In the quest for personalized cancer treatment, machine learning models have been developed to predict drug response based on tumor and drug features. However, the majority of algorithm development efforts have relied on cross-validation within a single study for assessing model accuracy. While cross-validation within a biological dataset is a crucial initial step, it often provides an overly optimistic estimate of prediction performance when applied to independent test sets. To address this limitation and offer a more rigorous assessment of model generalizability across different studies, machine learning is employed to analyze five publicly available cell line-based datasets: National Cancer Institute 60, Cancer Therapeutics Response Portal (CTRP), Genomics of Drug Sensitivity in Cancer, Cancer Cell Line Encyclopedia, and Genentech Cell Line Screening Initiative (gCSI). The analysis considers observed experimental variability across studies and explores estimates of prediction upper bounds. The study reports the performance results of various machine learning models, with a multitasking deep neural network demonstrating the best cross-study generalizability. Notably, models trained on CTRP yield the most accurate predictions on the remaining testing data, and gCSI stands out as the most predictable among the cell line datasets included in the study. Through experiments and simulations on partial data, two key lessons emerge: (1) differences in viability assays can limit model generalizability across studies, and (2) drug diversity, rather than tumor diversity, plays a crucial role in enhancing model generalizability in preclinical screening.