readme.txt for Budish, Eric; Roin, Benjamin N. and Williams, Heidi (forthcoming) Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials, American Economic Review.

----- Description of raw data files

Online Appendix B provides a detailed description of the datasets used in our paper; briefly, the datasets are:

(1) The Surveillance, Epidemiology, and End Results (SEER) data, compiled by the National Cancer Institute (NCI). We use both the SEER cancer registry data and the SEER population data. The raw TXT data files are available via a research data agreement; see https://www.seer.cancer.gov/seertrack/data/request/ for details.

(2) Year-age-gender-specific period life expectancy data for 1973-2006, drawn from the National Center for Health Statistics (NCHS) files posted at http://www.cdc.gov/nchs/products/life_tables.htm. For 2000-2006, digitized files are available from NCHS. For 1973-1999, the data was entered by the firm Digital Divide Data (http://www.digitaldividedata.org/) and was funded by NIA Grant Number T32-AG000186 to the NBER. No data is available for 1979 nor 1981.

(3) The Physician Data Query (PDQ) Cancer Clinical Trials Registry, compiled by the National Cancer Institute (NCI)'s cancer clinical trials database. The raw XML data files are available via a research licensing agreement; see http://www.cancer.gov/licensing for details.

(4) Web-scraped data on clinical trial lengths for the available subset of the PDQ clinical trials that are also registered in ClinicalTrials.gov, a service of the US National Institutes of Health (NIH).

(5) Data on the 71 US Food and Drug Administration (FDA) approved oncology drugs from 1990-2002, drawn from Johnson, Williams and Pazdur (2003). For 39 of these 71 drug approvals, we were able to hand-collect data on whether a surrogate endpoint was used, as well as the cancer and stage for which the drug was approved, from the Drugs@FDA database (http://www.accessdata.fda.gov/scripts/cder/drugsatfda/). Full reference for Johnson, Williams and Pazdur (2003): Johnson, John, Grant Williams, and Richard Pazdur, "End points and United States Food and Drug Administration approval of oncology drugs," Journal of Clinical Oncology, 2003, 21, 1404–1411.

Online Appendix B provides details on how we extracted, constructed, and combined these raw data files.

----- Datasets and programs used to estimate the final models

The datasets needed to replicate the final models in the paper are:
- cancer_stage_fda.dta
- cancer_stage_sponsor.dta
- cancer_stage.dta
- PDQ_s.dta
- PDQ_trials.dta
- SEER_cs_mortality.dta
- SEER_csy.dta
- stage.dta

The scripts needed to replicate the final models in the paper (which can be executed in Stata 13) are:

1) final_research_analysis.do:
replicates all estimates related to research and development (R&D) outcomes
2) final_survival_analysis.do:
replicates all estimates related to patient survival outcomes

These files should be executed from the same directory as the above list of .dta files. Three additional notes:
a) The log estimates referenced in footnote 37 are generated in final_research_analysis.do
b) The figure referenced in footnote 49 is generated in final_survival_analysis.do
c) Figure 6(a) is an illustration (not based on data) and hence is not generated by these files