We would like to thank the reviewers for their constructive feedback. We will correct the misprints raised and include the suggestions for improving the readability of the paper accordingly.

1 Reviewer #1

The experiment section is short. We will optimize the space and move some information about the experimental setup from Appendices A.5 and A.6 to Section 5 of the main paper.

Releasing an implementation. We will definitely release the code that was provided in the supplementary materials accompanying this submission.

2 Reviewer #2

The experiments are brief. We will add more information about the experimental setup as well as the interpretation of the results from Appendices A.5 and A.6 to Section 5 of the main paper.

Worst-case complexity of Algorithm 1 and runtime in Figure 7. In the worst-case, the complexity of Algorithm 1 still grows exponentially. This is a common issue in other parametric programming applications such as regularization paths. However, fortunately, it has been well-recognized that this worst case rarely happens in practice. In Figure 7, the runtime is almost linear thanks to the efficiency of the pruning lemmas. However, in the worst-case—which rarely happens in practice—our algorithm still has quadratic complexity.

Notations are complicated. Thanks for the suggestion to simplify the notation as well as the recommendation of providing description in language. We will do our best to improve the readability of the paper.

Relation to prior work. In several previous studies, similar arguments to Lemmas 1-3 have been made for a single specific value of z. However, the main difference in our work is that our Lemmas 1-3 are for all $z \in \mathbb{R}$, which is technically much more challenging and requires parametric programming techniques to conduct the pruning efficiently. Regarding the related works on Bayesian change point detection, we thank the reviewer for the suggestions. We will provide some appropriate discussions in the related work part in Section 1 of the paper.

Additional suggestions from the reviewer. Regarding the caption of Figure 1, as the reviewer pointed out, it should be "(C, D, F)" instead of "(C, D)". In regards to the explanation of $\Delta\mu$ in Figure 4 and 5, we will move the explanation from Appendix A.5 to the main paper. In terms of referring the reader to some review papers, we have already cited Truong et al (2019). We will also refer to Aminikhanghahi & Cook (2017) and Van den Burg & Williams (2020).

Broader impact statement. We are not saying that we solely rely on p-values for making medical decisions. With the valid p-values that we introduced in the paper, it is guaranteed that the probability of making false decisions is properly controlled, and valid p-values can be used as one of many other possible criteria for making medical decisions.

3 Reviewer #3

Dense writing. We will simplify the writting as much as possible.

Computation time is reported, but only on itself, not in comparison to other methods. The baseline methods are faster than the proposed method, but they have much lower power than the proposed method. Besides, SMUCE—one of the baseline methods—is only asymptotic and can not theoretically guarantee to properly control the false positive rate (FPR). The reason why we include the computational time of the proposed method in the paper is that we want to demonstrate our method not only has higher power but also is practically useful. However, based on the suggestion of the reviewer, we will also consider providing the computational time of the baseline methods in Section 5 or in Appendix A.5 of the revised version.

4 Reviewer #4

Real world datasets with ground-truth. Thanks for sharing with us a selection of real world datasets with ground-truth. We will investigate and apply our method to them.

Suggestion of adding one more internal comparison partner. Thanks for the suggestion. We will clearly explain how we can correct the multiplicity from the selective p-values. Actually, we have already considered Bonferroni correction in Figure 1. Once the selective p-values are obtained, we can treat these as nominal p-values, and any multiple testing adjustment such as Bonferroni correction can be applied as long as they are nominal p-values [4].