PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Characteristics of non-randomized studies of pharmacologic treatment: a cross-sectional study
AUTHORS	Yaacoub, Sally; Porcher, Raphael; Pellat, Anna; Bonnet, Hillary; Tran, Viet-Thi; Ravaud, Philippe; Boutron, Isabelle

VERSION 1 - REVIEW

REVIEWER NAME	Ju, Chengsheng
REVIEWER AFFILIATION	UCL School of Pharmacy, Pharmacy Practice and Policy
REVIEWER CONFLICT OF INTEREST	I receive funding from Pharmacy Research UK, University College London, National Institute for Health and Care Research - University College London Hospital, and CureParkinson's to support my research.
DATE REVIEW RETURNED	15-Apr-2024

GENERAL COMMENTS	Yaacoub et al. reported a cross-sectional study characterising the contemporary practice of comparative effectiveness research on pharmacologic interventions using real-world data. The findings are of interest to pharmacoepidemiologists and other clinical researchers. I have a number of comments and questions: 1. Introduction: “The literature underscores particular issues in non-randomized studies occasionally centering on specific conditions (11 -14). However, the collective issues of heterogeneity and limitations in the conduct, analysis, and reporting of non-randomized studies, in a representative sample, remain understudied.” This sentence needs to be revised. Ref 14 is a systematic review covering all trial emulation studies from all clinical specialties. And I doubt if the sample included in the current study, which were randomly selected 200 records between a time span of 3 months, is a “representative sample” of non-randomised comparative effectiveness studies. 2. Methods – Eligibility criteria: does the comparator arm need to be of a pharmacologic treatment? 3. Methods – Study selection: what is the rationale of selecting 200 records? 4. Methods – Data extraction: “We developed and piloted a standardized data extraction form inspired by …” Is this extraction from the one in Appendix 3? 5. Methods – Epidemiology: how was “presence of statistician or methodologist” determined? The information can be found in Table 1 but should be included in the methods section. 6. Methods – Reporting characteristics: reporting of follow-up, i.e., the definition of start and end of follow-up could also be important for a cohort study. 7. Methods – Risk of bias by design: this section is only applicable to cohort studies. To minimise bias in case-control studies, there are other special considerations (10.1136/bmj-2022-072346). Did the authors consider cohort designs and case-control designs

	separately? 8. Methods – Risk of bias by design: Appendix 5 (designs to mitigate immortal time bias) was not properly referenced in the main text. 9. Appendix 6: 35/747 reports are not available which is a large proportion. What is the reason? 10. Results: it would be helpful to provide the list of included studies. 11. Results – Risk of bias by design: “but biases by design were only completely addressed in three reports using the methods listed in Appendix 5” The methods in Appendix 5 are only for addressing immortal time bias. In fact the landmark analysis also has the “misalignment of three time-points” and is at risk of potential selection bias (10.1093/ckj/sfaa242). 12. Results – Risk of bias by design: grace period – the authors need to be clearer on what kind of grace period will induce immortal time bias. Grace period is commonly used in clone-censor-weight design without inducing immortal time bias. 13. Discussion – Implications for practice: adopting the target trial emulation framework is not enough, as the previous systematic review has already pointed out that the trial emulation framework has been inconsistently applied. The key is how to correctly apply the framework, possibly by promoting some of the bias-resistant designs as listed in Appendix 5? 14. Discussion: “only 11% had a comparator, accounted for confounding factors and had no risk of bias by design identified (or had all biases addressed).” This should go into the Key Messages. 15. Table 2: some of the proportions are apparently wrong, e.g., Causal contrast/Estimand section. 16. I don’t really understand Figure 1. Adding a caption could be helpful.

REVIEWER NAME	Tri Yuniar, Cindra
REVIEWER AFFILIATION	Department of Pharmacology and Clinical Pharmacy, School of Pharmacy, Institut Teknologi Bandung, Bandung 40132, Indonesia
REVIEWER CONFLICT OF INTEREST	None
DATE REVIEW RETURNED	23-Apr-2024

GENERAL COMMENTS	1. On line 25/26, the author should add the example of a non-randomized study that has a significant role in pharmacology decision-making. 2. The introduction didn't explain about the non-randomized reporting issues 3. In the study selection, the author mentioned 'one reviewer assessed ...', and the next sentence explains about the third reviewer. Is there a second reviewer? What is the rule of the second reviewer? How to make an agreement between the reviewer? 4. Please mention the reporting guidelines used in the study. It will be better to make a specific chart 5. There are no keywords or specific terms used for safety information. How did the author make sure the screening process was included the safety studies?

REVIEWER NAME	Ioannides, Anne

REVIEWER AFFILIATION	Imperial College London, School of Public Health
REVIEWER CONFLICT OF INTEREST	None
DATE REVIEW RETURNED	25-Apr-2024

GENERAL COMMENTS	The authors aimed to assess the degree to which biases are present, and epidemiological parameters are followed, in non-randomised studies on pharmacologic treatment; this is a highly relevant and very important research question, particularly as more real-world evidence is used in medical research. The authors demonstrated descriptively that, among 200 studies meeting full study criteria, there were several types of biases and reporting inconsistencies across the papers within the sample. Strengths of the paper: 1. The authors have addressed a topic that is highly relevant and for which establishing more stringent frameworks will be very important going forward. 2. Additionally, the paper describes a large range of characteristics of the studies that it evaluated - I commend the authors on the breadth of characteristics that they extracted. 3. I commend the authors on constructing their guideline framework, and I think that Appendix 3 and Appendix 4 are very helpful streamlining tools when thinking about non-randomised epi studies. I have a few comments for consideration. 1. I agree with the authors that adoption of trial emulation framework in non-randomised observational research would strengthen the evidence generated from these types of studies. To strengthen this claim, I recommend (1) reporting the proportion of studies within this research that are trial emulation studies; and (2) repeating some of the key descriptive metrics from this paper on target trial emulation studies only, and those not, to see whether there is a (descriptive) difference in the epidemiological approaches, reporting, and/or biases. If these differences are already known from previous literature, I recommend discussing differences in a bit more detail in the discussion (in addition to sentence four of the Comparison to available literature section of the discussion). 2. Given the granularity of assessment of the included studies (as per Appendix 3), perhaps the authors would consider adding some additional descriptive statistics to the supplementary information (one example includes that causal language was screened for – it would be interesting to see what proportion of papers use causal language inappropriately). 3. Why did the authors elect to not reference/list the studies that they included within the study? A similar study on reporting in trial emulation (Hansford et al 2023, PMID: 37755828) with similar methodology has done so. 4. Could authors please clarify why they selected the sampling period of June to August 2022? 5. Please could authors clarify what is meant by a “representative” sample of reports? What is the frame of reference that makes these articles representative?

	6. Authors stated that they did not use study design labels reported by authors, but instead used the designs as per Appendix 2. This was interesting to me - what motivated this decision – did you find that there was a discrepancy between what people reported they did and what they actually did? If so, how many? Minor comments: 7. Figure 1: Might it be possible to add panels and/or a legend to Figure 1? It took me a while to fully understand what each individual part of the figure/the axes were demonstrating. 8. Figure 2: I like the concept of the figure but I found the “tiny segment slices” slightly overwhelming – I think either a regular paneled stacked bar chart, or (if you would prefer the ring design) even making the rings solid colours with written %-values would make the figure a bit more digestible for the reader. 9. Appendix 8: what do the different colours mean? Please include a legend.

REVIEWER NAME	Hoffmann, Sabine
REVIEWER AFFILIATION	Ludwig Maximilian University Munich Institute of Medical Information Processing Biometrics and Epidemiology
REVIEWER CONFLICT OF INTEREST	None
DATE REVIEW RETURNED	10-May-2024

GENERAL COMMENTS	This is a well written manuscript and I commend the authors on their effort to evaluate the reporting and biases of non-randomized studies of pharmacologic treatments. In my view, it is particularly important to be aware of the high prevalence of risk of bias by design and the low rate of registration in these studies. Concerning the incompleteness of reporting, however, I am wondering whether it is fair to say that the reporting of a study is incomplete if it fulfilled all items of the corresponding reporting guideline. In particular, the key study elements you consider for reporting are different from the items in checklists to be reported in non-interventional pharmacoepidemiological studies including the REPORT guideline if I am not mistaken. I am not saying that the items 3) treatment deviations and 4) causal contrasts/estimand are not valuable and I think that it is an excellent idea to use validation of codes or algorithms applied to select the population, but if there are no reporting guidelines to tell authors that they should report this, can we blame them for not doing it? Shouldn’t we first change the reporting guidelines and then use these as benchmark to evaluate the completeness of reporting of individual studies? If we include items that some authors might not even have been aware of (although they make sense) any estimation on the completeness of reporting seems somewhat arbitrary. Apart from this major comment I only have a number of minor comments: - I am slightly uneasy with your use of the term “epidemiology” throughout the manuscript. Epidemiology is often defined as the study and analysis of the distribution and determinants of health-related states. You refer to “Epidemiology of non-randomized studies” when you talk about the general characteristics of the study, the study design, the research question, type of data used, sources, funding source, setting etc. Non-randomizes studies are

	arguably no health-related state and I am wondering whether it would be clearer for readers to refer to the general characteristics rather than to the epidemiology of non-randomized studies. - p. 4 line 21: When you mention the surge of real-world data in recent years in the sentence “The prominence of non-randomized studies has increased in recent years, specifically with the surge of real-world data” you cite a paper that is almost 30 years old (Black et al. 1996). Maybe somewhat naively, I would have expected a more recent reference to this statement. - p. 7 line 8: You mention that data were extracted in duplicate for 20% of studies. Did you quantify the level of agreement between the raters for these studies? - p. 10 line 36: The difference in the number of participants included and analysed seems to be quite large. Can you say something about the reasons why on average 30% of included participants were not included in the analysis? - p. 13 line 50: I would suggest to add “evaluating the reporting quality of cohort studies using real-world data” after “one study” - p. 14 line 25: The sentence “and out of which only five described solutions to mitigate theses biases” seems incomplete - p. 15 line 6 In the sentence “The identified biases by design can be avoided through thorough planning and explicit reporting to align the three time-points” it is not fully clear which time points you are referring to (because they were only mentioned much earlier in the manuscript), so I would suggest clarifying this. - Figure 2: Would it be possible to replace the 1 (at least 1 risk of bias by design) by another abbreviation? It took me some time to understand that the misalignment of the numbers under the barchart is explained by the fact that 2 - 5 correspond to a specific bias whereas 1 refers to the summary (at least 1 risk of bias by design). - The summary table of the risk of bias by design in Appendix 4 is very useful and in my view it would be valuable to include it in the main text of the paper if space restriction allows it.

REVIEWER NAME	Wallach, Joshua
REVIEWER AFFILIATION	Emory University Rollins School of Public Health, Department of Epidemiology
REVIEWER CONFLICT OF INTEREST	Funding from FDA, Arnold Ventures, Johnson & Johnson, and the NIH. Former consulting fees from Hagens Berman Sobol Shapiro LLP and Dugan Law Firm APLC
DATE REVIEW RETURNED	27-May-2024

GENERAL COMMENTS	In this manuscript by Yaacoub et al., the authors set out to examine comparative non-randomized studies that assess the effectiveness and/or safety of pharmacological treatments. In particular, they randomly selected 200 reports published June-August 2022 and evaluated the risk of bias by design using a piloted form inspired by reporting guidelines. Overall, the authors found that 71% were “at risk of bias by design; 27% due to inclusion of prevalent users, 31% due to post-treatment eligibility criteria, and 43% due to immortal time periods and 23% due to classification of treatment”. This was clearly a tremendous amount of work, and the authors should be praised for their effort and focus on transparency (preregistration, detailed searches and supplement). Furthermore, this is an interesting and important topic - given the growing use of real-world evidence and observational research methods, it is important to have an understanding of how this research is being

	conducted. Overall, I do think there are opportunity to enhance the focus and reporting of the manuscript. Therefore, I have focused on big-picture recommendations. 1) Focus: At times, I found the manuscript rather difficult to follow. For instance, at the Abstract-level, it is unclear what the study will evaluate/focus on until the results are reported (and once the results are reported, some of the terms may be new to readers - it may make sense to define some of the characteristics and then shorten the results). This is also the case for the Introduction section of the manuscript. After reading the Introduction, I wasn't certain what the authors were going to do/evaluate. Although the authors note that "the collective issues of heterogeneity and limitations in the conduct, analysis, and reporting of non-randomized studies, in a representative sample, remain understudied", it is unclear what will be done. However, after reading the Methods and Results, it felt like the authors were trying to do too much (this is hinted at in the title, which includes 'epidemiology, reporting and biases'). Perhaps it would make sense to reframe this manuscript as an evaluation of the 'characteristics of comparative effectiveness and safety studies of pharmacologic treatments'. The use of the word 'epidemiology' in the title and text is rather confusing. Furthermore, I would avoid the use of the phrase 'risk of bias'. Given that there is no formal risk of bias evaluation tool (e.g., RoB 2.0 for trials, and NOS for observational studies), I think it would be more transparent if the items evaluated were classified based on what they were trying to capture / the characteristics related to the target trial emulation framework. Then, based on the grouping of certain characteristics, the manuscript could be reworked to present the information across these groupings. In the Methods section, there are also opportunities to justify the characteristics considered. 2) Transparency characteristics: While some of the transparency characteristics included are important across all study designs, I suspect that they will not be the most interesting to the general reader compared to the other study design characteristics. Perhaps it would make sense to move this part of the table to the supplement and minimize the focus/reporting on these (it is unclear how data sharing would work for these types of studies. For instance, if authors are conducing analyses of administrative claims or EHR data, the chances that data can be shared are low). 3) This manuscript contains a lot of information (e.g., Table 1 is 3 pages long), and there are opportunities to simplify. For instance, Table 1 could be split across multiple tables based on the characteristics (e.g., those related to transparency – registration, protocol, data sharing, access to the codes, etc.). In reading the Summary of Findings in the Discussion section, it is hard to pinpoint the takeaway findings as it jumps between several of the many characteristics evaluated. I suspect that most readers will be interested in the study design characteristics (vs. protocol, data sharing, etc) 3) There are opportunities to simplify the Figures (e.g., the figure on Page 30 is difficult to follow – in particular, the top part of the page). The figure on page 31 is also not intuitive. 4) For the Discussion, there are several limitations missing: sample is from a specific period in 1 year; the abstractions were not all done in duplicate; etc.

REVIEWER NAME	Davies, Neil
REVIEWER AFFILIATION	University of Bristol
REVIEWER CONFLICT OF INTEREST	None
DATE REVIEW RETURNED	12-Jun-2024

GENERAL COMMENTS	bmjmed-2024-000932 entitled "Epidemiology, reporting and biases of non-randomized studies of pharmacologic treatment: a cross-sectional study" This is an interesting, well-conducted study that reviews the reporting of non-randomized pharmacotherapy studies. The authors randomly selected papers published between June 2022 and August 2022. Using their search terms, they got 26,123 hits. The findings are concerning but consistent with my view of the literature. Most published studies exhibit some key, widely known limitations. Is a random sample necessarily the best way to review the literature? A random sample places the same weight on all papers, given that your databases are likely to include a very high volume of very low-impact journals that are likely predatory. Do we necessarily care about the properties of papers on average? Rather than papers published in credible journals? The paper currently has very little substantive analysis beyond presenting means. For example, is there any evidence that the quality of reporting differs by the type or impact factor of the journal? Or whether having a statistician on the author list or government funding improves or worsens the quality of reporting? What was the rationale for including 200 papers? Did you run a power calculation, and why is this enough? Do you need to report standard errors with your results? Please could you archive the code used to clean and analyse your data on GitHub or similar? Page 16l6- It might be worth restating the three-time points here. I do not understand what is being presented in Figure 1. Could this either be clarified or revised?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Dr. Chengsheng Ju, UCL School of Pharmacy

Comments to the Author

Yaacoub et al. reported a cross-sectional study characterising the contemporary practice of comparative effectiveness research on pharmacologic interventions using real-world data. The findings are of interest to pharmacoepidemiologists and other clinical researchers. I have a

number of comments and questions:

Thank you for the positive feedback.

1. Introduction: "The literature underscores particular issues in non-randomized studies occasionally centering on specific conditions (11-14). However, the collective issues of heterogeneity and limitations in the conduct, analysis, and reporting of non-randomized studies, in a representative sample, remain understudied." This sentence needs to be revised. Ref 14 is a systematic review covering all trial emulation studies from all clinical specialties. And I doubt if the sample included in the current study, which were randomly selected 200 records between a time span of 3 months, is a "representative sample" of non-randomised comparative effectiveness studies.

Thank you for the comment. We wanted to highlight that in most cases, the literature focuses on a single 'issue', for example reporting (Ref 14), or on specific conditions, for example diabetes (Ref 11).

We modified the sentence as follows:

> "The literature underscores particular issues in non-randomized studies such as reporting and occasionally centers on specific conditions (11-14)."

To select our sample, we followed a similar methodology used in other studies, e.g., MJ Page et al. (10.1371/journal.pmed.1002028), where they searched for systematic reviews indexed throughout one calendar month, then they extracted data from a random sample of 300 of the included SRs, which were selected using the random number generator in Microsoft Excel.

We clarify that our sample is a representative sample of reports indexed in MEDLINE throughout the manuscript. We are happy to change the term to 'non-selective sample', if the editor deems necessary.

2. Methods – Eligibility criteria: does the comparator arm need to be of a pharmacologic treatment?

Thank you for the comment. The comparator can be standard of care (or no treatment). When the comparator is an active treatment, then it had to be a pharmacologic treatment. We clarified this in the text as follows:

> "We included reports of non-randomized studies:
>
> - conducted in humans;
> - aimed to assess effectiveness and/or safety of pharmacologic treatment(s);
> - with a comparator arm (e.g., active pharmacologic comparator, standard of care or no treatment)…"

3. Methods – Study selection: what is the rationale of selecting 200 records?

Thank you for the comment. No formal sample size calculation was performed because of the descriptive nature of this study. However, 200 reports allow for estimating percentages with a reasonable precision (half-width of a 95% confidence interval). For instance, this sample size permits estimating percentages of 10% and 50% with precisions of 4.2% and 6.9%, respectively.

4. Methods – Data extraction: "We developed and piloted a standardized data extraction form inspired by …" Is this extraction from the one in Appendix 3?

You are right. We have missed to add that to the data extraction section. We corrected the sentence as follows:

> "We developed and piloted a standardized data extraction form (Appendix 3) inspired by RECORD…"

5. Methods – Epidemiology: how was "presence of statistician or methodologist" determined? The information can be found in Table 1 but should be included in the methods section.

Thank you for the comment. We clarified this as follows:

> "We extracted study characteristics (e.g., medical area, region, presence of statistician or methodologist determined from author affiliation or explicit reporting)."

6. Methods – Reporting characteristics: reporting of follow-up, i.e., the definition of start and end of follow-up could also be important for a cohort study.

The reviewer makes a great point. We extracted data on the length of follow-up and when follow-up starts (as a study time point). For example, we capture if follow-up starts after treatment assignment. However, we did not extract data on the explicit definitions of the start and end of follow-up in each study.

7. Methods – Risk of bias by design: this section is only applicable to cohort studies. To minimise bias in case-control studies, there are other special considerations (10.1136/bmj-2022-072346). Did the authors consider cohort designs and case-control designs separately?

Thank you for the comment and for the insightful reference. The assessment of the possible biases was applicable to both cohort and case-control studies as it was based on the target trial emulation framework. Both study designs can be used in trial emulations, where a structured protocol should be specified as highlighted in the reference shared above (10.1136/bmj-2022-072346) and by the team of MA Hernan (10.1093/ije/dyaa144). In fact, similar considerations for time point alignment are relevant to case-control studies and similar issues arise such as post-treatment eligibility (10.1016/j.jclinepi.2016.04.014). There is an inherent issue of case-control studies that they evaluate the treatment at the time of the outcome of interest, i.e., 'look back' for the treatment. Rasouli et al (10.1136/bmj-2022-072346) describe methods to minimize biases through trial emulations using nested case-control design, focusing on prevalent user bias and bias due to inappropriate adjustment for covariates. However, we did not assess the latter, which is more of an issue in case-control studies, as we only focused on four biases related to time point misalignment: bias due to inclusion of prevalent users (selection bias), bias due to post-treatment eligibility (selection bias), bias due to immortal time periods and bias due to classification of treatment arms.

We highlighted this point as a limitation in the Discussion section:

> "One limitation is that we did not assess all biases that might have been present, particularly those that are more relevant to case-control studies such as inappropriate adjustment for covariates, as we only focused on possible biases related to time point misalignment."

8. Methods – Risk of bias by design: Appendix 5 (designs to mitigate immortal time bias) was not properly referenced in the main text.

Thank you for the comment. We referenced it more clearly as follows:

> "The methods to address bias due to immortal time period and due to classification of treatment arms is provided in Appendix 5."

9. Appendix 6: 35/747 reports are not available which is a large proportion. What is the reason?

Thank you for the comment. Due to our institution's subscriptions that have been cut in the last years, we were not able to retrieve the full-texts of these reports.

10. Results: it would be helpful to provide the list of included studies.

Thank you for the suggestion. We had not shared the list of included studies as we planned to have all our dataset publicly available on Open Science Framework, upon publication. We now include a list of included studies in Appendix 7. We will still publish the complete dataset on OSF, upon publication.

We added the following sentence to the Results:

> "Appendix 6 presents the flowchart and Appendix 7 presents the list of included reports of non-randomized studies."

11. Results – Risk of bias by design: "but biases by design were only completely addressed in three reports using the methods listed in Appendix 5" The methods in Appendix 5 are only for addressing immortal time bias. In fact the landmark analysis also has the "misalignment of three time-points" and is at risk of potential selection bias (10.1093/ckj/sfaa242).

The reviewer makes a great point. The bias due to inclusion of prevalent users and bias due to post-treatment eligibility can only be addressed by changing the eligibility criteria (i.e., only include incident users and not based on post-treatment criteria) as summarized in Appendix 4. On the other hand, bias due to immortal time periods and due to classification of treatment arms, can be addressed by statistical methods, as presented in Appendix 5. We clarified in the Methods section that Appendix 5 only presents methods to address risk of bias due to immortal time period and due to classification of treatment arms, as follows:

> "The methods to address bias due to immortal time period and due to classification of treatment arms is provided in Appendix 5."

We agree with the reviewer that landmark analysis might also introduce prevalent user bias. We clarified this in Appendix 5 as follows:

> "May introduce prevalent user bias"

12. Results – Risk of bias by design: grace period – the authors need to be clearer on what kind of grace period will induce immortal time bias. Grace period is commonly used in clone-censor-weight design without inducing immortal time bias.

Thank you for the comment. We clarify what we mean by a grace period in Appendix 4 as follows: 'a period between when participants meet eligibility to when they initiate the treatment. If not adequately

addressed in the analysis, the presence of a grace period would impose risk of bias due to immortal time period and due to classification of treatment'.

We considered that grace periods used with clone-censor-weighting did not induce bias, as it was considered as 'addressed'. We clarified this in Appendix 5 as follows:

> "The grace period introduced from clone-censor-weight approach was labeled as 'presence of grace period', but was considered as 'addressed' i.e., did not induce bias."

13. Discussion – Implications for practice: adopting the target trial emulation framework is not enough, as the previous systematic review has already pointed out that the trial emulation framework has been inconsistently applied. The key is how to correctly apply the framework, possibly by promoting some of the bias-resistant designs as listed in Appendix 5?

The reviewer makes a great point. Even when the target trial emulation framework is adopted, it has not been adequately applied. We highlight this point as follows:

> "Moreover, when applicable, researchers should adopt and properly apply the target trial emulation framework, which highlights the need to clearly define the research question, have a well-designed protocol with explicitly stated components and implement appropriate statistical analysis methods. Tools should be developed in order to facilitate and guide the planning of studies, urging researchers to explicitly define the research question and comprehensively state the components of the study along with determining an appropriate statistical analysis plan."

14. Discussion: "only 11% had a comparator, accounted for confounding factors and had no risk of bias by design identified (or had all biases addressed)." This should go into the Key Messages.

Thank you for the suggestion. Our key messages box now includes the following:

> "Only 11% of non-randomized studies with an objective to assess the effectiveness and/or safety of pharmacologic treatment had a comparator, accounted for confounding factors and had no biases related to time point misalignment (or had all biases addressed)."

15. Table 2: some of the proportions are apparently wrong, e.g., Causal contrast/Estimand section.

Thank for the comment. We corrected the proportions in the table.

16. I don't really understand Figure 1. Adding a caption could be helpful.

Thank you for the comment. We now provide a detailed caption as follows:

> "Each horizontal line corresponds to one included report. A specific color was attributed to each of the six key study elements. The color band show which of these items (labelled at the top) were reported for each included report. The 200 included reports of non-randomized studies were sorted according to the total number of reported items, in decreasing order. The diagram on the left shows the distribution of the total number of reported items across the 200 included reports. The diagram at the bottom shows the proportion of reports that reported each element."

Reviewer: 2

Dr. Cindra Tri Yuniar, Department of Pharmacology and Clinical Pharmacy, School of Pharmacy, Institut Teknologi Bandung, Bandung 40132, Indonesia

Comments to the Author

1. On line 25/26, the author should add the example of a non-randomized study that has a significant role in pharmacology decision-making.

Thank you for the suggestion. We highlighted how non-randomized studies inform recommendations in clinical practice guidelines by providing an example from a study done on cardiology guidelines. We added the following sentence:

> "In fact, a study summarizing the levels of evidence supporting cardiology clinical practice guidelines, found that 40% of 6329 recommendations were supported by level of evidence B (i.e., supported by data from observational studies or a single RCT), out of which only few were from randomized evidence (11)."

2. The introduction didn't explain about the non-randomized reporting issues

Thank you for the comment. We now clearly mention that non-randomized studies also suffer from inadequate reporting. We modified the Introduction as follows:

> "However, non-randomized studies are susceptible to inadequate reporting, as well as to numerous limitations related to their design and analysis choices, which could provide biased effect estimates (11). Several guidelines have been developed for reporting nonrandomized studies (12-14), while the target trial emulation framework has been developed to overcome avoidable methodological pitfalls of traditional causal analysis of observational data, thus reducing the risk of bias (15)."

> "The literature underscores particular issues in non-randomized studies such as inadequate reporting and occasionally centers on specific conditions (16-19)."

3. In the study selection, the author mentioned 'one reviewer assessed ...', and the next sentence explains about the third reviewer. Is there a second reviewer? What is the rule of the second reviewer? How to make an agreement between the reviewer?

The reviewer makes a great point. We clarified that a second reviewer conducted the 20% screening. We modified the sentence as follows:

> "One reviewer assessed the eligibility of abstracts and full-texts with 20% done in duplicate and independently by a second reviewer. Any disagreements were resolved by discussion between the two reviewers or by a third reviewer."

4. Please mention the reporting guidelines used in the study. It will be better to make a specific chart.

Thank you for the comment. Although we conducted a cross-sectional study, the methodology was mostly consistent with a systematic review. For that reason, we shared a filled PRISMA checklist during submission. However, we do not to explicitly state that we use the PRISMA checklist for reporting given the discrepancy in study design. We are happy to add a statement in the manuscript to if the editor deems necessary.

5. There are no keywords or specific terms used for safety information. How did the author make sure the screening process was included the safety studies?

Thank you for the comment. The reviewer is right, we do not have specific terms for safety. We also do have specific ones for effectiveness. In fact, we developed our search to be general, with terms not specific to effectiveness or safety, and rather to be specific to 'study design' and 'pharmacologic treatment'. We believe that safety studies would also be captured in our search strategy, specifically since we include keywords such as 'surveillance', 'vigilance', 'claims', etc.

Reviewer: 3
Dr. Anne Ioannides, Imperial College London
Comments to the Author
The authors aimed to assess the degree to which biases are present, and epidemiological parameters are followed, in non-randomised studies on pharmacologic treatment; this is a highly relevant and very important research question, particularly as more real-world evidence is used in medical research. The authors demonstrated descriptively that, among 200 studies meeting full study criteria, there were several types of biases and reporting inconsistencies across the papers within the sample.

Strengths of the paper:
1. The authors have addressed a topic that is highly relevant and for which establishing more stringent frameworks will be very important going forward.

2. Additionally, the paper describes a large range of characteristics of the studies that it evaluated - I commend the authors on the breadth of characteristics that they extracted.

3. I commend the authors on constructing their guideline framework, and I think that Appendix 3 and Appendix 4 are very helpful streamlining tools when thinking about non-randomised epi studies.

Thank you for the positive feedback.

I have a few comments for consideration.

1. I agree with the authors that adoption of trial emulation framework in non-randomised observational research would strengthen the evidence generated from these types of studies. To strengthen this claim, I recommend (1) reporting the proportion of studies within this research that are trial emulation studies; and (2) repeating some of the key descriptive metrics from this paper on target trial emulation studies only, and those not, to see whether there is a (descriptive) difference in the epidemiological approaches, reporting, and/or biases. If these differences are already known from previous literature, I recommend discussing differences in a bit more detail in the discussion (in addition to sentence four of the Comparison to available literature section of the discussion).

Thank you for the suggestion. In fact, we identified only two studies that explicitly state being trial emulation studies. For this reason, we are not able to compare the characteristics of trial emulations and that of other studies. We now report in Table 1 of the Results section that two studies were trial emulation studies as follows:

> "Two reports were of target trial emulation studies as reported by authors"

2. Given the granularity of assessment of the included studies (as per Appendix 3), perhaps the authors would consider adding some additional descriptive statistics to the supplementary information (one example includes that causal language was screened for – it would be interesting to see what proportion of papers use causal language inappropriately).

Thank you for the suggestion. We added the findings on the use of causal language in the subsection 'Study design and data' and in Appendix 8, as follows:

"In the abstract, one third of the reports used causal language (n=69, 35%)."

3. Why did the authors elect to not reference/list the studies that they included within the study? A similar study on reporting in trial emulation (Hansford et al 2023, PMID: 37755828) with similar methodology has done so.

Thank you for the suggestion. We had not shared the list of included studies as we planned to have all our dataset publicly available on Open Science Framework, upon publication. We now include a list of included studies in Appendix 7. We will still publish the complete dataset on OSF, upon publication.

We added the following sentence to the Results:

"Appendix 6 presents the flowchart and Appendix 7 presents the list of included reports of non-randomized studies."

4. Could authors please clarify why they selected the sampling period of June to August 2022?

Our approach was pragmatic. Not restricting to a specific time period resulted in hundreds of thousands of hits, which would have made the sorting in random order much more complex for study selection. For that reason, we made the arbitrary decision to restrict the sampling period to the most recent three months (June to August, 2022) before running the search on September 29, 2022.

5. Please could authors clarify what is meant by a "representative" sample of reports? What is the frame of reference that makes these articles representative?

Thank you for the comment. We believe that the random sorting of the search hits, allowed us to have a representative sample of reports indexed in MEDLINE. We now clarify that the representative sample is of 'MEDLINE-indexed' reports, throughout the manuscript.

We are happy to change the term to 'non-selective sample', if the editor deemed necessary.

6. Authors stated that they did not use study design labels reported by authors, but instead used the designs as per Appendix 2. This was interesting to me - what motivated this decision – did you find that there was a discrepancy between what people reported they did and what they actually did? If so, how many?

Thank you for the comment. In fact, we chose this approach as the reports used heterogeneous terms to describe the study design, which in many instances was not informative. Example of terms include: retrospective/prospective study, registry-based study, observational study, etc. However, we did not compare what was reported to what was judged.

Minor comments:
7. Figure 1: Might it be possible to add panels and/or a legend to Figure 1? It took me a while to fully understand what each individual part of the figure/the axes were demonstrating.

Thank you for the comment. We now provide a detailed caption as follows:

"Each horizontal line corresponds to one included report. A specific color was attributed to each of the six key study elements. The color band show which of these items (labelled at the

top) were reported for each included report. The 200 included reports of non-randomized studies were sorted according to the total number of reported items, in decreasing order. The diagram on the left shows the distribution of the total number of reported items across the 200 included reports. The diagram at the bottom shows the proportion of reports that reported each element."

8. Figure 2: I like the concept of the figure but I found the "tiny segment slices" slightly overwhelming – I think either a regular paneled stacked bar chart, or (if you would prefer the ring design) even making the rings solid colours with written %-values would make the figure a bit more digestible for the reader.

Thank for the suggestion. We edited the figure and added the following caption:

"This figure summarizes the possible biases for each report of non-randomized studies. Each spoke represents one report. The bricks are a visual representation of the possible bias related to time point misalignment, red indicating possible bias; yellow, unable to assess; and green, no bias. Every concentric circle represents one of the biases with 'bias due to inclusion of prevalent users' being the furthest circle from the center, and 'bias in classification of treatment arms', the central circle. The most external circle represents an overview of the possible biases for each report (red when at least one possible bias exists, yellow when unable to assess and green when no bias exists). The histogram is a representation of the summary of the possible biases, per bias, for the 200 reports."

9. Appendix 8: what do the different colours mean? Please include a legend.

Thank you for the comment. We now provide a detailed legend in Appendix 10 (previously Appendix 8) as follows:

"Brown: relevant to bias due to prevalent users; Blue: relevant to bias due to post-treatment eligibility; Green: relevant to bias due to immortal time periods"

Reviewer: 4
Dr. Sabine Hoffmann, Ludwig Maximilian University Munich Institute of Medical Information Processing Biometrics and Epidemiology
Comments to the Author

This is a well written manuscript and I commend the authors on their effort to evaluate the reporting and biases of non-randomized studies of pharmacologic treatments. In my view, it is particularly important to be aware of the high prevalence of risk of bias by design and the low rate of registration in these studies.

Thank you for the positive feedback.

Concerning the incompleteness of reporting, however, I am wondering whether it is fair to say that the reporting of a study is incomplete if it fulfilled all items of the corresponding reporting guideline. In particular, the key study elements you consider for reporting are different from the items in checklists to be reported in non-interventional pharmacoepidemiological studies including the REPORT guideline if I am not mistaken. I am not saying that the items 3) treatment deviations and 4) causal contrasts/estimand are not valuable and I think that it is an excellent idea to use validation of codes or algorithms applied to select the population, but if there are no reporting guidelines to tell authors that they should report this, can we blame them for not doing it? Shouldn't we first change the reporting guidelines and then use these as benchmark to evaluate the completeness of reporting of individual studies? If we include items that some authors might not even have been aware of (although they make sense) any estimation on the completeness of reporting seems somewhat arbitrary.

Thank you for the comment. We did consider multiple reporting guidelines for observational studies (e.g., STROBE, RECORD), and even for randomized controlled trials (e.g., CONSORT). Treatment deviations and causal contrast/estimand have been listed in the reporting guidelines (RECORD-PE item 7.1.g and STROBE/RECORD/RECORD-PE item 9), but perhaps not as explicitly. Since all items are equally important in reporting guidelines, we chose to highlight only specific ones, which we consider as key study elements. We highlighted this point in the Discussion section as follows:

> "Although all items of reporting guidelines are important, we highlight only specific ones, considering them as key study elements. In general, the reports of non-randomized studies insufficiently report these key study elements and elements that increase confidence in the results (e.g., use of validation studies, eligibility to any treatment arm)."

Apart from this major comment I only have a number of minor comments:

- I am slightly uneasy with your use of the term "epidemiology" throughout the manuscript. Epidemiology is often defined as the study and analysis of the distribution and determinants of health-related states. You refer to "Epidemiology of non-randomized studies" when you talk about the general characteristics of the study, the study design, the research question, type of data used, sources, funding source, setting etc. Non-randomizes studies are arguably no health-related state and I am wondering whether it would be clearer for readers to refer to the general characteristics rather than to the epidemiology of non-randomized studies.

Thank you for the suggestion. We changed the term 'Epidemiology' to 'General characteristics' throughout the manuscript.

- p. 4 line 21: When you mention the surge of real-world data in recent years in the sentence "The prominence of non-randomized studies has increased in recent years, specifically with

the surge of real-world data" you cite a paper that is almost 30 years old (Black et al. 1996). Maybe somewhat naively, I would have expected a more recent reference to this statement.

The reviewer makes an excellent point. We have updated the references to the following:

Dang A. Real-World Evidence: A Primer. Pharmaceut Med. 2023 Jan;37(1):25-36. doi: 10.1007/s40290-022-00456-6. Epub 2023 Jan 5. PMID: 36604368; PMCID: PMC9815890.

Franklin, J. M., & Schneeweiss, S. (2017). When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials?. Clinical pharmacology and therapeutics, 102(6), 924–933. https://doi.org/10.1002/cpt.857

- p. 7 line 8: You mention that data were extracted in duplicate for 20% of studies. Did you quantify the level of agreement between the raters for these studies?

Thank you for the comment. The data was extracted in duplicate (independently for 20% and 80% as data verification) and disagreements were resolved by discussion or by a third reviewer. We were able to discuss, calibrate, and agree on what and how data extraction should be done, while verifying the decisions made, however we did not quantify the level of agreement between the reviewers. In addition, based on the recommendations of Cochrane Handbook for Systematic Reviews of Interventions (Chapter 7), the use of statistical measures of agreement (such as kappa statistics) to describe the extent to which assessments by multiple authors were the same is not recommended; it is more important that reasons for any disagreement are explored and resolved.

Reference: Boutron I, Page MJ, Higgins JPT, Altman DG, Lundh A, Hróbjartsson A. Chapter 7: Considering bias and conflicts of interest among the included studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.4 (updated August 2023). Cochrane, 2023. Available from www.training.cochrane.org/handbook.

- p. 10 line 36: The difference in the number of participants included and analysed seems to be quite large. Can you say something about the reasons why on average 30% of included participants were not included in the analysis?

Thank you for the comment. Indeed, the difference is large. We did not extract such information from the included reports, however one example is excluding eligible participants from the analysis if they did not complete the follow-up or those with missing data (e.g., covariates for propensity score).

- p. 13 line 50: I would suggest to add "evaluating the reporting quality of cohort studies using real-world data" after "one study"

Thank you for the suggestion. We modified the sentence as follows:

> "One study evaluating the reporting quality of cohort studies using real-world data documented limited transparency with only 24% of studies having an available study protocol and 20% having available raw data."

- p. 14 line 25: The sentence "and out of which only five described solutions to mitigate these biases" seems incomplete

Thank you for the comment. We modified the sentence as follows:

> “One study found that 25% of studies were at high risk of selection and/or immortal time bias, and only five of these studies described solutions to mitigate these biases.”

- p. 15 line 6 In the sentence “The identified biases by design can be avoided through thorough planning and explicit reporting to align the three time-points” it is not fully clear which time points you are referring to (because they were only mentioned much earlier in the manuscript), so I would suggest clarifying this.

Thank you for the comment. We clarified the three time-points, as follows:

> “The identified biases by design can be avoided through thorough planning and explicit reporting to align the three time points. Although aligning the time point of eligibility, treatment assignment and start of follow-up might sometimes be challenging, many approaches have been proposed in the target trial emulation framework.”

- Figure 2: Would it be possible to replace the 1 (at least 1 risk of bias by design) by another abbreviation? It took me some time to understand that the misalignment of the numbers under the barchart is explained by the fact that 2 - 5 correspond to a specific bias whereas 1 refers to the summary (at least 1 risk of bias by design).

Thank for the suggestion. We edited the figure and added the following caption:

> “This figure summarizes the possible biases for each report of non-randomized studies. Each spoke represents one report. The bricks are a visual representation of the possible bias related to time point misalignment, red indicating possible bias; yellow, unable to assess; and green, no bias. Every concentric circle represents one of the biases with ‘bias due to inclusion of prevalent users’ being the furthest circle from the center, and ‘bias in classification of treatment arms’, the central circle. The most external circle represents an overview of the possible biases for each report (red when at least one possible bias exists, yellow when unable to assess and green when no bias exists). The histogram is a representation of the summary of the possible biases, per bias, for the 200 reports.”

- The summary table of the risk of bias by design in Appendix 4 is very useful and in my view it would be valuable to include it in the main text of the paper if space restriction allows it.
Thank you for the positive feedback. Unfortunately, we currently reached the table/figure limit. We are happy to add it to the main text, if allowed by the editor.

Reviewer: 5
Dr. Joshua Wallach, Emory University Rollins School of Public Health
Comments to the Author

In this manuscript by Yaacoub et al., the authors set out to examine comparative non-randomized studies that assess the effectiveness and/or safety of pharmacological treatments. In particular, they randomly selected 200 reports published June-August 2022 and evaluated the risk of bias by design using a piloted form inspired by reporting guidelines. Overall, the authors found that 71% were "at risk of bias by design; 27% due to inclusion of prevalent users, 31% due to post-treatment eligibility criteria, and 43% due to immortal time periods and 23% due to classification of treatment".

This was clearly a tremendous amount of work, and the authors should be praised for their effort and focus on transparency (preregistration, detailed searches and supplement). Furthermore, this is an interesting and important topic - given the growing use of real-world evidence and observational research methods, it is important to have an understanding of how this research is being conducted. Overall, I do think there are opportunity to enhance the focus and reporting of the manuscript. Therefore, I have focused on big-picture recommendations.

Thank you for the positive feedback.

1) Focus: At times, I found the manuscript rather difficult to follow. For instance, at the Abstract-level, it is unclear what the study will evaluate/focus on until the results are reported (and once the results are reported, some of the terms may be new to readers - it may make sense to define some of the characteristics and then shorten the results). This is also the case for the Introduction section of the manuscript. After reading the Introduction, I wasn't certain what the authors were going to do/evaluate. Although the authors note that "the collective issues of heterogeneity and limitations in the conduct, analysis, and reporting of non-randomized studies, in a representative sample, remain understudied", it is unclear what will be done. However, after reading the Methods and Results, it felt like the authors were trying to do too much (this is hinted at in the title, which includes 'epidemiology, reporting and biases'). Perhaps it would make sense to reframe this manuscript as an evaluation of the 'characteristics of comparative effectiveness and safety studies of pharmacologic treatments'. The use of the word 'epidemiology' in the title and text is rather confusing. Furthermore, I would avoid the use of the phrase 'risk of bias'. Given that there is no formal risk of bias evaluation tool (e.g., RoB 2.0 for trials, and NOS for observational studies), I think it would be more transparent if the items evaluated were classified based on what they were trying to capture / the characteristics related to the target trial emulation framework. Then, based on the grouping of certain characteristics, the manuscript could be reworked to present the information across these groupings. In the Methods section, there are also opportunities to justify the characteristics considered.

Thank you for the comment and suggestions.

We edited the abstract according to your suggestions.

We edited the Introduction to allude to the main issues (inadequate reporting and biases), as follows:

> "However, non-randomized studies are susceptible to inadequate reporting, as well as to numerous limitations related to their design and analysis choices, which could provide biased

effect estimates (11). Several guidelines have been developed for reporting nonrandomized studies (12-14), while the target trial emulation framework has been developed to overcome avoidable methodological pitfalls of traditional causal analysis of observational data, thus reducing the risk of bias (15)."

"The literature underscores particular issues in non-randomized studies such as inadequate reporting and occasionally centers on specific conditions (16-19)."

We now use different terms: 'general characteristics' (previously 'epidemiology') and time point alignment and possible related biases (previously 'risk of bias by design'). We do not use 'risk of bias' anymore, rather we use 'bias related to time point misalignment'.

We edited all text and figures in the main text and supplementary material to reflect these changes. For example, we edited the 'Objective' as follows:

"Our objective was to examine the characteristics of comparative non-randomized studies that assess the effectiveness and/or safety of pharmacologic treatment(s). We focused on general characteristics, reporting characteristics and time point alignment and possible related biases."

The title is now 'Characteristics of non-randomized studies of pharmacologic treatment: a cross-sectional study '.

2) Transparency characteristics: While some of the transparency characteristics included are important across all study designs, I suspect that they will not be the most interesting to the general reader compared to the other study design characteristics. Perhaps it would make sense to move this part of the table to the supplement and minimize the focus/reporting on these (it is unclear how data sharing would work for these types of studies. For instance, if authors are conducing analyses of administrative claims or EHR data, the chances that data can be shared are low).

Thank you for the comment. We have shortened Table 1 in the main text and moved additional results, including transparency characteristics, to Appendix 8.

3) This manuscript contains a lot of information (e.g., Table 1 is 3 pages long), and there are opportunities to simplify. For instance, Table 1 could be split across multiple tables based on the characteristics (e.g., those related to transparency – registration, protocol, data sharing, access to the codes, etc.). In reading the Summary of Findings in the Discussion section, it is hard to pinpoint the takeaway findings as it jumps between several of the many characteristics evaluated. I suspect that most readers will be interested in the study design characteristics (vs. protocol, data sharing, etc)

Thank you for the comment. We have shortened Table 1 in the main text and moved additional results, including transparency characteristics, to Appendix 8.

In the Summary of Findings, we present more findings on the study design, as follows:

"The majority of the reports were of cohort studies conducted in patients with chronic diseases. They commonly compared treatment initiation to usual care/no treatment or other active treatments, assessing both effectiveness and safety outcomes. The majority of reports used routinely collected data, however half were monocentric and conducted in tertiary settings."

3) There are opportunities to simplify the Figures (e.g., the figure on Page 30 is difficult to follow – in particular, the top part of the page). The figure on page 31 is also not intuitive.

Thank you for the comment. We provided a detailed caption for Figure 1 as follows:

> "Each horizontal line corresponds to one included report. A specific color was attributed to each of the six key study elements. The color band show which of these items (labelled at the top) were reported for each included report. The 200 included reports of non-randomized studies were sorted according to the total number of reported items, in decreasing order. The diagram on the left shows the distribution of the total number of reported items across the 200 included reports. The diagram at the bottom shows the proportion of reports that reported each element."

We edited Figure 2 and added the following caption:

> "This figure summarizes the possible biases for each report of non-randomized studies. Each spoke represents one report. The bricks are a visual representation of the possible bias related to time point misalignment, red indicating possible bias; yellow, unable to assess; and green, no bias. Every concentric circle represents one of the biases with 'bias due to inclusion of prevalent users' being the furthest circle from the center, and 'bias in classification of treatment arms', the central circle. The most external circle represents an overview of the possible biases for each report (red when at least one possible bias exists, yellow when unable to assess and green when no bias exists). The histogram is a representation of the summary of the possible biases, per bias, for the 200 reports."

4) For the Discussion, there are several limitations missing: sample is from a specific period in 1 year; the abstractions were not all done in duplicate; etc.

Thank you for the suggestion. We added the following to the Limitations:

> "In addition, data extraction was done in duplicate and independently for only 20% of reports, whereas it was only verified in 80% and we only focus on studies indexed in MEDLINE in a specific period of time (three months in the year 2022)."

Reviewer: 6
Dr. Neil Davies, University of Bristol
Comments to the Author
bmjmed-2024-000932 entitled "Epidemiology, reporting and biases of non-randomized studies of pharmacologic treatment: a cross-sectional study"

This is an interesting, well-conducted study that reviews the reporting of non-randomized pharmacotherapy studies.
The authors randomly selected papers published between June 2022 and August 2022. Using their search terms, they got 26,123 hits. The findings are concerning but consistent with my view of the literature. Most published studies exhibit some key, widely known limitations.

Thank you for the positive feedback.

Is a random sample necessarily the best way to review the literature? A random sample places the same weight on all papers, given that your databases are likely to include a very high volume of very low-impact journals that are likely predatory. Do we necessarily care about the properties of papers on average? Rather than papers published in credible journals?

Thank you for the comment. We focus on a random sample in order to have a general overview of the characteristics of the non-randomized studies assessing the effectiveness and/or safety of pharmacologic treatments. These same studies could be included in systematic reviews for example, since in evidence synthesis the impact factor or reputation of the journal is usually not taken into consideration.

The paper currently has very little substantive analysis beyond presenting means. For example, is there any evidence that the quality of reporting differs by the type or impact factor of the journal? Or whether having a statistician on the author list or government funding improves or worsens the quality of reporting?

This is a great point, however it is beyond the scope of our study, as we aimed to examine the general characteristics (previously 'epidemiology'), reporting characteristics and time point alignment and possible related biases (previously 'biases by design'). We did not plan to compare these characteristics based on specific factors (e.g., journal impact factor, presence of a statistician, etc).

What was the rationale for including 200 papers? Did you run a power calculation, and why is this enough? Do you need to report standard errors with your results?

Thank you for the comment. No formal sample size calculation was performed because of the descriptive nature of this study. However, 200 reports allow for estimating percentages with a reasonable precision (half-width of a 95% confidence interval). For instance, this sample size permits estimating percentages of 10% and 50% with precisions of 4.2% and 6.9%, respectively.

Please could you archive the code used to clean and analyse your data on GitHub or similar?

Thank you for the comment. All datasets and code will be publicly available on Open Science Framework when the study is published.

Page 16l6- It might be worth restating the three-time points here.

Thank you for the suggestion. We included them as follows:

> “Although aligning the time points of eligibility, treatment assignment and start of follow-up might sometimes be challenging, many approaches have been proposed in the target trial emulation framework.”

I do not understand what is being presented in Figure 1. Could this either be clarified or revised?

Thank you for the comment. We provided a detailed caption as follows:

> “Each horizontal line corresponds to one included report. A specific color was attributed to each of the six key study elements. The color band show which of these items (labelled at the top) were reported for each included report. The 200 included reports of non-randomized studies were sorted according to the total number of reported items, in decreasing order. The diagram on the left shows the distribution of the total number of reported items across the 200 included reports. The diagram at the bottom shows the proportion of reports that reported each element.”