To enable personalized cancer treatment, machine learning models have been developed to predict drug response as a function of tumor and drug features. However, most algorithm development efforts have relied on cross-validation within a single study to assess model accuracy. While an essential first step, cross-validation within a biological data set typically provides an overly optimistic estimate of the prediction performance on independent test sets. To provide a more rigorous assessment of model generalizability between different studies, we use machine learning to analyze five publicly available cell line-based data sets: National Cancer Institute 60, ancer Therapeutics Response Portal (CTRP), Genomics of Drug Sensitivity in Cancer, Cancer Cell Line Encyclopedia and Genentech Cell Line Screening Initiative (gCSI). Based on observed experimental variability across studies, we explore estimates of prediction upper bounds. We report performance results of a variety of machine learning models, with a multitasking deep neural network achieving the best cross-study generalizability. By multiple measures, models trained on CTRP yield the most accurate predictions on the remaining testing data, and gCSI is the most predictable among the cell line data sets included in this study. With these experiments and further simulations on partial data, two lessons emerge: (1) differences in viability assays can limit model generalizability across studies and (2) drug diversity, more than tumor diversity, is crucial for raising model generalizability in preclinical screening.