

**1. A description of which author(s) handled the data and conducted the analyses.**

Data handling and programming were primarily performed by Yuxiang Zheng.

**2. A detailed description of how the raw data were obtained or generated, including data sources, the specific date(s) on which data were downloaded or obtained, and the instrument used to generate the data (e.g., for surveys or experiments). We recommend that more than one author is able to vouch for the stated *source* of the raw data.**

The paper used data collected from various sources.

- 1) Financial data: We obtained financial information from Compustat (downloaded Oct 2022).
  - 2) Corporate scientific publication data: Corporate scientific publication data is shared by Arora, Belenzon, and Sheer (2021) and is available from Lia Sheer's website: <https://www.liasheer.com>. (downloaded Oct 2022). We supplement this data source with information on author affiliation and scientific publication fields from OpenAlex, which is available from <https://openalex.org> (downloaded Aug 2023).
  - 3) Patent data: We obtained patent data from multiple resources: (1) linkage of patents and U.S. public companies is shared by Kogan, Papanikolaou, Seru, and Stoffman (2017) and is available from <https://github.com/KPSS2017/Technological-Innovation-Resource-Allocation-and-Growth-Extended-Data> (downloaded Oct 2022). (2) Patent technology class is obtained from PatentsView and is available from [patentsview.org](https://patentsview.org) (downloaded Oct. 2022). (3) Patent global family is from Google Patents and is available from Google's BigQuery (downloaded Nov. 2022). (4) Patent rejection decision is obtained from Patent Application Office Actions Research Dataset and is available from <https://bulkdata.uspto.gov/data/patent/office/actions/bigdata/2017/> (downloaded Nov. 2022)
  - 4) Patent-to-paper citations: We obtained the data from the Reliance-on-Science database available from <https://relianceonscience.org> (downloaded Jul. 2023).
- 3. If the data are obtained from an organization on a proprietary basis, the authors should *privately* provide the editors with contact information for a representative of the organization who can confirm data were obtained by the authors. The editors would not make this information publicly available. The authors should also provide information to the editors about the data sharing agreement with the organization (e.g., non-disclosure agreements, any restrictions imposed by the organization on the authors, such as restrictions to publish certain results). In particular, the authors should indicate if an organization or data provider imposes restrictions on the publication of the results, has not given the authors full control of the relevant data, requires that the results have to be reviewed or approved prior to public release of the paper or publication. This information should be provided to the editors upon submission.**

Not applicable.

**4. A complete description of the steps necessary to collect and process the data used in the final analyses reported in the paper. For experimental and survey papers, we require information about the instructions and instruments used to generate the data, subject eligibility and/or selection, as well as any exclusion criteria. The full set of instructions and instruments can be provided in the online appendix.**

For further details see Section 4 of the manuscript and Online Appendix.

**5. After downloading or obtaining the raw data, all manipulations of the data should be done via computer programs. The code for these manipulations should be included in the code submitted upon acceptance (see below). No manipulations of raw data can take place manually or outside the computer code provided. If compliance with this requirement is not feasible, the authors need to explain and disclose any manipulations of the raw data (e.g., manually created variables or file conversions). When feasible, we also encourage the authors to share the code that downloads the data.**

Attached to this datasheet please find the replication package which includes codes to be run on Python for manipulations of raw data (“2\_Sample and Variables Construction.py”).

**6. The computer programs (i.e., code) used to (1) convert the raw data into the final dataset used in the analysis, (2) to execute the statistical or econometric analysis, and (3) to generate the tables or to produce the output used in constructing tables of the manuscript. A brief description that enables other researchers to understand and run the code should be provided. The purpose of this requirement is to facilitate replication and to help other researchers understand in detail how the raw data were processed, the final sample was formed, variables were defined, outliers were treated, and which commands were used in the analysis, etc. This code or programming is in most circumstances not proprietary. However, we recognize that some parts of the code or data generation process may be proprietary, including from the authors’ perspective. Therefore, instead of disclosing the proprietary portion of the code or program, researchers can provide a detailed step-by-step description of the code or the relevant parts of the code such that it enables other researchers to arrive at the same results that the authors obtained and presented in their manuscript. In such cases, the authors should inform the editors upon initial submission, so that the editors can consider an exemption allowing the step-by-step description. Whenever feasible, authors are required to provide the identifiers (e.g., CIK, CUSIP) for their final sample. Authors should consult our FAQ Sheet on the JAR website for further details.**

Attached to this datasheet please find the replication package which includes codes to be run on Python and Stata for manipulations of raw data and execution of statistical analysis (“2\_Sample and Variables Construction.py” and “3\_Regressions.do”). A list of sample firm identifiers (permno) is also attached (“FirmList.xlsx”).

Note that Stata packages change over time which can cause slight differences in results. For example, the REGHDFE package is regularly updated, and occasionally these updates can have material impacts such as: (1) changes to the syntax of a REGHDFE call; (2) changes to the calculations of clustered standard errors. Slight differences are to be expected, depending on vintage.

**7. A comprehensive log file that shows the execution of the entire code. This log file should cover all the steps that convert the raw data into a final dataset and the execution of all statistical and econometric analyses presented in the tables of the manuscript. The portion of the log file that shows proprietary code or data may be masked. In this case, the reader should be referred to the step-by-step description provided as per the requirements in Item 6.**

Please see attached the log file that shows the execution of the entire code (“LogFile.smcl”).

**8. An assurance that the data and programs will be maintained by at least one author (usually the corresponding author) for at least six years, consistent with National Science Foundation guidelines.**

We assure that we will maintain the data and programs for at least six years, consistent with National Science Foundation guidelines.