We thank all reviewers (**R1**, **R2**, **R4**) for helpful and insightful comments!

Matching Algorithm (R1). One major concern for **R1** is that we do not provide an algorithm to find the optimal matching (an NP-hard problem). We use the state-of-the-art method, "PICEF", of [15] (Dickerson et al.) to find optimal matchings. PICEF takes only fractions of a second for realistic exchange graphs (including those in our experiment). Our proposed edge-query methods (`Greedy` and `MCTS`) both use PICEF as a low- or no-cost matching oracle. While this is a good assumption on realistic kidney exchange graphs, this is not true on all instances because that problem is NP-hard.

Benchmark Methods, and $\Delta^{\texttt{MAX}}$ (R1, R2). In short, $\Delta^{\texttt{MAX}}$ compares edge-query methods against the benchmark of max-weight matching (without edge queries). To clarify, we compare our methods against two benchmark policies: the max-weight matching (policy $M^{\texttt{MAX}}(\cdot)$, which does not consider edge uncertainty), and the stochastic matching policy of [16] (Dickerson et al.) (which maximizes expected matching weight). As the reviewers point out, these policies do not query edges before constructing the final matching—which is the present reality in all fielded kidney exchanges. We compare against these policies to illustrate the improvement due to querying edges; we agree that it is unfair to compare an edge-query method to policy $M^{\texttt{MAX}}(\cdot)$ (which does not query edges). In other words, we compare against these benchmark policies to (a) show the potential gain from querying edges, and (b) compare different edge-query policies to each other. In our final submission we will clarify which policies are benchmarks, and we will elaborate on our reasons for including them.

Random Performs Well (Table 1) (R1). We believe this comment is in reference to the column labelled P_{90} in Table 1. These number are the 90th percentile of $\Delta^{\texttt{MAX}}$ over all exchange graphs, meaning that Random gives this improvement (about 50%) in only 10% of all graphs. On the other hand, by looking at the column labelled P_{10} (the 10th percentile) `Greedy` and `MCTS` achieve a similar improvement (40%) for 90% of all graphs. In summary: Random works well on about 10% of all graphs, while `Greedy` and `MCTS` perform well on *most* graphs.

Complexity & Runtime Analysis (R1). Thanks for bringing this up. We will include the following: Our methods rely on an "oracle" to solve the NP-hard kidney exchange matching problem. We can report the *number of calls* to this oracle for each method as a measure of complexity. Both benchmark methods (max-weight matching and failure-aware [16]) as well as `IIAB` (Blum et al. [10]) use exactly one oracle call; i.e., they are $O(1)$. Both `Greedy` and `MCTS` use a fixed number of samples (M) to evaluate the objective of an edge set. `Greedy` evaluates the objective of an edge set exactly Γ times; thus, `Greedy` is $O(M \cdot \Gamma)$. Finally, `MCTS` can in theory visit all potential edge sets of size at most Γ (i.e., an exhaustive search), which is $O(M \cdot \sum_{\gamma=1}^{\Gamma} \binom{|E|}{\gamma})$. Since this version of `MCTS` is intractable in both runtime and memory, we impose reasonable limits (see § 3 of our paper).

Focusing on Edge Selection vs. Edge Failure Probabilities (R2). We agree that our results suggest that the edge query problem is easy in practice: perhaps our most important finding is that *in theory* edge selection is hard (both non-monotonic and non-submodular, as well as NP-hard), while *in practice* this problem is easy (a greedy heuristic is nearly optimal). Importantly, we see our paper as *complementary* to ongoing research on edge failure distributions. We are aware of that research, and it informs our own investigations. In summary, several research questions address edge-existence uncertainty in kidney exchange: one is *Which edges fail and why?* (**R2**'s suggestion); another is, *Given an edge-failure model, how do we improve the outcome of kidney exchange?* (the question we address in our paper). These are complementary questions that both improve kidney exchange.

Algorithmic contribution (R2). We focus on motivating and formalizing the problem of selecting edges to query–and we prove that this is not a theoretically easy task via Prop. 2.1 and 2.2. Yet, we demonstrate that simple algorithms work well on real data. We strongly believe that our contributions are novel, and will serve as a foundation for future work. Indeed, we are the first (to our knowledge) to formalize this edge selection problem, either in the single-stage or multi-stage setting. Further, we are the first to formalize and test simple algorithms for this setting; the only similar work is that of Blum et al. [10], which considers a less-realistic setting (discussed in §1 of our paper).

Clarity of Exposition (R1, R2, R4). We thank all reviewers for clarity suggestions. In particular, we will clarify our experimental sections (4.1 and 4.2) to better separate our experimental design from our results. Finally, we will clarify details of our algorithms in the main paper. Currently the appendix describes all algorithms in detail; we will move some important details to the paper, including the UCB bounds used by our `MCTS` method and a discussion of the main hyperparameters of `MCTS` and the values we chose. This will be possible because an extra page is allowed in the final version.