PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Has open data arrived at the British Medical Journal (BMJ)? – an observational study
AUTHORS	Rowhani-Farid, Anisa; Barnett, Adrian

VERSION 1 - REVIEW

REVIEWER	Genevieve Pham-Kanter Assistant Professor Department of Health Management and Policy Drexel University Dornsife School of Public Health, USA
REVIEW RETURNED	29-Mar-2016

GENERAL COMMENTS	Thank you for giving me the opportunity to review "Has open data arrived at the British Medical Journal (BMJ)? an observational study" (MS ID bmjopen-2016-011784). The purpose of this paper is to assess the degree to which datasets used by articles published in the BMJ are publicly available. The authors randomly selected 160 articles published in the BMJ between 2009 and 2015; identified the subset of articles that indicated that their datasets were available; and contacted the authors of these articles to request the datasets. The authors were able to obtain 14% of the datasets used by articles that indicated that the underlying data were available, and 4% of datasets used in the original 160 articles sampled. The main strength of the paper is that it presents an interesting update on the availability of datasets from articles published in a medical journal that encourages data sharing. The figure, tables, and methodological description were also clear and helpful. The main weaknesses of the paper are that it aggregates different concepts and measures that are important to keep analytically separate and—given the methodology used to acquire datasets—may over-reach a bit with its conclusions. Overall, the paper adds to the literature on the degree to which data underlying medical journal publications are easily accessible. Major Comments: 1. The authors emphasize that only 31% of BMJ articles included a statement indicating that the raw data were publicly available; that only 14% of the datasets for articles that indicated data availability were successfully obtained; and that overall, only 4.4% of the datasets from the original sample of articles were obtained. These are very low rates, but it is important to note that BMJ policy governs only a fraction of the articles the authors sampled. BMJ

	policy governs only data sharing for clinical trials (and until 2015, for only a subset of clinical trials), so it would be helpful for the authors to separate data availability for those studies governed by the BMJ data sharing policy at any given time versus those for whom data sharing is voluntary, rather than combining them in the 3 statistics above. The authors do note on p. 4 and briefly in the Discussion section that the purpose of their paper is to examine data sharing defined broadly, but it seems substantively important to report in detail the data availability for these two groups separately. The question of whether there is full compliance with existing requirements seems separate and distinct from the question of whether researchers are voluntarily making their data available. Both are important questions to answer, are analytically distinct, and have very different policy implications. 2. The way in which the authors attempted to secure the datasets may have contributed to low rates of data acquisition. In particular, the authors requested the data by e-mail instead of interceding through BMJ and did not follow through with additional work that may have been required to obtain data access, including additional applications for data from articles whose statements explicitly indicated that data were available from external sources but would require additional application ("given the large amount of time it would take to apply, and because there was no guarantee we would gain access to the data" (p. 5)) . Unless the authors were sent the data directly through email, no questions asked, or the authors could download the data directly from a public website, the data were classified as unavailable. I'm not sure this is a reasonable standard to use to classify data as being unavailable—classified as not being easily acquired, yes—but not necessarily classified as being unavailable to the public. Given the nature of health research, only some types of studies and some types of data would be amenable to the easily uploadable data model. There are many reasons for lack of data sharing—some attributable to investigator personal and professional motives, others attributable to the research environment (and specific field norms and competitiveness), and others attributable to the nature and source of the data. The authors' binary rule of sharing or not based on whether the dataset is instantaneously obtainable does not acknowledge the complexity underlying data sharing decisions. 3. Related to point #2, the authors might want to report (a breakdown using) some kind of hierarchy in terms of how easily acquired the datasets are as well as characteristics related to ease of acquisition. This additional information may require more effort on their part to try to acquire the data so that they have a more refined measure of ease of acquisition. There is some information available in the Supplemental Appendix but this should both be more elaborated upon—the authors should make at least some additional effort in obtaining the data—as well as consolidated (i.e. the authors should condense/summarize the table). 4. I'm not as fussed as the authors about the additional requirements sometimes imposed to obtain data. It is not always possible to include these details in a data sharing statement. Sometimes these requirements are imposed by one's home

	institution, and sometimes they are barriers created by investigators driven by not-terribly-noble motives. This is why having a sense of the hierarchy of difficulty in obtaining data sets (referred to in point #3)—from, say, (1) downloadable instantaneously, to (3) requires some paperwork, to (10) basically impossible—seems important. The authors do raise an important point about the hidden requirements, so it would be interesting to pursue this idea further. Minor Comments: 1. In the Article Summary box, the authors state that "Our study quantified data sharing policy compliance," but as I mentioned in Point #1 in the Major Comments, it seems that most of the articles they sampled were not governed by the data sharing policy, so the study was arguably not assessing compliance (unless they were referring to the policy of having a data sharing statement, but the authors do not report the rate of compliance with this statement policy). A similarly equivocal use of the term "compliance" appears in the first full paragraph of p. 10. 2. In the abstract and in the methods section, the authors should specify the N of their sampling frame, i.e. the number of articles from which they sampled 160 articles. They should also indicate somewhere in their abstract the years (2009-2015) their sample covered.

REVIEWER	Serryth Colber Royal United Hospital NHS Foundation Trust, Bath UK
REVIEW RETURNED	07-Apr-2016

GENERAL COMMENTS	Well done on putting this interesting paper together - very well written. What changes to you recommend in view of your outcomes? Do you recommend that all authors submit their raw data to the publishers as part of the submission process?.... and that the publishers make the data available to the reader upon request? Would this be a sensible or practical solution to the difficulties with getting a response form the authors? Can you think of any other solutions? - please add to the paper as a minor revision. Well done once again.

REVIEWER	Giovanni Destro Bisol Università di Roma "La Sapienza"
REVIEW RETURNED	26-Apr-2016

GENERAL COMMENTS	This is an useful study which will could help BMJ improve data sharing policies. However, the sample size is really small, and the authors should try to make an effort to make it larger. Not less importantly, they should be more aware of the abundant literature on empirical and conceptual aspects of data sharing. There are some points which need to be revised: 1. The first paragraph of the introduction is wrong: Open data and data sharing are NOT interchangeable terms. In fact, Open data includes concepts like accessibility, assessability, useability and

	intelligibility (See Boulton et al. 2012). The authors should clearly state on which of these aspects their study has been focused. 2. Science 2.0 should not be confused with Open data 3. What is a strong sharing policy? This point seems to have been underconsidered: please look at Piwowar HA, Chapman WW (2008) Proceedings of the ELPUB 2008 Conference on Electronic Publishing. Toronto (ON), 25–27 June. 4. When comparing their results, the authors consider just two studies. The first one is practically useless, being based on the analysis of just 10 papers (Savage and Vickers 2009). They seem to ignore the existence of empirical studies on data accessibility (e.g. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0037552, of which I am the senior author). 5. The study cited above also discusses how data sharing policies may be enforced and provide examples of good practice.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Genevieve Pham-Kanter

Thank you for giving me the opportunity to review "Has open data arrived at the British Medical Journal (BMJ)? an observational study" (MS ID bmjopen-2016-011784).

The purpose of this paper is to assess the degree to which datasets used by articles published in the BMJ are publicly available. The authors randomly selected 160 articles published in the BMJ between 2009 and 2015; identified the subset of articles that indicated that their datasets were available; and contacted the authors of these articles to request the datasets. The authors were able to obtain 14% of the datasets used by articles that indicated that the underlying data were available, and 4% of datasets used in the original 160 articles sampled.

The main strength of the paper is that it presents an interesting update on the availability of datasets from articles published in a medical journal that encourages data sharing. The figure, tables, and methodological description were also clear and helpful. The main weaknesses of the paper are that it aggregates different concepts and measures that are important to keep analytically separate and—given the methodology used to acquire datasets—may over-reach a bit with its conclusions. Overall, the paper adds to the literature on the degree to which data underlying medical journal publications are easily accessible.

Major Comments: 1. The authors emphasize that only 31% of BMJ articles included a statement indicating that the raw data were publicly available; that only 14% of the datasets for articles that indicated data availability were successfully obtained; and that overall, only 4.4% of the datasets from the original sample of articles were obtained.

These are very low rates, but it is important to note that BMJ policy governs only a fraction of the articles the authors sampled. BMJ policy governs only data sharing for clinical trials (and until 2015, for only a subset of clinical trials), so it would be helpful for the authors to separate data availability for those studies governed by the BMJ data sharing policy at any given time versus those for whom data sharing is voluntary, rather than combining them in the 3 statistics above.

The authors do note on p. 4 and briefly in the Discussion section that the purpose of their paper is to examine data sharing defined broadly, but it seems substantively important to report in detail the data availability for these two groups separately. The question of whether there is full compliance with existing requirements seems separate and distinct from the question of whether researchers are voluntarily making their data available. Both are important questions to answer, are analytically distinct, and have very different policy implications.

RESPONSE:
Thank you for these thoughtful comments. We have added text that examines the data availability of those papers specifically bound by the BMJ data sharing policy: randomised controlled trials (RCTs). Out of the 160 (we have reduced the sample size to 157 as 3 articles had data available in the article itself) randomly sampled papers, 50 had data available, 21 of which were RCTs. Every RCT stated in their data sharing statements that data is available – 1 dataset was freely available on Dryad, leaving 20 RCTs which were emailed to request their data. 13/20 did not respond to our email, 4/20 gave responses that were consistent with their data sharing statements and made their data available (one of which was unverifiable). Hence 5/21 RCTs made their datasets available to our team, a data sharing rate of 24%, which is still low. 29/50 articles were not bound by the policy, but indicated their data is available, only 2 made their data available to us. The sharing rate for papers not bound by the BMJ data sharing policy is 2/29 (7%).

2. The way in which the authors attempted to secure the datasets may have contributed to low rates of data acquisition. In particular, the authors requested the data by e-mail instead of interceding through BMJ and did not follow through with additional work that may have been required to obtain data access, including additional applications for data from articles whose statements explicitly indicated that data were available from external sources but would require additional application ("given the large amount of time it would take to apply, and because there was no guarantee we would gain access to the data" (p. 5)). Unless the authors were sent the data directly through email, no questions asked, or the authors could download the data directly from a public website, the data were classified as unavailable. I'm not sure this is a reasonable standard to use to classify data as being unavailable—classified as not being easily acquired, yes—but not necessarily classified as being unavailable to the public. Given the nature of health research, only some types of studies and some types of data would be amenable to the easily uploadable data model. There are many reasons for lack of data sharing—some attributable to investigator personal and professional motives, others attributable to the research environment (and specific field norms and competitiveness), and others attributable to the nature and source of the data. The authors' binary rule of sharing or not based on whether the dataset is instantaneously obtainable does not acknowledge the complexity underlying data sharing decisions.

RESPONSE:
This is a fair point and we did acknowledge that we did not use the BMJ as a broker to negotiate access for the datasets nor did we apply for datasets. We now present our data in three categories: easily available, potentially available, and not available:
Data easily available = 7/50
Data potentially available = 8/50
Data not available = 35/50

3. Related to point #2, the authors might want to report (a breakdown using) some kind of hierarchy in terms of how easily acquired the datasets are as well as characteristics related to ease of acquisition. This additional information may require more effort on their part to try to acquire the data so that they have a more refined measure of ease of acquisition. There is some information available in the

Supplemental Appendix but this should both be more elaborated upon—the authors should make at least some additional effort in obtaining the data—as well as consolidated (i.e. the authors should condense/summarize the table).

RESPONSE:
We agree about the further reporting and this is related to our response on point #2. In terms of characteristics related to ease of acquisition, the key characteristic is whether the paper was an RCT and so bound by the policy. Here is the RCT breakdown for the 21/50 RCT that indicated that data is available:

Data easily available = 5/21
Data potentially available = 0/21
Data not available = 16/21

Given that none of the RCT datasets were 'potentially available', there is no need for us to pursue the additional avenue of using the BMJ as a broker as the data sharing policy does not cover those datasets that were potentially available.

We have not made additional effort to obtain the data as we only had ethical approval to e-mail the authors and not to ask the BMJ to intervene.

We have removed the additional information in the Supplemental Appendix and all of the results are analysed and summarised in the body of text of our article. The Appendix is our dataset which we have submitted for reproducibility and data sharing purposes.

4. I'm not as fussed as the authors about the additional requirements sometimes imposed to obtain data. It is not always possible to include these details in a data sharing statement. Sometimes these requirements are imposed by one's home institution, and sometimes they are barriers created by investigators driven by not-terribly-noble motives. This is why having a sense of the hierarchy of difficulty in obtaining data sets (referred to in point #3)—from, say, (1) downloadable instantaneously, to (3) requires some paperwork, to (10) basically impossible—seems important. The authors do raise an important point about the hidden requirements, so it would be interesting to pursue this idea further.

RESPONSE:
We understand that sometimes there are additional requirements and barriers for data sharing, but we are still firm believers of transparency and full disclosure of all policies that limit access to data. It would not take much extra time to add this information to the data sharing statement. An example could read: "We agree to share all data and computer code. However, our University's Data Sharing policy states that data is only available on university premises". To us, the most transparent data sharing is freely available, easily accessible raw data that is downloadable from an online data depository such as Dryad.

Minor Comments:

1. In the Article Summary box, the authors state that "Our study quantified data sharing policy compliance," but as I mentioned in Point #1 in the Major Comments, it seems that most of the articles they sampled were not governed by the data sharing policy, so the study was arguably not assessing compliance (unless they were referring to the policy of having a data sharing statement, but the authors do not report the rate of compliance with this statement policy). A similarly equivocal use of the term "compliance" appears in the first full paragraph of p. 10.

RESPONSE:
We agree with your point above and have changed the bullet point in the Article Summary box of the article by removing the words “policy compliance”.
2. In the abstract and in the methods section, the authors should specify the N of their sampling frame, i.e. the number of articles from which they sampled 160 articles. They should also indicate somewhere in their abstract the years (2009-2015) their sample covered.

RESPONSE:
We have added: “160 randomly sampled articles from 2009 to 2015”.

Reviewer: 2
Reviewer Name: Serryth Colber

Well done on putting this interesting paper together - very well written. What changes to you recommend in view of your outcomes? Do you recommend that all authors submit their raw data to the publishers as part of the submission process?.... and that the publishers make the data available to the reader upon request? Would this be a sensible or practical solution to the difficulties with getting a response form the authors? Can you think of any other solutions? - please add to the paper as a minor revision. Well done once again.

RESPONSE:
Thank you for the positive feedback. Data sharing is easily accessible, freely available raw data to ensure research reproducibility and transparency. We support open data in health and medical research, which applies to all type of data, not just from randomised controlled trials. The open data movement is quite new and, as our study demonstrates, data sharing policies are not yet fully effective. Getting a response from authors is a barrier so we are proposing that authors submit their raw data on a publicly available data depository such as Dryad. Some recent guidelines recommend that authors deposit their raw data as a requirement to getting published (http://www.nature.com/sdata/policies/data-policies). We do not think that using the publishers as a broker to negotiate access to raw data is practical or consistent with fully transparent data sharing.

Reviewer: 3
Reviewer Name: Giovanni Destro Bisol

This is an useful study which will could help BMJ improve data sharing policies. However, the sample size is really small, and the authors should try to make an effort to make it larger. Not less importantly, they should be more aware of the abundant literature on empirical and conceptual aspects of data sharing.

There are some points which need to be revised:
1. The first paragraph of the introduction is wrong: Open data and data sharing are NOT interchangeable terms. In fact, Open data includes concepts like accessibility, assessability, useability and intelligibility (See Boulton et al. 2012). The authors should clearly state on which of these aspects their study has been focused.

RESPONSE:
Thank you for pointing this out – the literature defines the terms open data as the ‘the growing movement to disseminate datasets along with their published articles’ Warr (2014) (doi: 10.1007/s10822-013-9705-z.) and Wikipedia defines data sharing as the act of sharing raw data.

The purpose of open data is easily accessible, freely available, raw data to promote research transparency and integrity. Our study is focussed on both open data and data sharing and does not fragment the two concepts. Open data is the term given to the cultural shift, and data sharing is the act of making raw data available – and our study focusses on both. We have changed our wording in the introduction and removed the word “interchangeable”.

2. Science 2.0 should not be confused with Open data

RESPONSE: We agree these are separate but related concepts.
The literature discusses the current paradigm shift in health and medical research and uses terms such as Science 2.0 to refer to the ‘scientific revolution’. Wikipedia also defines Science 2.0 as “similar to the open research and open science movements”. Open data and data sharing are now being considered as fundamental elements of the shift towards health and medical research that is verifiable, reproducible and transparent.

3. What is a strong sharing policy? This point seems to have been underconsidered: please look at Piwowar HA, Chapman WW (2008) Proceedings of the ELPUB 2008 Conference on Electronic Publishing. Toronto (ON), 25–27 June.

RESPONSE:
We will cite this paper – thank you for sharing it with us. The BMJ has been a pioneer of data sharing in health and medicine and has a strong data policy compared with other high ranking health and medical journals.

4. When comparing their results, the authors consider just two studies. The first one is practically useless, being based on the analysis of just 10 papers (Savage and Vickers 2009). They seem to ignore the existence of empirical studies on data accessibility (e.g. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0037552, of which I am the senior author).

RESPONSE:
We considered studies that were similar to ours, namely in health and medical research. Savage and Vickers 2009 paper is a highly cited paper that despite its small sample size, presents a prevalent issue – namely, that there is a gap between data sharing policy compliance and real practice.

Thank you for highlighting your empirical study. We have cited the study, however, the cultural of data sharing and open data in genetics is different to health and medical research as the principles outlined in Bermuda in 1996 have guided the cultural of openness in genetics. The issue of closed data is more prevalent in health and medical research, and there are few empirical studies to show the gap between policies and practice.

5. The study cited above also discusses how data sharing policies may be enforced and provide examples of good practice.

RESPONSE:

We agree that the health and medical research field has a lot to learn from the field of genomics. The article outlined some key points about enforcing data sharing policies and the findings of the paper have shaped the current data sharing movement.

VERSION 2 – REVIEW

REVIEWER	Serryth Colbert Royal United Hospitals NHS Foundation Trust Bath UK
REVIEW RETURNED	08-Jul-2016

GENERAL COMMENTS	The authors addressed the comments made from the first revision successfully. I am happy to recommend publication of this paper

REVIEWER	Giovanni Destro Bisol University of Rome "La Sapienza"
REVIEW RETURNED	03-Jul-2016

GENERAL COMMENTS	The authors have carefully revised the MS. I believe it is now acceptable for publication.