We thank all reviewers for your insightful and constructive comments. Please see our responses below.

[Reviewer #1] Please see our responses to your main questions below.

- ***Well-identified subgroups***: Theorem 2 guarantees that confidence intervals (CIs) will achieve the required finite sample coverage for the ITE estimates in each subgroup, regardless of how accurate or inaccurate the underlying ITE model is. If the CIs exhibit large overlap across the constructed subgroups (because, for example, the ITE estimator was inaccurate due to covariate imbalance), we can conclude that the constructed subgroups are not well-identified. Conversely, if the CIs have little or no overlap across subgroups, we can conclude that the subgroups are well-identified. Given the theoretical guarantee of the CIs in R2P, the subgroups are robust if well-identified. This is an important advantage of R2P, which we will highlight more clearly in the revised paper.
- ***Why not relying on a specific ITE estimator is better***: In recent years, many ITE estimators have been proposed. However, no one ITE estimator is consistently the best in all settings. Furthermore, these ITE estimators are non-interpretable black-box models. One of the main contributions of R2P is that it divides units into subgroups with respect to an interpretable tree-structure, and provides subgroup coverage guarantees for the ITE estimates in each subgroup. R2P can be combined with any existing ITE estimator, enabling it to play a vital role in producing trustworthy and interpretable ITE estimates in practice.

[Reviewer #2] R2P produces confidence intervals that achieve the required coverage guarantee for the ITE estimates in each subgroup with respect to an interpretable tree-structure (Theorem 2). This provides upper and lower bounds for the ITE within each subgroup. Taken together, the coverage guarantee and interpretability of R2P can guide the user to develop more effective interventions and/or improve the design of further experiments. We will demonstrate this point clearly in the revised paper, leveraging the superior empirical results of R2P compared to previous methods.

[Reviewer #3] We have summarized your main questions and provided our responses below.

- ***Baselines with powerful ITE estimators***: Grouping the units based on the quantiles of the estimated ITEs fails to satisfy the essential requirement of subgroup analysis: interpretability. The estimates from a black-box ITE estimator are non-interpretable. Similarly, the subgroups defined based on the estimated quantiles give no explanation (in terms of input covariates) regarding why the units are assigned to a particular subgroup. Previous state-of-the-art subgroup analysis methods are all interpretable but are tied to one particular estimator: decision tree. While compatible with any black-box ITE estimators, R2P constructs easy-to-interpret subgroups based on the tree-structure and partition rules of the covariates. In addition, the confidence intervals of R2P achieve coverage guarantees for the ITE estimates in each subgroup. We agree that the performance improvement of R2P comes both from a better way to construct subgroups *and* the use of a better estimator. However, this is one of the key advantages of R2P: it is able to use *any* ITE estimator. In the revised paper, we will replicate the same experiment for R2P using *different* ITE estimators; this should give insight into the source of gain.
- ***Table 2 of the paper, “Normalized V^{in}”***: In reporting normalized comparisons across methods, we normalize V^{in} by dividing by V^{pop}, the variance within the entire population. Because the normalizer V^{pop} is the same for all methods and R2P achieves the smallest V^{in} (Table 1 of the paper), R2P also achieves the best normalized V^{in} (Table 2 of the paper). We will clarify this in the revised paper.

Table R1: Average overlap across subgroups on Synthetic dataset B.

R2P	CCT	CT-A	CT-H	CT-L
0.14±.03	0.63±.15	0.44±.09	0.60±.16	2.27±.55

- ***False Discovery***: There is no perfect performance metric for subgroup analysis. The optimal ground-truth of subgroups depends on multiple objectives, including homogeneity, heterogeneity, and the number of subgroups. In the literature, the usual metric used is variance, rather than ground-truth, because greater heterogeneity across subgroups and homogeneity within each subgroup generally imply well-discriminated subgroups. As one metric for evaluating false discovery, we can use the overlap of treatment effects across subgroups, as in Fig. 3 of the paper. For this, we suggest *average overlap of treatment effects across subgroups* over 50 simulations. Table R1 here shows that R2P performs best for Synthetic Dataset B. We would also like to direct you to our response to Reviewer 1 (well-identified subgroups), in which we highlight how the confidence intervals in R2P can help avoid false discovery.
- ***Questions (1-3)***: **(1)** We will add brief explanations regarding datasets A and B in the main text, as per the reviewer's suggestions. **(2)** The hyperparameter λ balances the impact of homogeneity and the width of the confidence intervals, while γ controls regularization. (Experiments in the Supplementary Material demonstrate the impact of λ and γ.) The choice of λ should be made according to the user's prioritization of performance metrics (e.g., V^{in}, V^{across}, and the width of confidence intervals). Alternatively, given performance metrics, both λ and γ can be tuned via cross-validation. **(3)** RMSE is the appropriate metric to evaluate an ITE estimator. However, R2P is not a method for ITE estimation and should not be evaluated on that basis; it is a method for subgroup analysis and should be evaluated as such.
- ***Additional feedback***: (*Question on lines 200-201*) Smaller subgroups mean smaller sample sizes; smaller sample sizes lead to wider confidence intervals (for example, the confidence interval for a sample of size 1 would be infinite). Treatment effects across identified subgroups can be compared via V^{across}. (*Question on Figure 3*) The average overlap in Table R1 shows that the subgroups identified by R2P are well-discriminated over 50 simulations.