DOI:
10.1017/S003329171700294

Document Version
Peer reviewed version

Link to publication record in King's Research Portal

Citation for published version (APA):
Stahl, D. R., & Pickles, A. (2018). Fact or Fiction: Reducing the Proportion and Impact of False Positives. *Psychological medicine*, *48*(7), 1084-1091. https://doi.org/10.1017/S003329171700294

Main Text Word Count: 4598

Title: Fact or Fiction: Reducing the Proportion and Impact of False Positives

Daniel Stahl PhD and Andrew Pickles PhD

Department of Biostatistics and Health Informatics,

Institute of Psychiatry, Psychology and Neuroscience,

King's College London,

London SE5 8AF

Acknowledgement

Our work was in part supported by the National Institute for Health Research (NIHR) Biomedical Research Centre for Mental Health at the South London and Maudsley NHS Foundation Trust and Institute of Psychiatry, Psychology & Neuroscience, Kings College London. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health

Abstract

False positive findings in science are inevitable, but are they particularly common in psychology and psychiatry? The evidence that we review suggests that while not restricted to our field the problem is acute. We describe the concept of researcher 'degrees-of-freedom' to explain how many false-positive findings arise, and how the various strategies of registration, pre-specification, and reporting standards that are being adopted both reduce and make these visible. We review possible benefits and harms of proposed statistical solutions, from tougher requirements for significance, to Bayesian and machine learning approaches to analysis. Finally we consider the organisation and methods for replication and systematic review in psychology and psychiatry.

Introduction

When regarded from the positivist perspective of the scientific method, psychology and neuroscience should be an accumulative, iterative, self-correcting endeavour. While mistakes are inevitable they will typically be identified by the onward march of the research process. However, like all human activity, psychology and neuroscience research is also a social process. It relies on many inter-personal and organisational processes and displays all the social phenomena we observe elsewhere, such as fashions, concern with status and reputation, group-identification, collective judgements, social norms, competitive and defensive actions (Edmonds *et al.*, 2011). We are active participants in this complexity though we may sporadically rail against its norms and culture.

While the importance attached to scientific rigour is undoubtedly more than lip-service, the needs of careers, funders, scientific institutions and commercial stakeholders have priorities in addition to or other than the generation of reliable and unbiased findings. That published reports in psychology were over-represented by positive tests of hypotheses has long been known. Already in 1959 and again with colleagues in 1995 Sterling had reported that almost all statistical hypotheses in contemporary psychology journals were significant (Sterling, 1959, Sterling *et al.*, 1995) and Fanelli (2010) could not show an improvement in this century, finding more than 90% of published hypotheses were described as positive in spite of typical studies being insufficiently powered for such “success” (Bakker *et al.*, 2012).

Pashler and Wagenmaker (2012) describe this implausible reality as a “crisis of confidence” and examples of published reports of studies of premonition having achieved statistical significance and survived peer review seemingly particularly damning (Bem, 2011). The Open Science Foundations Reproducibility Project (https://osf.io/ezcuj/) attempted to provide a more objective assessment of the reliability of findings by attempting to replicate studies. Repeating almost 100 studies published in three high-ranking psychology journals, they found that less than half of the reported results could be replicated, the figure being somewhat higher for formal experimental studies than observational (Aarts et al., 2015). This and many other studies confirmed Ioannidis’ claim in his seminal paper (Ioannidis, 2005) that more than half of claimed positive discoveries in medical and psychological studies are false positives. Figure 1 explains Ioannidis' reasoning behind his assumption, now strongly supported by empirical evidence, and why underpowered studies are a major problem for the replication crisis.

However, false positive results from underpowered studies ought to be detected and self-corrected by the "vigorous and uncompromising scepticism " of the scientific method (Sagan, 1995). However, self-correction isn't always swift and it may need time and effort to uproot belief in false positive reports. Pickles (2009) has described a common natural history for many "findings". It begins with a small study that finds a significant association and an editor persuaded of its interest. Many small studies follow with the great majority failing to replicate the finding. Everyone knows that small studies have low power so the lack of replication comes as no surprise. Few are published but preference is given to those that do find a significant association, apparently confirming the interesting finding. Often studies are of different design and non-significant results are explained by small conceptual differences in the design. Eventually, a large study fails to replicate what is now an apparently well-established finding, and because of the study's size and the fact this failure is now seen as overturning a received wisdom, publication in a prominent journal follows. Such frequent diversions into cul-de-sacs, with research following these false trails, slows the real progress that we could be making. With research funding in psychology and mental health being so disproportionately low (Luengo-Fernandez *et al.*, 2015), researchers in these areas, in particular, cannot afford such diversions. Moreover, corrupting incentives persist or are increasing. We struggle with replication of simple main effects, so the high-level of enthusiasm for stratified or personalized medicine in which effects are shown in some isolated sub-group serves to exacerbate the problem, and the expectation that researchers show "impact" encourages clinical dissemination, commercialisation and incorporation into health guidelines prior to proper corroboration.

As the above suggests, we have moved from expressions of concern by a few lone voices, through a period of more widespread unease and some embarrassment, to a series of more objective assessments and ad hoc proposals and initiatives. However, the problem is many facetted, complicated in technical, sociological and economic terms, and insidious. To change the way we do science requires a form of political movement, so perhaps it is unsurprising therefore that we now have manifestos, for example the Manifesto for

Reproducible Science (Munafò *et al.*, 2017) that promotes a set of these proposals. Interestingly, while arguing for their broad adoption by researchers, institutions, funders and journals, the authors make explicit that there is also the need for the iterative evaluation and improvement of such proposals. As in any other area of policy, what may sound sensible could nonetheless have unintended and possibly negative consequences that will need evaluation.

In this review, while we highlight a fuller understanding of the statistical and methodological issues and suggestions as to how to improve the unsatisfactory situation, we also underscore how statistics alone is not sufficient. Both as investigators and reviewers we need to reassess what we should be valuing, and how that influences what we do. Our review falls into two major parts. We begin by a description of researcher "degrees-of-freedom"(Simmons *et al.*, 2011), the recognised and unrecognised flexibility of the researcher to "chase significance". From this a whole series of practices - both methodological and organizational - have been proposed to make this flexibility visible and then, where required, to curtail it. We discuss methodological solutions to the problem including machine learning methods and explain how many current applications can still mislead.

Fraud, False Positives and Researcher degrees-of-freedom

While there are cases of deliberate fraud, where data have been made-up or intentionally incorrect analysis presented, these remain fortunately relatively rare. Nonetheless a recent meta-analyses shows worryingly that about 2% of scientists admitted to having fabricated, falsified or modified data (Fanelli, 2009). But these represent the extreme end of a much larger problem where researchers may be turning a blind eye to what they know is not fully rigorous science or may be entirely unaware that what they have done results in misrepresentation.

We tend to think about the choices available to a researcher as being concerned with the study design and the instruments and schedule to be used in the measurement protocol. However, once the data are collected, the researcher still has a vast number of choices to be made before a report of findings is finalized. These choices imply alternatives and the extent to which the researcher can pick one or another, dependent upon how appealing are the results of such choices, and are referred to as degrees of freedom. Perhaps the most familiar are those associated with having multiple risk variables and multiple outcomes. A single p-value is interpretable against the standard probability threshold of .05 if it is the only test being undertaken. With multiple risks and multiple outcomes we face the problem of multiple comparisons, for which Bonferroni proposed a well-known correction, tightening the required threshold the more comparisons the researcher makes (Bender and Lange, 2001).

However, before users of multiple-comparison adjustment methods feel too self-righteous, we should consider some of the other researcher degrees of freedom (Simmons *et al.*, 2011). There may be multiple alternative definitions or constructions of even single risk factors or outcome disorders. Over how long/what follow-up period should these be measured? What transformations can be applied and what does the researcher do with potential outliers? What participants should be included or excluded? What statistical test should be used? Will additional data be collected if a finding is not quite significant such that there is an implicit series of interim analyses? Will univariate pre-screening be performed to reduce the number of variables? (Gelman and Loken, 2014) describe how these researcher degrees of freedom substantially increase the multiple testing problem but may not feel like fishing for significant results to the researcher. Exploiting these degrees-of-freedom can dramatically increase the actual false positive rate and rarely yields replicable results (Harrell, 2015, Simmons *et al.*, 2011). Many researchers are not aware of the consequences of such subtle types of fishing and they need to be more aware of the negative impact of ignoring researcher's degrees of freedom. Researchers should list their use of degrees of freedoms to determine the potential for bias e.g. by using a checklist (Wicherts *et al.*, 2016) and if possible correct for them in their analyses (Harrell 2015). However, given the current state of affairs, a paper from 72 researchers from around the world (Benjamin *et al.*, 2017) have proposed the blunt approach of reducing the standard alpha level for claiming significance from the gold standard of 0.05 to 0.005, with consequent implications for larger sample size to maintain power.

The null-hypothesis testing framework of Fisher, Neyman and Pearson works well if correctly applied, especially in double blinded placebo controlled randomized trial with a single outcome. However, it becomes more problematic with increasing numbers of often unplanned tests and choices, giving p values that are too small, confidence intervals too narrow and the final model from model selection too complex (Freedman, 1983, Harrell, 2015). Some have proposed Bayesian hypothesis testing instead, using the Bayes' factor - the ratio of the evidence for the alternative compared to the null (Ly *et al.*, 2016, Rouder *et al.*, 2009). Johnson (2013) showed that the significance level of 0.05 generally correspondents to a Bayes Factor of 5 or less, a range generally regarded as providing only modest evidence. Unsurprisingly therefore, given the use of a more demanding criterion, a Bayesian re-analysis of Bem's extrasensory perception (ESP) study by Wagenmaker and colleagues found little evidence for ESP (Wagenmakers *et al.*, 2011) and Wetzels and colleagues (2011) reanalysis of 259 psychology articles found only "anecdotal evidence" for 70% of the findings previously claimed as significant. Thus these arguments for a Bayesian approach as a solution appear to share much with the proposal for a stricter 0.005 criterion for nominal significance and may not escape the limitations of dichotomous thinking (Cumming, 2014).

Fixing upon an even lower alpha level may create new problems and could prevent innovation in science. Many sound research studies would become infeasible where funding envelopes and participant availability is fixed. Small but well-designed studies are already not expected to produce significant results at an alpha level of 0.05 and yet can still provide important insight for future research (see e.g. Parmar *et al.* (2016) for a discussion in the context of clinical trials). Making the practice of science more rigorous by reducing researcher degrees of freedom is therefore also necessary.

Prespecification, Registration and Oversight

An explicit consideration of these and similar questions have become routine for those working in pharmaceutical trials, where both the financial incentive and the potential harms to patients are large and obvious, and drug regulatory authorities and independent trialists have been formalizing procedures to defend against false positives.

In trials, the intention to undertake a study is made public by registration (e.g. at www.isrctn.com or www.clinicaltrials.gov), the details of the intended design and measures published in the Trial Protocol, and sometimes too the Statistical Analysis Plan, that has been first approved by the independent Data Monitoring Committee as a fair and complete specification of the intended analysis. These documents reduce the degrees-of-freedom of the researcher, laying the foundation for effective disclosure and oversight, and for a direct comparison of the study report against the originally intended study (http://compare-trials.org/). For the field of psychology and psychiatry as a whole, registration also helps improve the visibility of the evidence base for otherwise unpublished negative findings (Munafo and Black, 2017).

Any investigator first experiencing this process finds it challenging. Firstly, it soon becomes apparent just how many research questions had been previously poorly defined and how many decisions, some large and many small, had been left implicit, postponed or simply not made. Secondly, each of these decisions corresponds to giving up a researcher degree-of-freedom, and many find this quite painful. In particular, committing to one definition of risk, outcome and one analysis is seen as "risking" missing important positive findings. Of course those "missed" findings would have been much more likely to be a false positive than any positive finding from a pre-specified test of a well-founded research hypothesis.

It should also be understood that pre-specification is not intended to bind investigators come what may. Departures from the prespecification can, and often should be made. But when, how and the reasons why should be explicit in the study publication (Simmons *et al.*, 2011) so that HARKing (Hypotheses after the results are known (Kerr *et al.*, 1999), the practice of presenting results of posthoc or other explorative analyses as research hypotheses, can be prevented without denying the researcher the right to present explorative, novel findings.

Reporting and Reviewing

First published in 1996 the Consolidated Standards of Reporting Trials (CONSORT) has become a model for more than 350 reporting guidelines that now span laboratory measurement to epidemiology, and now helpfully collated by the Equator Network (www.equator-network.org). Reporting guidelines ease the writing of a protocol for pre-registrations and a report of a study drafted following these guidelines should allow an independent researcher to reproduce the published findings. The fact that such a report would exceed the word length of traditional journals is, with the use of online supplements, now of little consequence.

Pre-specification and reporting guidelines make the task of the reviewer much simpler – clarity of research question, appropriateness of design and the evidence of rigorous implementation should now be obvious. However, it is important to recognize, especially for editors, that the combination of pre-specification and reporting standards that press for full disclosure leave the authors exposed to unconstructive criticism and constrained in their ability to respond, for example to requests for unplanned revised analysis. Publishing reviewer's comments and responses may also be useful to encourage constructive criticism.

Organization of Research

Description and awareness of researcher's degrees of freedom, while good practice, does not solve the problem of underpowered studies. University and college based undergraduate and postgraduate psychology is hugely popular, and the common training vehicle is the small scale structured experiment. Schaller (2016) suggests that the tested hypotheses are often little more than personal opinion, rarely being based on formal deduction from well-founded theory (e.g. Festinger and Hutte, 1954) and often with no or an over-optimistic consideration of likely effect-size, both increasing the chances of a false positive. Small enough for completion by a single student or scientist rehearsing basic study design and with greater emphasis on methodology to achieve high internal validity, little priority is given to external validity and generalizability. A simple way to improve the situation was suggested by Daniel Kahneman (cited in Yong, 2012) by turning undergraduate and graduate projects into replication studies.

In some areas of research increased scale can offer advantages over and above a reduction in false positives, such as where industrial scale brings the opportunity for industrial processing with improvements in standardisation, quality control and economy. But for studies involving, say, neuro-imaging, face-to-face interviewing or coding of observational data, as yet there seem few such economies. Thus, outside of genetics and a small number of epidemiological studies, truly large studies are rare in our field and more common has been the formation of consortia. A frequently adopted model is one where contributing groups agree a common set of core measures that will enable large sample analysis of a limited number of key research questions, while retaining scope for site specific independent extensions. Such consortia have the advantages of being potentially sustained by a mixture of core and site-specific funding, and can also provide a framework for intellectual collaboration, training and exchange. However, they can also incur significant costs with respect to much time-consuming discussion about process, agreements, standards, publication rights and so on. For scientists used to the rapid pursuit of flights of scientific imagination this can be anathema. However, incorporating a wider brief, including training and a sharing of scientific ideas, can make such organisations both productive and healthy environments for early career and experienced researchers alike. The BASIS (British Autism Sibling Infant Study) Consortium (http://www.basisnetwork.org) is one such which is more focussed on protocol sharing and making a core study sample available to a wider group of researchers who undertake largely independent research.

Machine Learning and Cross-Validation

Increasing study size or establishing consortium is not always feasible and current methodology needs to be optimized as well. Increasingly journals already encourage the use of new statistics by focusing on confidence intervals and effect sizes (Eich, 2014, Giofrè *et al.*, 2017) or the use of Bayesian statistics (Dienes, 2011, Gallistel, 2015).

An interesting alternative statistical approach to reducing the problem of false positive results is assessing prediction accuracy of unseen cases as a measure of evidence is based on work by statisticians in the 1970s (Allen, 1974, Browne, 1975, Geisser, 1975, Harrell *et al.*, 1996, Stone, 1974) and nowadays widely used in the field of machine learning. The contribution of machine learning to look at inference as a search through a space of possible hypotheses to identify one or more supported by the data (Hunter, 2017) is an important one, and unlike the frequentist and Bayesian approaches, it allows the inclusion of data-preprocessing steps, such as variable transformation, variable selection or imputation of missing data within the model selection process (Kuhn and Johnson, 2013) and can thus account for some of the problems of researcher's degrees of freedom. The prediction capability of a model for independent, unseen data is different to "classical" inferential statistics where efficient unbiased estimation of parameters is the goal. This means that the best predictive model may differ from the best explanatory model (Hastie *et al.*, 2009, Hoerl and Kennard, 1970, Sober, 2006). The first minimizes the prediction error of unseen data while the second minimizes prediction error of the training data set (Shmueli, 2010, Shmueli and Koppius, 2011).

Less well understood is that model selection and model assessment are two separate goals and cannot be assessed on the same unseen data set (Hastie *et al.*, 2009, Varma and Simon, 2006)). This is often ignored in the machine learning community (Cawley and Talbot, 2010). If we want to perform model selection we would need to randomly divide the dataset into three parts: a training set to fit the models, a validation set to estimate prediction error, and a test set to assess the generalization error of the final chosen model (Hastie et al., 2009) This three-way split-sample approach may be feasible for big-data but is usually not possible in medical research and is anyway inefficient and potentially unreliable (Harrell, 2015, Steyerberg, 2009). Better alternatives are nested cross-validation or bootstrap validation (see supplementary procedures). Being more complex and computationally demanding these are often avoided, simpler two-split cross-validation being used instead with users often unaware that this provides incomplete correction (Harrell, 2015, Hastie *et al.*, 2009, Stone, 1974). Also, the fashion to maximize prediction by using Support Vector Machines or Deep Learning should also be questioned, as these offer poor interpretability as compared to regularized methods and simpler models may perform better in practical replication studies (Hand, 2006).

External Validation, Replication and Triangulation

Cross-validation can deliver sound internal validation. For external validation to confirm the generalizability of findings replication with a new study sample of sufficient size to have good power for a plausible effect size is the gold standard. Journals need to accept replication studies and these should be linked to the original studies with the original paper. However, three highly influential journals refused to publish a (negative) replication of Bem's ESP study because of their policy not publishing straight replication studies (Gelman and O'Rourke, 2014, Yong, 2012). Even 2017, only 3% of 1151 psychology journals welcome scientists to submit replication studies (Martin and Clarke, 2017). Pre-registration of replication studies with a commitment to publish them irrespective of the outcome would help. Still better would be the encouragement of complementary studies to partner, sometimes in reciprocal agreements, enabling primary and replication studies to appear in the same or linked reports. The Dutch cohort study Generation-R and the UK Avon Longitudinal Study of Pregnancy and Childhood (ALSPAC), have examined the consistency of the effects of parental depression and anxiety during pregnancy (Van Batenburg-Eddes *et al.*, 2013), finding maternal symptoms increased attentional problems in both studies, but neither convincingly excluding the possibility of a confounder induced association. The ability to nominate a respected study as a replication partner within a research funding proposal is an important strength, emphasizing the willingness of the primary study to put its findings under immediate independent testing. Outside of genetic studies, however, such partnering remains rare.

The term replication presupposes much about the equivalence of the primary and replication populations. For example, sometimes primary study samples for developing a classifier are selected as pure cases and non-cases, whereas the replication sample might include more every-day patients. Discrepancies can occur even when the causal process of interest is the same in the two samples, but the pattern of confounders is different. It is therefore recommended to quantify the predictive accuracy of a prediction model in different samples from the same or similar target populations or domains (Debray et al., 2015). This needs also to be considered for "classical" statistical modelling. External validation studies may range from temporal (e.g., sample from the same hospital or primary care practice only later in time), to geographical (e.g., sample from different hospital, region, or even country), to validations across different medical settings (e.g., from secondary to primary care setting or vice versa) or different target populations or domains (e.g., from adults to children) with increasingly different study samples or case mix between development and validation samples. But is it exact/direct replication that is being sought, where as far as possible everything is held identical to the original study, or is it reproducibility (Drummond, 2009) – the replication of the concept, where some aspects of the study are allowed or designed to be different? The latter is more appealing to funders and editors in that it has novel elements, but the associated researcher degrees of freedom that comes with these must be made visible and controlled. Triangulation could be considered a special case of a reproducibility study, where the differences in the studies relate to the differences in the likely impact or measurement of confounders rather than the core variables (Lawlor *et al.*, 2016)

Meta-Analysis/Systematic Review

Systematic review, the exhaustive collation of research findings, their structured evaluation against indicators of quality, and their formal quantitative combination has become the accepted basis for recommendations from organisations such as NICE (National Institute for Clinical Excellence www.nice.org) charged with formulating evidence based recommendations as to good clinical practice for the National Health Service in England. The best method available to establish consensus in multiple studies addressing the same issue is meta-analysis (Cumming, 2014). In therapeutic research systematic review has been led by the Cochrane collaboration that started in 1993. However at the end of 2014, of the more than five thousand Cochrane Reviews, only 2% (119) were within the areas of developmental, psychosocial and learning problems. It appears that many reviews fail to identify sufficient studies of eligible quality to justify a formal review panel. This cannot be considered satisfactory.

The small number of eligible studies also does not allow an adequate assessment of publication or other biases, such as selective analyses or outcome reporting towards significant studies (Ioannidis, 2008, Ioannidis and Karassa, 2010, Sterne *et al.*, 2011). Based on their expected power, Tsilidis and colleagues (2013) examined animal studies of neurological disorders finding that almost twice as many than expected showed a significant result. The discrepancy was larger in studies with small sample sizes. Even in randomized clinical trials for the efficacy of psychotherapies for major depressive disorder an excess of significant studies is evident (49 % expected versus 58% reported, Flint et al., 2015). Although researchers are nowadays fully aware of the file drawer problem it seems still to exists. Finally, results are also sensitive to study inclusion criteria e.g. meta-analyses of moderating effect of serotonin transporter genotype on the association between stressful life events and depression resulted in different conclusions based on different (but sensible) inclusion criteria (Taylor and Munafò, 2016). Psychiatric studies may be particular prone to this problem because of the variety of outcomes measures of similar cognitive deficits and of types of control treatments or groups.

Thus even results from Cochrane based meta-analyses need to be treated with care, and biases towards positive results assessed (Ioannidis and Trikalinos, 2007), perhaps using the online p value analyser (www.p-curve.com) of Simonsohn et al (2014) and used by Taylor and Munafo (2016). Schuit et al (2015) suggest a simple correction based on the number of studies, the number of significant results and the sample size of studies to keep the nominal alpha error at the pre-specified (i.e. 5%) level without loss of power. The use of network meta-analyses (Caldwell, 2014, Caldwell *et al.*, 2005) which allows the comparison of a network of related studies addressing a common condition or outcome should also be encouraged.

We have already noted that there are instances where small scale studies remain essential (Parmar et al. (2016)). Small trials of good quality may also be desirable from the point of view of generalization, with results of meta-analyses involving several smaller independent studies potentially being more reliable than performing one large study as long as all studies are available for a meta-analysis (Cappelleri *et al.*, 1996, Contopoulos-Ioannidis *et al.*, 2005).

Conclusions

Statisticians are professional sceptics. That more than half of published findings are not reproducible came as no surprise to us. The environment in which scientists work, namely what we teach, the training and career opportunities, data and publication infrastructures, and individual and institutional incentives all require careful monitoring and, in many cases, revision. We would encourage modernization of the curriculum for research methods to get away from presenting context-free statistical tests, to highlight instead the systematic and where possible pre-specified approaches to each step of the research process, and statistical modelling and analysis methods that more properly account for and describe uncertainty – the latter requiring further development work from statisticians. In spite of the claims of some, machine learning alone will not solve the problem, and badly done will make the problem worse. Well done, which is currently not that easy, it has a valuable place in our overall endeavour. Statistical learning, the study of properties of learning algorithm from the perspective of statistical theory can serve as a unifying of statistical modelling and machine learning (Hastie et al 2009). Perhaps we might even need to change the way research excellence is defined, away from the number of high impact publications and high h-index towards a replication index which measures how often other scientists could replicate their results as Chamber and Sumner (2012) suggested in a newspaper article.

Of course, the problem does not stop in the scientific community. Dumas-Mallet et al (2017) reported that newspapers also selectively cover mainly exciting positive findings of lifestyle risk factors studies, often triggered by exaggerated press releases from Universities (Sumner *et al.*, 2014). Here psychiatry performed poorly with only ~25% of studies being supported by later meta-analyses. Even more worryingly, newspapers rarely inform the public about the subsequent null finding. The crisis of replication is thus is not only a problem of science but also of the media, as the autism and vaccination “debate” illustrated only too well.

Figure 1 and figure legend

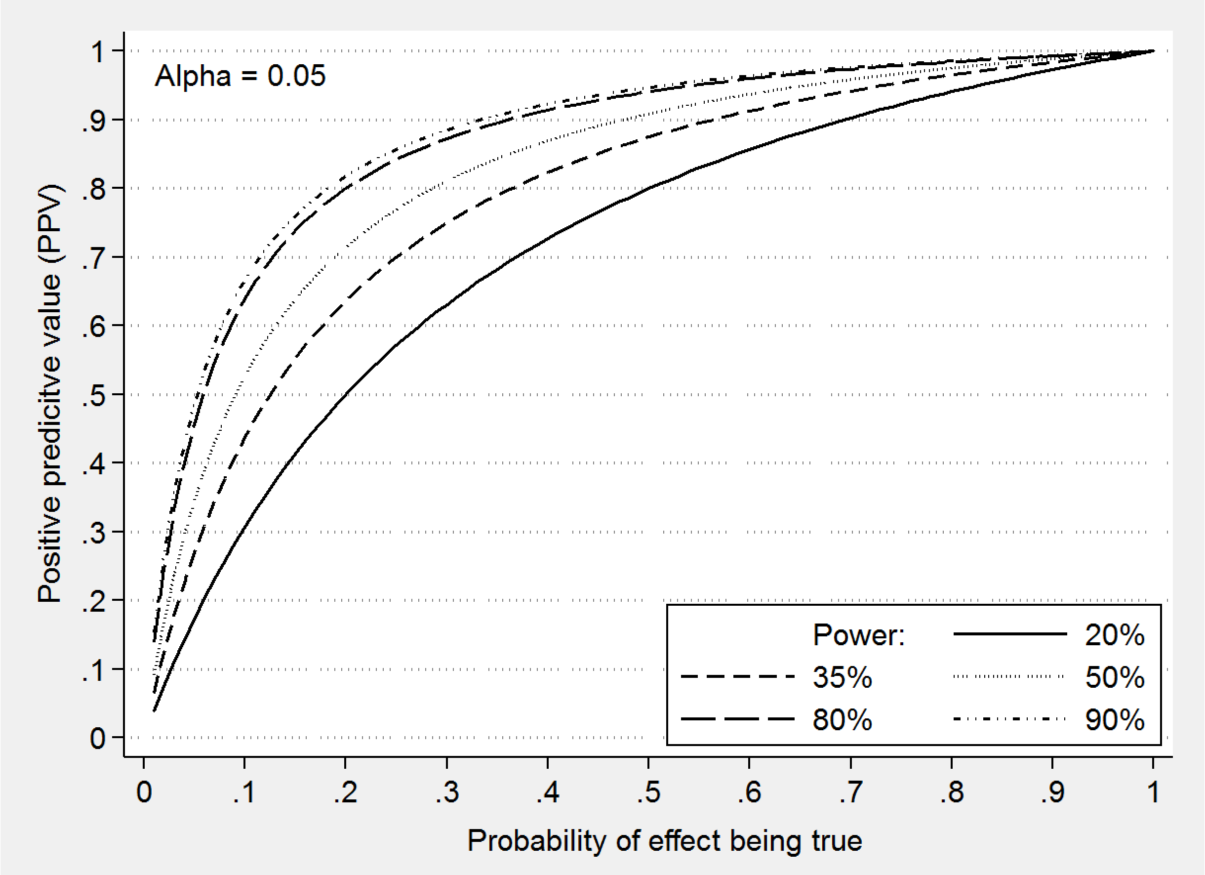

Figure 1: Why more than 50% of published research findings are false

Ioannidis (2005) claimed that more than 50% of published “significant” research findings in natural sciences are false. As others before him he explained that the p-value of a statistical test does not allow making probabilistic statements about the underlying validity of an effect. For that we need to make assumptions about the prior probability that an effect exists. The probability that a statistical test with a positive “significant” result is a true effect is the positive predicted value (PPV). The PPV depends on our prior probability, the power of a test (the probability of a test to detect a specified effect with a given sample size) and the alpha error.

Assume we are assessing 1000 risk factors for a mental health problem and 10% of the risk factors are real (but we do not know which ones). Our statistical tests have got 80% power to get a significant result if a specific effect exists and we usually specify an alpha error level of 0.05 or 5%. We would detect 80% of our 100 risk factors and thus obtain 80 true positive test results. 20 risk factors would not be detected (false negative results). The remaining

900 factors are not risk factors but we would expect that 5% or 45 would be falsely declared as such (false positives). We would therefore obtain 80+45 = 125 positive results but only 80 (64%) are true positives and about 1/3rd of the positive results are false positives, far more than our alpha error of 5% would suggest.

However, the PPV also depends on the power (and therefore the sample size) and the alpha error. Meta-analyses have shown that typically studies in psychology and psychiatry often have little power: an estimate for animal studies with neurological diseases is 20 % power (Tsilidis et al., 2013) and 35% for psychological studies (Bakker et al., 2012) although low power is not restricted to these research domains (Dumas-Mallet et al., 2017a). The Figure shows the PPV (Y axis) against the probability that a hypothesis is true for different powered studies (25%, 35%, 50%, 80% and 90 %) at an alpha error of 5%. Even without publication bias towards positive results the PPV of positive statistical results is far less than we may have expected. It underlines the importance of conducting and presenting a power analyses where feasible.

References

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., ..., Zuni, K. & Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* **349**.

Allen, D. M. (1974). Relationship between Variable Selection and Data Augmentation and a Method for Prediction. *Technometrics* **16**, 125-127.

Bakker, M., van Dijk, A. & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science* **7**, 543-554.

Bem, D. J. (2011). Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. *Journal of Personality and Social Psychology* **100**, 407-425.

Bender, R. & Lange, S. (2001). Adjusting for multiple testing - when and how? *Journal of Clinical Epidemiology* **54**, 343-349.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E., ... & Tingley, D. (2017). Redefine Statistical Significance. *Human Nature Behavior*.

Browne, M. W. (1975). Comparison of Single Sample and Cross-Validation Methods for Estimating Mean Squared Error of Prediction in Multiple Linear-Regression. *British Journal of Mathematical & Statistical Psychology* **28**, 112-120.

Caldwell, D. M. (2014). An overview of conducting systematic reviews with network meta-analysis. *Systematic Reviews* **3**, 109-109.

Caldwell, D. M., Ades, A. E. & Higgins, J. P. T. (2005). Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ* **331**, 897-900.

Cappelleri, J. C., Ioannidis, J. P. A., Schmid, C. H., deFerranti, S. D., Aubert, M., Chalmers, T. C. & Lau, J. (1996). Large trials vs meta-analysis of smaller trials - How do their results compare? *Jama-Journal of the American Medical Association* **276**, 1332-1338.

Cawley, G. C. & Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research* **11**, 2079-2107.

Chamber, C. & Sumner, P. (2012). Replication is the only solution to scientific fraud. In *Guardian*. Guardian: London, UK.

Contopoulos-Ioannidis, D. G., Gilbody, S. M., Trikalinos, T. A., Churchill, R., Wahlbeck, K., Ioannidis, J. P. A. & Project, E.-P. (2005). Comparison of large versus smaller randomized trials for mental health-related interventions. *American Journal of Psychiatry* **162**, 578-584.

Cumming, G. (2014). The New Statistics. *Psychological Science* **25**, 7-29.

Debray, T. P. A., Vergouwe, Y., Koffijberg, H., Nieboer, D., Steyerberg, E. W. & Moons, K. G. M. (2015). A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology* **68**, 280-289.

Dienes, Z. (2011). Bayesian Versus Orthodox Statistics: Which Side Are You On? *Perspectives on Psychological Science* **6**, 274-290.

Drummond, C. (2009). Replicability is not Reproducibility: Nor is it Good Science. In *Proceedings of the Twenty-Sixth International Conference on Machine Learning*, p. 4. ICML: Montreal, Canada.

Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F. & Munafò, M. R. (2017a). Low statistical power in biomedical science: a review of three human research domains. *Royal Society Open Science* **4**, 160254.

Dumas-Mallet, E., Smith, A., Boraud, T. & Gonon, F. (2017b). Poor replication validity of biomedical association studies reported by newspapers. *Plos One* **12**.

Edmonds, B., Gilbert, N., Ahrweiler, P. & Scharnhorst, A. (2011). Simulating the Social Processes of Science. *Jasss-the Journal of Artificial Societies and Social Simulation* **14**.

Eich, E. (2014). Business Not as Usual. *Psychological Science* **25**, 3-6.

Fanelli, D. (2009). How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *Plos One* **4**.

Fanelli, D. (2010). "Positive" Results Increase Down the Hierarchy of the Sciences. *Plos One* **5**.

Festinger, L. & Hutte, H. A. (1954). An Experimental Investigation of the Effect of Unstable Interpersonal Relations in a Group. *Journal of Abnormal and Social Psychology* **49**, 513-522.

Flint, J., Cuijpers, P., Horder, J., Koole, S. L. & Munafo, M. R. (2015). Is there an excess of significant findings in published studies of psychotherapy for depression? *Psychological Medicine* **45**, 439-446.

Freedman, D. A. (1983). A Note on Screening Regression Equations. *American Statistician* **37**, 152-155.

Gallistel, R. (2015). Bayes for beginners: Probability and likelihood. In *Observer*.

Geisser, S. (1975). Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association* **70**, 320-328.

Gelman, A. & Loken, E. (2014). The Statistical Crisis in Science. *American Scientist* **102**, 460-465.

Gelman, A. & O'Rourke, K. (2014). Discussion: Difficulties in making inferences about scientific truth from distributions of published p-values. *Biostatistics* **15**, 18-23.

Giofrè, D., Cumming, G., Fresc, L., Boedker, I. & Tressoldi, P. (2017). The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices. *PLOS ONE* **12**, e0175583.

Hand, D. J. (2006). Classifier Technology and the Illusion of Progress. 1-14.

Harrell, F. (2015). *Regression Modeling Strategies*. Springer: New York, USA.

Harrell, F. E., Lee, K. L. & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361-387.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer-Verlag: New York.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge Regression - Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55-&.

Hunter, L. (2017). An introduction to machine learning for statisticians University of Colorado.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Plos Medicine* **2**, 696-701.

Ioannidis, J. P. A. (2008). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of Evaluation in Clinical Practice* **14**, 951-957.

Ioannidis, J. P. A. & Karassa, F. B. (2010). The need to consider the wider agenda in systematic reviews and meta-analyses: breadth, timing, and depth of the evidence. *British Medical Journal* **341**.

Ioannidis, J. P. A. & Trikalinos, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *CMAJ : Canadian Medical Association Journal* **176**, 1091-1096.

Johnson, V. E. (2013). Uniformly most powerful Bayesian tests. *Ann. Statist.* **41**, 1716-1741.

Kerr, N. L., Niedermeier, K. E. & Kaplan, M. F. (1999). Bias in jurors vs bias in juries: New evidence from the SDS perspective. *Organizational Behavior and Human Decision Processes* **80**, 70-86.

Kuhn, M. & Johnson, K. (2013). *Applied Predcitive Modelling*. Springer-Verlag: New York.

Lawlor, D. A., Tilling, K. & Davey Smith, G. (2016). Triangulation in aetiological epidemiology. *International Journal of Epidemiology* **45**, 1866-1886.

Luengo-Fernandez, R., Leal, J. & Gray, A. (2015). UK research spend in 2008 and 2012: comparing stroke, cancer, coronary heart disease and dementia. *BMJ Open* **5**.

Ly, A., Verhagen, J. & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology* **72**, 19-32.

Martin, G. N. & Clarke, R. M. (2017). Are Psychology Journals Anti-replication? A Snapshot of Editorial Practices. *Frontiers in Psychology* **8**.

Munafo, M. R. & Black, S. (2017). Personality and Smoking Status: A Longitudinal Analysis (vol 9, pg 397, 2007). *Nicotine & Tobacco Research* **19**, 129-129.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J. & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour* **1**, 0021.

Parmar, M. K. B., Sydes, M. R. & Morris, T. P. (2016). How do you design randomised trials for smaller populations? A framework. *BMC Medicine* **14**, 183.

Pashler, H. & Wagenmakers, E. J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science* **7**, 528-530.

Pickles, A. (2009). What Clinicians Need to Know about Statistical Issues and Methods. In *Rutter's Child and Adolescent Psychiatry* (ed. M. Rutter, B. D.V.M., D. S. Pine, S. Scott, S. Stevenson, E. Taylor and A. Thapar), pp. 111-122. Wiley: Oxford.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review* **16**, 225-237.

Sagan, C. (1995). *The demon-haunted world : science as a candle in the dark*. Random House,: New York

Schaller, M. (2016). The empirical benefits of conceptual rigor: Systematic articulation of conceptual hypotheses can reduce the risk of non-replicable results (and facilitate novel discoveries too). *Journal of Experimental Social Psychology* **66**, 107-115.

Schuit, E., Roes, K. C., Mol, B. W., Kwee, A., Moons, K. G. & Groenwold, R. H. (2015). Meta-analyses triggered by previous (false-)significant findings: problems and solutions. *Systematic Reviews* **4**, 57.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science* **25**, 289-310.

Shmueli, G. & Koppius, O. R. (2011). Predictive Analytics in Information Systems Research. *Mis Quarterly* **35**, 553-572.

Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* **22**, 1359-1366.

Simonsohn, U., Nelson, L. D. & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *J Exp Psychol Gen* **143**.

Sober, E. (2006). Parsimony. In *The Philosophy of Science: An Encyclopaedia* (ed. S. Sarkar and J. Pfeifer), pp. 531-538. Routledge: Oxford.

Sterling, T. D. (1959). Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance - or Vice Versa. *Journal of the American Statistical Association* **54**, 30-34.

Sterling, T. D., Rosenbaum, W. L. & Weinkam, J. J. (1995). Publication Decisions Revisited - the Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice-Versa. *American Statistician* **49**, 108-112.

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rücker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D. & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* **343**.

Steyerberg, E. W. (2009). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer: New York.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **36**, 111-147.

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Venetis, C. A., Davies, A., Ogden, J., Whelan, L., Hughes, B., Dalton, B., Boy, F. & Chambers, C. D. (2014). The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ : British Medical Journal* **349**.

Taylor, A. E. & Munafò, M. R. (2016). Triangulating meta-analyses: the example of the serotonin transporter gene, stressful life events and major depression. *BMC Psychology* **4**, 23.

Tsilidis, K. K., Panagiotou, O. A., Sena, E. S., Aretouli, E., Evangelou, E., Howells, D. W., Salman, R. A. S., Macleod, M. R. & Ioannidis, J. P. A. (2013). Evaluation of Excess Significance Bias in Animal Studies of Neurological Diseases. *Plos Biology* **11**.

Van Batenburg-Eddes, T., Brion, M. J., Henrichs, J., Jaddoe, V. W. V., Hofman, A., Verhulst, F. C., Lawlor, D. A., Smith, G. D. & Tiemeier, H. (2013). Parental depressive and anxiety symptoms during pregnancy and attention problems in children: a cross-cohort consistency study. *Journal of Child Psychology and Psychiatry* **54**, 591-600.

Varma, S. & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7**, 91-91.

Wagenmakers, E. J., Wetzels, R., Borsboom, D. & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology* **100**, 426-432.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J. & Wagenmakers, E.-J. (2011). Statistical Evidence in Experimental Psychology. *Perspectives on Psychological Science* **6**, 291-298.

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M. & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology* **7**.

Yong, E. (2012). Bad Copy. *Nature* **485**, 298-300.