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.

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open. The paper was subsequently accepted for publication at BMJ Open.

One reviewer who previously reviewed this paper from another journal declined to publish his comment alongside with the article.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Data Sharing through a NIH Central Database Repository: A cross-sectional survey of BioLINCC users
AUTHORS	Ross, Joseph; Ritchie, Jessica; Finn, Emily; Desai, Nihar; Lehman, Richard; Krumholz, Harlan; Gross, Cary

VERSION 1 - REVIEW

REVIEWER	Joshua David Wallach PhD Candidate in Epidemiology and Clinical Research and researcher at the Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA After completing and submitting the first review for this paper (the BMJ), I was interviewed for a postdoctoral fellowship by Dr. Ross. My initial comments and recommendations were made before any interview had been scheduled. I have never published with any of the authors and shared all of this information with the editors of BMJ Open.
REVIEW RETURNED	05-Jun-2016

GENERAL COMMENTS	As I stated in the first review of the manuscript for the BMJ, this paper is a logical follow up to the 2012 BMJ article, "Sharing of clinical trial data among trialists: a cross sectional survey," by a group of similar authors. This study aims to understand the experiences of investigators who requested and received access to clinical research data from BioLINCC. Considering the growing effort to encourage reproducibility and transparency practices in the scientific community, including public access and use of raw data, the overall scope of this project is novel and the findings are significant to a broad clinical audience. Furthermore, the article is well written and the methods described are appropriate. The primary concern that I brought up during the first review was the (low) overall survey response rate. This updated version of the manuscript addresses my primary concern and many of the minor concerns that all four of the reviewers shared. Having studied the decision email from the BMJ, it is clear the all four peer reviewers thought favorably of the manuscript. Below are some new comments and some of the comments from the previous review. For the comments that have already been addressed in the current manuscript draft, I did not update the page

	and line numbers. 1. Addressed: Page 3, Line 38, Abstract: The opening of the results should state that there were 536 investigators who requested and received access to clinical research data from BioLINCC between 2007 and 2014. This gives a better sense of what the n= “441 potential respondents” actually means. **The authors updated the abstract and this addition has increased transparency. 2. Addressed: Page 3, lines 54-57, Abstract: The authors stated, “commonly cited reasons were data too complicated to use (n=5).” I think the word “commonly” should be avoided, considering the rather low n. ** The word "commonly" has been removed from the updated manuscript. 3. In the introduction, the authors state that their goal is to “understand…experiences with clinical research data…, as well as perceptions of the value, importance, and challenges of accessing data through BioLinCC..” While these are all important features, little justification is provided for the other sociodemographic characteristics collected (e.g. age, gender, and ethnicity). The authors should justify why these factors were studied and how they believe they are related to “perceptions of value, importance, and challenges.” *** Since the sociodemographic characteristics appear to be collected for descriptive purposes only, this could be prominently stated in the methods section. Under each “Survey Domain” listed, a brief justification of the importance of each question would be informative to the general clinical audience. 4. Page 5, line 37-40, Article Summary: There appears to be an out of place "." between the words "overcome" and "low." 5. Addressed: Page 10, line 31, Results: The authors state “Survey participation requests were thus sent to 485 eligible respondents, 44 of whom were subsequently excluded because of invalid contact information (n=31).” Here it is unclear what each of the numbers mean (n=44 and what is n=31). *** This concern has been clarified in the manuscript. 6. Addressed: Page 10, line 47-51, Results: The P-values reported appear to be from a chi-square test. A Fisher’s exact test may be more appropriate due to some variables having sparse data. *** The authors have included a statement about using a Fisher's exact test 7. Addressed: Page 10, line 29-38, Results: To follow up from comment #1 above, it is important to present both the response rate (44.2%) and the fact that out of the 536 investigators believed to have requested and received access to clinical research data, information was ultimately collected from 36.3%. *** The authors have included a statement about both response

	rates. 8. Page 11, line 54, Results: I agree with the initial response from the committee that there is no need to give p-values for the differences between responders and non-responders. These are not individual hypotheses that are being tested. 9. Page 14, line 33, Results. The authors may want to consider providing the actual number instead of "Of the 50%." 10. Addressed: In the survey, many questions were “check all that apply” (new research, replication research, or other). In the results section, it is not always clear when the survey respondents had the option to “select all that apply.” Considering the importance of study replication and validation, it would be extremely informative to report both how many selected ONLY replication research and how many respondents selected replication and any other reason. *** These questions/concerns have been addressed. 11. Addressed: Page 16, lines 15+, Discussion: The low response rate is my primary concern... Not only is there the possibility of social desirability bias, but also of recall bias (2007 is now almost 10 years ago). I appreciate that the authors spend a significant portion of the paper discussing this limitation. But it may also be worth mentioning the specific response rate from the two previous articles completed by Rathi et al (The value in this paper is very close to the 46% from “Predictors of clinical trial data sharing: exploratory analysis of a cross-sectional survey” from Rathi et al. 2014). The authors could also discuss response rates from different fields of study in order to provide a greater perspective of the relative magnitude of the rate reported in this study. *** There is now a discussion in the text about recall bias and a statement about response rates from other surveys. 12. Addressed: Page 16, line 49, Discussion: The authors state that they used a “limited survey scope to reduce response burden.” While the survey scope may have been limited to a certain area, I think this phrase may underemphasize the burden of a 50-item survey. *** The part about the limited survey scope has been removed. 13. Page 16/17, Discussion: It is worth mentioning some of the other repositories and discussing why the findings may or may not be applicable to the experience of investigators obtaining data from other repositories. Without further information, this limitation may either underemphasize or even overemphasize the scope of this limitation. *** This limitation was not addressed in detail. The discussion section would be strengthened if it included some mention of the other repositories. 14. Addressed: Page 30, Figure 2 & 3: To make this figure clearer, the authors should clarify that this comes from a “check all that apply question.” ***The figures have been updated

REVIEWER	Matthew Sydes MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, UCL, London, UK
REVIEW RETURNED	11-Jul-2016

GENERAL COMMENTS	===MAJOR=== 1. :: Section :: Introduction :: Comment :: More information is needed about what BioLINCC is. Does it contain RCT data only, or just non-RCT data, or both (and in what proportion)? What sort of studies are in there? How far back do the data go? Did the patients explicitly consent to this? Are they only US data? How do the characteristics of the studies that have been requested differ from the studies that have not been requested? Etc 2. :: Section :: Introduction :: Comment :: How similar are the datasets in BioLINCC? Is there any reason to think that the experience with one dataset would translate to an experience with another dataset? 3. :: Section :: Results :: Comment :: The low response rate is a major weakness, albeit one that the authors recognise. How can the reader be reassured that the respondents are representative of all applicants in terms of the aims of the study? The manuscript provides reassurance over geography and the number of datasets, but there it is easy to imagine that people who took the trouble to respond have particularly good or poor responses with the system. The fact that more recent applicants were more likely to respond is also a little troubling. ===MODERATE=== 4. :: Section :: Methods :: Comment :: Was the invitation also sent to applicants who had been unsuccessful in their request for data? 5. :: Section :: Results :: Comment :: Returning to the low response rate, reminders were in the same modality as the invitation (email). Could an alternative modality of reminder have been included as this might have improved response rates. 6. :: Section :: Results :: Text ref :: "Insufficient time for primary data collection (n=64; 33%)" :: Comment :: Presumably this motivation is taken from a list. I would not have predicted this. It makes me think all the more that the reader needs to understand BioLINCC. 7. :: Section :: Results :: Text ref :: "Fewer than one in five (n=36; 18%) respondents indicated that they had contacted the original study investigators to

	obtain data prior to requesting the data from BioLINCC. … Among the 20 (56%) respondents who indicated that the original study investigator denied their request, the most common response given by the original investigator was to direct the respondent to BioLINCC (n=11; 55%)." :: Comment :: Interesting if investigators had the option to access data directly from investigators rather than from BioLINCC. It seems an odd situation if some investigators denied access to the data so the applicants went behind their backs to get the data from somewhere else. Is this what is meant here? It would be interesting to know what datasets stored in BioLINCC had been shared directly by investigators during this period without reference to BioLINCC. I imagine this data cannot easily be collected or included, but it would be interesting to know why some people did not go to BioLINCC. 8. :: Section :: Results :: Comment :: It would be useful to understand by year of application whether research projects have yet been published. 9. :: Section :: Results :: Comment :: What sort of studies have been shared out of BioLINCC? ===TRIVIAL/MINOR=== 10. :: Section :: Methods :: Text ref :: "We conducted a cross-sectional survey from May to August 2015" :: Comment :: There is quite a delay between the survey and the submission. Could more up to date information have been used? 11. :: Section :: Discussion :: Comment :: Going forward, could the BioLINCC team mandate that applicants should complete a survey on their experience at the end of an unsuccessful application or after 6 months from a successful application? This would provide detailed information for the future which may be more complete. 12. :: Section :: References :: Comment :: I am a little disappointed not to see both the MRC CTU at UCL experiences and the CSDR experience of sharing trial data listed among the references as large studies that have discussed motivations of applicants for data and have consider other data resources. 13. :: Section :: Table 1 :: Comment :: For Request Year, does this have 1df? The default in Stata gives 8df and the p-value matches that shown here. Is this an issue and is it an issue for other parts of this table where the categories are ordered?

REVIEWER	William Wood University of North Carolina, Chapel Hill, USA
REVIEW RETURNED	12-Jul-2016

GENERAL COMMENTS	This is a thoughtful and interesting analysis of the way in which investigators have used an open NHLBI-maintained data repository to conduct research. The methods and analysis are clear and the conclusions drawn from this investigation are appropriate. As a very minor comment, the fourth bullet in the article summary contains a sentence fragment and should be revised. To orient readers who may not be familiar with BioLINCC, the authors might consider providing examples of the kinds of data that are contained within this repository, and the kinds of peer-reviewed, published analyses that survey respondents performed after accessing these data. It would also be of some interest to know the impact factors and citations associated with the publications that resulted from the BioLINCC analyses. Lastly, the authors might address some of the concerns that have been raised regarding data-sharing projects, and whether the survey respondents were asked about their awareness to and mitigation of these concerns? Overall this is a nice paper which will contribute to the literature.

VERSION 1 – AUTHOR RESPONSE

Reviewer #1:

General Comments: As I stated in the first review of the manuscript for the BMJ, this paper is a logical follow up to the 2012 BMJ article, “Sharing of clinical trial data among trialists: a cross sectional survey,” by a group of similar authors. This study aims to understand the experiences of investigators who requested and received access to clinical research data from BioLINCC. Considering the growing effort to encourage reproducibility and transparency practices in the scientific community, including public access and use of raw data, the overall scope of this project is novel and the findings are significant to a broad clinical audience. Furthermore, the article is well written and the methods described are appropriate. The primary concern that I brought up during the first review was the (low) overall survey response rate. This updated version of the manuscript addresses my primary concern and many of the minor concerns that all four of the reviewers shared.

Having studied the decision email from the BMJ, it is clear the all four peer reviewers thought favorably of the manuscript.

Below are some new comments and some of the comments from the previous review. For the comments that have already been addressed in the current manuscript draft, I did not update the page and line numbers.

Specific comments:

1. Addressed: Page 3, Line 38, Abstract: The opening of the results should state that there were 536 investigators who requested and received access to clinical research data from BioLINCC between 2007 and 2014. This gives a better sense of what the n= “441 potential respondents” actually means.

**The authors updated the abstract and this addition has increased transparency.

2. Addressed: Page 3, lines 54-57, Abstract: The authors stated, “commonly cited reasons were data too complicated to use (n=5).” I think the word “commonly” should be avoided, considering the rather

low n.

** The word "commonly" has been removed from the updated manuscript.

3. In the introduction, the authors state that their goal is to “understand…experiences with clinical research data…, as well as perceptions of the value, importance, and challenges of accessing data through BioLinCC.” While these are all important features, little justification is provided for the other sociodemographic characteristics collected (e.g. age, gender, and ethnicity). The authors should justify why these factors were studied and how they believe they are related to “perceptions of value, importance, and challenges.”

*** Since the sociodemographic characteristics appear to be collected for descriptive purposes only, this could be prominently stated in the methods section. Under each “Survey Domain” listed, a brief justification of the importance of each question would be informative to the general clinical audience.

Response: Thank you for this comment. We have modified the text to include justification of the importance of the questions (Survey Domains, pages 9-11).

“We used multiple response and yes/no questions to assess investigators’ primary research purpose and reasons for requesting data from BioLINCC. Multiple response questions were also used to determine the primary research objective, funding used to support the project, and other details of the planned research project. Knowing what these clinical research data are being used for will help tailor future data sharing efforts to the needs of investigators.”

“We used yes/no questions to determine whether original study investigators were contacted prior to or after requesting data through BioLINCC to obtain the data or to collaborate. These were followed by multiple response questions to determine why collaborations were sought, whether the requests for data or collaboration were approved, and reasons for not approving. Answers to these questions could potentially demonstrate the value of a data resource such as BioLINCC.”

“Multiple response, yes/no, and Likert-type questions were used to obtain information regarding investigator’s experience using BioLINCC, including whether the data were suitable and useful for their project. Knowledge gained from these questions can help to improve BioLINCC and other data sharing efforts.”

“We used multiple response and yes/no questions to characterize the completion stage of investigators’ projects. For those that did not complete their project, multiple response and yes/no questions were used to ascertain reasons why the project was incomplete. For those with completed projects, we used multiple response and yes/no questions to determine whether the final project differed from the pre-specified project as well as to obtain publication information. Multiple choice and multiple response questions were used to identify any funding sources and whether using the data from BioLINCC aided in any future grant applications. It is important to demonstrate not only that these data are being requested, but that they are also being used to potentially generate new knowledge to advance science and public health.”

“Respondents were asked to characterize their primary employer and career status using multiple choice questions, including whether they had ever been closely involved (as Principal or Co-Investigator) in the conduct of a randomized controlled trial and/or ever deposited clinical trial data in the BioLINCC repository. Respondent sociodemographic characteristics, including age, gender, and ethnicity, were also collected. While these characteristics were collected for descriptive purposes only, age, along with the professional characteristics collected, are of importance to demonstrate the value of the availability of BioLINCC data to investigators who are in certain stages of their career.”

4. Page 5, line 37-40, Article Summary: There appears to be an out of place "." between the words "overcome" and "low."

Response: Thank you for this comment. We have modified the text to clarify (Article Summary, page 5):

5. Addressed: Page 10, line 31, Results: The authors state “Survey participation requests were thus sent to 485 eligible respondents, 44 of whom were subsequently excluded because of invalid contact information (n=31).” Here it is unclear what each of the numbers mean (n=44 and what is n=31).

*** This concern has been clarified in the manuscript.

6. Addressed: Page 10, line 47-51, Results: The P-values reported appear to be from a chi-square test. A Fisher’s exact test may be more appropriate due to some variables having sparse data.

*** The authors have included a statement about using a Fisher's exact test

7. Addressed: Page 10, line 29-38, Results: To follow up from comment #1 above, it is important to present both the response rate (44.2%) and the fact that out of the 536 investigators believed to have requested and received access to clinical research data, information was ultimately collected from 36.3%.

*** The authors have included a statement about both response rates.

8. Page 11, line 54, Results: I agree with the initial response from the committee that there is no need to give p-values for the differences between responders and non-responders. These are not individual hypotheses that are being tested.

Response: As we explained in our initial response to the editorial committee, it is standard practice in survey research articles to compare responders to non-responders, using statistical tests and reporting the p values, despite no survey being powered explicitly to detect differences between responders and non-responders.

9. Page 14, line 33, Results. The authors may want to consider providing the actual number instead of "Of the 50%."

Response: Thank you for this comment. We have modified the text (Results, page 14):

Of the 97 respondents (50% of total) who have not yet completed their proposed projects, 84% (n=81) explained that they planned to complete their project; 65% (n=63) indicated that their project is in analysis/manuscript draft phase, while 28% (n=27) explained that they have thus far been too busy with other responsibilities to complete the research project using the data from BioLINCC and 13% (n=13) reported that lack of funding to support the project was a problem (Figure 3).

10. Addressed: In the survey, many questions were “check all that apply” (new research, replication research, or other). In the results section, it is not always clear when the survey respondents had the option to “select all that apply.” Considering the importance of study replication and validation, it would be extremely informative to report both how many selected ONLY replication research and how many respondents selected replication and any other reason.

*** These questions/concerns have been addressed.

11. Addressed: Page 16, lines 15+, Discussion: The low response rate is my primary concern... Not only is there the possibility of social desirability bias, but also of recall bias (2007 is now almost 10 years ago). I appreciate that the authors spend a significant portion of the paper discussing this limitation. But it may also be worth mentioning the specific response rate from the two previous articles completed by Rathi et al (The value in this paper is very close to the 46% from "Predictors of clinical trial data sharing: exploratory analysis of a cross-sectional survey" from Rathi et al. 2014). The authors could also discuss response rates from different fields of study in order to provide a greater perspective of the relative magnitude of the rate reported in this study.

*** There is now a discussion in the text about recall bias and a statement about response rates from other surveys.

12. Addressed: Page 16, line 49, Discussion: The authors state that they used a "limited survey scope to reduce response burden." While the survey scope may have been limited to a certain area, I think this phrase may underemphasize the burden of a 50-item survey.

*** The part about the limited survey scope has been removed.

13. Page 16/17, Discussion: It is worth mentioning some of the other repositories and discussing why the findings may or may not be applicable to the experience of investigators obtaining data from other repositories. Without further information, this limitation may either underemphasize or even overemphasize the scope of this limitation.

*** This limitation was not addressed in detail. The discussion section would be strengthened if it included some mention of the other repositories.

Response: We appreciate the reviewer's continued interest in our clarifying how our findings may be generalizable to the experience of investigators making use of other repositories. While we initially felt constrained by space limitations to further expound on this issue, we have now added a new paragraph of text in which we address this issue (Discussion, pages 18-19):

"Second, our study was limited to investigators who had received data from BioLINCC and our findings may not be applicable to the experience of investigators obtaining data from other repositories. There is currently great interest and scrutiny of existing clinical trial data sharing efforts,21-24 many of which require submission of a research proposal, as does BioLINCC, but which nearly always only make data available via a virtual, secure data sharing environment, as opposed to BioLINCC which provides de-identified data directly to approved researchers. One recently study evaluated how many clinical trials were publicly available to the research community through 3 open access data sharing platforms: ClinicalStudyDataRequest.com, the Yale University Open Data Access (YODA) Project, and the Supporting Open Access for Researchers (SOAR) Initiative, finding that while more than 3000 trials were available, only 15.5% had been requested by a limited number of investigators.25 The authors concluded that data sharing efforts are being underutilized, implicitly questioning the value of continued resource investment. However, the results of our survey of BioLINCC users suggests this conclusion may be premature, as use of data from these open access platforms can be expected to grow with time, although more remains to ensure the use of these data, and the successful completion and publication of the resulting research, to justify the investments being made in data sharing."

14. Addressed: Page 30, Figure 2 & 3: To make this figure clearer, the authors should clarify that this comes from a "check all that apply question."

***The figures have been updated

Reviewer #2

Specific comments:

===MAJOR===

1.
:: Section :: Introduction
:: Comment :: More information is needed about what BioLINCC is. Does it contain RCT data only, or just non-RCT data, or both (and in what proportion)? What sort of studies are in there? How far back do the data go? Did the patients explicitly consent to this? Are they only US data? How do the characteristics of the studies that have been requested differ from the studies that have not been requested? Etc

Response: We appreciate the reviewer's interest in more information about BioLINCC. We have added information to our introductory text to provide further background, and we continue to cite to key references for interested readers to learn more (Introduction, page 7). However, it was beyond the scope of this paper to examine how studies that were and were not requested differed with respect to key characteristics.

"While most of these data sharing efforts have been relatively newly established, the U.S. National Heart, Lung, and Blood Institute (NHLBI) of the NIH established a formal data repository in 2000, now managed by the Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC), to facilitate access to, maximize the scientific value of, and promote the availability and use of the biorepository, data repository and other NHLBI-funded population-based biospecimen and data resources by investigators worldwide.12 13 The BioLINCC data repository includes individual level data on more than 580,000 participants from over 110 Institute supported clinical trials and observational studies, beginning as far back as the 1980s. Each data set is prepared independently by the NHLBI-funded investigator to comply with specific requirements and data standards, with oversight by BioLINCC, including provision of baseline, interim visit, ancillary study and outcome data for clinical trials and provision of all examination and ancillary study data, along with follow-up information, for epidemiology studies."

2.
:: Section :: Introduction
:: Comment :: How similar are the datasets in BioLINCC? Is there any reason to think that the experience with one dataset would translate to an experience with another dataset?

Response: NHLBI provides guidelines for how datasets should be prepared prior to submission to the BioLINCC repository. However, this does not ensure exact concordance in the preparation of all datasets; thus, experience may differ among datasets. We have addressed this issue in the preceding response.

3.
:: Section :: Results
:: Comment :: The low response rate is a major weakness, albeit one that the authors recognise. How can the reader be reassured that the respondents are representative of all applicants in terms of the aims of the study? The manuscript provides reassurance over geography and the number of datasets, but there it is easy to imagine that people who took the trouble to respond have particularly good or

poor responses with the system. The fact that more recent applicants were more likely to respond is also a little troubling.

Response: We appreciate the reviewer's concern. We achieved a response rate of 44%, which compares favorably with other surveys of physicians and researchers. Moreover, we used several mechanisms to prospectively improve response rates, including a web-based survey platform for ease of completion, we employed several reminder contacts, including three e-mails and at least one telephone contact, and we offered financial incentives for participation. We clearly note the survey response rate as a limitation of our article, and explain how it may have biased our results, if at all (Discussion, pages 17-18).

===MODERATE===
4.
:: Section :: Methods
:: Comment :: Was the invitation also sent to applicants who had been unsuccessful in their request for data?

Response: The invitation was only sent to those who successfully obtained data from BioLINCC. No information was available for investigators who had not been successful in obtaining data from BioLINCC, although based on personal communication with BioLINCC staff, nearly every application for data is approved.

5.
:: Section :: Results
:: Comment :: Returning to the low response rate, reminders were in the same modality as the invitation (email). Could an alternative modality of reminder have been included as this might have improved response rates.

Response: In addition to email contact, non-respondents were contacted by telephone to solicit their participation up to twice per week, but no more than once per day, until one contact was made. This process is described in our manuscript (Methods, page 8).

6.
:: Section :: Results
:: Text ref :: "Insufficient time for primary data collection (n=64; 33%)"
:: Comment :: Presumably this motivation is taken from a list. I would not have predicted this. It makes me think all the more that the reader needs to understand BioLINCC.

Response: As described in our Methods, survey items were presented in multiple response, Likert scale, and open-ended formats; many of the multiple response questions enabled respondents to select multiple answers, including this response flagged by the reviewer. We provided this item because, in our background preparation of the survey, the lack of time and resources to collect data were frequently mentioned by investigators who wanted to make use of shared clinical trial data for their own research projects.

7.
:: Section :: Results
:: Text ref :: "Fewer than one in five (n=36; 18%) respondents indicated that they had contacted the original study investigators to obtain data prior to requesting the data from BioLINCC. … Among the 20 (56%) respondents who indicated that the original study investigator denied their request, the most common response given by the original investigator was to direct the respondent to BioLINCC (n=11; 55%)."

:: Comment :: Interesting if investigators had the option to access data directly from investigators rather than from BioLINCC. It seems an odd situation if some investigators denied access to the data so the applicants went behind their backs to get the data from somewhere else. Is this what is meant here? It would be interesting to know what datasets stored in BioLINCC had been shared directly by investigators during this period without reference to BioLINCC. I imagine this data cannot easily be collected or included, but it would be interesting to know why some people did not go to BioLINCC.

Response: We agree with the reviewer, it would be interesting to better understand the communications between investigators requesting to use shared clinical trial data and the investigators who originally collected that data. Presumably, all data made available via BioLINCC could also be requested directly from the original investigators – those investigators are generally well known in the field. However, a survey to address this issue is beyond the scope of the current paper.

8.
:: Section :: Results
:: Comment :: It would be useful to understand by year of application whether research projects have yet been published.

Response: We agree with the reviewer and have added this information to the results (page 14):

“Half of all respondents (n=98; 50%) reported that their projects have been completed, of which 67% (n=66) have been published. Respondents who had requested data prior to 2012 were more likely to have completed their project when compared with those who had requested data in 2012 or afterwards (73% versus 44%; p=0.008). However, among those who completed their project, rates of publication did not differ among those who had requested data prior to 2012 and those who had requested data in 2012 or afterwards (63% versus 69%; p=0.57).”

9.
:: Section :: Results
:: Comment :: What sort of studies have been shared out of BioLINCC?

Response: Unfortunately, we did not collect this information. However, a list of publications that have resulted from use of this shared data is made available on the BioLINCC website: https://biolincc.nhlbi.nih.gov/publications/. We have revised the text to include this information (Discussion, page 17):

“Moreover, even among completed projects, only two-thirds were published. While BioLINCC maintains an updated list of publications that have resulted from use of this shared data,14 mechanisms should be established to ensure that results from research made possible through data sharing are publicly disseminated, either through publication or through a results reporting initiative similar to ClinicalTrials.gov.”

===TRIVIAL/MINOR===

10.
:: Section :: Methods
:: Text ref :: "We conducted a cross-sectional survey from May to August 2015"
:: Comment :: There is quite a delay between the survey and the submission. Could more up to date information have been used?

Response: We appreciate the reviewer’s criticism, but note that our article was originally submitted to the BMJ in February of 2016, less than 6 months after survey administration.

11.
:: Section :: Discussion
:: Comment :: Going forward, could the BioLINCC team mandate that applicants should complete a survey on their experience at the end of an unsuccessful application or after 6 months from a successful application? This would provide detailed information for the future which may be more complete.

Response: This is a useful suggestion that we will convey to our colleagues at BioLINCC.

12.
:: Section :: References
:: Comment :: I am a little disappointed not to see both the MRC CTU at UCL experiences and the CSDR experience of sharing trial data listed among the references as large studies that have discussed motivations of applicants for data and have consider other data resources.

Response: We appreciate the reviewer's comment. We had included several citations to the MRC CTU at UCL (citations 10 and 11), but have added additional citations that describe these efforts, as well as the experience with CSDR. Our response to reviewer #1's 13th comment is also pertinent to this comment.

13.
:: Section :: Table 1
:: Comment :: For Request Year, does this have 1df? The default in Stata gives 8df and the p-value matches that shown here. Is this an issue and is it an issue for other parts of this table where the categories are ordered?

Response: We analyzed the data ordinally, as presented in the table, using Fischer-Exact testing. Alternatively, we could have collapsed this data into 2 categories, requests submitted prior to 2012 and those submitted in 2012 and afterwards. We would still have found that respondents were more likely to have submitted requests in 2012 and afterwards (77% versus 60%; $p < 0.001$). We retained the Table as it was originally submitted in order to provide added detail for interested readers.

Reviewer #3

General Comments: This is a thoughtful and interesting analysis of the way in which investigators have used an open NHLBI-maintained data repository to conduct research. The methods and analysis are clear and the conclusions drawn from this investigation are appropriate.

As a very minor comment, the fourth bullet in the article summary contains a sentence fragment and should be revised.

Response: Thank you for this comment. We have modified the text to clarify (Article Summary, page 5):

To orient readers who may not be familiar with BioLINCC, the authors might consider providing examples of the kinds of data that are contained within this repository, and the kinds of peer-reviewed, published analyses that survey respondents performed after accessing these data. It would also be of some interest to know the impact factors and citations associated with the publications that resulted from the BioLINCC analyses. Lastly, the authors might address some of the concerns that have been raised regarding data-sharing projects, and whether the survey respondents were asked

about their awareness to and mitigation of these concerns?

Overall this is a nice paper which will contribute to the literature.

Response: We appreciate this comment and have made several revisions to the article in response, as explained in response to reviewer #2's 1st comment (BioLINCC details), reviewer #2's 9th comment (BioLINCC publications), and reviewer #1's 13th comment (other data sharing projects), wherein these issues were addressed.

VERSION 2 – REVIEW

REVIEWER	Joshua David Wallach Stanford University I was contacted by the BMJ to review the original version of this manuscript before I had met Dr. Ross and Dr. Krumholz. I completed the initial review without any competing interests. Since that time, I interviewed for a position at Yale University, reporting to Dr. Ross. This position has an official start date of January 3rd, 2017. I have remained as transparent as possible about this competing interest. In particular, I emailed the editorial office before both revisions to clarify the status of this conflict. My initial review for the paper, prior to my knowledge of the position/interview at Yale, was generally positive. I still believe that I can provide an impartial review.
REVIEW RETURNED	26-Aug-2016

GENERAL COMMENTS	I appreciate the opportunity to review this manuscript for the third time. The updated manuscript now addresses all of the major concerns that were discussed during the first and second reviews. Based on the manuscript and author's response section, it is also clear that the authors have considered the major suggests from the other reviews. In particular: Page 7, line 11, introduction: This new section provides useful information for a general medical audience. Now the reader can have a better understanding of the BioLINCC data repository. I appreciate the new paragraph of the text (Discussion, pages 18-19) that discusses how the findings may be generalizable to the experience of investigators making use of other repositories. I believe that this is an important addition to the text. One minor comment: On page 14, the authors have also added information "by year of application." While I agree that this is interesting information, the authors might want to make note of this analysis in the methods section. Especially since this may be a secondary (post-hoc) analysis. Overall, I believe that this is now a strong paper that will be an important contribution to the literature.