4) In lecture 1 we quickly outlined some of the big challenges to causal inference when using observational data (see slide 41, “There should be strong effort to show the two groups are similar…”). These challenges include: inclusion/exclusion of observations, observational units that may be completely missing (censored, survival bias), missing data, imbalances in observed data, and imbalances in unobserved data. We’ll address each of these at different points in the course. But let’s focus on the decision to include/exclude observations. What we’re doing when matching – i.e., removing observations that do not have adequate counterparts in the contrast group – may seem a bit subversive. The intuition is: why “throw away” data? I think there are two reasons people worry about “throwing away data.” First, it seems like limiting the kinds of observations in our study we may be losing the ability to generalize our conclusions to a wider swath of the population. The counter to that is: yes, we are trading off the ability to generalize (i.e., external validity) for the ability to make stronger claims about a candidate causal effect (i.e., internal validity). The second concern is that it seems like more data is better. Formulate a response to this concern. (Note: OMG, this question seems so nebulous. Yup. That’s how this works; you’re playing Big Kid academics now. We made sure to mention this argument during lecture 01, so you know it. It’s a common statistical argument nowadays. If you want to read your way out of this one… here’s a good paper.)

Answer: From this paper, section 8, page 232: “…as with all statistical methods, a bias-variance trade-off exists for matching. If we drop many observations during [matching], and balance is not substantially improved, the mean squared error (or other mean-variance summary) of the estimated causal effect might actually increase. Users must pay close attention to this trade-off during the process of matching, but unfortunately no precise rules exist for how to make these choices... Of course, dropping observations does not necessarily mean that [matching] is worse since improving balance can also increase efficiency, and in any event including imbalanced observations requiring extrapolation in a parametric analysis merely produces false precision. So although estimated standard errors may increase in some cases with [matching], they would likely be more accurate. Moreover, in many situations, eliminating observations far from the rest of the data as matching does will reduce heterogeneity and thereby further reduce variance.”