

**Data description sheet for**  
**“Corporate Litigation, Governance, and the Role of Law Firms”**

Frank Allen Ferrell, Alberto Manconi, Ekaterina Neretina, William Powley, and Luc Renneboog

In this document, we provide information regarding the construction of our dataset and the sources of data used in the JAR publication entitled “Corporate Litigation, Governance, and the Role of Law Firms”, following the Journal of Accounting Research Data Policy.

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

Alberto Manconi and Ekaterina Neretina carried out the data collection and the empirical analysis.

**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 raw data on lawsuits and law firms were obtained from the following sources:

- *ISS Securities Class Action Services* (downloaded on May 14, 2020).
- *Advisen Master Significant Cases and Actions Database* (downloaded on January 31, 2020).
- *Audit Analytics Litigation* (downloaded on May 28, 2018).
- *Stanford Securities Class Action Clearinghouse* (downloaded on May 14, 2020).
- *Federal Court Cases Integrated Database* (this dataset contains information on lawsuits, but not law firms; downloaded on May 21, 2018).
- The list of law firms included in *Legal500* rankings and *Legal500* practice areas (downloaded on October 5, 2020).

To match defendant firms with Compustat GVKEY identifiers, and define publicly listed companies among the plaintiffs, we use the names of publicly listed firms from CRSP, Compustat, the WRDS SEC Analytics Suite, the Federal Court Cases Integrated Database, and the SEC Electronic Data Gathering, Analysis, Retrieval (EDGAR) system, and WRDS Subsidiaries (the matches were obtained throughout July-December 2020). The law firm characteristics were obtained from the Bureau van Dijk (BvD) database (downloaded on June 9, 2024), and from the Internet Archive/Wayback Machine (downloaded throughout June-September 2024). Moreover, the paper uses an extensive list of control variables that were downloaded from the above-mentioned data sources, and also additional data sources including MSCI (formerly KLD), WRDS Financial Ratios Suite, IBES, BoardEx, Thomson Reuters 13F (throughout 2019-2020).

Part of the data come from a confidential databased on insurance premiums from a leading insurance provider, as discussed in the next section of this document.

**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.**

The confidential information about litigation insurance pricing used in Table 7 (columns 3-6) of the paper and Online Appendix Figure C.2 was obtained from a leading insurance provider, as described in the Data Description Sheet included with the original submission. The authors have provided to the editors the contact information of the representative of the organization and the data sharing agreement upon submission.

Additionally, the data used in the main analysis and the baseline tests are available from the sources listed in point 2. of this document. Those sources (Audit Analytics Litigation, ISS Securities Class Action Services, Advisen Master Significant Cases and Actions Database) require a subscription. The Federal Court Cases Integrated Database is available through the WRDS platform (<https://wrds-www.wharton.upenn.edu/pages/get-data/federal-judicial-center/>). The Stanford Securities Class Action Clearinghouse is available at the URL: <https://securities.stanford.edu/>.

**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 paper data section (II) and Appendix B provide a description of the data sources and variable construction. Additionally, the replication package includes the log file Ferrell, Manconi, Neretina, Powley, & Renneboog – data construction log.log, which contains the complete set of instructions from the original Stata program used to merge the various sources and construct the final dataset on which most tests are run. The only omission is the suppressed output of commands that standardize specific entity names (e.g., law firms, defendant corporations, plaintiffs), which cannot be disclosed due to confidentiality agreements and the subscription-based nature of the underlying data. All other steps are documented exactly as implemented, enabling reconstruction of the dataset from the original sources. A pseudo dataset (mock\_dataset.dta), mirroring the structure of the final dataset but containing only random noise and mock identifiers (see point 6), is also provided for reference. As explained in point 6.

below, we also provide the full list of Compustat GVKEY identifiers for the defendant corporations in the Stata data file `gvkey_identifiers.dta`.

We use Stata to run the tests presented in the paper. The replication package includes a Stata `.do` file called `tests_mockdata.do`, which performs all the tests on the mock dataset. Separately, we have provided the editors with log files showing how we assemble the data and the output of the tests on the real dataset.

**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.**

In order to comply with data confidentiality agreements and protect proprietary information, we provide in the replication package the log file `Ferrell, Manconi, Neretina, Powley, & Renneboog - data construction log.log`, which was generated by running the original program that constructs all the data. The original program contains commands that standardize the names of individual law firms, defendants, and plaintiffs (e.g., correcting misspellings). Because the underlying datasets used in the analysis are subscription-based and contain proprietary information, these commands cannot be made public. When running the program, we therefore suppress the printout of these name-standardization commands to avoid revealing entity identities. The provided log file documents the complete sequence of operations and transformations, excluding only these name-standardization steps, thus preserving full transparency of the data construction process while maintaining confidentiality.

**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.**

To produce the final dataset used in the analysis, the following pieces of code are provided in the replication package:

- (i) `kim_skinner.sas`. This piece of SAS code is designed to run on the WRDS server, to obtain the variables used in our paper following Kim and Skinner (2012). It requires as inputs two files: `crsp_with_gvkey.sas7bdat` (containing links from Compustat's GVKEY identifiers to CRSP's PERMNO identifiers, directly obtained from the CRSP/Compustat Merged Database) and `ncusip_link.sas7bdat` (containing links from CRSP's PERMNO identifiers to the historical CUSIP code, directly obtained from the CRSP Monthly Stocks database).
- (ii) `transparency.sas`. This piece of SAS code is designed to run on the WRDS server, to obtain the part of the "transparency and liquidity" characteristics used as additional control variables in Tables 4 and 5.B. It requires as inputs the file `crsp_with_gvkey.sas7bdat` (containing links from Compustat's GVKEY identifiers to CRSP's PERMNO identifiers, directly obtained from the CRSP/Compustat Merged Database).
- (iii) `appendix_figure_C2.do` produces Appendix Figure C.2.

Each observation in the final data corresponds to one lawsuit, against one publicly listed U.S. firm, brought by one or more plaintiff law firms on a given date and resolved on a given date. This data structure is mimicked in the pseudo data set, provided along with the replication package, `mock_dataset.dta` (see point 4. above). Because a nontrivial portion of the data require a subscription or are confidential (see point 3. above), we cannot disclose the full set of defendant identifiers matched to the lawsuit dates. We provide, however, the full set of Compustat GVKEY identifiers for the defendant corporations in our data, in the Stata file `gvkey_identifiers.dta`, included in the replication package.

**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.**

We include in the replication package the following log files:

- (i) Ferrell, Manconi, Neretina, Powley, & Renneboog - data construction `log.log`. This log file shows the procedure used to assemble the dataset used in the analysis. Some portions, which contain proprietary data, are masked (these parts are indicated by comments in the log). The main program is written in Stata. Some of the data construction steps require running lines of commands in Python.

The corresponding lines of commands are provided in the relevant places of the log as comments. As it is evident from the log, they use and produce the intermediate data files used in the Stata code.

- (ii) `kim_skinner.log`. This log file reports the construction of the Kim and Skinner (2012) control variables, obtained running a piece of SAS code on the WRDS server.
- (iii) `transparency.log`. This log file reports the construction of several control variables related to the defendant corporation's transparency, obtained running a piece of SAS code on the WRDS server.
- (iv) `Ferrell, Manconi, Neretina, Powley, & Renneboog - replication on real data.log`. This log file reproduces all the results reported in the paper (with the exception of Appendix Figure C.2, see below).
- (v) `Ferrell, Manconi, Neretina, Powley, & Renneboog - replication Appendix Figure C2 (confidential data).log`. This log file refers to the replication of Appendix Figure C.2, which is based on the confidential dataset from an insurance provider.
- (vi) `Ferrell, Manconi, Neretina, Powley, & Renneboog - replication on mock data.log`. This log file reports the results from running the paper's replication code on the mock data set (`mock_dataset.dta`) provided in the replication package.

**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 will maintain all data and programs for at least six years.