Written evidence submitted by The Royal Statistical Society (RIN0085)

1. Introduction

1.1. The Royal Statistical Society (RSS) is a membership body for statisticians and data analysts, and is a charity with six key strategic goals, one of which is to promote and support the strength of statistics as an academic and research discipline.

1.2. The Royal Statistical Society has a strong tradition of contributing to aspects of research integrity, including objectivity, competencies, protocols and study-design, and statistical power in respect of a plausible effect size. These have been covered in reports that include *Official Statistics: Counting with Confidence; Statistics and Statisticians in Drug Regulation in the United Kingdom; Performance Indicators: Good, Bad and Ugly; Statistical Issues in First-in-Man Studies;* and *Data Capture - for the Public Good* [1]. RSS Fellows and Honorary Fellows have sought to assure research integrity, from statistical reporting standards for contributors to medical journals in the early 1980s through to today's suite of checklists for the competent reporting on a range of study-designs - ranging from randomised controlled trials (CONSORT), through observational studies (STROBE) to meta-analyses (Cochrane Collaboration and Risk of Bias).

1.3. Statistical science is an essential element of scientific discovery and progress. Our discipline is explicitly concerned with the evaluation of evidence, and so is crucially relevant to issues of the reliability of findings and problems of reproducibility.

1.4. As the POST report [2] makes clear, lack of scientific integrity runs from deliberate fraud to poor practice. Although statistical science can certainly help in the detection of fraud, in this submission we focus on improving the practice and publication of science.

Section 1 references:

[1] Moore, P. *et al.* (1991) 'Official statistics: counting with confidence', *J. R. Statist. Soc. A.* 154(1): 23-44, available at: https://www.jstor.org/stable/2982692

Moore, P. (1991) 'Statistics and statisticians in drug regulation in the United Kingdom', *J. R. Statist. Soc. A.* 154(3): 413-419, available at: https://www.jstor.org/stable/2983151

RSS (2003) *Performance indicators: good, bad and ugly* [PDF], available from: http://www.rss.org.uk/Images/PDF/publications/rss-reports-performance-monitoring-public-services-2003.pdf

RSS (2007) *Statistical issues in first-in-man studies* [PDF], available from: http://www.rss.org.uk/Images/PDF/publications/rss-reports-statistical-issues-first-in-man-studies-2007.pdf

Bird, S. (2011) 'Data capture for the public good: a matter of trust, or of science and public understanding?' [PDF], *Int. Statistical Inst.: Proc. 58th World Statistical Congress, 2011, Dublin (Session IPS119)*, available from: http://2011.isiproceedings.org/papers/450425.pdf

[2] Bunn, S. & Auckland, C. (2017) 'Integrity in Research' [PDF], *POSTNOTE 544*. London: The Parliamentary Office of Science and Technology. Available from: http://researchbriefings.parliament.uk/ResearchBriefing/Summary/POST-PN-0544

2. Recommendations

2.1. To help improve scientific integrity and improve reproducibility, we believe that the UK's system for research and science funding needs to support more skilled statistical instructors who work across disciplines. Mechanisms to address statistical integrity are most advanced for medicine and clinical trials, but models developed there should be applied more widely to other fields of research.

2.2. The RSS endorses the recommendations made in 2016 by the Interacademy Partnership for Health in its *Call for Action to Improve the Reproducibility of Biomedical Research* (see Appendix 1).

2.3. The RSS emphasises the importance of written, peer-reviewed and ethically approved study protocols, and their preregistration, for both experimental and non-experimental studies.

2.4. The RSS notes that, in general, there are no protocol obligations on those who conduct secondary analyses, other than those imposed by their professional codes of conduct (see here[3] for RSS's code of professional conduct). To support the ethics and integrity of data science, RSS has called for a national Council for Data Ethics [4] which should engage in depth with the public's understanding of, and tolerance for, new developments in data science.

2.5 To support research integrity and integrity in official statistics, the RSS has called for legislation in England, Wales & Northern Ireland to require the timely registration of fact-of-death, within 8 days of death having been ascertained, as in Scotland [5].

2.6. The RSS advises that an updated, explanatory check-list on the reporting of statistical methods in peer-reviewed journals is needed, to complement the equator-suite on study-designs. This should include a clear distinction between exploratory and confirmatory significance tests.

2.7. The RSS encourages editors affiliated to the Committee on Publication Ethics to promote peer-reviewed publication of replication studies, especially those which adequately test record linkage "discoveries".

Section 2 references:

[3] Royal Statistical Society (2014) *Code of Conduct* [PDF], available from: http://www.rss.org.uk/Images/PDF/join-us/RSS-A5-Code-of-Conduct-2014.pdf

[4] See: 'RSS welcomes science committee's 'Big Data Dilemma' report', *StatsLife*, 17 February 2016 https://www.statslife.org.uk/news/2685-rss-welcomes-science-committee-s-big-data-dilemma-report; and RSS (2016) *The Opportunities and Ethics of Big Data* (PDF), available from: http://www.rss.org.uk/Images/PDF/influencing-change/2016/rss-report-opps-and-ethics-of-big-data-feb-2016.pdf

[5] See Appendix in RSS (2016) *Response to Department of Health consultation on reforms to Death Certification in England and Wales* [PDF], available from: http://www.rss.org.uk/Images/PDF/influencing-change/2016/RSS-response-to-DH-consultation-on-reforms-to-Death-Certification-June-2016.pdf

3. Scientific integrity, reproducibility and statistics

3.1 The POST report identifies many concerns about scientific integrity, and specifically mentions the need for appropriate statistical analysis. But statistical science goes well beyond simply the calculations done on data, and in particular encompasses appropriate study design and reporting. We therefore recommend: **To help improve scientific integrity and improve reproducibility, we believe that the UK's system for research and science funding needs to support more skilled statistical instructors who work across disciplines. Mechanisms to address**

statistical integrity are most advanced for medicine and clinical trials, but models developed there should be applied more widely to other fields of research.

3.2. As the Interacademy Partnership for Health has made clear, there is no single cause of the problems concerning reproducibility [6]. However statistical issues are a dominant feature, which include

- Lack of pre-specification of design and analysis, allowing researchers freedom to 'tweak' their analysis and reporting to enhance the impact of their results
- Selective reporting of a few findings drawn from a large number of exploratory analyses. This is the phenomenon of 'p-hacking' - selecting what to report on the basis of 'statistical significance' in order to find an impressive result, omitting to say that there had been 'multiple bites at the cherry'
- Publication bias in favour of positive studies

3.3. **The RSS endorses the recommendations made in 2016 by the Interacademy Partnership for Health in its *Call for Action to Improve the Reproducibility of Biomedical Research* (see Appendix 1).**

Section 3 references:

[6] Interacademy Partnership for Health (2016) *A call for action to improve the reproducibility of biomedical research* [PDF], London: Academy for Medical Sciences. Available from: https://acmedsci.ac.uk/file-download/41599-57f7204459be7.pdf

4. The conduct of science: protocols and preregistration

4.1. Experiments on human subjects are, of course, regulated, and registration of studies is deemed necessary. Yet despite such protection against bad practice, problems still exist. Drug regulatory agencies, such as the Medicine and Health Care Products Regulatory Agency (MHRA), the Food and Drug Administration Agency (FDA) and the European Medicines Agency (EMA) require pre-specification of statistical analysis and this principle is enshrined in an International Conference on Harmonisation (ICH) guideline [7]. The AllTrials campaign (http://www.alltrials.net/) is striving to improve transparency in this area. The Cochrane Collaboration (http://www.cochrane.org/) which systematically examines evidence about health care interventions has experience of the shortcomings exhibited by some clinical trials. The Cochrane Handbook for systematic reviews of interventions (http://handbook.cochrane.org/) discusses the issue of publication bias.

4.2. In multi-centre studies, there will usually be a study protocol which explains not only why the study is being conducted (its rationale) but also what is to be done and where, when, how precisely, and by whom. Detailed methods include whether and how the **principle of randomisation** *(see Performance Indicators: Good, Bad and Ugly)* has been applied in treatment assignment or in sampling. The study protocol also includes a statistical analysis plan which sets out formally the study's primary and secondary outcomes; the a priori plausible effect size, as hypothesised; any stratification factors to be taken into account at randomisation

or analysis; and the study's statistical power (or precision) in discerning the effect sizes of prior interest.

4.3. Randomised controlled trials (RCT) are a special case because preregistration of RCTs, usually around the time of ethical approval, is a prerequisite for their subsequent publication in peer-reviewed journals, which includes citation of the RCT's assigned ISRCT Number.

4.4. The RCT's approved protocol is made available to peer-publication-reviewers, who are expected to compare methods, outcomes and statistical analysis plan as described by the protocol and in the submitted paper. This comparison helps to guarantee against selective reporting or emphasis that is data-inspired yet has not been acknowledged as such.

4.5. More generally, interested scientists can use the trial's acronym or ISRCT Number to access the approved protocol from the research-team during the trial's conduct (**open protocol principle**, see *Statistical Issues in First-in-Man Studies*) or simply to look up its registration details which typically include: rationale; intervention and control groups; primary outcome; secondary outcome(s); stratification or minimization factors; target effect size, numbers to be randomised and associated statistical power.

4.6. A study protocol is essential for RCTs (whether single-centre or multi-centre); usual for record linkage studies that are without consent, as these need independent approvals; recommended by RSS for performance indicators; and generally required for peer review by funders of research cohorts or for consented epidemiological studies on human subjects.

4.7. Major studies such as UK Biobank have a properly constituted committee to receive and peer review protocols from external research teams seeking approved for their novel research proposal to access the data and/or biological resources provided by Biobank's participants. The responsibility of such a committee, besides for research integrity, is also to ensure that the access sought in de novo protocols is consistent with the permissions given by the study-volunteers.

4.8. There are no study protocol obligations on secondary analysts of Open Access datasets and so there are no guarantees against their data dredging rather than their testing of a priori specified hypotheses. Nor are there ready guarantees against false discoveries.

4.9. Reproducibility is even more important for "discoveries" from record linkage studies. More or less subtle biases in how administrative and other non-research datasets were collected or overwritten (without the overwriting being date-stamped) may influence the potential for linkage-across datasets and may distort the reliability of data fields within datasets. Some such biases may remain undetected by the research team because of their limited access to source data and because the linkage routines have been programmed by others.

4.10. For such reasons as above, "discoveries" from record linkage studies require validation, for example by repeating a record linkage study for a different era but within the same jurisdiction; or in a different jurisdiction; or by adopting a different study method to test the inferences which follow from the initial "discovery".

4.11. Exploratory data analysis has value, but its exploratory nature needs to be acknowledged, not hidden. Empirical findings from exploratory data analysis, if vindicated in other settings or by formal experimentation, can lead to international endorsement and global change (such as the World Health Organization's change in 2001 to a low osmolarity oral rehydration solution).

4.12. The RSS emphasises the importance of written, peer-reviewed and ethically approved study protocols, and their preregistration, for both experimental and non-experimental studies.

Section 4 references

[7] International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) (1998) *Statistical Principles for Clinical Trials E9*, 5 February 1998, available from: http://www.ich.org/products/guidelines/efficacy/efficacy-single/article/statistical-principles-for-clinical-trials.html

5. The reporting of science

5.1 Accurate reporting is essential for scientific integrity. Progress has been made in trying to improve scientific reporting in healthcare by producing reporting guidelines through the 'Enhancing the Quality and Transparency of health Research' initiative (www.equator-network.org). These lessons need to be transferred across to other disciplines so that what has occurred in any study becomes absolutely clear to the reader of any research report.

5.2 The RSS advises that an updated, explanatory checklist on the reporting of statistical methods in peer-reviewed journals is needed, to complement the equator suite on study designs. This should include a clear distinction between exploratory and confirmatory significance tests. RSS Fellows are well placed to contribute to the development and tailoring of such checklists for specific scientific disciplines.

5.3. The RSS recognises that statistical reporting standards for contributors to scientific journals should be updated to be user friendly and apposite, including for disciplines beyond healthcare. Of statistics Myron Tribus said*: "If Experimentation be the Queen of Sciences, then Statistics is Guardian of the Royal Virtue".* The overly strict word limits which some journals impose on authors represent a real risk for authors that detailed description of their methods is short circuited to the detriment of reproducibility.

5.4. The RSS encourages editors affiliated to the Committee on Publication Ethics to promote peer-reviewed publication of replication studies especially those which adequately test record linkage "discoveries".

5.5. The *AllTrials* campaign expects that all RCTs should be published within one-year of completing the last scheduled follow-up for the last-randomised patient. The AllTrials expectation allows neither for publication delays nor for the late registration of inquest deaths in England, Wales and Northern Ireland where research teams may have to allow for two-year delays to registration. For example, to be almost sure of ascertaining information on all drug

related deaths or suicides within 12 weeks of release from prison in England, the N-ALIVE pilot trial applied a delay of more than 12 months [8]. The RSS has published 10 reasons against the late registration of deaths in England, Wales and Northern Ireland [9]. To support research integrity, including in official statistics, **the RSS calls for legislation in England, Wales & Northern Ireland to require the timely registration of fact-of-death, within 8 days of death having been ascertained, as in Scotland** [10].

5.6. The RSS recognises that calculation errors can usually be rectified, whereas errors in study-design may be irremediable and seriously compromise inferences. Statistical analysis plans, as set out in study protocols, aim to avert other transgressions against sound statistical thinking. Such transgressions may be occasioned by intent, incompetence or naivety. No amount of training can eradicate intended transgressions but a robust scientific culture may help to identify perpetrators. Appendix 2, which is not exhaustive, lists a dozen transgressions against sound statistical thinking.

Section 5 references

[8] Mahesh, K. Parmar, B. Strang, J. Choo, L. Meade, A.M. & Bird, S.M. (2017) Randomized controlled pilot trial of naloxone-on-release to prevent post-prison opioid overdose deaths, *Addiction 112*(3): 502–515, available at http://onlinelibrary.wiley.com/doi/10.1111/add.13668

[9] See Appendix in RSS (2016) *Response to Department of Health consultation on reforms to Death Certification in England and Wales* [PDF], available from: http://www.rss.org.uk/Images/PDF/influencing-change/2016/RSS-response-to-DH-consultation-on-reforms-to-Death-Certification-June-2016.pdf

[10] See Bird, S.M. (2013) 'Editorial: Counting the dead properly and promptly', *J. R. Statist. Soc. A*, available at: http://www.mrc-bsu.cam.ac.uk/wp-content/uploads/Editorial-on-late-registration-of-deaths-JRSSA.pdf

6. Transparency

6.1. In an ideal world, open access to data would allow external validation of any claim. But the downside of open access is its abrogation of the protections that prior approval and registration of study protocols affords, and could lead to ill-founded disputes in areas of contested science.

6.2. Program access is essential for full audit but is underemphasised in research governance, perhaps because much data checking (e.g. of date sequences in linked datasets), with back-and-forth data queries, may be required before a database passes logical checks to begin formal analysis.

6.3. External requirements imposed on scientists, if they are not also of direct benefit to the conduct of science, represent a cost on scientists' time – and one that is not repaid in how robustly research is conducted. Good practices which enhance the conduct of science are adopted by good scientists and should be embedded in their scientific culture. Documentation of good practice, if it is different from the documentation that scientists need for their own scientific audit, may not be cost efficient.

6.4. Study protocols and applications for research funding are typically research-in-confidence, but may be open access after approval. There is a risk in immediate openness, as research priority may be ceded if interested other parties, in effect, copy the application without necessarily having

the competency to put it into practice. However, immediate open access allows international colleagues to embark on replication studies which will typically be needed – for example, for even pre-hypothesised record linkage discoveries based on administrative data.

March 2017

Appendix 1

Recommendations, p. 3 in Interacademy Partnership for Health (2016) *A call for action to improve the reproducibility of biomedical research* [PDF], London: Academy for Medical Sciences. Available from: https://acmedsci.ac.uk/file-download/41599-57f7204459be7.pdf

"In signing this statement, IAP for Health member academies recognise that:

• It is critically important for the progress of science that the reproducibility of research is optimal. Where policies to improve national and global health are concerned, they must be based on the best available evidence – the value of research and the efficient use of resources can only be maximised through the most robust science.

• There is no single cause of irreproducibility and a number of measures are required to address it. These measures will rely on multiple actions from many stakeholders. For example:

- Universities and research institutions should embrace a culture change that rewards robust methods as much as novel findings, particularly when making decisions about career progression. Institutions should encourage the use of quality-enhancing infrastructures (e.g. electronic laboratory notebooks, quality assessment systems), as well as expert advice (e.g. in biostatistics).
- Funders should use their position at the start of the research process to set the tone for reproducible research, for example by rigorously assessing experimental design to minimise bias and improve statistical power.
- Publishers and journal editors should enable greater openness and transparency in methods, results and data; and be willing to publish replications and neutral or negative ('null') results from adequately powered studies. They should take steps to ensure that peer review focuses on the quality of the science rather than the excitement generated by the results. This may include measures to reduce the potential for bias, for example by implementing blinded peer review in which reviewers do not know the names or affiliations of authors.
- Researchers should take responsibility for portraying their results accurately, alongside science communicators where relevant, and engage in open communication and dialogue around replication attempts.

At country level, IAP for Health member academies should consider this issue within their own leaderships to establish the most effective role they can play in efforts to improve reproducibility, including by:

• Raising awareness about the challenge of irreproducibility and the possible causes – initially among their elected Fellows, who have an important leadership role to play; but then extending to the broader biomedical research community, including early career researchers, and wider society.

• Meeting national stakeholders to raise awareness and discuss measures that should be taken to improve research practice. These will include leaders within research funding agencies, publishers, institutions and professional bodies. Where possible, IAP for Health member academies will look to coordinate discussions among these stakeholders as well.

• Promoting the importance of an environment and culture for research that values the robustness of studies as much as their originality.

• Working to ensure that the biomedical research community is engaged in discussions as solutions are developed and implemented.

• Supporting education and training around optimal standards of research design and integrity. Science is a global endeavour and reproducibility presents a global challenge, which must be addressed through collaboration and cooperation. Therefore, at a regional and global level:

• IAP for Health member academies, including regional networks of academies, should work together to draw attention to this issue and promote measures to improve research practices, and share experiences of their own efforts in these endeavours.

• IAP for Health, working with its member academies at a national level, should join the efforts of the international science community to encourage discussions among partners, including international research funders and publishers/editors, about how to address this issue – seeking opportunities to facilitate these discussions, where appropriate."

Appendix 2: Transgressions against Sound Statistical Thinking

The checking of statistical calculations from published tables is useful but can be insufficient without direct access to the source-data. Detected errors may be minor, a matter of rounding say, rather than substantive. More troublesome errors of commission relate to study design (and may be irrecoverable) or to lapses in statistical thinking (and may undermine inferences).

A2.1. Change of primary outcome: preregistration of RCT-protocols has been particularly successful in identifying when authors have switched emphasis away from their a priori primary outcome to a different outcome which they have chosen to highlight post-hoc.
A2.2. Multiplicity: unless the study protocol has been quite specific about the analysis plan for each secondary outcome, a multiplicity of secondary outcomes can expand into a myriad of analysis possibilities and post-hoc emphases. The problem arises especially if the secondary outcomes are based on questionnaires, each with a host of subscales; or if outcomes are to be evaluated at a series of follow-up times.
A2.3. How explanatory factors are coded matters: the influence of age on the logarithm of clients' risk of methadone-specific death could be assumed to linearly increasing, or differentiated by pre-defined age-groups (under 25 years, 25-34, 35044, 45+ years) or tested objectively by fitting a distinct indicator variable for each of the upper four quintiles of age. Each approach has merit but they test differently the influence of age. Readers cannot be expected to know which coding has been adopted unless text and tables make this explicit.
A2.4. Do regression coefficients change when particular explanatory variables are included/excluded: Nor can readers know how other regression coefficients in the authors'

model have altered (if at all) when the influence of age is (versus is not) accounted for – unless the alternative models are reported explicitly.

A2.5. False discovery: statistical methods have been developed, and are now applied, which protect against false discovery rates in genome wide association and other genetic studies.

A2.6. Substitution of subjects: naivety may explain, but does not excuse, substitutions for experimental subjects who withdraw after they have been randomised. In unblinded studies, there is a risk that withdrawals may be influenced by the identifiable randomly-assigned treatment. Even in masked studies, the withdrawal rate matters because it reflects on the overall study-design: and should be duly reported.

A2.7. Imputation for missing data: some statisticians argue that the best thing to do with missing data is not to have any – by dint of designing studies well so that missing data rarely arises. However, research teams are also called upon to analyse policy-critical data which were not collected to research standards. Statistical methods have been developed, and are in use, which make different plausible assumptions about how the missing data has occurred. The chosen assumption can then be exploited to "impute" for the missing values and to derive estimates (with uncertainty) which properly account for the extent of missingness and which, if the missingness-assumption was correctly surmised, make substantially better use of the data than if the analysis had included only those subjects with complete data.

A2.8. Transparency about prior assumptions: imputation for missing data is only one example of the need for analysts to be transparent about the assumptions underlying their analysis, the sensitivity of results to those assumptions and the rationale (empirical; or expert judgement) for the chosen assumptions. The modelling of cost-effectiveness, as in submissions to NICE, typically calls for extrapolation of survival times beyond the follow-up horizon for the patients who participated in RCTs. See, for example, how the cost-effectiveness of screening men for abdominal aortic aneurysms was estimated by Thompson et al. (*BMJ 2009; 338: b2307;* and *British Journal of Surgery 2012; 99: 1649-1656).*

A2.9. Outliers: outliers may be unduly influential. Statistical methods have been developed, and are in use, which downgrade their influence. However, good statistical practice recognises that outliers exist, draws readers' attention to them, and explains how such outliers have been dealt with at analysis.

A2.10. Selective reporting about past studies: systematic review of *all* past studies, as recommended by the Cochrane Collaboration, obviates selective reporting about past studies but Turner et al. (2012) have shown that essentially the same inferences can be drawn, and time saved, if authors focussed their review on only those RCTs that were designed to have at least 50% power to discern modest change (such as relative risk of 1.3). Statistical peer review of applications for RCT-funding should mean that substantially under-powered RCTs did not pass muster in the 21st century; and were not funded.

A2.11. Selective reporting when setting new results in context: The format for many peer-reviewed articles is Introduction, Methods, Results, Discussion. Editors tend to discourage the repetition of uncertainty intervals in Discussion, which allows authors to cite central estimates – from their and other studies – without qualification (e.g. by standard error, number studied or confidence interval) and without putting the new results in the proper context of how much they add to the information accumulated thus far (eg when meta-analysed).

A2.12. Data-preparation: Much statistical work is put into effect by programming, firstly, of logical check on data and then to configure the data to conform to the data entry requirements for specific statistical software. Unintended errors may intrude at this data preparation stage, unless suitably

prudent safeguards or checks are made to ensure that the data configuration has worked as the analyst intended.