Protect us from poor-quality medical research

ESHRE Capri Workshop Group

Published in:
Human Reproduction

DOI:
10.1093/humrep/dey056

Publication date:
2018

Document Version
Peer reviewed version

Protect us from poor quality medical research

Ioannidis JP, Bhattacharya S, Evers JLH, Van der Veen F, Somigliana E, Barrat C, Bontempi G, Crosignani PG on behalf of the ESHRE Capri Workshop Group[1,*]

Short title: Poor quality medical research

*Correspondence should be addressed to: P.G. Crosignani, IRCCS Ca' Granda Foundation Maggiore Policlinico Hospital, Via M. Fanti, 6, 20122 Milano, Italy, e-mail: piergiorgio.crosignani@unimi.it.

This is a pre-copyedited, author-produced version of an article accepted for publication in Human Reproduction following peer review. The version of record Ioannidis, J.P.A, et al. (2018) 'Protect us from poor-quality medical research', Human Reproduction 33:5, pp.770-776, is available online at: https://doi.org/10.1093/humrep/dey056.

[1] The list of the ESHRE Capri Workshop Group contributors is given in the Appendix.

Abstract

Much of published medical research is apparently flawed, cannot be replicated and/or has limited or no utility. This paper presents an overview of the current landscape of biomedical research, identifies problems associated with common study designs and considers potential solutions. Randomized clinical trials, observational studies, systematic reviews and meta-analyses are discussed in terms of their inherent limitations and potential ways of improving their conduct, analysis and reporting. The current emphasis on statistical significance needs to be replaced by sound design, transparency and willingness to share data with a clear commitment towards improving the quality and utility of clinical research.

Key words: medical research / randomized trial / observational study / systematic review / statistical significance

Introduction

Much of published medical research is apparently flawed, cannot be replicated and/or has limited or no utility. Poor medical research has long been called a scandal (Altman 1994). Even though there have been some improvements in many research practices over time, some of the new opportunities in medical research create also more and more complex challenges on how to avoid and deal with poor research. The curricula of most medical schools do not prioritise conduct and interpretation of medical research. This creates a problem for future clinicians who wish to practice evidence based medicine, one which is compounded by the unreliability of much of published clinical research. Doctors need methodological training in order to critically appraise the quality of available evidence instead of taking all published literature on trust (Ioannidis et al, 2017).

The present manuscript is based on an ESHRE Capri Workshop held in September 2017. The workshop and the resulting manuscript tried to define the main current problems underlying poor biomedical research, with emphasis on examples that would be relevant for reproductive medicine in particular; analyze the main causes; and propose changes that would solve some of these problems. This has major implications not only for research, but also for the conduct of medicine and for medical outcomes that depend on research evidence.

We recognize upfront that perfectly reliable/credible and useful research is clearly an unattainable utopia. However, there are many ways in which the existing situation can be improved. In the following sections, we overview challenges in credibility and utility that affect medical research at large and

then focus on specific challenges that are more specific for some key types of influential studies: clinical trials and clinical research; big data and large observational studies; and systematic reviews and meta-analyses.

OVERVIEW OF CHALLENGES IN CREDIBILITY AND UTILITY OF MEDICAL RESEARCH

Most biomedical research studies are of poor quality

Overall, it has been estimated that 85% of research funding is wasted, by inappropriate research questions, by irrelevant endpoints, by faulty study design and flawed execution, by poor reporting and by non-publication (MacLeod et al., 2014, Moher et al., 2016).

Yet, credibility of biomedical research is an essential pre-requisite for evidence based medical decision-making. *Reliability* and *credibility* refer to how likely the results of a study are to be true. *Accuracy* refers to the difference between the observed results and the "truth". *Reproducibility* of methods implies that use of the same methods and tools on the same data and samples will generate the same results. Reproducibility of results denotes the ability to generate comparable results in a new study using methods which are similar to those in the original study. Finally, reproducibility of inferences indicates the ability to reach similar conclusions when different individuals read the same results (Goodman et al., 2016). Apart from these essential attributes, a highly desirable characteristic of preclinical and clinical research is *utility*, i.e. clinical usefulness.

The elusive P-value

Although reliability and utility are critical, most research studies primarily aim to obtain and present significant results. *Significance* itself can be

conceptual, clinical, and statistical - each carrying a very specific meaning. Statistical significance (typically expressed through P-values obtained from null hypothesis testing) is almost ubiquitous in the biomedical literature. An overwhelming majority of published papers claim to have found (statistically and/or conceptually) significant results. An empirical evaluation of all abstracts published in Medline (1990-2015) reporting P-values showed that 96% reported statistically significant results. In-depth analysis of close to 1 million full-text papers in the same time-period identified a similarly high proportion with statistically significant P-values (Chavalarias et al., 2016). Simulation studies have shown that in the absence of a pre-specified protocol and analysis plan, analytical manipulation can produce almost any desired result as a spurious artefact (Patel et al., 2015). Multiple analyses of the same dataset can lead to results which demonstrate variations in both magnitude and direction of effect, occasionally leading to a Janus phenomenon where different analyses of the same data provide conflicting results to the same question (Patel et al., 2015).

While these problems are most prevalent in observational studies, even experimental research is not immune from them. Small and biased randomized trials can produce unreliable results. Large treatment effects produced by trials with modest sample sizes and questionable quality often disappear when the same interventions are tested in large populations by well conducted trials (Pereira et al., 2012). The literature is replete with ways of assessing quality and the risk of bias in clinical trials and other types of studies. Empirical studies have shown that deficiencies in study characteristics that reflect low quality, or high risk of bias, can lead on

average to inflated treatment effects (Savović et al., 2012). However, as the effect of quality shows large between-trial and between-topic heterogeneity, the impact of poor design in a single study cannot be accurately assessed. A low quality study should lead to greater uncertainty, but we cannot just use a correction factor to get a clean, "corrected" result.

So far we have focused too much on P-values. The P value suggests a black-and-white distinction that is elusive (Farland et al., 2016). Effect sizes and confidence intervals are to be preferred in studies in the context of clinically relevant questions, biological plausibility, good study design and conduct. Interpretation of data should be performed in view of prior knowledge, and should preferably lead to the generation of a scientific theory. Our goal should be to perform relevant studies (for which collective equipoise is mandatory) that have adequate power (Braakhekke et al., 2017). Their findings should be placed in the context of broader research agendas and the updated evidence should be used to inform clinical practice.

The research landscape changes

The landscape of clinical research is also being transformed by an increasing volume of studies from outside Europe and the USA. There is some evidence that published results from developing countries without an established tradition of clinical research tend to report larger estimates of benefits for medical interventions (Panagiotou, 2013), even in multi-centre randomised trials (De Denus et al., 2017).

Commercial sponsors may design research in ways to maximize the chances of success of a new discovery, especially where large markets are involved. In these circumstances trials may not necessarily be of lower quality but the

133 questions may be defined and the analyses pre-specified in such ways as to
134 yield favourable conclusions. For example, 96.5% of non-inferiority trials in
135 2011 resulted in conclusions that favoured a new drug or intervention
136 (Flacco et al., 2015).
137 The advent of big data (see below) allows for more ambitious analyses but
138 most available data are of questionable quality and the chance of uncovering
139 genuine effects is low because of high risk of bias. Bias is separate from
140 random error; while random error affects the precision of the signal and big
141 data diminish the random error, bias may create signals that don't exist or
142 may inflate signals or cause signals in the entirely wrong direction. The
143 availability of big data has been perceived as the dawn of a new paradigm,
144 which liberates researchers from some of the more stringent aspects of
145 scientific rigour such as a clear hypothesis, pre-planned analysis, validation
146 and replication - but this is wrong. Hype surrounding new technologies can
147 sway the best academic institutions and innovative entrepreneurs, leading to
148 false expectations about what new tools and massive data can deliver
149 (Lipworth et al., 2017).
150 **Utility**
151 Finally, utility is an attribute that seems to have been overlooked by much of
152 medical research. It comprises the following key elements (Ioannidis, 2016c):
153 having a real problem to fix; appropriate anchoring of the question within
154 the context of prior evidence; substantial prospects of acquiring relevant new
155 information from the new study (irrespective of the direction of its results);
156 pragmatism; patient-centeredness ("what the patient wants"); value for
157 money; feasibility; and transparency (including protection from bias). For a

full discussion of these 8 features of useful research see a previous discussion (Ioannidis, 2016c). Most studies published even in the very best journals meet only a minority of these features (Ioannidis, 2016c).

Conflicts of interest

While recent years have seen major improvements in reporting of conflicts of interest, many continue to go unreported, and there is a growing realisation that non-financial conflicts may have a bigger impact than previously imagined. High-level evidence synthesis (e.g. systematic reviews and meta-analyses) and guidelines may help streamline some of the uncertainty surrounding the available evidence and facilitate medical decision-making. However, these tools also have their weaknesses (Clinical Practice Guidelines We Can Trust, 2011). As an example, a series of red flags has been proposed for guidelines (Lenzer et al., 2013), suggesting caution for those planning to use them in clinical practice. Some of these red flags are difficult to detect, e.g., when a committee for a guideline does not seem to have any major conflicts of interest among its members, but the selection of the members has been pre-emptively biased in favour of a particular recommendation, based on their known views on a subject.

There are many proposed solutions to improve research practices

While the challenges listed above are considerable, there is also a large body of research that has identified examples of good practice and highlighted ways of bypassing problems (Ioannidis, 2014; Munafò et al., 2017). Solutions need to be tailored to the type of study design and the questions being asked. For example, for clinical trials, preregistration of protocols and detailed description of outcomes, adoption of reporting standards, data

sharing, multi-site trials with careful selection of sites, involvement of methodological experts, appropriate regulatory oversight, and containment of conflicts of interest can all be helpful. There are still many unanswered questions about who needs to lead these positive changes in research practices: whether it is the responsibility of investigators, institutions, funders, journals, professional society, the industry, or other stakeholders. There is healthy debate on how best to protect the biomedical literature from preventable bias and error.

SPECIAL CONSIDERATIONS FOR SPECIFIC, INFLUENTIAL TYPES OF MEDICAL RESEARCH

A. CLINICAL TRIALS AND CLINICAL RESEARCH

Clinical relevance of selected outcomes

Outcomes for effectiveness studies should be relevant. Efficacy and mechanistic studies can be used judiciously to inform the best conduct of effectiveness trials with relevant outcomes. Standardization of outcomes is useful for both effectiveness and efficacy studies. Many specialties are reaching consensus on what are the core outcomes that are worth prioritizing. For example, the CROWN initiative aims at developing core outcome measures in woman's health (Core Outcomes in Women's Health (CROWN) Initiative, 2014).

Multiplicity issues in clinical research

If researchers perform many analyses, some will turn out to be statistically significant purely by chance, yielding false-positive results. Multiple testing might represent a particular problem in infertility treatment: due to the

multistage nature of many treatments, many outcomes may be reported in a study.

Registration

Registration of clinical research has become more common, especially for clinical trials, but still many trials are not pre-registered. Ideally, before carrying out a clinical trial, its full study design, including all primary and secondary outcomes (e.g. number of oocytes obtained per woman randomised, or cumulative live birth rate after three completed cycles of ART treatment), should be pre-specified and the trial registered in a WHO approved clinical trial registry, together with the latest approved version of the protocol (COMPare, 2017). In the absence of registration (or with incomplete details about registration), it is not possible to tell what goals, objectives, design aspects, or analyses were pre-specified versus post-hoc explorations.

Reporting of pre-specified outcomes

Once the trial is finished, the trial report should present *all* pre-specified outcomes. When reported outcomes differ from those pre-specified, this must be declared in the report, along with an appropriate explanation (COMPare, 2017). Changing endpoints of a study after the analysis of the data has occurred may denote scientific misconduct, especially if the change is instigated by the lack of significance in the primary outcome, but not in some arbitrary subordinate outcomes (COMPare, 2017). This is popularly known as P-hacking, data dredging, cherry picking, snooping, significance chasing, or the Texas sharpshooter fallacy (Evers, 2017).

Reporting guidance exists for randomized trials (CONOSRT), as well as for other types of clinical research, e.g. STARD (for diagnostic test studies), PRISMA (for meta-analyses) and IMPRINT (the latter specifically for fertility trials). These guidance documents aim to improve the quality and completeness of clinical research reports (Glasziou et al., 2014). It is very disturbing that comparisons of protocols with publications in major medical journals revealed that most studies had at least one primary endpoint changed, introduced, or omitted (Chan et al., 2004; Chan et al., 2014; Glasziou et al., 2014).

Power considerations in clinical trials

Lack of sufficient power is a major problem across various types of studies, including randomized trials in diverse disciplines and reproductive medicine is no exception. Differences in live birth rates of 3-5% may still be clinically relevant to detect, but hardly any trials in the field have sufficient sample size for this. Therefore, one should be careful in interpreting confidence intervals. Some trials where it is concluded that "the intervention had no effect" may in fact offer no conclusive information about whether the treatment is effective or not. Moreover, small trials are more likely to generate exaggerated effects and even false-positive spurious effects.

BIG DATA AND LARGE OBSERVATIONAL DATASETS

Database linkage: maximum temptation meets maximum opportunity

Sources of health care data include governments, healthcare providers, insurers, registries of specific conditions, treatments and medical devices, as well as registers of births and deaths. Increasingly, data are available in electronic formats and can be linked with other health, social, geographical

and education data to create massive datasets incorporating complex longitudinal records with large-scale population coverage and long-term follow-up. Medical records can provide demographic information, lifestyle choices, clinical findings, laboratory and imaging results, treatment details and outcomes. Ability to link sociodemographic and clinical details with genomic, proteomic, and metabolomic data could potentially allow physicians to deliver precision medicine (Peek et al., 2014) for individual patients. Routinely collected health data can also allow a real-world evaluation of treatment outcomes.

While opportunities seem to abound in theory, there are many serious limitations to big data and large observational datasets. Here we discuss some of the key ones.

Problems with information

The event-based nature of routinely collected health data is a potential limitation, as important problems or treatments not resulting in hospital contacts may be missing. Inaccuracies in the data can occur due to mistakes in data entry and lack of appropriate checks. Routine data are also likely to contain a minimum set of variables and many key confounders such as body weight, height, smoking status, alcohol intake and socio-economic status may be missing. Many historical datasets lack a planned schema, which can create problems during analysis (Jorm, 2015) although others have detailed metadata (Ayorinde et al., 2016). Finally, data is often missing in a non-random fashion thus introducing the possibility of bias. While some ways of dealing with missing data (Jagsi et al., 2014) are better than others, missingness may be difficult to address with high confidence.

Ethical challenges

Major concerns arise around the use of routinely collected data to answer questions for which the data were not originally collected. These concerns involve lack of informed consent, possible identification of subjects during linkage procedures (even after anonymisation), the dilemma of dealing with detected individual risks in an anonymised (rather than anonymous) population who could potentially be identified and informed and individuals in very small categories of groups with unusual conditions. Instead of widely open use of big data, it may be required to employ data safe havens where access is limited to trained staff and safe release of data after rigorous checks to minimise risks of identification (Lea et al., 2016).

Difficulties in linkage

Linkage presents a common technical challenge which could introduce significant error if done incorrectly. The most accurate is the deterministic method using a unique identifier, such as the personal identity number in the Nordic countries and the community health index (CHI) number in Scotland (Ayorinde et al., 2016). Where this is not feasible, probabilistic methods based on characteristics such as name, date of birth, geographical location have been used but this approach can result in errors.

Dealing with confounding in large observational datasets

All large observational datasets are prone to confounding that can cause spurious associations. For example, in the context of fertility data, age is a common confounder which influences the choice of treatment as well as its outcome. Often choice of therapy is usually based on preference, predicted response, or other non-random selection features which can impact on

306 outcomes (Jagsi et al., 2014). For example, as women with more severe
307 endometriosis may be more likely to receive surgery than women with less
308 severe disease, the outcome of surgical treatment may appear to be worse than
309 medical alternatives. Methods such as propensity score matching, propensity
310 score stratification, inverse probability of treatment weighting and
311 instrumental variable analysis, which uses counterfactuals to try to
312 approximate a randomized design situation, try to address this problem.
313 Although some reviews (Anglemyer et al., 2014) suggest that there is limited
314 evidence for significant differences in health care outcomes between
315 observational studies and randomised trials, other studies show that further
316 refinements in analysis need to be made in order to achieve the same degree
317 of accuracy (McGale et al., 2016). Empirical evaluations suggest that routinely
318 collected data are not yet used to their maximal potential utility (Hemkens et
319 al., 2016a), and they tend to generate inflated treatment effects even when
320 sophisticated propensity score methods are used (Hemkens et al., 2016b).

321 **Overpowered big data**

322 Studies based on large datasets can have sample sizes that are so large that
323 they detect very small and clinically unimportant effect sizes. Such studies
324 should be interpreted appropriately according to their clinical significance.
325 Highly statistically significant results may still represent pure chance findings
326 (Peek et al., 2014). With small effects, bias or confounding cannot be excluded.
327 Interpretation must therefore be cautious, despite whatever statistical
328 significance.

329 **Personalised medicine prospects**

Advances in computational infrastructures for dealing with big datasets and the related explosion in data science methodology, lead to speculations that the future of life sciences is likely to be dominated by systems which can ingest and sift through large volumes of -omics data to generate reliable information for individualised decision making (e.g. personalised [precision] medicine). However, these expectations have yet to be fully realized. A naïve expectation of accurate predictions from inherently flawed and incomplete data could turn out to be no more than blind faith in fool's gold (Khoury et al., 2014; Lipworth et al., 2017). Personalised medicine is an interesting concept but it meets with many conceptual (Senn, 2016) and practical difficulties in making it work.

SYSTEMATIC REVIEWS AND META-ANALYSES

A prolific industry of meta-analyses

Most hierarchies of evidence place well-conducted systematic reviews (SRs) and meta-analyses (MAs) at the top of the evidence pyramid and these publications have grown in volume as well as influence. As of mid-2017, nearly 100,000 published meta-analysis articles were indexed in PubMed with over 1000 new ones indexed every month (Ioannidis, 2016a). There are also approximately 250,000 published SRs in PubMed, with another 2500 new ones indexed every month. In many fields there are more SRs than primary studies (Prior et al., 2017) and, in many situations, SRs have replaced experience and clinical acumen in terms of driving clinical decision making. This has not gone unnoticed by individuals and groups with vested interests (financial or non-financial) who have used them as tools to

354 influence practice in favour of their preferred drugs and interventions
355 (Ioannidis, 2016b).

356 **Most SRs and MAs are not very useful and many are not useful at all**

357 A common conclusion of many systematic reviews, particularly those that
358 address questions on effective treatment is that primary evidence is lacking,
359 suboptimal or unreliable. This statement alone has some utility, because it
360 can still help calibrate the level of uncertainty in decision making and may
361 suggest avenues of new research. However, very often the primary data
362 feeding into SRs and MAs are so unreliable that these may have a more
363 important role in detecting bias rather than uncovering the truth. SRs and
364 MAs may also help identify gaps in the use of patient-relevant outcomes
365 where multiple studies exist but outcomes that matter are not addressed.

366 **The global profile of SRs and MAs**

367 The profile of SRs and MAs has changed over the last decade, with
368 increasing numbers of Mas now being generated in China. Most of these MAs
369 are unreliable, or misleading (especially the bulk-produced meta-analyses of
370 candidate gene associations). Moreover, there is a new large portfolio of MAs
371 conducted by contractor companies that are commissioned and paid by the
372 industry (Schuit and Ioannidis, 2016). Only a small proportion of these MAs
373 are published and publication bias may be related to the results of the MAs
374 and the interests of the sponsor. An online search suggested that over 100
375 service-offering companies perform SRs and MAs (Schuit and Ioannidis,
376 2016).

377 **Redundancy in SRs and MAs**

A recent evaluation suggested that only about 3% of current MAs are both methodologically sound and clinically useful (Ioannidis, 2016a). There is a lot of redundancy and large numbers of SRs and MAs continue to be conducted on some topics without clear evidence for the additional value of the newer publications, e.g. in the area of urinary derived versus recombinant FSH treatment (Van Wely et al, 2011).

More sophisticated MA designs

Even for more sophisticated forms of evidence synthesis such as network MAs, an empirical evaluation identified 28 publications on the same topic, each including part of the available evidence with inconsistent conclusions (Naudet et al., in press). Registration of MAs at the protocol stage, e.g. in registries like PROSPERO, may be helpful, but it is unclear whether this alone can create a more efficient, transparent and, ultimately, a more accurate compilation of all the available facts (Tricco et al., 2016; Moher et al., 2014).

An increasing number of MAs have been able to use individual participant data. These require more resources to perform compared with MAs of aggregated data, but they have a number of advantages in terms of being able to clean the data, standardize definitions, outcomes and co-variates across studies, and can explore subgroup differences in a more reliable fashion (Simmonds et al., 2005). Apart from higher costs, their disadvantage includes incomplete retrieval of data, potentially leading to bias, if some trials with specific directions of effect are missed. As results from randomized trials and other types of studies become more readily available, it may be easier to perform comprehensive MAs using individual-level data in

the future. Using advanced meta-analysis methods requires statistical and methodological competence that if often currently lacking in reviewers undertaking such analyses using software that they don't fully understand how they function.

Systematic reviews and meta-analyses in the future

Despite limitations, SRs and MAs will continue to be indispensable for summarizing the evidence and understanding its biases, strengths, and weaknesses. Moving forward, hopefully there will be more MAs in the future which use optimal methods for systematic searches, retrieving, analysing and reporting data. It is also likely that there will be more MAs that will use either networks or individual-level data or both, allowing for more informative analyses and data syntheses. Eventually, MAs may be planned as prospective exercises, i.e. designed contemporaneously with primary evaluative studies with a clear a priori plan of combining results from primary studies on completion (Ioannidis, 2017). This approach may help to minimize some of the biases that exist in retrospective data synthesis.

THE FUTURE

Given the challenges described above, it is probably not surprising that most medical research shows poor reproducibility of methods and results. Some of the problems are increasingly recognized by the scientific community. A 2016 Nature survey showed that more than two-thirds of scientists believed that there is a reproducibility problem (Baker, 2017). Replicability is a benchmark of scientific quality; authors should always try to replicate their own results and provide sufficiently detailed instructions for others to do so. While research fraud is uncommon, the temptation to cut corners prompts

many authors to indulge in poor scientific practices (Tanksalva, 2017). The "publish or perish" attitude favours hasty, low quality, incomplete research with the aim of maximising the number of papers from a single research project (salami slicing). There is also a temptation to sensationalize results. Incentive structures for rewarding research, e.g. publication, funding, promotion, and tenure, need to pay more attention to quality and reproducibility of the work produced.

Investigators can learn from studies which cannot be replicated. Adoption of reporting standards will help, as will multi-site trials, involvement of methodological experts, appropriate regulatory oversight, and transparency about conflicts of interest. As gatekeepers, journals can offer high quality peer review (which should include proper statistical/methodological review, as appropriate). Prospective trial registration is not enough, full protocols should also be published, and data should be shared.

Finally, many changes will require emphasis on education, including training at medical schools (physicians should be sensitized to strengths and weaknesses of the evidence that affects their practices) and training of researchers in methodological competence.

ACKNOWLEDGEMENTS

The secretarial assistance of Mrs Simonetta Vassallo is gratefully acknowledged.

AUTHORS' ROLES

Lecturers, chairmen and discussants contributed to the preparation of the final manuscript.

FUNDING

The meeting was organized by the European Society of Human Reproduction and Embryology with an unrestricted educational grant from Institut Biochimique S.A. (Switzerland).

CONFLICT OF INTEREST

None declared.

APPENDIX

Members of the ESHRE Capri Workshop Group: D.T. Baird (Centre for Reproductive Biology, University of Edinburgh, UK), C. Barratt (Division of Molecular & Clinical Medicine, School of Medicine, University of Dundee, Ninewells Hospital and Medical School, Dundee, UK), S. Bhattacharya (Professor of Reproductive Medicine, Head of Division of Applied Health Sciences and Director Institute of Applied Health Sciences, School of Medicine and Dentistry, University of Aberdeen, Aberdeen Maternity Hospital, Foresterhill, Aberdeen, UK), G. Bontempi (co-Head of the Machine

Learning Group, Département d'Informatique, Université Libre de Bruxelles, Bruxelles, Belgium), P.G. Crosignani (IRCCS Ca' Granda Foundation, Maggiore Policlinico Hospital, Milano, Italy), P. Devroey (AZ-VUB, Centre for Reproductive Medicine, Brussels, Belgium), K. Diedrich (Klin. Frauenheilkunde und Geburtshilfe, Univ. zu Lubeck, Lubeck, Germany), J.L.H. Evers (Dept. Obstet. Gynecol., Maastricht University Medical Centre, Maastricht, The Netherlands), R. G. Farquharson (Liverpool Women's Hospital, Department of OB/GYN, Liverpool, UK),) L.R. Fraser (Centre for Reproduction, Endocrinology & Diabetes, School of Biomedical Sciences, New Hunt's House, Kings College London, Guy's Campus, London, UK), J.P.M. Geraedts (AZ Maastricht, Klinische Genetica, Maastricht, The Netherlands), L. Gianaroli (S.I.S.M.E.R., Bologna, Italy), J.P. Ioannidis (Departments of Medicine, of Health Research and Policy, of Biomedical Data Science, and of Statistics, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, USA), C. La Vecchia (Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Milan, Italy), K. Lundin (Reproductive Medicine, Sahlgrenska University Hospital, Gothenburg, Sweden), C. Magli (S.I.S.M.E.R., Bologna, Italy), E. Negri (Department of Biomedical and Clinical Sciences, Università degli Studi di Milano, Milano, Italy), E. Somigliana (Clinica Ostetrica e Ginecologica, IRCCS Ca' Granda Foundation, Maggiore Policlinico Hospital, Milano, Italy), A. Sunde (University Hospital, Dept. Obstet. Gynecol., Trondheim, Norway), J.S. Tapanainen (University of Helsinki, Department of Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland), B.C. Tarlatzis (Infertility & IVF Center, Geniki

Kliniki, Thessaloniki, Greece), F. van der Veen (Academic Medical Centre, University of Amsterdam, Reproduction Medicine, Amsterdam, The Netherlands), A. Van Steirteghem (Centre for Reproductive Medicine, Universitair Ziekenhuis Vrije Universiteit Brussel, Belgium), A. Veiga (Reproductive Medicine Service, Dexeus Women's Health, Barcelona, Spain).

References

Altman DG. The scandal of poor medical research. BMJ 1994;**308**:283.

Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev. 2014 Apr 29;(4):MR000034.

Ayorinde AA, Wilde K, Lemon J, Campbell D, Bhattacharya S. Data Resource Profile: The Aberdeen Maternity and Neonatal Databank (AMND). Int J Epidemiol 2016;**45**:389-394.

Baker M, at www.nature.com, accessed August 21st, 2017.

Braakhekke M, Mol F, Mastenbroek S, Mol BW, van der Veen F. Equipoise and the RCT. Hum Reprod. 2017;**32**:257-260.

Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004;**291**:2457-2465.

Chan AW, Krleza-Jerić K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. CMAJ 2014;**171**:735-740.

Chavalarias D, Wallach J, Li A, Ioannidis JPA. Evolution of reporting of p-values in the biomedical literature, 1990-2015. JAMA 2016;**315**:1141-1148.

Clinical Practice Guidelines We Can Trust. Institute of Medicine (US) Committee on Standards for Developing Trustworthy Clinical Practice Guidelines; Graham R, Mancher M, Miller Wolman D, Greenfield S, Steinberg E, editors. Washington (DC): National Academies Press (US); 2011.

533 COMPare Trials Project. Goldacre B, Drysdale H, Powell-Smith A, Dale A,
534 Milosevic I, Slade E, Hartley P, Marston C, Mahtani K, Heneghan C.
535 www.COMPare-trials.org, 2017, last accessed 23 May 2017
536 Core Outcomes in Women's Health (CROWN) Initiative. The CROWN
537 Initiative: Journal editors invite researchers to develop core outcomes in
538 women's health. Hum Reprod 2014;**29**:1349-1350.
539 De Denus S, O'Meara E, Desai AS, Claggett B, Lewis EF, Leclair G, Jutras M,
540 Lavoie J, Solomon SD, Pitt B, Pfeffer MA, Rouleau JL. Spironolactone
541 Metabolites in TOPCAT - New Insights into Regional Variation. N Engl J
542 Med 2017;**376**:1690-1692.
543 Evers JL. The Texas scharpshooter fallacy. Hum Reprod 2017;**32**:1363.
544 Farland LV, Correia KF, Wise LA, Williams PL, Ginsburg ES, Missmer SA. P-
545 values and reproductive health: what can clinical researchers learn from
546 the American Statistical Association? Hum Reprod. 2016;**31**:2406-2410.
547 Flacco ME, Manzoli L, Boccia S, Capasso L, Aleksovska K, Rosso A, Scaioli
548 G, De Vito C, Siliquini R, Villari P, Ioannidis JP. Head-to-head randomized
549 trials are mostly industry-sponsored and almost always favour the
550 industry sponsor. J Clin Epidemiol 2015;**68**:811-820.
551 Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, Michie S,
552 Moher D, Wager E. Reducing waste from incomplete or unusable reports
553 of biomedical research. Lancet 2014;**383**:267-276.
554 Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility
555 mean? Sci Transl Med 2016;**8**:341ps12.

556 Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JP. Current use of
557 routinely collected health data to complement randomized controlled
558 trials: a meta-epidemiological survey. CMAJ Open. 2016a;**4**(2):E132-140.
559

560 Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JP. Agreement of
561 treatment effects for mortality from routinely collected data and
562 subsequent randomized trials: meta-epidemiological survey. BMJ
563 2016b;**352**:i493.
564 Ioannidis JP. How to make more published research true. PLoS Med
565 2014;**11**:e1001747.
566 Ioannidis JP. The mass production of redundant, misleading, and conflicted
567 systematic reviews and meta-analyses. Milbank Q. 2016a;**94**:485-514.
568 Ioannidis JP. Evidence-based medicine has been hijacked: a report to David
569 Sackett. J Clin Epidemiol 2016b;**73**:82-86.
570 Ioannidis JP. Why most clinical research is not useful. PLoS Med. 2016c,
571 **13**:e1002049.
572 Ioannidis JP. Meta-analyses can be credible and useful: a new standard.
573 JAMA Psychiatry 2017;**74**:311-312.
574 Ioannidis JPA, Stuart ME, Brownlee S, Strite SA. How to survive the medical
575 misinformation mess. Eur J Clin Invest 2017;**47**:795-802.
576 Jagsi R, Bekelman JE, Chen A, Chen RC, Hoffman K, Shih YC, Smith BD,
577 Yu JB. Considerations for observational research using large data sets in
578 radiation oncology. Int J Radiat Oncol Biol Phys 2014;**90**:11-24.
579 Jorm L. Routinely collected data as a strategic resource for research: priorities
580 for methods and workforce. Public Health Res Pract 2015;**25**:e2541540.
581 Khoury MJ, Ioannidis JP. Medicine. Big data meets public health. Science
582 2014;**346**:1054-1055.
583 Lea NC, Nicholls J, Dobbs C, Sethi N, Cunningham J, Ainsworth J, Heaven M,
584 Peacock T, Peacock A, Jones K, Laurie G, Kalra D. Data Safe Havens and

585 Trust: Toward a Common Understanding of Trusted Research Platforms for
586 Governing Secure and Ethical Health Research. JMIR Med Inform
587 2016;**4**:e22.
588 Lenzer J, Hoffman JR, Furberg CD, Ioannidis JP; Guideline Panel Review
589 working group. Ensuring the integrity of clinical practice guidelines: a tool
590 for protecting patients. BMJ 2013;**347**:f5535.
591 Lipworth W, Mason PH, Kerridge I, Ioannidis JP. Ethics and epistemology in
592 big data research. J Bioeth Inq 2017 Mar 20. doi: 10.1007/s11673-017-
593 9771-3. [Epub ahead of print]
594 MacLeod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JP, Al-
595 Shahi Salman R, Chan AW, Glasziou P. Biomedical research: increasing
596 value, reducing waste. Lancet 2014;**383**:101-104.
597 McGale P, Cutter D, Darby SC, Henson KE, Jagsi R, Taylor CW. Can
598 Observational Data Replace Randomized Trials? J Clin Oncol
599 2016;**34**:3355-3357.
600 Moher D, Booth A, Stewart L. How to reduce unnecessary duplication: use
601 PROSPERO. BJOG 2014;**121**:784-786.
602 Moher D, Glasziou P, Chalmers I, Nasser M, Bossuyt PM, Korevaar DA,
603 Graham ID, Ravaud P, Boutron I. Increasing value and reducing waste in
604 biomedical research: who's listening? Lancet 2016;**387**:1573-1586.
605 Munafò MR, Bishop DV, Button KS, Chambers C, Nosek B, Percie du Sert N,
606 Simonsohn U, Wagenmakers E-J, Ware JJ, Ioannidis JPA. A manifesto for
607 reprodubile science. Nature Human Behaviour 2017;**1**:0021.

608 Naudet F, Schuit E, Ioannidis JPA. Overlapping network meta-analyses on
609 the same topic: survey of published studies. Int J Epidemiol 2017 Aug 3.
610 doi: 10.1093/ije/dyx138. [Epub ahead of print].
611 Panagiotou OA, Contopoulos-Ioannidis DG, Ioannidis JP. Comparative effect
612 sizes in randomised trials from less developed and more developed
613 countries: meta-epidemiological assessment. BMJ 2013;**346**:f707.
614 Patel C, Burford B, Ioannidis JP. Assessment of vibration of effects due to
615 model specification can demonstrate the instability of observational
616 associations. J Clin Epidemiol 2015;**68**:1046-1058.
617 Peek N, Holmes JH, Sun J. Technical challenges for big data in biomedicine
618 and health: data sources, infrastructure, and analytics. Yearb Med Inform
619 2014;**9**:42-47.
620 Pereira TV, Horwitz RI, Ioannidis JP. Empirical evaluation of very large
621 treatment effects of medical interventions. JAMA 2012;**308**:1676-1684.
622 Prior M, Hibberd R, Asemota N, Thornton JG. Inadvertent P-hacking among
623 trials and systematic reviews of the effect of progestogens in pregnancy? A
624 systematic review and meta-analysis. BJOG 2017;**124**:1008-1015.
625 Savović J, Jones HE, Altman DG, Harris RS, Jűni P, Pildal J, Als-Nielsen B,
626 Balk EM, Gluud C, Gluud LL, Ioannidis JP, Schulz KF, Beynon R, Welton
627 NJ, Wood L, Moher D, Deeks JJ, Sterne JAC. Influence of reported study
628 design characteristics on intervention effect estimates from randomized
629 controlled trials: combined analysis of meta-epidemiologic studies. Ann
630 Intern Med. 2012;**157**:429-438.

631 Schuit E, Ioannidis JPA. Network meta-analyses performed by contracting
632 companies and commissioned by industry. Systematic Reviews
633 2016;**5**:198.
634 Senn S. Mastering variation: variance components and personalised
635 medicine. Stat Med 2016;**35:**966-77.
636 Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson
637 SG. Meta-analysis of individual patient data from randomized trials: a
638 review of methods used in practice. Clin Trials 2005;**2**:209-217.
639 Tanksalva S at www.clarivate.com/blog, accessed August 21st, 2017.
640 Tricco AC, Cogo E, Page MJ, Polisena J, Booth A, Dwan K, MacDonald H,
641 Clifford TJ, Stewart LA, Straus SE, Moher D. A third of systematic reviews
642 changed or did not specify the primary outcome: a PROSPERO register
643 study. J Clin Epidemiol 2016;**79**:46-54.
644 Van Wely M, Kwan I, Burt AL, Thomas J, Vail A, Van der Veen F, Al-Inany
645 HG. Recombinant versus urinary gonadotrophin for ovarian stimulation in
646 assisted reproductive technology cycles. Cochrane Database Syst Rev.
647 2011 Feb 16;(2):CD005354. doi: 10.1002/14651858.CD005354.pub2.
648 Review. PubMed PMID: 21328276.
649