Reviewer 2: We thank reviewer 2 for their positive comments and appreciate the helpful errata spotted. We agree the "Additional feedback" discussion is beyond the scope but look forward to having this important discussion.

Reviewer 3: "So in summary, the point of the paper is the improve runtime with an approximate algorithm." We strongly disagree with this casting, and the hope that the reviewer might reconsider in light of our comments. Improving runtime is not the key focus of the paper. There currently exists no algorithm to compute MCR for random forests whatsoever (and so nothing to improve runtime or accuracy over!). The approximations made in the technique we introduce are there so that calculation of MCR across a Rashomon set of RF models is physically tractable for the first time (along with accompanying proofs). This is the key contribution.
"The primary problem I have with this paper is that it does not compare with a good baseline.": We also disagree with this statement. The suggested MCR+ "rough equivalent" is inherently flawed and does not provide a sound measure of the upper bound of a variable when interactions exist amongst the input variables. E.g, in the synthetic XOR example each variable by itself predicts no better than chance, the same as a constant predictor. Therefore, the "rough equivalent MCR+" would return all zeros where the known truth is all 0.5 (see paper Fig 1). A method to measure upper bounds of variable importances in real world situations, which typically do contain variable interactions, is a key contribution of the work. More subtly the suggested MCR- baseline, hold-one-out VI, also does not offer a MCR baseline. As we note (although clearly opaquely) on lines 92-94 algorithm reliance methods (holding out variables) are not only a function of models that fit the data well while MCR methods are. The meaning and implication of this difference and why it matters (and therefore why hold-one-out should not be an MCR baseline) is strongly made in Fisher et al.'s (paper [7]) establishment of MCR in their 60-page JML paper both theoretically (§3.2) and empirically (§9.1). While a full discussion remains outside of the scope of this paper, given length constraints, we propose the inclusion of a slightly longer description of their argument to make it clearer why this should not be considered a baseline.

Reviewer 4 and 5 question whether the found "equivalent" models will retain the same predictive performance on new data as the reference model (RM), i.e. remain in the $\epsilon-$Rashomon set around the RM as we tend to the population. We agree this is an extremely interesting issue. However, the proposed method is based (inc. proofs and empirical evidence for MCR convergence) on constructing $\epsilon-$Rashomon sets ($\epsilon-$sets) based on in-sample (fit) equivalence. As correctly pointed out by the reviewers, making the claim that one can find $\epsilon-$sets based on generalized performance equivalence would require significantly more theoretical and empirical investigations. Its inclusion, in addition to developing and evaluating the proposed approach, however would take the work well beyond the length and scope possible in a NeurIPS paper. Therefore, we wish to note: (1) in general this is an open problem for MCR with Fisher et. al not fully addressing this issue but rather providing a proof (valid for our work) that in-sample estimation is sufficient under large sample sizes and a reference model with correctly selected complexity (their §4.1). (2) What is proposed corresponds to the real world use case where MCR would be computed based on a reference model trained on the full dataset as part of a fit(via CV to determine model complexity)-refit(on full data) methodology to model building. Notably, this is the underpinning use-case when the use of training data is motivated (instead of a test set) for computing traditional permutation importance (c.f. Interpretable Machine Learning by Molnar, 2020). (3) While we realise that our technique focuses on fit $\epsilon-$sets equivalence that we strongly believe this still makes an important statistical contribution in the road towards MCR. We thank the reviewers for highlighting this need for clarity and the importance of this discussion. We will include this clarification and discussion if accepted.

Reviewer 4 suggests the inclusion of (1) an ablation experiment to understand the importance of the two steps (2) further simulation studies With regard to (1) we have these results, as we did examined them ourselves, and simply left them out due to space. If accepted we will inject these as part of our additional page allowance. In brief: surrogates account for a change in the permutation importance by 0-13% while the majority (remainder) of the change comes from the second transform. With regards to (2), further simulation studies were run as part of the work and we agree that these provide additional insights. However, we feel that their inclusion would not provide significant additional insight worthy of the removal of other analysis/points given the available space. Finally, with respect to boosted trees - developing an MCR method for this class is of interest and part of our future work and something we're interested in discussing. Unfortunately, the approach does not directly transfer and is outside the scope of this work.

Reviewer 5: We agree the comparison to SHAP is not entirely fair. Line 90-92 attempts to indicate this. We are happy to adjust/extend the wording to clarify the use case differences between SHAP and MCR. With regard to the discussion of the cancer results: The fact that MCR- is 0 for all variables indicates that, there is at least one model in the Rashomon set (set of equally performant models) that does not rely on this variable to make predictions (although other models do, as indicated by the non-zero MCR+) and that the set is non-trivially large. Therefore, in certain contexts, i.e. when fitting a model and undertaking VIM to consider potential causal factors, fitting a single model would not provide the full picture. We agree, however, that in some cases (as mentioned in the introduction) this doesn't matter. If accepted we will briefly extend the discussion of the cancer results to ensure the interpretation is clear. We thank Reviewer 5 (as with all other reviewers) for the additional errata and pointing out lines requiring minor clarification (i.e. line 273) which we agree to.