**Reviewer 1:** The paper intends to describe a standardized framework for risk based assessment of treatment of treatment effect heterogeneity in observational databases. The paper describes a general workflow of several steps that should standardize the way to do such experiments (outcomes prediction) on the OMOP standardized model, as part of the OHDSI project. Eventually the framework is demonstrated on a use case of the effect of ACE inhibitors versus beta blockers on a set of 9 outcomes.  
  
Generally the paper is not well organized, while its in general easy to read. However, the structuring and descriptions are not detailed enough, and in general it is not clear what are the contributions of the paper. Perhaps having better structure would have made it easier to understand, or simply stating the contributions clearly. Here are some general comments, followed by more specific comments.

The paper describes a framework for the process of outcomes prediction, given a population that is exposed to two types of treatments, and implemented a risk assessment model. However, what is new here? In general, the flow seems quite proper, but then what is so special here, and brings anything new to the readers? initially it seemed that it is the use of OMOP, which intends to enable easily performing the same experiment on several EHR databases that are in different institutes, and perhaps even different countries, but the demonstration is not on different databases.

Moreover, the entire paper doesn’t refer explicitly to the use of OMOP.

Thank you for pointing that out. We now also explicitly mention the OMOP-CDM requirement in the following places:

* In the last paragraph of the introduction we write: *“Our publicly available package provides an out-of-the-box solution for implementing such analyses at scale within the OHDSI network, taking advantage of the OMOP-CDM”*
* In the first paragraph of Section 2 we write: *“The proposed framework defines 5 distinct steps that enable a standardized approach for risk-based assessment of treatment effect heterogeneity for databases mapped to the OMOP-CDM”*
* In Section 2.2.we added: *“It is required that the databases are mapped to the OMOP-CDM”*

The implemented code which is public on OHDSI's website, which is perhaps the contribution of the paper, is mentioned only at the end in the discussion, but it is not referred to explicitly along the paper.

We agree that our R-package should be mentioned a lot earlier, as it is one of the main contributions of our work. We added the following phrase to the first paragraph of Section 2:

*“We developed an open-source R-package for the implementation of the proposed framework and made it publicly available. The source code can be found at https://github.com/OHDSI/RiskStratifiedEstimation”*  
  
Here are some more specific comments.  
Introduction  
As mentioned earlier, the contributions of the paper are not clear. Please list the contributions of the paper at the end of the introduction section explicitly.  
Since JBI is a methodological journal, it will be good to focus on the contributions in that perspective.  
It will be good to write the introduction as a motivation for the listed contributions.

Indeed, the contributions of our paper were not very clear from the introduction. We have revised the final paragraph of the introduction to:

*“We aimed to develop a framework for implementing risk-based assessment of treatment heterogeneity in high-dimensional observational data, extending the existing guidelines of the RCT setting. Our publicly available package provides an out-of-the-box solution for implementing such analyses at scale within the OHDSI network, taking advantage of the OMOP-CDM”.*  
  
Background  
- there is no background section in this paper. Please write and cover the relevant literature, so it is clear where your contribution comes into. Will be good to refer to papers that deal with this problem, as well as the statistics behind it, as well as about OHDSI and OMOP.

We agree that the background provided in the introduction may have been quite brief. Literature on heterogeneity of treatment effect analyses has been expanding quite rapidly in the past decade. We decided to provide a high-level overview of the different categorizations that can be considered in treatment effect heterogeneity analyses by modifying the second paragraph of the introduction in the following way:

*“In recent years, a large number of methods has been developed for the assessment of HTE, mainly in the RCT setting. Earlier work suggested separating HTE analyses into exploratory, confirmatory, descriptive and predictive. Exploratory analyses focus on hypothesis generation, confirmatory analyses test subgroup effect hypotheses, descriptive analyses aim at facilitating future synthesis of subgroup effects and predictive analyses predict probabilities of benefit or harm in individual patients. Predictive HTE approaches can be further subdivided into risk modeling, treatment effect modeling and optimal treatment regime methods, based on the reference class used for defining patient similarity when making individualized predictions or recommendations. We focus on “risk modeling” approaches where patients are divided into risk strata using either an existing or an internally developed risk prediction model. Risk-stratum-specific estimates provide an overview of the evolution of treatment effects with increasing risk both on the relative and the absolute scale. Recently, systematic guidance on the application of such methods has been developed”.*  
  
Materials and Methods  
This section should explain and define the entire methodology. When figures or tables are presented, refer to them in details in the text. Currently for example figure 1 is described only in its caption, but not in the text, and it is not detailed enough. figure 1 is too schematic and it is hard to understand at this stage what it describes.  
what is the purpose of figure 1? to describe in high level the framework? make it clear please.

The general flow that is described in the section seems like any other data science project that is published in the literature - what is new here? make it clear and focus on that with details.

Thank you for pointing out this issue. We agree that the novelty of our framework may have been somewhat obscured in this section. There are two main contributions our framework is attempting. First, extend the existing framework described in the Predictive Approaches to Treatment Effect Heterogeneity (PATH) statement to the observational setting. Of course, the requirements for this endeavor should not be taken lightly and our contribution provides more of an initial step towards that. Second, develop a tool for implementing this standardized framework that is compatible with the OMOP-CDM in order to carry out such analyses at a very large scale in a uniform and reproducible way.

Currently this section is written in high level. Define properly the cohorts and how they are constructed, as well as the matching process - and always highlight how your framework is better than any alternative (which has to be introduced already in the introduction, and referred to in the background as well).

We agree that cohort definitions are presented at a high level. Especially in the case of cohort definitions, within OHDSI the term is used in a slightly different manner than in many manuscripts in the field. Whereas “cohort definition” usually describes the definition of a set of clinical codes (e.g. ICD-9/ICD-10, NDC, etc), the term is here used to describe more of a logic for implementing this code set. Code sets are used to define both inclusion that are applied on the patients to determine who fulfills them and for how long and cohort exit criteria that determine when a patient no longer qualifies for cohort membership. We have added the following text to the first paragraph of Section 2.1:

*“Cohort definitions are crucial for this step of the framework. We define a cohort as the set of patients who satisfy one or more inclusion criteria for a duration of time. A cohort within the OHDSI setting is more than a set of specific clinical codes, providing a definition of a logic for how to use that code set. All cohort definitions consist of: an entry event, i.e. the time a patient enters a cohort; a set of inclusion criteria applied to the initial event cohort to further restrict the set of people, resulting in the construction of the construction of the qualifying cohort; cohort exit criteria that terminate the patient's presence in the cohort. Cohort definitions are transportable, meaning they can be implemented in any database, provided that it is mapped to the OMOP-CDM”.*

We also agree with Reviewer 1 that the matching process is very briefly described in Section 2. However, there is multitude of ways both for estimating propensity scores and adjusting for them that can be considered. This is an active field of research with no “one-size-fits-all” solutions. For example, we opted for using a large-scale propensity score estimation process, where a large set of covariates is considered in a LASSO logistic regression for performing variable selection—currently the default within OHDSI (REF Tian). However, other approaches have also been suggested in the literature from using expert opinion (REF) to ranking covariates for their “confounding plausibility”(REF). The modeling approach is also another issue with many answers. We use LASSO logistic regression but any other machine learning approaches can also be considered (REF Austin). These issues are out of our manuscripts scope. So we decided to only present what is currently the default option in our package, which also is the default in OHDSI. We have modified the second paragraph of Section 3.2. and made it the last paragraph of Section 2.2:

*“More specifically, we first estimate propensity scores using LASSO logistic regression and a large set of baseline covariates including demographics, drug exposures, diagnoses, measurements and medical devices. We match patients 1-1 using a caliper, i.e. the maximum distance that is acceptable for any match. The default value we use is 0.2 on the standardized logit scale for the propensity scores. Other methods of fitting the propensity scores, such as random forest and others (REF) can also be considered”.*

Additionally, it will be good to make the model more structured. It is currently designed that the outcome are in a period of time, relative to the treatment (in the use case it is two years) but when the outcome appears in more than two years? is there this flexibility?  
Meanwhile, it seems that the framework uses case-control study design. It will be good to make it clear that this is indeed the design to make things clear.  
  
Evaluation  
it is expected to see an Evaluation section that is before the Results. This section should state the research questions, and the corresponding experimental plan, as well as the data and evaluation metrics.

The software package developed for implementing our framework is designed to run all analyses at once—both calculation of evaluation metrics and estimation of treatment effects. The researcher assessing treatment effect heterogeneity has the responsibility of following proper procedure for validating their results. The software ensures that a large part of widely accepted evaluation approaches are available and cannot in any way replace expert knowledge. That said, we totally agree with Reviewer 1 that evaluation should precede the results. In the Evaluation section we described the required process for evaluating both the developed prediction model and the effect estimation process within the derived risk strata. We decided to remove the Evaluation section and move its two components to the relevant remaining sections (Prediction step and Estimation step). Therefore in section 2.3, regarding the prediction step we added:

*“After model development, a performance overview of the derived prediction models including discrimination and calibration both in the propensity score matched subset, the entire population and separately for treated and comparator patients should also be reported. This is important to ensure that no overfitting of the prediction model in one of the cohorts has occurred. In addition, the performance of the prediction models is directly related to our ability to single out patient subgroups where treatment may be highly beneficial or unsafe. Kent et al demonstrated that the event rate and the discriminative ability of the prediction model can predict very well the distribution of predicted risk. Lower event rate and higher c-statistic (given good calibration) result in high risk heterogeneity, thus making estimated average treatment effects uninformative. In this case, risk stratified analysis of HTE can be more effective in singling out patient subgroups that stand to benefit (or be harmed) most by treatment in question.”*

Also, in section 2.4 regarding the estimation step we added:

“Before focusing on the results of the estimation process we need to evaluate if adequate covariate balance was achieved within each risk stratum accounting for measured confounding. Common approaches include evaluation of the overlap of propensity score distributions and calculation of standardized covariate differences before and after propensity score adjustment.”  
  
Results  
The results section describes the experiment, which is to some extent the experimental plan, although the reader is not informed by the goals of the evaluation, except a demonstration. The steps that were described in general in the Methods are presented here with the use case. The data is three databases from IBM, however, it doesn’t demonstrate the power of OMOP in the sense that these are different institutes' databases, and even better from different countries, to demonstrate its advantages. However, the advantages of OMOP were demonstrated already in previous papers.

Thank you for pointing this out. Despite having data on 3 databases (CCAE, MDCR and MDCD) we only presented the results from one of them (CCAE) and placed the rest in the supplementary material. We agree with the Reviewer that the power of OMOP, cannot be fully understood in this way. For that reason we decided to provide the risk-stratified results across all databases, both on the relative and the absolute scale. For that reason we made the following changes in subsection “Presentation of results” of the Results section:

1. We replaced the figures showing effect estimates across risk strata on the relative and the absolute scale in CCAE with 2 different figures. The first (Figure 4) shows the relative effect estimates in all 3 databases within quarters of predicted acute MI risk—estimated using an internally developed risk prediction model within each database—for all the 9 outcomes of interest. The second (Figure 5) shows the absolute effect estimates in all 3 databases within quarters of predicted acute MI risk for all the 9 outcomes of interest.
2. We modified the text of the subsection to reflect the shift of focus from one database to all the databases of the analyses.

Chart, scatter chart

Description automatically generated

***Figure 4****: Overview of heterogeneity of ACE-inhibitors treatment on the relative scale (hazard ratios) within strata of predicted risk of acute MI. Values below 1 favor ACE inhibitors, while values above 1 favor beta blockers.*

Chart, scatter chart

Description automatically generated

***Figure 5****: Overview of heterogeneity of ACE-inhibitors treatment on the absolute scale within strata of predicted*

*risk of acute MI. Estimates of absolute treatment effect are derived as the difference in Kaplan-Meier estimates at*

*730 after inclusion. Values above 0 favor ACE inhibitors, while values below 0 favor beta blockers.*

The results are shown in several figures. What is the purpose of these results? to show the demonstrated analyses results? this is purpose of the paper? what is the baseline? to what is it compared? what is the alternative? Meanwhile, when showing the results' figures, please refer to them in the text.  
  
  
Discussion  
only here the code package of the project is mentioned. is this the contribution of the paper? to make it available to the readers? then make it clear why is it worth using it, in comparison to the alternatives.  
it is mentioned in several places that the proposed framework is standardized - how? what makes the use of this framework more standard than the alternatives? for that you have to discuss the alternatives.  
At the end of the discussion it is said that a proof of concept was demonstrated. what is the concept? what is the proof? it is said that it is easily applicable and highly informative - in what respect? how is it demonstrated in the paper?  
  
In summary, there is more to refer to, but as it can be seen the paper has to be rewritten and get a better structure (and add a background section, and an Evaluation section, and a clear motivation in the Introduction section).

**Reviewer 2:**

General Comments  
1. Topic is of importance: valid statistical approaches for determining and quantifying HTE are essential, making the topic of this manuscript timely and important. In this regard, the steps outlined by the authors will help researchers adopt a principled approach. Another strength of the manuscript is the inclusion of a worked example.  
2. Lack of specificity: inference for HTE in observational data is challenging, requiring precise statements about the causal parameter of interest, the representativeness of the sample of the target population, the explicit assumptions made and the validity of these assumptions, among other things. The manuscript does not provide enough detail and precision for the key features of causal inference.  
3. Multiplicity: the authors examine 9 outcomes and, for each, produce 4 treatment effectiveness estimates, without any control for multiplicity of inference.  
4. Example: the example is a good one but requires a more thorough analysis, description and validation of assumptions, vulnerabilities in the datasets they utilize due to incompleteness or ascertainment issues, and quantitative assessment to sensitivity to unmeasured confounding.

Detailed Comments  
1. Abstract: please be clearer about the "OHDSI methods library". Are you referring to methods unique to OHDSI or are you referring to a set of generic statistical models?

We agree that “OHDSI methods library” may be somewhat unclear. What we are referring to is the latter, i.e. a set of open-source tools for implementing standardized frameworks within the OHDSI setting. For example, there are standardized frameworks for performing patient-level prediction, population-level effect estimation etc. We changed this part of the abstract as follows:

*“...OHDSI library of open-source tools…”*

2. Approach  
a. Step 1: it would be useful to be more precise in the specific question so that the reader understands the causal question. As currently described, the authors goal is "to compare". Is the question to determine the difference in outcomes for a randomly selected person from a population when treated with A versus when treated with the control treatment? Or is it, for a randomly selected person from the population of treated persons, what is the difference in outcomes when treated with A versus when treated with the control treatment?

We agree with Reviewer 2 that we should have made our research questions clearer. We think that these issues are more relevant to the *Estimation* part (step 4) and not in the *General definition of the problem* part (step 1). In step 1 we are more interested general definitions on what the treatment and the comparator cohorts are and what outcomes we are interested in studying. Within our framework, average treatment effects and average treatment effects on the treated can be assessed simultaneously, depending on the problem and the researcher’s objectives. We made the following changes:

1. We added to section 2.4:

*“Effect estimation may be focused on the difference in outcomes for a randomly selected person from the risk stratum (average treatment effect) or for a randomly selected person from the treatment cohort within the risk stratum receiving the treatment under study (average treatment effect on the*

*treated)”.*

1. We added to section 3.4:

*“Our aim was to estimate the average treatment effects on the relative and the absolute scale within strata of predicted acute MI risk”.*

b. Step 2: identification of the database is a critical step and thus it would be helpful for the authors to provide more guidance to the reader. It would be important to know the sampling frame for the database, the completeness of information included in the database, the reliability of data entry, etc. While some of these features may be outside the scope of the paper, noting their importance for estimation of HTE is critical for the reader.

We agree that we may have been somewhat brief on the presentation of the database identification step. This was done as there are no universal inclusion rules and database selection can be very often problem-specific. That said, we agree with the reviewer that some more general considerations could be presented. We have modified section 2.2 in the following way:

*“The aim of this step is the inclusion of databases that represent the patient population of interest. It is required that the databases are mapped to the OMOP-CDM. The inclusion of multiple databases potentially increases the generalizability of results. Furthermore, the cohorts should preferably have adequate sample size with adequate follow-up time to ensure precise effect estimation, even within smaller risk strata. Other issues that may be of importance for database inclusion are the depth of data capture (the precision at which measurements, lab tests, conditions are recorded), the reliability of data entry and many more, also depending on the task at hand”.*

c. Step 3:  
i. Kindly define "target cohort" - I would have thought this is the population to which inferences apply. However, I believe the authors mean the "sample". This use of "target" is different from that commonly adopted in causal inference and may cause confusion.

We agree with Reviewer 2 that “target cohort” should be better explained as it is not a standard term. What we mean is the initial cohort (a set of patients that satisfy one or more inclusion criteria for a period of time) from which the population on which the prediction model will be developed. This population (the “sample”) is defined by applying further restrictions on the target cohort, e.g. after excluding patients that had the outcome of interest before being included in the cohort. We have made the following modifications to better clarify this issue:

* We added in the first paragraph of Section 2.3:

*“This prediction framework requires the definition of two essential cohorts: a target cohort, i.e. a set of patients hat satisfy one or more inclusion criteria for a duration of time, and an outcome cohort”.*

* We added in the second paragraph of Section 2.3:

*“Further restrictions can be applied on the target cohort to construct the final population on which the prediction model will be developed (e.g. exclude patients with a prior outcome in their history, before being included in the target cohort)”.*

ii. Prediction: it seems you are using the data twice: once to get the predicted risk for patients and then again to estimate HTE. This seems incorrect - please justify this step, indicating how standard errors and estimates are impacted.

We agree with Reviewer 2 that using the data twice, once for predicting risk and then for estimating HTE may be problematic. Ideally, we would like to use an external and well-validated model for risk stratifying the study population. This was already pointed out in the PATH statement (REF), that provided guidance for performing risk-based assessment of HTE in the RCT setting. However, when such a model is not available an earlier simulation study (REF Burke) demonstrated that internally developed prediction models produce relatively unbiased estimates of treatment effect across the spectrum of risk. In our framework we followed the guidance provided in (REF Burke) and the PATH statement for using internal models: develop the prediction model on the combined treatment and control arms—not only on the control arm as this will result in more bias (REF: Burke; Abadie)—blinded to treatment. We highlighted all these in the Discussion section by adding the following:

*“Ideally, externally derived and adequately validated prediction model would be preferred for analyzing treatment effect heterogeneity. In the absence of such prediction models an internally-developed risk prediction model can be considered. Earlier simulations of RCT studies have shown that internal models developed on the combined treatment and control arms blinded to treatment gave relatively unbiased estimates of treatment effect across the spectrum of risk ”.*

iii. Propensity score: I believe your goal is to balance observed characteristics of persons in the two treatment arms. Therefore, can you provide some justification for use of matching and for use of the propensity score? Would you recommend other approaches? Why not weight?

That is correct. The goal of PS matching before carrying out the development of the prediction model is to balance the patient characteristics between the 2 treatment arms. As we are presenting our framework within the observational setting, treatment is not administered at random. Therefore, substantial imbalances between patient characteristics of different treatment arms that are also risk factors may exist. This can generate spurious treatment-risk interactions when estimating treatment effects within risk strata, leading to suboptimal evaluation of risk-based HTE. We selected matching on the PS because the matched subset of patients facilitates a straightforward setting within which the prediction model will be developed.

As the prediction model is developed on a subset of the original pooled treatment and comparator cohorts, we need to make sure that it has adequate performance on the entire population. Also, we need to make sure that it was not massively overfit in one treatment arm. We can compare the performance of the model in the treatment and the comparator cohorts separately to have a better idea on its behavior. We have added the following in the last paragraph of section 2.3:

*“After model development, a performance overview of the derived prediction models including discrimination and*

*calibration both in the propensity score matched subset, the entire population and separately for treated and comparator patients should also be reported. This is important to ensure that no overfitting of the prediction model in one of the cohorts has occurred”*

In most approaches, assessment of overlap and of covariate balance are two critical steps that should be taken, and therefore, these should be discussed? Moreover, it seems that machine learning could be used to estimate the treatment assignment mechanism - why is machine learning only discussed in the context of the prediction algorithm? Why balance treatment arms now?  
iv. Multiple outcomes: do you estimate different risk strata for each outcome?  
v. Poolability of datasets: the error in estimation may differ across databases. Kindly indicate how poolability of the datasets is determined.  
d. Step 4:  
i. Why quartiles? Perhaps comment on how much variation in the predictions is required to create a fixed number of risk strata.  
ii. Are you re-estimating the propensity score again, within each stratum? I may be confused on this point. Please provide justification for balancing within stratum again. How does this impact the estimator (that is, two design effects of matching)?

Our risk stratified approach for evaluating HTE can be considered as an alternative subgroup analysis. Instead of using the values of specific covariates for stratifying the population we use a summary score, i.e. an estimate of the baseline risk. At the moment, best practices for performing subgroup analyses within the observational setting are not in place (Wang et al 2017 [<https://doi.org/10.1002/pds.4328>]). However, the results from simulation studies have shown that using subgroup-specific propensity score models generally result in lower bias than using the overall propensity score model (Izem et al 2020 [<https://doi.org/10.1080/10543406.2020.1730868>]; Wang et al 2017).

e. Step 5: covariate balance should occur prior to estimation of treatment effects.

This issue was also raised by Reviewer 1. We agree with both reviewers that any evaluation steps need to be performed before any estimation is carried out. We have removed the evaluation section and added relevant parts to the prediction and estimation sections. We provided more details in our response to Reviewer 1.

3. Results:  
a. Step 1:  
i. Definition of problem: please state if you are interested in the ATT, ATE, CATE, CATT, etc.

We agree with Reviewer 2 that the target population for causal inference should be clearer. We have already described the changes we made in both sections 2 and 3 earlier in our response to Reviewer 2.

ii. I do not understand what is meant by "one year of follow-up before treatment initiation". This restriction seems to be one related to a database rather than to defining a population. For instance, I would have expected some age or comorbidity restrictions. You may be requiring one-year of "continuous" enrollment but this seems to belong to the database section.

We agree with Reviewer 2. We have moved the follow-up requirement to the section related to database selection. Therefore, we added to section 3.2:

*“We required that all included patients have at least one year of follow-up before initiating treatment with an ACE inhibitor or a beta blocker”.*

b. Step 2: the sampling frame for MarketScan should be described, the completeness of the data (does MarketScan Medicaid data include those under managed care arrangements or just fee for service), are you looking a dually-eligible individuals, what are the pharmacy benefits (for instance, the Medicare beneficiaries need to be enrolled in Part D), how many diagnosis codes are used (Medicaid typically records fewer than Medicare which will lead to differential ascertainment), etc.  
c. Step 3:  
i. Define what constituted censoring (e.g., disenrollment, aging out of Medicaid, etc.).  
ii. Death is a competing risk for the outcomes reported; how was that handled?  
iii. How was covariate balance and overlap assessed?

We agree with Reviewer 2 that proper evaluation of covariate balance before and after stratification on the PS and of the density overlap of the PS was not readily available in the main manuscript, although they were presented in the supplementary material. We have included Figures 2 and 3 presenting the overlap of PS densities and the covariate balance before and after stratification within all risk strata across the 3 databases.

Histogram

Description automatically generated with medium confidence

***Figure 2****: Preference score distributions for the evaluation of heterogeneity of the effect of ACE inhibitors compared*

*to beta blockers on acute MI based on quarters of predicted acute MI risk. The preference score is a transformation*

*of the propensity score that adjusts for prevalence differences between populations.*

Graphical user interface

Description automatically generated

***Figure 3****: Patient characteristic balance for ACE inhibitors and beta blockers before and after stratification on the*

*propensity scores. Each dot represents the standardized difference of means for a single covariate before (x-axis)*

*and after (y-axis) stratification*

d. Estimation:  
i. Was the PH assumption met?  
ii. The matched design effects should be included in the estimation step (e.g., strata = pairs).  
iii. The confidence intervals should be adjusted for multiplicity (at least 9 outcomes and 4 estimates for each outcome).  
iv. An analysis assessing how sensitive findings are to unmeasured confounders should be included.

We agree with Reviewer 2 that unmeasured confounding is a serious concern. We have already elaborated on this issue in the Discussion section. We presented the use of negative controls (treatment-outcome pairs for which null effect has been established) as a potential sensitivity analysis for residual confounding. These negative controls can also be used to recalibrate our relative effect estimates (hazard ratios) for unobserved confounding. However, there is not yet a certain approach for evaluating unobserved confounding both on the relative and the absolute scale. We strongly believe that further research is required on this regard, however we feel that it falls outside the scope of the current work.

4. Reference 28: please supply the authors.  
  
  
  
**Reviewer 3:** This study attempted to develop a framework and tool to facilitate the treatment effect heterogeneity analysis to the observational databases based on the OMOP CDM. It is a meaningful study toward extending the application of the OMOP CDM.  
  
Major Comments:  
1. My major concern is the research hypothesis regarding this study. In this study, the authors hypothesized that the HTE analysis could be implemented on the observational databases. However, the HTE analysis is designed for the RCT study. The observational and RCT studies differ with the data collection, data distribution, data quality, etc. It is very hard to say whether the HTE analysis is suitable for the observational databases. It would be better if the authors could provide more evidence to support their hypothesis, such as adding more references or conducting a comparative study on the RCT data.

We share many of the concerns Reviewer 3 has expressed here. Various sources of confounding, both observed and unobserved, data quality, precision of data collection and many other issues may introduce bias to our evaluation of HTE in observational databases. The original framework on which we based our own (Kent et al 2020) was developed on the RCT setting, in which many of the previous problems do not exist. However, with the increasing availability of large observational healthcare databases we believe that assessment of HTE can benefit from the large-scale coverage of patient populations.

To address some of the many issues observational research we based our framework on current guidelines for prediction modeling and population-level effect estimation. Within OHDSI software packages following best practices for carrying out these tasks have been developed and extensively used. Our contribution is a combination of these methods for evaluating HTE, by estimating relative and absolute treatment effects within strata of predicted risk of an outcome of interest.

Finally, we agree with Reviewer 3 that comparing our results to an RCT-based HTE evaluation would provide additional support for our approach. However, we could not find any specific comparative effectiveness studies that could be transported to the observational setting. Very often the outcomes studied were based on very specific measurements that are not recorded int the databases we had available at the time.  
   
2. Introduction Section. The authors should add some review contents and references related to their study. And it would be better to provide some more details to emphasize the novelty of this study compared with others and the significance of their research.

We agree with Reviewer 3 that more information on the background of our work should have been included. This issue was raised by Reviewer 1, as well. We added a paragraph in the introduction section regarding this issue. For more information please refer to our response to Reviewer 1.  
  
3. Methods Section. It would be better if the authors move some contents from the Results Section to here to make the method easier to understand by the readers. For example, in subsection 2.3, the authors described the prediction as "Currently, the available options are regularized logistic regression, random forest…", but they didn't mention which specific method they used to build the predictive model here. The description of the specific estimation method is also illustrated in the Result section but not in subsection 2.4.

We agree with this suggestion. We moved the following paragraph from section 3.3 to section 2.3:

“*More specifically, we first estimate propensity scores using LASSO logistic regression and a large set*

*of baseline covariates including demographics, drug exposures, diagnoses, measurements and medical*

*devices. We match patients 1-1 using a caliper, i.e. the maximum distance that is acceptable for any*

*match. The default value we use is 0.2 on the standardized logit scale for the propensity scores*”.

We do not explain which predictive model we used in the Methods section, because our aim here is to present all the available methods that a researcher can use, if they use our framework along with the software that implements it. Multiple risk models can be considered, and different risk stratification schemes can be evaluated. However, in the following section (Results) where we present an implementation of our method, the specific prediction model used (LASSO logistic regression) is presented.

4. Results Section, Subsection 3.4. It would be better for the authors to provide more details about the estimation results here.

We wanted the presentation of the results to be a separate step of our framework, as the output can be quite rich. Therefore, in estimation step we only describe the estimation process to be followed within risk strata for the relative and absolute effect estimates. However, we think that evaluation of the effect estimation process (e.g. covariate balance before and after PS adjustment, PS density overlap etc) belong to this section. Therefore we have included the figures and the relevant text that we described in our response to Reviewer 2.  
  
5. Results Section, Subsection 3.5. The authors only show the prediction results of the acute MI in the CCAE here. Other results are listed in the supplementary files. In my opinion, it would be better to show more results, such as the prediction in the other two databases in the main manuscript, to demonstrate the robustness of their methods and increase the significance of their study.

We agree with Reviewer 3. We have made changes in our manuscript to attempt to deal with the concerns raised. We provided more details in our response to Reviewer 1.  
  
Minor Comments:  
1. Page 8. Section 3.3, line 28 - 30. The sentence "We chose a time horizon of 2 years after inclusion into the target cohort…" is duplicated her  
  
  
  
**Reviewer 4:** I enjoyed reading this paper and had to work to find constructive comments to improve on it.  
  
I think it's a very important contribution, given that heterogeneity of treatment effect has been until now largely under-investigated and for which methods such as the approach suggested here are much needed. As a biostatistician with great interest in risk prediction modeling, it coincides with many of my interests.  
  
I have only the following small suggestions to improve the presentation of the manuscript.  
  
1. In abstract should spell out OMOP before using it, other terms are spelled out.  
  
2. In the legend for Fig 1 and related text in manuscript, the term 'quartile' would be more appropriate than 'quarter'. For item C suggest 'the prediction model to the entire population'. Also for item D in this legend suggest 'We separate into risk subgroups'.  
  
3. page 6 line 4: quartiles instead of quarters  
  
4. On page 7, lines 2-3 suggest the following: 'As a proof of concept, ... to beta blockers. The former are among the most ... '  
  
5. page 7 line 17; suggest you consistently refer to 'efficacy' outcome rather than 'main' outcomes, as the latter suggests the safety outcomes are less important, and the two names are interchanged throughout manuscript.  
  
6. page 8 section 3.3: lines 18 - 20 repeat the prior sentence.  
  
7. page 8 line 26: can you clarify whether you then present average effects from among the five quintiles that you specify within the four quartiles of predicted risk?  
  
8. table 1 column title" quarter <- quartile  
  
9. page 9 line 12 main <- efficacy  
  
10. Fig 2: suggest putting the event rate below the other two and invert its direction, this will increase interpretability and graphically suggests that the absolute and relative risks build off of the treatment/comp event rates within each hof the predictive risk quartiles.  
  
Good luck, look forward to seeing in print.