S1 Appendix. Bootstrapped inference procedure

We developed a bootstrapped inference procedure that allows us to formally test the difference in performance between sample size adjusted dedupe and three comparison algorithms: default dedupe, fastLink, and Name Match. To do so, we generate distributions of the difference in performance between relevant pairs of algorithms using the following bootstrap procedure.

We begin by bootstrapping the experimental dataset E 1,000 times. For each of the 1,000 bootstrap samples E_s, we compute the performance metrics of linking dataset E_s to the administrative dataset D using each algorithm (sample size adjusted dedupe, default dedupe, fastLink, and Name Match). Comparing the performance metrics of sample size adjusted dedupe to the performance metrics of the comparison algorithms for each bootstrap sample gives us the empirical distribution of the difference in performance between linking algorithms. We complete this process for each linking context considered, i.e., size of administrative database D and the label budget provided to dedupe.

These empirical distributions of performance differences allow us to measure statistical significance for three important questions:

- Does sample size adjusted dedupe perform better than default dedupe?
- Does sample size adjusted dedupe perform better than fastLink?
- Does Name Match perform better than sample size adjusted dedupe?

We consider the performance of "sample size"-adjusted dedupe to be significantly better than default dedupe if the difference in performance is less than zero at the 95th percentile of the empirical distribution for total error.

As described in Results section of the main manuscript, we ran each linking algorithm five times for each linking specification in order to understand the variability of the different linking tools on a given dataset. Because of this, we are able to generate 25 measures of significance for each combination of sample size adjusted dedupe and comparison algorithm. In other words, we can determine whether there is a significant difference in performance

between each of the five sample size adjusted dedupe runs and each of the five comparison algorithm runs (yielding 25 total significance measures). In Tables 2 and 3 of the main manuscript, we report the share of significance tests for which there was a significant difference in total error between sample size adjusted dedupe and comparison algorithms default dedupe, fastLink, and Name Match. The results of this bootstrapping procedure were also used to generate the 95% confidence intervals shown in Fig 3 of the main manuscript.