PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Status, use and impact of sharing Individual Participant Data from clinical trials: a scoping review
AUTHORS	Ohmann, Christian; Moher, David; Siebert, Maximilian; Motschall, Edith; Naudet, Florian

VERSION 1 – REVIEW

REVIEWER	Taylor, Rod University of Exeter, Institute of Health Research
REVIEW RETURNED	03-Apr-2021

GENERAL COMMENTS	This paper provides a scoping review that aims to summarise the empirical evidence around for impact of IPD data sharing strategies for clinical trials from a group of recognised authors in this area. Given that IPD data sharing is one of the key challenges currently facing the global clinical trial community to ensure we maximise research efficient/output minimise research waste, this paper is an important one and likely to be great interest to the readers of BMJ Open and highly cited. This paper also provides a very useful source of the empirical literature on the impact of IPD data sharing strategies. However, there are a couple of key areas for the consideration by the author team. 1. Context – the authors should include in the introduction section an overview/taxonomy of current data sharing initiatives with relevant timeline. Not only would this help set the context and scope for this paper for reader but also aid them navigate the remainder of the paper, especially results section. The authors refer to the “impact” of data sharing initiatives, but in its current form the manuscript does not provide a hint for an association (at very least a temporal one) between metrics reported and the launch of the “causative” initiatives. 2. Results – the results section is currently long and rather unwieldy for the reader. It would help the readability of the paper if the authors can revisit this section, especially in terms of inclusion of point 1 above. This section is focussed on summarising results from the individual studies identified from the literature but one misses, for instance with respect to actual data sharing, the big picture: how data sharing is implemented in practice? Is direct contact with the trialists more effective than gathering through a repository or platform? What about use of protected computing environments for the analyses? In some cases statistical codes are allowed to be run by the data owner internally and only results

	shared with the applicant. Would these practices qualify as data sharing? Other comments 3. Abbreviations - the authors introduce a number of abbreviations (e.g. CDSR/YODA) without statement in full on first citation. 4. Figures – I found the figures of mixed value. If they stay are as main figures within the paper, the authors need to revisit these graphics to in order make them more accessible to the reader. Figure 2: clear explanation of type of study; Figure 3: what does the % axes mean? Not clear what the n/N data represents; Figure 6: CDSR/Yoda – label so clearer (orange = DCRD & blue = YODA?). grouping Grouping don't make sense to me: isn't 'meta-analysis' is nor a form of 'secondary analysis'. ? 5. Grammar & language – needs checked/corrected throughout e..g Discussion – summary of evidence: "….a bit less…." 6. "So far only some IPD meta-analyses have been planned as part of data sharing initiatives and a few reported." Clarify what is meant by this statement.

REVIEWER	Hong, Kyungwan University of Maryland Baltimore, Department of Pharmaceutical Health Services Research
REVIEW RETURNED	22-Apr-2021

GENERAL COMMENTS	Thank you for the opportunity to review this manuscript. The authors thoroughly reviewed an extensive amount of literature and summarized the previous/current practice of individual participant data sharing. The authors also shed light on the discrepancy between the attitude towards data sharing among different stakeholders in biomedical research and some areas of improvement. The used methodology seems sound overall, and the results support the discussion and argument. Therefore, this study will be valuable to researchers and policymakers who would like to study and understand the current landscape of clinical trial data sharing. However, this manuscript lacks key details and prevents readers from understanding the core concept more clearly. Also, I think this manuscript requires a bit of polishment. I believe several things should be addressed in a revised version. Please find my specific comments as follows: Abstract: -Page 3, Line 23-24: If word counts are allowed, how about adding the published year range of those 93 studies? Strengths and limitations of this study: -Page 4, Line 6-7: Instead of "most important," I think the term "major" fits better here. Introduction: -Throughout the introduction: I think clinical trial data sharing and IPD data sharing need to be differentiated as clinical trial data may also refer to protocols, clinical study reports, statistical analysis plans, and case report forms.

	Method: -Page 6, line14: The authors stated, even unpublished reports were eligible for this study. However, I am confused about how the authors did data extraction on those unpublished reports (e.g., basic information on the paper). Did the authors use ClinicalTrials.gov or other trial registries to extract such information? Or those "unpublished reports" mean "pre-print" or "non-indexed"? Clarification is needed. -Page 7, Line 20, Platform: What about other data-sharing platforms, especially platforms operated by individual pharmaceutical companies? -Page 7, line 44-45: Please provide the rationale why you chose these journals. Is it based on the impact factor? Why were other major medical journals such as the NEJM, JAMA, and/or the Lancet not included/contacted? Results: -Throughout the results: I think the results section is nicely written and thorough. However, some paragraphs are too long (especially those on pages 10 and 13). How about splitting some of those long paragraphs or make the results section a bit more concise to improve the readability? -Page 10, Line 8-9: The authors stated, "there was a high risk of bias." Is this statement based on a formal risk of bias statement? I am confused as the methods section did not provide details on the risk of bias assessment. Discussion/Limitations/Conclusions: -Page 22, Line 8-11: The authors suggested that potential reasons for the limited IPD data sharing are the lack of widely accepted repositories for non-commercial clinical trials and insufficient incentives and benefits related to data sharing. Those are valid reasons, but it is important to note that different stakeholders in biomedical research have different concerns about IPD data sharing… Therefore, I think this section needs a bit more detail. -Page 22, Line 11-12: Can you elaborate further on "One issue is that not for all projects the publications from secondary are regularly updated so that statistics may be biased."? -Any thoughts on different pharmaceutical companies' initiatives to share their clinical trial IPD? Tables/Figures/Supplementary materials: -Tables: I would recommend putting table titles on the top of tables, not the bottom. -Figure 1: Consider providing a version with a higher resolution -Figure 3: Consider merging or filling with a different color in the rows that represent "For trialists," "For publishers/funders," "For trial participants," and so on to differentiate them better and improve readability. -Figure 4: What's the difference between data sharing policy vs. data sharing requirement? -Figure 6: Nice looking figure, but I would recommend putting the "N=" in the boxes of meta-analyses, methodological, re-analyses, and secondary analyses to improve the readability.

	-Supplementary material 3: Maybe put “et al.” next to each author or change the column title to “the first author” instead of just “author”? -Supplementary material 3: What do you mean by “Broader” under the type of shared material? I would recommend using a footnote to explain what that term means. -Supplementary material 3: What are those “0”s in front of “IPD,” starting from Vassar M? -Supplementary material 4: If I understood correctly, I would recommend revising the last column title to “assessment period.” On a similar note, were the first two studies by Ross et al. cross-sectional studies? I am a bit confused by the assessment time of those two articles. They should be periods, not a single month unless those two studies only evaluated published studies in those months (e.g., 6/2017 and 8/2018). Please double-check them. Also, please check the availability of other grammatical errors to strengthen your argument and clarify your delivery. Some of the errors that I identified are following: -Throughout the manuscript: Although the abbreviations are explained on page 31, I would still recommend describing what different acronyms stand for (only the first one). For example, “YODA,” “CSDR,” and “NIH” were used without explaining what those acronyms mean (Page 5, Line 45-46) Also, “PhRMA/EFPIA,” “IOM,” and “RDA” on page 7, line 11-12. -Throughout the manuscript: Please use consistent terms. For example, “ClinicalTrials.gov”/”CT.gov” and “ViVli”/“vivli” were used interchangeably. -Page 5, Line 21-22: Describe what “IPD” stands for here, not on Page 5, line 27-28. -Page 5, Line 28-29: It should be “COVID-19.” -Page 5, Line 45-46: It should be “CSDR” (There are a total of 5 “CDSR” in the manuscript. Correct these as well) -Page 6, Line 57-58: Add “,” (comma) after “not)” -Page 7, Line 9-10: Capitalize “I” in the initiative. -Page 7, Line 45-46: Add “)” after F1000Research. -Page 10, Line 42-44: I would stick to one format of “percent.” For example, “4%” and “nine percent” were used here. -Page 11, Line 28: “For the use case HIV.” Did you mean “For the case of HIV”? -Page 14, Line 15-16: Acronym “US” was already used, but “The United States” appears. -Page 18, Line 12-13: It should be “GlaxoSmithKline.”

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Prof. Rod Taylor, University of Exeter

Context – the authors should include in the introduction section an overview/taxonomy of current data sharing initiatives with relevant timeline. Not only would this help set the context and scope for this paper for reader but also aid them navigate the remainder of the paper, especially

results section. The authors refer to the “impact” of data sharing initiatives, but in its current form the manuscript does not provide a hint for an association (at very least a temporal one) between metrics reported and the launch of the “causative” initiatives.

We agree with the reviewer and have reworded the introduction by providing key dates for the most important initiatives. It gives now a better idea of the timeline. We already refer in the introduction to the fact that these initiavives are early initiatives. We have deleted a reference to NIH (in 2003 it recognized data sharing value) as this encompass various initiatives that can hardly be summarized without complexifying the text.

Results – the results section is currently long and rather unwieldy for the reader. It would help the readability of the paper if the authors can revisit this section, especially in terms of inclusion of point 1 above. This section is focussed on summarising results from the individual studies identified from the literature but one misses, for instance with respect to actual data sharing, the big picture: how data sharing is implemented in practice? Is direct contact with the trialists more effective than gathering through a repository or platform? What about use of protected computing environments for the analyses? In some cases statistical codes are allowed to be run by the data owner internally and only results shared with the applicant. Would these practices qualify as data sharing?

This comment was difficult to address without changing the whole structure of the manuscript. In this paper results are structured according to the five outcome domains and different stakeholder

groups. For the main subgroups, the results are summarised in figures. Then, each study is summarised in in the text. To make the text more readable, we have included subheadings of the sections. In addition, we have edited the text by a native speaker to make it clearer. This was all what we could do from the viewpoint of the authors.

Other comments
Abbreviations - the authors introduce a number of abbreviations (e.g. CDSR/YODA) without statement in full on first citation.

Has been performed throughout the manuscript.

Figures – I found the figures of mixed value. If they stay are as main figures within the paper, the authors need to revisit these graphics to in order make them more accessible to the reader. Figure 2: clear explanation of type of study;

Thank you, this has been added in the legend.

Figure 3: what does the % axes mean? Not clear what the n/N data represents;

We suspect that the comment rather refers to figure 4 and 5. This is indeed a good point. The percentage axis refers to the percentage identified for the outcome of interest described in the paper. We have changed the axis label. It reads now: “Percentage for the outcome”. The n/N correspond to the numbers of case with the outcome/number of cases in each reference. The legend has been edited.

Figure 6: CDSR/Yoda – label so clearer (orange = DCRD & blue = YODA?). grouping Grouping don’t make sense to me: isn’t ‘meta-analysis’ is nor a form of ‘secondary analysis’. ?

Of course, meta-analysis is a form of secondary analysis. However, these studies pool different independent studies together. Secondary analysis are rather exploratory analysis of a dataset resulting from one study. These categories are usually used in all the references we found and we have decided to keep these categories. We have added a legend: "- Blue: YODA, - Red: CSDR"

Grammar & language – needs checked/corrected throughout e..g Discussion – summary of evidence: "….a bit less…."

The manuscript has been checked and improved by a native English speaker.

"So far only some IPD meta-analyses have been planned as part of data sharing initiatives and a few reported." Clarify what is meant by this statement.

This has been clarified and the text has been changed into "So far only some IPD meta-analyses have been specified as reason for data sharing requests to repositories and a few reported".

Reviewer: 2
Dr. Kyungwan Hong, University of Maryland Baltimore

Abstract:
-Page 3, Line 23-24: If word counts are allowed, how about adding the published year range of those 93 studies?

Thank you for this comment. We have added this information in the abstract.
Strengths and limitations of this study:
-Page 4, Line 6-7: Instead of "most important," I think the term "major" fits better here.

We changed this according to the reviewer's suggestion.

Introduction:
-Throughout the introduction: I think clinical trial data sharing and IPD data sharing need to be differentiated as clinical trial data may also refer to protocols, clinical study reports, statistical analysis plans, and case report forms.

This is an important point and was clarified at the beginning of the introduction.

Method:
-Page 6, line14: The authors stated, even unpublished reports were eligible for this study. However, I am confused about how the authors did data extraction on those unpublished reports (e.g., basic information on the paper). Did the authors use ClinicalTrials.gov or other trial registries to extract such information? Or those "unpublished reports" mean "pre-print" or "non-indexed"? Clarification is needed.

Thank you, we have clarified this point in the manuscript.
-Page 7, Line 20, Platform: What about other data-sharing platforms, especially platforms operated by individual pharmaceutical companies?

Thank you for this comment. In theory, such platforms may fit in our definition of platforms but we did not find such platforms. In general, pharmaceutical industry used Vivli, CSDR and YODA. In case they don't use this, they proceed with a dedicated team for data sharing. Data files are directly shared without a

dedicated platform and there is no in general no specific metrics. We therefore did not change our text (but we corrected a typo in Initiatives).

-Page 7, line 44-45: Please provide the rationale why you chose these journals. Is it based on the impact factor? Why were other major medical journals such as the NEJM, JAMA, and/or the Lancet not included/contacted?

NEJM, JAMA and/or The Lancet had no data sharing policy before the ICMJE policy was in place in 2018. We clarified this point in the method section.

Results:
-Throughout the results: I think the results section is nicely written and thorough. However, some paragraphs are too long (especially those on pages 10 and 13). How about splitting some of those long paragraphs or make the results section a bit more concise to improve the readability?

As explained above (see reviewer 1), we have included subheadings of the sections to make the text more readable. In addition, we have edited the text by a native speaker to make it clearer.

-Page 10, Line 8-9: The authors stated, “there was a high risk of bias.” Is this statement based on a formal risk of bias statement? I am confused as the methods section did not provide details on the risk of bias assessment.

This has been clarified in the text.

Discussion/Limitations/Conclusions:
-Page 22, Line 8-11: The authors suggested that potential reasons for the limited IPD data sharing are the lack of widely accepted repositories for non-commercial clinical trials and insufficient incentives and benefits related to data sharing. Those are valid reasons, but it is important to note that different stakeholders in biomedical research have different concerns about IPD data sharing… Therefore, I think this section needs a bit more detail.

Thank you for this remark. In general, we wanted the discussion to be as short as possible as the results are rather detailed. We have slightly edited this part.

-Page 22, Line 11-12: Can you elaborate further on “One issue is that not for all projects the publications from secondary are regularly updated so that statistics may be biased.”?

Thank you for this comment. Our text was indeed rather unclear. We have tried to reformulate this part of the text.

-Any thoughts on different pharmaceutical companies’ initiatives to share their clinical trial IPD?

These initiatives exist but it is very hard to get some information/metrics. We have added a few words about this in our limitation section.

Tables/Figures/Supplementary materials:
-Tables: I would recommend putting table titles on the top of tables, not the bottom.

Has been performed.

-Figure 1: Consider providing a version with a higher resolution

We have tried to provide a better version (it is in pdf format with a high resolution). If it does not work, we will be pleased to work with the managing editor in order to improve the resolution.

-Figure 3: Consider merging or filling with a different color in the rows that represent “For trialists,” “For publishers/funders,” “For trial participants,” and so on to differentiate them better and improve readability.

For clarity purpose, we suppressed the colours but have added lines in the table. We hope that it is clearer now. We believe that all the information in this table is important because it details all the outcomes that are used in this literature.

-Figure 4: What’s the difference between data sharing policy vs. data sharing requirement?

Thanks, the clarification was needed. A data sharing policy can be either “supportive” (encouraging data sharing) or “mandatory” (i.e. data sharing requirement). Most paper are about encouraging policies. In DeVito 2018, we could extract some information about mandatory policies. We have labelled this as data sharing requirement. We have edited the legend.

-Figure 6: Nice looking figure, but I would recommend putting the “N=” in the boxes of meta- analyses, methodological, re-analyses, and secondary analyses to improve the readability.

Thank you for this suggestion. We made the required changes.

-Supplementary material 3: Maybe put “et al.” next to each author or change the column title to “the first author” instead of just “author”?

Thank you, this was done.

-Supplementary material 3: What do you mean by “Broader” under the type of shared material? I would recommend using a footnote to explain what that term means.

Thank you, this was done. “Broader: the definition is not solely restricted to IPD and can cover other type of additional material (e.g. study protocol, code, etc).”
-Supplementary material 3: What are those “0”s in front of “IPD,” starting from Vassar M?

It was an error as it should have been avoided. This has been corrected. Thank you.

-Supplementary material 4: If I understood correctly, I would recommend revising the last column title to “assessment period.” On a similar note, were the first two studies by Ross et al. cross- sectional studies? I am a bit confused by the assessment time of those two articles. They should be periods, not a single month unless those two studies only evaluated published studies in those months (e.g., 6/2017 and 8/2018). Please double-check them.

This is rather the date when the results from the repository were assessed. We therefore edit the column title with: “Date of assessment”

Also, please check the availability of other grammatical errors to strengthen your argument and clarify your delivery. Some of the errors that I identified are following:

-Throughout the manuscript: Although the abbreviations are explained on page 31, I would still recommend describing what different acronyms stand for (only the first one). For example, “YODA,” “CSDR,” and “NIH” were used without explaining what those acronyms mean (Page 5, Line 45-46) Also, “PhRMA/EFPIA,” “IOM,” and “RDA” on page 7, line 11-12.
-Throughout the manuscript: Please use consistent terms. For example, “ClinicalTrials.gov”/”CT.gov” and “ViVli”/“vivli” were used interchangeably.
-Page 5, Line 21-22: Describe what “IPD” stands for here, not on Page 5, line 27-28.
-Page 5, Line 28-29: It should be “COVID-19.”
-Page 5, Line 45-46: It should be “CSDR” (There are a total of 5 “CDSR” in the manuscript. Correct these as well) -Page 6, Line 57-58: Add “,” (comma) after “not)”
-Page 7, Line 9-10: Capitalize “I” in the initiative.
-Page 7, Line 45-46: Add “)” after F1000Research.
-Page 10, Line 42-44: I would stick to one format of “percent.” For example, “4%” and “nine percent” were used here.
-Page 11, Line 28: “For the use case HIV.” Did you mean “For the case of HIV”?

-Page 14, Line 15-16: Acronym “US” was already used, but “The United States” appears.
-Page 18, Line 12-13: It should be “GlaxoSmithKline.”

All these points have been taken into consideration and corrected.

Reviewer: 1
Competing interests of Reviewer: None

Reviewer: 2
Competing interests of Reviewer: The Laura and John Arnold Foundation funds the RIAT Support Centre (no grant number), which supports the salaries of Kyungwan Hong.

Kyungwan Hong's projects were supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award U01FD005946 totaling US$5,000 with 100 per cent funded by FDA/HHS. The project contents are those of Kyungwan Hong and do not necessarily represent the official views of, nor an endorsement by, FDA/HHS, or the U.S. Government.

VERSION 2 – REVIEW

REVIEWER	Taylor, Rod University of Exeter, Institute of Health Research
REVIEW RETURNED	10-Jul-2021

GENERAL COMMENTS	the authors has responded well to reviewer's comments and update the manuscript accordingly

REVIEWER	Hong, Kyungwan

	University of Maryland Baltimore, Department of Pharmaceutical Health Services Research
REVIEW RETURNED	06-Jul-2021

GENERAL COMMENTS	Summary: Thank you for the opportunity to re-review this manuscript. This manuscript thoroughly reviewed extensive literature and summarized the previous/current practice of individual participant data sharing. The authors also identified the discrepancy between the attitude towards data sharing and the rate of actual data sharing, shedding light on some areas of improvement in data transparency and open science. The used methodology seems sound overall, and the results support the discussion and argument. Therefore, this study will be valuable to researchers and policymakers who would like to understand the current landscape of clinical trial data sharing. Compared to the previous version, this newly submitted version has improved a lot in delivery and readability. I also thank the authors for taking my comments/suggestions seriously and addressing them accordingly. I only have a few minor comments for the authors' consideration. Please find my specific comments as follows: Revisions: Results (Page 16, Metrics of actual re-use): I was wondering if you explored how long it takes from submitting a data sharing request to receiving the requested data. If so, can you please mention that? From my past experience, the time it takes to receive the data can be burdensome. Discussion (Page 21, "In short, the pressure by publisher…"): Consider adding your own suggestions/recommendations for publishers and funders to improve their engagement to share data. Discussion (Page 23, "Research output from shared data should have an impact…"): Can you elaborate further on "impact"? I completely agree that more studies are needed to be done with shared clinical trial data, but that does not necessarily lead to a more powerful impact in biomedical research. Therefore, I am struggling to follow your point here. What are your suggestions to improve the impact of research projects using shared clinical trial data? Also, please check the availability of other minor grammatical errors to strengthen your argument and clarify your delivery. Some of the errors that I identified are following: -Throughout the manuscript: Please use consistent terms. For example, "ClinicalTrials.gov"/"clinicaltrials.gov" -Throughout the manuscript: Please make the referencing consistent. For example, references were mostly mentioned after a period (.xx), but sometimes references were mentioned before a period (xx.)

	-Introduction, Rationale (Page 5), acronym IPD should be described in the first paragraph, not in the third paragraph. (e.g., Besides the individual participant data (IPD) sets) -Method (Page 7, in the box) “project Data Sphere” --> “Project Data Sphere” -Results (Page 13), please add percentage after “280/2003”. -Results (Page 16, “Between 2013 and 2015 177”), please add a comma (,) between 2015 and 177. -Results (Page 17), Add period (.) after “154 (66%) were approved”. -Results (Page 18), acronyms PLCO and NLST should be described on Page 18, instead of Page 19. -Discussion (Page 21,22). Please use consistent terms: Annals of Internal Medicine vs. Ann Intern Med.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2

Dr. Kyungwan Hong, University of Maryland Baltimore Comments to the Author:

Revisions:

Results (Page 16, Metrics of actual re-use): I was wondering if you explored how long it takes from submitting a data sharing request to receiving the requested data. If so, can you please mention that? From my past experience, the time it takes to receive the data can be burdensome.

The reviewer is right, this is an important point, however, the time from submitting a data sharing request to receiving the requested data was not systematically investigated in the review. We have added a statement in “strengths and limitations of this study” and a small section in the “discussion” with a few references.

Discussion (Page 21, “In short, the pressure by publisher…”): Consider adding your own suggestions/recommendations for publishers and funders to improve their engagement to share data.

Thank you for this comment. We have a piece in preparation that details all the proposed changes. We think that listing all these ideas in this paper would be out of scope. We however agree to add a few words about this in the “discussion”.

Discussion (Page 23, “Research output from shared data should have an impact…”): Can you elaborate further on “impact”? I completely agree that more studies are needed to be done with shared clinical trial data, but that does not necessarily lead to a more powerful impact in biomedical research. Therefore, I am struggling to follow your point here. What are your suggestions to improve the impact of research projects using shared clinical trial data?

Thank you for this comment. We have tried to be more explicit.

Also, please check the availability of other minor grammatical errors to strengthen your argument and clarify your delivery. Some of the errors that I identified are following:

-Throughout the manuscript: Please use consistent terms. For example, “ClinicalTrials.gov”/“clinicaltrials.gov”

Has been corrected.

-Throughout the manuscript: Please make the referencing consistent. For example, references were mostly mentioned after a period (.xx), but sometimes references were mentioned before a period (xx.)

This was standardised throughout the manuscript.

-Introduction, Rationale (Page 5), acronym IPD should be described in the first paragraph, not in the third paragraph. (e.g., Besides the individual participant data (IPD) sets) -Method (Page 7, in the box) “project Data Sphere” --> “Project Data Sphere”

-Results (Page 13), please add percentage after “280/2003”.

-Results (Page 16, “Between 2013 and 2015 177”), please add a comma (,) between 2015 and 177.

-Results (Page 17), Add period (.) after “154 (66%) were approved”.

-Results (Page 18), acronyms PLCO and NLST should be described on Page 18, instead of Page 19.

-Discussion (Page 21,22). Please use consistent terms: Annals of Internal Medicine vs. Ann Intern Med.

All the points described above have been corrected.

VERSION 3 – REVIEW

REVIEWER	Hong, Kyungwan University of Maryland Baltimore, Department of Pharmaceutical Health Services Research
REVIEW RETURNED	19-Jul-2021

GENERAL COMMENTS	Thank you for the opportunity to re-review this manuscript. The authors addressed all of my suggested comments thoughtfully and carefully. I have no further comments, and this study will be a valuable resource to researchers and policymakers who seek to understand the current landscape of individual participant data sharing.