

Data Description Sheet
“Informing Entrepreneurs? Initial Public Offerings and New Business Formation”

Manuscript JOAR-2022-262

John M. Barrios (Yale University & NBER); Jung Ho Choi (Stanford University); Yael V. Hochberg (Rice University & NBER); Jinhwan Kim (Stanford University); Miao Liu (Boston College)

July 7, 2026

This statement is prepared in accordance with the Journal of Accounting Research Data and Code Sharing Policy. Each of the eight items in the policy is addressed in turn below.

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

All authors participated in handling the data and conducting the analyses. Jung Ho Choi, Jinhwan Kim, and Miao Liu reviewed all data and all analyses. Research assistants Jin Deng, Lingyu Gu, Mohammadhossein Shafina, Jerry Wen, and Charles Zhang assisted in this process under the authors’ supervision. More than one author can vouch for the source of each raw dataset described in Item 2.

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 can vouch for the stated source of the raw data.

The empirical analysis combines licensed (proprietary), publicly available, and author-generated (survey) data. All raw data were obtained or downloaded during the 2022–2024 data-construction period. For each source, the table below reports the provider, the basis on which the data were accessed, the coverage used in the paper, and the period during which the data were obtained or downloaded.

#	Data	Provider	Access basis	Coverage used	Obtained / downloaded
1	IPO data (filing/issue/withdrawal dates, issuer name, ZIP, CUSIP, SIC, NASDAQ return)	Refinitiv (Thomson Reuters) SDC Platinum, New Issues database	Licensed	1988:Q1–2016:Q4	2022–2024
2	New business registrations (county-quarter)	Startup Cartography Project (SCP)	Public	1988–2016 (50 states/territories through 2014; 47 thereafter)	2022–2024
3	Establishment births (county–two-digit NAICS–year)	U.S. Census Bureau, Statistics of U.S. Businesses (SUSB)	Public	2000–2015	2022–2024
4	Public firm locations / headquarters (ZIP, fiscal year)	CRSP and Compustat (via WRDS)	Licensed	As matched to sample period	2022–2024
5	Personal income per capita (county-year)	U.S. Bureau of Economic Analysis (BEA)	Public	1969–2018	2022–2024
6	County population (county-year)	U.S. Census Bureau	Public	1970–2017	2022–2024

7	Private-sector payroll and employment (county-quarter; used for volatility measures)	U.S. Bureau of Labor Statistics, Quarterly Census of Employment and Wages (QCEW)	Public	1990–2016	2022–2024
8	Entrepreneurship-related internet search volume	Google Trends	Public	2004:Q1–2016:Q4	2022–2024
9	EDGAR server log files (IP, date, CIK, form type, documents downloaded)	SEC EDGAR log files	Public	2003–2016	2022–2024
10	S-1 registration statement full text	SEC EDGAR Archives	Public	Sample IPO filings	2022–2024
11	IP-address-to-location mapping	MaxMind (GeoLite / commercial database)	Licensed / commercial	As of download	2022–2024
12	Startup organization names/locations (for entrepreneur EDGAR matching)	Crunchbase	Licensed	As of download	2022–2024
13	Company news events (media coverage; supplementary analysis)	S&P Capital IQ Key Developments	Licensed	1988–2018	2022–2024
14	Small Business Administration loan data	U.S. Small Business Administration (SBA)	Public	As used in build	2022–2024
15	Geocoding / place-to-county and ZIP-to-county crosswalks	Public geocoding resources (e.g., Census/HUD crosswalks)	Public	As of download	2022–2024

Survey data (instrument used to generate the data). To complement the archival evidence, the authors fielded a survey of 503 current and prospective entrepreneurs via the Prolific platform during the 2023–2024 period. Respondents were recruited through Prolific from a range of industries, with and without prior entrepreneurial experience; more than half held a college degree or higher. The survey instrument asked about awareness of IPOs, how such events affected respondents’ likelihood of starting a business, the types of information respondents found useful, and the sources from which they obtained it, with responses measured on a five-point Likert scale. The final sample consists of 503 respondents. The study was determined to be exempt from full institutional review board review. The complete survey instrument and participant demographics are reported in the Online Appendix (see Item 4).

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, and any restrictions imposed by the organization on the authors). 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 must be reviewed or approved prior to public release of the paper or publication.

Several datasets used in the paper were obtained by a licensed or proprietary basis: Refinitiv (Thomson Reuters) SDC Platinum (New Issues), CRSP, Compustat, Crunchbase, S&P Capital IQ Key Developments, and MaxMind. The remaining datasets (SCP, SUSB, BEA, Census population, BLS QCEW, Google Trends, SEC EDGAR log files and Archives, and SBA) are publicly available.

The CRSP and Compustat data were obtained through Stanford University’s WRDS (Wharton Research Data Services) subscription under standard academic terms; the other licensed datasets were accessed through institutional subscriptions or licenses at the authors’ universities. For each licensed source, the authors can provide the editors with the contact information of a representative of the data provider (or the relevant institutional subscription route) who can confirm that the authors obtained the data, together with information about the applicable data-sharing agreement:

Licensed source	Provider representative / access route (private to editors)	Data-sharing agreement / restrictions
Refinitiv SDC Platinum	Institutional subscription; provider representative available to the editors on request	Standard subscription; academic use
CRSP	Stanford University WRDS subscription (institutional)	Standard WRDS subscription; academic use
Compustat	Stanford University WRDS subscription (institutional)	Standard WRDS subscription; academic use
Crunchbase	Institutional/licensed access; provider representative available to the editors on request	Standard license; academic use
S&P Capital IQ Key Developments	Institutional/licensed access; provider representative available to the editors on request	Standard license; academic use
MaxMind	Licensed/commercial database; provider representative available to the editors on request	Standard license

To the best of the authors’ knowledge, no data provider has imposed any restriction on the publication of the results, required prior review or approval of the paper before public release, or withheld from the authors full control of the relevant data. The licensed data were used solely under standard academic/subscription terms. The authors can provide the applicable licenses for each licensed dataset upon request.

4. A complete description of the steps necessary to download, obtain or collect as well as 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.

The steps necessary to download, obtain, and process each dataset used in the final analyses are described in Section 2 (“Data Sources and Key Variables”) of the manuscript and in Appendix A (“Variables Definition”). The sample screens applied to the IPO data (excluding financial firms [SIC 6000–6999], unit offerings, closed-end funds including REITs, ADRs, limited partnerships, SPACs, and spin-offs) are described in the paper. The README file included in the replication package documents the corresponding data-construction sequence script-by-script, from raw source files to the final regression samples (see Item 6).

For the survey evidence (Section 5 and Table 11 of the manuscript), the Online Appendix section “OA Survey Instruments and Survey Participant Demographics” reports the complete survey instrument, participant recruitment, eligibility and selection (current and prospective entrepreneurs recruited via the Prolific platform), and participant demographics.

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.

After the raw data were downloaded or obtained, all manipulations of the data were performed via the computer programs (Stata, Python, R, and shell scripts) included in the replication package. No manipulation of the raw data took place manually or outside the provided code. The package README explains the software environment, the required user-written packages, and the order in which the scripts must be run. File paths in the scripts are retained from the original working environment for record-keeping purposes, those replicating the paper will need to update these paths to reflect their local directory structure before running the code.

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.

The replication package submitted with these final files contains the code used to (1) convert the raw data into the final analysis datasets, (2) execute the econometric analyses, and (3) generate the tables and figures of the manuscript. The package is organized as follows:

Build/Data Construction: five Stata scripts (1_build.do, 2_IPO.do, 3_Public_Firm_Num.do, 4_Merge_County_IPO_Issuance.do, 5_Birth_Clean.do), run in order, that process each raw source (new business registrations, SDC IPO data, Compustat/CRSP public-firm counts, BEA income per capita, Census population, Google Trends, SUSB establishment births, SBA loans) and merge them into the county-quarter working dataset.

Build/Regression Sample Construction: six Stata scripts that construct the regression samples used in the analyses: the stacked difference-in-differences samples, the county-industry sample, the propensity-score-matched sample, the QCEW-based uncertainty/volatility measures, and the supplementary media-coverage dataset.

Build/Regression Sample Construction/Edgar_Crunchbase: the Stata, Python, R, and shell code that processes the SEC EDGAR server logs, geocodes IP addresses using MaxMind, matches downloads to Crunchbase startup organizations, and computes the S-1 textual measures. A separate README in the S1 subfolder documents the S-1 text-processing chain.

Analysis/BCHKL_regression_final.do: This produces the tables and figures reported in the manuscript, using the datasets constructed in the Build steps as inputs. All regression output is written to Output/Regression_Final.xlsx and to separate image files for the figures.

A README.md file at the root of the package provides a brief description that enables other to understand and run the code. The software requirements (Stata version 16 or later, Python 3.7 or later, and R 4.0 or later, with the required user-written packages listed), the execution order, a description of each script, and a mapping from every table and figure in the manuscript to the dataset(s) and code that produce it. Note that Stata user-written packages (e.g., reghdfe) are updated over time, and slight differences in results can arise depending on package vintage.

Identifiers for the final sample. The file IPO_CUSIP.csv in the replication package provides the CUSIP identifiers for the final IPO sample which is based on 11,039 IPO events (completed and withdrawn) from the first quarter of 1988 through the fourth quarter of 2016, after applying the sample screens described in the paper. These identifiers allow other researchers to reconstruct the IPO sample from a licensed SDC subscription.

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.

The replication package includes log files documenting the execution of the code end-to-end. Each data-construction and sample-construction script in the Build folders is accompanied by its paired .log file (including edgar_crunchbase_data.log and the log files for the S-1 text-processing chain), and Analysis/analysis.log records the execution of the statistical and econometric analyses presented in the tables and figures of the manuscript. Table 11 reports descriptive tabulations of the survey responses described in Item 2.

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.

Consistent with National Science Foundation guidelines and JAR policy, John M. Barrios (the corresponding author) will maintain the data and computer programs used in this study for at least six years following publication.