We thank the reviewers for their valuable input on how to improve our manuscript. We are heartened by the general consensus on the importance of the problem and the theoretical grounding of our method. The reviewers rightly indicate a need for a stronger motivating example. We will retain the example of college advising (lines 20-22), but **replace** the parole example with one from child abuse hotline screening. Call workers must decide which allegations coming in to the child abuse hotline should be investigated. The worker relies on (immediate risk) information communicated during the call and an algorithmic risk score that summarizes (longer term) risk based on historical administrative data. The call is recorded but is not used as a predictor for three reasons: 1) inadequacy of existing case management software to run speech/NLP models on calls in realtime; 2) model interpretability; 3) need to maintain distinction between immediate risk (as may be conveyed during the call) and longer-term risk the model seeks to estimate. (We will discuss the parole example in the Broader Impact section to highlight the need for caution in applying the methodology. The criminal justice system is complex and involves multiple decision-making points, presenting opportunities for misuse.)

• Using the extra page permitted for the camera-ready, we will include an empirical example on real child welfare data. We use our evaluation procedure (§4) since we will not have ground-truth outcomes. MSE with 95% confidence intervals (CI) for our DR method is $0.248(0.245, 0.250)$; PL: $0.249(0.246, 0.251)$; TCR: $0.265(0.262, 0.269)$. This suggests the PL/DR models could help workers better identify at-risk children.

• Our characterization of TCR as standard practice was informed by discussions with government agencies about their decision support systems. As R3 observes, one might consider feature imputation, but imputation may not desirable or feasible: e.g., using *imputed* protected attributes may still be impermissible; and the speech (call) data in the child welfare call is too high-dimensional to impute. Imputation using the predictors will also estimate the same biased target as TCR since $\mathbb{E}(Y^a \mid V) = \mathbb{E}(Y^a \mid V, f(V))$. The revision will provide this discussion with relevant citations.

• We would like to clarify that Theorem 3.1 describes the conditions under which our method is optimal. Theorem 3.1 decomposes the error of DR prediction method into the error of an oracle with with access to the true nuisance functions and a product of nuisance terms, which if small enough, imply that the DR method achieves the same error rate as an oracle and therefore inherits the optimality (such as minimax) of the oracle. We will elaborate on this in the revision and provide examples for clarity. Reviewers noted that while we observed the expected behavior for LASSO, random forests (RFs) showed TCR performing on par or better than PL and DR. For our sparse linear data, RFs have higher error than LASSO (compare Fig. 1a to 1c). The RF estimation error dominates the confounding error. This example shows that depending on the context and modeling choices, TCR may outperform the counterfactually valid approaches. Our evaluation procedure (Alg. 5) can assess this for a given setting. We will add this to the discussion in lines 222-226.

R1 We thank R1 for the useful suggestions. For the MSE estimator proposed in §4, we provided empirical results in §C.3 that assessed how often it identifies the best model. The revision will include coverage results: On 100 simulations for a test size of 1000, the MSE estimator with 95% CI covered the true MSE 94 times for the DR approach and 93 times for the PL. We will expand Broader Impact to include examples that show when this method should *not* be used.

R2 We appreciate R2's detailed comments and will make all clarifications suggested. R2 raises a good question about our interpretation of Corollary 3.1. For a simple example, consider the case where $k_\nu \approx k_\mu \approx k_\pi$. When $d_v \ll d$, the second term of the PL bound dominates the error whereas the first term of the DR bound dominates in high-dimensional settings. The revision will clarify this. We thank R2 for noting the comprehensiveness of our empirical analysis.

R3 We thank R3 for noting that our proposed approach is more general than our articulated motivation suggests. The revision will describe how our approach is applicable to settings such as selective labels. While selection bias literature such as survey inference use doubly-robust approaches, existing theory does not cover the prediction setting. We hope that our theoretical contributions help fill this gap. R3 aptly identifies the connection between causal inference and missing data. Indeed the fundamental problem of causal inference is one of missing data: that we only observe one potential outcome (Holland 1986). In our setting, we have additional missingness in the features available for prediction, features which crucially affect the decisions and hence are confounders. We believe the missing confouders problem merits special attention. For the case in which the missing features only affect the outcome, with infinite data regressing the outcome on the available predictors is optimal. This is not true for our setting, where the confounding bias persists (Prop 3.2). We will include this discussion in a paragraph that relates our work to privileged learning.

We were glad to see R3's interest in the optimality of our method and hope our discussion above is useful. We used simulations to explore whether our method performs well without sample-splitting, using the **full** training sample (lines 202-206). The methods performed as expected for LASSO, even though our theory as presented does not cover this setting. Theory that does not rely on sample-splitting typically requires strong empirical process assumptions. We would like to clarify that the experiments held sample size fixed (line 211).

R4 We thank R4 for finding our contributions valuable and for suggesting a comparison to a model of decisions. A model of child welfare screening decisions achieves an estimated MSE for the adverse outcome of $0.356(0.353, 0.359)$. We will include this comparison to our method (see above) to demonstrate the value in predicting downstream outcomes.