DEVELOPMENT OF STRATEGIES FOR ASSESSING REPORTING IN BIOMEDICAL RESEARCH: MOVING TOWARD ENHANCING REPRODUCIBILITY

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

2016

By

Oscar Flórez-Vargas
School of Computer Science

Contents

Word count: 45,000

List of Figures

Thesis page *(Paper page)*

Chapter Two

Chapter Three

Chapter Four

Chapter Five

Chapter Six

List of Tables

DEVELOPMENT OF STRATEGIES FOR ASSESSING REPORTING IN BIOMEDICAL RESEARCH:
MOVING TOWARD ENHANCING REPRODUCIBILITY.

A thesis submitted to the University of Manchester
for the Doctor of Philosophy degree
in the Faculty of Engineering and Physical Sciences.

By
Oscar Flórez-Vargas
2016

Abstract

The idea that the same experimental findings can be reproduced by a variety of independent approaches is one of the cornerstones of science's claim to objective truth. However, in recent years, it has become clear that science is plagued by findings that cannot be reproduced and, consequently, invalidating research studies and undermining public trust in the research enterprise. The observed lack of reproducibility may be a result, among other things, of the lack of transparency or completeness in reporting. In particular, omissions in reporting the technical nature of the experimental method make it difficult to verify the findings of experimental research in biomedicine. In this context, the assessment of scientific reports could help to overcome – at least in part – the ongoing reproducibility crisis.

In addressing this issue, this Thesis undertakes the challenge of developing strategies for the evaluation of reporting biomedical experimental methods in scientific manuscripts. Considering the complexity of experimental design – often involving different technologies and models, we characterise the problem in methods reporting through domain-specific checklists. Then, by using checklists as a decision making tool, supported by miniRECH – a spreadsheet-based approach that can be used by authors, editors and peer-reviewers – a reasonable level of consensus on reporting assessments was achieved regardless of the domain-specific expertise of referees. In addition, by using a text-mining system as a screening tool, a framework to guide an automated assessment of the reporting of bio-experiments was created. The usefulness of these strategies was demonstrated in some domain-specific scientific areas as well as in mouse models across biomedical research.

In conclusion, we suggested that the strategies developed in this work could be implemented through the publication process as barriers to prevent incomplete reporting from entering the scientific literature, as well as promoters of completeness in reporting to improve the general value of the scientific evidence.

Declaration

The published papers constituting Chapters 3 and 5 were previously submitted by Michael Bramhall in support of an application for a doctoral degree at the University of Manchester in 2015. No other portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgments

I wish to thank my supervisors, Professors Andy Brass and Robert Stevens for their excellent guidance, patience and continuing support provided throughout this PhD project. I am extremely grateful for making my PhD such a positive experience – the best since I am enrolled in academia; so thank you. It is a privilege to have you both as my mentors. I would also like to thank Doctors Sheena Cruickshank, Harry Noyes, Suzanne Embury, and Goran Nenadic, who have been ready with help and advice throughout the various stages of preparing the manuscripts for publication. I know I'm a bit overwhelming when it comes to writing, so many thanks for tolerating the bombardment of emails to which I submitted all of you with several drafts of the manuscripts. Many improvements were made to the manuscripts from their feedback. Thank you also to my many (anonymous) paper reviewers, and my thesis examiners, Prof John Keane and Dr Helen Parkinson. Their comments have helped me to significantly refine my work.

I also owe a particular debt of gratitude to a few collaborators who were kind enough to lend important support during this project: Michael Bramhall, not only for sharing the pain of being a biologist adapting to the world of computer science, but also for his intellectual and proofreading support on the manuscripts – I couldn't ask more!; Binling Jin, with whom I made the miniRECH a reality; George Karystianis, who helped me a lot with the text mining strategy and who supported me during the more stressful times; and Diego Pérez, for teaching me the basics of the R programming language and helping me with several statistical queries.

I would also like to extend my warmest thanks to all my friends and colleagues from the School of Computer Science for their support and motivation and for being there to share both technical and social matters. Thank you so much for your sincere friendship.

I hope you had fun with the Colombian Latin flavour that runs in my veins. Francis, I appreciate your tips on English for Spanish speakers. I extend my heartfelt thanks to all my friends back home for their encouragement in completing my PhD studies. Thank you for sharing this friendship over the years with your visits, calls or messages. *Sin duda me hicieron sentir en casa.*

My deepest thanks are of course, and as always, for my parents, Olga and Julio, who have lent me their full support throughout this time. They always have believed in me, and never have stopped supporting my interest in science – even if it means to be away from them. *¡Gracias por ser mi mayor motivación!*

Finally, I would like to thank Colciencias [*Instituto Colombiano para la Ciencia y la Tecnologia, Francisco Jose de Caldas*] for contributing to the fulfilment of my studies through a scholarship supported by the Colombian government that provided me with full financial assistance during three years.

Dedication

To my parents [Olga and Julio] and brothers [Julio, Carlos and Nelson], and also to the memory of my best friend, William, who passed away during my PhD studies.

The Author

Oscar Flórez-Vargas is a Colombian citizen who came to the United Kingdom to study a *Ph.D.* in Computer Science at the University of Manchester in 2012, the outcome of which is this final Thesis. Currently, Oscar holds an *M.Sc.* in Biochemistry from the National University of Colombia – graduated with honours in 2008, and received a *B.Sc.* in Microbiology from the Industrial University of Santander in 2004.

He has worked as a lecturer and as a researcher in biomedical sciences in Colombia. Most of his academic career has been focused on studying the relationships between genes and different human diseases. He has published 15 peer-reviewed articles, since 2004, when he first began publishing as an academic. In addition, he has been an ad hoc reviewer for 4 peer-reviewed journals and he has acted as an expert evaluator of grant applications for Colciencias – the Colombian Council for Science.

Rationale of Format

This research project has produced important strategies into the methods for assessing the quality of reporting biomedical experiments and its role in reproducibility. The successful development of these strategies has had the collaboration of other research groups of the University of Manchester such as Manchester Immunology Group; Manchester Institute of Biotechnology; Information Management Group, as well as the School of Biological Science of the University of Liverpool. In this context, the alternative format allows the inclusion of collaborative research, and reduces the potential intellectual property conflict. Since the methodological applications developed in this project could contribute to mitigate the issue of reporting which is widely related to the ongoing irreproducibility crisis, publishing journal articles was desirable to report our findings to the scientific community. The alternative format also affords better preparation for writing and formatting based on external editing guidelines. Therefore, and considering all above arguments together, the alternative format thesis becomes appropriate for this research project.

Chapter One

Introduction

"Non-reproducible single occurrences are of no significance to science."

— *Sir Karl Popper*
Philosopher of Science

1.1 Motivation

The idea that the same experimental findings can be reproduced by a variety of independent approaches is one of the cornerstones of science's claim to objective truth. However, there is a growing concern both inside and outside the scientific community over the lack of reproducibility of many published scientific findings (Anon 2013e, d, c). An examination of this problem suggests that it can be attributed largely to the lack of transparency in reporting (Moher et al. 2008, Landis et al. 2012), although, misconduct and honest mistakes will naturally have an influence. In this regards, stakeholders in the life science research are paying greater attention to improve the quality of reporting in scientific studies (GBSI 2013).

An increasing number of reports have found discrepancies in basic and preclinical published studies across biomedical disciplines, which highlight a significant problem in the development of new therapies to treat diseases. For example, in the framework of scientific relations between academia and industry, the identification of potential drug candidates typically happens in academic research laboratories, whereas the drug development efforts of these new drug candidates are carried out by pharmaceutical companies (Begley and Ellis 2012, Freedman and Inglese 2014). However, and although this is not always a straightforward matter, the increasing reports of discrepancies in preclinical publications has led to a re-evaluation of the reliance on academic research by the pharmaceutical industry. Clinical trials in oncology, for instance, have the highest failure rate compared with other therapeutic areas: only 5% of agents that have shown promising anticancer activity in preclinical studies are licensed after demonstrating sufficient efficacy in posterior clinical studies, which contrasts with 20% for cardiovascular diseases (Hutchinson and Kirk 2011). These rate differences could be explained by the complexity of cancer as a disease process and the problem of anti-tumour drug resistance (Volm and Efferth 2015, Alaoui-Jamali, Dupre, and Qiang 2004).

In the above context, a few years ago, two pharmaceutical companies – based on their in-house target validation programs – revealed that major findings in a set of published oncology papers could be reproduced for less than a quarter (Prinz, Schlange, and Asadullah 2011, Begley and Ellis 2012). In one of the reports, C. Glenn Begley and Lee M. Ellis at Amgen – an American company headquartered in California, examined 53 studies that they considered landmark in the basic science of cancer. Nevertheless, the scientific findings were confirmed in only six (11%) of cases (Begley and Ellis 2012). In the other report, Florian Prinz and his colleagues at Bayer HealthCare – a German company headquartered in Berlin, reported that they were successfully able to reproduce the scientific findings in ~20–25% of 67 examined projects (Prinz, Schlange, and Asadullah 2011). Discrepancies in over-expression of certain genes in specific tumour types as well as decreased cell proliferation via ribonucleic acid interference, or RNAi were among the findings that could not be reproduced. In addition, when findings could not be reproduced, in both experiences an attempt was made to replicate exactly the results. This was carried out by co-operating closely with the original researchers in order to ensure that their experimental techniques matched those used the first time round – occasionally even in the laboratory of the original investigator. However, neither of these results could be replicated.

So why do so many biomedical publications contain research that cannot be reproduced? Although there is no clear consensus as to what constitutes a reproducible study, various factors contribute to the irreproducibility problem, involving all actors of the publishing process. While the authors are responsible for the information declared in their manuscripts, journal editors and peer-reviewers have a key role in determining the quality of a manuscript by preventing mistaken and non-reproducible evidence. In this context, and considering that scholarly articles are still the most effective means for archiving and disseminating scientific knowledge, the quality of the content found in any scientific publication should be guaranteed. By providing good reporting in publications, the scientific biomedical community can benefit greatly from translational discoveries: a task facilitated by the technology developed by the computer science community, *e.g.*, the Semantic Web through Ontologies.

In the domain of biomedical sciences, the vast amount of useable knowledge held in scientific publications and curated databases has transformed biological and medical sciences to data-driven sciences. As a consequence, scientists in fields such as genomics, proteomics, metabolomics, and clinical medicine face the challenge of coming up with novel ways of data utilisation to acquire new biomedical knowledge (Shadbolt, Hall, and Berners-Lee 2006, Cheung et al. 2009). Drawing meaningful inferences needs the proper integration of information. Nevertheless, the complexity of the biological phenomena, along with the heterogeneity arising from their experimental demonstrations, requires that the data integration task to be done in the framework of the metadata, *i.e.*, experimental design and protocols. The usefulness of linking data and metadata in an experimental context to discovery has been proved by Adam and Eve, two robot scientists able to perform independent bio-experiments to test hypotheses and interpret findings without human guidance (King et al. 2004). Most of these experiments are performed in the yeast *Saccharomyces cerevisiae*, an organism that scientists use to model more complex living systems, and the link between the experimental data and metadata produced by these robots is conducted by using the ontology EXACT2 (Soldatova et al. 2014). Thus, Adam and Eve record in great detail their bio-experiments so those experiments can faithfully be reproduced.

While Adam and Eve are living in a scientific Garden of Eden – where all experimental variables can be perfectly controlled, monitored, and recorded to allow reproducibility, we are still living in a scientific world that can be messy – where it is hard to know what to make of irreproducible experimental findings. To overcome the ongoing reproducibility crisis, the scientific community needs to take the step of caring about transparency in reporting data and metadata (Landis et al. 2012, Goodman et al. 2014). In order to help scientists in taking this step, this research will focus on the development of strategies that can facilitate the assessment of reporting experimental methods through the publication process. Providing a high-quality description of the method(s) is the necessary condition to know when a finding is due to the intrinsic properties of the biological systems or whether it is related to differences in experimental designs and protocols. Thus, the step of transparency in reporting toward enhancing reproducibility reminds the famous words of Neil Armstrong – *"That's one small step for man; one giant leap for mankind."*

1.2 Hypothesis, Aim and Objectives

Basic and preclinical research is fundamental because it provides the base on which future studies are built. These two research fields constitute the so-called biomedical framework, which is a fast-growing interdisciplinary area of science that involves the investigation of biological processes and their translation for human benefit. The rapid development of technology for biomedical research has generated, among other things, an enormous amount of experimental information from *"-omics"* technologies, *e.g.,* genomics, proteomics, and metabolomics. Currently, the volumes of information produced in high-throughput experiments from *In Vitro* and *In Vivo* models can be, to some extent, manageable. One of the biggest challenges in computational biomedical research is trying to integrate and analyse the vast amount of complex experimental information produced by these technologies in an interpretable way to exploit its full potential, thus drawing meaningful conclusions.

In recent years, however, it has become clear that biomedical science is plagued by findings that cannot be reproduced (Begley and Ellis 2012). There is, therefore, an urgent need to address this problem because it directly affects the outcomes generated by the integration and analysis processes. Indeed, the fact that the findings of biomedical experimentations carried out in laboratory models cannot always be reproducible with certainty, has led to doubt as to whether experiments should be considered as a source of knowledge for clinical evaluation (Prinz, Schlange, and Asadullah 2011). According to the evidence that will be presented in this Thesis, one of the roots of the reproducibility problem is in the current structure of scientific publishing, which does not provide a high-quality description framework of the methods, *i.e.* the meta-data. The lack of such a framework makes it difficult to effectively transfer and use knowledge of the complex experimental methods needed to adequately judge data for use and re-use, especially when data from more than one source are used together.

The general scientific community agrees that a proper reporting of how the experiments are done and what information is obtained from them will help greatly to overtake the ongoing reproducibility crisis (Landis et al. 2012, McNutt 2014). Nevertheless, and despite attempts to improve the reporting, the problem still persists (Kilkenny et al. 2009, Witwer 2013). Therefore, there is a need to promote a change of scientific culture towards the care of scientific data and metadata (Goodman et al. 2014). We hypothesised that this change could be achieved by generating strategies to facilitate the assessment of reporting through the publication process. Thus scientists (*i.e.*, authors, editors and peer-reviewers) will appreciate its usefulness [hopefully] and as a consequence its direct application to ensure the quality of experimental information to be published, which will represent a step forward in enhancing reproducibility.

1.2.1 Aim

The overall aim of this Thesis is to create strategies for the assessment of reporting of biomedical experimental methods. This will not only be beneficial for improving reproducibility, but will also be useful for integrating data and metadata and, so, gaining a better understanding of the natural phenomena in the context of the experimental evidence. In order to address this aim the following objectives were undertaken.

1.2.2 Objectives

i. Investigate the status of methods reporting in biomedical experimental models (*in vitro* and *in vivo*) that enable reproducibility and, therefore, comparability of results in order to facilitate translational discoveries by aggregating data in a metadata context from multiple experiments. This objective is addressed in Chapters Three and Five by the creation of checklists outlining the essential components that should be reported in experimental models of complex diseases for two domain-specific biomedical communities and its subsequent use in assessing the reporting via systematic review.

ii. Develop a consensus framework to guide scientists in verifying their judgments regarding scholarly reports. This objective was addressed in Chapter Four by the creation of a general spreadsheet-based tool named miniRECH for assessing quality of scientific reporting (both pre- and post-publication) by using checklists as templates. The tool-model developed in this work was designed to operate in Microsoft® Excel since MS Excel is widely used in the biomedical community.

iii. Develop an approach that allows an automatic assessment of reporting key information involved in biomedical research. This objective was addressed in Chapter Six by the creation of a text-mining strategy that aimed to assess the reporting of sex and age in mouse experiments across biomedical literature. Both sex and age are two inextricably linked factors that play key roles in the interpretation of experimental findings and thus in their reproducibility.

1.3 Thesis structure

This Thesis is submitted with permission from the School of Computer Science at the Faculty of Engineering and Physical Sciences in the alternative format. As a result, the major chapters within are organised into the structure of a research paper. Specifically, the rest of this Thesis is structured as follows:

- Chapter 2 provides the required background for this Thesis.
- Chapter 3 conducts an initial investigation to determine the status of methods reporting in published parasitic diseases experiments that should enable a valid comparison of research findings.
- Chapter 4 details the miniRECH spreadsheet-based tool, a framework to guide scientists via checklist in verifying their judgments regarding reporting biomedical experiments.
- Chapter 5 uses miniRECH to assess the reporting of experimental method in animal models of inflammatory diseases, with impact on the translation of basic research into clinical research.
- Chapter 6 develops a text-mining system as a survey technique for automatically assessing the reporting of key information that will allow scientists to improve the reproducibility of the findings presented in a scientific paper.
- Chapter 7 summarises and discusses the contributions of this work and provides some final concluding remarks.

1.4 Full publication list

The work done as part of this Thesis has been published as follows:

- Flórez-Vargas O, Bramhall M, Noyes H, Cruickshank S, Stevens R, Brass A. The quality of methods reporting in parasitology experiments. *PLoS One*. 2014; 9(7):e101131. doi: 10.1371/journal.pone.0101131.

  Author contributions: OF-V, RS and AB conceived and designed the experiments. OF-V, MB, HN and SC performed the experiments. OF-V, MB, HN and SC analysed the data. OF-V wrote the manuscript. MB, HN, SC, RS and AB edited the manuscript.

- Flórez-Vargas O, Bramhall M, Jin B, Pérez D, Cruickshank S, Embury S, Stevens R, Brass A. miniRECH: a spreadsheet-based tool for assessing the quality of methods reporting in scientific manuscripts. *Journal of Biomedical Informatics* (submitted).

  Author contributions: OF-V, SE, RS and AB conceived and designed the study. OF-V and BJ developed the tool. OF-V and MB performed the experiments. OF-V and DP performed the statistical analysis. OF-V wrote the manuscript. MB, SE, SC, RS and AB edited the manuscript.

- Bramhall M, Flórez-Vargas O, Stevens R, Brass A, Cruickshank S. Quality of methods reporting in animal models of colitis. *Inflammatory Bowel Diseases*. 2015; 21(6):1248-59. doi: 10.1097/MIB.0000000000000369.

  Author contributions: MB, RS, AB and SC conceived and designed the study. MB and OF-V carried out the experiments and analysed the data. MB wrote the manuscript. OF-V, RS, AB and SC edited the manuscript.

- Flórez-Vargas O, Brass A, Karystianis G, Bramhall M, Stevens R, Cruickshank S, Nenadic G. Bias in the reporting of sex and age in biomedical research on mouse models. *eLife*. 2016; 5: e13615. doi: 10.7554/eLife.13615.

> *Author contributions:* OF-V, AB, RS, SC and GN conceived and designed the study. OF-V, GK and MB implemented the system. OF-V and GK performed the experiments, analysed the data, and drafted the manuscript. AB, RS and GN supervised the study.

In addition, parts of this work were published as blog posts:

- Van Noorden R. (2016). Scientists still fail to record age and sex of lab mice. [Article] *Nature News*. Available at: http://www.nature.com/news/scientists-still-fail-to-record-age-and-sex-of-lab-mice-1.19500 [Accessed 07 Mar. 2016].
- Dunford Z. (2016). Groundbreaking text mining project highlights 'gender gap' in scientific research. [Press release] *The University of Manchester News*. Available at: http://www.manchester.ac.uk/discover/news/text-mining-project-highlights-gender-gap/ [Accessed 07 Mar. 2016].
- Flórez-Vargas O and Bramhall M. (2015). Misleading reporting is damaging scientific research. [Blog] *Manchester Policy Blogs: Science and Technology*. Available at: http://blog.policy.manchester.ac.uk/sci-tech/2015/06/misleading-reporting-is-damaging-scientific-research/ [Accessed 15 Dec. 2015].
- Moore C. (2015). Quality of Methods Reporting in Animal Models of Colitis Generally Deficient [Blog] *IBD News Today*. Available at: http://ibdnewstoday.com/2015/05/01/quality-methods-reporting-animal-models-colitis-generally-deficient/ [Accessed 06 Jul. 2016].
- Stevens R. (2014). Being a credible virtual witness. [Blog] *Robert Stevens' Blog*. Available at: https://robertdavidstevens.wordpress.com/2014/09/19/being-a-credible-virtual-witness/ [Accessed 15 Dec. 2015].

And as a poster and a presentation

- Presentation: Flórez-Vargas O. Reproducibility in biomedical research. *Research Seminars*. 2014. At The University of Manchester, School of Computer Science.

- Poster: Flórez-Vargas O., Brass A., Stevens R. Are methods in tropical infectious diseases models properly described? *Research Seminars*. 2013. At The University of Manchester, School of Computer Science.

- Poster: Bramhall M., Flórez-Vargas O., Stevens R., Wilson J., Cruickshank S., Brass A. Quality of methods reporting in colitis experiments and the subsequent impact on the development of a gut knowledge base. *Gut*. Jun 06, 2014; 63(Suppl 1):e3 doi:10.1136/gutjnl-2014-307263.LB6. At British Society of Gastroenterology Annual General Meeting.

Chapter Two

Background

> *"Experimental observations*
> *are only experience carefully*
> *planned in advance, and*
> *designed to form a secure basis*
> *of new knowledge."*
>
> — *Sir Ronald Fisher*
> *Bio-statistician*

2.1 The scientific experiment

In the context of *[the]* scientific method, an experiment is an organised series of steps to validate or reject a hypothesis – which is an explanation about a phenomenon in the natural world. Experiments provide insight into cause-and-effect relationship by determining whether the outcomes are actually caused by the manipulation of a particular variable, or some other factors may be attributed to the effects seen in the experiment. While a cause-and-effect relationship may be plausible to establish in some scientific areas such as physics and chemistry – because a good experimental design can neutralise potentially confounding variables, the complexity of the natural phenomena under investigation makes it difficult to establish a natural order in the life science arena – often interacting across different environments and time scales. Therefore, describing in detail the steps taken in any experiment becomes an important task to identify the factors involved in the experimental outcome via reproducibility, facilitating both interpretation and validation of the results (Landis et al. 2012).

The development of science via empiricism has a great history of at least 2500 years back to the ancient civilizations, which deserves some attention in order to understand how experiments have built our scientific knowledge. In Greece, Aristotle (384–322 B.C.E.) – a philosopher regarded as the father of science, formulated the basic principles of scientific epistemology. In the *Organon* [Greek: Ὄργανον], particularly in *Prior Analytics* and *Posterior Analytics*, he established an objective method for acquiring knowledge through reason by building upon what is already known. This 'proto-scientific method' involved making meticulous measurements and observations about almost everything, which represented the foundation of empiricism. During the Golden age of Islam: between the 10th and 14th centuries, Muslim scientists used experimental approaches to distinguish between competing scientific theories. The Arab physicist Ibn al-Haytham, in his *Book of Optics* [Arabic: كتاب المناظر], for instance, presented experimentally funded arguments on vision to prove the intromission theory to be correct and the extramission theory to be incorrect, concluding that rays of light are emitted from objects rather than from the eyes.

In the European Renaissance, the knowledge and methodological insights of the Greeks and the Muslims scientists gained considerable traction among European scholars. During the Scientific Revolution – the period covering the 16th and 17th centuries, the names of Francis Bacon and Isaac Newton are frequently mentioned regarding experimental science. Francis Bacon (1561–1626) – an English philosopher, emphasised the importance of the study of nature through experimental methods. His book *Novum Organum Scientiarum* (1620) promotes that all scientific knowledge comes from a process where an initial observation leads to the discovery of a certain pattern that can be translated to a general theory. Sir Isaac Newton (1642–1726) – an English mathematician, claimed that by combining deductive logic from a given statement with inductive reasoning from empirical observation, it is possible to derive a tentative premise known as a 'hypothesis' which needed to be tested by experiments in order to be validated. His book *Philosophiæ Naturalis Principia Mathematica* (1687) describes the mixture of both methods. The acquisition of knowledge via experimental investigation led to the foundation of The Royal Society in 1660 – a learned society for science and possibly the oldest such society still in existence. This body ruled that experimental evidence always supersedes theoretical evidence – one of the foundations of modern science.

In the 20th century, the role of experimental investigation as an engine for generating knowledge was strengthened by philosophers of science. Probably the most famous of these was Sir Karl Raimund Popper (1902–1994). Popper – an Austrian-British philosopher, proposed the principle of falsifiability to scientific discovery. In his book *The Logic of Scientific Discovery* (1934), Popper argues that no number of experiments can ever prove a theory, but a single experiment can contradict one. This is often shown by the example of swans – the assumption *"all swans are white"* seemed to be true until black swans were discovered and so this assumption was refuted. By adopting a methodology based on falsifiability, Popper suggests that if a theory is falsifiable, it is scientific, and if not, then it is unscientific. In this context, Einstein's theories are regarded as scientific since they are empirically testable, whereas those that have no potential for falsification were regarded as pseudoscientific, *e.g.,* astrology.

Accordingly, scientific theories became the foundation of scientific knowledge in modern science. These theories, in turn, are based on assumptions which are taken as postulates for such theories. A theory for antibiotic resistance, for instance, might depend on the theory of evolution by natural selection as an assumption for adaptation and speciation (Laehnemann et al. 2014); the compelling evidence on the evolution theory was published in the book *On the Origin of Species* in 1859 by Charles Darwin (1809–1882) – an English naturalist. In this way, falsifying a theory requires that its assumptions be demonstrably true. However, as in the example, assumptions are often theoretical statements, which – according to Popper – are almost impossible to verify. Therefore, if one cannot verify an assumption, one will not be able to falsify a theory. Regarding this matter, Carl Gustav Hempel (1905–1997) – a German philosopher, declared one of the most useful properties of scientific theories: *"the statements [assumptions] constituting a scientific explanation must be capable of empirical test."* In this context, the empirical tests used for testing the falsifiability of a theoretical scientific statement requires a certain level of universality. This means that the same theory may be tested by multiple scientists through different approaches and achieve similar answers. Thus, reproducibility is the requirement of universality.

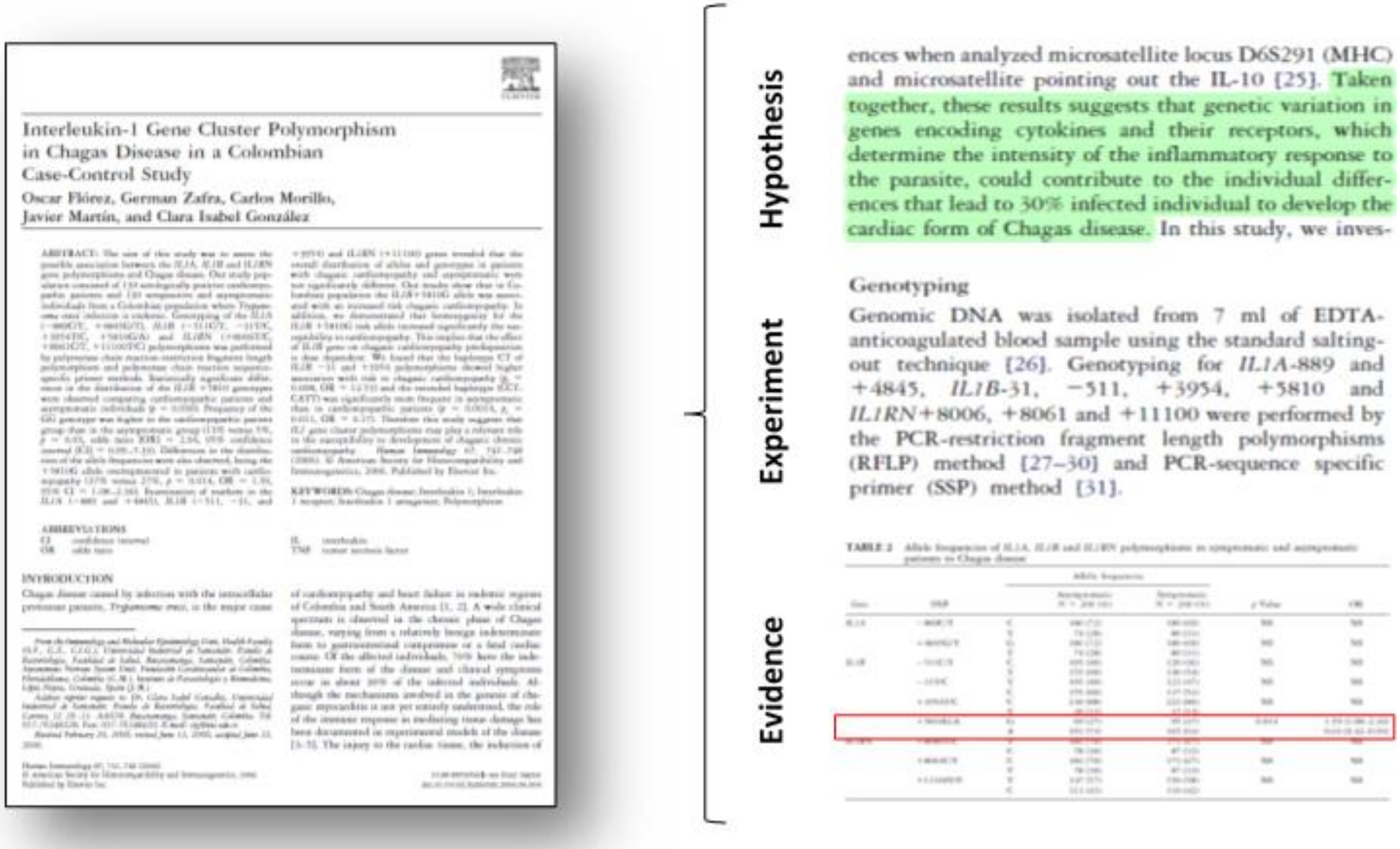
Now, in the 21st century, experimental science and therefore the knowledge built upon it are facing a reproducibility crisis. Reproducibility in science is as important as any new hypothesis or discovery; without reproducibility, there is no science. In this regards, Karl Popper stated *"Non-reproducible single occurrences are of no significance to science."* What does it mean to make an experiment be reproducible? In biomedicine – probably the scientific field with the major concern – many factors can account for irreproducible findings: from the biological complexity of the experimental model itself to the technical aspects of conducting an experiment, including its analysis. Since nobody can tell which protocol variables are going to matter in an experiment, the scientific community encourages scientists to record all that are possible in order to enhance the likelihood of finding them, *e.g.*, by the integration of experimental information. The issue is whether the level of annotation of the experimental context consigned in the biomedical literature is currently good enough to overtake the ongoing reproducibility crisis (Landis et al. 2012, Goodman et al. 2014).

2.2 The scientific paper

Scholarly publishing is a process by which new knowledge is created and disseminated. While it is difficult, if not impossible, to know the precise date of the first scientific like report – but it would surely date from early civilisations, it was 350 years ago when the pioneering journal dedicated to scientific endeavour was created: the *Journal des Sçavans*, published in Paris on January 5th, 1665. This journal, established by the editor Jean-Denis de Sallo, contained information not only on scientific matters, but also on matters of art, as well as legal and ecclesiastical judgments (Fjällbrandt 1997). Shortly later, member of the Royal Society – based on a review of the Sallo's early issues – decided to create a more philosophical type of serial publication for reporting the scientific material presented at meetings of the Royal Society. Then, the first volume of the journal *Philosophical Transactions* was published in London on March 6th, 1665 (Fjällbrandt 1997); providing the model for almost 30,000 scientific journals today. Since then, and mainly after the 19th century, the exchange of scientific knowledge has been carried out via scientific articles.

A scientific article is a collection of arguments, backed up by experiments and evidence, used to support a hypothesis (Figure 1). By acting as a *"virtual witness"* of scientists' activities, *i.e.*, through detailed description of the experimental design, methods and analysis, a scientific paper plays an important role in judging the scientific merits of the experimental work (Clark 2014, Fjällbrandt 1997). In this way, virtual witnessing required that experimentalists, after physically congregating to perform experiments, communicate the findings and interpretations to others, as well as the experimental technologies used, including both instruments and protocols (Clark 2014). Therefore, the accuracy of reporting about what was done in a scientific investigation is fundamental for validating its findings, and then repeating and reproducing the scientific work. This is particularly important in the life science arena due to the complexity of biological systems, *i.e.*, they are dynamic across temporal and spatial dimensions. Thus, both data and metadata need to be understood in the context of the experimental models and the proceedings themselves in order to verify the findings and so determine whether there is any conclusive evidence. In this regards, some journals, including *Nature* (Anon 2013c), have abolished space restrictions on the Methods section towards enhancing reporting of life-sciences research.

Through the research process – from the idea to the publication, reporting plays an important role, *e.g.*, writing grant proposals; design of experimental protocols; data and metadata recording and analysis; and manuscript preparation – just to mention some of them. In addition, several reporting quality checkpoints have been identified in the research process, which are clearly illustrated in Figure 2. Nevertheless, only a few of them have been reported to be broadly used; such is the case of the peer review process for both grant and manuscript submission (GBSI 2013).



Interleukin-1 Gene Cluster Polymorphism in Chagas Disease in a Colombian Case-Control Study

Oscar Flórez, German Zafra, Carlos Morillo, Javier Martín, and Clara Isabel González

ences when analyzed microsatellite locus D6S291 (MHC) and microsatellite pointing out the IL-10 [25]. Taken together, these results suggests that genetic variation in genes encoding cytokines and their receptors, which determine the intensity of the inflammatory response to the parasite, could contribute to the individual differences that lead to 30% infected individual to develop the cardiac form of Chagas disease. In this study, we inves-

Genotyping

Genomic DNA was isolated from 7 ml of EDTA-anticoagulated blood sample using the standard salting-out technique [26]. Genotyping for *IL1A*-889 and +4845, *IL1B*-31, −511, +3954, +5810 and *IL1RN*+8006, +8061 and +11100 were performed by the PCR-restriction fragment length polymorphisms (RFLP) method [27–30] and PCR-sequence specific primer (SSP) method [31].

Figure 1. Example of a scientific article. From Flórez O., et al. *Human Immunology* (2006); 67: 741-748 (Florez et al. 2006).

Moving beyond the issues related to the publication process itself (it is not a matter of this Thesis), which have been highlighted by Peter A. Lawrence in a *Nature Commentary* entitled *The Politics of Publications* (Lawrence 2003) – where publishing indicators became more important than the scientific evidence itself, the decision about publication of a paper is the outcome of an often lengthy feedback between authors, editors and reviewers. Therefore, the guarantee of the quality of reporting of any research study relies on these three stakeholders. In this regard, there is an implicit assumption of many readers of scientific journals – mainly those with high impact factor – that assume that if a study is of sufficient quality to pass the scrutiny of rigorous reviewers and editors, it must be true and therefore its findings reproducible (Loscalzo 2012). However, the evidence has shown otherwise. In Stem Cell research – a rapidly developing field with implications that can revolutionise medicine, two misconduct cases have received copious amounts of attention (Kennedy 2006, Obokata et al. 2014). In both cases the misconduct came to light when other scientists were trying to reproduce the promising findings reported in the most prestigious journals: *Nature* and *Science*, which has led to questioning of the editorial and peer-review processes.

Considering that the publication of a scientific article is not just the culmination of a research process but it is also the start-point of new research – as exemplified in Figure 2 for the academic life science, the reproducibility of its findings become the major yardstick by which knowledge is built. How easy is it to reproduce a published scientific finding? In an attempt to quantify the reproducibility by estimating the time required to reproduce a computational biology paper, the research team of Philip E. Bourne, at the University of California San Diego, estimated the overall time to reproduce the method as 280 hours for a novice with minimal expertise in bioinformatics (Garijo et al. 2013). This is an important amount of time considering not only that it was an *in silico* work (in theory, more easily reproducible than *in vitro* or *in vivo* works), but also that they were attempting to reproduce their own results. Regarding the desiderata for reproducibility of published findings in scientific papers, the main observation from P. E. Bourne's work relied in the quality of reporting.

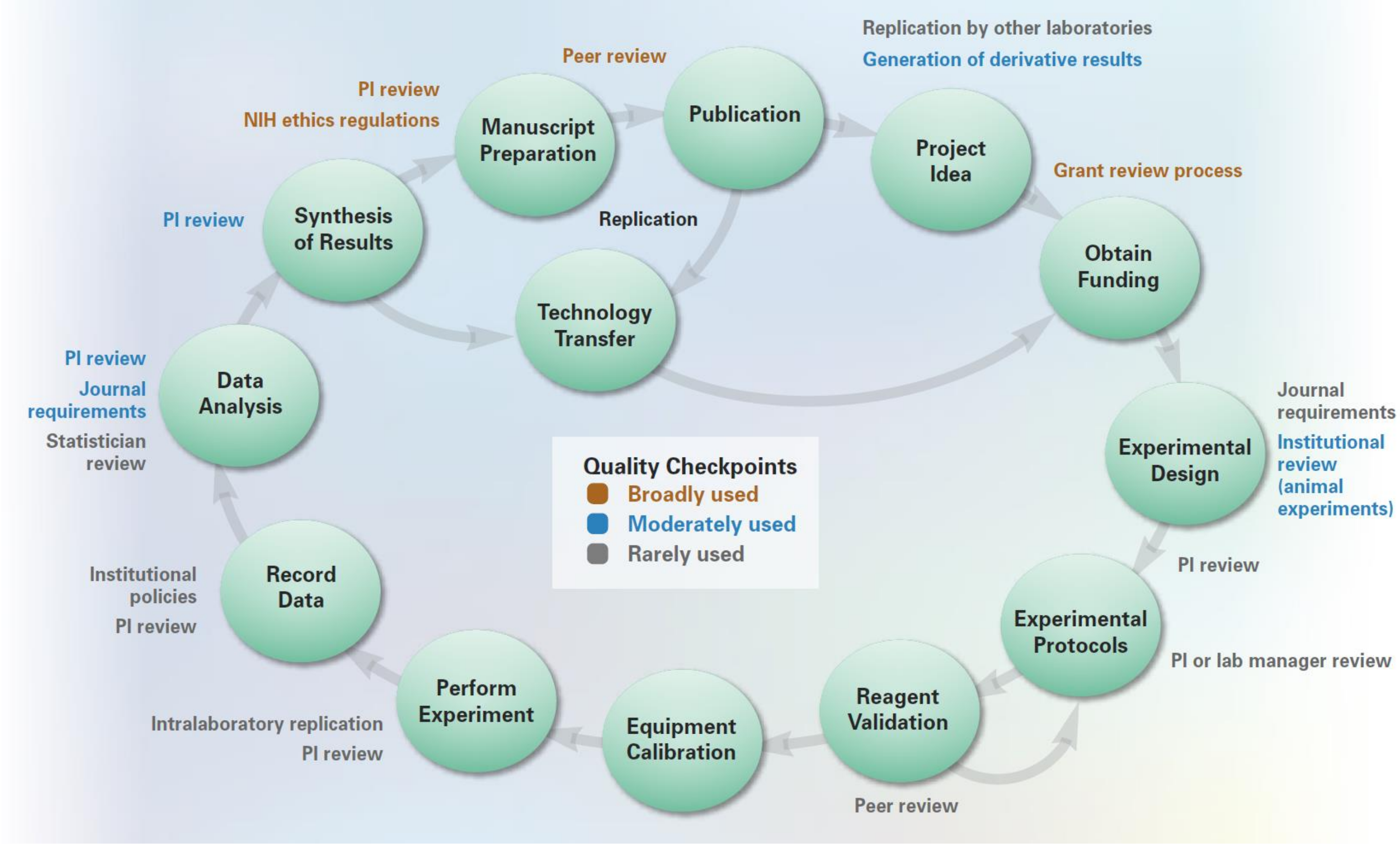


Figure 2. Academic life science research process. Green circles indicate common steps in the life science research process. Adjacent color-coded text describes current/traditional quality checkpoints. Image from The Case for Standards (GBSI 2013).

2.3 On experimental irreproducibility

In more traditional sciences such as biology, chemistry or physics there is an intuition to believe that the reproducibility of an experiment means that it can be replicated. However, there is an important distinction between these two terms that needs to be considered: on the one hand, *replicability* describes the ability to obtain an identical result when an experiment is carried out under exactly identical conditions, whereas, on the other hand, *reproducibility* refers to a phenomenon which can recur even when experimental conditions may vary to some degree (Drummond 2009). In this way, replicability reflects the technical precision of a specific experiment, while reproducibility reflects the fundamental accuracy of an experimental observation. While the replicability and reproducibility of experimental findings represent a successful assessment of the scientific evidence, scientists are mainly interested in the reproduction of findings rather than the replication of experimental results, which implies robustness of the original findings.

The observed lack of reproducibility may be a result, among other things, of the lack of transparency in reporting: including over-interpretation, misinterpretation, or even falsification of data and metadata (van der Worp and Macleod 2011, Moher et al. 2008, Landis et al. 2012). In the following paragraphs, and regarding the lack of transparency in reporting biomedical research, I will be presenting some of the most important causes and consequences to the irreproducibility problem outlined by the scientific community. While this Thesis focuses on the reporting as one of the important sources of experimental irreproducibility – taking into account the good faith of scientists, it is worth mentioning other sources of irreproducibility where the scientist does not act in good faith. After all, the number of papers retracted for honest errors and irreproducible findings is about 40% of retractions, whereas those attributed to misconduct overtake this number by four points (Van Noorden 2011).

2.3.1 Methodology – “tricky” experimental details not stated

Experimental procedures are expected to be described in enough detail to allow repetition of the experiment. In the biomedical field, for instance, the Uniform Guidelines of the International Committee of Medical Journal Editors state that "*the authors should include technical information in sufficient detail to allow the experiment to be repeated by other workers*" (International Committee of Medical Journal Editors 2013) – even though many journals are unlikely to accept manuscripts that precisely present already published findings (Casadevall and Fang 2010). Lack of novelty, indeed, is one of the most prominent reasons for manuscript rejection (Ali 2010).

By far, the main cause of irreproducible outcomes lies in the failure to report technical aspects of the conduct of an experiment. Providing a high-quality description of the experimental method is important not only to replicate and reproduce, but also to compare and integrate data and metadata and, hence, facilitating translational discoveries. The reporting issue probably stems, at least in part, from the current structure of scientific publishing, which does not provide a high-quality description of the method and it is not sufficient to effectively transfer knowledge of complex experimental methods. Nonetheless, sometimes the experimental details omitted by the authors are due to the lack of knowledge of their importance for comprehending the findings, from wanting to hide key information from their competition, or simply because they forgot to do so. Whatever the reason, the Methods section of a manuscript is frequently incomplete, superficial, and/or vague, making it difficult to know how a study was actually performed (Bolli 2015).

While it is true that the Author Guidelines of several journals commonly state that *"[...] detailed descriptions of methods already published should be avoided; a reference number can be provided to save space, with any new addition or variation stated"*, some authors rely on the practice of citing a previous paper, which in turn cites a previous paper, and so on. At the end of this domino effect there is the possibility of discovering that the original paper does not actually describe the method that it is supposed to describe (Bolli 2015).

There are simple and complex experiments and everything in between, but the important part is that they all should be clearly described in detail to allow them to be repeated successfully. Sometimes the success of an experiment is predicated on the technical minutia, *i.e.*, the “tricks”; which are described rarely in published articles. Examples of this technique minutia include, among others, reagents (*e.g.,* different brands), tissue culture plastic (*e.g.,* surface treated or non-treated to facilitate cell attachment and growth), and housing conditions (*e.g.*, light cycle schedules and bedding of mice). Such subtle technical differences could explain, at least in part, the difficulties in reproducing experimental findings.

In 2010, for instance, Deepak Srivastava and colleagues at the Gladstone Institute of Cardiovascular Disease in California, reported in *Cell* that a combination of three developmental transcription factors (*i.e.*, Gata4, Mef2c, and Tbx5) can efficiently reprogram dermal fibroblasts directly into differentiated cardiomyocyte-like cells (Ieda et al. 2010). However, in an attempt to reproduce this finding, three papers published in 2012 – from three independent research groups – reported complete (Song et al. 2012) and partial (Protze et al. 2012) success, as well as unsuccessful reproduction (Chen et al. 2012). These differences in reproducibility could be explained by technical differences that can directly or indirectly influence the outcome. A recent example shows that, indeed, differences of a technical nature would be responsible for the discordance in drug response measurements from two large-scale pharmacogenomic studies, even though genomic data were well correlated (Haibe-Kains et al. 2013).

2.3.2 Study design and statistical analysis – chasing the p-value

Probably the most common reason for lack of reproducibility is stemmed by inappropriate study design and statistical analysis, which results in false research findings. Regarding this matter, John Ioannidis, an epidemiologist at Stanford University, was the first to point out this issue in his 2005 paper that has a deliberately provocative title: *Why most published research findings are false* (Ioannidis 2005). In a framework of research discoveries based on statistical significance, typically for p-value less than 0.05, J. Ioannidis provides evidence of simulations using the formulas developed for the influence of power, ratio from true to non-true relationships, and bias. The results of these simulations showed that for most situations (>50%) that may be characteristic of specific study designs and settings (*e.g.* small group size), it is more likely for a research claim to be false than true (Ioannidis 2005). Nevertheless, a different rate of false-positive results (14%, SD 1%) was estimated by using a false discovery rate (FDR) technique adapted from genomic studies (Jager and Leek 2014). This result was calculated by mining 5,322 significant p-values from the abstracts of 77,430 papers in 5 major journals across a decade; suggesting that, to the contrary, the medical literature remains a reliable record of scientific progress (Jager and Leek 2014). However, it is worth mentioning that 5 journals do not represent the whole literature, and also the p values in abstracts do not represent the whole experimental design.

In turn, the ninety-five-percent boundary introduced by Ronald Fisher and the need for researchers to pass this statistical test would be contributing to the problem of lack of reproducibility – mainly derived from the significance chasing issue (Ware and Munafo 2015). In this respect, J. Ioannidis stated *"[…] research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles should be interpreted based only on p-values"* (Ioannidis 2005). This is, indeed, a current issue regarding incentives to publish findings: there is bias towards publishing positive statistically significant outcomes (Emerson et al. 2010), as it is easier to get 'positive' results accepted in 'good' journals, whereas negative results that are more likely to be true-negative results are disappearing (Ioannidis 2011, Fanelli 2012). Moreover, negative results can remain unpublished because researchers prefer not to submit them and/or because journal editors and peer reviewers are more likely to reject them (Fanelli 2010).

The chance to win the 'significance lottery' is linked to the study design in which the scientists have freedom in deciding how to collect, analyse and interpret data. One common practice is to perform several statistical analyses and/or data eligibility criteria and then only report those that produce significant results (Cumming 2014, Head et al. 2015). This flexibility has, therefore, helped them become 'lucky' in reaching their *a priori* hypothesis; *i.e.,* by scrutinising the data and reporting only that part of a dataset that yields significant results. This could be achieved by deciding which outliers to exclude, when to stop collecting data, or whether to include covariates. Converging evidence for this effect showed that the frequency of papers declaring significant statistical support for their *a priori* formulated hypotheses increased by 22% between 1990 and 2007 (n = 4656, $p < 0.001$) (Fanelli 2012).

2.3.3 Research misconduct – the dark side of scientific layout

Another reason for lack of reproducibility is due to findings of research misconduct. The Wellcome Trust defined scientific misconduct as fabrication, falsification, plagiarism or deception in proposing, carrying out or reporting results of research or deliberate, dangerous or negligent deviations from accepted practices in carrying out research (The Wellcome Trust 2002). The biomedical sciences seems particularly affected by research fraud (Goodstein 2010); indeed, the 'Hall of Infamy' in science includes a considerable number of biologists and clinicians (Broad and Wade 1982).

The Summerlin's faked 'transplants' is a famous case in point to exemplify in this regard (Broad and Wade 1982). In 1973, William T. Summerlin – an immunologist at the Sloan Kettering Institute of Cancer Research in New York, reported that he could transplant tissue grafts from one species of mouse to another without immunosuppression (Summerli.Wt et al. 1973). According to his finding, it is possible as long as the tissue to be transplanted is placed in culture medium for some time prior to grafting. The experimental evidence that led to this conclusion was supported by the use of mice of different colours: an area of skin taken from a black mouse and transplanted to a white mouse. However, one researcher after another reported an inability to replicate and reproduce the transplants. The lack of replicability and reproducibility in this very case was due to the fact that Summerlin used a black felt-tip pen to darken a transplanted skin patch in the white mice. This 'finding' was exposed as a fraud in 1974 when a laboratory assistant washed off the black ink by using a ball of cotton soaked in alcohol. The scandal was just as great that this scientist is credited with starting biomedical research misconduct.

Most scientists think that research misconduct is uncommon. However, in an anonymous survey on suspected research misconduct – conducted by the Gallup

Organization and funded by the Office of Research Integrity (Wells 2008), principal investigators of NIH-funded research grants were asked a single question: *"In the past three academic years, how many times have you observed or had other direct evidence of researchers in your department (or equivalent organizational unit) allegedly committing research misconduct (falsification, fabrication, or plagiarism) in proposing, performing, or reviewing research, or in reporting research results?"* The report estimated that 1.5% of all research conducted each year would be fraudulent. This would represent about 4659 research misconduct incidents per year by considering the number of scientists supported by NIH (about 155,000).

In addition, the natural scientists Daniele Fanelli, at the University of Edinburgh, carried out the first meta-analysis of surveys that have asked scientists directly whether they have committed or know of a colleague who committed research misconduct (Fanelli 2009). The findings of this meta-analysis showed that about 2% of scientists admitted to have fabricated, falsified or modified data or results at least once, and one-third admit to having engaged in other questionable research practices. In the same meta-analysis, a decrease in admission rates was observed over the years in self-reports but not in non-self-reports. According to Fanelli, this trend might suggests that scientists have become less likely to admit misconduct for themselves rather than less likely to commit it.

2.3.4 Irreproducibility – consequences

The inability to reproduce experimental findings in life sciences and biomedical research has resulted in the retraction of published manuscripts and, consequently, invalidating research breakthroughs and/or discontinuing clinical trials. Two PubMed database surveys from 2000 to 2010 found that the number of retracted articles has risen approximately 10-fold, whereas the number of new papers rose by only about 40% during that time (Steen 2011b, a). According to these surveys, this is consistent with the hypothesis that fraud is increasing more rapidly than scientific mistakes or publications overall. Taking a broad interpretation of this hypothesis, it seems possible that some retractions for mistakes actually represents fraudulent articles. This interpretation is because of the correlated incidence of retractions due to fraud and scientific mistakes (Steen 2011b, a). In numbers, approximately 44% of retracted papers are attributed to misconduct; 28% to honest error; 11% to irreproducible findings; and 17% to other or unstated reasons (Van Noorden 2011). In this context, irreproducible research can delay scientific progress due to the waste of valuable resources (money and time) when researchers try to replicate or build on by fraudulent and misleading claims.

While those numbers support the assumption that scientific misconduct is the main source of retractions (Fang, Steen, and Casadevall 2012), it has also been argued that the rising number of retractions is most likely to be an evidence of the commitment of the scientific community to remove invalid findings from the literature (Fanelli 2013). Be that as it may, the dramatically growing number of retractions in recent years has provided enough material for websites such as Retraction Watch – a blog created by two medical journalists, Adam Marcus and Ivan Oransky, that monitors and reports on retractions from scientific journals [http://retractionwatch.com]. This blog not only has demonstrated that retractions are more common than was previously thought, but also that there are scientists who hold a significant number of retractions. At the top of this list is Yoshitaka Fujii – a Japanese researcher in anaesthesiology, formerly of Toho University in Tokyo. Dr Y. Fujii stands alone as the record-holder for the most retractions by a single author: 183 over two decades of 'research' (Cyranoski 2012).

The lack of reproducibility of published findings not only has been identified as a major issue in science from a pure scientific perspective, but also because it has caused an economic drain that the scientific community can no longer afford. In the United States of America, for instance, it has been estimated that about 28 billion dollars per year is spent on basic and preclinical research that is not reproducible (Freedman, Cockburn, and Simcoe 2015). In addition, taking into account the financial, legal and ethical consequences of the scientific misconduct as a source of irreproducibility, it was estimated that the direct costs of a single case approach US $525,000 (Michalek et al. 2010). This cost would exceed US $100 million if it were to apply to all of the allegations of misconduct reported in the Office of Research Integrity (n =217 cases) or US $2 billion considering the incidents estimated per year according to the Gallup Organization survey (n =4659 cases) (Michalek et al. 2010, Wells 2008).

The impact of irreproducible findings and the influence of the mass media have spread this concern beyond the scientific community, potentially undermining public trust in the research enterprise and the science's image as a whole. In this regard, a couple of years ago, *The Economist* ran a provocative cover story entitled *How Science Goes Wrong* (Anon 2013a). By using the Ronald Reagan's mantra for nuclear agreements: *"trust, but verify"*, this article stated that *"modern scientists are doing too much trusting and not enough verifying – to the detriment of the whole of science, and of humanity."* Similarly, *The New York Times* has also reported on reproducible research; making headlines such as *New Truths That Only One Can See* (Johnson 2014). Therefore, this highlights the importance of publishing reliable research, among other things, because most research is paid for by tax payers, so public trust is essential.

2.4 Standards framework for enhancing reporting

In the life sciences arena, scientists seem to be facing a phenomenon that metaphorically resembles the problem that arose during the building of the Tower of Babel: when God confused the tongues of the people, the builders began speaking different languages and, according to Genesis in the Bible (Genesis 2010) *"[...] they stopped building the city and began to scatter throughout the face of the earth."* Thus, the absence of standards frameworks for performing and reporting research has delayed scientific progress, and generated huge volumes of irreproducible findings.

The above scenario has been presented by the Global Biological Standards Institute (GBSI), Washington, D.C. GBSI commissioned an independent organisation to interview almost 60 stakeholders across the life science community – including academia, industry, government body and other non-for-profit organisations – to evaluate the quality of R&D methodologies, identify areas of concern, and establish recommendations for adopting standards. The report *The Case for Standards in Life Science Research: seizing opportunities at a time of critical need* identified a variety of factors that contribute to life science research irreproducibility, most of them can be traced to the absence of a unifying standards framework (GBSI 2013). In this report, the stakeholders interviewed agreed that there is a need for more standards in life science research, especially now when biology has entered a new era with complex and multidisciplinary approaches for information processing frameworks and high-throughput experiments.

The concept of standards is not new and has been the foundation of progress in science and technology, *e.g.*, the Internet. According to the International Organization of Standardization (ISO), a standard is *"[...] a document that provides requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose."* In daily clinical practice, for instance, the adoption of standards has been essential to reduce variability and improve quality of diagnostic test results, such as blood banking (Green, Allard, and Cardigan 2015).

The GBSI describes two categories of standards (GBSI 2013): 1) Material Standards – well-characterised physical substances, such as chemical or biological reagents, which are used for assay validation and calibration, or in generation of scientific evidence; 2) Written Consensus Standards – documents outlined by community agreement that describe optimal practices. This category includes standards related to analytical and procedural elements for bio-assays, *e.g.*, the guidelines developed to assist clinical laboratories with the performing and reporting of next-generation sequencing analysis (Rehm et al. 2013). In the following paragraphs, I will focus on the written consensus standards for the biomedical community since academic papers are a major way that scientists communicate their findings and ideas.

Over the past decade, researchers have published sets of minimum information that should be reported in biological and medical literature, in order to improve the quality of research publications and of the research process itself, *e.g.*, the minimum information guidelines group, MIBBI (Taylor et al. 2008). These checklists have been developed by experts in particular fields, and have evolved over time to capture only the most essential considerations and, thus, to improve the reporting of biomedical experiments. The hope is that at least by reporting this minimal information, it will allow the readers not only to unambiguously interpret and critically evaluate the conclusions reached, but also potentially compare and reproduce the findings. The definition of standardised experimental descriptions via minimum information standards has greatly facilitated the interchange of data between laboratories (Brazma 2001).

To know what makes an experimental detail essential for reproducing a particular finding, Picasso's Bull would help to explain the concept of minimal information. Pablo Picasso created 'Bull' around the Christmas of 1945. The artist deconstructs the process of drawing a bull in a series of increasingly simplistic lithographs, capturing the bare essence of the beast (Figure 3). By following the same approach, the scientific biomedical community have stripped back complex experimental methods and protocols to reveal the bare essence required for a particular experiment. These essential elements have been documented in guidelines and checklists. Just as the bull missing its horns is no longer recognisable as a bull, the absence of an essential experimental component means one cannot correctly interpret the experiment.

The minimum information checklist or guidelines promote transparency in experimental reporting, enhance accessibility to both data and metadata, and support effective quality assessment, which increases the general value of the scientific evidence. In order to prompt authors to disclose technical and statistical information in their manuscripts, and to encourage referees to consider aspects important for research reproducibility, Nature publishing group created a checklist for life science articles which, although it is not exhaustive, takes into account experimental and analytical design elements that are crucial for the interpretation of research results but are often reported incompletely (Anon 2013b). In addition, organisations such as the Centre for Open Science have encouraged journals to adopt the use of standards for helping promoting reproducibility via reporting and transparency (Nosek et al. 2015).

Some standard initiatives, such as the Minimum Information About a Microarray Experiment (MIAME) (Brazma et al. 2001) and the Minimum Information About a Proteomics Experiment (MIAPE) (Taylor et al. 2007) – just to mention some of them, are being adopted by many journals as a requirement for publishing, such as *Nature Genetics* or the *Journal of Proteomics*. The hope is that harmonization of data annotation will facilitate interoperability between genomics, proteomics and metabolomics data sources, enabling the modelling of comprehensive interaction networks and the elucidation of emergent system-wide properties.

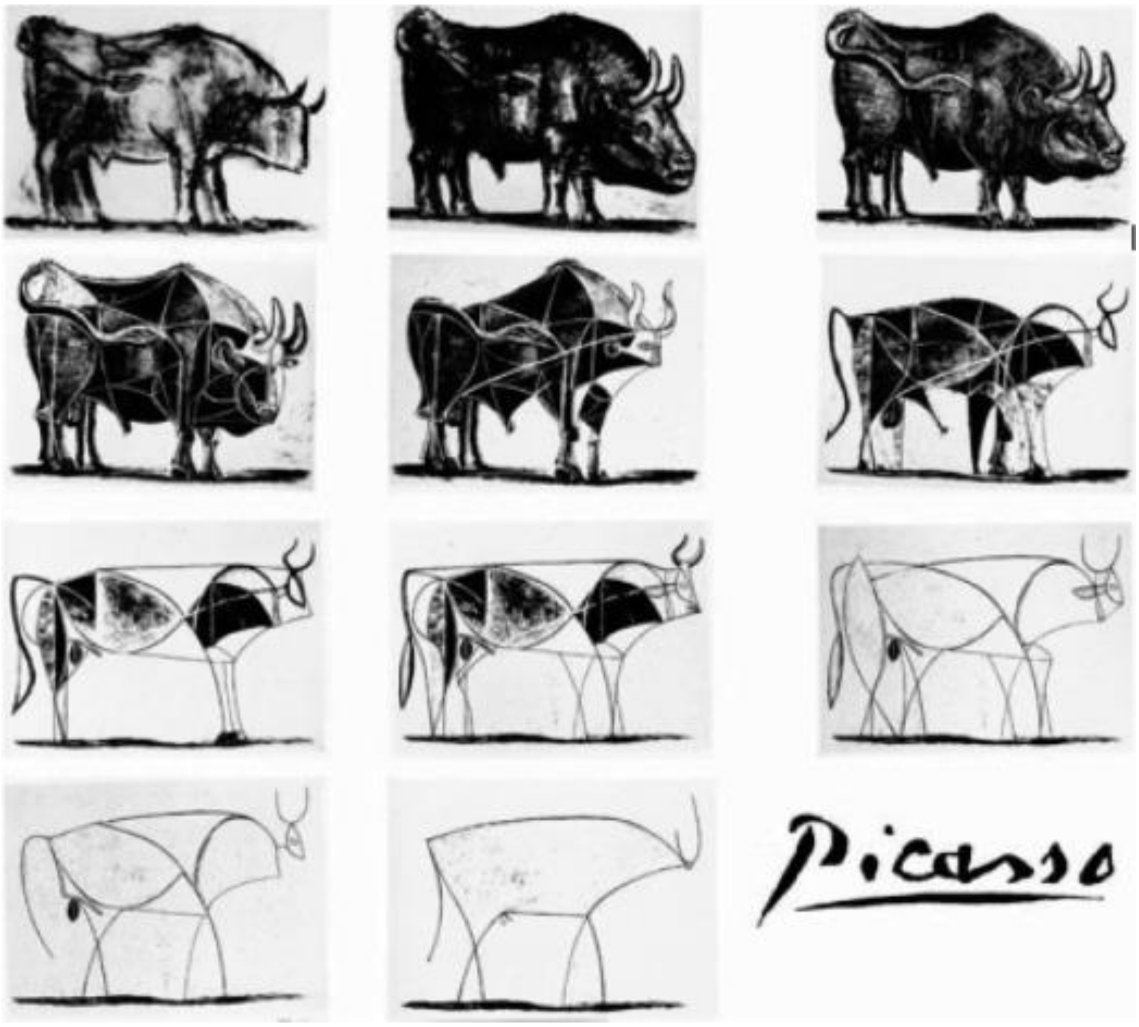


Figure 3. Picasso distils the essence of a bull. Estate of Pablo Picasso/Artists Rights Society (ARS), New York, NY. Imagen from (Gray, Young, and Waytz 2012). © Copyright Policy – open-access by Open-i service of the National Library of Medicine.

In the above context, initiatives such as the BioSharing catalogue [http://biosharing.org] and the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network [http://www.equator-network.org] maintain Web-based sites, freely accessible resources for guidelines, providing straightforward access to extant checklists for basic and clinical research. Both initiatives include guidelines for the description of experiments, as well as data and metadata from these experiments such as the Animals in Research: Reporting In Vivo Experiments (ARRIVE) (Kilkenny et al. 2010). This checklist includes, among other things, the number and specific characteristics of animals used (including species, strain, sex, and genetic background); details of housing and husbandry; and the experimental, statistical, and analytical methods (Kilkenny et al. 2010). All of these factors affect the performance and interpretation of an experiment.

Furthermore, the data and metadata collected by standard checklists have been the basis for the establishment of repositories that take and disseminate such information. This is the case of ArrayExpress (Rustici et al. 2013), one of the major international repositories for high-throughput functional genomics data from both microarray and high-throughput sequencing studies, where data are collected in conformity to the Minimum Information About a Microarray Experiment (MIAME) and the Minimum Information About a Sequencing Experiment (MINSEQE) standards.

Mandatory use of checklists is increasingly a requirement when submitting an article to biomedical scientific journals. The Nature Reproducibility Initiative (Nature 2013), for instance, intend to ensure good reporting of research by reminding authors to disclose sufficient technical and statistical information through an 18-step checklist. Nevertheless, despite the relative improvements when the reporting checklists are endorsed and enforced by journals – particularly by those with higher impact factors, the completeness of reporting remains suboptimal (Baker et al. 2014, Smith et al. 2015). This raises the question of whether the checklists are being ignored. The answer to this question involves all actors of the publishing process. In simple terms, the authors need to be more strongly encouraged to use reporting checklists during the preparation of the manuscripts (Baker et al. 2014), and both the peer-reviewers and editors need to pay more attention to the benefits of using these checklists as powerful management tools to aid decision making (Shamseer et al. 2012, Arnold et al. 2015, Gawande 2010).

2.5 On the importance of improving metadata reporting in scientific studies to increase the value of existing data

Biosciences are in transition from reductionist sciences to integrative sciences, *i.e.*, a systems approach to biology (Zamer 2011). One of the topics of interest, for instance, is the identification of genetic similarities among complex diseases, *e.g.*, autoimmune diseases (Cotsapas et al. 2011). While the reductionist approach has been responsible for the progress seen in biology during the last decades, it offers no methods to obtain a deeper comprehension of the properties of a specific biological phenomenon and the underlying mechanism (Sauer, Heinemann, and Zamboni 2007) (Figure 4).

Thanks to the remarkable developments in computational approaches in recent years, the life-science community will be able to perform a high-level analysis and, so, by using integrative approaches to gain a better understanding of the nature of systems. Nevertheless, an accurate integration of biological data should be accomplished based on metadata, *i.e.*, the experimental method, and, so, it will be possible to draw meaningful inferences via translational discoveries. However, the lack of transparency in reporting the methods used may severely affect the general value of the scientific evidence (Landis et al. 2012, Goodman et al. 2014).

"The term data integration refers to the situation where, for a given system, multiple sources (and possible types) of data are available and we want to study them integratively to improve knowledge discovery" (Gomez-Cabrero et al. 2014). In this context, an increased integration of computational approaches in biomedicine and other domains has led to the development of methods that hinges on mathematical models for the integration of sets of data and metadata, and thus supporting reproducible research (Antezana et al. 2011, Bechhofer et al. 2013). In the following paragraphs, and in order to exemplify the available technology in this area, I will be presenting ontologies as one of the most advanced approaches that can greatly facilitate the integration process by reducing the semantic heterogeneities of data and metadata (Bodenreider and Stevens 2006), and also an infrastructure developed to support data and metadata annotation of the multi-assay experiments (Rocca-Serra et al. 2010). These strategies allow the collection, organisation, exploration, sharing and reuse of information. However, the lack of transparency in reporting could affect the exploitation of the full potential and benefits offered by these kind of technologies and tools.

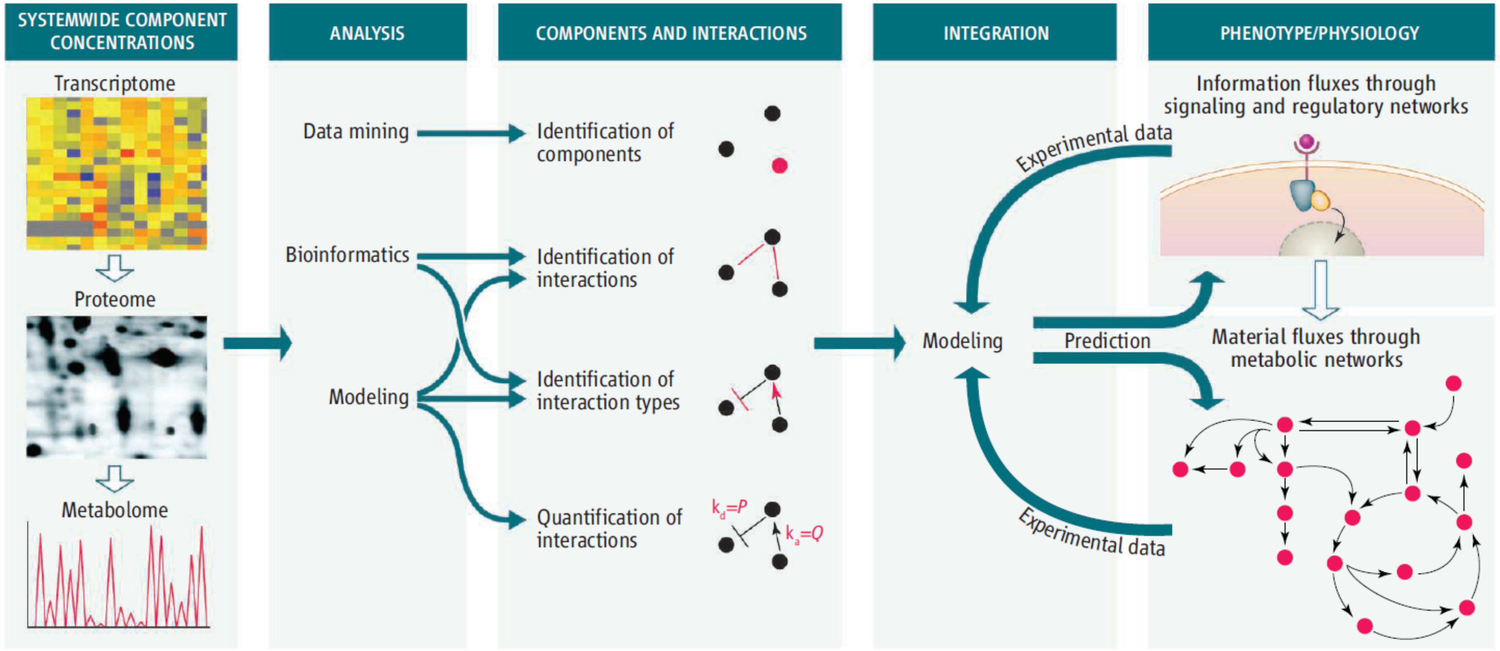

SYSTEMWIDE COMPONENT CONCENTRATIONS
ANALYSIS
COMPONENTS AND INTERACTIONS
INTEGRATION
PHENOTYPE/PHYSIOLOGY
Transcriptome
Proteome
Metabolome
Data mining
Bioinformatics
Modeling
Identification of components
Identification of interactions
Identification of interaction types
Quantification of interactions
k_d=P
k_a=Q
Experimental data
Modeling
Prediction
Experimental data
Information fluxes through signaling and regulatory networks
Material fluxes through metabolic networks

Figure 4 The diagram gives an example for inferring component interactions using computational methods. Taken from Getting Closer to the Whole Picture (Sauer, Heinemann, and Zamboni 2007).

2.5.1 Management of data and metadata: Ontologies

An ontology is normally defined as *"a formal explicit specification of a shared conceptualisation"* (Gruber 1993). Ontologies describe the types of entity in a domain and the relationships between those entities (Stevens et al. 2007). Medicine and life sciences are not only characterised by very large domains, but they are also some of the fields with the largest amount of research on ontologies. In fact, currently there are around 350 bio-ontologies on the NCBO BioPortal (Whetzel et al. 2011, Musen et al. 2012) and a number of databases contain information about genes and proteins, as well as their sequences, functions and expression profiling in several experimental models.

In order to manage the large and complex amount of data produced by biomedical experiments (*e.g.*, information from high-throughput technologies such as genomics, transcriptomics, proteomics, and metabolomics) research communities have developed bio-ontologies, such as the Ontology for Biomedical Investigations (OBI), to represent experiment data and metadata. OBI represents a biomedical experiment as a process that involves experimental materials (*e.g.*, whole organisms, organs and cells) in several sub-processes. These experimental materials are represented as subclasses of the Basic Formal Ontology (BFO) class material entity which, in turn, is an independent continuant. Material entities are entities that are spatially extended and persist through time (*e.g.*, organism and test tube). In addition, material entities can bear roles (*e.g.*, study subject role) and qualities (*e.g.*, weight). A schema of the main components of OBI and their relations in the modelling of experiments is presented in Figure 5.

Ontologies of this type expresses conceptual information about pre-analytical, analytical and post-analytical conditions that take place in a biomedical experiment (Brinkman et al. 2010, Malone et al. 2010), as well as offering important advantages in terms of description, annotation and integration of datasets. The management and representation of this information is important in order to determine the relationship between quantities and qualities of inputs and outputs in a biomedical experiment, which represents a significant improvement to metadata analysis because the result is understood in the context of the processes that were used to drive it. Understanding the method by which data were produced is important for the interpretation, comparison and integration of those data. In fact, the current advances in this area promise to ease identification of experimental factors that might influence the findings, and thus supporting reproducible research (Bechhofer et al. 2013).

Bio-ontologies have shown their usefulness in areas such as intelligent database integration, knowledge structuring and modelling (Bodenreider and Stevens 2006). The efficiency of these components is essential to facilitate translational discoveries by processing the data with reasoning engines. As a consequence, this advance has meant that biomedical researchers use ontologies to annotate their data (Bodenreider and Stevens 2006). In the context of data management, ontologies play an important role due to the fact that they provide an expressive and well-defined representation format that can be consistently and unambiguously interpreted.

2.5.2 Annotation of data and metadata: the ISA infrastructure

The adoption of standard formats, minimum information guidelines and terminologies for the annotation of experimental data and metadata is a crucial step, especially when considering data integration and sharing aiming at later reuse. Annotation is a time-consuming task that must be supported by software tools, which should also enable querying, reasoning and analysing the data. The Investigation/Study/Assay (ISA) Infrastructure (Rocca-Serra et al. 2010) was the first one source available desktop software suite to support both experimentalists and curators in the annotation and local management of multi-assay experiments. The software suite comprises several platform-independent Java-based software components for local use. The components work both as stand-alone applications and as a unified system.

The ISA tools are open source [http://isatab.sourceforge.net/index.html] and follow a modular architecture. The ISA-Tab format, for instance, was designed to address the pressing need of reporting and communicating data and metadata from biomedical and life sciences studies employing combinations of omics technologies along with more conventional methodologies (Sansone et al. 2008). The ISA-Tab includes the description of the contextual information of experiments such as the sample characteristics, the technology and measurement types, the parameters of the instrument used, among other things. In addition, the ISA-Tab assists in formatting data and metadata for submission to public repositories, in compliance with emerging minimum information reporting standards, which is crucial for the reproducibility and the comparability of the experiments and posterior data reuse.

An example of how ISA-Tab can be used in Trypanosomiases investigations at experimental level: Suppose there are files from 22 studies in which the gene expression due to *T. cruzi* infection was evaluated in both cells, e.g. cardiomyocytes, and animal models, e.g. heart tissue. Each study file describes the data read out for experimental infection using two *T. cruzi* strains and evaluated between 1-10 days post-infection. Suppose 12 out of the 22 studies were carried out via microarray and the remaining 10 studies via 2D electrophoresis and mass spectrometry. Each microarray slide contained 2400 genes spotted in duplicate and each 2D electrophoresis-mass spectrometry assay led to the identification of 120 differentially regulated proteins. Thus, on the one hand, each microarray assay file contains 2400 rows and, on the other hand, each 2D electrophoresis-mass spectrometry assay file contains 120 rows; describing the gene expression state measured in each time per each *T. cruzi* strain. Figure 6 depicts a simplified representation of the ISA-Tab files in this example.

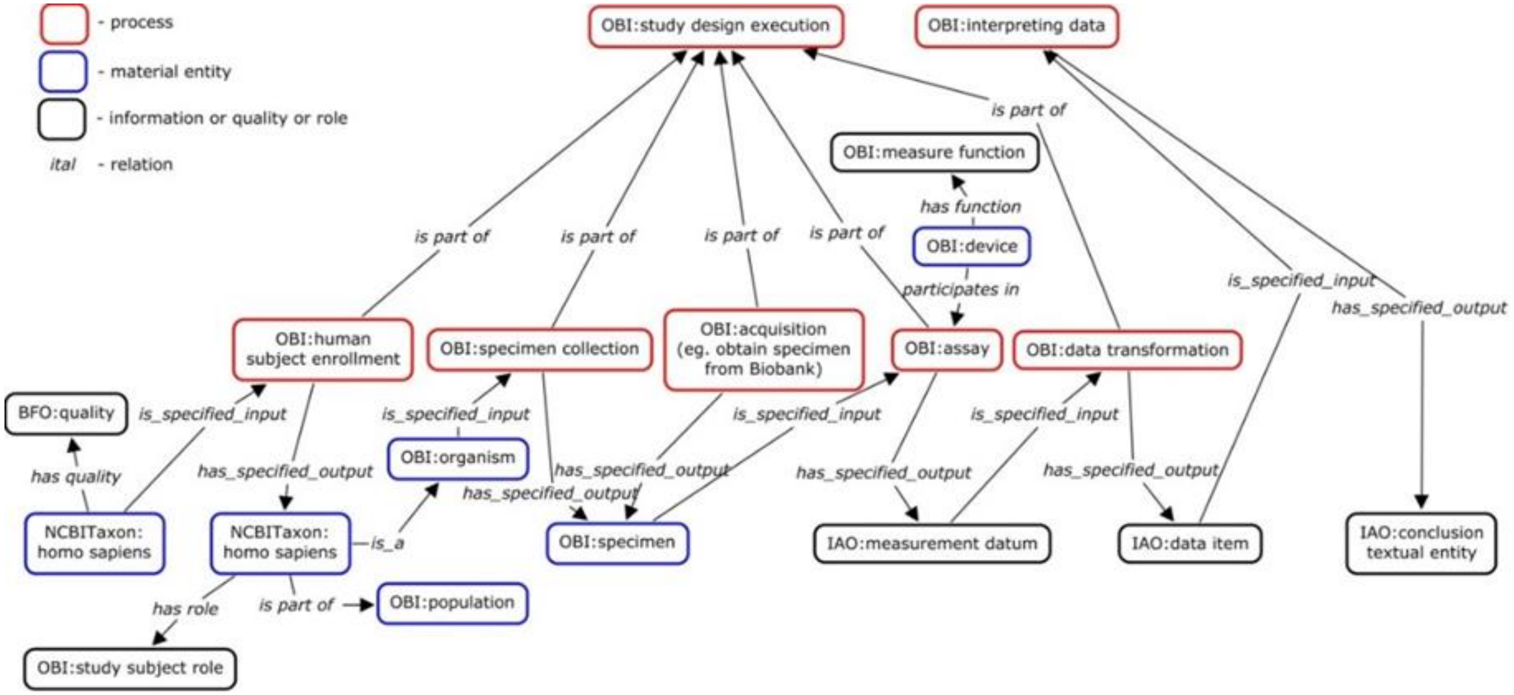


Figure 5. Main components of OBI and their relations in the modelling of experiments. Taken from Jie Zheng, University of Pennsylvania; 3rd International Conference on Biomedical Ontology, 2012.

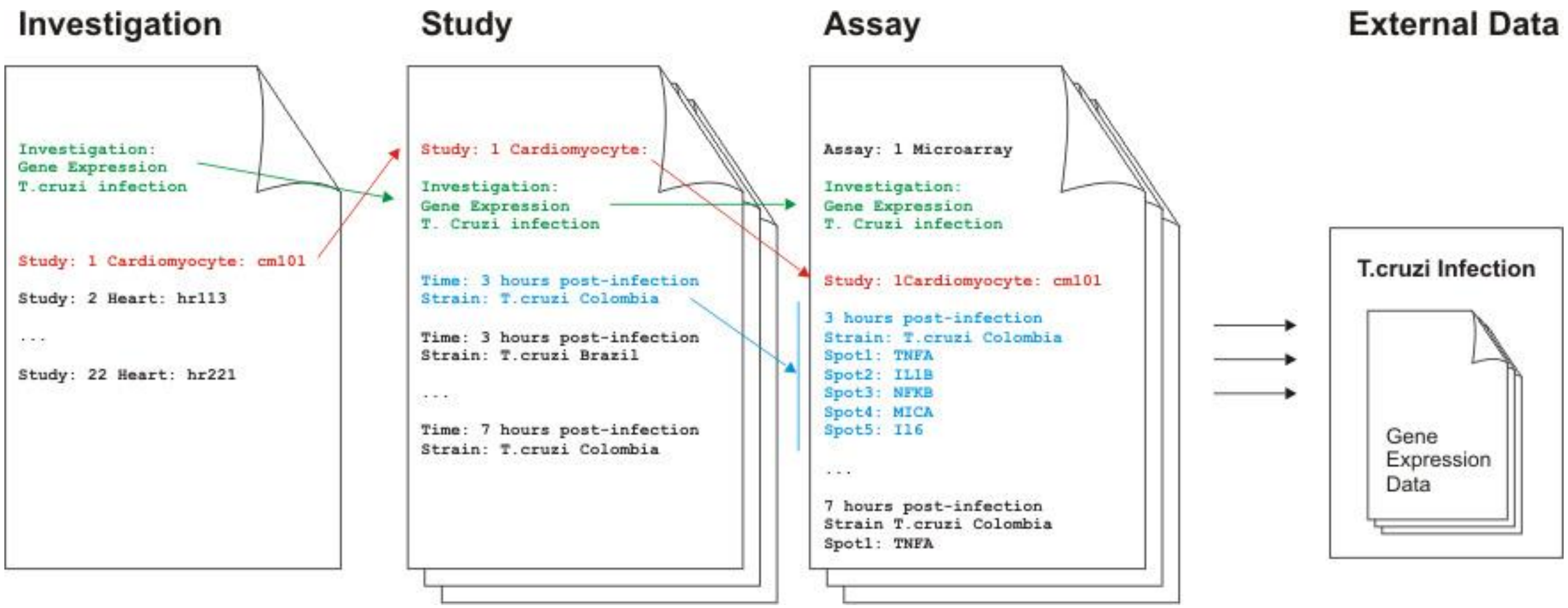


Figure 6. Simplified representation of the ISA-Tab files in an investigation about gene expression due to *Trypanosma cruzi* infection.

2.6 Summary

There is a growing concern in the scientific community over the lack of reproducibility of many published scientific findings. An examination of this problem suggests that it can be attributed, among other things, to the lack of transparency in reporting. In particular, omissions in reporting the technical nature of the experimental method make it difficult to understand and verify the findings of a research, as well as to draw meaningful inferences via translational discoveries by carrying out data integration in a metadata context. As a response to this issue, the minimum information standards community has developed guidelines as an attempt to improve the quality of scientific reporting in biosciences. However, the completeness and accuracy of reporting remains suboptimal – even when the reporting checklists are endorsed by journals. Therefore, there is an urgent need to address this problem. In this Thesis, and in order to promote a change of scientific culture towards the care of scientific data and metadata, will be presented two proof-of-concept applications for the assessment of reporting of biomedical experimental methods through the publication process. These strategies should help to prevent incomplete reporting from entering the literature.

Chapter Three

The status of methods reporting in published parasitic diseases experiments

The content of this chapter was published in the journal *PLoS One*; full citation:

Flórez-Vargas O, Bramhall M, Noyes H, Cruickshank S, Stevens R, Brass A. The quality of methods reporting in parasitology experiments. *PLoS One*. 2014; 9(7):e101131.

The starting point for the development of any assessment strategy is the analysis of the perceived needs and the potential users. In this sense, infectious diseases constitute an interesting area for development of an assessment instrument because they are one of the most common causes of morbidity and mortality worldwide and the microbiologist' community is a big one.

Among the infectious diseases, the so-called neglected tropical diseases require special attention since they are difficult and costly to manage; the disease burden is poorly understood; and they have relatively low investment in research and development (World Health Organization 2010). These infections affect humans and animals, usually with fatal consequences unless treated. Many studies have been carried out to explore their physiopathology, as well as their genetic susceptibility. Nevertheless, a considerable part of this evidence is controversial; probably stemming, at least in part, from differences in pre-analytical, analytical and post-analytical factors, as well as experimental design and data analysis [*e.g.,* see (Vespa, Cunha, and Silva 1994) and (Cummings and Tarleton 2004) for *Trypanosoma cruzi* infection].

Those controversial experimental evidence could be solved by knowing how the data were produced. This is because among a lot of possible explanations which must be considered when the results differ, the ones that are related to differences in experimental protocol are the most likely reasons to be found between two or more experiments. However, in order to find these differences, the experiments need to be compared and integrated with other sources of information to determine whether there is any conclusive evidence. Nonetheless, any two or more experiments are comparable and integrable to the extent that they provide a minimum set of information that describes the particular experiment.

Accordingly, we have created a checklist that included the minimum information that should be provided when describing infectious disease experiments, and which one impacts on its experimental models, both *In Vitro* and *In Vivo*. This checklist was created based on principles of replicability and reproducibility which state that published scientific literature discloses all necessary and relevant information to allow the experiment to be repeated (Drummond 2009). This information not only is useful to replicate and reproduce experiments, but also to compare and integrate those experiments. Therefore, and due to the complexity of the infectious diseases, making explicit the methodological context of an infectious disease experiment under the principles of replicability and reproducibility is a basic and important requirement to understand the pathogenesis of the disease in spite of the limitations of the models.

In this chapter we evaluated the reported information on experimental methods in published infectious diseases experiments that should enable a valid comparison of research findings.

OPEN ACCESS Freely available online

PLOS | ONE

The Quality of Methods Reporting in Parasitology Experiments

Oscar Flórez-Vargas[1], Michael Bramhall[1], Harry Noyes[2], Sheena Cruickshank[3], Robert Stevens[1], Andy Brass[1,3]*

1 Bio-health Informatics Group, School of Computer Science, University of Manchester, Manchester, United Kingdom, **2** School of Biological Science, University of Liverpool, Liverpool, United Kingdom, **3** Manchester Immunology Group, Faculty of Life Science, University of Manchester, Manchester, United Kingdom

Abstract

There is a growing concern both inside and outside the scientific community over the lack of reproducibility of experiments. The depth and detail of reported methods are critical to the reproducibility of findings, but also for making it possible to compare and integrate data from different studies. In this study, we evaluated in detail the methods reporting in a comprehensive set of trypanosomiasis experiments that should enable valid reproduction, integration and comparison of research findings. We evaluated a subset of other parasitic (*Leishmania*, *Toxoplasma*, *Plasmodium*, *Trichuris* and *Schistosoma*) and non-parasitic (*Mycobacterium*) experimental infections in order to compare the quality of method reporting more generally. A systematic review using PubMed (2000–2012) of all publications describing gene expression in cells and animals infected with *Trypanosoma spp* was undertaken based on PRISMA guidelines; 23 papers were identified and included. We defined a checklist of essential parameters that should be reported and have scored the number of those parameters that are reported for each publication. Bibliometric parameters (impact factor, citations and h-index) were used to look for association between Journal and Author status and the quality of method reporting. Trichuriasis experiments achieved the highest scores and included the only paper to score 100% in all criteria. The mean of scores achieved by *Trypanosoma* articles through the checklist was 65.5% (range 32–90%). Bibliometric parameters were not correlated with the quality of method reporting (Spearman's rank correlation coefficient <-0.5; $p>0.05$). Our results indicate that the quality of methods reporting in experimental parasitology is a cause for concern and it has not improved over time, despite there being evidence that most of the assessed parameters do influence the results. We propose that our set of parameters be used as guidelines to improve the quality of the reporting of experimental infection models as a pre-requisite for integrating and comparing sets of data.

Citation: Flórez-Vargas O, Bramhall M, Noyes H, Cruickshank S, Stevens R, et al. (2014) The Quality of Methods Reporting in Parasitology Experiments. PLoS ONE 9(7): e101131. doi:10.1371/journal.pone.0101131

Editor: Sylvie Bisser, INSERM U1094, University of Limoges School of Medicine, France

Received January 17, 2014; **Accepted** June 3, 2014; **Published** July 30, 2014

Funding: This research is supported by a scholarship 'Francisco José de Caldas' from Colciencias to Oscar Flórez-Vargas and a co-funding scholarship from EPSRC and Epistem Ltd awarded to Michael Bramhall. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: This manuscript has not been published and is not under consideration for publication elsewhere. This research was supported by a scholarship 'Francisco José de Caldas' from Colciencias to Oscar Flórez-Vargas and a co-funding scholarship from EPSRC and Epistem Ltd awarded to Michael Bramhall. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

* Email: andy.brass@manchester.ac.uk

Introduction

In this study, we evaluated the reported information on experimental methods in published infectious disease experiments that should enable a valid comparison of research findings. It has been claimed that most published research findings are false [1] and concern about this is spreading beyond the scientific community, making the cover of The Economist recently [2], and potentially undermining public trust in science. Amongst the scientific community there is a growing concern over the related problem of lack of reproducibility [3,4]. The depth and detail of reported methods directly contributes to the replicability, reproducibility and comparability of experimental work. Replicability is the exact repetition of an experiment to obtain the same results, reproducibility is the repetition of an experiment with small modifications, e.g. the changes that will inevitably occur when conducting the same experiment in different laboratories [5,6]. If results are replicable but not reproducible they may be of little value since they are likely to be idiosyncratic to the precise conditions used and further inference from the results will be problematic. Comparability is essential to facilitate translational discoveries by making it possible to aggregate data from multiple experiments in a single meta-analysis and answering questions not addressed by the original investigators. The information reported in the Materials & Methods section of an article plays a fundamental role in achieving this aim. In the biomedical field, for instance, the Uniform Guidelines of the International Committee of Medical Journal Editors state that authors should include technical information in sufficient detail to allow the experiment to be repeated by other workers [7]. However, the guidelines are not strictly adhered to and, consequently, the lack of methodological information can make the tasks of replicating, reproducing or comparing results by non-specialists in a field problematic.

Over the past decade sets of minimum items of information have been published that should be reported about a dataset or an experimental process [8]. This allows readers not only to unambiguously interpret and critically evaluate the conclusions reached, but also to potentially replicate, reproduce and compare the experiments. The minimum information checklist or guidelines seek to promote transparency in experimental reporting, enhance accessibility to data and support effective quality assessment, which increases the general value of data, and therefore of the scientific evidence. In this sense, some standard initiatives, such as the Minimum Information About a Microarray Experiment (MIAME) [9] and the Minimum Information About a Proteomics Experiment (MIAPE) [10], have been adopted by several journals, such as Nature Genetics or the Journal of Proteomics, as a requirement for publication.

To address the issue of reproducibility in the context of biomedical experiments, we looked at experimental infection models with a particular focus on the trypanosomiases, which are a widespread group of complex infectious diseases caused by flagellated protozoa of the genus *Trypanosoma*. These infections affect humans and animals, often with fatal consequences unless treated. In humans, African (sleeping sickness) and American (Chagas disease) trypanosomiases are responsible for considerable morbidity and mortality, affecting millions of people every year [11–13]. Moreover, human economic welfare in Africa is also affected by these diseases due to loss of livestock production [14]. The outcome of infection with both American and African trypanosomes depends on both the host and parasite genetic background as well as on environmental variation [15–17]. In addition, the trypanosomiases have been labelled as "neglected" because their study hovers in the margins of international health; there is a smaller investment in their research and development and as a result they are less well understood. Hence, an important task is to integrate and compare data from their studies in order to augment the value of this data.

Many studies have been carried out to explore the physiopathology of sleeping sickness and Chagas disease, as well as their genetics. At the time of writing, a PubMed search from 2000–2013 retrieved 1558 and 4248 journal articles containing the MeSH (Medical Subject Headings) terms "Trypanosomiasis, African" and "Chagas disease", respectively. Despite the large amount of published research, our understanding of the underlying mechanisms involved in these diseases is still limited. It is likely that this can be partly explained by the inherent difficulty in making direct comparisons between the results of independent *Trypanosoma* infection experiments.

Currently we have data from studies carried out in experimental models of trypanosomiasis. However, a considerable part of this evidence is controversial or contradictory; probably stemming from differences in pre-analytical, analytical and post-analytical variables, as well as experimental design and data analysis. In Chagas diseases, for instance, the role played by the Th17 immune response, T regulatory cells and Nitric Oxide may be critical to the outcome of infection [18–20] or these immune factors may have opposing effects or not be required [21–23]. Therefore, it is important to know how the data were produced in order to deal not only with the biological complexity of these diseases, but also to permit the replicability, reproducibility and, especially in the case of contradictory results, the comparability of research findings. In order to assess how easy it would be to replicate, reproduce or compare experiments we have undertaken a systematic review of all publications describing gene expression experiments in model organisms infected with these parasites. We have defined a list of essential parameters describing the parasite, the host and the infection that should be reported and for each experiment we have scored the number of those parameters that are reported. In order to determine whether our findings can be generalised to other diseases we have used the same method to assess a subset of papers on *Leishmania*, *Toxoplasma* and *Plasmodium*. A subset of papers that utilised the intestinal helminth parasite *Trichuris muris* or *Schistosoma sp.* were used as a comparative control in order to determine the relevance of the checklist in a non-protozoan parasite infection model. In addition, a subset of papers from a non-parasitic infection model (*Mycobacterium*) were used in order to determine whether this issue is unique to parasitology or has wider implications.

Results

Search strategy

A total of 23 papers on *Trypanosoma* experiments were identified for inclusion in the review. The search in PubMed provided a total of 5878 references with the MeSH term "Trypanosomiasis", of which 104 were related with terms "Genes" and "Proteins", 35 with "Microarray Analysis", and 27 with "Proteomics". After adjusting for duplicates 163 remained. The abstracts of these papers were reviewed manually and 139 were discarded because they did not meet the selection criteria (Figure 1 and Table 1). The remaining 23 references [24–46] were the corpus of papers identified that reported on gene expression profiling in the host due to an experimental *Trypanosoma* infection. A subset of 10 articles each of the closely related protozoan parasites *Leishmania* [47–56], *Toxoplasma* [57–66] and *Plasmodium* [67–76] were included for comparison. In addition, 10 articles of *Trichuris* [77–86] and *Schistosoma* [87–96] parasitic worm experiments, and 10 articles of *Mycobacterium* [97–106] experiments as a non-parasitic infection model were included in order to contrast the quality of method reporting in *Trypanosoma* experiments to other models and to determine the applicability of the checklist to different experimental systems.

Quality of method reporting

To assess the quality of method reporting in *Trypanosoma* experiments, each paper was checked for reporting of information in three domains: the parasite, the host and the experimental infection. The scores are listed in Tables S1, S2 and S3. A mean of 65.5% (SD = 15.12%) of the information required to reproduce an experiment was reported in this set of papers. No article met all criteria that should be reported in a *Trypanosoma* experiment according to our checklist (range 32–90%), although two studies [27,41] scored at 100% out of the available criteria for the parasite and host domains (Tables S1 and S2). The number of articles that met all criteria was higher in the parasite domain (6 out of 23 articles), however the number of criteria met by all the articles was higher in the host domain (7 out of 12 criteria) (Figure 2, Tables S1, S2 and S3). In the experimental infection domain, the inoculum was the only criteria met by all articles, whereas the viability criteria for both cells and parasites were not met in full by any of the studies (Table S3).

Bibliometric indices

Different journals have different criteria for publication in order to enhance the quality of research and to prevent publication of poor findings. However, these safeguards are not always successful; limited space for the method section or forms of bias in the peer review process are some of the issues that have generated serious discussion in several scientific journals [107]. Thus, to discover whether there was an association between bibliometric parameters and the quality of method reporting in *Trypanosoma* experiments, the journal impact factor, the h-index of the corresponding author

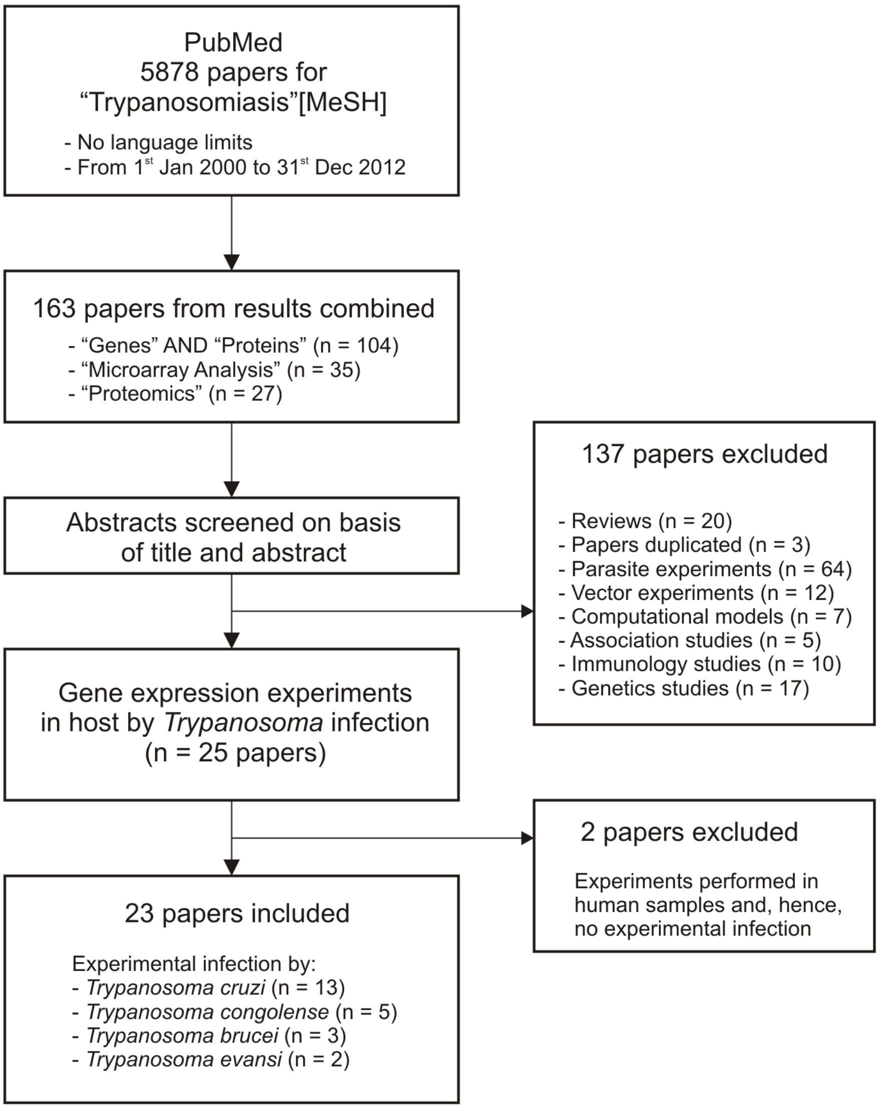

PubMed
5878 papers for
"Trypanosomiasis"[MeSH]
- No language limits
- From 1st Jan 2000 to 31st Dec 2012
163 papers from results combined
- "Genes" AND "Proteins" (n = 104)
- "Microarray Analysis" (n = 35)
- "Proteomics" (n = 27)
Abstracts screened on basis of title and abstract
137 papers excluded
- Reviews (n = 20)
- Papers duplicated (n = 3)
- Parasite experiments (n = 64)
- Vector experiments (n = 12)
- Computational models (n = 7)
- Association studies (n = 5)
- Immunology studies (n = 10)
- Genetics studies (n = 17)
Gene expression experiments in host by Trypanosoma infection (n = 25 papers)
2 papers excluded
Experiments performed in human samples and, hence, no experimental infection
23 papers included
Experimental infection by:
- Trypanosoma cruzi (n = 13)
- Trypanosoma congolense (n = 5)
- Trypanosoma brucei (n = 3)
- Trypanosoma evansi (n = 2)

Figure 1. Study selection process for *Trypanosoma* studies.
doi:10.1371/journal.pone.0101131.g001

and the number of citations of the article were compared with the scores for the quality of method reporting. No correlation was observed between method reporting scores and impact factor or h-index (Figures 3A and 3B). However, a significant negative correlation was observed when the scores for method reporting were correlated with the number of citations of the article obtained from Google Scholar ($r = -0.42$; $p = 0.044$, $n = 23$) but not with citations from the Web of Sciences ($r = -0.35$; $p = 0.105$, $n = 23$) (Figure 3C). Interpretation of this observation is confounded by the tendency of older papers to have more citations (Google Scholar: $r = -0.40$; $p = 0.057$, $n = 23$; Web of Sciences: $r = -0.42$; $p = 0.046$, $n = 23$; Figure 3D). There was no correlation between the quality of method reporting and the year of publication, which remained constant during the last 12 years (Figure 4).

In order to identify relations between the quality of methods reporting in *Trypanosoma* experiments and the experience of the journal with publishing papers about trypanosomiasis, we compared the scores achieved for the articles (arithmetic mean was calculated for two or more papers) with the number of articles about trypanosomiasis in the journal in which the articles were published. This comparison showed that the number of articles published in any one journal about trypanosomiasis was not associated with an increase in the quality of methods reporting. The journals with most and fewest articles published about trypanosomiasis between 2000 and 2012 were the American Journal of Tropical Medicine and Hygiene with 172 papers and Genes and Immunity with only three papers (Table S4). Nonetheless, the article that received the lowest score in the reported information (32%) was published in the American Journal of Tropical Medicine and Hygiene [40], whereas the mean score for articles published in Genes and Immunity [36] was almost double this value (60%) (Figure 5).

Comparison with other parasitic diseases

In order to test whether our observations about the quality of method reporting were a general phenomenon or whether they were specific to trypanosomiasis we evaluated 10 articles each on *Leishmania*, *Toxoplasma* and *Plasmodium*; these diseases were chosen because they are also complex and considered public health issues. As in the articles about *Trypanosoma* experiments,

Table 1. Studies characteristics in trypanosomiasis: parasite species, experimental infection models and aims of the studies.

Author, year and journal	Parasite	Infection model	Aim
Amin et al., 2010 Am J Trop Med Hyg	*T. b. brucei*	Mouse	Discover genes differentially expressed in brain of mice at the early and late stages of *T. b. brucei* infection.
Chessler et al., 2009 J Immunol	*T. cruzi*	Mouse	Examine the initial host-parasite interaction in vivo by monitoring changes in global host mRNA levels at the site of intradermal infection of mice with *T. cruzi*.
Costales et al., 2009 BMC Genomics	*T. cruzi*	Cell line	Investigate the impact of intracellular *T. cruzi* infection on host cell gene expression.
Garg at al., 2004 Biochem J	*T. cruzi*	Mouse	Characterise the cardiac metabolic response to *T. cruzi* infection and progressive disease severity.
Genovesio et al., 2011 PLoS One	*T. cruzi*	Cell line	Search for human cell factors that play a role during infection by the protozoan parasite *T. cruzi*.
Goldenberg et al., 2009 Microbes Infect	*T. cruzi*	Primary culture (Cardiomyocytes)	Examine gene profiling of *T. cruzi*-infected cardiac myocytes.
Graefe et al., 2006 PLoS One	*T. cruzi*	Mouse	Analyse genome wide expression differences in the spleen at the point at which the immune response diverges between susceptible and resistant mice, and then match the genomic localisation of differential expressed genes with mapped susceptibility loci.
Hashimoto et al., 2005 Int J Parasitol	*T. cruzi*	Cell line	Report the time-course of transcriptional changes in apoptosis-related genes responsive to Fas stimulation in *T. cruzi* infected cells.
Hill et al., 2005 Vet Immunol Immunopathol	*T. congolense*	Cattle	Investigate the transcriptional response of susceptible cattle to trypanosome infection.
Kierstein et al., 2006 Genes Immun	*T. congolense*	Mouse	Explore the ability of more integrated analysis of genetics of trypanotolerance underlying the response to infection and identify pathways involved in trypanotolerance.
Li et al., 2009 Parasitol Res	*T. evansi*	Mouse	Investigate the global gene expression in the liver and spleen of mice after infection with *T. evansi*.
Li et al., 2011 Exp Parasitol	*T. b. brucei*	Mouse	Examine the effects of *T. b. brucei* infection on the liver and spleen of mice at the molecular level.
Lopez et al., 2008 J Immunol	*T. b. rhodesiense*	Mouse, primary culture and cell line	Define the spectrum of host innate immune response genes that are induced during early trypanosome infection in macrophages *ex vivo* as well as macrophages treated *in vitro* with sVSG.
Manque et al., 2011 Infect Immun	*T. cruzi*	Primary culture (Cardiomyocytes)	Characterise the global response of murine cardiomyocytes after infection by trypomastigotes in a carefully controlled progression.
Meade et al., 2009 Mol Immunol	*T. congolense*	Cattle	Determine the expression levels of AMP and APP genes in PBMC isolated from trypanotolerant and trypanosusceptible cattle experimentally infected with *T. congolense*.
Mekata et al., 2012 Parasite Immunol	*T. evansi*	Mouse	Determine what kinds of inflammatory molecules play roles in the pathogenicity of *T. evansi* infection.
Mukherjee et al., 2003 Parasitol Res	*T. cruzi*	Mouse	Identify genes that could contribute to cardiac remodelling as a result of *T. cruzi* infection.
Mukherjee et al., 2008 Genomics	*T. cruzi*	Mouse	Report the patterns of gene expression during the development of murine chagasic heart disease, encompassing several time points in the transition from acute to chronic disease.
Noyes et al., 2009 PLoS One	*T. congolense*	Mouse	Assess the parameters that influence anaemia in murine *T. congolense* infections using mouse strains that differ in their susceptibility to trypanosomiasis.
O'Gorman et al., 2009 BMC Genomics	*T. congolense*	Cattle	Catalogue and analyse gene expression changes in PBMC from trypanotolerant and trypanosusceptible cattle following an experimental challenge with *T. congolense*.
Soares et al., 2010 J Infect Dis	*T. cruzi*	Mouse	Determine alterations in gene expression in the myocardium of mice chronically infected with *T. cruzi*.
Soares et al., 2011 Cell Cycle	*T. cruzi*	Mouse	Evaluate the efficacy of transplantation of BMC to restore the normal transcriptome in the myocardium of mice chronically infected with *T. cruzi*.
Tanowitz et al., 2011 Cell Cycle	*T. cruzi*	Primary culture (Endothelial cells)	Determine the potential molecular mechanisms by which the parasite-derived TXA_2 modulates Chagas disease progression and limits collateral damage to organs.

doi:10.1371/journal.pone.0101131.t001

no article about *Leishmania*, *Toxoplasma* and *Plasmodium* experiments met all criteria that should be reported on our checklist, although one publication on *Leishmania* [49] scored 100% for the parasite and host domains (Table S5 and S6). There was no significant difference in the percentage of reported information between *Trypanosoma*, *Leishmania*, *Toxoplasma* and

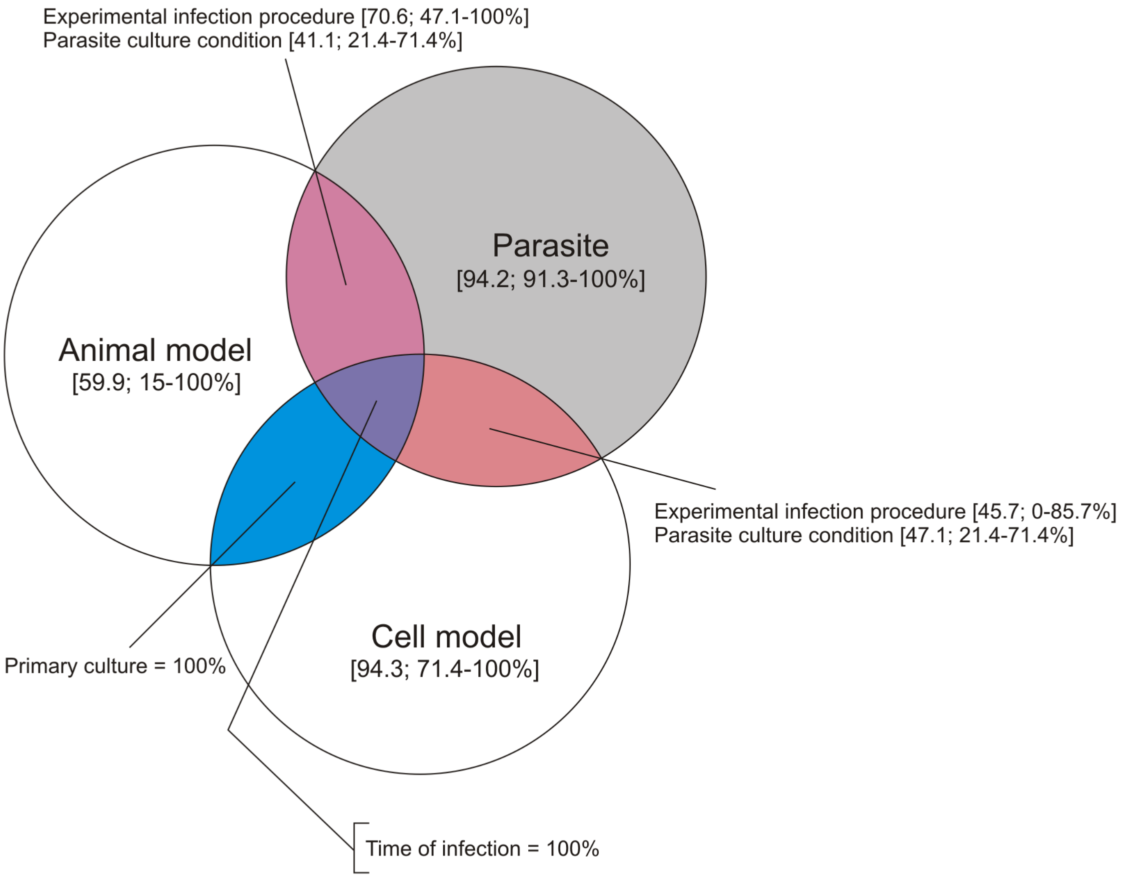

Experimental infection procedure [70.6; 47.1-100%]
Parasite culture condition [41.1; 21.4-71.4%]
Parasite
[94.2; 91.3-100%]
Animal model
[59.9; 15-100%]
Experimental infection procedure [45.7; 0-85.7%]
Parasite culture condition [47.1; 21.4-71.4%]
Cell model
[94.3; 71.4-100%]
Primary culture = 100%
Time of infection = 100%

Figure 2. Venn diagram summarising the quality of methods reporting in the three domains of *Trypanosoma* experiments. The average and range of percentages scored of the quality of methods reporting is shown in brackets.
doi:10.1371/journal.pone.0101131.g002

Plasmodium experiments (Figure 6). The lowest scores were found in the host domain in *Leishmania* and *Toxoplasma* experiments (20%, Table S5). *Plasmodium* experiments had the lowest score in the parasite domain (25%, Table S6) and *Leishmania* had the lowest score in the experimental infection domain (30%, Table S7). No *Toxoplasma* or *Plasmodium* experiment met all of the criteria in any domain (Table S5, S6 and S7).

In contrast to all protozoan parasite experiments, the quality of method reporting in the helminth model of infection by *Trichuris muris* showed the highest scores in all three domains (Figure 6). One *Trichuris muris* experiment [83] successfully scored 100% in all three domains. *Trichuris muris* experiments reported significantly more information than *Trypanosoma* ($p<0.001$), *Plasmodium* ($p<0.001$), *Schistosoma* ($p<0.001$), *Leishmania* ($p<0.01$) and *Toxoplasma* ($p<0.01$) experiments (Figure 6). However, the other helminth model, *Schistosoma sp.*, scored poorly with the second lowest mean reported information (61.16%). *Mycobacterium* (mean reported information 73.96%), the non-parasitic bacterial infection model, scored more highly than *Trypanosoma* (mean reported information 65.46%) but this was not significant.

Validation of scoring methods

The papers from *Trypanosoma* experiments were initially scored by the first and second authors. A specialist in trypanosomiasis then independently scored these papers. The evaluation made by the trypanosomiasis specialist scored 61.6% for the number of criteria from the checklist met in the corpus of articles, whereas a strict evaluation scored 65%. These evaluations scored 63.8% and 64.9% respectively after reviewing the results of both examinations. A linear correlation test (Figure 7A) showed a strong and significant linear correlation between the scores ($r^2 = 0.96$; $p<0.0001$); suggesting that the checklist items measure a common domain and that the personal opinion of the coder does not have an important impact on the scores. In addition, a Bland-Altman test (Figure 7B) was used to verify the agreement between the two evaluations. This analysis showed a good concordance as 16 points were on the line of no difference and 21 fell within the 95% limits of agreement (mean =0.80 and SD: ±2.91), verifying that the scoring was consistent between the evaluators.

Discussion

In order to draw conclusions about the quality of method information reported in articles and its impact on the replicability, reproducibility and comparability of experimental work, we have selected trypanosome infection models as a focus of study. Trypanosomiasis as a complex disease is an appropriate example to understand the importance of the subtlety of experimental variables in the outcome of the modelled disease. Our results indicate that the quality of method information reported in articles about experimental infection with *Trypanosoma spp* is a cause for concern and it has not shown improvement over time, despite there being evidence that most of these variables do influence the results.

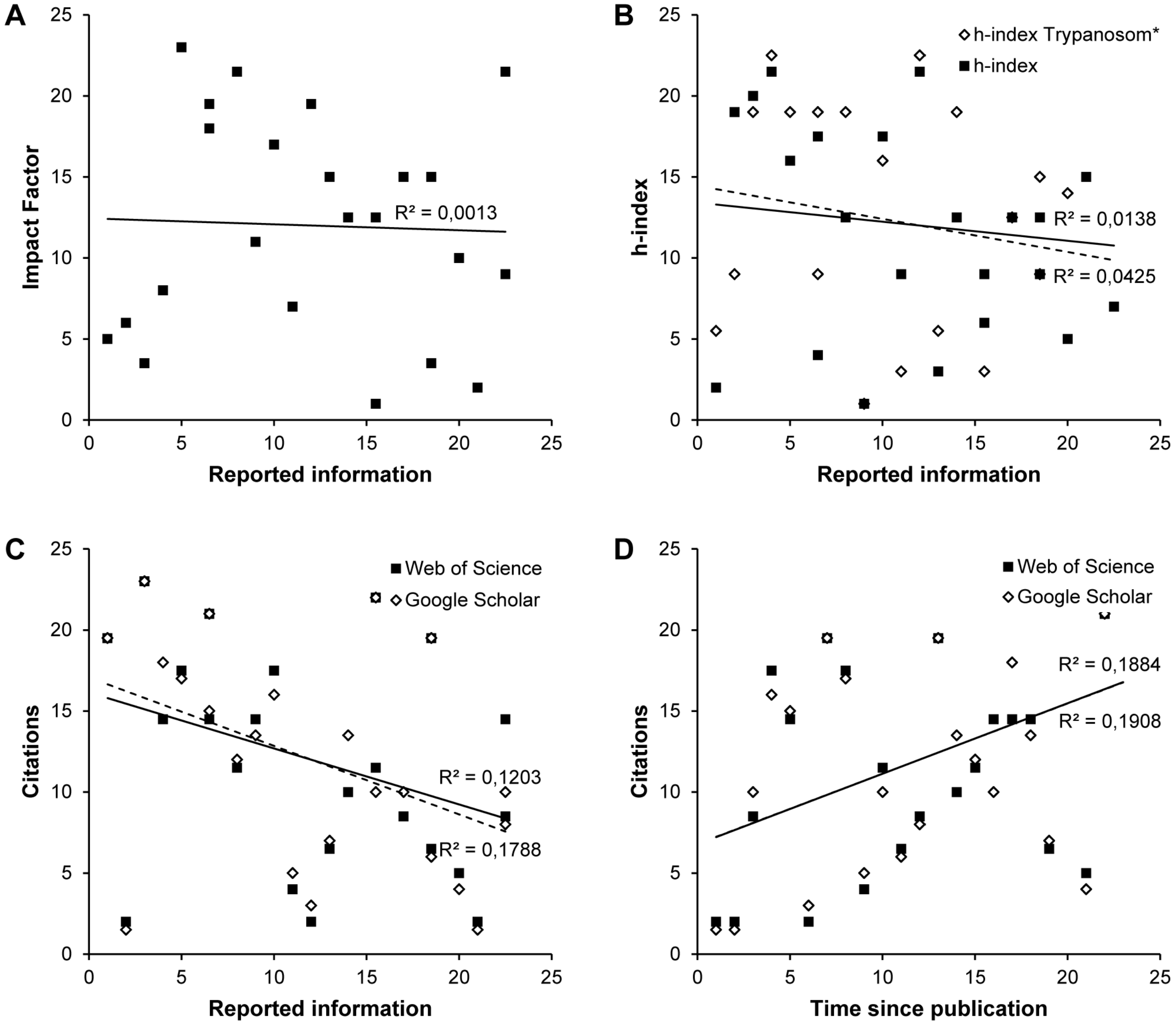


Figure 3. Scatter plots showing the relationship between the quality of methods reporting and the bibliometric indices. Journal impact factor in which the papers were published (A), h-index of the corresponding author (B), and number of citations that the articles have received in other publications (C). Spearman's rank correlation coefficient *r* is shown alongside the regression lines. The figure shows that there is no correlation between the quality of methods reporting and impact factor [$r = -0.04$, $p = 0.868$]. A similar result is shown with h-index, which was searched using the full name of the corresponding author [$r = -0.12$, $p = 0.593$; continuous line] and then filtered by the topic Trypanosom* [$r = -0.21$, $p = 0.345$; broken line]. There is a weak but significant correlation between the quality of methods reporting and the number of citations recorded by Google Scholar [$r = -0.42$, $p = 0.044$; broken line], but not by Web of Science [$r = -0.35$, $p = 0.105$; continuous line]. In order to find out if this association is due to a causal effect of the time of publication, a correlation between the number of citations and the time of publication was done (D), and also a weak but significant correlation was shown with the records of Web of Science [$r = 0.42$, $p = 0.046$; continuous line], but not with Google Scholar [$r = 0.40$, $p = 0.057$; broken line].
doi:10.1371/journal.pone.0101131.g003

Many studies have demonstrated the genetic diversity of *Trypanosoma* species [108,109], as well as the diversity of outcome associated with different parasite strains [17]. The classically described differences in humans infected with different subspecies of *T. brucei* or lineages of *T. cruzi* are well recognized. *T. brucei rhodesiense* causes acute disease and *T. brucei gambiense* causes a more chronic infection [110]. *T. b. gambiense* is divided into two groups which differ in phenotype including pathology [111]. In addition, the cardiomyopathy and digestive forms of Chagas' disease have been associated with *T. cruzi* lineage I and *T. cruzi* lineage II respectively [112]. Strain differences have also been observed in the three major strains of *Toxoplasma*, which vary greatly in their virulence and infection outcome [113]. In addition, isolates of *Trichuris muris* not only differ in virulence but can also trigger changes in the immune response elicited in susceptible hosts [114]; whereas eggs from different strains of *Schistosoma mansoni* cause specific granulomatous responses [115]. Consequently, reporting genus and species of the parasite is not enough; the parasite strain must be reported and if the parasite is a new isolate, it should be characterized.

Virulence of the parasite in all stages of its life cycle plays an important role in the outcome of infection. For example, the failure of laboratory experiments to develop successful malaria vaccines has been attributed to the failure of models to include a

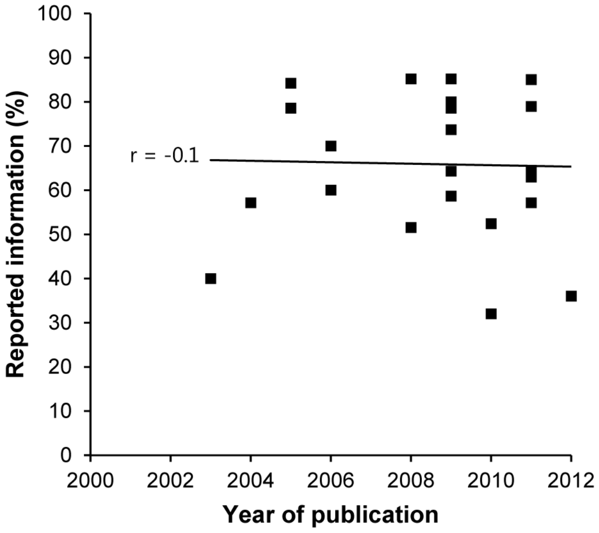


Figure 4. Scatter plots between the reported information in *Trypanosoma* experiments and year of publication. The figure shows that there is no correlation [$p = 0.711$] and that between 2000 and 2012 the quality of methods reporting has remain constant (arithmetic mean =65.5%).
doi:10.1371/journal.pone.0101131.g004

skin stage, which is deemed integral to suppressing host immunity and initiating tolerance to the parasite [116]. In *T. cruzi*, several factors have been implicated in the formation of the infective metacyclic stages. Long-term axenic cultures of *T. cruzi* exhibit a lower capacity to transform into metacyclic trypomastigotes, in comparison to those maintained by alternate invertebrate/vertebrate passages [117]. In addition, the infectivity of *T. cruzi* clones is modified when it is grown in different hosts; a clone passaged through mice has been shown to be more virulent to mice and guinea pigs than the same clone passaged through guinea pigs, the virulence of which remained unchanged [118]. Infection route has also been shown to exert significant impact on the overall course and outcome of infection. In Chagas disease, for instance, the outbreaks associated with food/beverage consumption display severe clinical features in comparison with those of patients that have been infected with *T. cruzi* by vector transmission [119]; a phenomenon that has been associated with the sylvatic biodemes and genotypes of *T. cruzi* [120,121]. In addition, in *Toxoplasma* infections, mice may be susceptible or resistant to infection depending on whether an oral or intraperitoneal challenge is used [122].

Since gender and the corresponding sex steroids affect the immune response [123,124] it is important to specify the gender of experimental animals used. Sex-differences have been demonstrated previously in several experimental infections. For example, in BALB/k mice, males are more resistant to *Toxoplasma gondii* than females [125]. Conversely, in BALB/c mice lacking IL-4, and C57BL/6 p55$^{-/-}$ or p75$^{-/-}$ mice, it is the female mice that are better at expelling *Trichuris muris* than males [84]. However, only 70% of *Trypanosoma* studies reported the sex of animals used in the experimental infection and only 25% reported the gender of animals used to maintain parasite stocks (Tables S1 and S2). In experimental trypanosome infections a gender-related effect has

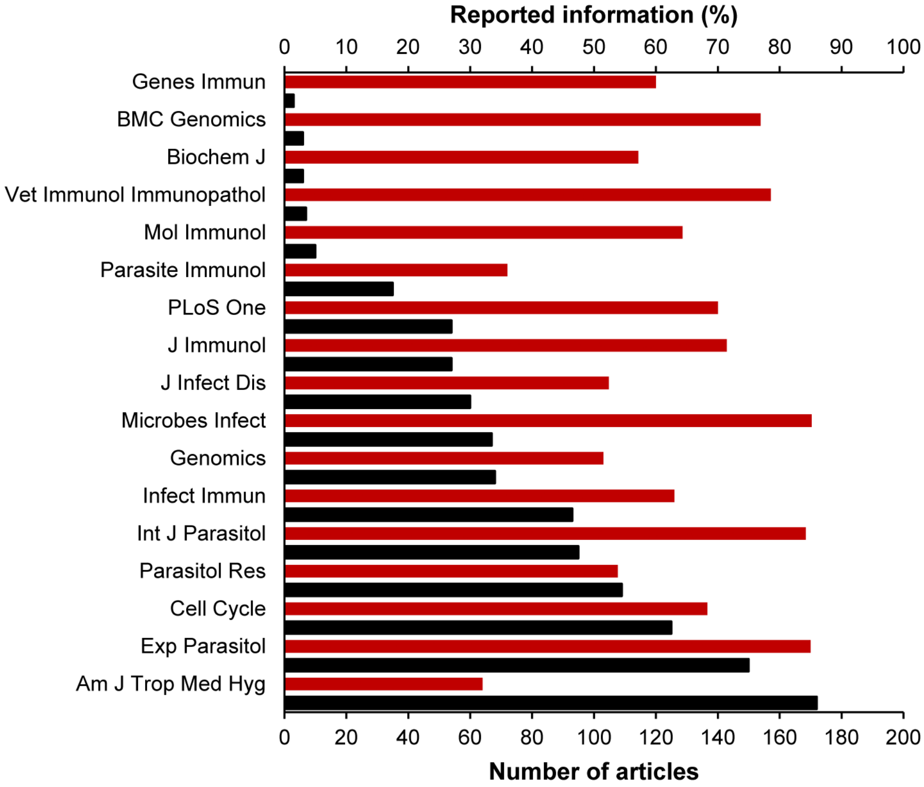


Figure 5. Diagram of articles about Trypanosomiasis[MeSH] published between 2000 and 2012. Number of articles published per journal (black bars) and the percentage of methods reporting (red bars). The figure shows that the quality of method reporting is not related with the number of papers published by any one of the journals.
doi:10.1371/journal.pone.0101131.g005

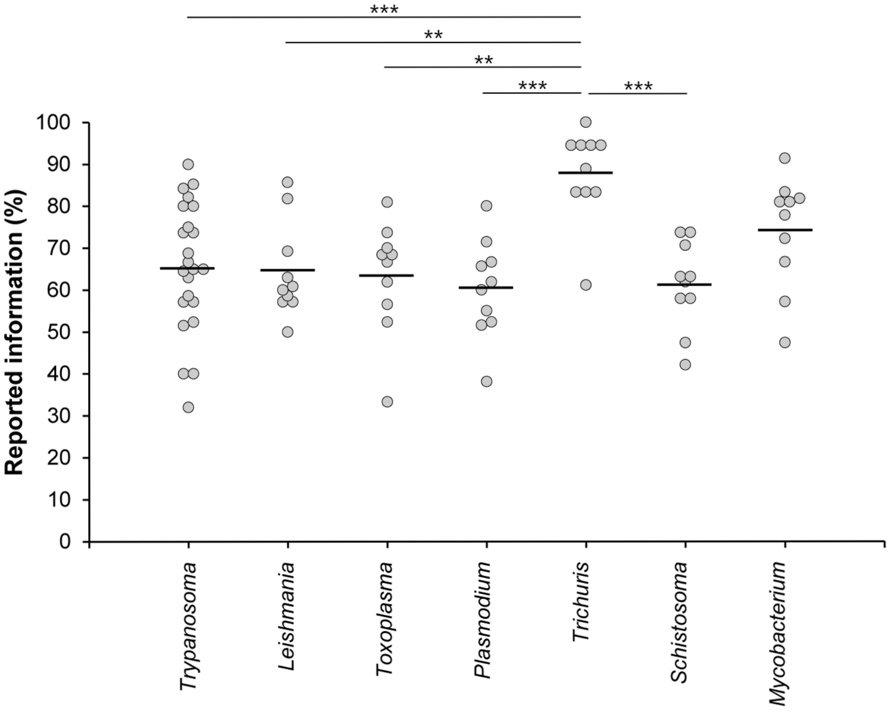


Figure 6. Box-percentile plot to compare the quality of methods reporting in parasitology experiments. Articles about "Trypanosomiasis"[MeSH]; "Leishmaniasis"[MeSH]; "Toxoplasmosis"[MeSH]; "Malaria"[MeSH]; "Trichuris"[MeSH]; "Schistosoma"[MeSH] and "Tuberculosis"[MeSH]. The figure shows that the experimental model of colitis induced by *Trichuris* had the highest scores, followed by tuberculosis, *Trypanosoma*, *Toxoplasma*, *Leishmania*, *Plasmodium* and *Schistosoma* experiments. P values less than 0.01 and 0.001 are represented by ** and *** respectively.
doi:10.1371/journal.pone.0101131.g006

been shown: using BALB/c mice infected with a natural dose of vector-derived metacyclic trypomastigotes of *T. cruzi* (100 parasites/mouse) the peak of parasitaemia in males was about four-fold higher than that in females [126]. Similarly, an experimental infection with a strain of *T. brucei brucei* at 50% of the mouse lethal dose showed that the female were more trypanotolerant than the males and there was no evidence that this was X-linked [127,128]. Housing conditions and social environment also affect the course of experimental trypanosome infections. For example, the parasitaemia levels vary according to whether the animals are kept individually or in a group due to pheromones of the opposite sex [126,129]. Furthermore, hormonal profiles during the oestrous cycle are not only modified by the parasite; such as *T. congolense* [130], but also by the light/dark cycle conditions [131].

In the case of contradictory results, the reporting of the essential parameters that describe a parasitic experimental infection can help to determine the nature of their discrepancies. To exemplify this issue, we have chosen two papers published in the journal Infection and Immunity that were undertaken to assess the role of Nitric Oxide (NO) in immunity to *T. cruzi* infection and their experiments showed contradictory results. Vespa *et al*. claim that NO is involved in control of *T. cruzi*-induced parasitaemia [20], whereas Cummings *et al*. claim that NO is not required for control of *T. cruzi* in the acute or chronic stages of the infection [23]. However, although these studies were carried out using female mice on a C57BL/6 background, the experimental infections were performed using different *T. cruzi* strains, which could explain, at least in part, the differences in their findings: mice infected with 10^4 trypomastigotes of the Y strain showed peak parasitaemia at day 8 that decreased thereafter [20], whereas mice infected with 10^3 trypomastigotes of the Brazil strain showed a peak at day 30 and decreased thereafter [23]. Moreover, although both infections were performed with blood-derived trypomastigotes none of them reported species, gender and age of the animals used to culture the parasite; important parameters that modified the infectivity of *T. cruzi* [117,118]. In addition, there is experimental evidence that shows significant differences among parasitaemia curves between older and younger BALB/c mice infected with a long-term mouse-passaged clone of the *T. cruzi* isolate TolAc1; higher parasitaemia levels were observed in older animals (31-day-old) with lower inoculum (3×10^4 trypomastigotes) than younger animals (8-day-old) with higher inoculum (9×10^4 trypomastigotes) [118]. However, the age of the animals used to evaluate the role of NO in the control of *T. cruzi* infection was reported by Vespa *et al*. but not by Cummings *et al*. [20,23]. Thus, these and other conditions that could also influence the parasitaemia and, hence, the researched outcome should be reported in order to understand the complexity of these parasitoses.

Although the information collected through the checklist should be reported for all *Trypanosoma* experiments, some information could be inferred from the characteristics of the experimental processes, although this depends on the level of expertise of observers (i.e. non-experts and experts). In this way, a factor such as the stage of the parasite used for a *T. cruzi* infection could be easily inferred by an expert since he/she knows that the infectious stage is the trypomastigote. Moreover, both experts and non-experts could also infer many details of the conditions used in cell cultures by assuming experimenters have opted for the most commonly used parameters. For example temperature and CO_2 atmosphere are usually set to 37°C and 5% of CO_2. However, neither experts nor non-experts could infer the species and strain

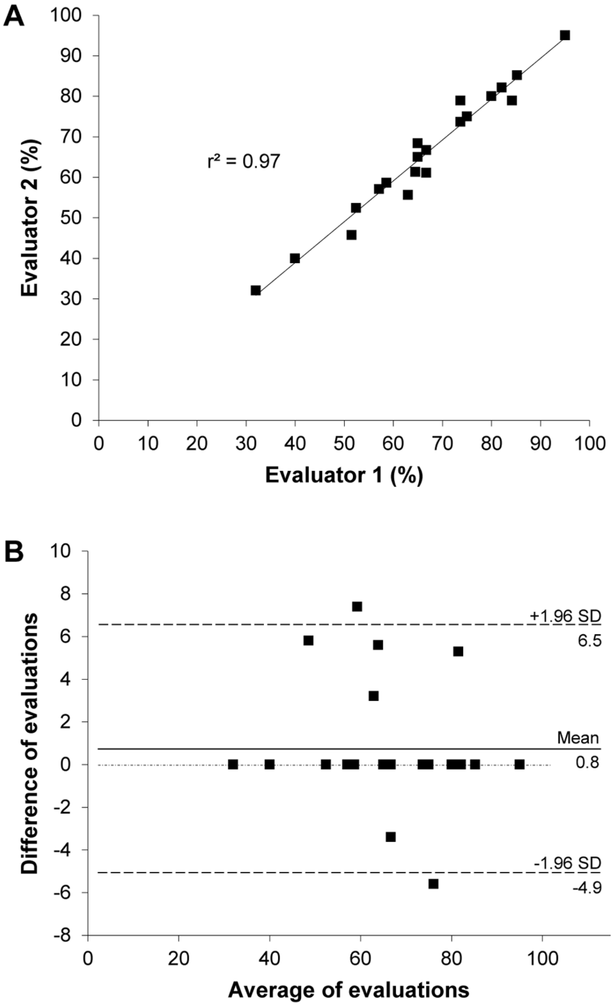


Figure 7. Linear correlation (A) and Bland-Altman (B) plots between scores of method reporting in *Trypanosoma* experiments. Evaluation based strictly on what was explicitly included in the published paper (Evaluator 1) and on interpretations and assumptions determined by an expert in the field (Evaluator 2).
doi:10.1371/journal.pone.0101131.g007

of the parasite; age and gender of the host; or the inoculum used in the infection assays, among others. Thus the validation of data becomes a difficult or impossible task when there is not only not enough information about the method used, but also most of the missing information cannot be inferred, even by an expert.

Providing a high-quality description of the experimental method is important not only to replicate and reproduce, but also to compare and integrate that data and, hence, facilitate translational discoveries. The issues found in reporting methods probably stem, at least in part, from the current structure of scientific publishing, which is not adequate to effectively communicate complex experimental methods. This problem has been recognised, with some journals already introducing editorial measures and methods checklists in order to improve the quality of methods reporting [132].

For the field of trypanosomiasis we have created a checklist to guide parasitologists in reporting *Trypanosoma* experiments (see Annex 1). This checklist included the minimum information that should be provided when describing the parasite, host and infection aspects of those experiments. Our checklist does not cover aspects inherent to each possible experimental assay such as those derived from omics and conventional technologies. In these cases, the BioSharing catalogue [133] should be consulted for checklists: e.g. the Minimum Information About a Microarray Experiment (MIAME) and Proteomics Experiment (MIAPE); and the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE). Moreover, there are other guidelines such as the Minimum Information About a Cellular Assay (MIACA) and the Animals in Research: Reporting *In Vivo* Experiments (ARRIVE) that provide detailed descriptions of experiments performed on cell and animal models.

In conclusion, it has become clear that biomedical science is plagued by findings that cannot be reproduced and/or compared; and the parasitology community is no stranger to this, as has been shown by this study. Nevertheless, the scientific community that works on trypanosomiases is small and many of them know each other personally so in principle it should be possible to change the way that *Trypanosoma* experiments are reported. However, it is important that the scientific community as a whole is engaged with that process. Finally, the checklist has been demonstrated to be applicable to several different infection models and could be implemented to improve the quality of methods reporting for all infection experiments in principle.

Materials and Methods

Search strategy

The method of the literature review follows the recommendations outlined in the PRISMA guidelines [134]. A protocol was designed to identify the method information reported in published articles that utilised experimental infection with *Trypanosoma* species, where the effects on gene expression –transcriptomics and proteomics– of the host were studied. Criteria in three domains were evaluated: characteristics and culture conditions of the parasite, characteristics and maintenance conditions of the host and the infection procedure. The protocol used here for capturing data has not been previously published.

The literature search was conducted using Medline via PubMed. The database was searched in April 2013 for articles that were published between 1st January, 2000 and 31st December,

Table 2. Search terms used in PubMed.

Search	Terms
Search 1	"Genes"[MeSH] AND "Trypanosomiasis"[MeSH]
Search 2	"Proteins"[MeSH] AND "Trypanosomiasis"[MeSH]
Search 3	"Microarray Analysis"[MeSH] AND "Trypanosomiasis"[MeSH]
Search 4	"Proteomics"[MeSH] AND "Trypanosomiasis"[MeSH]

doi:10.1371/journal.pone.0101131.t002

Table 3. Checklist for the reporting of *Trypanosoma* experiments.

Topic	Item#	Description	Does it meet?
Parasite information			
General	1	Identify the species of the parasite	
	2	Identify the strain of the parasite	
	3	Identify the stage of the parasite used	
Culture conditions for parasites grown *in vivo*	4	Identify the species and strain of the animal	
	5	Describe the age of the animal	
	6	Describe the gender of the animal	
	7	Identify the parasite collection sample	
Culture conditions for parasites grown *in vitro*	8	Identify the cell type	
	9	Describe the culture medium used	
	10	Describe the supplements and antibiotics used	
	11	Describe the temperature and CO_2 atmosphere of the culture	
Time of growing	12	Describe the time of growing of the parasite prior to infection	
Host information			
Animals	13	Identify the species and strain of the animal	
	14	Describe the age of the animal	
	15	Describe the gender of the animal	
	16	Describe the housing conditions (light/dark cycle)	
	17	Describe the method of sacrifice	
Cell	18	Identify the cell type	
	19	In primary culture, identify the organ/tissue from which cells come	
	20	In primary culture, describe the method of purification of the cells	
	21	Describe the culture medium used	
	22	Describe the supplements and antibiotics used	
	23	Describe the temperature and CO_2 atmosphere of the culture	
	24	Describe the time of growing of the cells prior to infection	
Experimental infection information			
Animal	25	Describe the inoculum –parasites per animal- used	
	26	Describe the way of inoculation	
	27	Describe the medium of inoculation	
	28	Report the parasitaemia and the time in which the parasitaemia was measured	
	29	Report the mortality of the animals post-infection	
Cell	30	Report the purity of the primary culture	
	31	Report the viability of cells prior infection	
	32	Describe the ratio –parasites per cell- used	
	33	Report the percentage of infected cells	
Parasite	34	Report the viability of parasites prior infection	
	35	Describe the purity of infective forms of the parasite	
	36	Describe the time course (length) of infection	

doi:10.1371/journal.pone.0101131.t003

2012 using the MeSH (Medical Subject Headings) terms as they appear in Table 2. The PubMed Identifier (PMID) numbers were used to identify those articles that were common between "Genes" AND "Trypanosomiasis" and "Proteins" AND "Trypanosomiasis". The search was not limited by study design or by language of publication. The year 2000 was chosen because it was the year in which the first rough draft of the human genome was completed [135,136] and these data were used in many fields of medicine including infectious disease. In addition, we chose to focus on gene expression profiling in the host due to an experimental *Trypanosoma* infection because it provides the broadest evidence about the molecular physiopathology of trypanosomiasis.

In order to compare the quality of method reporting in *Trypanosoma* experiments with the reporting of other parasitic disease infections we collected a subset of *Leishmania*, *Toxoplasma* and *Plasmodium* experimental infection models, since diseases produced by them are also complex and considered public health issues. In addition, as a comparative control of methods reporting in experimental infections, we sought two models of worm infection: one with a simple life cycle (*Trichuris muris*) and another with a complex life cycle (*Schistosoma sp*.); requiring adaptation for survival in fresh water as free-living forms and as parasites in snail intermediate and vertebrate definitive hosts. In addition, we assessed tuberculosis infectious models in order to have a general idea about the quality of method reporting in non-parasitic infection models. Tuberculosis was chosen because it is probably one of the most studied infectious disease.

The same search strategy was carried out where the MeSH term "Trypanosomiasis" was replaced with the following MeSH terms: "Leishmaniasis", "Toxoplasmosis", "Malaria", "*Trichuris*", "*Schistosoma*" and "Tuberculosis". To avoid selection bias, the articles were randomly ordered and the first 10 articles for each extra parasitosis and the non-parasitic infection model (*Mycobacterium*) that described gene expression profiling in the host due to an experimental infection were selected.

Study selection was made by one reviewer and checked independently by a second reviewer, any disagreement was resolved by consensus or by discussion with a third reviewer. Only primary research papers were included in the search. The titles and abstracts of articles were reviewed and analysed in detail to filter out those in which the experiments were performed on the parasite or on vector insects and keep those done on the host. This corpus of articles was then used to confirm eligibility and to extract data.

Structure definition and data extraction

A checklist that contains the minimum information required about the parasite, host and infection to describe an experiment carried out with any *Trypanosoma* species was elaborated by experts in the field of trypanosomiasis research and it is presented in Table 3. Pre-analytical variables in the methods were prioritised in this list because they are critical for interpretation of the results. The terms were classified into three domains according to their roles in a *Trypanosoma* experiment: the host, the parasite and the infection. A data extraction sheet was developed to annotate the information reported in the methods and results sections. Data extraction and quality assessment were carried out by one author and checked by a second reviewer, and inconsistencies were discussed by both reviewers and consensus reached.

Bibliometric indices

Bibliometric parameters were used to determine if they were associated with the quality of method reporting. The impact factor (IF) of each journal was retrieved from the Institute for Scientific Information (ISI) Web of Knowledge's Journal Citation Reports database science edition 2011. The number of citations was measured by the total recorded for each article by Thomson Scientific's Web of Science and Google Scholar in May 2013. For each corresponding author, the h-index was obtained through Thomson Scientific's Web of Science using a citation window up to one year before the article was published. The h-index was searched in two different ways: first, using the full name of the corresponding author and second, filtering the result by topic, using the term "Trypanosom*". The number of articles published for each journal about trypanosomiasis was sought in PubMed using the short name of the journals and the MeSH term "Trypanosomiasis". The search was filtered by time; from 1st January, 2000 to 31st December, 2012.

Validity of scoring methods

An expert in trypanosomiasis tested the quality of reported information on *Trypanosoma* experiments. The expert scored the corpus of articles using the checklist that contains the minimum information required to describe a *Trypanosoma* experiment (Table 3). This evaluation was based strictly on what was explicitly included in the published paper and its results are presented throughout this article. The validity of this assessment was tested based on its agreement with another evaluation based on interpretations and assumptions determined by another expert in the field in order to avoid bias of the retrieval results by interpretation.

Statistical analysis

For each article, the percentage of reported information in each article domain was obtained by direct counting. Linear and Spearman's rank correlations and Bland-Altman comparison were calculated using STATA software [137] and the equivalence of between scores obtained by the evaluators was determined by a correlation test. Comparisons between experimental infection models were performed using a one-way ANOVA in GraphPad PRISM 4 software [138].

Supporting Information

Table S1 Quality measures of the studies that failed to fulfil any one of data of minimal information about the parasite in *Trypanosoma* experiments.
(PDF)

Table S2 Quality measures of the studies that failed to fulfil any one of data of minimal information about the host in *Trypanosoma* experiments.
(PDF)

Table S3 Quality measures of the studies that failed to fulfil any one of data of minimal information about the experimental infection in *Trypanosoma* experiments.
(PDF)

Table S4 Bibliometric indices in reporting *Trypanosoma* experiments.
(PDF)

Table S5 Quality measures of the studies that failed to supply any one of the criteria for minimal information about the parasite in *Leishmania*, *Toxoplasma*, *Plasmodium*, *Trichuris*, *Schistosoma* and *Mycobacterium* experiments.
(PDF)

Table S6 Quality measures of the studies that failed to supply any one of the criteria for minimal information about the host in *Leishmania*, *Toxoplasma*, *Plasmodium*, *Trichuris*, *Schistosoma* and *Mycobacterium* experiments.
(PDF)

Table S7 Quality measures of the studies that failed to supply any one of the criteria for minimal information about the experimental infection in *Leishmania*, *Toxoplasma*, *Plasmodium*, *Trichuris*, *Schistosoma* and *Mycobacterium* experiments.
(PDF)

Checklist S1 PRISMA Checklist.
(DOC)

Author Contributions

Conceived and designed the experiments: OF-V RS AB. Performed the experiments: OF-V MB HN SC. Analyzed the data: OF-V MB HN SC. Wrote the paper: OF-V MB HN SC RS AB.

References

1. Ioannidis JP (2005) Why most published research findings are false. PLoS Med 2: e124.
2. Anon (19 Oct 2013) Unreliable research: Trouble at the lab. The Economist. Available: http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble. Accessed 20 October 2013.
3. Anon (14 Ags 2012) The Reproducibility Initiative. Available: https://www.scienceexchange.com/reproducibility. Accessed 17 January 2014.
4. Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten Simple Rules for Reproducible Computational Research. PLoS Comp Biol 9: e1003285.
5. Casadevall A, Fang FC (2010) Reproducible science. Infect Immun 78: 4972–4975.
6. Drummond C (2009) Replicability is not Reproducibility: Nor is it a good science. Paper presented at: Evaluation Methods for Machine Learning Workshop at the 26th International Conference on Machine Learning; June 2009; Montreal, Quebec, Canada. Available at: http://cogprints.org/7691/7/ICMLws09.pdf.
7. International Committee of Medical Journal Editors Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publications. http://www.icmje.org/#prepare.
8. Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, et al. (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 26: 889–896.
9. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, et al. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29: 365–371.
10. Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK, Jr., et al. (2007) The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol 25: 887–893.
11. Simarro PP, Diarra A, Ruiz Postigo JA, Franco JR, Jannin JG (2011) The human African trypanosomiasis control and surveillance programme of the World Health Organization 2000-2009: the way forward. PLoS Negl Trop Dis 5: e1007.
12. Schmunis GA, Yadon ZE (2010) Chagas disease: a Latin American health problem becoming a world health problem. Acta Trop 115: 14–21.
13. Hotez PJ, Dumonteil E, Betancourt Cravioto M, Bottazzi ME, Tapia-Conyer R, et al. (2013) An Unfolding Tragedy of Chagas Disease in North America. PLoS Negl Trop Dis 7: e2300.
14. Kristjanson PM, Swallow BM, Rowlands GJ, Kruska RL, de Leeuw PN (1999) Measuring the costs of African animal trypanosomosis, the potential benefits of control and returns to research. Agr Syst 59: 79–98.
15. Goodhead I, Archibald A, Amwayi P, Brass A, Gibson J, et al. (2010) A comprehensive genetic analysis of candidate genes regulating response to *Trypanosoma congolense* infection in mice. PLoS Negl Trop Dis 4: e880.
16. Vasconcelos RH, Montenegro SM, Azevedo EA, Gomes YM, Morais CN (2012) Genetic susceptibility to chronic Chagas disease: an overview of single nucleotide polymorphisms of cytokine genes. Cytokine 59: 203–208.
17. Goodhead I, Capewell P, Bailey JW, Beament T, Chance M, et al. (2013) Whole-genome sequencing of *Trypanosoma brucei* reveals introgression between subspecies that is associated with virulence. mBio 4: e00197-00113-e00197-00113.
18. da Matta Guedes PM, Gutierrez FR, Maia FL, Milanezi CM, Silva GK, et al. (2010) IL-17 produced during *Trypanosoma cruzi* infection plays a central role in regulating parasite-induced myocarditis. PLoS Negl Trop Dis 4: e604.
19. de Araujo FF, da Silveira AB, Correa-Oliveira R, Chaves AT, Adad SJ, et al. (2011) Characterization of the presence of Foxp3(+) T cells from patients with different clinical forms of Chagas' disease. Hum Pathol 42: 299–301.
20. Vespa GN, Cunha FQ, Silva JS (1994) Nitric oxide is involved in control of *Trypanosoma cruzi*-induced parasitemia and directly kills the parasite in vitro. Infect Immun 62: 5177–5182.
21. Miyazaki Y, Hamano S, Wang S, Shimanoe Y, Iwakura Y, et al. (2010) IL-17 is necessary for host protection against acute-phase *Trypanosoma cruzi* infection. J Immunol 185: 1150–1157.
22. Kotner J, Tarleton R (2007) Endogenous CD4(+) CD25(+) regulatory T cells have a limited role in the control of *Trypanosoma cruzi* infection in mice. Infect Immun 75: 861–869.
23. Cummings KL, Tarleton RL (2004) Inducible nitric oxide synthase is not essential for control of *Trypanosoma cruzi* infection in mice. Infect Immun 72: 4081–4089.
24. Mukherjee S, Belbin TJ, Spray DC, Iacobas DA, Weiss LM, et al. (2003) Microarray analysis of changes in gene expression in a murine model of chronic chagasic cardiomyopathy. Parasitol Res 91: 187–196.
25. Mukherjee S, Nagajyothi F, Mukhopadhyay A, Machado FS, Belbin TJ, et al. (2008) Alterations in myocardial gene expression associated with experimental *Trypanosoma cruzi* infection. Genomics 91: 423–432.
26. Garg N, Gerstner A, Bhatia V, DeFord J, Papaconstantinou J (2004) Gene expression analysis in mitochondria from chagasic mice: alterations in specific metabolic pathways. Biochem J 381: 743–752.
27. Hashimoto M, Nakajima-Shimada J, Ishidoh K, Aoki T (2005) Gene expression profiles in response to Fas stimulation in *Trypanosoma cruzi*-infected host cells. Int J Parasitol 35: 1587–1594.
28. Goldenberg RC, Iacobas DA, Iacobas S, Rocha LL, da Silva de Azevedo Fortes F, et al. (2009) Transcriptomic alterations in *Trypanosoma cruzi*-infected cardiac myocytes. Microbes Infect 11: 1140–1149.
29. Chessler AD, Unnikrishnan M, Bei AK, Daily JP, Burleigh BA (2009) *Trypanosoma cruzi* triggers an early type I IFN response in vivo at the site of intradermal infection. J Immunol 182: 2288–2296.
30. Costales JA, Daily JP, Burleigh BA (2009) Cytokine-dependent and-independent gene expression changes and cell cycle block revealed in *Trypanosoma cruzi*-infected host cells by comparative mRNA profiling. BMC Genomics 10: 252.
31. Soares MB, de Lima RS, Rocha LL, Vasconcelos JF, Rogatto SR, et al. (2010) Gene expression changes associated with myocarditis and fibrosis in hearts of mice with chronic chagasic cardiomyopathy. J Infect Dis 202: 416–426.
32. Soares MB, Lima RS, Souza BS, Vasconcelos JF, Rocha LL, et al. (2011) Reversion of gene expression alterations in hearts of mice with chronic chagasic cardiomyopathy after transplantation of bone marrow cells. Cell Cycle 10: 1448–1455.
33. Manque PA, Probst CM, Pereira MC, Rampazzo RC, Ozaki LS, et al. (2011) *Trypanosoma cruzi* infection induces a global host cell response in cardiomyocytes. Infect Immun 79: 1855–1862.
34. Genovesio A, Giardini MA, Kwon YJ, de Macedo Dossin F, Choi SY, et al. (2011) Visual genome-wide RNAi screening to identify human host factors required for *Trypanosoma cruzi* infection. PLoS One 6: e19733.
35. Tanowitz HB, Mukhopadhyay A, Ashton AW, Lisanti MP, Machado FS, et al. (2011) Microarray analysis of the mammalian thromboxane receptor-*Trypanosoma cruzi* interaction. Cell Cycle 10: 1132–1143.
36. Kierstein S, Noyes H, Naessens J, Nakamura Y, Pritchard C, et al. (2006) Gene expression profiling in a mouse model for African trypanosomiasis. Genes Immun 7: 667–679.
37. Li SQ, Reid SA, Fung MC, Inoue N, Lun ZR (2009) Analysis of gene expression profiles in the liver and spleen of mice infected with *Trypanosoma evansi* by using a cDNA microarray. Parasitol Res 104: 385–397.
38. Noyes HA, Alimohammadian MH, Agaba M, Brass A, Fuchs H, et al. (2009) Mechanisms controlling anaemia in *Trypanosoma congolense* infected mice. PLoS One 4: e5170.
39. O'Gorman GM, Park SD, Hill EW, Meade KG, Coussens PM, et al. (2009) Transcriptional profiling of cattle infected with *Trypanosoma congolense* highlights gene expression signatures underlying trypanotolerance and trypanosusceptibility. BMC Genomics 10: 207.
40. Amin DN, Ngoyi DM, Nhkwachi GM, Palomba M, Rottenberg M, et al. (2010) Identification of stage biomarkers for human African trypanosomiasis. Am J Trop Med Hyg 82: 983–990.
41. Li SQ, Luckins A, Lun ZR (2011) *Trypanosoma brucei brucei*: A comparison of gene expression in the liver and spleen of infected mice utilizing cDNA microarray technology. Exp Parasitol 128: 256–264.
42. Mekata H, Konnai S, Mingala CN, Abes NS, Gutierrez CA, et al. (2012) Kinetics of regulatory dendritic cells in inflammatory responses during *Trypanosoma evansi* infection. Parasite Immunol 34: 318–329.
43. Meade KG, O'Gorman GM, Hill EW, Narciandi F, Agaba M, et al. (2009) Divergent antimicrobial peptide (AMP) and acute phase protein (APP) responses to *Trypanosoma congolense* infection in trypanotolerant and trypanosusceptible cattle. Mol Immunol 47: 196–204.
44. Hill EW, O'Gorman GM, Agaba M, Gibson JP, Hanotte O, et al. (2005) Understanding bovine trypanosomiasis and trypanotolerance: the promise of functional genomics. Vet Immunol Immunopathol 105: 247–258.
45. Lopez R, Demick KP, Mansfield JM, Paulnock DM (2008) Type I IFNs play a role in early resistance, but subsequent susceptibility, to the African trypanosomes. J Immunol 181: 4908–4917.
46. Graefe SE, Streichert T, Budde BS, Nurnberg P, Steeg C, et al. (2006) Genes from Chagas susceptibility loci that are differentially expressed in *T. cruzi*-resistant mice are candidates accounting for impaired immunity. PLoS One 1: e57.
47. Park AY, Hondowicz BD, Scott P (2000) IL-12 is required to maintain a Th1 response during *Leishmania major* infection. J Immunol 165: 896–902.
48. Kinjyo I, Inoue H, Hamano S, Fukuyama S, Yoshimura T, et al. (2006) Loss of SOCS3 in T helper cells resulted in reduced immune responses and hyperproduction of interleukin 10 and transforming growth factor-beta 1. J Exp Med 203: 1021–1031.
49. Guerfali FZ, Laouini D, Guizani-Tabbane L, Ottones F, Ben-Aissa K, et al. (2008) Simultaneous gene expression profiling in human macrophages infected with *Leishmania major* parasites using SAGE. BMC Genomics 9: 238.
50. Ehrchen JM, Roebrock K, Foell D, Nippe N, von Stebut E, et al. (2010) Keratinocytes determine Th1 immunity during early experimental leishmaniasis. PLoS Pathog 6: e1000871.
51. Biswas A, Bhattacharya A, Kar S, Das PK (2011) Expression of IL-10-triggered STAT3-dependent IL-4Ralpha is required for induction of arginase 1 in visceral leishmaniasis. Eur J Immunol 41: 992–1003.
52. Bertholet S, Debrabant A, Afrin F, Caler E, Mendez S, et al. (2005) Antigen requirements for efficient priming of CD8+ T cells by Leishmania major-infected dendritic cells. Infect Immun 73: 6620–6628.

53. Brunner C, Sindrilaru A, Girkontaite I, Fischer KD, Sunderkotter C, et al. (2007) BOB.1/OBF.1 controls the balance of TH1 and TH2 immune responses. EMBO J 26: 3191–3202.

54. Filippi C, Hugues S, Cazareth J, Julia V, Glaichenhaus N, et al. (2003) CD4+ T cell polarization in mice is modulated by strain-specific major histocompatibility complex-independent differences within dendritic cells. J Exp Med 198: 201–209.

55. Jayakumar A, Widenmaier R, Ma X, McDowell MA (2008) Transcriptional inhibition of interleukin-12 promoter activity in Leishmania spp.-infected macrophages. J Parasitol 94: 84–93.

56. Vivarini Ade C, Pereira Rde M, Teixeira KL, Calegari-Silva TC, Bellio M, et al. (2011) Human cutaneous leishmaniasis: interferon-dependent expression of double-stranded RNA-dependent protein kinase (PKR) via TLR2. FASEB J 25: 4162–4173.

57. Gail M, Gross U, Bohne W (2001) Transcriptional profile of *Toxoplasma gondii*-infected human fibroblasts as revealed by gene-array hybridization. Mol Genet Genomics 265: 905–912.

58. Knight BC, Kissane S, Falciani F, Salmon M, Stanford MR, et al. (2006) Expression analysis of immune response genes of Muller cells infected with *Toxoplasma gondii*. J Neuroimmunol 179: 126–131.

59. Ju CH, Chockalingam A, Leifer CA (2009) Early response of mucosal epithelial cells during *Toxoplasma gondii* infection. J Immunol 183: 7420–7427.

60. Okomo-Adhiambo M, Beattie C, Rink A (2006) cDNA microarray analysis of host-pathogen interactions in a porcine in vitro model for *Toxoplasma gondii* infection. Infect Immun 74: 4254–4265.

61. Zhou DH, Yuan ZG, Zhao FR, Li HL, Zhou Y, et al. (2011) Modulation of mouse macrophage proteome induced by *Toxoplasma gondii* tachyzoites in vivo. Parasitol Res 109: 1637–1646.

62. Watford WT, Hissong BD, Durant LR, Yamane H, Muul LM, et al. (2008) Tpl2 kinase regulates T cell interferon-gamma production and host resistance to Toxoplasma gondii. J Exp Med 205: 2803–2812.

63. Tato CM, Villarino A, Caamano JH, Boothby M, Hunter CA (2003) Inhibition of NF-kappa B activity in T and NK cells results in defective effector cell expansion and production of IFN-gamma required for resistance to Toxoplasma gondii. J Immunol 170: 3139–3146.

64. Fux B, Rodrigues CV, Portela RW, Silva NM, Su C, et al. (2003) Role of cytokines and major histocompatibility complex restriction in mouse resistance to infection with a natural recombinant strain (type I-III) of Toxoplasma gondii. Infect Immun 71: 6392–6401.

65. Fang R, Nie H, Wang Z, Tu P, Zhou D, et al. (2009) Protective immune response in BALB/c mice induced by a suicidal DNA vaccine of the MIC3 gene of Toxoplasma gondii. Vet Parasitol 164: 134–140.

66. Desolme B, Mevelec MN, Buzoni-Gatel D, Bout D (2000) Induction of protective immunity against toxoplasmosis in mice by DNA immunization with a plasmid encoding Toxoplasma gondii GRA4 gene. Vaccine 18: 2512–2521.

67. Ylostalo J, Randall AC, Myers TA, Metzger M, Krogstad DJ, et al. (2005) Transcriptome profiles of host gene expression in a monkey model of human malaria. J Infect Dis 191: 400–409.

68. Carapau D, Kruhofer M, Chatalbash A, Orengo JM, Mota MM, et al. (2007) Transcriptome profile of dendritic cells during malaria: cAMP regulation of IL-6. Cell Microbiol 9: 1738–1752.

69. Miu J, Hunt NH, Ball HJ (2008) Predominance of interferon-related responses in the brain during murine malaria, as identified by microarray analysis. Infect Immun 76: 1812–1824.

70. Albuquerque SS, Carret C, Grosso AR, Tarun AS, Peng X, et al. (2009) Host cell transcriptional profiling during malaria liver stage infection reveals a coordinated and sequential set of biological events. BMC Genomics 10: 270.

71. Delic D, Dkhil M, Al-Quraishy S, Wunderlich F (2011) Hepatic miRNA expression reprogrammed by *Plasmodium chabaudi* malaria. Parasitol Res 108: 1111–1121.

72. Rosanas-Urgell A, Martin-Jaular L, Ricarte-Filho J, Ferrer M, Kalko S, et al. (2012) Expression of non-TLR pattern recognition receptors in the spleen of BALB/c mice infected with Plasmodium yoelii and Plasmodium chabaudi chabaudi AS. Mem Inst Oswaldo Cruz 107: 410–415.

73. Randall LM, Amante FH, McSweeney KA, Zhou Y, Stanley AC, et al. (2008) Common strategies to prevent and modulate experimental cerebral malaria in mouse strains with different susceptibilities. Infect Immun 76: 3312–3320.

74. Oakley MS, McCutchan TF, Anantharaman V, Ward JM, Faucette L, et al. (2008) Host biomarkers and biological pathways that are associated with the expression of experimental cerebral malaria in mice. Infect Immun 76: 4518–4529.

75. Lovegrove FE, Pena-Castillo L, Mohammad N, Liles WC, Hughes TR, et al. (2006) Simultaneous host and parasite expression profiling identifies tissue-specific transcriptional programs associated with susceptibility or resistance to experimental cerebral malaria. BMC Genomics 7: 295.

76. Delahaye NF, Coltel N, Puthier D, Barbier M, Benech P, et al. (2007) Gene expression analysis reveals early changes in several molecular pathways in cerebral malaria-susceptible mice versus cerebral malaria-resistant mice. BMC Genomics 8: 452.

77. Betts J, deSchoolmeester ML, Else KJ (2000) *Trichuris muris*: CD4+ T cell-mediated protection in reconstituted SCID mice. Parasitology 121 Pt 6: 631–637.

78. Bickle Q, Helmby H (2007) Lack of galectin-3 involvement in murine intestinal nematode and schistosome infection. Parasite Immunol 29: 93–100.

79. Cliffe LJ, Humphreys NE, Lane TE, Potten CS, Booth C, et al. (2005) Accelerated intestinal epithelial cell turnover: a new mechanism of parasite expulsion. Science 308: 1463–1465.

80. Humphreys NE, Worthington JJ, Little MC, Rice EJ, Grencis RK (2004) The role of CD8+ cells in the establishment and maintenance of a *Trichuris muris* infection. Parasite Immunol 26: 187–196.

81. Villarino AV, Artis D, Bezbradica JS, Miller O, Saris CJ, et al. (2008) IL-27R deficiency delays the onset of colitis and protects from helminth-induced pathology in a model of chronic IBD. Int Immunol 20: 739–752.

82. Massacand JC, Stettler RC, Meier R, Humphreys NE, Grencis RK, et al. (2009) Helminth products bypass the need for TSLP in Th2 immune responses by directly modulating dendritic cell function. Proc Natl Acad Sci U S A 106: 13968–13973.

83. Dixon H, Blanchard C, Deschoolmeester ML, Yuill NC, Christie JW, et al. (2006) The role of Th2 cytokines, chemokines and parasite products in eosinophil recruitment to the gastrointestinal mucosa during helminth infection. Eur J Immunol 36: 1753–1763.

84. Hepworth MR, Hardman MJ, Grencis RK (2010) The role of sex hormones in the development of Th2 immunity in a gender-biased model of Trichuris muris infection. Eur J Immunol 40: 406–416.

85. Hasnain SZ, Thornton DJ, Grencis RK (2011) Changes in the mucosal barrier during acute and chronic *Trichuris muris* infection. Parasite Immunol 33: 45–55.

86. Svensson M, Russell K, Mack M, Else KJ (2010) CD4+ T-cell localization to the large intestinal mucosa during *Trichuris muris* infection is mediated by G alpha i-coupled receptors but is CCR6- and CXCR3-independent. Immunology 129: 257–267.

87. Burke ML, McGarvey L, McSorley HJ, Bielefeldt-Ohmann H, McManus DP, et al. (2011) Migrating Schistosoma japonicum schistosomula induce an innate immune response and wound healing in the murine lung. Mol Immunol 49: 191–200.

88. Zhang M, Gao Y, Du X, Zhang D, Ji M, et al. (2011) Toll-like receptor (TLR) 2 and TLR4 deficiencies exert differential in vivo effects against Schistosoma japonicum. Parasite Immunol 33: 199–209.

89. Singh KP, Gerard HC, Hudson AP, Boros DL (2006) Differential expression of collagen, MMP, TIMP and fibrogenic-cytokine genes in the granulomatous colon of Schistosoma mansoni-infected mice. Ann Trop Med Parasitol 100: 611–620.

90. Perry CR, Burke ML, Stenzel DJ, McManus DP, Ramm GA, et al. (2011) Differential expression of chemokine and matrix re-modelling genes is associated with contrasting schistosome-induced hepatopathology in murine models. PLoS Negl Trop Dis 5: e1178.

91. de Oliveira Fraga LA, Torrero MN, Tocheva AS, Mitre E, Davies SJ (2010) Induction of type 2 responses by schistosome worms during prepatent infection. J Infect Dis 201: 464–472.

92. de la Torre-Escudero E, Valero L, Perez-Sanchez R, Manzano-Roman R, Oleaga A (2012) Proteomic identification of endothelial cell surface proteins isolated from the hepatic portal vein of mice infected with Schistosoma bovis. J Proteomics 77: 129–143.

93. Bystrom J, Dyer KD, Ting-De Ravin SS, Naumann N, Stephany DA, et al. (2006) Interleukin-5 does not influence differential transcription of transmembrane and soluble isoforms of IL-5R alpha in vivo. Eur J Haematol 77: 181–190.

94. Burke ML, McManus DP, Ramm GA, Duke M, Li Y, et al. (2010) Co-ordinated gene expression in the liver and spleen during Schistosoma japonicum infection regulates cell migration. PLoS Negl Trop Dis 4: e686.

95. Angyalosi G, Neveu R, Wolowczuk I, Delanoye A, Herno J, et al. (2001) HLA class II polymorphism influences onset and severity of pathology in Schistosoma mansoni-infected transgenic mice. Infect Immun 69: 5874–5882.

96. Ray D, Nelson TA, Fu CL, Patel S, Gong DN, et al. (2012) Transcriptional profiling of the bladder in urogenital schistosomiasis reveals pathways of inflammatory fibrosis and urothelial compromise. PLoS Negl Trop Dis 6: e1912.

97. Xu Y, Xie J, Li Y, Yue J, Chen J, et al. (2003) Using a cDNA microarray to study cellular gene expression altered by Mycobacterium tuberculosis. Chin Med J (Engl) 116: 1070–1073.

98. Volpe E, Cappelli G, Grassi M, Martino A, Serafino A, et al. (2006) Gene expression profiling of human macrophages at late time of infection with Mycobacterium tuberculosis. Immunology 118: 449–460.

99. Silver RF, Walrath J, Lee H, Jacobson BA, Horton H, et al. (2009) Human alveolar macrophage gene responses to Mycobacterium tuberculosis strains H37Ra and H37Rv. Am J Respir Cell Mol Biol 40: 491–504.

100. Sharbati J, Lewin A, Kutz-Lohroff B, Kamal E, Einspanier R, et al. (2011) Integrated microRNA-mRNA-analysis of human monocyte derived macrophages upon Mycobacterium avium subsp. hominissuis infection. PLoS One 6: e20258.

101. Ragno S, Romano M, Howell S, Pappin DJ, Jenner PJ, et al. (2001) Changes in gene expression in macrophages infected with Mycobacterium tuberculosis: a combined transcriptomic and proteomic approach. Immunology 104: 99–108.

102. Orlova MO, Majorov KB, Lyadova IV, Eruslanov EB, M'Lan C E, et al. (2006) Constitutive differences in gene expression profiles parallel genetic patterns of susceptibility to tuberculosis in mice. Infect Immun 74: 3668–3672.

103. Magee DA, Taraktsoglou M, Killick KE, Nalpas NC, Browne JA, et al. (2012) Global gene expression and systems biology analysis of bovine monocyte-

derived macrophages in response to in vitro challenge with Mycobacterium bovis. PLoS One 7: e32034.

104. Maddocks S, Scandurra GM, Nourse C, Bye C, Williams RB, et al. (2009) Gene expression in HIV-1/Mycobacterium tuberculosis co-infected macrophages is dominated by M. tuberculosis. Tuberculosis (Edinb) 89: 285–293.
105. Keller C, Lauber J, Blumenthal A, Buer J, Ehlers S (2004) Resistance and susceptibility to tuberculosis analysed at the transcriptome level: lessons from mouse macrophages. Tuberculosis (Edinb) 84: 144–158.
106. Beisiegel M, Mollenkopf HJ, Hahnke K, Koch M, Dietrich I, et al. (2009) Combination of host susceptibility and Mycobacterium tuberculosis virulence define gene expression profile in the host. Eur J Immunol 39: 3369–3384.
107. Sugimoto CR, Zhang G, Cronin B (2013) Bias in peer review. J Am Soc Inform Sci Tech 64: 2–17.
108. Zingales B, Miles MA, Campbell DA, Tibayrenc M, Macedo AM, et al. (2012) The revised *Trypanosoma cruzi* subspecific nomenclature: rationale, epidemiological relevance and research applications. Infect Genet Evol 12: 240–253.
109. Majiwa PA, Hamers R, Van Meirvenne N, Matthyssens G (1986) Evidence for genetic diversity in *Trypanosoma* (Nannomonas) *congolense*. Parasitology 93 (Pt 2): 291–304.
110. Barrett MP, Burchmore RJ, Stich A, Lazzari JO, Frasch AC, et al. (2003) The trypanosomiases. Lancet 362: 1469–1480.
111. Capewell P, Clucas C, DeJesus E, Kieft R, Hajduk S, et al. (2013) The TgsGP gene is essential for resistance to human serum in *Trypanosoma brucei gambiense*. PLoS Pathog 9: e1003686.
112. Miles MA, Cedillos RA, Povoa MM, de Souza AA, Prata A, et al. (1981) Do radically dissimilar *Trypanosoma cruzi* strains (zymodemes) cause Venezuelan and Brazilian forms of Chagas' disease? Lancet 1: 1338–1340.
113. Saeij JP, Boyle JP, Boothroyd JC (2005) Differences among the three major strains of *Toxoplasma gondii* and their specific interactions with the infected host. Trends Parasitol 21: 476–481.
114. Johnston CE, Bradley JE, Behnke JM, Matthews KR, Else KJ (2005) Isolates of *Trichuris muris* elicit different adaptive immune responses in their murine host. Parasite Immunol 27: 69–78.
115. Zuim NR, Allegretti SM, Linhares AX, Magalhaes LA, Zanotti-Magalhaes EM (2012) A Study of the Granulomatous Responses Induced by Different Strains of Schistosoma mansoni. Interdiscip Perspect Infect Dis 2012: 953524.
116. Guilbride DL, Guilbride PD, Gawlinski P (2012) Malaria's deadly secret: a skin stage. Trends Parasitol 28: 142–150.
117. De Lima AR, Navarro MC, Arteaga RY, Contreras VT (2008) Cultivation of *Trypanosoma cruzi* epimastigotes in low glucose axenic media shifts its competence to differentiate at metacyclic trypomastigotes. Exp Parasitol 119: 336–342.
118. Perez Brandan C, Padilla AM, Diosque P, Basombrio MA (2006) *Trypanosoma cruzi*: infectivity modulation of a clone after passages through different hosts. Exp Parasitol 114: 89–93.
119. Shikanai-Yasuda MA, Carvalho NB (2012) Oral transmission of Chagas disease. Clin Infect Dis 54: 845–852.
120. Camandaroba EL, Pinheiro Lima CM, Andrade SG (2002) Oral transmission of Chagas disease: importance of *Trypanosoma cruzi* biodeme in the intragastric experimental infection. Rev Inst Med Trop Sao Paulo 44: 97–103.
121. Ramirez JD, Montilla M, Cucunuba ZM, Florez AC, Zambrano P, et al. (2013) Molecular epidemiology of human oral Chagas disease outbreaks in Colombia. PLoS Negl Trop Dis 7: e2041.
122. Johnson AM (1984) Strain-dependent, route of challenge-dependent, murine susceptibility to toxoplasmosis. Z Parasitenkd 70: 303–309.
123. Schuurs AH, Verheul HA (1990) Effects of gender and sex steroids on the immune response. J Steroid Biochem 35: 157–172.
124. Klein SL (2012) Immune cells have sex and so should journal articles. Endocrinology 153: 2544–2550.
125. Roberts CW, Cruickshank SM, Alexander J (1995) Sex-determined resistance to *Toxoplasma gondii* is associated with temporal differences in cytokine production. Infect Immun 63: 2549–2555.
126. Schuster JP, Schaub GA (2001) Experimental Chagas disease: the influence of sex and psychoneuroimmunological factors. Parasitol Res 87: 994–1000.
127. Turay AA, Nwobu GO, Okogun GR, Igwe CU, Adeyeye K, et al. (2005) A comparative study on the susceptibility of male and female albino mice to *Trypanosoma brucei brucei*. J Vector Borne Dis 42: 15–20.
128. Greenblatt HC, Rosenstreich DL (1984) *Trypanosoma rhodesiense* infection in mice: sex dependence of resistance. Infect Immun 43: 337–340.
129. Schuster JP, Schaub GA (2001) *Trypanosoma cruzi*: the development of estrus cycle and parasitemia in female mice maintained with or without male pheromones. Parasitol Res 87: 985–993.
130. Mutayoba BM, Gombe S, Kaaya GP, Waindi EN (1988) Trypanosome-induced ovarian dysfunction. Evidence of higher residual fertility in trypanotolerant small East African goats. Acta Trop 45: 225–237.
131. Giammanco S, Ernandes M, La Guardia M (1997) Effects of environmental lighting and tryptophan devoid diet on the rat vaginal cycle. Arch Physiol Biochem 105: 445–449.
132. Nature (2013) Reporting Checklist For Life Sciences Articles. In: checklist.pdf, editor: Nature Publishing Group.
133. Maguire E, Gonzalez-Beltran A, Rocca-Serra P, Sansone S BioSharing. http://biosharing.org/.
134. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, et al. (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 6: e1000100.
135. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
136. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.
137. StataCorp. Stata Statistical Software: Release 10.College Station, TX, USA: StataCorp LP.
138. GraphPad. GraphPad Prism version 4.0 for Windows. La Jolla California USA, www.graphpad.com.

Chapter Four

Spreadsheet-based tool for assessing the biomedical scientific reporting

The content of this chapter was submitted to the *Journal of Biomedical Informatics*:

Flórez-Vargas O, Bramhall M, Jin B, Pérez D, Cruickshank S, Embury S, Stevens R, Brass A. miniRECH: a spreadsheet-based tool for assessing the quality of manuscripts reporting in bio-experiments.

Consistent with our findings in the previous chapter, we found that the quality of method information reported in articles about experimental infection with *Trypanosoma spp* is a cause for concern and it has not shown improvement over time, despite there being evidence that most of these variables do influence the outcomes. Inadequate reporting of key aspects of experimental design, including both instruments and protocols could act as a barrier to translation by preventing replication or inclusion in meta-analysis.

Since publication in 2014, our checklist and the findings found through our methodological strategy have been having an important impact not only on the parasitology community (Gulin, Rocco, and Garcia-Bournissen 2015, Klein and Roberts 2015), but also on the general scientific community [*e.g.*, see (Koch et al. 2015, Paradis et al. 2015)]; including the domain of metadata quality within the computer science studies (Tahvildari 2015). However, our checklist has not been yet endorsed by journals, which it makes difficult to verify its adherence. Nonetheless, for the reporting checklists that have been endorsed by journals – particularly by those with high impact factor, the completeness and accuracy of reporting remains suboptimal (Baker et al. 2014, Witwer 2013). This raises the question of whether the checklists are being ignored.

The current structure of reporting checklists does not allow scientists to perform assessments of scientific reports in a reproducibility context, since background assumptions of referees are not properly controlled. This issue is due in part to the fact that experts may consider some checklist items to be more important than others (Whiting, Harbord, and Kleijnen 2005). Perhaps, this is part of the reason behind the suboptimal information completeness and accuracy related with checklists endorsed by journals (Baker et al. 2014, Witwer 2013). Therefore, there is a need to develop a framework to achieve a consensus for assessing the quality of reporting.

In this chapter we presented miniRECH – which stands for **mini**mal **RE**porting **CH**ecklist; a general spreadsheet-based tool for assessing quality of scientific reporting (both pre- and post-publication) by using checklists as templates. The miniRECH framework was developed to operate in Microsoft® Excel, since MS Excel is used widely by the biomedical community. Through our evaluation process, we presented evidence that miniRECH has an important impact on the decision-making process; by helping both experts and non-experts in verify their judgments, and non-experts in producing judgments that approximate to the ones given by experts.

A guide describing a step-by-step process for scoring a bio-experiment on trypanosomiasis is provided in the Appendix section.

miniRECH: a spreadsheet-based tool for assessing the quality of methods reporting in scientific manuscripts

Oscar Flórez-Vargas[1], Binling Jin[2], Michael Bramhall[1], Diego Pérez[3], Robert Stevens[1], Andy Brass[1], Sheena Cruickshank[4], Suzanne Embury[2*]

1 Bio-health Informatics Group, School of Computer Science, University of Manchester, Manchester, United Kingdom.
2 Information Management Group, School of Computer Science, University of Manchester, Manchester, United Kingdom.
3 Statistics and its Applications Group, School of Mathematics, University of Manchester, Manchester, United Kingdom.
4 Manchester Immunology Group, Faculty of Life Science, University of Manchester, Manchester, United Kingdom.
* suzanne.m.embury@manchester.ac.uk

Abstract

Despite the effort made by the scientific community in addressing the ongoing reproducibility crisis in science, which aims at developing guidelines to improve the quality of reporting – as one of the sources of irreproducibility, the completeness and accuracy of reporting remains suboptimal. In particular, omissions in reporting the technical nature of the experimental method reduces transparency and make it difficult to understand and verify the findings of a research investigation. In an attempt to help prevent incomplete method reporting from entering the literature, we developed miniRECH – an Excel spreadsheet tool that provides a consensus framework for assessing the methods reporting in a scientific manuscript via checklists. By using a 10 point Likert scale, nine evaluators were asked to place 10 published studies in rank order considering the possibility of being able to reproduce the findings documented in each paper. Three of the nine evaluators were scientists involved in research on the topic of the articles (experts), whereas the remaining six evaluators were considered as non-experts in the field. The ranking order was performed before and after using miniRECH with a domain-specific checklist. The evaluation process conducted in this study showed that miniRECH has an important impact on the decision-making process; by helping both experts and non-experts in verifying their judgments regarding the completeness and accuracy of the information suggested by a checklist, and non-experts in producing judgments that approximate to the ones given by experts. The design of miniRECH offers two main features: firstly, by using checklists as templates, it helps prevent overlooking key information. And secondly, by using a scoring system based on scientific community feedback, it establishes a baseline level of the quality of reporting. We propose that miniRECH be considered as a strategy for improvement of the reporting of scientific reports.

Funding: This research is supported by a "Francisco José de Caldas" scholarship from Colciencias to O. Flórez-Vargas and a co-funding scholarship from EPSRC and Epistem Ltd awarded for M. Bramhall awarded to A. Brass and S. Cruickshank. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

The assessment of scientific reports both pre- and post-publication is an integral part of the scientific process. For over 350 years, since its development by the Royal Society in 1665, manuscripts have been subjected to the peer-review process for publication [1], and this process is being currently used by almost all scientific journals. Peer reviewers are responsible for critically assessing the quality of manuscripts within scientific standards [2], *i.e.*, identifying methodological flaws and ensuring that the reporting of research work is as truthful and accurate as possible [3].

However, despite the use of peer-review, several studies have demonstrated serious defects in the way biological and medical research are reported [4-6], which should be captured by the peer-review process [7,8]. In particular, the technical nature of the scientific method plays an important role in understanding and verifying the results and conclusions of research [9]. The omissions in reporting methods – particularly in the field of the life sciences – have been targeted as one of the main causes of the ongoing reproducibility crisis [10,11]. This relationship between reporting and reproducibility is to some extent ironic, considering that the peer-review process was developed to increase the likelihood that other researchers could replicate reviewed findings. This issue of method reporting is particularly significant when one consider that increasingly the scientific community are conducting cross-disciplinary research. So, scientific reports may not be used by experts in particular fields who would be more likely to know the methods and be able to draw inferences about what methodological issues could affect the reproducibility of the study.

As a response to this issue, the minimum information standards community has developed and implemented checklists and guidelines as an attempt to improve the quality of scientific reporting in biosciences [12], *e.g.*, Minimum Information about a Genotyping Experiment – MIGEN [13] and Preferred Reporting Items for Systematic reviews and Meta-Analyses – PRISMA (Liberati et al., 2009). These kinds of checklists are maintained in initiatives such as the BioSharing catalogue [http://biosharing.org] and the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network [http://www.equator-network.org].

Checklists explicitly define the essential criteria that should be considered for a given task in a particular area [14,15]. In biomedical research, for instance, it has been claimed that checklists can help to ensure transparency and consistency in reporting data and metadata from bio-experiments, which enhances the comprehensiveness of the scientific evidence and the reproducibility of its findings [16]. In this way, mandatory use of checklists is increasingly a requirement when submitting an article to biomedical scientific journals. However, despite the relative improvements when the reporting checklists are endorsed by journals, the completeness and accuracy of reporting remains suboptimal [17,18]. For example, in *Nature* journals the incidence of reporting of animal characteristics that influence experimental outcomes such as sex and age increased only by twofold (~80%) two years after endorsement of the ARRIVE guidelines [17]. Moreover, when 127 articles that reported microarray experiments were examined, 93 (73%) of them were judged to be MIAME noncompliant despite being published in journals with stringent policies regarding use of the MIAME checklist such as PLoS One and *Blood* [18].

Why is there an underreporting in the completeness and accuracy of information regarding checklists? The answer to this question will involve all actors of the publishing process, who are responsible for maintaining the highest scientific standards of publication, ensuring that the work is reported correctly, *i.e.*, ethically controlling the integrity in the writing, editing and publication process. While the authors need to be more strongly encouraged to use reporting checklists during the preparation of the manuscripts [19], both the peer-reviewers and editors need to pay more attention to the benefits of using these kinds of checklists as powerful management tools to aid decision making [19-21].

Considering that the peer-review process is an important part of the quality control mechanism that is used to determine what is published, and what is not [22], one of the main problems in the decision-making processes via checklists is the degree of experience of the peer reviewers. Such experience may have a considerable influence on their judgment of a checklist's completeness and accuracy of information. This could be because some reviewers could infer information about something that is not expressly stated by looking at the characteristics of the thing being assessed (*e.g.*, the mouse weight regarding its age), and also because some of them may consider that there are some checklist items that are more important than others [23]. These factors, therefore, create a need to develop a framework to achieve a consensus for judging the completeness and accuracy of the information suggested by checklists. This framework should ease the implementation and use of checklists.

To develop the consensus framework, we hypothesized that a democratic weighting system for each item in a checklist may be useful in producing a score that approximates the rating given by a community of experts in a given field; a democratic system ensures that the majority opinion informs the weighting, rather than the opinion of any single expert. Our goal in this study is to present evidence that a checklist in a consensus framework could be used as a decision-making tool by helping referees to verify their judgments regarding scientific reports. This framework should be important for assessing the quality of reporting of scholarly manuscripts by peer-reviewers due to their different scientific knowledge and expertise. In addition, such a checklist framework could be used retrospectively to assess reproducibility and study validity by a wide range of reviewers.

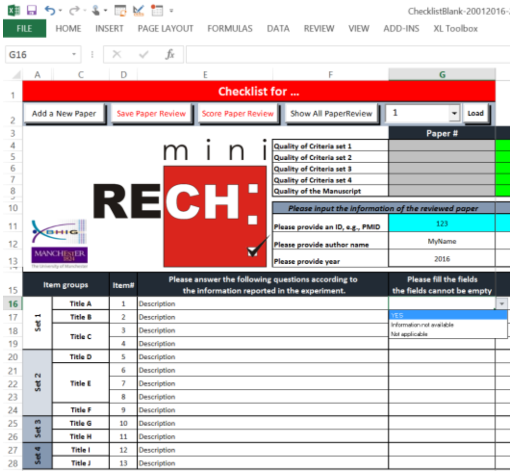
In order to demonstrate the usefulness of our consensus framework, we used as a model the checklist for animal models of colitis [24]. This is a domain-specific checklist with a structured questionnaire of 42 "*yes*", "*information not supplied*", or "*not applicable*" questions, so it is expected not only that it improves the verification process considerably, but also that it reduces the assumptions based on the referees expertise.

Methods

miniRECH spreadsheet

miniRECH, which stands for **mini**mal **RE**porting **CH**ecklist, is a general spreadsheet-based tool for assessing quality of scientific reporting (both pre- and post-publication) by using checklists as templates (Figure 1). The miniRECH model developed in this work was designed to operate in Microsoft® Excel since MS Excel is routinely used by the biomedical community. In a miniRECH template, any checklist is structured as a list of questions which should be answered "yes", "information not supplied", or "not applicable". A user's guide describing a step-by-step process for scoring a bio-experiment on trypanosomiasis is provided in the File S1.

To use a checklists based on miniRECH, stakeholders can either download an existing spreadsheet-checklist miniRECH formatted, or adapt a new checklist by modifying an already existing one, from the



Checklist for ...

	Paper #
Quality of Criteria set 1	
Quality of Criteria set 2	
Quality of Criteria set 3	
Quality of Criteria set 4	
Quality of the Manuscript	

Please input the information of the reviewed paper

Please provide an ID, e.g., PMID	123
Please provide author name	MyName
Please provide year	2016

Item groups		Item#	Please answer the following questions according to the information reported in the experiment.	Please fill the fields the fields cannot be empty
Set 1	Title A	1	Description	
	Title B	2	Description	YES Information not available Not applicable
	Title C	3	Description	
		4	Description	
Set 2	Title D	5	Description	
	Title E	6	Description	
		7	Description	
		8	Description	
	Title F	9	Description	
Set 3	Title G	10	Description	
	Title H	11	Description	
Set 4	Title I	12	Description	
	Title J	13	Description	

Figure 1. A miniRECH-enabled spreadsheet showing drop-down list selection.

Bio-Health Informatics group web site (http://www.cs.manchester.ac.uk/bhig/areas-and-projects/). Moreover, the MS Excel spreadsheet with its open-access Visual Basic macro scripts are made available as open source for download from GitHub (https://github.com/miniRECH).

Item weighting

We began developing the miniRECH weighting framework by asking a group of a minimum of three experts for feedback regarding existing checklists, via an individual online survey. Thus we have received feedback from biomedical scientists and epidemiologists regarding guidelines such as ARRIVE [25] and PRISMA [26], respectively. For example, in the case of the checklist for reporting bio-experiments using animal models of colitis the expert group included stakeholders from academia, research and industry [24]. The experts assigned a ranking for each criterion based on whether they perceived the item to be essential for repeating the experiment, and the importance of each criterion to the replicability and reproducibility of the experiment. Ranking was determined via a scale of one to three, with 1 being "highest importance" and 3 being "lowest importance". The majority vote was used to allocate weighting to criteria (a copy of this survey is available in the File S2).

Criteria were then assigned a weight via a combination of two factors: whether the item was considered essential (Y/N), and whether the item was determined to be of low, medium or high importance (L/M/H) for repeating, replicating and reproducing the findings. Weighted scores were conveniently allocated as follows: (Y=5 or N=2) and (L=3 or M=4 or H=5). At the end, by summing the two factors each criterion received a score between 5 and 10, which was then used to determine the weight. For example, an item considered non-essential (*i.e.* =2) and with low importance (*i.e.* =3) will score a 5 in our weighting system. In checklists that include animals as experimental models, *e.g.,* the ARRIVE guidelines for reporting animal research [25], the species and strain of the animal were two items that scored 10 since they were identified as essentials and highly important to contribute to the replicability and reproducibility of the findings.

Checklists scoring

In a miniRECH template, each item receives a weighted score as stated above if the criterion is present or not applicable, and zero if the criterion is absent. These scores are then used to assess the quality of scientific reporting as a percentage of the sum of all scores, both overall and by sections, *e.g.* introduction, methods, results, and discussion.

Evaluation

We hypothesize that the miniRECH framework is a facilitator to achieve a level of consensus among the judgments of referees on a topic of interest regarding the quality of reporting about such a topic.

To test this hypothesis, we used as a model the checklist for animal models of colitis [24]. This is a domain-specific checklist with a structured questionnaire of 42 "*yes*", "*information not supplied*", or "*not applicable*" questions, so it is expected not only that it increases the verification process considerably, but also that it reduces the assumptions based on the referees expertise. Considering the differences among referees regarding the scientific training in colitis, a total of two groups were created: a group of experts, who were three PhD scientists deeply involved in research on animal models of colitis, and a group of six non-experts, who have a biomedical background with different levels of experience (holding either a BSc., an MSc. or a PhD degree).

The evaluators were asked to place 10 studies on animal models of colitis in rank order, using a 10 point Likert scale from poor (one) to excellent (10) based on a holistic judgement of the quality of methods reporting of the studies, and considering the possibility of being able to replicate and reproduce the findings documented in the scientific paper (Figure 2). As soon as this task was done, the evaluators were asked to use the checklist for animal models of colitis to independently verify the completeness and accuracy of information reported in the 10 studies (Table 1). By using the scores obtained on the miniRECH (Table 2), we ranked the 10 studies and built a comparative table with the two ranks. Then, finally, the evaluators were asked again to place the 10 studies in rank order considering their prior judgments and the rank generated by using miniRECH (Figure 2).

The scientific articles included in this study were initially identified by our group in a study published recently [24]. Briefly, we conducted a systematic search following the recommendations of the PRISMA guidelines [26]. The literature search was conducted via PubMed in June 2014 using MeSH (Medical Subject Headings) terms and text strings. Those articles that conducted a microarray on colonic tissues were selected.

Table 1. Articles included for assessment in this study.

PMID	Author	Year	Journal
15973123	Abad *et al.*	2005	*Inflamm Bowel Dis*
16917233	Guzman *et al.*	2006	*Inflamm Bowel Dis*
17982090	Wu *et al.*	2007	*J Immunol*
19133689	Hansen *et al.*	2009	*Inflamm Bowel Dis*
19228061	Larrosa *et al.*	2009	*J Agric Food Chem*
19450596	Kiela *et al.*	2009	*Gastroenterology*
19560465	Zhou *et al.*	2009	*Gastroenterology*
20923862	Fang *et al.*	2011	*Physiol Genomics*
22865203	Reikvam *et al.*	2012	*Eur J Immunol*
23226271	Kremer *et al.*	2012	*PLoS One*

Table 2. Percentage total scores from the assessment via miniRECH for each study.

PMID	Evaluators								
	E1	E2	E3	NE1	NE2	NE3	NE4	NE5	NE6
15973123	72.97	70.27	70.27	72.38	62.16	63.51	67.56	67.56	67.56
16917233	72.7	70	67.29	72.11	67.29	61.89	62.7	70	21.97
17982090	78.37	77.56	72.97	77.74	66.75	64.86	72.97	72.16	79.18
19133689	75.67	67.56	70.27	75.06	75.4	76.75	72.43	74.86	75.4
19228061	92.7	83.51	90	92.76	90	90.81	84.86	85.4	79.45
19450596	80.27	52.16	70.54	79.62	64.32	74.32	72.43	71.35	64.59
19560465	82.97	67.29	67.56	82.3	67.56	69.72	82.43	79.45	74.86
20923862	67.29	67.29	64.59	66.75	61.89	68.64	67.56	59.45	70.81
22865203	94.86	89.45	86.48	93.56	90.81	85.67	94.86	91.35	88.1
23226271	89.19	82.43	83.78	90.08	81.08	85.94	86.48	87.83	78.64

* E = Expert and NE = Non-Expert

Statistical analysis

Due to the limited number of evaluators in each group (experts and non-experts), nonparametric methods were considered most appropriate. In this kind of situation, nonparametric methods show an advantage over their counterpart as they do not require the assumption of any parametric distribution in the data, *e.g.*, Normal distribution, allowing to achieve more robust conclusions. In all the analyses, we used ranked data rather than the raw data since it is more robust to variations in the extreme scores. Considering all the points above, a nonparametric Wilcoxon rank test was performed for the ranking pre- and post- miniRECH to determine the difference in the ranking between experts and non-experts. In the case of correlations between rankings, a Spearman's rank order correlation coefficient was used. A Bland-Altman analysis was used to assess the level of agreement between the experts and non-experts to compare their performance in ranking papers pre- and post- miniRECH; a range of agreement was defined as mean bias ±2 SD. In order to measure the correlation within the groups, an intraclass correlation coefficient was computed. The intraclass correlation takes into account the total variance which may be decomposed into two different sources of variability, the expert and the non-expert groups. As the groups differ in size, the mean of the variances for each source of variability was considered. Finally the intraclass correlation for each group was computed as the ratio between the mean of the variances of each group and the sum of the variances of the total sources of variability. All the statistical analysis was carried out by using the package MASS from the R language [27], and the minimum level of significance was defined at $p< 0.05$.

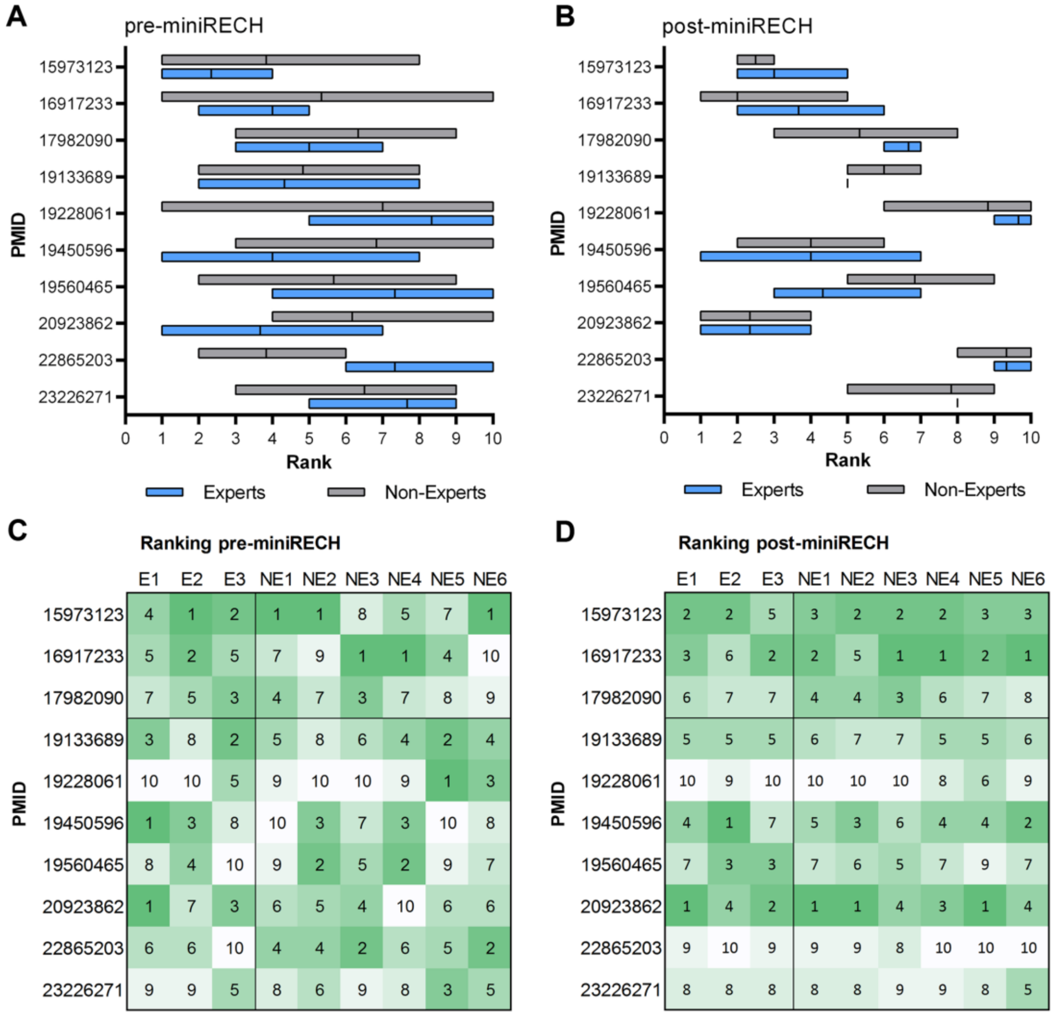
Results

Ranking the quality of studies pre- and post- miniRECH

The holistic ranking shows a low level of consistency between experts and non-experts (Figure 2). There are papers, for instance, that were included in all the 10 ranks by the non-experts, such as PMID 16917233 and 19228061, whereas the experts ranked these two papers within shorter ranges, *i.e.*, between three and five levels of ranking (Figure 2A). There is an evident improvement in the consistency of ranking after the evaluators used miniRECH: the experts were more consistent than the non-experts (Figure 2B). Nevertheless, no statistically significant difference between experts and non-experts based on a pre-miniRECH or post-miniRECH ranking was noted. This suggests that, despite the level of expertise of a group of referees about a particular topic, there is a wide disagreement among knowledgeable referees for assessing the quality of a set of scientific reports.

The miniRECH improves the level of consensus among the judgments of evaluators

In order to determine the global impact of the miniRECH approach on the consensus among the judgments of evaluators we performed Spearman correlations and Bland-Altman analyses.



Ranking pre-miniRECH

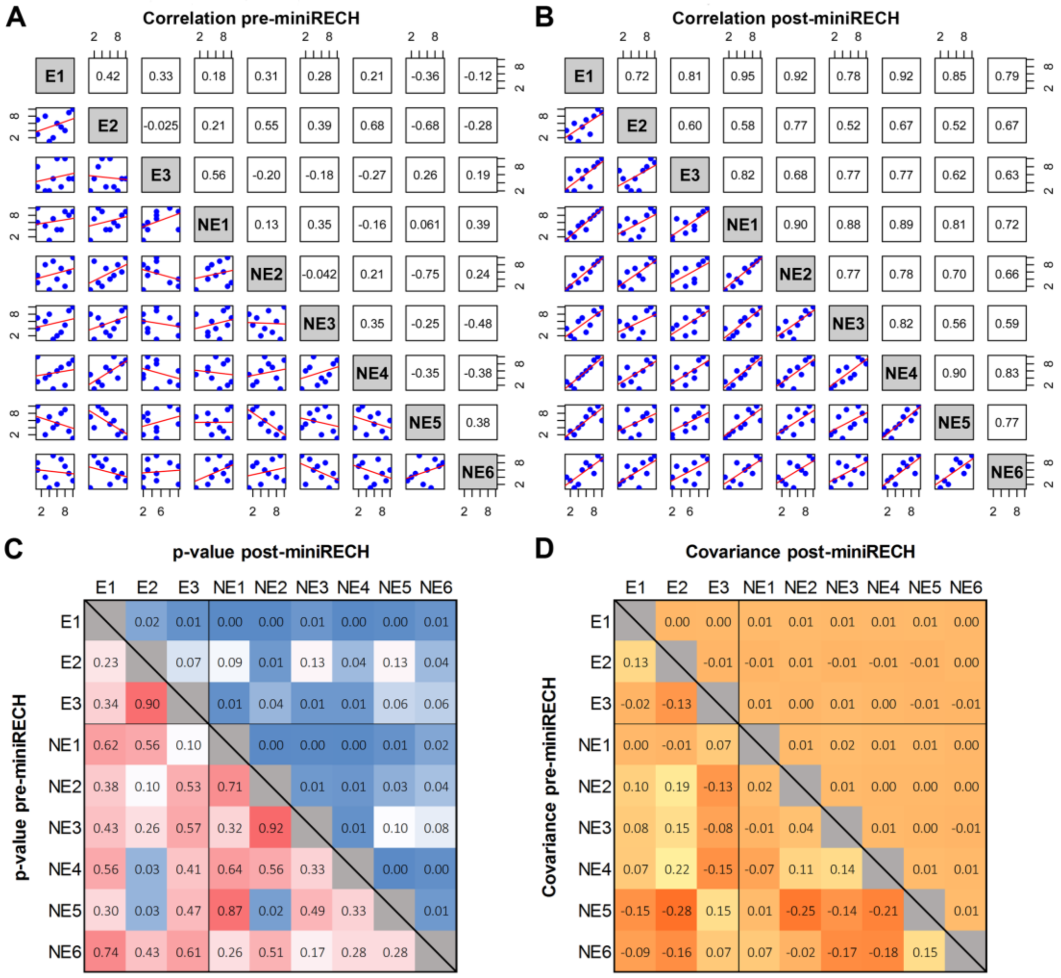
PMID	E1	E2	E3	NE1	NE2	NE3	NE4	NE5	NE6
15973123	4	1	2	1	1	8	5	7	1
16917233	5	2	5	7	9	1	1	4	10
17982090	7	5	3	4	7	3	7	8	9
19133689	3	8	2	5	8	6	4	2	4
19228061	10	10	5	9	10	10	9	1	3
19450596	1	3	8	10	3	7	3	10	8
19560465	8	4	10	9	2	5	2	9	7
20923862	1	7	3	6	5	4	10	6	6
22865203	6	6	10	4	4	2	6	5	2
23226271	9	9	5	8	6	9	8	3	5

Ranking post-miniRECH

PMID	E1	E2	E3	NE1	NE2	NE3	NE4	NE5	NE6
15973123	2	2	5	3	2	2	2	3	3
16917233	3	6	2	2	5	1	1	2	1
17982090	6	7	7	4	4	3	6	7	8
19133689	5	5	5	6	7	7	5	5	6
19228061	10	9	10	10	10	10	8	6	9
19450596	4	1	7	5	3	6	4	4	2
19560465	7	3	3	7	6	5	7	9	7
20923862	1	4	2	1	1	4	3	1	4
22865203	9	10	9	9	9	8	10	10	10
23226271	8	8	8	8	8	9	9	8	5

Figure 2. Ranking the quality of studies pre- and post- miniRECH. The range of rankings and its mean derived from the expert and non-expert assessments are shown in bars over the graphs [A] and [B]. The corresponding Likert scale rankings raw data is presented in [C] and [D]: ranking changes from dark green (poor or 1) to white (excellent or 10). The figure shows an improvement in the consistency of ranking after the evaluators used miniRECH. However, no statistically significant difference between experts and non-experts based on a pre-miniRECH or post-miniRECH ranking was noted. E= Expert and NE= Non-Expert.

The Spearman rank correlations were used to determine the degree of relationship among the ranking given by the evaluators pre- and post- using miniRECH (Figure 3A and 3B). Results showed that there were more positive significant correlations among evaluators after than before using miniRECH (Figure 3C). Regarding the level of expertise on the topic, at least each non-expert showed a significant ranking correlation post-miniRECH with at least one out of the three experts (Figure 3C). In addition, the variation of correlation coefficients among evaluators was lower after (range from -0.012 to 0.016) than before (range from -0.28 to 0.22) using miniRECH (Figure 3D). The higher variation pre-miniRECH indeed was observed when the correlations between experts and non-experts were compared (Figure 3D).



A — Correlation pre-miniRECH

	E1	E2	E3	NE1	NE2	NE3	NE4	NE5	NE6
E1	E1	0.42	0.33	0.18	0.31	0.28	0.21	-0.36	-0.12
E2		E2	-0.025	0.21	0.55	0.39	0.68	-0.68	-0.28
E3			E3	0.56	-0.20	-0.18	-0.27	0.26	0.19
NE1				NE1	0.13	0.35	-0.16	0.061	0.39
NE2					NE2	-0.042	0.21	-0.75	0.24
NE3						NE3	0.35	-0.25	-0.48
NE4							NE4	-0.35	-0.38
NE5								NE5	0.38
NE6									NE6

B — Correlation post-miniRECH

	E1	E2	E3	NE1	NE2	NE3	NE4	NE5	NE6
E1	E1	0.72	0.81	0.95	0.92	0.78	0.92	0.85	0.79
E2		E2	0.60	0.58	0.77	0.52	0.67	0.52	0.67
E3			E3	0.82	0.68	0.77	0.77	0.62	0.63
NE1				NE1	0.90	0.88	0.89	0.81	0.72
NE2					NE2	0.77	0.78	0.70	0.66
NE3						NE3	0.82	0.56	0.59
NE4							NE4	0.90	0.83
NE5								NE5	0.77
NE6									NE6

C — p-value post-miniRECH (upper) / p-value pre-miniRECH (lower)

	E1	E2	E3	NE1	NE2	NE3	NE4	NE5	NE6
E1		0.02	0.01	0.00	0.00	0.01	0.00	0.00	0.01
E2	0.23		0.07	0.09	0.01	0.13	0.04	0.13	0.04
E3	0.34	0.90		0.01	0.04	0.01	0.01	0.06	0.06
NE1	0.62	0.56	0.10		0.00	0.00	0.00	0.01	0.02
NE2	0.38	0.10	0.53	0.71		0.01	0.01	0.03	0.04
NE3	0.43	0.26	0.57	0.32	0.92		0.01	0.10	0.08
NE4	0.56	0.03	0.41	0.64	0.56	0.33		0.00	0.00
NE5	0.30	0.03	0.47	0.87	0.02	0.49	0.33		0.01
NE6	0.74	0.43	0.61	0.26	0.51	0.17	0.28	0.28	

D — Covariance post-miniRECH (upper) / Covariance pre-miniRECH (lower)

	E1	E2	E3	NE1	NE2	NE3	NE4	NE5	NE6
E1		0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.00
E2	0.13		-0.01	-0.01	0.01	-0.01	-0.01	-0.01	0.00
E3	-0.02	-0.13		0.01	0.00	0.01	0.00	-0.01	-0.01
NE1	0.00	-0.01	0.07		0.01	0.02	0.01	0.01	0.00
NE2	0.10	0.19	-0.13	0.02		0.01	0.00	0.00	0.00
NE3	0.08	0.15	-0.08	-0.01	0.04		0.01	0.00	-0.01
NE4	0.07	0.22	-0.15	-0.07	0.11	0.14		0.01	0.01
NE5	-0.15	-0.28	0.15	0.01	-0.25	-0.14	-0.21		0.01
NE6	-0.09	-0.16	0.07	0.07	-0.02	-0.17	-0.18	0.15	

Figure 3. Correlations among rankings based on evaluators' assessment pre- and post- miniRECH. The scatter-plot matrices show the increasing positive correlation among evaluators regarding the assessment of the quality of reporting by using the miniRECH approach [A and B]. Spearman's rank order correlation coefficients and the data dispersion around the trend line are shown respectively above and below the diagonal of each scatter-plot matrix. In [C] and [D] are presented the associated significance levels and the variation of correlation coefficients for the scatter-plot matrices: pre-miniRECH (lower triangle) and post-miniRECH (upper triangle). Considering, for instance, the expert 1 (E1) with the non-expert 1 (NE1), the scatter plots show that the rank correlation between these two evaluators is weak and non-significant pre-miniRECH (r= 0.18, p= 0.62; cov= 0.00), but it is statistically significant post-miniRECH (r= 0.95, p= 0.00; cov=0.01). The p values are gradient-shaded from blue (p < 0.01) to red (p > 0.1) to indicate high to low significance [C]. The changes in the covariance matrix are displayed in a colour gradient from dark-orange (lowest variation) to light-yellow (highest variation) [D]. Overall, the figure shows that there is more uniformity than variation among evaluators after they have used miniRECH than before they used it, as evidenced by the positive significant correlations. E= Expert and NE= Non-Expert.

By using the average of the rankings given by the experts and non-experts for each paper, we found that the group of experts, but not the group of non-experts, showed a statistically significant correlation (p= 0.0005) between the ranking order given before and after using miniRECH (Figure 4A). Result suggests that the level of expertise of a group of evaluators with respect to a particular topic has an impact on the decision-making process. In addition, the Bland-Altman analysis of this data indicated a good concordance between experts and non-experts groups when miniRECH was used (Figure 4B and 4C). This is because the 95% limits of agreement between the two evaluators groups were narrow for post-miniRECH (ranged from -2.226 to 2.426) and relatively wide for pre-miniRECH (ranged from -4.259 to 3.792). Moreover, points in the post-miniRECH assessment are mainly clustered around the line of no difference, whereas in the pre-miniRECH assessment these points do appear to be of slightly greater variability in extremes (Figure 4B and 4C).

Finally, and in order to determine the individual impact of miniRECH on the expert assessments towards a peer-review consensus, we performed a pairwise comparison using the Wilcoxon's test between pre- and post-miniRECH rankings by comparing the rank correlations for each expert regarding the rank correlations from the non-expert group. The results of this analysis showed that two out of the three experts had a level of consensus significantly superior by using the miniRECH approach (Z= 0, p= 0.0049). Additionally, the intraclass correlation was calculated separately for experts and non-experts, for pre-miniRECH a positive intraclass correlation was found (r=0.514) for experts. Also a positive intraclass correlation was found (r = 0.485) for non-experts. For post-miniRECH a positive intraclass correlation for expert was found (r =0.490) and for non-expert was (r = 0.509), but no statistically significant differences were detected between pre- and post-miniRECH when their intraclass correlations were compared.

Discussion

In this article, we have presented miniRECH – a consensus framework for assessing the reporting of scientific studies via checklists. The hope is that by providing a tool that researchers can use for improving the description of the scientific process performed during any investigation, other researchers will be able to replicate and reproduce its findings, or compare and integrate them. In this context, a spreadsheet-based tool seems adequate for carrying out this task since it can handle both text and numbers [28]. Spreadsheets not only have been successfully used for describing experiments¸ *e.g.*, RightField [29], ISA Software [30] and MAGE-TAB [31], but also for developing decision support tools due to their ability to calculate information [32,33].

Several studies have reported on the rates of failure of peer reviewers to detect significant methodological errors in manuscripts [34,35]. On studies facing this issue, some have attempted to explore whether interventions can improve the peer review performance [7,8,36]. It has been observed, for instance, that written feedback provided by editors to peer reviewers did not improve the quality of subsequent reviews [36]. In addition, other studies showed that reviewers who underwent training (face-to-face and self-taught) led only to some improvement in performance on errors detection relative to those who had no training [8]; reviewers reported – on average – only 3 out of 9 major errors focused on methodological weaknesses, inaccurate reporting of data and unjustified conclusions [7]. In this scenario, a checklist could provide a means against failure by reminding reviewers the minimum necessary information to be explicitly reported on a report.

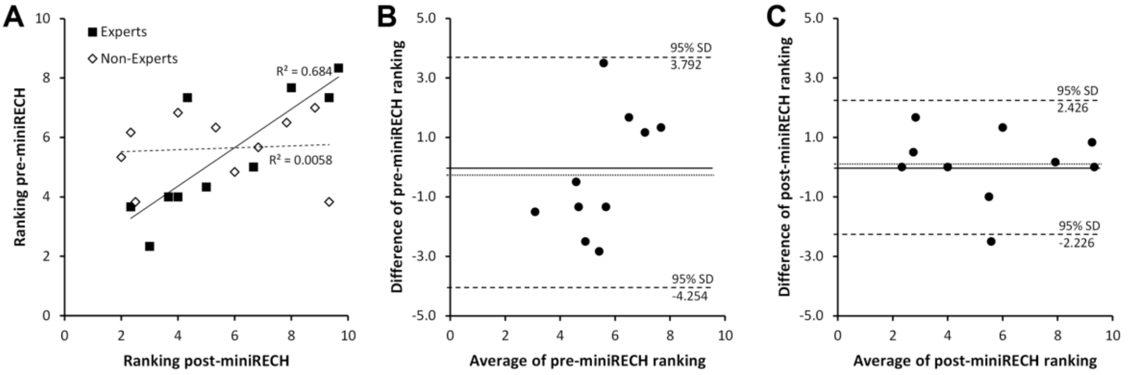


Figure 4. Correlation and agreement analysis between rankings from the experts and non-experts pre- and post-miniRECH. Spearman's rank correlation analysis showed a statistically significant correlation (r= 0.9147, p= 0.0005) between the ranking order from the experts group before and after using miniRECH, but not between the ranking order from the non-experts group (r= 0.1337, p= 0.7127) [A]. Bland-Altman plots revealed a better agreement between the ranking order from the experts and non-experts groups using miniRECH [B and C]. These results suggest that miniRECH is useful in producing scores that approximate the judgments given by experts.

Through our evaluation process, we have presented evidence that miniRECH has an important impact on the decision-making process (Figures 2, 3 and 4); by helping both experts and non-experts in verifying their judgments, and non-experts in producing judgments that approximate to the ones given by experts. This is particularly useful considering that the poor error detection rate observed in previous studies was not due to over-demanding expectation of reviewers, since the manuscripts assessed in these studies were general articles that apply to all areas of health care [7,8], *i.e.*, trials of medical records and communication activities. In contrast, while all reviewers that participated in our study have knowledge about the biomedical sciences, it is certainly true that the set of papers assessed were on a topic outside the field of expertise of one of the evaluator groups.

In addition, and considering the complexity of modern experiments – usually involving the combination of different technologies, the peer-review process requires multidisciplinary efforts for scrutinising their methods and findings. This, in turn, can lead to the possibility of judging a scientific paper on a biased background knowledge toward those things with which a reviewer is familiar, thus affecting the decision-making process, *e.g.*, recommending the acceptance of a manuscript. In fact, it was observed that the proportion of recommendations for rejection of manuscripts was higher for reviewers who found errors than for those who did not [7]. In this context, miniRECH attributes degrees of objectivity to evaluators based on the needs of a particular scientific area, which are explicitly stated in a reporting checklist.

> As stated by the non-expert participant 5 – a BSc research assistant, *"... regarding the initial ranking order, I created my own criteria list based on what I know about the techniques used in the paper, for instance, independent validation of gene expression levels. However, it did not include most of the details of the experimental model since I was not familiar with it."*

However, the current structure of reporting checklists does not allow scientists to perform assessments of scientific reports in a reproducibility context, since background assumptions of reviewers are not properly controlled.

> As non-expert participant 2 said – a former PhD researcher, *"... conducting the final ranking [post-miniRECH], I was expecting a considerable agreement with regard the initial ranking [pre-miniRECH], since I took into account all the information reported in each paper. However, it was not like that. So, I realised about the importance of this approach because it avoids bias for estimating the merits of a scientific report."*

This is perhaps part of the reason behind the suboptimal information completeness and accuracy related with checklists endorsed by journals [17,18]. By using a scoring system based on the opinions from members of a scientific community about each item included in a checklist, our miniRECH approach achieves a reasonable level of consensus between experts and non-experts in a specific domain. Therefore, this consensus framework can be used as a decision-making tool by helping reviewers get better at spotting missing and erroneous information about a particular experiment despite their different scientific knowledge and expertise regarding that particular experiment. In this way, miniRECH can help to ensure transparency in assessing the reporting of scientific studies toward enhancing the reproducibility of experimental findings, as well as improving the use and reuse of the reported information in such fields. Here is the comment from the non-expert participant 4 – an MSc researcher:

> *"... it [miniRECH] facilitates the objective evaluation of the scientific literature by considering the needs of a particular research community."*

Indeed, in the context of assessing the quality of published findings, and the methods by which they have been reached, the miniRECH tool could also be used for selecting scientific articles for systematic reviews and meta-analysis; as the study selection process is one of the main sources of bias due to lack of objectivity [26]. Therefore, by using miniRECH will be possible to include only publications with a particular quality score. In this regards, the miniRECH template enables all the articles to be assessed at a glance (see miniRECH guideline in the Supporting information). The usefulness of our approach for evaluating the quality of scientific reports was demonstrated by assessing published experiments on animal models of colitis [24].

Despite the overall positive performance of our approach, there are several limitations to consider. We restricted the study to colitis, which is a very specific domain, but it allowed us to compare the assessment performance between experts and non-experts in a particular topic. However, this approach should be able to be applied to other scientific domains, including those outside the biomedical arena and which are also facing the reproducibility issue, *e.g.*, computer science [37,38]. In addition, we cannot ensure that those papers ranked in the top of our Likert scale are reproducible; yet, they provide enough information for testing its validity. On the other hand, and considering that our primary aim was provide a 'proof of concept' approach for assessing the quality of methods reporting in scientific manuscripts, we have simulated the usual situation where at least three peer reviewers are appointed to assess the merits of a paper. Nevertheless, further validation to support the reliability of this method is needed, *e.g.*, by including a bigger sample of evaluators in a real peer-review process. Finally, last but not least, most of the evaluators (particularly the non-expert group) had no experience as reviewers. However, the good level of consensus achieved between experts and non-experts when miniRECH was used suggests that this approach can be used for training reviewers. It seems possible that training could have better impact on younger reviewers than those reviewers who have been reviewing for a long time [7,8,39].

In conclusion, the miniRECH tool is conceptually consistent with the framework for replicable and reproducible scientific research and, therefore, it should be considered as a strategy for improvement the reporting of scientific reports. This design of miniRECH offers two main features: firstly, by using checklists as templates, it prevents peer reviewers, even the experienced, against overlooking key information. And secondly, by using a scoring system based on scientific community feedback, it establishes a baseline level of the quality of reporting; which is, to some extent, comparable to the standards of a particular scientific community. This is particularly important in the peer-review process as a vital part of the quality control mechanism to determine what is published, and what is not; making these processes as helpful to authors as possible and preventing incomplete reporting from entering the literature.

Supporting information

File S1 miniRECH user's guide.

File S2 Survey for scoring the checklist criteria.

Acknowledgements

The authors want to thanks Clara Sanchez, Rocio Meneses, Sergio Gómez, Jhon Artunduaga, Inés Hernández and Szu-Wei Huang for accepting to take part in this work and their very useful comments.

Author contributions

OF-V, RS, AB and SE conceived the study. OF-V and JB developed the spreadsheet tool. OF-V, MB and SC designed and performed the experiments. OF-V and DP analysed the data. OF-V, MB and DP drafted the manuscript. SE, RS and AB; supervised the project. All authors discussed the findings and implications and commented on the manuscript at all stages.

References

1. Kronick DA (1990) Peer review in 18th-century scientific journalism. JAMA 263: 1321-1322.
2. Gannon F (2001) The essential role of peer review. EMBO Rep 2: 743.
3. Voight ML, Hoogenboom BJ (2012) Publishing your work in a journal: understanding the peer review process. International journal of sports physical therapy 7: 452-460.
4. Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, et al. (2009) Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PLoS One 4: e7824.
5. Chan AW, Altman DG (2005) Epidemiology and reporting of randomised trials published in PubMed journals. Lancet 365: 1159-1162.
6. Florez-Vargas O, Bramhall M, Noyes H, Cruickshank S, Stevens R, et al. (2014) The quality of methods reporting in parasitology experiments. PLoS One 9: e101131.
7. Schroter S, Black N, Evans S, Godlee F, Osorio L, et al. (2008) What errors do peer reviewers detect, and does training improve their ability to detect them? J R Soc Med 101: 507-514.
8. Schroter S, Black N, Evans S, Carpenter J, Godlee F, et al. (2004) Effects of training on quality of peer review: randomised controlled trial. BMJ 328: 673.

9. van der Worp HB, Macleod MR (2011) Preclinical studies of human disease: Time to take methodological quality seriously. Journal of Molecular and Cellular Cardiology 51: 449-450.
10. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, et al. (2012) A call for transparent reporting to optimize the predictive value of preclinical research. Nature 490: 187-191.
11. McNutt M (2014) Journals unite for reproducibility. Science 346: 679.
12. Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, et al. (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 26: 889-896.
13. Huang J, Mirel D, Pugh E, Xing C, Robinson PN, et al. (2011) Minimum Information about a Genotyping Experiment (MIGEN). Stand Genomic Sci 5: 224-229.
14. Hales BM, Pronovost PJ (2006) The checklist--a tool for error management and performance improvement. J Crit Care 21: 231-235.
15. Gawande A (2010) The Checklist Manifesto: How To Get Things Right. Great Britain: Profile Books Ltd.
16. GBSI (2013) The Case for Standards in Life Science Research: Seizing Opportunities at a Time of Critical Need. Washington, D.C.: Global Biological Standards Institute. 44 p.
17. Baker D, Lidster K, Sottomayor A, Amor S (2014) Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. PLoS Biol 12: e1001756.
18. Witwer KW (2013) Data submission and quality in microarray-based microRNA profiling. Clin Chem 59: 392-400.
19. Fuller T, Pearson M, Peters J, Anderson R (2015) What affects authors' and editors' use of reporting guidelines? Findings from an online survey and qualitative interviews. PLoS One 10: e0121585.
20. Shamseer L, Stevens A, Skidmore B, Turner L, Altman DG, et al. (2012) Does journal endorsement of reporting guidelines influence the completeness of reporting of health research? A systematic review protocol. Syst Rev 1: 24.
21. Roberts J (2010) Reporting Policies and Headache. Headache 50: 345-347.
22. Manchikanti L, Kaye AD, Boswell MV, Hirsch JA (2015) Medical journal peer review: process and bias. Pain Physician 18: E1-E14.
23. Whiting P, Harbord R, Kleijnen J (2005) No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 5: 19.
24. Bramhall M, Florez-Vargas O, Stevens R, Brass A, Cruickshank S (2015) Quality of methods reporting in animal models of colitis. Inflamm Bowel Dis 21: 1248-1259.
25. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG (2010) Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol 8: e1000412.
26. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, et al. (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 6: e1000100.
27. Team RC (2015) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
28. Juluru K, Eng J (2015) Use of Spreadsheets for Research Data Collection and Preparation:: A Primer. Acad Radiol 22: 1592-1599.
29. Wolstencroft K, Owen S, Horridge M, Krebs O, Mueller W, et al. (2011) RightField: embedding ontology annotation in spreadsheets. Bioinformatics 27: 2021-2022.
30. Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, et al. (2010) ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics 26: 2354-2356.
31. Rayner TF, Rocca-Serra P, Spellman PT, Causton HC, Farne A, et al. (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics 7: 489.
32. Bujkiewicz S, Jones HE, Lai MC, Cooper NJ, Hawkins N, et al. (2011) Development of a transparent interactive decision interrogator to facilitate the decision-making process in health care. Value Health 14: 768-776.
33. Shakespeare TP, Gebski VJ, Thiagarajan A, Jay Lu J (2006) Development of a spreadsheet for the calculation of new tools to improve the reporting of the results of medical research. Med Inform Internet Med 31: 121-127.
34. Godlee F, Gale CR, Martyn CN (1998) Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomized controlled trial. JAMA 280: 237-240.
35. Nylenna M, Riis P, Karlsson Y (1994) Multiple blinded reviews of the same two manuscripts. Effects of referee characteristics and publication language. JAMA 272: 149-151.
36. Callaham ML, Knopp RK, Gallagher EJ (2002) Effect of written feedback by editors on quality of reviews: two randomized trials. JAMA 287: 2781-2783.
37. Peng RD (2011) Reproducible research in computational science. Science 334: 1226-1227.
38. Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten simple rules for reproducible computational research. PLoS Comput Biol 9: e1003285.
39. Smith R (2006) Peer review: a flawed process at the heart of science and journals. J R Soc Med 99: 178-182.

Chapter Five

Assessing the reporting of laboratory animal models via miniRECH

The content of this chapter was published in the journal *Inflammatory Bowel Diseases*; full citation:

Bramhall M, Flórez-Vargas O, Stevens R, Brass A, Cruickshank S. Quality of methods reporting in animal models of colitis. *Inflamm Bowel Dis*. 2015; 21(6):1248-59.

In order to demonstrate the usefulness of miniRECH in assessing the quality of biomedical reporting, we considered the laboratory animals as models of humans in the context of translational research.

In recent years, an important debate has taken place about the extrapolation of findings from models to human beings. In a review by Niall Shanks *et al* on predictiveness of models for humans stated that "*[…] trans-species extrapolation is impossible vis-à-vis drug response and disease research especially when analysed in lights of the standards society today demands*" (Shanks, Greek, and Greek 2009). They are not the only ones concerned about the predictive power of experimental models for human conditions, this concern is also shared by others (Denayer, Stöhr, and Van Roy 2014, Cox 2015). Even though the position regarding the possibility of extrapolating from models to human beings is to some extent pessimistic, it is important in biomedical experimentation and should function as a reminder not to be naïve regarding extrapolation. Whether experimental models are good or bad models for human conditions is a philosophical and scientific debate that must be conducted with all the parties concerned in another scenario. Nevertheless, the presentation of this controversy is of critical importance in this scenario because the problem of extrapolation could be stemmed, to some extent, by the lack of reproducibility of experimental findings (Pound et al. 2004, Perrin 2014).

The success or failure of modelling human phenomena depends upon the validity of the model, and this validity is strongly related to the context in which the model is being used: the model and the thing being modelled. Therefore, a detailed description of the models and procedures carried out in them is fundamental for understanding the problem of extrapolation (van der Worp et al. 2010). Accordingly, there is an important need for identifying the factors associated with successful reproduction of basic science and translation to medical applications.

In an attempt to identify some of the factors that may influence the extrapolation from experimental models to human beings, we developed a checklist with a set of items that must be included in scientific articles when reporting bio-experiments using animal models of inflammatory bowel diseases (IBD) as a case study. IBD are a spectrum of multifactorial, chronic inflammatory diseases of the digestive tract, typically involving some degree of colitis. Inflammation, particularly the chronic type, is a complex and poorly understood pathway with important clinical significance both in terms of quality of life and financial impact. Therefore, it is vitally important that the animal experiments that inform almost all clinical practice are conducted rigorously and published in enough detail for others to benefit from and build upon.

In this chapter we aim to assess the current quality of methods reporting in published experiments in a subsection of available colitis models – at least 60 established IBD models are currently being used (Mizoguchi 2012). The assessment of the colitis models was performed via miniRECH by adapting to it a domain-specific checklist of essential and desirable reported methods criteria for these animal models.

Quality of Methods Reporting in Animal Models of Colitis

Michael Bramhall, MSc, Oscar Flórez-Vargas, MSc,* Robert Stevens, PhD,* Andy Brass, PhD,* and Sheena Cruickshank, PhD*[†]

Background: Current understanding of the onset of inflammatory bowel diseases relies heavily on data derived from animal models of colitis. However, the omission of information concerning the method used makes the interpretation of studies difficult or impossible. We assessed the current quality of methods reporting in 4 animal models of colitis that are used to inform clinical research into inflammatory bowel disease: dextran sulfate sodium, interleukin-$10^{-/-}$, CD45RB$^{\text{high}}$ T cell transfer, and 2,4,6-trinitrobenzene sulfonic acid (TNBS).

Methods: We performed a systematic review based on PRISMA guidelines, using a PubMed search (2000–2014) to obtain publications that used a microarray to describe gene expression in colitic tissue. Methods reporting quality was scored against a checklist of essential and desirable criteria.

Results: Fifty-eight articles were identified and included in this review (29 dextran sulfate sodium, 15 interleukin-$10^{-/-}$, 5 T cell transfer, and 16 TNBS; some articles use more than 1 colitis model). A mean of 81.7% (SD = ±7.038) of criteria were reported across all models. Only 1 of the 58 articles reported all essential criteria on our checklist. Animal age, gender, housing conditions, and mortality/morbidity were all poorly reported.

Conclusions: Failure to include all essential criteria is a cause for concern; this failure can have large impact on the quality and replicability of published colitis experiments. We recommend adoption of our checklist as a requirement for publication to improve the quality, comparability, and standardization of colitis studies and will make interpretation and translation of data to human disease more reliable.

(Inflamm Bowel Dis 2015;21:1248–1259)

Key Words: IBD, colitis, methods, animal models, checklist

Inflammatory bowel diseases (IBD) are a spectrum of multifactorial, chronic inflammatory diseases of the digestive tract, typically involving some degree of colitis. The etiology of IBD is still unclear, but genome-wide association studies have provided >160 contraindicated genetic loci for IBD susceptibility.[1] By knocking out or interfering with a number of these IBD-associated genes in animals (e.g., interleukin [IL]-$10^{-/-}$, IL-$2^{-/-}$, STAT$3^{-/-}$),[2] many of the symptoms, pathology, pathways, and histological features of IBD can be accurately reproduced in rodent models.[3] Mouse models have advanced our understanding of IBD and provided strong evidence of links between genetic predisposition and the loss of microbial tolerance in the onset of chronic colitis; as exemplified by genetically susceptible mice failing to develop colitis when housed in germ-free conditions.[4]

In order for the vast quantities of data derived from animal experimentation to be translated reliably into human studies, published experiments must be reported in sufficient detail to allow accurate comparison, reproduction, replication, and interpretation.[5] The ARRIVE (Animal Research: Reporting of In Vivo Experiments) guidelines suggest that reporting omissions prevent readers from reaching useful conclusions.[6] Recent work by the Reproducibility Initiative has highlighted the obstacles that can arise when repeating experimental work if the materials and methods have been insufficiently described in published articles.[7] In addition, this problem has become increasingly relevant due to the surge in interdisciplinary research, where experts from clinical or nonbiology backgrounds may be responsible for curating, managing, and analyzing data derived from laboratory experiments, and these individuals may not be able to identify or infer the missing details from experimental methods that could impact on data quality.

In recent years, a number of methods reporting guidelines and checklists have been developed, with a focus on a particular type of protocol (e.g., the minimum information guidelines group, MIBBI[8]) or a general theme, such as the ARRIVE guidelines for

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.ibdjournal.org).

Received for publication December 22, 2014; Accepted January 27, 2015.

From the *Bio-Health Informatics Group, School of Computer Science, University of Manchester, Manchester, United Kingdom; and †Manchester Immunology Group, Faculty of Life Science, University of Manchester, Manchester, United Kingdom.

Supported by a co-funding scholarship from EPSRC and Epistem Ltd awarded to M. Bramhall and a "Francisco José de Caldas" scholarship from Colciencias to O. Flórez-Vargas.

The authors have no conflicts of interest to disclose.

A. Brass and S. Cruickshank contributed equally to this study.

Reprints: Sheena Cruickshank, PhD, Faculty of Life Sciences, R4.004 AV Hill Building, University of Manchester, Oxford Road, Manchester M13 9PT, United Kingdom (e-mail: sheena.cruickshank@manchester.ac.uk).

DOI 10.1097/MIB.0000000000000369

Published online 29 April 2015.

experiments using animal models.[6] These interventions have largely been successful in raising awareness of flawed methods reporting within the scientific literature, gaining the support of journals, publishing houses, and members of the scientific community.[5,6,8,9] In several cases, publishers have implemented stricter guidelines for methods quality, introduced broad checklists, and removed limitations on word counts for methods reporting.[10] However, there is still a lag between implementation of these measures and adherence to them.[11]

We recently examined the quality of methods reporting in parasitology experiments,[12] highlighting the need for domain-specific guidelines: bespoke checklists tailored by experts that can be used to assess and improve the methods reporting quality within their community. These checklists can be implemented before the point of publication, acting as a barrier to prevent incomplete methods from entering the literature, and also as a review tool for nonexperts when assessing article quality postpublication. Animal models of colitis are numerous, with at least 60 established IBD models currently being used.[2] These models use diverse methods, and the exact mechanics of colitis induction (and the IBD they best model) are poorly understood in some cases. In this article, we aim to briefly summarize the types of colitis model that IBD researchers have at their disposal, highlight some of the problems that experimenters face in producing reliable and robust data from these models, and assess the current quality of methods reporting in published experiments in a subset of available colitis models; scoring them against a checklist of essential and desirable reported methods criteria. The selected criteria cover key aspects that can affect the outcome of colitis in animal models.

We have included checklist criteria relating to 3 broad areas. First, animal sex, age, origin, and housing is considered, which can affect the severity of inflammation, the balance of microbiota in the gut (e.g., strain, diet, acclimation), and animal stress levels (e.g., temperature, animals per cage), and therefore, collectively modulate the severity of induced colitis.[13–18] Second, factors pertaining to the colitis model, such as genetic modification of animals, origin of chemicals[19] and dosing should be recorded in order for the experiment to be repeatable under the same conditions. Finally, criteria relating to the measurement of colitis, time course of the experiment, and clinical monitoring of animals during the experiment should be reported as standard to determine the success of colitis induction and provide means by which similarity between experiments can be determined for inclusion into systematic reviews and meta-analyses.

Animal Models of IBD

Animal models of colitis have a number of distinct advantages over clinical data when it comes to determining the cause and prevention of IBD. For example, by controlling the onset of inflammation in the laboratory, the failures of immune tolerance, susceptibility genes, and specific proinflammatory pathways involved in triggering colitis can be identified more easily than in a patient admitted with progressive disease and potential comorbidities. Anticolitic preventative measures may also be tested before symptoms occur in an animal model, an impossible task in current treatment of human IBD, where new patients usually only present once the disease reaches clinical significance. The pathway of inflammation can also be accurately modulated in laboratory models to emulate acute or chronic disease depending on the strain of animal used, the mechanism of induction and the use of intervals between deliveries of proinflammatory stimulus.

Although the range of IBD models is diverse, they can be broadly categorized into 4 groups: chemically induced, biologically induced, genetic (including congenic and genetically modified animals), and cell transfer models. We have chosen a cross-section of colitis models to assess methods reporting quality in this field: dextran sulfate sodium (DSS), IL-10 knockout (IL-10$^{-/-}$), CD4^{+} CD45RB$^{\text{high}}$ T cell transfer, and 2,4,6-trinitrobenzene sulfonic acid (TNBS). In addition to animal housing conditions having an impact on the microbiota composition, which itself has a major impact on colitis models, different colitis models have specific criteria that influences their reproducibility as summarized below.

DSS-induced Colitis Model

DSS is one of the most commonly used inducers of colitis in animal models, thanks largely to the ease of use and potentially short turnaround times for obtaining results. DSS is typically administered in the drinking water of mice or rats at a dose dependent on the strain of animal, the severity of inflammation desired, and the length of the experiment. Acute and resolving inflammation usually occurs after a single continuous exposure to DSS in drinking water over a week or less, whereas repeated exposure punctuated with recovery periods results in chronic inflammation. The exact mechanism by which DSS induces colitis is still poorly understood, but its primary mode of action seems to chemically interfere with gut mucosa barrier integrity, allowing luminal antigens access to the lamina propria and the proinflammatory cells within.[20] Other factors that can influence the severity and susceptibility of exposure to DSS are the manufacturer and molecular weight of DSS,[19] the strain of animal used (C3H/HeJ and BALB/c mice show increased susceptibility), gender (males are more susceptible), and whether animals are raised in germ-free or specific pathogen-free environments.[20]

IL-10$^{-/-}$ Chronic Colitis Model

IL-10 is an anti-inflammatory cytokine that functions to prevent excessive inflammatory and autoimmune pathology.[21] Genome-wide association studies and clinical observations have identified IL-10 as a susceptibility gene for both Crohn's disease and ulcerative colitis.[22] By employing a number of genetic mechanisms, IL-10 or its receptor have been knocked out or functionally impaired to create several murine animal systems for the study of inflammation. IL-10$^{-/-}$ mice housed under normal conditions develop chronic inflammation in the gut, but mice will remain healthy when housed under germ-free conditions or with a defined selected microbiota and administration of antibiotics can prevent the onset of colitis in IL-10$^{-/-}$ mice.[21] Consequently, to standardize microbial influence on triggering colitis in the IL-$10^{-/-}$ model, specific enteric microbes such as *Enterococcus faecalis* or *Helicobacter hepaticus* may be used as an inoculum for mice that have been raised in germ-free housing.

T Cell Transfer Colitis Model

The T cell transfer model builds on the understanding that T lymphocytes play a pivotal role in the onset of colitis: mediating between antigen presenting cells and generating targeted immune responses to commensal enteric bacteria. In this model, naive T cells ($CD4^+$ $CD45RB^{high}$ or $CD4^+$ $CD62L^+$) are adoptively transferred from wild-type mice into genetically identical mice lacking T cells and B cells (e.g., SCID or $RAG^{-/-}$ mice). The onset of symptoms occurs 2 weeks after T cell transfer in the recipient mice, with pancolitis present from 4 weeks.[23] Due to the extraction, isolation, purification, and injection of adoptive T cells, this model requires a much more complex and labor-intensive protocol than many other IBD models. Factors that influence the resulting colitis include the strain of animal used, the number and viability of T cells transferred, and the presence of B cells in the recipient animals.[23]

TNBS-induced Colitis Model

TNBS is a chemical administered rectally in the form of an enema to mice or rats. TNBS is administered in combination with ethanol, which disrupts the mucous barrier, and it is generally thought that TNBS induces colitis by haptenating proteins within the gut, causing them to become preferential targets for immune cells.[24] As with other chemically induced colitis models, the severity of TNBS-induced colitis depends largely on the dosage applied and the strain of animal used.[24]

Scope of this Study

A vast amount of clinical and experimental IBD data are available for access: a PubMed search for the Medical Subject Headings (MeSH) term "inflammatory bowel diseases"[MeSH] from the year 2000 to present returns 30,931 articles. Researchers and health professionals cannot possibly hope to consult all the data to make decisions, so we are becoming increasingly reliant on meta-analyses and combinatory repositories to inform translation from animal experiments to clinical practice: it is vitally important that these processes are built on reliable foundations. This leads us to a pressing need to annotate and accurately record experiments from disparate sources, and this information is often lacking—not only does this prevent construction of well-founded knowledge-base systems, but it also prevents others from fully understanding the validity of results in the context of the experimental setting. How can a reader know whether 2 experiments are comparable if the methods from each experiment are not explicitly clear? In addition, geographical and language barriers or the use of nondomain experts may prevent the fluid exchange of tacit knowledge, resulting in subtle, yet important, omissions when describing experiments.[25]

To determine whether experiments in the field of primary colitis research are reported with adequate clarity and detail for replication, reproduction, and comparison, we defined a checklist of essential parameters that must be included and desirable parameters that ought to be included when describing experimental animal colitis. We then conducted a PubMed search to obtain a corpus of articles using DSS, IL-$10^{-/-}$, T cell transfer, or TNBS colitis models for assessing with the checklist. To gather a manageable number of results, we limited the search to studies published after 2000 that conducted a microarray on colitic tissues.

MATERIALS AND METHODS

A systematic search was performed following the recommendations of the PRISMA guidelines.[26] Relevant search terms were selected to identify published articles that used 1 (or more) of 4 animal models of colitis: DSS, IL-$10^{-/-}$, T cell transfer, or TNBS. The search was narrowed down to select only those articles that conducted a microarray on colonic tissues. Assessed criteria were divided into 3 sections in a protocol: aspects relating to the animal and its housing conditions, description of the model of perturbation used and criteria describing the assessment of colitis and the experimental design. The protocol used here for assessing criteria has not been previously published.

The literature search was conducted using PubMed in June 2014 and included articles published in English from January 1, 2000 to of June 1, 2014. The search terms included MeSH (Medical Subject Headings) terms and text strings, as outlined in Table 1. The year 2000 was selected as the cutoff due to the emergence of high-throughput analytical techniques becoming more commonplace after the publication of the first draft of the human genome. The DSS model was chosen as this is the most commonly used colitis model.[19] We also selected TNBS as a comparative chemical inducer of colitis, IL-$10^{-/-}$ to represent genetically modified colitis models, and T cell transfer as an example of a model that requires additional, more complex steps in its methods. Biologically induced colitis models, where bacterial or helminthic challenge is used to induce colitis, were not specifically included in this study. However, a number of IL-$10^{-/-}$ articles did include bacterial induction, where a specific cocktail of common murine bacterial strains were used to inoculate germ-free IL-$10^{-/-}$ mice (the checklist is capable of handling biologically induced colitis models). In addition, *Trichuris muris*–induced colitis, while not universally accepted as an IBD model, bears many phenotypic and transcriptional similarities to more traditional IBD models.[27] However, we chose not to include the *T. muris* infection model in this review as it was covered to some degree in our previous methods quality article.[12]

Inclusion Criteria

Primary research articles published in English, within the date constraints, that were returned in the PubMed search were considered for inclusion based on the title and abstract. Reviews, meta-analyses, and experiments that did not use any of the 4 chosen models were excluded. In addition, articles that conducted microarrays on human tissue or primary cell culture tissue only

TABLE 1. PubMed Search Terms Used for Each Colitis Model Included in the Systematic Review

Model	Search Terms
DSS	(Microarray[tw] OR “Microarray Analysis”[Mesh]) AND (“Dextran Sulfate”[Mesh] Dextran sulphate sodium [tw] OR Dextran sulfate sodium [tw] OR DSS [tw]) AND (Inflammatory Bowel Disease* [tw] OR IBD [tw] OR Crohn* Disease [tw] OR Ulcerative Colitis [tw] OR Coliti* [tw] OR Intestin* inflammat* [tw] OR Disease model* [tw] OR “Inflammatory Bowel Diseases”[MeSH] OR “Crohn Disease”[Mesh] OR “Colitis, Ulcerative”[Mesh] OR “Colitis”[MeSH] OR “Inflammation”[MeSH] OR “Disease Models, Animal”[Mesh])
IL-10$^{-/-}$	(Microarray[tw] OR “Microarray Analysis”[Mesh]) AND (IL-10 [tw] OR IL10 [tw] OR IL-10KO [tw] OR IL10KO [tw] OR Interleukin 10 [tw] OR Interleukin 10 [tw] OR “Interleukin-10”[Mesh]) AND (Inflammatory Bowel Disease* [tw] OR IBD [tw] OR Crohn* Disease [tw] OR Ulcerative Colitis [tw] OR Coliti* [tw] OR Intestin* inflammat* [tw] OR Disease model* [tw] OR “Inflammatory Bowel Diseases”[MeSH] OR “Crohn Disease”[Mesh] OR “Colitis, Ulcerative”[Mesh] OR “Colitis”[MeSH] OR “Inflammation”[MeSH] OR “Disease Models, Animal”[Mesh])
T cell transfer	(Microarray[tw] OR “Microarray Analysis”[Mesh]) AND (Adoptive transfer[tw] OR T cell transfer[tw] OR CD45RB[tw] OR CD45RBhigh[tw] OR “Antigens, CD45”[Mesh] OR “Adoptive Transfer”[Mesh]) AND (Inflammatory Bowel Disease* [tw] OR IBD [tw] OR Crohn* Disease [tw] OR Ulcerative Colitis [tw] OR Coliti* [tw] OR Intestin* inflammat* [tw] OR Disease model* [tw] OR “Inflammatory Bowel Diseases”[MeSH] OR “Crohn Disease”[Mesh] OR “Colitis, Ulcerative”[Mesh] OR “Colitis”[MeSH] OR “Inflammation”[MeSH] OR “Disease Models, Animal”[Mesh])
TNBS	(Microarray[tw] OR “Microarray Analysis”[Mesh]) AND (2,4,6- Trinitrobenzenesulfonic acid [tw] OR Trinitrobenzene sulphonic acid [tw] OR Trinitrobenzene sulfonic acid [tw] OR TNBS [tw] OR “Trinitrobenzenesulfonic Acid”[Mesh]) AND (Inflammatory Bowel Disease* [tw] OR IBD [tw] OR Crohn* Disease [tw] OR Ulcerative Colitis [tw] OR Coliti* [tw] OR Intestin* inflammat* [tw] OR Disease model* [tw] OR “Inflammatory Bowel Diseases”[MeSH] OR “Crohn Disease”[Mesh] OR “Colitis, Ulcerative”[Mesh] OR “Colitis”[MeSH] OR “Inflammation”[MeSH] OR “Disease Models, Animal”[Mesh])

Terms were chosen to cover both PubMed MeSH (Medical Subject Headings) and related strings to ensure that articles would still be captured even if they lacked correct subject heading annotations.

were also excluded, along with articles that were based on microarray data from a previous study. We also excluded combined colitis and carcinogenesis models. The resulting corpus of articles was assessed using the bespoke methods reporting checklist for animal models of colitis.

Checklist

A checklist of essential criteria that must be included and nonessential criteria that are useful to include when reporting the results of animal models of colitis was drawn up (Table 2), with additional input by experts in the field of colitis research. Articles were assessed on whether they included each criterion within the published article, supplementary methods, or relevant cited articles. For each criterion, an article received a weighted score if the criterion was present or not applicable, and zero if the item was absent. Total scores for all criteria were tallied to provide a final percentage score for successfully reported criteria. Data extraction and assessment was conducted by one reviewer, and half of the articles were randomly selected and scored blind by the second reviewer. Inconsistencies were discussed by both reviewers until a consensus was reached.

Weighting

Weight per item was determined in consultation with 3 colitis experts (Table 2). Criteria were assigned a weight by a combination of 2 factors: whether the item was considered essential (Y/N), and whether the item was determined to be of low, medium, or high importance (L/M/H). Weighted scores were allocated as follows: (Y = 5 or N = 2) and (L = 3 or M = 4 or H = 5). Therefore, each criterion received a score between 5 and 10, which was then used to determine the weight as a percentage of the sum of all scores. Where disagreement occurred in allocating weighting to criteria, the majority vote was used.

Journal Impact Factor

Journal impact factor (IF) was retrieved from the Institute of Scientific Information (ISI) Journal Citation Reports (JCR) database 2013.

Confirmation of Impartiality in Scoring of Studies

Half of all articles accepted were randomly selected and scored using the checklist by the second reviewer. Differences between scores were assessed using a Bland–Altman comparison and linear correlation to determine whether any reviewer bias was present.

Statistical Analysis

Data were analyzed by two-way analysis of variance, Bland–Altman correlation, and linear correlation using GraphPad Prism version 6.05 (Windows) and 6.0f (Mac), GraphPad Software, La Jolla CA, www.graphpad.com.

Ethical Considerations

There are no ethical considerations.

TABLE 2. Checklist of Essential and Desirable Criteria and the Weighting Applied to Each Criterion for Reporting Methods in Animal Models of Colitis

Group	Subgroup	No.	Item	Essential	Importance	Score	Weight, %
Information about the animal	Animals	1.1	Is the species of animal identified? (e.g., mouse)	Yes	High	10	2.7
		1.2	Is the strain of animal identified? (e.g., C57BL/6)	Yes	High	10	2.7
		1.3	Is the age of the animal described? (e.g., 12 wks old)	Yes	High	10	2.7
		1.4	Is the gender of the animal described? (e.g., male)	Yes	High	10	2.7
		2.1	Is the source of animals defined? (e.g., name of supplier or bred in facility)	Yes	High	10	2.7
		2.2	Were animals acclimated to local microbiota? (e.g., housed in identical conditions at least 7 d before experiment start)	Yes	High	10	2.7
	Animal housing conditions	3.1	Is the light/dark cycle described? (e.g., 12 hours light/dark)	No	High	7	1.89
		3.2	Is the temperature described? (e.g., 25°C)	No	Low	5	1.35
		3.3	Is the humidity described? (e.g., 40%–45%)	No	Low	5	1.35
		3.4	Is the food/water described? (e.g., regular chow)	Yes	Medium	9	2.43
		3.5	Is the number of animals per cage described? (e.g., 3 mice per cage)	No	Low	5	1.35
Information about the colitis model	Genetically modified animals	4.1	Is the genetic modification identified? (e.g., IL-$10^{-/-}$)	Yes	High	10	2.7
		4.2	Is the background strain of the animal described? (e.g., BALB/c)	Yes	High	10	2.7
	Chemically induced colitis model (e.g., DSS)	5.1	Is the chemical used to induce colitis specified? (e.g., DSS)	Yes	High	10	2.7
		5.2	Is the molecular weight of the chemical specified? (e.g., 36–50 kDa) (DSS only)	Yes	High	10	2.7
		5.3	Is the supplier of the chemical identified? (e.g., Sigma Aldrich)	Yes	Low	8	2.16
		5.4	Is the method of induction described? (e.g., dissolved in drinking water)	No	High	7	1.89
		5.5	Is the dosage used described? (e.g., 2% wt/vol)	Yes	High	10	2.7
		5.6	Is the medium of inoculation described? (e.g., TNBS in ethanol)	Yes	Medium	9	2.43
	Biologically induced colitis model (e.g., bacterial infection)	6.1	Is the species of organism identified? (e.g., *Helicobacter pylori*)	Yes	High	10	2.7
		6.2	Is the strain of organism identified? (e.g., PMSS1)	Yes	High	10	2.7
		6.3	Are the culture conditions described? (e.g., animal passage or cell culture)	No	Medium	6	1.62
		6.4	Is parasitemia/colonization adequately assessed? (e.g., colon homogenized and plated for colony counting)	Yes	High	10	2.7
		6.5	Is the method of inoculation described? (e.g., oral gavage)	Yes	High	10	2.7
		6.6	Is the dosage used described? (e.g., 10^8 cells)	Yes	High	10	2.7
	Adoptive transfer colitis model (e.g., T cell transfer)	7.1	Is the cell type being transferred described? (e.g., $CD4^+$ $CD45RB^{high}$)	Yes	High	10	2.7
		7.2	Is the species of the donor animal identified? (e.g., mouse)	Yes	High	10	2.7

TABLE 2 (*Continued*)

Group	Subgroup	No.	Item	Essential	Importance	Score	Weight, %
		7.3	Is the strain of the donor animal identified? (e.g., C57BL/6)	No	High	7	1.89
		7.4	Is the gender of the donor animal described? (e.g., male)	No	Medium	6	1.62
		7.5	Is the number of cells transferred specified? (e.g., 4 × 10^5)	Yes	High	10	2.7
		7.6	Is the purity of cells transferred specified? (e.g., >95%)	No	High	7	1.89
		7.7	Is the viability of cells confirmed before transfer? (e.g., via 7-AAD staining during FACS)	Yes	High	10	2.7
		7.8	Is the method of cell transfer described? (e.g., intraperitoneal injection)	Yes	High	10	2.7
Information about the experimental design	Experiment design	8.1	Is the time course of the experiment described? (e.g., mice killed after 7 d exposure to DSS)	Yes	High	10	2.7
		8.2	Is the method of euthanasia described? (e.g., cervical dislocation)	No	Medium	6	1.62
		8.3	Is animal weight loss reported? (e.g., as daily % of starting weight)	Yes	High	10	2.7
		8.4	Is mortality reported? (e.g., survival curve)	Yes	High	10	2.7
	Colitis monitoring and scoring	9.1	Is colitis monitored clinically? (e.g., disease activity index)	No	High	7	1.89
		9.2	Is colitis scored histologically? (e.g., H&E stain)	Yes	High	10	2.7
		9.3	Is microbiota diversity/population assessed? (e.g., 16S rRNA sequencing)	No	High	7	1.89
		9.4	Is colon length or weight measured after being killed?	Yes	Medium	9	2.43
		9.5	Is the section of gut for analysis identified? (e.g., proximal colon)	Yes	High	10	2.7

For the 4 subsections within "Information about the colitis model," only the relevant subsections were required. Weights are determined by points attributed to whether the criterion is deemed essential (Yes = 5 or No = 2) plus the level of importance (High = 5, Medium = 4, or Low = 3). The weight for each criterion is then calculated as the percentage of the sum of all scores.

RESULTS

Search Strategy

A total of 58 unique studies were identified for inclusion in the review (see Fig., Supplemental Digital Content 1, http://links.lww.com/IBD/A789). Six of the included articles were applicable to more than 1 of the colitis models and were subsequently included in the datasets for every relevant model (29 DSS,[28–56] 15 IL-$10^{-/-}$,[36,49,50,57–68] 5 T cell transfer,[56,69–72] and 16 TNBS[35,56,61,73–85]; for details of all included studies see Table, Supplemental Digital Content 2, http://links.lww.com/IBD/A790). Duplicate articles were only included once in summary analyses where data from all models are combined. The PubMed searches returned 256 unique articles (54 DSS, 146 IL-$10^{-/-}$, 42 T cell transfer, and 21 TNBS), 188 of which were rejected based on the title and abstract. A further 10 articles were excluded after assessing the full text of the article, leaving a corpus of 58 articles for analysis.

Quality of Methods Reporting

Each article was assessed for inclusion of the criteria outlined in the quality checklist, which was subdivided into 3 domains: animal, model, and experiment—correlating with subject, perturbation and outcome. The mean weighted score across all colitis models was 81.7% (SD = ±7.038) of criteria reported. By model, articles using the DSS model had the highest quality of methods reporting (mean = 83.30%, SD = ±7.019), and the lowest quality was observed in articles using the T cell transfer model (mean = 73.19%, SD = ±5.328): significantly lower than DSS ($P \leq 0.01$) and IL-$10^{-/-}$ ($P \leq 0.05$) colitis models (Fig. 1A). Individually, the article with the lowest mean score was 64.05% (T cell transfer model[72]), and the highest recorded was 94.86% (DSS model[52]).

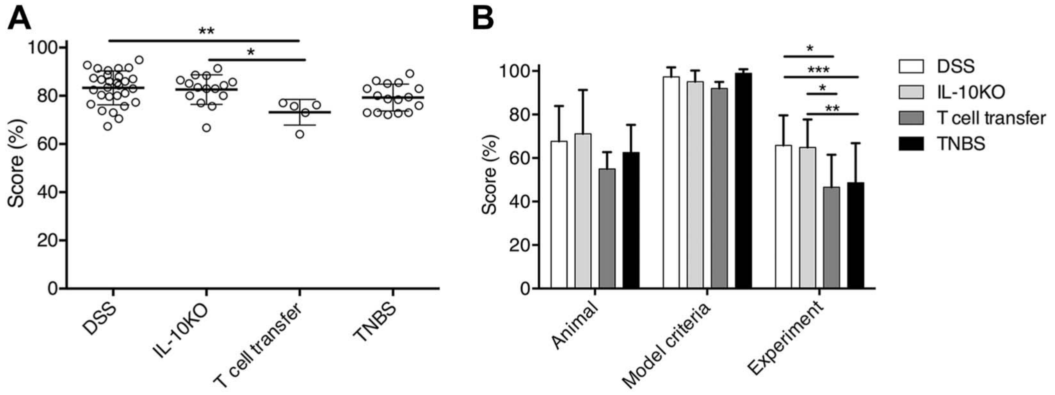


FIGURE 1. A, Overall scores (percent criteria reported) for the quality of methods reporting for each colitis model included in this review. The T cell transfer model scored significantly lower than DSS ($P \leq 0.01$) and IL-$10^{-/-}$ ($P \leq 0.05$) colitis models. n = 29 (DSS), 15 (IL-$10^{-/-}$), 5 (T cell transfer), and 16 (TNBS). Analysis by two-way ANOVA. B, Methods reporting quality (percent criteria reported) for each of the 3 subsections of the quality reporting checklist. Criteria relating to the model subsection scored higher than the animal and experimental design subsections. Within the experimental design subsection, DSS and IL-$10^{-/-}$ scored significantly higher than both T cell transfer ($P \leq 0.05$) and TNBS ($P \leq 0.001$ and $P \leq 0.01$, respectively) colitis models. n = 29 (DSS), 15 (IL-$10^{-/-}$), 5 (T cell transfer), and 16 (TNBS). Analysis by two-way ANOVA. ANOVA, analysis of variance.

No article reported 100% of all of the criteria on our checklist but 1 article (DSS model[39]) of all the 58 articles assessed successfully reported all essential criteria for every domain.

The best reported domain was the model itself (mean = 95.80%, SD = ±3.018), followed by animal criteria (mean = 64.05%, SD = ±6.992) and experiment criteria (mean = 56.44%, SD = ±10.225). Looking at scores per domain by colitis model, IL-$10^{-/-}$ had the highest quality for the animal domain (mean = 70.99%, SD = ±20.194), TNBS had the highest quality for the model domain (mean = 98.94%, SD = ±1.914), and DSS had the highest quality for the experiment domain (mean = 65.78%, SD = ±13.810). The T cell transfer model had the lowest mean scores for all 3 domains (animal = 54.95%, SD = ±7.770; model = 92.00%, SD = ±2.937; experiment = 46.58%, SD = ±14.908) (Fig. 1B). For full details of methods reporting quality for each included study see Tables, Supplemental Digital Content 3-14, http://links.lww.com/IBD/A935, http://links.lww.com/IBD/A936, http://links.lww.com/IBD/A937, http://links.lww.com/IBD/A938, http://links.lww.com/IBD/A939, http://links.lww.com/IBD/A940, http://links.lww.com/IBD/A941, http://links.lww.com/IBD/A942, http://links.lww.com/IBD/A943, http://links.lww.com/IBD/A944, http://links.lww.com/IBD/A945, and http://links.lww.com/IBD/A946.

DSS-induced Colitis Model

For DSS colitis, the most poorly reported criteria for the animal domain were food/water, acclimation, animal gender, and animal age (44.83%, 41.38%, 31.03%, and 20.69% of articles failed to report the criteria, respectively). When describing the DSS model itself, 9 articles (31.03%) failed to provide any information about the molecular weight of the DSS used, and 17.24% of articles did not provide information about the supplier of the DSS chemical (Fig. 2). A more detailed examination of the reporting of molecular weight of DSS revealed that of the 20 articles (68.97%) that proved information about the molecular weight of DSS, only 5 (17.24%) used the correct units of measurement: of the remaining 15 articles, 13 (44.83%) provided no units and 2 (6.90%) used incorrect units. Of the 29 articles that used DSS colitis, 24 (82.76%) failed to correctly report the nature of the DSS molecule that they used to induce colitis. The worst reported essential criteria in the experiment design domain were mortality reporting, colon length/weight measurements, animal weight loss, and colitis scoring by histology (72.41%, 51.72%, 20.69%, and 10.34% of articles failed to report these criteria, respectively).

IL-$10^{-/-}$ Chronic Colitis Model

In the animal domain, the criteria most poorly reported in the articles using the IL-$10^{-/-}$ model were very similar to those missing in the DSS model: acclimation, gender, and food/water were the most commonly absent essential criteria (46.67%, 40%, and 33.33% of articles failed to report, respectively). For the IL-$10^{-/-}$ model itself, measurement of bacterial colonization in the gut was poorly reported when specific bacterial inoculation was used to induce

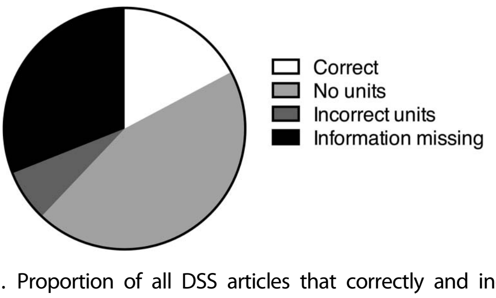


FIGURE 2. Proportion of all DSS articles that correctly and incorrectly described the molecular weight of the DSS used in the experiment. Correct reporting of DSS was only described in 17.24% of articles, and no information at all was provided in 31.03% the studies assessed (n = 29).

colitis (53.33% failed to report criteria). In addition, 26.67% of IL-$10^{-/-}$ articles did not specify the strain(s) of bacteria used to induce colitis. The worst reported criteria relating to the experimental design were mortality reporting and colon weight/length measurements, which were both absent in 66.67% of articles.

T Cell Transfer Colitis Model

For articles using the T cell transfer model, the worst reported criteria in the animal domain were food/water and acclimation (100% and 80% of articles failed to report these criteria, respectively). Gender of animals used was also not specified in 1 of the 5 T cell transfer articles (20%). When describing the T cell transfer model itself, none of the 5 articles described how viability of T cells transferred was measured or whether it was measured at all. For the experimental design, no article using T cell transfer reported mortality of animals used, 60% of articles failed to report colon length/weight measurements, and 40% of articles failed to report animal weight during the experiment.

TNBS-induced Colitis Model

Articles using TNBS to induce colitis were the worst for reporting whether animals had been acclimated (87.5% of articles failed to report this criterion). Also, food/water supply and age of animals used was missing in 50% and 25% of articles, respectively. The TNBS model itself was well reported, although 18.75% of articles failed to report the supplier of the TNBS. Similar to the other colitis model, the worst reported essential criteria in the experiment design domain for TNBS were mortality reporting, colon length/weight measurements, animal weight loss, and colitis scoring by histology (75%, 75%, 43.75%, and 37.5% of articles failed to report these criteria, respectively).

More Recent Articles Have Higher Methods Reporting Quality

Overall scores have significantly improved year on year ($P = 0.037$, $r^2 = 0.075$). T cell transfer is the only model to have a drop in methods reporting quality over time, but this is not significant. DSS and IL-$10^{-/-}$ show a trend toward improved methods reporting quality over time and TNBS overall reporting quality has significantly improved with time ($P = 0.0036$, $r^2 = 0.4659$) (Fig. 3A). The improvement in TNBS reporting quality over time has largely come from a significant improvement in the experiment domain ($P = 0.0203$, $r^2 = 0.3285$) (Fig. 3B).

Journal IF Has No Relation to Methods Reporting Quality

IF was not observed to have a significant impact on methods reporting quality in animal models of colitis (Fig. 3C). When broken down into domains, there was a slight negative

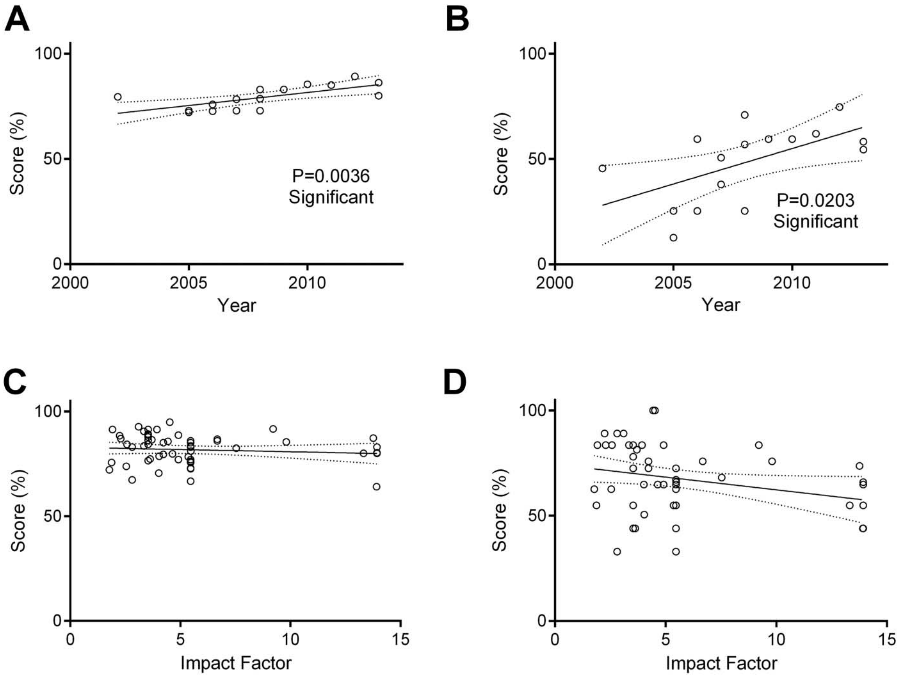


FIGURE 3. A, A significant positive correlation ($P \leq 0.01$, $r^2 = 0.47$) is seen between overall methods reporting quality score (%) and year of publication in studies using TNBS-induced colitis. B, The source of this correlation comes largely from the strong positive correlation ($P \leq 0.05$, $r^2 = 0.33$) between reporting quality (%) and year of publication within the experimental design subsection in TNBS colitis papers (n = 16). C, IF of the journal of publication had no impact on the overall quality of methods reporting. D, By subdomain, a nonsignificant negative correlation between reduced methods reporting quality and increased IF was observed in the animal domain ($P = 0.0536$, $r^2 = 0.07$) (n = 58). Analyses by linear correlation.

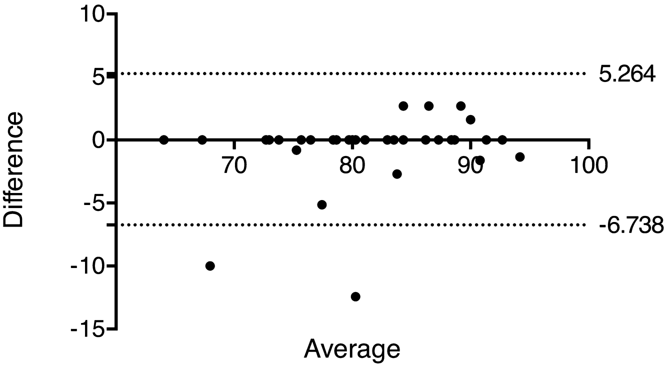


FIGURE 4. Bland–Altman plot to assess agreement between 2 experimenters in scoring articles with the minimum information checklist (n = 33). Articles were scored by the second marker, representing at least half the articles assessed for each model. Difference in scores is not significantly different from zero ($P = 0.149$, $r^2 = 0.066$).

correlation between IF and quality score in the animal domain, but this was not significant ($P = 0.0536$, $r^2 = 0.06488$) (Fig. 3D).

Verification of Consistency in Scoring of Studies

The second examiner scored 33 of the 58 articles included in the review (DSS = 14, IL-$10^{-/-}$ = 8, T cell transfer = 3, and TNBS = 8). Differences in scores for the 2 examiners were assessed through a Bland–Altman plot (Fig. 4). Difference in scores between examiners did not differ significantly from zero ($P = 0.149$, $r^2 = 0.066$) suggesting that there was no bias in scoring, and articles were scored consistently with the minimum information checklist.

DISCUSSION

Chronic inflammation is a complex and poorly understood pathway with important clinical significance both in terms of quality of life and financial impact. It is vitally important that the animal experiments that inform almost all clinical practice are conducted rigorously and published in enough detail for others to benefit from and build upon, which would be in agreement with the principles stated in the 3 Rs (replace, reduce, and refine).[86] To examine the quality of methods reporting in animal models of colitis and determine the potential impact on reliability, replicability, and comparability of studies in this field, we have assessed 4 commonly used animal models of colitis: DSS, IL-$10^{-/-}$, T cell transfer, and TNBS. Our results indicate that although these models score well against a checklist of essential criteria, there are still a variety of fundamental criteria that are repeatedly omitted. It is also encouraging to see an improvement over time, even if this effect is quite small. However, the fact that only 1 article from a corpus of 58 reported all essential criteria is a huge cause for concern, 98.3% of articles included in this analysis failed to include sufficient information to accurately repeat the experiment.

In the United Kingdom, death as an endpoint in animal experiments is to be avoided wherever possible.[87] However, mortality and morbidity does occur from time to time and for a variety of reasons, and this should be reported as it will have a significant impact on the data produced and the results of statistical analyses. A statement referring to animal mortality, even if no animal died during the experiment, was one of the worst reported essential criteria from the checklist across all 4 colitis models included in this analysis (48 of 58 articles, 82.76%, failed to include this criterion). Most animal models of colitis are not expected to cause significant morbidity or death, but the lack of reporting, even to confirm that no unexpected deaths occurred, is problematic. When results from animal experiments fail to disclose mortality, bias may be introduced, giving an overly optimistic estimate of the efficacy of the intervention.[88] For example, without adverse event reporting being enforced, there is no obligation for researchers to declare mice that die during an animal study, but failing to declare this information potentially puts the safety of animals and people in future trials at risk.[89] We are not suggesting that the studies included in this review are deliberately obscuring potentially harmful results, and we assume a lack of adverse event reporting reflects an absence of adverse events to report. However, without such a declaration, we cannot say for certain either way. Consequently, animal experiments should align more closely with clinical practice in this regard and declare adverse reactions as a matter of course.[90]

The key role of gut microbiota in the onset and severity of chronic colitis is well defined.[14] Thus, it was surprising that more than half of the studies (63.79%) failed to describe how animals had been acclimated to ensure potential differences in microbiota had been accounted for and controlled. In addition, very few articles specified the use of littermate controls, which would be the ideal gold-standard for controlling baseline equivalence in microbiota populations. It is insufficient to assume animals obtained from the same supplier or reared within the same experimental facility will harbor equivalent microbial populations, as differences can and do exist even within rooms or across facilities.[91] Simple tools to characterize microbiota are available,[16] and, ideally, these should be used to improve standardization and tighten controls within experiments. Alternatively, cohousing or litter mate controls reduce the likely impact of the environment. Additionally, acclimation serves to compensate for stresses involved in transporting animals. Moving cages to a new location in the same facility can have stressful effects on animals lasting several weeks, ultimately influencing immune responses in experimental conditions.[18] Movement of animals should be kept to a minimum and laboratory animals require up to 7 days for changes in immune and endocrine parameters to return to baseline before experimental procedures begin[92]; needless to say, these details should be declared in the methods of the study write-up.

Another key factor in determining microbial consistency is diet, with various dietary factors influencing the growth of different bacterial populations in the gut.[15,93] Again, over half of the studies (53.45%) in our analysis failed to define the chow fed to experimental animals, a factor that can have significant effects on the severity of induced colitis and the microbiota present in the gut.[15] Better standardizations are required for studies where gut microbiota can influence results, and colonization of laboratory animals with defined microbial populations would introduce a new level of control in these experiments.[94]

Reporting the gender of animals was one of few criteria where the quality differed depending on the animal model used, with 9 DSS studies and 6 IL-$10^{-/-}$ studies failing to report animal gender compared with just 1 study each from the T cell transfer and TNBS-induced models. The role of gender in inflammation is well established, with females (in both mice and humans) being more susceptible to developing autoimmune diseases and mounting a more pronounced inflammatory response than males.[17,95] In addition, sex differences also occur within animal models of colitis: male mice are more susceptible to DSS colitis, for example.[19] Failing to describe the gender of animals in an experiment relying on inflammation obscures vital information when trying to infer meaning from the results and prevents data from different studies from being reliably compared.

A number of criteria relating to animal housing were considered to be nonessential in our checklist, yet temperature, humidity, light/dark cycle, and the number of animals per cage were repeatedly omitted from the methods of between 50% and 100% of the studies assessed, depending on the model used. Temperature in particular can affect the immune system of mice, with low temperatures triggering immunosuppressive responses.[96] Many studies are conducted where animal facilities are kept at "room temperature" (19–22°C) to suit the experimenters but not necessarily the animals that they house: wild mice spend daytime inactive, nesting at 30 to 32°C and are therefore experiencing cold stress in the majority of animal facilities.[96,97] Also, in addition to behavioral and immunological changes,[98] mice housed alone will have to endure cooler conditions that mice housed in groups. Severity of colitis in the DSS model is strongly linked to the strain of animal used and the specifications of the DSS itself. Large molecular weight DSS (≥500 kDa) fails to bypass the mucous barrier and does not induce colitis,[99] whereas smaller preparations of DSS (5–40 kDa) elicit colitic responses in a spectrum of disease severity.[19] Although DSS is commonly prepared at around 40 kDa, not all experimenters obtain DSS from the same supplier or at the same molecular weight. That only 5 of the 29 DSS articles accurately reported the molecular weight of DSS with the appropriate units is problematic. The presence of arbitrary numbers with no denomination specified or with clearly incorrect units resulting in claims of molecular weight out by orders of magnitude (e.g., kDa instead of Da, or vice versa) in published studies is poor. The increased number of interdisciplinary, nondomain specialists involved in curating and annotating datasets for inclusion in meta-analyses means that this sort of information must be included within the methods of published articles. Authors of studies cannot assume that everyone accessing their study has the expertise to be able to infer the fine details of the protocols they used. Thus, these sorts of errors appearing in the literature suggest potential shortcomings in submission, peer review, and journal editing processes. It is often the responsibility of submitting authors to ensure that there are no errors in a submitted manuscript but peer reviewers ought to be spotting these errors before an article gets to print.

We recommend the continued uptake of methods quality checklists to assist authors and publishers with inclusion of all the relevant methods details that are required to fully interpret data and integrate results into larger analyses. We have provided a domain-specific checklist that can be used in the assessment of methods reporting in any colitis model, and we think this will aid translation of discoveries in animal models into human studies. However, we are aware that by including only microarray studies, we are focusing on a subset of published colitis research. Methods reporting quality for animal models of colitis in general may not reflect the results we have reported here. Also, we have not attempted to address the diversity of experimental design within models or the choice of statistical tests and power calculations used in analysis of data in this field, both of which will impact the feasibility of comparing data from colitis models. It is worth noting that, although all the studies in this review detailed the numbers of mice used per group, none of the studies included any statistical measure of power to justify the number of animals used. This is of concern, as power calculations are important for assessing the validity of statistical tests applied to the data generated and to limit unnecessary use of animals in research.[6,100]

In conclusion, we have demonstrated that the quality of methods reporting in modeling colitis, while generally appearing high, has serious flaws with long-ranging impact on the translation of primary research into clinical research of IBD. Automated methods, such as computerized histology scoring,[101] may become more commonplace in future, assisting experimenters in standardizing their methods, but more needs to be done to promote and enforce existing guidelines. Animal experimenters have an onus to follow the 3 Rs (replace, reduce, and refine), and better reporting of studies will add value to experimental data produced by animal studies.[86] Implementation of our colitis methods checklist would improve the quality of publications in this field, ensuring animal models, and the data they produce are used effectively to fulfill their maximum usefulness. The pipeline from basic science to clinical practice is filled with examples where success in the laboratory fails to translate into human subjects and improving methods reporting would be an excellent starting point in rectifying this problem at very little cost or effort.

ACKNOWLEDGMENTS

We wish to thank Cath Booth and James Wilson at Epistem, Dr. Joanne Pennock and Professor Werner Muller at the University of Manchester for support in this study.

REFERENCES

1. Van Limbergen J, Radford-Smith G, Satsangi J. Advances in IBD genetics. *Nat Rev Gastroenterol Hepatol.* 2014;11:372–385.
2. Mizoguchi A. Animal models of inflammatory bowel disease. In: Conn PM, ed. *Progress in Molecular Biology and Translational Science*. London, United kingdom: Academic Press; 2012;105:263–320.
3. Goyal N, Rana A, Ahlawat A, et al. Animal models of inflammatory bowel disease: a review. *Inflammopharmacology*. 2014;22:219–233.
4. Kamada N, Seo SU, Chen GY, et al. Role of the gut microbiota in immunity and inflammatory disease. *Nat Rev Immunol.* 2013;13:321–335.
5. Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet*. 2014;383: 267–276.

6. Kilkenny C, Browne WJ, Cuthill IC, et al. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol.* 2010;8:e1000412.
7. Noorden RV. Parasite test shows where validation studies can go wrong [Nature News web site]. 2014. Available at: http://www.nature.com/news/parasite-test-shows-where-validation-studies-can-go-wrong-1.16527?WT.mc_id=TWT_NatureNews. Accessed December 19, 2014.
8. Taylor CF, Field D, Sansone SA, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol.* 2008;26:889–896.
9. 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.
10. Raising standards. *Nat Genet.* 2013;45:467.
11. Baker D, Lidster K, Sottomayor A, et al. Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS Biol.* 2014;12:e1001756.
12. Flórez-Vargas O, Bramhall M, Noyes H, et al. The quality of methods reporting in parasitology experiments. *PLoS One*. 2014;9:e101131.
13. Bonaz BL, Bernstein CN. Brain-gut interactions in inflammatory bowel disease. *Gastroenterology*. 2013;144:36–49.
14. Guinane CM, Cotter PD. Role of the gut microbiota in health and chronic gastrointestinal disease: understanding a hidden metabolic organ. *Therap Adv Gastroenterol.* 2013;6:295–308.
15. Hekmatdoost A, Feizabadi MM, Djazayery A, et al. The effect of dietary oils on cecal microflora in experimental colitis in mice. *Indian J Gastroenterol.* 2008;27:186–189.
16. Hufeldt MR, Nielsen DS, Vogensen FK, et al. Variation in the gut microbiota of laboratory mice is related to both genetic and environmental factors. *Comp Med.* 2010;60:336–347.
17. Ngo ST, Steyn FJ, McCombe PA. Gender differences in autoimmune disease. *Front Neuroendocrinol.* 2014;35:347–369.
18. Olfe J, Domanska G, Schuett C, et al. Different stress-related phenotypes of BALB/c mice from in-house or vendor: alterations of the sympathetic and HPA axis responsiveness. *BMC Physiol.* 2010;10:2.
19. Perše M, Cerar A. Dextran sodium sulphate colitis mouse model: traps and tricks. *J Biomed Biotechnol.* 2012;2012:13.
20. Low D, Nguyen DD, Mizoguchi E. Animal models of ulcerative colitis and their application in drug research. *Drug Des Devel Ther.* 2013;7: 1341–1357.
21. Iyer SS, Cheng G. Role of interleukin 10 transcriptional regulation in inflammation and autoimmune disease. *Crit Rev Immunol.* 2012;32: 23–63.
22. Khor B, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. *Nature*. 2011;474:307–317.
23. Ostanin DV, Bao J, Koboziev I, et al. T cell transfer model of chronic colitis: concepts, considerations, and tricks of the trade. *Am J Physiol Gastrointest Liver Physiol.* 2009;296:135–146.
24. Wirtz S, Neufert C, Weigmann B, et al. Chemically induced mouse models of intestinal inflammation. *Nat Protoc.* 2007;2:541–546.
25. A little knowledge. *Nature*. 2014;514:139–140.
26. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* 2009;6:e1000097.
27. Levison SE, McLaughlin JT, Zeef LA, et al. Colonic transcriptional profiling in resistance and susceptibility to trichuriasis: phenotyping a chronic colitis and lessons for iatrogenic helminthosis. *Inflamm Bowel Dis.* 2010;16:2065–2079.
28. Benight NM, Stoll B, Marini JC, et al. Preventative oral methylthioadenosine is anti-inflammatory and reduces DSS-induced colitis in mice. *Am J Physiol Gastrointest Liver Physiol.* 2012;303:G71–G82.
29. Breynaert C, Dresselaers T, Perrier C, et al. Unique gene expression and MR T2 relaxometry patterns define chronic murine dextran sodium sulphate colitis as a model for connective tissue changes in human Crohn's disease. *PLoS One*. 2013;8:e68876.
30. Cho JY, Chi SG, Chun HS. Oral administration of docosahexaenoic acid attenuates colitis induced by dextran sulfate sodium in mice. *Mol Nutr Food Res.* 2011;55:239–246.
31. Cho JY, Hwang JK, Chun HS. Xanthorrhizol attenuates dextran sulfate sodium-induced colitis via the modulation of the expression of inflammatory genes in mice. *Life Sci.* 2011;88:864–870.
32. Chung SH, Park YS, Kim OS, et al. Melatonin attenuates dextran sodium sulfate induced colitis with sleep deprivation: possible mechanism by microarray analysis. *Dig Dis Sci.* 2014;59:1134–1141.
33. Coburn LA, Gong X, Singh K, et al. L-arginine supplementation improves responses to injury and inflammation in dextran sulfate sodium colitis. *PLoS One*. 2012;7:e33546.
34. Fang K, Bruce M, Pattillo CB, et al. Temporal genomewide expression profiling of DSS colitis reveals novel inflammatory and angiogenesis genes similar to ulcerative colitis. *Physiol Genomics.* 2011;43:43–56.
35. Hamilton MJ, Sinnamon MJ, Lyng GD, et al. Essential role for mast cell tryptase in acute experimental colitis. *Proc Natl Acad Sci U S A.* 2011; 108:290–295.
36. Hansen JJ, Holt L, Sartor RB. Gene expression patterns in experimental colitis in IL-10-deficient mice. *Inflamm Bowel Dis.* 2009;15:890–899.
37. Hontecillas R, Horne WT, Climent M, et al. Immunoregulatory mechanisms of macrophage PPAR-gamma in mice with experimental inflammatory bowel disease. *Mucosal Immunol.* 2011;4:304–313.
38. Iizuka Y, Okuno T, Saeki K, et al. Protective role of the leukotriene B4 receptor BLT2 in murine inflammatory colitis. *FASEB J.* 2010;24:4678–4690.
39. Jia Q, Ivanov I, Zlatev ZZ, et al. Dietary fish oil and curcumin combine to modulate colonic cytokinetics and gene expression in dextran sodium sulphate-treated mice. *Br J Nutr.* 2011;106:519–529.
40. Kabashima K, Saji T, Murata T, et al. The prostaglandin receptor EP4 suppresses colitis, mucosal damage and CD4 cell activation in the gut. *J Clin Invest.* 2002;109:883–893.
41. Kellermayer R, Balasa A, Zhang W, et al. Epigenetic maturation in colonic mucosa continues beyond infancy in mice. *Hum Mol Genet.* 2010;19:2168–2176.
42. Kiela PR, Laubitz D, Larmonier CB, et al. Changes in mucosal homeostasis predispose NHE3 knockout mice to increased susceptibility to DSS-induced epithelial injury. *Gastroenterology*. 2009;137:965–975, 975e1–975e10.
43. Lagishetty V, Misharin AV, Liu NQ, et al. Vitamin D deficiency in mice impairs colonic antibacterial activity and predisposes to colitis. *Endocrinology*. 2010;151:2423–2432.
44. Lara-Villoslada F, Debras E, Nieto A, et al. Oligosaccharides isolated from goat milk reduce intestinal inflammation in a rat model of dextran sodium sulfate-induced colitis. *Clin Nutr.* 2006;25:477–488.
45. Larrosa M, Yanez-Gascon MJ, Selma MV, et al. Effect of a low dose of dietary resveratrol on colon microbiota, inflammation and tissue damage in a DSS-induced colitis rat model. *J Agric Food Chem.* 2009;57:2211–2220.
46. Lee H, Ahn YT, Lee JH, et al. Evaluation of anti-colitic effect of lactic acid bacteria in mice by cDNA microarray analysis. *Inflammation*. 2009; 32:379–386.
47. Lopez-Dee ZP, Chittur SV, Patel B, et al. Thrombospondin-1 type 1 repeats in a model of inflammatory bowel disease: transcript profile and therapeutic effects. *PLoS One*. 2012;7:e34590.
48. Mannick EE, Cote RL, Schurr JR, et al. Altered phenotype of dextran sulfate sodium colitis in interferon regulatory factor-1 knock-out mice. *J Gastroenterol Hepatol.* 2005;20:371–380.
49. Mizoguchi E. Chitinase 3-like-1 exacerbates intestinal inflammation by enhancing bacterial adhesion and invasion in colonic epithelial cells. *Gastroenterology*. 2006;130:398–411.
50. Mizoguchi E, Xavier RJ, Reinecker HC, et al. Colonic epithelial functional phenotype varies with type and phase of experimental colitis. *Gastroenterology*. 2003;125:148–161.
51. Nakajima A, Wada K, Katayama K, et al. Gene expression profile after peroxisome proliferator activator receptor-gamma ligand administration in dextran sodium sulfate mice. *J Gastroenterol.* 2002;37(suppl 14):62–66.
52. Reikvam DH, Derrien M, Islam R, et al. Epithelial-microbial crosstalk in polymeric Ig receptor deficient mice. *Eur J Immunol.* 2012;42: 2959–2970.
53. Sainathan SK, Bishnupuri KS, Aden K, et al. Toll-like receptor-7 ligand imiquimod induces type I interferon and antimicrobial peptides to ameliorate dextran sodium sulfate-induced acute colitis. *Inflamm Bowel Dis.* 2012;18:955–967.
54. Sainathan SK, Hanna EM, Gong Q, et al. Granulocyte macrophage colony-stimulating factor ameliorates DSS-induced experimental colitis. *Inflamm Bowel Dis.* 2008;14:88–99.
55. Schaible TD, Harris RA, Dowd SE, et al. Maternal methyl-donor supplementation induces prolonged murine offspring colitis susceptibility in

association with mucosal epigenetic and microbiomic changes. *Hum Mol Genet.* 2011;20:1687–1696.

56. te Velde AA, de Kort F, Sterrenburg E, et al. Comparative analysis of colonic gene expression of three experimental colitis models mimicking inflammatory bowel disease. *Inflamm Bowel Dis.* 2007;13:325–330.
57. Barnett MP, McNabb WC, Cookson AL, et al. Changes in colon gene expression associated with increased colon inflammation in interleukin-10 gene-deficient mice inoculated with Enterococcus species. *BMC Immunol.* 2010;11:39.
58. de Buhr MF, Mahler M, Geffers R, et al. Cd14, Gbp1, and Pla2g2a: three major candidate genes for experimental IBD identified by combining QTL and microarray analyses. *Physiol Genomics.* 2006;25:426–434.
59. Edmunds SJ, Roy NC, Davy M, et al. Effects of kiwifruit extracts on colonic gene and protein expression levels in IL-10 gene-deficient mice. *Br J Nutr.* 2012;108:113–129.
60. Hemmerling J, Heller K, Hormannsperger G, et al. Fetal exposure to maternal inflammation does not affect postnatal development of genetically-driven ileitis and colitis. *PLoS One*. 2014;9:e98237.
61. Huang Z, Shi T, Zhou Q, et al. miR-141 regulates colonic leukocytic trafficking by targeting CXCL12beta during murine colitis and human Crohn's disease. *Gut*. 2014;63:1247–1257.
62. Knoch B, Barnett MP, Cooney J, et al. Dietary oleic acid as a control fatty acid for polyunsaturated fatty acid intervention studies: a transcriptomics and proteomics investigation using interleukin-10 gene-deficient mice. *Biotechnol J.* 2010;5:1226–1240.
63. Knoch B, Barnett MP, McNabb WC, et al. Dietary arachidonic acid-mediated effects on colon inflammation using transcriptome analysis. *Mol Nutr Food Res.* 2010;54(suppl 1):S62–S74.
64. Knoch B, Barnett MP, Zhu S, et al. Genome-wide analysis of dietary eicosapentaenoic acid- and oleic acid-induced modulation of colon inflammation in interleukin-10 gene-deficient mice. *J Nutrigenet Nutrigenomics.* 2009;2:9–28.
65. Kuo SM, Chan WC, Hu Z. Wild-type and IL10-null mice have differential colonic epithelial gene expression responses to dietary supplementation with synbiotic Bifidobacterium animalis subspecies lactis and inulin. *J Nutr.* 2014;144:245–251.
66. Reiff C, Delday M, Rucklidge G, et al. Balancing inflammatory, lipid, and xenobiotic signaling pathways by VSL#3, a biotherapeutic agent, in the treatment of inflammatory bowel disease. *Inflamm Bowel Dis.* 2009; 15:1721–1736.
67. Roy N, Barnett M, Knoch B, et al. Nutrigenomics applied to an animal model of inflammatory bowel diseases: transcriptomic analysis of the effects of eicosapentaenoic acid- and arachidonic acid-enriched diets. *Mutat Res.* 2007;622:103–116.
68. Russ AE, Peters JS, McNabb WC, et al. Gene expression changes in the colon epithelium are similar to those of intact colon during late inflammation in interleukin-10 gene deficient mice. *PLoS One*. 2013;8:e63251.
69. Brudzewsky D, Pedersen AE, Claesson MH, et al. Genome-wide gene expression profiling of SCID mice with T-cell-mediated Colitis. *Scand J Immunol.* 2009;69:437–446.
70. Fang K, Zhang S, Glawe J, et al. Temporal genome expression profile analysis during T-cell-mediated colitis: identification of novel targets and pathways. *Inflamm Bowel Dis.* 2012;18:1411–1423.
71. Kristensen NN, Olsen J, Gad M, et al. Genome-wide expression profiling during protection from colitis by regulatory T cells. *Inflamm Bowel Dis.* 2008;14:75–87.
72. Rivollier A, He J, Kole A, et al. Inflammation switches the differentiation program of Ly6Chi monocytes from antiinflammatory macrophages to inflammatory dendritic cells in the colon. *J Exp Med.* 2012;209:139–155.
73. Abad C, Juarranz Y, Martinez C, et al. cDNA array analysis of cytokines, chemokines, and receptors involved in the development of TNBS-induced colitis: homeostatic role of VIP. *Inflamm Bowel Dis.* 2005;11:674–684.
74. Billerey-Larmonier C, Uno JK, Larmonier N, et al. Protective effects of dietary curcumin in mouse model of chemically induced colitis are strain dependent. *Inflamm Bowel Dis.* 2008;14:780–793.
75. Brenna O, Furnes MW, Drozdov I, et al. Relevance of TNBS-colitis in rats: a methodological study with endoscopic, histologic and transcriptomic [corrected] characterization and correlation to IBD. *PLoS One*. 2013;8:e54543.
76. Guzman J, Yu JG, Suntres Z, et al. ADOA3R as a therapeutic target in experimental colitis: proof by validated high-density oligonucleotide microarray analysis. *Inflamm Bowel Dis.* 2006;12:766–789.
77. Kremer B, Mariman R, van Erk M, et al. Temporal colonic gene expression profiling in the recurrent colitis model identifies early and chronic inflammatory processes. *PLoS One*. 2012;7:e50388.
78. Mariman R, Kremer B, van Erk M, et al. Gene expression profiling identifies mechanisms of protection to recurrent trinitrobenzene sulfonic acid colitis mediated by probiotics. *Inflamm Bowel Dis.* 2012;18:1424–1433.
79. Martinez-Augustin O, Merlos M, Zarzuelo A, et al. Disturbances in metabolic, transport and structural genes in experimental colonic inflammation in the rat: a longitudinal genomic analysis. *BMC Genomics.* 2008;9:490.
80. Nur T, Peijnenburg AA, Noteborn HP, et al. DNA microarray technology reveals similar gene expression patterns in rats with vitamin a deficiency and chemically induced colitis. *J Nutr.* 2002;132:2131–2136.
81. Rivera E, Flores I, Appleyard CB. Molecular profiling of a rat model of colitis: validation of known inflammatory genes and identification of novel disease-associated targets. *Inflamm Bowel Dis.* 2006;12:950–966.
82. Wu F, Chakravarti S. Differential expression of inflammatory and fibrogenic genes and their regulation by NF-kappaB inhibition in a mouse model of chronic colitis. *J Immunol.* 2007;179:6988–7000.
83. Yamamoto S, Isuzugawa K, Takahashi Y, et al. Intestinal gene expression in TNBS treated mice using genechip and subtractive cDNA analysis: implications for Crohn's disease. *Biol Pharm Bull.* 2005; 28:2046–2053.
84. Zhou W, Cao Q, Peng Y, et al. FoxO4 inhibits NF-kappaB and protects mice against colonic injury and inflammation. *Gastroenterology*. 2009; 137:1403–1414.
85. Zwiers A, Fuss IJ, Leijen S, et al. Increased expression of the tight junction molecule claudin-18 A1 in both experimental colitis and ulcerative colitis. *Inflamm Bowel Dis.* 2008;14:1652–1659.
86. NC3Rs. The 3Rs [NC3Rs web site]. Available at: http://www.nc3rs.org.uk/the-3rs. Accessed 23 October, 2014.
87. Home Office. *Guidance on the Operation of the Animals (Scientific Procedures) Act 1986*. London, United Kingdom: HMSO; 2014.
88. Clarke M. Standardising outcomes for clinical trials and systematic reviews. *Trials*. 2007;8:39.
89. Couzin-Frankel J. When mice mislead. *Science*. 2013;342:922–925.
90. Check Hayden E. Misleading mouse studies waste medical resources. *Nature News*. 26 March, 2014.
91. Rogers GB, Kozlowska J, Keeble J, et al. Functional divergence in gastrointestinal microbiota in physically-separated genetically identical mice. *Sci Rep.* 2014;4:5437.
92. Obernier JA, Baldwin RL. Establishing an appropriate period of acclimatization following transportation of laboratory animals. *ILAR J.* 2006; 47:364–369.
93. Tremaroli V, Backhed F. Functional interactions between the gut microbiota and host metabolism. *Nature*. 2012;489:242–249.
94. Laying better plans for mice. *Nat Biotechnol.* 2013;31:263.
95. Fish EN. The X-files in immunity: sex-based differences predispose immune responses. *Nat Rev Immunol.* 2008;8:737–744.
96. Karp CL. Unstressing intemperate models: how cold stress undermines mouse modeling. *J Exp Med.* 2012;209:1069–1074.
97. Kokolus KM, Capitano ML, Lee CT, et al. Baseline tumor growth and immune control in laboratory mice are significantly influenced by subthermoneutral housing temperature. *Proc Natl Acad Sci U S A.* 2013;110: 20176–20181.
98. Rabin BS, Lyte M, Epstein LH, et al. Alteration of immune competency by number of mice housed per cage. *Ann N Y Acad Sci.* 1987; 496:492–500.
99. Kitajima S, Takuma S, Morimoto M. Histological analysis of murine colitis induced by dextran sulfate sodium of different molecular weights. *Exp Anim.* 2000;49:9–15.
100. Charan J, Kantharia ND. How to calculate sample size in animal studies? *J Pharmacol Pharmacother.* 2013;4:303–306.
101. Kozlowski C, Jeet S, Beyer J, et al. An entirely automated method to score DSS-induced colitis in mice by digital image analysis of pathology slides. *Dis Model Mech.* 2013;6:855–865.

Chapter Six

A text-mining strategy for assessing the reporting across biomedical research

The content of this chapter was published in the journal *eLife*; full citation:

Flórez-Vargas O, Brass A, Karystianis G, Bramhall M, Stevens R, Cruickshank S, Nenadic G. Bias in the reporting of sex and age in biomedical research on mouse models. *eLife*. 2016; 5:e13615.

Our previous works have shown that fundamental criteria of experimental methods are repeatedly omitted in laboratory models of infectious diseases [see Chapter 3 (Florez-Vargas et al. 2014)] and inflammation [see Chapter 5 (Bramhall et al. 2015)]. The sex and age of the experimental model, for instance, are some of those experimental factors that were poorly reported. In fact, and despite the endorsement by over 300 research journals of the ARRIVE guidelines – Animal Research: Reporting of *In Vivo* Experiments (Kilkenny et al. 2010), the reporting of these two biological variables (which are always available to researchers) is not done properly (Baker et al. 2014). In animal-based biomedical research, both sex and age affect the disease phenotypes; modifying their susceptibility, presentation and response to treatment (Arnold 2010).

This scenario suggests that the '*human factor*' plays a significant role in the reporting of key experimental factors of any biomedical study; where either authors, editors and peer-reviewers overlooked important experimental variables. Therefore, there is a need to develop a framework to guide an automated assessment of the reporting of bio-experiments in submitted biomedical manuscripts; acting as a barrier to prevent incomplete reporting from entering the literature. In this regard, text mining (TM) techniques offer the potential to ease this problem.

TM is a multidisciplinary field that includes others such as natural language processing (NLP) and machine learning (Hotho, Numberger, and Paab 2005). TM aims to assist researchers in analysing the scientific literature through automated processing of text. In the last decade, there has been a significant amount of research in the identification of targeted biomedical information in the scientific literature via TM (Cohen and Hersh 2005, Fleuren and Alkema 2015). In comparison with other TM applications that are focusing on the recognition of complex biomedical entities and their shared relationships, *e.g.*, gene variants associated with drug response (Garten, Coulet, and Altman 2010), our approach addresses a significantly more diverse literature space and questions the reporting of what should be the standard information in biomedical research regarding laboratory animals as models for human diseases.

A TM framework typically involves a number of distinct phases. The first step is referred to as information retrieval: a process to locate relevant textual resources for a given subject of interest and is typically done by querying bibliographic databases with a set of keywords. In our TM approach, we used the PubMed Central Open Access subset for this assignment, which contains over one million full-text articles to date. The second step is referred to as information extraction: a process that recognises of terms (or concepts), *e.g.*, disease names, genes, etc., as well as identify complex relationships between those entities, *e.g.*, disease and drug interaction, while associating them to the subject of interest from unstructured textual data (Hahn et al. 2012). Information extraction can be based on patterns, machine learning techniques, statistical analyses or automated reasoning (Rebholz-Schuhmann, Oellrich, and Hoehndorf 2012). All the information identified through a TM approach has the purpose of providing well targeted data for further analysis and mining and potentially (and ideally) the discovery of new knowledge (Shatkay H, Feldman R. Mining the biomedical literature in the genomic era: an overview (Shatkay and Feldman 2003).

In our TM approach, we designed and implemented generic rule based on the biomedical text lexico-syntactical patterns in order to identify sex and age mentions of the mice studied. The rules were created and applied via the GATE framework (Cunningham et al. 2013). In particular, we used tokenisation in order to capture these experimental factors, *i.e.*, sex and age of the mice. This is a process of breaking a stream of text up into words, but also punctuation, or other elementary linguistically plausible units called tokens (Comeau et al. 2014). Thus, the developed system based on text mining captures these experimental factors by matching tokens in text against lists of commonly used phrases, *e.g.*, ten C57BL/6 *"female"* mice (6-8-*"weeks old"*). This strategy has shown to be effective in the clinical and biomedical domains (Savova et al. 2010, Ramesh et al. 2012).

In this chapter, we presented a proof of concept implementation of a TM system as a survey analysis technique on full text articles for assessing the reporting of sex and age in mouse models across biomedical research.

Bias in the reporting of sex and age in biomedical research on mouse models

Oscar Flórez-Vargas[1], Andy Brass[1*], George Karystianis[2], Michael Bramhall[1], Robert Stevens[1], Sheena Cruickshank[3], Goran Nenadic[2,4]

[1] 1 Bio-health Informatics Group, School of Computer Science, University of Manchester, Manchester, United Kingdom. [2] Text Mining Group, School of Computer Science, University of Manchester, Manchester, United Kingdom. [3] Manchester Immunology Group, Faculty of Life Science, University of Manchester, Manchester, United Kingdom. [4] Manchester Institute of Biotechnology, University of Manchester, Manchester, United Kingdom.

Abstract Lack of accurate method reporting is one of the primary causes of irreproducibility in biomedical research. In animal-based biomedical research, both sex and age affect the disease phenotypes; modifying their susceptibility, presentation and response to treatment. Here we look at these two variables by using text-mining across available full text articles that report investigations where mice were the focus of the study. We found that, although there is an improvement during the last two decades, the lack of reporting of these variables is still a concern; only about 50% of the papers published in 2014 stated these variables. In addition, we observed a sex-bias variability according to the field of study. We hope that this text-mining strategy can be taken as a starting point for future more focused assessment of literature, both in preclinical and clinical studies, and thus impact on the reproducibility of findings and on future study validity.

***For correspondence:** andy.brass@manchester.ac.uk. School of Computer Science, University of Manchester, Kilburn building, Oxford Road, Manchester, M13 9PL

Competing Interests: The authors have no competing interests to declare.

Introduction

Studies using animal models are important tools in experimental biomedical sciences for understanding the physiopathological and therapeutic basis of human diseases. In doing this, the results of preclinical studies carried out in animal models provide not only a rationale for justifying clinical evaluation, but also a source of interpretations of unsuccessful translations during clinical development (*Kimmelman and Anderson, 2012*). Nevertheless, historically, the translation of scientific findings from animal models to humans is far from straightforward. Statistically, more than 80% of potential therapeutics fail in human clinical trials after being successful in animal models (*Perrin, 2014*). This uncomfortable truth is the fundamental reason why animal research is a cause of concern and, therefore, it needs to improve and become more reliable and reproducible (*van der Worp et al., 2010*). In fact, this has led to doubts as to whether experimental models should be considered as a source of knowledge for clinical evaluation (*Perel et al., 2007*).

The failure to translate from experimental models to human beings stems from various factors, where the reproducibility of the findings plays an important role (*Collins and Tabak, 2014; Freedman et al., 2015*). In this way, there is a growing concern over the lack of reproducibility in biomedical studies; a large proportion (75-90%) of the preclinical research findings published in top-ranked journals cannot be replicated (*Begley and Ellis, 2012; Prinz et al., 2011*). The observed lack of reproducibility may be a result, among other things, of the lack of transparency in reporting biomedical research (*Landis et al., 2012; Moher et al., 2008; van der Worp and Macleod, 2011*). In the United States, for instance, it has been estimated that about US $28 billion per year is spent on preclinical research that is not reproducible; where the reporting is one of the most common reasons (*Freedman et al., 2015*).

The Uniform Guidelines of the International Committee of Medical Journal Editors state that authors should include technical information in sufficient detail to allow the experiment to be repeated by other workers (*International Committee of Medical Journal Editors, 2013*). This is vitally important in animal experimentation, where a detailed description of any animal model is not only in agreement with the principles of the 3Rs (Replacement, Reduction and Refinement) (*Burden et al., 2015*), but also plays a fundamental role in the interpretation of the data and reproducibility of the findings derived from the animal model used to generate such data. In this context, the ARRIVE (Animal Research: Reporting In Vivo Experiments) guidelines were developed to improve consistency in reporting animal research (*Kilkenny et al., 2010*). However, there is still a lag in the implementation of these guidelines (*Baker et al., 2014*).

In experiments using animals, for instance, the sex and age of the mice should be reported because they influence the outcomes (*Diedrich et al., 2007; Wizemann and Pardue, 2001*). Both sex and age of organisms are among the variables that affect morphological, physiological, immunological and behavioral parameters and, hence, they are important in reporting both basic science and clinical research. These variables are inextricably linked: it has been proposed that under natural conditions sexual selection has profound effects on the lifespan of organisms (*Bale and Epperson, 2015; Maklakov and Lummaa, 2013*). Considering some taxa exceptions, the general conclusion is that in many animals (including humans), males have shorter lifespans than females (*Clutton-Brock and Isvaran, 2007*). Furthermore, from an evolutionary standpoint, these sex differences in lifespan depends to a great extent on sexually dimorphic life-history strategies (*Maklakov and Lummaa, 2013*), *e.g.* mating systems, and on genetic architecture; including both the sex chromosomes (*Nguyen and Disteche, 2006*) and the mitochondrial DNA (*Gemmell et al., 2004*).

Regarding preclinical and clinical studies, sex and age play key roles in disease phenotypes; modifying their susceptibility, presentation and response to treatment (*Arnold, 2010*). Some pathologies exhibit a clear sexual dimorphism (*Ober et al., 2008*). Using stroke as an example, it is known that its incidence is higher in men than women during their lifespan (*Mozaffarian et al., 2015*). However, recent evidence suggests that after the age of 60 years and thus post-menopause, women have more severe strokes than men (*Dehlendorff et al., 2015*). In the case of animal models, sex- and age-dependent differences in protein expression profiles were observed in the heart proteome of female and male C57BL/6 mice of two distinct age groups (14 and 100 weeks) (*Diedrich et al., 2007*). This evidence implies that sex differences must be studied across the entire lifespan in order to bring new insights into the pathogenesis of the diseases and identify targets for new drugs for both sexes and different times of life. Guidelines, such as ARRIVE (*Kilkenny et al., 2010*), have been developed because of the 3Rs (*Burden et al., 2015*) to highlight the importance of such biological factors in animal experiments, and these have been endorsed by journals with the aim of improving the reporting of bioscience research.

In this study, we have used large scale text mining to evaluate the reporting of information about mouse sex and age as “bibliomarkers” of method reporting quality in a set of over 15 thousand full-text articles. In the last decade, there has been a significant amount of research in the identification of targeted biomedical information in the scientific literature via text-mining (TM) (*Cohen and Hersh, 2005; Fleuren and Alkema, 2015*). In particular, efforts have been made to recognize protein and gene names in text (*Settles, 2005*) or other biomedical entities of interest such as electronic health records (*Meystre et al., 2008*). In comparison with other TM applications that are focusing on the recognition of complex biomedical entities and their shared relationships, our approach addresses a significantly more diverse literature space and questions the reporting of what should be standard information in biomedical research regarding laboratory animals as models for human diseases. Based on syntactic rules and simple dictionary matching, we extracted key characteristics in mouse-based models such as sex and age in order to comprehend the standards of information reporting to assess the possibility of reproducing mouse experiments. Previous work has shown that fundamental criteria of experimental methods are repeatedly omitted in laboratory models (*Bramhall et al., 2015; Florez-Vargas et al., 2014*). In light of this, our investigation looked at sex and age as two important factors across available full text articles that report investigations where mice were the focus of the study. We investigate questions of whether sex and age of mice is reported, the use of each sex in different types of research area, and the field of analysis for each area.

Results

System evaluation and data

We evaluated the TM system on a set of 50 full-text articles randomly selected from our corpus of study (Supplementary file 1) by comparing its performance with the manual annotations of the same papers performed by two biomedical experts. The F-scores that resulted from this evaluation were around 92% for both sex and age (Table 1), which indicates good quality of the results (*Ananiadou et al., 2006*).

Table 1. Evaluation of the performance of the text mining system

Characteristics	True-positives	True-negatives	False-positives	False-negatives	Precision (%)	Recall (%)	F-score (%)
Sex	29	16	3	2	90.6	93.5	92.0
Age	31	14	1	4	96.8	88.5	92.4

A total of 50 articles were used as the data set to evaluate the performance of the text mining system (Supplementary file 2D). The precision (P), calculated as TP/(TP+FP), determines the accuracy of the system in recognizing desirable terms. The recall (R), calculated as TP/(TP+FN), produces the coverage of the system. F-score is the harmonic mean of precision and recall and it is calculated as 2*P*R/(P+R).

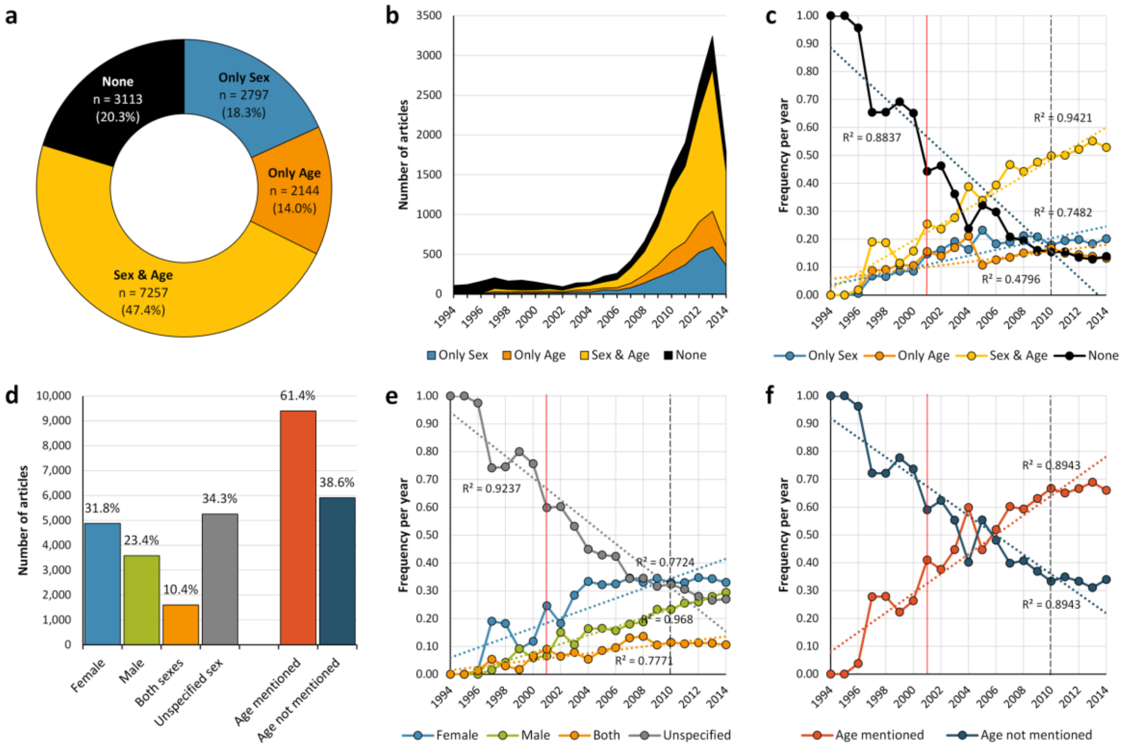


Figure 1. General distribution and historical change of reporting and non-reporting of sex and/or age in mouse-model experiments. Pie-chart (**a**) showing an overview of the reporting and non-reporting (none) of sex only, age, or both sex and age in a set of 15,311 studies published between 1994 and 2014 by stating the number and percentage of articles in each portion. The chronological change of the reporting and non-reporting is displayed both in a stacked area plot (**b**) and a scatter plot after normalization [per articles/year] (**c**). The chronological changes show that most of the articles assessed were published during the last decade (**b**), and that the improvement of reporting of these two biological factors started before, and not after, the US Institute of Medicine report in 2001 (*Wizemann and Pardue, 2001*) [indicated with a vertical red line] or the introduction of ARRIVE guideline (*Kilkenny et al., 2010*) [indicated with a vertical black dashed line] (**c**). Bar-chart (**d**) showing the number and percentage of articles reporting/not reporting of sex by sex [females only, males only, or both sexes either by mixing or separating them] or age. The chronological change of the reporting and non-reporting of sex by sex (**e**), and age (**f**), is displayed in scatter plots after normalization [per articles/year].

A total of 15,311 full-text articles from the PubMed Central Open Access subset as of February 2015 were processed in this study. These articles correspond to 7.15% and 27.85% of mouse experimentation articles retrieved by the same query in PubMed and PubMed Central, respectively. This corpus of documents were published between 1994 and 2014, of which 50.1% were published after 2011 (n= 7671) (Figure 1) Seventy journals out of the 628 analyzed covered 30 or more articles of the corpus (Figure 1-figure supplement 1), which corresponds to 81.05% of papers retrieved. *PLOS ONE* contained the highest number of articles (n= 5574, 36.41%), followed by *The Journal of Experimental Medicine* (n= 931, 6.08%), and *The Journal of Cell Biology* (n= 363, 2.37%).

Reporting of sex and age

The general and historical reporting of sex and age as experimental variables in mouse models is presented in Figure 1. Overall, from 1994 to 2014, about a fifth of papers did not report either the sex or the age of the mouse used in the study (Figure 1a and 1b). Figure 1c shows that the

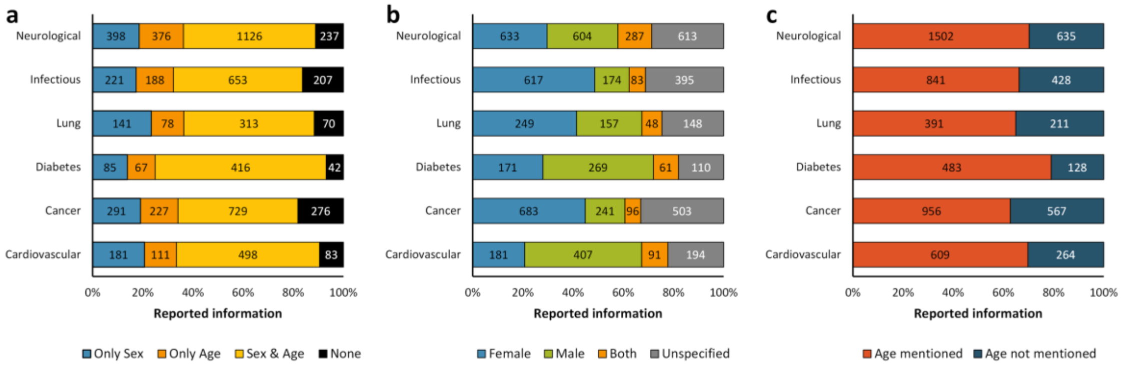


Figure 2. Distribution of reporting of the sex and age in mouse model of a group of diseases. The reporting of these variables was assessed for six groups of diseases from the top 10 causes of death according to the W.H.O. This analysis was performed in the set of 14,225 articles published from 2001, when the US Institute of Medicine report was published (*Wizemann and Pardue, 2001*) and when the non-reporting of sex and age together dropped about 50% – avoiding misinterpretations [**Figure 1c**], to 2014. The distribution is presented in stacked bar charts that illustrate the percentage of the reporting and non-reporting for both biological variables overall (**a**) and discriminated by variable: sex (**b**) and age (**c**); stating the number of articles corresponding to each percentage inside the stacks. A two-way ANOVA without replication was performed to assess the difference in reporting of the sex [p = 0.005] and age [p = 0.028] for each disease, indicating that the reporting and non-reporting of these biological factors varies across these

frequency of articles reporting sex and/or age in mice models has increased steadily during the last two decades, whereas missing information about these two experimental variables showed an important drop from 100% (no papers reported the sex and age of the mice in 1994 and 1995) to about 15% following a slope of approximately -0.045. Nevertheless, since 2012, the percentage of articles reporting both factors had reached only about 50% of the papers published in those years.

When the sex of the mouse model is stated in the article, experiments performed with female mice were more frequently reported than experiments performed with male mice (31.84% vs. 23.38%, Binomial test $p< 0.001$; 95% IC: 56.60 – 58.71) (Figure 1d). Our results showed that, historically, female mice have been reported more often than male mice, reaching a plateau of about 33% since the last decade (2004 – 2014) (Figure 1e). In addition, the use of both sexes in mice experiments stratified by sex showed the lowest improvement over time (Figure 1e); with a maximum of about 10% of the articles since 2006. Reporting of mouse age improved steadily from 1999 to 2006 (Figure 1f), at which point age is reported more than 50% of the time; since 2010 age reporting has plateaued, with between 65 and 70% of articles each year mentioning the age of mice.

In order to identify whether there are general features common on reporting sex and age as experimental variables to any biomedical field, we assessed six main preclinical research topics as defined by their impact on human health (*WHO, 2014*), including: cardiovascular diseases; cancer; diabetes mellitus; lung diseases; infectious diseases; and neurological disorders. A two-way ANOVA without replication was performed to assess the difference in reporting sex and age for each field. Our results showed statistically significant differences, *i.e.* $p < 0.05$, indicating that the reporting of these experimental factors varies across biomedical fields (Figure 2). In identifying the sex and age of the mouse, for instance, studies on diabetes showed the highest frequency (68%), whereas studies on cancer showed the lowest frequency (48%) (Figure 2a). Studies on cancer reported the worst results regarding missing information about sex (33%) or age (37%) of

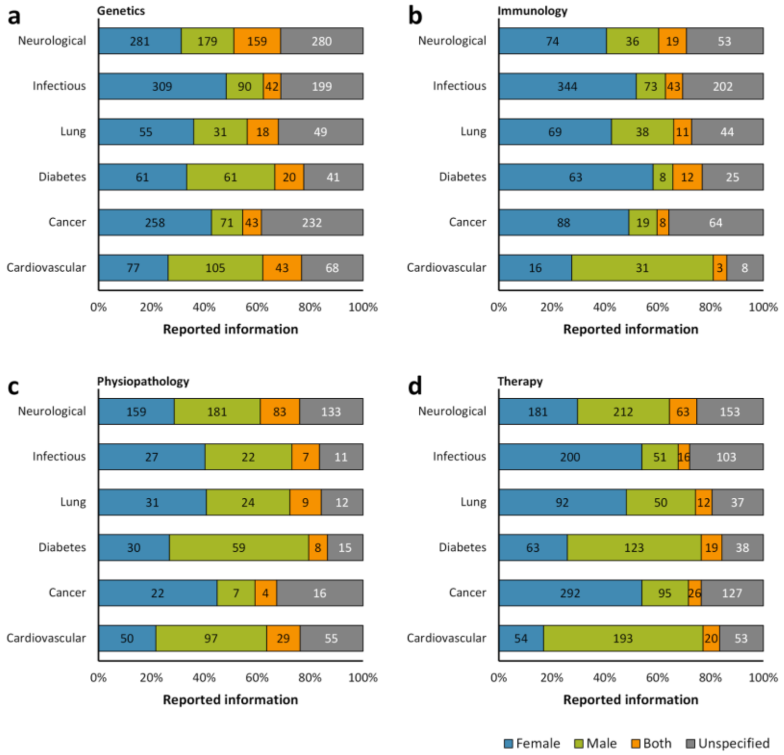


Figure 3. Distribution of reporting of the sex in mouse model of a group of diseases by research approach. The reporting of sex was assessed for each disease by the topic of research whether genetics (**a**), immunology (**b**), physiopathology (**c**), or therapy (**d**). This analysis was performed in the set of 14,225 articles published from 2001, when the US Institute of Medicine report[16] was published (*Wizemann and Pardue, 2001*) [**Figure 1c**], to 2014. The distribution is presented in stacked bar charts that illustrate the percentage of the reporting and non-reporting for the sex; stating the number of articles corresponding to each percentage inside the stacks. A two-way ANOVA without replication was performed to assess the difference in reporting of the sex for genetics [p = 0.0009], immunology [p = 0.0074], physiopathology [p < 0.0001], and therapy [p = 0.1165], indicating that the reporting and non-reporting of these biological factors varies across most of these biomedical approaches.

the mice used (Figure 2b and 2c). Overall, the best results in reporting sex and age were achieved by the studies on neurological disorders (Figure 2a, 2b and 2c).

For a more detailed analysis of sex-based reporting, the six groups of diseases were divided into four subgroups according to the characterization of the disease models via genetics, immunology, physiopathology and therapy. Our results suggest that there is a preference for studying the immunology of these diseases by using female mouse models, whereas there is a tendency to use male mouse models for studying their genetic basis (Figure 3a and 3b). Both in physiopathology and in therapy subgroups, male mice were more frequently studied in models of cardiovascular diseases, diabetes and neurological disorders, and female mice in models of cancer, lung diseases and infectious diseases (Figure 3c and 3d).

In order to further test whether the observations about the reporting of sex in the experimental mouse models were conserved even in specific cases, we focused the analysis on one particular disease per group as follows: myocardial ischemia (cardiovascular disease); diabetes mellitus type 2 (diabetes); chronic obstructive pulmonary disease (lung disease); Alzheimer's (neurological disorder). Three diseases were included in the case of infectious diseases that are among the most frequently reported causes of death world-wide (*WHO, 2014*), *i.e.* tuberculosis, HIV and malaria. Melanoma was included for the cancer group since it is a highly aggressive and notoriously

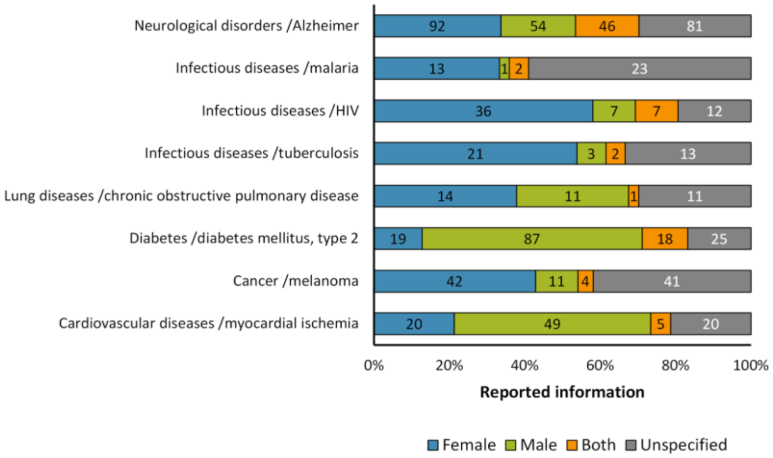


Figure 4. Distribution of reporting of the sex in mouse model of diseases. The graph shows the reporting in particular diseases. All these diseases that are among the most frequently reported causes of death world-wide or commonly used models. The distribution is presented in stacked bar charts that illustrate the percentage of the reporting and non-reporting for the sex; stating the number of articles corresponding to each percentage inside the stacks. This analysis was performed in a set of 791 articles; see Figure 1–source data 1.

chemoresistant form of cancer; making it a widely used tumor model (*Herlyn and Fukunaga-Kalabis, 2010*). Overall, our results suggest that in most cases there is a similar pattern of reporting as that found for the biomedical fields assessed to which these diseases belong (Figure 4).

Bibliometric parameters were used to determine if they were associated with the quality of method reporting. We used as journal metrics both the journal impact factor from the Institute for Scientific Information (ISI) Web of Knowledge's Journal Citation Report (2014), and h-index from the SCImago Journal and Country Rank (2014). No correlation was observed between the reporting of sex or age as experimental variables and the journal impact factor and h-index of the 70 journals that covered 30 or more articles of the corpus (Figure 5).

Discussion

By applying TM as a survey analysis technique on full texts of all available articles in the PubMed Central Open Access subset as of February 2015 we evaluated over 15 thousand papers that used the mouse as an animal model for the study of human biology. Therefore, this analysis constitutes the largest analysis of the quality of mouse experiment reporting, providing the strongest evidence about sex and age bias through biomedical research to date. Nevertheless, this analysis does not represent the entire biomedical literature; not all journals are found in the PubMed Central Open Access subset and some of the journals that deposit their complete contents into PubMed Central include some of their articles in the Open Access subset. This is undoubtedly a limitation of our study. For this survey, we have selected the mouse as a model because of all animal models the mouse is probably the most comprehensive and well-characterized model in life sciences. Researchers rely on mouse models to mimic human disease conditions for several reasons. One of the main reasons is that mouse and human genomes are genetically similar – about 90% of human genes have direct orthologues with mice (*Yue et al., 2014*). Moreover, as animal models, mice are convenient due to their small size, short lifespan (up to two years), and quick generation time; three weeks for gestation and from 6 to 8 weeks to reach sexual maturity.

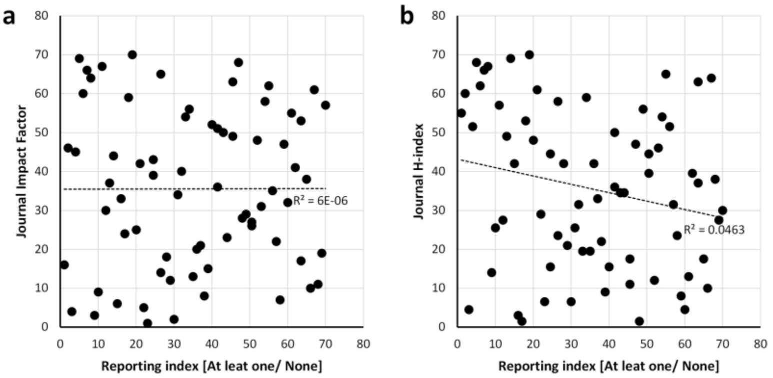


Figure 5. Scatter plots showing the relationship between the reporting and the bibliometric indices. Journal impact factor in which the papers were published (**a**) and h-index of journals (**b**). Spearman's rank correlation coefficient r square is shown alongside the regression lines. The scatter plots show that there is no correlation between the reporting and impact factor [r =0.002, p = 0.984] data from the Journal Citation Report (year 2014) and journal h-index [r =-0.215, p = 0.073] data from the SCImago Journal and Country Rank (year 2014). Analysis conducted on the 70 journals that published 30 or more articles of the 15,311 studies returned by searching the PubMed Central Open Access subset as of February 2015.

Therefore, they can be easily housed and maintained, can be genetically manipulated to define gene function in a whole body system and a large number of mice can be studied in a relatively short period of time. This, for instance, allows scientists to study cell/cell interactions in the tissue environment and thus cause and effect relationships in a controlled situation.

Despite the implications for interpretation and reproducibility of experimental findings, the sex and age of the experimental subjects are often not recorded in scientific reports (*Kilkenny et al., 2009*). In agreement with previous reports, the evidence presented in this study showed that the lack of reporting of key methodological parameters in mouse experiments is still a cause of concern; only about half of the papers published in 2014 stated both sex and age of the mice as experimental variables (Figure 1c). The reason why these variables are not described is unclear, since this simple information is always available to researchers. We do not believe it is a space issue, because in about 40 characters of text it is possible to describe them, including mice number and

Table 2. Summary of the data set used in this study.

Sets of articles	Number of articles	Task	File
Data 1	15,311	Corpus for assessing reporting of the sex and age of the mice	Supplementary file 1*
Data 2	40	Creating the text-mining rules	Supplementary file 2A
Data 3	40	Manual inspection for finding the location of the mention of the sex and age of the mice	Supplementary file 2B
Data 4	70	Enhancing the performance of the text-mining rules	Supplementary file 2C
Data 5	50	Evaluating the text-mining system	Supplementary file 2D

*Supplementary file 1 also contains data sets of the six groups of diseases analyzed (cardiovascular diseases; cancer; diabetes mellitus; lung diseases; infectious diseases; and neurological disorders), as well as of the different approaches to assessing the disease models (i.e. genetics, immunology, physiopathology and therapy), and the disease example for each of the six disease groups.

mouse strain, *e.g.* ten C57BL/6 female mice (6-8-weeks old). Whilst an improvement in the reporting of mouse sex and age has been observed over time, this is not solely attributable to the introduction of journal guidelines, as improvements were present prior to ARRIVE publication in 2010 (*Kilkenny et al., 2010*). In fact, a follow-up study in 2012 showed that while sex and age reporting had improved post-ARRIVE, journals that enforced the ARRIVE guidelines as a condition of publication still failed to publish sex and age in all cases (*Baker et al., 2014*). The observed improvements may therefore be a result of a growing recognition of the importance of sex and age as experimental factors that may affect study outcomes, resulting in a movement towards better reported experiments despite, not because of, the introduction of stricter journal guidelines.

An analysis of the scientific literature leads to the general conclusion that the males in both human and other animals are studied much more than their female counterparts. This conclusion is based mainly on the results of two studies that manually surveyed a set of biomedical articles (*Beery and Zucker, 2011; Taylor et al., 2011*). However, our results showed otherwise in mouse-based models: 31.84% and 23.38% of all papers assessed were on studies performed on female and male mice, respectively (Figure 1d). This could be explained by some practical advantages of using female rather than male mice: they are cheaper; less aggressive to each other and to experimenters; and they are smaller, requiring less weight-administered drug. In addition, the apparent contradiction between this observation and the previous reports might be related to the sample size and study design; our sample size was the largest to date and we surveyed a much broader range of disciplines. In addition, we focused the survey on mouse models, whereas many more species were included in the other reports (*Beery and Zucker, 2011; Taylor et al., 2011*), *e.g.* cat, dog, monkey etc. Nevertheless, although in both studies about 50% (*Taylor et al., 2011*) and 80% (*Beery and Zucker, 2011*) of documents relied on rodent models, *i.e.* mouse and rat, information regarding sex bias by species was not assessed; making comparison with our results difficult. Knowing the sex bias for each particular experimental model is fundamental in the era of decision-making towards reproducible science, which will optimize the design of future studies to fulfil the gap of information regarding sex differences in the model under study.

In preclinical studies, furthermore, we noted an important sex- and age-bias in mouse-based disease models (Figure 2b and 2c). Among the main preclinical research topics assessed, we observed the strongest male-bias in cardiovascular disease models (2.25:1) and the strongest female-bias in infectious disease models (3.54:1) (Figure 2b). This situation still persists: between 2012 and 2014, about 70% and 77% of research articles assessed on these two disease models are still biased towards male and female mice, respectively. These pathologies and many others, exhibit important sexual dimorphisms, which are not only inherent to genetic differences, but also to hormonal influence (*Case et al., 2013; Gilks et al., 2014*). For example, in the study of hypertension, one of the major risk factors for cardiovascular disease, a greater increase in blood pressure was reported in gonad-intact XY males than XX females using the four core genotype in the MF1 mouse model. However, the mean arterial pressure was greater in gonadectomized XX mice compared with XY mice regardless of whether the mice were born with testes or with ovaries (*Ji et al., 2010*). On the other hand, in the case of infectious diseases, females have a more robust immune system than males – both the innate and adaptive immune responses, which makes them less susceptible to developing many infections (mainly Th1-type infections), although it increases the risk of developing autoimmune diseases due to their trend to develop a stronger pro-inflammatory response (*Pennell et al., 2012*). Interestingly, we also observed that the sex-bias could change in a particular disease mouse model according to the biomedical study. Diabetes disease mouse models exemplified this situation. From a global point of view, this disease was found to be male-biased (1.57:1) (Figure 2b). However, in studies related with the immunology of diabetes, there was a strong female-bias (7.87:1) (Figure 3b); a change that remains in the study of diabetes mellitus type 2 (Figure 4).

In order to balance sex of animals and cells in preclinical studies, the National Institutes of Health (NIH) have proposed a multi-dimensional initiative, which includes, among other things, extramural training on experimental design and data analysis by sex (*Clayton and Collins, 2014*). Regarding this initiative, new ideas have been proposed to achieve, and sustain, the sex balance in biomedical research (*McCullough et al., 2014*). In this context, our study provides an implementation of TM to assess reporting of experimental factors. By knowing where there is imbalance for a particular variable, it is possible to address it in a cost-effective manner. This not only directly contributes to the comparability of experimental work, but also to the reproducibility of findings. To address this problem some journals are already introducing editorial measures and methods checklists in order to improve the quality of scientific reporting (*Nature, 2013*). Nevertheless, whilst journal checklists may make reference to species strain, sex and age of animals, most of these checklists focus on statistical analysis to ensure repeatability, which could lead to a biased analysis if it is not made based on biological factors that modify the outcomes. In addition, by checking with the laboratory that conducted the experiment in question it is possible to fix some reproducibility problems; implying a need to adopt more-uniform standards within particular fields. Toward the same direction, we hope that our TM strategy can be taken as a starting point for future more focused assessment of literature; targeting a wider array of characteristics in preclinical and clinical studies. Its potential implementation would enable a straightforward pathway when it comes to reporting key information involved in preclinical and clinical research – *e.g.* by entering it into the publication cycle as a pre-screening test for submitted manuscripts, which will have an important positive impact on several fronts of the biomedical domain, including the reproducibility of experimental findings and the accuracy of meta-analysis.

Methods

Search strategy and data

A literature search was carried out in Medline via PubMed in order to identify research articles that deal with mouse experimentation. The database was searched in March 2015 for articles that were published between 1st January, 1994 and 31st December, 2014 using the terms as they appear in Figure 1–source data 1. To ensure maximum specificity in the search, searching was limited to articles where the MeSH (Medical Subject Headings) "Mouse" term indicated the major focus of the article; moreover the keywords "Mouse" or "Mice" had to be stated in the title. This also prevented articles that made only passing references to mouse work from entering the dataset and ensured a high quality corpus for analysis. The search was restricted to English language papers and to research articles (excluding review articles). In addition, to obtain full text articles, we restricted the PubMed search to include only those in PubMed Central by adding the special term "pubmed pmc[sb]" in the query. The PubMed Identifiers (PMID) were then converted to the respective PubMed Central (PMC) reference numbers which were acquired by querying the PubMed Central Open Access subset as of February 2015, which contains over one million full-text articles to date.

In order to assess particular areas in which there is strong scientific interest world-wide, we analyzed experiments performed in mouse models for six groups of diseases from the top 10 causes of death according to the W.H.O. in high, low and middle income countries (*WHO, 2014*).

The six disease groups were as follows: cardiovascular diseases; cancer; diabetes mellitus; lung diseases; infectious diseases; and neurological disorders. Some causes of death did not apply for our study, *e.g.* road injury. HIV/AIDS, tuberculosis and other infections, for instance, were included in the infectious diseases group. A group for cancer was created in a similar way. An example disease for each of the six disease groups was also included. In addition, as there are different approaches to assessing disease models according to the research field, *e.g.* immunology, genetics etc., each of these areas were divided into a series of subgroups by using the Subheading MeSH terms "genetics", "immunology", "physiopathology" and "therapy" (Figure 1–source data 1). These four approaches were chosen because of their importance for understanding the molecular and physiological basis of diseases, as well as for developing novel therapeutic agents for their treatment. These subjects were used to find if these disease models are being assessed consistently by sex and age.

In 2001 the US Institute of Medicine report (*Wizemann and Pardue, 2001*) concluded that sex matters in diseases and response to therapy; we therefore decided to explore any changes before and after the report by selecting articles between 1994 and 2014. This time span allows us to assess the impact of this report on the reporting of this experimental factor. In order to avoid misinterpretation due to low number of papers prior to 2001, the analysis for groups and subgroups was applied to articles published after 1st January 2001.

Sex and Age identification: data sets

The TM approach involved the design and implementation of generic rule-based patterns, which identify age and sex mentions in text. The rules were based on lexical patterns engineered from a sample of 40 full-text articles manually selected from PubMed through a thematic query of interest as follows: "Mice"[Mesh] AND (mouse[ti] OR mice[ti]) AND "animals"[MeSH Terms:noexp] AND Journal Article[ptyp] AND English[lang]. The first 40 papers that mentioned the sex and/or age of the mice were selected (Supplementary file 2A).

The *age* rules were based on lexical patterns mentioning age clues, *e.g.* "*aged 3 to 8 weeks old*". Similarly, the *sex* rules were designed around word matching aiming to identify male, female or both sexes in mice, *e.g.* "*mice of either sex were used*".

The rules were created and applied via GATE (*Cunningham et al., 2013*) for Windows version 8.1; an open source free software enabling the design and implementation of information extraction systems in unstructured text with the crafted rules following its notation (https://gate.ac.uk/) . The number of crafted rules was 12 for sex and 18 for age. Figure 1–source data 2 presents examples of rules for both the sex and age whereas Supplementary file 3 displays all the utilized rules for the two characteristics.

The generated TM results were then integrated at the document level. In cases where several different candidate mentions for a single characteristic, *i.e.* sex or age, are recognized in a given document, we 'unified' them to get document level annotations using the following approach: if multiple mentions of different lengths occur, the longest is selected (usually the most informative) aiming to have one mention for both the sex and age per document, and where mentions are of the same length, the first one is chosen.

Since our method focuses on the recognition of age and sex at the mention level per document, we hypothesize that it is highly unlikely for researchers to report key information about animal models that they did not use. In order to further support this hypothesis, 40 full-text articles were randomly selected from our corpus and through manual inspection, we concluded that indeed, if there are mentions in text (particularly in the Method section) of specific age and sex (together) these are attributed to the mice used in the animal experiments and no further mentions were reported (Supplementary file 2B). The randomness was modelled by using the "=RANDBETWEEN()" function in Microsoft Office Excel for Windows version 2013 as follows: according to the TM results, each paper of the corpus of articles with a positive mention of the sex and/or age of the mice was assign a random number from 1 to 40. The first 40 papers identified with the random number 1 were selected.

Finally, to further enhance the performance of the rules, we applied this strategy to a development set of 70 full-text documents (Supplementary file 3C). These articles were randomly selected from our corpus by using the "=RANDBETWEEN()" function in Microsoft Office Excel for Windows version 2013; assigning to each paper a random number from 1 to 5. After sorting by the "Year" column, the first five papers identified with the random number 1 were selected by each year group. The mentions of age and sex in both corpus were manually identified and reviewed by the first author, who has a background in the field of biomedical research. A summary of the data sets used in this study is presented in Table 2.

System evaluation

The TM system's performance was evaluated at the document level by considering whether the returned mentions were correctly the sex and age of the mice studied. In order to create an evaluation dataset, 50 full-text articles were randomly selected from our corpus of study (Supplementary file 3D) and were manually double-annotated for both the age and the sex by the first and fourth authors due to their biomedical expertise. There was no disagreement between the manual annotations performed by two biomedical experts. The randomness was modelled by using the "=RANDBETWEEN()" function in Microsoft Office Excel for Windows version 2013 as follows: a random number from 1 to 50 was assigned to each paper. The first 50 papers identified with the random number 1 were selected.

Precision (P), Recall (R) and F-score were calculated for both the age and the sex using the standard metrics (*Ananiadou et al., 2006; Hotho et al., 2005*), which rely on the number of true- and false-positive (TP and FP), and true- and false-negative (TN and FN) cases. The precision (P), calculated as TP/(TP+FP), determines the accuracy of the system in recognizing desirable terms. The recall (R), calculated as TP/(TP+FN), produces the coverage of the system. Often, there is an inverse relationship between precision and recall; when an increase occurs in precision, a simultaneous decrease is observed in recall and vice versa. Therefore, the F-score was also used for evaluating the performance of information extraction systems due to its harmonic mean of precision and recall and it is calculated as 2*P*R/(P+R). Table 1 shows the results of the evaluation set at the document level.

Despite the overall positive performance of our TM system, there were some results that lead to false-positive and false-negative results due to the relatively complex expressions. False-negative results regarding age mentions occurred because the rules are based on syntactical patterns that require a numeric range between specific time units, *i.e.*, days, weeks and months. For example, in the sentence *"Nineteen animals, including males and females, of ages from postnatal day (P) 7 to several months were deeply anesthetized by isoflurane and decapitated"* (*Arbogast et al., 2013*), age is not mentioned as a range concept of days (or weeks or months) but as *"postnatal days to several months"* without indicating the exact number of months. Cases like this suggest that an extension of the current rule set could lead to an improvement towards the system's performance. False-negative results regarding sex mentions occurred because the rules for the sex recognition is rather straightforward with a simple dictionary matching (minimal), which, as a consequence, does not enable the identification of the sex through inference, *e.g.* when sex-specific proxy elements are mentioned, such as pregnancy. For example, in the sentence *"Primary mouse mammary epithelial (PMME) cells were isolated from 15-d timed-pregnant CD-1 mice"* (*Lin et al., 1995*) are expected to be missed since the sex of the mice used in this experiment is female and is being inferred by the word *"pregnant"*.

On the other hand, the application of a dictionary approach generated interestingly few false-positives in the sex recognition. This is because the system identified words like *"male"* or *"female"* early in text, whereas in the actual experiment the scientists did not report any specific sex for the selected model. For example, in the sentence *"The colony of animals carrying the Pak1ip1mray allele is maintained by crossing male carriers with FVB/NJ females. All embryos presented in the phenotypic analysis of this study were produced from carriers crossed for at least four generations onto an FVB/NJ background"* (*Ross et al., 2013*), the sex of the embryos was not established even though the findings relied on them. Other cases were: "Epithelial cells were derived from tracheas of 3-weeks old Gprc5a mice" and "by peritoneal into 8–12 weeks old C56Bl/6 mice". Cases like these suggest that the implementation of a more sophisticated system that could target common syntactical patterns observed in text (similar to those for the characteristic of age) will contribute to an improvement of the precision and performance of the system. This could explain why sex had the lower precision (90.6%) of the two analyzed factors (Table 1). On the contrary, there was only one false positive (referring to the embryonic stage of the mice) although the real age could not be recognized directly due to not being explicitly expressed; *"Genomic DNA and pooled total RNAs were isolated from CRL2196 cells and from various tissues, ages and lineages of mice as indicated, using standard methods and Trizol (Invitrogen), respectively"* (*Li et al., 2014*). The more refined rules led to an increased precision of 96.8% (Table 1).

Although our TM protocol does produce reliable results, the returned results are merely an indication of how TM can be used to improve issues such as the under-reporting of key information in mouse based studies. There is room to improve the applied TM strategy. Crafting more flexible rules for the capture of age and including more specific ones for the recognition of sex could improve the generated results and reveal a clearer picture of the reporting of these variables in the biomedical field. While the variety of the observed common lexical patterns was not wide in the training and development sets (Supplementary files 2A and 2C), a larger set could reveal other patterns that could help increase the recall. Nevertheless, the F-measure of 92% (Table 1) gives enough confidence in using this automated method to assess the incidence of reporting sex and age in biomedical articles.

Statistical analysis

The frequencies of reporting of sex and age by articles were determined in Microsoft Office Excel 2013 for Windows. Differences in reporting of sex and age of mice in multiple models of diseases, as well as the use of each sex by the topic of research for each disease were assessed by two-way ANOVA without replication. An index of the reporting for each journal was calculated by dividing the number of articles that report the sex and/or age of the mouse by the number of articles that do not report any of these biological variables. Spearman's rank correlations were calculated between the reporting index and impact factor from the Journal Citation Report, and h-index journal from the SCImago Journal and Country Rank. All statistical analysis was performed by using the GraphPad Prism software for Windows version 6.05, La Jolla CA, (www.graphpad.com). Graphical representation of the data was performed using Microsoft Office Excel for Windows version 2013.

Acknowledgement

The authors want to thank Dr Geraint Duck for his valuable suggestions into making this method automatic.

Additional information

Funding: This research is supported by a "Francisco José de Caldas" scholarship from Colciencias to O. Flórez-Vargas and a co-funding scholarship from EPSRC and Epistem Ltd awarded to M. Bramhall. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions: O.F.-V., A.B., R.S., S.C. and G.N. conceived and designed the study. O.F.-V., G.K. and M.B. implemented the system. O.F.-V. and G.K. performed the experiments, analyzed the data, and drafted the manuscript. A.B., R.S., S.C. and G.N. supervised the project. All authors have read and approved the final manuscript.

Additional information

Figure 1–figure supplement 1. Reporting of sex or age in mouse-model experiments by journal.

Figure 1–source data 1. PubMed search terms used for each disease group and their approaches.

Figure 1–source data 2. Example rules for identification of sex and age.

Supplementary file 1. Corpus for assessing reporting of the sex and age of the mice.

Supplementary file 2A. Set of articles for creating the text-mining rules.

Supplementary file 2B. Set of articles for finding the location of the mention of the sex and age of the mice.

Supplementary file 2C. Set of articles for enhancing the performance of the text-mining rules.

Supplementary file 2D. Set of articles for evaluating the text-mining system.

Supplementary file 3. Rules used to identify the sex and age of experimental mouse models.

References

Ananiadou, S., D.B. Kell, and J. Tsujii. 2006. Text mining and its potential applications in systems biology. *Trends Biotechnol* 24:571-579. doi: 10.1016/j.tibtech.2006.10.002.

Arbogast, P., M. Glosmann, and L. Peichl. 2013. Retinal cone photoreceptors of the deer mouse Peromyscus maniculatus: development, topography, opsin expression and spectral tuning. *PLoS One* 8:e80910. 10.1371/journal.pone.0080910.

Arnold, A.P. 2010. Promoting the understanding of sex differences to enhance equity and excellence in biomedical science. *Biology of sex differences* 1:1. doi: 10.1186/2042-6410-1-1.

Baker, D., K. Lidster, A. Sottomayor, and S. Amor. 2014. Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS Biol* 12:e1001756. 10.1371/journal.pbio.1001756.

Bale, T.L., and C.N. Epperson. 2015. Sex differences and stress across the lifespan. *Nat Neurosci* 18:1413-1420. doi: 10.1038/nn.4112.

Beery, A.K., and I. Zucker. 2011. Sex bias in neuroscience and biomedical research. *Neuroscience and biobehavioral reviews* 35:565-572. doi: 10.1016/j.neubiorev.2010.07.002.

Begley, C.G., and L.M. Ellis. 2012. Drug development: Raise standards for preclinical cancer research. *Nature* 483:531-533. 10.1038/483531a483531a.

Bramhall, M., O. Florez-Vargas, R. Stevens, A. Brass, and S. Cruickshank. 2015. Quality of methods reporting in animal models of colitis. *Inflamm Bowel Dis* 21:1248-1259. doi: 10.1097/MIB.0000000000000369.

Burden, N., K. Chapman, F. Sewell, and V. Robinson. 2015. Pioneering better science through the 3Rs: an introduction to the national centre for the replacement, refinement, and reduction of animals in research (NC3Rs). *J Am Assoc Lab Anim Sci* 54:198-208.

Case, L.K., E.H. Wall, J.A. Dragon, N. Saligrama, D.N. Krementsov, M. Moussawi, J.F. Zachary, S.A. Huber, E.P. Blankenhorn, and C. Teuscher. 2013. The Y chromosome as a regulatory element shaping Immune cell transcriptomes and susceptibility to autoimmune disease. *Genome Research* 23:1474-1485. doi: 10.1101/gr.156703.113.

Clayton, J.A., and F.S. Collins. 2014. Policy: NIH to balance sex in cell and animal studies. *Nature* 509:282-283. doi: 10.1038/509282a.

Clutton-Brock, T.H., and K. Isvaran. 2007. Sex differences in ageing in natural populations of vertebrates. *Proc Biol Sci* 274:3097-3104. doi: 10.1098/rspb.2007.1138.

Cohen, M.A., and R.W. Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in Bioinformatics* 6:57-71. doi: 10.1093/bib/6.1.57.

Collins, F.S., and L.A. Tabak. 2014. Policy: NIH plans to enhance reproducibility. *Nature* 505:612-613.

Cunningham, H., V. Tablan, A. Roberts, and K. Bontcheva. 2013. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Comput Biol* 9:e1002854. doi: 10.1371/journal.pcbi.1002854.

Dehlendorff, C., K.K. Andersen, and T.S. Olsen. 2015. Sex Disparities in Stroke: Women Have More Severe Strokes but Better Survival Than Men. *J Am Heart Assoc* 4:doi: 10.1161/JAHA.115.001967.

Diedrich, M., J. Tadic, L. Mao, M.A. Wacker, G. Nebrich, R. Hetzer, V. Regitz-Zagrosek, and J. Klose. 2007. Heart protein expression related to age and sex in mice and humans. *Int J Mol Med* 20:865-874. doi: 10.3892/ijmm.20.6.865.

Fleuren, W.W., and W. Alkema. 2015. Application of text mining in the biomedical domain. *Methods* 74:97-106. doi: 10.1016/j.ymeth.2015.01.015.

Florez-Vargas, O., M. Bramhall, H. Noyes, S. Cruickshank, R. Stevens, and A. Brass. 2014. The quality of methods reporting in parasitology experiments. *PLoS One* 9:e101131. doi: 10.1371/journal.pone.0101131.

Freedman, L.P., I.M. Cockburn, and T.S. Simcoe. 2015. The Economics of Reproducibility in Preclinical Research. *PLoS Biol* 13:e1002165. 10.1371/journal.pbio.1002165.

Gemmell, N.J., V.J. Metcalf, and F.W. Allendorf. 2004. Mother's curse: the effect of mtDNA on individual fitness and population viability. *Trends Ecol Evol* 19:238-244. doi: 10.1016/j.tree.2004.02.002.

Gilks, W.P., J.K. Abbott, and E.H. Morrow. 2014. Sex differences in disease genetics: evidence, evolution, and detection. *Trends Genet* 30:453-463. doi: 10.1016/j.tig.2014.08.006.

Herlyn, M., and M. Fukunaga-Kalabis. 2010. What is a good model for melanoma? *J Invest Dermatol* 130:911-912. doi: 10.1038/jid.2009.441.

Hotho, A., A. Numberger, and G. Paab. 2005. A Brief Survey of Text Mining. *LDV Forum – GLDV Journal for Computational Linguistics and Language Technology* 20:19–62.

International Committee of Medical Journal Editors. 2013. Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publications. In.

Ji, H., W. Zheng, X. Wu, J. Liu, C.M. Ecelbarger, R. Watkins, A.P. Arnold, and K. Sandberg. 2010. Sex chromosome effects unmasked in angiotensin II-induced hypertension. *Hypertension* 55:1275-1282. doi: 10.1161/HYPERTENSIONAHA.109.144949.

Kilkenny, C., W.J. Browne, I.C. Cuthill, M. Emerson, and D.G. Altman. 2010. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol* 8:e1000412. 10.1371/journal.pbio.1000412.

Kilkenny, C., N. Parsons, E. Kadyszewski, M.F.W. Festing, I.C. Cuthill, D. Fry, J. Hutton, and D.G. Altman. 2009. Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals. *Plos One* 4:doi: 10.1371/journal.pone.0007824.

Kimmelman, J., and J.A. Anderson. 2012. Should preclinical studies be registered? *Nat Biotechnol* 30:488-489. doi: 10.1038/nbt.2261.

Landis, S.C., S.G. Amara, K. Asadullah, C.P. Austin, R. Blumenstein, E.W. Bradley, R.G. Crystal, R.B. Darnell, R.J. Ferrante, H. Fillit, R. Finkelstein, M. Fisher, H.E. Gendelman, R.M. Golub, J.L. Goudreau, R.A. Gross, A.K. Gubitz, S.E. Hesterlee, D.W. Howells, J. Huguenard, K. Kelner, W. Koroshetz, D. Krainc, S.E. Lazic, M.S. Levine, M.R. Macleod, J.M. McCall, R.T. Moxley, K. Narasimhan, L.J. Noble, S. Perrin, J.D. Porter, O. Steward, E. Unger, U. Utz, and S.D. Silberberg. 2012. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* 490:187-191. 10.1038/nature11556.

Li, J., M. Kannan, A.L. Trivett, H. Liao, X. Wu, K. Akagi, and D.E. Symer. 2014. An antisense promoter in mouse L1 retrotransposon open reading frame-1 initiates expression of diverse fusion transcripts and limits retrotransposition. *Nucleic Acids Res* 42:4546-4562. 10.1093/nar/gku091.

Lin, C.Q., P.J. Dempsey, R.J. Coffey, and M.J. Bissell. 1995. Extracellular matrix regulates whey acidic protein gene expression by suppression of TGF-alpha in mouse mammary epithelial cells: studies in culture and in transgenic mice. *J Cell Biol* 129:1115-1126.

Maklakov, A.A., and V. Lummaa. 2013. Evolution of sex differences in lifespan and aging: Causes and constraints. *Bioessays* 35:717-724. doi: 10.1002/bies.201300021.

McCullough, L.D., G.J. de Vries, V.M. Miller, J.B. Becker, K. Sandberg, and M.M. McCarthy. 2014. NIH initiative to balance sex of animals in preclinical studies: generative questions to guide policy, implementation, and metrics. *Biology of sex differences* 5:15. doi: 10.1186/s13293-014-0015-5.

Meystre, M.S., K.G. Savova, C.K. Kipper-Schuler, and F.J. Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Methods of Information in Medicine* 47:128-144.

Moher, D., I. Simera, K.F. Schulz, J. Hoey, and D.G. Altman. 2008. Helping editors, peer reviewers and authors improve the clarity, completeness and transparency of reporting health research. *Bmc Med* 6:10.1186/1741-7015-6-13.

Mozaffarian, D., E.J. Benjamin, A.S. Go, D.K. Arnett, M.J. Blaha, M. Cushman, S. de Ferranti, J.P. Despres, H.J. Fullerton, V.J. Howard, M.D. Huffman, S.E. Judd, B.M. Kissela, D.T. Lackland, J.H. Lichtman, L.D. Lisabeth, S. Liu, R.H. Mackey, D.B. Matchar, D.K. McGuire, E.R. Mohler, 3rd, C.S. Moy, P. Muntner, M.E. Mussolino, K. Nasir, R.W. Neumar, G. Nichol, L. Palaniappan, D.K. Pandey, M.J. Reeves, C.J. Rodriguez, P.D. Sorlie, J. Stein, A. Towfighi, T.N. Turan, S.S. Virani, J.Z. Willey, D. Woo, R.W. Yeh, M.B. Turner, C. American Heart Association Statistics, and S. Stroke Statistics. 2015. Heart disease and stroke statistics--2015 update: a report from the American Heart Association. *Circulation* 131:e29-322. doi: 10.1161/CIR.0000000000000152.

Nature. 2013. Reporting Checklist For Life Sciences Articles. *http://www.nature.com/authors/policies/checklist.pdf*

Nguyen, D.K., and C.M. Disteche. 2006. Dosage compensation of the active X chromosome in mammals. *Nature Genetics* 38:47-53. doi: 10.1038/ng1705.

Ober, C., D.A. Loisel, and Y. Gilad. 2008. Sex-specific genetic architecture of human disease. *Nat Rev Genet* 9:911-922. doi: 10.1038/nrg2415.

Pennell, L.M., C.L. Galligan, and E.N. Fish. 2012. Sex affects immunity. *J Autoimmun* 38:J282-291. doi: 10.1016/j.jaut.2011.11.013.

Perel, P., I. Roberts, E. Sena, P. Wheble, C. Briscoe, P. Sandercock, M. Macleod, L.E. Mignini, P. Jayaram, and K.S. Khan. 2007. Comparison of treatment effects between animal experiments and clinical trials: systematic review. *BMJ* 334:197. doi: 10.1136/bmj.39048.407928.BE.

Perrin, S. 2014. Preclinical research: Make mouse studies work. *Nature* 507:423-425. doi: 10.1038/507423a.

Prinz, F., T. Schlange, and K. Asadullah. 2011. Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov* 10:712. doi: 10.1038/nrd3439-c1.

Ross, A.P., M.A. Mansilla, Y. Choe, S. Helminski, R. Sturm, R.L. Maute, S.R. May, K.K. Hozyasz, P. Wojcicki, A. Mostowska, B. Davidson, I.E. Adamopoulos, S.J. Pleasure, J.C. Murray, and K.S. Zarbalis. 2013. A mutation in mouse Pak1ip1 causes orofacial clefting while human PAK1IP1 maps to 6p24 translocation breaking points associated with orofacial clefting. *PLoS One* 8:e69333. 10.1371/journal.pone.0069333.

Settles, B. 2005. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* 21:3191-3192. doi: 10.1093/bioinformatics/bti475.

Taylor, K.E., C. Vallejo-Giraldo, N.S. Schaible, R. Zakeri, and V.M. Miller. 2011. Reporting of sex as a variable in cardiovascular studies using cultured cells. *Biology of sex differences* 2:11. doi: 10.1186/2042-6410-2-11.

van der Worp, H.B., D.W. Howells, E.S. Sena, M.J. Porritt, S. Rewell, V. O'Collins, and M.R. Macleod. 2010. Can animal models of disease reliably inform human studies? *PLoS Med* 7:e1000245. doi: 10.1371/journal.pmed.1000245.

van der Worp, H.B., and M.R. Macleod. 2011. Preclinical studies of human disease: Time to take methodological quality seriously. *Journal of Molecular and Cellular Cardiology* 51:449-450. 10.1016/j.yjmcc.2011.04.008.

WHO. 2014. The top 10 causes of death. *[Internet] (accessed on 15 Jun 2015 available from: http://www.who.int/mediacentre/factsheets/fs310/en/)*

Wizemann, T.M., and M.L. Pardue. 2001. Exploring the Biological Contributions to Human Health: Does Sex Matter? Board on Health Sciences Policy, Institute of Medicine, Washington (DC).

Yue, F., Y. Cheng, A. Breschi, J. Vierstra, W. Wu, T. Ryba, R. Sandstrom, Z. Ma, C. Davis, B.D. Pope, Y. Shen, D.D. Pervouchine, S. Djebali, R.E. Thurman, R. Kaul, E. Rynes, A. Kirilusha, G.K. Marinov, B.A. Williams, D. Trout, H. Amrhein, K. Fisher-Aylor, I. Antoshechkin, G. DeSalvo, L.H. See, M. Fastuca, J. Drenkow, C. Zaleski, A. Dobin, P. Prieto, J. Lagarde, G. Bussotti, A. Tanzer, O. Denas, K. Li, M.A. Bender, M. Zhang, R. Byron, M.T. Groudine, D. McCleary, L. Pham, Z. Ye, S. Kuan, L. Edsall, Y.C. Wu, M.D. Rasmussen, M.S. Bansal, M. Kellis, C.A. Keller, C.S. Morrissey, T. Mishra, D. Jain, N. Dogan, R.S. Harris, P. Cayting, T. Kawli, A.P. Boyle, G. Euskirchen, A. Kundaje, S. Lin, Y. Lin, C. Jansen, V.S. Malladi, M.S. Cline, D.T. Erickson, V.M. Kirkup, K. Learned, C.A. Sloan, K.R. Rosenbloom, B. Lacerda de Sousa, K. Beal, M. Pignatelli, P. Flicek, J. Lian, T. Kahveci, D. Lee, W.J. Kent, M. Ramalho Santos, J. Herrero, C. Notredame, A. Johnson, S. Vong, K. Lee, D. Bates, F. Neri, M. Diegel, T. Canfield, P.J. Sabo, M.S. Wilken, T.A. Reh, E. Giste, A. Shafer, T. Kutyavin, E. Haugen, D. Dunn, A.P. Reynolds, S. Neph, R. Humbert, R.S. Hansen, M. De Bruijn, L. Selleri, A. Rudensky, S. Josefowicz, R. Samstein, E.E. Eichler, S.H. Orkin, D. Levasseur, T. Papayannopoulou, K.H. Chang, A. Skoultchi, S. Gosh, C. Disteche, P. Treuting, Y. Wang, M.J. Weiss, G.A. Blobel, X. Cao, S. Zhong, T. Wang, P.J. Good, R.F. Lowdon, L.B. Adams, X.Q. Zhou, M.J. Pazin, E.A. Feingold, B. Wold, J. Taylor, A. Mortazavi, S.M. Weissman, J.A. Stamatoyannopoulos, M.P. Snyder, R. Guigo, T.R. Gingeras, D.M. Gilbert, R.C. Hardison, M.A. Beer, B. Ren, and E.C. Mouse. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* 515:355-364. doi: 10.1038/nature13992.

Chapter Seven

Summary & Discussion

7.1 An overview of scientific and scholarly contributions of this work

Basic and preclinical research is fundamental because it provides the base on which future studies are built. These two research fields constitute the so-called biomedical framework, which is a fast-growing interdisciplinary area of science that involves the investigation of biological processes and their translation for human benefit. However, the observed lack of reproducibility in this research may be a result, among other things (*e.g.*, such as misconduct), of the lack of transparency in reporting biomedical research (van der Worp and Macleod 2011, Landis et al. 2012). In particular, the technical nature of the scientific method plays a vital role in achieving this transparency, since it provides the basis for drawing conclusions (van der Worp and Macleod 2011, Landis et al. 2012).

The work presented in this Thesis is an attempt to create strategies to support the biomedical research community in assessing the reporting of experimental research in order to improve the reproducibility of its findings. The core of this work is based upon the notion of the checklist as an effective method in surmounting failure in reporting (Gawande 2010). Checklists make explicit the minimum expected information that should be supplied for reproducing a scientific work. Considering that most research stakeholders come to the conclusion, that either the use of checklists is insufficiently required by publishing journals, or that the implementation of checklists in final articles is unsatisfactory (Baker et al. 2014, Fuller et al. 2015, GBSI 2013, Shamseer et al. 2012), we developed a framework for improving the use of checklists as powerful management tools to aid decision making, *i.e.*, by helping scientists in verifying their judgments regarding scientific reports – spotting missing and erroneous information about a particular experiment. This framework was developed as a spreadsheet-based tool considering as this format seems adequate for carrying out this task since it can handle both text and numbers (Juluru and Eng 2015). In fact, spreadsheets not only have been successfully used for describing experiments¸ *e.g.*, RightField (Wolstencroft et al. 2011), ISA Software (Rocca-Serra et al. 2010) and MAGE-TAB (Rayner et al. 2006), but also for developing decision support tools due to their ability to calculate information (Bujkiewicz et al. 2011, Shakespeare et al. 2006). In addition, by using text mining technology it will be possible to automate the reporting checking process with regards to aspects of method reporting.

The strategies created are listed as follows:

- miniRECH – a spreadsheet-based tool for assessing manually the quality of scientific reporting via a checklist. The miniRECH model developed in this work was designed to operate in Microsoft® Excel since MS Excel is routinely used by the biomedical community.
- An automatic text-mining system as a survey technique for automatically assessing the quality of the scientific literature by targeting key experimental characteristics in biomedical research.
- The usefulness of these strategies is demonstrated by several case studies, which have been published in scientific journals.

Checklists can help to ensure transparency and consistency in reporting data and metadata from scientific studies, which enhances the comprehensiveness of the scientific evidence and the reproducibility of its findings (GBSI 2013). There is evidence that endorsement of some checklists as MIAME, ARRIVE and CONSORT by journals increases the completeness of reporting even if reporting remains suboptimal (Baker et al. 2014, Witwer 2013, Turner et al. 2012). For example, in *Nature* journals the incidence of reporting of animal characteristics that influence experimental outcomes such as sex and age increased only by twofold (~80%) two years after the endorsement of the ARRIVE guidelines (Baker et al. 2014). In addition, our experimental evidence showed that only about 50% of the papers published in 2014 reported these two variables; where approximately 80% of papers assessed were published in journals that endorsed ARRIVE guidelines and/or stated the reporting of sex and age in the author guidelines (Florez-Vargas et al. 2016).

However, for some prospective authors, journal requirements for providing a relevant checklist can feel like yet another hurdle along the journey to publication. In addition, there are a wider range of barriers and factors (*e.g.*, professional culture, journal or regulatory agencies' policies, inability to find reporting guidelines, etc.) that influence the likelihood of using reporting guidelines by both authors and editors (Fuller et al. 2015). Nevertheless, in response to these matter, the start-up company Penelope has developed an automatic tool to assess reporting information stated in checklists (Penelope 2016). Moreover, there are also other tools for helping researchers in writing a randomized trial report (the COBWEB – Consort-based WEB tool) (Barnes et al. 2015) and systematic reviews – Review Manager (RevMan) (Cochrane Schizophrenia Group 2016).

Checklists can play a number of roles – in publication they can support paper writing, paper refereeing, in the experimental process they can be used to ensure relevant detail is collected when the experiment is run or to support effective metadata capture for data repositories. How checklists are implemented and the supporting tooling will be a function of the purpose to which they are put.

Supporting authors – checklists early in the publication process: a checklist supports authors where and when it should be used by them – at the time of manuscript writing. Examples come from specific domains. In the medical clinical trial literature, for instance, the CONSORT guidelines have been developed (Moher, Hopewell, et al. 2010, Schulz et al. 2010). Although the evidence suggests that journal endorsement of CONSORT may benefit the completeness of reporting of randomised clinical trials they publish, the completeness of reporting of trials remains sub-optimal (Turner et al. 2012). In this context was developed the COBWEB – Consort-based WEB tool (Barnes et al. 2015). By guiding the authors through a series of questions based on the CONSORT guidelines via a formatted Word document, the tool ensures that a paper's first draft includes many of the key requirements for reporting trials (Barnes et al. 2015). When this tool was tested in a randomised trial of 41 students tasked with writing the methods section of trials based on real trial protocols, the method sections from those who used the tool were more completely reported than those who didn't used the tool (Barnes et al. 2015). This use case is well-supported by simple text documents.

Supporting publishers – checklists at the point of paper submission or after, to look at quality: an example of these is the ARRIVE guidelines (Kilkenny et al. 2010). However the thesis has shown that although the guidelines are there – they do not lead to compliance. In this context, the the QUAlity and Transparency Of health Research (EQUATOR) network [http://www.equator-network.org] is working with the start-up company Penelope to develop a web tool that aims to help authors identify relevant reporting guidelines more intuitively (Penelope 2016). This tool screens manuscripts for common reporting errors and helps researchers improve their work before submitting it to a journal; this includes highlighting potentially relevant checklists but goes further by identifying other commonly missed or incompletely reported pieces of information that are required for publication of a research article, such as citations, tables, and ethics statements, and by even scrutinising p-values.

In terms of assessing paper quality, supporting publication – an MS Excel strategy works well. In this way, we have presented evidence that miniRECH has an important impact on the decision-making process; by helping both experts and non-experts in verifying their judgments, and non-experts in producing judgments that approximate to the ones given by experts. In addition, it is hoped that the implementation of our text mining strategy can be used by authors and publishers; by ensuring that manuscripts adhere to subject specific reporting checklists. It not only could help authors make their work more valuable, but also vastly reducing processing times by peer-reviewers and editors and, as a consequence, costs for publishers.

Supporting experimentation: MIAME (Minimum Information About a Microarray Experiment) represents a good example of this. The MIAME guidelines outline the minimum information that should be included when describing a microarray experiment to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment (Brazma et al. 2001). This guideline is being widely accepted by the scientific community. In fact, supply MIAME compliant data is encouraged by the two major public databases of microarray data: the Gene Expression Omnibus (GEO) database at the National Center for Biotechnology Information (NCBI) (Edgar and Barrett 2006) and ArrayExpress at the European Bioinformatics Institute (EBI) (Rustici et al. 2013). In order to help scientists produce MIAME-compliant descriptions of their experiments MicroArray and Gene Expression – MAGE-TAB was developed (Rayner et al. 2006). This is a simple tab-delimited, spreadsheet-based format, which can be used for annotating and communicating microarray data in a MIAME compliant fashion.

Considering that there are many repositories which lack metadata information regarding the data they contain. Checklists have also the potential to add as rules for database loading. An effort has been taken by applying the ARRIVE guidelines to the International Mouse Phenotyping Consortium (IMPC) (Karp et al. 2015), where a detailed explanation of the experiments played an important role during the implementation process of applying checklists to databases. More recent work has highlighted the potential of capturing the data automatically using workflows – supported with ontologies (Maccagnan et al. 2010). Workflows allow the description and the orchestration of complex processes. In this context, checklists could be used to build standard workflows for describing laboratory protocols which, in turn, could be extended by the scientists according to their experimental design. By adopting standard workflows based on checklists, it should be possible to guarantee that the minimum information about a particular experimental methodology is considered. A starting point for adapting checklists to laboratory protocols could be made by using the COW (Combining Ontologies with Workflows) software tool (Maccagnan et al. 2010). This approach could not only deal with the complexity of the experimental process via workflows, but also with the formalisation of data and metadata via ontologies.

To summarise, there are many ways in which checklists can be used and many technologies by which they can be delivered. What does this mean for miniRECH? It is there to support good practice for authors, and for expert and non-experts to assess quality. MS Excel meets this need. It is not currently focussed on data loading in repositories – so it is not appropriate for the database or workflow options. However, a miniRECH model could be adapted in order to meet these purposes.

Regarding the temporal nature of checklists, it is worth mentioning that checklists capture the state of biological knowledge at a particular time. That knowledge changes. In inflammation, for example, it is now known that time of day (Haspel et al. 2014) and microbiota (Curtis et al. 2015) are important – that wasn't the case 5 or 10 years ago. If checklists are to stay current they will need to adapt to the new knowledge. As a consequence we will need to version control the checklists – and state the versions used for any application. This isn't specific to checklists – it is true of all knowledge resources.

Checklists come with costs. They need community involvement to create. They are another hurdle an author has to negotiate. They add to the refereeing costs. To be useful the benefit of using them will have to outweigh this cost. This process is made rather more complex because of the multiple stakeholders involved. For example the cost of applying the checklist falls on the author – the benefit might be seen as accruing to wider community. A very recent paper has shown that current scientific writing practice is encouraging bad practice 'by accident': *"The natural selection of bad science"* (Smaldino and Mcelreath 2016). In this model of competing science laboratories the rewards flow to those that publish often, not those that publish accurately, *i.e.*, "the most powerful incentives in contemporary science actively encourage, reward and propagate poor research methods and abuse of statistical procedures".

Considering the prior context, checklists have been taken up widely in areas where the cost of making a mistake is high. Thus, checklists made their way into medicine from industry, where they have been used for quality and safety assurance of processes and products, especially those carrying high risk (Gawande 2010). The WHO Surgical Safety Checklist is the most globally relevant example of a medical checklist (Haynes et al. 2009), which has been demonstrated a consistent effect on the reduction of postoperative complications and mortality (Bergs et al. 2014) – which represents an important fact in the cost-benefit of using checklists in the clinical arena. This carries across into areas of clinical trials research, where guidelines are used more widely. The reputational and patient risk of a bad study can be very high, *e.g.*, the Andrew Wakefield scandal around the safety of vaccines (Rao and Andrade 2011). For checklists to be used more widely we therefore need to consider two aspects. The first is making them "cheaper" to use. That is where we hope tools like miniRECH can make a difference. The second is to change the community attitude to the importance of accurate reporting and to change the incentives that propagate poor methods. That is outside the scope of what this thesis has considered.

Considering the above scenario, there is an urgent need for a scientific cultural change towards the use of checklists as an effective part of the strategy to improve science. This change should be based not only on improving the engagement to use checklists, but also to develop new ones – especially domain-specific checklists. However, this is a difficult challenge. Drivers can come from research funders (in terms of data policies), publishers (such as MIAME), regulators (as in medicine) or a perceived need in the community (such as the maths guidelines (IM2C 2016)). Most successful projects have used a mix of these. Nonetheless, reporting checklists are now starting to be used as an adjunct in developing educational courses in the design and conduct of health research (Moher, Schulz, et al. 2010).

While there is no single best or correct approach for developing reporting checklists, they need to be developed using robust and widely accepted methodologies if reporting checklists are to be useful and more widely disseminated. In this way, it has been proposed 18 steps have been proposed to facilitate the reporting checklists development in health research (*e.g.*, identify the need for a checklist; conduct a Delphi exercise; generate a list of items for consideration at a face-to-face meeting; etc.) (Moher, Schulz, et al. 2010). What an "optimal" checklist would like is then a function of its purpose – whether supporting paper writing, refereeing or data repository creation. As described previously, each of these use cases could be supported by different technologies – from simple text documents through to automated workflows, and as a function of whether the user is a general scientist, an expert or indeed an automated system. However, in all cases it would be expected that the principle of “minimal information” should hold. The intention should be to make the cost of using the checklist as low as possible.

In general terms, there is no information about the costs required to develop a reporting checklist. Even if funding is available, most developers limit their fiscal requests to cover only the main reporting guideline meeting (Moher, Schulz, et al. 2010). Nonetheless, the benefits of checklists on quality and safety assurance in medicine and industry – as previously mentioned – can also be relevant into the scientific arena by improving the completeness, accuracy and clarity of published research and thus reduce wasted effort and enhance reproducibility. This is particularly important in the light of the state of scientific research today, where there is an urgent need to be productive to have a successful career which, as a consequence, may increase the reporting of false findings and reduce the room for testing the work of others (Oxenham 2016). At present the funding for this kind of work is coming in terms of developer and user time, time that is being volunteered. For such a “volunteer” model to be viable the benefits must be widely accepted by the community.

A checklist is useful if it improves the quality of reporting. To achieve this it needs to exist, to be used, to be useful, and to be shown to have had an impact. The first is easy to assess. The second – usage – can be monitored from tracking citations to the standard in the literature – there are papers being published that clearly state they have used the checklist as part of their standard method. The third – effectiveness – is rather harder. There are a number of studies which clearly show that the use of checklists can improve the quality of paper writing, there are fewer studies which have assessed the use of checklist for some of the other tasks in which they have been applied. The final criteria – the impact of a checklist – is much harder to study. In this thesis we have used a text-mining strategy to show that reporting of sex and age have improved – but there is still much that needs to be done in terms of techniques for checking compliance. Measuring the real impact of checklists is clearly an important area for future study and would make a good potential follow on project.

There are many checklists available. Currently, the most comprehensive sources of information about reporting guidelines in medicine and biomedicine such as the EQUATOR network [http://www.equator-network.org] and BioSharing [http://biosharing.org], list several different reporting checklists. We have only looked at a small number of checklists in a small number of potential application areas. Although the work has demonstrated the potential usefulness of the checklists in assessing paper quality and supporting better reporting, there are other applications of checklists we have not explored – particularly around there more automated use. For a checklist to be useful we need a clear understanding of the cost/benefit balance that must be struck. This means that in many cases – where the cost of generating the checklist is high and the benefits of using it are perceived to be limited, this would not be the technology of choice for improving standards. Checklists have the best chance of flourishing when the cost of developing and using them is outweighed by cost of inadequate reporting. It is not clear yet that we have a good understanding of where the balance between these lies – and so it isn't completely clear to know which areas of science would most benefit from a checklist strategy. However, what we can do is continue to develop methodologies – such as miniRECH – to reduce the costs of using the technology. If we could better understand the drivers within science which lead to poor reporting (Smaldino and Mcelreath 2016) we might also be able to increase the perceived cost to researchers of poor reporting in general – thereby generating a more robust, accurate and repeatable scientific literature. Reporting guidelines can be also used by peer reviewers and editors to strengthen manuscript review, as well as research funders that can benefit from introducing reporting guidelines into the research application system. Therefore, there are enormous potential benefits of good reporting.

7.2 Limitations of this work

The miniRECH strategy, it is hoped, would ease the way of using checklists by providing a framework in which authors, editors and peer-reviewers could identify at glance key missing information in the manuscripts prior to publication – particularly considering the complexity of modern experiments: usually involving the combination of different technologies. In this way, the peer-review process requires multidisciplinary efforts for scrutinising their methods and findings. Nevertheless, optimisation of our miniRECH framework should be done in order to improve the engagement of this strategy. In addition, considering the human factor as a constraint to use checklists for improving reporting (Fuller et al. 2015), our text mining approach offers an alternative in this regard as it is an automatic method that could be implemented by journals in order to assess the reporting information in the submitted manuscripts. Nevertheless, it should be extended to a wider range of experimental characteristics in order to guarantee that all information requested by a checklist is assessed during the text mining process.

Our assessing strategies and their optimisations represent an important starting point for improving the reporting of scientific information toward enhancing reproducibility. While scientists may not be interested in using these strategies in their daily research practice, our strategies could be, at least in part, a way to highlight the importance of increasing the quality of reporting toward improving reproducibility. It is expected, for example, that the rate of retraction by human errors can be reduced which, in turn, could improve the likelihood that the experimental findings can be reproduced.

Although this Thesis successfully provides two strategies for assessing the reporting of the scientific biomedical literature, there are at least three inherent limitations of the current work.

- Other scientific literature search engines could improve the accuracy of our assessment findings.

While there are other search engines available, for the purpose of this Thesis, we decided to examine the literature on biomedicine by using PubMed – Public/Publisher MEDLINE – as it is considered to be the most widely used search engine in biomedical literature. In addition, for the literature assessment via text-mining we searched in the PubMed Central Open Access subset, which contains over one million full-text articles to date. Therefore, any conclusions that we draw about the current state of the quality of reporting in biomedical experiments do not represent the entire biomedical literature. Nevertheless, our findings do suggest that an appreciable proportion of scientific journal articles have issues with the reporting of experimental details to allow reproducibility of the scientific work, as well as to have a rigorous description of the scientific model for integrating data and metadata in the context of the experimental evidence.

- The adaptation of miniRECH to different operating systems can improve the usability of this application tool.

The miniRECH tool was designed to operate in Microsoft® Excel, since MS Excel is used widely by the biomedical community. However, due to the differences between platforms, *e.g.*, Excel for Mac and Excel for Windows, the Excel VBA script written on Windows does not work properly on Macintosh. Therefore, there is a need to adapt the miniRECH codes to work on several operating systems for allowing users to have a wider opportunity to use the miniRECH tool. Some strategies that could be carried out in order to solve this issue are discussed in the next section.

- The implementation of a more sophisticated system that could target common syntactical patterns observed in text and the extension of the current rule set could lead to an improvement of the text mining system.

Although our text-mining method does produce reliable results, the returned results are merely an indication of how text mining can be used to improve issues such as the under-reporting of key information in animal based studies.

In particular, despite the overall positive performance of our system, there were some cases that lead to false-positive and false-negative results due to the lack of rule flexibility. For example, in the sentence *"Nineteen animals, including males and females, of ages from postnatal day (P) 7 to several months were deeply anesthetized by isoflurane and decapitated"*, age is mentioned as a range concept of not days (or weeks or months) without indicating the exact numbers of each age or sex. This case demonstrates an example of a false-negative result due to the fact that the rules used are based on syntactical patterns that require a numeric range between specific time units like days, weeks and months.

On the other hand, the application of a dictionary approach generated interestingly few false-positives in the sex recognition. This is because the system identified words like male or female early in the text, whereas in the actual experiment the scientists did not report any specific sex for the selected model. For example, in the sentence *"The colony of animals carrying the Pak1ip1mray allele is maintained by crossing male carriers with FVB/NJ females. All embryos presented in the phenotypic analysis of this study were* *produced from carriers crossed for at least four generations onto an FVB/NJ background"*, the sex of the embryos was not established even though the findings relied on them.

Crafting more rules for the capture of other experimental information requested by checklists could improve the generated results and reveal a clearer picture of the reporting of the experimental variables in the biomedical field. Nonetheless, this text mining strategy can be extended to other scientific fields which lack comprehensive information about the quality of reporting information, e.g., mathematics, engineering, etc. The outcome of its application across the scientific published knowledge would let us know a clear picture of the scientific fields that need more attention and, as a consequence, develop strategies for improving reporting in specific scientific communities.

7.3 Future work

There are several directions we hope to take this work in the future, including but not limited to the following:

- A checklist generator for specific scientific manuscripts.

Several checklists have been created by the minimum information guidelines group, MIBBI (Taylor et al. 2008). However, despite the mandatory use of checklists as a requirement when submitting an article to biomedical scientific journals, the completeness and accuracy of reporting remains suboptimal (Baker et al. 2014, Witwer 2013). This could be due to the complexity of experimental design, often involving different technologies and models, which would require more than one checklist to describe the study. This, in turn, might impact on the likelihood of using reporting guidelines by authors and editors (Fuller et al. 2015). Therefore, it would be of great benefit to create a repository of interoperable checklists that allows scientists to generate a single checklist that includes a list of all the factors and attributes that should be reported when considering the publication of their experimental findings. This repository could be fed from initiatives such as the BioSharing catalogue [http://biosharing.org]. The checklists included in this catalogue have been developed by experts in particular fields, and have evolved over time to capture only the most essential considerations about biomedical experiments.

- Domain-specific checklists.

Following the previous rationale – where a study involves multiple technologies and models, the scientific community should also be encouraged to focus attention on development of domain-specific checklists, *i.e.*, as those we have created here. Domain-specific checklists have the ability to catch more specific information with regard to the experimental complexity. For example, checklists such as (MIAME) (Brazma et al. 2001) and (MIAPE) (Taylor et al. 2007) – just to mention some of them, were developed for describing genomic and proteomic experiments. Nevertheless, they do not catch all the information on the experimental context under which a genomic or proteomic experiment was carried out, *e.g.*, the infection process or the monitoring of colitis development, which were included in our checklists. Therefore, the development, implementation and use of domain-specific checklists will help to improve the reporting of multiple methods in an experimental design context. This is especially important for use and re-use of information from several sources.

- Optimisation of the miniRECH interface.

The quality of the interaction between software and its users is an important factor to be considered to ensure its usability. Considering our miniRECH framework, this might improve the likelihood of using reporting checklists by authors and editors (Fuller et al. 2015). Therefore, it is important to make a study of the needs of all stakeholders in order to develop a checklist tool that genuinely meets the needs of the scientific community and that can benefit the progress of science. In this context, our miniRECH prototype could be used as a starting point for capturing these requirements. This could be achieved by following an agile software development method. Agile methods are a group of human-oriented adaptive and flexible development methods based on iterative and incremental techniques (Losada, Urretavizcaya, and Fernandez-Castro 2013). Agile methods not only provide constant feedback from end users against which we can validate and steer decisions, but also provide an excellent mechanism to conduct formal reviews of a product before it is released. The evaluation of the miniRECH interface could be achieved by gathering representatives of the scientific community into a focus group, providing the chance to react fast as their requirements are elicited.

- A Web application for miniRECH.

With a web-based application, the miniRECH itself needs only be developed for a single operating system, achieving an important level of usability that benefits the progress of science. Therefore, there is no need to develop and test it on all possible operating system versions and configuration.

- Completing checklists with natural language processing techniques.

Development of an automatic framework to fill checklists constitutes an interesting and challenging strategy for improving the completeness and accuracy of experimental data and metadata requested by checklists. In order to do so, natural language processing techniques could play an important role in this context (Hirschberg and Manning 2015). These techniques can be applied to extract experimental features on the basis of which to describe and cluster words, avoiding manual entries (Uzuner, Solti, and Cadag 2010). Currently, a new PhD-student in our research group is developing a strategy for filling some of our checklists via a text mining approach.

- Combining checklists with ontologies.

In order to assist scientists in describing their experiments, some generic software tools such as RightField (Wolstencroft et al. 2011) and recently XperimentR (Tomlinson et al. 2013) have been developed. These software tools use standardised ontologies to annotate the details of the experiments. The combination of checklists with bio-ontologies can greatly facilitate the integration process; by reducing the semantic heterogeneities of data and metadata, facilitating the collection, organisation, exploration, sharing and reuse of information (Bodenreider and Stevens 2006). However, due to the current lack of information reported in the biomedical scientific literature, it is only possible to represent some entities. The Ontology Engineering Group of the Universidad Politécnica de

Madrid, Spain, reached the same conclusions after modelling our checklists for *Trypanosoma* experiments into an ontology for experimental protocols. Therefore, there is a need for describing experimental details in order to ensure that experimental findings can be easily interpreted and, as a consequence, integrated (Parkinson et al. 2009).

- A framework to amalgamate data and metadata via checklists.

Both data and metadata collected by standard checklists have been the basis for the establishment of repositories that take and disseminate such information. This is the case of ArrayExpress (Rustici et al. 2013), one of the major international repositories for high-throughput functional genomics data from both microarray and high-throughput sequencing studies, where data are collected in conformity to the Minimum Information About a Microarray Experiment (MIAME) and the Minimum Information About a Sequencing Experiment (MINSEQE) standards. However, there are many repositories which lack the metadata information regarding the data they contain. Therefore, there is a need to create a framework in which data and metadata are described under the same criteria. An effort has been taken by applying the ARRIVE guidelines to the International Mouse Phenotyping Consortium (IMPC) (Karp et al. 2015). In this context, checklists could offer a starting point in how experiments should be described for linking data and metadata. By creating this framework, the scientific community will be able to use and reuse the data in a metadata context, which will improve the general value of the scientific evidence.

- Implementation of checklists through the experimental process

Following the previous rationale, it is also important to implement checklists through the experimental process. This is because laboratory protocols are an integral part of the research and, therefore, they are decisive in enabling reproducibility. An effort has been taken to represent laboratory protocols by implementing workflows (Maccagnan et al. 2010). By adopting standard workflows based on checklists would be possible to guarantee that the minimum information about a particular experimental methodology is considered. A starting point for adapting checklists to laboratory protocols could be made by using the COW (Combining Ontologies with Workflows) software tool (Maccagnan et al. 2010). This approach could not only deal with the complexity of the experimental process via workflows, but also with the formalisation of data and metadata via ontologies.

- Automatic reporting assessments via text mining technology

We hope that our text mining strategy can be taken as a starting point for future more focused assessment of literature; targeting a wider array of characteristics in preclinical and clinical studies. Its potential implementation would enable a straightforward pathway when it comes to reporting key information involved in preclinical and clinical research – *e.g.* by entering it into the publication cycle as a pre-screening test for submitted manuscripts, which will have an important positive impact on several fronts of the biomedical domain, including the reproducibility of experimental findings and the accuracy of meta-analysis.

- Evaluation of the success of our strategies for assessing reporting

There are several approaches that could be considered to assess the success of our strategies. On the one hand, in the case of miniRECH, it would be useful to carry out a study in which both authors and peer-reviewers are involved. By using the manuscripts ready to be submitted to a journal, two study groups could be created: 1) authors who would assess the information reported in their manuscripts (and improve it accordingly) before submission, 2) peer-reviewers who would assess the information reported in manuscripts (without a prior assessment by authors). A group (both authors and peer-reviewers) with no intervention of miniRECH would be considered as a control group. By keeping the normal feedback between authors and peer-reviewers through editors, this approach would help not only assess the success of our strategies, but also identify the actor (authors, peer-reviewers, or both) in the publication process who would be playing the main role in guaranteeing the quality of reporting information of a paper. Therefore, miniRECH should be adapted according to their needs in order to increase its usability. On the other hand, in the case of the text-mining approach – and considering the impact that it has had on the scientific community, it is hoped that this analysis drives journals and authors to include the descriptions of age and sex in the biomedical literature. Therefore, it would be good if this analysis could be rerun between three and five years from now to see if the situation has changed.

7.4 Conclusion

The assessment of scientific reports both pre- and post- publication has become an integral part of the scientific process, particularly in the field of the life sciences. This will not only be beneficial for improving reproducibility, but will also be helpful to avoid incomplete reporting from entering the literature and to select articles for evaluation by systematic reviews and meta-analyses. This Thesis, therefore, focused on the development of strategies for assessing the reporting of bio-experiments. A high-quality description of the bio-experiments will be useful for the integration of data and metadata and, so, gaining a better understanding of the natural phenomena in the context of the experimental evidence.

The initial hypothesis for this work was that by generating strategies to facilitate the assessment of reporting through the publication process, scientists (*i.e.*, authors, editors and peer-reviewers) will appreciate its usefulness (hopefully) and as a consequence its direct application to ensure the quality of experimental information to be published, which will represent a step forward in enhancing reproducibility. In this way, this Thesis provides both a manual and an automatic method for the assessment of the scientific reporting of bio-experiments. This was done through a spreadsheet-based tool and a text mining approach, respectively.

Although our strategies were used to assess the quality of experimental methods reporting in some biomedical fields, they can be adapted in order to cover a broader range of biomedical and life science subjects; *i.e.*, targeting a wider array of characteristics in experimental research. In addition, as a result of our assessments, this Thesis provides evidence about the current state of the reporting in biomedical publications.

We hope that other researchers will see the utility of our strategies to help promoting reproducibility via reporting and therefore apply them in their daily research practice, improving the general value of the scientific evidence.

References

Alaoui-Jamali, M. A., I. Dupre, and H. Qiang. 2004. "Prediction of drug sensitivity and drug resistance in cancer by transcriptional and proteomic profiling." *Drug Resist Updat* 7 (4-5):245-55. doi: 10.1016/j.drup.2004.06.004.

Ali, J. 2010. "Manuscript rejection: causes and remedies." *J Young Pharm* 2 (1):3-6. doi: 10.4103/0975-1483.62205.

Anon. 2013a. "How science goes wrong." *The Economics*, 19 Oct 2013, 13.

Anon. 2013b. "Raising standards." *Nature Genetics* 45 (5):467-467. doi: 10.1038/ng.2621.

Anon. 2013c. "Reducing our irreproducibility." *Nature* 496 (7446):398-398.

Anon. 2013d. "The Reproducibility Initiative." Available: https://www.scienceexchange.com/reproducibility. Accessed 17 January 2014.

Anon. 2013e. "Unreliable research: Trouble at the lab." *The Economist*, 19 Oct 2013, 26-30.

Antezana, E., W. Blonde, A. Venkatesan, B. De Baets, V. Mironov, and M. Kuiper. 2011. "Semantic systems biology: enabling integrative biology via semantic web technologies." Proceedings of the International Conference on Web Intelligence, Mining and Semantics, Sogndal, Norway.

Arnold, A. P. 2010. "Promoting the understanding of sex differences to enhance equity and excellence in biomedical science." *Biol Sex Differ* 1 (1):1. doi: doi: 10.1186/2042-6410-1-1.

Arnold, S. F., M. Stenzel, D. Drolet, and G. Ramachandran. 2015. "Using checklists and algorithms to improve qualitative exposure judgment accuracy." *J Occup Environ Hyg*:1-36. doi: 10.1080/15459624.2015.1053892.

Baker, D., K. Lidster, A. Sottomayor, and S. Amor. 2014. "Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies." *PLoS Biol* 12 (1):e1001756. doi: 10.1371/journal.pbio.1001756.

Barnes, C., I. Boutron, B. Giraudeau, R. Porcher, D. G. Altman, and P. Ravaud. 2015. "Impact of an online writing aid tool for writing a randomized trial report: the COBWEB (Consort-based WEB tool) randomized controlled trial." *BMC Med* 13:221. doi: 10.1186/s12916-015-0460-y.

Bechhofer, S., I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, et al. 2013. "Why linked data is not enough for scientists." *Future Generation Computer Systems-the International Journal of Grid Computing and Escience* 29 (2):599-611.

Begley, C. G., and L. M. Ellis. 2012. "Drug development: Raise standards for preclinical cancer research." *Nature* 483 (7391):531-3. doi: 10.1038/483531a483531a.

Bergs, J., J. Hellings, I. Cleemput, O. Zurel, V. De Troyer, M. Van Hiel, J. L. Demeere, D. Claeys, and D. Vandijck. 2014. "Systematic review and meta-analysis of the effect of the World Health Organization surgical safety checklist on postoperative complications." *Br J Surg* 101 (3):150-8. doi: 10.1002/bjs.9381.

Bodenreider, O., and R. Stevens. 2006. "Bio-ontologies: current trends and future directions." *Brief Bioinform* 7 (3):256-74. doi: 10.1093/bib/bbl027.

Bolli, R. 2015. "Reflections on the Irreproducibility of Scientific Papers." *Circ Res* 117 (8):665-6. doi: 10.1161/CIRCRESAHA.115.307496.

Bramhall, M., O. Florez-Vargas, R. Stevens, A. Brass, and S. Cruickshank. 2015. "Quality of methods reporting in animal models of colitis." *Inflamm Bowel Dis* 21 (6):1248-59. doi: doi: 10.1097/MIB.0000000000000369.

Brazma, A. 2001. "On the importance of standardisation in life sciences." *Bioinformatics* 17 (2):113-4.

Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, et al. 2001. "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data." *Nat Genet* 29 (4):365-71. doi: 10.1038/ng1201-365.

Brinkman, R. R., M. Courtot, D. Derom, J. M. Fostel, Y. He, P. Lord, J. Malone, H. Parkinson, B. Peters, P. Rocca-Serra, et al. 2010. "Modeling biomedical experimental processes with OBI." *J Biomed Semantics* 1 Suppl 1:S7. doi: 10.1186/2041-1480-1-S1-S7.

Broad, W.J., and N. Wade. 1982. *Betrayers of the truth: fraud and deceit in the halls of science*. New York, NY, USA: Simon and Schuster.

Bujkiewicz, S., H. E. Jones, M. C. Lai, N. J. Cooper, N. Hawkins, H. Squires, K. R. Abrams, D. J. Spiegelhalter, and A. J. Sutton. 2011. "Development of a transparent interactive decision interrogator to facilitate the decision-making process in health care." *Value Health* 14 (5):768-76. doi: 10.1016/j.jval.2010.12.002.

Casadevall, A., and F. C. Fang. 2010. "Reproducible science." *Infect Immun* 78 (12):4972-5. doi: 10.1128/IAI.00908-10.

Chen, J. X., M. Krane, M. A. Deutsch, L. Wang, M. Rav-Acha, S. Gregoire, M. C. Engels, K. Rajarajan, R. Karra, E. D. Abel, et al. 2012. "Inefficient reprogramming of fibroblasts into cardiomyocytes using Gata4, Mef2c, and Tbx5." *Circ Res* 111 (1):50-5. doi: 10.1161/CIRCRESAHA.112.270264.

Cheung, K. H., H. R. Frost, M. S. Marshall, E. Prud'hommeaux, M. Samwald, J. Zhao, and A. Paschke. 2009. "A journey to Semantic Web query federation in the life sciences." *BMC Bioinformatics* 10. doi: 10.1186/1471-2105-10-S10-S10.

Clark, T. W. 2014. "Sociotechnical architecture for biomedical communication on the Web of argument and data." Doctor of Philosophy, Faculty of Engineering and Physical Sciences, School of Computer Science, The University of Manchester (eScholarID:234124).

Cochrane Schizophrenia Group, . 2016. "RevMan HAL v 4.0 Frequently Asked Questions." Accessed Accessed 12/07/2016. http://schizophrenia.cochrane.org/revman-hal-v-40-frequently-asked-questions.

Cohen, M. A., and R. W. Hersh. 2005. "A survey of current work in biomedical text mining." *Briefings in Bioinformatics* 6 (1):57-71. doi: doi: 10.1093/bib/6.1.57.

Comeau, D. C., H. Liu, R. Islamaj Dogan, and W. J. Wilbur. 2014. "Natural language processing pipelines to annotate BioC collections with an application to the NCBI disease corpus." *Database (Oxford)* 2014. doi: 10.1093/database/bau056.

Cotsapas, C., B. F. Voight, E. Rossin, K. Lage, B. M. Neale, C. Wallace, G. R. Abecasis, J. C. Barrett, T. Behrens, J. Cho, et al. 2011. "Pervasive sharing of genetic effects in autoimmune disease." *PLoS Genet* 7 (8):e1002254. doi: 10.1371/journal.pgen.1002254.

Cox, T. C. 2015. "Utility and limitations of animal models for the functional validation of human sequence variants." *Mol Genet Genomic Med* 3 (5):375-82. doi: 10.1002/mgg3.167.

Cumming, G. 2014. "The new statistics: why and how." *Psychol Sci* 25 (1):7-29. doi: 10.1177/0956797613504966.

Cummings, K. L., and R. L. Tarleton. 2004. "Inducible nitric oxide synthase is not essential for control of *Trypanosoma cruzi* infection in mice." *Infect Immun* 72 (7):4081-9. doi: 10.1128/IAI.72.7.4081-4089.2004.

Cunningham, H., V. Tablan, A. Roberts, and K. Bontcheva. 2013. "Getting more out of biomedical documents with GATE's full lifecycle open source text analytics." *PLoS Comput Biol* 9 (2):e1002854. doi: doi: 10.1371/journal.pcbi.1002854.

Curtis, A. M., C. T. Fagundes, G. Yang, E. M. Palsson-McDermott, P. Wochal, A. F. McGettrick, N. H. Foley, J. O. Early, L. Chen, H. Zhang, et al. 2015. "Circadian control of innate immunity in macrophages by miR-155 targeting Bmal1." *Proc Natl Acad Sci U S A* 112 (23):7231-6. doi: 10.1073/pnas.1501327112.

Cyranoski, D. 2012. "Retraction record rocks community." *Nature* 489 (7416):346-7. doi: 10.1038/489346a.

Denayer, T., T. Stöhr, and M. Van Roy. 2014. "Animal models in translational medicine: Validation and prediction." *New Horizons in Translational Medicine* 2 (1):5-11. doi: 10.1016/j.nhtm.2014.08.001.

Drummond, C. 2009. "Replicability is not Reproducibility: Nor is it a good science." Paper presented at: Evaluation Methods for Machine Learning Workshop at the 26th International Conference on Machine Learning, Montreal, Quebec, Canada, June 2009.

Edgar, R., and T. Barrett. 2006. "NCBI GEO standards and services for microarray data." *Nat Biotechnol* 24 (12):1471-2. doi: 10.1038/nbt1206-1471.

Emerson, G. B., W. J. Warme, F. M. Wolf, J. D. Heckman, R. A. Brand, and S. S. Leopold. 2010. "Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial." *Arch Intern Med* 170 (21):1934-9. doi: 10.1001/archinternmed.2010.406.

Fanelli, D. 2009. "How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data." *PLoS One* 4 (5):e5738. doi: 10.1371/journal.pone.0005738.

Fanelli, D. 2010. ""Positive" results increase down the Hierarchy of the Sciences." *PLoS One* 5 (4):e10068. doi: 10.1371/journal.pone.0010068.

Fanelli, D. 2012. "Negative results are disappearing from most disciplines and countries." *Scientometrics* 90 (3):891-904. doi: 10.1007/s11192-011-0494-7.

Fanelli, D. 2013. "Why growing retractions are (mostly) a good sign." *PLoS Med* 10 (12):e1001563. doi: 10.1371/journal.pmed.1001563.

Fang, F. C., R. G. Steen, and A. Casadevall. 2012. "Misconduct accounts for the majority of retracted scientific publications." *Proc Natl Acad Sci U S A* 109 (42):17028-33. doi: 10.1073/pnas.1212247109.

Fjällbrandt, N. 1997. "Scholarly Communication: Historical Development and New Possibilities." Proceedings of the IATUL Conferences, Trondheim, Norway.

Fleuren, W. W., and W. Alkema. 2015. "Application of text mining in the biomedical domain." *Methods* 74:97-106. doi: doi: 10.1016/j.ymeth.2015.01.015.

Florez-Vargas, O., M. Bramhall, H. Noyes, S. Cruickshank, R. Stevens, and A. Brass. 2014. "The quality of methods reporting in parasitology experiments." *PLoS One* 9 (7):e101131. doi: doi: 10.1371/journal.pone.0101131.

Florez-Vargas, O., A. Brass, G. Karystianis, M. Bramhall, R. Stevens, S. Cruickshank, and G. Nenadic. 2016. "Bias in the reporting of sex and age in biomedical research on mouse models." *Elife* 5. doi: 10.7554/eLife.13615.

Florez, O., G. Zafra, C. Morillo, J. Martin, and C. I. Gonzalez. 2006. "Interleukin-1 gene cluster polymorphism in chagas disease in a Colombian case-control study." *Hum Immunol* 67 (9):741-8. doi: 10.1016/j.humimm.2006.06.004.

Freedman, L. P., I. M. Cockburn, and T. S. Simcoe. 2015. "The Economics of Reproducibility in Preclinical Research." *PLoS Biol* 13 (6):e1002165. doi: 10.1371/journal.pbio.1002165.

Freedman, L. P., and J. Inglese. 2014. "The increasing urgency for standards in basic biologic research." *Cancer Res* 74 (15):4024-9. doi: 10.1158/0008-5472.CAN-14-0925.

Fuller, T., M. Pearson, J. Peters, and R. Anderson. 2015. "What affects authors' and editors' use of reporting guidelines? Findings from an online survey and qualitative interviews." *PLoS One* 10 (4):e0121585. doi: 10.1371/journal.pone.0121585.

Garijo, D., S. Kinnings, L. Xie, L. Xie, Y. Zhang, P. E. Bourne, and Y. Gil. 2013. "Quantifying reproducibility in computational biology: the case of the tuberculosis drugome." *PLoS One* 8 (11):e80278. doi: 10.1371/journal.pone.0080278.

Garten, Y., A. Coulet, and R. B. Altman. 2010. "Recent progress in automatically extracting information from the pharmacogenomic literature." *Pharmacogenomics* 11 (10):1467-89. doi: 10.2217/pgs.10.136.

Gawande, A. 2010. *The Checklist Manifesto: How To Get Things Right*. Great Britain: Profile Books Ltd.

GBSI. 2013. The Case for Standards in Life Science Research: Seizing Opportunities at a Time of Critical Need. Washington, D.C.: Global Biological Standards Institute.

Genesis, 11:1-9. 2010. *The Holy Bible*. 1611 ed. Vol. 6, *King James Version*. Edinburgh, UK: Hendrickson Publishers.

Gomez-Cabrero, D., I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegner. 2014. "Data integration in the era of omics: current and future challenges." *BMC Syst Biol* 8 Suppl 2:I1. doi: 10.1186/1752-0509-8-S2-I1.

Goodman, A., A. Pepe, A. W. Blocker, C. L. Borgman, K. Cranmer, M. Crosas, R. Di Stefano, Y. Gil, P. Groth, M. Hedstrom, et al. 2014. "Ten simple rules for the care and feeding of scientific data." *PLoS Comput Biol* 10 (4):e1003542. doi: 10.1371/journal.pcbi.1003542.

Goodstein, David L. 2010. *On fact and fraud : cautionary tales from the front lines of science*. Princeton, N.J.: Princeton University Press.

Gray, K., L. Young, and A. Waytz. 2012. "Mind Perception Is the Essence of Morality." *Psychol Inq* 23 (2):101-124. doi: 10.1080/1047840X.2012.651387.

Green, L., S. Allard, and R. Cardigan. 2015. "Modern banking, collection, compatibility testing and storage of blood and blood components." *Anaesthesia* 70 Suppl 1:3-9, e2. doi: 10.1111/anae.12912.

Gruber, T. R. 1993. "A Translation Approach to Portable Ontology Specifications." *Knowledge Acquisition* 5 (2):199-220. doi: 10.1006/knac.1993.1008.

Gulin, J. E., D. M. Rocco, and F. Garcia-Bournissen. 2015. "Quality of Reporting and Adherence to ARRIVE Guidelines in Animal Studies for Chagas Disease Preclinical Drug Research: A Systematic Review." *PLoS Negl Trop Dis* 9 (11):e0004194. doi: 10.1371/journal.pntd.0004194.

Hahn, U., K. B. Cohen, Y. Garten, and N. H. Shah. 2012. "Mining the pharmacogenomics literature--a survey of the state of the art." *Brief Bioinform* 13 (4):460-94. doi: 10.1093/bib/bbs018.

Haibe-Kains, B., N. El-Hachem, N. J. Birkbak, A. C. Jin, A. H. Beck, H. J. Aerts, and J. Quackenbush. 2013. "Inconsistency in large pharmacogenomic studies." *Nature* 504 (7480):389-93. doi: 10.1038/nature12831.

Haspel, J. A., S. Chettimada, R. S. Shaik, J. H. Chu, B. A. Raby, M. Cernadas, V. Carey, V. Process, G. M. Hunninghake, E. Ifedigbo, et al. 2014. "Circadian rhythm reprogramming during lung inflammation." *Nat Commun* 5:4753. doi: 10.1038/ncomms5753.

Haynes, A. B., T. G. Weiser, W. R. Berry, S. R. Lipsitz, A. H. Breizat, E. P. Dellinger, T. Herbosa, S. Joseph, P. L. Kibatala, M. C. Lapitan, et al. 2009. "A surgical safety checklist to reduce morbidity and mortality in a global population." *N Engl J Med* 360 (5):491-9. doi: 10.1056/NEJMsa0810119.

Head, M. L., L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions. 2015. "The extent and consequences of p-hacking in science." *PLoS Biol* 13 (3):e1002106. doi: 10.1371/journal.pbio.1002106.

Hirschberg, J., and C. D. Manning. 2015. "Advances in natural language processing." *Science* 349 (6245):261-266.

Hotho, A., A. Numberger, and G. Paab. 2005. "A Brief Survey of Text Mining." *LDV Forum – GLDV Journal for Computational Linguistics and Language Technology* 20 (1):19–62.

Hutchinson, L., and R. Kirk. 2011. "High drug attrition rates--where are we going wrong?" *Nat Rev Clin Oncol* 8 (4):189-90. doi: 10.1038/nrclinonc.2011.34.

Ieda, M., J. D. Fu, P. Delgado-Olguin, V. Vedantham, Y. Hayashi, B. G. Bruneau, and D. Srivastava. 2010. "Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors." *Cell* 142 (3):375-86. doi: 10.1016/j.cell.2010.07.002.

IM2C, . 2016. "Reporting on mathematical models." Accessed Accessed 13/07/2016. https://www.immchallenge.org.au/supporting-resources/reporting-on-mathematical-models.

International Committee of Medical Journal Editors. 2013. "Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publications." http://www.icmje.org/#prepare.

Ioannidis, J. P. 2005. "Why most published research findings are false." *PLoS Medicine* 2 (8):e124. doi: 10.1371/journal.pmed.0020124.

Ioannidis, J. P. 2011. "An epidemic of false claims. Competition and conflicts of interest distort too many medical findings." *Sci Am* 304 (6):16.

Jager, L. R., and J. T. Leek. 2014. "An estimate of the science-wise false discovery rate and application to the top medical literature." *Biostatistics* 15 (1):1-12. doi: 10.1093/biostatistics/kxt007.

Johnson, G. 2014. "New truths that only one can see." *The New York Times*, 20 Jan 2014. Accessed 10 Nov 2015. Available: http://www.nytimes.com/2014/01/21/science/new-truths-that-only-one-can-see.html. Accessed 10 October 2015.

Juluru, K., and J. Eng. 2015. "Use of Spreadsheets for Research Data Collection and Preparation:: A Primer." *Acad Radiol* 22 (12):1592-9. doi: 10.1016/j.acra.2015.08.024.

Karp, N. A., T. F. Meehan, H. Morgan, J. C. Mason, A. Blake, N. Kurbatova, D. Smedley, J. Jacobsen, R. F. Mott, V. Iyer, et al. 2015. "Applying the ARRIVE Guidelines to an In Vivo Database." *PLoS Biol* 13 (5):e1002151. doi: 10.1371/journal.pbio.1002151.

Kennedy, D. 2006. "Editorial retraction." *Science* 311 (5759):335. doi: 10.1126/science.1124926.

Kilkenny, C., W. J. Browne, I. C. Cuthill, M. Emerson, and D. G. Altman. 2010. "Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research." *PLoS Biol* 8 (6):e1000412. doi: 10.1371/journal.pbio.1000412.

Kilkenny, C., N. Parsons, E. Kadyszewski, M. F. Festing, I. C. Cuthill, D. Fry, J. Hutton, and D. G. Altman. 2009. "Survey of the quality of experimental design, statistical analysis and reporting of research using animals." *PLoS One* 4 (11):e7824. doi: 10.1371/journal.pone.0007824.

King, R. D., K. E. Whelan, F. M. Jones, P. G. K. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell, and S. G. Oliver. 2004. "Functional genomic hypothesis generation and experimentation by a robot scientist." *Nature* 427 (6971):247-252. doi: 10.1038/nature02236.

Klein, S. L., and C. W. Roberts. 2015. "Epilogue: Future of Sex and Gender-Based Studies in Infectious Diseases." In *Sex and Gender Differences in Infection and Treatments for Infectious Diseases*, 389-393. Springer International Publishing.

Koch, M., P. Riss, W. Umek, and E. Hanzal. 2015. "The explicit mentioning of reporting guidelines in urogynecology journals in 2013: A bibliometric study." *Neurourol Urodyn*. doi: 10.1002/nau.22726.

Laehnemann, D., R. Pena-Miller, P. Rosenstiel, R. Beardmore, G. Jansen, and H. Schulenburg. 2014. "Genomics of rapid adaptation to antibiotics: convergent evolution and scalable sequence amplification." *Genome Biol Evol* 6 (6):1287-301. doi: 10.1093/gbe/evu106.

Landis, S. C., S. G. Amara, K. Asadullah, C. P. Austin, R. Blumenstein, E. W. Bradley, R. G. Crystal, R. B. Darnell, R. J. Ferrante, H. Fillit, et al. 2012. "A call for transparent reporting to optimize the predictive value of preclinical research." *Nature* 490 (7419):187-191. doi: 10.1038/nature11556.

Lawrence, P. A. 2003. "The politics of publication." *Nature* 422 (6929):259-61. doi: 10.1038/422259a.

Losada, B., M. Urretavizcaya, and I. Fernandez-Castro. 2013. "A guide to agile development of interactive software with a "User Objectives"-driven methodology." *Science of Computer Programming* 78 (11):2268-2281.

Loscalzo, J. 2012. "Irreproducible experimental results: causes, (mis)interpretations, and consequences." *Circulation* 125 (10):1211-4. doi: 10.1161/CIRCULATIONAHA.112.098244.

Maccagnan, A., M. Riva, E. Feltrin, B. Simionati, T. Vardanega, G. Valle, and N. Cannata. 2010. "Combining ontologies and workflows to design formal protocols for biological laboratories." *Autom Exp* 2:3. doi: 10.1186/1759-4499-2-3.

Malone, J., E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson. 2010. "Modeling sample variables with an Experimental Factor Ontology." *Bioinformatics* 26 (8):1112-8. doi: 10.1093/bioinformatics/btq099.

McNutt, M. 2014. "Journals unite for reproducibility." *Science* 346 (6210):679.

Michalek, A. M., A. D. Hutson, C. P. Wicher, and D. L. Trump. 2010. "The costs and underappreciated consequences of research misconduct: a case study." *PLoS Med* 7 (8):e1000318. doi: 10.1371/journal.pmed.1000318.

Mizoguchi, Atsushi. 2012. "Animal Models of Inflammatory Bowel Disease." In *Progress in Molecular Biology and Translational Science*, edited by P. Michael Conn, 263-320. Academic Press.

Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gotzsche, P. J. Devereaux, D. Elbourne, M. Egger, D. G. Altman, and Group Consolidated Standards of Reporting Trials. 2010. "CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials." *J Clin Epidemiol* 63 (8):e1-37. doi: 10.1016/j.jclinepi.2010.03.004.

Moher, D., K. F. Schulz, I. Simera, and D. G. Altman. 2010. "Guidance for developers of health research reporting guidelines." *PLoS Med* 7 (2):e1000217. doi: 10.1371/journal.pmed.1000217.

Moher, D., I. Simera, K. F. Schulz, J. Hoey, and D. G. Altman. 2008. "Helping editors, peer reviewers and authors improve the clarity, completeness and transparency of reporting health research." *Bmc Medicine* 6. doi: 10.1186/1741-7015-6-13.

Musen, M. A., N. F. Noy, N. H. Shah, P. L. Whetzel, C. G. Chute, M. A. Story, B. Smith, and Ncbo Team. 2012. "The National Center for Biomedical Ontology." *Journal of the American Medical Informatics Association* 19 (2):190-195. doi: 10.1136/amiajnl-2011-000523.

Nature. 2013. Reporting Checklist For Life Sciences Articles. edited by checklist.pdf: Nature Publishing Group.

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, et al. 2015. "SCIENTIFIC STANDARDS. Promoting an open research culture." *Science* 348 (6242):1422-5. doi: 10.1126/science.aab2374.

Obokata, H., Y. Sasai, H. Niwa, M. Kadota, M. Andrabi, N. Takata, M. Tokoro, Y. Terashita, S. Yonemura, C. A. Vacanti, et al. 2014. "Retraction: Bidirectional developmental potential in reprogrammed cells with acquired pluripotency." *Nature* 511 (7507):112. doi: 10.1038/nature13599.

Oxenham, S. 2016. "Evolutionary forces are causing a boom in bad science." *New Scientist*. Accessed Accessed 12/07/2016. https://www.newscientist.com/article/2096542-evolutionary-forces-are-causing-a-boom-in-bad-science/.

Paradis, E. G., L. T. Pinilla, B. P. Holder, Y. Abed, G. Boivin, and C. A. A. Beauchemin. 2015. "Impact of the H275Y and I223V Mutations in the Neuraminidase of the 2009 Pandemic

Influenza Virus In Vitro and Evaluating Experimental Reproducibility." *Plos One* 10 (5). doi: 10.1371/journal.pone.0126115.

Parkinson, H., M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, et al. 2009. "ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression." *Nucleic Acids Res* 37 (Database issue):D868-72. doi: 10.1093/nar/gkn889.

Penelope. 2016. "Penelope: Automated Scientific Scrutiny." Accessed Accessed 12/07/2016. http://www.peneloperesearch.com/.

Perrin, S. 2014. "Preclinical research: Make mouse studies work." *Nature* 507 (7493):423-5. doi: doi: 10.1038/507423a.

Pound, P., S. Ebrahim, P. Sandercock, M. B. Bracken, I. Roberts, and Group Reviewing Animal Trials Systematically. 2004. "Where is the evidence that animal research benefits humans?" *BMJ* 328 (7438):514-7. doi: 10.1136/bmj.328.7438.514.

Prinz, F., T. Schlange, and K. Asadullah. 2011. "Believe it or not: how much can we rely on published data on potential drug targets?" *Nat Rev Drug Discov* 10 (9):712. doi: 10.1038/nrd3439-c1.

Protze, S., S. Khattak, C. Poulet, D. Lindemann, E. M. Tanaka, and U. Ravens. 2012. "A new approach to transcription factor screening for reprogramming of fibroblasts to cardiomyocyte-like cells." *J Mol Cell Cardiol* 53 (3):323-32. doi: 10.1016/j.yjmcc.2012.04.010.

Ramesh, B. P., R. Prasad, T. Miller, B. Harrington, and H. Yu. 2012. "Automatic discourse connective detection in biomedical text." *J Am Med Inform Assoc* 19 (5):800-8. doi: 10.1136/amiajnl-2011-000775.

Rao, T. S., and C. Andrade. 2011. "The MMR vaccine and autism: Sensation, refutation, retraction, and fraud." *Indian J Psychiatry* 53 (2):95-6. doi: 10.4103/0019-5545.82529.

Rayner, T. F., P. Rocca-Serra, P. T. Spellman, H. C. Causton, A. Farne, E. Holloway, R. A. Irizarry, J. Liu, D. S. Maier, M. Miller, et al. 2006. "A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB." *BMC Bioinformatics* 7:489. doi: 10.1186/1471-2105-7-489.

Rebholz-Schuhmann, D., A. Oellrich, and R. Hoehndorf. 2012. "Text-mining solutions for biomedical research: enabling integrative biology." *Nat Rev Genet* 13 (12):829-39. doi: 10.1038/nrg3337.

Rehm, H. L., S. J. Bale, P. Bayrak-Toydemir, J. S. Berg, K. K. Brown, J. L. Deignan, M. J. Friez, B. H. Funke, M. R. Hegde, E. Lyon, et al. 2013. "ACMG clinical laboratory standards for next-generation sequencing." *Genet Med* 15 (9):733-47. doi: 10.1038/gim.2013.92.

Rocca-Serra, P., M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, S. Harris, W. Hide, O. Hofmann, et al. 2010. "ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level." *Bioinformatics* 26 (18):2354-6. doi: 10.1093/bioinformatics/btq415.

Rustici, G., N. Kolesnikov, M. Brandizi, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, J. Ison, M. Keays, et al. 2013. "ArrayExpress update--trends in database growth and links to data analysis tools." *Nucleic Acids Res* 41 (Database issue):D987-90. doi: 10.1093/nar/gks1174.

Sansone, S. A., P. Rocca-Serra, M. Brandizi, A. Brazma, D. Field, J. Fostel, A. G. Garrow, J. Gilbert, F. Goodsaid, N. Hardy, et al. 2008. "The first RSBI (ISA-TAB) workshop: "can a simple format work for complex studies?"." *OMICS* 12 (2):143-9. doi: 10.1089/omi.2008.0019.

Sauer, U., M. Heinemann, and N. Zamboni. 2007. "Getting closer to the whole picture." *Science* 316 (5824):550-1. doi: 10.1126/science.1142502.

Savova, G. K., J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. 2010. "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications." *J Am Med Inform Assoc* 17 (5):507-13. doi: 10.1136/jamia.2009.001560.

Schulz, K. F., D. G. Altman, D. Moher, and Consort Group. 2010. "CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials." *BMC Med* 8:18. doi: 10.1186/1741-7015-8-18.

Shadbolt, N., W. Hall, and T. Berners-Lee. 2006. "The Semantic Web revisited." *Ieee Intelligent Systems* 21 (3):96-101. doi: 10.1109/Mis.2006.62.

Shakespeare, T. P., V. J. Gebski, A. Thiagarajan, and J. Jay Lu. 2006. "Development of a spreadsheet for the calculation of new tools to improve the reporting of the results of medical research." *Med Inform Internet Med* 31 (2):121-7. doi: 10.1080/14639230600551397.

Shamseer, L., A. Stevens, B. Skidmore, L. Turner, D. G. Altman, A. Hirst, J. Hoey, A. Palepu, I. Simera, K. Schulz, et al. 2012. "Does journal endorsement of reporting guidelines influence the completeness of reporting of health research? A systematic review protocol." *Syst Rev* 1:24. doi: 10.1186/2046-4053-1-24.

Shanks, N., R. Greek, and J. Greek. 2009. "Are animal models predictive for humans?" *Philos Ethics Humanit Med* 4:2. doi: 10.1186/1747-5341-4-2.

Shatkay, H., and R. Feldman. 2003. "Mining the biomedical literature in the genomic era: an overview." *J Comput Biol* 10 (6):821-55. doi: 10.1089/106652703322756104.

Smaldino, P. E., and R. Mcelreath. 2016. "The natural selection of bad science." *arXiv*:1-41.

Smith, T. A., P. Kulatilake, L. J. Brown, J. Wigley, W. Hameed, and S. Shantikumar. 2015. "Do surgery journals insist on reporting by CONSORT and PRISMA? A follow-up survey of 'instructions to authors'." *Ann Med Surg (Lond)* 4 (1):17-21. doi: 10.1016/j.amsu.2014.12.003.

Soldatova, L. N., D. Nadis, R. D. King, P. S. Basu, E. Haddi, V. Baumle, N. J. Saunders, W. Marwan, and B. B. Rudkin. 2014. "EXACT2: the semantics of biomedical protocols." *BMC Bioinformatics* 15 Suppl 14:S5. doi: 10.1186/1471-2105-15-S14-S5.

Song, K., Y. J. Nam, X. Luo, X. Qi, W. Tan, G. N. Huang, A. Acharya, C. L. Smith, M. D. Tallquist, E. G. Neilson, et al. 2012. "Heart repair by reprogramming non-myocytes with cardiac transcription factors." *Nature* 485 (7400):599-604. doi: 10.1038/nature11139.

Steen, R. G. 2011a. "Retractions in the scientific literature: do authors deliberately commit research fraud?" *J Med Ethics* 37 (2):113-7. doi: 10.1136/jme.2010.038125.

Steen, R. G. 2011b. "Retractions in the scientific literature: is the incidence of research fraud increasing?" *J Med Ethics* 37 (4):249-53. doi: 10.1136/jme.2010.040923.

Stevens, R., M. E. Aranguren, K. Wolstencroft, U. Sattler, N. Drummond, M. Horridge, and A. Rector. 2007. "Using OWL to model biological knowledge." *International Journal of Human-Computer Studies* 65 (7):583-594. doi: 10.1016/j.ijhcs.2007.03.006.

Summerli.Wt, C. Broutbar, R. B. Foanes, R. Payne, O. Stutman, L. Hayflick, and R. A. Good. 1973. "Acceptance of Phenotypically Differing Cultured Skin in Man and Mice." *Transplantation Proceedings* 5 (1):707-710.

Tahvildari, D. 2015. "Semantic Support for Recording Laboratory Experimental Metadata: A Study in Food Chemistry." *Semantic Web: Latest Advances and New Domains, Eswc 2015* 9088:783-794. doi: 10.1007/978-3-319-18818-8_51.

Taylor, C. F., D. Field, S. A. Sansone, J. Aerts, R. Apweiler, M. Ashburner, C. A. Ball, P. A. Binz, M. Bogue, T. Booth, et al. 2008. "Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project." *Nat Biotechnol* 26 (8):889-96. doi: 10.1038/nbt.1411nbt.

Taylor, C. F., N. W. Paton, K. S. Lilley, P. A. Binz, R. K. Julian, Jr., A. R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E. W. Deutsch, et al. 2007. "The minimum information about a proteomics experiment (MIAPE)." *Nat Biotechnol* 25 (8):887-93. doi: 10.1038/nbt1329.

The Wellcome Trust. 2002. "Handling allegations of research misconduct." Last Modified 2005 Accessed 19 Oct 2015. http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002756.htm.

Tomlinson, C. D., G. R. Barton, M. Woodbridge, and S. A. Butcher. 2013. "XperimentR: painless annotation of a biological experiment for the laboratory scientist." *BMC Bioinformatics* 14:8. doi: 10.1186/1471-2105-14-8.

Turner, L., L. Shamseer, D. G. Altman, K. F. Schulz, and D. Moher. 2012. "Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review." *Syst Rev* 1:60. doi: 10.1186/2046-4053-1-60.

Uzuner, O., I. Solti, and E. Cadag. 2010. "Extracting medication information from clinical text." *Journal of the American Medical Informatics Association* 17 (5):514-518.

van der Worp, H. B., D. W. Howells, E. S. Sena, M. J. Porritt, S. Rewell, V. O'Collins, and M. R. Macleod. 2010. "Can animal models of disease reliably inform human studies?" *PLoS Med* 7 (3):e1000245. doi: doi: 10.1371/journal.pmed.1000245.

van der Worp, H. B., and M. R. Macleod. 2011. "Preclinical studies of human disease: Time to take methodological quality seriously." *Journal of Molecular and Cellular Cardiology* 51 (4):449-450. doi: 10.1016/j.yjmcc.2011.04.008.

Van Noorden, R. 2011. "Science publishing: The trouble with retractions." *Nature* 478 (7367):26-8. doi: 10.1038/478026a.

Vespa, G. N., F. Q. Cunha, and J. S. Silva. 1994. "Nitric oxide is involved in control of *Trypanosoma cruzi*-induced parasitemia and directly kills the parasite in vitro." *Infect Immun* 62 (11):5177-82.

Volm, M., and T. Efferth. 2015. "Prediction of Cancer Drug Resistance and Implications for Personalized Medicine." *Front Oncol* 5:282. doi: 10.3389/fonc.2015.00282.

Ware, J. J., and M. R. Munafo. 2015. "Significance chasing in research practice: causes, consequences and possible solutions." *Addiction* 110 (1):4-8. doi: 10.1111/add.12673.

Wells, J. A. 2008. Observing and Reporting Suspected Misconduct in Biomedical Research. The Gallup Organization and The Office of Research Integrity

Whetzel, P. L., N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen. 2011. "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications." *Nucleic Acids Research* 39:W541-W545. doi: 10.1093/Nar/Gkr469.

Whiting, P., R. Harbord, and J. Kleijnen. 2005. "No role for quality scores in systematic reviews of diagnostic accuracy studies." *BMC Med Res Methodol* 5:19. doi: 10.1186/1471-2288-5-19.

Witwer, K. W. 2013. "Data submission and quality in microarray-based microRNA profiling." *Clin Chem* 59 (2):392-400. doi: 10.1373/clinchem.2012.193813.

Wolstencroft, K., S. Owen, M. Horridge, O. Krebs, W. Mueller, J. L. Snoep, F. du Preez, and C. Goble. 2011. "RightField: embedding ontology annotation in spreadsheets." *Bioinformatics* 27 (14):2021-2. doi: 10.1093/bioinformatics/btr312.

World Health Organization, WHO. 2010. "Neglected tropical diseases: Innovative and Intensified Disease Management (IDM)." http://www.who.int/neglected_diseases/disease_management/en/.

Zamer, W. E. 2011. "A Cohesive Biology of Organisms Is on the Horizon." *Bioscience* 61 (11):848-849. doi: 10.1525/bio.2011.61.11.3.

Appendix

A.1. Supplementary files Chapter 3

Table S1. Quality measures of the studies that failed to fulfil any one of data of minimal information about the parasite in *Trypanosoma* experiments.

				Culture conditions of Trypanosomes										
	Parasite information			Parasites from animals				Parasites from cells						
Articles	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**	**P12**	**Total**	**%**
Amin et al., 2010	✓	✓	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2/12	16.7%
Chessler et al., 2009	✓	NA	✓	*	*	*	*	✓	✓	✓	✓	NA	6/8	75%
Costales et al., 2009	✓	✓	✓	*	*	*	*	✓	✓	✓	✓	NA	7/8	87.5%
Garg at al., 2004	✓	✓	✓	*	*	*	*	✓	✓	✓	NA	NA	6/8	75%
Genovesio et al., 2011	✓	✓	✓	*	*	*	*	✓	✓	✓	✓	NA	7/8	87.5%
Goldenberg et al., 2009	✓	✓	✓	*	*	*	*	✓	✓	✓	✓	✓	8/8	100%
Hashimoto et al., 2005	✓	✓	✓	*	*	*	*	✓	✓	✓	✓	✓	8/8	100%
Hill et al., 2005	✓	✓	✓	✓	*	*	*	*	*	*	*	*	4/4	100%
Kierstein et al., 2006	✓	✓	✓	✓	NA	NA	✓	*	*	*	*	*	5/7	71.4%
Li et al., 2009	✓	✓	✓	✓	✓	✓	✓	*	*	*	*	*	7/7	100%
Li et al., 2011	✓	✓	✓	✓	✓	✓	✓	*	*	*	*	*	7/7	100%
Lopez et al., 2008	✓	✓	✓	✓	✓	✓	✓	*	*	*	*	*	7/7	100%
Manque et al., 2011	✓	✓	✓	*	*	*	*	✓	NA	NA	NA	NA	4/8	50%
Meade et al., 2009	✓	NA	✓	✓	*	*	*	*	*	*	*	*	3/4	75%
Mekata et al., 2012	✓	✓	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1/12	8.3%
Mukherjee et al., 2003	✓	✓	✓	NA	NA	NA	NA	NA	NA	NA	NA	NA	3/12	25%
Mukherjee et al., 2008	✓	✓	✓	NA	NA	NA	NA	NA	NA	NA	NA	NA	3/12	25%
Noyes et al., 2009	✓	✓	✓	✓	✓	NA	✓	*	*	*	*	*	6/7	85.7%
O'Gorman et al., 2009	✓	✓	✓	✓	NA	NA	*	*	*	*	*	*	4/6	66.7%
Soares et al., 2010	✓	✓	✓	*	*	*	*	✓	NA	NA	NA	NA	4/8	50%
Soares et al., 2011	✓	✓	✓	*	*	*	*	✓	NA	NA	NA	NA	4/8	50%
Graefe et al., 2006	✓	✓	✓	✓	NA	NA	NA	*	*	*	*	*	4/7	57.1%
Tanowitz et al., 2011	✓	✓	✓	✓	NA	NA	✓	✓	✓	✓	✓	✓	10/12	83.3%
Total	23/23	21/23	21/23	10/14	4/12	3/12	6/11	10/14	7/14	7/14	6/14	3/14		
%	100%	91.3%	91.3%	71.4%	33.3%	25%	54.5%	71.4%	50%	50%	42.9%	21.4%		

Criteria: P1 (species), P2 (strain), P3 (stage), P4 (species and strain of animal), P5 (age), P6 (gender), P7 (parasite collection sample); P8 (cell type), P9 (culture medium), P10 (supplements and antibiotics), P11 (temperature and CO_2 atmosphere), and P12 (time of growing of the parasite prior to infection).

✓: meets the criteria

NA: information not available

*****: not applicable

Table S2. Quality measures of the studies that failed to fulfil any one of data of minimal information about the host in *Trypanosoma* experiments.

	Characteristics and culture conditions of the host models												
	Animal model					**Cell model**							
Articles	**H1**	**H2**	**H3**	**H4**	**H5**	**H6**	**H7**	**H8**	**H9**	**H10**	**H11**	**H12**	**Total**	**%**
Amin et al., 2010	✓	✓	NA	NA	✓	*	*	*	*	*	*	*	3/5	60%
Chessler et al., 2009	✓	NA	NA	NA	NA	✓	✓	✓	✓	*	*	NA	5/10	50%
Costales et al., 2009	*	*	*	*	*	✓	✓	✓	✓	*	*	✓	5/5	100%
Garg at al., 2004	✓	✓	✓	NA	NA	*	*	*	*	*	*	*	3/5	60%
Genovesio et al., 2011	*	*	*	*	*	✓	✓	✓	✓	*	*	✓	5/5	100%
Goldenberg et al., 2009	✓	✓	✓	NA	✓	✓	✓	✓	✓	✓	✓	✓	11/12	91.7%
Graefe et al., 2006	✓	✓	NA	NA	NA	*	*	*	*	*	*	*	2/5	40%
Hashimoto et al., 2005	*	*	*	*	*	✓	✓	✓	✓	*	*	✓	5/5	100%
Hill et al., 2005	✓	✓	✓	NA	*	*	*	*	*	*	*	*	3/4	75%
Kierstein et al., 2006	✓	✓	NA	NA	NA	*	*	*	*	*	*	*	2/5	40%
Li et al., 2009	✓	✓	NA	✓	✓	*	*	*	*	*	*	*	4/5	80%
Li et al., 2011	✓	✓	✓	✓	✓	*	*	*	*	*	*	*	5/5	100%
Lopez et al., 2008	✓	✓	✓	NA	NA	✓	✓	✓	✓	✓	✓	*	9/11	81.8%
Manque et al., 2011	✓	✓	✓	NA	NA	✓	✓	✓	✓	✓	✓	✓	10/12	83.3%
Meade et al., 2009	✓	✓	✓	NA	*	*	*	*	*	*	*	*	3/4	75%
Mekata et al., 2012	✓	✓	✓	NA	NA	*	*	*	*	*	*	*	3/5	60%
Mukherjee et al., 2003	✓	✓	✓	NA	NA	*	*	*	*	*	*	*	3/5	60%
Mukherjee et al., 2008	✓	✓	✓	NA	NA	✓	✓	✓	✓	✓	✓	*	9/11	81.2%
Noyes et al., 2009	✓	✓	NA	NA	✓	*	*	*	*	*	*	*	3/5	60%
O'Gorman et al., 2009	✓	✓	✓	NA	*	*	*	*	*	*	*	*	3/4	75%
Soares et al., 2010	✓	NA	✓	NA	NA	*	*	*	*	*	*	*	2/5	40%
Soares et al., 2011	✓	✓	✓	✓	NA	*	*	*	*	*	*	*	4/5	80%
Tanowitz et al., 2011	✓	NA	✓	NA	NA	✓	✓	✓	✓	✓	✓	NA	8/12	66.7%
Total	20/20	17/20	14/20	3/20	5/17	9/9	9/9	9/9	9/9	5/5	5/5	5/7		
%	100%	85%	75%	15%	29.4%	100%	100%	100%	100%	100%	100%	71.4%		

Criteria: H1 (species and strain), H2 (age); H3 (gender), H4 (light and dark cycle), H5 (method of sacrifice), H6 (cell type), H7 (culture medium), H8 (supplements and antibiotics), H9 (temperature and CO_2 atmosphere), H10 (organ or tissue which takes the primary culture), H11 (method of purification for establishing primary culture), and H12 (time of growing of the cell prior infection).
✓: meets the criteria
NA: information not available
*****: not applicable

Table S3. Quality measures of the studies that failed to fulfil any one of data of minimal information about the experimental infection in *Trypanosoma* experiments.

	Characteristics of the experiment													
	Animal models					**Cellular models**				**Parasite**				
Articles	**I1**	**I2**	**I3**	**I4**	**I5**	**I6**	**I7**	**I8**	**I9**	**I10**	**I11**	**I12**	**Total**	**%**
Meade et al., 2009	✓	✓	✓	NA	NA	*	*	*	*	NA	✓	✓	5/8	62.5%
Amin et al., 2010	✓	✓	NA	NA	NA	*	*	*	*	NA	NA	✓	3/8	37.5%
Chessler et al., 2009	✓	✓	✓	✓	NA	*	NA	✓	NA	NA	NA	✓	6/11	54.5%
Costales et al., 2009	*	*	*	*	*	*	NA	NA	✓	NA	NA	✓	2/6	33.3%
Garg at al., 2004	✓	✓	NA	NA	NA	*	*	*	*	NA	NA	✓	3/8	37.5%
Genovesio et al., 2011	*	*	*	*	*	*	NA	✓	NA	NA	NA	✓	3/6	33.3%
Goldenberg et al., 2009	*	*	*	*	*	✓	NA	✓	✓	NA	NA	✓	4/7	57.1%
Graefe et al., 2006	✓	✓	✓	✓	✓	*	*	*	*	NA	✓	✓	7/8	87.5%
Hashimoto et al., 2005	*	*	*	*	*	*	NA	✓	✓	NA	NA	✓	3/6	50%
Hill et al., 2005	✓	✓	✓	NA	NA	*	*	*	*	NA	✓	✓	4/6	62.5%
Kierstein et al., 2006	✓	✓	✓	✓	NA	*	*	*	*	NA	✓	✓	6/8	75%
Li et al., 2009	✓	✓	✓	✓	✓	*	*	*	*	NA	✓	✓	7/8	87.5%
Li et al., 2011	✓	✓	NA	NA	NA	*	*	*	*	NA	✓	✓	4/8	50%
Lopez et al., 2008	✓	✓	✓	✓	✓	NA	NA	*	*	NA	✓	✓	7/10	70%
Manque et al., 2011	*	*	*	*	*	NA	NA	✓	✓	NA	NA	✓	3/7	42.9%
Mekata et al., 2012	✓	✓	NA	✓	✓	*	*	*	*	NA	NA	✓	5/8	62.5%
Mukherjee et al., 2003	✓	NA	NA	✓	✓	*	*	*	*	NA	NA	✓	4/8	50%
Mukherjee et al., 2008	✓	✓	NA	NA	✓	✓	NA	*	*	NA	NA	✓	5/10	50%
Noyes et al., 2009	✓	✓	✓	✓	✓	*	*	*	*	NA	✓	✓	7/8	87.5%
O'Gorman et al., 2009	✓	✓	✓	NA	NA	*	*	*	*	NA	✓	✓	5/8	62.5%
Soares et al., 2010	✓	✓	NA	✓	✓	*	*	*	*	NA	NA	✓	5/8	62.5%
Soares et al., 2011	✓	✓	✓	NA	NA	*	*	*	*	NA	NA	✓	4/8	50%
Tanowitz et al., 2011	*	*	*	*	*	NA	NA	✓	NA	NA	NA	✓	2/7	28.6%
Total	17/17	16/17	10/17	9/17	8/17	2/5	0/9	6/7	4/7	0/23	9/23	23/23		
%	100%	94.1%	58.8%	52.9%	47.1%	40%	0%	85.7%	57.1%	0%	39.1%	100%		

Criteria: I1 (inoculum –parasite per animal), I2 (route of inoculation), I3 (medium of inoculation), I4 (parasitaemia and time post infection when parasitaemia was measured), I5 (mortality of animals post infection), I6 (purity of primary culture), I7 (viability of the cells prior to infection), I8 (ratio –parasites per cell), I9 (percentage infected cells), I10 (viability of the parasite prior to infection), I11 (purity of the infective form of the parasite), and I12 (duration of infection).

✓: meets the criteria

NA: information not available

*: not applicable

Table S4. Bibliometric indices in reporting *Trypanosoma* experiments.

Journal			Citations			Author		
Name	**IF**	**Topic**	**Reference**	**Web of Science**	**Google Scholar**	**Corresponding author**	**h-index**	**h-index***
Am J Trop Med Hyg	2.592	172	Amin et al., 2010	13	17	Daniel N. Amin	3	3
Biochem J	4.897	6	Garg at al., 2004	20	19	N. Garg	6	5
BMC Genomics	4.073	6	Costales et al., 2009	7	12	Barbara A. Burleigh	14	13
BMC Genomics	4.073	6	O'Gorman et al., 2009	8	9	David E. MacHugh	13	2
Cell Cycle	5.359	125	Soares et al., 2011	10	13	Milena Soares	20	13
Cell Cycle	5.359	125	Tanowitz et al., 2011	0	1	H. B. Tanowitz	33	20
Exp Parasitol	2.122	150	Li et al., 2011	0	0	Zhao-Rong Lun	15	7
Genes Immun	3.872	3	Kierstein et al., 2006	10	12	S. Kierstein	1	1
Genomics	3.019	68	Mukherjee et al., 2008	10	16	H. B. Tanowitz	33	20
Infect Immun	4.165	93	Manque et al., 2011	11	14	Gregory A. Buck	20	12
Int J Parasitol	3.393	95	Hashimoto et al., 2005	2	2	J. Nakajima-Shimada	8	8
J Immunol	5.788	54	Chessler et al., 2009	8	11	Barbara A. Burleigh	14	13
J Immunol	5.788	54	Lopez et al., 2008	10	9	Donna M. Paulnock	12	5
J Infect Dis	6.410	60	Soares et al., 2010	11	15	Milena Soares	19	13
Microbes Infect	3.101	67	Goldenberg et al., 2009	6	8	David Spray	48	5
Mol Immunol	2.897	10	Meade et al., 2009	1	3	David E. MacHugh	13	2
Parasite Immunol	2.601	35	Mekata et al., 2012	0	0	Kazuhiko Ohashi	23	5
Parasitol Res	2.149	109	Li et al., 2009	3	4	Zhao-Rong Lun	14	5
Parasitol Res	2.149	109	Mukherjee et al., 2003	37	51	H. B. Tanowitz	25	13
PLoS One	4.092	54	Genovesio et al., 2011	6	9	Lucio H. Freitas-Junior	14	7
PLoS One	4.092	54	Graefe et al., 2006	3	5	Sebastian Graefe	5	3
PLoS One	4.092	54	Noyes et al., 2009	13	17	Jan Naessens	13	10
Vet Immunol Immunopathol	2.076	7	Hill et al., 2005	27	40	David E. MacHugh	9	2

Topic: articles published per journal about "Trypanosomiasis"[MeSH]
*****: h-index filtered by topic using the term Trypanosom*

Table S5. Quality measures of the studies that failed to supply any one of the criteria for minimal information about the parasite in *Leishmania*, *Toxoplasma*, *Plasmodium*, *Trichuris*, *Schistosoma* and *Mycobacterium* experiments.

					Culture conditions of *Leishmania, Toxoplasma, Plasmodium, Trichuris, Schistosoma* and *Mycobacterium*										
		Parasite information			**Parasites from animals**				**Parasites from cells**						
Articles	**Model**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**	**P12**	**Total**	**%**
Park et al., 2000	L	✓	✓	✓	*	*	*	*	*	✓	✓	NA	NA	5/7	71.4%
Filippi et al., 2003	L	✓	✓	✓	*	*	*	*	*	✓	✓	NA	✓	6/7	85.7%
Bertholet et al., 2005	L	✓	✓	✓	*	*	*	*	*	✓	✓	NA	✓	6/7	85.7%
Kinjyo et al., 2006	L	✓	✓	✓	✓	NA	NA	NA	*	✓	✓	NA	NA	6/11	54.5%
Brunner et al., 2007	L	✓	✓	✓	✓	NA	NA	NA	*	✓	✓	✓	NA	7/11	63.6%
Guerfali et al., 2008	L	✓	✓	✓	*	*	*	*	*	✓	✓	✓	✓	7/7	100%
Jayakumar et al., 2008	L	✓	✓	✓	✓	NA	NA	✓	*	✓	✓	✓	✓	9/11	81.8%
Ehrchen at al., 2010	L	✓	✓	✓	*	*	*	*	*	✓	✓	NA	NA	5/7	71.4%
Biswas et al., 2011	L	✓	✓	✓	*	*	*	*	*	✓	✓	NA	NA	5/7	71.4%
de Carvalho et al., 2011	L	✓	✓	✓	*	*	*	*	*	✓	✓	✓	✓	7/7	100%
Desolme et al., 2000	T	✓	✓	✓	✓	NA	NA	✓	*	*	*	*	NA	5/8	62.5%
Gail et al., 2001	T	✓	✓	✓	*	*	*	*	✓	✓	✓	NA	NA	6/8	75%
Fux et al., 2003	T	✓	✓	✓	✓	NA	✓	✓	*	*	*	*	✓	7/8	87.5%
Tato et al., 2003	T	✓	✓	✓	✓	NA	NA	✓	*	*	*	*	NA	5/8	62.5%
Okomo et al, 2006	T	✓	✓	✓	*	*	*	*	✓	✓	✓	✓	NA	7/8	87.5%
Knight et al., 2006	T	✓	✓	✓	*	*	*	*	✓	✓	✓	NA	✓	7/8	87.5%
Watford et al., 2008	T	✓	✓	✓	NA	NA	NA	NA	*	*	*	*	NA	3/8	37.5%
Ju et al., 2009	T	✓	NA	✓	✓	NA	NA	✓	✓	✓	✓	NA	NA	7/12	58.3%
Fang et al., 2009	T	✓	✓	✓	✓	NA	NA	*	*	*	*	*	✓	5/7	71.4%
Zhuo et al., 2011	T	✓	✓	✓	NA	NA	NA	NA	*	*	*	*	NA	3/8	37.5%
Ylostalo et al., 2005	P	✓	NA	✓	✓	NA	NA	✓	*	*	*	*	NA	4/8	50%
Lovergrove et al., 2006	P	✓	✓	✓	✓	NA	NA	*	*	*	*	*	NA	4/7	57.1%
Delahaye et al., 2007	P	✓	✓	✓	✓	NA	NA	✓	*	*	*	*	✓	6/8	75%
Carapau et al., 2007	P	✓	NA	✓	NA	NA	NA	NA	*	*	*	*	NA	2/8	25%
Miu et al., 2008	P	✓	✓	✓	NA	NA	NA	NA	*	*	*	*	NA	3/8	37.5%
Randall et al., 2008	P	✓	✓	✓	NA	NA	NA	✓	*	*	*	*	NA	4/8	50%
Oakley et al., 2008	P	✓	✓	NA	NA	NA	NA	NA	*	*	*	*	NA	2/8	25%
Albuquerque et al, 2009	P	✓	✓	✓	✓	NA	NA	✓	*	*	*	*	NA	5/8	62.5%
Delic et al., 2011	P	✓	NA	✓	✓	NA	NA	NA	*	*	*	*	NA	3/8	37.5%
Rosanas et al., 2012	P	✓	✓	✓	NA	NA	NA	✓	*	*	*	*	NA	4/8	50%
Betts et al., 2001	C	✓	✓	✓	✓	✓	✓	*	*	*	*	*	✓	7/7	100%
Humphreys et al., 2004	C	✓	✓	✓	✓	✓	✓	*	*	*	*	*	✓	7/7	100%
Cliffe et al.,2005	C	✓	✓	✓	✓	✓	✓	*	*	*	*	*	✓	7/7	100%
Dixon et al., 2006	C	✓	✓	✓	✓	✓	✓	*	*	*	*	*	✓	7/7	100%
Bickle et al., 2007	C	✓	✓	✓	✓	✓	✓	*	*	*	*	*	✓	7/7	100%
Villarino et al., 2008	C	✓	NA	✓	NA	NA	NA	*	*	*	*	*	✓	3/7	42.9%
Massacand et al. 2009	C	✓	✓	✓	✓	✓	✓	*	*	*	*	*	✓	7/7	100%
Svensson et al. 2009	C	✓	✓	✓	✓	✓	✓	*	*	*	*	*	✓	7/7	100%
Hepworth et al. 2009	C	✓	✓	✓	✓	✓	✓	*	*	*	*	*	✓	7/7	100%
Hasnain et al. 2010	C	✓	✓	✓	✓	✓	✓	*	*	*	*	*	✓	7/7	100%
Angyalosi et al. 2001	S	✓	✓	✓	✓	NA	*	*	*	*	*	*	Y	4/5	80%
Byström et al. 2006	S	✓	✓	✓	✓	NA	*	*	*	*	*	*	NA	4/6	66.7%
Singh et al. 2006	S	✓	✓	✓	✓	NA	*	*	*	*	*	*	NA	4/6	66.7%
Burke et al. 2010	S	✓	✓	✓	NA	NA	*	*	*	*	*	*	NA	3/6	50%
de Oliveira et al. 2010	S	✓	✓	✓	Y	NA	*	*	*	*	*	*	✓	4/5	80%
Burke et al. 2011	S	✓	✓	✓	NA	NA	*	*	*	*	*	*	NA	3/6	50%
Perry et al. 2011	S	✓	✓	✓	NA	NA	*	*	*	*	*	*	NA	3/6	50%
Zhang et al. 2011	S	✓	✓	✓	✓	NA	*	*	*	*	*	*	NA	4/6	66.7%
Ray et al. 2012	S	✓	NA	✓	✓	NA	NA	*	*	*	*	*	✓	4/7	57.1%
de la Torre et al. 2012	S	✓	NA	✓	NA	NA	*	*	*	*	*	*	NA	2/6	33.3%
Ragno et al., 2001	TBC	✓	✓	*	*	*	*	*	*	✓	*	✓	NA	4/5	80%
Xu et al., 2003	TBC	✓	✓	*	*	*	*	*	*	✓	✓	✓	✓	6/6	100%
Keller et al., 2004	TBC	✓	✓	*	*	*	*	*	*	✓	✓	✓	✓	6/6	100%
Volpe et al., 2006	TBC	✓	✓	*	*	*	*	*	*	✓	*	NA	✓	4/5	80%
Orlova et al., 2006	TBC	✓	✓	*	*	*	*	*	*	NA	NA	NA	NA	2/6	33.3%
Silver et al., 2009	TBC	✓	✓	*	*	*	*	*	*	✓	✓	✓	✓	6/6	100%
Maddocks et al., 2009	TBC	✓	✓	*	*	*	*	*	*	✓	*	NA	NA	3/5	60%
Beisiegel et al., 2009	TBC	✓	✓	*	*	*	*	*	*	✓	✓	NA	✓	5/6	83.3%
Sharbati et al., 2011	TBC	✓	✓	*	*	*	*	*	*	✓	✓	✓	✓	6/6	100%
Magee et al., 2012	TBC	✓	✓	*	*	*	*	*	*	✓	✓	✓	✓	6/6	100%
Total		60/60	53/60	49/50	27/39	9/40	10/31	10/18	4/4	23/24	20/21	11/24	28/59		
%		100%	88.3%	98.0%	69.2%	22.5%	32.3%	55.6%	100%	95.8%	95.2%	45.8%	47.5%		

L = *Leishmania*, T = *Toxoplasma*, P = *Plasmodium*, C = colitis induced by *Trichuris*, S = *Schistosoma* and TBC = tuberculosis. Criteria: P1 (species), P2 (strain), P3 (stage), P4 (species and strain), P5 (age), P6 (gender), P7 (parasite collection sample); P8 (cell type), P9 (culture medium), P10 (supplements and antibiotics), P11 (temperature and CO_2 atmosphere), and P12 (time of growing of the parasite prior to infection).

✓: meets the criteria

NA: information not available

*: not applicable

Table S6. Quality measures of the studies that failed to supply any one of the criteria for minimal information about the host in *Leishmania*, *Toxoplasma*, *Plasmodium*, *Trichuris*, *Schistosoma* and *Mycobacterium* experiments.

		Characteristics and culture conditions of the host models													
		Animal model					**Cell model**								
Articles	**Model**	**H1**	**H2**	**H3**	**H4**	**H5**	**H6**	**H7**	**H8**	**H9**	**H10**	**H11**	**H12**	**Total**	**%**
Park et al., 2000	L	✓	✓	✓	NA	NA	✓	✓	✓	NA	✓	NA	*	7/11	63.6%
Filippi et al., 2003	L	✓	✓	NA	NA	NA	*	*	*	*	*	*	*	2/5	40%
Bertholet et al., 2005	L	✓	NA	NA	NA	NA	*	*	*	*	*	*	*	1/5	20%
Kinjyo et al., 2006	L	✓	✓	NA	NA	NA	*	*	*	*	*	*	*	2/5	40%
Brunner et al., 2007	L	✓	✓	NA	NA	NA	*	*	*	*	*	*	*	2/5	40%
Guerfali et al., 2008	L	*	*	*	*	*	✓	✓	✓	✓	✓	✓	✓	7/7	100%
Jayakumar et al., 2008	L	*	*	*	*	*	✓	*	*	✓	✓	✓	NA	4/5	80%
Ehrchen at al., 2010	L	✓	✓	✓	NA	NA	*	*	*	*	*	*	*	3/5	60%
Biswas et al., 2011	L	✓	✓	✓	NA	NA	✓	✓	✓	✓	✓	✓	NA	9/12	75%
de Carvalho et al., 2011	L	✓	NA	NA	NA	NA	✓	✓	NA	✓	✓	✓	✓	7/12	58.3%
Desolme et al., 2000	T	✓	✓	✓	NA	NA	*	*	*	*	*	*	*	3/5	60%
Gail et al., 2001	T	*	*	*	*	*	✓	✓	✓	NA	*	*	NA	3/5	60%
Fux et al., 2003	T	✓	✓	NA	NA	✓	*	*	*	*	*	*	*	3/5	60%
Tato et al., 2003	T	✓	✓	NA	NA	NA	*	*	*	*	*	*	*	2/5	40%
Okomo et al, 2006	T	*	*	*	*	*	✓	✓	✓	✓	*	*	NA	4/5	80%
Knight et al., 2006	T	*	*	*	*	*	✓	✓	✓	NA	*	*	NA	3/5	60%
Watford et al., 2008	T	✓	NA	✓	NA	NA	*	*	*	*	*	*	*	2/5	40%
Ju et al., 2009	T	*	*	*	*	*	✓	✓	✓	NA	*	*	NA	3/5	60%
Fang et al., 2009	T	✓	✓	✓	NA	NA	*	*	*	*	*	*	*	3/5	60%
Zhuo et al., 2011	T	✓	NA	NA	NA	NA	*	*	*	*	*	*	*	1/5	20%
Ylostalo et al., 2005	P	✓	NA	NA	NA	*	*	*	*	*	*	*	*	1/4	25%
Lovergrove et al., 2006	P	✓	✓	✓	NA	NA	*	*	*	*	*	*	*	3/5	60%
Delahaye et al., 2007	P	✓	✓	✓	NA	*	*	*	*	*	*	*	*	3/4	75%
Carapau et al., 2007	P	✓	NA	NA	NA	NA	✓	✓	✓	NA	✓	✓	*	6/11	54.5%
Miu et al., 2008	P	✓	✓	✓	NA	✓	*	*	*	*	*	*	*	4/5	80%
Randall et al., 2008	P	✓	✓	✓	NA	✓	*	*	*	*	*	*	*	4/5	80%
Oakley et al., 2008	P	✓	✓	✓	NA	NA	*	*	*	*	*	*	*	3/5	60%
Albuquerque et al, 2009	P	✓	✓	✓	✓	NA	✓	✓	✓	✓	✓	✓	✓	11/12	91.7%
Delic et al., 2011	P	✓	NA	✓	NA	✓	*	*	*	*	*	*	*	3/5	60%
Rosanas et al., 2012	P	✓	✓	✓	✓	NA	*	*	*	*	*	*	*	4/5	80%
Betts et al., 2001	C	✓	✓	NA	✓	NA	*	*	*	*	*	*	*	3/5	60%
Humphreys et al., 2004	C	✓	✓	NA	✓	NA	*	*	*	*	*	*	*	3/5	60%
Cliffe et al.,2005	C	✓	✓	✓	✓	NA	*	*	*	*	*	*	*	4/5	80%
Dixon et al., 2006	C	✓	✓	✓	✓	✓	*	*	*	*	*	*	*	5/5	100%
Bickle et al., 2007	C	✓	✓	✓	✓	NA	*	*	*	*	*	*	*	4/5	80%
Villarino et al., 2008	C	✓	✓	✓	✓	NA	*	*	*	*	*	*	*	4/5	80%
Massacand et al. 2009	C	✓	NA	NA	✓	NA	*	*	*	*	*	*	*	2/5	40%
Svensson et al. 2009	C	✓	✓	NA	✓	NA	*	*	*	*	*	*	*	3/5	60%
Hepworth et al. 2009	C	✓	✓	✓	✓	NA	*	*	*	*	*	*	*	4/5	80%
Hasnain et al. 2010	C	✓	✓	✓	✓	NA	*	*	*	*	*	*	*	4/5	80%
Angyalosi et al. 2001	S	✓	✓	NA	✓	NA	*	*	*	*	*	*	*	3/5	60%
Byström et al. 2006	S	✓	NA	NA	NA	NA	*	*	*	*	*	*	*	1/5	20%
Singh et al. 2006	S	✓	✓	✓	✓	NA	*	*	*	*	*	*	*	4/5	80%
Burke et al. 2010	S	✓	✓	✓	✓	NA	*	*	*	*	*	*	*	4/5	80%
de Oliveira et al. 2010	S	✓	N	N	✓	NA	*	*	*	*	*	*	*	2/3	66.7%
Burke et al. 2011	S	✓	✓	✓	NA	NA	*	*	*	*	*	*	*	3/5	60%
Perry et al. 2011	S	✓	✓	✓	NA	NA	*	*	*	*	*	*	*	3/5	60%
Zhang et al. 2011	S	✓	✓	✓	Y	NA	*	*	*	*	*	*	*	3/4	75%
Ray et al. 2012	S	✓	✓	✓	NA	NA	*	*	*	*	*	*	*	3/5	60%
de la Torre et al. 2012	S	✓	NA	NA	NA	✓	*	*	*	*	*	*	*	2/5	40%
Ragno et al., 2001	TBC	*	*	*	*	*	✓	*	*	✓	✓	NA	NA	3/5	60%
Xu et al., 2003	TBC	*	*	*	*	*	✓	✓	NA	✓	✓	✓	NA	5/7	71.4%
Keller et al., 2004	TBC	✓	✓	✓	✓	NA	✓	✓	NA	✓	✓	NA	✓	9/12	75%
Volpe et al., 2006	TBC	✓	NA	NA	*	*	✓	✓	✓	✓	✓	✓	✓	8/10	80%
Orlova et al., 2006	TBC	*	*	*	*	*	✓	✓	NA	✓	✓	NA	NA	4/7	57.1%
Silver et al., 2009	TBC	✓	NA	NA	*	*	✓	✓	✓	✓	✓	*	✓	7/9	77.8%
Maddocks et al., 2009	TBC	✓	NA	NA	*	*	✓	✓	✓	✓	✓	NA	✓	7/10	70%
Beisiegel et al., 2009	TBC	✓	NA	NA	✓	NA	*	*	*	*	*	*	*	2/5	40%
Sharbati et al., 2011	TBC	✓	NA	NA	*	*	✓	✓	✓	✓	✓	✓	✓	8/10	80%
Magee et al., 2012	TBC	✓	✓	✓	✓	*	✓	✓	✓	✓	✓	✓	✓	11/11	100%
Total		51/51	35/50	28/50	19/46	6/44	20/20	18/18	14/18	15/20	16/16	10/15	9/18		
%		100%	70%	56%	41.3%	13.6%	100%	100%	77.8%	75%	100%	66.7%	50%		

L = *Leishmania*, T = *Toxoplasma*, P = *Plasmodium*, C = colitis induced by *Trichuris*, S = *Schistosoma* and TBC = tuberculosis. Criteria: H1 (species and strain), H2 (age); H3 (gender), H4 (light and dark cycle), H5 (method of sacrifice), H6 (cell type), H7 (culture medium), H8 (supplements and antibiotics), H9 (temperature and CO_2 atmosphere), H10 (organ or tissue which takes the primary culture), H11 (method of purification for establishing primary culture), and H12 (time of growing of the cell prior to infection).

✓: meets the criteria

NA: information not available

*****: not applicable

Table S7. Quality measures of the studies that failed to supply any one of the criteria for minimal information about the experimental infection in *Leishmania*, *Toxoplasma*, *Plasmodium*, *Trichuris*, *Schistosoma* and *Mycobacterium* experiments.

		Characteristics of the experiment													
		Animal models					**Cellular models**				**Parasite**				
Articles	**Model**	**I1**	**I2**	**I3**	**I4**	**I5**	**I6**	**I7**	**I8**	**I9**	**I10**	**I11**	**I12**	**Total**	**%**
Park et al., 2000	L	✓	✓	NA	✓	NA	NA	*	*	*	NA	✓	✓	5/9	55.6%
Filippi et al., 2003	L	✓	✓	NA	NA	NA	✓	*	*	*	NA	NA	✓	4/9	44.4%
Bertholet et al., 2005	L	✓	✓	NA	✓	NA	*	NA	✓	NA	✓	✓	✓	7/11	63.6%
Kinjyo et al., 2006	L	✓	✓	NA	✓	NA	*	*	*	*	NA	NA	✓	4/8	50%
Brunner et al., 2007	L	✓	✓	✓	✓	NA	*	*	*	*	NA	Y	✓	5/7	71.4%
Guerfali et al., 2008	L	*	*	*	*	*	✓	NA	✓	✓	NA	NA	✓	4/7	57.1%
Jayakumar et al., 2008	L	*	*	*	*	*	*	NA	✓	✓	✓	✓	✓	5/6	83.3%
Ehrchen at al., 2010	L	✓	✓	✓	NA	NA	*	*	*	*	NA	NA	✓	4/8	50%
Biswas et al., 2011	L	✓	✓	NA	NA	NA	NA	NA	*	*	NA	NA	✓	3/10	30%
de Carvalho et al., 2011	L	*	*	*	*	*	NA	NA	✓	✓	NA	✓	✓	4/7	57.1%
Desolme et al., 2000	T	✓	✓	NA	✓	✓	*	*	*	*	NA	✓	✓	6/8	75%
Gail et al., 2001	T	*	*	*	*	*	*	NA	✓	✓	NA	✓	✓	4/6	66.7%
Fux et al., 2003	T	✓	✓	✓	✓	✓	*	*	*	*	NA	✓	✓	7/8	87.5%
Tato et al., 2003	T	✓	✓	NA	✓	✓	*	*	*	*	NA	✓	✓	6/8	75%
Okomo et al, 2006	T	*	*	*	*	*	*	NA	✓	NA	NA	✓	✓	3/6	50%
Knight et al., 2006	T	*	*	*	*	*	*	NA	✓	NA	NA	✓	✓	3/6	50%
Watford et al., 2008	T	✓	✓	NA	✓	✓	*	*	*	*	NA	✓	✓	6/8	75%
Ju et al., 2009	T	*	*	*	*	*	*	NA	✓	NA	NA	✓	✓	3/6	50%
Fang et al., 2009	T	✓	✓	✓	NA	✓	*	*	*	*	NA	✓	✓	6/8	75%
Zhuo et al., 2011	T	✓	NA	NA	NA	NA	*	*	*	*	NA	✓	✓	3/8	37.5%
Ylostalo et al., 2005	P	✓	✓	✓	✓	NA	*	*	*	*	NA	✓	✓	6/8	75%
Lovergrove et al., 2006	P	✓	✓	✓	NA	NA	*	*	*	*	NA	✓	✓	5/8	62.5%
Delahaye et al., 2007	P	✓	✓	✓	✓	✓	*	*	*	*	NA	✓	✓	7/8	87.5%
Carapau et al., 2007	P	✓	✓	✓	✓	NA	✓	NA	✓	NA	NA	✓	✓	8/12	66.7%
Miu et al., 2008	P	✓	✓	✓	NA	NA	*	*	*	*	NA	NA	✓	4/8	50%
Randall et al., 2008	P	✓	NA	✓	✓	✓	*	*	*	*	NA	✓	✓	6/8	75%
Oakley et al., 2008	P	✓	✓	NA	✓	NA	*	*	*	*	NA	NA	NA	3/8	37.5%
Albuquerque et al, 2009	P	✓	✓	NA	NA	NA	NA	NA	✓	NA	NA	✓	✓	5/12	41.7%
Delic et al., 2011	P	✓	✓	✓	✓	✓	*	*	*	*	NA	✓	✓	7/8	87.5%
Rosanas et al., 2012	P	✓	✓	✓	✓	✓	*	*	*	*	NA	✓	✓	7/8	87.5%
Betts et al., 2001	C	✓	✓	NA	✓	*	*	*	*	*	✓	*	✓	5/6	83.3%
Humphreys et al., 2004	C	✓	✓	NA	✓	*	*	*	*	*	✓	*	✓	5/6	83.3%
Cliffe et al.,2005	C	✓	✓	✓	✓	*	*	*	*	*	✓	*	✓	6/6	100%
Dixon et al., 2006	C	✓	✓	✓	✓	*	*	*	*	*	✓	*	✓	6/6	100%
Bickle et al., 2007	C	✓	✓	✓	✓	*	*	*	*	*	✓	*	✓	6/6	100%
Villarino et al., 2008	C	✓	✓	NA	✓	*	*	*	*	*	NA	*	✓	4/6	66.7%
Massacand et al. 2009	C	✓	✓	✓	✓	*	*	*	*	*	✓	*	✓	6/6	100%
Svensson et al. 2009	C	✓	✓	✓	✓	*	*	*	*	*	✓	*	✓	6/6	100%
Hepworth et al. 2009	C	✓	✓	✓	✓	*	*	*	*	*	✓	*	✓	6/6	100%
Hasnain et al. 2010	C	✓	✓	✓	✓	*	*	*	*	*	✓	*	✓	6/6	100%
Angyalosi et al. 2001	S	✓	✓	NA	✓	✓	*	*	*	*	NA	✓	✓	6/8	75%
Byström et al. 2006	S	✓	✓	NA	NA	NA	*	*	*	*	NA	✓	✓	4/8	50%
Singh et al. 2006	S	✓	✓	NA	NA	NA	*	*	*	*	NA	✓	✓	4/8	50%
Burke et al. 2010	S	✓	✓	✓	Y	NA	*	*	*	*	NA	✓	✓	5/7	71.4%
de Oliveira et al. 2010	S	✓	✓	✓	NA	NA	*	*	*	*	NA	✓	✓	5/8	62.5%
Burke et al. 2011	S	✓	✓	NA	✓	NA	*	*	*	*	NA	✓	✓	5/8	62.5%
Perry et al. 2011	S	✓	✓	NA	✓	NA	*	*	*	*	NA	✓	✓	5/8	62.5%
Zhang et al. 2011	S	✓	✓	✓	✓	NA	*	*	*	*	NA	✓	✓	6/8	75%
Ray et al. 2012	S	✓	✓	Y	NA	NA	*	*	*	*	NA	✓	✓	4/7	57.1%
de la Torre et al. 2012	S	✓	✓	NA	NA	NA	*	*	*	*	NA	✓	✓	4/8	50%
Ragno et al., 2001	TBC	*	*	*	*	*	*	NA	✓	NA	✓	*	✓	3/5	60%
Xu et al., 2003	TBC	*	*	*	*	*	*	NA	✓	✓	NA	*	✓	3/5	60%
Keller et al., 2004	TBC	*	*	*	*	*	NA	✓	✓	✓	✓	*	✓	5/6	83.3%
Volpe et al., 2006	TBC	*	*	*	*	*	✓	NA	✓	✓	✓	*	✓	5/6	83.3%
Orlova et al., 2006	TBC	*	*	*	*	*	NA	NA	✓	✓	NA	*	✓	3/6	50%
Silver et al., 2009	TBC	*	*	*	*	*	✓	NA	✓	✓	NA	*	✓	4/6	66.7%
Maddocks et al., 2009	TBC	*	*	*	*	*	NA	NA	✓	NA	NA	*	✓	2/6	33.3%
Beisiegel et al., 2009	TBC	✓	✓	✓	✓	NA	*	*	*	*	✓	*	✓	6/7	85.7%
Sharbati et al., 2011	TBC	*	*	*	*	*	✓	NA	✓	NA	✓	*	✓	4/6	66.7%
Magee et al., 2012	TBC	*	*	*	*	*	✓	NA	✓	NA	✓	*	✓	4/6	66.7%
Total		44/44	42/44	23/43	30/43	10/34	7/14	1/20	19/19	9/19	17/60	32/39	59/60		
%		100%	95.5%	53.5%	69.8%	29.4%	50%	5%	100%	47.4%	28.3%	82.1%	98.3%		

L = *Leishmania*, T = *Toxoplasma*, P = *Plasmodium*, C = colitis induced by *Trichuris*, S = *Schistosoma* and TBC = tuberculosis. Criteria: I1 (inoculum –parasite per animal), I2 (way of inoculation), I3 (medium of inoculation), I4 (parasitaemia and post infection time in which parasitaemia was measured), I5 (mortality of animals post infection), I6 (purity of primary culture), I7 (viability of the cells prior infection), I8 (ratio –parasites per cell), I9 (percentage infected cells), I10 (viability of the parasite prior infection), I11 (purity of the infective form of the parasite), and I12 (duration of infection).

✓: meets the criteria

NA: information not available

*****: not applicable

A.2. Supplementary files Chapter 4

miniRECH:

A consensus framework for scoring the quality of methods reporting via checklists

Authors

Oscar Flórez-Vargas, Binling Jin, Michael Bramhall, Robert Stevens, Andy Brass and Suzanne Embury

School of Computer Science
The University of Manchester
Kilburn building
Oxford Road
Manchester
M13 9PL
United Kingdom

miniRECH philosophy

“*If you cannot measure it, you cannot improve it*.” – Lord Kelvin

The goal of miniRECH [minimal REporting CHecklist] is to provide a tool for performing metadata quality assessment via checklist. The term metadata refers to descriptive information about data and which is essential for appropriate interpretation of a given data set. There are several checklists that have been developed by the ‘minimum information’ community to address adequacy of reporting metadata from a range of data types including genomics, transcriptomics, proteomics, and metabolomics. In this way, we created miniRECH based on the recommendation to develop minimum information guidelines of the Minimum Information for Biological and Biomedical Investigations (MIBBI) project[1]. The miniRECH is a flexible spreadsheet format useful not only for new submissions to scientific journals, but also for scoring the published scientific literature. Thus, the miniRECH will also help to make systematic reviews and meta-analyses of publications more accurate; reducing selection bias.

About miniRECH-*Trypanosoma*

The main purpose of adapting miniRECH to *Trypanosoma* experiments, for instance, is to provide a tool for performing metadata quality assessment for this kind of experiments in order to improve the quality of reporting *Trypanosoma* experimental details, helping to make these experiments more reproducible and comparable and allowing better re-use of experimental results[2]. Moreover, we strongly recommend to all scientists in the parasitology community and to editors of journals publishing *Trypanosoma* studies to take a closer look at the contents of this checklist.

System requirements

- Windows XP/Vista/7/8
- Microsoft Excel 2007, 2010 or 2013.

[1] Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, et al. (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 26: 889-896.

[2] Flórez-Vargas O, Bramhall M, Noyes H, Cruickshank S, Stevens R, Brass A. (2014). The quality of methods reporting in parasitology experiments. PLOS one 9(7): e101131.

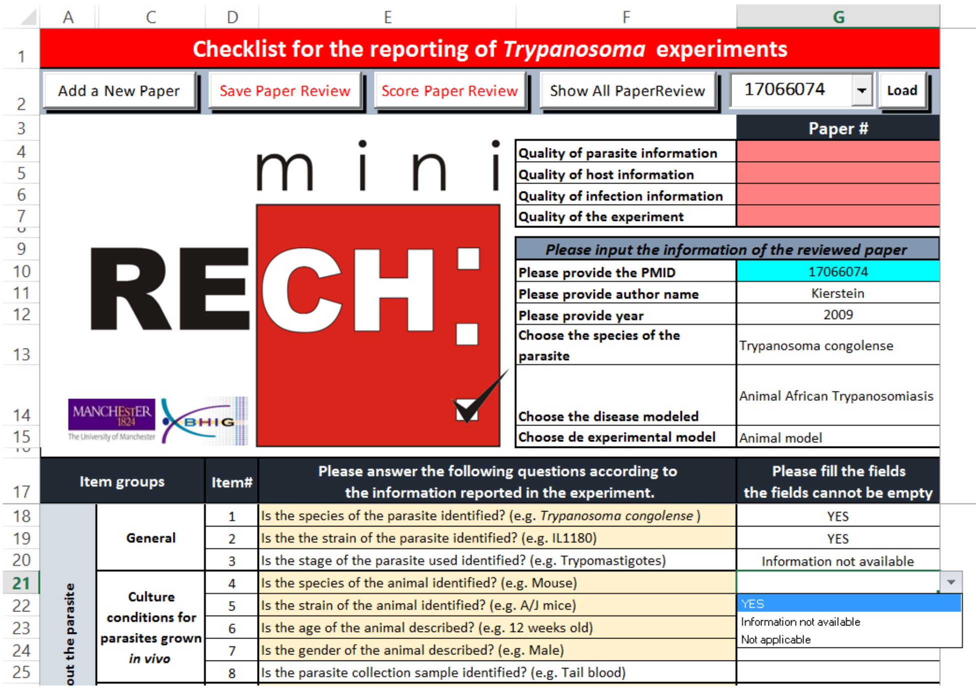
OPENING THE SPREADSHEET

1. Start Microsoft Excel.

2. When opening the spreadsheet, Excel will warn you than some active content in this document has been disabled. You will need to ‘Enable content’ otherwise nothing will work.

3. After a short amount of time, a new spreadsheet will be visible with a blank checklist (**Figure 1**).

4. You might also want to save your spreadsheet to a file on your PC. You can browse your hard disk and save your spreadsheet. You may want to rename the file in order to not overwrite your original file and maintain the source history.



Checklist for the reporting of *Trypanosoma* experiments

	Paper #
Quality of parasite information	
Quality of host information	
Quality of infection information	
Quality of the experiment	
Please input the information of the reviewed paper	
Please provide the PMID	17066074
Please provide author name	Kierstein
Please provide year	2009
Choose the species of the parasite	Trypanosoma congolense
Choose the disease modeled	Animal African Trypanosomiasis
Choose de experimental model	Animal model

Item groups	Item#	Please answer the following questions according to the information reported in the experiment.	Please fill the fields the fields cannot be empty
General	1	Is the species of the parasite identified? (e.g. *Trypanosoma congolense*)	YES
	2	Is the the strain of the parasite identified? (e.g. IL1180)	YES
	3	Is the stage of the parasite used identified? (e.g. Trypomastigotes)	Information not available
Culture conditions for parasites grown *in vivo*	4	Is the species of the animal identified? (e.g. Mouse)	
	5	Is the strain of the animal identified? (e.g. A/J mice)	
	6	Is the age of the animal described? (e.g. 12 weeks old)	
	7	Is the gender of the animal described? (e.g. Male)	
	8	Is the parasite collection sample identified? (e.g. Tail blood)	

Figure 1. A section of the Spreadsheet.

SCORING AN ARTICLE

1. Ensure that you have an active spreadsheet.

2. Press the “**Add a New Paper**” button shown in **Figure 2**. This button creates an empty column representing the checklist for the article.

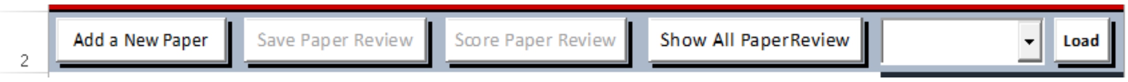
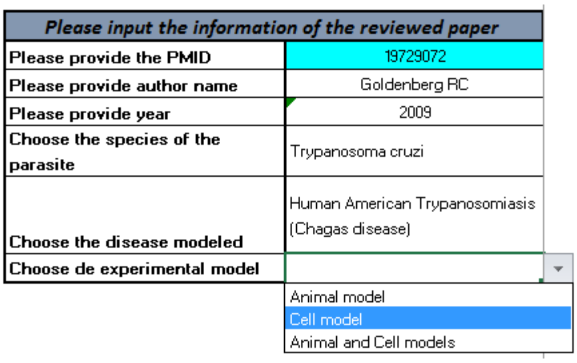


Figure 2. Menu buttons

3. Manually enter the PMID (PubMed Identifier), author and year of the article to be scored. Alternatively, you may enter a unique identifying number of your choosing for each article.

4. Enter the parasite species, disease and experimental model by using the drop-down lists as shown in **Figure 3**.

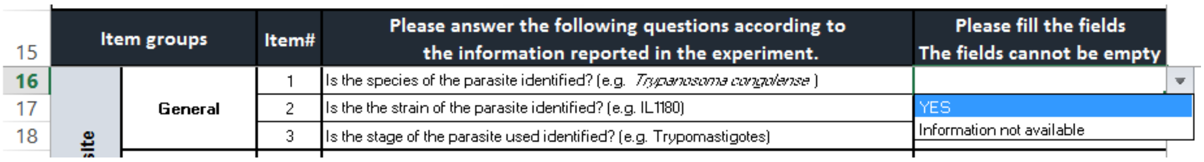


Please input the information of the reviewed paper	
Please provide the PMID	19729072
Please provide author name	Goldenberg RC
Please provide year	2009
Choose the species of the parasite	Trypanosoma cruzi
Choose the disease modeled	Human American Trypanosomiasis (Chagas disease)
Choose de experimental model	

Animal model
Cell model
Animal and Cell models

Figure 3. Fields for paper categorisation

5. Using the drop-down lists as shown in **Figure 4** fill the empty fields in the checklist according to what is reported in the article. Fields must not be left empty. You will be shown a maximum of three options in these fields: '**Yes**' when the article meets the item requested or '**Information not available**' when it does not. You should choose '**Not applicable**' for those items that do not apply for a particular experiment. For example, an experiment was conducted to determine the mechanisms controlling anaemia in *T. congolense* infected mice[3]. If the species of the parasite is identified, you should chose '**Yes**' in the item 1; if the gender of the mice is not reported, you should chose '**Information not available**' in the item 19. Finally, due to the fact that the experiment was carried out in mice, the information about Cell as experimental model (items 22-30 and 36-39) should be '**Not applicable**'.



	Item groups		Item#	Please answer the following questions according to the information reported in the experiment.	Please fill the fields The fields cannot be empty
15					
16	site	General	1	Is the species of the parasite identified? (e.g. *Trypanosoma congolense*)	
17			2	Is the the strain of the parasite identified? (e.g. IL1180)	YES
18			3	Is the stage of the parasite used identified? (e.g. Trypomastigotes)	Information not available

Figure 4. Drop-down lists for paper assessment

6. After filling the checklist, the "**Save Paper Review**" and "**Score Paper Review**" buttons become active as shown in **Figure 5**. Press the "**Save Paper Review**" button and then press "**Score Paper Review**" button.

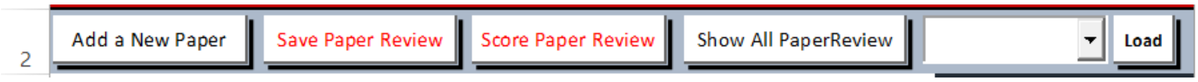
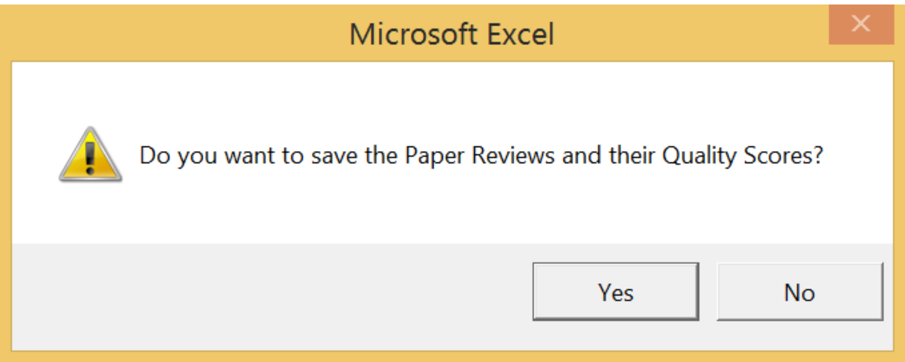


Figure 5. Active menu buttons

7. When scoring the paper, this selection will prompt you to save the paper review and their quality scores as shown in **Figure 6**. You should click 'Yes' if you want so. Otherwise click 'No'. When you decide to save the paper review, the attributes will be permanently saved until you decide to edit them; otherwise they will be temporarily saved.

[3] Noyes HA, Alimohammadian MH, Agaba M, Brass A, Fuchs H, Gailus-Durner V, Hulme H, Iraqi F, Kemp S, Rathkolb B, Wolf E, de Angelis MH, Roshandel D, Naessens J. Mechanisms controlling anaemia in Trypanosoma congolense infected mice. PLoS One. 2009;4(4):e5170.

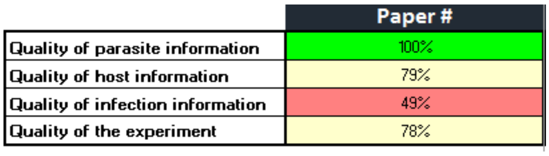


Microsoft Excel

Do you want to save the Paper Reviews and their Quality Scores?

Figure 6. Warning message: save the paper

8. Having scored the article, the quality of reported information in it will appear in qualitative and quantitative terms for each part of the experiment as shown in **Figure 7**.
    - Scores < 50% are represented with red cells.
    - Scores between 50% and 80% are represented with yellow cells.
    - Scores > 80% and < 90% are represented with light-green cells.
    - Scores > 90% are represented with bright-green.



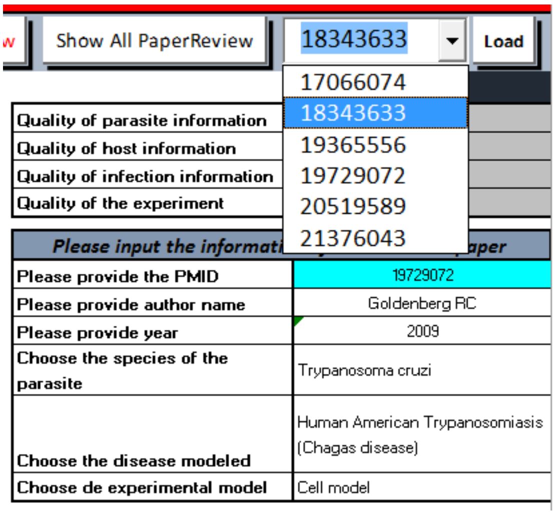
	Paper #
Quality of parasite information	100%
Quality of host information	79%
Quality of infection information	49%
Quality of the experiment	78%

Figure 7. Scoring results

9. Any time you can press "**Show All Paper Review**" button as shown in **Figure 2** and **5** to expand in the right side portion of the checklist all articles have been scored.

EDITING AN EXISTING ARTICLE

1. Select the article you want edit from the drop-down list as shown in **Figure 8**.



Quality of parasite information	
Quality of host information	
Quality of infection information	
Quality of the experiment	

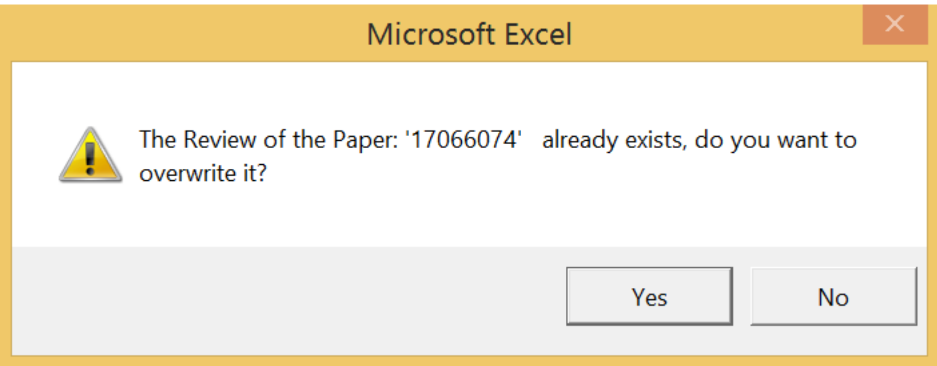
Please input the informati	*aper*
Please provide the PMID	19729072
Please provide author name	Goldenberg RC
Please provide year	2009
Choose the species of the parasite	Trypanosoma cruzi
Choose the disease modeled	Human American Trypanosomiasis (Chagas disease)
Choose de experimental model	Cell model

Figure 8. Drop-down list for paper reviewed

2. Load the filled checklist of the article by pressing the "**Load**" button next to this drop-down list as shown in **Figure 8**.

3. You can edit any field you want by using the drop-down lists.

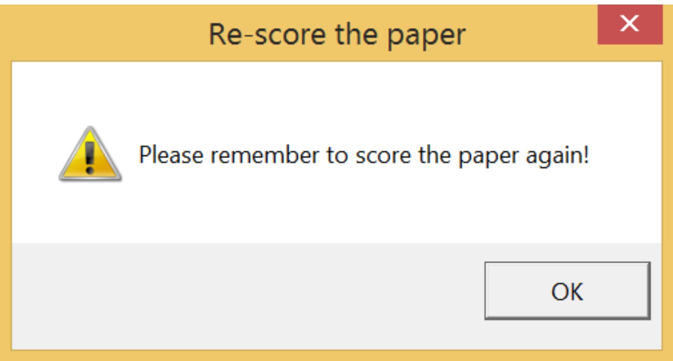
4. Press the "**Save Paper Review**" button to save the modifications to the checklist of the article. If the PMID number was not modified, then this selection will prompt you to overwrite the paper as shown in **Figure 9**. You should click 'Yes' if you want to do so. Otherwise click 'No'.


Microsoft Excel

The Review of the Paper: '17066074'   already exists, do you want to overwrite it?

Figure 9. Warning message: overwrite the paper

5. You will be advised to re-score the paper as shown in **Figure 10**. Press 'OK' and follow the steps mentioned in the "Scoring an article" section for this procedure.


Re-score the paper

Please remember to score the paper again!

Figure 10. Warning message: score the paper

The miniRECH repository

https://github.com/miniRECH

A collection of checklists for assessing the quality of methods reporting in selected biomedical fields.
This repository also includes the scripts necessaries to create a new miniRECH checklist in an easy way.

			Checklist for animal models of colitis	Please indicate at each item what you believe should be reported in an scientific article using animal models of colitis in order to reproduce and replicate its experiments. **Option A** *"This item must be provided"* or **Option B** *"This item is not demanded; however if it is presented, it would improve the reporting of the experiments"* . Please write **A** or **B** according to the option selected.	Please rank each item according to their contribution in the replicability and reproducibility of experimental work. Please write the value accordingly: where 1st = "highest importance"; and 3rd = "lowest importance".
		#	**Please answer the following questions according to the information reported in the experiment.**	**Category**	**Rank**
Information about the animal	**Animals**	1	Is the species of animal identified? (e.g. mouse)		
		2	Is the strain of animal identified? (e.g. C57BL/6)		
		3	Is the age of the animal described? (e.g. 12 weeks old)		
		4	Is the gender of the animal described? (e.g. Male)		
		5	Is the source of animals defined? (e.g. name of supplier or bred in facility)		
		6	Were animals acclimated to local microbiota? (e.g. housed in identical conditions at least 7 days prior to experiment start)		
	Animal housing conditions	7	Is the light/dark cycle described? (e.g. 12 hours light/dark)		
		8	Is the temperature described? (e.g. 25 °C)		
		9	Is the humidity described? (e.g. 40-45 %)		
		10	Is the food/water described? (e.g. regular chow)		
		11	Is the number of animals per cage described? (e.g. 3 mice per cage)		
Information about the colitis model	**Generically modified**	12	Is the genetic modification identified? (e.g. IL-10-/-)		
		13	Is the background strain of the animal described? (e.g. BALB/c)		
	Chemically induced colitis model (e.g. DSS)	14	Is the chemical used to induce colitis specified? (e.g. DSS)		
		15	Is the molecular weight of the chemical specified? (e.g. 36-50 kDa; DSS only)		
		16	Is the supplier of the chemical identified? (e.g. Sigma Aldrich)		
		17	Is the method of induction described? (e.g. dissolved in drinking water)		
		18	Is the dosage used described? (e.g. 2% w/v)		
		19	Is the medium of inoculation described? (e.g. TNBS in ethanol)		
	Biologically induced colitis model (e.g. bacterial infection)	20	Is the species of organism identified? (e.g. H. pylori)		
		21	Is the strain of organism identified? (e.g. PMSS1)		
		22	Are the culture conditions described? (e.g. animal passage/cell culture)		
		23	Is parasitaemia/colonisation adequately assessed? (e.g. colon homogenised and plated for colony counting)		
		24	Is the method of inoculation described? (e.g. oral gavage)		
		25	Is the dosage used described? (e.g. 10^8 cells)		
	Adoptive transfer colitis model (e.g. T cell transfer)	26	Is the cell type being transferred described? (e.g. CD4+CD45RB5high)		
		27	Is the species of the donor animal identified? (e.g. mouse)		
		28	Is the strain of the donor animal identified? (e.g. C57BL/6)		
		29	Is the gender of the donor animal described? (e.g. Male)		
		30	Is the number of cells transferred specified? (e.g. 4x10^5)		
		31	Is the purity of cells transferred specified? (e.g. >95%)		
		32	Is the viability of cells confirmed prior to transfer? (e.g. via 7-AAD staining during FACS)		
		33	Is the method of cell transfer described? (e.g. intraperitoneal injection)		
Information about the experimental design	**Experimental design**	34	Is the time course of the experiment described? (e.g. mice sacrificed after 7 days exposure to DSS)		
		35	Is the method of euthanasia described? (e.g. cervical dislocation)		
		36	Is animal weight loss reported? (e.g. as daily % of starting weight)		
		37	Is mortality reported? (e.g. survival curve)		
	Colitis monitoring and scoring	38	Is colitis monitored clinically? (e.g. disease activity index)		
		39	Is colitis scored histologically? (e.g. H&E stain)		
		40	Is microbiota diversity/population assessed? (e.g. 16S rRNA sequencing)		
		41	Is colon length or weight measured after sacrifice?		
		42	Is the section of gut for analysis identified? (e.g. proximal colon)		

A.3. Supplementary files Chapter 5

Table S1. Summary information about all included studies.

Author	Year	Journal	Colitis model(s)	Aim
Abad *et al.*	2005	Inflamm Bowel Dis	TNBS	Analyze the expression of several mediators related to the inflammatory cascade in colitic and vasoactive intestinal peptide-treated animals.
Barnett *et al.*	2010	BMC Immunol	IL-10$^{-/-}$	Characterize changes in colonic gene expression levels in Il10$^{-/-}$ and C57BL/6J mice resulting from oral bacterial inoculation with 12 *Enterococcus faecalis* and *faecium* (EF) strains, complex intestinal flora, or a mixture of the two.
Benight *et al.*	2012	Am J Physiol Gastrointest Liver Physiol	DSS	Investigate the anti-inflammatory properties of the anti-inflammatory, Methylthioadenosine in models of intestinal inflammation.
Billerey-Larmonier *et al.*	2008	Inflamm Bowel Dis	TNBS	Investigate the effect of dietary curcumin in colitis induced by TNBS in NKT-deficient SJL/J mice and BALB/c mice.
Brenna *et al.*	2013	PLoS One	TNBS	Study the correlation between endoscopic, histologic and gene expression alterations at different timepoints after colitis induction in a rat model of colitis and compare rat and human IBD mucosal transcriptomic data to evaluate whether TNBS colitis is an appropriate model of IBD.
Breynaert *et al.*	2013	PLoS One	DSS	Investigate changes in connective tissue in a chronic murine model resulting from repeated cycles of DSS ingestion, to mimic the relapsing nature of the human disease.
Brudzewsky *et al.*	2009	Scand J Immunol	T cell transfer	Employ a murine model of IBD to identify pathways and genes, which may play a key role in the pathogenesis of IBD and could be important for discovery of new disease markers inhuman disease.
Cho *et al.*	2011	Life Sci	DSS	Investigate the effects of the anti-inflammatory, xanthorrhizol in a mouse model of DSS-induced colitis.
Cho *et al.*	2011	Mol Nutr Food Res	DSS	Investigate the effects of oral administration of pure docosahexaenoic acid (DHA) and the therapeutic agent sulfasalazine on chemically induced colitis in mice, and analyzed the expression levels of DHA-responsive genes in colonic tissue using cDNA arrays.
Chung *et al.*	2014	Dig Dis Sci	DSS	Investigate the effects of sleep deprivation and melatonin on inflammation. Additionally investigate genes regulated by sleep deprivation and melatonin.
Coburn *et al.*	2012	PLoS One	DSS	Investigate the effect of L-Arganine on DSS induced colitis.
de Buhr *et al.*	2006	Physiol Genomics	IL-10$^{-/-}$	Identify candidate genes for colitis resistance/susceptibility in two strains of IL-10$^{-/-}$ mice by a combination of QTL mapping and microarray analyses.
Edmunds *et al.*	2012	Br J Nutr	IL-10$^{-/-}$	Examine whether kiwifruit extracts have immune-modulating effects *in vivo* against inflammatory processes in IL-10-gene deficient mice.
Fang *et al.*	2011	Physiol Genomics	DSS	Investigate temporal changes in genome expression profiles in the DSS colitis model, using whole genome expression profile analysis during the development of DSS colitis in comparison with ulcerative colitis (UC) specimens to identify novel and common responses during disease.
Fang *et al.*	2012	Inflamm Bowel Dis	T cell transfer	Identify changes in whole genome expression profiles using the CD4^{+} CD45RB^{high} T-cell transfer colitis model compared to genome expression differences from Crohn's disease (CD) tissue specimens.
Guzman *et al.*	2006	Inflamm Bowel Dis	TNBS	Assess the protective effect of the Adenosine A3 receptor agonist N(6)-(3-iodobenzyl)-adenosine-5-N-methyluronamide on gene dysregulation and injury in a rat chronic model of TNBS-induced colitis.
Hamilton *et al.*	2011	Proc Natl Acad Sci U S A	DSS and TNBS	Evaluate the differences between C57BL/6 mouse lines that differ in their expression of mast cell protease-6 and mast cell protease-7 in DSS and TNBS-induced colitis
Hansen *et al.*	2009	Inflamm Bowel Dis	DSS and IL-10$^{-/-}$	Compare gene expression profiles in cecal specimens from specific pathogen-free IL-10$^{-/-}$ mice with colitis and normal wild-type mice.
Hemmerling *et al.*	2014	PLoS One	IL-10$^{-/-}$	Investigate fetal exposure to maternal inflammation in genetically driven ileitis and colitis in response to maternal inflammation using susceptible and disease-free mice.
Hontecillas *et al.*	2011	Mucosal Immunol	DSS	Characterize the mechanisms underlying the beneficial effects of macrophage PPAR-c in DSS-induced colitis.
Huang *et al.*	2013	Gut	IL-10$^{-/-}$ and TNBS	Examine miRNA level in colon tissues and study the potential functions of miRNAs that regulate pathological genes during the inflammation process in TNBS-induced and IL-10$^{-/-}$ chronic colitis mice compared to CD patients.
Iizuka *et al.*	2010	FASEB J	DSS	Clarify the role of low-affinity leukotriene B4 (BLT) receptors in intestinal inflammation via DSS-induced colitis in mice lacking either BLT1 or BLT2.
Jia *et al.*	2011	Br J Nutr	DSS	Explore the combined effects of fish oil and curcumin on DSS-induced colitis in C57BL/6 mice.
Kabashima *et al.*	2002	J Clin Invest	DSS	Examine the roles of prostanoids in DSS-induced colitis in mice deficient in each of the eight types and subtypes of prostanoid receptors.
Kellermayer *et al.*	2010	Hum Mol Genet	DSS	Assess developmental changes in colitis susceptibility during the physiologically relevant period of childhood in mice, and concurrent changes in DNA methylation and gene expression in murine colonic mucosa.
Kiela *et al.*	2009	Gastroenterology	DSS	Investigate the role of NHE3 in maintaining mucosal integrity using DSS-induced colitis wild-type and NHE3$^{-/-}$ mice.
Knoch *et al.*	2010	Biotechnol J	IL-10$^{-/-}$	Examine colonic transcriptomic and proteomic profiles associated with colitis development in IL-10$^{-/-}$ and C57BL/6 mice fed either a linoleic acid-rich corn oil diet or an oleic acid-rich diet.
Knoch *et al.*	2010	Mol Nutr Food Res	IL-10$^{-/-}$	Investigate the effect of arachidonic acid on colonic inflammation in IL-10$^{-/-}$ mice.
Knoch *et al.*	2009	J Nutrigenet Nutrigenomics	IL-10$^{-/-}$	Test the effect of dietary eicosapentaenoic acid on intestinal inflammation using IL-10$^{-/-}$ mice.
Kremer *et al.*	2012	PLoS One	TNBS	Evaluate the development of pathology in conjunction with gene expression in the colon in response to chronic TNBS challenge.

Kristensen *et al.*	2008	Inflamm Bowel Dis	T cell transfer	Identify pathways of importance for immune regulation in the genome-wide expression profile in the inflamed rectum of SCID mice with CD4^{+} T cell transfer colitis and in the uninflamed rectum of mice protected from colitis by T reg cells.
Kuo *et al.*	2014	J Nutr	IL-10$^{-/-}$	Test the effect of symbiotic (prebiotic plus probiotic) supplements on colitis in wild-type and IL-10-deficient mice.
Lagishetty *et al.*	2010	Endocrinology	DSS	Test the hypothesis that impaired vitamin D status predisposes to IBD using vitamin D-deficient or vitamin D-sufficient diets followed by treatment with.
Lara-Villoslada *et al.*	2006	Clin Nutr	DSS	Evaluate the effect of oligosaccharides from goat milk in a rat model of DSS induced colitis.
Larrosa *et al.*	2009	J Agric Food Chem	DSS	Ascertain whether resveratrol can exert anti-inflammatory activity in a rat model of DSS-induced colitis.
Lee *et al.*	2009	Inflammation	DSS	Evaluate the anti-colitic effect of lactic acid in DSS colitis.
Lopez-Dee *et al.*	2012	PLoS One	DSS	Ascertain possible functions and evaluate potential therapeutic effects of Thrombospondin-1 type 1 repeats in inflammatory bowel disease.
Mannick *et al.*	2005	J Gastroenterol Hepatol	DSS	Examine the role of interferon regulatory factor-1 in DSS colitis to determine if absence of the gene would protect against colitis.
Mariman *et al.*	2012	Inflamm Bowel Dis	TNBS	Evaluate the efficacy of probiotics in the recurrent TNBS-induced colitis model and gain more insight into protective mechanisms.
Martínez-Augustin *et al.*	2008	BMC Genomics	TNBS	Characterize the TNBS-induced rat colitis model at the genomic level using a longitudinal approach.
Mizoguchi	2006	Gastroenterology	DSS and IL-10$^{-/-}$	Characterize the functional role of the Chitinase 3-like-1 molecule and its involvement in the dysregulation of host/microbial interaction in colitis.
Mizoguchi *et al.*	2003	Gastroenterology	DSS and IL-10$^{-/-}$	Investigate the impact of colonic crypt elongation during chronic and acute colitis.
Nakajima *et al.*	2002	J Gastroenterol	DSS	Perform a global analysis of differential gene expression during DSS colitis following administration of peroxisome proliferator activator receptor-gamma (PPARγ).
Nur *et al.*	2002	J Nutr	TNBS	Compare the expression profiles of rat models of vitamin A deficiency and induced colitis.
Reiff *et al.*	2009	Inflamm Bowel Dis	IL-10$^{-/-}$	Identify important signaling pathways and transcription factors relevant to gut inflammation and anti-inflammatory probiotics in the IL-10 knockout mouse model.
Reikvam *et al.*	2012	Eur J Immunol	DSS	Compare gene expression of wild-type and polymeric Ig receptor knockout mice in response to DSS-induced colitis.
Rivera *et al.*	2006	Inflamm Bowel Dis	TNBS	Identify differentially expressed genes in the TNBS-induced rat model of experimental colitis and compare gene expression profiles with that reported in patients.
Rivollier *et al.*	20 12	J Exp Med	T cell transfer	Investigate dendritic cell and macrophage populations involved in homeostasis in the colon during T cell transfer colitis.
Roy *et al.*	2007	Mutat Res	IL-10$^{-/-}$	Investigate transcriptomic changes in IL-10$^{-/-}$ and C57BL/6j mice inoculated with complex intestinal microflora and/or pure cultures of *Enterococcus faecalis* and *E. faecalis*.
Russ *et al.*	2013	PLoS One	IL-10$^{-/-}$	Investigate the molecular changes that occur in early and late inflammation stages in colonic epithelium in the IL-10 knockout mouse model of inflammatory bowel diseases.
Sainathan *et al.*	2012	Inflamm Bowel Dis	DSS	Test the hypothesis that Toll-like receptor-7 agonists have therapeutic efficacy in an acute DSS-induced colitis model.
Sainathan *et al.*	2008	Inflamm Bowel Dis	DSS	Study the effects of granulocyte macrophage colony stimulating factor (GM-CSF) in the DSS-induced acute murine colitis model to identify the possible mechanisms of how GM-CSF induces clinical response and remission in patients with active Crohn's disease.
Schaible *et al.*	2011	Hum Mol Genet	DSS	Study the effects of maternal methyl-donor diet supplementation on offspring colitis susceptibility and colonic mucosal DNA methylation and gene expression changes in DSS-induce murine colitis.
te Velde *et al.*	2007	Inflamm Bowel Dis	DSS, T cell transfer and TNBS	Compare the gene expression profiles of DSS-, TNBS- and T cell transfer-induced murine colitis models with the gene expression profiles of clinical IBD patients.
Wu & Chakravarti	2007	J Immunol	TNBS	Elucidate inflammatory signals that regulate fibrosis by investigating gene expression changes underlying chronic inflammation and fibrosis in TNBS-induced murine colitis.
Yamamoto *et al.*	2005	Biol Pharm Bull	TNBS	Identify gene transcripts associated with the onset of inflammation in the intestine of TNBS-treated mice.
Zhou *et al.*	2009	Gastroenterology	TNBS	Examine the role of forkhead box transcription factor O4 in intestinal mucosal immunity and inflammatory bowel disease using the TNBS-induced mouse model.
Zwiers *et al.*	2008	Inflamm Bowel Dis	TNBS	Identify candidate genes that confer resistance/susceptibility to TNBS-induced colitis in mice.

Table S2. Checklist scoring regarding animals and housing for all 29 DSS papers included in the systematic review.

	Information about the animal												
DSS model	**Animals**						**Animal housing**					**Section quality**	**Section score (%)**
Author/Date	1.1	1.2	1.3	1.4	2.1	2.2	3.1	3.2	3.3	3.4	3.5		
Benight NM (2012)	Y	Y	Y	Y	Y	Y	Y	N	N	Y	N	G	83.52
Breynaert C (2013)	Y	Y	Y	Y	Y	N	N	N	N	N	N	KCM	54.95
Cho JY (2010)	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	N	KCM	83.52
Cho JY (2011)	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	N	KCM	83.52
Chung SH (2014)	Y	Y	Y	N	Y	N	Y	Y	Y	N	N	KCM	62.64
Coburn LA (2012)	Y	Y	Y	Y	Y	N	Y	N	N	Y	N	KCM	72.53
Fang K (2010)	Y	Y	Y	N	N	N	N	N	N	N	N	KCM	32.97
Hamilton MJ (2010)	Y	Y	Y	Y	Y	Y	N	N	N	Y	N	G	75.82
Hansen JJ (2009)	Y	Y	N	N	Y	Y	N	N	N	N	N	KCM	43.96
Hontecillas R (2010)	Y	Y	Y	N	Y	Y	Y	Y	N	N	N	KCM	68.13
Iizuka Y (2010)	Y	Y	Y	Y	Y	Y	N	N	N	N	N	KCM	65.93
Jia Q (2011)	Y	Y	Y	Y	Y	Y	Y	N	N	Y	N	G	83.52
Kabashima K (2002)	Y	Y	Y	Y	Y	Y	Y	N	N	N	N	KCM	73.63
Kellermayer R (2010)	Y	Y	Y	Y	Y	Y	N	N	N	Y	N	G	75.82
Kiela PR (2009)	Y	Y	Y	N	Y	Y	N	N	N	Y	N	KCM	64.84
Lagishetty V (2010)	Y	Y	Y	N	Y	Y	N	N	N	Y	N	KCM	64.84
Lara-Villoslada F (2005)	Y	Y	N	Y	Y	Y	Y	Y	N	Y	Y	KCM	83.52
Larossa M (2009)	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	KCM	89.01
Lee H (2009)	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	N	KCM	83.52
Lopez-Dee ZP (2012)	Y	Y	Y	N	Y	N	N	N	N	N	N	KCM	43.96
Mannick EE (2005)	Y	Y	Y	N	Y	N	N	N	N	N	N	KCM	43.96
Mizoguchi E (2003)	Y	Y	N	N	Y	Y	N	N	N	N	N	KCM	43.96
Mizoguchi E (2006)	Y	Y	Y	Y	Y	N	N	N	N	N	N	KCM	54.95
Nakajima A (2003)	Y	Y	Y	Y	Y	N	N	N	N	Y	N	KCM	64.84
Reikvam DH (2012)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	G	100.00
Sainathan SK (2007)	Y	Y	Y	Y	Y	N	Y	N	N	Y	N	KCM	72.53
Sainathan SK (2011)	Y	Y	Y	Y	Y	N	Y	N	N	N	N	KCM	62.64
Schaible TD (2011)	Y	Y	Y	Y	Y	Y	N	N	N	Y	N	G	75.82
te Velde AA (2007)	Y	Y	Y	Y	Y	N	N	N	N	N	N	KCM	54.95

G: Good
KCM: KCM

Table S3. Checklist scoring regarding the colitis model for all 29 DSS papers included in the systematic review.

	Information about the colitis model																							
DSS model	**GM**		**Chemically induced**						**Biologically induced**						**Adoptive transfer**								**Section quality**	**Section score (%)**
Author/Date	4.1	4.2	5.1	5.2	5.3	5.4	5.5	5.6	6.1	6.2	6.3	6.4	6.5	6.6	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8		
Benight NM (2012)	*	*	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Breynaert C (2013)	*	*	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Cho JY (2010)	*	*	Y	N	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	KCM	95.0
Cho JY (2011)	*	*	Y	N	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	KCM	95.0
Chung SH (2014)	*	*	Y	N	N	N	Y	N	*	*	*	*	*	*	*	*	*	*	*	*	*	*	KCM	83.0
Coburn LA (2012)	Y	Y	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Fang K (2010)	*	*	Y	N	N	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	KCM	91.0
Hamilton MJ (2010)	Y	Y	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Hansen JJ (2009)	Y	Y	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Hontecillas R (2010)	Y	Y	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Iizuka Y (2010)	Y	Y	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Jia Q (2011)	*	*	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Kabashima K (2002)	Y	Y	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Kellermayer R (2010)	*	*	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Kiela PR (2009)	Y	Y	Y	N	N	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	KCM	91.0
Lagishetty V (2010)	*	*	Y	N	N	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	KCM	91.0
Lara-Villoslada F (2005)	*	*	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Larossa M (2009)	*	*	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Lee H (2009)	*	*	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Lopez-Dee ZP (2012)	*	*	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Mannick EE (2005)	Y	Y	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Mizoguchi E (2003)	Y	Y	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Mizoguchi E (2006)	Y	Y	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Nakajima A (2003)	*	*	Y	N	N	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	KCM	91.0
Reikvam DH (2012)	Y	Y	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	KCM	100
Sainathan SK (2007)	Y	Y	Y	N	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	KCM	95.0
Sainathan SK (2011)	*	*	Y	N	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	KCM	95.0
Schaible TD (2011)	*	*	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
te Velde AA (2007)	*	*	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	Y	Y	Y	Y	Y	Y	N	Y	KCM	95.0

G: Good
KCM: KCM

Table S4. Checklist scoring regarding the experimental design for all 29 DSS papers included in the systematic review.

	Information about the experimental design												
DSS models	**Experimental design**				**Colitis monitoring & scoring**					**Section quality**	**Section score (%)**	**Overall quality**	**Overall score (%)**
Author/Date	8.1	8.2	8.3	8.4	9.1	9.2	9.3	9.4	9.5				
Benight NM (2012)	Y	Y	Y	Y	Y	Y	N	N	Y	KCM	79.75	KCM	91.62
Breynaert C (2013)	Y	Y	Y	N	Y	Y	N	Y	Y	KCM	78.48	KCM	84.32
Cho JY (2010)	Y	Y	Y	N	Y	Y	N	Y	Y	KCM	78.48	KCM	88.65
Cho JY (2011)	Y	N	Y	N	Y	Y	N	Y	Y	KCM	70.89	KCM	87.03
Chung SH (2014)	Y	N	Y	Y	N	Y	N	N	Y	KCM	63.29	KCM	73.78
Coburn LA (2012)	Y	N	Y	Y	N	Y	N	Y	Y	G	74.68	KCM	87.84
Fang K (2010)	Y	N	N	N	Y	Y	N	N	Y	KCM	46.84	KCM	67.30
Hamilton MJ (2010)	Y	N	Y	N	Y	Y	N	N	Y	KCM	59.49	KCM	85.41
Hansen JJ (2009)	Y	N	Y	N	N	Y	N	N	Y	KCM	50.63	KCM	75.68
Hontecillas R (2010)	Y	Y	N	N	Y	Y	N	N	Y	KCM	54.43	KCM	82.43
Iizuka Y (2010)	Y	N	Y	N	N	Y	N	N	Y	KCM	50.63	KCM	81.08
Jia Q (2011)	Y	N	Y	Y	N	Y	N	Y	Y	G	74.68	G	90.54
Kabashima K (2002)	Y	N	Y	N	Y	Y	N	Y	Y	KCM	70.89	KCM	87.30
Kellermayer R (2010)	Y	N	Y	N	N	Y	N	Y	Y	KCM	62.03	KCM	85.95
Kiela PR (2009)	Y	Y	Y	Y	N	Y	N	N	Y	KCM	70.89	KCM	80.27
Lagishetty V (2010)	Y	N	Y	N	Y	Y	Y	N	Y	KCM	68.35	KCM	79.73
Lara-Villoslada F (2005)	Y	Y	Y	N	Y	Y	N	Y	Y	KCM	78.48	KCM	91.35
Larossa M (2009)	Y	Y	Y	N	N	Y	Y	Y	Y	KCM	78.48	KCM	92.70
Lee H (2009)	Y	Y	Y	N	Y	Y	N	Y	Y	KCM	78.48	KCM	91.35
Lopez-Dee ZP (2012)	Y	Y	N	N	Y	Y	N	N	Y	KCM	54.43	KCM	76.49
Mannick EE (2005)	Y	Y	Y	N	N	Y	N	N	Y	KCM	58.23	KCM	77.30
Mizoguchi E (2003)	Y	N	Y	N	Y	Y	N	Y	Y	KCM	70.89	KCM	80.00
Mizoguchi E (2006)	Y	N	Y	Y	Y	Y	N	N	Y	KCM	72.15	KCM	82.97
Nakajima A (2003)	Y	N	N	N	N	N	N	N	Y	KCM	25.32	KCM	70.54
Reikvam DH (2012)	Y	Y	Y	Y	Y	N	Y	N	Y	KCM	75.95	KCM	94.86
Sainathan SK (2007)	Y	Y	Y	N	Y	Y	N	Y	Y	KCM	78.48	KCM	85.95
Sainathan SK (2011)	Y	Y	Y	N	Y	Y	N	Y	Y	KCM	78.48	KCM	83.51
Schaible TD (2011)	Y	Y	N	Y	N	N	Y	Y	Y	KCM	65.82	KCM	86.76
te Velde AA (2007)	Y	N	N	N	N	Y	N	N	Y	KCM	37.97	KCM	72.97

G: Good
KCM: KCM

Table S5. Checklist scoring regarding animals and housing for all 15 IL-10-/- papers included in the systematic review.

	Information about the animal												
IL-10$^{-/-}$ model	**Animals**						**Animal housing**					**Section quality**	**Section score (%)**
Author/Date	1.1	1.2	1.3	1.4	2.1	2.2	3.1	3.2	3.3	3.4	3.5		
Barnett MP (2010)	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	KCM	89.01
de Buhr MF (2006)	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	KCM	89.01
Edmunds SJ (2011)	Y	Y	Y	Y	Y	Y	Y	N	N	Y	N	G	83.52
Hansen JJ (2009)	Y	Y	Y	N	Y	N	N	N	N	N	N	KCM	43.96
Hemmerling J (2014)	Y	Y	Y	N	Y	Y	Y	Y	N	Y	Y	KCM	83.52
Huang Z (2013)	Y	Y	Y	Y	Y	N	N	N	N	N	N	KCM	54.95
Knoch B (2009)	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	N	KCM	83.52
Knoch B (2010a)	Y	Y	Y	Y	Y	Y	N	N	N	Y	Y	G	81.32
Knoch B (2010b)	Y	Y	Y	Y	Y	N	N	N	N	Y	N	KCM	64.84
Kuo SM (2014)	Y	Y	Y	Y	Y	Y	N	N	N	Y	N	G	75.82
Mizoguchi E (2003)	Y	Y	N	N	Y	Y	N	N	N	N	N	KCM	43.96
Mizoguchi E (2006)	Y	Y	Y	Y	Y	N	N	N	N	N	N	KCM	54.95
Reiff C (2009)	Y	Y	Y	N	N	N	N	N	N	N	N	KCM	32.97
Roy N (2007)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	G	100
Russ AE (2013)	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	N	KCM	83.52

G: Good
KCM: KCM

Table S6. Checklist scoring regarding the colitis model for all 15 IL-10-/- papers included in the systematic review.

	Information about the colitis model																							
IL-10$^{-/-}$ model	GM		Chemically induced						Biologically induced						Adoptive transfer								Section quality	Section score (%)
Author/Date	4.1	4.2	5.1	5.2	5.3	5.4	5.5	5.6	6.1	6.2	6.3	6.4	6.5	6.6	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8		
Barnett MP (2010)	Y	Y	*	*	*	*	*	*	Y	Y	Y	N	Y	Y	*	*	*	*	*	*	*	*	KCM	95
de Buhr MF (2006)	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Edmunds SJ (2011)	Y	Y	*	*	*	*	*	*	Y	Y	Y	N	Y	Y	*	*	*	*	*	*	*	*	KCM	95
Hansen JJ (2009)	Y	Y	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Hemmerling J (2014)	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Huang Z (2013)	Y	Y	Y	*	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Knoch B (2009)	Y	Y	*	*	*	*	*	*	Y	N	Y	N	Y	Y	*	*	*	*	*	*	*	*	KCM	90
Knoch B (2010a)	Y	Y	*	*	*	*	*	*	Y	Y	Y	N	Y	Y	*	*	*	*	*	*	*	*	KCM	95
Knoch B (2010b)	Y	Y	*	*	*	*	*	*	N	N	Y	N	Y	Y	*	*	*	*	*	*	*	*	KCM	85
Kuo SM (2014)	Y	Y	*	*	*	*	*	*	Y	Y	N	N	Y	Y	*	*	*	*	*	*	*	*	KCM	92
Mizoguchi E (2003)	Y	Y	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Mizoguchi E (2006)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	G	100
Reiff C (2009)	Y	N	*	*	*	*	*	*	Y	N	N	Y	Y	Y	*	*	*	*	*	*	*	*	KCM	87
Roy N (2007)	Y	Y	*	*	*	*	*	*	Y	N	Y	N	Y	Y	*	*	*	*	*	*	*	*	KCM	90
Russ AE (2013)	Y	Y	*	*	*	*	*	*	Y	Y	Y	N	Y	Y	*	*	*	*	*	*	*	*	KCM	95

G: Good
KCM: KCM

Table S7. Checklist scoring regarding the experimental design for all 15 IL-10-/- papers included in the systematic review.

	Information about the experimental design												
IL-10$^{-/-}$ model	**Experimental design**				**Colitis monitoring & scoring**					**Section quality**	**Section score (%)**	**Overall quality**	**Overall score (%)**
Author/Date	8.1	8.2	8.3	8.4	9.1	9.2	9.3	9.4	9.5				
Barnett MP (2010)	Y	Y	Y	Y	N	Y	N	N	Y	KCM	70.89	KCM	88.38
de Buhr MF (2006)	Y	Y	N	N	N	N	N	N	Y	KCM	32.91	KCM	82.97
Edmunds SJ (2011)	Y	Y	Y	N	Y	Y	N	N	N	KCM	54.43	KCM	83.51
Hansen JJ (2009)	Y	N	Y	N	N	Y	N	N	Y	KCM	50.63	KCM	75.68
Hemmerling J (2014)	Y	Y	Y	N	N	Y	Y	Y	Y	KCM	78.48	KCM	91.35
Huang Z (2013)	Y	N	Y	N	Y	Y	N	Y	N	KCM	58.23	KCM	80.00
Knoch B (2009)	Y	Y	Y	Y	N	Y	N	N	Y	KCM	70.89	KCM	84.32
Knoch B (2010a)	Y	Y	Y	Y	N	Y	N	N	Y	KCM	70.89	KCM	86.49
Knoch B (2010b)	Y	Y	Y	Y	N	Y	N	N	Y	KCM	70.89	KCM	77.03
Kuo SM (2014)	Y	Y	Y	N	Y	Y	N	Y	Y	KCM	78.48	KCM	85.14
Mizoguchi E (2003)	Y	N	Y	N	Y	Y	N	Y	Y	KCM	70.89	KCM	80.0
Mizoguchi E (2006)	Y	N	Y	Y	Y	Y	N	N	Y	KCM	72.15	KCM	82.97
Reiff C (2009)	Y	Y	N	N	N	Y	Y	N	Y	KCM	54.43	KCM	66.76
Roy N (2007)	Y	Y	Y	N	N	Y	N	N	Y	KCM	58.23	KCM	85.68
Russ AE (2013)	Y	Y	Y	N	Y	Y	N	Y	Y	KCM	78.48	KCM	88.65

G: Good
KCM: KCM

Table S8. Checklist scoring regarding animals and housing for all 16 TNBS papers included in the systematic review.

	Information about the animal												
TNBS model	**Animals**						**Animal housing**					**Section quality**	**Section score (%)**
Author/Date	1.1	1.2	1.3	1.4	2.1	2.2	3.1	3.2	3.3	3.4	3.5		
Huang Z (2013)	Y	Y	Y	Y	Y	N	N	N	N	N	N	KCM	54.95
Brenna A (2013)	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	N	KCM	83.52
Kremer B (2012)	Y	Y	Y	Y	Y	N	Y	Y	N	Y	N	KCM	78.02
Mariman R (2011)	Y	Y	Y	Y	Y	N	Y	N	N	Y	N	KCM	72.53
Hamilton MJ (2010)	Y	Y	Y	Y	Y	Y	N	N	N	Y	N	G	75.82
Zhou W (2009)	Y	Y	Y	Y	Y	Y	N	N	N	N	N	KCM	65.93
Martinez-Augustin O (2008)	Y	Y	N	Y	N	N	Y	N	N	Y	N	KCM	50.55
Zwiers A (2008)	Y	Y	Y	Y	Y	N	N	N	N	N	N	KCM	54.95
Billerey-Larmonier C (2008)	Y	Y	Y	Y	Y	N	N	N	N	Y	N	KCM	64.84
Wu F (2007)	Y	Y	Y	Y	Y	N	N	N	N	N	N	KCM	54.95
te Velde AA (2007)	Y	Y	Y	Y	Y	N	N	N	N	N	N	KCM	54.95
Rivera E (2006)	Y	Y	N	Y	Y	N	Y	Y	N	Y	N	KCM	67.03
Guzman J (2006)	Y	Y	N	N	Y	N	N	N	N	N	N	KCM	32.97
Yamamoto S (2005)	Y	Y	Y	Y	Y	N	Y	N	N	N	N	KCM	62.64
Abad C (2005)	Y	Y	Y	Y	Y	N	N	N	N	N	N	KCM	54.95
Nur T (2002)	Y	Y	N	Y	Y	N	Y	Y	Y	Y	N	KCM	72.53

G: Good
KCM: KCM

Table S9. Checklist scoring regarding the colitis model for all 16 TNBS papers included in the systematic review.

	Information about the colitis model																							
TNBS model	**GM**		**Chemically induced**						**Biologically induced**						**Adoptive transfer**								**Section quality**	**Section score (%)**
Author/Date	4.1	4.2	5.1	5.2	5.3	5.4	5.5	5.6	6.1	6.2	6.3	6.4	6.5	6.6	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8		
Huang Z (2013)	Y	Y	Y	*	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Brenna A (2013)	*	*	Y	*	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Kremer B (2012)	*	*	Y	*	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Mariman R (2011)	*	*	Y	*	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Hamilton MJ (2010)	Y	Y	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Zhou W (2009)	Y	Y	Y	*	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Martinez-Augustin O (2008)	*	*	Y	*	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Zwiers A (2008)	*	*	Y	*	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Billerey-Larmonier C (2008)	*	*	Y	*	N	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	KCM	96
Wu F (2007)	*	*	Y	*	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
te Velde AA (2007)	*	*	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	Y	Y	Y	Y	Y	Y	N	Y	KCM	95
Rivera E (2006)	*	*	Y	*	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Guzman J (2006)	*	*	Y	*	N	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	KCM	96
Yamamoto S (2005)	*	*	Y	*	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Abad C (2005)	*	*	Y	*	Y	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	100
Nur T (2002)	*	*	Y	*	N	Y	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	KCM	96

G: Good
KCM: KCM

Table S10. Checklist scoring regarding the experimental design for all 16 TNBS papers included in the systematic review.

	Information about the experimental design												
TNBS model	**Experimental design**				**Colitis monitoring & scoring**					**Section quality**	**Section score (%)**	**Overall quality**	**Overall score (%)**
Author/Date	8.1	8.2	8.3	8.4	9.1	9.2	9.3	9.4	9.5				
Huang Z (2013)	Y	N	Y	N	Y	Y	N	Y	N	KCM	58.22	KCM	80
Brenna A (2013)	Y	Y	Y	N	Y	Y	N	N	N	KCM	54.43	KCM	86.21
Kremer B (2012)	Y	N	Y	Y	N	Y	N	Y	Y	G	74.68	KCM	89.18
Mariman R (2011)	Y	N	Y	N	N	Y	N	Y	Y	KCM	62.02	KCM	85.13
Hamilton MJ (2010)	Y	N	Y	N	Y	Y	N	N	Y	KCM	59.49	KCM	85.40
Zhou W (2009)	Y	N	N	Y	Y	Y	N	N	Y	KCM	59.49	KCM	82.97
Martinez-Augustin O (2008)	Y	Y	Y	N	N	N	N	Y	Y	KCM	56.96	KCM	78.64
Zwiers A (2008)	Y	N	N	N	N	N	N	N	Y	KCM	25.31	KCM	72.97
Billerey-Larmonier C (2008)	Y	Y	Y	Y	N	Y	N	N	Y	KCM	70.88	KCM	82.97
Wu F (2007)	Y	N	N	Y	N	Y	N	N	Y	KCM	50.63	KCM	78.37
te Velde AA (2007)	Y	N	N	N	N	Y	N	N	Y	KCM	37.97	KCM	72.97
Rivera E (2006)	Y	N	N	N	N	N	N	N	Y	KCM	25.31	KCM	75.94
Guzman J (2006)	Y	N	Y	N	Y	Y	N	N	Y	KCM	59.49	KCM	72.70
Yamamoto S (2005)	Y	N	N	N	N	N	N	N	N	KCM	12.65	KCM	72.16
Abad C (2005)	Y	N	N	N	N	N	N	N	Y	KCM	25.31	KCM	72.97
Nur T (2002)	Y	Y	Y	N	N	N	N	N	Y	KCM	45.56	KCM	79.45

G: Good
KCM: KCM

Table S11. Checklist scoring regarding animals and housing for all 5 T cell transfer papers included in the systematic review.

	Information about the animal												
T cell model	**Animals**						**Animal housing**					**Section quality**	**Section score (%)**
Author/Date	1.1	1.2	1.3	1.4	2.1	2.2	3.1	3.2	3.3	3.4	3.5		
Brudzewsky D (2009)	Y	Y	Y	Y	Y	N	N	N	N	N	N	KCM	54.95
Fang K (2011)	Y	Y	Y	Y	Y	N	N	N	N	N	N	KCM	54.95
Kristensen NN (2007)	Y	Y	Y	Y	Y	Y	N	N	N	N	N	KCM	65.93
Rivollier A (2012)	Y	Y	Y	N	Y	N	N	N	N	N	N	KCM	43.96
te Velde AA (2007)	Y	Y	Y	Y	Y	N	N	N	N	N	N	KCM	54.95

G: Good
KCM: KCM

Table S12. Checklist scoring regarding the colitis model for all 5 T cell transfer papers included in the systematic review.

	Information about the colitis model																						
T cell model	**GM**		**Chemically induced**						**Biologically induced**						**Adoptive transfer**								**Section quality**	**Section score (%)**
Author/Date	4.1	4.2	5.1	5.2	5.3	5.4	5.5	5.6	6.1	6.2	6.3	6.4	6.5	6.6	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8		
Brudzewsky D (2009)	*	*	*	*	*	*	*	*	*	*	*	*	*	*	Y	Y	Y	Y	Y	Y	N	Y	KCM	95
Fang K (2011)	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	Y	Y	Y	Y	Y	N	N	Y	KCM	91.5
Kristensen NN (2007)	*	*	*	*	*	*	*	*	*	*	*	*	*	*	Y	Y	Y	Y	Y	Y	N	N	KCM	90
Rivollier A (2012)	Y	Y	*	*	*	*	*	*	*	*	*	*	*	*	Y	Y	Y	N	Y	N	N	Y	KCM	88.5
te Velde AA (2007)	*	*	Y	Y	Y	Y	Y	Y	*	*	*	*	*	*	Y	Y	Y	Y	Y	Y	N	Y	KCM	95

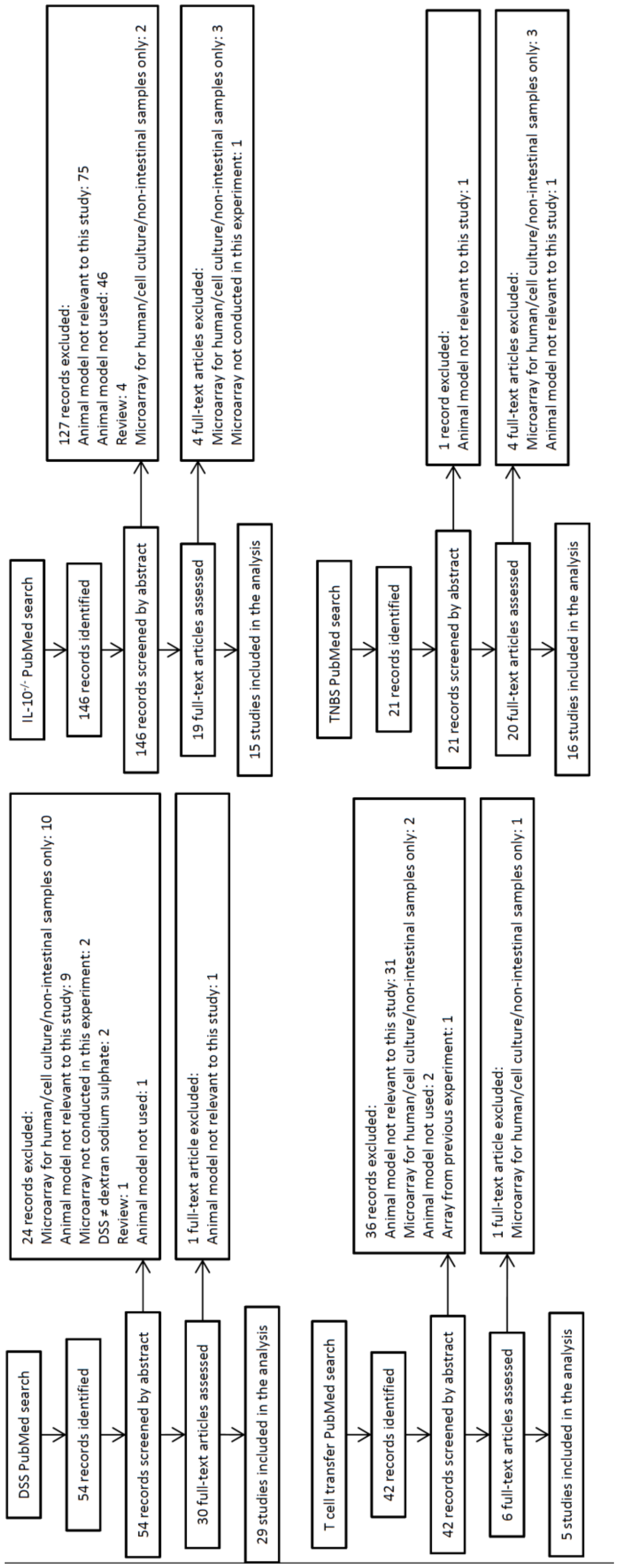
G: Good
KCM: KCM

Table S13. Checklist scoring regarding the experimental design for all 5 T cell transfer papers included in the systematic review.

	Information about the experimental design											
T cell model	**Experimental design**				**Colitis monitoring & scoring**					**Section qualit**	**Section score (%)**	**Overall quality**	**Overall score (%)**
Author/Date	8.1	8.2	8.3	8.4	9.1	9.2	9.3	9.4	9.5				
Brudzewsky D (2009)	Y	N	Y	N	N	Y	N	N	Y	KCM	50.63	KCM	75.68
Fang K (2011)	Y	N	Y	N	N	Y	N	Y	Y	KCM	62.02	KCM	76.22
Kristensen NN (2007)	N	Y	Y	N	N	Y	N	Y	Y	KCM	56.96	KCM	77.03
Rivollier A (2012)	Y	N	N	N	N	N	N	N	Y	KCM	25.31	KCM	64.05
te Velde AA (2007)	Y	N	N	N	N	Y	N	N	Y	KCM	37.97	KCM	72.97

G: Good
KCM: KCM

Figure S1. Study selection process for DSS, IL-10-/-, T cell transfer and TNBS colitis models.


DSS PubMed search
54 records identified
54 records screened by abstract
30 full-text articles assessed
29 studies included in the analysis
24 records excluded:
Microarray for human/cell culture/non-intestinal samples only: 10
Animal model not relevant to this study: 9
Microarray not conducted in this experiment: 2
DSS ≠ dextran sodium sulphate: 2
Review: 1
Animal model not used: 1
1 full-text article excluded:
Animal model not relevant to this study: 1
IL-10-/- PubMed search
146 records identified
146 records screened by abstract
19 full-text articles assessed
15 studies included in the analysis
127 records excluded:
Animal model not relevant to this study: 75
Animal model not used: 46
Review: 4
Microarray for human/cell culture/non-intestinal samples only: 2
4 full-text articles excluded:
Microarray for human/cell culture/non-intestinal samples only: 3
Microarray not conducted in this experiment: 1
T cell transfer PubMed search
42 records identified
42 records screened by abstract
6 full-text articles assessed
5 studies included in the analysis
36 records excluded:
Animal model not relevant to this study: 31
Microarray for human/cell culture/non-intestinal samples only: 2
Animal model not used: 2
Array from previous experiment: 1
1 full-text article excluded:
Microarray for human/cell culture/non-intestinal samples only: 1
TNBS PubMed search
21 records identified
21 records screened by abstract
20 full-text articles assessed
16 studies included in the analysis
1 record excluded:
Animal model not relevant to this study: 1
4 full-text articles excluded:
Microarray for human/cell culture/non-intestinal samples only: 3
Animal model not relevant to this study: 1

A.4. Supplementary files Chapter 6

Figure 1 – source data 1. PubMed search terms used for each disease group and their approaches.

Groups	Search terms	Number of articles
Main corpus	("Mice"[Mesh] AND (mouse[ti] OR mice[ti])) AND (Journal Article[ptyp] NOT Review[ptyp] AND ("1994/01/01"[PDAT] : "2014/12/31"[PDAT]) AND "animals"[MeSH Terms:noexp] AND English[lang]) AND pubmed pmc[sb]	15,311
Subsets	("Mice"[Mesh] AND (mouse[ti] OR mice[ti])) AND (Journal Article[ptyp] NOT Review[ptyp] AND ("2001/01/01"[PDAT] : "2014/12/31"[PDAT]) AND "animals"[MeSH Terms:noexp] AND English[lang]) AND pubmed pmc[sb]	14,225
Cardiovascular diseases	AND "Cardiovascular Diseases"[Mesh]	873
Genetics	AND "Cardiovascular Diseases/genetics"[Mesh]	293
Immunology	AND "Cardiovascular Diseases/immunology"[Mesh]	58
Physiopathology	AND "Cardiovascular Diseases/physiopathology"[Mesh]	231
Therapy	AND "Cardiovascular Diseases/therapy"[Mesh]	320
Myocardial Ischemia	AND "Myocardial Ischemia"[Mesh]	94
Cancer	AND "Neoplasms"[Mesh]	1,523
Genetics	AND "Neoplasms/genetics"[Mesh]	604
Immunology	AND "Neoplasms/immunology"[Mesh]	179
Physiopathology	AND "Neoplasms/physiopathology"[Mesh]	49
Therapy	AND "Neoplasms/therapy"[Mesh]	540
Melanoma	AND "Melanoma"[Mesh]	98
Diabetes Mellitus	AND "Diabetes Mellitus"[Mesh]	611
Genetics	AND "Diabetes Mellitus/genetics"[Mesh]	183
Immunology	AND "Diabetes Mellitus/immunology"[Mesh]	108
Physiopathology	AND "Diabetes Mellitus/physiopathology"[Mesh]	112
Therapy	AND "Diabetes Mellitus/therapy"[Mesh]	243
Diabetes type 2	AND "Diabetes Mellitus, Type 2"[Mesh]	149
Lung diseases	AND "Lung Diseases"[Mesh]	602
Genetics	AND "Lung Diseases/genetics"[Mesh]	153
Immunology	AND "Lung Diseases/immunology"[Mesh]	162
Physiopathology	AND "Lung Diseases/physiopathology"[Mesh]	76
Therapy	AND "Lung Diseases/therapy”[Mesh]	191
Diabetes type 2	AND "Pulmonary Disease, Chronic Obstructive”[Mesh]	37
Neurological disorders	AND "Nervous System Diseases"[Mesh]	2,137
Genetics	AND "Nervous System Diseases/genetics"[Mesh]	899
Immunology	AND "Nervous System Diseases/immunology"[Mesh]	182
Physiopathology	AND "Nervous System Diseases/physiopathology"[Mesh]	556
Therapy	AND "Nervous System Diseases/therapy"[Mesh]	609
Alzheimer	AND "Alzheimer Disease”[Mesh]	273
Infectious diseases	AND (“microbiology” [Subheading] OR “virology” [Subheading] OR “Parasitology” [Subheading])	1,269
Physiopathology	AND ("microbiology" [Subheading] OR "virology" [Subheading] OR "Parasitology" [Subheading]) AND "physiopathology" [Subheading]	67
Genetics	AND ("microbiology" [Subheading] OR "virology" [Subheading] OR "Parasitology" [Subheading]) AND "genetics" [Subheading]	640
Immunology	AND ("microbiology" [Subheading] OR "virology" [Subheading] OR "Parasitology" [Subheading]) AND "immunology" [Subheading]	662
Therapy	AND ("microbiology" [Subheading] OR "virology" [Subheading] OR "Parasitology" [Subheading]) AND "therapy" [Subheading]	370
Tuberculosis	AND ("microbiology" [Subheading] NOT "virology" [Subheading] NOT "Parasitology" [Subheading]) AND "Tuberculosis"[Mesh]	39
HIV	AND ("virology" [Subheading] NOT "Parasitology" [Subheading]) AND "HIV"[Mesh]	62
Malaria	AND ("Parasitology" [Subheading]) AND "Malaria"[Mesh]	39

Terms were chosen to cover both PubMed MeSH (Medical Subject Headings) and related strings to ensure that articles would still be captured even if they lacked correct subject heading annotations.

Figure 1 – source data 2. Example rules for identification of sex and age.

<table>
<tr><th>Characteristics</th><th>Rules</th><th colspan="24">Phrase</th></tr>
<tr><td rowspan="7">Sex</td><td>Abstract rule</td><td colspan="12">Gender (adjective)</td><td colspan="12">Mice (noun phrase)</td></tr>
<tr><td>Rule example</td><td colspan="12">({Token.string==~"(?i)male"})</td><td colspan="12">{Token.string==~"(?i)mice"}</td></tr>
<tr><td>Male mice 6-8-wk-old</td><td colspan="12">Male</td><td colspan="12">mice</td></tr>
<tr><td colspan="25"></td></tr>
<tr><td>Abstract rule</td><td colspan="6">Mice (noun phrase)</td><td colspan="6">Preposition</td><td colspan="6">Conjunction</td><td colspan="6">Gender</td></tr>
<tr><td>Rule example</td><td colspan="6">{Token.string==~"(?i)mice"}</td><td colspan="6">{Token.string ==~"(?i)of"}</td><td colspan="6">({Token.string ==~"(?i)either"}</td><td colspan="6">{Token.string ==~"(?i)sex"})</td></tr>
<tr><td>Mice of either sex were used</td><td colspan="6">Mice</td><td colspan="6">of</td><td colspan="6">either</td><td colspan="6">Sex</td></tr>
<tr><td colspan="26"></td></tr>
<tr><td rowspan="7">Age</td><td>Abstract rule</td><td colspan="6">Mice (noun phrase)</td><td colspan="6">Verb</td><td colspan="3">Numeric dictionary</td><td colspan="3">Any token</td><td colspan="3">Numeric dictionary</td><td colspan="3">Age (noun phrase)</td></tr>
<tr><td>Rule example</td><td colspan="6">{Token.string==~"(?i)mice"}</td><td colspan="6">{Token.string=="were"}</td><td colspan="3">((numbers)</td><td colspan="3">{Token}[0,1]</td><td colspan="3">(numbers)?</td><td colspan="3">(age))</td></tr>
<tr><td>Mice were 4 wk old</td><td colspan="6">mice</td><td colspan="6">were</td><td colspan="3">4</td><td colspan="3"></td><td colspan="3"></td><td colspan="3">wk old</td></tr>
<tr><td colspan="25"></td></tr>
<tr><td>Abstract rule</td><td colspan="4">Numeric dictionary</td><td colspan="4">Any token</td><td colspan="4">Numeric dictionary</td><td colspan="4">Age (noun phrase)</td><td colspan="4">Any token</td><td colspan="4">Mice (noun phrase)</td></tr>
<tr><td>Rule example</td><td colspan="4">((numbers)</td><td colspan="4">{Token}[0,1]</td><td colspan="4">(numbers)?</td><td colspan="4">(age))</td><td colspan="4">{Token}[0,2]</td><td colspan="4">{Token.string==~"(?i) mice"}</td></tr>
<tr><td>Generally 3-4 months old healthy adult mice</td><td colspan="4">3</td><td colspan="4">-</td><td colspan="4">4</td><td colspan="4">months old</td><td colspan="4">healthy adult</td><td colspan="4">mice</td></tr>
</table>

Examples show both an “abstract” description of the rule and the applied GATE notation. Rule components in highlighted text are the extracted (target) text that denote the mention of interest; the rest of the rule (if any) specifies the context. The rules use explicit matching of tokens (e.g., {Token.string ==~ "(?i)male"} matches the string ‘male’) and vocabularies that contain mentions of specific dictionaries. For example, (age) matches variations of the expressed age (e.g., ‘wk old’, ‘months old’) and (numbers) contains multiple numbers in both arithmetic and lexical forms. The ‘?’ at the end of certain rule components suggests ‘if any’, whereas {Token}[0,1] matches up to the given number of tokens, if any.

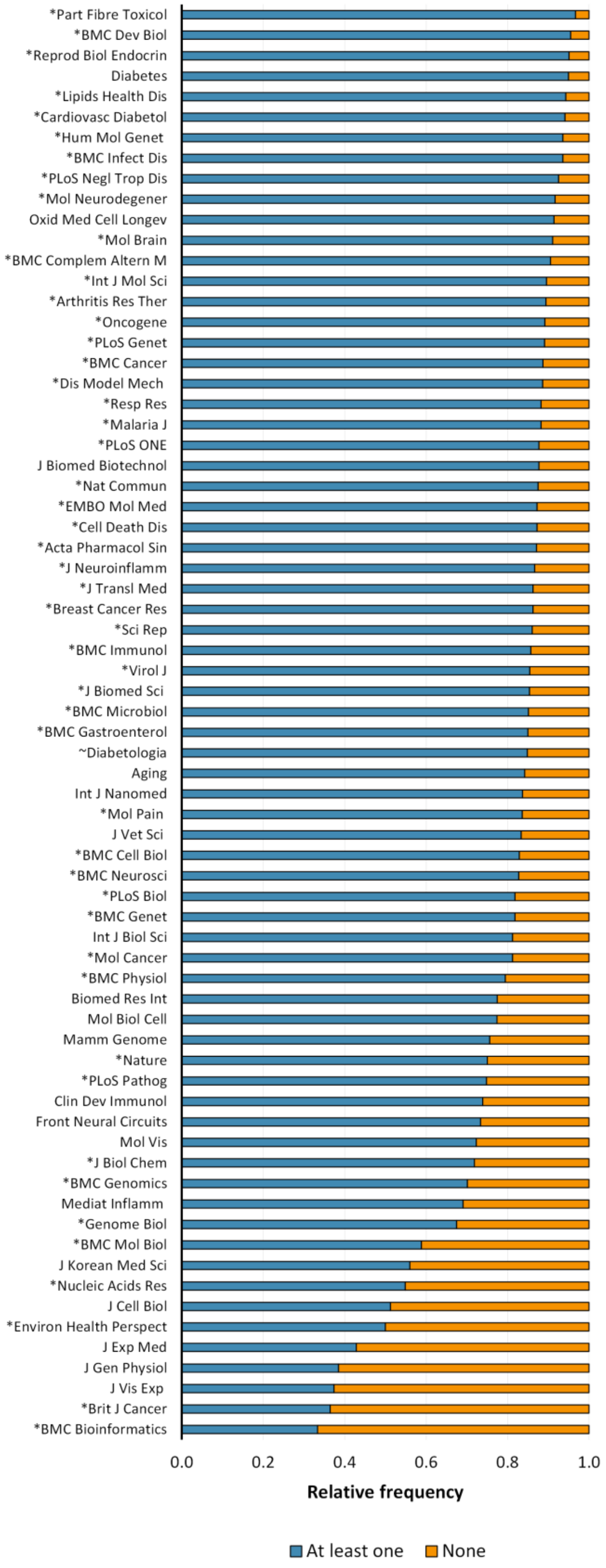


Figure 1 – figure supplement 1. Reporting of sex or age in mouse-model experiments by journal. The figure shows the top 70 journals from a total of 628 journals in which were published 30 or more articles of the corpus; corresponding to 81.05% of papers assessed. The journals are organised in descending order of the reporting of sex or age (i.e. at least one) as experimental variables. *Journals that endorsed ARRIVE and ~Journals that stated the reporting of sex and age in the author guidelines.

Supplementary File 2A

SET OF ARTICLES FOR CREATING THE TEXT-MINING RULES

The documents were manually extracted from PubMed using the query "Mice"[Mesh] AND (mouse[ti] OR mice[ti]) AND Journal Article[ptyp] AND English[lang].

PMID	Journal	Year
15642986	Clin Diagn Lab Immunol	2005
19254734	Physiol Behav	2009
21185930	J Ethnopharmacol	2011
21190827	J Nutr Biochem	2011
21193983	Psychopharmacology (Berl)	2011
21199659	J Immunol Methods	2011
21218482	J Sci Food Agric	2011
24015257	PLoS One	2013
24534203	Cancer Lett	2014
24646876	Immunobiology	2014
24736856	J Antibiot (Tokyo)	2014
24776490	Behav Pharmacol	2014
24871354	J Nat Prod	2014
24887420	PLoS One	2014
25069986	Infect Immun	2014
25201301	Br J Nutr	2014
25217696	Blood	2014
25218594	Cancer Lett	2014
25224570	Cancer Lett	2014
25231351	Am J Physiol Regul Integr Comp Physiol	2014
25234596	Biochem Biophys Res Commun	2014
25245810	Infect Immun	2014
25246326	Exp Parasitol	2014
25261995	Nat Med	2014
25267834	Infect Immun	2014
25268558	J Toxicol Environ Health A	2014
25273880	Am J Physiol Cell Physiol	2014
25280587	BMC Complement Altern Med	2014
25282357	Nat Med	2014
25287930	Infect Immun	2014
25288643	J Med Microbiol	2014
25288806	J Biol Chem	2014
25303897	Exp Mol Pathol	2014
25308446	Metabolism	2014
25318387	BMC Complement Altern Med	2014
25320354	Am J Physiol Renal Physiol	2014
25355549	BMC Complement Altern Med	2014
25367573	Immunity	2014
25283970	Prostate	2015
25347995	J Pharmacol Exp Ther	2015

Supplementary File 2B

SET OF ARTICLES FOR FINDING THE LOCATION OF THE MENTION OF THE SEX AND AGE OF THE MICE

The documents were randomly extracted from our corpus by using the "=RANDBETWEEN()" function in Microsoft Office Excel for Windows version 2013, and manually inspected in order to determine in which part of the article the sex and age of the mice were mentioned.

PMID	Journal	Year
8976197	J Exp Med	1996
8976195	J Exp Med	1996
8976192	J Exp Med	1996
9049243	J Cell Biol	1997
9034144	J Exp Med	1997
9008713	J Cell Biol	1997
9808781	Dev Biol	1998
16172261	J Exp Med	2005
17900358	BMC Neurosci	2007
17683525	BMC Neurosci	2007
17592641	J Transl Med	2007
19020657	PLoS One	2008
18688274	PLoS Pathog	2008
18568131	Mol Vis	2008
19765281	Arthritis Res Ther	2009
20405019	PLoS One	2010
20368974	PLoS One	2010
20098691	PLoS One	2010
23272179	PLoS One	2012
23237483	BMC Immunol	2012
22802958	PLoS One	2012
21765465	Oncogene	2012
24278473	PLoS One	2013
24212843	Clinics (Sao Paulo)	2013
24194903	PLoS One	2013
24147098	PLoS One	2013
24098534	PLoS One	2013
23966857	PLoS Pathog	2013
23903059	Exp Anim	2013
23762356	PLoS One	2013
23667681	PLoS One	2013
23613811	PLoS One	2013
23516562	PLoS One	2013
23451234	PLoS One	2013
23326190	Int J Nanomedicine	2013
23321513	Br J Cancer	2013
23302418	BMC Neurosci	2013
23286586	J Biomed Sci	2013
24995344	J Immunol Res	2014
24455991	J Cell Mol Med	2014

Supplementary File 2C

SET OF ARTICLES FOR ENHANCING THE PERFORMANCE OF THE TEXT-MINING RULES

The documents were randomly extracted from our corpora by using the "=RANDBETWEEN()" function in Microsoft Office Excel for Windows version 2013. Five documents were used for each year from 2001 to 2014.

PMID	Journal	Year
11305942	Genome Biol	2001
11304550	J Exp Med	2001
11748281	J Exp Med	2001
11785668	Dev Immunol	2001
11737881	BMC Complement Altern Med	2001
12021255	J Cell Biol	2002
12401133	BMC Cell Biol	2002
12163565	J Exp Med	2002
12198088	J Gen Physiol	2002
11956298	J Exp Med	2002
14623911	J Exp Med	2003
12925704	J Cell Biol	2003
12860970	J Cell Biol	2003
12932298	Reprod Biol Endocrinol	2003
12771178	J Exp Med	2003
15483348	J Korean Med Sci	2004
14728723	BMC Neurosci	2004
15302899	J Exp Med	2004
15534693	PLoS Biol	2004
15154615	Clin Dev Immunol	2004
16250671	PLoS Med	2005
15998448	Genome Biol	2005
16293190	BMC Infect Dis	2005
16033648	BMC Dev Biol	2005
16079067	Environ Health Perspect	2005
16571105	BMC Genet	2006
16563162	Mol Cancer	2006
16502487	Yonsei Med J	2006
17069643	BMC Gastroenterol	2006
17069661	Virol J	2006
17406675	PLoS ONE	2007
17683579	BMC Cancer	2007
17266762	Genome Biol	2007
17220887	Nat Neurosci	2007
17605779	BMC Dev Biol	2007
18371231	BMC Genomics	2008
18716442	J Vet Sci	2008
18547429	Lipids Health Dis	2008
18307760	Respir Res	2008
18789160	BMC Cell Biol	2008
19750022	Toxicol Mech Methods	2009
19557135	PLoS One	2009
19129917	PLoS One	2009
19296832	BMC Microbiol	2009
19221395	J Exp Med	2009
20525357	BMC Biol	2010
20041326	Cancer Chemother Pharmacol	2010
20689830	PLoS One	2010
20796285	BMC Neurosci	2010
21171988	BMC Genomics	2010
21818344	PLoS One	2011
21799730	PLoS One	2011

21412423	PLoS One	2011
22163031	PLoS One	2011
21439091	Malar J	2011
21954065	Dis Model Mech	2012
22235288	PLoS One	2012
22275470	Dis Model Mech	2012
23087911	Front Cell Infect Microbiol	2012
22859963	PLoS One	2012
23967191	PLoS One	2013
23536174	Sci Rep	2013
23451234	PLoS One	2013
24317954	Oncotarget	2013
23519026	Dis Model Mech	2013
24466007	PLoS One	2014
25077564	BMC Genomics	2014
24924430	Dis Model Mech	2014
24638941	Int J Mol Med	2014
24559113	BMC Complement Altern Med	2014

Supplementary File 2D

SET OF ARTICLES FOR EVALUATING THE TEXT-MINING SYSTEM

The documents were randomly extracted from our corpus by using the "=RANDBETWEEN()" function in Microsoft Office Excel for Windows version 2013, and manually double-annotated for both the age and the sex by two biomedical experts.

PMID	Journal	Year
8145050	J Exp Med	1994
7744960	J Cell Biol	1995
8924761	Dev Immunol	1995
8879219	J Exp Med	1996
9064345	J Exp Med	1996
10880524	J Exp Med	2000
11532190	BMC Cell Biol	2001
16800892	BMC Biotechnol	2006
18280460	Biochem Pharmacol	2008
19127268	Br J Cancer	2009
19255868	Biogerontology	2009
20041218	PLoS Genet	2009
19850720	Nucleic Acids Res	2010
19920212	Physiol Genomics	2010
20084100	PLoS Genet	2010
20107508	PLoS ONE	2010
20167811	Neuro Oncol	2010
20169060	PLoS ONE	2010
20405007	PLoS ONE	2010
20686609	PLoS ONE	2010
20532624	Transgenic Res	2011
21464968	PLoS ONE	2011
21492450	BMC Neurosci	2011
22428884	J Environ Sci Health B	2012
22520439	J Neuroinflammation	2012
22532835	PLoS ONE	2012
22547652	J Exp Med	2012
22675511	PLoS ONE	2012
22892315	Mol Brain	2012
22906987	Lab Invest	2012
22952733	PLoS ONE	2012
23049968	PLoS ONE	2012
23194061	Reprod Biol Endocrinol	2012
23233794	Mol Vis	2012
23316291	J Am Heart Assoc	2012
23341968	PLoS ONE	2013
23935987	PLoS ONE	2013
23936125	PLoS ONE	2013
23942071	Br J Cancer	2013
23991183	PLoS ONE	2013
24386094	PLoS ONE	2013
24459328	Mediators Inflamm	2013
24273196	J Lipid Res	2014
24361736	Neuroscience	2014
24493738	Nucleic Acids Res	2014
24500039	Med Sci Monit Basic Res	2014
24621297	Aging Cell	2014
24833816	Mediators Inflamm	2014
24877142	Biomed Res Int	2014
25092975	Int J Nanomedicine	2014

Supplementary File 3

RULES USED TO IDENTIFY THE SEX AND AGE OF EXPERIMENTAL MOUSE MODELS

Our rules were created and applied via GATE –General Architecture for Text Engineering; an open source free software enabling the design and implementation of information extraction systems in unstructured text with the crafted rules following its notation.

Rules for the identification of sex

```
(
{Token.string ==~ "(?i)male"}
|
{Token.string ==~ "(?i)female"}
|
{Token.string ==~ "(?i)males"}
|
{Token.string ==~ "(?i)females"}
|
({Token.string ==~ "(?i)female"}{Token.string == "/"}{Token.string ==~
"(?i)male"})
|
({Token.string ==~ "(?i)male"}{Token.string == "/"}{Token.string ==~
"(?i)female"})
|
({Token.string ==~ "(?i)male"}{Token.string ==~
"(?i)and"}{Token.string ==~ "(?i)female"})
|
({Token.string ==~ "(?i)female"}{Token.string ==~
"(?i)and"}{Token.string ==~ "(?i)male"})
|
(
({Token.string ==~"(?i)age"}|{Token.string ==~"(?i)sex"})
({Token.string == "-"})?
{Token.string ==~"(?i)and"}
({Token.string ==~"(?i)sex"}|{Token.string ==~"(?i)age"})
({Token.string == "-"})?
{Token.string ==~"(?i)matched"}
)
|
(
 ({Token.string ==~"(?i)age-"}|{Token.string ==~"(?i)sex-"})
{Token.string ==~"(?i)and"}
({Token.string ==~"(?i)sex-"}|{Token.string ==~"(?i)age-"})
{Token.string ==~"(?i)matched"}
)
|
(
({Token.string ==~"(?i)age-"}|{Token.string ==~"(?i)sex-"})
{Token.string ==~"(?i)and"}
({Token.string ==~"(?i)sex-matched"}|{Token.string ==~"(?i)age-
matched"})
)
|
(
{Token.string ==~"(?i)mice"}
{Token.string ==~"(?i)of"}
({Token.string ==~"(?i)both"}|{Token.string==~"(?i)either"})
({Token.string ==~"(?i)sexes"}|     {Token.string ==~"(?i)gender"})
)
```

Frozen lexical expression used as anchors inside the rules for the identification of age

```
Macro: gender
(
{Token.string ==~"(?i)male"}|{Token.string
==~"(?i)female"}|{Token.string ==~"(?i)males"}|{Token.string
==~"(?i)females"}
)

Macro: weeks
(
{Token.string ==~"(?i)mts"}
|
{Token.string ==~"(?i)months"}
|
{Token.string ==~"(?i)days"}
|
{Token.string ==~"(?i)week"}
|
{Token.string ==~"(?i)wk"}
|
{Token.string ==~"(?i)weeks"}
|
{Token.string ==~"(?i)wks"}
|
{Token.string ==~"(?i)month-old"}
|
{Token.string ==~"(?i)months-old"}
|
{Token.string ==~"(?i)mts-old"}
|
{Token.string ==~"(?i)week-old"}
|
{Token.string ==~"(?i)weeks-old"}
|
{Token.string ==~"(?i)wks-old"}
|
{Token.string ==~"(?i)wk-old"}
|
{Token.string ==~"(?i)day-old"}
|
{Token.string ==~"(?i)days-old"}
|
{Token.string ==~"(?i)months"}{Token.string =="-"}{Token.string
==~"(?i)old"}
|
{Token.string ==~"(?i)month"}{Token.string =="-"}{Token.string
==~"(?i)old"}
|
{Token.string ==~"(?i)mts"}{Token.string =="-"}{Token.string
==~"(?i)old"}
|
{Token.string ==~"(?i)week"}{Token.string =="-"}{Token.string
==~"(?i)old"}
|
{Token.string ==~"(?i)wk"}{Token.string =="-"}{Token.string
==~"(?i)old"}
|
{Token.string ==~"(?i)weeks"}{Token.string =="-"}{Token.string
==~"(?i)old"}
|
{Token.string ==~"(?i)wks"}{Token.string =="-"}{Token.string
```

```
==~"(?i)old"}
|
{Token.string ==~"(?i)days"}{Token.string =="-"}{Token.string
==~"(?i)old"}
|
{Token.string ==~"(?i)day"}{Token.string =="-"}{Token.string
==~"(?i)old"}
|
{Token.string ==~"(?i)month"}{Token.string ==~"(?i)old"}
|
{Token.string ==~"(?i)mts"}{Token.string ==~"(?i)old"}
|
{Token.string ==~"(?i)months"}{Token.string ==~"(?i)old"}
|
{Token.string ==~"(?i)wks"}{Token.string ==~"(?i)old"}
|
{Token.string ==~"(?i)weeks"}{Token.string ==~"(?i)old"}
|
{Token.string ==~"(?i)wk"}{Token.string ==~"(?i)old"}
|
{Token.string ==~"(?i)d"}{Token.string ==~"(?i)old"}
|
{Token.string ==~"(?i)week"}{Token.string ==~"(?i)old"}
|
{Token.string ==~"(?i)weeks"}{Token.string ==~"(?i)of"}{Token.string
==~"(?i)age"}
|
{Token.string ==~"(?i)wks"}{Token.string ==~"(?i)of"}{Token.string
==~"(?i)age"}
|
{Token.string ==~"(?i)week"}{Token.string ==~"(?i)of"}{Token.string
==~"(?i)age"}
|
{Token.string ==~"(?i)wk"}{Token.string ==~"(?i)of"}{Token.string
==~"(?i)age"}
|
{Token.string ==~"(?i)month"}{Token.string ==~"(?i)of"}{Token.string
==~"(?i)age"}
|
{Token.string ==~"(?i)months"}{Token.string ==~"(?i)of"}{Token.string
==~"(?i)age"}
|
{Token.string ==~"(?i)mts"}{Token.string ==~"(?i)of"}{Token.string
==~"(?i)age"}
|
{Token.string ==~"(?i)days"}{Token.string ==~"(?i)of"}{Token.string
==~"(?i)age"}
)

Macro: whole_string_age
(
{Token.string ==~"(?i)five-week-old"} |{Token.string ==~"(?i)six-week-
old"}|{Token.string ==~"(?i)five-weeks-old"}|{Token.string
==~"(?i)six-weeks-old"}|{Token.string ==~"(?i)two-week-
old"}|{Token.string ==~"(?i)three-week-old"}|{Token.string
==~"(?i)seven-week-old"}|{Token.string ==~"(?i)eight-week-
old"}|{Token.string ==~"(?i)nine-week-old"}|{Token.string ==~"(?i)ten-
week-old"}|{Token.string ==~"(?i)five-week"} |{Token.string
==~"(?i)six-week"}|{Token.string ==~"(?i)five-weeks"}|{Token.string
==~"(?i)six-weeks"}|{Token.string ==~"(?i)two-week"}|{Token.string
==~"(?i)three-week"}|{Token.string ==~"(?i)seven-week"}|
      {Token.string ==~"(?i)eight-week"}|{Token.string ==~"(?i)nine-
```

```
week"}|{Token.string ==~"(?i)ten-week"}
)

Macro: wholte_string_age2
(
{Token.string ==~"(?i)five-"} |{Token.string ==~"(?i)six-
"}|{Token.string ==~"(?i)two-"}|{Token.string ==~"(?i)three-
"}|{Token.string ==~"(?i)seven-"}|  {Token.string ==~"(?i)eight-"}|
      {Token.string ==~"(?i)nine-"}|      {Token.string ==~"(?i)ten-"}
)

Macro: numbers
(
{Token.string ==~"[0-9]"}|{Token.string==~"[0-9]+"}|{Token.string
==~"(?i)one"}|{Token.string ==~"(?i)two"}|{Token.string
==~"(?i)three"}|{Token.string ==~"(?i)four"}|{Token.string
==~"(?i)five"}|{Token.string ==~"(?i)six"}|{Token.string
==~"(?i)seven"}|{Token.string ==~"(?i)eight"}|{Token.string
==~"(?i)nine"}|{Token.string ==~"(?i)ten"}|{Token.string
==~"(?i)eleven"}|{Token.string ==~"(?i)twelve"}|{Token.string
==~"(?i)thirteen"}|{Token.string ==~"(?i)fourteen"}
)

Macro: link
(
{Token.string ==~"(?i)to"}|{Token.string =="-"}|{Token.string =="-"}
)
```

Rules for the identification of age

```
(
({Token.string==~"(?i)embryos"}):age
)
|
(
{Token.string==~"(?i)mice"}
({Token})[0,2]
{Token.string==~"(?i)age"}
{Token.string==~"(?i)of"}
((numbers)
({Token})[0,1]
(numbers)?
(weeks)):age
)
|
(
{Token.string==~"(?i)mice"}
({Token})[0,1]
({Token.string==~"(?i)aged"})?
((numbers)
({Token})[0,1]
(numbers)?
(weeks)):age
)
|
(
{Token.string ==~"(?i)mice"}
{Token.string ==~"(?i)were"}
{Token.string ==~"(?i)used"}
{Token.string ==~"(?i)for"}
{Token.string ==~"(?i)experiments"}
```

```
{Token.string ==~"(?i)at"}
((numbers)
({Token})[0,1]
(numbers)?
(weeks)):age
)
|
(
{Token.string ==~"(?i)mice"}
{Token.string ==~"(?i)aged"}
{Token.string ==~"(?i)between"}
((numbers)
({Token})[0,1]
(numbers)?
(weeks)):age
)
|
(
{Token.string ==~"(?i)mice"}
{Token.string=="("}
{Token.string ==~"(?i)average"}
{Token.string ==~"(?i)age"}
((numbers)
({Token})[0,1]
(numbers)?
(weeks)):age
)
|
(
{Token.string ==~"(?i)mice"}
{Token.string ==~"(?i)were"}
{Token.string ==~"(?i)used"}
{Token.string ==~"(?i)before"}
((numbers)
({Token})[0,1]
(numbers)?
(weeks)):age
)
|
(
((numbers)
({Token})[0,1]
(numbers)?
(weeks)):age
({Token})[0,2]
{Token.string==~"(?i)mice"}
)
|
(
((numbers)
({Token})[0,1]
(numbers)?
(weeks)):age
({Token})[0,1]
{Token.string ==~"(?i)C"}
{Token.string ==~"57"}
{Token.string ==~"(?i)bl"}
{Token.string =="/"}
{Token.string ==~"6"}
({Token})[0,1]
{Token.string ==~"(?i)mice"}
)
```

```
|
(
{Token.string==~"(?i)aged"}
((numbers)
({Token})[0,1]
(numbers)?
(weeks)
({Token})[0,1]):age
({Token.string==~"(?i)were"}|{Token.string==~"(?i)are"})
)
|
(
((numbers)
({Token})[0,1]
{Token.string==~"(?i)to"}
(numbers)
({Token})[0,1]
(numbers)?
(weeks)):age
(gender)
)
|
(
(gender)
({Token})[0,1]
((numbers)
({Token})[0,1]
(numbers)?
(weeks)):age
)
|
(
((numbers)
({Token})[0,1]
(numbers)?
(weeks)):age
(gender)
)
|
(
(wholte_string_age):age
({Token})[0,1]
(gender)
)
|
(
((wholte_string_age2)
({Token})
(numbers)
({Token})[0,1]
(numbers)?
(weeks)):age
(gender)
)
|
(
((whole_string_age2)
({Token})
(wholte_string_age)
({Token})[0,1]
{Token.string ==~"(?i)old"}):age
(gender)
)
|
(
{Token.string ==~"(?i)were"}
({Token.string ==~"(?i)purchased"}|{Token.string=~"(?i)used"})
({Token})[0,1]
((numbers)
({Token})[0,1]
(numbers)?
(weeks)):age
)
|
(
{Token.string==~"(?i)aged"}
{Token.string==~"(?i)to"}
((numbers)
({Token})[0,1]
(numbers)?
(weeks)):age
)
```