Reproducibility and reliability of biomedical research: improving research practice

Symposium report, October 2015

The Academy of Medical Sciences

BBSRC bioscience for the future

MRC Medical Research Council

wellcome trust

This report is a summary of a symposium held in April 2015 and reflects the views expressed, but does not necessarily represent the views of all participants, the Academy of Medical Sciences, BBSRC, MRC or Wellcome Trust. Contributions by the steering group and symposium attendees were made purely in an individual capacity and not as representatives of, or on behalf of, their affiliated universities, hospitals, organisations or associations.

We are most grateful to Professor Dorothy Bishop FRS FBA FMedSci and members of the steering group who led this project and all individuals who contributed to the symposium and report. The report has been approved for publication by the four coordinating organisations.

All web references were accessed in October 2015.

Reproducibility and reliability of biomedical research: improving research practice

Contents

Summary

Recent reports in both the general and scientific media show there is increasing concern within the biomedical research community about the lack of reproducibility of key research findings. If too many results are irreproducible, it could hinder scientific progress, delay translation into clinical applications and waste valuable resource. It also threatens the reputation of biomedical science and the public's trust in its findings. To explore how to improve and optimise the reproducibility of biomedical research, the Academy of Medical Sciences, the Biotechnology and Biological Sciences Research Council (BBSRC), the Medical Research Council (MRC) and the Wellcome Trust held a small symposium in April 2015.

The process of scientific research involves conducting experiments to test and/or generate a hypothesis. Results of these experiments are collected and analysed, and then shared with the wider research community through publication. Science progresses as hypotheses are further generated and tested, building on existing findings. Such progress requires that studies are rigorous and the findings reproducible.

Sometimes the results of an experiment may not be reproducible, i.e. when the study is repeated under similar conditions, the same results are not obtained. While a study may be poorly conducted, or even on rare occasions fraudulent, irreproducibility could happen for many legitimate reasons. For example, in biomedical research, it might be due to the natural variability in biological systems or to small changes in conditions. Consequently, there is acceptance in the scientific community that some irreproducibility will occur, but there are concerns about its current scale.

Views differ on how serious an issue irreproducibility is. Systematic efforts to reproduce samples of published results are underway in some fields in an attempt to quantify irreproducibility. The outcomes notwithstanding, symposium participants agreed that by considering how to improve reproducibility, we can ensure that biomedical research is as efficient and productive as possible. This report summarises the discussions at the symposium about potential causes of irreproducibility and ways in which it might be counteracted. Key messages from the symposium are summarised below and in Figure 1:

- **There is no single cause of irreproducibility.** In some cases, poor experimental design, inappropriate analysis and questionable research practices can lead to irreproducible results. Some examples of poor practice are highlighted in Figure 1. Cultural factors, such as a highly competitive research environment and the high value placed on novelty and publication in high-profile journals, may also play a part.
- **There are a number of measures that might improve reproducibility** (which are also represented in Figure 1), such as:
  - **Greater openness and transparency** – in terms of both **methods** and **data**, including publication of null or negative results.
  - Better use of input and advice from other experts, for example through **collaboration** on projects, or on parts of projects.
  - **Reporting guidelines** to help deliver publications that contain the right sort of information to allow other researchers to reproduce results.
  - **Post-publication peer review** to encourage continued appraisal of previous research, which may in turn help improve future research.
  - **Pre-registration** of protocols and plans for analysis to counteract some of the practices that undermine reproducibility in certain fields, such as the post-hoc cherry-picking of data and analyses for publication.
  - Better use of standards and quality control measures, and increased use of **automation** in some cases.
- **A 'one size fits all' approach is unlikely to be effective** and in most cases, no single measure is likely to work in isolation. It will take time to identify and implement the most effective solutions.
- **Overarching factors – both 'top-down' and 'bottom-up' – will drive the implementation of specific measures and ultimately enhance reproducibility. These include:**
  - The **environment and culture** of biomedical research. Robust science and the validity of research findings must be the primary objective of the incentive structure. These should be valued above novel findings and publications in high-impact journals.
  - The need to **raise awareness among researchers** about the importance of reproducibility and how to achieve it.
  - The role for **continuing education and training that improves research methods and statistical knowledge;** this should be targeted at individuals across career stages.
  - **Advice from experts in statistics and experimental design** being made more widely available and sought at the beginning of a project.
  - **Technology and infrastructure** to help deliver better reproducibility. This might include shared virtual lab environments and new tools for data capture and sharing.
  - **Talking openly** within the research community about challenges of delivering reproducible results. Scientists and science communicators, including press officers, have a duty to **portray research results accurately.**
  - The need for a **global approach**, in all senses: funding bodies, research institutions, publishers, editors, professional bodies, and individual researchers must act together to identify and deliver solutions – and they will need to do so at an international level. Cultural change will take time.
- **Measures to improve reproducibility should be developed in consultation with the biomedical research community and evaluated** to ensure that they achieve the desired effects. They should not unnecessarily inhibit research, stifle creativity, or increase bureaucracy

Figure 1: Reproducibility and the conduct of research

Improving reproducibility will ensure that research is as efficient and productive as possible. This figure summarises aspects of the conduct of research that can cause irreproducible results, and potential strategies for counteracting poor practice in these areas. Overarching factors can further contribute to the causes of irreproducibility, but can also drive the implementation of specific measures to address these causes. The culture and environment in which research takes place is an important 'top-down' overarching factor. From a 'bottom-up' perspective, continuing education and training for researchers can raise awareness and disseminate good practice.

Possible strategies

Open data
Openly sharing results and the underlying data with other scientists.

Pre-registration
Publicly registering the protocol before a study is conducted.

Collaboration
Working with other research groups, both formally and informally.

Automation
Finding technological ways of standardising practices, thereby reducing the opportunity for human error.

Open methods
Publicly publishing the detail of a study protocol.

Post-publication review
Continuing discussion of a study in a public forum after it has been published (most are reviewed before publication).

Reporting guidelines
Guidelines and checklists that help researchers meet certain criteria when publishing studies.

1. Introduction

Reproducible and reliable studies are crucial for all scientific endeavours. There has been a growing unease about the reproducibility of much biomedical research, with failures to replicate findings noted in high-profile scientific journals, as well as in the general and scientific media.[1,2,3,4] Lack of reproducibility hinders scientific progress and translation, and threatens the reputation of biomedical science.

Poor reproducibility threatens the reputation of biomedical science and the public's trust in its findings. The consequences of a lack of reproducibility could also impact on the translational pathway. However, the process of scientific endeavour is complex and indeed, in biomedical research, the biological systems used are also complex. It is therefore difficult to isolate the contribution of irreproducibility to, for example, the challenges in translating science into clinical applications. Nonetheless, there seems to be broad agreement that there are issues that should be addressed. The Academy of Medical Sciences, Wellcome Trust, Medical Research Council (MRC) and Biotechnology and Biological Sciences Research Council (BBSRC) held a 1.5 day symposium on 1-2 April 2015 to explore the challenges and opportunities for improving the reproducibility and reliability of pre-clinical biomedical research in the UK.

The goals of the symposium were to:

- Discuss aspects of basic and translational biomedical research that may contribute to problems with reproducibility and reliability.
- Explore the role of incentives for researchers, both academic and commercial.
- Understand the role of training and how it may contribute to the solutions.
- Explore current efforts by key stakeholders in biomedical sciences for improving reproducibility and reliability.
- Discuss challenges around reproducibility and robustness in other disciplines and examine how lessons learned from these disciplines might be applied in biomedical sciences.

The aim was to highlight current initiatives, good practice, and potential solutions for non-clinical biomedical research. In clinical research, issues of reproducibility have been much discussed, especially with regard to clinical trials, and progress has been made in developing solutions. The symposium drew on examples from clinical research, but the principal focus was on non-clinical biomedical research.

Irreproducibility can arise for various reasons, including the deliberate fabrication or falsification of data. However, evidence suggests that this kind of scientific misconduct is less common than questionable or unsatisfactory research practices.[5, 6] While it may not always be easy to separate the two, the meeting organisers felt that discussion of deliberate fraud should be excluded from this short symposium. Box 1 notes some key definitions used in this report. For the purposes of the symposium, we focused on irreproducible research where results cannot be replicated because of:

- Poor experimental design, methodologies and/or practices.
- Inappropriate statistical analysis.
- Incomplete reporting of research studies, including methodological details.

The symposium took place at the Wellcome Trust and was attended by about 80 delegates, many of whom had already been involved in the reproducibility debate. They included academics from a range of disciplines and career stages; representatives from journals and publishers; research funders; journalists and sciencemedia representatives; and individuals

Box 1:Key definitions used in this report [7]

Results are regarded as **reproducible** when an independent researcher conducts an experiment under similar conditions to a previous study, and achieves commensurate results.

A **replication study** is designed to test reproducibility. Although it should be similar to the original study, a replication study need not be identical in terms of methods. Indeed, a perfect copy of a study may be neither feasible nor desirable. The key goal is to establish that the original results can be repeated by an independent researcher using similar methods and analysis. In many cases, researchers would anticipate that findings should generalise beyond the specific conditions of the original study.

In some areas of science, such as studies of humans, there is substantial uncontrolled variation. Here the task is to identify a meaningful signal against a background of noise, and statistics are used to inform the evaluation of results. **Reliability** has a formal meaning in this context, referring to the extent of measurement error associated with a result. In many areas of biomedical science, we would not expect exactly the same result to be obtained in a replication study; the goal is rather to specify, and ideally minimise, measurement error, so we can quantify how far the result is likely to be reproducible.


400
300
200

from industry. The symposium was developed by a steering group chaired by Professor Dorothy Bishop FRS FBA FMedSci (see Annex I for a list of members). The agenda for the symposium is in Annex II and a full list of participants is included in Annex III. Participants were predominantly UK-based, but there were also attendees from Australia, Belgium, Canada, France, Germany, the Netherlands, Singapore and the USA. Their contributions confirmed that irreproducibility is a global issue and requires international coordination to address it. It is hoped that examining the challenges and initiatives in the UK may provide a catalyst for solutions across the global research community.

The symposium featured presentations and other contributions from experts across disciplines and sectors, and allowed for ample discussion among attendees. This report covers the emerging themes and key discussion points from the meeting, and does not necessarily reflect the views of the organisations that hosted the symposium, nor of individual participants. We anticipate that this report will have relevance to researchers at all career stages across a wider range of scientific disciplines, scientific publishers, research institutions and universities, funding bodies, and the pharmaceutical and biotechnology industries.

Chapter 2 explores the scale and nature of the problems associated with the lack of research reproducibility. Chapter 3 discusses lessons that can be learned from disciplines within biomedical research and beyond. Chapter 4 considers potential strategies aimed at improving research practice and the reproducibility of pre-clinical biomedical research, while Chapter 5 focuses on how to talk about reproducibility with those outside the scientific community. Chapter 6 brings together conclusions and next steps.

References

1. Prinz F, *et al*. (2011). *Believe it or not: how much can we rely on published data on potential drug targets?* Nature Reviews Drug Discovery **10**, 712.
2. Van Noorden R (2011). *Science publishing: the trouble with retractions.* Nature, **478,** 26-28.
3. The Economist (2013). *Unreliable research: trouble at the lab.* http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble
4. The Lancet (2014). *Research: increasing value, reducing waste.* http://www.thelancet.com/series/research
5. Fanelli D (2009). *How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data.* PLOS ONE **4(5).** doi: 10.1371/journal.pone.0005738
6. John LK, Loewenstein G & Prelec D (2012). *Measuring the prevalence of questionable research practices with incentives for truth telling.* Psychological Science. doi: 10.1177/0956797611430953
7. The use of these terms within the biomedical research community varies slightly, but the definitions noted here reflect the use of these terms throughout this report.

2. What is the scale of the problem?

Overview

- It is difficult to quantify the level of irreproducibility in the published literature, but symposium participants agreed that there is a need to improve the reproducibility and reliability of studies.
- There is no single cause of irreproducibility. Key problems include: poor experimental design, such as underpowered studies and a lack of methods to address potential bias; lack of training; an incentive structure that disproportionately rewards novel, positive results over robust methods and the encouragement of brevity in reporting.
- Questionable research practices, such as overstating a research finding, failing to publish or only partially publishing results, and cherry-picking data or analyses, all contribute to irreproducibility.

Concern about the irreproducibility of biomedical research

The scientific method involves conducting experiments to test and/or generate a hypothesis. The results of these experiments are then analysed and may be published; the wider research community may test the hypothesis further, and build on or translate the findings. Even rigorously conducted studies will yield a proportion of published results that cannot subsequently be reproduced, but which will ultimately be corrected through the process of science. In biomedical research it is not possible to fully control for all the natural variability in biological systems and there is an acceptance that some irreproducibility will occur. But there is increasing concern that the number of findings in the literature that cannot be reproduced is higher than it should be. These concerns are highlighted by recent reports in both the general and scientific media.[8, 9, 10, 11] This is not confined to a particular area of research and has been reported across biomedical science and beyond.[12]

It is difficult to quantify the exact level of irreproducibility in the published literature and so far only limited data are available. The Reproducibility project: cancer biology and the Reproducibility project: psychology are attempting to independently replicate selected results from 50 papers in cancer biology and 100 studies in psychology, respectively.[13, 14] The pharmaceutical industry, among others, has called for reproducibility to be improved and has reported on efforts to investigate the scale of the issue. For example, a group of researchers from Bayer HealthCare examined 67 early stage in-house projects, where published results were being used for target identification or validation, to establish how reproducible the original results were. They found that the published data were completely in line with the in-house data in only about a quarter of cases.[15] Similarly, Amgen attempted to confirm the results of 53 'landmark' studies. They were selected on the basis that they were reporting something completely new; only 11% of the scientific findings from these studies were confirmed.[16] Both papers highlight the impact on the drug discovery pipeline if results published in the literature, which form the basis of drug discovery programmes, are not reproducible.

Causes and factors associated with lack of reproducibility

In his presentation, Marcus Munafò, Professor of Biological Psychology at the University of Bristol, described the key causes and factors leading to irreproducibility in biomedical research. There are many underlying causes, including poor study design, poor statistical practices, inadequate reporting of methods, and problems with quality control. These 'bottom-up' problems are compounded by 'top-down' influences including poor training, and a research culture and career structure that incentivises novelty and publication.

In many biomedical fields, the default approach to evaluating research findings uses a statistical method of null hypothesis significance testing (NHST; see Box 2). Some of the problems of irreproducibility arise when researchers apply these methods inappropriately. Because it is so widely adopted, this report focuses on NHST, but it is important to note that there is considerable debate between statisticians about the optimal approach to evaluating findings.[17]

In 2005, Professor John Ioannidis published a seminal paper entitled "Why most published research findings are false", which Professor Munafò outlined at the meeting. It highlighted six key factors influencing the validity of research:[18]

Box 2: Null hypothesis significance testing (NHST)

Null hypothesis (H_0) and alternative hypothesis (H_1): Much of biomedical science takes a hypothesis-testing approach in which two competing hypotheses are defined, one of which is a null hypothesis (**null hypothesis significance testing** or NHST). For example, a researcher might be investigating whether a specific gene is related to a specific personality trait, say anxiety. The null hypothesis is the theory that there is no effect – in this case, this gene is not related to anxiety. The alternative hypothesis is that there is an effect – that this gene is related to anxiety. Incorrectly rejecting the alternative hypothesis and accepting the null (for example, saying there is no effect of the gene on anxiety when really there is) is a **false negative**. Incorrectly accepting the alternative hypothesis and rejecting the null (saying there is an effect of the gene on anxiety when really there is not) is a **false positive**.

P-values and significance thresholds: In the NHST model, researchers estimate the probability of getting the observed results if the null hypothesis is true. In the example above, they would be assessing the likelihood that the results show a statistically significant relationship between the gene and anxiety when in reality there is no relationship. Researchers decide how certain they want to be and select a significance threshold. The convention in most biomedical sciences is to use a significance threshold of p=0.05, which translates into a 5% probability of obtaining an effect at least as large as that observed if the null hypothesis were true. This means that when adopting a significance threshold of p=0.05 a researcher accepts that 1 time in 20 the results will support the alternative hypothesis when in reality the null hypothesis is true i.e. it will be a 'significant' **false positive**. Some irreproducibility is inevitable in this context, with the rate of irreproducible results depending on the p-value adopted.

1. False discovery rate and small sample sizes[19]

The false discovery rate is the expected proportion of false positives in a set of significant results. To estimate the false discovery rate we need to specify the proportion of hypotheses that are tested that are actually true. Figure 2 considers a scenario where 1,000 hypotheses are tested of which just 100 are true. If the average power of the study is 80%, then in 1,000 studies, 100 true associations will exist, 80 (or 80%) of which will be detected (see Box 3 for a definition of power). In addition, if we used a 0.05 significance level, then 45 (or 5%) of the remaining 900 non-associations will wrongly be considered as significant.[20] In this case, the false discovery rate (i.e. the proportion of false positives) is 45 out of 125, or 36%. It is noteworthy that this is higher than the false positive rate of 5% that most disciplines consider to be acceptable. This problem is worsened when small sample sizes are used. For a given effect size, the smaller the study, the lower the power – i.e. the lower its ability to detect a predicted effect – and therefore the higher the false discovery rate. In practice, it has been estimated that the median power – the overall probability that the statistical tests employed will rightly reject the null hypothesis – in studies in some fields is as low as 20%, or even less.[21] Using the assumptions in the example above, this would give a false discovery rate of 69%. This problem could be addressed by adopting a much more stringent level of significance; e.g. in the above example, with p of .001, only one false positive would be expected.[22]

Box 3: Statistical power and sample size

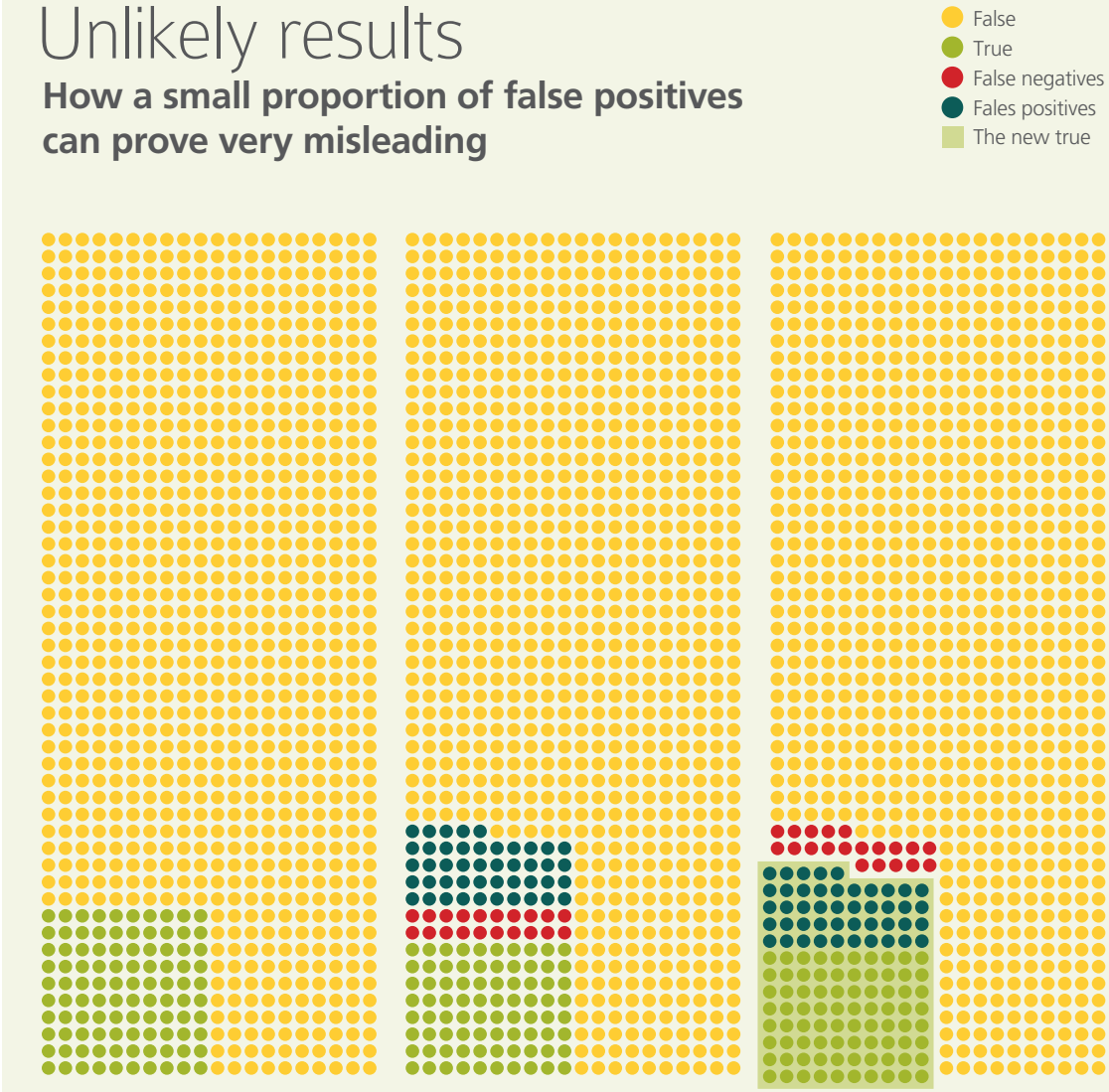
Statistical power: Statistical power refers to the ability of an analysis to detect an effect, if the effect exists. Power is related to the size of the sample (e.g. the number of participants in the example gene study in Box 2) and the size of the effect (e.g. the strength of the relationship between the gene and anxiety). If the power of a set of 10 studies is 0.8, and in all of those studies there is a true effect, then we would expect 8 out of 10 of these studies to detect this effect – the other two will be false negatives. It is intuitive that small studies with low power will be less able to detect small effects, and will be more prone to false negatives. However, as explained in the text, underpowered studies also have a higher false discovery rate compared to well-powered studies (meaning a larger proportion of findings are false positives). This increase in false positive findings is far less intuitive and not widely appreciated.

Sample size calculation: If a researcher knows what size of effect they are expecting to find, it is possible to calculate the size of sample they would need, for a certain level of power. When calculating a sample size, the researcher needs to estimate several unknown parameters, depending on the analysis they are planning to run. Sometimes these parameters may have been reported in previous studies, but if not then this can give researchers a lot of leeway in their power calculations, making them far less useful.[23]

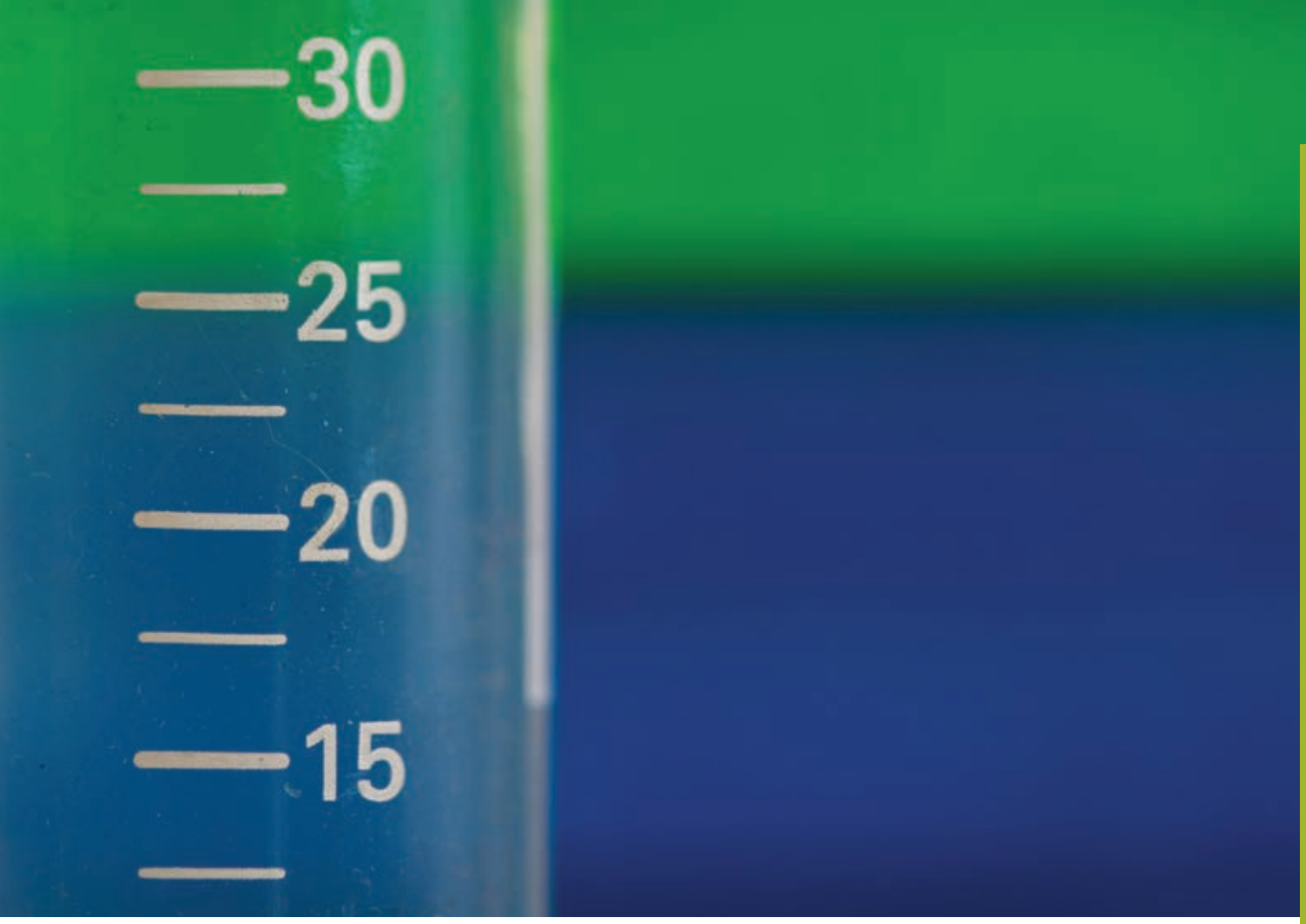
2. Small effect sizes [24]

In many scientific fields, a large proportion of the most easily observed phenomena – the socalled 'low hanging fruit' – have already been discovered. Consequently, researchers are now investigating much more subtle effects, which are generally more difficult to detect. As a study's power is also related to the true effect size of the phenomenon under investigation, research findings are more likely to be true in scientific disciplines with large effects. Effects that are initially thought to be large typically decline with repeated testing, meaning that even when power calculations are performed, effect sizes and statistical power are often overestimated. This is another reason why the false discovery rate in science is higher than often recognised.[25]

Figure 2: Unlikely results



Source:
The Economist (2013). *Unreliable research: Trouble at the lab.*
http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble



3. Exploratory analyses [26]

As further highlighted at the meeting by Dr Katherine Button, NIHR Postdoctoral Fellow at the University of Bristol, researchers are under pressure to publish the results of their research. 'Positive results' that tell a story are more likely to be published than negative, null or inconclusive results (and there are relevant pressures on both journals and researchers).[27,28,29] The motivation to find a significant result can lead researchers to interrogate datasets in multiple ways until a 'positive result' is found. This practice – variously termed p-hacking or data dredging – increases the likelihood that significant findings will be spurious (see Box 4). Indeed, the more analyses that were not originally planned for a study, the less likely the research findings are to be true. Exploratory analyses can be very informative, but they must be presented as such, because they follow a very different process to those in which a hypothesis is set and the study specifically designed to test it (this is discussed further in relation to genomics in Chapter 3). Presenting exploratory analyses as if they were hypothesistesting is misleading and can distort the evidence base for further studies. This kind of 'Hypothesising After Results Known', also known as HARKing or retrofitting hypotheses, contributes to the bias towards positive results in the published literature.

4. Flexible study designs

A related point is that many researchers treat study designs as flexible, modifying them as they progress, so as to increase the likelihood of finding a 'positive result' that can be published. However, the greater the flexibility in study design, the less likely the research findings are to be true, as bias is likely to be introduced. Researchers should be clear about the study aims and analyses before a study is carried out and these should be accurately reported in publications. The lack of transparency in the analytical pathway, combined with flexibility in the way data are presented, can lead to a false sense of certainty in the results and to inappropriate confidence in the reported effects. For instance, in a typical neuroscience study using functional magnetic resonance imaging, there are thousands of different analytic approaches that could be used with the same dataset.

Box 4: P-hacking and HARKing

- **P-hacking:** P-hacking refers to the practice of running multiple tests, looking for a statistic that surpasses the threshold for statistical significance, and reporting only this. The problem is that by running multiple analyses, a researcher will increase the likelihood of finding a statisticallyt significant result by chance alone. For example, if a researcher was studying the relationship between a gene and a battery of 20 different personality questionnaires (all filled in by multiple participants) and did not adjust their significance threshold to take into account the fact that they are running so many tests, we would expect at least one of the personality questionnaires to have a statistically significant relationship to the gene at the 0.05 level, even if in reality there is no relationship. The likelihood that none of the variables will reach the 0.05 level of significance is 1 – 0.95 to the power of N, where N is the number of measures. So with 10 measures, there is a 40% chance that at least one measure will be 'significant'; with 20 measures this rises to 64%. There are various ways of correcting for this issue of multiple comparisons, but they are often ignored by researchers. P-hacking is discussed in Chapter 3 in relation to the multiple comparisons problem in genomics.

- **HARKing:** P-hacking is often coupled with HARKing, i.e. hypothesising after the results are known – here, the researcher invents a plausible-sounding explanation for the result that was obtained, after the data have been inspected.

5. Conflicts of interest and introduction of bias

A different kind of bias is introduced by conflicts of interest. The most commonly discussed cause of potential bias is the source of funding for research and it is important that this is transparently reported in studies. However, non-financial factors, such as commitment to a scientific belief or career progression, can also introduce bias, and researchers should be aware of these when designing and analysing their studies to ensure appropriate controls are in place. Bias is often unconscious, and at the meeting Dr Button and Malcolm Macleod, Professor of Neurology and Translational Neuroscience at the University of Edinburgh, both noted factors that should be controlled for to minimise it; this has been particularly explored in relation to clinical trials. These include:

- Selection bias: systematic differences between baseline characteristics of the groups that are compared.
- Performance bias: systematic differences in how groups are treated in the experiment, or exposure to other factors apart from the intervention of interest.
- Detection bias: systematic differences between groups in how outcomes are determined.
- Attrition bias: systematic differences between groups in withdrawals from a study, which can lead to incomplete outcome data.
- Reporting bias: systematic differences between reported and unreported findings.
- Experimenter bias: subjective bias towards a result expected by the experimenter.
- Confirmation bias: the tendency to search for, interpret, or favour information in a way that confirms one's preconceptions or hypotheses.[33, 34]

6. High-profile scientific fields [35]

The scientific community is incentivised to generate novel findings and publish in journals with a high-impact factor, which can introduce further bias. Some evidence suggests that journals with a high-impact factor are particularly likely to overestimate the true effect of research findings.[36] Studies reported in these journals are arguably at the cutting edge of science, and therefore it might be expected that a proportion of these findings will be subsequently disproved, as research methodologies advance and we gain more knowledge of related fields of research. Authors themselves might drive publication bias, at least as much as journals. They have few incentives to publish so-called 'negative data', and may on balance feel it is not worth devoting time and effort to publish 'negative results' in lower ranking journals at the expense of publishing other 'positive results' in a higher profile journal.[37, 38]

Other factors

Professor Mark J Millan, Director of Innovative Pharmacology at the Institut de Recherches, Servier, highlighted how irreproducibility may be due to poor study design and lack of standardisation, randomisation, blinding and/or automation. Researchers themselves can lack appropriate training in research methods. This can result in a lack of rigour and control for bias, and in insufficient experimental validation. All of these factors increase the likelihood of obtaining false positive or false negative results. Inadequate training and understanding of statistics and data analysis can lead to underpowered studies with sample sizes that are too small, ill- or un-defined end-points, and inappropriate handling and interpretation of data (see Boxes 2 and 4).

Professor Millan also stressed that research papers may lack sufficient information about the materials and methods or data analysis, making it impossible to replicate or reproduce them. Researchers may refuse to disclose full results, including the underlying raw data, and there may be difficulty in accessing the appropriate core materials, such as antibodies, cell lines, genetically modified animals and pharmacological agents. Dr Lawrence A Tabak, Principal Deputy Director at the National Institutes of Health in the USA, outlined additional contributors to irreproducibility, such as problems with the reagents themselves – and authentication of cell lines – and consideration of sex as a biological variable.[39, 40, 41]

The culture and nature of science

While the questionable research practices outlined above often arise through ignorance rather than a deliberate attempt to distort findings, Professor Munafò stressed that they are still damaging to the scientific endeavour. Such questionable practices constitute a grey area and there is evidence that their prevalence is surprisingly high in at least some research fields.[42] These practices most likely stem from a culture in which researchers feel they need to publish novel findings while simultaneously being seen as productive. The incentive structure within which scientists operate is seen as rewarding positive results (e.g. by publication in high-impact journals) over robust methods. Researchers may unconsciously be working to succeed in an environment that can have 'unexpected and potentially detrimental effects on the ethical dimensions of scientists' work'.[43]

Dr Button echoed these points and added that such practices can affect all stages of scientific careers. She highlighted how competition for faculty positions has increased year-on-year. The number of PhDs awarded in science and engineering has been rising steadily since the 1980s, while the number of faculty positions has remained relatively constant.[44] Key predictors for career progression are mainly associated with publication records, including the number of publications or the impact factor of the relevant journal.[45] This hyper-competitive research environment might further exacerbate the problem of reporting only positive findings over negative or inconclusive results.[46] The pressure to attract funding and publish regularly in journals with a high-impact factor may not ultimately incentivise the best scientific practice, especially when a researcher's salary is dependent on factors such as securing grant funding.

More generally, science may be self-correcting over the long-term, but this can only be the case if science is done rigorously. In other words, studies need to be replicated, all research findings (positive and negative) need to be published in a timely fashion, and published findings need to be challenged for this process to function properly.[47]

The emphasis on novelty means that direct replication is surprisingly rare in biomedical science, even in areas where it is not prohibitively expensive or time-consuming. This means that, rather than science being self-correcting, whole lines of research can develop, based on shaky foundations. Bias extends to citation practices as well as publication practices, with cherry-picking of findings that support a researcher's own position, and failure to consider the whole body of evidence.[48] Studies which report positive results that fit a particular theoretical position are preferentially cited, often for years after they have been discredited, leading to a false, distorted sense of certainty.[49]

Responding proportionately

It is important to find the right balance. If researchers are to improve biomedical science, the community must embrace a 'no-blame' culture that encourages them to identify factors that work against reproducibility in their own field. Failure to replicate a study should not be seen as discrediting a researcher's reputation. Rather it should prompt a discussion about the source of the discrepancy in results, so that it can be resolved. A study may not be reproducible for many reasons and lack of reproducibility does not necessarily mean that a result is wrong. Furthermore, focusing solely on reproducibility runs the risk of giving weight to work that may be reproducible without being meaningful; it may be trivial or may suffer from some basic flaw in design or analysis.

As the community attempts to deal with the current problems, care should be taken to ensure science does not become unnecessarily bureaucratic and cumbersome, hampering the creative scientific process that should be encouraged. One example concerns the introduction of guidelines and checklists: these can enhance reproducibility, but they may do more harm than good if extended into fields that do not fit neatly into the kinds of boxes that are used, or which involve exploratory or innovative research.[50] It will be important that the research community is engaged in developing and evaluating the effectiveness of solutions – and that unintended consequences are avoided.

References

8. The Economist (2013). *Problems with scientific research: how science goes wrong.* http://www.economist.com/news/leaders/21588069-scientific-research-has-changed-world-now-it-needs-change-itself-how-science-goes-wrong

9. The Economist (2013). *Unreliable research: trouble at the lab.* http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble

10. Ioannidis JP (2005). *Why most published research findings are false.* PLOS Medicine **2(8),** e124.

11. The Lancet (2014). *Research: increasing value, reducing waste.* http://www.thelancet.com/series/research

12. Bell M & Miller N (2013). *A replicated study on nuclear proliferation shows the critical necessity of reviewing accepted scientific results.* http://blogs.lse.ac.uk/impactofsocialsciences/2013/11/05/reproducing-social-science-nuclear-proliferation/

13. Since the symposium, the *Reproducibility project: psychology* published its results. The mean effect size of the replication studies was half the magnitude of the mean effect size of the original studies; and 36% of the replications had significant results, compared to 97% of the original studies. Open Science Collaboration (2015). *Estimating the reproducibility of psychological science.* Science **349(6251).**

14. For more information on the *Reproducibility project: cancer biology:* osf.io/e81xl/

15. Prinz F, *et al.* (2011). *Believe it or not: how much can we rely on published data on potential drug targets?* Nature Reviews Drug Discovery **10,** 712.

16. Begley CG & Ellis LM (2012). *Drug development: raise standards for pre-clinical cancer research.* Nature **483,** 531–533.

17. Cumming G (2014). *The new statistics: why and how.* Psychological Science, **25(1),** 7-29.

18. Ioannidis JP (2005). *Why most published research findings are false.* PLOS Medicine **2(8),** e124.

19. *Ibid.*

20. Sterne JA & Davey Smith G (2001). *Sifting the evidence – what's wrong with significance tests?* The BMJ **322(7280),** 226-31.

21. Button K, *et al.* (2013). *Power failure: why small sample size undermines the reliability of neuroscience.* Nature Reviews Neuroscience **14,** 365-376.

22. Colquhoun D (2014). *An investigation of the false discovery rate and the misinterpretation of p-values.* Royal Society Open Science **1(3).** doi: 10.1098/rsos.140216

23. Drummond G, *et al.* (2015). T*he fickle p-value generates irreproducible results.* Nature Methods **12(3),** 179-185.

24. Ioannidis JP (2005). *Why most published research findings are false.* PLOS Medicine **2(8),** e124.

25. Button K, *et al.* (2013). *Power failure: why small sample size undermines the reliability of neuroscience.* Nature Reviews Neuroscience **14(5),** 365-376.

26. Ioannidis JP (2005). *Why most published research findings are false.* PLOS Medicine **2(8)** e124.

27. Fanelli D (2011). *Negative results are disappearing from most disciplines and countries.* Scientometrics **90(3),** 891-904.

28 Dwan K, *et al.* (2013). *Systematic review of the empirical evidence of study publication bias and outcome reporting bias — an updated review.* PLOS ONE **8(7),** e66844.

29. Kühberger A, Fritz A & Scherndl T (2014). *Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size.* PLOS ONE **9(9),** e105825.

30. Ioannidis JP (2005). *Why most published research findings are false.* PLOS Medicine **2(8),** e124.

31. Carp J (2012). *On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments.* Frontiers in Neuroscience **6.** doi: 10.3389/fnins.2012.00149

32. Ioannidis JP (2005). *Why most published research findings are false.* PLOS Medicine **2(8),** e124.

33. Higgins JP, *et al.* (2011). *The Cochrane collaboration's tool for assessing risk of bias in randomised trials.* The BMJ **343,** d5928.

34. Higgins J & Green S (2011). *Introduction to sources of bias in clinical trials.* Cochrane Handbook for Systematic Reviews of Interventions. http://handbook.cochrane.org/chapter_8/8_4_introduction_to_sources_of_bias_in_clinical_trials.htm

35. Ioannidis JP (2005). *Why most published research findings are false.* PLOS Medicine **2(8),** e124.

36. Munafò MR, *et al.* (2009). *Bias in genetic association studies and impact factor.* Molecular Psychiatry **14,** 119-120.

37. Littner Y, *et al.* (2005). *Negative results and impact factor. A lesson from neonatology.* Archives of Pediatrics and Adolescent Medicine **159,** 1036-1037.

38. Dwan K, *et al.* (2008). *Systematic review of the empirical evidence of study publication bias and outcome reporting bias.* Public Library of Science ONE **3,** e3081.

39. Masters JR (2013). *Cell-line authentication: end the scandal of false cell lines.* Nature **492,** 186.

40. Lorsch JR, Collins FS & Lippincott-Schwartz J (2014). *Fixing problems with cell lines.* Science **346,** 1452-1453.

41. Clayton JA & Collins FS (2014). *NIH to balance sex in cell and animal studies.* Nature **509,** 282-283.

42. John LK, Loewenstein G & Prelec D (2012). *Measuring the prevalence of questionable research practices with incentives for truth telling.* Psychological Science **23,** 524-532.

43. Martinson BC, Anderson MS & de Vries R (2005). *Scientists behaving badly.* Nature **435,** 737-738.

44. Schillebeeckx M, Maricque B & Lewis C (2013). *The missing piece to changing the university culture.* Nature Biotechnology **31(10),** 938-941.

45. van Dijk D, Manor O & Carey LB (2014). *Publication metrics and success on the academic job market.* Current Biology **24,** R516–R517. This study also showed that gender and university rankings can also influence career progression.

46. Mueck L (2013). *Report the awful truth!* Nature Nanotechnology **8(10),** 693-695.

47. Ioannidis JP (2012). *Why science is not necessarily self-correcting.* Perspectives on Psychological Science **7,** 645-654.

48. Greenberg SA (2009). *How citation distortions create unfounded authority: analysis of a citation network.* The BMJ **339,** b2680.

49. Bastiaansen JA, de Vries YA & Munafò MR (2015). *Citation distortions in the literature on the serotonin-transporter-linked polymorphic region and amygdala activation.* Biological Psychiatry (in press).

50. One such instance was the introduction of rules around microarray experiments, in which some publishers required three independent biological replicas. However, this did not account for all the ways in which microarrays might be used – for example, where authors used microarrays to identify some genes in which expression changed in an interesting way, validated those few genes, and performed detailed functional analyses of their role.

3. What can we learn from disciplines within biomedical sciences and beyond?

Overview

- Reproducibility is not a challenge only for biomedical research, but issues and solutions may differ from field to field. There will not be a 'one size fits all' solution to reproducibility problems.
- The environment in which research is carried out is a critical factor in addressing reproducibility. Collaborative working can help address some problems that affect reproducibility, such as underpowered studies.
- Open and reusable data are important and the data from a project can be considered a 'deliverable', just as a publication is.
- The incentives for researchers are important. The structure of the scientific environment can lead to grant chasing and questionable research practices.
- Technology or infrastructure such as using shared virtual lab environments, and new tools for data capture and sharing, can support reproducibility.
- Some technological advances involve increased complexity of methods and analysis. This can lead to problems unless matched by understanding of that complexity, such as the high risk of false positives when dealing with complex, multivariate datasets.

Individual scientific fields have their own standard methods, approaches and cultures, and they are affected differently by issues of reproducibility and reliability. To identify some lessons that different fields might learn from one another, we heard from disciplines within biomedical science and fields beyond it. Key themes that emerged from across the fields are discussed in more detail at the end of the chapter.

Examples from biomedical fields

Genomics

Genomics is the study of the structure and function of genomes (the complete set of genes in an organism). The techniques used when this field first emerged were at the cutting edge of science and inevitably faced challenges, including ensuring reproducibility of results. Jonathan Flint FMedSci, Professor of Neuroscience and Wellcome Trust Principal Fellow at the University of Oxford, described how the field of genetics has developed robust approaches to dealing with false positive findings. He began by illustrating the difficulties that have beset the field with the example of candidate gene association studies.

Candidate gene association studies consider whether differences between versions of the same gene in different individuals (called alleles) are related to particular traits, such as anxiety or neuroticism. With decreasing costs of genotyping and improved technology, candidate gene studies became easy to do, and this led to many such studies being published. However, subsequent replications typically failed, and meta-analyses (combining information across multiple studies) found little evidence of meaningful relationships between the candidate genes and traits studied.[51]

By contrast, more recent genome-wide association studies (GWAS) have led to robust discoveries. GWAS is a method of interrogating association between phenotypes (observable traits) and common genetic differences across the entire genome. To avoid the increased risk of spurious findings due to running multiple comparisons, researchers set appropriately high significance thresholds. When testing up to a million or more variants for a particular trait the threshold is approximately $p < 0.5 \text{ X} 10\text{-}8$, rather than the $p < 0.05$ so often used in candidate gene studies (Box 4 provides more detail on p-hacking, which is related to the multiple comparisons problem). Consequently, the probability that a given result is a false positive is much lower than in candidate gene studies.

As well as applying stringent thresholds, GWAS have several methodological advantages over older candidate gene studies. Independent replication is seen as essential to confirm evidence of a genetic effect and novel findings are considered with scepticism until they have been reproduced. Furthermore, there is routine implementation of quality control and exclusion of confounding factors. When a paper is peer reviewed before publication these aspects are checked, and findings that could be due to confounding or poor quality control are unlikely to be published. Indeed, there have been instances where quality control problems have led to the retraction of GWAS papers, including some in relatively high-profile journals.[52]

Findings from GWAS have shown that complex behaviours are often influenced by multiple genes and the effects of each gene can be very small.[53] This further highlights the weaknesses of the candidate gene studies that preceded GWAS, which were typically underpowered to detect these effects. Nonetheless, the problems of candidate gene studies are not widely appreciated by non-geneticists, and such studies are still being conducted and published in neuroscience and psychology journals, even though the concerns surrounding the high false discovery rate persist.[54]

Neuroscience

The field of neuroscience, particularly neuroimaging, has played a large role in bringing issues of reproducibility and replicability to the attention of the wider scientific community. Brain imaging has seen huge investment,[55] but the literature in this field faces major challenges in terms of reproducibility and reliability. Dr Jean-Baptiste Poline, a research scientist at the University of California, Berkeley, described why this field has seen such challenges.

Neuroimaging studies are expensive, so they tend to use small numbers of participants, and seldom undertake replications. In the rare cases when replications are performed, the results ofprevious studies may not be reproduced.[56] It seems likely this is largely due to the factors noted in Chapter 2, for example

- Small sample sizes: one paper estimated that the median power of brain imaging studies was just 8%.[57]
- Flexible study designs: the range and complexity of data analysis techniques in neuroimaging studies leaves room for 'flexible analysis', where researchers may run many analyses and then selectively report the analyses that give a positive result.[58]
- Culture and incentive structures: in the field of brain imaging, data are not commonly shared, or if they are, data sharing is sometimes traded for authorship on a paper. The current culture of chasing publications and grants encourages neuroscientists not to collaborate, even though the reliability of neuroimaging results could be improved with more data sharing, transparency and collaboration.[59]

Dr Poline argued that data and code should be considered deliverable outputs of a funded research project, i.e. a resource that should be shared. Indeed, some funding bodies now require that researchers make their data open.[60,61] However, sharing imaging data requires not only the willingness to share, but also the technical expertise and the infrastructure to make sharing possible and to ensure that data are genuinely 'reusable' by others (see Box 5 for an example from neuroimaging). In cases such as this, the infrastructure and tools would need to be developed and maintained, and to be most effective, would need to continue beyond the life-cycle of most grants, which would require longer-term funding mechanisms.

The challenges of data sharing have, of course, received much attention. Dr Poline also suggested that improving the technical skills of brain imaging researchers could make data sharing easier and could also help prevent mistakes due to the use of poorly tested code. Testing and checking code is time consuming, but it is necessary to have confidence in the results. Improved collaboration would also encourage checking by peers.

Brain imaging is one field that has been rather slow to recognise the problems of poor reproducibility. Solutions to the issues in this field are still being developed and will require wide implementation, but increased awareness of the scale of the problem in neuroscience and neuroimaging studies could catalyse change – not only in this field, but beyond it.[62]

Box 5: International Neuroinformatics Coordinating Facility (ICNF) Task force on neuroimaging data sharing[63]

One barrier to sharing neuroimaging data is a lack of the technical infrastructure that makes data easily sharable. One example of efforts to address this is the International Neuroinformatics Coordinating Facility (INCF) Task force on neuroimaging data sharing, which is working with neuroimaging software developers to store results in a common format.[64] The common elements can be then analysed using results obtained from different software packages. The task force has also developed software to make it easier to share certain types of imaging data.[65] XNAT is a public, open image repository within an international organisation that can host and manage imaging data. There is also a counterpart Electrophysiology task force that supports the sharing of electrophysiological data.

Health informatics

Health informatics is the study of methods and technologies for maximising the utility of information for understanding, protecting and improving human health. In the context of big data in biomedical and public health research, informatics is central to tackling the basic problem of scale: that simply pouring more data into conventional research pipelines will compound the problem of non-reproducible findings. Iain Buchan, Co-Director of the National Farr Institute of Health Informatics Research and Clinical Professor in Public Health Informatics at the University of Manchester, described a research climate with a tsunami of data; a blizzard of non-reproducible findings; and a relative drought in human resource and expertise. These factors interact to compound major gaps in the evidence base. He highlighted the discovery gap between population-based and biological studies, for example where a crude clinical classification, such as asthma versus no-asthma, may be too aggregated to uncover important underlying endotypes (or clinical subgroups with distinctive biological mechanisms) of disease risk or treatment outcome. Professor Buchan showed the need for a more joined-up modelling approach whereby a typical fixed classification, such as asthma diagnosis, would be replaced with statistical models of phenotypes, say a latent class model of wheeze trajectory, i.e. a statistical model of subgroups hidden in the data that have distinctive patterns of wheezing over time. Subsequent analyses of biological data resources to search for explanatory mechanisms would, therefore, become more sensitive and specific. In order to achieve this, scientists from different disciplines and organisations need to collaborate in a common digital laboratory. Professor Buchan gave the example of the MRC-funded Study Team for Early Life Asthma Research (STELAR) consortium, which set up such a lab: Asthma e-Lab.[66]

STELAR assembles the data, methodology and disease-specific expertise from five birth cohort studies into one 'team science' network. Combined, STELAR has data on over 14,000 children. Preliminary work used atypical methodology called 'model-based machine learning' to generate hypotheses from data: for example, a subgroup of children with patterns of early allergy predicting subsequent asthma strongly. So the STELAR consortium built Asthma e-Lab to share not only data but also statistical scripts/computational algorithms and conversations over emerging analyses. Investigators log in to a Facebook-like environment where they are notified when colleagues use data or update models on which they are collaborating. Data extracts, statistical scripts, results, comments on results and manuscripts are gathered around research questions. Deeper descriptions of variables that arise in the conduct of research are fed back to the original data sources so that the learning is reused in other studies. Weekly teleconferences support the online 'team science' interaction. This way of working has led to research questions being addressed with more data sources and a greater variety of analytic approaches – producing rapid replication of important findings and more robust modelling across heterogeneous populations.

STELAR's work provides clear examples of how large-scale collaborations and good technological infrastructure can provide reproducible insights with clinical application. The shared digital environment of the e-Lab provided researchers with the tools to efficiently analyse their data and replicate their findings. The field of health informatics that produced the general e-Lab methodology is applying it to other domains of study and linking them, for example to better understand multimorbidity and to couple science with healthcare.[67] There is now an opportunity to better link health and bio-informatics to increase the reproducibility and reliability of biomedical research, particularly in discovering endotypes of disease risk and treatment outcome.

Industry

As noted in Chapter 2, the pharmaceutical industry had a prominent role to play in initially highlighting the poor reproducibility of certain biomedical studies and drawing attention to the implications for translational research and clinical development.[68] Professor Mark J Millan, Director of Innovative Pharmacology at the Institut de Recherches, Servier, described the distinctive aspects of the pre-clinical and clinical research and development process in industry. There is a different array of challenges in industry, compared to research in academia, because processes vary from target characterisation through drug discovery to clinical development.

Professor Millan explained that industry needs results that are not only reproducible and robust, but also therapeutically relevant, in that findings can be used clinically for the benefit of patients. This poses unique challenges. For pharmaceutical companies, in the past, most target validation was done using pharmacological tools (such as molecules that activate or inhibit target receptors). Many complementary agents were available for each target and openly shared so that a broad and robust set of data could be collected to inform clear conclusions. However, today most new potential targets emerge from academic research

using more complex, less standardised techniques like inducible knock-out mice or optogenetics. This relates back to the challenge of moving beyond the 'low hanging fruit' that was noted in Chapter 2. Replicating such highly specialised academic work in industry is more time-consuming and the resources needed (antibodies or mouse lines, for example) not as readily available. Nonetheless, industry must confirm the new and promising findings emerging from high-tech studies using novel agents that can ultimately be translated into therapies for use in patients.

The commercial nature of the pharmaceutical industry also leads to some unique problems regarding intellectual property constraints and limitations on data disclosure. These can hamper independent replications of internal findings. For example, it is difficult to disclose pre-clinical data on a novel agent that has not yet been patented.

Other challenges are common to research in the academic sector, such as a bias against publishing negative results. However, the cause may be different; in academia, negative results may remain unpublished because they are not perceived to be exciting, but in pharmaceutical research, a project may be dropped before completion if it does not support further development. For industry, this saves time and money since pursuing an unproductive target and project is costly, but it means that negative results may not be shared.

Professor Millan noted the value of academia and industry working together to address irreproducibility. He described aspects of data recording and checking in industry that might also be valuable in academia; for example, all data are digitally recorded and traceable, data points are counter-signed, and there are regular inspections of lab books and data. Professor Millan also felt that there may be aspects of the reward systems (e.g. salary, bonuses and promotion) in industry that help because a wider range of activities are rewarded, aside from high-impact publications, such as setting up and validating new screening procedures, taking out patents, or effective project management.

Researchers in industry are concerned with improving reproducibility and reliability of pre-clinical academic research, insofar as industry relies on the findings of academia to highlight new targets. Furthermore, pre-clinical research is now commonly being outsourced by industry to academia, underlining the importance of improving reliability. Translational research linking pre-clinical findings to patients is crucial for industry; it is essential to identify robust biomarkers as medication targets, if they are to deliver efficacious treatments.

Clinical trials

Clinical trials are generally conducted to a very high standard, and are subject to a high level of scrutiny and regulation. Attendees highlighted recent further improvements in the conduct of trials, such as increased transparency of trial conduct and reporting, which could also help to improve the quality of pre-clinical research.

Funders of clinical trials impose certain criteria on research that can improve reliability and reproducibility of clinical trials. Applicants for funding from some sources are expected to have a statistical expert on their team, or at least to have consulted a statistician; most funders also require power calculations to show that the trial will be sufficiently powered to have a reasonable prospect of demonstrating whether or not a treatment works.[69] All trials in the UK now have an electronic trail, as ethical approval for using NHS patients is done through IRAS (Integrated Research Application System). Researchers are expected to have reviewed existing literature to ensure that a new trial will add information and not just confirm what is already known. Clinical trials should be registered before the first patient is recruited, so even if the results of a trial are negative, there will be a record of the existence of a trial, which can be found by other medical researchers. Researchers performing clinical trials are increasingly required by funders, ethical approval committees and journals, to make their data and publications open.[70, 71, 72]

The benefit of publishing all findings, whether positive or negative, is that it reduces publication bias. As noted in relation to pre-clinical research in industry, where studies are cut short due to negative findings, they may still hold useful information for the wider research community, but there is little financial incentive to seeing them published.

There may also be lessons from how the attrition of patients in a study is reported. It is common for clinical trials to present information on how many patients were initially recruited, how many consented, how many were followed up, and so on. This makes clear to the reader why the final dataset is the size it is and it can also help discourage researchers from presenting partial datasets.[73]


455043-0000

Pre-clinical studies rarely document attrition to such a degree, but it can be particularly important if drop-outs are non-random; for instance, if severely affected animals do not survive to outcome assessment this may give a falsely optimistic measurement of the treatment effect.

Applying some of the regulation that exists within clinical trials to non-clinical biomedical research may improve reproducibility and reliability, but there are several key differences worth noting. For example, clinical trials are less exploratory, as by the time a clinical trial is being run there is usually substantial pre-clinical research that has led to that trial being performed. In contrast, pre-clinical biomedical work usually tests hypotheses with less previous evidence. This makes both a systematic appraisal of existing knowledge and more rigorous regulation of individual experiments more challenging. In addition, clinical trial design will have many layers of review while non-clinical biomedical research may follow experiments where only one or two individuals design and test the hypothesis. Clearly, requiring experiment-by-experiment regulation and systematic review of existing evidence would be overly burdensome; nevertheless, current incentives may be driving researchers to spend insufficient time defining their research question, and making them feel pressured to rush on to the next new thing. Time spent in specifying a good experimental design should be encouraged and rewarded in both clinical trials and non-clinical biomedical research. Researchers may need to consider doing fewer, higher quality experiments with adequate replications to ensure key findings are reproducible.

Examples from fields beyond biomedical science

Each field has its own scientific culture and distinctive statistical and methodological procedures. In one session, participants considered whether there might be lessons to be learned from two areas: particle physics (a field that uses large datasets) and manufacturing (where standards and consistent results are paramount).

Big data in particle physics

Tony Weidberg, Professor of Particle Physics at the University of Oxford and collaborator on the ATLAS project, described some of these practices in his own field.[74] First, analysis of new data is typically performed blind, in that some key pieces of information about the data are hidden from researchers performing the analysis. Second, using simulated or partial data, researchers design and refine an analysis, and then apply this analysis to the real data, without altering it. This prevents them from introducing bias into the analysis by tweaking it after the data have been examined.

A third point is that the significance threshold in particle physics is extremely stringent. For a result to be considered 'significant' it must be above 5 sigma, which means that the probability of seeing a positive result when the null hypothesis is actually true is 1 in 3,500,000. The significance thresholds and confidence intervals used in biomedical sciences are substantially less stringent than those used in particle physics research.[75]

Both blind analyses and stringent significance thresholds, such as those used in particle physics, may be useful in improving the reliability of biomedical research, but key differences between the fields must be considered. Achieving higher p-values would not be pragmatic for many biomedical fields, where effect sizes are often small and some degree of uncontrolled variation is inevitable; nevertheless it was agreed that a good case could be made for using a higher threshold than the commonly used $p < 0.05$.

A fourth point was that the scientific culture in the field of particle physics has a much bigger emphasis on collaboration. The ATLAS experiment (studying the Higgs Boson) involves several thousand researchers and is one of the largest collaborative efforts in the physical sciences. Papers are collaboratively read before publication and checked for errors by everyone, including junior members of the team who are encouraged to give critical feedback.[76] One of the effects of these large-scale collaborations is that authorship on papers is not such a key part of a scientist's CV, as some papers have hundreds of authors. Instead, individuals are credited with what they brought to a particular project (for example, a bug they identified and fixed, or code they wrote for analysing data). This kind of recognition system can help avoid the 'publish or perish' culture. Finally, although there is competition to be the first to make a discovery, the findings are not generally considered conclusive until they have been replicated by an independent experiment.

Manufacturing

Poor reliability in the manufacturing industry leads to fragile and slow technology transfer, with the result that companies get low returns on their research and development investments. Dr Matt Cockerill, Founding Managing Director (Europe) at Riffyn Inc., discussed the challenges faced and possible lessons for biomedical science.[77]

In manufacturing, one must be able to replicate products and processes with a high degree of certainty that the results or product will be the same. This requires unambiguous process specifications, capturing all the relevant parameters, and using the captured data to perform root cause analysis to identify the reasons for fluctuations in performance. This has been made possible on an industrial scale using technology. Many processes in manufacturing are digitally controlled and automated using Computer Aided Design or Computer Aided Manufacturing (CAD/CAM). Adoption of this technology has helped to develop unambiguous digital specifications of products and reliable processes.

In academia, published methods do not always provide enough detail for replication. Tools do exist that allow computational analyses to be stored and shared, but these are not standardised across different fields. New companies are emerging for outsourced labs, in which experiments are run automatically using robots.[78] To run an experiment using these labs, researchers send a specification of their experimental design (analogous to a CAD). Such tightly specified experimental procedures are very practical for reuse and replication, both by the original research group and by independent labs. This acts to reduce the amount of unreliability caused by underspecified methods sections.

Other tools are being developed that allow researchers to automatically capture data from the lab environment, and to detect modifying factors (such as the effect of temperature, or batch number of a particular product). Other tools can automate laboratory procedures to ensure consistency. Use of such tools makes methods more tightly specified and more easily replicable. However, in manufacturing, engineers know the product they want to build, whereas in science there is arguably more uncertainty and exploration.

Finally, Dr Cockerill noted that the team-based approach to manufacturing may also be an important factor for achieving high standards.

Emerging themes from the case studies

It is clear that there is not a 'one size fits all' solution to the problems of irreproducibility in these diverse fields. However, several themes emerged from the talks and discussion around these case studies, which may usefully point to ideas that individual fields could take forward and adapt to create solutions appropriate to their own challenges.

First, the research environment appears to be an important factor, including whether research is undertaken by large-scale collaborations or individual research groups. Both styles of research have benefits and drawbacks. An individualistic approach may work better than large-scale collaboration for some fields, and there may be concerns that recognition of the hard work of particular individuals may be lost in large collaborations. Nonetheless, for some areas, such as the GWAS in genomics, the joining up of huge birth cohorts in public health informatics, and the great undertakings of the ATLAS experiment in particle physics, collaboration has been crucial. Without it, these fields would not have been able to collect enough data or do the analyses necessary for significant breakthroughs.

A related topic is that of open and reusable data. This may be especially useful in fields where data collection is very expensive, such as neuroimaging, or where there is a large dataset with important longitudinal aspects, such as the birth cohorts in the STELAR group. Having open data, or data that can be shared with others, means that even if a project is not carried out in formal collaboration with other groups, datasets from individual studies can effectively still be used in a collaborative spirit to increase power and reproducibility. Similarly, open data or a transparent record of the work that has been undertaken combats positive publication bias, and this has been an important development in the conduct of clinical trials.

Incentives and the research culture were also discussed in relation to several fields. We heard from three fields in which the reality of poor reproducibility is perhaps more acutely felt. In industry and manufacturing,


ACADEMIC
SURGERY

poor reproducibility has commercial implications and leads to lost investments, while in clinical trials it affects the risk to participating volunteers and also to patients, who may receive treatments based on unreliable evidence. In these three fields, there are tangible consequences for poor reproducibility, but what are the implications of irreproducibility for a pre-clinical researcher? As discussed by Dr Poline in relation to neuroscience, and further in other chapters, current research culture may incentivise poor research practices because careers in academia depend on publication and grants. It was suggested that academics are not incentivised to be right in the way that industry, manufacturers and clinical trial researchers are.

The research culture should also create an environment that enables and encourages critiquing and challenging work, and rewards individuals for doing so. Concerns were raised that in the current research culture early career researchers and students may fear pointing out potential drawbacks or mistakes in more senior researchers' work. The research culture needs to value reproducibility as well as novel findings.

Technology and infrastructure also emerged as themes across the disciplines we heard from. Technological advances have made it more straightforward to share data and collaborate, and easier to replicate experiments and complex analyses. Public health informatics research has benefited from e-Labs, and clinical trials are commonly logged through a universal electronic system. Fields outside biomedical science, such as manufacturing, have been revolutionised by computer-aided technology which helps to specify processes and products in precise detail, and ensures these processes are reproduced with a high degree of similarity. These approaches are starting to be tested in biomedical sciences. Long-term technological infrastructure for data sharing was highlighted as a need for the field of neuroimaging and potentially in imaging more broadly. Further discussion and examples of data-sharing platforms can be found in Chapter 4.

On the negative side, technology has arguably made questionable research practices easier. For example, in genomics technological advances made certain studies easier to conduct, but this breakthrough was not accompanied by a matched understanding of the risk of false positive discovery. Similarly in neuroscience, technology makes it is possible to run many neuroimaging analyses, adjusting certain parameters or trying slightly different approaches each time, but researchers' understanding of the impact of these questionable research practices in many cases seems lacking. Technology has made 'p-hacking' a quick process, compared to the length of time that running so many analyses would have taken a few decades ago.

This point relates to another common theme from these case studies – statistical practices.
The quite different fields of genomics and particle physics reported the benefits of having stringent statistical thresholds, and the highly flexible analyses of neuroimaging are in contrast to the blind analysis approaches of particle physics. To some extent this problem can be addressed by increased collaboration, which allows researchers to capitalise on the statistical and technical skills of colleagues, as illustrated by the machine learning algorithms employed by the Asthma e-Lab. Such collaboration also tends to encourage more thorough checking of codes and analyses before publication.

It is clear that reproducibility and reliability are issues in all scientific fields, commercial and academic, biomedical and beyond. Potential solutions and current initiatives that span disciplines will also be discussed in the next chapter.

References

51. Munafò MR & Flint J (2004). *Meta-analysis of genetic association studies.* Trends in Genetics **20(9),** 439-444.

52. Ledford H (2011). *Paper on genetics of longevity retracted.* Nature News. http://www.nature.com/news/2011/110721/full/news.2011.429.html

53. Munafò MR & Flint J (2011). D*issecting the genetic architecture of human personality.* Trends in Cognitive Sciences **15(9),** 395-4000.

54. Munafò MR (2014). *Where are the genes for psychological traits?* http://www.in-mind.org/blog/post/where-are-the-genes-for-psychological-traits

55. For example, in the US, the NIH currently funds projects involving brain imaging studies for about $400 million. This number is estimated using a search for 'functional or diffusion MRI' or 'EEG-MEG' with the NIH reporter tool: http://projectreporter.nih.gov/reporter.cfm

56. Bennett CM & Miller MB (2010). *How reliable are the results from functional magnetic resonance imaging?* Annals of the New York Academy of Sciences **1191(1),** 133-155.

57. Button K, *et al.* (2013). *Power failure: why small sample size undermines the reliability of neuroscience.* Nature Reviews Neuroscience **14,** 365-376.

58. Carp J (2012).*The secret lives of experiments: methods reporting in the fMRI literature.* NeuroImage **63,** 289-300.

59. Pernet C & Poline JB (2015). *Improving functional magnetic resonance imaging reproducibility.* GigaScience **4(15).** doi: 10.1186/s13742-015-0055-8

60. The NIHR's policy on open access can be found here: http://www.nihr.ac.uk/policy-and-standards/nihr-policy-on-open-access-for-its-funded-research.htm

61. RCUK's common principle on data policy can be found here: http://www.rcuk.ac.uk/research/datapolicy/

62. Since the symposium, the Stanford Center for Reproducible Neuroscience, was established, which aims to provide researchers with tools to do better science: http://reproducibility.stanford.edu/

63. Details on the INCF Task force on neuroimaging are available here: http://www.incf.org/activities/our-programs/datasharing/neuroimaging-task-force

64. For further information see https://github.com/INCF/nidm

65. Available at: http://xnat.incf.org/

66. Custovic A, *et al.* (2015). *The Study Team for Early Life Asthma Research (STELAR) consortium, Asthma e-lab: team science bringing data, methods and investigators together.* Thorax 2015 **70,** 799–801.

67. Ainsworth J & Buchan I (2015). *Combining health data uses to ignite health system learning.* Methods of Information in Medicine (in press).

68. Prinz F, *et al.* (2011). *Believe it or not: how much can we rely on published data on potential drug targets?* Nature Reviews Drug Discovery **10,** 712.

69. For example, see the guidance on planning a trial outlined on the NIHR's Clinical trials toolkit: http://www.ct-toolkit.ac.uk/routemap/trial-planning-and-design

70. The NIHR's policy on open access can be found here: http://www.nihr.ac.uk/policy-and-standards/nihr-policy-on-open-access-for-its-funded-research.htm

71. The HRA's policy on transparency within research can be found here: http://www.hra.nhs.uk/about-the-hra/our-plans-and-projects/transparency/

72. The European Union clinical trials regulation, which becomes applicable in 2016, states that *'All clinical trials should be registered in the EU database prior to being started.'* http://ec.europa.eu/health/files/eudralex/vol-1/reg_2014_536/reg_2014_536_en.pdf

73. Clinical trialists are encouraged to utilise the CONSORT format for their clinical trial reports. CONSORT's guidance on reporting the attrition of participants can be found here: http://www.consort-statement.org/checklists/view/32-consort/99-participant-flow

74. Details on the ATLAS project can be found here: http://atlas.ch/

75. Lamb E (2015). *5 sigma what's that?* http://blogs.scientificamerican.com/observations/five-sigmawhats-that/

76. The Academy is exploring the challenges that early career researchers face when participating in team science in biomedicine: http://www.acmedsci.ac.uk/policy/policy-projects/team-science/

77. Details on Riffyn can be found here: http://www.riffyn.com/

78. Hayden E (2014). *The automated lab.* Nature **516,** 131–132.

4. Strategies to improve research practice and the reproducibility of biomedical research

Overview

- Some of the strategies that will help improve reproducibility relate to **policies and practices that are already in place**, but need to be further embedded or implemented.
- Sophisticated **statistical knowledge** is needed by researchers at all career stages, and statistics experts may need to be called in more often to help on grant reviews and in editorial boards. Increased requirements for power analysis may help researchers become more familiar with the reproducibility issues surrounding weakly powered studies.
- **Methodology in animal studie**s may be improved with the better implementation of existing guidelines; incentives from funders might also have a role. At present, publications describing animal studies pay insufficient attention to the reporting of measures (e.g. randomisation, blinding) to reduce the risk of the biases outlined in earlier chapters.

- **Standards and quality control** are important, particularly in research using cell lines and tissues, where misidentification of lines and reagents presents a real challenge. Investing in tools for rapid identification and the development of 'gold standard' cell lines may help, as would better reporting of cell lines in published articles.
- **Continuing education and training** may be one way to address the issues and there is evidence of an unmet need among researchers in training around certain aspects of research, such as 'laboratory leadership'.
- Different models of publishing may present some opportunities for improvement. **Pre-publication registration of protocols** has the potential to prevent some questionable research practices and **post-publication peer review** allows for continued appraisal of previous research.
- The Open Science Framework is one practical example of a means of making it easier for researchers to document and register the workflow of their experiments, improving **openness and transparency.**
- **Competition** is embedded in the research system; the community needs to find ways to retain the positive impacts of this while avoiding the negative ones.
- The **whole scientific community needs to be engaged** in addressing reproducibility, especially around issues that relate to scientific culture, such as incentives. Measures to counteract irreproducibility must be proportionate.
- The strategies outlined consider the research environment in the UK, but it was acknowledged that **reproducibility is a global issue**, and examples such as the National Institutes of Health in the US were noted, which has a programme in place to improve reproducibility.

One purpose of the symposium was to propose strategies to address the irreproducibility of biomedical research. Having set out the extent of the problem (Chapter 2) and explored what the research community might learn from other disciplines from either within or beyond the biomedical sciences (Chapter 3), this chapter summarises the proposals that were put forward by participants to improve the reproducibility of research. These included measures involving: the conduct of research; training; publishing scientific findings; openness and transparency; and culture and incentives. This chapter does not aim to represent the views of the meeting sponsors or a consensus among participants, but rather to present an overview of the discussions, which can be used as a basis for further discussions and plans.

Conduct of research

In a two-day meeting it was impossible to cover all aspects of research methods. Our focus was on how the following aspects might be improved: statistics, animal studies, and the use of standards and quality control.

Statistics

There was much discussion of the pressing need to provide opportunities for biomedical researchers to continually update their knowledge of statistics and quantitative skills. While many researchers acknowledge the limits of their statistical understanding and seek expert statistical support where they need it (which might be provided by research institutions or funding bodies), this is not always the case, and it can be difficult to keep up-to-date in this complex and changing field. There is often a need for expert statistical advice, particularly before designing and carrying out an experiment. Researchers should, at the bare minimum, understand the statistical concepts and assumptions they are employing. Continuing education in statistics should be provided to researchers at all career levels, using examples and case studies that are relevant to the researcher, so that they are relatable and easily transferable into practice. Working with simulated datasets is a good way of training researchers, since this can give unique insights into the ease with which 'significant' results can appear if one engages in data-dredging.

It was suggested there might be a greater role for statistics experts on grant review panels and editorial boards. In journals there are already instances of such practice. For example, Nature has committed to examine statistics more closely, demanding precise descriptions of statistics, and commissioning expert input from statisticians when deemed necessary by the referees and editors.[79] There should also be less emphasis on p-values in publications. As noted in Chapter 2, p-values are often misunderstood by researchers who misinterpret a p-value of less than 0.05 as meaning that results are true.

Dr Katherine Button, NIHR Postdoctoral Fellow at the University of Bristol, stressed that there was a lack of awareness around statistical power, particularly in laboratory studies using animals and human participants. Power analysis can provide information on how reliable statistical analyses are, and on the appropriate sample size to use. Sample size is often based on historic precedent (i.e. sample sizes commonly used in similar studies) and little thought is given to the likely effect size of the results. As discussed in Chapter 2, this is problematic because a subtle effect that is hard to detect requires a larger study than one where a more pronounced effect is under investigation. There was some debate as to the value of a priori power analyses and sample size calculations, as these are often based on assumptions that can be easily manipulated in order to obtain a more manageable (and affordable) sample size. However, most participants agreed that even with the limitation of power analysis, it is good practice to consider how the sample size will affect the likelihood of detecting an effect of a given size.

Animal studies

Malcolm Macleod, Professor of Neurology and Translational Neuroscience at the University of Edinburgh, outlined how bias in experimental design (e.g. selection bias in non-randomised studies, performance and detection bias in non-blinded studies, low sample size and inappropriate animal models), in data analysis and in data availability can confound the conclusions drawn from animal studies. In a survey of publications from five UK leading universities (determined by the 2008 Research Assessment Exercise), randomisation, sample size calculation, blinding and exclusion criteria were seldom reported. Only 1 of 1,173 publications described all four, and 68% did not mention any of these.[80] This indicates that these parameters were not viewed as important in the reporting of results, which suggests that either they were not done, or that peer review did not highlight their absence.

There are various guidelines for improving experimental design and the reporting of animal studies, including the Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines from the National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs); the core reporting standards published following a 2012 meeting organised by the US National Institute of Neurological Disorders and Stroke; and publication policies such as Nature's checklist of reporting standards.[81, 82, 83] Participants cautioned against assuming that additional guidelines are needed; the sense of many was that efforts should focus on the implementation of the guidelines that are already in existence, and collecting evidence about which guideline components were most important. Many felt there was insufficient quality assurance of animal studies (achieved in clinical trials through monitoring) and that there should be additional resources to secure better compliance with guidelines.[84] One option might be to introduce

a traffic light or badge system highlighting complete compliance, compliance with minimum standards and non-compliance, which could sit alongside research publications and provide further incentive for adherence to guidelines.

Funders may be able to provide incentive for improvements in experimental design and data analysis of animal studies by emphasising the need for rigour in grant applications. For example, inclusion of plans for randomisation, sample size calculations and blinding could be made conditions of funding. Adherence to these conditions could be monitored through publications (although this may have significant resource implications). Funders could also support the development of a standard format for study protocols and increase available information about ongoing research; for example, efforts to publish lay abstracts when funding is awarded are valuable as they allow the research community to see the full range of funded studies.[85] Institutions could also play a role, for example by carrying out periodic audits of performance and providing training opportunities for scientists at all career stages. Such measures would help to ensure that animal studies are carried out to the highest standards of best practice. Researchers should be held to account, and should proactively make protocols and data available for reuse and scrutiny.

As noted above, sample size is critical to reproducibility and it is important that the validity of animal studies is not compromised by small sample size. Power calculations can help here and it may sometimes be helpful for funders and institutions to provide guidance in experimental design and statistical procedures. Research using animals must be as efficient as possible. There is a risk that efforts to reduce the number of animals used in research might compromise the scientific outcome of the study by reducing its power, and so making the results less reliable. In the future, researchers may carry out fewer but larger studies, which would entail increased collaboration between research groups. The Multi-centre Pre-clinical Animal Research Team (Multi-PART) is currently developing a framework for multi-centre animal studies to help improve translational success of research studies.[86]

Standards and quality control

There are some areas of research where establishing community standards could provide a relatively straightforward way to address some of the issues associated with reproducibility. For example, a simple minimum standard for data sharing that is straightforward and usable could be designed to assist researchers in making their data available. Experiments involving cells and tissues might particularly benefit from some of these initiatives. Indeed, the Minimum Information About a Microarray Experiment (MIAME)[87] could be further developed to define a core set of parameters required in publications to enable replication.

Tools and reagents would likewise benefit from standardisation on a global scale. Notable examples include antibodies and cell lines. An issue that is common to biological research is the validation and quality control of research reagents and cell lines, and their provenance. This was discussed in a break-out group. Evidence suggests that over 400 cell lines used widely across the globe have been misidentified since the 1960s.[88] This presents a challenge to reproducibility, as researchers may be attempting to reproduce results in cell lines that are thought to be the same, but are not, due to cell line misidentification [89] or contamination. These issues are further confounded by genetic drift.[90] The genetic makeup of cells is known to change over time to adapt to the conditions and pressures in which they are cultured. This means that researchers ostensibly working on the same cell lines in two different laboratories may in fact be working on cells that behave very differently owing to differences in genetic makeup.

A number of solutions are possible, for instance investment in tools that allow rapid, efficient and easy genotyping. A central source of 'gold standard' cell lines might be useful to ensure that very similar, if not identical, cell lines could be used by researchers anywhere in the world.[91] Where these exist, researchers should make better use of them. There might also be a role for journals here, in mandating that the source of cell lines be recorded, and perhaps also the results of DNA analysis, where appropriate. Another proposal was that researchers should consider carrying out their research in multiple cell lines and primary cells where possible to determine the broader applicability of their research, when appropriate. Identification and validation of the wide range of antibodies used in research was highlighted as another area where simple measures could help increase reproducibility. For example, many commercial suppliers market more than one antibody against a particular target and while the source of the antibody may be cited in publications, its exact identity is often not specified. Publishers could easily request this information and require that appropriate validation of antibody reactivity and specificity has been undertaken. The progress made in cataloguing the characteristics of the huge range of antibodies used in research, particularly monoclonal antibodies, driven in part by the rise in commercial antibodies, was noted as a positive step forward and could serve as a good example

for other areas. Comparable issues arise with the use of other biological research reagents and similar measures around reporting the source and identity of reagents could help to address these problems.

It is important that all new standards are developed with extensive input from the biomedical community, so that they are relevant and amenable to the diversity of research they might encompass. New standards should be piloted within the research community; new tools to assist with standardisation might increase the uptake of, and adherence to, any new standards that are developed.

Continuing education

Over the course of the meeting, continuing education of researchers at all career levels, including principal investigators, was highlighted as a priority. The format of training courses will have to be carefully considered as the training needs of researchers will be different depending on experience, seniority and area of research.

Dr Veronique Kiermer, Director of Author and Reviewer Services at Nature Publishing Group, outlined Nature's interest in the potential of training courses to address issues associated with irreproducibility. This effort stemmed from concerns around the increase in the number of retractions and corrigenda across its 18 journals over the last decade, many of which were deemed preventable had the research been conducted and reported with more rigour.[92] In 2011, an article in Nature reported that the number of retraction notices had increased 10-fold in a decade, while the literature only expanded by 44%.[93] The scale of the problem indicates that there may be endemic problems in keeping up-to-date with statistical literacy and experimental design. Management of data, people and laboratories has also become considerably more elaborate over the years and many researchers are inadequately prepared to handle this complexity.

Most researchers have already received a high level of training, but many participants agreed that there are specific issues related to reproducibility that need to be addressed. Dr Kiermer told the meeting that a recent Nature Publishing Group survey of almost 1,700 researchers reflected this feeling, revealing an unmet appetite for training in soft skills, such as laboratory leadership and mentoring, and core skills including experimental design and good research practice, data presentation, data management, and peer review. The survey results informed the development of a pilot Laboratory leadership workshop aimed at helping researchers enhance the professionalism and integrity of their laboratory practices and outputs (see Box 6 for further details). Dr Kiermer reported that the workshop received positive feedback and was felt to add significant value, even 18 months after the meeting.

To reach a wider audience, training courses like these would benefit from being translated into a more scalable format, such as an e-learning tool that could be widely distributed and accessed via the internet. Symposium participants also highlighted that it will be important for those who undertake such training courses to be given the time and space for implementing what they have learnt, and to be supported by their institutions to do so.

The impact of training courses will, however, need to be evaluated to assess their effectiveness in driving a change in practice. One suggestion was to assess the publications of participants as a mechanism of monitoring whether the lessons from training courses were being translated into practice.

Nature's survey revealed that lack of funding to attend training courses was a key barrier to accessing the relevant courses to fulfil researchers' needs. This was echoed by many of the participants who considered support and funding for such initiatives to be essential in the future. It would be a false economy to not provide such services, as the potential gains from updating and enhancing researchers' skills far exceed the investment in impact-focused training, in terms of economics and intellectual consequences.

Box 6: Nature's Laboratory leadership workshop pilot

Nature Publishing Group piloted a workshop in 2013 aimed at early career researchers. Lab management and leadership were particular themes, as currently there is little training available to new principal investigators in these areas. Since 2005, Nature has been recognising mentorship through its annual mentoring awards, which provided insights into what good mentorship is. These were used in this training workshop.

Two pilots have taken place, targeting (i) postdoctoral researchers on the cusp of becoming principal investigators and (ii) graduate students. The pilots explored laboratory practices that produce rigorous research outputs, and aimed to develop skills and strategies to guide a research group to adopt these practices. There were sessions on mentoring, data management, analysis and presentation, good practice in determining authorship, and peer review. The aim was not to provide solutions to all of the diverse issues covered, but to raise awareness and encourage people to seek further training and guidance. Sessions drew on case studies based on real-life events, derived from Nature's own database of retractions and corrections. As well as facilitators, the workshop also included academics who could provide first-hand experience of lab management.

Participants were engaged and the real-world examples were deemed particularly effective. Participants' immediate evaluations of the workshop were positive, and they were also followed up 18 months later to see if the training had had any impact on their longer term practices. For those who responded, data management, statistics, and a sense of responsibility for the working of their team were all aspects of research practice that the workshop had continued to influence.

Publishing scientific findings

Publishing the results of research is an essential part of the scientific process. It is the main means by which results can be shared with the wider scientific community and allows experiments to be reproduced, validated or indeed challenged. Crucially, it allows research to progress by building on the studies that have already been conducted.

A researcher's publication record is regarded as a measure of scientific merit that influences career opportunities and further funding decisions. Publication records are also important for academic institutions as a whole as they are taken as indicators of institutions' research capabilities and are important for securing funding for the entire organisation.

However, current publication models have come under criticism for not fulfilling their role properly in terms of providing appropriate scrutiny of results, access to the underlying data and sufficient information in order for studies to be replicated. Furthermore, by focusing on novelty of findings rather than reproducibility, journals can distort the research base. New publication models are emerging. Discussion at the symposium particularly considered two such models and their potential value in addressing some of the issues associated with irreproducible findings: protocol pre-registration, and commenting as a form of post-publication peer review.

Protocol pre-registration

Chris Chambers, Professor of Cognitive Neuroscience at the Cardiff University Brain Research Imaging Centre, proposed pre-registration of protocols as a means to improve reproducibility. Pre-registration is an approach that is widely adopted in clinical trials literature. It aims to ensure that null trials do not disappear from the record, and to prevent authors changing their hypotheses and outcome measures after the results are obtained. The 2013 Declaration of Helsinki now requires registration of all studies involving human participants before recruitment of the first subject.[95] It is possible to pre-register any study through a depository such as the Open Science Framework, described below. The model described by Professor Chambers, which was first adopted by the journal Cortex (see Box 7), goes further by having peer review and in-principle acceptance of a registered protocol before the research outcomes are known, thereby redressing the balance in favour of the hypothesis and the quality of the methods used, as opposed to the results produced. The decision about whether to accept a paper is not dependent on whether the results are statistically significant, novel or are considered to have 'impact'. The potential benefits of this model include:

- **Encouraging good methodology**. In the traditional publishing model, peer review takes place after the study has been conducted, so if the reviewers have concerns about experimental procedures they cannot be taken into account. With pre-registration, peer review of protocols is a constructive process, from which researchers can learn and refine their experimental methodology before doing the study – this can improve specific studies and have a long-term impact on researchers' future work.
- **Making a clear distinction between exploratory and hypothesis-testing analyses**. With pre-registration, p-hacking and HARKing (hypothesising after results are known) are not possible, because the hypothesis, analyses and end points are pre-defined. This does not preclude investigators from including additional, exploratory analyses in the paper, but these cannot be presented as if they were predicted in advance.
- **Guaranteed publication**. Publication is almost certainly guaranteed once a protocol has been accepted, which may have a number of benefits – for example for early career researchers who are building their publication record. It also eliminates publication bias by increasing the reporting of null or inconclusive results. Meaningless null results are, however, minimised, because high statistical power is required.

Image courtesy of Understanding Animal Research (www.uar.org.uk/).

Applicability of pre-registration is currently being explored across various disciplines. Professor Chambers agreed with participants that pre-registration of protocols is not a universal panacea: it does not lend itself to fields which do not involve hypotheses testing, inferential statistical analysis or where experiments are done to develop protocols.

Pre-registration has not been without its critics. Three points that have been raised are concerns about added bureaucracy, intellectual property, and impact on scientific creativity. Professor Sophie Scott FMedSci, Wellcome Senior Fellow at the UCL Institute of Cognitive Neuroscience, expressed reservations about the time and resources required to use new publication models such as this – and a number of participants agreed. It was noted that pre-registration involves at least one round of peer review before studies can begin, which adds to the multitude of other tasks required of researchers, and might simply not be feasible for some contexts, such as student projects. In response, it was pointed out that pre-registration removes the delays that usually occur once a study is completed, because it is not necessary to go through rounds of peer review, which often involve more than one journal. Because peer review occurs before the study commences, pre-registration also avoids the problem where a methodological problem may only

Box 7: Pre-registration of protocols at *Cortex*

Pre-registration describes a different way of publishing research articles, in which the introduction, hypotheses, methods and analysis plans are sent to the journal before the experiment is carried out (a 'Stage 1 submission'). This submission is then subjected to peer review before data are collected. If the study is judged to be of good merit, researchers receive an 'in principle acceptance'. There is a strong guarantee of publication at this point, regardless of whether the study's findings are positive or negative. At this initial peer review stage, reviewers can also suggest changes to the planned experiment and researchers can be invited to revise and resubmit. Submitted studies must also meet statistical power requirements, and include sufficient a priori power analyses or Bayesian sampling and analysis strategies.[98]

Once the research has been done, a manuscript containing the results and discussion is submitted for a second round of peer review. Time-stamped raw data files and certification from all authors that the data were collected after the provisional acceptance are required to ensure that researchers do not 'pre-register' a study that has already been conducted. At this stage, the peer-review process is essentially a quality control mechanism to ensure that the approved protocol has been followed and that the conclusions are justified by the data. If this is the case, the report is published. Any additional unregistered analyses can be published alongside the Registered Report and clearly marked as exploratory or post-hoc analyses that were not in the original pre-registration.

To further encourage exploratory research, the journal Cortex is launching Exploratory Reports in the near future. The purpose of this feature is to generate hypotheses rather than test them; as such there will be no hypothesis testing or p-values.

A list of responses to frequently asked questions about Registered Reports is available at http://bit.ly/1fzG6aN.

be detected by a peer reviewer after the study is completed, rendering the study unpublishable. The protection of intellectual property was also raised as an issue – in both academic and commercial settings. This is addressed, however, by the fact that a dated, registered account of the protocol is openly available. Perhaps the most serious concern was that pre-registration of protocols was seen as counter to the nature of research and would stifle scientific creativity. In response, it was noted that in the model used by Cortex, and all other journals that currently offer Registered Reports, there is no restriction on additional analyses that go beyond the agreed protocol: the only requirement is that such exploratory analyses are clearly distinguished from the analyses that were pre-registered to test a specific hypothesis. Finally, Professor Chambers noted that pre-registration is not being proposed as mandatory across biomedical research, but rather as an optional route to publication that offers benefits to authors while improving reproducibility.

Post-publication peer review

There are various forms of post-publication peer review, including formal approaches (letters to editors, journal commentaries and critiques) and more informal channels, such as blogs and social media. Some journals provide online comment features, but these are still relatively rare. Ms Hilda Bastian, Editor at PubMed Commons, described PubMed Commons as an example of a platform for comments that was developed in response to a demand from the scientific community (see Box 8). This is an open channel for commenting on publicly-funded research.

This kind of commenting could be a powerful tool, if used to its full potential by the scientific community, as it enables rapid scientific discussion about research, almost in real time, thereby enriching and updating the scientific record by author and community curation. Not all journals allow unrestricted letters to editors. Furthermore, questions, which are almost inevitable even with detailed reports, can be raised with authors in an open forum, which may provide an extra incentive for authors to respond, compared with existing communication channels like e-mail. PubMed Commons has, for instance, been used to request data and code from authors.

To date, the overall use of PubMed Commons has been low, although the quality of the comments is high. Three possible reasons for this were discussed:

- **Poor awareness of the tool or ongoing discussions.** One solution would be for notifications to be sent to authors when their papers are being discussed, a feature which could be added to reference management systems.
- **Poor incentives for researchers to allocate time to reviewing and commenting on publications.** Doreen Cantrell CBE FRS FRSE FMedSci, Professor of Cellular Immunology at the University of Dundee, noted during the panel discussion that commenting is analogous to discussion at scientific meetings, and is a good innovation that researchers should engage with. However she pointed out that time pressures can make it very difficult for them to get involved, especially when there are few incentives to do so. The British Medical Journal was cited as an example of a journal that has a vibrant commenting system, which was developed within a strong community.
- **Fear of potential consequences.** Comments on PubMed Commons are attributed, which may discourage some researchers from taking part, particularly early career researchers who may feel that leaving critical comments on more senior researchers' work could impact on their career prospects. Anonymity, however, does not seem to provide a solution, because experience indicates that discussion can become destructive and abusive on platforms where commenters can hide their identity. Ms Bastian also argued that anonymity is not the major barrier to participation. A culture shift is needed where constructive criticism is seen as a positive, so that researchers who comment effectively are given credit for this.

Participants generally felt that there is a lot of potential for this type of post-publication peer review to improve scientific debate. However, to fulfil its promise, commenting needs to become more normative and be better incentivised.

There are currently limited mechanisms for modifying the published literature to note that a study is not reproducible. There are legitimate reasons why this might be the case, but at present, the options for recording this are limited and usually involve retracting a paper. Dr Damian Pattinson, Editorial Director of the Public Library of Science journal, PLOS ONE, noted that retraction is usually perceived as the result of

misconduct, and this has implications for both the authors and the journal. Authors who discover honest mistakes in their work and move to correct them should not suffer reputational damage. Dr Pattinson suggested that a broader selection of options to update the literature as science progresses might be needed. Post-publication commenting could play a role, but other options also need to be considered.

Box 8: Post-publication peer review and PubMed Commons

Post-publication peer review can take many forms and refers to a variety of activities, some of which are already embedded in the scientific community. Post-publication peer review ranges in formality, from informal discussion of papers at journal clubs or on social media (including Twitter and blogs) to more formal options like systematic reviews and letters to the journal editor. Commenting is also another form of post-publication peer review.

PubMed Commons is a forum set up by the US National Institutes of Health (NIH), in which authors of publications on PubMed Commons can comment on published papers. The pilot of this initiative began in October 2013 and comments were made visible to the public in December 2013. In March 2015, there were nearly 9,000 members with close to 3,000 comments on over 2,400 articles. So far, while the quality of comments is high, the overall number of comments has been low, and only a few authors have responded to comments on their work.

PubMed Commons is now investing in a pilot programme for journal clubs to be able to post the discussion of papers on PubMed Commons. The intellectual effort that goes into journal club discussion is a valuable source of post-publication peer review, but these efforts are not currently being captured.

Pre-publication peer review

The conventional form of peer review, where manuscripts are sent out to expert reviewers for their views on the experimental results and validity of the conclusions before publication, was discussed in a break-out group. Those involved felt that tools such as proformas and checklists could help improve the quality of reviews. The system might be improved through careful training and guidance for editors and reviewers to ensure that they are properly enacting the policies of the associated funding bodies and journals. Specialist categories of peer reviewers would be helpful for identifying whether all the relevant aspects of a study have been appropriately reviewed (for example statistics, methodologies, and so on).

Openness and transparency

Low levels of openness and transparency contribute to the issues of irreproducibility. Indeed, the lack of detail in the materials and methods section of a publication, poor availability of research materials or analysis tools, the lack of data sharing, and the withholding of code are all factors that can prevent a study from being accurately replicated. Various strategies have been developed to address some of these issues, including:

- **Journals providing additional space for expanded materials and methods**. Nature, for example, has abolished restrictions on the length of the methods section in their journals.[99] It was noted, however, that reporting of full methods is not obligatory and few authors make use of this extra space.
- **Journals mandating that the data underlying findings are made available in a timely manner**. This is already required by certain publishers such as the Public Library of Science (PLOS)[100] and it was agreed by many participants that it should become more common practice.[101]
- **Funders requiring that data be released in a timely fashion**. Many funding agencies require that data generated with their funding be made available to the scientific community in a timely and responsible manner.[102, 103, 104]
- **The development of platforms for data sharing**. The Center for Open Science was introduced by Dr Courtney Soderberg, Statistical Consultant at the Center for Open Science in the US, as an example of such a platform.[105] It provides a free, open source infrastructure to support open research practices, thereby increasing the ability to run replications, check for computational reproducibility and determine the robustness of the statistical analysis. This Open Science Framework (OSF) enables researchers to make all the research inputs and outputs – including materials, methods, data and code – freely available (see Box 9).[106]

However, participants questioned whether these practices went far enough. To encourage high-quality openness and transparency, 'top-down' and 'bottom-up' approaches will be needed. These should ideally be aligned globally to prevent duplication of effort and ensure worldwide harmonisation of practices.

Possible 'top-down' approaches include institutions and funders requiring researchers to deposit data in a repository, mandating publication of protocols, or better rewarding and recognising data sharing. One suggestion was that funders withhold a proportion of the grant award pending data publication or deposition.[107] 'Bottom-up' approaches driven by researchers might gather more momentum if there were examples of good practice or if the benefits of data sharing were clearly demonstrated. Such benefits include networking, increased research outputs and the efficiency of research that builds on previous datasets. This might help increase adherence to funding agencies' data access policies by allaying researchers' concerns around the reuse of data.

Delivering greater openness and access to data will require investment in infrastructure, data curation and greater interoperability of datasets. Data-sharing initiatives have seen the burgeoning of repositories for data deposition. This has become an issue in itself, as although the data may be available, they may be impossible to find or impossible to use if there is not good curation. Ensuring that data are searchable, retrievable and easy to locate is vital. On the other hand, some participants noted that some datasets will never be used once they are published, so a balance is needed between making data accessible and ensuring the best use of researchers' time. The benefits to researchers and society will need to be clearly demonstrated if the case is to be made for more investment in this area.

Culture and incentives

The culture embedded within the scientific community was cited regularly throughout the meeting as a barrier to improving research reproducibility. The Nuffield Council on Bioethics recently undertook a review into the culture of scientific research in the UK, the findings of which echoed many of the issues that were being raised by participants as contributing to irreproducible results.[110] Professor Ottoline Leyser CBE FRS, Director of the Sainsbury Laboratory at the University of Cambridge, who chaired the review, presented the relevant findings. Input from the scientific community was collected by means of an online survey (with responses from 970 individuals), and via 15 discussion events and evidence-gathering meetings with key stakeholders. The resulting report had limitations inasmuch as the survey and discussion participants were self-selecting and therefore not necessarily representative, though the

Box 9: The Center for Open Science and the Open Science Framework

The mission of the Center for Open Science (COS) is to increase the openness, integrity, and reproducibility of scientific research. One of the Center's big projects has been the development of the Open Science Framework (OSF). The OSF is a free, open source web platform that supports project management, collaboration, archiving, and sharing of scientific workflows and output. Any type of data file can be uploaded onto the OSF and investigators can choose to make them publicly available. Collaborators can be added to a project, granting them access to all of the files (both private and public). The OSF provides a unique, persistent URL for every project, user, component and file, enabling materials to be found and cited.

The tool tracks different versions of uploaded files but also allows researchers to create 'registrations' (time-stamped read-only versions) of certain aspects of a project so that they cannot be further modified. This may be useful for registration of protocols, recording statistical analysis plans or setting hypotheses and ensuring these are not further modified. By tracking changes over time, documenting how data are collected and analysed, and the choices made by the researchers regarding data collection and statistical analyses, the OSF provides greater clarity on how the results were achieved. Such information is valuable for determining the validity of a study, given that flexibility in data collection, analysis and reporting increases the likelihood of false positive findings.[108]

This platform also allows unpublished studies to be made available and provides alternative metrics including citation, download counts, and number of views to monitor impact and gauge interest in or reuse of a project and its data. The COS aims to incentivise good practice by issuing badges to acknowledge open practices. Currently, three types of badges can be displayed on published articles, which recognise open data, open material and pre-registration. Four journals have agreed to display these badges and a number of other journals have expressed an interest. By offering visual recognition of open practices, it is hoped that more researchers will be encouraged to do the same. Badges were adopted by Psychological Science in 2014 and the first badge appeared in May 2014. In the first three months of 2015, approximately 41% of empirical articles published in Psychological Science received badges for openly sharing their data.[109]

composition of respondents broadly reflected the distribution of the research community.
The report highlighted that competition is embedded in the research system. While some aspects of competition are positive, for example because it can bring out the best in people, it can also:

- Encourage poor quality research practices, such as using less rigorous research methods, cutting corners, and hurrying to publish without carrying out appropriate replications or scrutiny of work.
- Hinder collaboration and sharing of data and methodologies.
- Reward self-promotion and 'headline chasing' attitudes.

Symposium participants questioned whether competition was exacerbated by the increased numbers of PhD studentships on offer, which has not been accompanied by a rise in the number of permanent positions available. From the very beginning of a scientist's career, the 'publish or perish' attitude is engrained.

The strong pressure to publish in high-profile journals was found to be driven by perceptions of the assessment criteria of research. Publishing in high-impact factor journals was widely thought to be the most important element in determining whether researchers obtain funding, jobs and promotions. Research Excellence Framework (REF) panels were instructed not to make any use of journal impact factors in assessing the quality of research outputs, but there is a perception that they still used them, which many participants felt adds to the pressure to publish in high-impact journals. It was suggested that the focus on publication-based metrics within the research community means there is little recognition of other types of outputs, such as making data and software available, submitting patents, and so on. High-impact journals, by their very nature, seek to publish cutting-edge, novel research. Consequently, null, negative or inconclusive data, replication studies or refutations are not published or prioritised by researchers. Symposium participants suggested that such results are often published in abstracts in conferences or PhD theses and that there would be merit in making these publicly available, in a format that is searchable and citable.

The 2012 San Francisco Declaration on Research Assessment (DORA), warns against using journal-based metrics as 'a surrogate measure of the quality of individual research articles, to assess an individual scientist's contributions, or in hiring, promotion, or funding decisions'. Symposium participants viewed this as a positive development, and were keen to encourage more UK organisations to sign up to this statement. To rebalance the research assessment system, a wide range of assessment criteria need to be employed. Policies should be clearly communicated and training should be provided so that these can be followed.

Professor Leyser reported a widely held belief among researchers that data sharing facilitates the dissemination of results, enabling research to be carried out more efficiently and allowing for greater scrutiny of findings. However, concerns about commercial sensitivities and practical challenges were raised. Seventy-one per cent of survey respondents thought the peer review system in the UK is having a positive or very positive effect overall on scientists in terms of encouraging the production of high-quality science. Many felt, however, that for the current peer review system to function properly, researchers involved in the peer review process needed to be given the time and recognition for this important aspect of their work. The Nuffield Council on Bioethics' project found mixed opinions about the best way to do this, from remuneration to including it among key criteria in career appraisals, which again reflected the broader symposium discussions.

Professor Dame Nancy Rothwell DBE FRS FMedSci, President and Vice-Chancellor of the University of Manchester, stressed that institutions must take their role in setting incentives seriously and participants agreed that they should do more to actively support the creation of conditions for ethical research conduct. In addition, continuing education in good research practice is important at all career levels. The development of training courses will have to be mindful of time pressures scientists are under. Codes of conduct, such as The concordat to support research integrity, can be helpful to encourage high-quality science and remind researchers of appropriate research practices.[112]

A key outcome of the Nuffield Council on Bioethics' report, which was echoed by participants, was to engage the whole community – including funding bodies, research institutions, publishers and editors, professional bodies, and individual researchers – in acting together to identify and deliver solutions. For an ethical research environment to be created, an open and collaborative research culture should be promoted in which research ethics are embedded, mentoring and career advice for researchers is provided, and an ethos of collective responsibility is nurtured.

Global nature of the issue

Throughout the symposium, participants noted the global nature of this issue and stressed that solutions will depend on international and cross-disciplinary cooperation to galvanise a change in practice. There were a number of international participants present at the symposium (see Annex III). Professor James L Olds, Assistant Director of the Directorate for Biological Sciences at the US National Science Foundation, highlighted that the issue of reproducibility is a key priority in the US and is being considered at the very highest levels of government. He noted the effect irreproducibility is having on the confidence of investors, including US taxpayers. Professor Olds described how the nature of science is changing, for example with the emergence of big data, and that such transformation presents the scientific community with a real opportunity for collaboration. In this regard, the biomedical sciences have a lot to learn from physics and other disciplines that have made great strides in collaborati on and have seized this opportunity to catalyse the future careers of young researchers.[113]

Dr Lawrence A Tabak, Principal Deputy Director at the US National Institutes of Health (NIH), outlined the NIH's plans to address reproducibility, which include efforts to:

- **Raise community awareness:** Workshops with journal editors, industry, professional societies and institutions have identified common areas for action. Over 135 journals have endorsed the Principles and guidelines for reporting pre-clinical research published by the NIH in November 2014.[114] These principles and guidelines aim to facilitate the interpretation and replication of experiments as carried out in the published study.
- **Enhance formal training:** The NIH is developing a series of training modules to address the underlying issues and is providing funding for around 20 projects aimed at developing training modules to enhance data reproducibility. The aim is for such resources to be openly available online. The NIH has also held targeted workshops on the interpretation of findings from novel experimental techniques, for example in imaging and structural biology.
- **Protect the quality of funded and published research by adopting more systematic review processes:** The NIH is running a series of pilots to address some of the underlying issues of irreproducibility focusing on, among others the evaluation of the scientific premise in grant applications; checklists and reporting guidelines; approaches to reduce perverse incentives to publish; supporting replication studies; and training. Via PubMed Commons, the NIH are also trying to catalyse interactions between scientists, thereby encouraging scientific discourse and challenging the status quo.[115]
- **Increase stability for investigators:** The NIH plans to award longer grants, targeted at individuals, thereby reducing the pressures researchers feel in the short term to secure funding.[116]

Proportionate solutions

While scientists may differ in their views of the scale of the problem, particularly given the self-correcting nature of science, there is general agreement that there are too many irreproducible results. However, participants stressed the diversity of opinion across the wider biomedical research community. In implementing solutions it is important to find the right balance: delivering improvement, but not introducing new difficulties that outweigh the benefits. Three kinds of concern were noted:

- Financial costs of new measures.
- Time demands that might take researchers away from front-line science.
- Dangers of introducing regulation that could unnecessarily stifle exploratory and creative science.

The effectiveness of proposed solutions needs careful evaluation, with monitoring of unintended negative consequences as well as benefits. Finding proportionate solutions is key to addressing the challenge of reproducibility.

References

79. For further information see http://www.nature.com/polopoly_fs/1.12852!/menu/main/topColumns/topLeftColumn/pdf/496398a.pdf

80. Macleod M, *et al.* (2015). *Risk of bias in reports of in vivo research: a focus for improvement.* PLOS Biology (in press).

81. ARRIVE guidelines. https://www.nc3rs.org.uk/sites/default/files/documents/Guidelines/NC3Rs%20ARRIVE%20Guidelines%202013.pdf

82. Landis S, *et al.* (2014). *A call for transparent reporting to optimize the predictive value of pre-clinical research.* Nature **490(7419),** 187-191.

83. Nature 2015 Reporting checklist for life sciences articles. http://www.nature.com/authors/policies/checklist.pdf

84. The NC3Rs have developed, and will shortly launch, a new Experimental Design Assistant, which is a free online resource to help researchers with the design and analysis of animal experiments. More information: https://www.nc3rs.org.uk/experimental-design-assistant-eda

85. For example, Research Councils UK publish lay abstracts via the Gateway to Research portal: http://gtr.rcuk.ac.uk/

86. For further information see http://cordis.europa.eu/project/rcn/109353_en.html

87. Brazma A, *et al.* (2001). *Minimum information about a microarray experiment (MIAME) – toward standards for microarray data.* Nature Genetics **29(4),** 365-71.

88. For further information see http://www.sciencemag.org/content/346/6216/1452 summary and references 2 and 3 therein.

89. Since the meeting, the Nature journals announced a new policy of increased scrutiny for cell line misidentification. See editorial: http://www.nature.com/news/announcement-time-to-tackle-cells-mistaken-identity-1.17316

90. Geraghty R, *et al.* (2014). *Guidelines for the use of cell lines in biomedical research.* British Journal of Cancer **111(6),** 1021-1046.

91. For example, the UK Stem Cell Bank is a repository for human embryonic stem cell lines.

92. Van Noorden R (2011). *Science publishing: the trouble with retractions.* Nature **478,** 26-28.

93. *Ibid.*

94. Details on these awards are available at http://www.nature.com/nature/awards/mentorship/

95. Coyne J (2013). *Revised ethical principles have profound implications for psychological research.* PLOS Blogs. http://blogs.plos.org/mindthebrain/2013/10/20/revised-ethical-principles-have-profound-implications-for-psychological-research/

96. The Royal Society's Open Science journal will be launching Registered Reports across all sciences, which will provide an indication of the wider uptake of this format and its potential value: https://royalsociety.org/news/2015/05/royal-society-open-science-to-tackle-publication-bias/ See also https://osf.io/8mpji/wiki/home/ for an evolving list of journals that have adopted Registered Reports.

97. A comparison of the features of the Registered Report format in journals that have so far adopted it is available at: https://osf.io/8mpji/wiki/2.%20Journal%20Comparison/

98. In contrast to null hypothesis significance testing (see Box 2), Bayesian hypothesis testing reveals how strongly the data support one hypothesis over another. See: Dienes Z (2014). *Using Bayes to get the most out of non-significant results.* Frontiers in Psychology **5.** doi: 10.3389/fpsyg.2014.00781

99. For further information see http://www.nature.com/news/announcement-reducing-our-irreproducibility-1.12852

100. For further information see http://journals.plos.org/plosone/s/materials-and-software-sharing

101. For further information see http://www.nature.com/news/data-access-practices-strengthened-1.16370

102. The MRC's data-sharing policy can be found here: http://www.mrc.ac.uk/research/research-policy-ethics/data-sharing/policy/

103. The Wellcome Trust's position on data management and sharing is available at: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Data-management-and-sharing/index.htm

104. The BBSRC's data-sharing policy can be found here: http://www.bbsrc.ac.uk/documents/data-sharing-policy-pdf/

105. For further information see http://centerforopenscience.org/osf/?gclid=CPjQlPbZicUCFSYYwwodY6IAHw

106. For further information see https://osf.io

107. Pampel H, *et al.* (2013). *Making research data repositories visible: the re3data.org registry.* PLOS ONE **8(11),** e78080.

108. Simmons JP, Nelson LD & Simonsohn U. (date is missing) *False positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant.* Psychological Science **22(11),** 1359-1366.

109. Since the symposium, data were released showing the following: between May 2014 and April 2015, 30.3% of empirical articles published in Psychological Science shared their data, and 32.8% shared their materials, compared to 4.4% data sharing and 18.2% materials sharing in the period from January 2012 to April 2014, before badges were implemented. See analysis published online: '*The effect of badges on availability of data and materials*': https://osf.io/rfgdw

110. For further information see http://nuffieldbioethics.org/project/research-culture/

111. For further information see http://www.ascb.org/dora-old/files/SFDeclarationFINAL.pdf

112. For further information see http://www.universitiesuk.ac.uk/highereducation/Documents/2012/TheConcordatToSupportResearchIntegrity.pdf

113. The Academy is exploring the challenges that early career researchers face when participating in team science in biomedicine: http://www.acmedsci.ac.uk/policy/policy-projects/team-science/

114. For further information see http://nih.gov/about/endorsing-jounals.htm

115. For further information see http://directorsblog.nih.gov/2014/08/05/pubmed-commons-catalyzing-scientist-to-scientist-interactions/

116. For further information see http://directorsblog.nih.gov/2014/07/17/3260/

5. Public trust – how do we talk about reproducibility?

Overview

- There is high public trust in scientists, but this must **continue to be earned.** Insofar as lack of reproducibility is a problem, the scientific community should talk about plans to tackle it, rather than trying to hide it from public view.
- Greater effort is needed to **raise awareness of the complexity of the scientific method,** noting that conflict is a natural part of the process. Scientists should not underestimate the general public's scientific understanding, but communicate the complexity of the issues rather than reducing the problem to a simple narrative that may be misleading.
- There is a need to emphasise that **irreproducible research can occur for many legitimate reasons** and this does not necessarily mean that research is fraudulent or poorly conducted.
- Even more than in academic publishing, **publication bias affects which stories are featured in the media.** The need for newsworthiness means replications and null results are almost never published in the popular press.
- It is the **joint responsibility** of researchers, journalists, science writers and press officers to ensure science is accurately reported in the media.

An important component of any discussion of reproducibility is how to talk about the issue – particularly in the context of a wider public debate. Discussion on this topic was led by an expert panel composed of: Ms Fiona Fox OBE, Director of the Science Media Centre; Dr Ian Sample, Science Editor at The Guardian; and Mr Ed Yong, a freelance science writer. Panellists and participants alike recognised that having an honest and transparent debate is the best way to ensure that the credibility of scientific enterprise is not undermined. However, talking about this issue openly must be accompanied by efforts to actively address it. Two themes of challenge emerged from the discussions:

- **Communicating the issue of irreproducibility itself.**
- **The tendency within media outlets to ascribe greater levels of certainty to scientific findings than is justified.**

Communicating the issue of irreproducibility

Discussions about reproducibility in research must be embedded in the broader context of how society interprets the scientific method. Surveys consistently show high levels of public trust in scientists compared to other professions, but this should not be taken for granted – it must be earned and maintained.[117, 118] As noted above, openness and transparency about this issue are important here, but so are efforts to raise awareness of the complexity of the scientific method, which builds upon existing knowledge through experimentation, discussion and, at times, conflicting results. Panel members highlighted the need to communicate and embrace the fact that disagreement is a natural part of the scientific process. Current scientific consensus is based on 'working truths' which are subject to revisions if evidence to the contrary is discovered. Related to this, Dr Sample noted the need to raise public understanding of the fragility of individual studies without undermining entire scientific disciplines.

Much of this report considers causes of irreproducibility that are not traditionally thought of as intentionally fraudulent, in the sense that they do not involve falsification or fabrication (plagiarism is often also noted as fraud, but is not so relevant here). However, while scientists might distinguish between questionable research practices and intentional fraudulent activity, this can be a fine line to draw, and the wider public would be concerned if they find that research funds are spent on poorly conducted studies, even if no deliberate fraud was intended. Efforts to communicate issues relating to reproducibility should recognise this, and should take care to convey the complexities and highly contested nature of science, including irreproducibility issues, without damaging the whole field of study or implying all irreproducible results arise from bad practice. It is important to emphasise that the way science is conducted has changed massively in the past few decades and the scientific method is continually developing to address new challenges as they arise.

Communicating findings that turn out to be irreproducible

Scientific stories generally get high prominence in the UK news agenda. However, science reporting can sometimes get distorted or sensationalised, often when stories move from being covered by specialist correspondents to general news journalists. This broader picture sets the context for efforts to talk about reproducibility. It was clear from the discussion that journalists and writers, press office staff and researchers themselves all have a shared responsibility to provide accurate and nuanced portrayal of research results.

For example, press office staff are often under pressure to deliver results by securing media attention for their researchers and institutions. This can sometimes lead to overstated claims in press releases.[119] A recent study found that exaggeration in news reports on health stories is strongly associated with exaggteration in academic press releases. If methodologically weak but sensational studies are highlighted, then the public may lose confidence in science when they are subsequently debunked. Press officers have a continuing duty to ensure that press releases are not a source of misinformation and Ms Fox suggested that a concordat for science press officers could be helpful in reinforcing these responsibilities. Scientists could arguably play more of a role in scrutinising relevant press releases to ensure that their results have not been misinterpreted. The discussion revealed that ultimately, press officers and scientists should cooperate to strike the right tone in press releases. One suggestion was that scientists and science press officers could use a traffic light system when publicising new studies: red could indicate preliminary, small studies; amber could be used for bigger, peer-reviewed studies that still need to be replicated; and green for the most rigorous studies.

Journalists too can overstate the significance of a study or leave out important caveats such as study limitations. Participants agreed that the media's job is not to boost public trust in science but to

provide an accurate depiction of science, but the media also has a duty to critique extraordinary claims that are unsupported by solid evidence and to be discerning in scrutinising claims made by individual studies.

Communicating uncertainty and 'grey areas' in science is challenging, but journalists and writers should avoid reducing complex topics to binary terms, and not assume that their readership is incapable of understanding such issues. In relation to reproducibility, a single failure to replicate a study does not automatically negate the whole field, but it is hard to convey this when there is a strong tendency to rely on simplistic narratives. It was generally agreed that choosing not to cover a story due to insufficient evidence can be a difficult decision that journalists could arguably make more often.

'Churnalism', where journalists overly rely on press releases, rather than original reporting, is problematic (though is arguably caused by the pressures placed on journalists) and can further compound issues stemming from exaggerated press releases. Mr Yong stressed that journalists and science writers should look beyond single press releases to generate stories, and should identify and contact sources who will provide nuanced views on the papers that they are asked to comment on. This is particularly applicable to interdisciplinary science where it is essential to contact a wide range of experts in multiple fields to understand the relevance of the research and to point out the studies' limitations. Ms Fox stressed that scientists again have an important enabling role to play here – in providing third-party statements that critique studies and put the science in context. Such comments might highlight issues such as poor study design, small sample size, inappropriate use of statistics and the need to treat findings with caution. One of the Science Media Centre's key roles is to facilitate the provision of such comments – it circulates third-party comments about new studies to journalists before an embargo lifts.

References

117. Ipsos MORI (2014). *Public attitudes to science 2014.* https://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-main-report.pdf

118. Wellcome Trust/Ipsos MORI (2012). *Wellcome Trust monitor: wave 2.* http://www.wellcome.ac.uk/About-us/Publications/Reports/Public-engagement/WTX058859.htm

119. Chambers CD, et al. (2014). *The association between exaggeration in health related science news and academic press releases: retrospective observational study.* The BMJ **349,** g7015.

6. Conclusions and next steps

Overview

- Providing further education in research methods may be a starting point in ensuring scientists are up-to-date with new developments and making them aware of issues surrounding irreproducibility.
- Due diligence by research funders can improve research standards.
- Greater openness and transparency is critical to addressing irreproducibility in research. Changes in policies by funders and publishers could drive notable changes in data-sharing practices. Pre-registration of protocols prevents one source of irreproducibility, i.e. post-hoc cherry-picking of data, which could be important in some fields.
- There needs to be a faster culture shift within the biomedical community from a sole focus on valuing novel findings and publishing in high-impact journals, to one that also rewards the quality of the study, including the reproducibility and validity of research findings.
- Addressing reproducibility need not only involve new measures, but also efforts to better embed and implement existing policies and practices.
- Care should be taken in communicating issues of reproducibility without undermining public trust. Science is one of the most powerful drivers of improvements in people's quality of life, and we must be clear that the goal is to make it even better, rather than to disparage scientists. Scientists and science communicators (including journalists and press office staff) have a duty to accurately portray research results.
- Irreproducibility of biomedical research is a global issue – tackling it requires a global approach with multi-stakeholder involvement.
- Wherever possible, strategies to address these issues should be supported by evidence of the effectiveness of these strategies, and that will require more 'research on research'.

Possible strategies for improving the reproducibility and reliability of biomedical research in the UK were outlined in Chapter 4, but to address this issue properly, a clear strategy will be required that includes 'top-down' measures from journals, funders, and research organisations, as well as 'bottom-up' ones from individual researchers and laboratories. This chapter summarises the key conclusions and next steps emanating from the discussions, which are based on areas where we believe there was a high level of consensus at the symposium. In many cases, there are examples of where the suggestions made in this chapter are being developed or implemented; however there may also be benefit in seeing such practice more widely embraced. In addition, many of the main actionable priorities discussed below will require a coordinated effort by multiple stakeholders. In fact, one of the most overwhelming conclusions of the symposium was that improving reproducibility requires a global approach – in every sense. It will require cooperation across disciplines, and between stakeholders, sectors and countries. Consequently, solutions will require a collaborative approach. Views differ on the precise scale of irreproducibility and the ways in which it should be tackled, but participants agreed that there is evidence that there is room for improvement. Solutions will, therefore, need to be both balanced and based on discussion across the biomedical research community.

In terms of the solutions themselves, the overarching message from the meeting was that there needs to be a shift from a sole focus on valuing innovation and novelty, to one that also rewards robust methodology and the reproducibility and validity of research findings. This will require a degree of culture change within the biomedical community. However, with a concerted effort from all the stakeholders, participants were optimistic this could be achieved in practice. In general, scientists are motivated by a love of science, and they want to do science as well as they can. There is real potential for implementing changes that could enable even greater benefits to be gained from scientific research, but all stakeholders will need to join forces to achieve this kind of change.

These conclusions do not represent an exhaustive list, nor do they necessarily represent the views of all the participants, but they are issues that the meeting sponsors will want to take forward in future dialogue with the wider research community.

Raising awareness and continuing education

It was clear from discussions that greater awareness of the issues around reproducibility is needed within the research community, accompanied by a wider discussion about how to address them. Training courses provide one vehicle for raising this awareness and creating the kind of environment that improves research. It is clear that such training should be delivered at all career levels, thereby ensuring that good practice permeates through the community from senior to more junior levels. Courses might nurture both soft skills like laboratory leadership and mentoring, as well as more traditional core skills (such as statistics, experimental design, and good research practice). Delivery will require a concerted effort particularly by funders and research institutions, but it is clear there is an appetite among scientists for this kind of intervention. In the case of funding agencies, they have a role in supporting the scientific community by facilitating the development of this kind of training and promoting it to researchers at all stages. For example, Professor Jacqueline Hunter CBE FMedSci, Chief Executive of the Biotechnology and Biological Sciences Research Council (BBSRC), noted that the BBSRC has been working to provide a broader range of skills through training. E-learning tools were considered by many to be a good format as they could reach a wider audience. Good practice guidelines for a range of subjects (experimental design, data analysis, data sharing, etc) could also be very useful. Institutions can support researchers by giving them the time and space to attend training courses and encouraging them to update their skills and be more constructively critical of other researchers' work. It is important that researchers feel comfortable acknowledging the limits of their understanding (for example in statistics) and seek support or additional educational opportunities where necessary.

Reproducibility considerations when funding science

There was strong agreement among symposium participants that due diligence at the time of funding can improve research standards. Professor Peter Weissberg FMedSci, Medical Director at the British Heart Foundation noted that funders have the first opportunity to review research proposals, meaning they can help to set the tone for the whole process. Research calls and application processes must place clear value on scientific rigour. For example, in some cases, it may be helpful for

funders to incentivise applicants to undertake a more extensive review and evaluation of prior literature, or to fund research where such analysis is the primary goal. To encourage robust methodology, some funders could require more detail in statistical analysis plans; power calculations will be appropriate for many disciplines. Funding agencies may be able to provide guidance here or engage statistical experts at the grant review stage – where they do so already, they might benefit from raising awareness about this activity. Where they exist, good practice guidelines often help researchers go through all the appropriate steps. It may be helpful for funders to review their policies to see if they fully address the issues to optimise reproducibility. Funders might also consider funding well-powered replications or proposals that foster collaborations to achieve the necessary degree of power.

While the symposium identified barriers to reproducibility, many of which are recorded in the academic literature, research into the scientific method will be necessary to fully understand areas of concern, examine how these can be counteracted and demonstrate the value of new initiatives. This kind of activity will be important to inform the development of new practices and safeguard against potential unintended consequences, but it will require funding support. For example, although collaboration was thought by many to be an attractive way to carry out better powered studies, there were concerns that large consortia can be cumbersome and inefficient, and individual contributions may go unrecognised. This last point highlights the need for balance in finding solutions – another element here will be balance between funding new research and funding replications or studies evaluating effectiveness of measures designed to improve reproducibility.

Improving openness and transparency

Fostering a culture of openness and transparency through data sharing was highlighted as a priority to tackle issues of irreproducibility in research. This particularly relates to the roles of funders, journals and publishers in enforcing public availability of research data and analyses.

Research funders can encourage timely data sharing through grant terms and conditions or targeted policies. For example, many funders have data access policies and expect researchers to follow them, but efforts to monitor adherence to such policies vary. Similarly, they may facilitate adherence through developing tools that enable data sharing, and by making it clear that researchers can enhance their reputation by sharing data.

It was felt that more could be done by journals and publishers to encourage publication of null, negative or inconclusive findings, and for data to be made publicly available. This would ensure that scientific findings are not skewed towards a positive representation. Dr Deborah Dixon, Vice-President, Global Publishing Director for Health Sciences at John Wiley & Sons noted that journals and publishers could contribute substantially by providing the tools so that published data are linked with associated articles and hence easy to find. As with funders, publishers could do more to enforce and monitor their policies on issues such as data sharing. Additional support could be provided by journals and publishers to incentivise good publication practice, including good practice guidelines for data sharing and analysis, tools to assist with peer review and recognition of compliance with open data policies, publication guidelines (e.g. ARRIVE guidelines), and so on.

At a minimum, journals should mandate that sufficient information for replication is made available in the final publication (which may need to be online). This includes information such as the source, identification, validation and quality control of cell lines and research reagents. Nature, for example, has already removed restrictions on the length of its methods sections to this effect. Peer reviewers should be asked to request further information where it is not available, and researchers encouraged to publish their protocols and provide a link in the final article. It is also important that researchers reference the breadth of the literature in their publications and avoid citing only the research that supports their findings.[120]

Participants acknowledged that data-sharing practices alone do not go far enough to encourage openness at all stages of the research process. There was strong feeling that more should be done to ensure that researchers are open about the original hypotheses of studies, where they exist. This will help us better understand the scale and causes of irreproducible studies, and how to address them. Pre-registration of study protocols and the use of platforms, such as the Open Science Framework (OSF), have the potential to provide a robust method for defending against post-hoc cherry-picking of data and hypotheses, which is an important source of irreproducibility in many disciplines. It was clear that pre-registration may not be

applicable to all aspects of biomedical science – for example, in basic discovery-led biomedical research where protocols are not well established and are developed alongside the experiments, or in the case of exploratory analysis where a study is not designed to test a specific hypothesis. It is important that the community embarks on further discussion to take into account the needs of different disciplines, and evaluates the impact of new initiatives to be aware of any unintended consequences. The more informal channels of post-publication peer review established by some publishers, such as commenting, are welcome and offer opportunities for further scrutiny of the published literature; it is hoped that they will be embraced by the scientific community.

Changes in journal policies could drive notable changes in practices, such as data sharing, given the current emphasis on publication record as a metric for career progression and securing funding. Professor Richard Horton FMedSci, Editor-in-Chief of The Lancet stressed that a coordinated effort by journals and publishers worldwide (and more broadly with the community) will be needed if major changes in publication policies are to be made. Indeed, there is a considerable choice of journals in which to publish and, therefore, researchers can easily circumvent publishing in certain journals if they do not agree with their publication and data access policies. Furthermore, it is worth remembering that publishing is ultimately a business, and publishers will be sensitive to potential financial implications of particular decisions.

Communicating the issue of irreproducibility

An honest and transparent debate is the best way to ensure that the credibility of the entire scientific enterprise is not undermined, but discussion of reproducibility must be accompanied by active efforts to address it. Care should be taken in communicating the issue of irreproducibility itself, to ensure that it is seen as a complicated phenomenon that has many causes and is not the same as fraud or malpractice. The broader issue of communicating science is also relevant because media outlets often tend towards reporting greater levels of certainty from scientific studies than might actually be the case, which adds to the challenge of talking about irreproducible studies.

Scientists and science communicators (including journalists and press office staff) have a shared responsibility to provide accurate and nuanced portrayal of research results. Press officers should work with scientists to ensure that press releases are not a source of misinformation; they could consider developing a concordat that sets out good practice. Journalists and writers should avoid overstating the significance of a study or leaving out important caveats such as study limitations. Scientists should embrace the opportunity to provide journalists with third-party comments on published research that help them report such limitations and set the science in context. All of these stakeholders should not shy away from talking about the complexity of the scientific method and the fact that science makes progress by continually challenging its own findings.

Valuing reproducible science

There was strong feeling that a culture shift within the scientific community, in which the reproducibility and robustness of research findings is valued as much as the need to generate novel findings, would go a long way in addressing issues of irreproducibility. This will particularly relate to the role of employers i.e. research institutions, in structuring individual incentives, such as hiring and promotion, to encourage good research practice. In addition, researchers themselves should not be complacent or underestimate the difference an individual can make in changing wider culture and research practices.

Symposium participants agreed that research institutions must encourage, reward and recognise behaviours that are conducive to good research practice. These include data sharing, publication of protocols, open data and collaboration among others. Such practices should inform professional opportunities and career progression, thereby reducing the focus on only a small number of criteria and in turn addressing the issue of perverse incentives that can drive poor practice. The Research Excellence Framework (REF) could be an important contributor to generating this kind of culture and a critical factor in culture shift. It was noted that many of the measures suggested would be difficult to implement in practice if they are not incentivised by the REF and this will need further thought. Good practice guidelines might help and many participants felt that institutions should sign the San Francisco Declaration on Research Assessment (DORA), which stipulates that they should not use journal-based metrics as a measure of a scientist's contributions, or in hiring or promotion decisions. In addition, institutions should give researchers enough time to engage properly with some of the measures noted in this chapter – such as writing up negative or inconclusive results before seeking the next grant. These kinds of measures would substantially alleviate the pressures researchers are under and should allow for better research practices.

It will be important that any measures do not introduce new problems, such as stifling creativity or introducing unnecessary bureaucracy. The idea of introducing appropriate measures within institutions would be to replace perverse incentives with a new structure that promotes reproducible research, while delivering greater freedom for researchers so that they can focus on science, innovation and robust research methodology. Getting this right will again require engagement across the community.

While much of the research environment is driven by funders, institutions and publishers, researchers themselves have a duty to be more rigorous, to value robust research methodology, and expect more from themselves and the teams within which they work. Principal investigators should instil an ethical research framework in their labs and encourage good research practice. Researchers must be willing to consider the best way of conducting their research. For example, in order to increase the power of their studies, collaboration might enhance a study, or researchers may need to be prepared to carry out fewer but perhaps larger studies. There may be areas of biomedical research that are well suited to more standardisation, e.g. through the kinds of protocols used in manufacturing, and researchers should consider these.

Research on research

The scientific endeavour is informed by evidence, and this applies as much to any proposed changes in the research landscape as it does to research itself. Many participants at the meeting noted that, while a number of plausible strategies had been proposed, the evidence for the effectiveness of these was limited. This concurs with the conclusions of The Lancet's series, Research: increasing value, reducing waste, that while many problems with biomedical research could be identified, well-evidenced interventions to address these did not exist.[121] There is ongoing research around, for instance, the effectiveness of different editorial strategies in improving the quality of published manuscripts, but further research will be required if improvement strategies are to be maximally effective with minimal burden to the community.

Concluding remarks

There should be an ethos of collective responsibility in addressing issues of reproducibility, as stakeholders from across the biomedical sciences community all have a role to play – and all must be involved in developing solutions, if they are to be accepted and integrated. There are areas where the different communities should come together, for example to achieve the following:

- Implementation of current guidelines, from funders, journals and institutions, aimed at improving robustness and reducing perverse incentives.
- Incentivising good practice including publication of null, negative or inconclusive results and making data more readily available.
- Developing standards for experiments, data sharing, clinical studies, cell lines and reagents with extensive input from the scientific community.
- Rewarding quality of the study over the results to counter the 'publish or perish' culture.

These issues are not confined to the area of biomedical sciences and we hope that some of the strategies put forward in this report will be more widely applicable. The UK will not be able to solve these issues alone, but it is hoped that by setting a good example, other countries will follow suit. Our aim is that this report should ignite further discussions and a call for international action to make much-needed progress in enhancing reproducibility in biomedical sciences.

References

120. Since the symposium, *Guidelines on Transparency and Openness Promotion (TOP)* were published to provide templates to enhance transparency in the science that journals publish, coordinated by the Center for Open Science. At the time of publication, they had been endorsed by over 100 journals: https://osf.io/9f6gx/

121. The Lancet (2014). *Research: increasing value, reducing waste.* http://www.thelancet.com/series/research


Cabinet 4
INFLOW

Annex I: Steering committee membership

Professor Dorothy Bishop FRS FBA FMedSci
(Chair), Professor of Developmental Neuropsychology, University of Oxford

Professor Doreen Cantrell CBE FRS FRSE FMedSci
Professor of Cellular Immunology, Wellcome Trust Principal Research Fellow, and Vice-Principal and Head of College, University of Dundee

Professor Peter Johnson FMedSci
Professor of Medical Oncology, University of Southampton

Professor Shitij Kapur FMedSci
Professor of Schizophrenia, Imaging and Therapeutics, King's College London

Professor Malcolm Macleod
Professor of Neurology and Translational Neuroscience, University of Edinburgh

Professor Caroline Savage FMedSci
VP and Head, Experimental Medicine Unit, GlaxoSmithKline

Dr Jim Smith FRS FMedSci
Director of Research, Francis Crick Institute, and Deputy CEO and Chief of Strategy, Medical Research Council

Professor Simon Tavaré FRS FMedSci
Director, Cancer Research UK Cambridge Institute, University of Cambridge

Professor Melanie Welham
Executive Director of Science, Biotechnology and Biological Sciences Research Council (BBSRC)

Dr John Williams
Head of Science Strategy, Performance and Impact, Wellcome Trust (committee member up to June 2015)

Dr Nicola Perrin
Head of Policy, Wellcome Trust (committee member since July 2015)

Annex II: Symposium programme

Day 1

Welcome and introduction
Professor Dorothy Bishop FRS FBA FMedSci, Professor of Developmental Neuropsychology, University of Oxford, and Chair, Symposium Steering Committee

Session 1: What is the scale of the problem?
Keynote: What is the scale of the problem?
Professor Marcus Munafò, Professor of Biological Psychology, University of Bristol

Session 2: Case studies from biomedical research and beyond
What can we learn from case studies from within biomedical research?

- Neuroscience: Dr Jean-Baptiste Poline, Research scientist, UC Berkeley
- Genomics: Professor Jonathan Flint FMedSci, Wellcome Trust Principal Fellow and Michael Davys Professor of Neuroscience, University of Oxford
- Public health informatics: Professor Iain Buchan, Clinical Professor in Public Health Informatics, University of Manchester

What can basic biomedical research learn from other areas?

- Big data use in particle physics: Professor Tony Weidberg, Professor of Particle Physics, University of Oxford
- Standards, checks and balances from areas such as manufacturing: Dr Matt Cockerill, Founding Managing Director (Europe), Riffyn

Session 3: What is the role of pre-registration of protocols and post-publication peer review?

- Pre-registration of protocols: Professor Chris Chambers, Professor of Cognitive Neuroscience, Cardiff University Brain Research Imaging Centre
- Post-publication peer review: Ms Hilda Bastian, Editor, PubMed Commons
- Panel discussion: led by Professor Chris Chambers, Ms Hilda Bastian, Professor Doreen Cantrell CBE FRS FRSE FMedSci (Professor of Cellular Immunology, University of Dundee) and Professor Sophie Scott FMedSci (Wellcome Senior Fellow, UCL Institute of Cognitive Neuroscience)

Session 4: Improving the research method

- Key issues with the current model: Dr Katherine Button, NIHR School for Primary Care Research (SPCR) Postdoctoral Fellow, University of Bristol
- Perspective from the pharmaceutical industry: Professor Mark Millan, Director of Innovative Pharmacology, Institut de Recherches, Servier
- How can we optimise the reproducibility of research using animals? Professor Malcolm Macleod, Professor of Neurology and Translational Neuroscience, University of Edinburgh
- What role might training courses play in addressing reproducibility? Dr Véronique Kiermer, Director of Author and Reviewer Services, Nature Publishing Group
- What is the role of open science in addressing reproducibility? Dr Courtney Soderberg, Statistical Consultant, Center for Open Science

Session 5: Break-out sessions

Break-out groups considered:

- What can be done to improve reproducibility in:
  – Research using cells and tissues
  – Research using animals
  – Experimental human studies
- How might improvement in the following processes contribute to better reproducibility:
  – Data storage, structuring and sharing
  – More robust statistical approaches and reporting
  – Peer review practices
- What might we learn from:
  – Clinical research

Informal discussion over buffet dinner, with remarks from Professor James L Olds, Assistant Director of the Directorate for Biological Sciences at the US National Science Foundation.

Day 2

Session 6: The role of culture and incentives

- The culture of scientific research: findings of the Nuffield Council on Bioethics' 2014 report: Professor Ottoline Leyser CBE FRS, Director, Sainsbury Laboratory, University of Cambridge
- Panel discussion: led by Professor Ottoline Leyser CBE FRS and Professor Dame Nancy Rothwell DBE FRS FMedSci, President and Vice-Chancellor, University of Manchester

Session 7: The role of funders, journals and publishers

- Perspective from the US National Institutes of Health: Dr Lawrence A Tabak, Principal Deputy Director, National Institutes of Health (via video)
- Panel discussion:
  – Professor Jacqueline Hunter CBE FMedSci, Chief Executive, Biotechnology and Biological Sciences Research Council
  – Dr Deborah Dixon, Vice-President and Global Publishing Director, Health Sciences, John Wiley & Sons
  – Dr Damian Pattinson, Editorial Director, PLOS ONE
  – Professor Richard Horton FMedSci, Editor-in-Chief, The Lancet
  – Professor Peter Weissberg FMedSci, Medical Director, British Heart Foundation

Session 8: Public trust – how do we talk about reproducibility issues?

- Panel discussion:
  – Fiona Fox OBE, Director, Science Media Centre
  – Dr Ian Sample, Science Editor, The Guardian
  – Ed Yong, Freelance Science Writer

Annex III: Symposium participants

Dr Bruce Altevogt, Senior Program Officer, Institute of Medicine

Ms Liz Bal, Journal Development Manager, BioMed Central

Ms Hilda Bastian, Editor, PubMed Commons

Dr Claus Bendtsen, Head of Computational Biology, AstraZeneca

Professor Dorothy Bishop FRS FBA FMedSci, Professor of Developmental Neuropsychology,University of Oxford

Dr Theodora Bloom, Executive Editor, BMJ

Mr Robi Blumenstein, President, CHDI Foundation

Ms Elizabeth Bohm, Senior Policy Adviser, Royal Society

Professor Iain Buchan, Clinical Professor in Public Health Informatics, University of Manchester

Professor Pat Buckley, Dean of Postgraduate Studies, University of South Australia

Dr Katherine Button, NIHR School for Primary Care Research (SPCR) Postdoctoral Fellow, University of Bristol

Professor Doreen Cantrell CBE FRS FRSE FMedSci, Professor of Cellular Immunology, Wellcome Trust Principal Research Fellow Vice-Principal and Head of College**,** University of Dundee

Professor Christina Chai, Dean of Postgraduate Studies, National University of Singapore

Dr Andrew Chalmers, Senior Lecturer, University of Bath

Sir Iain Chalmers FMedSci, Coordinator, James Lind Initiative

Professor Chris Chambers, Professor of Cognitive Neuroscience, Cardiff University Brain Research Imaging Centre

Dr An-Wen Chan Phelan, Scientist, Women's College Research Institute, University of Toronto

Dr Andy Clempson, Research Policy Manager, Association of Medical Research Charities

Professor John Climax, Co-founder, ICON plc

Dr Matt Cockerill, Founding Managing Director (Europe), Riffyn

Professor David Colquhoun FRS, Emeritus Professor of Pharmacology, University College London

Dr Claire Cope, Senior Policy Officer, Academy of Medical Sciences

Dr Michael Crumpton CBE FRS FMedSci, Fellow, Academy of Medical Sciences

Dr Chris DeFeo, Program Officer, Institute of Medicine**Dr Deborah Dixon**
Vice-President and Global Publishing Director, Health Sciences, John Wiley & Sons

Dr Orla Doyle, Postdoctoral Research Associate, King's College London

Dr Stephen Eglen, Senior Lecturer, Centre for Mathematical Sciences, University of Cambridge

Professor Jonathan Flint FMedSci, Wellcome Trust Principal Fellow and Michael Davys Professor of Neuroscience, University of Oxford

Ms Fiona Fox OBE, Director, Science Media Centre

Dr Andrew Furley,Senior Lecturer (BMS) and Faculty Head of Research Training, University of Sheffield

Dr Oliver Grewe, Program Director, Volkswagen Foundation

Ms Hannah Hobson, DPhil Student, Department of Experimental Psychology, University of Oxford

Professor Richard Horton FMedSci, Editor-in-Chief, The Lancet

Professor Martin Humphries FMedSci, Vice-President and Dean, Faculty of Life Sciences, University of Manchester

Professor Jacqueline Hunter CBE FMedSci, Chief Executive, Biotechnology and Biological Sciences Research Council (BBSRC)

Dr Mehwaesh Islam, Policy Officer, Academy of Medical Sciences

Professor Peter Johnson FMedSci, Professor of Medical Oncology, University of Southampton

Dr Cynthia Joyce, Chief Executive Officer, MQ: Transforming Mental Health

Professor Shitij Kapur FMedSci, Professor of Schizophrenia, Imaging and Therapeutics, King's College London

Professor Dermot Kelleher FMedSci, Vice-President (Health); Dean of the Faculty of Medicine, Imperial College London

Dr Véronique Kiermer, Executive Editor and Head of Researcher Services, Nature Publishing Group

Professor Ottoline Leyser CBE FRS, Director, Sainsbury Laboratory, University of Cambridge

Ms Catherine Luckin, Head of International, Academy of Medical Sciences

Professor Malcolm Macleod, Professor of Neurology and Translational Neuroscience, University of Edinburgh

Professor Susan Michie, Professor of Health Psychology, University College London

Professor Mark Millan, Director of Innovative Pharmacology, Institut de Recherches, Servier

Professor Marcus Munafò, Professor of Biological Psychology, University of Bristol

Dr Helen Munn, Executive Director, Academy of Medical Sciences

Dr Chris Munro, Research Assistant, University of Manchester**Professor James L Olds**
Assistant Director Directorate for Biological Sciences, US National Science Foundation

Dr Jonathan O'Muircheartaigh, Sir Henry Wellcome Postdoctoral Fellow, King's College London

Dr Damian Pattinson, Editorial Director, PLOS ONE

Dr Jean-Baptiste Poline, Research Scientist, UC Berkeley

Dr Rachel Quinn, Director, Medical Science Policy, Academy of Medical Sciences

Professor Humphrey Rang FRS FMedSci, President, British Pharmacological Society

Dr Frances Rawle, Head of Corporate Governance and Policy, Medical Research Council

Professor Dame Nancy Rothwell DBE FRS FMedSci, President and Vice-Chancellor, University of Manchester

Dr Ian Sample, Science Editor, The Guardian

Professor Caroline Savage FMedSci, VP and Head, Experimental Medicine Unit, GlaxoSmithKline

Professor Sophie Scott FMedSci, Wellcome Senior Fellow, UCL Institute of Cognitive Neuroscience

Dr Nathalie Percie du Sert, Programme Manager, NC3Rs

Dr Jim Smith FRS FMedSci, Director of Research; Deputy CEO and Chief of Strategy, Francis Crick Institute; Medical Research Council

Dr Courtney Soderberg, Statistical Consultant, Center for Open Science

Dr Thomas Steckler, Associate Director, Pharma R&D Quality and Compliance, Janssen Research & Development

Dr Lawrence A Tabak, Principal Deputy Director, National Institutes of Health

Professor Simon Tavaré FRS FMedSci, Director, Cancer Research UK Cambridge Institute, University of Cambridge

Dr Stuart Taylor, Publishing Director, Royal Society

Dr Elizabeth Wager, Publications Consultant, Sideview

Professor Tony Weidberg, Professor of Particle Physics, University of Oxford

Professor Peter Weissberg FMedSci, Medical Director, British Heart Foundation

Professor Melanie Welham, Executive Director of Science, Biotechnology and Biological Sciences Research Council (BBSRC)

Dr Jelte Wicherts, Associate Professor, Department of Methodology and Statistics, Tilburg University

Dr John Williams, Head of Science Strategy, Performance and Impact, Wellcome Trust

Mr Ed Yong, Freelance Science Writer

The Academy of Medical Sciences

Academy of Medical Sciences
41 Portland Place
London W1B 1QH

@acmedsci

+44 (0)20 3176 2150
info@acmedsci.ac.uk
www.acmedsci.ac.uk

Registered Charity No. 1070618
Registered Company No. 3520281

BBSRC
bioscience for the future

Biotechnology and Biological Sciences Research Council (BBSRC)
Polaris House
North Star Avenue
Swindon SN2 1UH

@BBSRC

+44 (0)1793 413200
webmaster@bbsrc.ac.uk
www.bbsrc.ac.uk

MRC Medical Research Council

Medical Research Council
Polaris House
North Star Avenue
Swindon SN2 1UH

@The_MRC

+44 (0)1793 416200
corporate@headoffice.mrc.ac.uk
www.mrc.ac.uk

wellcome trust

Wellcome Trust
Gibbs Building
215 Euston Road
London NW1 2BE

@wellcometrust

+44 (0)20 7611 8888
contact@wellcome.ac.uk
www.wellcome.ac.uk

Registered Charity No. 210183
Registered Company No. 2711000