Thanks to the reviewers for the helpful comments. We'll address them all in the paper & answer the key questions here.

Single-decision confounding (R1, R4) We completely agree that multi-decision confounding is also important, but note our model is already more realistic than sequential ignorability, which is assumed in almost all existing work on OPE. Under our model, an unobserved confounder can affect rewards or state transitions in other time steps, but only directly affect the action in a single time step. Since the state transition and action taken reveals new information, it is sometimes reasonable to assume that an unobserved confounder in one time step is implicitly observed in the next. We provide several sample scenarios where single decision of confounding is reasonable, including evaluation of automated sepsis treatment policies. If OPE is robust to confounding at multiple times, it should be robust to confounding in any one time, which we can check with our method. Therefore, our method provides nontrivial necessary conditions for reliable OPE estimates, and is a step towards ensuring more robust OPE, as we will clarify in the text.

Testing assumptions (R1) Testing assumptions is important, and some of the standard assumptions we make such as overlap can be tested statistically. However, sequential ignorability (SI) cannot be tested from off-policy data, and our method precisely allows analyzing sensitivity to violations to this untestable assumption. As such, the relaxations of SI we consider (bounded confounding in Assumption D, E) can be thought of as the extent to which SI is violated.

Looseness (R1): Our bound is tight when the final decision is confounded; assessing conditions for tightness in earlier timesteps appears challenging, and is left to future work. We will add this discussion in the final version.

Assump. D and relation to Z & B 2019 (R1) We briefly discussed this in Lines 76-80 and will highlight it further.

Flexibility of loss minimization and simplifying computation (R2, R4): While our notation is cumbersome due to the complexity of the sequential setting, our procedure is ultimately a loss minimization problem with importance sampling (IS) weights. We can minimize the loss using any model class (e.g. deep nets; not just linear models), with standard ML training algorithms. We only use the linearity assumption in our proof to guarantee statistical consistency. Consistency would hold with many other ML models, but we use a linear model for simpler exposition. The naïve bound in the supplement are simpler to compute, but are often too conservative.

Extending to multi-step confounding (R1, R2). When multiple decisions are confounded, we can still write down an optimization problem over likelihood ratios corresponding to each confounded decision. However, both the objective and equality constraints are nonconvex, and the problem is over infinite dimensional likelihood ratios. Our approach of exploiting convex duality to derive the loss minimization problem does not generalize to the multi-step confounding case. Tractable approximations in the general case is a interesting future direction, which we will discuss in more detail.

Interpreting Γ (R2): As we discuss in lines 190-192, for binary actions, the confounding model we study has a crisp interpretation: $\log \Gamma$ bounds the difference in log odds for two individuals with identical observed states, but different unobserved confounders. Our model is a natural extension of this to multi-action, sequential settings. In practice, a meaningful assessment of the robustness of OPE is the threshold level of Γ at which the findings of OPE no longer holds (e.g. bound on evaluation policy's reward become worse than that of benchmark's); we will emphasize this more.

Relation to Yadlowsky et al. (R2): Addressing a sequence of actions requires considering their dependence, which the methods in Yadlowsky et al. don't do; See lines 86-8, 267-9. Solving these directly enables new multi-step applications.

Overlap and guidlines/protocols (R1, R4): As long as the action space is finite, overlap is a standard assumption, even in continuous state spaces; continuous actions are beyond our paper's scope (we will clarify this). Indeed, overlap is a requirement for any OPE algorithm that does not make structural modeling assumptions. It is an intuitive condition, because lack of overlap implies that there are no trajectories in the data that follow the actions we wish to evaluate. Note that any stochasticity that only affect the rewards through the decision maker's actions creates overlap.

Guidelines are usually based on strong evidence, and so safe policies to be evaluated should follow these guidelines as well, again implying overlap. A common source of random variation occurs when clinicians make different decisions within the guidelines. Important opportunities for policy improvement are where there are large variations in clinical practices, e.g. guidelines that defer to practitioners, or lack thereof. We are happy to clarify these points in the main text.

Seq. ignorable eval policy (R4): The reviewer's point is a good one, e.g. for evaluation of updated clinical recommendations. However, we focus on automated policies that can only use recorded inputs due to the importance of making credible claims about AI systems. Even when influencing human decision makers, it is problematic in some cases to recommend they continue using unobserved variables to influence decision; e.g. when clinicians may be biased or influenced by spurious correlations. Also, the distribution of confounding variables may also shift over time, between populations and providers; using only observed variables could provide more robust recommendations.

Related work and clarity of writing (R1): We agree that Section 4 was very dense. In the final version, we will expand discussion of theoretical results, implications of our model, and connection to other methods with the additional space. Also, we will include discussion of the suggested references on bandits in our related work.