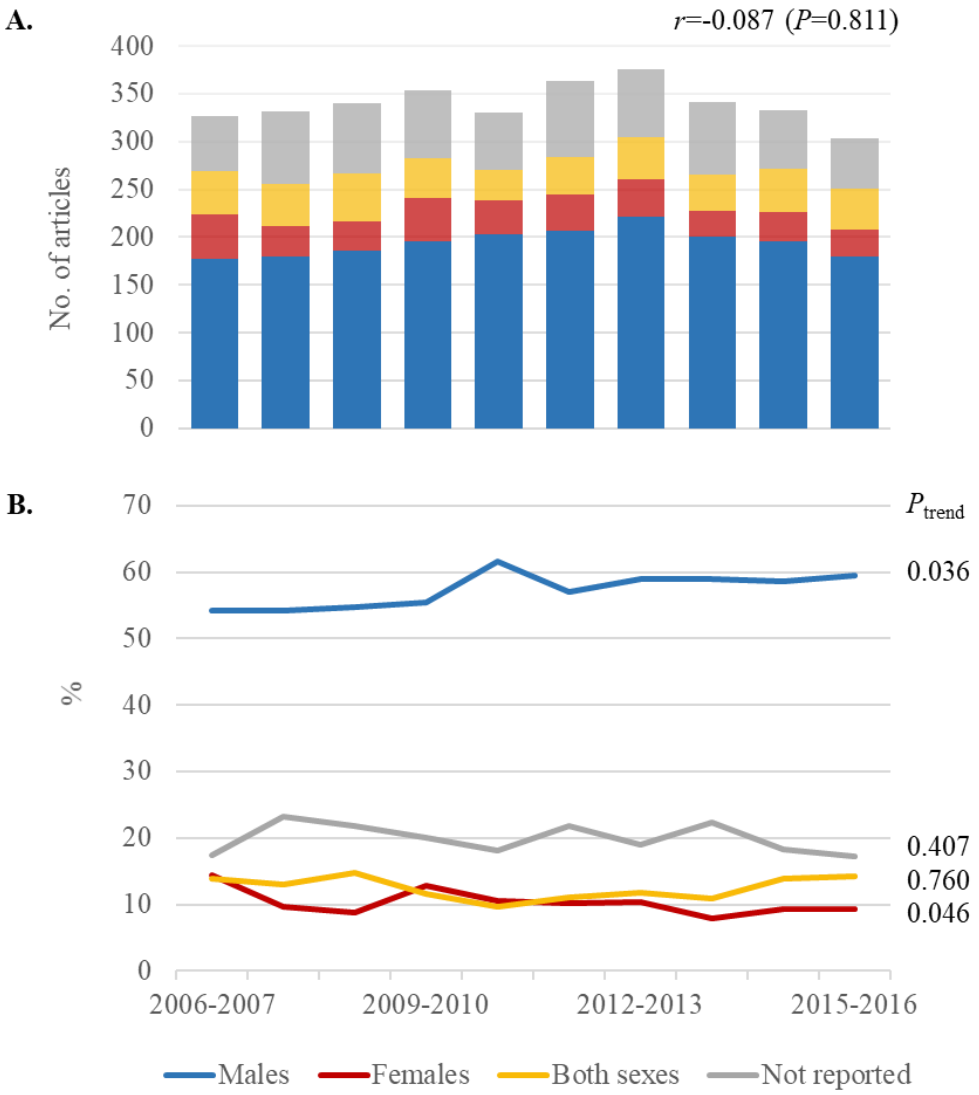


Figure. Temporal patterns in the sex of animals used in preclinical cardiovascular research over a 10-year period. *A.* Absolute number of preclinical studies published and sex of animals used as a function of time. *r* denotes Pearson's correlation coefficient for the total number of studies over time. *B.* Sex of animals used in preclinical studies over time (%). P_{trend} calculated for the proportion of studies using exclusively males, exclusively females, both sexes, and not reporting the sex of the animals used over time.

CHAPTER 5: TEMPORAL TRENDS IN WOMEN'S INVOLVEMENT IN PRECLINICAL CARDIOVASCULAR RESEARCH AND ITS ASSOCIATION WITH EXPERIMENTAL DESIGN

The following is a manuscript accepted for publication in *JACC Basic to Translational Science* that examines whether researcher sex is associated with differences in measures of methodological rigour, especially the inclusion of female animals in experiments and sex-based analyses in preclinical cardiovascular research. We additionally explored whether researcher sex was associated with citation counts as a surrogate for research impact as well as examined temporal patterns in female authorship and in research mentorships stratified by sex.

Female authorship in preclinical cardiovascular research: temporal trends and influence on experimental design

Alisha Labinaz[*,†,‡], Jeffrey A. Marbach MBBS[*,§], Richard G. Jung BSc[*,†,‖,¶], Robert Moreland MD[*,#], Pouya Motazedian MD[*,**], Pietro Di Santo MD[*,§,††], Aisling A. Clancy MD MSc[‡‡,§§], Zachary MacDonald MD[‖‖], Trevor Simard MD[*,†,§,‖], Benjamin Hibbert MD PhD[*,†,§,‖], F. Daniel Ramirez MD[*,§,††]

[*] CAPITAL Research Group, University of Ottawa Heart Institute, Ottawa, Ontario, Canada

[†] Vascular Biology and Experimental Medicine Laboratory, University of Ottawa Heart Institute, Ottawa, Ontario, Canada

[‡] Faculty of Science, University of Ottawa, Ottawa, Ontario, Canada

[§] Division of Cardiology, University of Ottawa Heart Institute, Ottawa, Ontario, Canada

[‖] Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, Ontario, Canada

[¶] Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada

[#] Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA

[**] Department of Medicine, University of Calgary Cumming School of Medicine, Calgary, Alberta, Canada

[††] School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada

[‡‡] Department of Obstetrics and Gynecology, University of Ottawa, Ottawa, Ontario, Canada

[§§] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

[‖] Department of Emergency Medicine, University of Ottawa, Ottawa, Ontario, Canada

Total word count: 3,359

Funding: RGJ: Canadian Institutes of Health Research (CIHR). FDR: CIHR, University of Ottawa Heart Institute Foundation, Ernest and Margaret Ford Research Endowed Fund, Cardiac Arrhythmia Network of Canada (CANet) as part of the Networks of Centres of Excellence. None of the authors have relevant conflicts of interest to disclose.

Address for correspondence:

Dr. Benjamin Hibbert

University of Ottawa Heart Institute

40 Ruskin Street, H-4238

Ottawa, Ontario, Canada K1Y 4W7

Tel.: (613) 696-7280 | Fax: (613) 696-7245

Email: bhibbert@ottawaheart.ca

Structured Abstract

Objectives: To determine whether researcher sex is associated with differences in methodological rigor, including considering sex as a biological variable, in preclinical cardiovascular science.

Background: Women's involvement in research has been proposed to enhance knowledge outcomes and scientific progress, in part via enhanced exploration of sex differences.

Methods: The associations of author sex with reported study design and citation counts were examined in 3,396 preclinical studies published over 10 years in 5 leading cardiovascular journals. Temporal trends in female authorship and patterns in mentorship relationships by sex were secondarily assessed.

Results: Women accounted for 24 ± 17% of authors per manuscript, with 16.0% of articles not including any female authors and none being solely authored by women. Women were first, senior, and both first and senior authors in 41.3%, 20.7%, and 11.0%, respectively. Female authorship increased over the period examined; however, the proportion of articles with first and senior authors of differing sex remained stable. Female authors were more likely to study female animals, to study animals of both sexes, and to report sex-specific results ($P<0.001$ for all), but author sex was not associated with other measures of methodological rigor or 60-month citation counts.

Conclusions: Women's involvement in preclinical cardiovascular research is positively associated with considering sex as a biological variable – a practice that is expected to inform and promote advances in women's health. Researcher sex is not associated with other measures

of experimental rigor or research impact. Limited mentorship opportunities may be contributing to women's underrepresentation in this field.

Condensed Abstract:

In this analysis of 3,396 preclinical cardiovascular studies, women were first, senior, and both first and senior authors in 41.3%, 20.7%, and 11.0%, respectively. Female authorship increased over a 10-year period. However, the proportion of studies with first and senior authors of differing sex was low and stable, suggesting that segregation by sex in mentorship relationships exists and persists. Female authors were more likely to consider sex as a biological variable, but author sex was not associated with other measures of experimental rigor or research impact, indicating that women's underrepresentation is not due to differences in research capacity or impact.

Abbreviations

CI: confidence interval

NIH: National Institutes of Health

OR: odds ratio

Introduction

The now outdated axiom that cardiovascular disease is a disease of men began to be meaningfully challenged in the 1980's, with considerable effort focused on adequately representing women in clinical trials. Central to many of these initiatives was (and continues to be) the recognition that fundamental yet poorly understood differences exist between men and women and that they could engender health disparities if ignored. However, this 'revolution' has not permeated preclinical stages of research, which serve to inform clinical trials. In fact, the preferential use of male animals and a lack of sex-disaggregated reporting is *increasing* in cardiovascular science (1) despite the National Institutes of Health's (NIH) emphasis on considering sex in preclinical research as important for advancing women's health (2) and despite its feasibility in most research settings (3). This bias has the potential to undermine advances made in clinical trial design by skewing our understanding of disease processes and the effectiveness of potential therapies (4).

It has been argued that women's participation in research enhances knowledge outcomes and scientific progress (5), in part via enhanced exploration of sex differences (6). If true, sex gaps in medicine and research – including the underrepresentation of women in fields such as cardiology (7) and the relative lack of adequate mentorship for women (8) – would not only have important societal implications but also meaningful scientific ramifications. Recognizing this possibility, the NIH has emphasized their commitment to diversity in the biomedical research workforce (including better representation of women) and identified assessing the impact of this diversity on the quality and outputs of research as one of four major diversity challenges facing the biomedical "ecosystem" (9).

We therefore examined a large body of leading preclinical cardiovascular research to explore potential differences in experimental rigor between male and female researchers, focusing on the inclusion of female animals in experiments and on analyses of sex differences. Temporal trends in female authorship and patterns in mentorship relationships by sex were examined as secondary analyses. We hypothesized that female authorship and mentorship relationships between men and women had both increased but that these would not be associated with increased consideration of sex as a biological variable or with other measures of experimental rigor.

Methods

As described (10), all preclinical studies published between July 2006 and June 2016 in *Circulation*; *Circulation Research*; *Hypertension*; *Stroke*; and *Arteriosclerosis, Thrombosis, and Vascular Biology* were reviewed. Studies were included if they reported original data from *in vivo* experiments using non-human mammals and proposed therapeutic applications or implications to specific human disorders. Pre-specified variables collected included the disease studied, the animal model(s) used and their sex, and whether any study result was reported by sex. Given the possibility that differences in female animal inclusion or sex-specific analyses may be attributable to broader differences in experimental design, we also analyzed whether animals were randomized, whether blinding was used (concealed allocation or blinded outcome assessment), and whether sample size/power estimations were performed.

Study author names were extracted from Scopus. Author sex was determined using the online database genderize.io, which included >216,000 first names across 79 countries and 89

languages when queried. The database determines the sex of a name with an associated certainty factor. First authors are generally considered to have executed most of the published work whereas senior authors are usually the study supervisors or principal investigators and are generally considered to have contributed most to study planning and design (11,12); therefore, these were selected as our primary analyses. Female authorship as a proportion of all authors per manuscript was examined as a secondary analysis (per 10% increase and dichotomized as ≥33% of authors). Scopus was also queried to determine citation counts at 60 months for articles published between July 2006 and June 2011, as described (10). *Post hoc* analyses of mentorship relationships by sex were performed, with senior authors designated as mentors and first authors as mentees. A certainty factor of ≥70% was used to assign author sex for all analyses, otherwise sex was considered 'unknown'. Articles with authors of 'unknown' sex were excluded in primary analyses but were included in secondary and *post hoc* analyses to minimize missing data. Sensitivity analyses using a certainty factor of ≥90% for author sex and of sex-specific reporting restricted to studies in which animals of both sexes were used were performed.

Categorical variables are reported as number (percentage) and were compared using χ^2 tests. Continuous variables are reported as mean ± standard deviation or median with interquartile range (25^{th} to 75^{th} percentiles) range (IQR) and were compared using *t* tests or Mann-Whitney *U* tests, respectively. Temporal trends were assessed using Cochran-Armitage tests or Spearman rank-order correlation (ρ) using 12-month intervals. The associations of author sex with factoring the sex of the animals studied in the reporting, design, and analysis of experiments and with the implementation of other study design elements were examined using χ^2 tests (reported as absolute percent differences in study characteristics) and via simple and multivariable logistic regression (reported as odds ratios [ORs] with 95% confidence intervals [CIs]). Citation count

comparisons were performed via stratification and multivariable linear regression. All analyses were performed using SAS 9.4 (SAS Institute, Inc., Cary, NC) using a two-tailed α level of 0.05 (corrected using the Bonferroni method to account for multiple comparisons when specified).

Results

Of 28,636 articles screened, 3396 met study inclusion criteria (**Figure 1**). Women accounted for 24 ± 17% of authors per manuscript, with 542 articles (16.0%) not including any female authors and none being authored solely by women. After excluding articles with authors of unknown sex, women were identified as first authors of 1016/2458, senior authors of 569/2749, and both first and senior authors of 235/2135 preclinical studies (41.3%, 20.7%, and 11.0%, respectively). The distribution of observed mentorship relationships differed from the expected distribution, with disproportionately high numbers of same-sex mentorships identified ($P<0.001$), including a relative 50% greater than expected frequency of male-male mentorships. No difference was observed in the number of co-authors of articles with female vs. male first or senior authors (9.6 vs. 9.6, $P=0.864$; and 9.2 vs. 9.5, $P=0.075$; respectively). Five-year citation counts were comparable between female and male authors (21 [13-31] vs. 20 [12-33], $P=0.509$ for first authors; 18 [12-30] vs. 20 [13-31], $P=0.170$ for senior authors). Author sex was similarly not predictive of citation counts in multivariable models adjusting for cardiovascular disease studied, publication date, and journal.

Temporal analyses over 10 years indicate that the proportions of female first and senior authors have each increased, with women accounting for between 32.3% and 46.3% of first authors and between 11.8% and 26.1% of senior authors (**Figure 2A**). There has been a corresponding

temporal increase in the proportion of manuscripts with women as both first and senior authors (range 5.5-14.9%, P_{trend}=0.001). The proportions of articles with first and senior authors of differing sex have not changed (range 28.5-34.3%, P_{trend}=0.663 for female first and male senior authorship; range 5.5-14.9%, P_{trend}=0.184 for male first and female senior authorship). Per article, the mean proportion of female authors has slightly increased (range 21.4-26.8%, P<0.001) (**Figure 2B**).

Female and male authors were equally likely to report the sex of animals used in preclinical experiments (815/1016 [80.2%] vs. 1150/1442 [79.8%], P=0.776 for first authors; 455/569 [80.0%] vs. 1767/2180 [81.1%], P=0.556 for senior authors). However, when the sex of the animals was reported, women were more likely to have used female animals in their experiments (292/815 [35.8%] vs. 313/1150 [27.2%], P<0.001 for first authors; 181/455 [39.8%] vs. 492/1767 [27.8%], P<0.001 for senior authors), to have used animals of both sexes (156/815 [19.1%] vs. 179/1150 [15.6%], P=0.038 for first authors; 98/455 [21.5%] vs. 262/1767 [14.8%], P<0.001 for senior authors), and to have reported sex-specific results (326/815 [40.0%] vs. 380/1150 [33.0%], P=0.002 for first authors; 199/455 [43.7%] vs. 597/1767 [33.8%], P<0.001 for senior authors). Similar findings were observed when female authorship was examined as a proportion of all authors per manuscript. In contrast, female authorship was not associated with randomization, blinding, or sample size estimation after correcting for multiple comparisons (**Figure 3**). The above differences persisted when female authorship was examined on an interval scale (per 10% increase) and after adjusting for pre-specified potential confounders (**Table 1**). All findings were also comparable in sensitivity analyses using a minimum certainty factor of 90% for author sex. When restricted to studies that included both male and female animals (N=335-421), female senior authorship (but not first or cumulative authorship) remained

associated with reporting sex-specific results after adjusting for pre-specified potential confounders (OR 2.2 [95% CI 1.1-4.4], *P*=0.036).

Discussion

Sex differences in innate scientific ability have long ago been refuted (13); however, career trajectories and personal and professional experiences often differ between male and female academics (8,13), which may influence researchers' priorities and practices. Although there are clear societal benefits to promoting sex inclusivity in research, its impact on the quality and impact of research remains unexplored (9). Our analysis of preclinical cardiovascular research demonstrates that [1] 41% of first authors and 21% of senior authors over a recent 10-year period have been women; [2] although female authorship is increasing, segregation by sex in mentorship relationships exists and persists; [3] when women influence experimental design and reporting or comprise a larger proportion of the research team, studies are more likely to include female animals and to explore sex-based differences, but author sex is not associated with other measures of methodological rigor or research impact.

Preclinical research using animal models often precedes and informs clinical trials. However, important and remediable methodological shortcomings remain prevalent in preclinical cardiovascular research (10), including a tendency to exclude females in animal experiments and to ignore the potential influence of sex on study outcomes (1). Women's perspectives have been suggested to uniquely promote scientific progress in general and advances in women's health in particular, in part due to their greater tendency to explore the effects of gender and sex (5,6). Our data suggest that women are indeed more attuned to considering sex as a biological variable in

preclinical experiments, but that male and female researchers are more alike than they are different when broader measures of scientific rigor and research impact are considered.

A corollary to the above findings is that the persistent underrepresentation of women in preclinical cardiovascular research is highly unlikely to be attributable to differences in research capacity or potential impact. Rather, systemic factors are probably influential. For instance, although modern scientific endeavours are increasingly reliant on research teams and networks – settings in which a diversity of viewpoints and experiences are sought and valued (6) – our analysis highlights a persistent predominance of same-sex mentorships. Given the relative paucity of available female mentors, this practice has the potential to perpetuate women's underrepresentation in the field and to hinder efforts to improve the status quo (7).

Our study has important limitations. The journals examined were selected based on their prominence in cardiovascular research, collective focus on a wide range of cardiovascular disorders, and endorsement of NIH guidelines on rigor and reproducibility (10). However, they may not be representative of all preclinical cardiovascular journals. Author sex was determined using an arbitrary certainty factor, which may have resulted in misclassification in a minority of cases. However, our results were comparable in sensitivity analyses using a more stringent criterion. Our analysis did not capture instances of multiple first or corresponding authors (equally contributing authors) and used author position as a proxy for mentor-mentee relationships. Presumed author sex may not reflect author gender, which may be a relevant distinction in our analysis. Our analysis also focused on experimental and reporting standards proposed by the NIH, which are not exhaustive.

Conclusions

Women's involvement in preclinical cardiovascular research is positively associated with considering sex as a biological variable – a practice that is expected to inform and promote advances in women's health. Researcher sex is not associated with other measures of experimental rigor or research impact. Limited mentorship opportunities may be contributing to women's underrepresentation in this field.

Perspectives

Clinical Competencies: In this analysis of 3,396 preclinical studies in 5 leading cardiovascular journals, female authorship was positively associated with considering sex as a biological variable, but not with other measures of methodological rigor or 60-month citation counts. Over a 10-year period, the proportion of studies with first and senior authors of differing sex was low and stable, suggesting that segregation by sex in mentorship relationships exists and persists. These data indicate that women's underrepresentation in preclinical cardiovascular research is not due to differences in research capacity or impact. Limited mentorship opportunities for women in preclinical cardiovascular research may be an important contributor.

Translational Outlook: Further study to identify and understand barriers to women's involvement in preclinical cardiovascular research is warranted. Enhancing mentorship opportunities for women should be explored as a potential strategy to improve the status quo.

References

1. Ramirez FD, Motazedian P, Jung RG et al. Sex Bias Is Increasingly Prevalent in Preclinical Cardiovascular Research: Implications for Translational Medicine and Health Equity for Women: A Systematic Assessment of Leading Cardiovascular Journals Over a 10-Year Period. Circulation 2017;135:625-626.
2. Clayton JA, Collins FS. Policy: NIH to balance sex in cell and animal studies. Nature 2014;509:282-3.
3. Beery AK. Inclusion of females does not increase variability in rodent research studies. Curr Opin Behav Sci 2018;23:143-149.
4. Klein SL, Schiebinger L, Stefanick ML et al. Opinion: Sex inclusion in basic research drives discovery. Proc Natl Acad Sci U S A 2015;112:5257-8.
5. Rotenstein LS, Jena AB. Lost Taussigs - The Consequences of Gender Discrimination in Medicine. N Engl J Med 2018;378:2255-2257.
6. Nielsen MW, Alegria S, Borjeson L et al. Opinion: Gender diversity leads to better science. Proc Natl Acad Sci U S A 2017;114:1740-1742.
7. Asghar M, Usman MS, Aibani R et al. Sex Differences in Authorship of Academic Cardiology Literature Over the Last 2 Decades. J Am Coll Cardiol 2018;72:681-685.
8. Soklaridis S, Zahn C, Kuper A, Gillis D, Taylor VH, Whitehead C. Men's Fear of Mentoring in the #MeToo Era - What's at Stake for Academic Medicine? N Engl J Med 2018.
9. Valantine HA, Collins FS. National Institutes of Health addresses the science of diversity. Proc Natl Acad Sci U S A 2015;112:12240-2.

10. Ramirez FD, Motazedian P, Jung RG et al. Methodological Rigor in Preclinical Cardiovascular Studies: Targets to Enhance Reproducibility and Promote Research Translation. Circ Res 2017;120:1916-1926.
11. Wren JD, Kozak KZ, Johnson KR, Deakyne SJ, Schilling LM, Dellavalle RP. The write position. A survey of perceived contributions to papers based on byline position and number of authors. EMBO Rep 2007;8:988-91.
12. Lariviere V, Desrochers N, Macaluso B, Mongeon P, Paul-Hus A, Sugimoto CR. Contributorship and division of labor in knowledge production. Soc Stud Sci 2016;46:417-435.
13. Ceci SJ, Ginther DK, Kahn S, Williams WM. Women in Academic Science: A Changing Landscape. Psychol Sci Public Interest 2014;15:75-141.

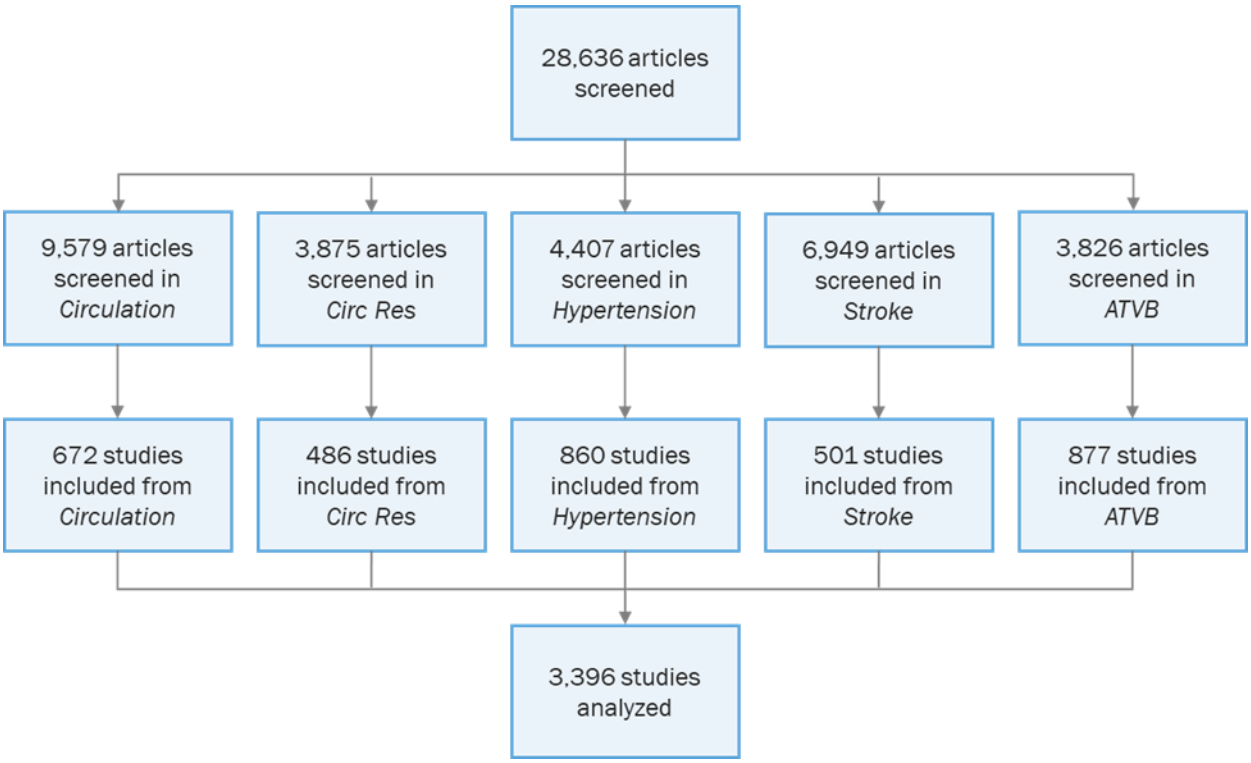


Figure 1: Literature search. *ATVB*: *Arteriosclerosis, Thrombosis and Vascular Biology*; *Circ Res*: *Circulation Research*. Modified from Ramirez *et al* (10).

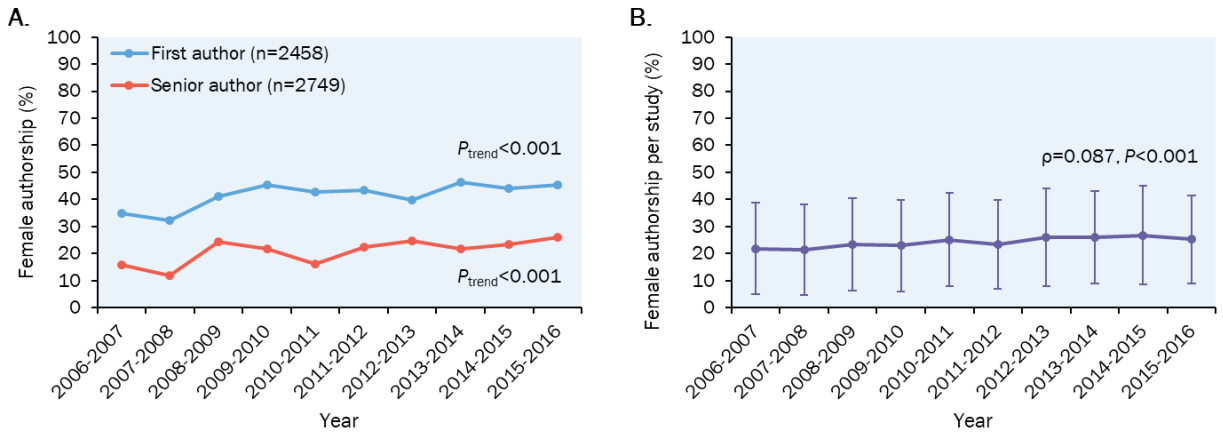


Figure 2: Female authorship in preclinical cardiovascular research. *A.* Proportion of women as first and senior authors. *B.* Mean proportion of women as authors per study. Error bars depict standard deviation.

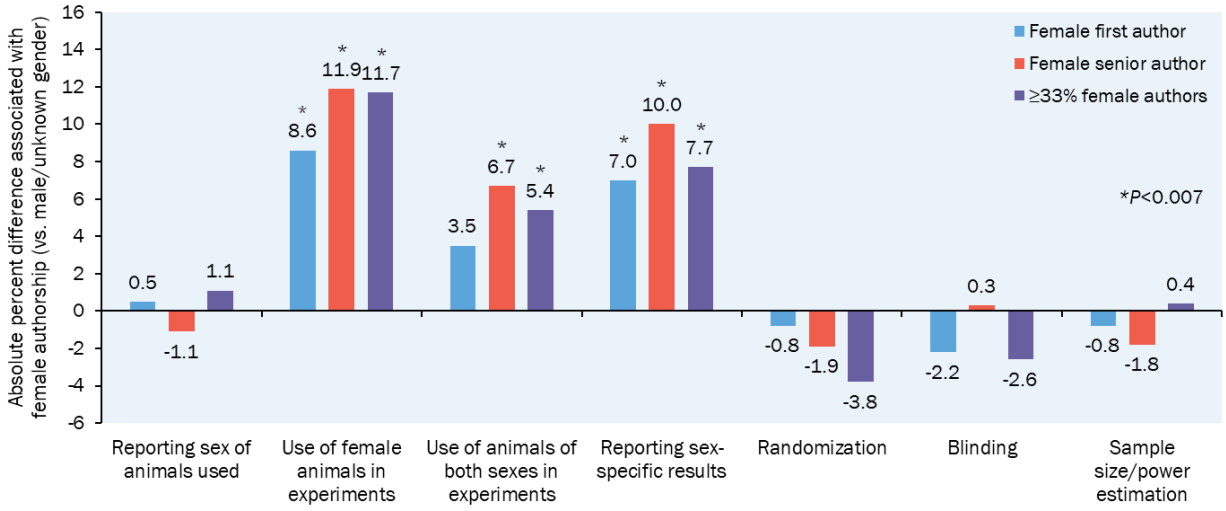
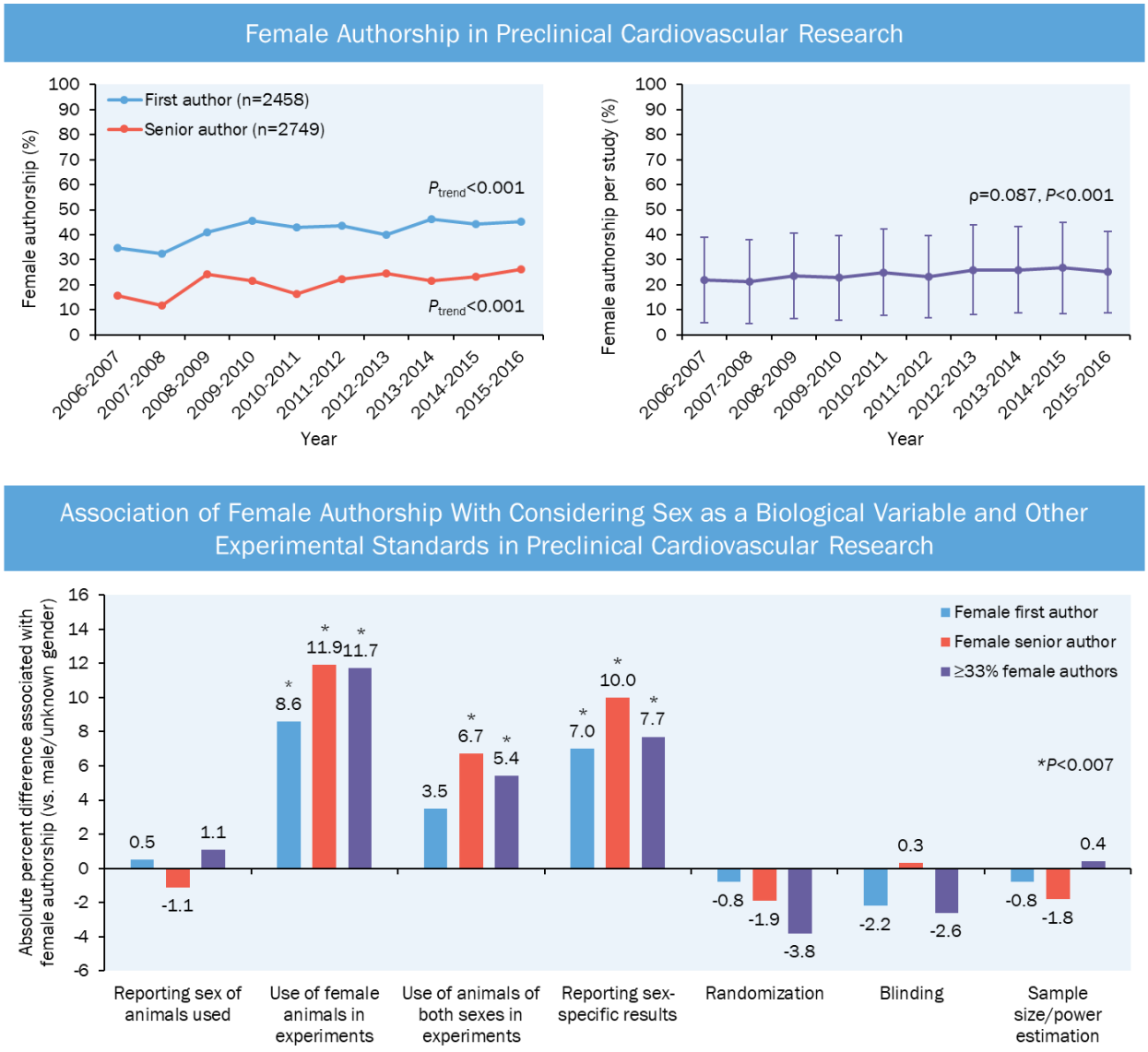


Figure 3: Association of female authorship with considering sex as a biological variable and other experimental standards in preclinical cardiovascular research. Absolute percent differences shown for first, senior, and ≥33% female authorship. *Corrected using the Bonferroni method to account for multiple comparisons (=0.05/7).



Visual Abstract: Female authorship in preclinical cardiovascular research and its association with considering sex as a biological variable and with other experimental standards.

Table 1. Association between female authorship and experimental design characteristics in preclinical cardiovascular studies

	*N**	Crude OR (95% CI)	Adjusted OR (95% CI)†	*P*†
Female first author				
Reporting sex of animals	2458	1.03 (0.84,1.26)	0.96 (0.77,1.19)	0.698
Inclusion of female animals‡	1965	1.49 (1.23,1.81)	1.60 (1.30, 1.96)	<0.001
Inclusion of animals of both sexes‡	1965	1.28 (1.01, 1.63)	1.29 (1.01, 1.66)	0.046
Sex-specific reporting of results‡	1965	1.35 (1.12, 1.63)	1.33 (1.10, 1.62)	0.004
Randomization	2458	0.95 (0.78, 1.16)	1.03 (0.83, 1.28)	0.776
Blinding	2458	0.91 (0.76, 1.08)	0.95 (0.79, 1.15)	0.608
Sample size estimation	2458	0.70 (0.41, 1.20)	NR	
Female senior author				
Reporting sex of animals	2749	0.93 (0.74, 1.18)	0.83 (0.65, 1.07)	0.153
Inclusion of female animals‡	2222	1.71 (1.38, 2.12)	1.81 (1.43, 2.28)	<0.001
Inclusion of animals of both sexes‡	2222	1.58 (1.22, 2.04)	1.58 (1.20, 2.08)	0.001
Sex-specific reporting of results‡	2222	1.52 (1.24, 1.88)	1.44 (1.16, 1.79)	0.001
Randomization	2749	0.90 (0.71, 1.12)	0.88 (0.68, 1.13)	0.308
Blinding	2749	1.02 (0.84, 1.23)	1.14 (0.92, 1.42)	0.219
Sample size estimation	2749	0.32 (0.13, 0.80)	NR	
Female authorship (per 10% increase)				
Reporting sex of animals	3396	1.04 (0.98, 1.09)	1.02 (0.97, 1.09)	0.435

Inclusion of female animals‡	2718	1.20 (1.14, 1.26)	1.20 (1.13, 1.26)	<0.001
Inclusion of animals of both sexes‡	2718	1.13 (1.07, 1.21)	1.12 (1.05, 1.20)	<0.001
Sex-specific reporting of results‡	2718	1.13 (1.08, 1.19)	1.11 (1.05, 1.17)	<0.001
Randomization	3396	0.95 (0.90, 1.00)	1.00 (0.94, 1.06)	0.949
Blinding	3396	0.99 (0.94, 1.03)	1.04 (0.99, 1.09)	0.173
Sample size estimation	3396	1.08 (0.94, 1.24)	NR	

*Refers to the total number of studies included in analyses. †Adjusted for disease studied, animal model(s) used, journal of publication, date of publication, and number of co-authors. ‡Analysis restricted to studies in which the sex of animals used was reported. Bonferroni-corrected α level of 0.007 used to define statistical significance (=0.05/7). NR: not reported due to small number of events per predictor variable. OR: odds ratio.

CHAPTER 6: ANALYSES OF JOURNAL INITIATIVES TO IMPROVE THE REPORTING AND METHODOLOGICAL QUALITY OF PRECLINICAL RESEARCH

The following is a manuscript submitted for publication that explored the hypothesis that journal initiatives such as those implemented by *Stroke* are associated with improvements in the reporting and methodological quality of published preclinical research.

Additional documentation related to this manuscript is provided in the following appendices:

Appendix J: *Stroke*'s 'Basic Science Checklist' (from Vahidy *et al*. *Stroke* 2016;47:2435-2438)

Appendix K: *Nature* journals' Life Sciences Reporting Summary (example)

Journal initiatives to enhance preclinical research: analyses of *Stroke*, *Nature Medicine*, *Science Translational Medicine*

F. Daniel Ramirez MD,[a,b,c] Richard G. Jung BSc,[c,d,e,f] Pouya Motazedian MD,[c,d,f,g] Dylan Perry-Nguyen BMSc,[c,d,f] Pietro Di Santo MD,[a,b,c] Zachary MacDonald MD,[f,h] Aisling A. Clancy MD MSc,[i,j] Alisha Labinaz,[c,d,k] Steven Promislow MD,[a,c] Trevor Simard MD,[a,c,d,e] Steeve Provencher MD MSc,[l,m] Sébastien Bonnet PhD,[l,m] Ian D. Graham PhD,[b,n] George A. Wells PhD,[b,o] Benjamin Hibbert MD PhD[a,c,d,e]

[a] Division of Cardiology, University of Ottawa Heart Institute, Ottawa, Ontario, Canada

[b] School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada

[c] CAPITAL Research Group, University of Ottawa Heart Institute, Ottawa, Ontario, Canada

[d] Vascular Biology and Experimental Medicine Laboratory, University of Ottawa Heart Institute, Ottawa, Ontario, Canada

[e] Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, Ontario, Canada

[f] Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada

[g] Department of Medicine, University of Calgary Cumming School of Medicine, Calgary, Alberta, Canada

[h] Department of Emergency Medicine, University of Ottawa, Ottawa, Ontario, Canada

[i] Department of Obstetrics and Gynecology, University of Ottawa, Ottawa, Ontario, Canada

[j] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

[k] Faculty of Science, University of Ottawa, Ottawa, Ontario, Canada

[l] Pulmonary Hypertension Research Group, Centre de recherche de l'Institut universitaire de cardiologie et de pneumologie de Québec, Université Laval, Québec City, Québec, Canada

[m] Department of Medicine, Université Laval, Québec, Québec City, Canada

[n] Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada

[o] Cardiovascular Research Methods Centre, University of Ottawa Heart Institute, Ottawa, Ontario, Canada

Social media handles: @fdanielramirez @benhibbertMDPhD

Cover title: Journal initiatives to enhance preclinical studies

Tables: 1

Figures: 5

Key words: research methodology, reproducibility, animal model, bias

Subject terms: animal models of human disease, translational studies, basic science research, quality and outcomes, statements and guidelines

Word count: 5,483

Corresponding author:

Benjamin Hibbert, MD PhD

University of Ottawa Heart Institute

40 Ruskin Street, H-4238

Ottawa, Ontario, Canada K1Y 4W7

Tel.: 613-696-7280

Fax: 613-696-7245

Email: bhibbert@ottawaheart.ca

Abstract

Background and Purpose: Preclinical research using animals often informs clinical trials. However, its value is dependent on its scientific validity and reproducibility, which are in turn dependent on rigorous study design and reporting. In 2011, *Stroke* introduced a 'Basic Science Checklist' to enhance the reporting and methodology of its preclinical studies. Except for *Nature* and *Science*, few publishing groups have implemented similar initiatives. We sought to estimate the impact of these journal interventions on the quality of their published reports.

Methods: All articles published in *Stroke*, *Nature Medicine*, and *Science Translational Medicine* over 9 to 18 years and in 2 control journals without analogous interventions over a corresponding 11.5 years were reviewed to identify reports of experiments in non-human mammals with proposed clinical relevance. The effect of journal interventions on the reporting and use of key study design elements (SDEs) was estimated via interrupted time series analyses.

Results: Of 33,009 articles screened, 4,162 studies met inclusion criteria. In the 3.5 to 12 years preceding each journal's intervention, the proportions of studies reporting and using key SDEs were stable except for blinding in *Stroke* and randomization in *Science Translational Medicine*, which were both increasing ($P_{\text{trend}} \leq 0.007$). Post-intervention, abrupt and often marked increases were seen in the reporting of randomization status (level change: +17% to +44%, $P \leq 0.005$), blinding (level change: +20% to +40%, $P \leq 0.008$), and sample size estimation (level change: 0% to +40%, $P \leq 0.002$ in 2 journals). Significant but more modest improvements in the use of these SDEs were also observed. These improvements were not seen in control journals.

Conclusions: Journal interventions such as *Stroke*'s author submission checklist can meaningfully improve the quality of published preclinical research and should be encouraged to

enhance study transparency and design. However, such interventions are alone insufficient to fully address widespread shortcomings in preclinical research practices.

Preclinical experiments using animal models are an integral part of biomedical research; however, their value is dependent on their scientific validity and reproducibility,[1, 2] which are in turn dependent on rigorous study design and reporting.[3, 4] Poorly designed or inadequately reported preclinical studies can therefore reduce the dividends from research investments,[5, 6] result in animals being unjustifiably subjected to harm,[2, 7, 8] and inappropriately spur or deter clinical trials in humans.[3, 9, 10] Randomization, blinding, sample size estimation, and considering sex as a biological variable are considered essential study design elements (SDEs) to improve the reproducibility and predictive value of preclinical experiments.[1, 8, 11-14] Randomization protects against selection bias; blinding protects against performance, detection, and attrition biases; sample size estimation ensures adequate statistical power, which improves effect estimate precision and protects against spurious conclusions; and considering sex as a biological variable increases the likelihood that results will be relevant to both men and women.[2, 5, 7, 14-16] These SDEs are routinely used in clinical trials, but are rarely implemented at the preclinical stage, even though animal and human studies are susceptible to similar risks of bias and threats to study validity.[8, 12, 14, 15]

Recently, the National Institutes of Health (NIH) and prominent journals, including *Nature* and *Science* publishing groups, have focused considerable effort on improving the rigor and reproducibility of preclinical research, particularly through enhanced experimental conduct and reporting standards.[1, 11, 13, 17] In May 2013, *Nature* journals introduced editorial measures that included a mandatory checklist to prompt authors to disclose information regarding "crucial" experimental SDEs[18] – a similar initiative to the 'Basic Science Checklist' introduced by *Stroke* in 2011.[19] *Science Translational Medicine* (*Science TM*) similarly began requiring that authors describe and defend their study designs, including disclosing whether key SDEs were

implemented.[20] Yet, robust data on the impact of such interventions are limited and at times contradictory.[12, 19, 21-23] We hypothesized that the introduction of targeted journal initiatives had improved the reporting and experimental design of published preclinical research.

Materials and Methods

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Study screening and selection

All studies published in *Stroke* between January 2000 and December 2017 were reviewed. The start date of January 2000 was selected as the first Stroke Academic Industry Roundtable (STAIR) guidelines on preclinical experiments were published exclusively in *Stroke* in December 1999,[24] arguably marking an editorial awareness of their goals and importance. The screening and study selection process for this journal has been described previously.[12] Briefly, studies were included if they described original research published as full manuscripts; reported results of *in vivo* experiments in non-human mammals; and described pathophysiology, genetics, and/or therapeutic interventions that were stated to be directly relevant to a specific disorder in humans. Studies on physiologic or genetic characteristics were included if potential therapeutic applications or implications were proposed in the article. These criteria are consistent with proposed definitions of preclinical ("confirmatory" or "proof-of concept") experiments.[2, 25, 26] All articles in *Stroke* were independently selected and data extracted using standardized case report

forms by two reviewers, including via review of all articles' supplementary materials. Discrepancies were resolved by consensus or by an independent adjudicator. Our group has reported 94.5% interrater agreement for study inclusion before resolution using this approach in a previous review of >28,000 articles.[12]

All studies published in *Nature Medicine* and *Science TM* between 2009 and 2017 were similarly reviewed, but the study selection process differed slightly for these journals as described in the supplemental methods. Data extraction was performed by two independent assessors as was done for *Stroke*. The starting year of 2009 for data collection was selected because the inaugural issue of *Science TM* was published in that year.

We previously examined the reported methodological quality of articles published in an additional four leading cardiovascular journals between July 2006 and June 2016 using before-after comparisons.[12] Two of these journals (*Circulation* and *Circulation Research*) publish on all aspects of cardiovascular medicine, allocate a substantial amount of space to preclinical research that is of similar reported methodological quality,[12] have among the highest journal impact factors in the cardiovascular field, and did not explicitly and publicly implement initiatives to improve the reporting or methodology of their manuscripts until October 2017.[17] Limited journal- and disease-specific data from those before-after comparisons have been reported previously.[12, 14, 27] The dataset from these journals was extended to December 2017, expanded in scope (see below), and combined into a single 'control' group to increase statistical power. Study selection and data extraction for these control journals were performed as was done for *Stroke*.

Diseases or disease systems of interest in each journal are listed in the supplemental methods. Studies on sex-specific diseases (for instance, prostate or ovarian cancer) were prospectively identified during data collection.

Data extraction

The date of publication, disease system studied, and animal model(s) used were collected for all articles. Disease systems studied and animal models used in <2% of reports in all journals were included in 'other' categories. Data on randomization, blinding (concealed allocation or blinded outcome assessment), sample size/power estimations, and the sex of the animals used were also collected. These data have been identified as minimum reporting requirements by the NIH.[2, 13] A distinction was made between the reporting of any given SDE and its reported use such that cases in which studies reported *not* randomizing animals to treatment groups or *not* using blinding, for instance, were captured. Similarly, studies that reported using a specific SDE, but its description revealed that it had not been properly executed, were captured (e.g. studies that reported being adequately powered because group comparisons yielded statistically significant results). The reporting and use of SDEs are therefore described separately. These elements were deemed to have been reported on or used if done for any experiment or animal model within a manuscript. Because of the small number of *Science TM* studies published in 2009 (the first issue was published in October of that year), studies from 2009 and 2010 were combined for that journal for all analyses. Similarly, studies published in control journals between July and December 2006 were included with articles published in 2007.

Given the volume of articles screened, journal-specific stylistic features (page layouts, article organization, color schemes, and appearance of figures), and the variability and volume of online supplementary materials, de-identification of articles was not performed and reviewers were not blinded to journal or date of publication during screening or data extraction, although they were blinded to scores from other reviewers. Articles were also reviewed in a predominantly chronological sequence rather than in a randomized fashion.

Statistical analysis

Journal-specific sample size calculations are described in the supplemental methods. Categorical variables are reported as number (%) and were compared via chi-square tests. Continuous variables are reported as mean ± standard deviation (SD) and were compared using *t*-tests. Interrater agreement for study inclusion for *Stroke* and control journals was calculated using percent agreement and Cohen's kappa statistic. Secular trends in the number of preclinical studies published over time were examined using Spearman rank correlation coefficients (ρ); trends in the proportions of studies reporting on or using SDEs were examined using Cochrane-Armitage trend tests. Randomization, blinding, and sample size estimation were specifically targeted in all initiatives studied; therefore, temporal changes in the proportions of studies reporting on and using these SDEs were evaluated as journal-specific interrupted time series with editorial initiatives introduced as step functions in segmented regression models and are reported as absolute cumulative means or absolute percentages with 95% confidence intervals (CIs). Data on sex reporting and sex of animals used were similarly but separately evaluated and excluded studies on sex-specific disorders. For temporal analyses, *Stroke* was examined in 12-month intervals whereas *Nature Medicine*, *Science TM*, and control journals were examined in 6-month intervals to render sample sizes similar per time point between journals. The influence of journal initiatives was also estimated by comparing reporting and methodological quality measures before versus after their implementation in each journal and via simple and multivariable modified Poisson regression models with robust error variance (sandwich estimation) to generate journal-specific crude and adjusted relative risks (RRs) with 95% CIs. Covariate category inclusion in multivariable models was pre-specified and included relevant disease system studied and animal model used. Within these covariate categories, specific disease system and animal

model covariates were selected via backward elimination using an inclusion criterion of $P<0.20$ to ensure adequate events per predictor variable. All analyses were performed using SAS 9.4 (SAS Institute Inc., Cary, NC) using a two-tailed α level of 0.05 to define statistical significance.

Results

Study selection and characteristics

Of 33,009 indexed articles, 4,162 met inclusion criteria and were analyzed (Figure I). Interrater agreement for study inclusion before resolution was 97.9% ($\kappa=0.82$). The number of relevant articles published yearly in *Nature Medicine* and in control journals remained stable over the period examined (105 ± 10, $\rho=0.234$, $P=0.544$ and 118 ± 20, $\rho=-0.378$, $P=0.252$), but increased and decreased in *Science TM* (108 ± 28, $\rho=0.976$, $P<0.001$) and *Stroke* (55 ± 13, $\rho=-0.634$, $P=0.005$), respectively.

In *Stroke* and in control journals, 95.4% of studies focused on cardiovascular diseases and 86.3% used rats or mice. In *Nature Medicine* and *Science TM*, the most commonly studied disease systems were oncology (24.8%), infectious diseases (12.9%), neurology (12.4%), endocrinology (9.9%), cardiovascular (8.8%), and 'other' (9.9%), with 78.8% using mice and 10.2% using multiple animal models. Psychiatry, anesthesia, obstetrics/gynecology, ophthalmology, and toxicology were each studied in <2% of studies in all journals and were therefore included in the 'other' category (Table I). One hundred eight articles were deemed to be sex-specific and were excluded from sex-based analyses.

Stroke

In the 12 years leading up to and including the year in which *Stroke* introduced its 'Basic Science Checklist', the proportions of studies that reported on and used blinding were increasing (P_{trend}=0.002 for both) whereas the reporting or use of randomization and sample size estimations were stable. Beginning in 2012, significant trend level and/or slope increases in the reporting and use of these SDEs were observed (Figure 1). No study reported not having randomized animals and two studies reported not having used blinding after the checklist was introduced (Table 1). Among studies that reported having randomized animals, the proportion describing the method of randomization increased from 6.0% to 16.4% (P<0.001). Temporal increases in the cumulative numbers of reported and used SDEs were observed (Figure II).

Nature Medicine

In the 4.5 years preceding *Nature Medicine*'s author submission checklist, the reporting of randomization, blinding status, or sample size estimation remained stable. After the journal implemented its checklist, significant improvements in the trend lines for the reporting and use of randomization, blinding, and sample size estimation were observed. Large numbers of articles reported *not* having implemented various SDEs as shown by diverging trend lines of the reporting and use of individual SDEs after the initiative (Figure 2) and as shown in Table 1. Among studies that reportedly randomized animals, the proportion describing the method used for randomization did not appreciably change following checklist implementation (2.0% vs. 5.7%, P=0.480). Corresponding temporal increases in the cumulative number of reported and used SDEs per article were detected (Figure II).

Science Translational Medicine

In the 3.5 years before *Science TM* revised its reporting requirements and policies, both the reporting and use of randomization were increasing (P_{trend}=0.004 and 0.007, respectively) whereas the reporting and use of blinding and sample size estimation were stable. Following the introduction of this journal's initiative, significant improvements were also observed in the reporting of randomization, blinding status, and sample size estimations and in the use of randomization and sample size estimations. As was seen in *Nature Medicine*, a substantial number of studies reported not having used specific SDEs, particularly blinding and power calculations (Figure 3, Table 1), but among studies that reported using randomization there was no change in the proportion describing how it was performed (0% vs. 4.8%, *P*=0.143). Temporal increases in the cumulative numbers of reported and used SDEs were also detected (Figure II).

Control journals

In contrast, in control journals, there were no temporal improvements in the reporting or use of randomization, blinding, or sample size calculations over a corresponding 11.5-year period (Figure 4), excepting a modest improvement in the reporting of sample size calculations beginning in 2012 when analyzed using *Stroke*'s initiative as the time series interruption (1% absolute trend line slope increase, 95% CI 0-2%, *P*=0.049) (data not shown). No changes in the cumulative numbers of reported or implemented SDEs were detected in these control journals (Figure II).

Sex reporting and inclusion

The proportions of articles published in *Stroke* that reported the sex of the animals used and included both males and females in experiments were stable prior to its checklist being introduced and did not improve after the initiative was introduced. The proportions of studies in *Nature Medicine* and *Science TM* that reported the sex of the animals used and included animals of both sexes were increasing prior to their respective interventions ($P_{trend} \leq 0.040$ for both in both journals), but also did not improve following the journals' initiatives. In control journals, prior to *Nature Medicine*, *Science TM*, or *Stroke*'s interventions, the proportions of articles reporting the sex of animals used were stable and the proportions reporting the use of both males and females were decreasing ($P_{trend} < 0.001$). A trend level increase (+10%, 95% CI 0-20%, $P=0.050$) and slope increase (+2%, 95% CI 0-3%, $P=0.014$) in the use of both sexes were observed in these journals when analyzed using *Stroke*'s initiative as the time series interruption, but not when analyzed using *Nature Medicine* or *Science TM*'s initiatives. A marked and persistent preferential use of male animals was evident in articles in *Stroke* and control journals (Figure 5).

Adjusted analyses

All temporal patterns persisted in before-after comparisons adjusting for the most relevant disease studied and animal model(s) used (Table 1).

Discussion

Time series analyses identified significant and marked improvements in the quality of published preclinical studies coinciding with the introduction of relevant initiatives in three journals – findings that persisted after accounting for secular trends in reporting and methodological

practices and that were not seen in influential journals without analogous interventions over the same period. These findings strongly support the notion that journal interventions can meaningfully influence the quality of published preclinical research.

Mechanistically, well-designed author submission checklists and analogous journal initiatives could improve the quality of published preclinical research in at least four ways: [1] by improving the reporting of study methods without affecting the methodology used (i.e. improving transparency for readers);[28] [2] by improving transparency for journal editors and peer reviewers, which could influence whether a manuscript is accepted for publication;[29] [3] by influencing authors' decision to submit to a particular journal (e.g. dissuading authors from submitting a manuscript if their study is felt to have methodological shortcomings that will be highlighted);[17] and [4] by influencing experimental study design practices among researchers who hope to publish in a particular journal.[28] However, few robust analyses of the impact of journal initiatives on published research quality exist. The Nature Publication Quality Improvement Project (NPQIP) Collaborative group examined the potential impact of *Nature* journals' checklist on the reporting quality of *in vivo* research in 448 articles using primarily a controlled before-after study design.[21] Our analysis of *Nature* journals' publications differs by focusing on *Nature Medicine* (considered to be focused on translational research) and by evaluating a larger sample size (n=949) over a longer period to increase our confidence that changes observed were due to the publishing group's intervention rather than secular trends in editorial practices or culture. Nevertheless, our data support their conclusions and expand upon them by demonstrating improvements in study quality after instituting reporting and methodological standards at other journals. The recently reported IICARus (Intervention to Improve Compliance with the ARRIVE guidelines) study[22] randomized manuscripts reporting *in*

vivo animal research submitted to *PLoS One* to journal-requested completion of the ARRIVE (Animal Research: Reporting of In Vivo Experiments)[29] checklist versus standard practice (no checklist). Contrasting with our results and those of the NPQIP study, the study found no difference in the proportion of articles fully complying with the ARRIVE guidelines. However, the primary outcome in IICARus was reporting compliance with *all* 38 items in the checklist, which authors may have deemed onerous or judged as unlikely to be enforced.

The journal interventions included in the present analysis improved the reporting and use of SDEs to varying degrees. Differences in the nature of the initiatives may account for some of this. For instance, the checklist introduced by *Nature* journals requires authors to describe whether and how specific SDEs were used.[18] In contrast, *Stroke*'s Basic Science Checklist, when first introduced, was a simple and brief questionnaire consisting of dropdown menus with limited options[19] (although it has since been updated to a more comprehensive version[30]). *Science TM* does not have a readily accessible checklist for authors, but instead provides instructions in its 'Information for Authors' section, which its Editor has stated are enforced.[20] Differences in how vigorously journals enforce their requirements and how carefully they screen for author adherence are also be expected to be influential, could fluctuate over time, and may be affected by how attuned reviewers and editors are to them.

The use of key SDEs overall remains low in preclinical research, even in recent years and despite the prominence of the journals examined – a fact that is more clearly brought to light with improved reporting. Based on reported methods, approximately 40-50% did not use any randomization, >50% did not use any blinding, and up to 75% did not ensure adequate statistical power. Additionally, consistent across journals was a lack of improvement in the inclusion of animals of both sexes after accounting for baseline secular trends. Despite the NIH requiring sex

inclusion plans in preclinical research funding applications and the importance placed on sex considerations in clinical research, existing journal reporting requirements place little emphasis on this aspect of study design for preclinical experiments.

There are several limitations to our study. The journals examined are widely considered leading general translational research journals in terms of impact metrics and reputation. It is possible that their initiatives are more likely to be accepted and adhered to by the research community. Furthermore, our quasi-experimental study design suggests but does not prove causal effects of their interventions. However, *Nature Medicine* and *Science TM* are among the few translational research journals that have endorsed the NIH guidelines on rigor and reproducibility[11] and communicated to the research community their initiatives with specific implementation dates,[18, 20] rendering them ideal case study candidates. In addition, we show a lack of meaningful improvement in the reporting or conduct of preclinical research practices in high-ranking journals that did not implement similar initiatives over the same period despite having similar impact factors to *Science TM*, while confirming previously suggested improvements in *Stroke* following the implementation of analogous initiatives.[12, 19] Furthermore, all initiatives addressed randomization, blinding, and sample size estimation, but minimally addressed considering sex as a biological variable, which corresponded to the patterns of improvements seen. Our analysis relied on published reports, therefore the risk of misclassification exists. For instance, the method of randomization used was rarely described, raising the possibility that certain articles that reported randomly assigning animals may have referred to haphazard assignment.

As gatekeepers of what is published, journal interventions can meaningfully influence the quality of preclinical research and are promising strategies to improve study transparency and design. However, although the studied initiatives were associated with substantial improvements in

reporting, improvements in the use of key SDEs were relatively modest, suggesting that additional strategies focused on improving preclinical study design are required to address widespread methodological shortcomings, including sex bias.

Acknowledgements: None

Sources of funding: FDR was supported by a Canadian Institutes of Health Research (CIHR) Frederick Banting & Charles Best Canada Graduate Scholarship, the University of Ottawa Heart Institute Foundation, the Ernest and Margaret Ford Research Endowed Fund, and the Cardiac Arrhythmia Network of Canada (CANet) as part of the Networks of Centres of Excellence. RGJ was supported by a CIHR Vanier Research Graduate Scholarship.

Disclosures: None.

References

1. Collins FS, Tabak LA. Policy: Nih plans to enhance reproducibility. *Nature*. 2014;505:612-613
2. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*. 2012;490:187-191
3. Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. Evidence for the efficacy of nxy-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke*. 2008;39:2824-2829

4. Macleod MR, O'Collins T, Howells DW, Donnan GA. Pooling of animal experimental data reveals influence of study design and publication bias. *Stroke*. 2004;35:1203-1208
5. Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*. 2014;383:166-175
6. Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. *PLoS Biol*. 2015;13:e1002165
7. van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, et al. Can animal models of disease reliably inform human studies? *PLoS Med*. 2010;7:e1000245
8. Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, Fry D, et al. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One*. 2009;4:e7824
9. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. *Nature*. 2012;483:531-533
10. Horn J, de Haan RJ, Vermeulen M, Luiten PG, Limburg M. Nimodipine in animal model experiments of focal cerebral ischemia: A systematic review. *Stroke*. 2001;32:2433-2438
11. National Institutes of Health. Principles and guidelines for reporting preclinical research.
12. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald ZD, Moreland R, et al. Methodological rigor in preclinical cardiovascular studies: Targets to enhance reproducibility and promote research translation. *Circ Res*. 2017;120:1916-1926
13. Clayton JA, Collins FS. Policy: Nih to balance sex in cell and animal studies. *Nature*. 2014;509:282-283

14. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald Z, Simard T, et al. Sex bias is increasingly prevalent in preclinical cardiovascular research: Implications for translational medicine and health equity for women: A systematic assessment of leading cardiovascular journals over a 10-year period. *Circulation*. 2017;135:625-626
15. Begley CG, Ioannidis JP. Reproducibility in science: Improving the standard for basic and preclinical research. *Circ Res*. 2015;116:116-126
16. Provencher S, Archer SL, Ramirez FD, Hibbert B, Paulin R, Boucherat O, et al. Standards and methodological rigor in pulmonary arterial hypertension preclinical and translational research. *Circ Res*. 2018;122:1021-1032
17. Bolli R. New initiatives to improve the rigor and reproducibility of articles published in circulation research. *Circ Res*. 2017;121:472-479
18. Reducing our irreproducibility. *Nature*. 2013;496:398
19. Minnerup J, Zentsch V, Schmidt A, Fisher M, Schabitz WR. Methodological quality of experimental stroke studies published in the stroke journal: Time trends and effect of the basic science checklist. *Stroke*. 2016;47:267-272
20. Kelner KL. Playing our part. *Sci Transl Med*. 2013;5:190ed197
21. The NPQIP Collaborative group. Did a change in nature journals' editorial policy for life sciences research improve reporting? *BMJ Open Science*. 2019;3:e000035
22. Hair K, Macleod M, Sena E, Collaboration I. A randomised controlled trial of an intervention to improve compliance with the arrive guidelines (iicarus). *BioRxiv*. 2018
23. Baker D, Lidster K, Sottomayor A, Amor S. Two years later: Journals are not yet enforcing the arrive guidelines on reporting standards for pre-clinical animal studies. *PLoS Biol*. 2014;12:e1001756

24. Stroke Therapy Academic Industry Roundtable. Recommendations for standards regarding preclinical neuroprotective and restorative drug development. *Stroke*. 1999;30:2752-2758
25. Unger EF. All is not well in the world of translational research. *J Am Coll Cardiol*. 2007;50:738-740
26. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biol*. 2014;12:e1001863
27. Ramirez FD, Hibbert B. Letter by ramirez and hibbert regarding article, "consideration of sex differences in design and reporting of experimental arterial pathology studies: A statement from the arteriosclerosis, thrombosis, and vascular biology council". *Arterioscler Thromb Vasc Biol*. 2018;38:e99-e100
28. Checklists work to improve science. *Nature*. 2018;556:273-274
29. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: The arrive guidelines for reporting animal research. *PLoS Biol*. 2010;8:e1000412
30. Vahidy F, Schabitz WR, Fisher M, Aronowski J. Reporting standards for preclinical studies of stroke therapy. *Stroke*. 2016;47:2435-2438

Table 1. Comparison of study design element reporting and use in preclinical studies before and after the implementation of relevant initiatives in *Stroke*, *Nature Medicine*, *Science Translational Medicine*, and control journals

	Before *n* (%)	After *n* (%)	Crude RR (95% CI)	*P*	Adjusted RR (95% CI)*	*P**
Stroke	*N*=715	*N*=269				
Study design element reporting						
Randomization	233 (32.6)	183 (68.0)	2.1 (1.8-2.4)	<0.0001	2.2 (2.0-2.6)	<0.0001
Blinding	338 (47.3)	218 (81.0)	1.7 (1.6-1.9)	<0.0001	1.7 (1.6-1.9)	<0.0001
Sample size estimation	40 (5.6)	70 (26.0)	4.7 (3.2-6.7)	<0.0001	5.5 (3.8-8.1)	<0.0001
Sex of animals used	601 (84.1)	246 (91.5)	1.1 (1.0-1.1)	<0.001	1.1 (1.0-1.1)	0.008
Study design element use						
Randomization	233 (32.6)	183 (68.0)	2.1 (1.8-2.4)	<0.0001	2.2 (2.0-2.6)	<0.0001
Blinding	336 (47.0)	216 (80.3)	1.7 (1.5-1.9)	<0.0001	1.7 (1.6-1.9)	<0.0001
Sample size estimation	14 (2.0)	55 (20.5)	10.4 (5.9-18.5)	<0.0001	14.1 (7.8-25.5)	<0.0001
Inclusion of both sexes	50 (7.0)	26 (9.7)	1.1 (1.0-1.2)	0.044	1.1 (1.0-1.2)	0.090
Nature Medicine	*N*=476	*N*=473				
Study design element reporting						
Randomization	50 (10.5)	356 (75.3)	7.2 (5.5-9.4)	<0.0001	7.1 (5.4-9.3)	<0.0001
Blinding	112 (23.5)	390 (82.5)	3.5 (3.0-4.1)	<0.0001	3.5 (3.0-4.1)	<0.0001
Sample size estimation	6 (1.3)	334 (70.6)	56.0 (25.2-124.3)	<0.0001	NR	
Sex of animals used	311 (68.5)†	414 (91.2)†	1.3 (1.2-1.4)	<0.0001	1.3 (1.2-1.4)	<0.0001
Study design element use						
Randomization	50 (10.5)	245 (51.8)	4.9 (3.7-6.5)	<0.0001	4.8 (3.7-6.4)	<0.0001
Blinding	112 (23.5)	226 (47.8)	2.0 (1.7-2.4)	<0.0001	2.0 (1.7-2.4)	<0.0001
Sample size estimation	1 (0.2)	145 (30.7)	145.9 (20.5-1038.6)	<0.0001	NR	
Inclusion of both sexes	98 (21.6)†	186 (41.0)†	1.5 (1.4-1.6)	<0.0001	1.5 (1.3-1.6)	<0.0001
Science Translational Medicine	*N*=290	*N*=574				

Study design element reporting						
Randomization	54 (18.6)	375 (65.3)	3.5 (2.7-4.5)	<0.0001	3.5 (2.7-4.4)	<0.0001
Blinding	88 (30.3)	404 (70.4)	2.3 (1.9-2.8)	<0.0001	2.3 (2.0-2.8)	<0.0001
Sample size estimation	14 (4.8)	274 (47.7)	9.9 (5.9-16.6)	<0.0001	NR	
Sex of animals used	175 (64.8)‡	370 (69.0)‡	1.1 (1.0-1.2)	0.238	1.1 (1.0-1.2)	0.284
Study design element use						
Randomization	53 (18.3)	335 (58.4)	3.2 (2.5-4.1)	<0.0001	3.2 (2.5-4.1)	<0.0001
Blinding	85 (29.3)	280 (48.8)	1.7 (1.4-2.0)	<0.0001	1.7 (1.4-2.1)	<0.0001
Sample size estimation	4 (1.4)	146 (25.4)	18.4 (6.9-49.3)	<0.0001	NR	
Inclusion of both sexes	47 (17.4)‡	121 (22.6)‡	1.2 (1.0-1.3)	0.022	1.1 (1.0-1.3)	0.035
Control journals	*N*=889§	*N*=476§				
Study design element reporting						
Randomization	169 (19.0)	93 (19.5)	1.0 (0.8-1.3)	0.813	1.1 (0.8-1.3)	0.614
Blinding	308 (34.7)	161 (33.8)	1.0 (0.8-1.1)	0.761	1.0 (0.8-1.1)	0.839
Sample size estimation	31 (3.5)	42 (8.8)	2.5 (1.6-4.0)	<0.0001	2.7 (1.7-4.1)	<0.0001
Sex of animals used	626 (70.9)‖	361 (76.3)‖	1.1 (1.0-1.1)	0.028	1.1 (1.0-1.1)	0.041
Study design element use						
Randomization	164 (18.5)	91 (19.1)	1.0 (0.8-1.3)	0.762	1.1 (0.9-1.3)	0.575
Blinding	307 (34.5)	159 (33.4)	1.0 (0.8-1.1)	0.676	1.0 (0.8-1.1)	0.765
Sample size estimation	8 (0.9)	20 (4.2)	4.7 (2.1-10.5)	<0.001	NR	
Inclusion of both sexes	143 (16.2)‖	76 (16.1)‖	1.0 (0.8-1.3)	0.952	0.9 (0.7-1.2)	0.613

NR: not reported due to small number of events per predictor variable; RR: relative risk

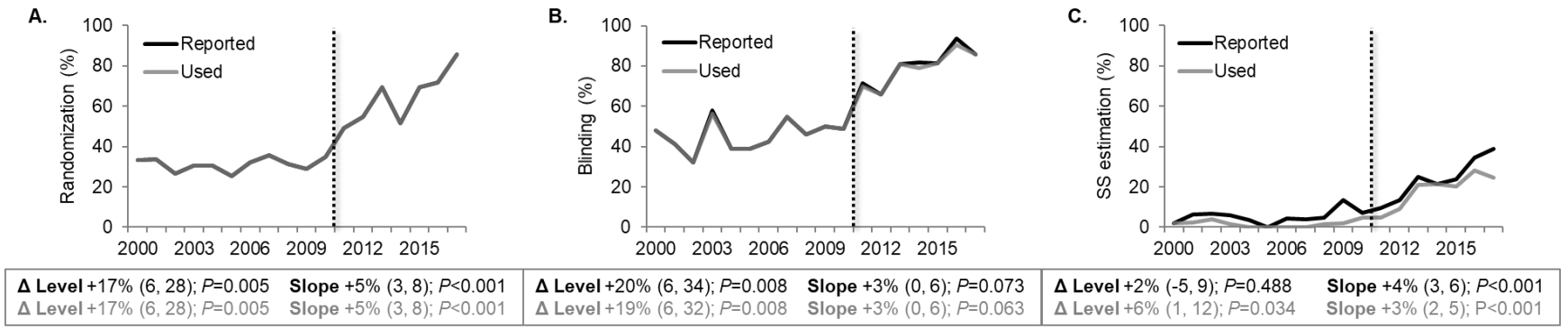
* Adjusted for most relevant disease system studied and animal model used

† *N*=454 and 454 before and after, respectively (excluding studies on sex-specific diseases)

‡ *N*=270 and 536 before and after, respectively (excluding studies on sex-specific diseases)

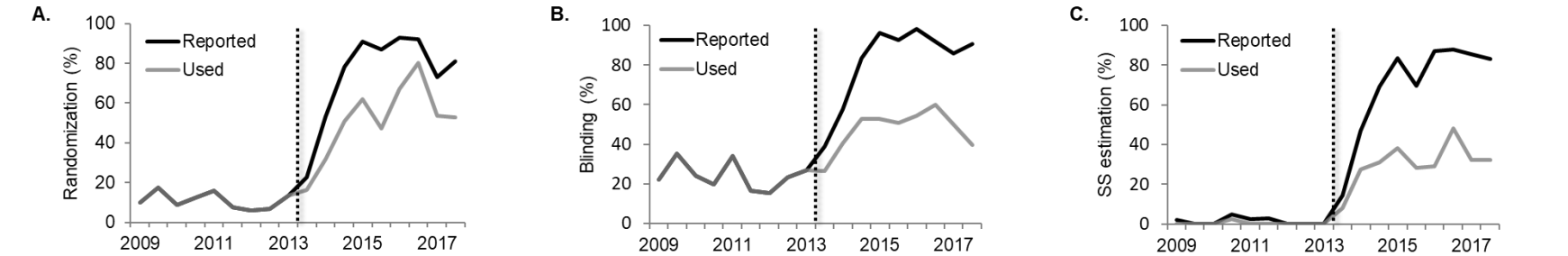
§ Before and after the introduction *Nature Medicine*/*Science TM*'s initiatives. All effect estimates were similar when analyzed as before vs. after the introduction of *Stroke*'s initiative (data not shown)

‖ *N*=883 and 473 before and after, respectively (excluding studies on sex-specific diseases)



Δ Level +17% (6, 28); *P*=0.005 **Slope** +5% (3, 8); *P*<0.001	**Δ Level** +20% (6, 34); *P*=0.008 **Slope** +3% (0, 6); *P*=0.073	**Δ Level** +2% (-5, 9); *P*=0.488 **Slope** +4% (3, 6); *P*<0.001
Δ Level +17% (6, 28); *P*=0.005 **Slope** +5% (3, 8); *P*<0.001	**Δ Level** +19% (6, 32); *P*=0.008 **Slope** +3% (0, 6); *P*=0.063	**Δ Level** +6% (1, 12); *P*=0.034 **Slope** +3% (2, 5); P<0.001

Figure 1. Reporting and use of randomization (*A*), blinding (*B*), and sample size (SS) estimation (*C*) in *Stroke*. Vertical interrupted line denotes the journal intervention. Δ Level/Slope: change in level/slope.



Δ Level +44% (16, 72); *P*=0.004 **Slope** +6% (0, 11); *P*=0.035	**Δ Level** +40% (16, 64); *P*=0.004 **Slope** +6% (1, 11); *P*=0.019	**Δ Level** +40% (17, 63); *P*=0.002 **Slope** +7% (3, 12); *P*=0.003
Δ Level +25% (2, 47); *P*=0.034 **Slope** +5% (1, 9); *P*=0.024	**Δ Level** +20% (2, 38); *P*=0.030 **Slope** +2% (-1, 6); *P*=0.181	**Δ Level** +22% (8, 35); *P*=0.004 **Slope** +2% (0, 5); *P*=0.075

Figure 2. Reporting and use of randomization (*A*), blinding (*B*), and sample size (SS) estimation (*C*) in *Nature Medicine*. Vertical interrupted line denotes the journal intervention. Δ Level/Slope: change in level/slope.

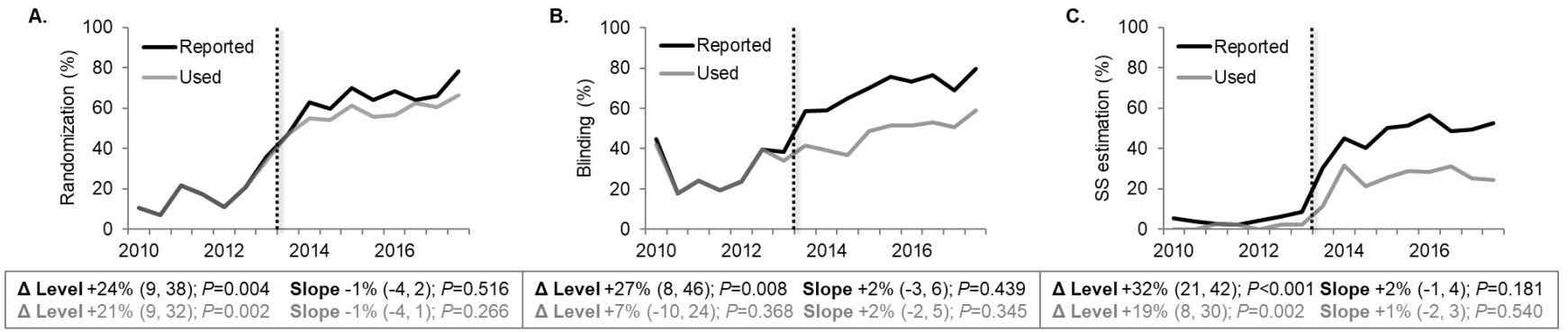


Figure 3. Reporting and use of randomization (*A*), blinding (*B*), and sample size (SS) estimation (*C*) in *Science Translational Medicine*. Vertical interrupted line denotes the journal intervention. Δ Level/Slope: change in level/slope.

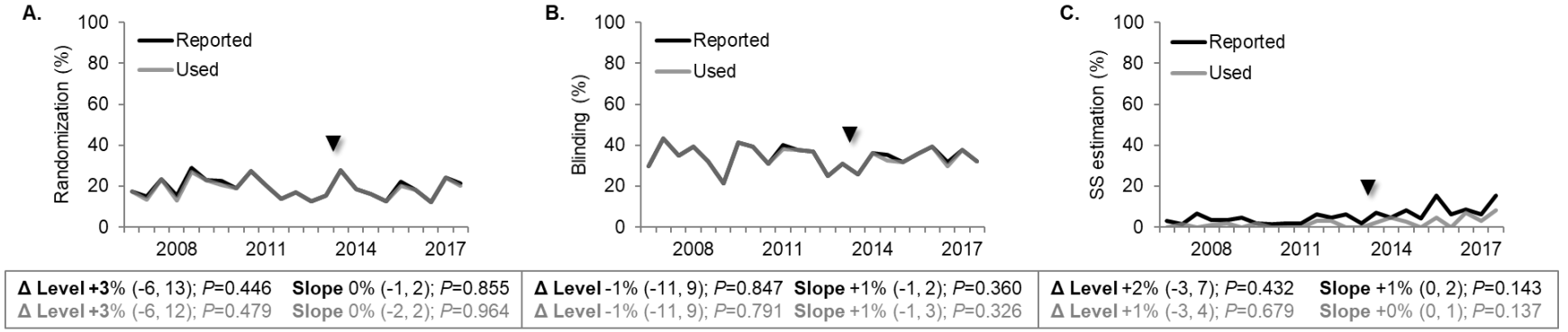


Figure 4. Reporting and use of randomization (*A*), blinding (*B*), and sample size (SS) estimation (*C*) in control journals. Analyzed using *Nature Medicine* and *Science TM*'s initiatives as the time series interruption (arrowhead) (similar results were obtained using *Stroke*'s initiative as interruption). Δ Level/Slope: change in level/slope.

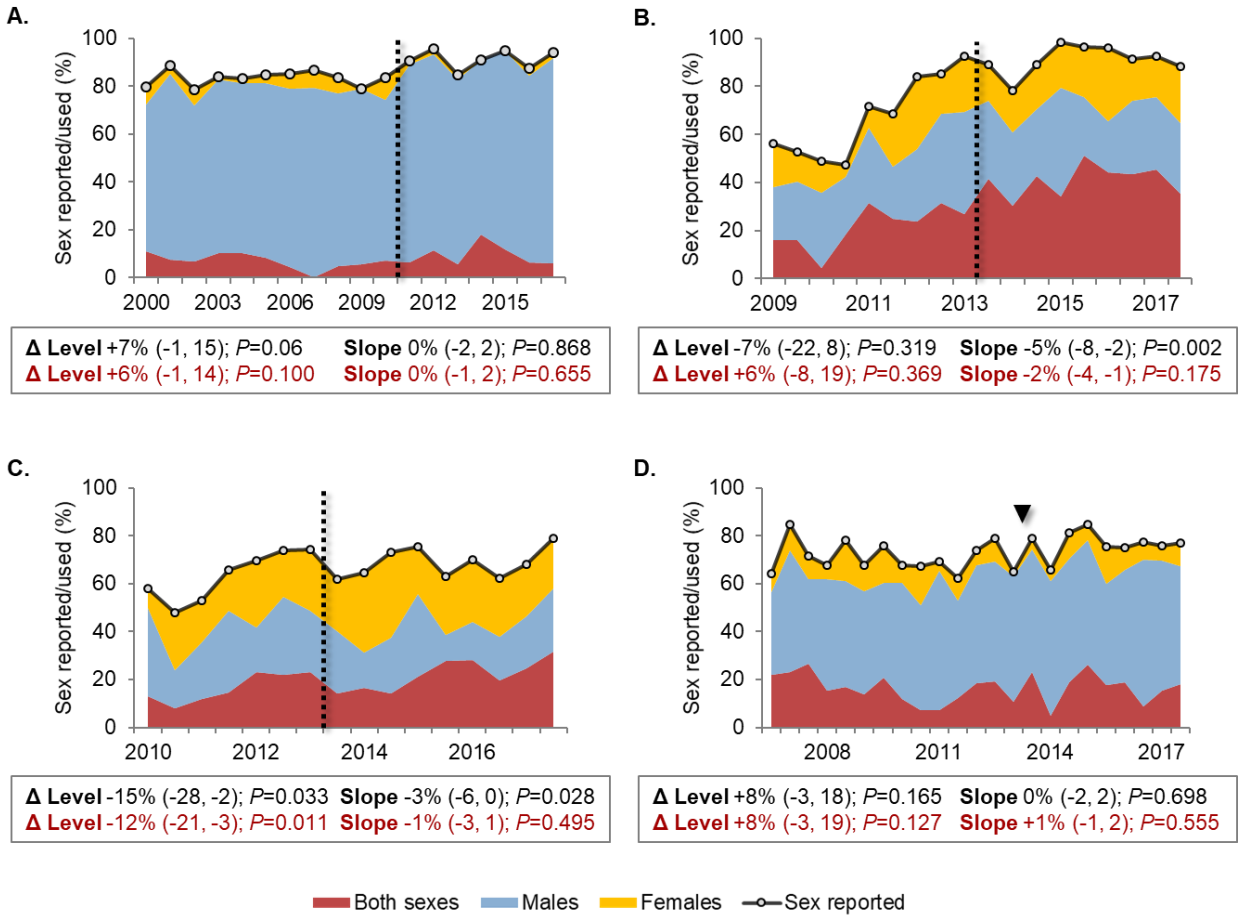


Figure 5. Sex reporting and sex of animals used in *Stroke* (*A*), *Nature Medicine* (*B*), *Science Translational Medicine* (*Science TM*) (*C*), and control journals (*D*). Vertical interrupted lines denote journal interventions. Results shown for sex reporting (black) and use of both sexes (red). Control journals were analyzed using *Nature Medicine* and *Science TM*'s initiatives as the time series interruption (arrowhead). Δ Level/Slope: change in level/slope.

SUPPLEMENTAL MATERIAL

Journal initiatives to enhance preclinical research: analyses of *Stroke*, *Nature Medicine*, *Science Translational Medicine*

F. Daniel Ramirez MD, Richard G. Jung BSc Pouya Motazedian MD, Dylan Perry-Nguyen BMSc, Pietro Di Santo MD, Zachary MacDonald MD, Aisling A. Clancy MD MSc, Alisha Labinaz, Steven Promislow MD, Trevor Simard MD, Steeve Provencher MD MSc, Sébastien Bonnet PhD, Ian D. Graham PhD, George A. Wells PhD, Benjamin Hibbert MD PhD

Supplemental Methods

Study selection process from Nature Medicine and Science Translational Medicine

All original research articles and letters reporting on experiments using non-human mammals in *Nature Medicine* and *Science Translational Medicine* (*Science TM*) were included for analysis, irrespective of stated or implied relevance to a human disease by the articles' authors. The reasoning for this selection strategy was that these journals explicitly define their scope or mission as focused on translational research with clinical implications (see https://www.nature.com/nm/about/aims and http://stm.sciencemag.org/site/misc/about.xhtml), therefore study relevance to human disease and clinical medicine was judged to have been decided by the respective journal editorial board. Given the substantial interrater agreement for study inclusion observed for *Stroke* articles and the simpler criterion used for article inclusion from *Nature Medicine* and *Science TM*, article selection for these journals was performed by a single reviewer. The appropriateness of article inclusion was then confirmed by a second independent reviewer.

Journal-specific diseases and disease systems of interest

Diseases of interest in Stroke and in control journals were cardiovascular in nature, including atherosclerosis and/or vascular homeostasis, arterial aneurysms and/or dissections, myocardial infarction, valvular disease, cardiomyopathy and/or heart failure, cardiac transplantation, pulmonary hypertension, cardiac arrhythmia, stroke, resuscitative medicine, hypertension, metabolic or endocrine diseases, and hematological disorders (including thrombosis). Studies

deemed to report on clinically relevant conditions but not falling into one of the pre-specified categories could be included at the journal reviewers' discretion (category 'other').

The most relevant disease systems studied or fields of medicine in these journals were categorized as cardiovascular (including stroke), respiratory, gastroenterology, nephrology, neurology (excluding stroke), allergy/immunology (including rheumatology), endocrinology, hematology (excluding hematological malignancies), infectious diseases, oncology (including hematological malignancies), obstetrics/gynecology, ophthalmology, surgery, dermatology, psychiatry, anesthesia, or toxicology.

Sample size calculations

Given the multiple factors influencing statistical power in interrupted time series analyses,[1] sample size calculations were based on comparisons of the prevalence of SDE use before versus after the introduction of journal initiatives. Based on an article inclusion rate of 50 per year from *Stroke* between July 2006 and June 2016 in a previous study[2] and preliminary estimates of 100 articles per year from *Nature Medicine* and *Science TM*, we estimated that 18 years of data from Stroke and 9 years from *Nature Medicine* and *Science TM* would afford 80% power to detect differences of ≤10.0% in the use of every SDE in every journal. The baseline incidences of the use of specific SDEs in *Nature Medicine* and *Science TM* (before their respective interventions were introduced) were assumed to be similar to those of *Circulation*.[2]

Supplemental Table

Table I. Study characteristics

	Stroke (*n*=984)	*Nature Medicine* (*n*=949)	*Science TM* (*n*=864)	Control journals (*n*=1,365)	Total (*n*=4,162)
Most relevant disease system					
Cardiovascular	972 (98.8)	81 (8.5)	79 (9.1)	1,270 (93.0)	2,402 (57.7)
Respiratory	0 (0)	28 (3.0)	35 (4.1)	0 (0)	63 (1.5)
Gastroenterology	0 (0)	20 (2.1)	25 (2.9)	0 (0)	45 (1.1)
Nephrology	0 (0)	30 (3.2)	9 (1.0)	2 (0.2)	41 (1.0)
Neurology	3 (0.3)	109 (11.5)	115 (13.3)	2 (0.2)	229 (5.4)
Allergy/immunology	0 (0)	72 (7.6)	55 (6.4)	0 (0)	127 (3.1)
Endocrinology	6 (0.6)	144 (15.2)	35 (4.1)	34 (2.5)	219 (5.3)
Hematology	2 (0.2)	38 (4.0)	15 (1.7)	34 (2.5)	89 (2.1)
Infectious diseases	0 (0)	103 (10.9)	131 (15.2)	0 (0)	234 (5.6)
Oncology	0 (0)	221 (23.3)	228 (26.4)	0 (0)	449 (10.8)
Dermatology	0 (0)	11 (1.2)	20 (2.3)	0 (0)	31 (0.7)
Surgery	0 (0)	11 (1.2)	19 (2.2)	0 (0)	35 (0.8)
Other	1 (0.1)	81 (8.5)	98 (11.3)	23 (1.7)	203 (4.9)
Animal model used					

Mouse	350 (35.6)	812 (85.6)	616 (71.3)	1,024 (75.0)	2,802 (67.3)
Rat	511 (51.9)	36 (3.8)	42 (4.9)	143 (10.5)	732 (17.6)
Pig	17 (1.7)	0 (0)	20 (2.3)	35 (2.6)	72 (1.7)
Rabbit	42 (4.3)	2 (0.2)	11 (1.3)	25 (1.8)	80 (1.9)
Dog	20 (2.0)	0 (0)	5 (0.6)	49 (3.6)	74 (1.8)
Primate	12 (1.2)	25 (2.6)	45 (5.2)	4 (0.3)	86 (2.1)
Other	20 (2.0)	1 (0.1)	13 (1.5)	16 (1.2)	50 (1.2)
Combination	12 (1.2)	73 (7.7)	112 (13.0)	69 (5.1)	266 (6.4)

Variables reported as number (%). *Science TM*: *Science Translational Medicine*.

Supplemental Figures and Figure Legends

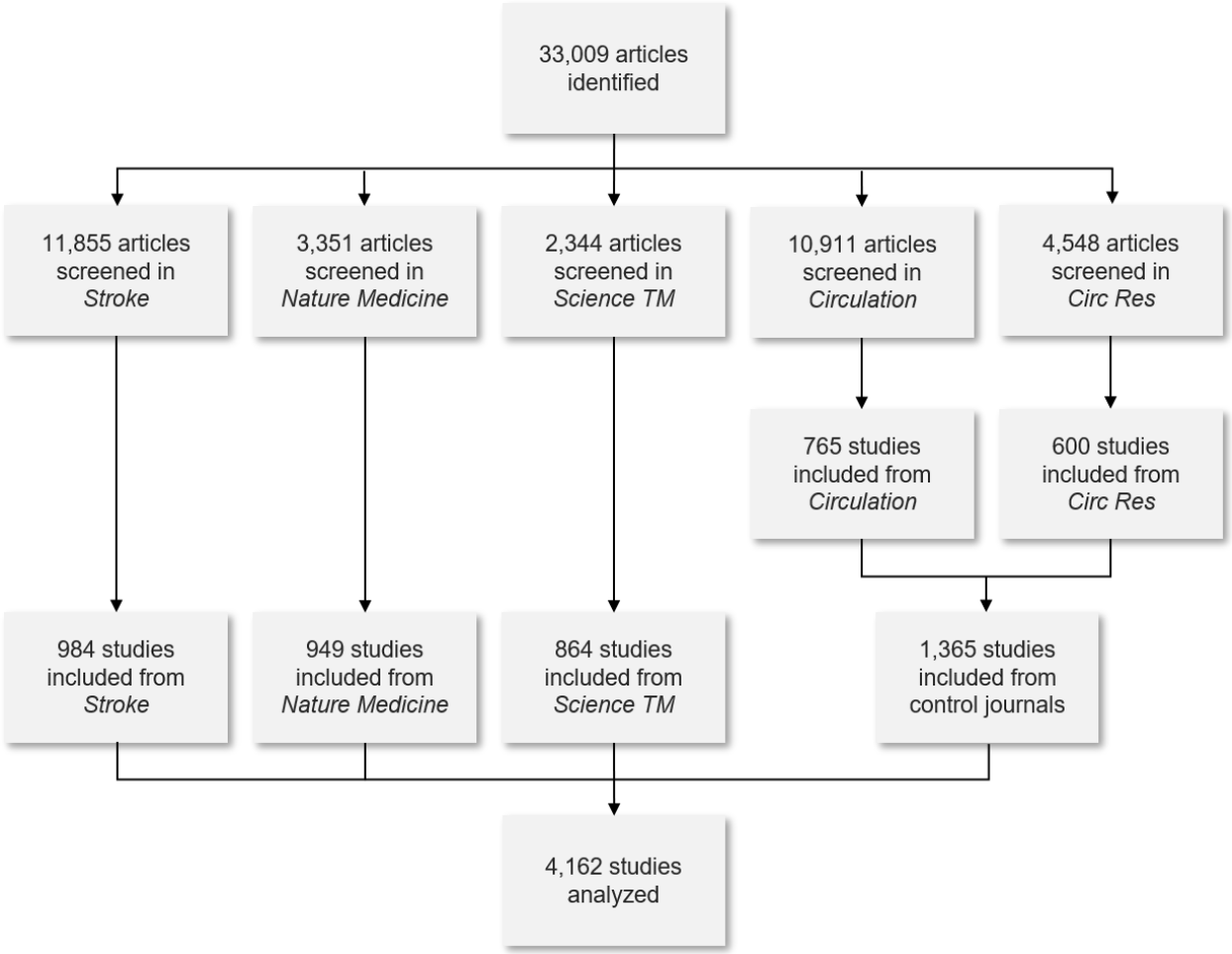


Figure I. Literature search and results. *Circ Res*: *Circulation Research*; *Science TM*: *Science Translational Medicine*.

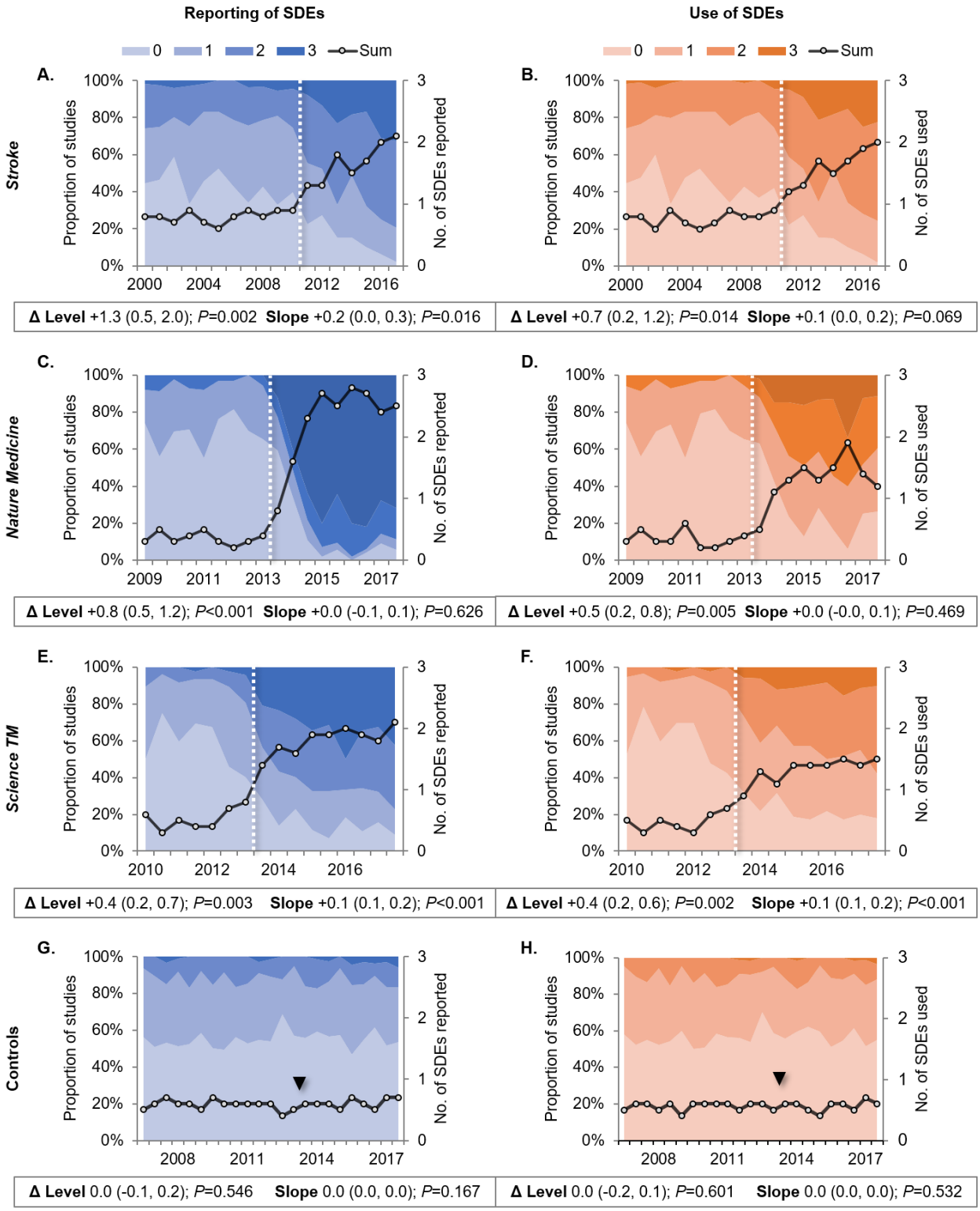


Figure II. Reporting and use of multiple study design elements (SDEs) in published preclinical studies. The proportions of studies reporting on zero to three SDEs in *Stroke* (*A*), *Nature*

Medicine (*C*), *Science Translational Medicine* (*Science TM*) (*E*), and control journals (*G*) over time are shown in shades of blue (randomization, blinding, and sample size estimation each contribute 1 point). The mean number of SDEs reported on over time is plotted as a black line. The proportion of studies using zero to three SDEs in *Stroke* (*B*), *Nature Medicine* (*D*), *Science TM* (*F*), and control journals (*H*) are shown in shades of orange (randomization, blinding, and sample size estimation each contribute 1 point). The mean number of SDEs used over time is plotted as a black line. Vertical interrupted lines denote interventions at each journal. Control journals were analyzed using *Nature Medicine* and *Science TM*'s initiatives as the time series interruption (arrowhead) (similar results were obtained using *Stroke*'s initiative as interruption). Δ Level/Slope: change in trend line level/slope.

References

1. Zhang F, Wagner AK, Ross-Degnan D. Simulation-based power calculation for designing interrupted time series analyses of health policy interventions. J Clin Epidemiol. 2011;64:1252-1261

2. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald ZD, Moreland R, et al. Methodological rigor in preclinical cardiovascular studies: Targets to enhance reproducibility and promote research translation. Circ Res. 2017;120:1916-1926

CHAPTER 7: DISCUSSION AND RECOMMENDATIONS

7.1 Summary of thesis objectives

The purpose of this thesis was to describe the reporting and methodological quality of preclinical research with a focus on cardiovascular science and to examine potential strategies to improve the status quo. The purpose of this final chapter is to tie together the extent and potential consequences of methodological shortcomings in preclinical cardiovascular research, including sex bias (Chapters 3 and 4); the potential influence of research team composition on measures of methodological rigor, specifically the sex of researchers (Chapter 5); and the role journals can play in improving the quality of published preclinical research (Chapter 6). It also briefly discusses temporal patterns in women's involvement in preclinical cardiovascular research and the possibility that limited mentoring opportunities may be hindering women's advancement in this field (Chapter 5).

7.2 Extent of the problem and clue to journals' influence

Preclinical studies using animals often precede and inform clinical trials yet standards in reporting and study design differ between these stages of research.[1, 2] In Chapter 3, we show that key study design elements (SDEs) are routinely ignored in preclinical cardiovascular research and that this practice does not appear to have improved in recent years, except possibly in stroke research.[3] On closer examination of stroke research included in our study, we noted that 87% had been published in *Stroke*, raising the question of whether stroke as a research topic and/or *Stroke* as a journal drove this improvement (or neither). *Post hoc* analyses suggested that both stroke as

a study subject and *Stroke* as a journal independently predicted more robust reported study design. Unbeknownst to our research group at the outset of this first study, *Stroke* had introduced a checklist in 2011 that asked authors to disclose important SDEs at the time of manuscript submission.[4] This initiative was intended to improve the rigor of preclinical stroke research and a limited analysis had previously suggested that it had had a modest effect.[5] Our exploratory analysis informed our subsequent hypotheses that journals act as 'gatekeepers' of the published literature and therefore can meaningfully influence the quality of published preclinical research, which we more robustly tested in Chapter 6.

7.3 Sex bias in preclinical cardiovascular research

Despite remarkable progress in preventive and interventional therapies, cardiovascular disease (CVD) remains a leading cause of morbidity and mortality in North America and abroad.[6-8] Age-adjusted mortality rates from CVD have consistently been higher in men yet the absolute number of cardiovascular deaths in the United States has been higher in women for 19 of the last 20 years studied,[9] unequivocally underscoring the impact of CVD irrespective of sex.

As previously highlighted, sex and gender bias in research was specifically identified as a barrier to advancing women's health since at least the 1980's, with the Report of the Public Health Service Task Force on Women's Health Issues underscoring that data obtained from studies using only male subjects may not be applicable to the female population.[10] In the early 1990's, the Revitalization Act mandated that women be included in NIH-funded clinical trials with accompanying guidelines cautioning that this effort could be undermined if the influence of sex and/or gender was ignored in earlier stages of research. This warning has gone largely ignored.

Although considerable progress with regards to sex inclusion in clinical research has been made,[11-13] relatively little attention has been paid to this issue in preclinical experiments[14-22] and a reluctance to study female animals has not only persisted but increased within the preclinical cardiovascular research community, as we show in Chapter 4.[23] When data from male-predominant preclinical experiments are used to inform clinical trials, a bias toward male-specific physiology and therapies can be perpetuated.[24]

7.4 Female authorship, research quality, and mentorship in preclinical cardiovascular science

The notion that researcher sex is associated with different viewpoints and research questions and that these differences can influence research quality underlies many of the arguments put forth to promote sex diversity in research teams.[25, 26] In Chapter 5 we examine the extent of women's underrepresentation in preclinical cardiovascular research and explore the possibility that it not only has societal implications, but also scientific ramifications. We demonstrate that female researchers continue to be underrepresented in preclinical cardiovascular science, especially as senior authors. We also present empirical data that suggest that despite being similar in most measures of scientific rigor and impact, female preclinical researchers are more attuned to the importance of considering sex as a biological variable than their male counterparts – a practice that is expected to promote advances in women's health. This finding persisted when women's involvement in studies was examined as first author, senior author, as a dichotomized proportion of authors (≥33%), and as an incremental proportion of authors (per 10%), as well as after adjusting for potential confounders.

Although female authorship is increasing in preclinical cardiovascular research, our secondary analysis suggests that mentorship relationships continue to be divided by sex: men seem to preferentially mentor men and women seem to preferentially mentor women. Given the relative paucity of senior female scientists in this field, this segregation may be problematic for women given the importance of mentorship for career advancement in academia.

7.5 Empirical evidence of the influence of journals on preclinical research quality

To test the hypothesis that journal initiatives such as those recently introduced by *Circulation Research* to improve the reporting and methodological quality of preclinical research, we undertook a more comprehensive analysis of three journals with similar editorial interventions (*Nature Medicine*, *Science Translational Medicine*, and *Stroke*), using two influential journals as controls (journals that had not implemented analogous initiatives over the study period). This analysis used a quasi-experimental design to account for secular trends in reporting and study design in each journal. We conclude that journal initiatives are indeed promising strategies to improve the quality of published preclinical research and that journals and editors are key stakeholders to engage in initiatives seeking to change research practices in this field.

7.6 Recommendations

1. Recognition of widespread methodological shortcomings in preclinical research is required to understand the scope of the problem, its potential implications, and to gauge progress.

2. Preclinical researchers should consider including animals of both sexes in experiments unless there are valid scientific reasons not to do so. Before attempts to 'translate' research to the clinical realm, sufficient female animals should be studied to detect whether relevant and meaningful sex differences exist. Not doing so has the potential to undermine progress in achieving adequate representation of women in clinical research.
3. Efforts to promote women's involvement in preclinical cardiovascular research should consider strategies to facilitate mentorship opportunities for female trainees and early career researchers.
4. Journals have an important role to play in improving the rigour and reproducibility of preclinical research. These influential stakeholders should be targeted in relevant initiatives; however, journals alone are unlikely to fully change research practices.

References

1. Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet. 2014;383(9912):166-75.
2. Collins FS, Tabak LA. Policy: Nih plans to enhance reproducibility. Nature. 2014;505(7485):612-3.
3. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald ZD, Moreland R, et al. Methodological rigor in preclinical cardiovascular studies: Targets to enhance reproducibility and promote research translation. Circ Res. 2017;120(12):1916-26.
4. Vahidy F, Schabitz WR, Fisher M, Aronowski J. Reporting standards for preclinical studies of stroke therapy. Stroke. 2016;47(10):2435-8.

5. Minnerup J, Zentsch V, Schmidt A, Fisher M, Schabitz WR. Methodological quality of experimental stroke studies published in the stroke journal: Time trends and effect of the basic science checklist. Stroke. 2016;47(1):267-72.
6. GBD Mortality and Causes of Death Collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: A systematic analysis for the global burden of disease study 2013. Lancet. 2015;385(9963):117-71.
7. GBD DALYs and Hale Collaborators. Global, regional, and national disability-adjusted life years (dalys) for 306 diseases and injuries and healthy life expectancy (hale) for 188 countries, 1990-2013: Quantifying the epidemiological transition. Lancet. 2015;386(10009):2145-91.
8. Kochanek KD, Murphy SL, Xu JQ, Tejeda-Vera B. Deaths: Final data for 2014. National vital statistics reports. Hyattsville, MD2016.
9. Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, et al. Heart disease and stroke statistics-2016 update: A report from the american heart association. Circulation. 2016;133(4):e38-360.
10. Women's health. Report of the public health service task force on women's health issues. Public Health Rep. 1985;100(1):73-106.
11. United States General Accounting Office. Women's health. Nih has increased its efforts to include women in research. Washington, DC.2000.
12. United States General Accounting Office. Women's health. Women sufficiently represented in new drug testing, but fda oversight needs improvement. Washington, DC2001.

13. United States Government Accountability Office. National institutes of health. Better oversight needed to help ensure continued progress including women in health research. Washington, DC2015.
14. Clayton JA, Collins FS. Policy: Nih to balance sex in cell and animal studies. Nature. 2014;509(7500):282-3.
15. Putting gender on the agenda. Nature. 2010;465(7299):665.
16. Check Hayden E. Sex bias blights drug studies. Nature. 2010;464(7287):332-3.
17. Wald C, Wu C. Biomedical research. Of mice and women: The bias in animal models. Science. 2010;327(5973):1571-2.
18. Zucker I, Beery AK. Males still dominate animal studies. Nature. 2010;465(7299):690.
19. Miller VM. In pursuit of scientific excellence: Sex matters. Am J Physiol Endocrinol Metab. 2012;302(9):E1025-6.
20. Shah K, McCormack CE, Bradbury NA. Do you know the sex of your cells? Am J Physiol Cell Physiol. 2014;306(1):C3-18.
21. Yoon DY, Mansukhani NA, Stubbs VC, Helenowski IB, Woodruff TK, Kibbe MR. Sex bias exists in basic science and translational surgical research. Surgery. 2014;156(3):508-16.
22. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: The arrive guidelines for reporting animal research. PLoS Biol. 2010;8(6):e1000412.
23. Mogil JS. Perspective: Equality need not be painful. Nature. 2016;535(7611):S7.
24. Klein SL, Schiebinger L, Stefanick ML, Cahill L, Danska J, de Vries GJ, et al. Opinion: Sex inclusion in basic research drives discovery. Proc Natl Acad Sci U S A. 2015;112(17):5257-8.

25. Nielsen MW, Alegria S, Borjeson L, Etzkowitz H, Falk-Krzesinski HJ, Joshi A, et al. Opinion: Gender diversity leads to better science. Proc Natl Acad Sci U S A. 2017;114(8):1740-2.
26. Valantine HA, Collins FS. National institutes of health addresses the science of diversity. Proc Natl Acad Sci U S A. 2015;112(40):12240-2.

Integrative Physiology

OPEN

Methodological Rigor in Preclinical Cardiovascular Studies

Targets to Enhance Reproducibility and Promote Research Translation

F. Daniel Ramirez, Pouya Motazedian, Richard G. Jung, Pietro Di Santo, Zachary D. MacDonald, Robert Moreland, Trevor Simard, Aisling A. Clancy, Juan J. Russo, Vivian A. Welch, George A. Wells, Benjamin Hibbert

Rationale: **Methodological sources of bias and suboptimal reporting contribute to irreproducibility in preclinical science and may negatively affect research translation. Randomization, blinding, sample size estimation, and considering sex as a biological variable are deemed crucial study design elements to maximize the quality and predictive value of preclinical experiments.**

Objective: **To examine the prevalence and temporal patterns of recommended study design element implementation in preclinical cardiovascular research.**

Methods and Results: **All articles published over a 10-year period in 5 leading cardiovascular journals were reviewed. Reports of in vivo experiments in nonhuman mammals describing pathophysiology, genetics, or therapeutic interventions relevant to specific cardiovascular disorders were identified. Data on study design and animal model use were collected. Citations at 60 months were additionally examined as a surrogate measure of research impact in a prespecified subset of studies, stratified by individual and cumulative study design elements. Of 28 636 articles screened, 3396 met inclusion criteria. Randomization was reported in 21.8%, blinding in 32.7%, and sample size estimation in 2.3%. Temporal and disease-specific analyses show that the implementation of these study design elements has overall not appreciably increased over the past decade, except in preclinical stroke research, which has uniquely demonstrated significant improvements in methodological rigor. In a subset of 1681 preclinical studies, randomization, blinding, sample size estimation, and inclusion of both sexes were not associated with increased citations at 60 months.**

Conclusions: **Methodological shortcomings are prevalent in preclinical cardiovascular research, have not substantially improved over the past 10 years, and may be overlooked when basing subsequent studies. Resultant risks of bias and threats to study validity have the potential to hinder progress in cardiovascular medicine as preclinical research often precedes and informs clinical trials. Stroke research quality has uniquely improved in recent years, warranting a closer examination for interventions to model in other cardiovascular fields. (*Circ Res*. 2017;120:1916-1926. DOI: 10.1161/CIRCRESAHA.117.310628.)**

Key Words: animal models ■ bias ■ biomedical research ■ cardiovascular diseases ■ reproducibility of results

Preclinical research and clinical therapy development exist in a symbiotic continuum yet are guided by different reporting standards, regulatory forces, and reward systems.[1,2] Consequent discrepancies in study designs and analytical methods can compromise scientific validity,[3] result in irreproducible results (particularly from preclinical experiments using animal models[2]), and may contribute to the high rate of attrition seen in early stages of clinical development.[4,5] Improved methodological rigor and transparent reporting practices have therefore been advocated to improve the predictive value of animal model data.[2,6–10]

Editorial, see p 1852
In This Issue, see p 1843
Meet the First Author, see p 1844

The issue of irreproducibility in preclinical research and its impact are well-recognized in academic circles[2,11–14] and

Original received January 12, 2017; revision received March 11, 2017; accepted March 31, 2017. In February 2016, the average time from submission to first decision for all original research papers submitted to *Circulation Research* was 15.4 days.

From the Division of Cardiology (F.D.R., P.M., R.G.J., P.D.S., T.S., J.J.R., B.H.), CAPITAL Research Group (F.D.R., P.M., R.G.J., P.D.S., Z.D.M.D., R.M., T.S., J.J.R., B.H.), Vascular Biology and Experimental Medicine Laboratory (R.G.J., T.S., B.H.), and Cardiovascular Research Methods Centre (G.A.W.), University of Ottawa Heart Institute, Ontario, Canada; and School of Epidemiology, Public Health and Preventive Medicine (F.D.R., V.A.W., G.A.W.), Department of Cellular and Molecular Medicine (R.G.J., T.S., B.H.), Department of Radiology (R.M.), Department of Obstetrics and Gynecology (A.A.C.), Bruyère Research Institute (V.A.W.), and Centre for Global Health (V.A.W.), University of Ottawa, Ontario, Canada.

The online-only Data Supplement is available with this article at http://circres.ahajournals.org/lookup/suppl/doi:10.1161/CIRCRESAHA.117.310628/-/DC1.

Correspondence to Benjamin Hibbert, MD, PhD, University of Ottawa Heart Institute, 40 Ruskin St, H-4238, Ottawa, Ontario, Canada K1Y 4W7. E-mail bhibbert@ottawaheart.ca

Novelty and Significance

What Is Known?

- Preclinical research often precedes and informs clinical trials.
- Randomization, blinding, sample size estimation, and considering sex as a biological variable are considered crucial study design elements to maximize the predictive value of preclinical experiments.

What New Information Does This Article Contribute?

- Key study design elements are rarely implemented in preclinical cardiovascular research.
- The implementation of these elements has not appreciably improved over the past decade, except in stroke research.
- Methodological shortcomings in preclinical experiments may be overlooked when researchers design subsequent studies.

Preclinical and clinical research are guided by different methodological standards. Resultant discrepancies in study designs can compromise scientific validity, give rise to experimental irreproducibility, and may contribute to failures in research translation. Randomization, blinding, sample size estimation, and considering sex as a biological variable have been identified as key study design elements to promote rigor and reproducibility of animal experiments and research relevance to both men and women. However, there are limited data on the extent of suboptimal methodological rigor in preclinical research. We identified and analyzed 3396 preclinical studies in leading cardiovascular journals over a 10-year period, including citation counts at 60 months as a surrogate measure of research impact for papers published in the first 5 years. We show that these design elements were rarely implemented—a situation that has not substantially improved over the past decade, except in stroke research. The individual and cumulative implementation of these design elements was not associated with increased citations. Together, these data indicate that methodological shortcomings are prevalent in preclinical cardiovascular research and suggest that they may be overlooked when researchers design subsequent studies.

Nonstandard Abbreviations and Acronyms	
CVD	cardiovascular disorder
NIH	National Institutes of Health

increasingly so among the lay public[15–17] and the biopharmaceutical industry.[13,18] In 2011, Bayer HealthCare published an analysis of 67 in-house target identification and validation projects in the fields of oncology, women's health, and cardiovascular disease over a 4-year period, reporting that the published data were reproducible in less than one third of cases and that in nearly two thirds inconsistencies either prolonged the validation process or led to project termination.[11] In 2012, Amgen similarly published the results of their attempt to reproduce 53 "landmark" studies in the fields of hematology and oncology, succeeding in only 6 cases even after contacting the original authors for guidance, exchanging reagents, and in certain instances repeating experiments in the laboratory of the original investigators.[12] Some findings could not be reproduced even by the original investigators in their own laboratories when the experiments were repeated in a blinded fashion.[19] Disconcertingly, some of the irreproducible research were deemed to have prompted clinical studies. Others have likewise described candidate therapies that have advanced to testing in humans despite irreproducible or inconsistent results in animals.[8,20,21]

Since the mid-1990s, pharmaceutical research and development productivity has declined with decreasing numbers of medications approved, increasing attrition rates at all stages of research, and lengthening development times despite rising expenditures.[4] This decline has been attributed, in part, to flawed preclinical research methodology, highlighting the importance of preclinical research for advancing clinical care, but equally portending a waning confidence in animal model data to identify promising therapeutic targets.[5,7,11,12,22,23] Cardiovascular research and development, in particular, has fared poorly in recent years relative to other disease categories, exhibiting the greatest decline as a percentage of total projects[4] and among the lowest success rates and likelihood of approval at all phases of clinical trials.[24] Moreover, research and development investment patterns are increasingly deterring incremental innovation (for instance, improving upon available effective therapies). Instead, increasing focus is being placed on novel disease targets with higher revenue potential but at higher risk of failure (eg, specific cancers).[4] Poor reproducibility in preclinical research coupled with this increasingly unfavorable landscape for successful cardiovascular therapy development could, therefore, substantially hinder progress in cardiovascular care.

In response to the above systematic issues and in collaboration with editors from major journals, funding agencies, and scientific leaders, the National Institutes of Health (NIH) proposed a set of reporting guidelines and funding policies to improve the reproducibility and rigor of preclinical research.[25] A core set of standards first proposed by the NIH's National Institute of Neurological Disorders and Stroke were adopted, which identified randomization, blinding, sample size estimation, and data handling as minimum reporting requirements to promote transparency in animal studies.[7] In addition, the NIH announced that it would require that sex be considered as a biological variable in applications for preclinical research funding.[26] These principles and guidelines have been endorsed by prominent academic societies, associations, and journals, including all American Heart Association (AHA) journals,[25] with evidence of editorial commitment to complying.[27,28] However, there are limited data on the extent of suboptimal methodological rigor or incomplete reporting in preclinical cardiovascular science and therefore no baseline from which to gauge progress. A commitment to improving the quality and impact of preclinical research and to maintaining the trust of public and private stakeholders requires transparency on current and future states of scientific practice.[13] We, therefore, reviewed all preclinical cardiovascular studies published in leading AHA journals over the past decade to determine the prevalence of randomization, blinding, sample size estimation, and inclusion of both sexes and trends in these practices over time.

Methods

As previously described,[29] all preclinical cardiovascular studies published in AHA journals with archives spanning at least 10 years were reviewed. Five journals met these criteria: *Circulation*; *Circulation Research*; *Hypertension*; *Stroke*; and *Arteriosclerosis, Thrombosis, and Vascular Biology* (*ATVB*). All reports published during a 10-year period (July 2006 to June 2016) were screened. Studies were included if they represented original research published as full manuscripts; reported results of in vivo experiments in nonhuman mammals; and described pathophysiology, genetics, or therapeutic interventions that were stated to be directly relevant to a specific cardiovascular disorder (CVD) in humans (see prespecified list of CVDs below). Studies on physiological and genetic characteristics were included if potential therapeutic applications or implications of the study findings were proposed in the article. These criteria are consistent with previously proposed definitions of preclinical ("confirmatory" or "proof-of-concept") experiments.[7,30,31] CVDs of interest included atherosclerosis or vascular homeostasis, arterial aneurysms or dissections, myocardial infarction, valvular disease, cardiomyopathy or heart failure, cardiac transplantation, pulmonary hypertension, cardiac arrhythmia, stroke, resuscitative medicine, hypertension, metabolic or endocrine diseases, and hematologic disorders (including thrombosis). Studies deemed to report on clinically relevant cardiovascular conditions but not falling into one of the prespecified categories could be included at the journal reviewers' discretion (category "other"). Each journal article was independently reviewed and data extracted using standardized case report forms by 2 reviewers, allowing for the assessments of inter-rater agreement. To permit our team to review a large volume of articles, we allowed for studies to be excluded as soon as it was clear that a manuscript violated any of our inclusion criteria, which was most often because of the model used (eg, zebrafish or humans) or because they were published in formats other than full manuscripts (eg, conference abstracts). The specific inclusion criterion/criteria that was/were deemed to have been violated for each excluded article, therefore, varied according to journal reviewer and were often multiple (eg, a nonmammalian animal model was used and no specific reference to therapeutic implications/applications was made). Discrepancies were resolved by consensus or by an independent adjudicator before building a final locked database for analysis.

Pre-specified data, including the date of publication, CVD investigated, animal model(s) used and their sex, whether animals were randomized to treatment groups, whether any blinding was implemented (concealed allocation or blinded outcome assessment), and whether a priori sample size/power estimations were performed were collected. Subgroup analyses restricted to studies of therapeutic interventions and by CVD studied were performed as were post hoc comparisons of these practices before and after the publication of NIH guidelines and policies for reporting preclinical research and the implementation of a "Basic Science checklist" by *Stroke*,[28] which is purported to have improved the quality and designs of preclinical studies published in that journal.[32]

Finally, the number of citations has been used as a surrogate measure of research impact and influence.[33–37] Therefore, Scopus (Elsevier) was queried to identify original research articles that cited preclinical cardiovascular studies published between July 2006 and June 2011. This 5-year period was selected as it ensured that each article had ≥5 years of citation data available in June 2016. Reviews, conference papers, editorials, letters, books/book chapters, and errata were excluded. Scopus was selected because it is the largest database available for citation analysis and retrieves a greater number of citations when compared with others.[38] Prespecified analyses of citations at 60 months after the index publication were performed, stratified by individual and cumulative study design elements and adjusted for journal of publication, year of publication, and CVD studied. Given the possibility that certain publications may have garnered attention in the short-term but could have ultimately been disproven or deemed irreproducible, a sensitivity analysis of citation counts at 36 months was also undertaken as this time point approximates contemporary mean time-to-retraction.[39]

Categorical variables are reported as number (%) and were compared via χ^2 tests. Continuous variables are reported as median (interquartile range) and were compared using Wilcoxon signed-rank or Kruskal–Wallis tests. Inter-rater agreement was calculated using Cohen κ statistic and percent agreement. Temporal patterns in the proportions of studies reporting randomization, blinding, and sample size estimations were evaluated via Cochrane–Armitage trend tests and journal-specific logistic regression models adjusting for CVD studied and animal model used when the number of events per predictor variable was adequate.[40–42] For the latter adjusted analyses, backward elimination was used for model building using a criterion of $P<0.20$ for specific CVD and animal model predictor variable inclusion. Associations between study design elements and citation counts were examined via stratification and multivariable linear regression. Non-normally distributed variables were log-transformed, when required. All analyses were performed using SAS 9.4 (SAS Institute Inc, Cary, NC) using a 2-tailed α level of 0.05 to define statistical significance.

Results

Study Selection and Characteristics

As previously described,[29] of 28 636 articles screened, 3396 met inclusion criteria and were analyzed (Figure 1). Inter-rater

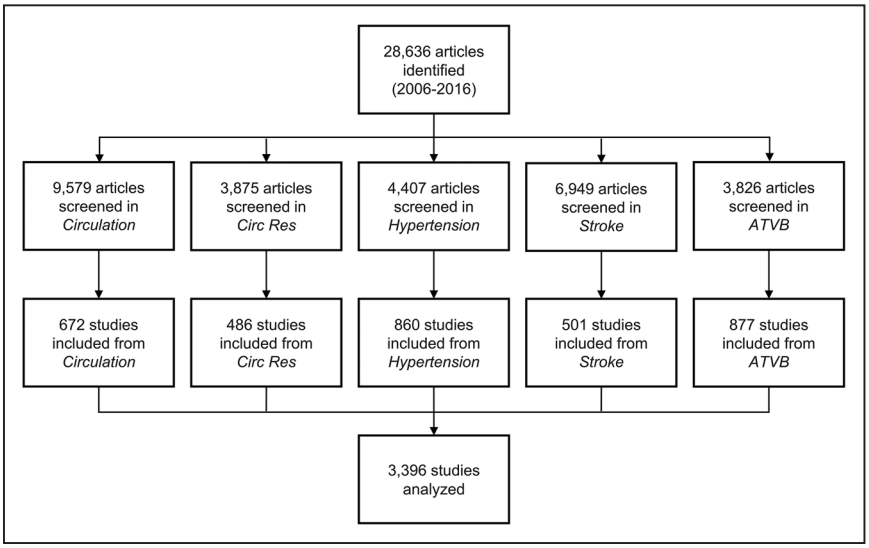


Figure 1. Literature search and results. *ATVB* indicates *Arteriosclerosis, Thrombosis and Vascular Biology*; and *Circ Res*, *Circulation Research*.

agreement for study inclusion before resolution was 94.5% (κ=0.72; 95% CI, 0.70–0.73). Atherosclerosis, hypertension, stroke, and cardiomyopathy/heart failure were the most commonly studied CVDs (range 14.2%–19.5%). Most other CVDs were the focus of <5.0% of studies. Ten studies on resuscitative medicine were identified and were included in the "other" category, which comprised 3.0% of studies. Nearly one third of studies examined a therapeutic intervention. Mice or rats were most often used by researchers (used in 89.8% of studies) whereas guinea pigs, gerbils, or hamsters were used in <0.2%. Multiple animal models were used in 2.8% of studies (Table).

Randomization

Randomization of animals was reported in 740 studies (21.8%) overall, but was more frequently noted in the subset examining therapeutic interventions (38.3% versus 15.0%, $P<0.0001$). When studies were stratified by CVD studied, significant differences in the proportions reporting randomization were noted ($P<0.0001$, range 5.6%–46.5%) with a lack of randomization predominating in all cases except stroke (Figure 2A). Significant differences in the proportions of studies reporting randomization were also observed when stratified by animal model used ($P<0.0001$, range 14.6%–45.0%) with a lack of randomization predominating in all cases except when pigs or a combination of animal models were used (Figure 2B).

Blinding

Blinding of treatment allocation or outcome assessment was reported in 1110 studies (32.7%) overall, but was also more frequently noted in the subset examining therapeutic interventions (41.9% versus 28.9%, $P<0.0001$). When stratified by CVD studied, significant differences in the proportions of studies reporting any blinding were observed ($P<0.0001$, range 10.9%–62.6%). Studies on stroke ranked highest among CVDs although arterial aneurysms/dissections and transplantation medicine had comparable numbers of blinded and nonblinded studies. For all other CVDs, studies with blinding formed the minority (Figure 3A). No significant difference in the proportions of studies reporting blinding were seen when stratified by animal model used ($P=0.369$, range 20.9%–36.8%; Figure 3B).

Table. Study Characteristics

	Circulation (n=672)	*Circ Res* (n=486)	*Hypertension* (n=860)	*Stroke* (n=501)	*ATVB* (n=877)	Total (n=3396)
Disease studied						
Atherosclerosis	119 (17.7)	113 (23.3)	14 (1.6)	2 (0.4)	385 (43.9)	633 (18.6)
Hypertension	34 (5.1)	35 (7.2)	577 (67.1)	1 (0.2)	14 (1.6)	661 (19.5)
Arterial aneurysm	16 (2.4)	21 (4.3)	15 (1.7)	24 (4.8)	62 (7.1)	138 (4.1)
Myocardial infarction	95 (14.1)	56 (11.5)	19 (2.2)	0 (0)	115 (13.1)	285 (8.4)
Cardiomyopathy/heart failure	173 (25.7)	154 (31.7)	135 (15.7)	0 (0)	21 (2.4)	483 (14.2)
Pulmonary hypertension	57 (8.5)	29 (6.0)	34 (4.0)	0 (0)	15 (1.7)	135 (4.0)
Arrhythmia	65 (9.7)	35 (7.2)	4 (0.5)	0 (0)	0 (0)	104 (3.1)
Valvular disease	9 (1.3)	4 (0.8)	0 (0)	0 (0)	14 (1.6)	27 (0.8)
Stroke	20 (3.0)	4 (0.8)	9 (1.1)	435 (86.8)	18 (2.1)	486 (14.3)
Transplantation medicine	28 (4.2)	15 (3.1)	0 (0)	31 (6.2)	32 (3.7)	106 (3.1)
Metabolic/endocrine disorders	19 (2.8)	10 (2.1)	23 (2.7)	2 (0.4)	75 (8.6)	129 (3.8)
Hematologic disorders	19 (2.8)	8 (1.7)	1 (0.1)	0 (0)	79 (9.0)	107 (3.2)
Other	18 (2.7)	2 (0.4)	4 (0.5)	6 (1.2)	47 (5.4)	102 (3.0)
Animal model(s) used						
Mouse	463 (68.9)	391 (80.5)	357 (41.5)	223 (44.5)	744 (84.8)	2178 (64.1)
Rat	84 (12.5)	44 (9.1)	443 (51.5)	237 (47.3)	63 (7.3)	872 (25.7)
Rabbit	10 (1.5)	15 (3.1)	16 (1.9)	18 (3.6)	23 (2.6)	82 (2.4)
Dog	37 (5.5)	12 (2.5)	13 (1.5)	5 (1.0)	0 (0)	67 (2.0)
Pig	84 (12.5)	4 (0.8)	9 (1.1)	3 (0.6)	19 (2.2)	60 (1.8)
Primate	3 (0.5)	1 (0.2)	2 (0.2)	3 (0.6)	8 (0.9)	17 (0.5)
Other	10 (1.5)	4 (0.8)	9 (1.1)	1 (0.2)	1 (0.1)	25 (0.7)
Combination	40 (6.0)	15 (3.1)	11 (1.3)	11 (2.2)	18 (2.1)	95 (2.8)
Therapeutic study	162 (24.1)	85 (17.5)	221 (25.7)	264 (52.7)	256 (29.2)	988 (29.1)

Values reported as number (%). *ATVB* indicates *Arteriosclerosis, Thrombosis, and Vascular Biology*; and *Circ Res, Circulation Research.*

Sample Size Estimation

A priori sample size estimations or power calculations were reported in 79 studies (2.3%). This aspect of study design was also more frequently reported in the subset examining therapeutic interventions (4.2% versus 1.6%, $P<0.0001$). Significant differences in the proportion of studies reporting sample size calculations were observed when stratified by CVD studied ($P<0.0001$, range 0%–9.5%) with stroke ranking highest (Figure 4A), but not when stratified by animal model used ($P=0.270$, range 0%–6.1%; Figure 4B).

Inclusion of Both Sexes

Sex bias prevalence and detailed temporal trends have been described previously.[29] After excluding studies that did not report the sex of the animals used, significant differences were noted in the proportions including animals of both sexes when stratified by CVD of interest ($P<0.0001$, range 2.4%–34.7%) and by animal model used ($P<0.0001$, range 6.7%–26.9%; Online Figure).

Temporal Patterns in Study Design Element Implementation

Over the past 10 years, there have been no significant changes in the proportions of studies reporting blinding or randomization (inclusion of both sexes has been reported previously[29]). There has been a significant increase in the proportion reporting a priori sample size estimations ($P_{trend}<0.0001$), although it remains below 7% (Figure 5). There was no substantial difference in the prevalence of any of these study design elements before and after the NIH principles and guidelines for reporting preclinical research were published in 2014 (range of differences 0.7%–3.6%). Post hoc calculations suggest that our sample sizes had ≥80% power to detect a difference of 5.0% for all study design elements.

CVD-specific temporal analyses suggest that preclinical studies on stroke are increasingly incorporating randomization, blinding, sample size estimations, and inclusion of both sexes—a pattern that is not seen in studies on atherosclerosis, hypertension, or cardiomyopathy/heart failure (the most

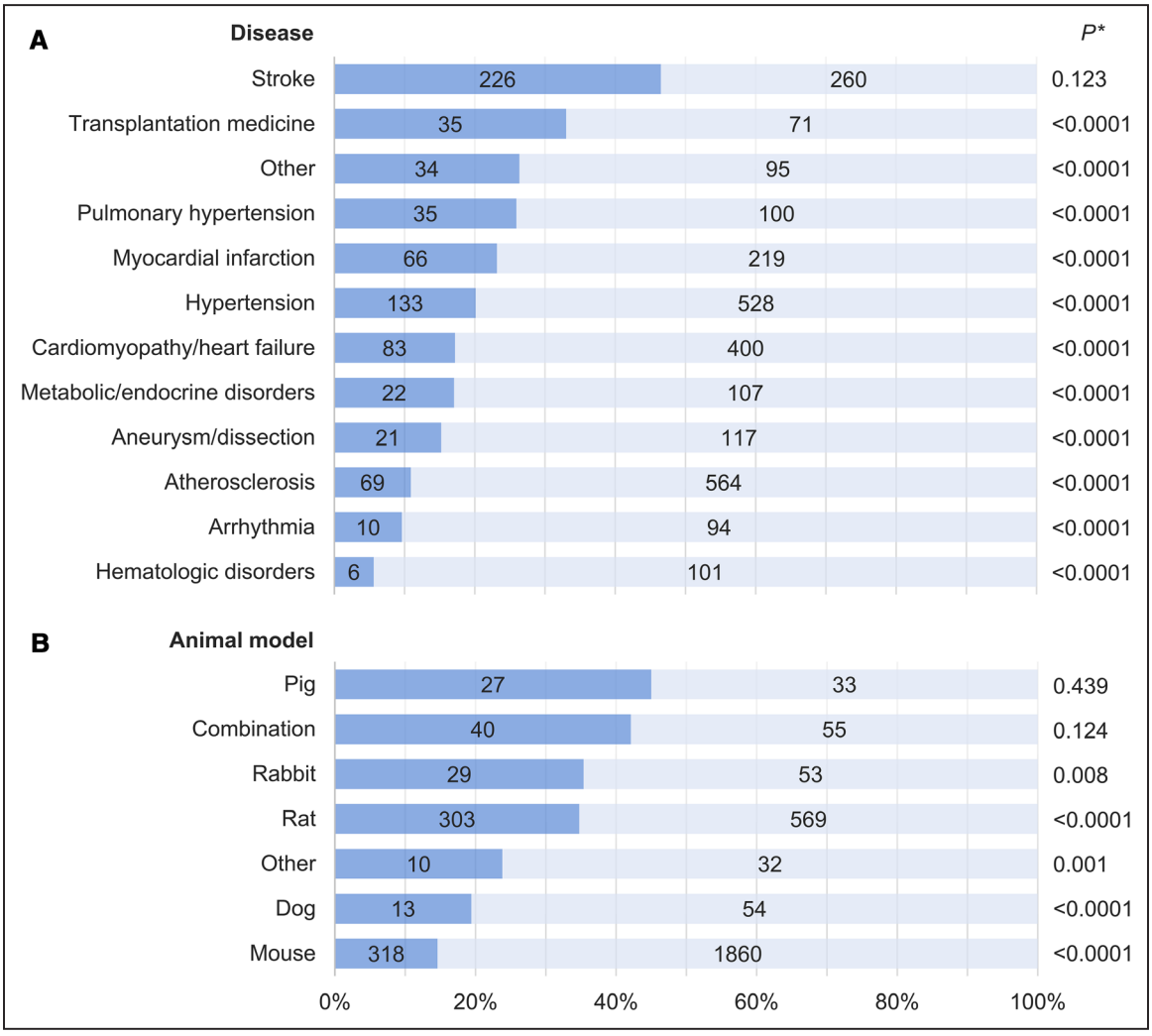


Figure 2. Randomization in preclinical cardiovascular research studies published over a 10-y period stratified by (A) disease studied and (B) species used. Dark blue corresponds to the proportion of studies implementing randomization; numbers in bars correspond to absolute numbers of studies. Valvular disease and resuscitative medicine studies were included in the "Other" category because of the small number of relevant publications. "Combination" refers to more than one animal species used within the same publication; animal models in the "Other" category included guinea pig, gerbil, and hamster. *For comparison of studies incorporating randomization vs not.

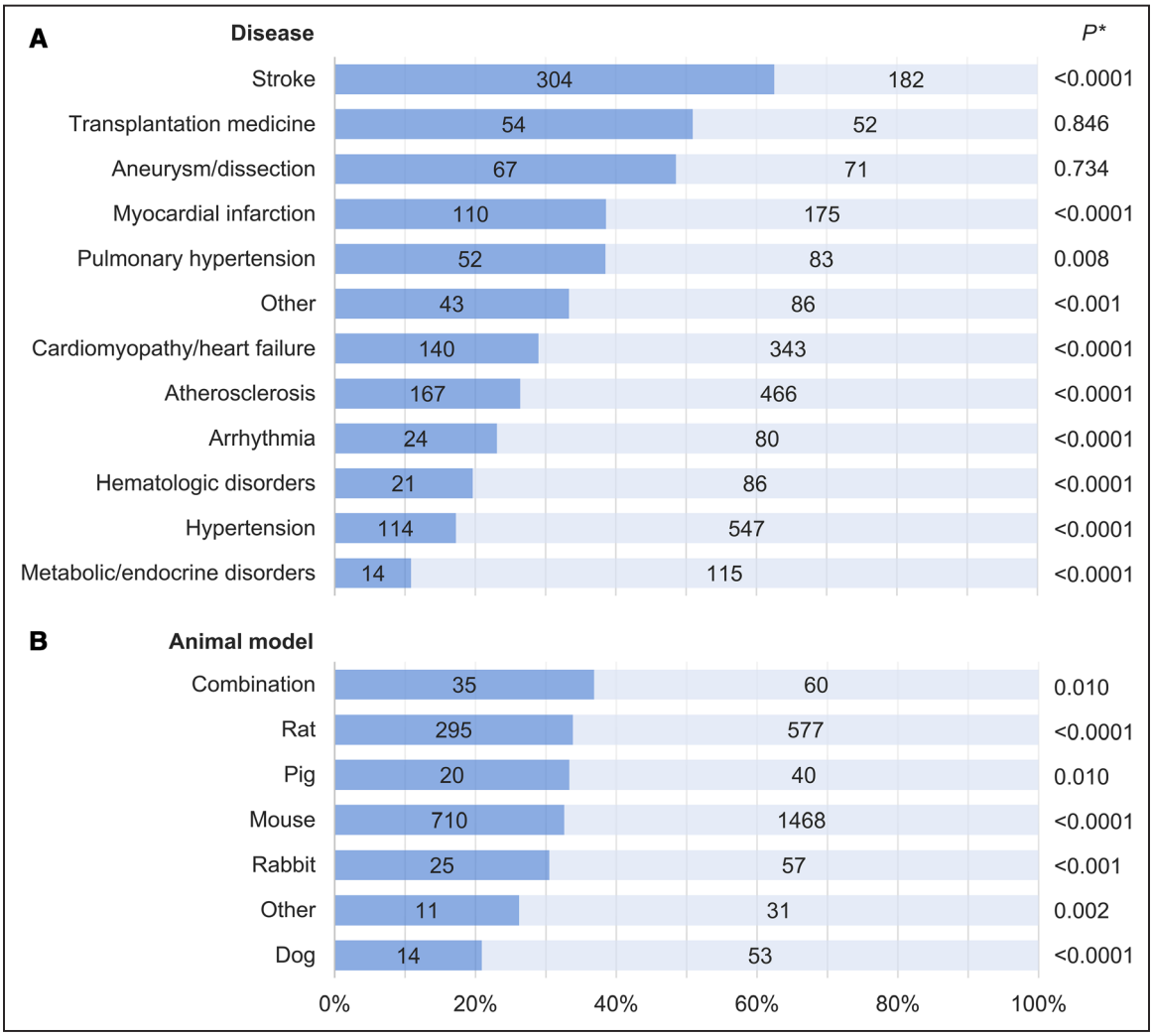


Figure 3. Blinding in preclinical cardiovascular research studies published over a 10-y period stratified by (A) disease studied and (B) species used. Dark blue corresponds to the proportion of studies implementing blinding; numbers in bars correspond to absolute numbers of studies. Valvular disease and resuscitative medicine studies were included in the "Other" category because of the small number of relevant publications. "Combination" refers to more than one animal species used within the same publication; animal models in the "Other" category included guinea pig, gerbil, and hamster. *For comparison of studies incorporating blinding vs not.

commonly studied CVDs). Studies on atherosclerosis have shown a modest increase in the proportion reporting sample size estimation in recent years, but an overall significant decrease in the proportion reporting blinding, whereas the proportion of studies on cardiomyopathy/heart failure that include both sexes has significantly decreased. No other significant changes were observed during the 10-year period (Figure 6).

Given the disproportionate amount of stroke research published in *Stroke*, which has reported improvements in the quality and design of published preclinical research following the introduction of their "Basic Science Checklist" in 2011,[32] CVD-adjusted and animal model-adjusted comparisons of study design implementation before and after this time point were performed for each journal. *Stroke* uniquely exhibited significant improvements in all measures of methodological quality (range of adjusted odds ratios 2.4–8.2, $P<0.0001$ for all study design elements; Online Table). These analyses also identified stroke as the CVD studied as an independent positive predictor of one or more study design element in every journal.

Citations According to Index Study Methodology

At 60 months, 41 441 articles citing the 1681 preclinical studies that were published between July 2006 and June 2011 were identified. The median citation count per preclinical study was 20 (interquartile range 13–31, range 0–131). Studies that implemented randomization, blinding, or sample size estimation had similar numbers of citations as those that did not; however, studies that included both males and females were cited less frequently (median 18 versus 20, $P=0.023$). The cumulative number of these study design elements was not associated with citation counts (Figure 7). No study implemented all 4 of the above design elements and only 20 studies (1.2%) incorporated 3; therefore, studies with ≥2 were grouped together for this analysis. The above associations (or lack thereof) between study design element(s) and number of citations persisted after adjusting for journal of publication, year of publication,

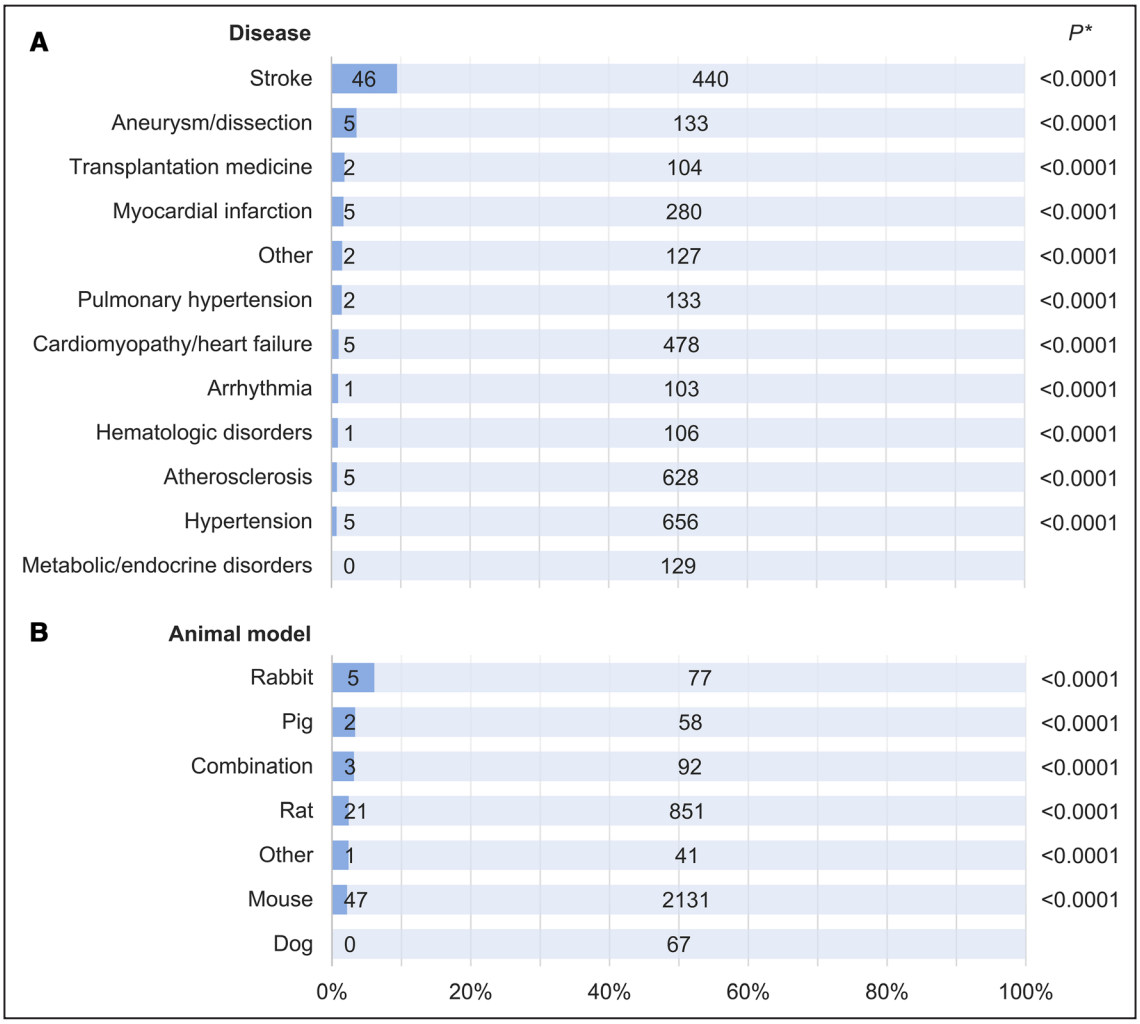


Figure 4. Sample size estimation in preclinical cardiovascular research studies published over a 10-y period stratified by (A) disease studied and (B) species used. Dark blue corresponds to the proportion of studies reporting sample size estimations/power calculations; numbers in bars correspond to absolute numbers of studies. Valvular disease and resuscitative medicine studies were included in the "Other" category because of the small number of relevant publications. "Combination" refers to more than one animal species used within the same publication; animal models in the "Other" category included guinea pig, gerbil, and hamster. *For comparison of studies incorporating sample size estimation vs not.

and CVD studied in multivariable regression models and all findings were comparable in sensitivity analyses of citation counts at 36 months.

Discussion

Preclinical cardiovascular research using animal models plays an integral role in advancing the care of patients afflicted with CVDs. However, its impact is contingent on its scientific validity, reproducibility, and relevance to human physiology and disease. It is understood that inherent limitations of using animals to model human diseases can undermine the predictive value of preclinical findings, rendering it difficult for even the most skilled scientists to make impactful discoveries.[12] However, this difficulty is compounded when methodological sources of bias are introduced. We systematically examined a continuous and large body of preclinical cardiovascular research to determine how often randomization, blinding, and a priori sample size estimation are incorporated and did so over a sufficiently long period to draw meaningful conclusions on the temporal patterns of these practices. We report that, overall, these design elements are rarely implemented in studies published in leading peer-reviewed cardiovascular journals, paralleling the previously reported prevalence of sex bias.[29] Furthermore, apart from a modest increase in the proportion of studies reporting sample size calculations, there has been no improvement in these practices over the past decade, although stroke research may be a notable exception. Finally, analyses of citation counts suggest that crucial methodological aspects of preclinical studies may be overlooked by cardiovascular researchers, affording potentially biased research comparable influence on scientific research efforts as methodologically more robust studies.

Poorly designed preclinical studies not only contribute to experimental irreproducibility[2,7] and wasted resources[43,44] but also may result in erroneous conclusions regarding the treatment effects,[9,21,23,45,46] which can ultimately spur or deter clinical trials in humans with consequent risk of harm.[12,20,47] The experimental design elements evaluated in our study are

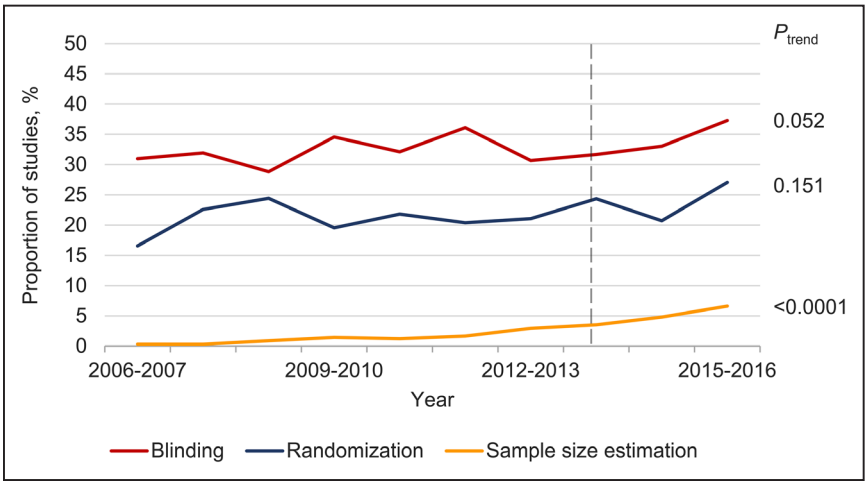


Figure 5. Temporal patterns in randomization, blinding, and sample size estimation in preclinical cardiovascular studies. Dashed line indicates the publication of the National Institutes of Health guidelines and policies for reporting preclinical research. Data for inclusion of both sexes have been reported previously.[29]

deemed crucial to improving the quality of preclinical research by addressing selection bias (randomization); minimizing performance, detection, and attrition bias (blinding); ensuring adequate statistical power and the ethical use of animals (sample size estimation); and promoting research relevance for both men and women (inclusion of both sexes).[1,6,7,10,13,26,48,49] Although these elements are routinely implemented in clinical trials as they have been shown to protect against bias and imprecision,[50–53] they have conspicuously not permeated preclinical experiments.[1,23,54–56] The high degree of experimental control that is possible in preclinical research (eg, via genetic and environmental homogeneity) can reduce variation and the sample sizes required relative to clinical trials; however, variations in injury or disease induction and the potential for persistent and unrecognized confounders represent important sources of bias.[6,31,57]

Furthermore, despite a pervasive belief among scientists that there is a reproducibility "crisis" in research that is attributable to methodological and reporting shortcomings,[14] our citation analyses suggest that greater methodological rigor in preclinical cardiovascular research does not translate into greater scientific influence. A similar observation was noted by Amgen in a study of 53 preclinical cancer research publications: studies that they could not adequately reproduce were cited as often, if not more often, than those that they could successfully reproduce, irrespective of the journal of

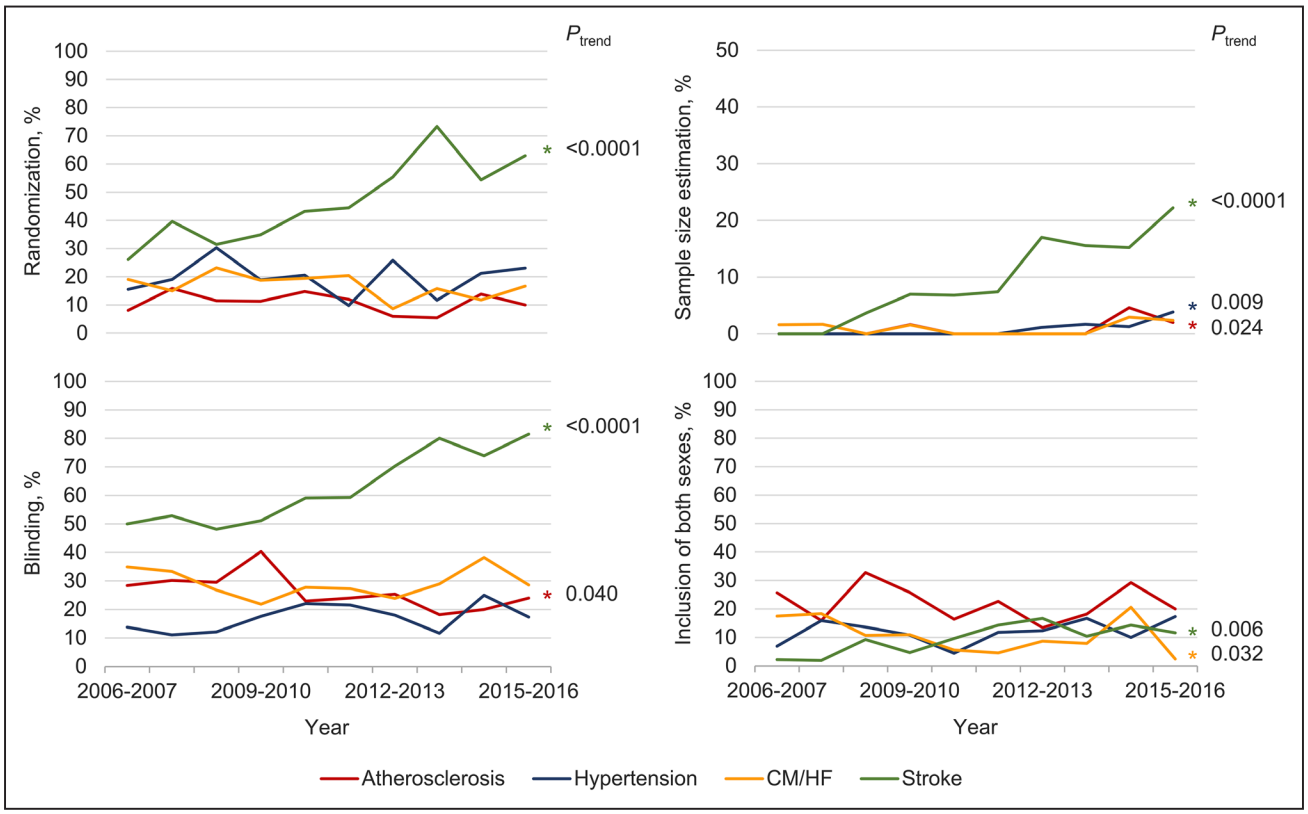


Figure 6. Temporal patterns in randomization, blinding, sample size estimation, and inclusion of both sexes in preclinical research for the most commonly studied cardiovascular diseases. Note the different scale of *y*-axis for sample size estimation. CM/HF indicates cardiomyopathy/heart failure.

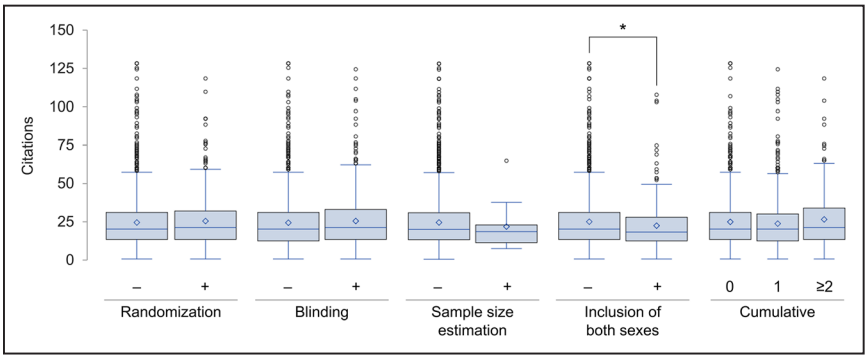


Figure 7. Box-whisker plot of the number of citations by methodological rigor of index preclinical study (n=1681). Diamond symbol plotted at mean. Outliers identified as beyond 1.5 interquartile range. Cumulative refers to sum of randomization, blinding, sample size estimation, and inclusion of both sexes (each contributing 1 point). *P<0.05.

publication's impact factor.[12] The suggestion that methodologically robust studies are given equal consideration as studies at greater risk of bias by cardiovascular researchers is troubling as it undermines the reputed "self-correcting" tenet of science and increases the risk of pursuing unfruitful avenues of research.[2,5,44,48]

Cogent analyses and arguably practical solutions to improve preclinical research quality (and secondarily to enhance research translation) have been proposed. However, evidence of changes in research practice and data on the impact of corrective actions are scarce.[13] Several guidelines and checklists have been developed to improve preclinical methodology[45,58,59] and reporting,[25,45,60–62] yet there is little indication that they are effecting change despite being widely endorsed.[43,58,63] Systematic reviews and meta-analyses have been advocated as safeguards to expose bias in preclinical research before embarking on clinical trials,[9,47,64,65] but they are limited by the internal validity of included studies[66] and are not routinely performed. Changes in research funding such as those undertaken by the NIH[2,26] may bolster these efforts; however, our analyses of animal studies in the cardiovascular sciences have not shown signs of substantial changes in research practices or reporting since they were announced.[29]

Journal editors and reviewers directly influence what is published thereby often serving as ultimate gatekeepers of research findings; yet, few journals effectively encourage the use of reporting guidelines.[67] At the time of writing, there exists considerable variation in author guidelines for reporting animal studies among leading cardiovascular journals with few requiring authors to report randomization, blinding, sample size estimation, or the sex of the animals used—requirements that could increase transparency and reinforce the importance of these study design elements.[13] Uniquely, however, *Stroke* implemented a "Basic Science Checklist" in 2011 (updated in 2016),[28] which includes all of these relevant elements, forms part of the manuscript submission process, and is evaluated by editors and reviewers. Over the 2.5 years after its introduction, Minnerup et al[32] noted improvements in randomization, blinding, and allocation concealment compared with the preceding 18 months. We noted significant improvements in the quality of preclinical stroke research, 90% of which was published in *Stroke*, raising the question of whether these improvements were due to journal editorial policies/culture or driven by changes in research practices among stroke researchers in general. Our post hoc analyses suggest that it is likely a combination of both: *Stroke* uniquely showed improvements in all study design elements even after adjusting for CVD studied and animal model used, but stroke as the CVD studied was identified as an independent predictor for at least one study design element for every journal examined. Our findings, therefore, corroborate those of Minnerup et al[32] and expand upon them by demonstrating that these improvements have continued into 2016, extend to sample size estimation (and to a lesser extent inclusion of both sexes), contrast with an extended preceding period of suboptimal reporting, and have not been paralleled in other prominent cardiovascular journals. Therefore, although not conclusive, our data suggest that journal editors and reviewers can exert considerable influence on preclinical research practices and reporting. Yet, our findings also suggest that stroke research has independently improved in quality—a finding that may be attributable to its community's early appreciation of the importance of methodological rigor in preclinical science and extensive involvement in efforts to improve research translation.[7,20,21,46,58,68–71]

Our study is not without limitations. The study sample comprised articles from 5 cardiovascular journals over 10 years, which may not fully represent preclinical cardiovascular research publications or practices. However, we deliberately selected AHA journals given their prominence in the field of cardiovascular medicine, established reputation, collective focus on a broad range of CVDs, and unanimous endorsement of the NIH guidelines on rigor and reproducibility. Given that none of the American College of Cardiology or European Society of Cardiology journals have endorsed the guidelines so far,[25] our data may actually overestimate the methodological quality of most preclinical cardiovascular research. There is no accepted definition for preclinical studies therefore criteria were developed for this study. It is, therefore, possible that not all relevant studies were included in our analysis. However, the inclusion criteria used, the number of journals reviewed, and the substantial inter-rater agreement for study inclusion support the validity of our results and render our data the best available on methodological rigor in preclinical cardiovascular research. Furthermore, our inclusion criteria are in line with previously proposed distinctions between "exploratory" and "confirmatory" preclinical research.[7,30,31] Our analysis considered blinding as a single and dichotomous variable, which limits detailed assessments of this design element, and did not examine all relevant potential sources of bias. Data handling and publication bias, for instance, are important factors that may contribute to irreproducibility in science.[2,6,7,9–11,13,43] As well, factors other than experimental

methodology can influence the predictive value of preclinical research, including how well an animal model reflects human physiology and disease,[6,9,45,72] the appropriateness of statistical analyses,[73] and a lack of standardization of definitions and surrogate markers,[74,75] which were not examined. Although article citation is the most commonly used measure of research impact, it is an imperfect indicator.[33,34] Finally, study design elements may have been implemented but not reported, which could result in underestimates of their prevalence. However, underreporting of measures of methodological rigor is believed to be low[49,65] and therefore unlikely to have significantly influenced our results.

Our analysis of preclinical research published in leading cardiovascular journals over the past 10 years demonstrates that methodological sources of potential bias and imprecision are prevalent, have not appreciably improved over time, and may be overlooked by researchers when basing subsequent studies. Concerted efforts to address this problem are urgently needed. Stroke research has uniquely shown substantial improvement in several measures of quality in recent years, warranting a closer examination to identify drivers of its success.

Disclosures

None.

References

1. Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, Schulz KF, Tibshirani R. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*. 2014;383:166–175. doi: 10.1016/S0140-6736(13)62227-8.
2. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. *Nature*. 2014;505:612–613.
3. Henderson VC, Kimmelman J, Fergusson D, Grimshaw JM, Hackam DG. Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. *PLoS Med*. 2013;10:e1001489. doi: 10.1371/journal.pmed.1001489.
4. Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical R&D. *Nat Rev Drug Discov*. 2011;10:428–438. doi: 10.1038/nrd3405.
5. Bunnage ME. Getting pharmaceutical R&D back on target. *Nat Chem Biol*. 2011;7:335–339. doi: 10.1038/nchembio.581.
6. van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, Macleod MR. Can animal models of disease reliably inform human studies? *PLoS Med*. 2010;7:e1000245. doi: 10.1371/journal.pmed.1000245.
7. Landis SC, Amara SG, Asadullah K, et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*. 2012;490:187–191. doi: 10.1038/nature11556.
8. Scott S, Kranz JE, Cole J, Lincecum JM, Thompson K, Kelly N, Bostrom A, Theodoss J, Al-Nakhala BM, Vieira FG, Ramasubbu J, Heywood JA. Design, power, and interpretation of studies in the standard murine model of ALS. *Amyotroph Lateral Scler*. 2008;9:4–15. doi: 10.1080/17482960701856300.
9. Perel P, Roberts I, Sena E, Wheble P, Briscoe C, Sandercock P, Macleod M, Mignini LE, Jayaram P, Khan KS. Comparison of treatment effects between animal experiments and clinical trials: systematic review. *BMJ*. 2007;334:197. doi: 10.1136/bmj.39048.407928.BE.
10. Muhlhausler BS, Bloomfield FH, Gillman MW. Whole animal experiments should be more like human randomized controlled trials. *PLoS Biol*. 2013;11:e1001481. doi: 10.1371/journal.pbio.1001481.
11. 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. doi: 10.1038/nrd3439-c1.
12. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. *Nature*. 2012;483:531–533. doi: 10.1038/483531a.
13. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res*. 2015;116:116–126. doi: 10.1161/CIRCRESAHA.114.303819.
14. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature*. 2016;533:452–454. doi: 10.1038/533452a.
15. Unreliable research: Trouble at the lab. *The Economist*. 2013.
16. Lehrer J. The truth wears off: Is there something wrong with the scientific method? *The New Yorker*. 2010.
17. Freedman DH. Lies, damned lies, and medical science. *The Atlantic*. 2010.
18. Osherovich L. Hedging against academic risk. *Science-Business eXchange*. 2011;4:doi:10.1038/scibx.2011.1416
19. Begley CG. Six red flags for suspect work. *Nature*. 2013;497:433–434. doi: 10.1038/497433a.
20. Horn J, de Haan RJ, Vermeulen M, Luiten PG, Limburg M. Nimodipine in animal model experiments of focal cerebral ischemia: a systematic review. *Stroke*. 2001;32:2433–2438.
21. Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke*. 2008;39:2824–2829. doi: 10.1161/STROKEAHA.108.515957.
22. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2:e124. doi: 10.1371/journal.pmed.0020124.
23. Hackam DG, Redelmeier DA. Translation of research evidence from animals to humans. *JAMA*. 2006;296:1731–1732. doi: 10.1001/jama.296.14.1731.
24. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. *Nat Biotechnol*. 2014;32:40–51. doi: 10.1038/nbt.2786.
25. National Institutes of Health. Principles and guidelines for reporting preclinical research. https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research. Accessed December 2, 2016.
26. Clayton JA, Collins FS. Policy: NIH to balance sex in cell and animal studies. *Nature*. 2014;509:282–283.
27. Daugherty A, Hegele RA, Mackman N, Rader DJ, Schmidt AM, Weber C. Complying with the national institutes of health guidelines and principles for rigor and reproducibility: refutations. *Arterioscler Thromb Vasc Biol*. 2016;36:1303–1304. doi: 10.1161/ATVBAHA.116.307906.
28. Vahidy F, Schäbitz WR, Fisher M, Aronowski J. Reporting standards for preclinical studies of stroke therapy. *Stroke*. 2016;47:2435–2438. doi: 10.1161/STROKEAHA.116.013643.
29. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald Z, Simard T, Clancy AA, Russo JJ, Welch V, Wells GA, Hibbert B. Sex bias is increasingly prevalent in preclinical cardiovascular research: implications for translational medicine and health equity for women: a systematic assessment of leading cardiovascular journals over a 10-year period. *Circulation*. 2017;135:625–626. doi: 10.1161/CIRCULATIONAHA.116.026668.
30. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biol*. 2014;12:e1001863. doi: 10.1371/journal.pbio.1001863.
31. Unger EF. All is not well in the world of translational research. *J Am Coll Cardiol*. 2007;50:738–740. doi: 10.1016/j.jacc.2007.04.067.
32. Minnerup J, Zentsch V, Schmidt A, Fisher M, Schäbitz WR. Methodological quality of experimental stroke studies published in the stroke journal: time trends and effect of the basic science checklist. *Stroke*. 2016;47:267–272. doi: 10.1161/STROKEAHA.115.011695.
33. Ranasinghe I, Shojaee A, Bikdeli B, Gupta A, Chen R, Ross JS, Masoudi FA, Spertus JA, Nallamothu BK, Krumholz HM. Poorly cited articles in peer-reviewed cardiovascular journals from 1997 to 2007: analysis of 5-year citation rates. *Circulation*. 2015;131:1755–1762. doi: 10.1161/CIRCULATIONAHA.114.015080.
34. Danthi N, Wu CO, Shi P, Lauer M. Percentile ranking and citation impact of a large cohort of National Heart, Lung, and Blood Institute-funded cardiovascular R01 grants. *Circ Res*. 2014;114:600–606. doi: 10.1161/CIRCRESAHA.114.302656.
35. Kaltman JR, Evans FJ, Danthi NS, Wu CO, DiMichele DM, Lauer MS. Prior publication productivity, grant percentile ranking, and topic-normalized citation impact of NHLBI cardiovascular R01 grants. *Circ Res*. 2014;115:617–624. doi: 10.1161/CIRCRESAHA.115.304766.
36. Lauer MS, Danthi NS, Kaltman J, Wu C. Predicting productivity returns on investment: thirty years of peer review, grant funding, and publication of highly cited papers at the National Heart, Lung, and Blood Institute. *Circ Res*. 2015;117:239–243. doi: 10.1161/CIRCRESAHA.115.306830.
37. Fabsitz RR, Papanicolaou GJ, Sholinsky P, Coady SA, Jaquish CE, Nelson CR, Olson JL, Puggal MA, Purkiser KL, Srinivas PR, Wei GS, Wolz M, Sorlie PD. Impact of National Heart, Lung, and Blood Institute-Supported Cardiovascular Epidemiology Research, 1998 to 2012. *Circulation*. 2015;132:2028–2033. doi: 10.1161/CIRCULATIONAHA.114.014147.

38. Kulkarni AV, Aziz B, Shams I, Busse JW. Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. *JAMA*. 2009;302:1092–1096. doi: 10.1001/jama.2009.1307.
39. Steen RG, Casadevall A, Fang FC. Why has the number of scientific retractions increased? *PLoS One*. 2013;8:e68397. doi: 10.1371/journal.pone.0068397.
40. Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. *J Clin Epidemiol*. 1995;48:1495–1501.
41. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol*. 1995;48:1503–1510.
42. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49:1373–1379.
43. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, Michie S, Moher D, Wager E. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet*. 2014;383:267–276. doi: 10.1016/S0140-6736(13)62228-X.
44. Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. *PLoS Biol*. 2015;13:e1002165. doi: 10.1371/journal.pbio.1002165.
45. Macleod MR, O'Collins T, Howells DW, Donnan GA. Pooling of animal experimental data reveals influence of study design and publication bias. *Stroke*. 2004;35:1203–1208. doi: 10.1161/01.STR.0000125719.25853.20.
46. Macleod MR, O'Collins T, Horky LL, Howells DW, Donnan GA. Systematic review and metaanalysis of the efficacy of FK506 in experimental stroke. *J Cereb Blood Flow Metab*. 2005;25:713–721. doi: 10.1038/sj.jcbfm.9600064.
47. Pound P, Ebrahim S, Sandercock P, Bracken MB, Roberts I; Reviewing Animal Trials Systematically (RATS) Group. Where is the evidence that animal research benefits humans? *BMJ*. 2004;328:514–517. doi: 10.1136/bmj.328.7438.514.
48. Clayton JA. Studying both sexes: a guiding principle for biomedicine. *FASEB J*. 2016;30:519–524. doi: 10.1096/fj.15-279554.
49. 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. doi: 10.1371/journal.pbio.1002273.
50. Savović J, Jones HE, Altman DG, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Ann Intern Med*. 2012;157:429–438. doi: 10.7326/0003-4819-157-6-201209180-00537.
51. Tzoulaki I, Siontis KC, Ioannidis JP. Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: meta-epidemiology study. *BMJ*. 2011;343:d6829.
52. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282:1061–1066.
53. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ*. 2006;174:469–476. doi: 10.1503/cmaj.050090.
54. Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, Fry D, Hutton J, Altman DG. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One*. 2009;4:e7824. doi: 10.1371/journal.pone.0007824.
55. Beery AK, Zucker I. Sex bias in neuroscience and biomedical research. *Neurosci Biobehav Rev*. 2011;35:565–572. doi: 10.1016/j.neubiorev.2010.07.002.
56. Yoon DY, Mansukhani NA, Stubbs VC, Helenowski IB, Woodruff TK, Kibbe MR. Sex bias exists in basic science and translational surgical research. *Surgery*. 2014;156:508–516. doi: 10.1016/j.surg.2014.07.001.
57. Reardon S. A mouse's house may ruin experiments. *Nature*. 2016;530:264. doi: 10.1038/nature.2016.19335.
58. Fisher M, Feuerstein G, Howells DW, Hurn PD, Kent TA, Savitz SI, Lo EH; STAIR Group. Update of the stroke therapy academic industry roundtable preclinical recommendations. *Stroke*. 2009;40:2244–2250. doi: 10.1161/STROKEAHA.108.541128.
59. Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW. SYRCLE's risk of bias tool for animal studies. *BMC Med Res Methodol*. 2014;14:43. doi: 10.1186/1471-2288-14-43.
60. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol*. 2010;8:e1000412. doi: 10.1371/journal.pbio.1000412.
61. Announcement: Reducing our irreproducibility. *Nature*. 2013;496
62. Hooijmans CR, Leenaars M, Ritskes-Hoitinga M. A gold standard publication checklist to improve the quality of animal studies, to fully integrate the Three Rs, and to make systematic reviews more feasible. *Altern Lab Anim*. 2010;38:167–182.
63. Baker D, Lidster K, Sottomayor A, Amor S. Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for preclinical animal studies. *PLoS Biol*. 2014;12:e1001756. doi: 10.1371/journal.pbio.1001756.
64. Sandercock P, Roberts I. Systematic reviews of animal experiments. *Lancet*. 2002;360:586. doi: 10.1016/S0140-6736(02)09812-4.
65. Hirst JA, Howick J, Aronson JK, Roberts N, Perera R, Koshiaris C, Heneghan C. The need for randomization in animal trials: an overview of systematic reviews. *PLoS One*. 2014;9:e98856. doi: 10.1371/journal.pone.0098856.
66. van Luijk J, Bakker B, Rovers MM, Ritskes-Hoitinga M, de Vries RB, Leenaars M. Systematic reviews of animal studies; missing link in translational research? *PLoS One*. 2014;9:e89981. doi: 10.1371/journal.pone.0089981.
67. Hirst A, Altman DG. Are peer reviewers encouraged to use reporting guidelines? A survey of 116 health research journals. *PLoS One*. 2012;7:e35621. doi: 10.1371/journal.pone.0035621.
68. Stroke Therapy Academic Industry Roundtable (STAIR). Recommendations for standards regarding preclinical neuroprotective and restorative drug development. *Stroke*. 1999;30:2752–2758
69. O'Collins VE, Macleod MR, Donnan GA, Horky LL, van der Worp BH, Howells DW. 1,026 experimental treatments in acute stroke. *Ann Neurol*. 2006;59:467–477. doi: 10.1002/ana.20741.
70. Crossley NA, Sena E, Goehler J, Horn J, van der Worp B, Bath PM, Macleod M, Dirnagl U. Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach. *Stroke*. 2008;39:929–934. doi: 10.1161/STROKEAHA.107.498725.
71. 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. doi: 10.1016/j.tins.2007.06.009.
72. Perrin S. Preclinical research: Make mouse studies work. *Nature*. 2014;507:423–425. doi: 10.1038/507423a.
73. García-Berthou E, Alcaraz C. Incongruence between test statistics and P values in medical papers. *BMC Med Res Methodol*. 2004;4:13. doi: 10.1186/1471-2288-4-13.
74. Johnson VE. Revised standards for statistical evidence. *Proc Natl Acad Sci U S A*. 2013;110:19313–19317. doi: 10.1073/pnas.1313476110.
75. Ioannidis JP. How to make more published research true. *PLoS Med*. 2014;11:e1001747. doi: 10.1371/journal.pmed.1001747.

Circulation Research
JOURNAL OF THE AMERICAN HEART ASSOCIATION

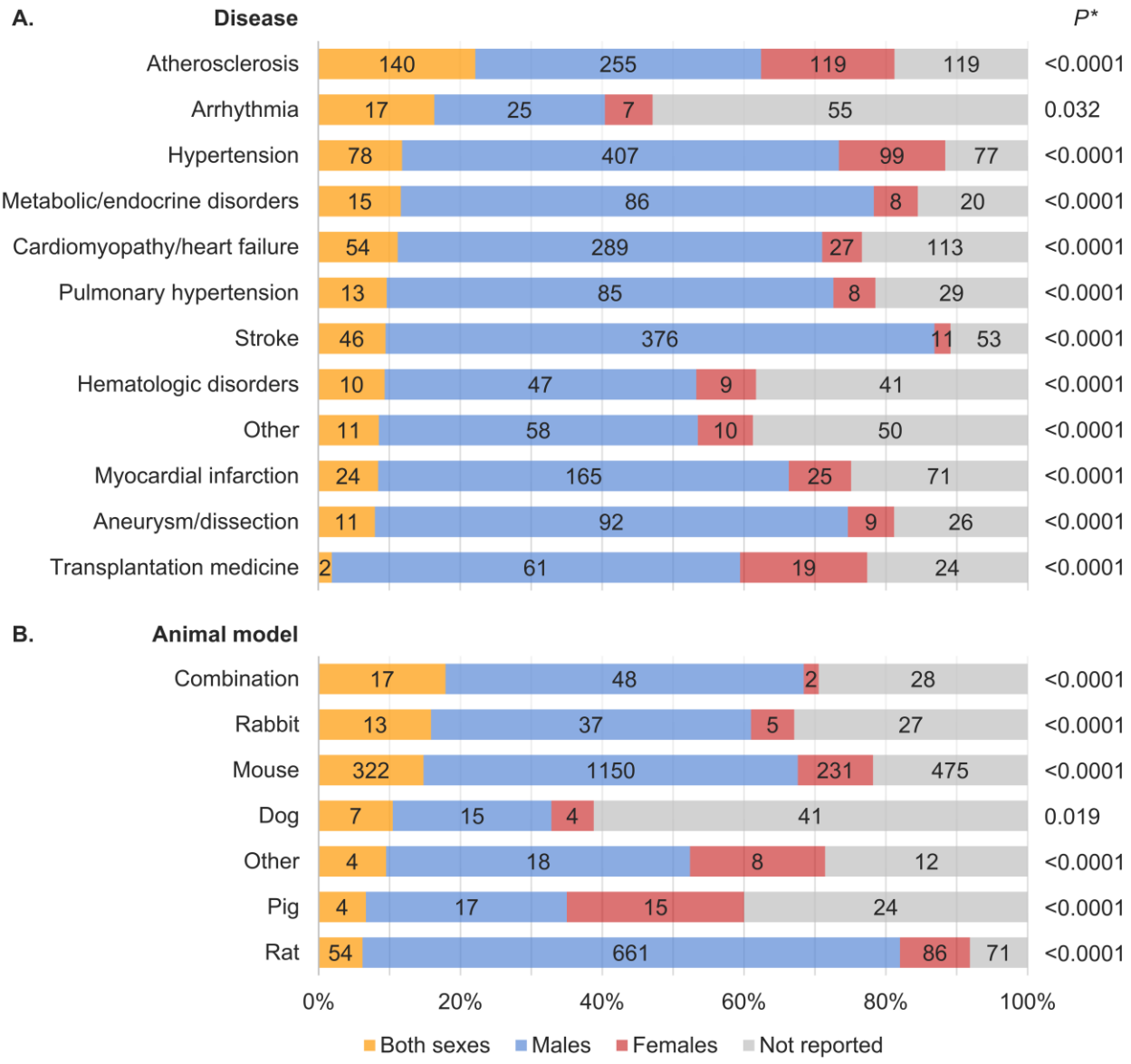
Methodological Rigor in Preclinical Cardiovascular Studies: Targets to Enhance Reproducibility and Promote Research Translation

F. Daniel Ramirez, Pouya Motazedian, Richard G. Jung, Pietro Di Santo, Zachary D. MacDonald, Robert Moreland, Trevor Simard, Aisling A. Clancy, Juan J. Russo, Vivian A. Welch, George A. Wells and Benjamin Hibbert

Circ Res. 2017;120:1916-1926; originally published online April 3, 2017;
doi: 10.1161/CIRCRESAHA.117.310628
Circulation Research is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 0009-7330. Online ISSN: 1524-4571

SUPPLEMENTAL MATERIAL



Online Supplemental Figure. Inclusion of both sexes in preclinical cardiovascular research studies published over a 10-year period stratified by (*A*) disease studied and (*B*) species used. Valvular disease and resuscitative medicine studies were included in the 'Other' category due to the small number of relevant publications. Numbers in bars correspond to absolute numbers of studies. 'Combination' refers to more than one animal species used within the same publication. *For comparison of proportion of studies including both sexes vs. males or females only.

Supplemental Table: Comparison of study design element implementation in preclinical studies before and after the implementation of the *Stroke* Basic Science Checklist, stratified by journal of publication

	Period 1* *n* (%)	Period 2* *n* (%)	Crude OR (95% CI)	*P*	Adjusted OR (95% CI)†	*P*†
Circulation	*n*=464	*n*=208				
Randomization	107 (23.1)	36 (17.3)	0.7 (0.5-1.1)	0.093	0.7 (0.4-1.1)	0.119
Blinding	169 (36.4)	59 (28.4)	0.7 (0.5-1.0)	0.042	0.7 (0.5-1.0)	0.043
Sample size estimation	7 (1.5)	5 (2.4)	1.6 (0.5-5.1)	0.422	NR	
Inclusion of both sexes	64 (13.8)	29 (13.9)	1.0 (0.6-1.6)	0.959	1.0 (0.6-1.6)	0.967
Circulation Research	*n*=303	*n*=183				
Randomization	35 (11.6)	29 (15.8)	1.4 (0.8-2.5)	0.176	1.4 (0.8-2.5)	0.261
Blinding	93 (30.7)	60 (32.8)	1.1 (0.7-1.6)	0.630	0.9 (0.6-1.4)	0.788
Sample size estimation	1 (0.3)	1 (0.3)	1.7 (0.1-26.7)	0.721	NR	
Inclusion of both sexes	57 (18.8)	33 (18.0)	0.9 (0.6-1.5)	0.830	1.0 (0.6-1.6)	0.937
Hypertension	*n*=485	*n*=375				
Randomization	104 (21.4)	81 (21.6)	1.0 (0.7-1.4)	0.956	1.2 (0.9-1.7)	0.298
Blinding	101 (20.8)	86 (22.9)	1.1 (0.8-1.6)	0.457	1.1 (0.8-1.5)	0.617
Sample size estimation	0 (0)	1 (0.3)	→∞ (0.0-∞)	0.946	NR	
Inclusion of both sexes	43 (8.9)	36 (9.6)	1.1 (0.7-1.7)	0.712	1.1 (0.7-1.7)	0.798
Stroke	*n*=316	*n*=185				
Randomization	120 (38.0)	119 (64.3)	2.9 (2.0-4.3)	<0.0001	3.2 (2.1-4.7)	<0.0001
Blinding	171 (54.1)	144 (77.8)	3.0 (2.0-4.5)	<0.0001	3.0 (2.0-4.5)	<0.0001
Sample size estimation	10 (3.2)	35 (18.9)	7.1 (3.4-14.8)	<0.0001	8.2 (3.7-18.4)	<0.0001
Inclusion of both sexes	15 (4.7)	20 (10.8)	2.4 (1.2-4.9)	0.012	2.4 (1.2-4.9)	<0.0001
ATVB	*n*=476	*n*=401				
Randomization	61 (12.8)	48 (12.0)	0.9 (0.6-1.4)	0.706	0.9 (0.6-1.4)	0.668
Blinding	130 (27.3)	97 (24.2)	0.8 (0.6-1.2)	0.293	0.7 (0.5-1.0)	0.026
Sample size estimation	2 (0.4)	10 (2.5)	6.1 (1.3-27.8)	0.021	NR	
Inclusion of both sexes	72 (15.1)	52 (13.0)	0.8 (0.6-1.2)	0.361	0.8 (0.6-1.3)	0.411

NR: not reported due to small number of events per predictor variable; OR: odds ratio

*Periods 1 and 2 correspond to before and after the date of implementation of the 'Basic Science Checklist' by *Stroke*, respectively

†Adjusted for cardiovascular disease studied and animal model used

Editorial

High Appraisal of Methodological Quality of Basic Science Articles Published in *Stroke*

Wolf-Rüdiger Schäbitz, MD; Jens Minnerup, MD; Marc Fisher, MD; Jaroslaw Aronowski, MD

See related articles, p 2341, p 2632

Translational failure from bench to bedside in medical research may be due, at least in part, to the suboptimal quality of experimental studies. For example, the factors such as lack of blinding and randomization in experimental studies may lead to an overestimation of treatment effects, contributes to irreproducibility of experimental data, and subsequently blocks translation into the clinic.[1] To overcome this dilemma, the Animal Research Reporting In vivo Experiments guidelines provided recommendations for design and reporting of animal experimental studies.[2] *Stroke* was a pioneer in the effort to improve the quality of preclinical research by establishing a requirement for the reporting of methodological quality of animal studies by implementation of the *Basic Science Checklist*, a prerequisite for experimental research later delineated by the landmark article of Landis et al.[3]

In a recent article, Ramirez et al[4] systematically examined the methodology of preclinical studies published in 5 major cardiovascular journals published by the American Heart Association: *Circulation*, *Circulation Research*, *Hypertension*, *Stroke*, and *Arteriosclerosis, Thrombosis, and Vascular Biology*, between July 2006 and June 2016. The authors sought to determine how often studies met basic quality indicators, such as randomization, blinding, and a priori sample size estimation/power calculation, and so forth. Studies reporting in vivo experiments in nonhuman mammals describing pathophysiology, genetics, and therapeutic interventions relevant to specific cardiovascular disorders were considered. Twenty-eight thousand six hundred and thirty-six articles were screened; 3396 met inclusion criteria. Overall, the results showed disappointingly poor study quality of articles published in these journals: overall randomization was reported in only 21.8%, blinding procedures in just 32.7%, and sample size estimation/power calculation was de facto absent (reported in only 2.3%). These numbers remained largely unchanged over the whole observation period of 10 years, except for basic science stroke studies. These studies, 90% of which were published in *Stroke*, show an encouraging and substantial increase in study quality indicators compared with the 4 journals. In the 2015/2016 examination period, randomization in preclinical stroke studies reached 65% versus ≤23% compared with atherosclerosis, hypertension, cardiomyopathy/heart failure; blinding surpassed 80% versus <30%, and sample size estimation/power calculation reached 23% versus <5%, respectively.

These positive findings support an earlier analysis made from basic science articles published in *Stroke* before and after the implementation of the Basic Science Checklist (first introduced in 2011,[5] updated in 2016[6]). The use of the checklist tool led to an increased reporting of inclusion and exclusion criteria definition, allocation concealment, and blinding and showed an increase of studies with the highest quality category of the checklist item sum score compared with the period before where no checklist existed.[5] An interesting question is certainly whether these improvements were because of the journal editorial policy or may even reflect a change in research practice in the basic science stroke community. A post hoc analysis by Ramirez et al[4] came to the conclusion that it is likely a combination of both: *Stroke* uniquely showed improvements in all study design elements even after adjusting for cardiovascular disorder studied and animal model used, but stroke as the cardiovascular disorder studied was identified as an independent predictor for at least one study design element for every journal examined.[4]

Interestingly, but not unexpected, the Ramirez et al[4] report of poor preclinical study quality not to impact on citations at 60 months. Article citation is a most commonly used measure of research impact but is obviously an imperfect indicator of study quality.[7] Despite the superior quality of basic science articles published in *Stroke,* its impact factor ranges below that of the cardiovascular journals, *Circulation* or *Circulation Research*. However, *Stroke* articles belong to the most accessed ones with >11 million downloads in 2016 and an higher article influence score, measuring the high relative importance of the journal on a per-article basis, compared with other neurological and cardiovascular journals.

Beside novelty and innovation, one of the most important issues in real life and in particular in research is quality. Obviously, *Stroke* basic science authors are on the right path, but further quality improvement of preclinical articles is needed. Well performed experimental studies will impact the field, alleviate the translational roadblock, and finally contribute to a better treatment of stroke patients. We anticipate that the recently updated preclinical checklist will lead to further improvements in the quality of preclinical research published by *Stroke* over the next few years.

From the Department of Neurology, EvKB-Bethel, Bielefeld (W.-R.S.), and Department of Neurology (J.M.), University of Münster, Germany; Department of Neurology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA (M.F.); and Department of Neurology, University of Texas Health Science Center, McGovern Medical School, Houston (J.A.).

Correspondence to Wolf-Rüdiger Schäbitz, MD, Department of Neurology, Bethel-EvKB, Burgsteig 13 33617, Bielefeld, Germany. E-mail wolf.schaebitz@evkb.de

(*Stroke*. 2017;48:2337-2338.
DOI: 10.1161/STROKEAHA.117.017717.)

***Stroke* is available at http://stroke.ahajournals.org**
DOI: 10.1161/STROKEAHA.117.017717

Disclosures

None.

References

1. Crossley NA, Sena E, Goehler J, Horn J, van der Worp B, Bath PM, et al. Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach. *Stroke*. 2008;39:929–934. doi: 10.1161/STROKEAHA.107.498725.
2. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol*. 2010;8:e1000412. doi: 10.1371/journal.pbio.1000412.
3. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*. 2012;490:187–191. doi: 10.1038/nature11556.
4. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald ZD, Moreland R, et al. Methodological rigor in preclinical cardiovascular studies: targets to enhance reproducibility and promote research translation. *Circ Res*. 2017;120:1916–1926. doi: 10.1161/CIRCRESAHA.117.310628.
5. Minnerup J, Zentsch V, Schmidt A, Fisher M, Schäbitz WR. Methodological quality of experimental stroke studies published in the Stroke Journal: time trends and effect of the basic science checklist. *Stroke*. 2016;47:267–272. doi: 10.1161/STROKEAHA.115.011695.
6. Vahidy F, Schäbitz WR, Fisher M, Aronowski J. Reporting standards for preclinical studies of stroke therapy. *Stroke*. 2016;47:2435–2438. doi: 10.1161/STROKEAHA.116.013643.
7. Minnerup J, Wersching H, Diederich K, Schilling M, Ringelstein EB, Wellmann J, et al. Methodological quality of preclinical stroke studies is not required for publication in high-impact journals. *J Cereb Blood Flow Metab*. 2010;30:1619–1624. doi: 10.1038/jcbfm.2010.74.

KEY WORDS: Editorials ■ animal models ■ basic science ■ randomization ■ stroke ■ study quality

Stroke
Journal of the American Heart Association

High Appraisal of Methodological Quality of Basic Science Articles Published in *Stroke*

Wolf-Rüdiger Schäbitz, Jens Minnerup, Marc Fisher and Jaroslaw Aronowski

Stroke. 2017;48:2337-2338; originally published online July 27, 2017;
doi: 10.1161/STROKEAHA.117.017717
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 0039-2499. Online ISSN: 1524-4628

Editorial

I'll Have the Rigor, but Hold the Mortis

Steven P. Jones

There is no rigorous definition of rigor

—Morris Kline, 20th century mathematician

Much has been written about a perceived lack of rigor in biomedical research,[1] and this deficiency is implicated in the inability to replicate results from one laboratory in another. Such discussions may often be based more on anecdotal information rather than systematic studies; however, it is becoming accepted that clearer methodologies could benefit the scientific community. More importantly, this perceived lack of rigor supposedly underlies, in part, the failure to translate laboratory findings to the clinic. But what does such rigor look like? There are multiple elements of experimental design that can be affected by rigor, including how a study was designed, conducted, analyzed, and interpreted. Whether and how such elements are made evident in published studies is the subject of Ramirez et al,[2] which stimulates a broader discussion on topics related to transparency, rigor, and reproducibility, all of which could ultimately relate to a common theme: bias.

Article, see p 1916

There are several salient features of Ramirez et al.[2] Not surprisingly, the authors found that 4 of 5 publications they evaluated did not report blinding. If the literature is replete with studies in which blinding was not used for the purpose of promoting experimental bias (which, I hope, it is not), it would erode one's confidence in the literature. Although it is safe to say that many of these unblinded studies were indeed unblinded for unknown reasons, I am not convinced that not reporting blinding equates to not blinding. That is, many authors may not have reported blinding although they incorporated blinding into their study execution. For some investigators, blinding has been a natural but unpublicized practice, like calibrating pH meters, pipets, and analytic balances—all things (some) experimentalists do, but do not tell people. So, what should change? Authors should explicitly report blinding or the lack thereof, and editors and reviewers should ask for and recognize when blinding may or may not make a difference. I think a generalized transition to blinding should happen sooner rather than later because it is a simple and valuable addition to the design of most studies.

There are other issues underlying the barriers to interlaboratory experimental replicability. It should be acknowledged that discrepant results between laboratories should not necessarily mean that at least one of the laboratories engaged in substandard practices (eg, not blinding). Although it is certainly possible that the results of 2 laboratories seem opposite from one another at the level of their conclusions, there is always an explanation. And while sometimes the explanation for the apparent contradiction may involve marginal practices, many times it does not. In fact, a transparent conversation between the 2 laboratories may often identify where one laboratory differed from the other. In a utopian scientific landscape, the discrepant parties could discuss it with one another, perhaps even before publishing their study. Such an approach would limit the inevitable divisions into factions of the devout and the heretical. It is, however, doubtful that collaboration across seemingly religious scientific denominations will soon become common practice. Because many scientists are chasing acceptance by glamor journals, whose focus is on items that could potentially be picked up by the lay press, there is no reward for collaborating with other scientists who may have used a different experimental design. In addition to the absence of reward, it seems that some publishing outlets actually relish in conflict and discord. Thus, some journals (certainly not most) seem to be most interested in grabbing attention, and just like much of the lay media, thrive on controversy, and fights. Such an ecosystem then encourages and enables studies that antagonize one another and provokes scientists into making grand claims about the implications of their study. After all, have not heart disease and cancer already been cured dozens of times, at least according to major media outlets? Ramirez et al[2] ask another important question in their study: Is recent cardiovascular literature improving in reporting of experimental rigor? The answer, simply put, is not really. The authors identify a general exception in the stroke literature, which seems to be driven by proactive decisions at the journal level. The authors also indicate that during the time period of study, the National Institutes of Health (NIH) issued guidance on related topics. Specifically, in 2014, the NIH published its guiding principles about rigor and reproducibility (https://www.nih.gov/research-training/rigor-reproducibility). Although much can happen in 3 years, this is a short time to expect a cultural change in the way scientists conduct (or at least report) studies. Although it is true that there may not have been a significant difference in the way literature has been reported since the NIH's guidance, perhaps such an expectation of causality (ie, between the NIH's 2014 statement and published studies) is premature. There is really little reason to think the 2014 publication of the NIH's principles about rigor and reproducibility would have yet changed what is occurring in the published literature—the time from grant preparation to study completion to publication is not short. I

The opinions expressed in this article are not necessarily those of the editors or of the American Heart Association.

From the Department of Medicine—Cardiovascular, University of Louisville, KY.

Correspondence to Steven P. Jones, PhD, Department of Medicine—Cardiovascular, University of Louisville, 580 S Preston St, Delia Baxter Bldg, Room 321F, Louisville, KY 40202. E-mail Steven.P.Jones@Louisville.edu

(*Circ Res*. 2017;120:1852-1854.
DOI: 10.1161/CIRCRESAHA.117.311114.)

***Circulation Research* is available at http://circres.ahajournals.org**
DOI: 10.1161/CIRCRESAHA.117.311114

agree with the premise—that the NIH's guidance will eventually have a positive impact on methodological rigor—but we have not yet seen the impact of the NIH's engagement of vigilance about these principles. Moreover, journals' lack of coherence or action on such expectations has been slow, and even lip-service provided on the topic is not widespread. It is, therefore, not surprising that there has not been a significant change in methodological reporting, except when mandated by a journal.

Journals—not funding agencies—have the greatest potential for immediate and significant impact on methodological rigor and reporting and by extension reproducibility. Journals must decide that such aspects of a study design/reporting figures prominently into their decision-making algorithm. Also, it is not simply that journals lack initiative in encouraging transparency and rigor. Some journals indirectly contribute to problems with methodological reporting by restricting their page allowances, and Methods are the first section to be cut or relegated to the supplement. Perhaps, the growing glut of journals should show restraint in their publication of an extraordinary number of positive studies at the expense of rejecting negative studies, particularly those performed rigorously. Likewise, scientists should be more careful and transparent in their reporting of methodologies. Unfortunately, many journals will likely continue to bias toward a flashy headline over a carefully conducted study. Thus, some level of reward for experimental rigor, or punition for lack thereof, must occur during the article adjudication process.

Again, the issues triggered by Ramirez et al[2] are important and stimulating on many levels. For example, let us consider the topic of reporting negative studies: is it ethical to continue a study that is, at an interim analysis, deemed futile? Clinical research usually involves an intermittent inspection of the data for safety, overwhelming benefit, or futility, and there are many examples of clinical studies being terminated for these reasons. Thus, should preclinical studies include interim looks at the data to test for futility (which come with a statistical penalty)? And, if the studies are deemed (statistically) to be futile, how would this be reported? Or, should the negativity of the interim assessment be ignored and the study completed? If so, are there ethical implications? These and other questions should be considered thoroughly before anything approaching a blanket mandate is issued.

In this brief space, there are far too many issues to address, particularly related to rigorous conduct of preclinical research. There is, however, another area of general discussion stimulated by the work of Ramirez et al[2]—that is, the translation of preclinical studies to clinical practice. It is safe to say there is no single type of laboratory study. Some investigations may effectively be final tests of an intervention, and these may be conducted in large animal models. In this condition, the most rigorous application of randomized controlled trial (RCT)–like approaches should be applied to increase intrinsic value. In fact, this issue has been addressed in the cardiovascular field with the CAESAR (Consortium for Preclinical Assessment of Cardioprotective Therapies) that had the goal of identifying promising infarct-sparing drugs by enhancing rigor and reproducibility in preclinical studies of infarct size reduction.[3] CAESAR was an example of conducting multisite, RCT-like preclinical studies; however, this type of approach is expensive and cumbersome, and there are other types of studies that do not lend themselves easily to RCT-like approaches. Applying RCT-like standards to studies focused on discovery and elucidating fundamental biological processes may not add the same level of value as applying an RCT-like standard would add to an in vivo study. In fact, placing RCT-like expectations on a discovery study could limit scientific progress because "rigor pushed too far is sure to miss its aim, however good, as the bow snaps that is bent too stiffly" (Friedrich Schiller). Thus, expectations must be appropriately calibrated for the goals of the study. At another time, we can discuss the (ab)use of statistics in preclinical studies, which was also highlighted by Ramirez et al.[2]

Rather than list all of the oft-cited preclinical issues that potentially limit translation to the clinic, consider the share of ownership of this failure from the clinical side. Too often the failure to translate is blamed solely on various preclinical deficiencies du jour—to wit, the present reproducibility crisis. It is potentially curious that preclinical research may be in a crisis because scientists are not reporting methodological details, implementing blinding, and adequately addressing statistical concerns. I am confident that multiple, specific examples of all combinations of such shortcomings exist, but they are not likely the sole cause for clinical failure of all interventions. If we rewind the clock several decades—when, presumably, preclinical work was much more reliable—we see the same deficiencies lamented by those describing a reproducibility crisis. That is, a priori statistical considerations were not made, blinding was not evident, etc. Yet, studies from such a seemingly lax (tongue in cheek) era translated to the clinic more frequently. So, perhaps more contemporary scientists are simply becoming less scrupulous because of temptations of fame, money, and other such corrupting pursuits.

I would like to mention one last element of the barrier of translation from the laboratory to the clinic. There is often ample evidence that interventions deemed ineffective in preclinical studies should also not have worked in the clinical environment. Many interventions taken to clinical trials had preclinical evidence to cast doubt on whether the intervention should have been expected to work. One example where this has been true is with cardioprotection (ie, infarct size reduction). Preclinical studies have shown protection of various agents, some of which were known in preclinical studies not to be effective if, for example, the duration of ischemia was extended significantly. Not surprisingly, their subsequent clinical trials failed to translate, and the preclinical studies were blamed; however, better design of the clinical trials would have allowed a more direct comparison with the preclinical work. The inherent variation and imprecision of the clinical world play a seemingly important, yet poorly defined, role as barriers to translation. A well-executed preclinical study will certainly involve a higher level of consistency, rigorous adherence to a protocol, and reporting of end points than a typical multicenter, phase III clinical trial. And, yet, it could still fail when translated to the clinic. Does this imply that clinical trials are fatally flawed? Of course not. The point here is that treatment of patients differs in heterogeneity and adherence to accepted clinical standards (think phase

IV, community-based experiences), whereas even poorly controlled preclinical studies would be more consistent (in terms of subject to subject variation). Germane to this topic, it is important to note that the majority of preclinical studies do not include risk factors (eg, diabetes mellitus) or polypharmacy—imagine treating heart failure animals, like patients, that is, with β-blockers, angiotensin-converting enzyme inhibitors, and aldosterone antagonists. Until the preclinical and clinical sides move closer to one another, the ownership of the failure to translate remains community property because we are part of a larger continuum in our pursuit of knowledge.

In closing, if there is a crisis in reproducibility, journals share with authors at least part of the responsibility (if you read this sentence, *Circulation Research* must be taking seriously their responsibility, too). It is abundantly clear that much more work and thought are needed to address these important topics discussed here and addressed by Ramirez et al.[2] Unfortunately, it is not clear who is sufficiently interested to pay, in terms of time and money, for such an effort. Unless and until funding bodies, and especially journals, establish clear sets of expectations, the perception of turmoil, uncertainty, and irreproducibility will persist. In the meantime, let us start being more rigorous—it will not kill us.

Sources of Funding

This study was supported by National Institutes of Health.

Disclosures

None.

References

1. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature*. 2016;533:452–454. doi: 10.1038/533452a.
2. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald ZD, Moreland R, Simard T, Clancy AA, Russo JJ, Welch VA, Wells GA, Hibbert B. Methodological rigor in preclinical cardiovascular studies: targets to enhance reproducibility and promote research translation. *Circ Res*. 2017;120:1916–1926. doi: 10.1161/CIRCRESAHA.117.310628.
3. Jones SP, Tang XL, Guo Y, et al. The NHLBI-sponsored Consortium for preclinicAl assESsment of cARdioprotective therapies (CAESAR): a new paradigm for rigorous, accurate, and reproducible evaluation of putative infarct-sparing interventions in mice, rabbits, and pigs. *Circ Res*. 2015;116:572–586. doi: 10.1161/CIRCRESAHA.116.305462.

Key Words: Editorials ■ attention ■ biomedical research ■ National Institutes of Health (U.S.) ■ publications

Circulation Research
JOURNAL OF THE AMERICAN HEART ASSOCIATION

I'll Have the Rigor, but Hold the Mortis

Steven P. Jones

Circ Res. 2017;120:1852-1854
doi: 10.1161/CIRCRESAHA.117.311114
Circulation Research is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 0009-7330. Online ISSN: 1524-4571

American Heart Association | American Stroke Association

The Next Best Thing in Cardiovascular Research

Highlights From the Basic Cardiovascular Sciences Scientific Sessions

Ryan L. Boudreau, PhD

"What will be the next best thing in cardiovascular research?" This was the loaded question recently posed in a panel discussion at the American Heart Association Council on Basic Cardiovascular Sciences (BCVS) 2017 Scientific Sessions, held in Portland, OR, from July 10 to 13. The panelists made clear the challenge to predict the next major breakthrough, with most scientists in the room hoping it would evolve from their own research endeavors. Despite our inability to predict the future and the notion that most substantial discoveries are made serendipitously, several notable highlights in the conference programming reiterated the paths that will inevitably lead to the next greatest research successes in the field (Figure).

Diverse and Multidisciplinary Studies

The meeting encompassed >400 scientific abstracts on an array of research topics, including cardiac fibrosis, oxidative stress, metabolism, arrhythmias, heart failure signaling, RNA biology, and novel therapies, with many studies spanning and integrating multiple disciplines involving molecular, genetic, and physiological methodologies. A number of important model systems were highlighted, including advances in human induced pluripotent stem cell–derived cardiomyocytes, new models of heart failure with preserved ejection fraction—an area of emerging interest considering the increasing prevalence of this morbid condition—and genetic-based minipig models of hypertrophic cardiomyopathy. Clearly, the field's interest in improving "bench-to-bedside" research continues to grow. This diversity in approaches and concepts signifies that our research community continues to cast broad nets in search of the next best thing, and panelists noted this as positive, acknowledging that many research avenues show promise but no single one shows enough to justify major refocusing of our conglomerate scientific efforts. Instead, continued collaboration and engagement is needed to understand how our broad research efforts interrelate and fit into the big picture.

Expanding "-Omics" and "Big Data" Integration Into Research

The field has made great strides in incorporating large-scale approaches and data sets along with advanced computational analyses to deeply interrogate cardiovascular disease. Heaps of data of varying kinds are pouring out of an array of biological systems, including expansive patient networks and animal and cellular models in the laboratory (relevant reviews[1,2]). Despite significant advances in defining genetic and environmental risks, cellular pathways of interest, and more thoroughly characterizing the heart's "-omics" (eg, transcriptomes, proteomes, metabolomes, and interactomes), sustained efforts are needed to continually expand the application of "big data" and related technologies more broadly throughout our strong base of basic science laboratories. On this front, meeting organizers arranged several new focused workshops to convey the power of integrating high-content -omics methodologies and bioinformatics into wet lab research. Perhaps one of the next best things in cardiovascular research will be the emergence of "hybrid" investigators with both wet lab and computational skills. With an understanding of the biological concepts, these scientists will open new doors by learning to program (eg, scripting in Perl or R languages), allowing them to efficiently manipulate, analyze, and integrate the breadth of big data, which can point to new hypotheses that they themselves can subsequently test in wet lab experiments. Indeed, many of the next

From the Department of Internal Medicine, University of Iowa College of Medicine, Iowa City, IA.

Correspondence to: Ryan L. Boudreau, PhD, Department of Internal Medicine, University of Iowa College of Medicine, Iowa City, IA 52242. E-mail: ryan-boudreau@uiowa.edu

J Am Heart Assoc. 2017;6:e007245. DOI: 10.1161/JAHA.117.007245.

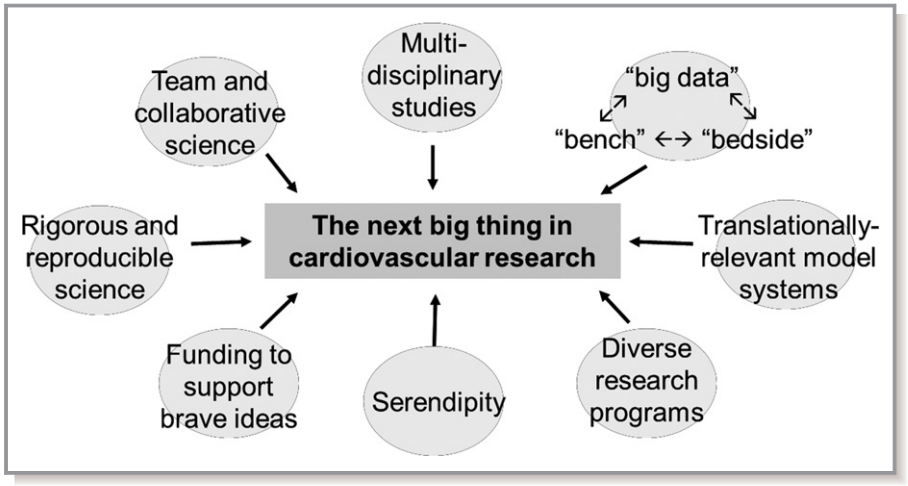


Figure. Schematic overview of key factors that will contribute to major discoveries in cardiovascular research.

greatest discoveries will likely result from effective scientific integration and translation across bedside, big data, and bench themes.

Scientific Rigor, Reproducibility, and Transparency

The conference also presented a special session on "Ethics, Open Science, Reproducibility, Unconscious Bias and All That Jazz," touching on important topics related to recent National Institutes of Health initiatives aimed at improving the overall quality of scientific research. A recent study published in *Circulation Research* revealed that a significant lack of methodological rigor (eg, not randomizing participants, blinding analyses, or adequately determining group sizes) is likely at the root of the field's plague of irreproducible data.[3] Considering the detrimental effects this has on science and the public's view of scientific research, investigators ardently need to take steps in continuing to discuss and work toward resolving problems of this nature, as was achieved in this session. The goal is to minimize the risk of building "houses of cards" based on weak and unreliable experimental evidence, while moving toward building scientific foundations on which the next best thing in cardiovascular research can be realized.

Brave Ideas

The BCVS 2017 Scientific Sessions keynote lecture was delivered by Dr Calum MacRae, who gave an overview of his team's strategy for moving toward the eradication of coronary heart disease. Last year, Dr MacRae and his collaborative network successfully competed for the One Brave Idea award (sponsored by the American Heart Association and Verily, with support from AstraZeneca), earning an impressive $75 million to conduct a multiphase exploration of patient cohorts and big data to identify critical genetic, cellular, and molecular factors that would allow future identification of patients at high risk of developing coronary heart disease later in life. Although most scientists in the room were dreaming of such funding, several agreed with Dr MacRae that this was "not enough" considering the scope of the studies and the breadth of the research team needed to tackle them. Dr MacRae noted that his intention is to further spread the resources among the research community to ensure that the ultimate benefit is bestowed on coronary heart disease patients. Only time will tell whether this unprecedented program and research endeavor will yield the next best thing in cardiovascular research. Nonetheless, the spirit of this award encourages others to think boldly and creatively about prevalent morbidities, to focus on patients, and to hope for similar future large-scale funding ventures to prevent these disease from devastating society.

In summary, the BCVS 2017 Scientific Sessions were marked by considerable enthusiasm and highlighted many exciting new findings and approaches in basic science and translational research, along with special-topic workshops and sessions that offered insight into some core components that will contribute to significant future breakthroughs in the field. Although nobody could confidently say what the next best thing in cardiovascular research will be, perhaps the answer for attendees is quite simply, "the upcoming BCVS Scientific Sessions."

Disclosures

None.

References

1. Scruggs SB, Watson K, Su AI, Hermjakob H, Yates JR III, Lindsey ML, Ping P. Harnessing the heart of big data. *Circ Res*. 2015;116:1115–1119.
2. Kim J. Big data, health informatics, and the future of cardiovascular medicine. *J Am Coll Cardiol*. 2017;69:899–902.
3. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald ZD, Moreland R, Simard T, Clancy AA, Russo JJ, Welch VA, Wells GA, Hibbert B. Methodological rigor in preclinical cardiovascular studies: targets to enhance reproducibility and promote research translation. *Circ Res*. 2017;120:1916–1926.

Key Words: basic science • big data • cardiovascular disease • multidisciplinary • One Brave Idea • rigor • translational science

APPENDIX E

Preclinical Studies Don't Regularly Adhere to Best Practices

Animal experiments published in a handful of cardiovascular journals mostly ignore NIH guidelines.

May 8, 2017
KERRY GRENS

1

WIKIMEDIA, RAMA

In 2014, the National Institutes of Health established guidelines for preclinical experimental design—hoping to encourage researchers to adopt best practices, such as randomization and the inclusion of both sexes of lab animals. Yet, in the years preceding and since, the majority of papers published in several cardiovascular journals show a widespread disregard for these standards, according to a recent analysis published in *Circulation Research*.

"If we're falling down at this early stage, we have very little hope of having a good translation rate," said Benjamin Hibbert, a cardiologist and researcher at the University of Ottawa who led the study.

The notable exception to the trend, Hibbert and his colleagues found, was one journal, *Stroke,* which saw a marked increase in the adherence to best practices following the implementation of a Basic Science Checklist for authors. The checklist, which was introduced in 2011 and updated last year, asks for

Checklist for authors. The checklist, which was introduced in 2011 and updated last year, asks for roughly the same standards as NIH's guidelines: randomization, blinding, sample size estimation, and explanations of which animals were included, among other expectations.

"It shows that [our] required preclinical checklist for treatment related animal experiments have improved the quality of our preclinical papers," Marc Fisher, *Stroke*'s editor-in-chief who developed the checklist, wrote in an email to *The Scientist*, "but we are still trying to make them even better."

In light of concerns about reproducibility and the quality of animal studies, Hibbert's team wanted to see how preclinical study design had changed in recent years. They gathered nearly 4,000 papers published from 2006 to 2016 in five journals under the umbrella of the American Heart Association: *Circulation, Circulation Research, Hypertension, Stroke,* and *Arteriosclerosis, Thrombosis, and Vascular Biology.* All of the papers included in the analysis reported the results of in vivo interventions on animals. Hibbert and his coauthors then picked through each paper to see how it complied with the four standards: randomization of animals, having a researcher blinded to the treatment groups, sample size estimation, and the inclusion of both sexes of model organism.

Of these four criteria, only sample size estimation became more common. But, as the researchers pointed out, it was reported in fewer than 7 percent of papers by the end of the decade of publishing they considered. Although it didn't change over time, blinding was more commonly reported than sample size estimation—in more than 35 percent of papers in 2016—and randomization was present in more than 25 percent of papers in 2016.

As far as the sex of the animals included, Hibbert and colleagues found 80 percent of papers disclosed the sex, with 71.6 percent of those studies using males only. Around 13 percent used only female animals, and roughly 15 percent used both (they published these data separately in February in *Circulation*).

"I thought, given the progress made in clinical trials [regarding the inclusion of females], that would percolate into preclinical researcher circles," said lead author Daniel Ramirez, a research fellow at the Ottawa Heart Institute. "But there was no signal we could find that males are being used exclusively less so than before. If anything it's increasing."

"This paper points out a number of methodological problems in the preclinical cardiovascular literature in various journals, including *Circulation Research*. The fact that we elected to publish this article is an eloquent testimony to the seriousness of our concern regarding rigor in preclinical research," Roberto Bolli, the editor-in-chief of *Circulation Research,* wrote in an email to *The Scientist*. He added that his journal is taking steps to boost the scientific rigor of the papers it publishes, including laying out expectations for experimental design.

Nathalie Percie du Sert, the experimental design program manager for the National Centre for Replacement, Refinement, and Reduction of Animals in Research (NC3Rs) based in London, said she is not surprised by the results. "Researchers don't realize that this can have an impact on their results, so they don't know enough how important those measures are."

Except for the inclusion of both sexes, *Stroke* outperformed other journals in the percentage of papers complying with the preclinical standards, and adherence became even better after the 2011 checklist rolled out. After circulating the checklist, for instance, 64 percent of papers randomized animals, nearly

rolled out. After circulating the checklist, for instance, 64 percent of papers randomized animals, nearly 78 percent included blinding of their treatment allocation, and close to 19 percent reported sample size estimation. That was up from roughly 38 percent, 54 percent, and 3 percent, respectively.

"It's a credit to *Stroke*," said Hibbert. "Implementing that checklist and incentivizing properly is the only way you're going to get people to comply."

It's also a credit to the stroke research community at large, noted Percie du Sert, who said the stroke field has been leading the charge for better preclinical design. Her organization has also been working to develop standards for proper experimental design. NC3Rs established the ARRIVE guidelines (for Animal Research: Reporting In Vivo Experiments) in 2010, and since then numerous journals have endorsed them. NC3Rs is currently funding a trial to see if requiring authors to fill out an ARRIVE checklist upon submitting a manuscript can effect positive change. And the organization has developed an online tool to help guide scientists through study design.

While Hibbert said he was encouraged by *Stroke*'s progress, one result was particularly discouraging: the citations given to papers of sub-par experimental design. He and his colleagues found no relationship between citations and papers that either adhered to the guidelines or didn't.

"We naively thought that people who are doing better science must be getting rewarded," said Hibbert. "It has no impact. People don't recognize high quality and low quality science."

F.D. Ramirez et al., "Methodological rigor in preclinical cardiovascular studies: Targets to enhance reproducibility and promote research translation," *Circ Res,* doi: 10.1161/CIRCRESAHA.117.310628, 2017.

Keywords:

animal research, cardiovascular disease, experimental design, NIH, preclinical, research integrity, stroke, waste

Editorial

New Initiatives to Improve the Rigor and Reproducibility of Articles Published in *Circulation Research*

Roberto Bolli

The rampant irreproducibility of scientific articles is arguably the most serious problem that bedevils biomedical research.[1–12] It is a veritable plague that undermines the credibility of published studies, hinders clinical translation of basic work, and impedes the progress of medicine. Disquietingly, the problem seems to be getting worse rather than better.[10] Although the causes of irreproducibility are multifarious, inadequate rigor is probably the most important[2,11]; thus, it will be impossible to augment reproducibility without augmenting methodological rigor. It is not the purpose of this editorial to revisit the nature, origins, mechanisms, and consequences of irreproducibility, all of which have been discussed innumerable times in recent years, both in the literature and in ad hoc workshops.[1–12] There has been enough discussion; now it is time for action.

Much of this action must come from editors. Publication of methodologically flawed or suboptimal research can be limited by promulgating, and diligently applying, higher standards during editorial evaluation of submitted work. The editors of *Circulation Research* believe that the journal has a responsibility to implement initiatives that promote more rigorous and, therefore, more reproducible scientific work. It is for this reason that *Circulation Research* has published several reviews, editorials, Viewpoints, and News & Views articles focused on this issue.[1,2,7–12] It is for this reason that the journal has joined other leading journals to promote reproducibility of biomedical research[4] by endorsing the National Institutes of Health principles and guidelines for reporting preclinical research,[6] and it is for this very reason that we have recently chosen to publish a study showing that only a minority of preclinical animal experiments reported in leading cardiovascular journals (including *Circulation Research*) adheres to basic criteria necessary to assure rigor: for example, it was found that during a 10-year span, blinding was reported in 33% of the articles examined, randomization in 22%, and a priori sample size calculation in 2%.[7] The most troubling aspect of this report is that, with the exception of stroke research, there has been no improvement in the level of rigor in the past decade.[7]

The purpose of this editorial is to announce new initiatives aimed at enhancing the rigor, transparency, and reproducibility—and, thus, the overall robustness—of articles published in our journal. Three distinct initiatives will be implemented: checklists for rigor, measures to improve transparency, and a policy on refutations.

Checklists for Rigor

As of October 1, 2017, *Circulation Research* will require that authors fill out checklists to document the methodological rigor of preclinical work submitted for publication. Different checklists will be used for studies in animals and for studies in vitro because the challenges are different in the 2 settings. Although methodological weaknesses plague all domains of research, they seem to be particularly conspicuous in animal studies.[2,11,13] Accordingly, major emphasis in the checklists will be given to the methodology and reporting of investigations in animals.

Checklists for Studies in Animals

Authors will be asked to complete 2 checklists: a short one on initial submission and a longer one if and when the article is resubmitted in revised form. These documents are modifications of similar documents that were recently implemented by *Stroke*[3] and have resulted in improved adherence to standards of rigor.[7,14] The short checklist (Table 1) is relatively simple and covers the basic methodological aspects that are considered essential to the validity of the work: description of animals used, use of randomization, use of blinding, specification of inclusion and exclusion criteria and whether such criteria were established a priori, reporting of animals excluded, and description of statistical methods. The long checklist (Table 2) is much more detailed, providing information that enables in-depth evaluation of the level of rigor (and, therefore, robustness) of the study.

The short checklist will be used by reviewers and editors to evaluate the article but will not be published. The long checklist will be published as an online supplement along with the article; its purpose is to assist not only reviewers and editors, but also readers, to fully evaluate the extent to which the work adheres to standards of rigor. By examining the long checklist, readers will be able to assess for themselves the strengths and weaknesses of the study and how robust the data and conclusions are. The long checklist is accompanied by an explanatory document (Table 3), modeled after that used by *Stroke*,[3,14] which clarifies the standards of rigor outlined in the checklist and assists the authors in reporting them. We have elected to use 2 checklists because most articles submitted to *Circulation Research* are rejected at first submission; thus, requiring detailed information for every article that is submitted would be unnecessarily burdensome for the authors.

Checklist for Studies In Vitro

For molecular and cellular studies performed in vitro, authors will be required to fill out only one checklist if and when the

The opinions expressed in this article are not necessarily those of the editors or of the American Heart Association.

From the Institute of Molecular Cardiology, University of Louisville, KY.
Correspondence to Roberto Bolli, MD, Institute of Molecular Cardiology, University of Louisville, ACB Third Floor, 550 S, Jackson St, Louisville, KY 40202. E-mail rbolli@louisville.edu

(*Circ Res*. 2017;121:472-479.
DOI: 10.1161/CIRCRESAHA.117.311678.)

***Circulation Research* is available at http://circres.ahajournals.org**
DOI: 10.1161/CIRCRESAHA.117.311678

article is resubmitted in a revised form. This document (checklist for in vitro studies; Table 4) will be provided as a downloadable PDF file on the journal webpage and will cover the fundamental aspects that determine the rigor and robustness of the work, including technical replicates, characteristics of the antibodies and cell lines used, data processing and conversions, presentation of Western blots and immunofluorescence images, etc. This checklist will not be published, but authors will be asked to include the corresponding information in the text or in an online supplement. As in the case of the animal studies, the checklist will be accompanied by an explanatory document (Table 5), modeled after that used by *Stroke*,[3,14] which clarifies the standards of rigor outlined in the checklist and assists the authors in reporting them.

Although the new checklists clearly enunciate the desired standards of rigor, it is important to stress that failure to comply with some of the criteria outlined in these lists will not automatically or necessarily lead to rejection. Each study is different, and the weight that rigor has in the editorial decision must be individualized for each article. In the case of animal work, rigor is particularly important in validation studies—studies designed to verify the validity of a hypothesis in vivo (eg, whether a type of cell therapy, previously suggested to alleviate left ventricular dysfunction in exploratory studies, does indeed have such an effect). In contrast, some of the standards outlined in Tables 1 and 2 may not apply to exploratory, hypothesis-generating, or mechanistic animal studies (ie, studies where the primary goal is to explore the effect of an intervention in vivo for the first time, to obtain initial evidence in support of a hypothesis, or to elucidate a molecular or cellular mechanism). There are also gray zones. Many preclinical studies contain a combination of in vitro and in vivo data; the importance of the in vivo component varies greatly, and in some cases, these data may be peripheral to the main goal of the study. Thus, decisions on acceptance or rejection will continue to be made on a case-by-case basis depending on the specific content and features of each submitted article. If adherence to particular standards outlined in Tables 1, 2, and 4 was not appropriate or meaningful, we encourage the authors to address these issues in the text and to describe why this was the case.

Measures to Improve Transparency

In addition to the checklists, and in keeping with the goal of enhancing transparency,[6] we ask authors to use online supplements to provide detailed experimental protocols and communicate any secondary results not reported in the main article, both for in vivo and in vitro studies. As stated previously on these pages,[2,15] we expect all articles to describe the methods in sufficient detail to enable others to reproduce the work; this is usually achieved by publishing a detailed Methods section online. As recommended by the National Institutes of Health guidelines, authors should use online data supplements to describe "any similar experimental results that were omitted from the reporting for any reason, especially if the results do not support the main findings of the study" and "any outcomes or conditions that were measured or used and are not reported in the Results section."[6]

Policy on Refutations

Circulation Research also adheres to the National Institutes of Health recommendation that if an article is published in our journal, the journal "assumes responsibility to consider publication of refutations of that article, according to usual standards of quality."[6] In other words, an article that contradicts previous findings will not be penalized for not being the first; if the article is of quality consistent with the standards of *Circulation Research*, it will receive the same priority as if it were the first.

Other Considerations

The CAESAR Experience

An excellent example of the importance of methodological rigor can be found in the field of cardioprotection (limitation of myocardial infarct size), where thousands of therapies have been claimed to reduce infarct size during the past 5 decades, yet few have been reproducibly effective and (with the exception of early reperfusion) none has been translated into a clinical therapy.[11] The major reason for this colossal fiasco has been insufficient methodological rigor and transparency.[11] This problem was addressed in a workshop organized by the NHLBI in June 2003, which resulted in the implementation of a multicenter consortium (CAESAR) that conducted studies of putative cardioprotective therapies with a level of rigor comparable with that of randomized clinical trials; for example, with the use of randomization, blinding, a priori sample size and power calculation, etc.[16,17] When these methods were implemented, the CAESAR investigators found that different laboratories using the same protocols were able to obtain similar results.[16] It was also found that several therapies previously claimed to reduce infarct size were reproducibly ineffective in 3 different species, suggesting that an increased level of rigor eliminated false-positive results.[18,19]

Impact of the New Initiatives

The new checklists, transparency criteria, and refutations policy introduced herein are the most important changes in the editorial policies of *Circulation Research* in many decades. Their consequences will be profound and far reaching. It will take several years for their full impact to be realized, but this impact will be, I think, positive.

By promoting methodological rigor and transparency, these new guidelines will further enhance the already high quality of the work published in our journal, thereby augmenting its reproducibility and scientific utility. I expect that this will result in fewer controversies; therefore, less time, money, and effort will be consumed to resolve apparent conflicts in the literature. As fewer biased or inaccurate conclusions will be published, the number of positive studies will decrease dramatically. This, in turn, will enable investigators to focus on those ideas or hypotheses that are true (or likely to be true) rather than try to reproduce those that are not, which will speed up progress. Translation of preclinical results to clinical therapies will be facilitated because investigators will concentrate their efforts on those (few) interventions that are reproducibly effective in animal models and, thus, are more likely to be effective in patients. Importantly, there will be a moratorium, or possibly even a reversal, of the erosion of the credibility of research caused by irreproducible results.

Change is often difficult, and it seems likely that the guidelines enunciated in this article will meet with some opposition.

For example, it will be argued that they constitute yet another burden on authors in our overregulated world. Although I personally agree that our society is overregulated and that excessive paperwork and regulations are oppressive and can stifle creativity, I am more concerned about publishing conclusions that may be biased or inaccurate. The damage caused by the introduction of erroneous ideas in the literature greatly outweighs the extra effort necessary to comply with the principles that underlie the checklists.

An unintended consequence may be that these new checklists will discourage authors from submitting articles to *Circulation Research*. It is hoped that this will not happen, but it is a risk that must be taken. The alternative (publishing conclusions that may be inaccurate) is much worse. We have made a deliberate effort to minimize the burden on authors. For example, the initial checklist for animal studies (Table 1) is quite simple. Only when an article is seriously considered for publication is the long, more burdensome checklist (Table 2) required.

Time-Table for Implementation

To give authors time to reformat their papers in accordance with the new guidelines, the checklists will not be required until October 1, 2017.

What will happen then? After the launch of the initiatives outlined herein, we do not foresee a sudden spike in the level of rigor of the papers that we publish, but rather, a gradual increase. Manuscripts submitted to *Circulation Research* cannot be expected to conform with the new guidelines overnight. Obviously, studies that have already been completed cannot be modified, and changing the design and conduct of ongoing or planned experiments will take time. In many cases, 2-3 years elapse from the time when a study is first conceived to its submission for publication to our journal. Accordingly, in the short term we do not expect that, to be acceptable for publication, papers will have to meet all or almost all of the criteria specified in the checklists. We do hope, however, that these criteria will motivate investigators to change their practice in order to achieve a higher level of rigor, such that, in time, the rigor criteria encapsulated in these checklists (or at least most of them) will be met.

If this happens, it will be a slow process; it would be unrealistic to assume otherwise. The important thing, in our opinion, is to start this process. While our ultimate goal is to augment rigor and reproducibility in those papers that are published in *Circulation Research*, our more immediate objective is to promote a shift in experimental procedures toward practices that are conducive to a higher level of scientific validity. We will not be surprised if, over the next several months, we will continue to receive and publish manuscripts that do not meet all the criteria listed in the checklists. We will, however, be disappointed if, a few years from now, the level of rigor of studies published in *Circulation Research* has not improved.

Enhancing rigor, important though it may be, is not the only motivation for our actions. Even if a study does not meet the rigor criteria listed herein, the checklists will be helpful to the readers because they will allow them to see at a glance how much a study deviated from the ideal level of rigor and, therefore, how robust its conclusions are. An additional immediate benefit of the checklists is that they will spur authors to better report the details of their studies, which will enhance reproducibility irrespective of the level of rigor. Even if the experimental procedures do not change, a more complete, organized, and detailed description of the experiment will make it easier to either reproduce or refute it.

Concluding Remarks

The editors of *Circulation Research* are committed to upholding the highest standards of rigor and responsibility in both the conduct and the reporting of preclinical research. We hope that the new checklists and transparency criteria will clearly communicate the desired standards of rigor to authors and investigators while facilitating the journal's overall mission of promoting robust, rigorous, and reproducible scientific inquiry and experimentation.

Although adding more requirements for publication of articles in *Circulation Research* is not desirable, in the final analysis, the editors do not have a choice. Publication of studies that are less than rigorous (and, thus, inconclusive or possibly even misleading) does not benefit anyone, nor does dissemination of concepts that may be biased or inaccurate. As Ramirez et al[7] have pointed out, there has been no progress in rigor and transparency during the past decade, even after the promulgation of the National Institutes of Health guidelines[6] and the publication of many high-profile articles on the topic. The status quo is not acceptable. If we, as editors and gate keepers of scientific reporting, do not take action to remedy this problem, who will?

In conclusion, I think the initiatives outlined herein will enhance the quality of research published in *Circulation Research* and facilitate progress in cardiovascular biology. At present, most animal studies do not meet all of the criteria of rigor listed in Tables 1 and 2. In time, however, I think that the principles that underlie these guidelines will permeate the scientific community and become ingrained in the design of cardiovascular research, to the point that complying with the criteria articulated in the checklists will come natural. It will take time, but I think it will happen. When that happens, there will be less hype, less glamor, and probably less media headlines, but there will also be more truth. And that—the pursuit of the truth—is the core mission of any scientific journal.

Disclosures

None.

References

1. Bolli R. Dandum semper est tempus: the crucial importance of (and increasing disregard for) the test of time. *Circ Res*. 2015;117:755–757. doi: 10.1161/CIRCRESAHA.115.307613.
2. Bolli R. Reflections on the irreproducibility of scientific papers. *Circ Res*. 2015;117:665–666. doi: 10.1161/CIRCRESAHA.115.307496.
3. Vahidy F, Schäbitz WR, Fisher M, Aronowski J. Reporting standards for preclinical studies of stroke therapy. *Stroke*. 2016;47:2435–2438. doi: 10.1161/STROKEAHA.116.013643.
4. Journals unite for reproducibility. *Nature*. 2014;515–517. doi:10.1038/515007a.
5. Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, Fry D, Hutton J, Altman DG. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One*. 2009;4:e7824. doi: 10.1371/journal.pone.0007824.
6. NIH. NIH Principles and Guidelines for Reporting Preclinical Research. https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research Accessed June 1, 2017.
7. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald ZD, Moreland R, Simard T, Clancy AA, Russo JJ, Welch VA, Wells GA,

Hibbert B. Methodological rigor in preclinical cardiovascular studies: targets to enhance reproducibility and promote research translation. *Circ Res*. 2017;120:1916–1926. doi: 10.1161/CIRCRESAHA.117.310628.
8. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res*. 2015;116:116–126. doi: 10.1161/CIRCRESAHA.114.303819.
9. Demer LL, Tintut Y, Nguyen KL, Hsiai T, Lee JT. Rigor and reproducibility in analysis of vascular calcification. *Circ Res*. 2017;120:1240–1242. doi: 10.1161/CIRCRESAHA.116.310326.
10. Williams R. Can't get no reproduction: leading researchers discuss the problem of irreproducible results. *Circ Res*. 2015;117:667–670. doi: 10.1161/CIRCRESAHA.115.307532.
11. Bolli R, Becker L, Gross G, Mentzer R Jr, Balshaw D, Lathrop DA; NHLBI Working Group on the Translation of Therapies for Protecting the Heart from Ischemia. Myocardial protection at a crossroads: the need for translation into clinical therapy. *Circ Res*. 2004;95:125–134. doi: 10.1161/01.RES.0000137171.97172.d7.
12. Jones SP. I'll have the rigor, but hold the mortis. *Circ Res*. 2017;120:1852–1854. doi: 10.1161/CIRCRESAHA.117.311114.
13. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the arrive guidelines for reporting animal research. *PLoS Biol*. 2010;8:e1000412.
14. Minnerup J, Zentsch V, Schmidt A, Fisher M, Schäbitz WR. Methodological quality of experimental stroke studies published in the stroke journal: time trends and effect of the basic science checklist. *Stroke*. 2016;47:267–272. doi: 10.1161/STROKEAHA.115.011695.
15. Bolli R. The new circulation research: a manifesto. *Circ Res*. 2010;106:216–226. doi: 10.1161/RES.0b013e3181d39f39.
16. Jones SP, Tang XL, Guo Y, et al. The NHLBI-sponsored consortium for preclinical assessment of cardioprotective therapies (CAESAR): a new paradigm for rigorous, accurate, and reproducible evaluation of putative infarct-sparing interventions in mice, rabbits, and pigs. *Circ Res*. 2015;116:572–586. doi: 10.1161/CIRCRESAHA.116.305462.
17. Lefer DJ, Bolli R. Development of an NIH consortium for preclinical assessment of cardioprotective therapies (CAESAR): a paradigm shift in studies of infarct size limitation. *J Cardiovasc Pharmacol Ther*. 2011;16:332–339. doi: 10.1177/1074248411414155.
18. Kukreja R, Tang XL, Lefer D, et al. Administration of sildenafil at reperfusion fails to reduce infarct size: results from the CAESAR cardioprotection consortium (LB650). *The FASEB Journal*. 2014;28(1 suppl). http://www.fasebj.org/content/28/1_Supplement/LB650.short.
19. Lefer D, Jones S, Steenbergen C, et al. Sodium nitrite fails to limit myocardial infarct size: results from the CAESAR cardioprotection consortium (LB645). *The FASEB Journal*. 2014;28(1 suppl). http://www.fasebj.org/content/28/1_Supplement/LB645.short.

KEY WORDS: Editorials ■ biomedical research ■ checklist ■ sample size ■ stroke

Table 1. Short (Initial) Checklist of Methodological and Reporting Aspects for Articles Submitted to *Circulation Research* That Report Studies of Experimental Interventions in Animals

Preclinical testing	Prevention of bias is important for experimental cardiovascular research. For studies where the primary objective is the preclinical testing of therapies, the following checklist items must be adhered to and clearly presented in the article. Study involves testing of therapeutic or diagnostic agent in animal models: Yes/No
Animals	Species, age, sex, strains, and sources of animals are described: yes/no
Randomization	Randomization and allocation concealment were performed: yes/no
Blinding	Blinding was performed: yes/no
Inclusions and exclusions	Specific criteria for inclusions and exclusions are specified: yes/no
	Criteria for inclusions and exclusions were set before the study: yes/no
Reporting of excluded animals	All animals excluded after randomization are reported: yes/no
Statistical methods	Statistical methods are described: yes/no

Table 2. Checklist of Methodological and Reporting Aspects for Articles Submitted to *Circulation Research* That Report Studies of Experimental Interventions in Animals

Methodological and Reporting Aspects	Description of Procedures
Study design	□ The experimental group(s) have been clearly defined in the article, including number of animals in each experimental arm of the study.
	□ An overall study timeline is provided.
	□ The protocol was prospectively written.
	□ The primary and secondary end points are specified.
	□ For primary end points, a description is provided as to how the type I error multiplicity issue was addressed (eg, correction for multiple comparisons was or was not used and why). (Note: correction for multiple comparisons is not necessary if the study was exploratory or hypothesis generating in nature).
	□ A description of the control group is provided including whether it matched the treated groups.
Inclusion and exclusion criteria	□ Inclusion and exclusion criteria for enrollment into the study were defined and are reported in the article.
	□ These criteria were set a priori (before commencing the study).
Randomization	□ Animals were randomly assigned to the experimental groups. If random assignment was not used, adequate explanation has been provided.
	□ Type and methods of randomization have been described.
	□ Allocation concealment was used.
	□ Methods used for allocation concealment have been reported.

(*Continued*)

Table 2. Continued

Methodological and Reporting Aspects	Description of Procedures
Blinding	□ Blinding procedures with regard to masking of group/treatment assignment from the experimenter were used and are described. The rationale for nonblinding of the experimenter has been provided, if such was not performed.
	□ Blinding procedures with regard to masking of group assignment during outcome assessment were used and are described.
	□ If blinding was not performed, the rationale for nonblinding of the people analyzing outcome has been provided.
Sample size and power calculations	□ Formal sample size and power calculations were conducted before commencing the study based on a priori determined outcome(s) and treatment effect(s), and the data are reported.
	□ If formal sample size and power calculation was not conducted, a rationale has been provided.
Data reporting	□ Baseline characteristics (species, sex, age, strain, chow, bedding, and source) of animals are reported.
	□ The number of animals in each group that were randomized, tested, and excluded and died is reported. If the experimentation involves repeated measurements, the number of animals assessed at each time point is provided for all experimental groups.
	□ Baseline data on assessed outcome(s) for all experimental groups are reported.
	□ Details on important adverse events and death of animals during the course of the experiment are reported for all experimental groups.
	□ Numeric data on outcomes are provided in the text or in a tabular format in the main article or as supplementary tables, in addition to the figures.
	□ To the extent possible, data are reported as dot plots as opposed to bar graphs, especially for small sample size groups.
	□ In the online supplemental material, methods are described in sufficient detail to enable full replication of the study.
Statistical methods	□ The statistical methods used for each data set are described.
	□ For each statistical test, the effect size with its standard error and *P* value is presented. Authors are encouraged to provide 95% confidence intervals for important comparisons.
	□ Central tendency and dispersion of the data are examined, particularly for small data sets.
	□ Nonparametric tests are used for data that are not normally distributed.
	□ 2-sided *P* values are used.
	□ In studies that are not exploratory or hypothesis generating in nature, corrections for multiple hypotheses testing and multiple comparisons are performed.
	□ In negative studies or null findings, the probability of a type II error is reported.
Experimental details, ethics, and funding statements	□ Details on experimentation, including formulation and dosage of therapeutic agent, site and route of administration, use of anesthesia and analgesia, temperature control during experimentation, and postprocedural monitoring, are described.
	□ Both male and female animals have been used. If not, the reason/justification is provided.
	□ Statements on approval by ethics boards and ethical conduct of studies are provided.
	□ Statements on funding and conflicts of interests are provided.

The above checklist is a modified version of a similar checklist promulgated by *Stroke* in 2011 and revised and expanded in 2016. The purpose of the checklist is to provide important metrics for the conduct and reporting of preclinical studies that involve assessment of experimental interventions in vivo. The criteria listed herewith are intended to (i) clearly communicate the desired standards of rigor to authors and investigators and (ii) facilitate the journal's overall mission of promoting robust, rigorous, and reproducible scientific inquiry and experimentation. We recognize the complexity and diversity of many studies, and the fact that different studies have different levels of emphasis on in vivo data. The information requested herein is meant to be used by reviewers and editors as an aid in assessing the value of the work, not as a rigid criterion for accepting or rejecting articles; thus, decisions on acceptance or rejection will be made on a case-by-case basis depending on the individual content and characteristics of each submitted article. In addition, we recognize that certain standards outlined above may not apply to exploratory, hypothesis-generating, or mechanistic animal studies. We encourage and welcome such studies; however, in these cases we encourage the authors to address the areas outlined here and describe why adherence to particular standards was not practical or meaningful. We promote the use of online supplements for providing detailed experimental protocols and communicating any secondary results not reported in the main article.

Table 3. Explanatory Accompaniment to the Checklist for Methodological and Reporting Features of Studies Involving Experimental Interventions in Animals

Study design	The protocol should be written before commencing the study. If the study is exploratory (eg, the first test of a new therapy or manipulation in vivo), or if it is primarily mechanistic (ie, its primary purpose is to elucidate mechanisms rather than establish efficacy), definition of primary and secondary end points is not necessary. However, if the study is confirmatory (ie, its purpose is to validate the effect of a therapy or intervention), primary and secondary end points should be defined and reported. Furthermore, if the study is confirmatory and if multiple groups are compared with respect to the primary end point, corrections for a type I error should be used.
Inclusion and exclusion criteria	A priori defined criteria to include and exclude animals need to be established and described. These may be various parameters like animals' species, sex, strain, weight, developmental state, source nomenclature, and genotype, drug or test naive, and so on. Any criteria used for inclusion/exclusion after the protocol is started (eg, loss of cardiac function after infarction) should also be reported.

(*Continued*)

Table 3. Continued

Randomization	Randomization is considered the gold standard for studies intended to test hypotheses and should be used for assignment of animals to control or experimental groups. Certain studies, such as those with only a single arm (intervention only) or those with exploratory or mechanistic aims (vide supra), may not need to use randomization. Authors are asked to provide clear explanations of such exceptions. The type and methods used to randomize animals to various study arms should be reported. a. Type of randomization: for example, simple or stratified. If stratified, please describe which variables/factors were used for randomization and why. It is also recommended that investigators report measures of successful randomization, by providing baseline characteristics of animals in various comparison groups. b. Process of randomization: please provide details of the process used for randomization. These may include (but are not limited to) use of electronically generated lists, sealed envelopes, coin toss, dice roll, etc. c. Other procedural details pertaining to randomization include describing the independence of team members that generate randomization sequences, perform randomization and group allocation, and conduct experimentation. Allocation concealment is an important aspect of randomization. If the study team member(s) performing randomization have knowledge of treatment assignment groups before selection of animals, then the allocation is not concealed, and a selection bias can be introduced. Allocation concealment can be achieved if deidentified codes are generated and provided to team members performing the randomization. This usually entails an independent team member producing randomization sequences and lists. The journal encourages use and reporting of such procedures.
Blinding	Blinding is essential to avoid bias and must be used whenever possible. Blinding or masking refers to a. Inability of the investigator(s) that administers the study intervention to discern treatment from the placebo/vehicle. This can be achieved by providing the investigator(s) with animals having codes for group allocation, as well as study drug/product and placebo using a concealment procedure. The journal strongly recommends using and reporting the blinding procedures used for administration of treatment/placebo, wherever possible. For experimentation that in itself is surgical in nature, masking of the investigator(s) performing the study intervention may not be possible. We encourage appropriate reporting of these aspects in the article. b. Besides the masking of intervention, it is also important that all outcomes be assessed in a blinded manner. Although blinding the investigator(s) who administers the treatment may not be possible in all instances, blinded assessment of functional (eg, echocardiographic), histological (eg, infarct size), and other outcomes is almost universally possible. This is achieved by using an effective tracking mechanism for the experimental animals and via independence of the study team members performing outcome ascertainment from those administering the study intervention. Authors are expected to provide an account of the experimental phases that were blinded to various investigator(s) and the steps undertaken to maintain blinding while transitioning from one experimental phase to the other.
Sample size and power calculations	In confirmatory studies involving formal hypothesis testing (vide supra), authors should provide an explanation of power and sample size calculations. The rationale for using the chosen sample size should be provided. The rationale for postulating a certain effect size of the intervention on an a priori-determined primary outcome(s) should be presented, along with the anticipated variability. The sources that form the bases of power and sample size calculations should be cited. We do recognize that in exploratory or mechanistic experiments (vide supra), a formal power and sample size calculation may not be possible or meaningful. These exploratory or mechanistic analyses should be presented as such, and an explanation of why power calculations were not formally performed should be provided.
Data reporting	Control groups may either be placebo concurrent, no-treatment concurrent, active-treatment concurrent, or dose/timing-comparison concurrent. This needs to be clearly stated. It is required that investigators use a concurrent control for each intervention group. Using singular control data across several experiments conducted during a course of time is similar to using historical controls in clinical research and is a potential source of error. The number of animals included and analyzed in intervention and control groups should be clearly reported. The authors may consider using a flow diagram (as supplemental material) showing the number of animals available for randomization, number of animals randomized to various treatment/control groups, number of animals in which baseline data were collected, number of animals included in the intervention, number of animals in which follow-up/outcomes were assessed, and number of animals that were lost at any stage during the experimental protocol. Animal characteristics, such as age, developmental stage, sex, species, genetic strain, and comorbidities, should be reported, along with relevant baseline data (eg, baseline left ventricular function). Number of deaths or other events leading to exclusion should be reported in all experimental groups. If none, please state so. When reporting results, investigators are encouraged to provide summary data as tables (ie, authors should not limit data to figures). At times, figures in themselves do not allow for precise reporting of measures of central tendency and spread for different time points. If data are presented as figures, tabulated results can be provided in supplementary materials. When sample sizes are small, dot plots are preferable to bar graphs because they provide more information on variability.
Statistical methods	Authors are required to provide details on the statistical methods used for each analysis and hypothesis tested. Descriptive analyses should use measures of central tendency and spread suitable to the distribution of the data. Measures of spread (eg, SD or SE) should be suitably used, reported, and interpreted. Use of 95% confidence intervals is important to assess the size of the effect of interventions. Nonparametric tests are more appropriate for data that are not distributed normally. Commonly used tests, such as the Student t test, might not be appropriate when the sample size in each group is small. The levels of significance used for hypothesis testing should be reported, and if any adjustments were done for multiple or repeated testing. Corrections for multiple hypotheses testing and multiple comparisons to avoid type I errors are essential, unless the study is exploratory or mechanistic in nature (vide supra). When the primary outcome of a study is negative, the probability of a type II error should be reported because sample sizes are often insufficient to rule out the null hypothesis with a high level of confidence.

(*Continued*)

Table 3. Continued

Experimental details, ethics, and funding statements	It is important to provide as many experimental details as necessary for interpretation, comparison, and, importantly, full replication of the study. These details include (but are not limited to) formulation and dosage of therapeutic agent, site and route of administration, use of anesthesia and analgesia, temperature control during experimentation, and postprocedural monitoring. Temperature is particularly important in open-chest models. Further details on primary and secondary outcomes and methods used for analysis of outcomes should also be provided or clearly referenced. Authors should give due consideration to the use of animals with both sexes. Assuming only males or females are used, reasons need to be provided. The journal requires that all studies reporting animal experimentation abide by the respective institutional animal welfare and ethics regulations and reporting of approvals. Funding and conflict of interest statements should also be provided. The funding of the study by a drug development or technology development industry should be clearly stated. Any financial interest of the authors in the product being tested should be clearly indicated.

The above table is a modified version of a similar table promulgated by *Stroke* in 2011 and revised and expanded in 2016. The purpose of the table is to provide important metrics for the conduct and reporting of preclinical studies that involve assessment of experimental interventions in vivo. The recommendations provided herewith are intended to (i) clearly communicate the desired standards of rigor to authors and investigators and (ii) facilitate the journal's overall mission of promoting robust, rigorous, and reproducible scientific inquiry and experimentation. We recognize the complexity and diversity of many studies, and the existence of different levels of emphasis on in vivo data in different studies. The recommendations outlined herein are meant to be used by reviewers and editors as guidelines, not as rigid criteria for accepting or rejecting articles; thus, decisions on acceptance or rejection will continue to be made on a case-by-case basis depending on the individual content and message of each submitted article. In addition, we recognize that certain standards outlined above may not apply to exploratory, hypothesis-generating, or mechanistic animal studies. We encourage and welcome such studies; however, in these cases, we encourage the authors to address the areas outlined here and describe why adherence to particular standards was not practical or meaningful. We promote the use of online supplements for providing detailed experimental protocols and communicating any secondary results not reported in the main article.

Table 4. Checklist of Methodological and Reporting Aspects for Articles Submitted to *Circulation Research* That Report Molecular and/or Cellular Studies In Vitro

Methodological and Reporting Aspects	Description of Procedures
Conduct of the study	□ Technical replicates, whenever applicable, such as RNA-sequencing and ChIP experiments, are described.
	□ Sufficient information about sample collection is provided to distinguish between independent biological data points and technical replicates.
	□ Specific characteristics of each antibody used, such as source, host, immunoglobulin fragment, titer, catalog number, and how the antibody was validated, are described.
	□ Specific characteristics of each cell line used, such as source, authentication, and passage number, are described.
	□ Any data processing and conversions are described.
Data reporting	□ In Western blots, all experimental and control groups are included in the same blot.
	□ Molecular size markers are included in all blots.
	□ Immunofluorescence imaging is processed similarly in the experimental and control groups.
	□ A control secondary antibody or IgG is included in immunofluorescence imaging.
	□ Scale bars have been added to all micrographs.
	□ All units of measurements are specified.
	□ To the extent possible, data are reported as dot plots as opposed to bar graphs, especially for small sample size groups.
	□ In the online supplemental material, methods are described in sufficient detail to enable full replication of the study.
Statistical methods	□ The statistical methods used for each data set are described.
	□ For each statistical test, the effect size with its standard error, 95% confidence interval (for important comparisons), and *P* value is presented.
	□ Central tendency and dispersion of the data are examined, particularly for small data sets.
	□ Nonparametric tests are used for data that are not normally distributed.
	□ 2-sided *P* values are used.
	□ In studies that are not exploratory in nature, corrections for multiple hypotheses testing and multiple comparisons are performed.

The purpose of this checklist is to provide important metrics for the conduct and reporting of molecular and cellular studies in vitro. The criteria listed herewith are intended to (i) clearly communicate the desired standards of rigor to authors and investigators and (ii) facilitate the journal's overall mission of promoting robust, rigorous, and reproducible scientific inquiry and experimentation. The information requested herein is meant to be used by reviewers and editors as an aid in assessing the value of the work, not as a rigid criterion for accepting or rejecting articles; thus, decisions on acceptance or rejection will be made on a case-by-case basis depending on the individual content and message of each submitted article. In addition, we recognize that certain standards outlined above may not apply to exploratory, hypothesis-generating, or mechanistic studies. In these cases, we encourage the authors to address the criteria outlined here and describe why adherence to particular standards was not practical or meaningful. We ask authors to use online supplements for providing detailed experimental protocols and communicating any secondary results not reported in the main article.

Table 5. Explanatory Accompaniment to the Checklist for Methodological and Reporting Features of Molecular and/or Cellular Studies In Vitro

Data reporting	When reporting results, investigators are encouraged to provide summary data as tables, whenever applicable (ie, authors should not limit data to bar graphs). At times, the data presented as bar graphs in themselves do not allow for precise reporting of measures of central tendency and spread for different time points. If data are presented as bar graphs, tabulated results can be provided in supplementary materials. When sample sizes are small, dot plots are preferable to bar graphs because they provide more information on variability. Please indicate that large datasets, such as RNA-Seq and other omics, will be deposited to the appropriate public databases.
Statistical methods	Authors are required to provide details on the statistical methods used for each analysis and hypothesis tested. Descriptive analyses should use measures of central tendency and spread suitable to the distribution of the data. Measures of spread (eg, SD or SE) should be suitably used, reported, and interpreted. The 95% confidence intervals are important to assess the size of the effect of interventions and should be specified for comparisons that are deemed important. Nonparametric tests are more appropriate for data that are not distributed normally. Commonly used tests, such as the Student t test, might not be appropriate when the sample size in each group is small. The levels of significance used for hypothesis testing, and if any adjustments were done for multiple or repeated testing, should be reported. Corrections for multiple hypotheses testing and multiple comparisons to avoid type I errors are essential, unless the study is exploratory or mechanistic in nature (vide supra). When the primary outcome of a study is negative, the probability of a type II error should be reported because sample sizes are often insufficient to rule out the null hypothesis with a high level of confidence.
Experimental details, ethics, and funding statements	It is important to provide as many experimental details as necessary for interpretation, comparison, and, importantly, full replication of the study. Any financial interest of the authors in the product being tested should be clearly indicated.

The above table is a modified version of a similar table promulgated by *Stroke* in 2011 and revised and expanded in 2016. The purpose of the table is to provide important metrics for the conduct and reporting of molecular and/or cellular studies in vitro. The recommendations provided herewith are intended to (i) clearly communicate the desired standards of rigor to authors and investigators and (ii) facilitate the journal's overall mission of promoting robust, rigorous, and reproducible scientific inquiry and experimentation. The recommendations outlined herein are meant to be used by reviewers and editors as guidelines, not as rigid criteria for accepting or rejecting articles; thus, decisions on acceptance or rejection will continue to be made on a case-by-case basis depending on the individual content and characteristics of each submitted article. We promote the use of online supplements for providing detailed experimental protocols and communicating any secondary results not reported in the main article.

Circulation Research
JOURNAL OF THE AMERICAN HEART ASSOCIATION

New Initiatives to Improve the Rigor and Reproducibility of Articles Published in *Circulation Research*

Roberto Bolli

Circ Res. 2017;121:472-479
doi: 10.1161/CIRCRESAHA.117.311678
Circulation Research is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 0009-7330. Online ISSN: 1524-4571

Review

Standards and Methodological Rigor in Pulmonary Arterial Hypertension Preclinical and Translational Research

Steeve Provencher, Stephen L. Archer, F. Daniel Ramirez, Benjamin Hibbert, Roxane Paulin, Olivier Boucherat, Yves Lacasse, Sébastien Bonnet

Abstract: **Despite advances in our understanding of the pathophysiology and the management of pulmonary arterial hypertension (PAH), significant therapeutic gaps remain for this devastating disease. Yet, few innovative therapies beyond the traditional pathways of endothelial dysfunction have reached clinical trial phases in PAH. Although there are inherent limitations of the currently available models of PAH, the leaky pipeline of innovative therapies relates, in part, to flawed preclinical research methodology, including lack of rigour in trial design, incomplete invasive hemodynamic assessment, and lack of careful translational studies that replicate randomized controlled trials in humans with attention to adverse effects and benefits. Rigorous methodology should include the use of prespecified eligibility criteria, sample sizes that permit valid statistical analysis, randomization, blinded assessment of standardized outcomes, and transparent reporting of results. Better design and implementation of preclinical studies can minimize inherent flaws in the models of PAH, reduce the risk of bias, and enhance external validity and our ability to distinguish truly promising therapies form many false-positive or overstated leads. Ideally, preclinical studies should use advanced imaging, study several preclinical pulmonary hypertension models, or correlate rodent and human findings and consider the fate of the right ventricle, which is the major determinant of prognosis in human PAH. Although these principles are widely endorsed, empirical evidence suggests that such rigor is often lacking in pulmonary hypertension preclinical research. The present article discusses the pitfalls in the design of preclinical pulmonary hypertension trials and discusses opportunities to create preclinical trials with improved predictive value in guiding early-phase drug development in patients with PAH, which will need support not only from researchers, peer reviewers, and editors but also from academic institutions, funding agencies, and animal ethics authorities.** (*Circ Res*. 2018;122:1021-1032. DOI: 10.1161/CIRCRESAHA.117.312579.)

Key Words: animals ■ bias ■ humans ■ hypertension, pulmonary ■ methodological rigor ■ reproducibility ■ rodentia

The number of scientific publications related to pulmonary hypertension (PH) listed on Pubmed has increased exponentially during the last 2 decades (Figure 1). This great interest in preclinical, translational, and clinical research has led to significant advances in our understanding of the pathophysiology[1] and the management[2] of group 1 PH, also called pulmonary arterial hypertension (PAH). Within this period, >10 drugs targeting the endothelin, NO, and prostacyclin pathways[3] were developed and commercialized, decreasing clinical worsening[4] and potentially short-term mortality.[5] However, long-term prognosis of patients with PAH remains poor.[6–8] Significant translational and therapeutic gaps between preclinical research and improved patients' outcomes thus remain. Unfortunately, few innovative therapies have reached clinical trial phases in PAH,[9–12] and drug development beyond the traditional pathways of endothelial dysfunction has been unsuccessful.

Since the beginning of the millennium, the number of compounds in development and total Research and Development expenditures have increased markedly.[13] Yet, during the same period, the average number of new drugs approved yearly by health authorities has declined.[14] Clinical drug development is notoriously arduous, and there are several possible explanations for the divergence of Research and Development spending and new product approval, including higher regulatory efficacy hurdles, the increased complexity and cost of clinical trials, and the inherent variation and imprecision of clinical studies because of patient heterogeneity.[15] As a result, <5% of high-impact basic science discoveries[16] and <10% of development paths in phase 1[14] are eventually approved by health

From the Pulmonary Hypertension Research Group, Institut universitaire de cardiologie et de pneumologie de Québec Research Center (S.P., R.P., O.B., S.B.), Institut universitaire de cardiologie et de pneumologie de Québec Research Center (S.P., R.P., O.B., Y.L., S.B.), and Department of Medicine (S.P., R.P., O.B., Y.L., S.B.), Université Laval, Québec, Canada; Department of Medicine, Queen's University, Kingston, Canada (S.L.A.); Division of Cardiology (F.D.R., B.H.), CAPITAL Research Group (F.D.R., B.H.), and Vascular Biology and Experimental Medicine Laboratory (B.H.), University of Ottawa Heart Institute, Ontario, Canada; and Department of Cellular and Molecular Medicine (B.H.) and School of Epidemiology and Public Health (F.D.R.) University of Ottawa, Ontario, Canada.

Correspondence to Sébastien Bonnet, PhD, Pulmonary Hypertension Research Group, Institut universitaire de cardiologie et de pneumologie de Québec Research Center, Laval University, Quebec City, Canada. E-mail sebastien.bonnet@criucpq.ulaval.ca

Circulation Research **is available at http://circres.ahajournals.org** **DOI: 10.1161/CIRCRESAHA.117.312579**

Nonstandard Abbreviations and Acronyms	
PAH	pulmonary arterial hypertension
PH	pulmonary hypertension

authorities. Such large-scale attrition of investigational drugs is potentially harmful to individuals in trials and consumes scarce human and material resources.[14,17,18]

The failure to test new classes of PAH therapy in humans relates, in part, to choices by Pharma, related to their willingness to take risks (including risks of repurposing approved agents), the role that their proprietary pipeline plays on selection of lead compounds for testing, and their diminished use of preclinical in-house testing of new therapies. Adding to this, there are inherent limitations of the currently available in vitro and animal models, which imperfectly mimic the full spectrum of the human disease.[19] This is not unique to PAH[20] because preclinical trials have recently been identified as being less than helpful in predicting human therapeutic response in diseases such as pulmonary fibrosis,[21] stroke,[22] and sepsis.[23] However, we would argue that in many translational models, the failing is more in the study design, implementation, and analysis than in the ability of the model to recapitulate the human disease, ultimately weakening our confidence in preclinical studies to identify promising therapeutic targets.[17] Bias in study design, analytical methods, and reporting practices may compromise scientific validity,[24] data reproducibility,[25] and ultimately hinder progress in patients' care. In response to these issues, the National Institutes of Health proposed a set of guidelines and funding policies as minimum reporting requirements to promote rigor, reproducibility, and transparency of preclinical research (Table).

Although these principles and guidelines have been endorsed by prominent academic societies, associations, and journals with evidence of editorial commitment to complying,[26–28] recent empirical evidence suggests that such rigor is still lacking in cardiovascular[29] and PH preclinical research.[30] The preclinical phase of research is critical to the decisions made about future drug development. The challenges and pitfalls of early-stage preclinical studies in PAH have been reviewed elsewhere,[19] including the selection and processing of human samples, novel culture technologies, and the advantages and disadvantages of currently available animal models of PH. The present article is an in-depth review addressing the needs for improved methodological rigor and reporting in translational PAH confirmatory research using practical strategies mimicking clinical trial methodologies to enhance reproducibility

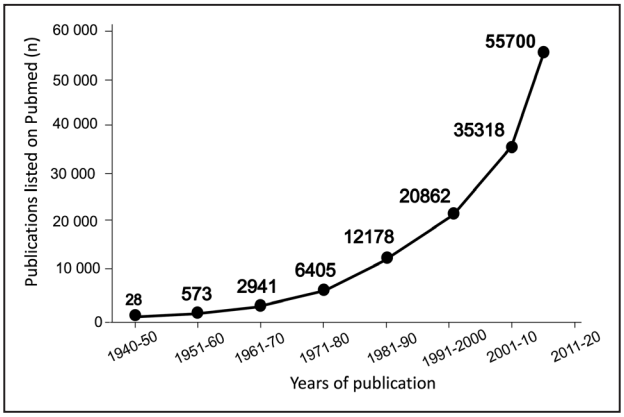


Figure 1. Number of publications related to pulmonary hypertension indexed on Pubmed per decade from 1940 to 2017.

Table. Key Principles and Guidelines for Reporting Preclinical Research Delivered by the National Institutes of Health

Rigorous statistical analysis
A section outlining the journal's policies for statistical analysis should be included in the Information for Authors, and the journal should have a mechanism to check the statistical accuracy of submissions.
Transparency in reporting
Journals should have no limit or generous limits on the length of Methods section (including online options) and use a checklist during editorial processing to ensure the reporting of key methodological and analytic information to reviewers and readers, including detailed reporting on the following:
How often each experiment was performed, and whether the results were substantiated by repetition under a range of conditions.
Sample collection to distinguish between independent biological data points and technical replicates.
Statistics, including the statistical test used, exact value of n, definition of center, and dispersion and precision measures.
Whether the samples were randomized, and method of randomization.
Whether experimenters were blind to group assignment and outcome assessment.
Whether an appropriate sample size was computed when the study was being designed and includes the statistical method of computation. If no power analysis was used, include how the sample size was determined.
Criteria that were used for exclusion of any data or subjects.
Similar experimental results that were omitted from the reporting for any reason, especially if the results do not support the main findings of the study.
Outcomes or conditions that were measured or used and are not reported in the Results section.
Data and material sharing
All datasets on which the conclusions of the article rely must be made available on request (where ethically appropriate) during consideration of the manuscript.
Deposition of datasets in public repositories is recommended.
Software should be shared (at the minimum description of whether the software is available and how it can be obtained).
Consideration of refutations
Journals should have a policy stating that if the journal publishes a paper, it assumes responsibility to consider publication of refutations of that paper, according to its usual standards of quality.
Best practice guidelines
Consider establishing best practice guidelines for the following:
Image-based data
Description of biological material with enough information to uniquely identify the reagents.

and minimize the risk of advancing a compound to clinical testing that is then proven to have no benefit or to have detrimental effects.

Defining the Programmatic Purpose of Research and Adapting Statistical and Methodological Rigor Accordingly

For obvious reasons, methodological expectations must be appropriately calibrated to the nature and goals of the study (Figure 2). The statistical and methodological rigor should thus be adapted according to the nature of the study. In exploratory research, investigators seek to develop pathophysiological theories and identify candidate cellular and molecular signaling pathways/targets implicated in PAH development or progression. Positive findings are generally followed by proof-of-concept studies in which the presence of an efficacy signal is hoped. At these stages, avoiding false-negative results by accepting less-rigorous study designs and statistical analyses may be reasonable to increase the sensitivity of detecting a potentially useful therapy at the expense of specificity.[31] However, even at an exploratory stage, attention is required to ensure appropriate sample size is accrued to avoid spurious results and enhance the likelihood of reproducible findings that can be externally validated. In confirmatory investigations, however, researchers seek to provide strong, reproducible, and detailed information using clinically relevant assays and end points. In many ways, confirmatory studies resemble clinical trials, providing data on aspects such as drug dosing, toxicity, and long-term effects of potential drug candidates to decide whether the drug could or should be tested in humans. Methodological prerequisites for this type of study differ, and the most stringent study design is essential.

Pitfalls in PAH Preclinical Methodological Design

A bias is a systematic error that results in a deviation from the truth and cannot be rectified via study replication or by increasing a study's statistical power. In contrast, imprecision relates to random error. Thus, studies may produce precise but biased results because of flaws in study design and execution. Conversely, a study may be free of significant bias but yield an incorrect effect estimate because of low statistical power (random error). Thoughtful subject eligibility criteria, sample size estimation, randomization, blinding, standardized outcome assessment, proper data handling, and transparent reporting methods address these issues and have profoundly improved the validity of clinical trial results over the years.[32] However, although these design and reporting elements are often considered only in clinical trials, preclinical research is not immune to these same methodological pitfalls. Preclinical scientists should, therefore, consider the impact of study design, data interpretation, and reporting to promote scientific reproducibility and transparency in confirmatory preclinical studies.

Eligibility Criteria: Matching Models to Human Manifestations of PAH

The first step in designing a clinical trial is to define the eligibility criteria with the aim of recruiting a study population representative of future patients to be treated while minimizing confounding effects. Care must thus be taken to enrol experimental animals in studies as patients are

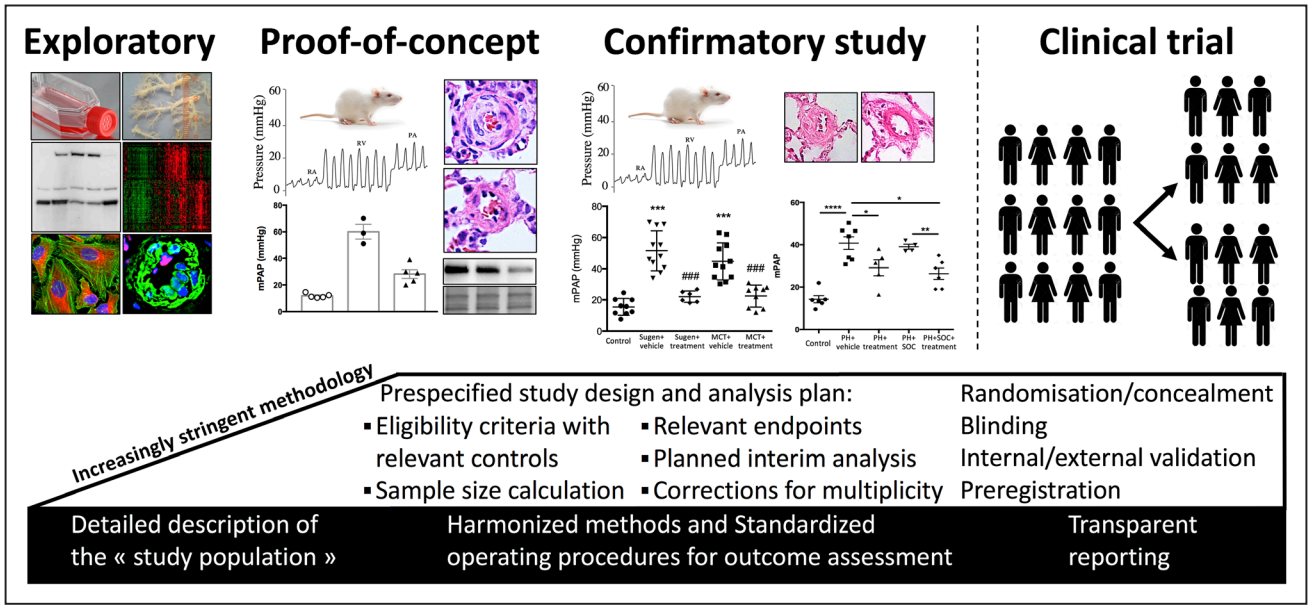


Figure 2. Adaptation of statistical and methodological rigor according to the nature of the study. There is no single type of laboratory study. For confirmatory studies, the most rigorous application of randomized controlled trial-like approaches should be applied, including prespecified study design and statistical analysis plan, describing eligibility criteria, sample size calculation, relevant study read-outs, interim analysis (if any), and corrections for multiple analyses, as well as randomization and concealment procedures, validation of the results, and potentially study preregistration. However, this type of approach may be expensive and cumbersome, and applying randomized controlled trial-like standards to studies focusing on discovery and elucidating fundamental biological processes may not add the same level of value. Thus, expectations and methodological rigor must be appropriately calibrated to the goals of the study. In all cases, however, the choice of an appropriate animal model that is representative of the human disease, the development of standardized operating procedures, blinded outcome assessment, and transparent reporting is critical.

enrolled in clinical trials with attention to reporting results from all animals studied and to prespecify inclusion and exclusion criteria for analysis. In addition to the substantial limitations of current animal models that are beyond the scope of the current review,[19,33] variations in disease induction and the potential for persistent and unrecognized confounders represent other important sources of bias. Even when using a well-recognized animal model of PH, there are considerable inconsistencies in animals' age and weight, how PH is induced (eg, route and dose of the inducing agent),[34,35] duration of hypoxia, and when the intervention is initiated and terminated.[30] This is particularly of concern because PH severity has been shown to increase progressively over time before spontaneously improving or even resolving when hypoxic models are returned to normoxic condition, as in weeks after low-dose monocrotaline[36] or Sugen[37,38] injections. Similarly, experimental PH is more sensitive to preventive strategies compared with reversal strategies,[30] likely influencing the predictability of the model. In studying a preclinical therapy for PAH, it is, therefore, reasonable to randomize animals to therapy when irreversible PH is expectedly fully established and following prior confirmation (eg, pulmonary artery acceleration time <25 ms as evidence of PH).

The high degree of experimental control that is possible in preclinical research can reduce the interindividual variability and, therefore, the sample sizes required relative to clinical studies. This is in marked contrast with the heterogeneous responses observed in most clinical trials as a result of the diverse phenotypic and genetic backgrounds in the patients.[12,39] In addition to the inconsistencies in animal models described above, there can be striking differences in cardiovascular responses to stimuli in different strains[40] because of both genetic and epigenetic differences in the animals. For these reasons, the rationale for choosing models should be stated,[41,42] and performing studies using >1 model is recommended, whenever possible, before proceeding to human studies. However, studies not only across different models of PH but also across animal strains are encouraged. The recent establishment of the hybrid mouse diversity panel may be extremely helpful in beginning to address some of these issues.[43] Ultimately, large animal models, such as pigs,[44,45] the neonatal chronic hypoxia calf model,[46,47] or nonhuman primates,[48] may share some common features of the human disease, including right heart failure and the development of complex vascular lung lesions. Indeed, preclinical testing in large animal models is often the last step before translating novel drug candidates to clinical trials, for both safety and efficacy assessment.[49] However, current large animal models have been limitedly exploited in PAH research because of incomplete recapitulation of the human phenotype or high mortality.[50,51] Human ex vivo model systems are also rapidly emerging strategies to study human disease pathogenesis and response to pharmacological intervention, including in PAH.[12] Similarly, engineered 3-dimensional lung tissue constructs that mimic complex tissue physiology have amazing potential for use as lung physiology or disease pathology models, but models requiring large number of cells, multiple cell populations, or use of specific extracellular components or natural scaffolds are often not suitable for the generation of standardized microfluidics-based high-throughput systems for drug discovery or toxicity testing.[52,53] Importantly, increasing methodological and reporting rigor of studies using inappropriate animal models is unlikely to better inform the scientific community about mechanisms and therapeutic interventions in human PAH. We would argue, however, that if the findings in a single robust model fully support extensive data acquired from human tissues and cells, adding further animal models may not be justified. Conversely, not all laboratories have access to human cells and tissues, and they may make fundamental discoveries with translational potential using animal cells and models but need to acknowledge the lack of corroborative human studies as a limitation.

The need to include women in clinical trials is now a well-established requirement.[54] However, analogous standards have not been equally enforced in preclinical stages of research. This is especially problematic in PAH where there is a significant female predominance in humans,[55] and preclinical data have confirmed the impact of female hormones on the disease.[56] Nevertheless, the vast majority of preclinical PH studies use male rodents (Figure 3).[30,57] Arguments put forward for the preferential use of males include variability attributable to fluctuating gonadal hormone levels, sample size implications of planning sex-based analyses, and cost.[58,59] Yet, evidence to support these concerns is limited.[60] Furthermore, there is an important distinction between adequately powering a study for detailed sex-based analyses of results and including sufficient animals of both sexes to detect large differences between them, if present.[59] Although the former is valuable, the latter may be sufficient in many instances.[61] Inferring experimental findings to both sexes when a single sex is studied could disadvantage women by biasing our understanding of disease processes toward male-predominant patterns.

Control groups may either be placebo concurrent, no-treatment concurrent, active-treatment concurrent, or dose/timing comparison concurrent. It is required, however, that investigators use a concurrent control for each intervention group. Importantly, new targets can be alternatives to the currently approved therapies in humans even if they do not have additive effects (eg, for nonresponders). However, the landscape of clinical trials in PAH dramatically changed during the last decade. Sequential or upfront combination therapy is now an emerging standard of care in PAH.[2,62] For ethical reasons, future compounds will, therefore, almost necessarily be tested on top of currently available therapies in clinical trials leading to drug approval. As for placebo-controlled trials in humans, the demonstration of additive or even synergic effects of novel therapeutic targets nowadays seems essential for confirmatory preclinical studies.

Therefore, a detailed characterization and reporting of animal traits at baseline (species, strain, age, sex, genetic modification status, housing conditions, etc), of appropriate controls (littermate, purchased, contemporaneous versus historical, etc) using animal characteristics that are representative of the human disease should thus be promoted for a better standardization of the experimental design, enhanced reproducibility, and greater predictive ability.

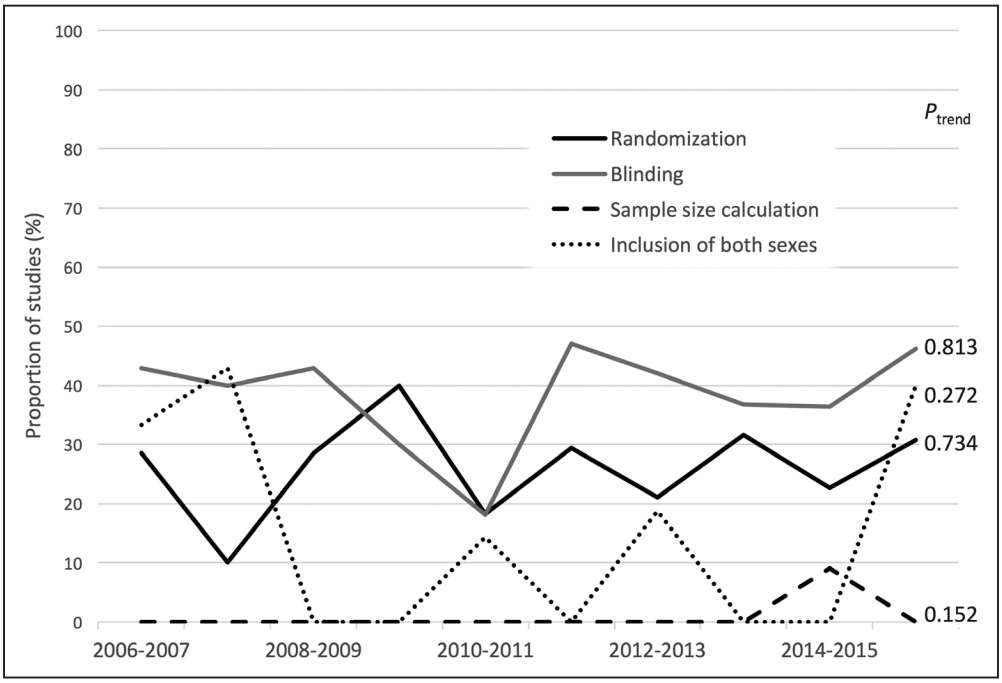


Figure 3. Temporal patterns in randomization, blinding, sample size estimation, and inclusion of both sexes in pulmonary hypertension preclinical research. Articles from Circulation; Circulation Research; Hypertension; Stroke; and Arteriosclerosis, Thrombosis, and Vascular Biology were screened from July 2006 to June 2016. P_{trend} calculated for the proportion of studies reporting the implementation of a specific study design element over time using Cochrane–Armitage trend tests. Adapted from Ramirez et al[29] with permission. Copyright © 2017, the American Heart Association.

Randomization and Allocation Concealment of Animals to Intervention

The starting point for an unbiased interventional study is the use of a mechanism that ensures that the same sort of participants receive each intervention. Although researchers may argue that preclinical interventions are generally performed on a homogeneous group of animals (eg, same strain), animals may have inherent differences when the intervention is introduced (eg, age, weight, and PH severity). Several interrelated processes need to be considered to allow proper balance between groups. First, random animal allocation, if properly implemented in a large enough sample, should balance characteristics that may influence response to treatment. As such, randomization is essential to minimizing bias. However, selection bias may arise if the random allocation is not concealed. Concealment refers to techniques used to implement the allocation sequence. If future assignment can be anticipated (ie, allocation is not concealed), then selection bias may be introduced as animals may be selected based on the upcoming intervention assignment. In clinical trials, there is empirical evidence that either inadequate generation[63] or concealment[64] of allocation sequence yield to exaggerated estimates of intervention effects. Similar observations were made for preclinical research.[65,66] Thus, the type and methods used to randomize animals to various study arms should be reported clearly. These may include (but are not limited to) the use of computerized randomization. Allocation concealment can be achieved if deidentified codes are generated and provided to team members performing group allocation. This usually entails an independent team member producing randomization sequences and lists. Finally, researchers should report measures of successful randomization by providing baseline characteristics of animals (eg, age, weight, and echocardiography parameters) in various comparison groups.

Blinding of Outcome Assessment

Blinding refers to the process by which the study personnel, including people assessing the outcomes, are kept unaware of intervention allocations. Lack of blinding in clinical trials is associated with exaggerated estimates of intervention effects,[64] especially when the outcome of interest is subjective.[67] Importantly, many apparently objective outcome measures in preclinical PAH studies remain subject to interpretation (eg, pulmonary artery remodeling and hemodynamic tracings). Thus, blinding procedures must be reported unless there is a clear rationale for the nonblinding of the experimenter (eg, surgical intervention) in which case this should be made explicit. Although blinding the investigator administering the treatment/intervention may not be possible in all instances, blinded assessment of imaging, hemodynamics, and histological outcomes is almost universally possible through independent team members performing outcome ascertainment. Even scientists of high integrity are enthusiastic supporters of their hypotheses, and unconscious bias can creep into evaluation of unblinded experiments.

Study Read-Outs and Interstudy Standardization in Preclinical PAH Studies

The primary end point of a study may be defined as the variable capable of providing the most relevant and convincing evidence directly related to the primary objective of the study. It is noteworthy that even with rigorous attention to inclusion of multiple models and both sexes and blinded assessments,

results may not have translational validity if the end point specified is not valid or is not measured using robust techniques. In preclinical confirmatory studies, this outcome measure should match the clinical realm using relevant measures (eg, comprehensive hemodynamics) because nonsurrogate outcomes, such as survival, may not be approved by ethical authorities in many countries. Because multiple end points are often necessary to adequately evaluate the effects of a therapy, other read-outs are generally used to provide supportive (secondary end points) or exploratory, hypothesis-generating information. Preclinical PH interventional studies most commonly include makers of pulmonary hemodynamics (eg, right ventricular systolic pressure and mean pulmonary artery pressure) and pulmonary artery remodeling. A major value of animal studies is the ability to ensure hemodynamic end points are correlated with histological, anatomic, and biochemical findings postmortem. Nonetheless, the use of the same core set of end points in confirmatory preclinical efficacy studies in PH would clearly facilitate the comparison of results. Likewise, the use of incomplete hemodynamics and premature termination of studies based on soft end points biases studies toward positive findings and should be avoided.

As for clinical studies, the primary end point of preclinical confirmatory PH studies must be decided before the study begins because this is a prerequisite for sample size calculation. The importance of prespecified sample size calculation, referred to as a power calculation, in confirmatory experiments cannot be overemphasized, although it is rarely performed, or at least is rarely reported, in preclinical PH studies (Figure 3). A small sample will result in an inconclusive study, whereas an unnecessary large sample size will accrue excessive cost—both scenarios expending limited resources (Figure 4). In both situations, inappropriate sample size can be judged unethical because exposing laboratory animals (as humans) to the possible risks associated with research is only justifiable if there is a realistic chance that the study will yield useful information. Although many researchers are tempted to perform interim analyses and subsequently increase the sample size as necessary, prespecified sample size calculation is mandatory to avoid the risk of false-positive results because of multiple analyses. The exploratory nature of some experiments makes formal power and sample size calculation impossible or meaningless; however, the rationale for selecting the primary outcome and the effect size of the intervention for which the study is powered should be provided in the Methods section of confirmatory experiments. The main factors that must be considered in formal hypothesis testing are the α-error, β-error, and effect size, as well as the attrition rate and SD for quantitative measurements. These values are either known from the literature or previous proof-of-concept studies or can be estimated by reasonable guesswork.

Similarly important is how the outcomes of interest are measured. In clinical trials, detailed standard operating procedures have been developed for measures that may influence study outcomes, including how the 6-minute walk test,[68] echocardiography,[69] and catheterization[70] are performed and how adverse events are reported.[71] Conversely, empirical work has confirmed marked heterogeneity in the methodologies used to assess study outcomes in preclinical PH studies.[30] Indeed, the assessment of pulmonary hemodynamics is frequently limited to right ventricular systolic pressure, whereas indirect markers of pulmonary vascular disease (eg, pulmonary vascular resistance) and right ventricular function (eg, cardiac output and end-diastolic pressure) would be readily available. It can be argued that because the most common confounder in terms of a diagnosis of group 1 PH is group 2 PH, the requirement in preclinical studies (as in clinical studies) should include measurement of left atrial or left ventricle end-diastolic pressure and calculation of pulmonary vascular resistance to exclude the possibility the observed PH is in fact because of left heart disease. This is a particular concern in genetically manipulated mouse models. In this regard, echocardiography alone

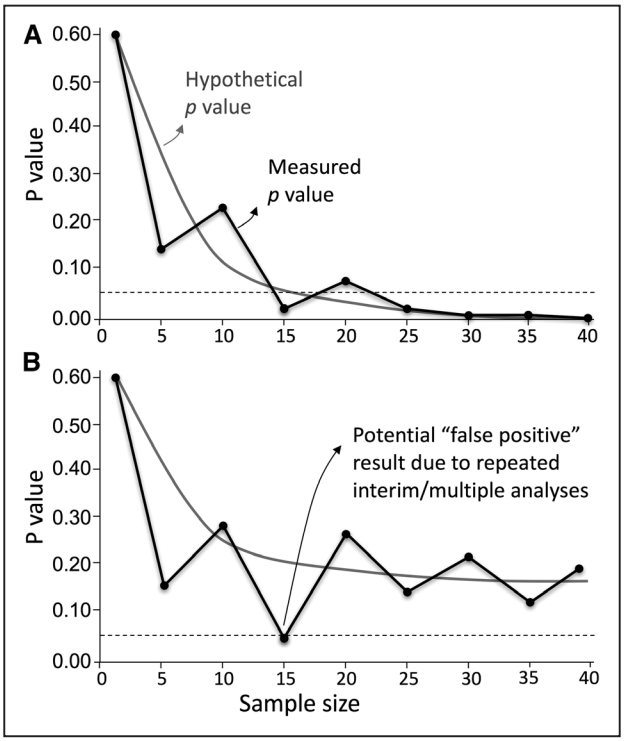


Figure 4. Precision of the effect estimate as the sample size increases. Statistical analyses are made on a study sample, which is a subset of a larger population that is collected and analyzed to make inferences. As the sample size increases, the precision of the estimates increases, assuming that the study sample was representative of the population of interest (ie, no sampling error). When comparing 2 groups as sample size increases, the calculated *P* value would theoretically stabilize toward its true value (gray lines). In practice, the *P* value erratically progresses, especially at lower sample size (black lines), particularly in the presence of outlier measurements. An underpowered study may be associated with type II error (false-negative result), whereas increasing the sample size beyond a certain point (eg, beyond 30 in the example above) would be unnecessary (**A**). Conversely, interim analyses may be associated with type I error (false-positive results). Indeed, using the conventional significance level of 5%, it is expected that 1 in 20 tests will be statistically significant even when there is no difference between the interventions being compared, then 0.95×0.05=0.0475 at second look (thus a total probability of 0.0975 after both analyses), assuming that the measures are independent. Thus, the nominal *P* values should be adjusted after conducting interim analyses if one does not want to compromise the validity or integrity of the final analyses (**B**).

is insufficient to exclude group 2 PH. How catheterization is performed is also frequently poorly described. However, hemodynamics should ideally be done with inhaled anesthesia in closed-chest animals to avoid artifactual reduction in pressures that accompany over sedation[72–74] or opening of the thorax (Figure 5). When conducted using rigorous attention to anesthesia, a head-to-head comparison showed that the results of hemodynamics assessment of the pulmonary circulation by high fidelity cardiac catheterization, cardiac magnetic resonance imaging, and echocardiography are similar (and reproducible) in rodent models.[75]

Significant heterogeneity also exists when describing pulmonary remodeling,[30] and there is no consensus on how to report adverse events, and the majority of preclinical studies fail to appropriately monitor for toxicity. To ensure comparability and reproducibility of results from different laboratories, as many variables as possible in terms of animal handling and testing need to be controlled for. Obviously, these elements cannot be unilaterally dictated but require a consensus process to take place, with experts in the field coming together to agree on best practice. Consensus and guidelines for a more harmonized, common effort in preclinical research have been developed for preclinical in vivo evaluation of pharmacological active drugs for stroke,[76] amyotrophic lateral sclerosis,[77] and multiple sclerosis[78] to name a few. Such efforts are clearly needed in PH.

Multiplicity and the Play of Chance: Interim Analyses and *P* Value Adjustments

Because multiple read-outs are necessary to fully evaluate pathophysiological pathways and evaluate all of the effects of interventions, multiple end points are frequently measured. However, every time statistical analyses are performed, the chance of falsely rejecting the null hypothesis is introduced. Conducting multiple tests of significance progressively increases the probability that a null hypothesis is rejected when the null hypothesis is actually true (ie, false-positive result). Using the conventional significance level of 5%, it is expected that 1 in 20 tests will be statistically significant even when there is no difference between the interventions being compared. Consideration must be given to controlling the risk of false-positive conclusions, and adjustment for multiplicity will typically be necessary. A simple solution to impose stricter control on the probability of getting a false-positive result for confirmatory studies, yet increasing the risk of type II errors (false-negative results). Several approaches have been proposed for addressing multiplicity, including the Bonferroni procedure, related stepwise approaches, and hierarchical procedures.[79–81] Similarly, interim analyses are also frequently used to incorporate what is learned during the course of a study. The simplest result of such an interim analysis is early stopping for futility or success or continuation of the study. Similar to multiple analyses, interim analyses increase the risk of falsely rejecting the null hypothesis (Figure 4). Therefore, there is a need to adjust the nominal *P* values after the conduct of interim analyses if one does not want to compromise the validity or integrity of subsequent analyses. As such, researchers should avoid unplanned interim analyses, and preliminary results should be presented without formal statistical analyses unless nominal *P* values have been adjusted accordingly. In an attempt to maintain the validity and integrity of the final analyses, collaboration with a statistician at the design stage is crucial. Regardless of which approach is used, the selected procedure must be prespecified in the statistical analysis plan before undertaking any analyses of the data.

Interpretation of Study Results and Reporting

Incomplete Outcome Data and Handling of Missing Data

Missing outcome data, because of attrition during the study or exclusions from the analysis, raise the possibility that the observed effect estimate is biased. Attrition and exclusions frequently occur in preclinical PH studies when animals die or are withdrawn from the experiments, assessment does not provide relevant data, data are lost or unavailable, or the animals are excluded for other reasons (eg, outliers). Comprehensive reporting of the number of animals in intervention and control groups that were available for randomization, that were randomized to various treatments/control groups and included in the intervention, as well as the number of animals for which baseline data were collected, follow-up/outcomes were assessed, and were lost at any stage during the experimental process should be detailed. Researchers should consider using a flowchart showing animals at each experimental step. A timeline of experimentation is also desirable to inform whether all animals within each experimental group were analyzed together or were a composite of >1 experiment conducted at different times.

For confirmatory studies, an intention-to-treat analysis may be considered as potentially the least biased way to estimate intervention effects in randomized trials.[82] In clinical trials, empirical evidence has suggested more exaggerated effect estimates from per-protocol analyses compared with intention-to-treat analysis of the same trials.[83] The same likely applies to preclinical studies. The risk of bias from incomplete

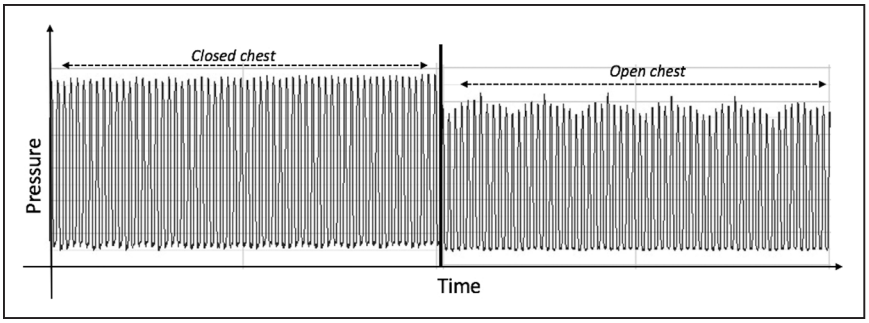


Figure 5. Representative tracings of the artifactual reduction in pressures in right ventricular pressure when opening the thorax.

outcome data depends on several factors, including the amount and distribution of missing data across intervention arms, the reasons for missing outcome data, and the impact of including or excluding the missing data on the outcome effect estimate. However, few studies can perform true intention-to-treat analyses without imputation. This involves making assumptions about the outcomes of animals for which no outcome was recorded. Examples include imputing the last observation carried forward in which the outcome measure is assumed to hold for all subsequent outcome assessment times in case of unreliable measures (eg, echocardiographic measures) or worst-case scenario outcomes in case of an animal's death. However, statistical analyses cannot entirely compensate for missing data, and imputation rules can also lead to serious biases unless conservative methods are used. Thus, where imputation is used, both the per-protocol and the intention-to-treat analyses should be presented, and the methods and assumptions for imputing data should be defined a priori and described in the Methods section of the article.

Interpretation of the Results: What Does the *P* Value Tell You (and Does Not Tell You)?

The high pressure to find low *P* values, combined with a common misunderstanding of how to correctly interpret *P* values, frequently distorts the interpretation of significant results.[84] A low *P* value (typically ≤0.05) is considered strong evidence against the null hypothesis. However, a *P* value of 0.05 is frequently incorrectly interpreted as meaning that there is 95% chance that the observed difference is true, rather than as indicating a 5% probability of the observed difference (or greater) if the null hypothesis was true. The *P* value is not the error rate or the likelihood of making a mistake by rejecting a true null hypothesis (type I error or false-positive rate). Indeed, Sellke et al[85] estimated that a *P* value of 0.05 corresponds to a false-positive rate of at least 23% (and typically close to 50%). Thus, a single statistically significant hypothesis test often provides insufficient evidence to confidently discard the null hypothesis. Study replication, especially by independent investigators, thus enhances the confidence that study results are true findings (Figure 6). Therefore, investing more time in replicating results (those of others, as well as our own) should be considered and incentivized. There is an obvious balance that must be achieved to meet the 3Rs of ethical animal research (ie, replacement, reduction, and refinement, which mandates to avoid unnecessary duplication with mini-mal incremental yield). One can make a compelling case that investigators should have latitude to design the studies as they wish and can afford in accordance with ethical standards, provided they draw conclusions from appropriately designed studies that will directly guide future drug development. Systematic reviews and meta-analyses of preclinical experiments could also play important roles in synthesizing these data and have been advocated as prerequisites to undertaking clinical trials.[86]

Publication Bias and Selective Reporting Bias in Preclinical PH Studies

Reporting biases arise when the dissemination of research findings is influenced by the nature and direction of results.

First study

	True effects	No real effects	Predictive value
Positive study	80	20	80/100 = 80%
Negative study	20	380	380/400 = 95%
Number of studies	100	400	500

Repeated study

	True effects	No real effects	Predictive value
Positive study	64	1	64/65 = 98.5%
Negative study	16	19	19/35 = 54.3%
Number of studies	80	20	100

Figure 6. True-positive rates according to prior probability of effectiveness. A *P* value of 0.05 is frequently incorrectly interpreted as meaning that there is 95% chance that the observed difference is true, rather than assuming that in the event that the null hypothesis was true, you would obtain the observed difference (or more) in 5% of studies. In fact, the false-positive rate depends on various factors, including the prevalence (or prior probability) of real effects, study power, and *P* value. Let's imagine that based on previous experiments, a compound has a 20% probability (100 of 500 studies) of being effective in diseased animals. A *P* value of 0.05 (ie, 20 of 400 false-positive studies) in the in vivo study powered at 80% (ie, 80 of 100 true-positive studies) would be associated with an error estimate of 20% (ie, a 20% risk of being a false-positive result) based on effect size and variability estimated a priori. Replicating the study will then minimize the risk of false-positive results to 1.5% because the prior probability of being effective is now 80%. Increasing the study power and the *P* value would also increase the true-positive rate if the sample size is determined accordingly a priori.

Publication bias generally occurs when entire studies are not published. There is a strong possibility that such studies are missing because of their uninteresting or unwelcome findings (ie, true publication bias), leading almost inevitably to major overstatements of efficacy.[87] More subtle publication biases arise when studies are published in obscure journals, rarely cited, inappropriately indexed in databases, or when publication is delayed. Interestingly, when authors were asked why they had not published their findings, the most frequent answer was that they were not interesting enough to merit publication,[88,89] and rejection by a journal was rarely mentioned as a reason for not publishing. Therefore, selective submission by the authors rather than selective recommendations by the reviewers or selective acceptance by the editors may be the dominant contributor to publication bias. Even within a published report, those analyses with statistically significant findings are more likely to be reported, the most commonly reported reason for nonpublication of results being lack of statistical significance or limited magnitude of effects. Some journals indirectly contribute to selective reporting by restricting their page allowances or relegating less-interesting findings to the Supplement section. For the readers, it is often difficult to determine whether this is because the outcome was not measured or because the outcome was not reported. This sort of within-study publication bias may be one of the most

substantial biases affecting results from individual studies,[90] including in preclinical PH research.[30] It is, therefore, imperative that analyses of confirmatory studies be determined by a prespecified statistical plan in which the primary, secondary, and hypothesis-generating measures are identified, along with the specific methods to correct for multiplicity and that the entire set of outcome measures planned a priori be reported.

Publication and selective reporting biases also prevent others from learning about negative study results. Full reporting of study findings is perhaps the most important aspect of conducting transparent preclinical research, with implications for animal ethics and research funding. Retaining negative findings also distorts the available body of evidence, which may lead to waste of time and resources for other researchers and humans being exposed to ineffective or toxic therapies. To minimize publication and reporting bias, study preregistration—a bias-thwarting strategy that requires researchers to formally document analysis plans before collecting results—was developed for clinical trials. The idea behind preregistration is that publicly listing the planned experimental design and analysis limits a researcher's ability to modify them afterward—a practice that is common among researchers. Indeed, many researchers engage in some form of (usually mild and unintentional) scientific malpractice at one point or another. As a result, the International Committee of Medical Journal Editors now considers only those clinical trials for publication that have been registered before the start of patient recruitment.[91] In preclinical studies, preregistration in a public repository at the study inception is a debated issue. Some advocate that it would minimize publication and reporting biases, prevent fishing or manipulation to achieve a desired result, and provide leverage for scholars who face result-oriented pressure from financial benefactors or funding agencies. However, many researchers fear loss of the ability to protect intellectual propriety. More importantly, preregistration relies strongly on the notion that science is always confirmatory. Although it is obvious that a finding is more convincing when it was predicted, breakthrough findings have been made through exploration with limited a priori hypotheses. Imposing the restrictions that are applied to clinical trials may thus discourage exploratory research, whereas some kind of preregistration should be considered to minimize publication/reporting bias for confirmatory studies.

Call for Changes in Preclinical PH Confirmatory Studies

Scientific irreproducibility is a growing concern among academics and in the general population.[92] Poorly designed preclinical confirmatory studies likely contribute to experimental irreproducibility, wasted resources, and erroneous conclusions about the treatment effects.[93] Practical solutions to improve preclinical research quality and research translation have been proposed,[27,28,32] including guidelines and checklists to improve methodology and reporting,[26,27] as well as multicentre preclinical studies,[94] systematic reviews, and meta-analyses.[86,95] Most of these recommendations are related to good scientific practice in general and are often assumed to have been implemented in high-ranked scientific research. Indeed, reported rigor is not the same as actual rigor—a distinction that sometimes is impossible to make for the readers. However, biased effect estimates from studies that do not report important study design elements have been demonstrated,[95–98] arguing for the importance of reported rigor. Nonetheless, despite widespread recognition of the importance of these experimental design elements, confirmatory preclinical PH studies rarely report having implemented stringent methodologies, and their study designs have not improved during the past decade (Figure 3).[29,96,99] Furthermore, citation counts suggest that researchers may overlook crucial methodological aspects of preclinical PH studies and that greater rigor does not necessarily translate into greater scientific influence (Figure 7). Some reviewers are also tolerant of studies in a small number of animals of a single sex and strains or with methodological flaws described above. Journal reviewers and editors directly serve as ultimate gatekeepers of research findings. However, few journals enforce methodological and reporting guidelines for confirmatory studies, despite being associated with significant improvements in methodological quality in other fields of preclinical research.[28,29,100]

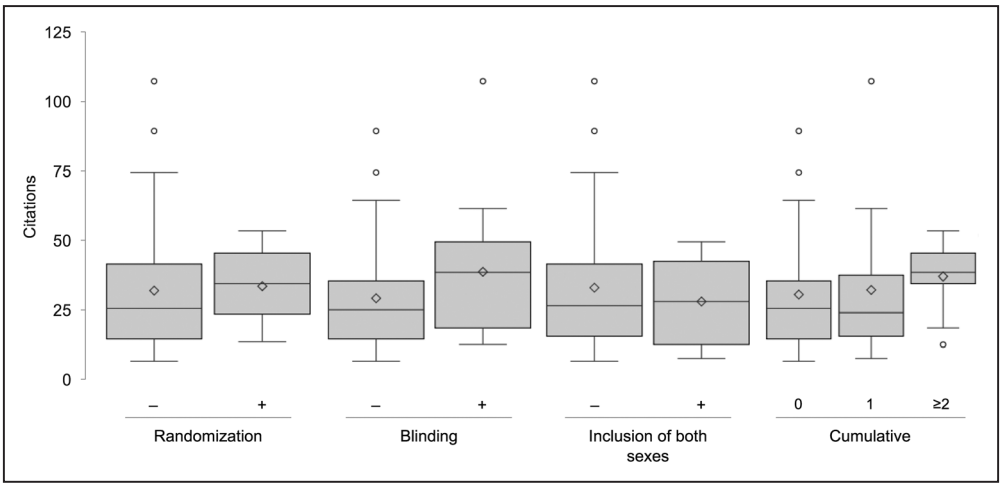


Figure 7. Box–whisker plot of the number of citations by methodological rigor of index preclinical pulmonary hypertension study (n=135). Diamond symbol plotted at mean. Outliers identified as beyond 1.5 interquartile range. Cumulative refers to sum of randomization, blinding, sample size estimation, and inclusion of both sexes (each contributing 1 point). Adapted from Ramirez et al[29] with permission. Copyright © 2017, the American Heart Association.

Implementing such requirements in preclinical PH research will involve a major paradigm shift for scientists, their institutions, journals, and funding agencies. For instance, many researchers were trained in an era where little attention was paid to the risk of bias in preclinical science. Although most universities offer courses in applied statistics and data analysis, there is a need for science-specific training in these aspects. Similarly, experimental rigor rather than flashy results should be prioritized during the article adjudication process. It should also be recognized that more robust preclinical studies generally require more time and involve more resources. Early-career scientists may, therefore, not be expected to publish as many high-methodologically rigorous articles, which should be taken into account by academic institutions and funding agencies. This transition should, therefore, be accompanied by changes in research funding, such as those undertaken by the National Institutes of Health.[101]

Conclusion

In preclinical PH research, methodological sources of potential bias and imprecision are prevalent and frequently overlooked by researchers. Although unlikely to be the sole cause for clinical failures, they are likely contributors to the significant discordance between preclinical and clinical results. Concerted efforts to address this problem are needed for more effective translation of preclinical research findings into sustainable improvements in patient outcomes. Rigorous study designs, methodological standardization, appropriate data interpretation, adopting statistical analysis plans, and transparent reporting of preclinical confirmatory studies are important prerequisites for this goal. Support from all relevant stakeholders, including researchers, peer reviewers, editors, academic institutions, funding agencies, animal ethics authorities, and the industry, is thus essential to fostering drug development in PH.

Acknowledgments

All authors have written and edited the article.

Disclosures

S. Provencher is a clinician scientist of the Fonds de Recherche en Santé du Québec and has received research grants from Actelion Pharmaceuticals and Boehringer Ingelheim. He has also received speaker fees from Actelion Pharmaceuticals.
S. Bonnet holds a Canadian Research Chair in translational research in pulmonary vascular diseases at Université Laval. He also received consultant fees from Janssen. The other authors report no conflicts.

References

1. Schermuly RT, Ghofrani HA, Wilkins MR, Grimminger F. Mechanisms of disease: pulmonary arterial hypertension. *Nat Rev Cardiol*. 2011;8:443–455. doi: 10.1038/nrcardio.2011.87.
2. Galiè N, Humbert M, Vachiery JL, et al. 2015 ESC/ERS guidelines for the diagnosis and treatment of pulmonary hypertension: the joint task force for the diagnosis and treatment of pulmonary hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS): endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT). *Eur Respir J*. 2015;46:903–975. doi: 10.1183/13993003.01032-2015.
3. Humbert M, Sitbon O, Simonneau G. Treatment of pulmonary arterial hypertension. *N Engl J Med*. 2004;351:1425–1436. doi: 10.1056/NEJMra040291.
4. Lajoie AC, Lauzière G, Lega JC, Lacasse Y, Martin S, Simard S, Bonnet S, Provencher S. Combination therapy versus monotherapy for pulmonary arterial hypertension: a meta-analysis. *Lancet Respir Med*. 2016;4:291–305. doi: 10.1016/S2213-2600(16)00027-8.
5. Galiè N, Manes A, Negro L, Palazzini M, Bacchi-Reggiani ML, Branzi A. A meta-analysis of randomized controlled trials in pulmonary arterial hypertension. *Eur Heart J*. 2009;30:394–403. doi: 10.1093/eurheartj/ehp022.
6. Humbert M, Sitbon O, Chaouat A, et al. Survival in patients with idiopathic, familial, and anorexigen-associated pulmonary arterial hypertension in the modern management era. *Circulation*. 2010;122:156–163. doi: 10.1161/CIRCULATIONAHA.109.911818.
7. Humbert M, Sitbon O, Yaïci A, et al; French Pulmonary Arterial Hypertension Network. Survival in incident and prevalent cohorts of patients with pulmonary arterial hypertension. *Eur Respir J*. 2010;36:549–555. doi: 10.1183/09031936.00057010.
8. Benza RL, Miller DP, Barst RJ, Badesch DB, Frost AE, McGoon MD. An evaluation of long-term survival from time of diagnosis in pulmonary arterial hypertension from the REVEAL Registry. *Chest*. 2012;142:448–456. doi: 10.1378/chest.11-1460.
9. Hoeper MM, Barst RJ, Bourge RC, et al. Imatinib mesylate as add-on therapy for pulmonary arterial hypertension: results of the randomized IMPRES study. *Circulation*. 2013;127:1128–1138. doi: 10.1161/CIRCULATIONAHA.112.000765.
10. Ghofrani HA, Morrell NW, Hoeper MM, Olschewski H, Peacock AJ, Barst RJ, Shapiro S, Golpon H, Toshner M, Grimminger F, Pascoe S. Imatinib in pulmonary arterial hypertension patients with inadequate response to established therapy. *Am J Respir Crit Care Med*. 2010;182:1171–1177. doi: 10.1164/rccm.201001-0123OC.
11. Granton J, Langleben D, Kutryk MB, Camack N, Galipeau J, Courtman DW, Stewart DJ. Endothelial NO-synthase gene-enhanced progenitor cell therapy for pulmonary arterial hypertension: the PHACeT trial. *Circ Res*. 2015;117:645–654. doi: 10.1161/CIRCRESAHA.114.305951.
12. Michelakis ED, Gurtu V, Webster L, Barnes G, Watson G, Howard L, Cupitt J, Paterson I, Thompson RB, Chow K, et al. Inhibition of pyruvate dehydrogenase kinase improves pulmonary arterial hypertension in genetically susceptible patients. *Sci Transl Med* 2017;9:eaao4583.
13. Pharmaceutical Research and Manufacturers of America. Annual report 2016. http://phrma-docs.phrma.org/sites/default/files/pdf/PhRMA_2016_Special_301_Submission.pdf.
14. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. *Nat Biotechnol*. 2014;32:40–51. doi: 10.1038/nbt.2786.
15. DiMasi JA, Feldman L, Seckler A, Wilson A. Trends in risks associated with new drug development: success rates for investigational drugs. *Clin Pharmacol Ther*. 2010;87:272–277. doi: 10.1038/clpt.2009.295.
16. Contopoulos-Ioannidis DG, Ntzani E, Ioannidis JP. Translation of highly promising basic science research into clinical applications. *Am J Med*. 2003;114:477–484.
17. Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical R&D. *Nat Rev Drug Discov*. 2011;10:428–438. doi: 10.1038/nrd3405.
18. Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. *PLoS Biol*. 2015;13:e1002165. doi: 10.1371/journal.pbio.1002165.
19. Bonnet S, Provencher S, Guignabert C, Perros F, Boucherat O, Schermuly RT, Hassoun PM, Rabinovitch M, Nicolls MR, Humbert M. Translating research into improved patient care in pulmonary arterial hypertension. *Am J Respir Crit Care Med*. 2017;195:583–595. doi: 10.1164/rccm.201607-1515PP.
20. Perrin S. Preclinical research: make mouse studies work. *Nature*. 2014;507:423–425. doi: 10.1038/507423a.
21. Moeller A, Ask K, Warburton D, Gauldie J, Kolb M. The bleomycin animal model: a useful tool to investigate treatment options for idiopathic pulmonary fibrosis? *Int J Biochem Cell Biol*. 2008;40:362–382. doi: 10.1016/j.biocel.2007.08.011.
22. Hwang IC, Kim YJ, Park JB, Yoon YE, Lee SP, Kim HK, Cho GY, Sohn DW. Pulmonary hemodynamics and effects of phosphodiesterase type 5 inhibition in heart failure: a meta-analysis of randomized trials. *BMC Cardiovasc Disord*. 2017;17:150. doi: 10.1186/s12872-017-0576-4.
23. Xiong B, Nie D, Cao Y, Zou Y, Yao Y, Tan J, Qian J, Rong S, Wang C, Huang J. Clinical and hemodynamic effects of endothelin receptor antagonists in patients with heart failure. *Int Heart J*. 2017;58:400–408. doi: 10.1536/ihj.16-307.
24. Henderson VC, Kimmelman J, Fergusson D, Grimshaw JM, Hackam DG. Threats to validity in the design and conduct of preclinical efficacy

studies: a systematic review of guidelines for in vivo animal experiments. *PLoS Med*. 2013;10:e1001489. doi: 10.1371/journal.pmed.1001489.
25. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2:e124. doi: 10.1371/journal.pmed.0020124.
26. Rigor and Reproducibility: Principles and Guidelines for Reporting Preclinical Research. https://www.Nih.Gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research.
27. Bolli R. New initiatives to improve the rigor and reproducibility of articles published in circulation research. *Circ Res*. 2017;121:472–479. doi: 10.1161/CIRCRESAHA.117.311678.
28. Vahidy F, Schäbitz WR, Fisher M, Aronowski J. Reporting standards for preclinical studies of stroke therapy. *Stroke*. 2016;47:2435–2438. doi: 10.1161/STROKEAHA.116.013643.
29. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald ZD, Moreland R, Simard T, Clancy AA, Russo JJ, Welch VA, Wells GA, Hibbert B. Methodological rigor in preclinical cardiovascular studies: targets to enhance reproducibility and promote research translation. *Circ Res*. 2017;120:1916–1926. doi: 10.1161/CIRCRESAHA.117.310628.
30. Sztuka K, Jasińska-Stroschein M. Animal models of pulmonary arterial hypertension: a systematic review and meta-analysis of data from 6126 animals. *Pharmacol Res*. 2017;125:201–214. doi: 10.1016/j.phrs.2017.08.003.
31. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biol*. 2014;12:e1001863. doi: 10.1371/journal.pbio.1001863.
32. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med*. 2010;152:726–732. doi: 10.7326/0003-4819-152-11-201006010-00232.
33. Stenmark KR, Meyrick B, Galie N, Mooi WJ, McMurtry IF. Animal models of pulmonary arterial hypertension: the hope for etiological discovery and pharmacological cure. *Am J Physiol Lung Cell Mol Physiol*. 2009;297:L1013–L1032. doi: 10.1152/ajplung.00217.2009.
34. Long L, Ormiston ML, Yang X, et al. Selective enhancement of endothelial BMPR-II with BMP9 reverses pulmonary arterial hypertension. *Nat Med*. 2015;21:777–785. doi: 10.1038/nm.3877.
35. Taraseviciene-Stewart L, Kasahara Y, Alger L, Hirth P, Mc Mahon G, Waltenberger J, Voelkel NF, Tuder RM. Inhibition of the VEGF receptor 2 combined with chronic hypoxia causes cell death-dependent pulmonary endothelial cell proliferation and severe pulmonary hypertension. *FASEB J*. 2001;15:427–438. doi: 10.1096/fj.00-0343com.
36. Ruiter G, Ying Wong Y, de Man FS, Louis Handoko M, Jaspers RT, Postmus PE, Westerhof N, Niessen HW, van der Laarse WJ, Vonk-Noordegraaf A. Right ventricular oxygen supply parameters are decreased in human and experimental pulmonary hypertension. *J Heart Lung Transplant*. 2013;32:231–240. doi: 10.1016/j.healun.2012.09.025.
37. Zungu-Edmondson M, Shults NV, Melnyk O, Suzuki YJ. Natural reversal of pulmonary vascular remodeling and right ventricular remodeling in SU5416/hypoxia-treated Sprague-Dawley rats. *PLoS One*. 2017;12:e0182551. doi: 10.1371/journal.pone.0182551.
38. de Raaf MA, Schalij I, Gomez-Arroyo J, Rol N, Happé C, de Man FS, Vonk-Noordegraaf A, Westerhof N, Voelkel NF, Bogaard HJ. SuHx rat model: partly reversible pulmonary hypertension and progressive intima obstruction. *Eur Respir J*. 2014;44:160–168. doi: 10.1183/09031936.00204813.
39. Benza RL, Gomberg-Maitland M, Demarco T, Frost AE, Torbicki A, Langleben D, Pulido T, Correa-Jaque P, Passineau MJ, Wiener HW, Tamari M, Hirota T, Kubo M, Tiwari HK. Endothelin-1 pathway polymorphisms and outcomes in pulmonary arterial hypertension. *Am J Respir Crit Care Med*. 2015;192:1345–1354. doi: 10.1164/rccm.201501-0196OC.
40. Jiang B, Deng Y, Suen C, Taha M, Chaudhary KR, Courtman DW, Stewart DJ. Marked strain-specific differences in the SU5416 rat model of severe pulmonary arterial hypertension. *Am J Respir Cell Mol Biol*. 2016;54:461–468. doi: 10.1165/rcmb.2014-0488OC.
41. Mitani Y, Mutlu A, Russell JC, Brindley DN, DeAlmeida J, Rabinovitch M. Dexfenfluramine protects against pulmonary hypertension in rats. *J Appl Physiol (1985)*. 2002;93:1770–1778. doi: 10.1152/japplphysiol.00500.2002.
42. Pullamsetti SS, Berghausen EM, Dabral S, Tretyn A, Butrous E, Savai R, Butrous G, Dahal BK, Brandes RP, Ghofrani HA, Weissmann N, Grimminger F, Seeger W, Rosenkranz S, Schermuly RT. Role of Src tyrosine kinases in experimental pulmonary hypertension. *Arterioscler Thromb Vasc Biol*. 2012;32:1354–1365. doi: 10.1161/ATVBAHA.112.248500.
43. Lusis AJ, Seldin MM, Allayee H, et al. The hybrid mouse diversity panel: a resource for systems genetics analyses of metabolic and cardiovascular traits. *J Lipid Res*. 2016;57:925–942. doi: 10.1194/jlr.R066944.
44. Courboulin A, Kang C, Baillard O, Bonnet S, Bonnet P. Increasing pulmonary artery pulsatile flow improves hypoxic pulmonary hypertension in piglets. *J Vis Exp*. 2015:e52571. doi: 10.3791/52571.
45. Boulate D, Perros F, Dorfmuller P, Arthur-Ataam J, Guihaire J, Lamrani L, Decante B, Humbert M, Eddahibi S, Dartevelle P, Fadel E, Mercier O. Pulmonary microvascular lesions regress in reperfused chronic thromboembolic pulmonary hypertension. *J Heart Lung Transplant*. 2015;34:457–467. doi: 10.1016/j.healun.2014.07.005.
46. Deng L, Blanco FJ, Stevens H, et al. MicroRNA-143 activation regulates smooth muscle and endothelial cell crosstalk in pulmonary arterial hypertension. *Circ Res*. 2015;117:870–883. doi: 10.1161/CIRCRESAHA.115.306806.
47. Wohrley JD, Frid MG, Moiseeva EP, Orton EC, Belknap JK, Stenmark KR. Hypoxia selectively induces proliferation in a specific subpopulation of smooth muscle cells in the bovine neonatal pulmonary arterial media. *J Clin Invest*. 1995;96:273–281. doi: 10.1172/JCI118031.
48. George MP, Champion HC, Simon M, Guyach S, Tarantelli R, Kling HM, Brower A, Janssen C, Murphy J, Carney JP, Morris A, Gladwin MT, Norris KA. Physiologic changes in a nonhuman primate model of HIV-associated pulmonary arterial hypertension. *Am J Respir Cell Mol Biol*. 2013;48:374–381. doi: 10.1165/rcmb.2011-0434OC.
49. Van der Velden J, Snibson KJ. Airway disease: the use of large animal models for drug discovery. *Pulm Pharmacol Ther*. 2011;24:525–532. doi: 10.1016/j.pupt.2011.02.001.
50. Boulate D, Arthur Ataam J, Connolly AJ, Giraldeau G, Amsallem M, Decante B, Lamrani L, Fadel E, Dorfmuller P, Perros F, Haddad F, Mercier O. Early development of right ventricular ischemic lesions in a novel large animal model of acute right heart failure in chronic thromboembolic pulmonary hypertension. *J Card Fail*. 2017;23:876–886. doi: 10.1016/j.cardfail.2017.08.447.
51. Rothman A, Wiencek RG, Davidson S, Evans WN, Restrepo H, Sarukhanov V, Mann D. Challenges in the development of chronic pulmonary hypertension models in large animals. *Pulm Circ*. 2017;7:156–166. doi: 10.1086/690099.
52. Nichols JE, Niles JA, Vega SP, Cortiella J. Novel in vitro respiratory models to study lung development, physiology, pathology and toxicology. *Stem Cell Res Ther*. 2013;4(suppl 1):S7. doi: 10.1186/scrt368.
53. Patel B, Gauvin R, Absar S, Gupta V, Gupta N, Nahar K, Khademhosseini A, Ahsan F. Computational and bioengineered lungs as alternatives to whole animal, isolated organ, and cell-based lung models. *Am J Physiol Lung Cell Mol Physiol*. 2012;303:L733–L747. doi: 10.1152/ajplung.00076.2012.
54. Pinn VW. Sex and gender factors in medical studies: implications for health and clinical practice. *JAMA*. 2003;289:397–400.
55. Humbert M, Sitbon O, Chaouat A, et al. Pulmonary arterial hypertension in France: results from a national registry. *Am J Respir Crit Care Med*. 2006;173:1023–1030. doi: 10.1164/rccm.200510-1668OC.
56. Austin ED, Lahm T, West J, Tofovic SP, Johansen AK, Maclean MR, Alzoubi A, Oka M. Gender, sex hormones and pulmonary hypertension. *Pulm Circ*. 2013;3:294–314. doi: 10.4103/2045-8932.114756.
57. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald Z, Simard T, Clancy AA, Russo JJ, Welch V, Wells GA, Hibbert B. Sex bias is increasingly prevalent in preclinical cardiovascular research: implications for translational medicine and health equity for women: a systematic assessment of leading cardiovascular journals over a 10-year period. *Circulation*. 2017;135:625–626. doi: 10.1161/CIRCULATIONAHA.116.026668.
58. Zucker I, Beery AK. Males still dominate animal studies. *Nature*. 2010;465:690. doi: 10.1038/465690a.
59. Clayton JA. Studying both sexes: a guiding principle for biomedicine. *FASEB J*. 2016;30:519–524. doi: 10.1096/fj.15-279554.
60. Prendergast BJ, Onishi KG, Zucker I. Female mice liberated for inclusion in neuroscience and biomedical research. *Neurosci Biobehav Rev*. 2014;40:1–5. doi: 10.1016/j.neubiorev.2014.01.001.
61. Klein SL, Schiebinger L, Stefanick ML, Cahill L, Danska J, de Vries GJ, Kibbe MR, McCarthy MM, Mogil JS, Woodruff TK, Zucker I. Opinion: sex inclusion in basic research drives discovery. *Proc Natl Acad Sci USA*. 2015;112:5257–5258. doi: 10.1073/pnas.1502843112.
62. Galiè N, Barberà JA, Frost AE, et al; AMBITION Investigators. Initial use of ambrisentan plus tadalafil in pulmonary arterial hypertension. *N Engl J Med*. 2015;373:834–844. doi: 10.1056/NEJMoa1413687.
63. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*. 1995;273:408–412.
64. Pildal J, Hróbjartsson A, Jørgensen KJ, Hilden J, Altman DG, Gøtzsche PC. Impact of allocation concealment on conclusions drawn from

meta-analyses of randomized trials. *Int J Epidemiol*. 2007;36:847–857. doi: 10.1093/ije/dym087.
65. Crossley NA, Sena E, Goehler J, Horn J, van der Worp B, Bath PM, Macleod M, Dirnagl U. Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach. *Stroke*. 2008;39:929–934. doi: 10.1161/STROKEAHA.107.498725.
66. Rooke ED, Vesterinen HM, Sena ES, Egan KJ, Macleod MR. Dopamine agonists in animal models of parkinson's disease: a systematic review and meta-analysis. *Parkinsonism Relat Disord*. 2011;17:313–320.
67. Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG, Gluud C, Martin RM, Wood AJ, Sterne JA. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ*. 2008;336:601–605. doi: 10.1136/bmj.39465.451748.AD.
68. ATS Committee on Proficiency Standards for Clinical Pulmonary Function Laboratories. ATS statement: guidelines for the six-minute walk test. *Am J Respir Crit Care Med*. 2002;166:111–117. doi: 10.1164/ajrccm.166.1.at1102.
69. Howard LS, Grapsa J, Dawson D, Bellamy M, Chambers JB, Masani ND, Nihoyannopoulos P, Simon R Gibbs J. Echocardiographic assessment of pulmonary hypertension: standard operating procedure. *Eur Respir Rev*. 2012;21:239–248. doi: 10.1183/09059180.00003912.
70. Bashore TM, Bates ER, Berger PB, et al; American College of Cardiology. Task Force on Clinical Expert Consensus Documents. American College of Cardiology/Society for Cardiac Angiography and interventions clinical expert consensus document on cardiac catheterization laboratory standards. A report of the American College of Cardiology Task Force on Clinical Expert Consensus Documents. *J Am Coll Cardiol*. 2001;37:2170–2214.
71. McLaughlin VV, Badesch DB, Delcroix M, Fleming TR, Gaine SP, Galiè N, Gibbs JS, Kim NH, Oudiz RJ, Peacock A, Provencher S, Sitbon O, Tapson VF, Seeger W. End points and clinical trial design in pulmonary arterial hypertension. *J Am Coll Cardiol*. 2009;54:S97–S107. doi: 10.1016/j.jacc.2009.04.007.
72. Bagshaw RJ, Cox RH. Baroreceptor reflexes and pulmonary hemodynamics during halothane and halothane-nitrous oxide anesthesia in the dog. *Anesth Analg*. 1981;60:701–709.
73. Rich GF, Roos CM, Anderson SM, Daugherty MO, Uncles DR. Direct effects of intravenous anesthetics on pulmonary vascular resistance in the isolated rat lung. *Anesth Analg*. 1994;78:961–966.
74. Maruyama K, Maruyama J, Yokochi A, Muneyuki M, Miyasaka K. Vasodilatory effects of ketamine on pulmonary arteries in rats with chronic hypoxic pulmonary hypertension. *Anesth Analg*. 1995;80:786–792.
75. Armstrong PW, Roessig L, Patel MJ, Anstrom KJ, Butler J, Voors AA, Lam CSP, Ponikowski P, Temple T, Pieske B, Ezekowitz J, Hernandez AF, Koglin J, O'Connor CM. A multicenter, randomized, double-blind, placebo-controlled trial of the efficacy and safety of the oral soluble guanylate cyclase stimulator: the VICTORIA trial. *JACC Heart Fail*. 2018;6:96–104. doi: 10.1016/j.jchf.2017.08.013.
76. Liu S, Zhen G, Meloni BP, Campbell K, Winn HR. Rodent stroke model guidelines for preclinical stroke trials (1st edition). *J Exp Stroke Transl Med*. 2009;2:2–27.
77. Ludolph AC, Bendotti C, Blaugrund E, et al. Guidelines for preclinical animal research in ALS/MND: a consensus meeting. *Amyotroph Lateral Scler*. 2010;11:38–45. doi: 10.3109/17482960903545334.
78. Moreno B, Espejo C, Mestre L, Suardiaz M, Clemente D, de Castro F, Fernández-Fernández Ó, Montalban X, Villoslada P, Guaza C; Spanish Network for MS. [Guidelines on the appropriate use of animal models for developing therapies in multiple sclerosis]. *Rev Neurol*. 2012;54:114–124.
79. Holm S. A simple sequentially rejective multiple test procedure. *Scand J of Stat*. 1979;6:65–70.
80. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988;75:800–802.
81. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*. 1988;75:383–396.
82. Newell DJ. Intention-to-treat analysis: implications for quantitative and qualitative research. *Int J Epidemiol*. 1992;21:837–841.
83. Porta N, Bonet C, Cobo E. Discordance between reported intention-to-treat and per protocol analyses. *J Clin Epidemiol*. 2007;60:663–669. doi: 10.1016/j.jclinepi.2006.09.013.
84. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31:337–350. doi: 10.1007/s10654-016-0149-3.
85. Sellke T, Bayarri M, Berger J. Calibration of p values for testing precise null hypotheses. *Am Stat*. 2001;55:62–71.
86. Sandercock P, Roberts I. Systematic reviews of animal experiments. *Lancet*. 2002;360:586. doi: 10.1016/S0140-6736(02)09812-4.
87. Sena ES, van der Worp HB, Bath PM, Howells DW, Macleod MR. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biol*. 2010;8:e1000344. doi: 10.1371/journal.pbio.1000344.
88. Decullier E, Lhéritier V, Chapuis F. Fate of biomedical research protocols and publication bias in France: retrospective cohort study. *BMJ*. 2005;331:19. doi: 10.1136/bmj.38488.385995.8F.
89. Weber EJ, Callaham ML, Wears RL, Barton C, Young G. Unpublished research from a medical specialty meeting: why investigators fail to publish. *JAMA*. 1998;280:257–259.
90. Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ*. 2005;330:753. doi: 10.1136/bmj.38356.424606.8F.
91. De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, Kotzin S, Laine C, Marusic A, Overbeke AJ, Schroeder TV, Sox HC, Van Der Weyden MB; International Committee of Medical Journal Editors. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *N Engl J Med*. 2004;351:1250–1251. doi: 10.1056/NEJMe048225.
92. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature*. 2016;533:452–454. doi: 10.1038/533452a.
93. Gonzalez Martin-Moro J. The science reproducibility crisis and the necessity to publish negative results. *Arch Soc Esp Oftalmol*. 2017;92:e75–e77.
94. Llovera G, Hofmann K, Roth S, et al. Results of a preclinical randomized controlled multicenter trial (pRCT): anti-CD49d treatment for acute brain ischemia. *Sci Transl Med*. 2015;7:299ra121. doi: 10.1126/scitranslmed.aaa9853.
95. Macleod MR, O'Collins T, Howells DW, Donnan GA. Pooling of animal experimental data reveals influence of study design and publication bias. *Stroke*. 2004;35:1203–1208. doi: 10.1161/01.STR.0000125719.25853.20.
96. Begley CG. Six red flags for suspect work. *Nature*. 2013;497:433–434. doi: 10.1038/497433a.
97. Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke*. 2008;39:2824–2829. doi: 10.1161/STROKEAHA.108.515957.
98. Macleod MR, O'Collins T, Horky LL, Howells DW, Donnan GA. Systematic review and metaanalysis of the efficacy of FK506 in experimental stroke. *J Cereb Blood Flow Metab*. 2005;25:713–721. doi: 10.1038/sj.jcbfm.9600064.
99. 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. doi: 10.1038/nrd3439-c1.
100. Minnerup J, Zentsch V, Schmidt A, Fisher M, Schäbitz WR. Methodological quality of experimental stroke studies published in the stroke journal: time trends and effect of the basic science checklist. *Stroke*. 2016;47:267–272. doi: 10.1161/STROKEAHA.115.011695.
101. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. *Nature*. 2014;505:612–613.

Circulation Research
JOURNAL OF THE AMERICAN HEART ASSOCIATION

Standards and Methodological Rigor in Pulmonary Arterial Hypertension Preclinical and Translational Research

Steeve Provencher, Stephen L. Archer, F. Daniel Ramirez, Benjamin Hibbert, Roxane Paulin, Olivier Boucherat, Yves Lacasse and Sébastien Bonnet

Circ Res. 2018;122:1021-1032
doi: 10.1161/CIRCRESAHA.117.312579
Circulation Research is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 0009-7330. Online ISSN: 1524-4571

Sex Bias Is Increasingly Prevalent in Preclinical Cardiovascular Research: Implications for Translational Medicine and Health Equity for Women

A Systematic Assessment of Leading Cardiovascular Journals Over a 10-Year Period

F. Daniel Ramirez, MD*
Pouya Motazedian, BSc*
Richard G. Jung, BSc
Pietro Di Santo, MD
Zachary MacDonald, BSc
Trevor Simard, MD
Aisling A. Clancy, MD, MSc
Juan J. Russo, MD
Vivian Welch, PhD
George A. Wells, PhD
Benjamin Hibbert, MD, PhD

Ensuring that women are adequately represented in clinical trials is recognized as essential for sex equity in health. However, the use of female animal models and sex-based reporting have not been equally enforced in preclinical stages of research, which often precede and inform clinical trials. In 2014, the National Institutes of Health announced that it would require that sex be considered as a biological variable in applications for preclinical research funding,[1] yet a reluctance to include female animal models in preclinical experiments persists. Inappropriately inferring experimental findings to both sexes when a single sex is studied or when sex is not specified has the potential to disadvantage women by skewing our understanding of disease processes toward male-predominant patterns and by reducing the likelihood of female-specific therapeutics advancing to the clinical realm. We therefore systematically examined all preclinical cardiovascular studies published in American Heart Association journals with archives spanning at least 10 years (*Circulation*, *Circulation Research*, *Hypertension*, *Stroke*, and *Arteriosclerosis, Thrombosis, and Vascular Biology* [*ATVB*]) for evidence of sex bias.

Full articles published between July 2006 and June 2016 and reporting original data from in vivo experiments in nonhuman mammals on pathophysiology, genetics, or therapeutic interventions directly relevant to a specific cardiovascular disorder in humans were included. Studies on physiological or genetic characteristics were included if they proposed potential therapeutic applications or implications of the study findings. Each journal article was independently reviewed and prespecified data were extracted by using standardized case report forms by 2 authors, including the cardiovascular disease investigated, animal model(s) used and their sex, whether study samples were sex matched (in studies using both sexes), whether at least 1 study result was reported by sex, and whether the use of a single sex was reported as a limitation (in single-sex studies). Interrater agreement was calculated using the Cohen κ-statistic and percentage agreement. Discrepancies were resolved by consensus or independent adjudication. Categorical variables were compared via χ^2 tests. Temporal patterns were evaluated via Pearson correlation or Cochrane-Armitage trend tests. All analyses were performed by using SAS 9.4 (SAS Institute Inc) with the use of an α-level of 0.05 to define statistical significance.

Of 28636 articles screened, 3396 met inclusion criteria and were analyzed. Interrater agreement for study inclusion before resolution was 94.5% (κ=0.72; 95% confidence interval, 0.70–0.73). The sex of the animals used was not reported in 20.0% of studies. Males were exclusively used in 71.6% of studies in which sex was reported, whereas females were exclusively used in 12.9% and both sexes in 15.5%. Sex matching of animals was reported in 17.1% of studies that included both sexes. Restricting this analysis to the 988 studies of therapeutic interventions did not appreciably change these distributions. When stratified by the cardiovascular disease studied, males were exclusively used significantly more often than females or both sexes in all cases, with the exception of atherosclerosis and arrhythmia. When stratified by the animal model

*Dr Ramirez and P. Motazedian contributed equally.

Correspondence to: Benjamin Hibbert, MD, PhD, University of Ottawa Heart Institute, 40 Ruskin Street, H-4238, Ottawa, ONT, Canada K1Y 4W7. E-mail bhibbert@ottawaheart.ca

Key Words: bias (epidemiology) ■ cardiovascular diseases ■ publications ■ research ■ sex

used, males were exclusively used significantly more often when mice, rats, rabbits, or a combination of animal models was used (accounting for 95.0% of studies).

The number of preclinical studies published yearly over the past decade has remained stable; however, there has been a significant increase in the proportion using exclusively males and a significant decrease in the proportion using exclusively females. There have not been significant changes in the proportions of studies using both sexes or reporting the sex of the animals used (Figure). These trends were unchanged when studies in which sex was not specified were excluded.

Sex-based reporting of results occurred in 35.1% of studies in which sex was reported using the liberal criterion of any results being reported in reference to the sex of the animals used at least once. There was no appreciable difference in the prevalence of sex-based reporting before and after the publication of planned National Institutes of Health policies to improve sex inclusion in preclinical research[1] (34.9% versus 35.7%, respectively). Of the 2297 studies in which only 1 sex was reportedly used, 1.6% mentioned this aspect of the study design as a potential limitation.

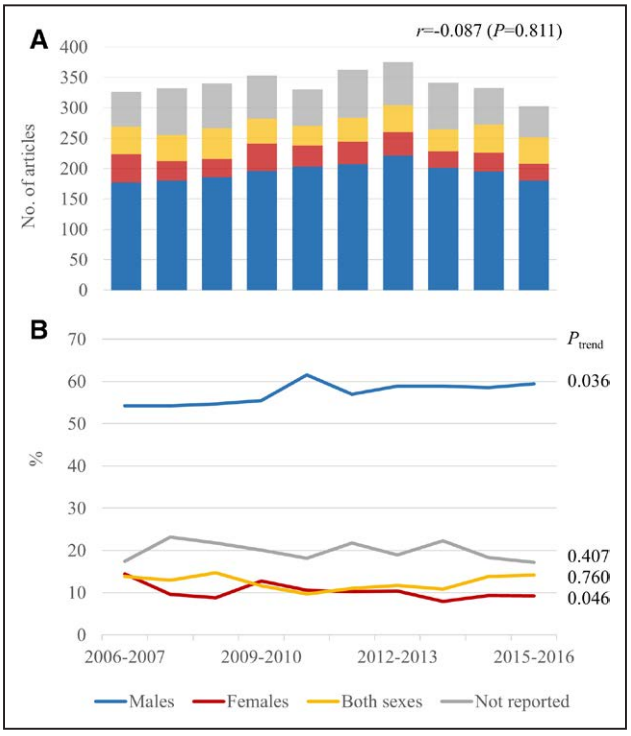


Figure. Temporal patterns in the sex of animals used in preclinical cardiovascular research over a 10-year period.

A, Absolute number of preclinical studies published and sex of animals used as a function of time. *r* denotes Pearson correlation coefficient for the total number of studies over time. **B,** Sex of animals used in preclinical studies over time (%). P_{trend} calculated for the proportion of studies using exclusively males, exclusively females, both sexes, and not reporting the sex of the animals used over time.

Although the need to include women in clinical trials is now a well-established requirement to enhance external validity, analogous standards have not gained a foothold in preclinical stages. Reasons put forward for the preferential use of males include increased variability attributable to fluctuating gonadal hormone levels throughout the estrous cycle and sample size implications of planning sex-based analyses.[2,3] Evidence to support these concerns is limited for most experimental settings, however.[3–5]

Our review of preclinical research published in leading cardiovascular journals over the past 10 years demonstrates that sex bias is prevalent and increasing, contrasting with advances made in clinical research designs and reporting. Concerted efforts on the part of researchers, journal editors, peer reviewers, universities, industry, professional organizations, and patient advocacy groups are urgently needed to address this discrepancy.

DISCLOSURES

None.

AFFILIATIONS

From Division of Cardiology, University of Ottawa Heart Institute, ON, Canada (F.D.R., P.D.S., T.S., J.J.R., B.H.); CAPITAL Research Group, University of Ottawa Heart Institute, ONT, Canada (F.D.R., P.M., R.G.J., P.D.S., Z.M., T.S., J.J.R., B.H.); School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, ONT, Canada (F.D.R., V.W., G.A.W.); Vascular Biology and Experimental Medicine Laboratory, University of Ottawa Heart Institute, ONT, Canada (R.G.J., T.S., B.H.); Department of Cellular and Molecular Medicine, University of Ottawa, ONT, Canada (R.G.J., T.S., B.H.); Department of Obstetrics and Gynecology, University of Ottawa, ONT, Canada (A.A.C.); Bruyère Research Institute, University of Ottawa, ONT, Canada (V.W.); Centre for Global Health, University of Ottawa, ONT, Canada (V.W.); and Cardiovascular Research Methods Centre, University of Ottawa Heart Institute, ONT, Canada (G.A.W.).

FOOTNOTES

Circulation is available at http://circ.ahajournals.org.

REFERENCES

1. Clayton JA, Collins FS. Policy: NIH to balance sex in cell and animal studies. *Nature*. 2014;509:282–283.
2. Zucker I, Beery AK. Males still dominate animal studies. *Nature*. 2010;465:690. doi: 10.1038/465690a.
3. Clayton JA. Studying both sexes: a guiding principle for biomedicine. *FASEB J*. 2016;30:519–524. doi: 10.1096/fj.15-279554.
4. Prendergast BJ, Onishi KG, Zucker I. Female mice liberated for inclusion in neuroscience and biomedical research. *Neurosci Biobehav Rev*. 2014;40:1–5. doi: 10.1016/j.neubiorev.2014.01.001.
5. Itoh Y, Arnold AP. Are females more variable than males in gene expression? Meta-analysis of microarray datasets. *Biol Sex Differ*. 2015;6:18. doi: 10.1186/s13293-015-0036-8.

Sex Bias Is Increasingly Prevalent in Preclinical Cardiovascular Research: Implications for Translational Medicine and Health Equity for Women: A Systematic Assessment of Leading Cardiovascular Journals Over a 10-Year Period

F. Daniel Ramirez, Pouya Motazedian, Richard G. Jung, Pietro Di Santo, Zachary MacDonald, Trevor Simard, Aisling A. Clancy, Juan J. Russo, Vivian Welch, George A. Wells and Benjamin Hibbert

Circulation. 2017;135:625-626
doi: 10.1161/CIRCULATIONAHA.116.026668
Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 0009-7322. Online ISSN: 1524-4539

Letter to the Editor

Letters to the Editor will be published, if suitable, as space permits. They should not exceed 1000 words (typed double-spaced) in length and may be subject to editing or abridgment.

Letter by Ramirez and Hibbert Regarding Article, "Consideration of Sex Differences in Design and Reporting of Experimental Arterial Pathology Studies: A Statement From the *Arteriosclerosis, Thrombosis, and Vascular Biology* Council"

To the Editor:

We read with interest the recent article titled "Consideration of Sex Differences in Design and Reporting of Experimental Arterial Pathology Studies: A Statement From the *Arteriosclerosis, Thrombosis, and Vascular Biology* Council" in which Robinet et al[1] review sex differences in arterial diseases in both humans and animal models. The authors provide suggestions for examining the influence of sex when designing and reporting preclinical arterial pathology experiments. They propose that both sexes should be included in experiments unless scientifically justified, the sex of the animals should be clearly reported, the studies should be adequately powered to observe sex differences, and the data should be analyzed and reported separately by sex. We commend the authors for their initiative and expect that their article will inform preclinical research in the field.

However, linked to the authors' recommendations is the perception that preclinical arterial disease researchers do not consider or report on sex as a biological variable—at least not sufficiently. Yet, data to support this are notably limited in their report. It has been suggested that cardiovascular researchers may already be more attuned to the importance of considering sex as a biological variable relative to researchers in other medical disciplines.[2] Without knowledge of preclinical arterial disease research and reporting practices, the importance and potential impact of the authors' suggestions are difficult to gauge.

In support of implementing, incentivizing, and possibly enforcing many of the proposals by Robinet et al, our group has recently shown that considering sex as a biological variable in preclinical experiments is indeed rare in cardiovascular research. We analyzed all preclinical studies published between July 2006 and June 2016 in leading American Heart Association journals, including *Arteriosclerosis, Thrombosis, and Vascular Biology* (*ATVB*). Studies were included if they represented original research published as full manuscripts, reported results of in vivo experiments in nonhuman mammals, and described findings with proposed therapeutic applications or implications to a specific cardiovascular disorder in humans. Two independent reviewers extracted all prespecified data from each article, including via review of all online supplementary material. Discrepancies were resolved by consensus or by an independent adjudicator before data analysis.[3,4] Of the 3396 articles included, 771 focused on atherosclerosis and arterial aneurysms/dissection, which are the subject of the article by Robinet et al.

Of these 771 preclinical arterial pathology studies, 447 (58.0%) were published in *ATVB*. Mice were most often used as the experimental animal models (90.1%). The sex of the animals was not specified in 145 studies (18.8%). When restricted to studies that did specify the sex of the animals, 347 (55.4%) used exclusively males, 128 (20.4%) used exclusively females, and 151 (24.1%) used both sexes.[3] Reporting of results by sex was observed in 290 (46.3%) using the liberal criterion of any results reported in reference to the sex of the animals used. Of the 475 studies that reported only using animals of one sex, 9 (1.9%) mentioned this aspect of the study design as a potential limitation. Temporal analyses demonstrate that, overall, the number of preclinical arterial pathology research articles published yearly in the cardiovascular journals examined did not change during the 10-year period studied (range, 65–96; r=0.121; P=0.738). However, the proportion of studies using exclusively males has increased over time (P_{trend}<0.0001), whereas the proportion of studies using exclusively females has decreased (P_{trend}<0.0001). The proportions of studies using animals of both sexes or not reporting the sex of the animals have not changed (P_{trend}=0.277 and 0.499, respectively). Finally, although we do not have data regarding adequate statistical power to detect sex differences specifically, more broadly, only 10 studies (1.3%) reported any sample size or a priori power calculation. Although temporal analyses of the proportion of studies reporting power estimations suggested a significant improvement over the 10-year period (P_{trend}=0.002), the maximum yearly proportion reporting this study design element was only 4.2%.

In summary, sex bias is not only prevalent but also increasing in preclinical arterial pathology research, as it is for preclinical cardiovascular research in general.[3,4] We stress that the journals selected for our analysis are among the most highly regarded in the cardiovascular sciences with editorial boards attuned to many of the issues raised by the National Institutes of Health,[5–7] rendering it possible (if not likely) that our estimates paint an optimistic picture of the current state of cardiovascular research.

Our data, therefore, add weight and a sense of urgency to the efforts of Robinet et al and the ATVB Council to encourage preclinical arterial pathology researchers to consider sex as a biological variable when designing and reporting experiments. We caution, however, that guidelines alone may not be sufficient to effect change.[8,9] Although journal editorial policies aimed at improving research methodology and reporting seem to be more effective (and have been pioneered by prominent journals such as *Nature*, *Stroke*, and *Circulation Research*),[3,10] it is likely that to realize meaningful progress in the methodological and reporting quality of preclinical studies, a concerted effort by researchers, academic institutions, funding agencies, and industry will be needed.

Disclosures

None.

(*Arterioscler Thromb Vasc Biol*. 2018;38:e99-e100. DOI: 10.1161/ATVBAHA.118.310942.)

F. Daniel Ramirez
Division of Cardiology
University of Ottawa Heart Institute
Ottawa, Canada

Benjamin Hibbert
Division of Cardiology
University of Ottawa Heart Institute
Ottawa, Canada

References

1. Robinet P, Milewicz DM, Cassis LA, Leeper NJ, Lu HS, Smith JD. Consideration of sex differences in design and reporting of experimental arterial pathology studies: a statement from the Arteriosclerosis, Thrombosis, and Vascular Biology Council. *Arterioscler Thromb Vasc Biol*. 2018;38:292–303. doi: 10.1161/ATVBAHA.117.309524.
2. Oertelt-Prigione S, Parol R, Krohn S, Preissner R, Regitz-Zagrosek V. Analysis of sex and gender-specific research reveals a common increase in publications and marked differences between disciplines. *BMC Med*. 2010;8:70. doi: 10.1186/1741-7015-8-70.
3. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald ZD, Moreland R, Simard T, Clancy AA, Russo JJ, Welch VA, Wells GA, Hibbert B. Methodological rigor in preclinical cardiovascular studies: targets to enhance reproducibility and promote research translation. *Circ Res*. 2017;120:1916–1926. doi: 10.1161/CIRCRESAHA.117.310628.
4. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald Z, Simard T, Clancy AA, Russo JJ, Welch V, Wells GA, Hibbert B. Sex bias is increasingly prevalent in preclinical cardiovascular research: implications for translational medicine and health equity for women: a systematic assessment of leading cardiovascular journals over a 10-year period. *Circulation*. 2017;135:625–626. doi: 10.1161/CIRCULATIONAHA.116.026668.
5. Daugherty A, Hegele RA, Mackman N, Rader DJ, Schmidt AM, Weber C. Complying with the National Institutes of Health Guidelines and Principles for Rigor and Reproducibility: refutations. *Arterioscler Thromb Vasc Biol*. 2016;36:1303–1304. doi: 10.1161/ATVBAHA.116.307906.
6. Vahidy F, Schäbitz WR, Fisher M, Aronowski J. Reporting standards for preclinical studies of stroke therapy. *Stroke*. 2016;47:2435–2438. doi: 10.1161/STROKEAHA.116.013643.
7. Bolli R. New initiatives to improve the rigor and reproducibility of articles published in Circulation Research. *Circ Res*. 2017;121:472–479. doi: 10.1161/CIRCRESAHA.117.311678.
8. Hirst A, Altman DG. Are peer reviewers encouraged to use reporting guidelines? A survey of 116 health research journals. *PLoS One*. 2012;7:e35621. doi: 10.1371/journal.pone.0035621.
9. Baker D, Lidster K, Sottomayor A, Amor S. Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS Biol*. 2014;12:e1001756. doi: 10.1371/journal.pbio.1001756.
10. Minnerup J, Zentsch V, Schmidt A, Fisher M, Schäbitz WR. Methodological quality of experimental stroke studies published in the Stroke journal: time trends and effect of the basic science checklist. *Stroke*. 2016;47:267–272. doi: 10.1161/STROKEAHA.115.011695.

Arteriosclerosis, Thrombosis, and Vascular Biology

JOURNAL OF THE AMERICAN HEART ASSOCIATION

Letter by Ramirez and Hibbert Regarding Article, "Consideration of Sex Differences in Design and Reporting of Experimental Arterial Pathology Studies: A Statement From the *Arteriosclerosis, Thrombosis, and Vascular Biology* Council"

F. Daniel Ramirez and Benjamin Hibbert

Arterioscler Thromb Vasc Biol. 2018;38:e99-e100
doi: 10.1161/ATVBAHA.118.310942
Arteriosclerosis, Thrombosis, and Vascular Biology is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 1079-5642. Online ISSN: 1524-4636

not required for publication in high-impact journals. *J Cereb Blood Flow Metab*. 2010;30:1619–1624. doi: 10.1038/jcbfm.2010.74.
7. Dirnagl U, Endres M. Found in translation: preclinical stroke research predicts human pathophysiology, clinical phenotypes, and therapeutic outcomes. *Stroke*. 2014;45:1510–1518. doi: 10.1161/STROKEAHA.113.004075.
8. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*. 2012;490:187–191. doi: 10.1038/nature11556.
9. Journals unite for reproducibility. *Nature*. 2014;515:7.
10. Macleod MR, Fisher M, O'Collins V, Sena ES, Dirnagl U, Bath PM, et al. Good laboratory practice: preventing introduction of bias at the bench. *Stroke*. 2009;40:e50–e52. doi: 10.1161/STROKEAHA.108.525386.
11. Lapchak PA. Recommendations and practices to optimize stroke therapy: developing effective translational research programs. *Stroke*. 2013;44:841–843. doi: 10.1161/STROKEAHA.112.680439.
12. Minnerup J, Zentsch V, Schmidt A, Fisher M, Schäbitz WR. Methodological quality of experimental stroke studies published in the stroke journal: time trends and effect of the basic science checklist. *Stroke*. 2016;47:267–272. doi: 10.1161/STROKEAHA.115.011695.
13. Food, Drug Administration HHS. International conference on harmonisation; choice of control group and related issues in clinical trials; availability. Notice. *Fed Regist*. 2001;66:24390–24391.
14. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. *Lancet*. 2002;359:614–618. doi: 10.1016/S0140-6736(02)07750-4.
15. Altman DG, Bland JM. Standard deviations and standard errors. *BMJ*. 2005;331:903. doi: 10.1136/bmj.331.7521.903.
16. Ahnstedt H, McCullough LD, Cipolla MJ. The importance of considering sex differences in translational stroke research. *Transl Stroke Res*. 2016;7:261–273. doi: 10.1007/s12975-016-0450-1.

KEY WORDS: Editorials ■ animal research ■ biomedical research ■ cerebrovascular circulation ■ checklist ■ consensus

Table 1. Checklist of Methodological and Reporting Aspects for Articles Submitted to *Stroke* Involving Preclinical Experimentation

Methodological and Reporting Aspects	Description of Procedures
Experimental groups and study timeline	☐ The experimental group(s) have been clearly defined in the article, including number of animals in each experimental arm of the study. ☐ An account of the control group is provided, and number of animals in the control group has been reported. If no controls were used, the rationale has been stated. ☐ An overall study timeline is provided.
Inclusion and exclusion criteria	☐ A priori inclusion and exclusion criteria for tested animals were defined and have been reported in the article.
Randomization	☐ Animals were randomly assigned to the experimental groups. If the work being submitted does not contain multiple experimental groups, or if random assignment was not used, adequate explanations have been provided. ☐ Type and methods of randomization have been described. ☐ Methods used for allocation concealment have been reported.
Blinding	☐ Blinding procedures have been described with regard to masking of group/treatment assignment from the experimenter. The rationale for nonblinding of the experimenter has been provided, if such was not feasible. ☐ Blinding procedures have been described with regard to masking of group assignment during outcome assessment.
Sample size and power calculations	☐ Formal sample size and power calculations were conducted based on a priori determined outcome(s) and treatment effect, and the data have been reported. OR A formal size assessment was not conducted and a rationale has been provided.
Data reporting and statistical methods	☐ Number of animals in each group: randomized, tested, lost to follow-up, or died have been reported. If the experimentation involves repeated measurements, the number of animals assessed at each time point is provided, for all experimental groups. ☐ Baseline data on assessed outcome(s) for all experimental groups have been reported. ☐ Details on important adverse events and death of animals during the course of experimentation have been provided, for all experimental arms. ☐ Statistical methods used have been reported. ☐ Numeric data on outcomes have been provided in text, or in a tabular format with the main article or as supplementary tables, in addition to the figures.
Experimental details, ethics, and funding statements	☐ Details on experimentation including stroke model, formulation and dosage of therapeutic agent, site and route of administration, use of anesthesia and analgesia, temperature control during experimentation, and postprocedural monitoring have been described. ☐ Different sex animals have been used. If not, the reason/justification is provided. ☐ Statements on approval by ethics boards and ethical conduct of studies have been provided. ☐ Statements on funding and conflicts of interests have been provided.

nature research

Corresponding author(s): Alexander Marson

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

No sample-size calculations were performed. Sample size was determined to be adequate based on the magnitude and consistency of measurable differences between groups.

2. Data exclusions

Describe any data exclusions.

On principle, data were only excluded for failed experiments, reasons for which included suboptimal activation and microbial contamination.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Replicate experiments were successful.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No randomization of mice. Mice analyzed were litter mates and sex-matched whenever possible.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Investigators were not blinded to mouse genotypes during experiments. Data reported for mouse experiments are not subjective but rather based on quantitative flow cytometry.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Flow cytometry data analyzed using FlowJo (v10).

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	IL2RA enhancer-edited mice will be available for distribution. No other unique materials were used in these studies.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	Supplementary Table 4 provided with manuscript contains information on all antibodies used in the study.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	Jurkat cells for CRISPRa screen obtained from Berkeley Cell Culture Facility. HuT78 and Jurkat (Fig. 2c,d) cells gift from Art Weiss. More information in Methods section.
b. Describe the method of cell line authentication used.	Short tandem repeat analysis was used to confirm the identity of the Jurkat and HuT78 T cell lines. Detailed information listed in Methods section.
c. Report whether the cell lines were tested for mycoplasma contamination.	Jurkat cells used for CRISPRa pooled and arrayed screens and HuT78 cells used for CRISPRa validation tested negative for mycoplasma. Detailed information listed in Methods section.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No cell lines used are listed in the database of commonly misidentified cell lines.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	Description of research mice used for experiments can be found in the relevant figure legends and Methods.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Human blood donors for experiments were anonymous.

nature research

Corresponding author(s): Alexander Marson

☐ Initial submission ☐ Revised version ☒ Final submission

Flow Cytometry Reporting Summary

Form fields will expand as needed. Please do not leave fields blank.

Data presentation

For all flow cytometry data, confirm that:

☒ 1. The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ 2. The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ 3. All plots are contour plots with outliers or pseudocolor plots.

☒ 4. A numerical value for number of cells or percentage (with statistics) is provided.

Methodological details

5. Describe the sample preparation.	Sample preparation listed in Methods
6. Identify the instrument used for data collection.	FACS Aria Fusion, Aria II, Fortessa and LSR II
7. Describe the software used to collect and analyze the flow cytometry data.	FACSDiva for collection and FlowJo (version 10) for analysis
8. Describe the abundance of the relevant cell populations within post-sort fractions.	1-3 million naive CD4+ T cells were sorted per mouse for activation experiments. Purity was assessed by staining for CD4 and other T cell markers, including CD69.
9. Describe the gating strategy used.	Relevant gating strategies shown in Supplementary Information

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information. ☒