This file provides information on the code and datasets used in the drug project-level phase-II development analyses. These include Figures 2, 3 and 4, Tables 6, 7, and Panel A of 8. Note that all datasets in the "Phase-II Development Tests" folder are pseudo data that include a randomized 1% of the original sample used in the corresponding analysis, and are provided for the purpose of illustrating data structure and testing the code and will not reproduce the results reported in the paper. Detailed information on the construction of the datasets is provided in the main paper and the Internet Appendix.

The files included in the "Phase-II Development" folder of the replication package are described as follows:

- "pseudo_phase2_main.dta" This pseudo data includes information on phase-II projects that experience BTD entry and those that do not. The data is used in the analyses of Figures 2, 3 and 4, and Table 6. The observational level is drug project-quarter. The variables in this data are described below.
- "pseudo_phase2_exantelevel.dta" This pseudo data includes information on a subsample of the phase-II projects in the sample above, however, in ex-ante leveled markets that have no approved products on the market and is used in the analysis of Table 7. The observational level is drug project-quarter. The variables in this data are described below.
- "pseudo_phase2_nevershocked.dta" This pseudo data includes information on phase-II projects owned by rivals that never experience BTD entry and is used in the analysis of Panel A of Table 8. The observational level is drug project-quarter. The variables in this data are described below.
- "Phase2_Dev - Replication Code.do" this file contains the code that replicates all the project-level phase-II development tests. Information on the sample and the statistical methods used in each analysis is provided below.

Common Information Across All (or several) Analyses

Relevant information on the sample, variables and statistical methods used in all (or several) analyses is presented here. Information that is exclusive to each Figure/Table is provided in the corresponding analysis section.

- Variable Definitions
    - Identifying Variables: The following variables are used in all, or most, of the analyses:
        - dp: drug project identifies
        - fqdate: the calendar quarterly date
        - realid: firm identifier
        - i4: the numerical identifier of the 2nd subchapter ICD-10 therapeutic market
        - i3: the numerical identifier of the 1st subchapter ICD-10 therapeutic market
    - Shock Variables:
        - btdshock: a dummy equal to one if a project is shocked by BTD entry, and zero otherwise.
        - sm: a dummy equal to one if a project is shocked by BTD entry, and the project uses a technology that is different from that used by the awarded drug.

        - smst: a dummy equal to one if a project is shocked by BTD entry, and the project uses the same technology as that used by the awarded drug.
    - Drug Characteristics
        - dp_grad: a dummy equal to one in the quarter that a phase-II project graduates to phase-III and equal to zero if it remains in phase-II development
        - phase2age: identifies the number of quarters since a project had reported the start of phase-II development.
    - Market and Firm-Market Characteristics:
        - comp_median: identifies whether a project is above or below the median level of market competition in the full sample.
        - exp_median: identifies whether the exposure of a firm to the focal project's market is above or below the median level of exposure in the full sample
        - lowcomp (hicomp): a dummy equal to one if a project is in a low (high) competition market below (above) the median level of competition in the full sample
        - lowexp (hiexp): a dummy equal to one if the firm's exposure to the focal project's market is below (above) the median level of exposure in the full sample
        - lncomp: the natural log of the market competition level

- Statistical Method
    - Note that all analyses use a cox proportional hazards model that has an analysis time equal to the number of quarters since the start of phase-II development. A success event is defined if a project graduates to phase-III (i.e., if dp_grad=1). The hazard function is defined at the beginning of each analysis section using the "stset" command.
    - Note that in some analyses, the hazard model is not estimated (particularly when estimating models with the SMST variable and using the subsamples that are partitioned by both competition and firm exposure). This is due to the lack of variation in the main variables. We have ensured that the code is accurate for all analyses.
    - to run the test or proportional hazards, run the Stata command "estat phtest"

Additional Information on the Analyses of Figures 2, 3 & 4

The tests in Figures 2, 3 and 4 use the "pseudo_phase2_main" sample. In addition to most of the variables described above, these tests also use year indicators that define a year's distance from the BTD event year. These shock year indicators are defined as follows:

- Figure 2:
    - The dummy variables: btdshock_yn5, btdshock_yn4, btdshock_yn3, btdshock_yn2 and btdshock_yn1 are equal to one in the fifth, fourth, third, second, and first year before the shock, respectively, and are equal to zero otherwise.
    - The dummy variables: btdshock_yp1, btdshock_yp2, btdshock_yp3, btdshock_yp4, and btdshock_yp5 are equal to one for the first, second, third, fourth and fifth year after the shock, respectively, and are equal to zero otherwise.
- Figure 3:

    - The dummy variables: smst_yn5, smst_yn4, smst_yn3, smst_yn2, smst_yn1, smst_yp1, smst_yp2, smst_yp3, smst_yp4, and smst_yp5 are defined similar to their btdshock counterparts only around the SMST event year.
- Figure 4:
    - The dummy variables: sm_yn5, sm_yn4, sm_yn3, sm_yn2, sm_yn1, sm_yp1, sm_yp2, sm_yp3, sm_yp4, and sm_yp5 are defined similar to their btdshock counterparts only around the sm event year.

Additional Information on the Analysis of Panel A of Table 8

The test in Panel A of Table 8 use the "pseudo_phase2_nevershocked" sample. In addition to most of the variables described above, these tests also use the following variables:

- hicomp_hiexp_rival: a dummy equal to one if the focal project's firm had recently been shocked in another market that had a competition level above the median level of competition in the full sample, and was highly exposed to that market with an exposure level that is above the median level of exposure in the full sample.
- hicomp_rival: a dummy equal to one if the focal project's firm had recently been shocked in another market that had a competition level above the median level of competition in the full sample.
- hiexp_rival: a dummy equal to one if the focal project's firm had recently been shocked in another market where the rival was highly exposed with an exposure level that is above the median level of exposure in the full sample.