

Deep learning-based ranking method for subgroup and predictive biomarker identification in patients

Corresponding Author: Dr Zihuan Liu

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)

The authors introduce a novel deep learning framework called DeepRAB that explores treatment effect heterogeneity and constructs individualized treatment rules (ITRs). DeepRAB leverages deep neural networks to model complex biomarker-causal effect relationships.

Strength:

1. DeepRAB incorporates model interpretability by enabling predictive biomarker identification. It utilizes a concrete autoencoder technique to select and rank important biomarkers influencing the treatment effects. This makes DeepRAB more interpretable compared to traditional black-box deep learning models.
2. Simulations under various scenarios demonstrate that DeepRAB outperforms existing approaches like causal forest, XGBoost with modified loss, and linear regression models in terms of power for subgroup exploration and capability to uncover predictive biomarkers. DeepRAB shows robustness even in the presence of weak effect signals and when detecting interactive treatment effects.
3. Application of DeepRAB to adalimumab (Humira) hidradenitis suppurativa clinical trial data showcases its practical usage. DeepRAB successfully identifies important predictive biomarkers like % reduction in AN count at week 12 and AN count at week 12, aligning with previous findings. It also boosts prediction performance for identifying patient subgroups with varying treatment responses.

Weakness:

1. Please provide a more detailed analysis of Figure 7 regarding insights into the similarities and differences in the interpretation results from different methods.
2. For interpretation results, since XGBoost and LRMC perform poorly regarding the prediction results, will their interpretation still make any sense?
3. Lack of error analysis or at least failure case studies.

Reviewer #2

(Remarks to the Author)

The author proposed DeepRAB that aimed at identifying predictive biomarkers that can boost subgroup identification and achieve precision medicine.

Though it is an interesting idea, the model developed and experiments may not be sufficient to support its claims.

1. The proposed network was a shallow MLP, with only relu/leaky relu as the nonlinear function. I would not term such a model as a "deep" model and the lack of non-linearity may lead to less robustness of the model.
2. Though the model structure had different configurations (e.g., different number of layers, different learning rate), the author failed to provide further details regarding how the model performed under different settings, thus it is hard to tell if the proposed model overfitted on the dataset consisting of 199 samples. Also, details regarding the implementation were missing. For example, if the learning being adjusted during training?

3. The authors predefined a few potential biomarkers and didn't analyze if the biomarkers are inner-correlated. Thus, the proposed model may be biased when selecting the "important" biomarkers.

4. The baseline selected were all non-deep learning approaches, The conclusions will be more convincing if their model is superior over other deep learning-based models. The author may consider to fine-tune some state-of-the-art deep learning-based classification models and compare with it.

5. Should the method section be placed before the acknowledgement?

Reviewer #3

(Remarks to the Author)

This work proposes DeepRAB, a deep learning-based method primarily for identifying predictive biomarkers, the variables that are informative for the heterogeneous treatment effect (HTE), and subgroups that will benefit from the treatment. These two research questions are important questions that have attracted many research endeavors. This work combines A-learning --- a framework for HTE estimation and subgroup identification --- and concrete autoencoder, a neural network structure that facilitates variable selection and dimension reduction, to select the informative variables regarding HTE as well as estimate HTE. This innovative approach directly targets HTE through A-learning and leverages the predictive power of neural networks. This work evaluates the proposed method through numerical experiments, including a simulation and a case study using the data set of two clinical trials, AbbVie Pioneer I and Pioneer II. The results demonstrate the advantages of DeepRAB over other methods.

Despite the significance of the addressed problems and the potential of the proposed approach, there are serious concerns about both the proposed approach and the numerical experiments, as listed below:

Major Comments:

1. Deep Learning is widely known to behave worse than classical machine learning methods on data with small sample sizes and only excels on data with large sample sizes. However, the sample sizes in clinical trials are usually small. For example, Pioneering I only includes 307 patients, and Pioneering II only includes 326 patients, where each of the patients has at least 15 attributes. Deep learning is prone to overfitting in this context. However, the case study does not seem to address this concern. The "Tunning" section seems to suggest that the authors use the entire data set to do cross-validation to select the tuning parameters and use the same data set for evaluation, which gives Figure 8: Proportion of patients achieving HiSCR. This makes the evaluation an in-sample test and inadequate to evaluate the true performance of the method. What should be done is to split the data into two parts: one for tuning and training, the other for testing. Additionally, it is recommended that the authors clarify their training and evaluation mechanisms and also report the structure and parameters of the resulting neural network to demonstrate that the resulting DeepRAB does not degenerate into models akin to linear regression or its variants.

2. Neural networks are known to lack theoretical guarantees and are therefore considered black boxes. Current theoretical investigations on deep learning, or neural networks in general, are based on simplified toy models, aiming to gain some initial understanding.

The authors claim that the DeepRAB is not a black box method, yet they provide no theoretical guarantees to support this claim.

3. The absence of theoretical justification of DeepRAB and the lack of ground truth in clinical trials emphasize the need for comprehensive simulations to evaluate the proposed method. However, the simulations conducted test only one function under limited settings, which does not convincingly reflect the method's overall performance.

a. The function that authors include in the simulation can be approximated infinitely well by a neural network, and it is likely to favor neural network-based methods. More diverse and representative functions need to be included in the simulation, including those that can be well approximated by other models (e.g., linear function, quadratic functions, staircase functions).

b. The signal-to-noise ratios in the current experiment do not include strong-signal regions (e.g., $\beta > 3$).

c. The sample sizes are not reported and presumably not varied.

4. Simulation results show unexpected patterns in the performances of XGboostML and Causal Forest persistently. For example, in Figure 3, XGboostML persistently singles out X6 for $\beta_0 > 0$; in Figure 5, XGboostML persistently singles out X5 for $\beta_0 > 0$. However, according to the true model, X4-X10 are of equal status, which means that for reasonable methods, X4-X10 should have the same probability of being chosen incorrectly and we should expect similar detection rates of X4-X10. In Figure 5, for $\beta_0 = 2$, Causal Forest singles out X4 persistently, which is also unexpected for the same reason. Could the authors explain this unexpected phenomenon or give some evidence that the code and simulation are correctly written and executed?

Minor Comments:

1. In the section "simulation framework", line 423 does not give a binary outcome; please write a rigorous formula. I assume the authors want to transform continuous values onto $[0, 1]$ to represent probabilities of taking 0 or 1, but the current version does not reflect that.

2. The authors claim to explore HTE but do not adequately attempt to estimate HTE.

3. There are grammatical errors and typos; I list some here.

a. Line 199: Replace "similarity" with "similarly".

b. Line 218: missing comma after "(Figure 6)".

c. Line 232: According to Figure 6, "DeepRAB on patients" should be "DeepRAB on responded patients".

- d. Line 242: Replace "comprised patients" with "comprised of patients".
- e. Line 379: Replace "with P = " with "with probability P= ".
- f. Line 380: Recheck the notation and the sentence.
- g. Line 387: Replace "and composed of" with "and are composed of".

Reviewer #4

(Remarks to the Author)

This paper introduces DeepRAB, a deep learning-based framework designed for subgroup identification and predictive biomarker discovery in precision medicine. The study addresses the need for personalized treatment strategies by leveraging deep learning algorithms to uncover treatment effect heterogeneity and identify important biomarkers. DeepRAB outperforms existing methods in subgroup exploration and biomarker discovery, as demonstrated through analyses of simulated data and application to a clinical trial dataset. The manuscript provides a clear research question, adequate background, and detailed methodology for replication. Results are presented accurately, with appropriate statistical analysis, discussion on experiments and limitations.

Here are a few points I would like to discuss with the authors :

1. Could the authors provide details on the number of learnable parameters in the proposed network? How does this compare with other existing methods in terms of model complexity?
2. What volume of data was generated during the simulation phase? Have the authors observed any overfitting phenomena in the training process on the synthetic data? Given that only 199 patients' data were involved in clinical validation, how do the authors address concerns about potential overfitting and limited generalizability? And this limited dataset is hard to validate the effectiveness and generalization ability of the proposed method.
3. The whole pipeline is not clear, especially for training. Could the authors provide a more detailed description of the entire pipeline, particularly for the training process? Specifically, is the model retrained after validation on synthetic data, or is the synthetic-trained model directly used for inference on clinical data?
4. Have the authors considered making the data and code involved in this research publicly available? The reproducibility of the method is very important.
5. In the Discussion section, the authors state that "DeepRAB method holds promising prospects for extension into neural imaging data modalities such as MRI, CT, and PET scans." Given the significant differences between biomarkers and imaging data, such as dimensionality, patterns and data size, what additional evidence or experiments could support this claim? May need more experiments to verify this statement.
6. There appears to be a typographical error in line 196: "a high detection rate of 98%" should be "a high detection rate of 93%".
7. How does DeepRAB compare to other state-of-the-art methods for subgroup identification and predictive biomarker discovery? While the current experiments focus on predictive biomarker selection, could the authors provide a more comprehensive comparison, particularly in subgroup identification?
8. Have ethical considerations related to the use of patient data or potential biases in the predictive biomarker identification process been addressed in the manuscript? Ethical considerations are essential in precision medicine research.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Thank the authors for the detailed reply. I don't have further questions.

Reviewer #2

(Remarks to the Author)

1. Could the authors update the rebuttal by citing the revised part of the manuscript, or at least listing Line numbers? It is not very easy to identify which part of the revision corresponding to which question.
2. The authors added two simulations with data biomarker data, X , generated from a multivariate normal distribution with a mean of 0, a variance of 2, and pairwise correlations of 0.2 among the biomarkers. ---- This is a very neat dataset, while the real-world data sample may not follow such a nice distribution. I know it's hard to analyze the real-world clinical data, could the author at least discuss the gap between the theoretical simulation data and practical clinical data?
3. There might not be an existing deep learning-based biomarkers identification method to compare with. But DNN + casual inference is not a new idea in medical domain. Could the authors briefly discuss those? For example, Learning Counterfactual Representations for Estimating Individual Dose-Response Curves

Reviewer #3

(Remarks to the Author)

The revised manuscript has significant improvement in quality, with added details on the neural network, training, and testing procedures, as well as the inclusion of more scenarios for evaluation. However, there remain several concerns and questions that need to be addressed, as outlined below:

1. Clarification on AUC Calculation

- The manuscript uses AUC (Area Under the Curve) to measure the efficacy of subgroup identification but does not provide sufficient details on its calculation. AUC is typically derived from the ROC curve, which evaluates a spectrum of classifiers. However, for a fixed classifier, only a single (true positive rate, false positive rate) pair exists. Thus, it is unclear how AUC was calculated in this context.
- Presumably, the spectrum of classifiers for DeepRAB is based on $\hat{f}$, while for other methods, it may rely on the estimated treatment effect. However, methods designed for subgroup identification may not necessarily excel in terms of AUC, as AUC does not directly assess subgroup identification.
- To ensure appropriate evaluation, the manuscript should report false positive rates, false negative rates, and variance across different test set trials for the classifiers' vanilla forms (e.g., thresholding at 0 for $\hat{f}$ in DeepRAB).

2. AUC in Null Scenarios

- For null scenarios, the true positive rate is undefined since the denominator (ground truth positives) is zero. Consequently, neither the ROC curve nor the AUC is well-defined in such cases. The manuscript needs to clarify how AUC values were computed under these circumstances.

3. Interpretation of Results in Simulation I and II

- In the section on comparative performance in Simulation I and II, the authors suggest that baseline models' high biomarker detection rates and inferior AUC performance may indicate overfitting. This interpretation is not convincing, especially since Simulation I involves a linear setting with 10 features and a sample size of 1000, where linear-model-based methods are unlikely to overfit.
- This empirical phenomenon might instead highlight issues with either the AUC calculation or its appropriateness as a criterion for evaluating subgroup identification, reinforcing the concerns raised in Points 1 and 2.

4. Specifying the Number of Biomarkers

- The structure of the neural network requires pre-specifying the number of biomarkers to be selected, but this information is typically unknown in practice. How should practitioners determine the number of biomarkers, and what are the consequences of incorrect specification?
- Additionally, other methods do not rely on knowing the number of biomarkers. Would evaluating under the case where the number of biomarkers is known introduce an unfair advantage for DeepRAB?

5. Figure 8 (Importance of Biomarkers)

- The x-axis in Figure 8 is not specified. The manuscript should explain how the importance of biomarkers is calculated.

6. Duplicate Figure Labels

- There are two figures labeled as Figure 8. This duplication should be corrected to avoid confusion.

Reviewer #4

(Remarks to the Author)

Please see attachment.

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

Thank you for adding more simulations which I concerned the most from the last two rounds reviews. I have no further questions for the authors.

Reviewer #3

(Remarks to the Author)

I do not have further questions.

Reviewer #4

(Remarks to the Author)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Dear Editor,

We wish to thank you and the four referees for their encouragement and insightful comments. Their valuable comments helped us improve the presentation and the clarity of our manuscript. In the following, we respond to the referees' comments point by point. The reviewers' comments are in *italic* followed by our response in a standard font.

Comments from Reviewer 1

1. The authors introduce a novel deep learning framework called DeepRAB that explores treatment effect heterogeneity and constructs individualized treatment rules (ITRs). DeepRAB leverages deep neural networks to model complex biomarker-causal effect relationships.

Strength:

- 1. DeepRAB incorporates model interpretability by enabling predictive biomarker identification. It utilizes a concrete autoencoder technique to select and rank important biomarkers influencing the treatment effects. This makes DeepRAB more interpretable compared to traditional black-box deep learning models.*
- 2. Simulations under various scenarios demonstrate that DeepRAB outperforms existing approaches like causal forest, XGBoost with modified loss, and linear regression models in terms of power for subgroup exploration and capability to uncover predictive biomarkers. DeepRAB shows robustness even in the presence of weak effect signals and when detecting interactive treatment effects.*
- 3. Application of DeepRAB to adalimumab (Humira) hidradenitis suppurativa clinical trial data showcases its practical usage. DeepRAB successfully identifies important predictive biomarkers like % reduction in AN count at week 12 and AN count at week 12, aligning with previous findings. It also boosts prediction performance for identifying patient subgroups with varying treatment responses.*

Response: Thank you for the positive feedback on our manuscript. We appreciate your recognition of its strengths, particularly in model interpretability, performance under various simulation scenarios, and practical application.

Weakness:

- 1. Please provide a more detailed analysis of Figure 7 regarding insights into the similarities and differences in the interpretation results from different methods.*

Response: Thank you for your suggestion. In the revised manuscript, we have expanded the discussion of similarities and differences in the interpretation results from different methods in Figure 7. The expanded analysis can be found in the section of real data of the revised manuscript.

- 2. For interpretation results, since XGBoost and LRMC perform poorly regarding the prediction results, will their interpretation still make any sense?*

Response: Thank you for the insightful question. In our original manuscript, we have demonstrated that XGBoostML and LRMC perform poorly in some simulation settings. This outcome is expected for LRMC, as it is a linear regression-based model and thus not well-suited for capturing non-linear relationships. Although XGBoost ML exhibits poor performance under conditions of weak predictive effect signals, we want to clarify that XGBoostML demonstrates a relatively robust performance in scenarios where the prognostic effect is absent, and the predictive effect is strong.

To further evaluate these two models' performance at different settings, we have included two additional simulation studies in the revised manuscript, which showcase that both XGBoost and LRMC can indeed perform adequately under specific conditions. These new simulations reveal that these models can be quite effective in certain settings, thereby supporting their utility in interpreting results from data.

In real-world applications, the true nature of associations between biomarkers (X) and outcomes (Y) often remains uncertain. Thus, employing a range of models to identify predictive biomarkers allows for a comprehensive comparative analysis. This approach helps to ascertain which biomarkers are consistently recognized across different methodologies, enhancing our understanding of their relevance amidst the complexities of real-world data. For instance, CHG_AN is consistently identified by DeepRAB, CF, and XGBoostML as one of the top biomarkers, affirming its importance in evaluating the continued therapeutic benefits of adalimumab. This finding is corroborated by existing literature^{1,2}, which further validate CHG_AN as a critical biomarker.

In conclusion, while XGBoostML and LRMC may encounter challenges in certain simulated environments, they can still offer significant insights in more favorable settings and real-world applications. We believe their contributions to biomarker identification in clinical data remain invaluable. We have added these changes to discussion section of revised manuscript.

3. Lack of error analysis or at least failure case studies.

Response: Thank you for your suggestion. We have expanded our examination of error analysis in the revised manuscript. Specifically, we have introduced scenarios where there is no association between predictive biomarkers and outcomes. These calculations have been conducted and documented in the section on Null Scenarios in the revised manuscript.

Comments from Reviewer 2

The author proposed DeepRAB that aimed at identifying predictive biomarkers that can boost subgroup identification and achieve precision medicine. Though it is an interesting idea, the model developed and experiments may not be sufficient to support its claims.

1. The proposed network was a shallow MLP, with only relu/leaky relu as the nonlinear function. I would not term such a model as a "deep" model and the lack of non-linearity may lead to less robustness of the model.

Response: Thank you for your feedback. We would like to emphasize that our paper investigates the numerical properties of neural networks where the architecture can include more than one layer. While there is no universally agreed-upon threshold defining the transition from shallow to deep learning, the term “deep” network is commonly used to describe neural networks with more than one hidden layer^{3,4}. In our manuscript, we discuss selecting the number of hidden layers from {1, 2, 3, 4, 5} in the tuning section, which justifies our use of the term “deep” learning. However, we acknowledge the concern regarding the use of only ReLU and leaky ReLU activation functions, which may limit the non-linearity and reduce the robustness of the model. In response to this point, we have enhanced our tuning procedure to include additional activation functions, specifically Sigmoid and Tanh. We have incorporated these changes in the tuning section of revised manuscript.

2. Though the model structure had different configurations (e.g., different number of layers, different learning rate), the author failed to provide further details regarding how the model performed under different settings, thus it is hard to tell if the proposed model overfitted on the dataset consisting of 199 samples. Also, details regarding the implementation were missing. For example, if the learning being adjusted during training?

Response: Thank you for your insightful suggestions. We have expanded the section of Tuning to include in-depth information about the CV process used for hyperparameter optimization. This includes a detailed description of how we selected the optimal learning rate and other critical parameters. Additionally, we conducted two new simulation studies to further evaluate the model’s performance across different sample sizes (400 and 1000), which more closely mirror real-world clinical trial data. These three simulations were designed to address scenarios where the ground truth is unknown, as in clinical settings, and to assess whether the model can generalize without overfitting. The results demonstrated that our model performs well across all scenarios and mitigates overfitting through the proposed CV approach. Further details can be found in the revised manuscript.

3. The authors predefined a few potential biomarkers and didn't analyze if the biomarkers are inner-correlated. Thus, the proposed model may be biased when selecting the "important" biomarkers.

Response: Thank you for your valuable suggestion. In light of your comments, we have redesigned our simulation studies to assess the robustness of our model in the presence of inter-correlated biomarkers. Specifically, we generated biomarker data, X , from a multivariate normal distribution with a mean of 0, a variance of 2, and pairwise correlations of 0.2 among the biomarkers. This design allows us to evaluate whether our model can accurately identify the true biomarkers under conditions where they are interrelated. Furthermore, we have enhanced the manuscript by adding two additional simulation studies using the same data generation scheme for X . The results from these simulations

demonstrate that our model consistently identifies the true biomarkers across all scenarios, affirming its capability to remain unbiased in the selection of important biomarkers, even when they exhibit mutual correlations. We have incorporated these changes in the section of Simulation framework of revised manuscript.

4. The baseline selected were all non-deep learning approaches, The conclusions will be more convincing if their model is superior over other deep learning-based models. The author may consider to fine-tune some state-of-the-art deep learning-based classification models and compare with it.

Response: Thank you for your comment. We want to clarify that our proposed deep learning model represents one of the first efforts to simultaneously perform subgroup identification and predictive biomarker selection. This dual capability is a key novelty of our manuscript. While we recognize the value in comparing our model against other state-of-the-art deep learning models, our extensive literature search did not identify any existing deep learning-based models specifically designed for this dual purpose in subgroup analysis. The baseline models we selected for comparison are the most commonly used methods in subgroup analysis, and they provide a meaningful benchmark for evaluating our model's performance. These have been widely validated and accepted in the field. We acknowledge the value in comparing our model to other deep learning approaches and intend to explore this in future work as suitable models become available. We have incorporated this into discussion section of revised manuscript.

5. Should the method section be placed before the acknowledgement?

Response: Thank you for your suggestion. We have moved the method section before results section.

Comments from Reviewer 3

This work proposes DeepRAB, a deep learning-based method primarily for identifying predictive biomarkers, the variables that are informative for the heterogeneous treatment effect (HTE), and subgroups that will benefit from the treatment. These two research questions are important questions that have attracted many research endeavors. This work combines A-learning --- a framework for HTE estimation and subgroup identification --- and concrete autoencoder, a neural network structure that facilitates variable selection and dimension reduction, to select the informative variables regarding HTE as well as estimate HTE. This innovative approach directly targets HTE through A-learning and leverages the predictive power of neural networks. This work evaluates the proposed method through numerical experiments, including a simulation and a case study using the data set of two clinical trials, AbbVie Pioneer I and Pioneer II. The results demonstrate the advantages of DeepRAB over other methods.

Despite the significance of the addressed problems and the potential of the proposed approach, there are serious concerns about both the proposed approach and the numerical experiments, as listed below:

Response: Thank you for your positive feedback. We appreciate your recognition of the significance of our work. We have carefully addressed your concerns as outlined below.

Major Comments:

1. Deep Learning is widely known to behave worse than classical machine learning methods on data with small sample sizes and only excels on data with large sample sizes. However, the sample sizes in clinical trials are usually small. For example, Pioneering I only includes 307 patients, and Pioneering II only includes 326 patients, where each of the patients has at least 15 attributes. Deep learning is prone to overfitting in this context. However, the case study does not seem to address this concern. The "Tunning" section seems to suggest that the authors use the entire data set to do cross-validation to select the tuning parameters and use the same data set for evaluation, which gives Figure 8: Proportion of patients achieving HiSCR. This makes the evaluation an in-sample test and inadequate to evaluate the true performance of the method. What should be done is to split the data into two parts: one for tuning and training, the other for testing. Additionally, it is recommended that the authors clarify their training and evaluation mechanisms and also report the structure and parameters of the resulting neural network to demonstrate that the resulting DeepRAB does not degenerate into models akin to linear regression or its variants.

Response: Thank you for your valuable feedback. We have addressed your concerns by providing additional details about the CV procedure in the tuning section of the revised manuscript. To clarify, in our simulation studies, we split

the dataset into separate training and testing sets within the CV framework to mitigate the risk of overfitting. However, in the case of real-world data, we used the entire dataset for cross-validation.

There are two key reasons for this approach. First, the true subgroup labels in real-world datasets are unknown. Therefore, our primary objective is to use the model to identify predictive biomarkers, and using the entire dataset for cross-validation helps achieve this. Ideally, validation would be performed using an independent dataset to confirm whether the model yields consistent results. Unfortunately, we do not have access to such a dataset at present.

However, we believe the robustness of our model is supported by the fact that, in our real-world analysis, DeepRAB, along with CF and XGBoostML, consistently identified CHG_{AN} as a significant biomarker. This finding is in line with prior studies^{2, 5}, suggesting that our model is not producing overly optimistic results, as it has been independently validated by previously published research.

Secondly, given the small sample size in our real-world dataset, we believe that including all available data in the CV procedure is the most appropriate approach. We have now provided further details about the tuning procedures, training, and evaluation mechanisms in the revised manuscript. Furthermore, we have reported the structure and parameters of the resulting neural network in the real-world data section.

2. Neural networks are known to lack theoretical guarantees and are therefore considered black boxes. Current theoretical investigations on deep learning, or neural networks in general, are based on simplified toy models, aiming to gain some initial understanding. The authors claim that the DeepRAB is not a black box method, yet they provide no theoretical guarantees to support this claim.

Response: Thank you for your feedback. We agree that neural networks are often considered black boxes due to their lack of theoretical guarantees, and we have accordingly revised the manuscript to use “non-transparent” in place of “black-box.” In our manuscript, we claim that DeepRAB is not a black box method because it incorporates feature selection, specifically predictive biomarker selection. This approach helps guide users by identifying which biomarkers are associated with treatment effects and provides insights into how the model makes decisions based on each biomarker. While we acknowledge that our current work does not provide theoretical guarantees, we believe that the interpretability of DeepRAB through feature selection is a significant step towards reducing the black box nature of neural networks. We recognize the importance of theoretical investigations and plan to explore the theoretical properties of DeepRAB in future work.

3. The absence of theoretical justification of DeepRAB and the lack of ground truth in clinical trials emphasize the need for comprehensive simulations to evaluate the proposed method. However, the simulations conducted test only one function under limited settings, which does not convincingly reflect the method's overall performance.

a. The function that authors include in the simulation can be approximated infinitely well by a neural network, and it is likely to favor neural network-based methods. More diverse and representative functions need to be included in the simulation, including those that can be well approximated by other models (e.g., linear function, quadratic functions, staircase functions).

Response: Thank you for your feedback. We have added two more simulations based on linear functions and quadratic functions, to provide a more comprehensive evaluation of our method. We have incorporated these changes in the revised manuscript.

b. The signal-to-noise ratios in the current experiment do not include strong-signal regions (e.g., $\beta > 3$).

Response: Thank you for your suggestion. We have incorporated all simulation scenarios with $\beta > 3$ to cover strong-signal regions in the revised manuscript. We have incorporated these changes in the revised manuscript.

c. The sample sizes are not reported and presumably not varied.

Response: Thank you for your observation. We apologize for not specifying the sample sizes in the original manuscript. In our simulations, we used a sample size of 1000. In the revised manuscript, we also consider an additional sample size of 400 across all simulation scenarios. We believe that these three simulation studies, using sample sizes of 400 and 1000, provide a comprehensive evaluation of the model's performance and better reflect real-world clinical trial data. We have incorporated these changes in the revised manuscript.

4. Simulation results show unexpected patterns in the performances of XGboostML and Causal Forest persistently. For example, in Figure 3, XGboostML persistently singles out X_6 for $\beta_{-0} > 0$; in Figure 5, XGboostML persistently singles out X_5 for $\beta_{-0} > 0$. However, according to the true model, X_4 - X_{10} are of equal status, which means that for reasonable methods, X_4 - X_{10} should have the same probability of being chosen incorrectly and we should expect similar detection rates of X_4 - X_{10} . In Figure 5, for $\beta_{-0} = 2$, Causal Forest singles out X_4 persistently, which is also unexpected for the same reason. Could the authors explain this unexpected phenomenon or give some evidence that the code and simulation are correctly written and executed?

Response: Thank you for your insightful observations. We have reviewed our code setup to ensure their accuracy. After re-executing the original simulations, we found consistent results, confirming that the observed patterns are not due to any coding errors.

There are several potential explanations for these patterns. First, in the original manuscript, we ran the simulations only 100 times, which may have introduced some bias in feature selection. To address this, we have increased the number of replications to 1000 in the revised manuscript, which provides more robust results.

Second, these patterns may be attributed to inherent biases in the feature selection process of different models. Tree-based models like XGboostML and CF can prioritize certain splits based on data distribution and impurity reduction criteria, which may lead to the persistent selection of specific variables, such as X_4 or X_5 .

Third, we have also explored additional tuning parameters for both XGboostML and CF to ensure the models reach their optimal performance.

In the revised manuscript, we have re-run all three simulations with 1000 replications, and now observe more equal feature selection across $X_5 - X_{10}$, addressing the previous concern. We also plan to make the simulation code available on GitHub to enhance transparency. We hope these changes clarify the observed phenomenon and provide confidence in the reliability of our results.

Minor Comments:

1. In the section "simulation framework", line 423 does not give a binary outcome; please write a rigorous formula. I assume the authors want to transform continuous values onto $[0, 1]$ to represent probabilities of taking 0 or 1, but the current version does not reflect that.

Response: Thank you for your feedback. We have included a clear formula for the binary outcome simulation.

2. The authors claim to explore HTE but do not adequately attempt to estimate HTE.

Response: Thank you for your feedback. We apologize for the confusion. Our primary goal is to explore ITR. We have clarified this in the revised manuscript by updating the wording accordingly.

3. There are grammatical errors and typos; I list some here.

a. Line 199: Replace "similarity" with "similarly".

Response: We have corrected the typo and replaced "similarity" with "similarly."

b. Line 218: missing comma after "(Figure 6)".

Response: We have added the missing comma after "(Figure 6)."

c. Line 232: According to Figure 6, "DeepRAB on patients" should be "DeepRAB on responded patients".

Response: We have revised the sentence to "DeepRAB on responded patients" as per your suggestion.

d. Line 242: Replace "comprised patients" with "comprised of patients".

Response: We have corrected the phrase to "comprised of patients."

e. Line 379: Replace "with P = " with "with probability P = ".

Response: We have replaced “with $P =$ ” with “with probability $P =$ ” for clarity.

f. Line 380: Recheck the notation and the sentence.

Response: We have reviewed and rechecked the notation and sentence for accuracy.

g. Line 387: Replace “and composed of” with “and are composed of”.

Response: We have corrected the phrasing to “and are composed of.”

Comments from Reviewer 4

This paper introduces DeepRAB, a deep learning-based framework designed for subgroup identification and predictive biomarker discovery in precision medicine. The study addresses the need for personalized treatment strategies by leveraging deep learning algorithms to uncover treatment effect heterogeneity and identify important biomarkers. DeepRAB outperforms existing methods in subgroup exploration and biomarker discovery, as demonstrated through analyses of simulated data and application to a clinical trial dataset. The manuscript provides a clear research question, adequate background, and detailed methodology for replication. Results are presented accurately, with appropriate statistical analysis, discussion on experiments and limitations.

Response: Thank you for your encouraging and positive feedback. We appreciate your recognition of the contributions made by DeepRAB in addressing treatment effect heterogeneity and predictive biomarker discovery.

Here are a few points I would like to discuss with the authors:

1. Could the authors provide details on the number of learnable parameters in the proposed network? How does this compare with other existing methods in terms of model complexity?

Response: Thank you for your comments. We have included detailed information about the number of learnable parameters in the Tuning section of the revised manuscript. DeepRAB’s complexity is driven by several factors, including the number of layers, nodes per layer, learning rate, and the number of training iterations. While DNN-based models have a greater number of learnable parameters compared to traditional machine learning methods like linear regression or tree-based models, it is designed to handle more complex relationships within the data, such as non-linear interactions and feature dependencies. In contrast, models like XGboostML or CF tend to have fewer parameters but rely on extensive tree-building, which can be computationally intensive depending on the dataset size. These details have been incorporated into the discussion section of the revised manuscript.

2. What volume of data was generated during the simulation phase? Have the authors observed any overfitting phenomena in the training process on the synthetic data? Given that only 199 patients' data were involved in clinical validation, how do the authors address concerns about potential overfitting and limited generalizability? And this limited dataset is hard to validate the effectiveness and generalization ability of the proposed method.

Response: Thank you for your thoughtful comments. We would like to clarify that the original manuscript utilized a sample size of $N = 1000$ during the simulation phase. We understand the concerns regarding potential overfitting, especially given that only 199 patient data points were available for clinical validation. Now, we have introduced two additional simulation cases with a sample size of $N = 400$, bringing the total to three simulation cases. This allows us to examine a broader range of scenarios to assess the model’s performance, including condition of smaller sample size more reflective of real-world clinical data. Our results demonstrate that DeepRAB remains robust across all sample sizes.

In addition, we have provided additional details on the CV process, which includes performance evaluation on testing data. These evaluations indicate no evidence of overfitting based on the given simulation data. On the other

hand, we conducted null case simulations where no true subgroups exist, and the model's random feature selection in these scenarios further supports that it is not overly optimistic in selecting features.

We acknowledge the difficulty of validating the model's performance in real-world data where true subgroups are unknown. However, we are encouraged by the fact that the biomarkers identified by DeepRAB have been validated by previous publications, reinforcing the model's effectiveness in this context. We have included these in the revised manuscript.

3. The whole pipeline is not clear, especially for training. Could the authors provide a more detailed description of the entire pipeline, particularly for the training process? Specifically, is the model retrained after validation on synthetic data, or is the synthetic-trained model directly used for inference on clinical data?

Response: Thank you for your suggestion. We have provided a more detailed description of the entire pipeline, focusing on the training process, in the revised manuscript.

4. Have the authors considered making the data and code involved in this research publicly available? The reproducibility of the method is very important.

Response: Thanks for your suggestion, we have made the code publicly available on GitHub.

5. In the Discussion section, the authors state that "DeepRAB method holds promising prospects for extension into neural imaging data modalities such as MRI, CT, and PET scans." Given the significant differences between biomarkers and imaging data, such as dimensionality, patterns and data size, what additional evidence or experiments could support this claim? May need more experiments to verify this statement.

Response: Thank you for your valuable suggestion. Our claim that the DeepRAB method holds promising prospects for extension into neural imaging data modalities, such as MRI, CT, and PET scans, is based on the successful application of deep learning techniques in these imaging areas. Deep learning has shown significant potential in handling the high dimensionality, complex patterns, and large data sizes characteristic of imaging data. Our current model could be adapted to incorporate conventional layers designed to address the specific features of imaging datasets, such as convolutional layers for pattern recognition and spatial hierarchy in images. However, we acknowledge that integrating DeepRAB with imaging data would require additional development and validation.

Given the complexity of designing and validating such a model, we consider this an area for future work. In the revised manuscript, we have adjusted our discussion to reflect this and provided more detailed insights into how the model could be extended and validated with imaging data in future studies.

6. There appears to be a typographical error in line 196: "a high detection rate of 98%" should be "a high detection rate of 93%".

Response: Thank you for identifying this. We have corrected it in the revised manuscript.

7. How does DeepRAB compare to other state-of-the-art methods for subgroup identification and predictive biomarker discovery? While the current experiments focus on predictive biomarker selection, could the authors provide a more comprehensive comparison, particularly in subgroup identification?

Response: Thank you for your comments. In the original manuscript, we used AUC as the primary metric to evaluate subgroup identification performance.

8. Have ethical considerations related to the use of patient data or potential biases in the predictive biomarker identification process been addressed in the manuscript? Ethical considerations are essential in precision medicine research.

Response: Thank you for highlighting the importance of ethical considerations. In the current study, no specific ethical considerations were addressed in the manuscript.

Reference

1. Huang X, *et al.* Patient subgroup identification for clinical drug development. *Statistics in medicine* **36**, 1414-1428 (2017).
2. Huang X, Tian L, Sun Y, Chatterjee S, Devanarayan V. Predictive signature development based on maximizing the area between receiver operating characteristic curves. *Statistics in Medicine* **41**, 5242-5257 (2022).
3. Schmidhuber J. Deep learning in neural networks: An overview. *Neural networks* **61**, 85-117 (2015).
4. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. *Neural computation* **18**, 1527-1554 (2006).
5. Huang X, Li H, Gu Y, Chan IS. Predictive biomarker identification for biopharmaceutical development. *Statistics in Biopharmaceutical Research* **13**, 239-247 (2021).

Dear Editor,

We wish to thank you and the three referees for their encouragement and insightful comments. Their valuable comments helped us improve the presentation and the clarity of our manuscript. In the following, we respond the referees' comments point by point. The reviewers' comments are in *italic* followed by our response in a standard font.

Reviewer #1 (Remarks to the Author):

Thank the authors for the detailed reply. I don't have further questions.

Response: Thank you for your recognition and for taking the time to review our work.

Reviewer #2 (Remarks to the Author):

1. Could the authors update the rebuttal by citing the revised part of the manuscript, or at least listing Line numbers? It is not very easy to identify which part of the revision corresponding to which question.

Response: Thank you for your suggestion, and we apologize for any inconvenience caused. In our updated rebuttal letter, we have explicitly listed the line numbers corresponding to each revision. We hope these additions clarify the changes and allow for a more straightforward review process. In addition, we have updated our initial rebuttal letter by providing the relevant line numbers for your earlier comments.

"1. The proposed network was a shallow MLP, with only relu/leaky relu as the nonlinear function. I would not term such a model as a "deep" model and the lack of non-linearity may lead to less robustness of the model.

Response: Thank you for your feedback. We would like to emphasize that our paper investigates the numerical properties of neural networks where the architecture can include more than one layer. While there is no universally agreed-upon threshold defining the transition from shallow to deep learning, the term "deep" network is commonly used to describe neural networks with more than one hidden layer^{1,2}. In our manuscript, we discuss selecting the number of hidden layers from {1, 2, 3, 4, 5} in the tuning section, which justifies our use of the term "deep" learning. However, we acknowledge the concern regarding the use of only ReLU and leaky ReLU activation functions, which may limit the non-linearity and reduce the robustness of the model. In response to this point, we have enhanced our tuning procedure to include additional activation functions, specifically Sigmoid and Tanh. We have incorporated these changes in the tuning section of revised manuscript (Line 308).

2. Though the model structure had different configurations (e.g., different number of layers, different learning rate), the author failed to provide further details regarding how the model performed under different settings, thus it is hard to tell if the proposed model overfitted on the dataset consisting of 199 samples. Also, details regarding the implementation were missing. For example, if the learning being adjusted during training?

Response: Thank you for your insightful suggestions. We have expanded the section of Tuning to include in-depth information about the CV process used for hyperparameter optimization. This includes a detailed description of how we selected the optimal learning rate and other critical parameters. Additionally, we conducted two new simulation studies to further evaluate the model's performance across different sample sizes (400 and 1000), which more closely mirror real-world clinical trial data (Line 212-234). These three simulations were designed to address scenarios where the ground truth is unknown, as in clinical settings, and to assess whether the model can generalize without overfitting. The results demonstrated that our model performs well across all scenarios and mitigates overfitting through the proposed CV approach. Further details can be found in the revised manuscript (Line 282-301).

3. The authors predefined a few potential biomarkers and didn't analyze if the biomarkers are inner-correlated. Thus, the proposed model may be biased when selecting the "important" biomarkers.

Response: Thank you for your valuable suggestion. In light of your comments, we have redesigned our simulation studies to assess the robustness of our model in the presence of inter-correlated biomarkers. Specifically, we generated biomarker data, X , from a multivariate normal distribution with a mean of 0, a variance of 2, and pairwise correlations of 0.2 among the biomarkers. This design allows us to evaluate whether our model can accurately identify the true biomarkers under conditions where they are interrelated. Furthermore, we have enhanced the manuscript by adding

two additional simulation studies using the same data generation scheme for X . The results from these simulations demonstrate that our model consistently identifies the true biomarkers across all scenarios, affirming its capability to remain unbiased in the selection of important biomarkers, even when they exhibit mutual correlations. We have incorporated these changes in the section of Simulation framework of revised manuscript (Line 216-218, 426-430)

4. The baseline selected were all non-deep learning approaches, The conclusions will be more convincing if their model is superior over other deep learning-based models. The author may consider to fine-tune some state-of-the-art deep learning-based classification models and compare with it.

Response: Thank you for your comment. We want to clarify that our proposed deep learning model represents one of the first efforts to simultaneously perform subgroup identification and predictive biomarker selection. This dual capability is a key novelty of our manuscript. While we recognize the value in comparing our model against other state-of-the-art deep learning models, our extensive literature search did not identify any existing deep learning-based models specifically designed for this dual purpose in subgroup analysis. The baseline models we selected for comparison are the most commonly used methods in subgroup analysis, and they provide a meaningful benchmark for evaluating our model's performance. These have been widely validated and accepted in the field. We acknowledge the value in comparing our model to other deep learning approaches and intend to explore this in future work as suitable models become available. We have incorporated this into discussion section of revised manuscript (Line 546-550)

5. Should the method section be placed before the acknowledgement?

Response: Thank you for your suggestion. We have moved the method section before results section (Line 100-315)

2. The authors added two simulations with data biomarker data, X , generated from a multivariate normal distribution with a mean of 0, a variance of 2, and pairwise correlations of 0.2 among the biomarkers. ---- This is a very neat dataset, while the real-world data sample may not follow such a nice distribution. I know it's hard to analyze the real-world clinical data, could the author at least discuss the gap between the theoretical simulation data and practical clinical data?

Response: Thank you for highlighting this important issue. We recognize that real-world clinical data often involve more intricate distributions (e.g., non-Gaussian behavior) and may contain missing values. To address this discrepancy and its implications for clinical practice, we have provided a detailed discussion of several approaches in the revised manuscript

1) Correlation analysis in real data

We performed a correlation analysis of our real clinical dataset and found that its covariate structure is considerably more intricate than our simulation design. Despite this complexity, our method successfully identified clinically meaningful biomarkers, which are consistent with regulatory documents such as the European Medicines Agency (EMA) Summary of Product Characteristics (SmPC) and the Canada Product Monograph. For instance:

EMA SmPC:

"In patients with at least a partial response to Humira 40 mg weekly at Week 12, the HiSCR rate at Week 36 was higher in patients who continued weekly Humira than in patients in whom dosing frequency was reduced or withdrawn."

Canada Product Monograph:

"In patients with at least a partial response ($\geq 25\%$ improvement in AN count) at Week 12, 57.1% achieved HiSCR at Week 24 with weekly dosing, compared to 51.4% on every-other-week dosing and 32.9% on placebo."

2) Clinical context and unknown subgroup labels

Real-world data inherently lack fully known subgroup labels, making it difficult to verify whether an identified biomarker is truly predictive. By collaborating closely with clinicians, we contextualize our model outputs within established medical knowledge.

In conclusion, real-world clinical data often require exploratory steps to better understand the dataset prior to modeling. More importantly, close collaboration with clinicians ensures that the identified biomarkers align with

evidence-based medical insights. Whenever feasible, validation on independent clinical datasets is also crucial to confirm the robustness of the results. Our findings indicate that the proposed method can still detect clinically relevant biomarkers, supported by references to regulatory documents and previous studies. We have incorporated these considerations into the revised manuscript (Line 523-535,566-574, 583-585).

3. *There might not be an existing deep learning-based biomarkers identification method to compare with. But DNN + casual inference is not a new idea in medical domain. Could the authors briefly discuss those? For example, Learning Counterfactual Representations for Estimating Individual Dose-Response Curves.*

Response: Thank you for your thoughtful comments. While DNN-based causal inference is indeed well-established, our primary contribution focuses on a deep learning framework specifically tailored for subgroup identification and predictive biomarker discovery—an aspect that remains comparatively underexplored.

The referenced work, “*Learning Counterfactual Representations for Estimating Individual Dose-Response Curves*”³, provides valuable insights on using DNNs to model outcomes (Y) under various treatments and dosages. In contrast, our approach focuses on simultaneously predicting the individual treatment rule (ITR), conceptualized through $\Delta Y = Y(1) - Y(0)$, enabling patient stratification and personalized treatment strategies. Furthermore, DeepRAB integrates feature-ranking mechanisms to identify predictive biomarkers, offering interpretability and actionable clinical insights, an aspect less emphasized in most DNN-based causal inference models.

We hope this clarification highlights the novelty of our framework and its importance in advancing DNN-causal inference for biomarker-driven precision medicine. We have added this discussion to the introduction section of revised manuscript (Line 71-73).

Reviewer #3 (Remarks to the Author):

The revised manuscript has significant improvement in quality, with added details on the neural network, training, and testing procedures, as well as the inclusion of more scenarios for evaluation. However, there remain several concerns and questions that need to be addressed, as outlined below:

1. Clarification on AUC Calculation

The manuscript uses AUC (Area Under the Curve) to measure the efficacy of subgroup identification but does not provide sufficient details on its calculation. AUC is typically derived from the ROC curve, which evaluates a spectrum of classifiers. However, for a fixed classifier, only a single (true positive rate, false positive rate) pair exists. Thus, it is unclear how AUC was calculated in this context. Presumably, the spectrum of classifiers for DeepRAB is based on $\hat{\theta}$, while for other methods, it may rely on the estimated treatment effect. However, methods designed for subgroup identification may not necessarily excel in terms of AUC, as AUC does not directly assess subgroup identification. To ensure appropriate evaluation, the manuscript should report false positive rates, false negative rates, and variance across different test set trials for the classifiers' vanilla forms (e.g., thresholding at 0 for $\hat{\theta}$ in DeepRAB).

Response: Thank you for these valuable comments. We have provided more detailed information regarding the calculation of AUC in subgroup identification.

In our simulation design, we know the true subgroup labels, when label = 1 if $Y(1) > Y(0)$ and label = 0 otherwise. Our model outputs $\hat{f}(X)$, which we interpret as a score indicating whether a patient is likely to belong to the treatment-benefiting subgroup. By varying the threshold on $\hat{f}(X)$ across its possible values, we obtain different pairs of true positive rates (TPR) and false positive rates (FPR), thereby generating an ROC curve. We then compute the AUC by integrating under this curve, in line with the evaluation strategies outlined in prior studies^{4,5}.

Since XGBoostML also incorporates an A-learning function into its loss, it follows the same approach as DeepRAB for computing the AUC. For other subgroup identification methods, the output often takes the form of an estimated treatment effect. To compute the AUC in a similar manner, we shift the decision boundary (i.e., the cutoff on estimated treatment effect) and record the corresponding true positive and false positive rates against the ground-truth subgroup labels. This process yields an ROC curve for each method, from which we derive the AUC. It is important to note that there is no requirement to always use a threshold of zero for the estimated treatment effect. While metrics such as True Positive Rate (TPR) and False Positive Rate (FPR) are valuable for classification tasks, there is no established rule that the threshold of zero for $\hat{f}(X)$ in DeepRAB and XGboostML, or for the estimated treatment effect in other models, is the best choice. Instead, summarizing performance across all potential thresholds using the AUC remains a widely accepted approach to evaluate model effectiveness without relying on a single decision threshold. We hope this clarifies our approach to AUC calculation and our rationale for comparing different methods on an ROC curve. We have incorporated these modifications into the revised manuscript (Line 252-265).

2. AUC in Null Scenarios

• For null scenarios, the true positive rate is undefined since the denominator (ground truth positives) is zero. Consequently, neither the ROC curve nor the AUC is well-defined in such cases. The manuscript needs to clarify how AUC values were computed under these circumstances.

Response: Thank you for raising this concern. We acknowledge that in a strict null scenario, the true positive rate is undefined. However, in our approach to null scenarios, the true subgroup does exist, but predictive biomarkers are absent ($\beta = 0$). We employ a ground truth by labeling a sample as positive ($\text{label}_i = 1$) if $Y_i(1) > Y_i(0)$ and negative $\text{label}_i = 0$ otherwise. Both $Y_i(1)$ and $Y_i(0)$ include the same prognostic term plus independent $N(0,1)$ error term, so $\Delta Y = Y_i(1) - Y_i(0)$ is purely random.

Although this labeling is random, it allows us to compute an ROC curve and AUC. Importantly, this approach acts as a safeguard against overfitting: if a method were to yield an AUC substantially greater than 0.5 in this scenario, it would suggest a spurious or biased detection of a subgroup. We have clarified this procedure in the revised manuscript (Line 441-444).

3. Interpretation of Results in Simulation I and II

• In the section on comparative performance in Simulation I and II, the authors suggest that baseline models' high biomarker detection rates and inferior AUC performance may indicate overfitting. This interpretation is not convincing, especially since Simulation I involves a linear setting with 10 features and a sample size of 1000, where linear-model-based methods are unlikely to overfit.

• This empirical phenomenon might instead highlight issues with either the AUC calculation or its appropriateness as a criterion for evaluating subgroup identification, reinforcing the concerns raised in Points 1 and 2.

Response: Thank you for highlighting this concern. As noted in our earlier explanation, we use AUC to measure subgroup classification performance consistent with prior studies^{4,5}. In general, disparity between high detection rates during training and lower AUC in testing data can indicate overfitting. However, upon re-examining our own results and wording, we agree that labeling this phenomenon as “overfitting” is not persuasive in our specific context, particularly since linear-model-based methods achieve robust performance in a relatively simple setting.

We observed that although all models demonstrate high detection rates for key biomarkers (e.g., DeepRAB at 98% on X_1 and 100% on X_2 100%, while other methods nearly reach 100% on both), yet the median AUC on the testing data is only slightly higher for DeepRAB. To avoid confusion, we have revised our discussion and removed explicit references to overfitting. Instead, we emphasize that while all models exhibit robust performance, DeepRAB achieves a marginally stronger median AUC, pointing to its effectiveness without implying overfitting. We have incorporated these modifications into the revised manuscript (Line 367-368).

4. Specifying the Number of Biomarkers

- *The structure of the neural network requires pre-specifying the number of biomarkers to be selected, but this information is typically unknown in practice. How should practitioners determine the number of biomarkers, and what are the consequences of incorrect specification?*

- *Additionally, other methods do not rely on knowing the number of biomarkers. Would evaluating under the case where the number of biomarkers is known introduce an unfair advantage for DeepRAB?*

Response: Thank you for this thoughtful question. As clarified in the manuscript, “we set the predefined number of biomarkers (k) equal to the total number of covariates (p)”, allowing the network to assign an importance ranking to all features. Hence, all methods do not strictly require prior knowledge of the exact number of biomarkers, they do rely on an understanding of how many biomarkers are truly relevant.

For the simulation studies, we selected the top two ranked biomarkers for each method, consistent with the knowledge that two biomarkers were truly relevant. In practice, however, the optimal number of biomarkers is rarely known in advance, and choosing the top M features may not be sufficient in real-world clinical contexts. Close collaboration with medical professionals is essential to determine whether an identified set of biomarkers aligns with established clinical insights or not. More importantly, examining whether the findings are supported by regulatory documents provides further validation. For instance, our method successfully identified clinically meaningful biomarker (CHG_AN), which are consistent with regulatory documents such as the European Medicines Agency (EMA) Summary of Product Characteristics (SmPC) and the Canada Product Monograph.

EMA SmPC:

“In patients with at least a partial response to Humira 40 mg weekly at Week 12, the HiSCR rate at Week 36 was higher in patients who continued weekly Humira than in patients in whom dosing frequency was reduced or withdrawn.”

Canada Product Monograph:

“In patients with at least a partial response ($\geq 25\%$ improvement in AN count) at Week 12, 57.1% achieved HiSCR at Week 24 with weekly dosing, compared to 51.4% on every-other-week dosing and 32.9% on placebo.”

Determining the exact number of true biomarkers for real-world data lies beyond the scope of this study. Our primary objective is to demonstrate DeepRAB’s capacity to robustly rank candidate biomarkers and facilitate feature selection. We have elaborated on these considerations in the revised manuscript (see Line 523-535, 559-565).

5. Figure 8 (Importance of Biomarkers)

- *The x-axis in Figure 8 is not specified. The manuscript should explain how the importance of biomarkers is calculated.*

Response: Thank you for noting this. We have clarified in the revised manuscript how the biomarker importance is determined for each method (Line 267-271).

6. Duplicate Figure Labels

- *There are two figures labeled as Figure 8. This duplication should be corrected to avoid confusion.*

Response: Thank you for bringing this to our attention. We have corrected the duplication of figure labels (Line 591, 629)

Reviewer #4 (Remarks to the Author):

This study introduces DeepRAB, an innovative deep learning framework for precision medicine. It aims to identify patient subgroups and discover predictive biomarkers, addressing the need for personalized treatment strategies. The authors' responses to the initial queries have provided valuable clarification regarding the methodology and results, effectively addressing the main concerns raised. However, upon further consideration, several additional issues warrant discussion to ensure a comprehensive understanding of the study's implications and limitations.

1. The quantity of generated data ($N = 1,000$) in the simulation scenarios is insufficient to adequately train or validate the effectiveness of the proposed method. Deep learning approaches typically require training datasets comprising substantially more than one thousand samples. Furthermore, the simplistic nature of the data generation mechanism may be easily learned or comprehended by neural networks, given their robust fitting capabilities. As both the training and test data are derived from the same mechanism, the model may exhibit artificially high performance on the test set due to overfitting on the implicit data generation mechanism. This implicit mechanism may diverge significantly from clinical scenarios, potentially leading to the aforementioned generalization issues.

Consequently, to further validate the model, it is imperative to utilize a larger sample size and implement more complex generation mechanisms. Of particular importance is the need to generate training and test samples using different mechanisms (equations) to simulate the discrepancies between simulated and real-world data. This approach would more accurately reflect the heterogeneity encountered in practical applications and provide a more robust assessment of the model's generalizability.

Response: Thank you for this valuable feedback. As clinical trials often enroll on the order of 100–300 participants for phase 2 studies and a few hundred to around 1,000 participants for phase 3⁸, we believe that simulating sample sizes of 400 and 1,000 aligns well with many real-world scenarios. In response to your earlier comment, we introduced a simulation with 400 participants to reflect smaller trials. Second, based on other reviewers' earlier feedback, we have included two additional simulations of varying complexity. Third, our study focuses on a single disease context, rather than applying the model across unrelated diseases (e.g., using liver cancer data and then evaluating the performance on lung cancer). Together, we believe that these three simulations capture the range of participant counts commonly seen in clinical research.

Notably, in our real data analysis, our results are consistent with findings from prior studies and demonstrate that our model successfully identifies CHG_{AN} as a clinically meaningful biomarker. This aligns not only with baseline methods but also with regulatory documents, including the European Medicines Agency (EMA) Summary of Product Characteristics (SmPC) and the Canada Product Monograph.

EMA SmPC:

“In patients with at least a partial response to Humira 40 mg weekly at Week 12, the HiSCR rate at Week 36 was higher in patients who continued weekly Humira than in patients in whom dosing frequency was reduced or withdrawn.”

Canada Product Monograph:

“In patients with at least a partial response ($\geq 25\%$ improvement in AN count) at Week 12, 57.1% achieved HiSCR at Week 24 with weekly dosing, compared to 51.4% on every-other-week dosing and 32.9% on placebo.”

We recognize that generating training and test samples using different mechanisms could better address the discrepancies between simulated and real-world data. However, applying our proposed method to data generated through differing mechanisms would also require advanced techniques such as domain randomization⁶. This falls outside the scope of our current work, as our primary objective is to identify subgroups and predictive biomarkers on a single disease context. Given the strong performance demonstrated in our simulations and the alignment of our real-world findings with regulatory guidance from the EMA SmPC and Canada Product Monograph, we believe our model exhibits robust performance under these constraints. Nonetheless, we have broadened our discussion in the revised manuscript to underscore the potential benefits of introducing separate data-generation mechanisms. (see Line 523-535,611-615).

2. The methodologies employed in the comparative analysis primarily consist of relatively early approaches, such as CF (2018), LRMO (2018), and Concrete Autoencoder (2019). Given the rapid advancements in the field, it is pertinent to inquire whether more recent and potentially superior techniques have emerged in the intervening years. If indeed such developments have occurred, it may be necessary to incorporate contemporary methodologies into the experimental design to ensure a more comprehensive and equitable comparison. This inclusion would not only

enhance the robustness of the study but also provide a more accurate representation of the current state-of-the-art in the field.

Response: Thank you for bringing this to our attention. We have explored an extended application of the Concrete Autoencoder for GWAS in 2022⁷, but we have not identified other widely adopted or substantially improved versions in the recent literature. We agree that developing advanced models would be valuable, and we have discussed these possibilities in our revised manuscript's discussion section (Line 608-611).

Reference

1. Schmidhuber J. Deep learning in neural networks: An overview. *Neural networks* **61**, 85-117 (2015).
2. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. *Neural computation* **18**, 1527-1554 (2006).
3. Schwab P, Linhardt L, Bauer S, Buhmann JM, Karlen W. Learning counterfactual representations for estimating individual dose-response curves. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2020).
4. Chen S, Tian L, Cai T, Yu M. A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics* **73**, 1199-1209 (2017).
5. Huang X, Tian L, Sun Y, Chatterjee S, Devanarayan V. Predictive signature development based on maximizing the area between receiver operating characteristic curves. *Statistics in Medicine* **41**, 5242-5257 (2022).
6. Tobin J, Fong R, Ray A, Schneider J, Zaremba W, Abbeel P. Domain randomization for transferring deep neural networks from simulation to the real world. In: *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE (2017).
7. Liu Z, Dai W, Wang S, Yao Y, Zhang H. Deep learning identified genetic variants for COVID - 19 - related mortality among 28,097 affected cases in UK Biobank. *Genetic Epidemiology* **47**, 215-230 (2023).
8. National Institute on Aging. (n.d.). What are clinical trials and studies? National Institutes of Health. Retrieved January 21, 2025, from <https://www.nia.nih.gov/health/clinical-trials-and-studies/what-are-clinical-trials-and-studies#:~:text=A%20Phase%20trial%20tests,whether%20a%20drug%20is%20effective>

Dear Editors,

Thanks for the constructive feedback and encouragement provided by you and the referees. Below, we address each referee's comments with the comments *italicized* and our responses following in standard font.

Reviewer #2 (Remarks to the Author):

Thank you for adding more simulations which I concerned the most from the last two rounds reviews. I have no further questions for the authors.

Response: Thank you for your thoughtful comments throughout the revision process and for taking the time to review our work.

Reviewer #3 (Remarks to the Author):

I do not have further questions.

Response: Thank you for your thoughtful comments throughout the revision process and for taking the time to review our work.