nature portfolio

Peer Review File

Empirical assessment of constraints and credibility when emulating a placebo-controlled trial with the clone-censor-weight approach

Corresponding Author: Dr Anna-Janina Stephan

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
OVERALL ASSESSMENT
This manuscript presents a methodologically rigorous and highly relevant empirical evaluation of the challenges inherent in emulating placebo-controlled trials using observational data. Recommendation: Minor Revision. The manuscript is of high quality and makes a valuable contribution to the literature. The recommended revisions are intended to enhance clarity, transparency, and the nuanced interpretation of findings.
1. Negative Control Outcome Assumptions
Comment: The choice of non-cardiovascular mortality as a negative control outcome is central to the bias correction strategy. While the authors acknowledge that this approach relies on the assumption of shared confounding with cardiovascular mortality, the discussion of this assumption's plausibility is somewhat brief. The assumption is strong and inherently untestable, and its violation could lead to over- or under-correction.
2. Interpretation of Benchmarking Results
Comment: The benchmarking exercise against TOPCAT is a major strength, but the interpretation of "benchmarking failure" for mortality outcomes requires more nuance. The authors note population differences (age, comorbidity) and shorter follow-up in their study, but these factors are presented primarily as limitations rather than as alternative explanations for the observed discrepancies.
3. Weight Model Misspecification
Comment: The supplementary materials (Tables S9-S12) provide valuable diagnostics showing that the pseudo-population sizes do not perfectly align with observed populations, suggesting some degree of weight model misspecification. While the authors briefly note this, they do not discuss the potential direction or magnitude of bias this might introduce.
4. Definition of "Net Bias"
Comment: The term "net bias" is used throughout the manuscript but is not explicitly defined when first introduced in the Methods section.
5. Language and Readability
Comment: Overall, the manuscript is well-written. However, a few sentences are lengthy and could be simplified for clarity.

Reviewer #2

(Remarks to the Author)
I really enjoyed reading this thoughtful and well-conducted study, and the findings are a timely reminder about the difficulties we have comparing users to non-users of medicines in pharmacoepidemiology. My comments are minor – overall I would recommend this for publication and congratulate the authors for a nice contribution to the literature.

Main comments:

1. You use CCW for the emulation which means that your treatment strategies include a grace period for starting treatment. However, the benchmarking trial appears to not have a grace period and instead follows people from randomization, which coincided with placebo / treatment initiation. Do you think part of the differences in findings might be due to the RCT and the emulation addressing different questions? E.g. “What is the risk of death after starting spironolactone” in the RCT compared

to “What is the 12-month risk of death after a HF hospitalisation if I start spironolactone within 6 months” in the emulation?

2. Related to the above – throughout the paper CCW is positioned as a method that is particularly suited for use vs. non-use comparisons because of the difficulties associated with assigning time zero for the non-user group (for example L41-45 in introduction, and L172 - 174 in the discussion). However, often the challenge with non-users isn’t indistinguishable time zero but multiple eligibility points which would lead us to consider a sequential trial design. It seems to me that in this study the indistinguishable time zero is not a result of the non-user comparison but the application of a grace period, which also changes the research question. I think the paper would benefit from clarifying the difference between indistinguishable time zero and multiple eligibility points, and from including a more explicit comparison of CCW and sequential trials (particularly in terms of the questions answered by each design).

Minor suggestions / considerations

3. Could you comment on the plausibility of the assumption of ‘consistent pattern of residual confounding’ for the NCO? Was there a specific confounder that was of concern you anticipated being captured, and how accurately do you think the NCO did this? Presumably something like severity of underlying CV disease would increase only CV specific mortality.

4. I wasn’t sure I followed the rationale for adjusting for time-updated HCRU - like number of HF hospitalisations - in an analysis where HF hospitalisation is the outcome: how did you distinguish these two? Even for other endpoints, e.g. death, I wasn’t sure whether preceding hospitalisation would act as a confounder - would spironolactone be discontinued on hospitalisation?

5. Could you expand the discussion of intercurrent events: what were the intercurrent events you considered (e.g. L345 in the methods – you mention that some intercurrent events did not result in censoring?). How did you handle death due to other causes in the analyses of cause-specific mortality?

6. How is date of death captured in Medicare after Dec 31 2016 without the linkage, and is this complete death information or restricted to e.g. in-hospital deaths?

7. Could you add more information how cause of death was defined: was this any contributing / underlying cause, or only the primary cause of death?

8. Could you expand on how you applied the 1-month time lag? Did you set events occurring within the first month to missing, start FU later on, or censor people who had events in this period? For safety endpoints, a 1-month time lag seems potentially too long.

9. The naïve analysis is an interesting demonstration, but the analysis strategy I see most often for these kinds of questions is to start FU at treatment for the treated and at diagnosis for the untreated (often still defined as never users) – effectively just discarding the time between diagnosis and treatment in the treated. There might not be scope for you to explore this, but I would be interested in seeing how biased this common (but still flawed) approach is.

10. I would consider summarizing the features of the TOPCAT trial in your TTE table (as a separate column). This would make also make the difference in the treatment strategies assessed clear.

11. P7 L170 – the constant hazard ratio assumption would presumably also apply to a CCW analysis applying a cox model?

Editorial

1. I couldn’t access the SAS code on OSF via the link in the paper (I think this is still behind a ‘permissions’ wall).

Reviewer #3

(Remarks to the Author)
This study looked at the effectiveness of spironolactone in heart failure with preserved ejection fraction, aiming to assess effectiveness in a broader population than the TOPCAT trial. While doing this, it aimed to assess how applying the clone-censor-weight approach within the target trial emulation framework compared to the TOPCAT Americas result and a negative control outcome, and further compare results with a number of approaches outside the target trial emulation framework. Approaches outside the TTE framework were found to be substantially biased, while even with clone-censor-weighting, residual confounding remained.

This is a novel and potentially important study for highlighting the utility and limitations of the clone-censor-weight approach when applied to the emulation of placebo-controlled trials.

There are a number of areas for clarification, as detailed below:

Introduction
1. Introduction line 43-47: please add an example here of what is meant by a “time zero” that is identifiable in both treatment

arms
2. Introduction line 49: In line 43, the remit/benefit of cloning, censoring and weighting is clearly defined i.e. the cloning, censoring and weighting can be applied when treatment strategies are indistinguishable at time zero. Here, a much broader statement is made about the clone-censor-weight approach which makes is sound like this approach is required in order to study a well-defined causal question that could be addressed by a corresponding RCT, when in fact there are a variety of approaches within the TTE framework that could allow investigators to "study a well-defined causal question that could be addressed by a corresponding RCT". Please either rephrase or remove this statement.
3. Introduction line 55: “investigator-induced design flaws” – does this mean the immortal bias mentioned previously? Please be specific.
4. Introduction line 56: please add the year of the trial
5. Introduction line 68: This is a very good opportunity to explain exactly why it is important to use clone-censor-weight for this question which would really help clarify the introduction i.e.: (1) there is a no-treatment group so the date of initiation of treatment cannot be used as a start date for this group (2) a start date of latest eligibility for trial is therefore used, but if this is used as time 0, there is immortal time bias as the treatment group by definition need to stay alive from HF diagnosis to treatment initiation so there will be no events during this period.
6. Introduction line 69: TOPCAT outcomes?
7. Introduction line 73: due to confounding as a result of unmeasured differences?
8. Introduction line 76: mortality as a negative control outcome?

Methods
1. Methods Table 1 follow-up start (time 0): “Same as target trial” is very vague and is absolutely critical for definition of a target trial i.e. what is the time 0? The time 0 for the hypothetical trial is treatment assignment, but the comparison is a non-treatment group (who weren't assigned to a placebo) so there is no treatment assignment date for the comparator group. It’s already been stated that a clone-censor-weight approach is being applied to account for issues with time 0 that could lead to immortal time bias so this needs to be clearly defined here. The way this is described here, this study could not be reproduced. In figure S13 this is heart failure diagnosis, so I assume this is an updating error and should be corrected to be in line with figure S13.
2. Methods line 323-326: Here, it seems like the index date (time 0) is a heart failure diagnosis, but this isn't specified in Table 1 - it just says treatment assignment, even though the comparator group don't get any treatment – same point as above.
3. Methods 342: “To avoid immortal time bias” - please elaborate further on what is meant here i.e. that the period between HD diagnosis time 0 and treatment initiation in the exposed group will be immortal and confer a survival advantage to the treated group.
4. Methods 343: replace “index” with “heart failure diagnosis index date” or similar for clarity

Discussion
1. Discussion line 165: Please add a brief description in brackets for what is meant by a "landmark analysis" or a "person-time analysis"
2. Discussion line 175: “even when design-induced bias” – as per the comments in the introduction, please be more specific about the design-induced bias that clone-censor-weight is attempting to overcome (immortal time bias), There are other design-induced biases that could be introduced that clone-censor-weight would not account for.
3. Discussion line 188-199: it is great to see a clear argument based on these results for the importance of benchmarking and how it can add confidence in results beyond emulation of a hypothetical target trial, and also the importance of analysing a negative control outcome. Please replace the term “index trial” with “reference trial”. This is more inline with the concept of “benchmarking” (which is always against a reference value), and “index” suggests that the trial is the first study ever to address a specific question (which often it isn’t). This terminology is also consistent with work published in nature by the same research group (Krüger, N., Schneeweiss, S., Desai, R.J. et al. Cardiovascular outcomes of semaglutide and tirzepatide for patients with type 2 diabetes in clinical practice. Nat Med 32, 342–352 (2026). https://doi-org.ezproxy2.lib.gla.ac.uk/10.1038/s41591-025-04102-x).
4. Discussion line 243: Could the authors add text to the final sentence that specifically reiterates the utility of benchmarking against a reference trial and negative control outcomes for aiding interpretation of results of non-user comparisons in observational data.

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
Thank you for the considered responses to my comments, which I feel have been well addressed. The point on the potential equvialence of estimands in sequential trials and CCW is well taken, and spurred some interesting discussions in my team!

Reviewer #3

(Remarks to the Author)
All comments have been addressed and I am happy with the updated manuscript

Point-by-point response to reviewers' comments

We would like to thank all reviewers for their valuable comments, and for taking the time to review our submitted manuscript so thoroughly. We would like to stress that we found all reviews very well-argued and of high quality, and feel that they have helped us further improve this manuscript. All our replies to reviewer comments can be found in blue font below.

Please note that we adapted the sequence of sections (from Introduction-Results-Discussion-Methods to Introduction-Methods-Results-Discussion) without tracking these changes to keep the actual content changes induced by reviewer comments visible thereafter. Deletions are marked with strikethrough font and highlighted in grey. Insertions are highlighted in yellow.

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

OVERALL ASSESSMENT

This manuscript presents a methodologically rigorous and highly relevant empirical evaluation of the challenges inherent in emulating placebo-controlled trials using observational data. Recommendation: Minor Revision. The manuscript is of high quality and makes a valuable contribution to the literature. The recommended revisions are intended to enhance clarity, transparency, and the nuanced interpretation of findings.

Thank you for the positive evaluation and helpful comments.

1. Negative Control Outcome Assumptions

Comment: The choice of non-cardiovascular mortality as a negative control outcome is central to the bias correction strategy. While the authors acknowledge that this approach relies on the assumption of shared confounding with cardiovascular mortality, the discussion of this assumption's plausibility is somewhat brief. The assumption is strong and inherently untestable, and its violation could lead to over- or under-correction.

We have amended the following paragraph in the discussion section based on this comment and a similar comment raised by reviewer 2 (comment 3):

"Lastly, it should be noted that net bias evaluation and correction rely on the untestable assumption of a consistent pattern of residual confounding for both the target effect estimate and the negative control outcome. *In this study, this would imply that unmeasured confounders bias the estimated effects in the same direction to a similar extent for both cardiovascular events and non-cardiovascular mortality. While this assumption is likely to hold for disease-unspecific confounders such as socioeconomic*

factors or general health behaviors, it might not always hold for heart failure related unmeasured confounders such as symptoms or disease severity. If spironolactone was differentially prescribed to patients with higher disease severity, and these factors are highly prognostic of cardiovascular mortality, such confounding by indication may be incompletely captured with non-cardiovascular mortality as negative control outcome"

2. Interpretation of Benchmarking Results

Comment: The benchmarking exercise against TOPCAT is a major strength, but the interpretation of "benchmarking failure" for mortality outcomes requires more nuance. The authors note population differences (age, comorbidity) and shorter follow-up in their study, but these factors are presented primarily as limitations rather than as alternative explanations for the observed discrepancies.

We amended the paragraph as follows:

"The benchmarking failure that we observed for the *standalone* mortality outcomes may therefore also be explained by *different follow-up times or characteristics of the underlying populations rather than a genuine failure to control residual bias*[27]*."*

3. Weight Model Misspecification

Comment: The supplementary materials (Tables S9-S12) provide valuable diagnostics showing that the pseudo-population sizes do not perfectly align with observed populations, suggesting some degree of weight model misspecification. While the authors briefly note this, they do not discuss the potential direction or magnitude of bias this might introduce.

We agree with the reviewer that an assessment of the potential direction or magnitude of bias introduced through the observed weight model misspecifications would be highly informative. However, this assessment is unfortunately not trivial: In our perspective, the most likely reason for the observed weight model misspecification is unavailability of information on additional covariates driving censoring and treatment adherence. Therefore, we unfortunately cannot confidently comment on the direction or magnitude of bias.

To help raise awareness for the possibility to conduct the weight model diagnostics, we added the following paragraph to the methods section:

"*As a diagnostic for potential weight model misspecification, we assessed the sample sizes of the weighted pseudo-population in the first twelve months of follow-up. Correctly specified weight models should result in a pseudo-population size that is (approximately) equal to the original study population before censoring when derived from unstabilized weights, and (approximately) equal to the original study population after censoring when derived from stabilized weights*[52]."

Additional remark: Due to adjustments in the sequence and layout of the Supplementary Materials, the information in question (former Supplementary Tables S9-S12) is now found in Supplementary Tables S16-S19.

4. Definition of "Net Bias"

Comment: The term "net bias" is used throughout the manuscript but is not explicitly defined when first introduced in the Methods section.

We have added the following definition to the methods sub-section on unmeasured confounders:

“Even after adjusting for observed confounders, unmeasured confounders can bias estimates. If different confounders distort estimated effects in opposing directions, they may partially cancel each other out. Net bias is the remaining deviation from the true causal effect under the combined influence of all unobserved confounders.”

5. Language and Readability

Comment: Overall, the manuscript is well-written. However, a few sentences are lengthy and could be simplified for clarity.

We tried to further simplify language in those instances where sentences seemed too convoluted.

Reviewer #2 (Remarks to the Author):

I really enjoyed reading this thoughtful and well-conducted study, and the findings are a timely reminder about the difficulties we have comparing users to non-users of medicines in pharmacoepidemiology. My comments are minor – overall I would recommend this for publication and congratulate the authors for a nice contribution to the literature.

Thank you for the positive feedback and valuable suggestions.

Main comments:

1. You use CCW for the emulation which means that your treatment strategies include a grace period for starting treatment. However, the benchmarking trial appears to not have a grace period and instead follows people from randomization, which coincided with placebo / treatment initiation. Do you think part of the differences in findings might be due to the RCT and the emulation addressing different questions? E.g. “What is the risk of death after starting spironolactone” in the RCT compared to “What is the 12-month risk of death after a HF hospitalisation if I start spironolactone within 6 months” in the emulation?

The reviewer is right to point out that in any target trial emulation necessary pragmatic adaptations of the study design (for example with regard to inclusion criteria for the study population) may inadvertently change the research question and estimand. We also agree that the inclusion criterion regarding a prior HF hospitalization before start of follow-up is one central factor in terms of comparability of study populations. The TOPCAT trial allowed inclusion of participants based on either Brain natriuretic peptide (BNP) levels (≥ 100 pg/ml or N-terminal pro-BNP ≥ 360 pg/ml in the last 60 days) or based on a heart failure hospitalization in the 12 months before start of follow-up. As we also point out in the discussion section, we were unable to emulate the BNP randomization stratum with Medicare data:

“First, only TOPCAT results for the heart failure hospitalization outcome were available separately by randomization stratum in the Americas subgroup. For all other outcomes, published TOPCAT Americas results may be partially driven by the BNP randomization stratum, which we could not emulate.”

However, to align our emulation as closely as possible to the TOPCAT trial with respect to HF hospitalization inclusion criteria, we required a prior heart failure hospitalization in the six months before index as inclusion criterion, while allowing both inpatient and outpatient HFpEF index diagnoses to define the start of follow-up in our observational study. We opted for a maximum of 6 instead of 12 months pre-baseline time span to last hospitalization to account for the additional 6-months grace period for treatment initiation after baseline. This will effectively ensure that all spironolactone initiators

included in our emulation have a HF hospitalization within 12 months prior to treatment initiation in our emulation, similar to the trial.

We also agree with the reviewer with regard to the different follow-up periods as a potential contributor to the observed differences in results, and added the following sentence (also following comment 2 by reviewer 1, which pointed in a similar direction).

"The benchmarking failure that we observed for the standalone mortality outcomes may therefore also be explained by different follow-up times or characteristics of the underlying populations rather than a genuine failure to control residual bias [29]"

2. Related to the above – throughout the paper CCW is positioned as a method that is particularly suited for use vs. non-use comparisons because of the difficulties associated with assigning time zero for the non-user group (for example L41-45 in introduction, and L172 - 174 in the discussion). However, often the challenge with non-users isn't indistinguishable time zero but multiple eligibility points which would lead us to consider a sequential trial design. It seems to me that in this study the indistinguishable time zero is not a result of the non-user comparison but the application of a grace period, which also changes the research question. I think the paper would benefit from clarifying the difference between indistinguishable time zero and multiple eligibility points, and from including a more explicit comparison of CCW and sequential trials (particularly in terms of the questions answered by each design).

Thank you for the opportunity to clarify this important aspect. We have now revised the introduction to explicitly note that both a CCW and a sequential trial approach are potential options for emulating placebo-controlled trials.

Note that we have provided a detailed tutorial on this very topic in our recent publication (Fu et al. BMJ 2026;392:e084909, https://doi.org/10.1136/bmj-2025-084909)[1]. In summary, we argue that one central requirement in the target trial framework is that, like in a randomized controlled trial, assessment of eligibility criteria, treatment assignment and start of follow-up should be anchored at a clinically relevant index event in patient history ("time zero"). However, in a real-world outpatient setting, filling a prescription and (presumably) initiating a new pharmaceutical treatment usually follow this clinically relevant index event with some time lag.

The investigator of an observational study could consider the following options: 1) As in our study, try to approximate the actual clinically relevant index event in patient history as well as possible, for example by using a physician contact (as assumed source of the prescription, and presumably closer in time to the clinically relevant event than the medication fill) as anchor, and add a grace period for treatment initiation, thereby accepting that assigned treatment strategies may be indistinguishable at index, with the advantage of avoiding introduction of immortal time in both groups; 2) start follow-up at prescription, and require occurrence of the clinically relevant event in the baseline period as an eligibility criterion (and potentially control for time from clinically relevant event to treatment initiation as a regression/matching variable). In

active-comparator new-user designs, this is often the straightforward decision. In a non-user comparison, this could be implemented through risk-set sampling into a separate trial at each time a treatment is initiated.

Option 1) reflects the clone-censor-weight approach, option 2) a sequential trial emulation approach.

Both the sequential trial, and the clone-censor-weight approach aim to estimate the effect of treatment *initiated within a pre-defined time window following a clinically relevant event, among individuals who experienced this event.*

In the sequential trial approach, the *clinically relevant event serves as an eligibility criterion*, with follow-up anchored at *treatment initiation for the exposed group or a* corresponding pseudo-date for the comparator group.

In contrast, the clone-censor-weight approach anchors the start of follow-up at the clinically relevant index event and implements the time-window for treatment initiation as a subsequent grace period.

In our perspective, the difference between the two designs is therefore primarily operational in how the eligibility criteria are defined and assessed. In both design options, the estimand is an average effect over the distribution of treatment initiation times observed within the grace period, or relative to the eligibility criterion, and is thus inherently dependent on that sample-specific distribution.

We added the following paragraph to the introduction to briefly summarize this:

"In the new-user active comparator design, a time zero that is identifiable in both groups is initiation of treatment (the treatment of interest in the exposed group, the comparator treatment in the control group). For target trial emulations of placebo-controlled trials, sequential trials and clone-censor-weighting represent viable operational strategies in the absence of clear time zero[5]. In sequential trials, a clinically relevant event may serve as an eligibility criterion assessed in some prespecified baseline period, while the follow-up is anchored at treatment initiation for exposed individuals and a corresponding timepoint for risk set sampled comparators. Cloning, censoring and weighting anchors follow-up for all treatment strategies at a clinically relevant event that is assumed to precede the treatment of interest (e.g. a healthcare contact or diagnosis), and introduces a subsequent grace period for treatment initiation, accepting that treatment strategies may be indistinguishable at time zero[6]"

Minor suggestions / considerations

3. Could you comment on the plausibility of the assumption of 'consistent pattern of residual confounding' for the NCO? Was there a specific confounder that was of concern you anticipated being captured, and how accurately do you think the NCO did

this? Presumably something like severity of underlying CV disease would increase only CV specific mortality.

We have amended the following paragraph in the discussion section based on this comment and a similar comment raised by reviewer 1 (comment 1):

"Lastly, it should be noted that net bias evaluation and correction rely on the untestable assumption of a consistent pattern of residual confounding for both the target effect estimate and the negative control outcome. *In this study, this would imply that unmeasured confounders bias the estimated effects in the same direction to a similar extent for both cardiovascular events and non-cardiovascular mortality. While this assumption is likely to hold for disease-unspecific confounders such as socioeconomic factors or general health behaviors, it might not always hold for heart failure related unmeasured confounders such as symptoms or disease severity. If spironolactone was differentially prescribed to patients with higher disease severity, and these factors are highly prognostic of cardiovascular events, such confounding by indication may be incompletely captured with non-cardiovascular mortality as negative control outcome*"

4. I wasn't sure I followed the rationale for adjusting for time-updated HCRU - like number of HF hospitalisations - in an analysis where HF hospitalisation is the outcome: how did you distinguish these two? Even for other endpoints, e.g. death, I wasn't sure whether preceding hospitalisation would act as a confounder - would spironolactone be discontinued on hospitalisation?

We apologize for the ambiguous wording in the original manuscript text. HF hospitalizations were only used as a time-varying variable in analyses of endpoints that did not comprise HF hospitalizations. The rationale for using time-updated HCRU variables in general was that each additional healthcare contact provides additional opportunity for treatment adaptations (in both directions, discontinuation and initiation). Also, frequency of healthcare contacts may affect risk of adverse outcomes (for example reduce it through closer patient monitoring). If off-label prescription of spironolactone also leads to increased subsequent healthcare contacts (for example to monitor potential adverse reactions, and this monitoring has additional beneficial effects on health outcomes), there is also the potential for treatment-confounder feedback, which we wanted to control. Generally, it might be worth adding that adjustment for time-varying covariates through IP weighting did not seem to strongly affect point estimates for outcomes, as illustrated by Supplementary Figure S14.

We adapted the respective wording as follows to improve clarity:

"[…] *and* number of heart failure hospitalizations, *where these were not part of the outcome definition*) […]"

5. Could you expand the discussion of intercurrent events: what were the intercurrent events you considered (e.g. L345 in the methods – you mention that some intercurrent

events did not result in censoring?). How did you handle death due to other causes in the analyses of cause-specific mortality?

We had originally listed the intercurrent events we considered in the methods subsection that described the hypothetical target trial:

“Both treatment strategies would allow for deviation from the assigned strategy at physician discretion after intercurrent events that pose either clinical contraindications to spironolactone use (a hospitalization for hyperkalemia, ESRD diagnosis, prescription of lithium) or make it potentially redundant (LVAD or heart transplant).”

However, we agree that it might be helpful to repeat the relevant intercurrent events in the section on treatment strategy emulation via the clone-censor-weight approach:

“After intercurrent events *(hyperkalemia hospitalization, ESRD diagnosis, lithium prescription, LVAD or heart transplant),* clones were not censored even if their observed treatment trajectories deviated from their assigned strategy, *as forcing them to stay on treatment would represent an unrealistic and potentially harmful hypothetical strategy”*

We also corrected a related small error in former Table 1 (now Supplementary Table S1), where ESRD had erroneously been listed as condition removing the HFpEF indication instead of a contraindication to spironolactone use.

In analyses of cause-specific mortality (or, more generally, in all analyses with competing risk), we opted for a hypothetical estimand and used inverse probability weighting to represent the risk of outcome in a hypothetical world where the competing event was absent. Therefore, in analyses of HF hospitalizations, all deaths were “eliminated” from the dataset through inverse probability of censoring weighting. In analyses of cardiovascular mortality and the composite outcome that comprised cardiovascular mortality, competing non-cardiovascular deaths were dealt with through IPCW. For non-cardiovascular mortality, competing cardiovascular deaths were dealt with through IPCW. For all-cause mortality and the composite outcome that comprised all-cause mortality, no competing risks existed, so here we did not have to use a hypothetical scenario to eliminate competing risk.

We acknowledge that this decision, though hinted at in the methods sub-section on IP weighting, could be made more explicit, and its implications for interpretation should be discussed. We originally provided the following information:

“Interval-specific conditional probabilities for the two types of censoring mechanisms (treatment deviation or administrative censoring due to loss to follow-up and death, where death was not part of the outcome definition) would be estimated using pooled logistic regression models […]”.

We now added the following paragraph:

“For competing events to each outcome of interest, this is conceptually equivalent to “eliminating” the competing event from the dataset through inverse probability of

censoring weighting, and produces to an estimand for a hypothetical scenario where competing events do not exist[13,14].”

6. How is date of death captured in Medicare after Dec 31 2016 without the linkage, and is this complete death information or restricted to e.g. in-hospital deaths?

All deaths (inpatient + outpatient) are captured administratively by CMS for all Medicare enrollees. We have reported this information to be highly complete (>99%) in a prior validation study (Desai J Am Heart Assoc 2020 Sep;9(17):e016906, doi: https://doi.org/10.1161/JAHA.120.016906)[2], also cited in the respective paragraph as reference no. 17 against the National Death Index NDI. We added the following sentence:

“Complete information on date of all-place all-cause mortality without the cause was available for the entire Medicare sample also beyond 2016.”

7. Could you add more information how cause of death was defined: was this any contributing / underlying cause, or only the primary cause of death?

We amended the following sentence:

“Information on *primary* cause of death was available through linkage with the National Death Index *(NDI)* until 12/2016[17]”

8. Could you expand on how you applied the 1-month time lag? Did you set events occurring within the first month to missing, start FU later on, or censor people who had events in this period? For safety endpoints, a 1-month time lag seems potentially too long.

As mentioned in the methods section, “Individuals with outcome events in the first follow-up interval were excluded from the respective outcome analysis, resulting in slightly different analysis datasets for each outcome.” These individuals did not contribute to the analysis as no follow-up was started thereafter.

We added a paragraph reflecting on the potential implications for safety endpoints to the discussion section:

“The fact that we introduced a 1-month time lag between treatment and outcome measurements may help explain why the risk for hyperkalemia hospitalizations (our secondary safety endpoint) was higher in TOPCAT compared to our study, as it led to the exclusion of events occurring very early after initiation. This time gap, though a common choice in target trial emulation, and a necessary trade-off with adequate time windows for measuring effectiveness outcomes, may thus have reduced our ability to capture early safety outcomes.”

9. The naïve analysis is an interesting demonstration, but the analysis strategy I see most often for these kinds of questions is to start FU at treatment for the treated and at diagnosis for the untreated (often still defined as never users) – effectively just discarding the time between diagnosis and treatment in the treated. There might not be scope for you to explore this, but I would be interested in seeing how biased this common (but still flawed) approach is.

What the reviewer refers to is investigator-induced bias through exclusion as compared to misclassification of immortal time. We agree that in principle the implications of both types of naïve analyses may be of interest, but eventually decided to only present one naïve analysis for the same reason also mentioned by the reviewer: We felt that presenting multiple types of flawed analyses would further shift the scope of the paper. Between exclusion and misclassification of immortal time, we eventually opted for the misclassification of immortal time because an early narrative review article on this topic listed more studies that misclassified immortal time than studies that excluded immortal time[3], and also for didactical reasons (we expected the bias due to misclassification to be larger than excluded immortal time[4]).

However, we agree that it might be useful to remind readers to stay alert for both types of immortal time bias, and therefore added a respective sentence to the introduction section:

"Starting follow-up either at diagnosis for untreated patients, and at subsequent treatment initiation for treated patients, or for both groups at diagnosis while deciding group assignment based on future information on treatment initiation, can introduce immortal time bias (through exclusion or misclassification of immortal time, respectively)."

10. I would consider summarizing the features of the TOPCAT trial in your TTE table (as a separate column). This would make also make the difference in the treatment strategies assessed clear.

The respective column was added to the TTE table. Due to the relatively large table size and to follow journal formatting requirements, the table was subsequently moved to the Supplementary Materials as Supplementary Table S1.

11. P7 L170 – the constant hazard ratio assumption would presumably also apply to a CCW analysis applying a cox model?

Thank you for this valid objection. We removed this argument and adapted the sentence to:

"A person-time analysis *does not identify a per-protocol effect of continued treatment and may be biased in the presence of treatment-confounder-feedback* ~~*when risk of the*~~

~~*outcomes varies over time and constant hazard ratios cannot be assumed during the follow-up time*~~*"*

Editorial

1. I couldn't access the SAS code on OSF via the link in the paper (I think this is still behind a 'permissions' wall).

Thank you for bringing this to our attention, and we apologize for the inconvenience. The SAS codes on OSF should now be publicly accessible.

Reviewer #3 (Remarks to the Author):

This study looked at the effectiveness of spironolactone in heart failure with preserved ejection fraction, aiming to assess effectiveness in a broader population than the TOPCAT trial. While doing this, it aimed to assess how applying the clone-censor-weight approach within the target trial emulation framework compared to the TOPCAT Americas result and a negative control outcome, and further compare results with a number of approaches outside the target trial emulation framework. Approaches outside the TTE framework were found to be substantially biased, while even with clone-censor-weighting, residual confounding remained.

This is a novel and potentially important study for highlighting the utility and limitations of the clone-censor-weight approach when applied to the emulation of placebo-controlled trials.

Thank you for the overall positive assessment, and the suggested modifications.

There are a number of areas for clarification, as detailed below:

Introduction

1. Introduction line 43-47: please add an example here of what is meant by a “time zero” that is identifiable in both treatment arms

We added the following examples:

“The target trial emulation framework was proposed to ensure that investigators define a clinically relevant event in patient history (e.g. a healthcare contact, diagnosis, or change in treatment regimen) as common start of follow-up for both groups to avoid immortal time bias”

“In the new-user active comparator design, a time zero that is identifiable in both groups is initiation of treatment (the treatment of interest in the exposed group, the comparator treatment in the control group). […] In sequential trials, […] follow-up [is] anchored at treatment initiation for exposed individuals or a corresponding pseudo date for risk set sampled comparators. Target trial emulation with cloning, censoring and weighting *anchors follow-up for all treatment strategies at a clinically relevant event that is assumed to precede the treatment of interest (e.g. a healthcare contact or diagnosis), and introduces a subsequent grace period for treatment initiation […]”*

2. Introduction line 49: In line 43, the remit/benefit of cloning, censoring and weighting is clearly defined i.e. the cloning, censoring and weighting can be applied when treatment strategies are indistinguishable at time zero. Here, a much broader

statement is made about the clone-censor-weight approach which makes is sound like this approach is required in order to study a well-defined causal question that could be addressed by a corresponding RCT, when in fact there are a variety of approaches within the TTE framework that could allow investigators to "study a well-defined causal question that could be addressed by a corresponding RCT". Please either rephrase or remove this statement.

We moved the sentence further towards the beginning of the introduction and rephrased it as follows:

"The target trial emulation framework was proposed to ensure that investigators define a clinically relevant event in patient history as common start of follow-up for both groups to avoid immortal time bias, while studying a well-defined causal question in observational data that could be addressed by a corresponding RCT."

3. Introduction line 55: "investigator-induced design flaws" – does this mean the immortal bias mentioned previously? Please be specific.

We adapted the wording to:

"[…] non-user comparisons may still be especially prone to confounding even when *immortal time bias is* avoided."

4. Introduction line 56: please add the year of the trial

We adapted the wording to:

"[…] we built on evidence generated by the *2006-2013* "Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist" (TOPCAT) trial […]"

5. Introduction line 68: This is a very good opportunity to explain exactly why it is important to use clone-censor-weight for this question which would really help clarify the introduction i.e.: (1) there is a no-treatment group so the date of initiation of treatment cannot be used as a start date for this group (2) a start date of latest eligibility for trial is therefore used, but if this is used as time 0, there is immortal time bias as the treatment group by definition need to stay alive from HF diagnosis to treatment initiation so there will be no events during this period.

We slightly restructured this part of the introduction to accommodate these suggestions, and added the following paragraphs:

"As there was no definitive standard of care treatment strategy for HFpEF in the study period, we followed the TOPCAT placebo comparison rationale which suggested that the most clinically relevant comparator in this population would be non-use. *Because the non-user group had no date of treatment initiation that could be used as a start date for follow-up, eligibility was anchored at a preceding healthcare encounter with a*

recorded HFpEF diagnosis. […] This bias stems from the fact that the treatment group will stay alive from HFpEF diagnosis to treatment initiation by definition. To avoid this bias, we emulated a target trial using cloning, censoring and weighting […]"

6. Introduction line 69: TOPCAT outcomes?

We adapted the wording to:

"[…] to compare initiation and sustained spironolactone use against non-use *with regard to outcomes previously evaluated in TOPCAT.*"

7. Introduction line 73: due to confounding as a result of unmeasured differences?

We adapted the wording to:

"Anticipating potential threats to validity due to *confounding through* unmeasured differences between active treatment and non-use groups, […]"

8. Introduction line 76: mortality as a negative control outcome?

We are not entirely sure we correctly interpreted this comment/suggestion. We really only used deaths where cause of death was not classified as cardiovascular (so the non-cardiovascular mortality) as negative control outcome, not all-cause mortality.

Methods

1. Methods Table 1 follow-up start (time 0): "Same as target trial" is very vague and is absolutely critical for definition of a target trial i.e. what is the time 0? The time 0 for the hypothetical trial is treatment assignment, but the comparison is a non-treatment group (who weren't assigned to a placebo) so there is no treatment assignment date for the comparator group. It's already been stated that a clone-censor-weight approach is being applied to account for issues with time 0 that could lead to immortal time bias so this needs to be clearly defined here. The way this is described here, this study could not be reproduced. In figure S13 this is heart failure diagnosis, so I assume this is an updating error and should be corrected to be in line with figure S13.

Thanks for catching this. We updated the wording in former Table 1 (now Supplementary Table S1 due to journal requirements) to "*At first qualifying heart failure diagnosis".*

2. Methods line 323-326: Here, it seems like the index date (time 0) is a heart failure diagnosis, but this isn't specified in Table 1 - it just says treatment assignment, even though the comparator group don't get any treatment – same point as above.

See above: Former Supplementary Figure S13 (now Supplementary Figure S1) and the Methods description are correct. We assume this comment is resolved through the correction made to Table 1 in response to the previous comment.

3. Methods 342: “To avoid immortal time bias” - please elaborate further on what is meant here i.e. that the period between HD diagnosis time 0 and treatment initiation in the exposed group will be immortal and confer a survival advantage to the treated group.

We added the following sentence:

“Assignment based on future treatment initiation is not an acceptable solution, as it creates an investigator-induced immortal time period between the index HF diagnosis and treatment initiation in the exposed group, which can bias results.”

4. Methods 343: replace “index” with “heart failure diagnosis index date” or similar for clarity

To improve clarity, we amended the respective sentence to:

“[…] we created 2 copies (clones) of each patient’s data, of which one was assigned to each possible strategy at the *index HF diagnosis* […]”

For consistency reasons, we also adapted several other instances in the manuscript to this wording.

Discussion

1. Discussion line 165: Please add a brief description in brackets for what is meant by a "landmark analysis" or a "person-time analysis"

We added the following brief explanations:

“Approaches such as a landmark analysis *(where follow-up for both groups is started at the end of a grace period, the “landmark”, following a clinically relevant index event, and assignment is decided based on treatment status at the landmark),* or a person-time analysis *(where follow-up starts at the clinically relevant index, but time-varying exposures allow for attributing unexposed person-time to control and exposed person-time to treatment conditions)* […]”

2. Discussion line 175: “even when design-induced bias” – as per the comments in the introduction, please be more specific about the design-induced bias that clone-censor-weight is attempting to overcome (immortal time bias), There are other design-induced biases that could be introduced that clone-censor-weight would not account for.

We adapted the wording to:

"Our investigation provides a reminder that even when design-induced *immortal time* bias *in the presence of a grace period* is avoided through the clone-censor-weight approach"

3. Discussion line 188-199: it is great to see a clear argument based on these results for the importance of benchmarking and how it can add confidence in results beyond emulation of a hypothetical target trial, and also the importance of analysing a negative control outcome. Please replace the term "index trial" with "reference trial". This is more inline with the concept of "benchmarking" (which is always against a reference value), and "index" suggests that the trial is the first study ever to address a specific question (which often it isn't). This terminology is also consistent with work published in nature by the same research group (Krüger, N., Schneeweiss, S., Desai, R.J. et al. Cardiovascular outcomes of semaglutide and tirzepatide for patients with type 2 diabetes in clinical practice. Nat Med 32, 342–352 (2026). https://doi-org.ezproxy2.lib.gla.ac.uk/10.1038/s41591-025-04102-x).

We adapted the wording in this, and, for consistency reasons, in two more instances.

4. Discussion line 243: Could the authors add text to the final sentence that specifically reiterates the utility of benchmarking against a reference trial and negative control outcomes for aiding interpretation of results of non-user comparisons in observational data.

We restructured the conclusion section to:

"In conclusion, our study demonstrated practical challenges when emulating *a placebo-controlled target trial for spironolactone in patients with HFpEF using observational data. While the principled clone-censor-weight approach avoided immortal time bias, our pre-specified net bias analyses that comprised of benchmarking against a reference trial (TOPCAT) and a negative control outcome (non-cardiovascular death), suggested potential residual bias. Our results thus point to the importance of incorporating guardrails to detect such bias, and the* heighted scrutiny that is needed when interpreting results from emulations of placebo-controlled target trials."

Further changes induced by editorial requirements:

- We adapted the title, first based on the journal's suggestion, and subsequently slightly refined it based on our own judgement within the journal requirements for title length and structure
- We amended the originally unstructured abstract of 150 words to a structured abstract within the required scope of 250 words. This allowed us to add some more specific information on results and further reduce the number of abbreviations used. We also changed the wording in the abstract's results and conclusions to present tense as per journal requirements.
- We added a 120-word plain language summary after the abstract as per journal requirements.
- As per journal requirements, we added a paragraph to the end of the introduction section that summarizes the major results and conclusions in the present tense.
- We changed the section header "Statistical analysis" to "Statistics and Reproducibility" as per journal requirements.
- We moved the summary paragraph from the beginning of the Discussion section to the end of the Discussion section as per journal requirements.
- We moved the definition of immortal time bias from the discussion to the introduction section.
- Figures were removed from the main manuscript document and are now provided as single jpg files.
- We changed the wording from "supplemental" to "supplementary" materials as per journal recommendations. Supplementary Texts are now labelled "Supplementary Notes"
- Due to the 1-page size limit for tables in the main manuscript, the target trial table (former Table 1) was moved to the supplementary materials (now supplementary Table S1).
- Bold formatting of table headings and underlined words were removed as per journal instructions.
- As tables in the main manuscript are not allowed to cover more than one page, we removed all descriptives from Table 2 for which no TOPCAT equivalent information was available. All removed information can however still be found in the more comprehensive Supplementary Tables S5-S8.
- As the size of former Table 1 covered more than one page, we moved it to Supplementary Materials
- Aggregate source data for main text figures were added as Supplementary Data in excel format as per journal requirements
- Former Supplementary Table S1 now split into smaller Supplementary Tables S3-S4 to comply with journal requirements of tables not exceeding 1 page length

- Former Supplementary Table S2 now split into smaller Supplementary Tables S5-S8 to comply with journal requirements of tables not exceeding 1 page length
- Former Supplementary Table S5, S7, S8 (now S11, S14, S15) reformatted to fit on one page, respectively
- Former Supplementary Table S6 now split into smaller Supplementary Tables S12-S13 to comply with journal requirements of tables not exceeding 1 page length
- Former Supplementary Table S17 now split into smaller Supplementary Tables S24-28 to comply with journal requirements of tables not exceeding 1 page length

References

1. Fu, E.L., Harhay, M.O., Schneeweiss, S., Desai, R.J. & Hernán, M.A. Starting right: aligning eligibility and treatment assignment at time zero when emulating a target trial. *bmj* **392**(2026).
2. Desai, R.J., Levin, R., Lin, K.J. & Patorno, E. Bias Implications of Outcome Misclassification in Observational Studies Evaluating Association Between Treatments and All-Cause or Cardiovascular Mortality Using Administrative Claims. *Journal of the American Heart Association* **9**, e016906 (2020).
3. Suissa, S. Immortal time bias in observational studies of drug effects. *Pharmacoepidemiology and drug safety* **16**, 241–249 (2007).
4. Duchesneau, E.D.*, et al.* The timing, the treatment, the question: comparison of epidemiologic approaches to minimize immortal time bias in real-world data using a surgical oncology example. *Cancer Epidemiology, Biomarkers & Prevention* **31**, 2079–2086 (2022).