PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Design differences explain variation in results between randomized trials and their non-randomized emulations
AUTHORS	Heyard, Rachel; Held, Leonhard; Schneeweiss, Sebastian; Wang, Shirley

VERSION 1 - REVIEW

REVIEWER 1	Perol, David; Centre Leon Berard, Clinical Research. Competing Interest: None
REVIEW RETURNED	07-Aug-2023

GENERAL COMMENTS	Target trial emulation aims to prevent biases by applying RCT methodological and design principles to observational data. Therefore, it is expected that design emulation differences are related to variation in results between RCT and RWD studies. This paper demonstrates what kind of study emulation characteristics could explain most of the observed variation by comparing results from RCT-RWD study pairs based on data from the well-known RCT-DUPLICATE initiative. The paper has many excellent strengths, including a rigorous selection of trials from the RCT DUPLICATE which could be used in the study, a relevant choice and description of numerical and categorical emulation difference characteristics, use of meta-regression methods to estimate heterogeneity, and use of cross-validated mean squared errors method to reduce the complexity of the meta-regression. Nevertheless, there are several issues which should be addressed to simplify and to make this paper more accessible to a large audience of readers, including researchers and clinicians interested in performing trial emulation. 1. Data and methods, emulation differences identified in RCT-DUPLICATE, p. 6-7: please explain more precisely how the composite indicator was defined and why some characteristics (start of follow-up in hospital etc.) had been chosen from the emulation difference list to consider a “close emulation”. 2. Results: this paragraph includes results of unequal weight, and should be simplify: - With an emphasis on the key elements from their work, i.e. Figures 3.1, 3.3 and 3.5; and tables 3.1, 3.2 and 3.3 - And by moving into the appendix the Figures 3.2. (tests for heterogeneity) and 3.4 (evolution of the MSE), which are not essential to the demonstration. Only a brief summary of these results could be included in the Results paragraph. - Minor comment: please give the figure number above each figure. 3. Discussion: this paragraph is too brief and warrants several comments: - How can the authors explain the low weight of the characteristic: "placebo comparator"? This characteristic is often presented as one of the major causes of emulation failure. Is this due to the specificity

	of the studies considered in their work? Or for other reasons? - I fully agree with tis statement: ". Therefore, emphasis has to be put on understanding where those differences come from, and the clinical or research question that is being asked by each study type". Can the authors provide a short guide to help the reader rank the relative importance of these differences? - As the authors point out, the explanatory emulation characteristics are context dependent, and some may not be appropriate for other therapeutic areas. Especially, the three study emulation characteristics found as main explanatory characteristics could be poorly relevant in other clinical areas like oncology, for instance. Therefore, can the authors specify in which therapeutic areas these characteristics would be useful to researchers wishing to carry out a trial emulation? And conversely, in which areas would these conclusions not be appropriate?

REVIEWER 2	Zullo, Andrew. Competing Interest: None
REVIEW RETURNED	19-Sep-2023

GENERAL COMMENTS	Heyard and colleagues leveraged data from the RCT-DUPLICATE consortium to examine if and how differences in study design explain variation in results between randomized controlled trials (RCTs) and the observational real-world evidence (RWE) studies that were designed to emulate them. The work is excellent because the authors have employed a rigorous approach to understanding which characteristics explain variation in effect estimate differences. It is a more comprehensive and systematic examination of why the findings of RCTs and RWE studies than related prior work in the field. I particularly appreciated that the code and data to reproduce the analyses and recompile this manuscript were available through Github; I also reviewed these materials. The authors found that the observed variation in effect estimates between the RCT and RWE study pairs in their study were explained by three emulation differences in their meta-regression model: (i) in-hospital start of treatment (which is unobservable in health insurance claims data), (ii) discontinuation of certain baseline therapies at randomization (which would often be an unusual care decision in clinical practice), and (iii) delayed onset of drug effects (which would be under ascertained in real-world clinical practice due to relatively short persistence with treatment). The conclusion of the authors that a substantial proportion of the observed variation between results from RCTs and RWE studies can be attributed to design emulation differences (rather than intractable confounding) is an appropriate interpretation of their findings, in my opinion. I also agree that their study serves as a well-done example of how meta-regression could be employed to better understand emulation differences. The current research is likely to be useful to researchers as they specify causal questions about treatments and design observational RWE studies to answer these questions, especially if existing or hypothetic RCTs are being emulated. I also believe that the research will be useful to researchers seeking to conduct their own studies of why RCTs and RWEs may produce discrepant findings for some research questions. I have some additional questions and/or comments which, if addressed, I believe will enhance the quality of this manuscript. Major questions and/or comments 1. Data and Methods: The authors provide a concise but helpful

	description of the RCT-DUPLICATE collection of trials. I understand that detailed information about the RCT-DUPLICATE is available in earlier papers or elsewhere, but I feel that providing the following additional details would make it easier for readers to follow this paper: • Page 5, Lines 51-56: The three RWE data sources used to emulate the randomized trials are specified, with only two of the databases containing the first year of data utilized. The first and last years of data utilized from all the three RWE database sources should be specified. • Page 5, Line 57: What is the rationale for focusing on hazard ratios as the measure of effect? Would the selection of alternative measures (e.g., risk ratios or risk differences) markedly change the inferences? An explanation for why only randomized trials where the hazard ratio was the primary measure of effect were included should be provided. • I would recommend adding a table with the names of the RCT-RWE pairs, research question the pair of studies addressed, their causal estimands (e.g., intention to treat, per-protocol, or both), and hazard ratios (of the pooled RWE and RCT pairs) to the supplemental material section. Doing so is likely to meaningfully further enhance complete transparency about the data and methods. I suspect that readers may appreciate having sufficient information readily available to them to understand the study components without having to search for prior papers that they may not be able to obtain full-text access to. 2. Data and Methods: What were the causal contrasts for each trial emulation? Were the hazard ratios calculated as observational analogues of the intention-to-treat or per-protocol estimands in the randomized studies? Do the authors believe that differences in the target estimands may also explain differences between the RCTs and RWEs? 3. Data and Methods: I would imagine that follow-up time is an important potential design attribute in the emulation of randomized studies. How closely was follow-up emulated in the RWE studies? Do the authors agree that the duration of follow-up is an important design attribute? And if so, why was follow-up duration not included as a covariate in the meta-analysis regression and thereby investigated as a potential source of differences in effect estimates between the RWE and RCT studies? 4. Data and Methods: Were both unadjusted and adjusted hazard ratios, or, only adjusted hazard ratios, from the RWE studies compared to the hazard ratios in the RCTs? 5. Data and Methods: The authors mention that the analysis is “exploratory” in several places in the manuscript. I did not understand what makes their analysis any more or less exploratory in nature than any other study. Could the authors clarify what exactly they mean that their analysis is “of an exploratory nature” and explain what would make their analysis (or a subsequent one) “decisive” or “definitive”? 6. Results: I would expect that, in addition to the differences in design, differences in the distributions of effect measure modifiers between the RCT and RWE study populations would account for some meaningful amount of the variation in effect estimate differences. Age and sex did not appear to result in notable

	decreases in heterogeneity. Could the authors comment on A) whether that might be because age and sex were either not effect measure modifiers for treatments examined in the studies that were included; B) whether other relevant effect measure modifiers are likely to have been omitted as explanatory characteristics; and C) whether it is likely that the relevant effect measure modifiers would not be common across all of the studies included in the analysis? 7. Discussion: In the results section of the abstract, the authors state that study emulation characteristics that explained most of the variation in the results between the RCTs and the RWE studies designed to emulate them were 1) delayed treatment effect, 2) discontinuation during run-in period and 3) in-hospital start of treatment. Readers may benefit from further details on how exactly emulation elements specific to these characteristics in the RWEs varied from the corresponding RCTs they emulated (e.g., (i) in-hospital start of treatment (which is unobservable in health insurance claims data), (ii) discontinuation of certain baseline therapies at randomization (which would often be an unusual care decision in clinical practice), and (iii) delayed onset of drug effects (which would be under ascertained in real-world clinical practice due to relatively short persistence with treatment). 8. Discussion: As the authors mention briefly, residual confounding could also still explain differences in effect estimates for the RWE-RCT pairs. Did the authors consider including a covariate for probability of residual confounding in the RWE-RCT pair in the meta-regression model? Minor questions and/or comments 1. Results, Table 3.2: To emphasize the takeaway that the closer the residual heterogeneity value is to 1, the more that characteristic explains part of the variations, as well as to facilitate readability of the data in this table, perhaps it could instead be arranged in ascending order of the residual heterogeneity values? 2. Results, Figure 3.2: For the histogram of the p-values from the Q-test for heterogeneity within the RCT-RWE pairs, adding a Y-axis label might be helpful for improving readers' understanding. I'm unsure of whether the Y-axis variable for this histogram is the same or different from the Y-axis variable of the histogram of the distribution of the observed standardized differences of the RCT-RWE study pairs, especially because these two panel graphs are examining two different x-axis variables. 3. Results, Figure 3.3: A very minor point, but the forest plot showing the difference in log hazard ratios observed in the RCT and the pooled RWE (as RWE-RCT) might benefit from a "Trial Name" label on the x-axis to identify what the x-axis represents. It might also be helpful to edit the title for this figure to specify that the forest plot shows the difference in hazard ratios on the log transformed scale. 4. Results, Figure 3.4: To enhance clarity, I think it would be helpful to label the minimum-MSE marked with a circle and the broken line at the value of the minimum MSE + 1 standard error of the minimum-MSE on the graph. Doing so will probably make it is easier for readers to understand what those are without referring to the figure legend or manuscript text, but this is a minor suggestion for

	the authors' consideration. Thank you for the opportunity to review this work.

REVIEWER 3	Danaei, Goodarz. Competing Interest: None
REVIEW RETURNED	25-Sep-2023

GENERAL COMMENTS	Summary: The authors have used the results of a large study on emulating 32 RCTs using real-world evidence to investigate the magnitude and the reason for differences between these sources of evidence. Considering that this line of research is fairly novel and that many other publications have only used one comparison, I think there is a wealth of information that has been used in this analysis. Therefore, my comments summarized below are more around presenting this information in a more understandable way for a clinical audience. As is, I think a lot of emphasis is on details of the analytical methods and quantitative results which are not as intuitive. Major comments 1- I think most clinicians would be interested in knowing whether the emulation of the trial gave 'similar' results, i.e., significant and beneficial or harmful effects, vs. non-significant results. Therefore, I suggest reporting at least one set of results that tells readers if the emulations overall agreed with the evidence from the trials. This could be done by reporting (e.g., in the second paragraph of Results) how many emulations had P values less than 0.05 or 0.1. 2- It would also be great if the authors gave a more intuitive interpretation of the standardized mean difference results, e.g., by taking either the pooled delta or one from one emulation and interpreting it. I would add the pooled delta and its confidence interval to the first sentence of the Results section both in the main text and in the Abstract. 3- Similarly, a more intuitive quantity of interest to report would be the proportion of variation in the difference between RCTs and RWEs that is explained by each (set of) item(s). Indeed, the first sentence of the last paragraph of the Discussion alludes to this quantity of interest but this has not been reported elsewhere in the manuscript. 4- The list of the study characteristics seems limited. For example, the complexity of the treatment regimen (dosing, frequency, duration), which is often hard to emulate, has not been incorporated. Other potential characteristics include eligibility criteria other than sex and age, and proportion of and reasons for loss to follow-up and follow-up duration. 5- I would also suggest separately specifically 'appropriate adjustment for potential confounders' in the RWE. How were the potential confounders chosen (e.g., using a DAG)? and how were they adjusted for (e.g., regression, propensity score, IPW)? 6- A more detailed explanation of the RWD for emulation would be helpful. I suppose these are all insurance claims databases, but the introduction also mentions EMR sources so it would be great to clarify this. 7- Did the source of data for RWD matter in how close the results of the emulation were? I suppose that different RCTs may be more closely emulated using one source vs. than others.

	Minor comments 8- Methods: page 6: please clarify the final number of trials in the analysis. 9- Table 2.1: the fourth item on the list seems to overlap with the first one. Please clarify. 10- Table 2.1: the fifth item: please clarify “therapy in RCT started in the hospital” 11- Table 2.1: the last item: “delayed effect” is one of the reasons for lower adherence in RWD compared with RCTs. why not directly capture differences in adherence between RCT and RWE? 12- Table 3.1 is rather thin. These numbers can be easily reported in the text. 13- Methods, page 7, paragraph 2: this part can be re-written to start with the simpler analysis of pooling the difference in logHR across trials and then explaining how heterogeneity was quantified. 14- Figure 1: I suggest changing the pattern/thickness of the CI95 bars for RCT vs. RWD so that readers can interpret the differences by examining whether one or both of them cross the diagonal (see comment #1). If neither cross, then the P value for the comparison would be below 0.05. 15- I would move both panels of Figure 3.2 to the Appendix. Please also add more detail to the x-axis label of the left panel (see comment #1 above about reporting the right panel in the text) . 16- Results: in a few sections text that belongs to Methods has been repeated here. E.g., page 8, lines 51-52 and page 9 lines 6-10. 17- Table 3.2: the second largest impact on reducing heterogeneity comes from the ‘run-in phase’ but this has not been mentioned in the Results or Discussion. Please add one or two sentences as this is one of the items that is almost impossible to emulate using RWD. 18- Table 3.3: Can the LOO MSE for different models be reported as a column? 19- Appendix: Delta(i) and Delta have not been introduced. Please add. 20- Appendix: line 33, please briefly justify truncating the SE at 1 for psi. 21- Appendix, last three lines: please clarify what was done if the model MSEs had overlapping CI95s. 22- Figure 3: I would suggest grouping studies and reporting the pooled difference in logHR for each set as opposed to sorting by study name. This would then include all the information presented on the left panel of Fig 5, which can now be removed. 23- Figure 5B can be a separate Figure. The x-axis labels have been misplaced to the right. 24- Figure 4: I would suggest moving this to the Appendix and removing the star on the fifth bar as the conclusion is to use a model with 3 characteristics. Please explain in the caption what the dashed line represents.

REVIEWER 4	Davies, Neil; University of Bristol. Competing Interest: None
REVIEW RETURNED	02-Nov-2023

GENERAL COMMENTS	This is a very interesting, carefully conducted and important study. It systematically assesses estimates from RWE and RCTs, and is one of the most rigorous attempts to do this to date. I have only minor comments:

	1. In the introduction it would be worth sketching out the limitations to RCTs and why evidence from EHR could be helpful, this is all very well known, but worth adding in this motivation. E.g. lack of follow-up, representative both in terms of sample and adherence, cost, timeliness, treatment effect heterogeneity, I'm sure there are others. 2. The pooled RWE analysis, are surprisingly imprecise. I would have expected far larger sample size and precision than the trials. Or am I misreading Figure 3.1? 3. In general, it'd be good to comment on and quantify (if possible?) a) the relative representativeness, b) sample sizes, and c) follow-up of the RWE vs RCT evidence. (You may have equalised some of these as part of your protocol). 4. Table 3.2 in the discussion about this it'd be good to clearly state to the reader which variables are most important, looks like you've evidence 3 can explain the residual heterogeneity, and the rest are less clear. 5. Are the 29 trials emulated representative of all trials? You touch on this in your discussion, but would it be worth sketching out how (if?) we could definitively answer this? 6. Pre-reregistration – I'm pretty sure this was all pre-reregistered, I may have missed this, but I didn't spot it in the manuscript. 7. Could you use a check list for EHR studies. E.g. code-EHR/STROBE or similar? Typos etc 1. P3 "estimates be2tween RCT-RWE st 2. P9 "All 210 = 1′024 possible" 3. Table 3.2 what is the order these factors are presented in? Could you order via residual heterogeneity?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Target trial emulation aims to prevent biases by applying RCT methodological and design principles to observational data. Therefore, it is expected that design emulation differences are related to variation in results between RCT and RWD studies. This paper demonstrates what kind of study emulation characteristics could explain most of the observed variation by comparing results from RCT-RWD study pairs based on data from the well-known RCT-DUPLICATE initiative. The paper has many excellent strengths, including a rigorous selection of trials from the RCT DUPLICATE which could be used in the study, a relevant choice and description of numerical and categorical emulation difference characteristics, use of meta-regression methods to estimate heterogeneity, and use of cross-validated mean squared errors method to reduce the complexity of the meta-regression.

Nevertheless, there are several issues which should be addressed to simplify and to make this paper more accessible to a large audience of readers, including researchers and clinicians interested in performing trial emulation.

1. Data and methods, emulation differences identified in RCT-DUPLICATE, p. 6-7: please explain more precisely how the composite indicator was defined and why some characteristics (start of follow-up in hospital etc.) had been chosen from the emulation difference list to consider a "close emulation".

Reply: We agree that more information on the definition of the composite indicator is needed. The sentence describing the indicator was therefore extended into a bullet list to improve readability. Further explanation on the reasons for choosing the specific characteristics was also added.

2. Results: this paragraph includes results of unequal weight, and should be simplify:
    - With an emphasis on the key elements from their work, i.e. Figures 3.1, 3.3 and 3.5; and tables 3.1, 3.2 and 3.3 and by moving into the appendix the Figures 3.2. (tests for heterogeneity) and 3.4 (evolution of the MSE), which are not essential to the demonstration. Only a brief summary of these results could be included in the Results paragraph.

Reply: Thank you for your suggestion. The results section was shortened and simplified in the proposed way: Figures 3.2 and 3.4 were moved to the paper supplement. The results are briefly summarized in the main text.

    - Minor comment: please give the figure number above each figure.

Reply: We submitted our Figures separately from the manuscript in ScholarOne Manuscripts and believe that this will be addressed in the final formatting step of the hopefully published version of our manuscript.

3. Discussion: this paragraph is too brief and warrants several comments:
    - How can the authors explain the low weight of the characteristic: "placebo comparator"? This characteristic is often presented as one of the major causes of emulation failure. Is this due to the specificity of the studies considered in their work? Or for other reasons?

Reply: It is indeed very likely that the results were influenced by the way placebo-controlled trials were emulated in RCT-DUPLICATE. As described in the supplement of Wang et al (2023), the team “chose not to conduct non-user comparisons because of expected differences between patients that are treated versus not treated on characteristics that are either unmeasured or poorly captured in clinical practice data.” Instead, they “identified active comparators that could proxy for placebo because they were expected to have no effect on the outcome of interest”. As noted in Wang et al (2023), the active comparator placebo proxies varied in terms of how well they proxied for placebo. For four trials, the placebo proxy was classified as of poor emulation quality, for the remainder, the quality was moderate. The heterogeneity in quality of placebo comparators across RCT-database pairs likely contributed to the low weight for this characteristic. A sentence was added in the first paragraph of the discussion: “This result might have been influenced by the quality of the placebo proxy that was used in emulation of placebo-controlled trials for RCT-DUPLICATE.”

    - I fully agree with this statement: “Therefore, emphasis has to be put on understanding where those differences come from, and the clinical or research question that is being asked by each study type”. Can the authors provide a short guide to help the reader rank the relative importance of these differences?

Reply: Thank you for this comment. We would like to give more guidance to the reader, however the results of the meta-regression are context-dependent and not generalisable. We added a sentence to clarify that our results inform understanding of how concordance in results is influenced by concordance in design, but the specific coefficients should not be interpreted as generalizable due to the highly selected sample of trials “Therefore, our results can inform understanding of how concordance in results between RCTs and database studies are influenced by concordance in design, but the specific coefficients should not be interpreted as generalizable due to the highly selected sample of trials.”

    - As the authors point out, the explanatory emulation characteristics are context dependent, and some may not be appropriate for other therapeutic areas. Especially, the three study emulation characteristics found as main explanatory characteristics could be poorly relevant in other clinical areas like oncology, for instance. Therefore, can the authors specify in which therapeutic areas these characteristics would be useful to researchers wishing to carry out a trial emulation? And conversely, in which areas would these conclusions not be appropriate?

Reply: We agree that this information would be great to have. However, more research in this area is needed to give a thorough answer to this interesting question you raise. Even though many of the included trials evaluate cardiovascular outcomes, our results are too limited to generalize. Additionally, the data we used was generated through emulations in insurance claims data. If instead, the RCTs were emulated with registry data or data from electronic health records other emulation characteristics were to be included. A sentence was added in the limitations paragraph of our Discussion section: “Further, the emulations were conducted within insurance claims data. RCTs emulated using registry data or data from electronic health records may be challenged with additional design emulation differences, for instance, challenges to defining “observable time” when using data from fragmented healthcare systems.”

Reviewer: 2

Heyard and colleagues leveraged data from the RCT-DUPLICATE consortium to examine if and how differences in study design explain variation in results between randomized controlled trials (RCTs) and the observational real-world evidence (RWE) studies that were designed to emulate them. The work is excellent because the authors have employed a rigorous approach to understanding which characteristics explain variation in effect estimate differences. It is a more comprehensive and systematic examination of why the findings of RCTs and RWE studies than related prior work in the field. I particularly appreciated that the code and data to reproduce the analyses and recompile this manuscript were available through Github; I also reviewed these materials. The authors found that the observed variation in effect estimates between the RCT and RWE study pairs in their study were explained by three emulation differences in their meta-regression model: (i) in-hospital start of treatment (which is unobservable in health insurance claims data), (ii) discontinuation of certain baseline therapies at randomization (which would often be an unusual care decision in clinical practice), and (iii) delayed onset of drug effects (which would be under ascertained in real-world clinical practice due to relatively short persistence with treatment). The conclusion of the authors that a substantial proportion of the observed variation between results from RCTs and RWE studies can be attributed to design emulation differences (rather than intractable confounding) is an appropriate interpretation of their findings, in my opinion. I also agree that their study serves as a well-done example of how meta-regression could be employed to better understand emulation differences. The current research is likely to be useful to researchers as they specify causal questions about treatments and design observational RWE studies to answer these questions, especially if existing or hypothetic RCTs are being emulated. I also believe that the research will be useful to researchers seeking to conduct their own studies of why RCTs and RWEs may produce discrepant findings for some research questions. I have some additional questions and/or comments which, if addressed, I believe will enhance the quality of this manuscript.

Major questions and/or comments

1. Data and Methods: The authors provide a concise but helpful description of the RCT-DUPLICATE collection of trials. I understand that detailed information about the RCT-DUPLICATE is available in earlier papers or elsewhere, but I feel that providing the following additional details would make it easier for readers to follow this paper:
    - Page 5, Lines 51-56: The three RWE data sources used to emulate the randomized trials are specified, with only two of the databases containing the first year of data utilized. The first and last years of data utilized from all the three RWE database sources should be specified.

Reply: Thank you for pointing this lack of reporting out. We now added both to all data sources.

    - Page 5, Line 57: What is the rationale for focusing on hazard ratios as the measure of effect? Would the selection of alternative measures (e.g., risk ratios or risk differences)

markedly change the inferences? An explanation for why only randomized trials where the hazard ratio was the primary measure of effect were included should be provided.

Reply: To run the analysis the estimates of all included studies have to be on the same scale, either by nature (as in our case) or through appropriate transformation. Alternative measures, like risk ratios or mean differences, can be used. If the included studies were investigating a mix of measures, those would have to be transformed to the same scale before the analysis. We added the following comment in the Discussion section: "Even though nearly all included studies focused on a hazard ratio for the primary result, the proposed analysis can be applied to studies investigating other outcome measures, *i.e.*, risk ratios or risk differences. However, the meta-regression analyses require that the estimates for all studies be on the same scale. Appropriate transformations could be applied to include studies whose primary analyses use a different scale."

- I would recommend adding a table with the names of the RCT-RWE pairs, research question the pair of studies addressed, their causal estimands (e.g., intention to treat, per-protocol, or both), and hazard ratios (of the pooled RWE and RCT pairs) to the supplemental material section. Doing so is likely to meaningfully further enhance complete transparency about the data and methods. I suspect that readers may appreciate having sufficient information readily available to them to understand the study components without having to search for prior papers that they may not be able to obtain full-text access to.

Reply: We agree with you that the reader of our manuscript should have access to all the relevant information for our study without relying on having access to other papers. Therefore we extended the information on the included RCT-RWE pairs and the results from the main RCT-DUPLICATE investigation in a table in the Supplement of our paper. Note that, as also mentioned in the paragraph before the table, all results used in our analysis originate from a "while on-treatment" analysis.

2. Data and Methods: What were the causal contrasts for each trial emulation? Were the hazard ratios calculated as observational analogues of the intention-to-treat or per-protocol estimands in the randomized studies? Do the authors believe that differences in the target estimands may also explain differences between the RCTs and RWEs?

Reply: The primary analysis in RCT-DUPLICATE was a "while on-treatment" analysis. This version of analysis was chosen because the persistence with treatment was generally shorter in clinical practice while adherence was higher in RCTs. Intention-to-treat analyses were done as a sensitivity analysis for each emulation, with the results summarized in the supplement of Wang et al. (2023). The expression "while on-treatment analysis" and the reason for the analysis were added in the Methods section (as well as before the new table in the supplement).

3. Data and Methods: I would imagine that follow-up time is an important potential design attribute in the emulation of randomized studies. How closely was follow-up emulated in the RWE studies? Do the authors agree that the duration of follow-up is an important design attribute? And if so, why was follow-up duration not included as a covariate in the meta-analysis regression and thereby investigated as a potential source of differences in effect estimates between the RWE and RCT studies?

Reply: We agree that duration of follow up is an important design attribute. This is indirectly captured by the covariate "delayed effects during follow up", which equals 1 if the RCT cumulative incidence curves suggested delayed or time-varying effects in the context of often substantially longer follow up for the RCT compared to the database study.

We have expanded our discussion to include the limitation that the characteristics that we included in our analysis are a subset of all possible characteristics that could be considered. However, differences in emulation of follow-up in the context of delayed or time-varying effects was identified as an important design emulation difference and contributor to divergence in RCT-RWE results.

4. Data and Methods: Were both unadjusted and adjusted hazard ratios, or, only adjusted hazard ratios, from the RWE studies compared to the hazard ratios in the RCTs?

Reply: Only adjusted hazard ratios from the RWE studies were compared to the adjusted hazard ratios in the RCTs. We added a note on this in our methods section: “The version of the hazard ratios that were adjusted for confounding through propensity score matching as described in [19] were used for the RWE-RCT comparisons.”

5. Data and Methods: The authors mention that the analysis is “exploratory” in several places in the manuscript. I did not understand what makes their analysis any more or less exploratory in nature than any other study. Could the authors clarify what exactly they mean that their analysis is “of an exploratory nature” and explain what would make their analysis (or a subsequent one) “decisive” or “definitive”?

Reply: We want to highlight the fact that our analysis is exploratory and not confirmatory, and that the conclusions drawn from our analysis have to be further scrutinized in follow-up studies. The RCT DUPLICATE data we are using in this analysis was not collected for this purpose. A follow-up study investigating the same research question might collect more specific data and information with an assumed relation to heterogeneity. In that sense, our study is only exploratory and limited from drawing strong conclusions. We added a footnote in the Data and Methods Section to briefly explain the difference between exploratory and confirmatory studies.

6. Results: I would expect that, in addition to the differences in design, differences in the distributions of effect measure modifiers between the RCT and RWE study populations would account for some meaningful amount of the variation in effect estimate differences. Age and sex did not appear to result in notable decreases in heterogeneity. Could the authors comment on A) whether that might be because age and sex were either not effect measure modifiers for treatments examined in the studies that were included; B) whether other relevant effect measure modifiers are likely to have been omitted as explanatory characteristics; and C) whether it is likely that the relevant effect measure modifiers would not be common across all of the studies included in the analysis?

Reply: The estimated effects that were examined were hazard ratios. The absence of notable heterogeneity related to age and sex on the multiplicative scale implies that there is heterogeneity on the additive scale. It is certainly possible that other important risk factors could be important effect modifiers and that these may not be common across all studies. The lack of commonality would preclude us from using these other potential risk factors in the meta-regression.

7. Discussion: In the results section of the abstract, the authors state that study emulation characteristics that explained most of the variation in the results between the RCTs and the RWE studies designed to emulate them were 1) delayed treatment effect, 2) discontinuation during run-in period and 3) in-hospital start of treatment. Readers may benefit from further details on how exactly emulation elements specific to these characteristics in the RWEs varied from the corresponding RCTs they emulated (e.g., (i) in-hospital start of treatment (which is unobservable in health insurance claims data), (ii) discontinuation of certain baseline therapies at randomization (which would often be an unusual care decision in clinical practice), and (iii) delayed onset of drug effects (which would be under ascertained in real-world clinical practice due to relatively short persistence with treatment).

Reply: Thank you for your suggestion to make the abstract easier to follow. We accepted the proposed changes.

8. Discussion: As the authors mention briefly, residual confounding could also still explain differences in effect estimates for the RWE-RCT pairs. Did the authors consider including a covariate for probability of residual confounding in the RWE-RCT pair in the meta-regression model?

Reply: No, we only included covariates collected and presented in the RCT-DUPLICATE JAMA paper. We do not have a good measure for the probability of residual confounding. In theory, such a covariate or others could be investigated. We have added such text to our Discussion section.

Minor questions and/or comments

1. Results, Table 3.2: To emphasize the takeaway that the closer the residual heterogeneity value is to 1, the more that characteristic explains part of the variations, as well as to facilitate readability of the data in this table, perhaps it could instead be arranged in ascending order of the residual heterogeneity values?

Reply: We agree with your comment; the table was adjusted accordingly.

2. Results, Figure 3.2: For the histogram of the p-values from the Q-test for heterogeneity within the RCT-RWE pairs, adding a Y-axis label might be helpful for improving readers' understanding. I'm unsure of whether the Y-axis variable for this histogram is the same or different from the Y-axis variable of the histogram of the distribution of the observed standardized differences of the RCT-RWE study pairs, especially because these two panel graphs are examining two different x-axis variables.

Reply: The y-axis was both times the density, but since this might not have been clear, the Y-axis was added to both plots. Note that this Figure was now added in the new supplement we created (and not Figure 3.2 anymore).

3. Results, Figure 3.3: A very minor point, but the forest plot showing the difference in log hazard ratios observed in the RCT and the pooled RWE (as RWE-RCT) might benefit from a "Trial Name" label on the x-axis to identify what the x-axis represents. It might also be helpful to edit the title for this figure to specify that the forest plot shows the difference in hazard ratios on the log transformed scale.

Reply: Thanks for these detailed comments on our Figues. An axis title "Trial Name" was added to the x-axis to help readability.

4. Results, Figure 3.4: To enhance clarity, I think it would be helpful to label the minimum-MSE marked with a circle and the broken line at the value of the minimum MSE + 1 standard error of the minimum-MSE on the graph. Doing so will probably make it easier for readers to understand what those are without referring to the figure legend or manuscript text, but this is a minor suggestion for the authors' consideration.

Reply: We agree with your suggestion and now added two labels, "min MSE" and "min MSE + 1 SE" to the plot.

Thank you for the opportunity to review this work.

Reviewer: 3

Summary:

The authors have used the results of a large study on emulating 32 RCTs using real-world evidence to investigate the magnitude and the reason for differences between these sources of evidence. Considering that this line of research is fairly novel and that many other publications have only used one comparison, I think there is a wealth of information that has been used in this analysis. Therefore, my comments summarized below are more around presenting this information in a more

understandable way for a clinical audience. As is, I think a lot of emphasis is on details of the analytical methods and quantitative results which are not as intuitive.

Major comments

1. I think most clinicians would be interested in knowing whether the emulation of the trial gave 'similar' results, i.e., significant and beneficial or harmful effects, vs. non-significant results. Therefore, I suggest reporting at least one set of results that tells readers if the emulations overall agreed with the evidence from the trials. This could be done by reporting (e.g., in the second paragraph of Results) how many emulations had P values less than 0.05 or 0.1.

Reply: Investigating agreement between RCT and the RWE emulation was the aim of the RCT-DUPLICATE initiative. As such, these results are included in the main RCP-DUPLICATE output paper (Wang et al. (2023)). Since another reviewer also asked for more information on the included study pairs we decided to add a table on the results of the initiative in the supplement to our paper. The last column in Table A.1 summarizes the results of when applying three agreement metrics: one based on statistical significance, one based on the effect estimates and one based on standardized differences.

2. It would also be great if the authors gave a more intuitive interpretation of the standardized mean difference results, e.g., by taking either the pooled delta or one from one emulation and interpreting it. I would add the pooled delta and its confidence interval to the first sentence of the Results section both in the main text and in the Abstract.

Reply: We computed the average delta, the difference in log-hazard ratios, and added the following sentence in the Results section: "The average difference in log hazard ratio over all included trials is estimated to be slightly negative (-0.015 with 95% confidence interval [-0.084; 0.054]), suggesting that, on average, the hazard ratio estimated using the RWD is larger than in the RCT." We additionally added a sentence on the direction of this difference in the Discussion. Since the discussion on the standardized differences was moved to the supplement we did not further interpret those results in the main text to avoid confusions.

3. Similarly, a more intuitive quantity of interest to report would be the proportion of variation in the difference between RCTs and RWEs that is explained by each (set of) item(s). Indeed, the first sentence of the last paragraph of the Discussion alludes to this quantity of interest but this has not been reported elsewhere in the manuscript.

Reply: This information is indirectly given through the heterogeneity which can be interpreted as a variance inflation parameter. As such, a covariate which reduces the heterogeneity parameter, explains part of the variation. We added R squared values to Table 3.2 which are indeed directly linked to the residual heterogeneity.

4. The list of the study characteristics seems limited. For example, the complexity of the treatment regimen (dosing, frequency, duration), which is often hard to emulate, has not been incorporated. Other potential characteristics include eligibility criteria other than sex and age, and proportion of and reasons for loss to follow-up and follow-up duration.

Reply: As previously mentioned in the comments and now clarified in our Discussion section, we added all covariates recorded as such by the RCT-DUPLICATE initiative in the meta-regression. Other covariates could be included and play an important role in decreasing the residual heterogeneity which should be investigated in follow-up studies.

5. I would also suggest separately specifically 'appropriate adjustment for potential confounders' in the RWE. How were the potential confounders chosen (e.g., using a DAG)? and how were they adjusted for (e.g., regression, propensity score, IPW)?

Reply: As mentioned now in the Methods section, the adjusted estimates of the hazard ratio were adjusted for confounding through 1:1 nearest neighbor propensity score matching on prespecified risk

factors for the outcome that are associated with exposure. These risk factors were chosen in consultation with clinical experts. All details of the protocols are now linked in the supplement of our paper.

6. A more detailed explanation of the RWD for emulation would be helpful. I suppose these are all insurance claims databases, but the introduction also mentions EMR sources so it would be great to clarify this.

Reply: All RWD used in the RCT-DUPLICATE emulations is from insurance claims data. In the Discussion section we now distinguish it from data from registries or EHR, so that the data source should be clearer now.

7. Did the source of data for RWD matter in how close the results of the emulation were? I suppose that different RCTs may be more closely emulated using one source vs. than others.

Reply: The RWE emulations were powered based on the pooled results (from multiple databases). Hence, comparing emulations from specific RWD databases in the meta-regression is therefore not advised but in theory possible.

Minor comments

8. Methods: page 6: please clarify the final number of trials in the analysis.

Reply: The sentence “In total 29 trials are included in the analysis.” was added to the Methods section. The final sample size was also added in the Methods part of the Abstract.

9. Table 2.1: the fourth item on the list seems to overlap with the first one. Please clarify.

Reply: We agree that they are related, while the covariate “Comparator emulation” goes further than a yes-no indicator. In general, we expect the covariates to be dependent, which is why the variable selection is so important for the meta-regression. A sentence was added in the Methods section: “Many of the included characteristics are suspected to be dependent.”

10. Table 2.1: the fifth item: please clarify “therapy in RCT started in the hospital”

Reply: The description was extended.

11. Table 2.1: the last item: “delayed effect” is one of the reasons for lower adherence in RWD compared with RCTs. why not directly capture differences in adherence between RCT and RWE?

Reply: Thank you for your suggestion. Delayed effect is not a reason for lower adherence, however, given low adherence in clinical practice, it may contribute to divergence in effect estimates in the presence of delayed effects. We felt that measuring whether there was likely to be divergence in results due to difference in adherence would be more informative than capturing lower adherence in clinical practice, which is unlikely to vary across the sample.

12. Table 3.1 is rather thin. These numbers can be easily reported in the text.

Reply: We believe the Table still increases readability and will help with the understanding of the next more complex table. Therefore we suggest keeping it.

13. Methods, page 7, paragraph 2: this part can be re-written to start with the simpler analysis of pooling the difference in logHR across trials and then explaining how heterogeneity was quantified.

Reply: We agree that the second paragraph in describing the statistical analysis has become rather long and complex. To help readability we split the paragraph into two (specifically after the method to quantify heterogeneity using a simple model is explained, and before the characteristics are introduced in a meta-regression).

14. Figure 1: I suggest changing the pattern/thickness of the CI95 bars for RCT vs. RWD so that readers can interpret the differences by examining whether one or both of them cross the diagonal (see comment #1). If neither cross, then the P value for the comparison would be below 0.05.

Reply: The thickness of the lines was increased in the first Figure. As for comment #1, we added the agreement criteria in the Table in the supplement.

15. I would move both panels of Figure 3.2 to the Appendix. Please also add more detail to the x-axis label of the left panel (see comment #1 above about reporting the right panel in the text).

Reply: The previous Figures 3.2 and 3.4 were moved into the supplement of the paper as suggested by another reviewer.

16. Results: in a few sections text that belongs to Methods has been repeated here. E.g., page 8, lines 51-52 and page 9 lines 6-10.

Reply: We agree with your observation and shortened, respectively deleted those sentences.

17. Table 3.2: the second largest impact on reducing heterogeneity comes from the 'run-in phase' but this has not been mentioned in the Results or Discussion. Please add one or two sentences as this is one of the items that is almost impossible to emulate using RWD.

Reply: We changed the ordering of the Table with increasing residual heterogeneity which should already help highlighting this.

18. Table 3.3: Can the LOO MSE for different models be reported as a column?

Reply: We decided against this to simplify the Table. The information on LOO MSE etc is in the Figure which is now in the supplement.

19. Appendix: Delta(i) and Delta have not been introduced. Please add.

Reply: Thank you for these detailed comments. We introduced both Deltas.

20. Appendix: line 33, please briefly justify truncating the SE at 1 for psi.

Reply: We rearranged the sentence and hope the reason for the truncation becomes clear.

21. Appendix, last three lines: please clarify what was done if the model MSEs had overlapping CI95s.

Reply: We did not use 95% CI to select the final model, but only the intervals MSE +/- 1 SE. For the final selection only the point estimate was used to find the smallest model and not its interval. For clarity we added a last sentence in the Variable selection paragraph.

22. Figure 3: I would suggest grouping studies and reporting the pooled difference in logHR for each set as opposed to sorting by study name. This would then include all the information presented on the left panel of Fig 5, which can now be removed.

Reply: The goal of Figure 3 is to show the observed differences in heterogeneity etc depending on the characteristic "Close emulation", while Figure 5 presents the predicted heterogeneity given the regression model. In that sense, the information is quite different, and we would prefer keeping both Figures.

23. Figure 5B can be a separate Figure. The x-axis labels have been misplaced to the right.

Reply: Thank you for pointing out the misplaced x-axis labels, we corrected this. See above for the first part of the comment.

24. Figure 4: I would suggest moving this to the Appendix and removing the star on the fifth bar as the conclusion is to use a model with 3 characteristics. Please explain in the caption what the dashed line represents.

Reply: The Figure was moved to the supplement. More information was added in the Figure and caption.

Reviewer: 4

bmjmed-2023-000709, entitled "Design differences explain variation in results between randomized trials and their non-randomized emulations."

This is a very interesting, carefully conducted and important study. It systematically assesses estimates from RWE and RCTs, and is one of the most rigorous attempts to do this to date.

I have only minor comments:

1. In the introduction it would be worth sketching out the limitations to RCTs and why evidence from EHR could be helpful, this is all very well known, but worth adding in this motivation. E.g. lack of follow-up, representative both in terms of sample and adherence, cost, timeliness, treatment effect heterogeneity, I'm sure there are others.

Reply: We believe that these limitations in using RWD over RCTs are already included in the introduction, when we mention the differences and give examples. To clarify, RCT-DUPLICATE did not use EHR, but insurance claims data. We specified this now in the introduction.

2. The pooled RWE analysis, are surprisingly imprecise. I would have expected far larger sample size and precision than the trials. Or am I misreading Figure 3.1?

Reply: One would indeed expect that the RWEs could use larger samples, and would therefore result in more precise estimates, with smaller confidence intervals. However, RCT-DUPLICATE aimed at emulating the RCTs as close as possible and powered the RWE analysis as such. The final pooled RWE studies resulted in at least as much power as used in the RCT (sometimes it was a bit higher, sometimes the initiative barely made it) .

3. In general, it'd be good to comment on and quantify (if possible?) a) the relative representativeness, b) sample sizes, and c) follow-up of the RWE vs RCT evidence. (You may have equalised some of these as part of your protocol).

Reply: All of this is in the RCT-DUPLICATE protocol(s), supplement to Wang et al. (2023). As the exact emulation method and intermediate results are not in the scope of this paper, we decided against adding this information. We did however add a Table summarizing some of the final results of RCT-DUPLICATE in out supplementent.

4. Table 3.2 in the discussion about this it'd be good to clearly state to the reader which variables are most important, looks like you've evidence 3 can explain the residual heterogeneity, and the rest are less clear.

Reply: The table was reordered as to start with the characteristic having most "influence" on the residual heterogeneity, which should already make it clearer. We added more discussion on the different characteristics and their respective relevance from other reviewer's comments.

5. Are the 29 trials emulated representative of all trials? You touch on this in your discussion, but would it be worth sketching out how (if?) we could definitively answer this?

Reply: The 29 included trials are definitely not representative of all trials. The RCT DUPLICATE selection is very specific. We already said that more research is needed here, if one wanted to generalize our findings - and now made this more explicit in the Discussion section.

6. Pre-reregistration – I'm pretty sure this was all pre-reregistered, I may have missed this, but I didn't spot it in the manuscript.

Reply: The RCT-DUPLICATE study was indeed pre-registered and all emulations are registered on clinicaltrials.gov., as described in Wang et al. (2023). We also added all the NCT numbers of the RCTs and emulations included in RCT-DUPLICATE in our supplement to further clarify. This particular follow-up analysis was post-hoc and purely explorative. As such it was not pre-registered. We however encourage other researchers, and might do so ourselves, to potentially confirm part of our generated hypothesis in a follow-up confirmatory and preregistered study.

7. Could you use a check list for EHR studies. E.g. code-EHR/STROBE or similar?

Reply: As our analysis is an observational study using secondary data from RCT-DUPLICATE we filled out the STROBE checklist and included it as supplement to our manuscript.

Typos etc

1. P3 “estimates be2tween RCT-RWE st
2. P9 “All 210 = 1′024 possible”
3. Table 3.2 what is the order these factors are presented in? Could you order via residual heterogeneity?

Reply: Thank you for pointing out typos. They were corrected. As suggested by another reviewer, we did change the order of the emulation characteristics in the Table. They now appear in increasing heterogeneity order.