



OPINION ARTICLE

Time for sharing data to become routine: the seven excuses for not doing so are all invalid [version 1; referees: 2 approved, 1 approved with reservations]

Richard Smith^{1,2}, Ian Roberts^{3,4}

¹ICDDR, B, Dhaka, Bangladesh

²Former editor, BMJ, London, UK

³Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK

⁴Clinical Trials Unit, London School of Hygiene & Tropical Medicine, London, UK

v1 First published: 29 Apr 2016, 5:781 (doi: [10.12688/f1000research.8422.1](https://doi.org/10.12688/f1000research.8422.1))
 Latest published: 29 Apr 2016, 5:781 (doi: [10.12688/f1000research.8422.1](https://doi.org/10.12688/f1000research.8422.1))

Abstract

Data are more valuable than scientific papers but researchers are incentivised to publish papers not share data. Patients are the main beneficiaries of data sharing but researchers have several incentives not to share: others might use their data to get ahead in the academic rat race; they might be scooped; their results might not be replicable; competitors may reach different conclusions; their data management might be exposed as poor; patient confidentiality might be breached; and technical difficulties make sharing impossible. All of these barriers can be overcome and researchers should be rewarded for sharing data. Data sharing must become routine.

Open Peer Review

Referee Status:

	Invited Referees		
	1	2	3
version 1 published 29 Apr 2016	 report	 report	 report

- 1 **Thomas Walley**, University of Liverpool UK
- 2 **Heather M Goodare**, Edinburgh Health Forum UK
- 3 **Gustav Nilsson**, Karolinska Institute Sweden

Discuss this article

Comments (2)

Corresponding author: Richard Smith (richardswsmith@yahoo.co.uk)

How to cite this article: Smith R and Roberts I. **Time for sharing data to become routine: the seven excuses for not doing so are all invalid [version 1; referees: 2 approved, 1 approved with reservations]** *F1000Research* 2016, 5:781 (doi: [10.12688/f1000research.8422.1](https://doi.org/10.12688/f1000research.8422.1))

Copyright: © 2016 Smith R and Roberts I. This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Competing interests: RS is a paid consultant to *F1000Research*, which requires submission of full data with research articles. IR works at LSHTM which received NIHR funds to set up a data sharing website (<https://ctu-app.lshtm.ac.uk/freebird/>).

First published: 29 Apr 2016, 5:781 (doi: [10.12688/f1000research.8422.1](https://doi.org/10.12688/f1000research.8422.1))

Good, well curated data are more valuable than the words authors write about them, but until now the main currency of science has been publications. With the World Wide Web sharing and publishing data is now possible, and researchers should be rewarded for doing so. Authors unfortunately have incentives not to share data and continue to find excuses for not doing so – but the excuses are poor. It's time for data sharing to become routine.

The value of data

Datasets are more valuable than papers because: they allow analyses to be replicated helping to avoid error, selective reporting and fraud; they can be used to answer other research questions; and they facilitate methodological research and the teaching and training of researchers. Papers, in contrast, rarely report the full data and are often “spun” to present results that flatter authors and please editors.

Patients are the main beneficiaries of data sharing

The main beneficiaries of sharing data are patients, the people who as taxpayers fund most research. They clearly have an interest in both the right conclusion being reached and in maximum value being squeezed from every dataset. Unfortunately many others in the research system do not have the same interest in the “truth.”

If we consider a clinical trial or indeed any study with clinical implications then the prime interest of the patients is that the results are “true” and that clinicians use them to improve their well-being. This means that the analyses should be accurate and replicable. Sadly the producers of research have interests apart from truth: researchers want high impact papers; universities want the same and lots of publicity too; editors and publishers want “good” publications that increase their impact factor; and funders want to show “value for money,” which may mean lots of publications regardless of their truth. Nobody is incentivised to share data, replicate results, and perhaps show the weak underbelly of science, which is why the scientific community has responded so poorly to allegations of misconduct¹.

By participating in clinical research patients make a gift to others, rather as those who give blood do. They and their gift, their data, should be treated with reverence. Their gift is not for individual researchers to use to advance their careers but for the wider scientific community and other patients. Their gift must be shared.

The seven incentives not to share

Because they are measured primarily by how much and where they publish, researchers are strongly incentivised to publish, preferably in high impact journals. There are not the same incentives to share data. Indeed, there are seven incentives (or excuses) not to share.

Firstly, data are the base for research articles, and one anxiety for researchers is that others will use their data to produce publications without having to go to the trouble of gathering them. They will be disadvantaged in the academic rat race, although if everybody shared data they could benefit from using data from others.

Secondly, other researchers might scoop them, perhaps even prevent them from achieving publication in a high impact journal. Funders

who require data sharing have responded to the anxiety of being scooped by allowing researchers to delay sharing their data. A better response would be to move away from “outsourcing” the judgement of the performance of researchers to publishers and for employers and funders to recognise that judging researchers is core business that should not be outsourced to the arbitrary and corrupted publishing process.

A third reason for not sharing data is a fear held by researchers that their conclusions will not be replicable. This is an ignoble reason because replicability is central to science. Some scientists may fear replication because they repeat experiments day after day and publish them only when they become “right.” This is unscientific and can lead to serious defects in the scientific evidence base.

One of us (IR) has made data from two large clinical trials available in the hope that somebody will replicate the analysis and confirm (or fail to confirm) the results (<https://ctu-app.lshtm.ac.uk/freebird/>)^{2,3}. Although the data have been used to answer many different questions, there has been no replication of the original trial results, probably because there is no incentive to do so - there ought to be. It surely makes economic sense for the millions spent on the trial to be backed up by the few thousands that would be needed to encourage replication. We hope that somebody will take up the challenge.

A fourth reason researchers may want to keep their data to themselves is to avoid their critics analysing the data and coming up with different or contrary results. Statisticians say that “if you torture the data they will confess,” but refusing to release data hands a victory to critics who will inevitably say “the researchers obviously have something to hide, they can't support their conclusions.” Uncomfortable as it may be, it's a better and more scientific strategy to enter “the market of ideas” and expect to show the correctness of your analysis and conclusions.

There is a legitimate worry about releasing data when researchers fear they may be sued. The problem here is that a battle in court is not a battle of evidence and data but a battle of showmen with a highly uncertain outcome. This is not a worry with most datasets, and perhaps when it is the data can be released in exchange for a legally binding commitment not to sue.

The authors of a major trial that showed the ineffectiveness of hydroxyethyl starch solutions for fluid resuscitation have declined to share their data^{4,5}. They say that there have been “repeated efforts to discredit” by critics who want “to protect their commercial interests.” The authors have declined even to allow a reanalysis by a third party. This cannot be in the interest of patients, who clearly want to know whether the treatment is ineffective or not, but the authors may have a legitimate worry about legal action.

The fifth and perhaps worst reason for not releasing data is that data management is often poor and sharing the data may expose horrible weaknesses, flaws, and inconsistencies in the data. Sadly this may be the commonest but least declared reason for not sharing data. That some universities dedicate more resources to media relations than research governance is disturbing but not surprising.

Making a big splash in the news can bolster grant income and student recruitment even when the informational content of the research is doubtful.

A sixth excuse for not sharing data that is available to those who do research with patients is patient confidentiality. One case of private information of a patient being exposed could, some researchers argue, bring data sharing to a halt. It is a “never event” that must be avoided even if huge benefits are foregone by not sharing data. Patient confidentiality must be guarded, and most of the time it’s easy to do so by anonymising data and removing data on, for example, place and time. It’s true that small risks remain because of rare conditions and events and because of “jigsawing” (combining datasets to break confidentiality), but these small risks can be explained to patients, who will almost always consent to their data being made available in anonymous form. With datasets that are already collected patients might be asked to give retrospective consent.

Patient confidentiality is the reason that authors of a controversial trial on treatment of chronic fatigue syndrome give for not sharing their data, but inevitably they look as if they are hiding something^{6,7}.

The final and probably weakest excuse researchers give for not sharing data is “technical reasons.” But this is a lame excuse—other areas of science—for example, physics, astronomy, and engineering—have shared datasets far larger and more complex than those produced in biomedical research. There are no insurmountable technical reasons to sharing and publishing data.

Reward authors for sharing data

Researchers should be rewarded not for publications but for producing large amounts of high quality data. Papers are a poor measure of the quantity or quality of research data. In terms of papers, a trial with 100 patients is the same as one with 10 000 patients, even though the informational content of the latter is 100 times the former. And despite the reverence for peer review, data quality is remarkably hard to judge from publications.

Funders of research and employers of researchers need to change the incentives for researchers to encourage data sharing, but researchers must also recognise the weakness of their excuses and contribute to the big advance in science that can come from sharing and publishing data.

Author contributions

Both authors contributed to the paper and have read and approved the final version.

Competing interests

RS is a paid consultant to *F1000Research*, which requires submission of full data with research articles. IR works at LSHTM which received NIHR funds to set up a data sharing website (<https://ctu-app.lshtm.ac.uk/freebird/>).

Grant information

The author(s) declared that no grants were involved in supporting this work.

References

1. Smith R: **Statutory regulation needed to expose and stop medical fraud.** *BMJ*. 2016; **352**: i293.
[PubMed Abstract](#) | [Publisher Full Text](#)
2. The CRASH-2 trial collaborators, Shakur H, Roberts I, *et al.*: **Effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with significant haemorrhage (CRASH-2): a randomised, placebo-controlled trial.** *Lancet*. 2010; **376**(9734): 23–32.
[PubMed Abstract](#) | [Publisher Full Text](#)
3. Edwards P, Arango M, Balica L, *et al.*: **Final results of MRC CRASH, a randomised placebo-controlled trial of intravenous corticosteroid in adults with head injury—outcomes at 6 months.** *Lancet*. 2005; **365**(9475): 1957–9.
[PubMed Abstract](#) | [Publisher Full Text](#)
4. Doshi P: **Data too important to share: do those who control the data control the message?** *BMJ*. 2016; **352**: i1027.
[PubMed Abstract](#) | [Publisher Full Text](#)
5. Myburgh JA, Finfer S, Bellomo R, *et al.*: **Hydroxyethyl starch or saline for fluid resuscitation in intensive care.** *N Engl J Med*. 2012; **367**(20): 1901–11.
[PubMed Abstract](#) | [Publisher Full Text](#)
6. Smith R: **QMUL and King’s college should release data from the PACE trial.** [Reference Source](#)
7. White PD, Goldsmith KA, Johnson AL, *et al.*: **Comparison of adaptive pacing therapy, cognitive behaviour therapy, graded exercise therapy, and specialist medical care for chronic fatigue syndrome (PACE): a randomised trial.** *Lancet*. 2011; **377**(9768): 823–36.
[PubMed Abstract](#) | [Publisher Full Text](#)

Open Peer Review

Current Referee Status:



Version 1

Referee Report 13 June 2016

doi:[10.5256/f1000research.9066.r14294](https://doi.org/10.5256/f1000research.9066.r14294)



Gustav Nilsson

Department of Clinical Neuroscience, Karolinska Institute, Stockholm, Sweden

This opinion piece describes and refutes seven arguments against sharing research data. The authors focus on clinical trials, but their reasoning is applicable to research with human participants in general.

In the ongoing conversation about open research data in scientific journals, arguments against open data are not always presented clearly and explicitly. The mere listing of counterarguments in a paper that can be referenced is therefore an important contribution.

The authors refute each argument against data sharing in a clear and coherent manner and their counterarguments are a valuable resource for researchers debating open data.

I have only one minor point of criticism: the statement in the last paragraph that a study with 10 000 participants has 100 times more information content than a study with 100 participants does not take into account the diminishing information content in consecutive dependent observations. I suggest this may be reworded.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: I am currently chairperson of the Open Badges committee at the Center for Open Science, which works to incentivise data sharing.

Referee Report 13 May 2016

doi:[10.5256/f1000research.9066.r13824](https://doi.org/10.5256/f1000research.9066.r13824)



Heather M Goodare

Edinburgh Health Forum, Edinburgh, UK

Thank you for asking me to comment on this paper. I can only speak from the point of view of patients and carers.

The main problem is that of confidentiality of data, and some patients are worried about this. The authors acknowledge that this could be a problem (a sixth excuse). Anonymising data is of course essential, but

'small risks remain'. Remember the case of the anonymous male with back problems who was written about in an American medical journal and turned out (without too much detective work) to be President Kennedy? Personally I don't care tuppence who knows that I have had breast cancer - it's in the public domain anyway. But some conditions people would not wish to be known about: abortions, STD, some mental illnesses, and so on. If data are anonymised that is usually sufficient safeguard, but in epidemiological studies unique postcodes are a giveaway.

I am a member of the Public Panel of the FARR Institute in Scotland, and we have debated the matter of Big Data at length. We have a system here called SHARE, where if you are happy for your data to be used for research you sign a form, obtainable from your GP's surgery. This also gives permission for the residue of blood samples taken for routine purposes to be used for research. Most people are happy, but some are not, even given the guarantee of anonymity.

However, this system gives permission for data culled from healthcare registries to be used for research: it does not as far as I know include data from trials already conducted. This to me is a new idea, and it raises different issues.

FARR talks about 'safe havens' for data, so that personal details cannot be shared and anonymity is guaranteed. It seems to me that if data already gathered for research are to be released to researchers other than the original investigators, this raises an entirely new issue. It would mean that consent forms should be revised so that they take account of the possibility that data will be shared with others at a later date.

It is important to make it clear that healthcare data are not the property of the researchers who have only borrowed them: they belong to the patient. Therefore, if data are to be made more widely available, the patient needs to give consent. This means that consent forms need to make this explicit, and all other data used for research, for instance epidemiological studies that do not require active co-operation from the patient, need to have blanket consent from patients, who should be encouraged to complete a SHARE form.

Personally, I like to know what researchers are going to do with my data. My husband and I were 'consulted' as members of a patient reference group about a stroke trial (he has had a stroke), and we both felt that it should not have gone ahead: the rest of the patient group thought so too, but it went ahead anyway. I don't know how it got through Ethics. The relevance for this paper is that patients do have a right to say what their data are going to be used for. If they don't approve of the trial, then they won't let their data be used. If they have given permission initially for a study that they approve of, and the proposal is to share the data further, should they not be given a say in what their data are to be used for subsequently? Once Big Pharma get their hands on the data who knows what will become of it.

This makes the sharing of data more complicated, but I believe it should be done, and that this article needs to take account of these issues.

I also have some minor copy-edit suggestions:

- Abstract: "[...]competitors **might** reach different conclusions[...]"

- The seven incentives not to share: “[...]for employers and funders to recognise that judging researchers is core business that should not be outsourced to the arbitrary and **corrupt** publishing process.”
- The seven incentives not to share: “This cannot be in the interest of patients, who clearly want to know whether the treatment is **effective** or not [...]”
- The seven incentives not to share: “There are no insurmountable technical reasons **for not** sharing and publishing data.”

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Referee Report 05 May 2016

doi:10.5256/f1000research.9066.r13665



Thomas Walley

Department of Health Services Research, University of Liverpool, Liverpool, UK

Data sharing has been an expectation and indeed a contractual obligation for all research funded by NIHR, the research arm of the NHS, for many years. This has meant that bona fides researchers can request access to study data for defined purposes and with a suitable protocol, which should not be unreasonably withheld, e.g. for purposes of IPD meta-analysis. This is not open but controlled access to the data. The arbiter of what is reasonable access to the data falls to the researcher in the first instance, then to his/her host institute, but ultimately to the funder who held the contract.

The recent consultation from the ICMJE (<http://www.nejm.org/doi/full/10.1056/NEJMe1515172>) will probably translate into a requirement that data sets be made available in a more transparent way, usually by host institutions, in some form of as yet undefined registry.

Why not open access? Smith and Roberts consider some of these issues:

Ownership of the data: this (and responsibility for curation and archiving) rests with the institute but subject to the terms of the contract. Inevitably however, a researcher will feel a degree of proprietary protectiveness towards data sets. Most of us are not as altruistic in this regard as Smith and Roberts would like. Given the incentives that exist in academia, some respect for the intellectual property that the researcher has created is inevitable, and usually an agreement to access the data either in collaboration or with due acknowledgement is an acceptable outcome for all.

Risks of confidentiality: many studies are not of the 20000 patients size that Roberts has made available: smaller studies, with geographically defined recruitment may mean that the patient is potentially identifiable, especially if complex sets of data – often collected in smaller studies but less likely in larger - can also be accessed. Regrettably, there are people who seem to thrive on breaking open data like this: I think that patient confidentiality requires us to ensure that the data remains anonymous, best achieved by limited rather than open access.

Poor data handling: making data available to others is not without substantial cost, at a time when most researchers are planning to move on to another study: e.g. labelling the files from complex data sets in clear manner understandable to those who have not lived and breathed it for several years. Hence collaborative access is an easier and less expensive solution, where possible. Archiving the data also poses problems – who will take responsibility for converting data from old systems or software.

NIHR have established a contractual obligation, but like most other funders, has not yet provided the level of funding to make this possible (except on one occasion to Roberts), nor a vehicle similar to the GSK-led clinicalstudydatarequest.com to facilitate this.

None of this is to argue against the principles that Smith and Roberts put forward, but only to point out that achieving their worthy aims will not be easy or as quick as it might seem. NIHR like other funders continue to work to support this aim. As part of this, the NIHR journals library is also considering what constitutes publication: perhaps a somewhat selective journal article, a detailed monograph as has been our practice (www.journalslibrary.nihr.ac.uk) or in the future, such a document with access to the data. These questions will not be quickly solved, and need much more debate to which this article by Smith and Roberts is a valuable contribution

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: I have framed this from an NIHR perspective and work for NIHR

Author Response (*F1000Research Advisory Board Member*) 10 May 2016

Richard Smith, Patients Know Best Ltd, UK

I'm grateful to Tom for giving a rapid and useful on our paper. No doubt he is right that it will take a longer time than we would like for data sharing to become routine.

Incentives are fundamental. At the moment incentives reward keeping data, but we must change the incentives. We argue that the data are more valuable than the papers that arise from them, and so funders of research should be thinking hard about how to reward the production of high quality data. At the moment huge value is being lost from data being locked away, and data have a longer lifespan than papers. At the very least funders should be willing to meet the costs of data sharing that Tom identifies.

The confidentiality risk is, I fear, exaggerated. The obvious response is for researchers to get consent from participants for data to be shared at the same time as minimising the risk of exposure. It used to be that doctors did not get consent from patients for the sharing of case reports, but now they have to—and few patients refuse. The risk of exposure from participation in a trial is way below that of a case report.

Competing Interests: I'm one of the authors of the paper, and my competing interests are included in the paper.

Discuss this Article

Version 1

Author Response (F1000Research Advisory Board Member) 10 May 2016

Richard Smith, Patients Know Best Ltd, UK

I'm grateful to Carolyn for commenting on our article, and I agree that "no parties to research are 'neutral.'" Indeed, with other colleagues I have written on "the fallacy of impartiality." (1)

Recognising that nobody, including the authors is neutral, is, as Carolyn writes, a strong argument for data sharing. She identifies two ways in which we can improve the reliability of research, and they are, of course, not mutually exclusive—but reanalysis of data may be the best.

1 Smith R, Feachem R, Feachem NS, Koehlmoos TP, Kinlaw H. The fallacy of impartiality: competing interest bias in academic publications. *Journal of the Royal Society of Medicine*. 2009;102(2):44-45. doi:10.1258/jrsm.2009.080400.

Competing Interests: I'm one of the authors of this paper and declared my competing interests in the paper.

Reader Comment 04 May 2016

Carolyn Wilshire, Victoria University of Wellington, New Zealand

"Statisticians say that "if you torture the data they will confess,"

I wish to comment on this quote, which has appeared in various forms in other articles written by they first author

If we take the quote at face value - to be true in some sense - then it does not raise a problem for data sharing. Rather, it raises problems with NOT sharing data.

Consider we have a group of primary researchers who collected the data, and another group, who are suspicious of its conclusions, and wish to examine the data for themselves. Who in this scenario is most powerfully motivated to "make the data confess"? Very probably, the primary researchers themselves.

Let's be realistic here. Researchers do not approach their data as neutral bystanders without investment. They come to it with a powerful set of beliefs. Many have invested years of their career into those beliefs. Like all human beings, they are convinced that there will be support for their view in the data somewhere - if only they can find it! So they explore all sort of variables and ways of measuring them. They look at "outliers" and maybe take a few out in various ways. They notice errors that work against their conclusion, but may fail to notice those that work in its favour. And so on. These practices are widespread, and need not indicate outright fraud. But they can - and often do - lead to significant distortion of the facts. Add to that the personal motives associated with a desire to get published and advance one's career, and we have the perfect recipe for data torturing.

In Psychology, we are only just becoming aware of the size of this problem, as various findings once thought to be secure have turned out to be unreproducible.

It is time to recognise that no parties to research are "neutral". All can be subject to bias. There are two ways to improve the reliability of our research. The first is to continue to question our methods and improve how we conduct research (through the use of pre-registration, reporting standards and guideline, consideration of the limits of inferential methods like null hypothesis testing). The second is to allow groups with different beliefs and motives to examine the same data. and for each to present their findings. Researchers can then evaluate both sets of conclusions.

Competing Interests: No competing interests were disclosed.
