ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Royal Statistical Society
Statistics in Society
Series A

Professor David Spiegelhalter

J. R. Statist. Soc. A (2017)
180, *Part* 4, *pp.* 949–965

Trust in numbers

David Spiegelhalter

University of Cambridge, UK

[*The address of the President delivered to* The Royal Statistical Society *on Wednesday, June 28th, 2017*]

Summary. Those who value quantitative and scientific evidence are faced with claims both of a reproducibility crisis in scientific publication and of a post-truth society abounding in fake news and alternative facts. Both issues are of vital importance to statisticians, and both are deeply concerned with trust in expertise. By considering the 'pipelines' through which scientific and political evidence is propagated, I consider possible ways of improving both the trustworthiness of the statistical evidence being communicated, and the ability of audiences to assess the quality and reliability of what they are being told.

Keywords: Media filters; Post-truth; Reproducibility; Trustworthiness

1. Introduction

This Presidential address gives me a fine opportunity to bring together two topical issues: first, the claims of a reproducibility crisis in science, which have led to concerns about the quality and reliability of at least parts of the scientific literature; second, the suggestion that we live in a 'post-truth' society abounding in fake news and alternative facts, in which emotional responses dominate evidence-informed judgement. These two topics have a close connection: both are associated with claims of a decrease in trust in expertise, and both concern the use of numbers and scientific evidence. They are therefore of vital importance to professional statisticians, or to any who analyse and interpret data.

A simple Internet search will reveal the daunting amount that has been written about these contested issues, and here I can give only a brief personal review of the evidence and the possible causes, focusing on the 'filters' that distort statistical evidence as it is passed through the information pipeline from the originators to its final consumption by the public. No single group can deal with these complex matters, but I shall argue that statisticians, and in particular the Royal Statistical Society, have an essential role both in improving the trustworthiness of statistical evidence as it flows through the pipeline, and in improving the ability of audiences to assess that trustworthiness. On statistical shoulders rests a great responsibility.

2. Reproducibility and replication

The idea of a 'reproducibility or replication crisis' might reasonably be said to date from John Ioannidis's 2005 paper which notoriously proclaimed 'Why most published research findings are false' (Ioannidis, 2005). Although initially concerned with the biomedical literature, the idea has since been applied particularly to psychology and other social sciences. Note that,

Address for correspondence: David Spiegelhalter, Statistical Laboratory, Centre for Mathematical Sciences, Wilberforce Road, Cambridge, CB3 0WB, UK.
E-mail: d.spiegelhalter@statslab.cam.ac.uk

 0964–1998/17/180949

although attempts have been made to define 'reproducibility' and 'replication' precisely (Leek and Jager, 2017), I feel that we should try to avoid giving yet more technical definitions to words in routine use. (We are all familiar with misunderstandings from using 'significance' as a technical term, but I had always thought of 'expected' as fairly innocuous. That was until a journalist labelled all hospital deaths above the expected level as 'unexpected'.) So I shall treat the terms interchangeably and distinguish when an entire study is repeated or when data are reanalysed).

The extent of this 'crisis' is contested. Ioannidis (2005) was based on modelling rather than empirical evidence: he argued that reasonable assumptions about the design of studies, biases in conduct, selection in reporting and the proportion of hypotheses investigated that were truly non-null meant a high rate of 'false discoveries', i.e. the proportion of published positive results that were actually null hypotheses that had been falsely rejected. In contrast, an analysis of published *p*-values (Jager and Leek, 2014) came up with an estimated false discovery rate of 14% in the mainstream medical literature, and a recent review (Leek and Jager, 2017) concluded 'We do not believe science is in the midst of a crisis of reproducibility, replicability, and false discoveries'.

So was the claim about false claims itself a false claim? This is strongly disputed by Ioannidis (2014), and a recent exercise (Szucs and Ioannidis, 2017) scraped nearly 30000 *t*-statistics and degrees of freedom from recent psychology and neuroscience journals, and on the basis of the observed effect sizes and low power concluded that

> 'Assuming a realistic range of prior probabilities for null hypotheses, false report probability is likely to exceed 50% for the whole literature'.

Some of this apparent disagreement will be due to different literatures: Jager and Leek (2014) examined abstracts from top medical journals with many randomized controlled trials and meta-analyses, which would be expected to be much more reliable than first claims of 'discoveries'. And even a 14% false discovery rate might be considered too high.

An alternative approach is purely empirical, in which the experiments behind past published claims are replicated by other teams of researchers: for example the effect of 'power posing', popularized in a Ted talk that has been viewed over 40 million times (Cuddy, 2012), has been subject to repeated failed replications (Ranehill *et al.*, 2015). The reproducibility project was a major exercise in which 100 psychology studies were replicated with higher power (Open Science Collaboration, 2015): whereas 97% of the original studies had statistically significant results, only 36% of the replications did. This was widely reported as meaning that the majority of the original studies were false discoveries, but Patil *et al.* (2016) pointed out that 77% of the new results lay within the 95% predictive interval from the original study, which corresponds to there not being a significant difference between the original and replication studies. This illustrates that the *difference between significant and not significant is often not significant* (Gelman and Stern, 2006). But it also means that 23% of original and replication studies had significantly different results.

Perhaps the main lesson is that we should stop thinking in terms of significant or not significant as determining a 'discovery', and instead focus on effect sizes. The reproducibility project found that replication effects were on average in the same direction as the originals but were around half their magnitude (Open Science Collaboration, 2015). This clearly displays the biased nature of published estimates in their literature, and strong evidence for what might be termed regression to the null.

3. What is the cause of this 'crisis'?

It is important to note that deliberate fabrications of data do occur but appear relatively rare. A

review estimated that 2% of scientists admitted falsification of data (Fanelli, 2009), and the US National Science Foundation and Office of Research Integrity deal with a fairly small number of deliberately dishonest acts (Mervis, 2017), although substantial numbers of cases must go undetected as it is generally difficult to check raw material. Computational errors are more common but can be detected by repeating analyses if the original data are available.

Rather than deliberate dishonesty or computational incompetence, the main blame has been firmly placed on a 'failure to adhere to good scientific practice and the desperation to publish or perish' (Begley and Ioannidis, 2015). The crucial issue is the quality of what is submitted to journals, and the quality of what is accepted, and deficits are a product of what have become known as 'questionable research practices'.

Fig. 1 shows the results of a survey of academic psychologists in the USA, which had a 36% response rate (John *et al.*, 2012). A very low proportion admitted falsification, but other practices that can severely bias outcomes were not only frequently acknowledged but also generally seen as defensible: for example the 50% who admitted selectively reporting studies gave an average score of 1.66 when asked whether this practice was defensible, where 0 ≡ no, 1 ≡ possibly and 2 ≡ yes. An Italian survey found similar rates, although the respondents were more inclined to agree that the practices were not defensible (Agnoli *et al.*, 2017).

These questionable research practices just involve experimentation. If we consider general observational biomedical studies and surveys, then there is a vast range of additional potential source of bias: these might include

(a) sampling things that are convenient rather than appropriate,
(b) leading questions or misleading wording,
(c) inability to adjust properly for confounders and to make fair comparisons,

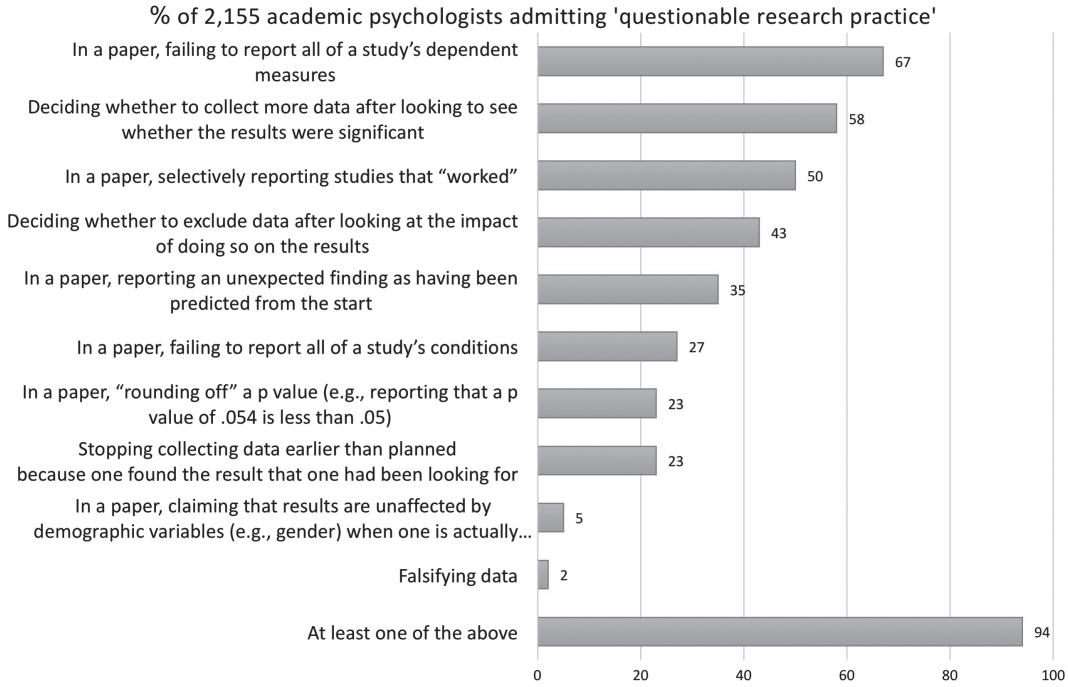

Fig. 1. Questionable research practices admitted by 2155 US academic psychologists (John *et al.*, 2012)

(d) too small a sample,
(e) inappropriate assumptions in a model and
(f) inappropriate statistical analysis.

And to these we might add many additional questionable practices concerned with interpretation and communication, which we shall return to later.

These are not just technical issues of, say, lack of adjustment of p-values for multiple testing. Many of the problems arise through more informal choices made throughout the research process in response to the data, say in selecting the measures to emphasize, choice of adjusting variables, cut points to categorize continuous quantities and so on: this has been described as the 'garden of forking paths' (Gelman and Loken, 2014) or 'researcher degrees of freedom' (Simmons *et al.*, 2011) and will often take place with no awareness that these are questionable research practices.

There have been strong arguments that the cult of p-values is fundamental to problems of reproducibility, and recent guidance from the American Statistical Association clearly revealed their misuse (Wasserstein and Lazar, 2016). Discussants called for their replacement or at least downplaying their pivotal role in delineating 'discoveries' through the use of arbitrary thresholds. We have already seen that p-values are fragile things that need to be handled carefully in replication studies—for example a study with $p = 0.05$ would only be predicted a 50% chance of obtaining $p < 0.05$ in a precise replication.

This is a complex issue, and in a recent article in *Significance* (Matthews *et al.*, 2017) I confessed that I liked p-values, and that they are good and useful measures of compatibility between data and hypotheses, but insufficient distinction is made between their informal use in exploratory analysis and their more formal use in confirmatory analyses that summarize the totality of evidence—perhaps they should be distinguished as p_{exp} and p_{con}.

Essentially there is too strong a tendency to use p-values to jump from selected data, to a claim about the strength of evidence, and to conclusions about the practical importance of the research. p-values 'do what they say on the tin', but people do not read the tin.

4. What gets into the scientific literature?

Questionable practices influence what is submitted to the scientific literature, and what finally appears depends on the publisher's ability to critique and select from what is presented to them. Ideally, peer review would weed out inadequate research and reporting, and recommend publication of good science regardless of the actual results. But we know that peer review is often inadequate, and there is an urge for the leading journals, to a varying amount across different disciplines, to publish newsworthy positive 'discoveries' and hence they produce a skewed resource.

We should not be surprised at this, since traditionally journals were set up to report new findings rather than the totality of evidence. Now there is an explosion in the amount of research and publishing opportunities, I would agree that

> 'most scientific papers have a lot more noise than is usually believed, that statistically significant results go in the wrong direction far more than 5% of the time, and that most published claims are overestimated, sometimes by a lot'

(Gelman, 2013). Although Gelman adds, more positively, that even though there are identifiable problems with individual papers, areas of science could still be moving generally in the right direction.

So what can be done? A group of prominent researchers recently published a 'manifesto for

Table 1. Proposals from the 'manifesto for reproducible science' (Munafò *et al*., 2017)

Theme	*Proposal*	*Stakeholders*
Methods	Protecting against cognitive biases	Journals, funders
	Improving methodological training	Funders, institutions
	Independent methodological support	Funders
	Collaboration and team science	Funders, institutions
Reporting and dissemination	Promoting study pre-registration	Journals, funders
	Improving the quality of reporting	Journals
	Protecting against conflicts of interest	Journals
Reproducibility	Encouraging transparency and open science	Journals, funders, regulators
Evaluation	Diversifying peer review	Journals
Incentives	Rewarding open and reproducible practices	Journals, funders, institutions

Table 2. Who is trusted as a source of medical research information?: responses from 1500 UK adults (Wellcome Trust, 2017)

Profession	*Trust completely or largely (%)*	*Trust very little or not at all (%)*
Doctors or nurses	64	6
University scientists	59	4
Medical research charities	37	11
Pharma scientists	32	16
Industry scientists	29	16
Journalists	3	59

reproducible science' whose recommendations for action are summarized in Table 1, together with the relevant stakeholders (Munafò *et al.*, 2017).

The list in Table 1 demonstrates the responsibility of a wide range of stakeholders and is firm in its commitment to transparency. Statistical science has a major role in many of these proposals, in particular in methodological training, improving reporting and peer review, and the sharing of data for reanalysis. However, the one important element that seems to be missing from Table 1 is the need for external commentary, critique and 'calling out' of poor practice, which are the responsibility of the entire scientific community, the media and the public—we shall return to this theme later.

In spite of extensive discussion about problems in the reliability of published science, these concerns do not seem to have fed into public opinion yet. A recent survey (Wellcome Trust, 2017) revealed the trust ratings shown in Table 2.

It is ironic that pharmaceutical scientists are given low levels of trust, although they work under far greater constraints than university scientists in terms of prespecification of design and analyses for regulators, and arguably produce more trustworthy analyses (personally I would trust the opinion of pharma statisticians on medical research far more than I would many health professionals). Journalists are given very low trust ratings in spite of being a major source of information to the public.

This introduces the idea that expressions of trust are not, in general, based on careful consideration of evidence but arise as an immediate response based on our gut feelings, which brings

us naturally to the way that we handle all the other numbers that deluge us as part of daily life, and in particular those that appear in the news.

5. Numbers in the news

Scientists are not the only people reporting claims based on statistical evidence. Politicians, non-governmental organizations and many other bodies are all competing for our attention, using numbers and science to provide an apparently 'objective' basis for their assertions. Technology has changed, encouraging an increasing diversity of sources to use on-line and social media to communicate, with few controls to ensure reliable use of evidence. This has led to suggestions that we are in a time of populist political discourse in which appeals to our emotion triumph over reason.

At the extreme, there are completely fabricated, demonstrably false facts that can be honestly labelled 'fake news'. But, as with science, I believe that deliberate fabrication is not the main issue: this will be better dealt with in the future by a combination of calling-out by fact checking organizations such as Full Fact (Full Fact, 2017), crowdsourcing on social media, automatic algorithms and possible regulation of social media sites: for example, Full Fact covered the recent election campaign in the *Evening Standard* as well as collaborating with Facebook on prominent advertising of 'Tips for spotting false news'.

As with science, a much bigger risk is manipulation and exaggeration through inappropriate interpretation of 'facts' that may be technically correct but are distorted by what we might call 'questionable interpretation and communication practices'. Fig. 2 provides a highly simplified view of the process by which we hear about statistical evidence as the end of a pipeline that starts with the originators of the data, and then goes through the 'authorities', then through their press and communication offices to the traditional media, and finally to us as individual members of society.

The questionable practices that have been adopted by some of the more ruthless press offices, communications teams and journalists include those in Table 3.

Many of these would be seen as defensible by those whose professional career depends on attracting readers, listeners or clicks, and it would be very interesting to conduct a survey of press officers and journalists to see how many of these they had used. But it is important to note that scientists might use these as well—in my experience some can, in spite of their proclaimed *caveats*, be too quick to jump to the wider implications of their work. When communicating

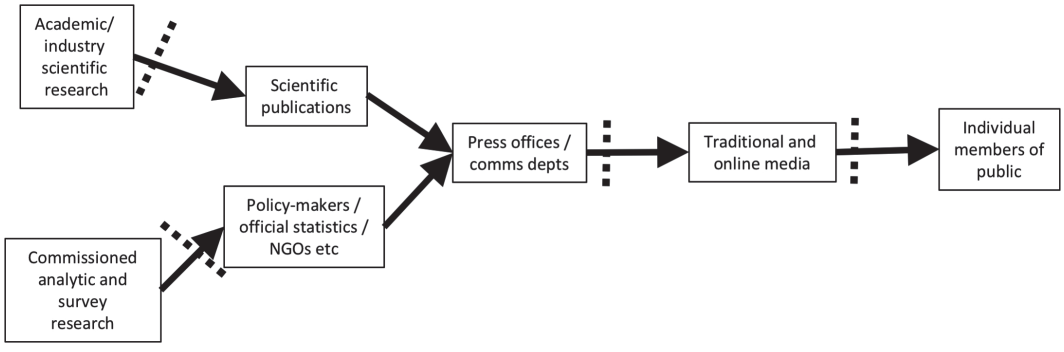

Fig. 2. Simplistic diagram of the traditional information flows from statistical sources through to the public: ⋰, ⋮, ⋱, filters arising from questionable research, interpretation and communication practices, such as selective reporting, lack of context and exaggeration of importance

Table 3. Some highly questionable interpretation and communication practices

Pick stories that go against current consensus
Promote stories regardless of research quality
Do not report uncertainties
Do not provide context or comparative perspective, such as a time series
Suggest a cause when only an association is observed
Exaggerate relevance and importance of findings
Claim that the evidence supports a particular policy
Provide only relative and not absolute risks
Use positive or negative framing depending on whether the aim is to reassure or frighten
Do not dwell on conflicts of interest or alternative views
Use a gripping but uninformative graphic
Write a headline which may have little connection to the story but will encourage clicks

subtle statistical evidence there seems an irresistible tendency to produce a simplifying narrative; we have seen X, it is because of Y, and so we should do Z.

This pipeline suggests that it is too easy to blame journalists for misreporting science. Press offices, and the journals and scientists themselves, can be to blame: a recent study found that, of 462 press releases from UK universities in 2011, 40% contained exaggerated advice, 33% contained exaggerated causal claims, 36% contained exaggerated inference to humans from animal research and the majority of exaggerations appearing in the press could be traced back to the press release (Sumner *et al.*, 2014). The same team found slightly more reassuring results in 534 press releases from major biomedical journals: causal claims or advice in a paper were exaggerated in 21% of corresponding press releases, although these exaggerations, which tended to be reported, did not produce more press coverage (Sumner *et al.*, 2016).

One of my favourite examples of imaginative storytelling by a communications team concerns a rather dull study which found that 10% of people carried a gene which protected them against high blood pressure (Newton-Cheh *et al.*, 2009), which was reframed negatively as nine in 10 people carrying a gene which *increases* the risk of high blood pressure: this duly received international press coverage (Devlin, 2009).

Another recent classic was a careful Swedish population study whose published abstract (Khanolkar *et al.*, 2016) said 'We observed consistent associations between higher socioeconomic position and higher risk of glioma'; the press release headlined with 'High levels of education linked to heightened brain tumour risk' (Medical Xpress, 2016) and the subeditor of the *Daily Mirror* finally turned it into 'Why going to university increases risk of getting a brain tumour' (Gregory, 2017).

The use of relative risks without absolute risks is a standard complaint and is explicitly warned against in British Broadcasting Corporation (BBC) statistical guidelines (British Broadcasting Corporation, 2017). It is known that relative risks, which are often referred to by the media as simply an 'increased risk' regardless of magnitude, are an effective way of making a story look more exciting, and this is not helped by the fact that odds, rate and hazard ratios are the standard output from most biomedical studies. The gripping headline 'Why binge watching your TV box sets could kill you' (Donnelly, 2016) arose from an epidemiological study that estimated an adjusted hazard ratio of 2.5 for a fatal pulmonary embolism associated with watching more than 5 h of television a night compared with less than 2.5 h (Shirakawa *et al.*, 2016). But careful scrutiny of the absolute rate in the high risk group (13 in 158000 person-years) could be translated as meaning that you can expect to watch more than 5 h of television a night for 12000

years before experiencing the event, which somewhat lessens the impact. The newspaper article was, as usual, much better than the headline, but whose fault is it not to insist on including this *perspective*: the journalist, the press office, the publication or the scientists?

I am unsure whether this misuse of statistical evidence is becoming worse. There are certainly more outlets promoting partial news, but mainstream Web sites, newspapers and television are perhaps under more scrutiny. What would fact checkers have found were they active say 30 years ago? My unsupported feeling is that it would have been even worse than now.

But in my darkest moods I follow what could be called the 'Groucho principle': because stories have gone through so many filters that encourage distortion and selection, the very fact that I am hearing a claim based on statistics is reason to disbelieve it.

6. Trust in expertise

When faced with stories about how the natural world or society works, we can rarely check them for ourselves. So *trust* is an inevitable element in dealing with statistical evidence, and therefore recent claims that there has been a decrease in trust in expertise is worth serious attention.

This claim is often associated with Michael Gove in the 'Brexit' campaign saying that people had had enough of experts, but it is important to quote him in full:

> 'people have had enough of experts from organisations with acronyms saying that they know what is best and getting it consistently wrong'

(Youtube, 2016). This sounds a little more reasonable and reflects recent high-profile failures to predict and control financial markets, and the lamentable quality of many political forecasts identified by the 'Good judgement project' (Tetlock and Gardner, 2015).

The evidence for such a decline is mixed. The Edelman trust barometer claims that trust is 'in crisis', and their poll shows that a 'person like yourself' is now as credible as a technical expert, and yet their data show an overall increase in trust in government, the media, business and non-governmental organizations since 2012 (Edelman, 2017). A recent YouGov poll showed scientists trusted by 71%, although this is 63% *versus* 83% depending on whether voting to leave or remain in the European Union, and scientists come fourth in a UK trust league table behind nurses, doctors and your own general practitioner (YouGov, 2017). Levels of trust in official statistics remain high and have increased (National Centre for Social Research, 2017): of those able to give an opinion in 2016,

(a) 90% trust the Office for National Statistics,
(b) 85% trust the statistics produced by the Office for National Statistics,
(c) 78% agree that official figures are accurate,
(d) 26% agree that the UK Government presents official figures honestly and
(e) 18% agree that newspapers present official figures honestly.

These figures look reassuring to official statistics, although not to government or the media, but might improve further were prerelease access of politicians and their advisors to official statistics abolished, which is an objective of a continuing campaign by the Royal Statistical Society (RSS). (A letter on this topic was published in *The Times* during the recent election campaign, with 114 signatories (including Baroness Onora O'Neill), reflecting the 114 people with prerelease access to labour market statistics†.)

†*Note added in proof*
This campaign was successful, with prerelease access abolished from July 1st, 2017.

In her Reith lectures, philosopher Onora O'Neill pointed out the undifferentiated nature of these questions about whom we trust, whose answers perhaps reflect a more general mood of suspicion (O'Neill, 2002). More important is *active* trust, judged by our actions, which display that we commonly put our trust in institutions that we profess not to trust. Crucially, she went on to say that nobody can just expect to be trusted—they need to demonstrate *trustworthiness*.

7. Improving trustworthiness

In a remarkable Ted talk (O'Neill, 2013), Onora O'Neill argued that, rather than aiming to build trust, the task should be to become trustworthy—this means demonstrating competence, honesty and reliability. But you also have to provide usable evidence that allows others to check whether you are trustworthy, which necessitates making yourself vulnerable to the other party. Although identifying *deception* as being a key breaker of trust, she emphasized the danger of too much focus on policing deliberate fraud—this is too low a bar. This reinforces the need to avoid excessive attention to deliberate dishonesty, say through data fabrication or demonstrably fake news, because it could distract attention from the more pressing problem of misleading, incompetent and unreliable use of evidence.

There seem to be three main ways of building more trustworthiness in the statistical evidence pipeline shown in Fig. 2: change the communication structure, improve the filters for the information being passed and improve the ability of audiences to check trustworthiness.

7.1. Changing the communication structure

There are increasing possibilities to bypass potentially distorting filters. These include direct-to-public communication through social media by scientists, agencies, statistical 'experts', and even US Presidents. Although these innovations open up exciting opportunities for direct communication, there is also the risk of bypassing filters that have a positive role in weeding out poor science and statistics, and this emphasizes even more the need for audiences to be able to appraise the reliability of what is being claimed.

7.2. Improving the filters

We have already seen the proposals in Table 1 for improving the reproducibility, and hence the trustworthiness, of published science. Many of these are concerned with transparency, but O'Neill has observed that transparency does not necessarily avoid deception (O'Neill, 2002). She recommended 'intelligent transparency', which requires information to be 'accessible, intelligible, assessable, and useable' (Royal Society, 2012). The crucial element is that audiences need to be able to inquire and not just to accept assurances on trust.

Many of the ideas that are listed in Table 1 would also serve to improve trustworthiness in evidence in general, such as training of professionals, improved reporting standards, openness and protection against conflicts of interest. Other measures that could enhance the reputation of scientific and statistical expertise might include clear demonstration of the following factors.

(a) *Uncertainty*: many have recommended a greater willingness to embrace uncertainty (Makri, 2017) and to display humility (Shafik, 2017). I strongly concur, but I would add that this does not mean a reluctance to speak out confidently when faced with clear false statements or beliefs: perhaps we need a form of *muscular uncertainty*.

(b) *Engagement*: it seems essential to have empathy with audiences, and in particular understanding their beliefs and concerns. As we shall see below, this can also allow some pre-emption of misunderstandings.

(c) *Impartiality*: trustworthiness can be demonstrated by meticulous avoidance of broader agendas, so that there is a clear demarcation between the description of the evidence and any potential policy recommendations. If scientists and statisticians are seen as advocates, then they must expect their objectivity to be questioned.

This final point is especially relevant to the history of the Royal Statistical Society, whose founding principles in 1834 included the pious assertion that

> 'The Statistical Society will consider it to be the first and most essential rule of its conduct to exclude carefully all opinions from its transactions and publications—to confine its attention rigorously to facts—and, as far as it may be found possible, to facts which can be stated numerically and arranged in tables'.

This 'essential rule' was immediately ignored by Fellows who made bold recommendations on the basis of flimsy data. Even contemporary commentators commented on the ambiguity of the term 'facts' (McConway, 2016), with its implication of unquestionable veracity and authority, whereas data do not exist in a vacuum and only acquire meaning and value in a context.

This means acknowledging that numbers do not speak for themselves and so entails a responsibility to provide interpretation and potential implications of data, but without slipping into advocacy or suggesting that the evidence mandates a particular decision without taking account more general societal values. The current RSS strapline—'Data, evidence, decisions' —explicitly recognizes the role of statistical science at all stages of this path: for example the RSS's data manifesto encourages the publication of the evidence behind policies (Royal Statistical Society, 2016) and so to 'show the working'. Interpretation can be provided in a clearly separate section of a report, as practised by the Office for National Statistics.

When it comes to the specific outputs from press offices and the media, the primary aim should be to avoid the sort of questionable communication practices that are listed in Table 3. This might be helped by the following strategies:

(a) construction and adoption of reporting guidelines, such as the simple list commissioned by the Levesen Inquiry (Fox, 2012) (the RSS made a major contribution to revised BBC guidelines on reporting statistics (British Broadcasting Corporation, 2017), and the BBC is also investing in data journalism—other broadcasters might follow their example);
(b) establishing close links between statisticians and journalists, although this is not without problems (McConway, 2016), and journalism training, such as carried out by the RSS;
(c) working with dedicated organizations such as the Science Media Centre (Science Media Centre, 2017) and Sense about Science (Sense about Science, 2017);
(d) encouraging good storytelling with data, with appropriate and attractive narratives and visualization.

Although there are many exhortations to turn numbers into stories, the process does carry risks. Stories need an arc and a well-rounded conclusion, which science rarely provides, and so it is tempting to oversimplify and overclaim. We need to encourage stories that are true to the evidence: its strengths, weaknesses and its uncertainties. We need, for example, to be able to say that a drug or another medical intervention is neither good nor bad, it has benefits and harms, that people might weigh them up in different ways and quite reasonably come to different conclusions. Journalists seem to shy away from such nuanced narratives but, say by including testimony from people with differing views, a good communicator should be able to make these stories vivid.

As an apparently rare example of such a story, Christie Aschwanden from FiveThirtyEight discussed the statistics about breast screening, and then said that she had decided to avoid the

procedure, whereas her smart friend, provided with the same evidence, had made the opposite decision (Aschwanden, 2015). This neatly asserts the importance of personal values and concerns, while still respecting the statistical evidence.

But it is not enough simply to invent lists of things that could be done to improve communication—we need active research into how best to do them. For example, how can we best communicate uncertainty about facts and the future without jeopardizing trust and credibility, and how can our techniques be tailored to audiences with different attitudes and knowledge? In addition, there seems a remarkable lack of research into different ways of communicating how policy decisions are expected to impact society.

7.3. Trust as feeling and trust as analysis

The concept of a 'dual process' in psychology has been popularized by Kahneman's image of thinking fast or slow: a rapid automatic non-conscious system 1, and a more considered conscious system 2 (Kahneman, 2011). This idea has proved useful in examining different attitudes to risk: Slovic *et al*. (2004) distinguished 'risk as feelings', our immediate personal reactions to perceived threats, from 'risk as analysis', the analytic approach that is more familiar to statisticians and actuaries.

Trust might be approached similarly. When we are the recipient of a claim, trust is generally viewed as a 'feeling', a product as much of whether we warm to the 'expert' than by careful consideration of what is said: we have seen how pharma scientists suffer mistrust through broad suspicion of the industry. This is often a good heuristic, but like all heuristics it can be gamed by manipulative persuaders. In the spirit of Kahneman, we might distinguish 'fast trust' and 'slow trust', with fast trust dominated by our feelings about the topic and whether we feel that the source shares our values and has our interests at heart. Slow trust is based on the type of considered inquiry and analysis that was encouraged by O'Neill.

But is it possible to move people from 'trust as feeling' to 'trust as analysis'? Can people be 'reasoned' out of gut feelings, when they have an emotional investment in an opinion and their 'motivated reasoning' means that they are not shifted by evidence? This is not a new debate: in 1682 the English poet John Dryden optimistically claimed 'A Man is to be cheated into Passion, but to be reason'd into Truth' (Dryden, 1682), but in 1721 Jonathan Swift presented a directly opposing view: 'Reasoning will never make a Man correct an ill Opinion, which by Reasoning he never acquired' (Swift, 1843).

An active area of research focuses on whether people's demonstrably inaccurate opinions can be corrected through provision of evidence. There are many studies of the 'backfire' effect, a form of confirmation bias, which says that simply correcting 'fake news' can end up reinforcing the very belief that is being countered. However, there is increasing evidence that misconceptions can to some extent be overcome by persuasive information (Spinney, 2017), studies show it is possible to protect pre-emptively ('inoculate') public attitudes about climate change against real world misinformation (van der Linden *et al.*, 2017; Cook *et al.*, 2017) and that good visualizations can improve 'immunity to misleading anecdote' (Fagerlin *et al.*, 2005). People do not like to be deceived.

7.4. Improving the assessment of trustworthiness

There appear to be two main ways of ensuring that trustworthiness can be properly assessed: training audiences in critical appraisal, and encouraging platforms dedicated to response and 'calling-out'. Possible routes are shown in Fig. 3.

Although each of the groups of 'assessors' will have different capacities and interests, rather similar principles should apply to whoever is considering the trustworthiness of statistical ev-

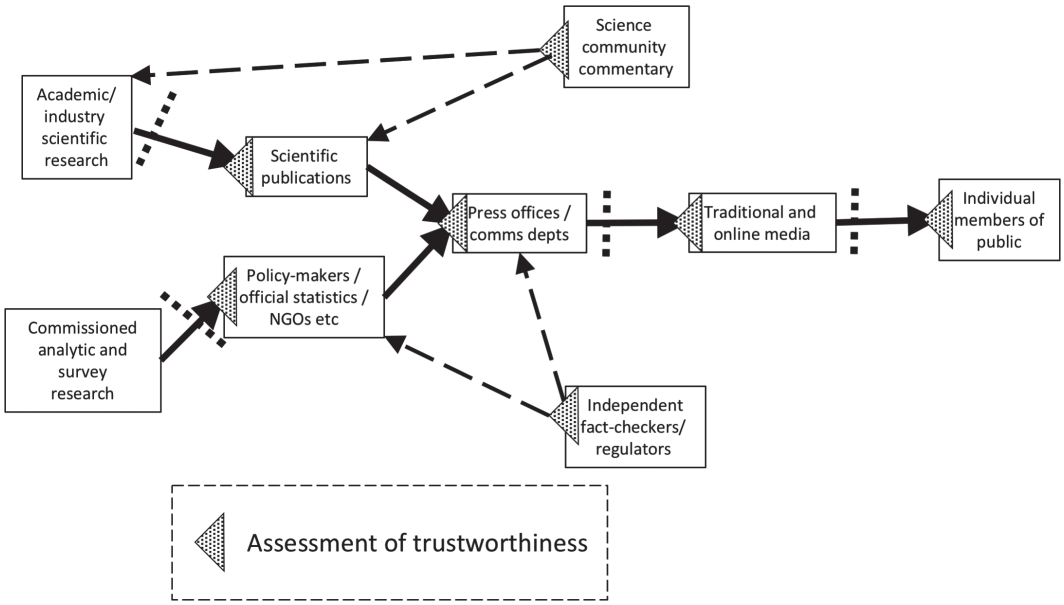

Fig. 3. Potential for the assessment of trustworthiness of statistical evidence

idence, whether it is policy professionals critiquing the impact assessments provided by their analysts, or patients confronted by information leaflets. For example, Stempra's (2017) 'Guide to being a press officer' emphasizes the need to be clear about the limitations of the study, Sense about Science (2017) have aimed directly at the public with their 'Ask for evidence' campaign and recent randomized trials in Africa have shown that families can be taught, by using comic books and audio lessons, to question claims made about medical treatments (Informed Health Choices, 2017).

Training can involve the development of teaching and assessment material in critical appraisal, provision of checklists and awareness of tricks illustrated with gripping examples that are relevant to the specific audience. I have already mentioned Facebook and Full Fact's checklist for detecting 'false news', and I am pleased that the RSS is active in creating a more general list of questions that can be adapted to specific circumstances. Three aspects of a story can be critiqued:

(a) questions about the *research*—i.e. the trustworthiness of the number itself ('internal validity');
(b) questions about the *interpretation*—i.e. the trustworthiness of the conclusions drawn (external validity);
(c) questions about the *communication*—i.e. the trustworthiness of the source and what we are being told ('spin').

Fig. 3 shows that fact checkers, blogs and official watchdogs such as the UK Statistics Authority can all publicly 'name-and-shame' bad practice in the use of statistics. In contrast, the corresponding opportunity for the scientific community to comment on publications is mainly limited to a myriad of personal blogs, because of the rather dysfunctional publication model that does not encourage even a moderated on-line discussion forum, even though the RSS has had published commentaries on papers for nearly two centuries. Short of retraction, there still seem to be few penalties for scientists indulging in questionable practices or slipping into advocacy.

The first step in good communication is to shut up and to listen, and the flow of trustworthy evidence will only improve when providers are aware that at least part of their audience is carefully monitoring the quality of what is delivered and will publicly call them out if they deviate too much from competence, honesty and reliability.

8. Conclusions

I have tried to bring together two related issues—lack of scientific reproducibility and dubious numbers in the news—by framing them both as threats to trust, which should be countered by improving trustworthiness. The pipeline through which we receive information, complete with a series of filters, applies to both contexts, and the possible measures to improve trustworthiness of what is published in both the scientific and the general media have much in common. Audiences need to be able to assess trustworthiness, and again the measures to improve their ability to do so are similar in both scientific and general media.

These are complex issues with many actors, of which the RSS is just one player, and we are fortunate that the UK has an active and collaborative ecology of organizations who are trying to improve the reliability of our science publications and the 'factfulness' of the media. Again, the RSS strapline, 'Data, evidence, decisions', has never been so pertinent, and it is a noble role to negotiate the delicate steps along that process. My personal heuristic is that statisticians are a trustworthy bunch, good and conscientious at their job, if a little nerdy. I believe that they should have a higher profile in promoting impartial evidence, and this means that at least some need to become better at converting their insights into accessible (and trustworthy) stories.

Of course this endeavour is not restricted to those who would label themselves professional statisticians and join the RSS. Hans Rosling, master statistical storyteller, was a public health physician, and there is a growing and vibrant community of people who analyse and communicate data. Those who use, and misuse, statistics come from a wide variety of backgrounds, and so the aim must be to promote the trustworthiness of statistically based claims and decisions, not just the trustworthiness of statisticians. Nevertheless, when making such an assessment it may be reasonable to take into account the professionalism of the source.

I hope that the RSS will continue to be at the forefront of both improving the trustworthiness of the numbers that are used in society, and the ability of audiences to assess that trustworthiness.

Acknowledgements

I am indebted to many for comments and encouragement on this diatribe, in particular Kevin McConway, Alex Freeman, Michael Blastland, Theresa Marteau and Iain Wilton. And I could not be working in this area at all without the unquestioning support of David Harding of Winton Capital Management.

References

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P. and Cubelli, R. (2017) Questionable research practices among Italian research psychologists. *PLOS One*, **12**, no. 3, article e0172792.

Aschwanden, C. (2015) Science won't settle the mammogram debate. *Five Thirty Eight*, Oct. 20th.

Begley, C. G. and Ioannidis, J. P. A. (2015) Reproducibility in science: improving the standard for basic and preclinical research. *Circln Res.*, **116**, 116–126.

British Broadcasting Corporation (2017) Reporting statistics. British Broadcasting Corporation, London. (Available from `http://downloads.bbc.co.uk/rmhttp/guidelines/editorialguidelines/pdfs/ReportingStatistics.pdf`.)

Cook, J., Lewandowsky, S. and Ecker, U. K. H. (2017) Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS One*, **12**, no. 5, article e0175799.
Cuddy, A. (2012) Your body language shapes who you are. (Available from http://www.ted.com/talks/amy_cuddy_your_body_language_shapes_who_you_are.)
Devlin, K. (2009) Nine in 10 people carry gene which increases chance of high blood pressure. *Telegraph*, Feb. 15th.
Donnelly, L. (2016) Why binge watching your TV box-sets could kill you. *Telegraph*, July 25th.
Dryden, J. (1682) Religio laici, or, A laymans faith a poem. (Available from http://name.umdl.umich.edu/A36673.0001.001.)
Edelman (2017) Trust barometer. (Available from http://www.edelman.com/trust2017/.)
Fagerlin, A., Wang, C. and Ubel, P. A. (2005) Reducing the influence of anecdotal reasoning on people's health care decisions: is a picture worth a thousand statistics? *Med. Decsn Makng*, **25**, 398–405.
Fanelli, D. (2009) How many scientists fabricate and falsify research?: a systematic review and meta-analysis of survey data. *PLOS One*, **4**, no. 5, article e5738.
Fox, F. (2012) 10 best practice guidelines for reporting science & health stories. (Available from http://webarchive.nationalarchives.gov.uk/20140122145147/http:/www.levesoninquiry.org.uk/wp-content/uploads/2012/07/Second-Submission-to-inquiry-Guidelines-for-Science-and-Health-Reporting.pdf.)
Full Fact (2017) Full Fact is the UK's independent factchecking organisation. Full Fact, London.
Gelman, A. (2013) Difficulties in making inferences about scientific truth from distributions of published p-values. (Available from http://andrewgelman.com/2013/09/26/difficulties-in-making-inferences-about-scientific-truth-from-distributions-of-published-p-values/.)
Gelman, A., and Loken, E. (2014) The statistical crisis in science. *Am. Scient.*, **102**, 460–465.
Gelman, A. and Stern, H. (2006) The difference between "significant" and "not significant" is not itself statistically significant. *Am. Statistn*, **60**, 328–331.
Gregory, A. (2017) Why going to university increases risk of getting a brain tumour. *Mirror Online*, June 20th.
Informed Health Choices (2017) Using evidence to change the world. (Available from http://www.informedhealthchoices.org/.)
Ioannidis, J. (2014). Discussion: Why 'An estimate of the science-wise false discovery rate and application to the top medical literature' is false. *Biostatistics*, **15**, 28–36.
Ioannidis, J. P. A. (2005) Why most published research findings are false. *PLOS Med.*, **2**, no. 8, article e124.
Jager, L. R. and Leek, J. T. (2014) An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, **15**, 1–12.
John, L. K., Loewenstein, G. and Prelec, D. (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.*, **23**, 524–532.
Kahneman, D. (2011) *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.
Khanolkar, A. R., Ljung, R., Talbäck, M., Brooke, H. L., Carlsson, S., Mathiesen, T. and Feychting, M. (2016) Socioeconomic position and the risk of brain tumour: a Swedish national population-based cohort study. *J. Epidem. Commty Hlth*, **70**, 1222–1228.
Leek, J. T. and Jager, L. R. (2017) Is most published research really false? *A. Rev. Statist. Appl.*, **4**, 109–122.
van der Linden, S., Leiserowitz, A., Rosenthal, S. and Maibach, E. (2017) Inoculating the public against misinformation about climate change. *Globl Chall.*, **1**, no. 2, article 1600008.
Makri, A. (2017) Give the public the tools to trust scientists. *Nat. News*, **541**, 261.
Matthews, R., Wasserstein, R. and Spiegelhalter, D. (2017) The ASA's *p*-value statement, one year on. *Significance*, **14**, 38–41.
McConway, K. (2016) Statistics and the media: a statistician's view. *Journalism*, **17**, 49–65.
Medical Xpress (2016) High levels of education linked to heightened brain tumor risk. (Available from https://medicalxpress.com/news/2016-06-high-linked-heightened-brain-tumor.html.)
Mervis, J. (2017) Data check: NSF sends Congress a garbled message on misconduct numbers. *Science*, Mar. 24th.
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. and Ioannidis, J. P. A. (2017) A manifesto for reproducible science. *Nat. Hum. Behav.,* **1**, no. 1, article 0021.
National Centre for Social Research (2017) Public confidence in official statistics. National Centre for Social Research, London. (Available from http://natcen.ac.uk/our-research/research/public-confidence-in-official-statistics/.)
Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M. D., Bochud, M., Coin, L., Najjar, S. S., Zhao, J. H., Heath, S. C., Eyheramendy, S., Papadakis, K., Voight, B. F., Scott, L. J., Zhang, F., Farrall, M., Tanaka, T., Wallace, C., Chambers, J. C., Khaw, K. T., Nilsson, P., van der Harst, P., Polidoro, S., Grobbee, D. E., Onland-Moret, N. C., Bots, M. L., Wain, L. V., Elliott, K. S., Teumer, A., Luan, J., Lucas, G., Kuusisto, J., Burton, P. R., Hadley, D., McArdle, W. L., Wellcome Trust Case Control Consortium, Brown, M., Dominiczak, A., Newhouse, S. J., Samani, N. J., Webster, J., Zeggini, E., Beckmann, J. S., Bergmann, S., Lim, N., Song, K., Vollenweider, P., Waeber, G., Waterworth, D. M., Yuan, X., Groop, L., Orho-Melander, M., Allione, A., Di Gregorio, A., Guarrera, S., Panico, S., Ricceri, F., Romanazzi, V., Sacerdote, C., Vineis, P., Barroso, I., Sandhu, M. S., Luben,

R. N., Crawford, G. J., Jousilahti, P., Perola, M., Boehnke, M., Bonnycastle, L. L., Collins, F. S., Jackson, A. U., Mohlke, K. L., Stringham, H. M., Valle, T. T., Willer, C. J., Bergman, R. N., Morken, M. A., Döring, A., Gieger, C., Illig, T., Meitinger, T., Org, E., Pfeufer, A., Wichmann, H. E., Kathiresan, S., Marrugat, J., O'Donnell, C. J., Schwartz, S. M., Siscovick, D. S., Subirana, I., Freimer, N. B., Hartikainen, A. L., McCarthy, M. I., O'Reilly, P. F., Peltonen, L., Pouta, A., de Jong, P. E., Snieder, H., van Gilst, W. H., Clarke, R., Goel, A., Hamsten, A., Peden, J. F., Seedorf, U., Syvänen, A. C., Tognoni, G., Lakatta, E. G., Sanna, S., Scheet, P., Schlessinger, D., Scuteri, A., Dörr, M., Ernst, F., Felix, S. B., Homuth, G., Lorbeer, R., Reffelmann, T., Rettig, R., Völker, U., Galan, P., Gut, I. G., Hercberg, S., Lathrop, G. M., Zelenika, D., Deloukas, P., Soranzo, N., Williams, F. M., Zhai, G., Salomaa, V., Laakso, M., Elosua, R., Forouhi, N. G., Völzke, H., Uiterwaal, C. S., van der Schouw, Y. T., Numans, M. E., Matullo, G., Navis, G., Berglund, G., Bingham, S. A., Kooner, J. S., Connell, J. M., Bandinelli, S., Ferrucci, L., Watkins, H., Spector, T. D., Tuomilehto, J., Altshuler, D., Strachan, D. P., Laan, M., Meneton, P., Wareham, N. J., Uda, M., Jarvelin, M. R., Mooser, V., Melander, O., Loos, R. J., Elliott, P., Abecasis, G. R., Caulfield, M. and Munroe, P. B. (2009) Genome-wide association study identifies eight loci associated with blood pressure. *Nat. Genet.*, **41**, 666–676.

O'Neill, O. (2002) *A Question of Trust: the BBC Reith Lectures 2002*. Cambridge: Cambridge University Press.

O'Neill, O. (2013) What we don't understand about trust. (Available from `https://www.ted.com/talks/onora_o_neill_what_we_don_t_understand_about_trust/transcript?language=en`.)

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science*, **349**, article. 4716.

Patil, P., Peng, R. D. and Leek, J. T. (2016) What should researchers expect when they replicate studies?: a statistical view of replicability in psychological science. *Perspect. Psychol. Sci.*, **11**, 539–544.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S. and Weber, R. A. (2015) Assessing the robustness of power posing: no effect on hormones and risk tolerance in a large sample of men and women. *Psychol. Sci.*, **26**, 653–656.

Royal Society (2012) Science as an open enterprise. *Report*. Royal Society, London. (Available from `https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/`.)

Royal Statistical Society (2016) Data manifesto. Royal Statistical Society, London. (Available from `http://www.rss.org.uk/RSS/Influencing_Change/Data_manifesto/RSS/Influencing_Change/Data_democracy_sub/Data_manifesto.aspx?hkey=5dd70207-82e7-4166-93fd-bcf9a2a1e496`.)

Science Media Centre (2017) Where science meets the headlines. Science Media Centre, London.

Sense about Science (2017) Because evidence matters. Sense about Science, London.

Shafik, M. (2017) In experts we trust? Bank of England, London. (Available from `http://www.bankofengland.co.uk/publications/Documents/speeches/2017/speech964.pdf`.)

Shirakawa, T., Iso, H., Yamagishi, K., Yatsuya, H., Tanabe, N., Ikehara, S., Ukawa, S. and Tamakoshi, A. (2016) Watching television and risk of mortality from pulmonary embolism among Japanese men and women: the JACC study (Japan Collaborative Cohort). *Circulation*, **134**, 355–357.

Simmons, J. P., Nelson, L. D. and Simonsohn, U. (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.*, **22**, 1359–1366.

Slovic, P., Finucane, M. L., Peters, E. and MacGregor, D. G. (2004) Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. *Risk Anal.*, **24**, 311–322.

Spinney, L. (2017) How Facebook, fake news and friends are warping your memory. *Nat. News*, **543**, 168.

Stempra (2017) Guide to being a press officer. Stempra. (Available from `https://stempra.org.uk/wp-content/themes/stempra/downloads/2017_stempra_guide_to_being_a_media_officer.pdf`.)

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Bott, L., Adams, R., Venetis, C. A., Whelan, L., Hughes, B. and Chambers, C. D. (2016) Exaggerations and caveats in press releases and health-related science news. *PLOS One*, **11**, no 12, article e0168217.

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Venetis, C. A., Davies, A., Ogden, J., Whelan, L., Hughes, B., Dalton, B., Boy, F. and Chambers, C. D. (2014) The association between exaggeration in health related science news and academic press releases: retrospective observational study. *Br. Med. J.*, **349**, article 7015.

Swift, J. (1843) *The Works of Jonathan Swift ...: Containing Interesting and Valuable Papers, not hitherto Published ... with Memoir of the Author*. London: Bohn.

Szucs, D. and Ioannidis, J. P. A. (2017) Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biol.*, **15**, no.3, article e2000797.

Tetlock, P. E. and Gardner, D. (2015) *Superforecasting: the Art and Science of Prediction*. Toronto: McClelland and Stewart.

Wasserstein, R. L. and Lazar, N. A. (2016) The ASA's statement on *p*-values: context, process, and purpose. *Am. Statistn*, **70**, 129–133.

Wellcome Trust (2017) Public views on medical research. Wellcome Trust, London. (Available from `https://wellcome.ac.uk/what-we-do/our-work/public-views-medical-research`.)

YouGov (2017) Leave voters are less likely to trust any experts—even weather forecasters. YouGov, London. (Available from `//yougov.co.uk/news/2017/02/17/leave-voters-are-less-likely-trust-any-experts-eve/`.)

Youtube (2016) Gove: Britons "Have had enough of experts". (Available from `http://www.youtube.com/watch?v=GGgiGtJk7MA`.)

Vote of thanks

Peter J. Diggle (*Lancaster University*)
Our President, Sir David Spiegelhalter, has chosen a topic that is central to our discipline and of enormous contemporary importance, both for science and for society at large. It is also one on which he can speak with great authority. Unsurprisingly, and not just out of respect for historical tradition, I find myself in agreement with almost everything that he says.

When applied to numbers, the notion of 'trust' can have many meanings. The first requirement for trust is that the numbers constituting a data set can be assumed to have been correctly recorded, which is already a big thing to ask in some fields of research even when there is no intent to deceive. See, for example, Baggerly and Coombes (2009) in the context of high throughput biology. One step along from this is that the numbers that are produced by the software used to analyse a set of data are technically correct. Beyond that, the question of trust becomes rather subtle. Have the data been analysed by using an appropriate statistical method? Two equally reputable statisticians can easily disagree on the answer to that question, so whose answer should the scientist who generated the data trust? Finally, even if the data have been checked meticulously, the software has been tested thoroughly, statisticians agree on the methodology and scientists agree on the findings, if those findings are used to justify a public policy decision, can we trust that this is done in good faith?

David says that he likes p-values. So do I, and I suspect that most statisticians use them routinely, even if they do not quote them in their final published analyses. Also like most statisticians, I am less enamoured of using the conventional 5% as a rigid definition of statistical significance, especially when coupled with the widespread practice of basing a sample size calculation on a requirement to achieve 80% power to detect a clinically significant difference between two treatments. In my first meeting of the Medical Research Council's then newly established Population and Systems Medicine Board, I asked of one grant proposal that we were evaluating, whose experimental programme involved a two-treatment comparison, 'Why are they using six mice?', to which I received the reply 'We always use six mice'. This answer came not from a naive research tyro but from another member of the Board, a distinguished biomedical research scientist. The answer puzzled me. The only explanation I can offer is that if the data are to be analysed by using a two-sample t-test and the clinically significant difference is defined to be 1.96 standard deviations of an individual measurement, six mice per group is the smallest number that meets the requirement for 80% power when testing at the 5% level of significance. Quite why one *should* define the clinically significant difference in this way completely escapes me. I also do not understand how such an experiment can be said to have reached a positive conclusion if the treatment effect is significant, when at the same time the lower limit of a 95% confidence interval (and, presumably, at least some version of a 95% posterior interval) for the effect size can be arbitrarily close to 0.

More broadly than this, and echoing David's call for a stronger focus on effect sizes, I would like to see power calculations replaced by sample size calculations that set out how precision of estimation varies with sample size over a range of plausible values of variance components and other unknown parameters that affect precision, e.g. serial correlation in longitudinal studies. This is surely both more useful and more realistic than quoting a single sample size that achieves an arbitrary power against what is often an equally arbitrary alternative.

It may still be true that 'Figures do not lie, but liars figure'. However, it needs no cynicism to acknowledge that in the presence of uncertainty even the most honest and unbiased person can reach a false conclusion. I think that one of the most difficult but important of current challenges for a society awash with data is to understand fully how decision makers handle uncertainty... because I do not think that they use decision theory, however much some of our colleagues might wish otherwise. In this context, I find encouragement in a recent editorial piece in the *British Medical Journal* (Macdonald, 2017) in which Helen Macdonald advocates that 'Doctors need to be able to discuss degrees of certainty, including numbers, with their patients'. Both in his address tonight and more generally in his superb work as the Winton Professor for the Public Understanding of Risk, David shows that he is very much a President for our times. It gives me very great pleasure indeed to propose the vote of thanks.

John Pullinger (*Office for National Statistics, Newport*) (© Crown copyright 2017)
I am delighted to have this opportunity to congratulate David on his penetrating analysis of a critical question for our discipline. Society can have available to it the best statistical evidence that it is scientifically possible to generate but if there is a lack of trust in numbers those findings will not make their mark. Without trust in numbers we are doomed to make choices, including fundamental choices about the future for our

families, our communities and our countries, in ignorance or, even worse, with a wilfully inaccurate world view in mind.

David unpacks two central responsibilities for us as a statistical community: first, among ourselves to be ruthlessly professional and to understand the limits of our own cleverness—we must get the supply side right—second, to help others to derive value from statistics—celebrating what is good and not standing for what is unacceptable. We have duty on the demand side also.

The assessment that we have heard of reproducibility and replication is both elegant and devastating. It should be etched into the practice of every statistician, indeed every scientist. It should certainly be etched into the practice of every journal editor and peer reviewer. The questionable research practices that David names should never find a place in respectable public places—no more sampling things that are convenient rather than appropriate; no more leading questions or misleading wording; no more failure properly to adjust for confounders; no more samples that are too small; no more inappropriate assumptions; no more inapt statistical analysis.

David reserves a particular scorn for use of *p*-values. His main lesson is that we should stop thinking in terms of significant or not significant as determining a discovery and instead focus on effect sizes: hear, hear. How difficult can it be? As he says, *p*-values do what they say on the tin, but people do not read the tin. His conclusion, drawing on the work of Munafò *et al*. (2017), highlights the role that statistical science can play through methodological training, improving reporting and peer review and sharing of data for re-analysis. There is a call to action for the Society here that I readily endorse.

So much for our role on the supply side: moving to the demand side is where David really gets going with his questionable interpretation and communication practices. Ridiculous examples always bring a smile in a statistical audience. Many really are hilarious but are also tragic. David's dirty dozen questionable interpretation and communication practices are both pervasive and corrosive to the effective functioning of a democratic society. We must all care about this. It matters.

David's solution is focused on improving trustworthiness. He identifies three main ways of building more trustworthiness in the statistical evidence pipeline: change the communication structure, improve the filters for the information being passed and improve the ability of audiences to check trustworthiness. He sees potential in social media to bypass potentially distorting filters but also recognizes that this could cut the other way as well. Social media can provide a direct way for us to receive both new insight and fake news.

He sees transparency as a way to improve the filters but again recognizes Onora O'Neill's caution that transparency does not necessarily avoid deception. A notorious crime is often one committed in plain sight.

He looks deeper then to other measures and makes three proposals for action: a greater willingness to embrace uncertainty, the importance of engagement and empathy with audiences and the meticulous avoidance of broader agendas to give assurance on impartiality. His definition of impartiality is one which I really like. I hope that he will amplify yet further on future occasions. Statistical impartiality should be a clear demarcation between the description of the evidence and any potential policy recommendations. If scientists and statisticians are seen as advocates then they must expect their objectivity to be questioned.

The final section of David's address deals with training and with naming and shaming when bad practice occurs. There are powerful levers here. As he says, people do not like to be deceived.

I support the assessment that he has made but I would have liked to have seen even more here. The data revolution is providing ever more opportunities for statistics. The potential for good and bad use of numbers is growing commensurately. As a community we need to step up and to take on the challenge. This is indeed a timely debate and the President of the Society is uniquely placed to bring it to wider public attention.

I would urge him not to stop his efforts with this address. He can build on the tradition of the Society since its creation as well as more recent endeavours such as the Getstats campaign and awards for statistical journalism. He can be confident of my support and I am sure of all of us in this room and the membership of the Society.

It gives me great pleasure to second the vote of thanks.

The vote of thanks was passed by acclamation.

References

Baggerly, K. A. and Coombes, K. R. (2009) Deriving chemo-sensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. *Ann. Appl. Statist.*, **3**, 1309–1334.

Macdonald, H. (2017) Navigating uncertainty. *Br. Med. J.*, **357**, article j2524.

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. and Ioannidis, J. P. A. (2017) A manifesto for reproducible science. *Nat. Hum. Behav.*, **1** no. 1, article 0021.