

Data Description Sheet

This document describes the data collection and analysis for “**Incidence, Risk, and Disclosure of Corporate Litigation: Insights from Federal Court Filings**” by **Mary Brooke Billings, Robert W. Holthausen, Christine Petrovits and Danye Wang**.

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

- **Danye Wang** handled extracting the raw data from all data sources and creating the main merged dataset. She also conducted all analyses in the paper.
- Although not directly involved in the handling of the data or the writing of the code to perform the analyses, **Mary Billings, Robert Holthausen, and Christine Petrovits** were involved in vetting data sources, validating the data, and understanding/guiding coding decisions as to measurement and other econometric issues. This included corroborating the existence of lawsuits and the coding of key variables (e.g., OUTCOME, DISCLOSE) for subsets of the main dataset using original source information (PACER, LAW360, SEC filings), helping with the manual review of “fuzzy” matches, and investigating differences among data sources (e.g., PACER, FJC, Audit Analytics).

2. A detailed description of how the raw data were obtained or generated, including data sources, the 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 main analyses in the paper use litigation data described in the attached **Online Appendix and Appendices A through D of the paper** (reproduced here) describe how the raw data were obtained, provide the specific sources of the data, and describe variable measurement.
- **We accessed the data on the following dates:**
 - PACER: June 2019, July 2019, December 2019, April 2021, September 2021, October 2022, December 2022. Accessed during 2019-2025 to manually review details of specific cases.
 - FJC: September 2019, October 2019, November 2022, April 2024.
 - Audit Analytics: May 2018 (litigation dataset), February 2019 (SEC comment letter data), March 2023 (SEC comment letter data), February 2023 (litigation dataset).
 - FASB: August 2019.
 - Compustat: August 2019, December 2022.
 - CRSP: October 2019, February 2020, October 2021, June 2022, March 2023, October 2023, August 2024.
 - IBES: January 2018, August 2023.
 - RavenPack: October 2019, February 2023, March 2023.
 - Advisen: September 2023
 - Stanford Securities Class Action Clearinghouse: January 2023 through June 2025.
 - SEC EDGAR (sec.gov): Accessed during 2019-2025 to manually review filings.

- **To access the BHPW Corporate Federal Litigation Data assembled and described in the paper and the Online Appendix please visit Wharton Research Data Services (WRDS) at <https://wrds-www.wharton.upenn.edu/pages/get-data/>.**

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 agreement, any restrictions imposed by the organization on the authors with respect to publishing certain results).

Not applicable, as all data sources have been identified in the paper.

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.

A complete description of the steps necessary to download, obtain, collect, and process the data used in the final analyses reported in the paper is provided in the response to #2 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.

All data manipulations are performed using computer programs, and the relevant code has been submitted (see response to #6).

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 SAS and Stata codes used to generate the final dataset and tables are provided. The CIK identifiers for the final sample are also included. The litigation data that we assemble and describe in the paper (and online appendix) will be provided through Wharton Research Data Services (WRDS) at <https://wrds-www.wharton.upenn.edu/pages/get-data/>.

The BHPW Corporate Federal Litigation Data assembled and described in the paper and the online appendix is accessible at Wharton Research Data Services (WRDS) at <https://wrds-www.wharton.upenn.edu/pages/get-data/>. There is no charge to download the litigation data. Requiring WRDS credentials facilitates the inclusion of proprietary firm identifiers (i.e., GVKEYs) with the data. We provide CIKs to those users without access to Compustat GVKEYs.

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.

Comprehensive log files are provided for the SAS programs and for the Stata programs used in different parts of the analysis.

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.

The authors agree to maintain the data and programs for at least six years.