Corporate Governance: An International Review

WILEY

ORIGINAL ARTICLE

Alternative Data as an External Governance Mechanism

Emma Li[1] | Zhengwei Wang[2] | Le Zhao[3]

[1]Department of Finance, Deakin University, Geelong, Australia | [2]PBC School of Finance, Tsinghua University, Beijing, China | [3]Business School, China Academy of Corporate Governance, Nankai University, Tianjin, China

Correspondence: Le Zhao (zhaole@nankai.edu.cn)

Received: 18 February 2024 | **Revised:** 4 January 2025 | **Accepted:** 10 January 2025

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 72402105).

Keywords: alternative data | corporate governance | financial statement fraud

ABSTRACT

Research Question/Issue: This study explores the relationship between the availability of alternative data and financial statement fraud among Chinese firms.

Research Findings/Insights: We first show that companies' online sales data, independently collected by a third-party data vendor, can detect financial statement fraud before 2018, before which such alternative data were not available to the public. Second, we find that the association between financial statement fraud and the proxy constructed from alternative data becomes weaker after the publicity of alternative data. Third, we find that the impact from alternative data availability on corporate financial statement fraud is greater among firms with weaker external governance, such as firms with less analyst coverage, as well as lower quality auditors.

Theoretical/Academic Implications: This study enhances the understanding of the relationship between information environment characteristics and corporate fraud. We suggest a new information intermediary, alternative data providers, and investigate the role that alternative data play in financial statement fraud. And we also add to the growing literature regarding the impact of technology on corporate fraud and corporate governance.

Practitioner/Policy Implications: This study proposes a new measure to capture company managers' hidden behavior and provide new insights regarding the usefulness of company-level online sales data, as provided by an independent entity in relation to financial statement fraud. These findings have important implications for investors, policymakers, and market participants in understanding the role of alternative data in detecting financial fraud and acting as an external governance mechanism.

1 | Introduction

Financial fraud is a critical area of research, particularly financial statement fraud, where firms engage in deceptive practices related to financial reporting, such as inflating profits or revenues. Such financial statement fraud has increased tremendously in recent years and has generated serious negative consequences for shareholders, employees and societies (Amiram et al. 2018; Dimmock, Gerken, and Graham 2018). At the same time, recent technology advancements have enabled third-party vendors to independently collect real-time, granular indicators of firms' fundamentals and have made the data available to different external parties, including investors, auditors and analysts (Kolanovic and Krishnamachari 2017). These alternative data sources can include point of online sales, consumer transactions, and satellite images (Froot et al. 2017; Monk, Prins, and Rook 2019; Agarwal, Qian, and Zou 2021; Kang, Stice-Lawrence, and Wong 2021; Liao, Cui, and Sun 2021), which are not within traditional data sources, such as SEC filings, financial statements, and press releases. Understanding the relationship between financial statement fraud and the availability of additional information in an economy is important because this alternative data may affect managerial behavior, and corporate decisions (Zhu 2019; Blankespoor et al. 2022).

The theoretical framework of Beyer et al. (2010) examines the relationship between information environment characteristics and financial statement fraud. Third-party information intermediation utilizes external sources to provide independent viewpoints, which helps reduce information asymmetry and increase transparency. As new information becomes available to the public, there is less room for mangers' manipulation. We suggest that alternative data providers act as new information intermediaries and explore their role in preventing financial statement fraud, indicating that their absence may lead to higher instances of earnings manipulation and managerial misconduct.

However, empirically testing this conjecture can be challenging due to the endogenous relationship between information availability and financial statement fraud. It is difficult to determine whether managers commit fraud in response to changes in information or if they act differently due to other unobserved motivations. To overcome this empirical challenge, we employ a quasi-natural experiment in China by leveraging an exogenous event: the availability of alternative online sales data to third parties. This approach allows us to test whether the association between financial statement fraud and alternative data diminishes when such data become more accessible.

In recent years in China, the expansion of mobile services and the growth of online payment systems have reduced data-collecting costs, leading independent companies to gather alternative data. By making alternative data available, external parties have been able to acquire information at a reduced cost. However, despite the increase in the use of these alternative data sources, research into their impact on financial markets, financial statement fraud, and corporate governance remains limited. Our study aims to bridge this gap, offering insights into how the accessibility of alternative data can potentially mitigate the occurrence of financial statement fraud, thereby enhancing the integrity of financial reporting and corporate governance practices.

In this study, we initially present evidence demonstrating a meaningful association between reported sales growth in financial reports and online sales growth, validating the informativeness of online sales data. Our findings reveal a significant and positive relationship between reported sales growth and online sales growth. These results suggest that online sales data offer valuable insights into overall sales performance, aligning with prior studies by Froot et al. (2017) and Liao, Cui, and Sun (2021). Next, we examine the association between financial statement fraud and the availability of alternative data sources through online consumer transactions. We show that online sales data independently collected by third-party vendors can detect financial misreporting and accounting fraud in the e-commerce industry. Traditionally, researchers have primarily used accruals, discretionary, and the mapping between revenue and other financial statement figures to identify financial reporting misbehavior. However, using financial statement data to isolate nondiscretionary and discretionary aspects of revenue can be challenging because financial statement measures can be manipulated by managers to disguise their misreporting (Brazel, Jones, and Zimbelman 2009). Recently, researchers have started using alternative data from a third-party company that may offer reliable guidance on the discretionary and nondiscretionary aspects of revenue (Allee, Baik, and Roh 2021; Chiu et al. 2023). We constructed a new proxy based on alternative data of online sales to detect accounting fraud. Before December 2018, the alternative data of online sales were not available to the public. We show that the difference between the actual online sales collected independently from third-party vendors and the revenue reported on financial statements is both economically and statistically significant in its association with financial statement fraud. One standard deviation increase of the difference between alternative data from online sales and actual revenue from financial statements is associated with a 1.08% increase of the probability of financial statement fraud. This result is economically important considering that the base rate of corporate fraud is 18.9%.

Second, we leverage an exogenous event to companies in December 2018, when the largest third-party financial data provider in China, WIND made alternative online sales data available to the public[1] and showed that availability of alternative data weakens the association between financial statement fraud and the proxy based on alternative data of online sales. The corporate information environment emerges endogenously because of information asymmetries and agency problems between entrepreneurs, managers, and investors. According to previous studies, information intermediaries play a crucial role in corporate governance (Chen, Harford, and Lin 2015). Information intermediaries can act as an external regulatory agency and play a significant role in studying areas in corporate governance, including fraud detection and insider trading (Chen et al. 2016; Dang et al. 2021; Ren, Zhong, and Wan 2021). When alternative data providers arise as another type of information source, and intermediaries and information acquisition costs decrease for the public, we show that corporate management is less likely to manipulate earnings: the difference between growth rate of sales from financial statements and actual growth rate of online sales collected from the alternative data provider is much smaller. In addition, we show that the availability of alternative data weakens the association between financial statement fraud and the difference between the actual online sales collected independently from third-party vendors and the revenue reported on financial statements.

Third, we examine the possible mechanism to explain the relationship between the availability of alternative data and financial statement fraud. Financial analysts and high-quality external auditors can play an important external governance role in scrutinizing management behavior and in deterring earnings management and corporate fraud (Chen, Harford, and Lin 2015; Chen et al. 2016; Fan and Wong 2005; Lennox and Pittman 2010). If alternative data providers arise and play a role as an external regulatory agency, we expect that firms suffering from severer agency problems and weaker external governance will be more likely affected by introduction of the alternative data. We find that the effects of alternative data are more pronounced among the firms with low-quality auditors or/and less analyst coverage. In addition, we investigate the market reaction to the introduction of alternative data sources. The positive market reaction to the introduction of alternative data sources provides indirect supporting evidence that investors value the information from this data source.

Our study makes three main contributions. The first contribution of our study is to investigate the relationship between information environment characteristics and financial statement fraud. Corporate governance can be shaped by many different factors including social media and managers' voluntary disclosure decisions. Zingales (2000) and Dyck, Volchkova, and Zingales (2008) first reveal how the media plays a key regulatory role in influencing corporate policies and guiding decision-making about resource distribution. The results of previous studies suggest that business-focused media play a positive role in detecting accounting fraud (Dyck, Morse, and Zingales 2010), exposing board inefficiency (Joe, Louis, and Robinson 2009), and preventing insider trading (Dai, Parwada, and Zhang 2015). In this paper, we propose a new information intermediary, alternative data providers, and investigate the role of alternative data play in the financial statement fraud and provide supporting evidence that alternative data providers can act as an external governance mechanism when corporate management are subject to weaker external governance protection.

Second, we add to the growing literature of the impact of technology on financial statement fraud and corporate governance. In recent years, multiple technological innovations have been assessed for their effects on capital markets. These include algorithms and high-frequency trading (Hendershott, Jones, and Menkveld 2011), robo-journalism (Blankespoor, deHaan, and Zhu 2018), and alternative data (Zhu 2019; Blankespoor et al. 2022). Our paper relates to Zhu (2019) and Blankespoor et al. (2022) in that it examines the role of alternative data; Zhu (2019) investigates its effect on stock price informativeness and managerial trading behavior in the United States, while Blankespoor et al. (2022) explore its impact on firms' disclosure choices based on credit and debit card sales data. However, our study differs from theirs by focusing on the influence of alternative data on managerial decision-making, specifically in the context of managerial financial fraud. Recently, several fintech companies have developed alternative data and complex algorithms to mitigate the limitations of traditional models and data when assessing borrowers' credit risks and their ability to repay loans. Several studies have documented that alternative data can overcome the information asymmetry that was critical in evaluating borrower risks, especially for the unbanked population and borrowers with limited credit history (Goldstein, Jagtiani, and Klein 2019; Croux et al. 2020). We examine the impact of alternative data on corporate governance in our paper. While prior studies find that alternative data sources can predict earnings and revenue (Froot et al. 2017; Agarwal, Qian, and Zou 2021; Liao, Cui, and Sun 2021; Dichev and Qian 2022), we investigate the question of how the introduction of alternative data affects corporate financial statement fraud, where the information acquisition cost is reduced for a set of external parties.

Third, in general, our findings relate to forensic economics research. The goal of this research line is to detect managers' unseen actions or decisions by comparing two separate measures (Allee, Baik, and Roh 2021; Chiu et al. 2023). These two separate measures capture the same economic activity, but managers' incentives can impact those two measures in different ways. Our paper proposes a new measure to capture managers' hidden behavior and provide new insights regarding the usefulness of company-level online sales data provided by an independent entity in relation to the financial statement fraud. Moreover, our study is also related to an emerging research field regarding the application of non-financial measures (NFMs) in a variety of scenarios. In this case, online sales in our study provide additional information, compared with financial statement measures, to external parties, highlighting the importance of this particular NFM for uncovering fundamentals (Brazel, Jones, and Zimbelman 2009).

2 | Theoretical Framework and Institutional Background

2.1 | Theoretical Framework

Our study enhances the understanding of the relationship between information environment characteristics and financial statement fraud. Figure 1 illustrates the broader theoretical framework of this relationship, as described by Beyer et al. (2010). The managerial decision-making regarding information disclosure and fraud is shaped endogenously by three main factors:

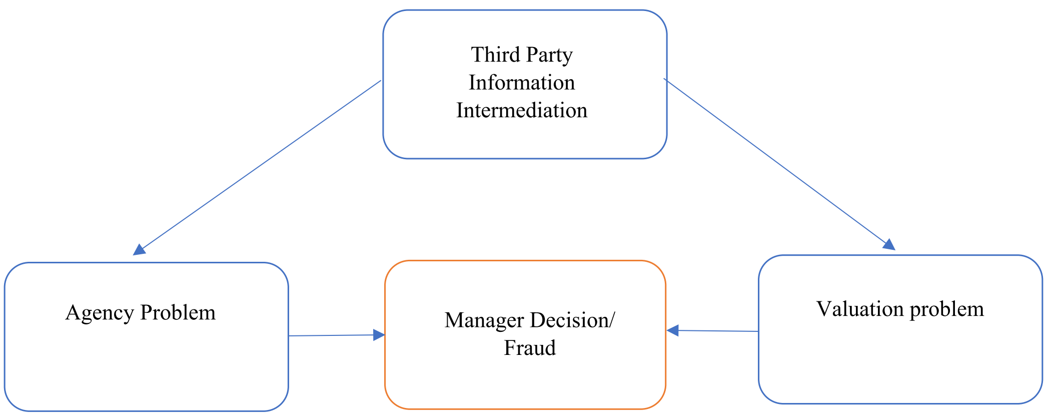

FIGURE 1 | Theoretical framework. [Colour figure can be viewed at wileyonlinelibrary.com]

1. Agency problem: The ex-post demand for accounting information arises from the separation of ownership and control, resulting in capital providers not having full decision-making rights.
2. Valuation problem: Ex-ante, firm managers typically possess more information about the expected profitability of current and future investments than outsiders. This information asymmetry complicates the assessment of investment opportunities' profitability by external capital providers.
3. Third-party information intermediation: The availability and cost-effectiveness of alternative mechanisms or information from external sources help address valuation and agency problems. External sources provide independent perspectives and data, reducing information asymmetry between stakeholders and the firm. These sources bring diverse expertise and methodologies, offering a broader range of insights into the firm's performance and prospects. This diversity reduces reliance on potentially biased internal information, enhancing transparency and improving the valuation of the firm's assets and operations. Additionally, by providing objective benchmarks and information, external sources help align the interests of managers (agents) with those of shareholders (principals), thereby reducing agency costs and improving corporate governance.

In contrast, the lack of cost-effective information is associated with conditions that enable fraud to occur without significant risk of detection (Murphy and Dacin 2011). For instance, insufficient external scrutiny may allow managers to manipulate financial statements. Agency theory (Jensen and Meckling 1976) suggests that financial analyst coverage can increase firm value by reducing agency costs through enhanced monitoring. Previous empirical studies have examined the role of analysts as third-party information intermediaries (Chen, Harford, and Lin 2015; Chen et al. 2016). We propose that alternative data providers serve as a new type of information intermediary and investigate their role in financial statement fraud. Our conjecture is that when alternative data providers are not publicly available, creating an information environment conducive to undetected fraud, corporate managers are more likely to engage in earnings manipulation or other misconduct.

2.2 | Institutional Background

In general, alternative data is defined as data that are not part of a financial statement or a report (Kolanovic and Krishnamachari 2017; Jagtiani and Lemieux 2019; Dichev and Qian 2022). The cost of data collection has gradually been reduced due to recent developments in machine learning, cloud computing, and artificial intelligence (AI). Several third-party vendors extract information from consumer online activities while others use website traffic or measure pedestrian traffic in stores. Some other companies rely on satellite imagery to count the number of cars in outdoor parking lots. Overall, alternative data are more focused on a consumer's footprint (Froot et al. 2017). Compared with traditional financial data usually released only after the end of the financial quarter, alternative data are updated or provided more frequently from third-party vendors. Some vendors update the data every day, while others update every month. In addition, these alternative data are argued to be relatively objective (Chiu et al. 2023) and is less likely to suffer corporate manipulation when compared with financial data provided by companies themselves.

The e-commerce market has grown tremendously. Between 2016 and 2020, e-commerce volume had a 12.2% annual growth rate in the United States and 9.3% in China. Most products and services can be provided through e-commerce in China, including grocery, electronics, computers, software, fashion, beauty and personal care, home and kitchen, health and household, and toys and games. Therefore, e-commerce sales data are generally representative of consumer industries and in line with the industries covered in the Deloitte report regarding Chinese consumer and retail industry.[2] But we want to point out two differences between online shopping and household consumption in general. First, in-hospital treatment is prohibited to sell through e-commerce platforms. Second, vehicles are usually purchased offline through the dealer and rarely on e-commerce platforms. Despite the above differences, and the fact that we cannot generalize our results to all industries, leveraging the availability of e-commerce sales data still provides us a good opportunity to examine the impact of alternative data on financial statement fraud in major consumer and retail industries.

China's weak legal enforcement and investor protection (Allen, Qian, and Qian 2005) may not have kept pace with its rapid economic growth, resulting in regulatory loopholes that allow firms to act opportunistically. Among Chinese listed firms, common forms of corporate fraud include false statements, material information omission, and embezzlement. The China Securities Regulatory Commission (CSRC), the primary regulator of the Chinese securities exchanges, investigates and disciplines corporate fraud. Besides regular reviews and random inspections, the CSRC responds to external whistleblower reports and complaints about alleged fraud. Upon identifying fraud, the CSRC's enforcement actions range from internal and public criticism to formal criminal prosecution.

Wind Information (WIND) is a leading provider of financial information services in China. WIND provides financial data to 90% of financial institutions in China, including securities companies, fund management corporations, insurance companies, banks, and investment firms. WIND started collecting e-commerce sales data since 2015; however, such data were not released to external parties until December 2018.

The public release of WIND data at the end of 2018 was unanticipated and straightforward, providing a valuable opportunity to leverage this event. Although we understand that WIND has been developing this product for some time, our search of media coverage found no discussion of these data release until December 2018. Second, online shopping constitutes a larger portion of consumption in China compared with the rest of the world. This dataset provides a substantial sample of firms, allowing us to track detailed online sales relative to accounting sales figures. While the impact of alternative data on online sales in other markets might not be as significant, we argue that the availability of such data should enhance corporate governance and influence managerial behavior. Practitioners can use alternative data for strategic advantage, and policymakers can leverage it for stronger governance.

This dataset covers more than 95% of the e-commerce platforms and companies in China, including Taobao, JD, Tmall, Yihaodian, Gome, Suning, and so on. The dataset has been updated monthly to users since December 2018, and the previous data collected since 2015 is also provided to all users. WIND collects online sales data from every transaction of every consumer on those platforms in real time. According to the sales records of each brand and each product from those e-commerce platforms, WIND then aggregates that data at the company/enterprise level. For example, if individual A buys an iPhone through Taobao and individual B buys an iPhone through JD, WIND will collect each sale of an iPhone through Taobao and through JD and aggregate both sales into the online sales for Apple. Therefore, we argue that this dataset can well reflect the real time online sales information from different companies on the e-commerce platforms, especially for the companies in consumer industries, where most of the sales revenue are generated through online platforms. Even for some companies with a lower proportion of online sales, the online sales data can be extrapolated for overall sales, as online sales and offline sales are usually closely correlated. In a recent study, Liao, Cui, and Sun (2021) find that online sales are significantly positively correlated with corporate operating income.

In addition, the dataset is subject to a lower possibility of corporate manipulation or systematic fraud as it is collected from real transactions of consumers through an independent third-party vendor. Companies tend to manipulate earnings or revenue through early or late earnings recognition when such third-party independent data are not available to the public (Allee, Baik, and Roh 2021). While the independent data vendor can pinpoint the exact time of each sale through the confirmation of receipt by consumers, we argue that online sales data are more objective and timely compared with financial data disclosed by the company. Our first conjecture is that, during the period when online sales data are not available to the public, the company has a higher probability of engaging in earnings manipulation or other misconduct, where we observe the growth rate of sales revenue disclosed in the company's annual report is significantly different from the growth rate of online sales. Therefore, we first construct an indicator *DIFF*, which is the difference between the growth rate of sales revenue disclosed in corporate annual reports and the growth rate of online sales; second, we examine whether this indicator *DIFF* can predict or identify financial statement fraud during the period when online sales data are not available to the public. We examine whether the greater the value of *DIFF*, the more likely that a company is going to conduct financial fraud. After 2018, the introduction of this additional information provides a unique experimental opportunity to examine the impact of alternative data directly extracted from e-commerce platforms on managerial behavior within those listed companies.

3 | Data, Sample, and Variable Construction

3.1 | Data and Sample

We compile data from two commercially available sources: CSMAR and WIND. China Stock Market and Accounting Research (CSMAR) database contains detailed financial accounting information and daily stock returns for all firms publicly traded in China since 1990, and it has been widely used in prior studies regarding Chinese firms (e.g., Fan, Wong, and Zhang 2007; Chen et al. 2012). CSMAR database also collects regulatory sanction and corporate fraud information, including all regulatory sanctions where the nature of fraud and fraud-perpetrating years are recorded (Wu and Wang 2018).

WIND collects detailed information for online sales from publicly traded companies that sell their products on e-commerce platforms. For each company and brand, the dataset includes total online sales, total number of products sold online, and average price of product sold online. Since December 2018, the dataset is available to the public and is updated at the beginning of each month and covers seven industries, including electronic appliances, food and beverages, medicine and biology, textiles and apparel, culture, education, sports and entertainment, building materials, and furniture and home furnishing.

For our main analysis, we obtained 233 companies with online sales information from WIND and identify those companies that have been listed before 2017 in CSMAR. Therefore, we can obtain financial data from 2017 and 2018, 2 years before and 2 years (2019 and 2020) after the event in December 2018. Over the sample period from 2017 to 2020, there are a total of 647 company-year observations (163 unique firms). Table 1 summarizes the sample selection process.

3.2 | Key Variable Construction

In this paper, we propose a new variable that aims to detect financial statement fraud by adopting an independent source of data from company-level online sales, to provide a trustworthy, timely, and cost-efficient measure of a company's economic activity. We construct a new variable *DIFF* using the difference between the growth rate of sales revenue disclosed in the company's annual report and the growth rate of online sales collected

TABLE 1 | Sample selection.

Sample selection criteria for main analysis	Observations (firm-year)
All A-share listed companies that are covered in WIND Online Sales database during our sample period from 2017 to 2020.	806
Less: Firms listed after 2017	(156)
Less: Firms with missing data for the main variables	(3)
Final sample	647

by an independent vendor. If a company's accounting performance grows faster (slower) than its real online sales activities, captured by a higher (lower) *DIFF*, it is possible that the firm is engaging in upward (downward) performance manipulation.[3] Traditionally, researchers have focused primarily on the use of key revenue related information from financial statements to measure earnings manipulation. However, it can be difficult to isolate discretionary and nondiscretionary components of earnings and revenue because managers may be motivated to twist key financial measures to cover their misbehavior. To isolate this impact, we identify the alternative data of online sales from an independent reporting agency WIND that provides credible insight into both the discretionary and nondiscretionary elements.

To assess financial statement fraud, we adhere to prior research (Wu and Wang 2018) and gather data from CSMAR. Information on regulatory sanctions and corporate fraud is sourced from the CSMAR database, detailing the type and year of fraud. Our focus is solely on violations linked to financial statement fraud, such as inflated profits and revenues, while excluding nonfinancial transgressions from our sample, such as illegal stock trading and unauthorized loan guarantees. The variable *FRAUD* is an indicator that equals one if a firm engaged in financial statement fraud in a given year, and zero otherwise.

3.3 | Descriptive Statistics

Table 2 shows descriptive statistics for our main variables. For the full sample, *FRAUD* has a mean of 0.189, suggesting that 18.9% of firm-years commit financial statement fraud and are detected by regulators. The variable *DIFF* has a mean (median) of −0.223 (−0.189), implying that sample firms' online sales growth rates are greater than their growth rates in financial statements on average. However, around 28% sample firms experience greater reported sales growth. The standard deviation of *DIFF* is 0.415, suggesting a large variation in the difference between reported sales growth and online sales growth. The mean value of reported sales growth (*GROWTH*) is 10.7%, while the corresponding value of online sales growth (*E_GROWTH*) is 31.6%, implying that online sales grow faster. We also note that around 10% sample firms report a loss, 40% are audited by small auditing firms, and only 4.6% receive modified audit opinion.

Furthermore, we assess the significance of online sales in relation to firms' fundamentals to validate our *DIFF* measure. In Appendix B, we present the results of the regression analysis, which examines the reported sales growth (*GROWTH*) in financial reports and its correlation with online sales growth (*E_GROWTH*). Our findings reveal a significant and positive relationship between *E_GROWTH* and *GROWTH*, indicating that online sales data contain valuable and relevant information about the overall sales performance of the business. These results align with previous studies from Froot et al. (2017) and Liao, Cui, and Sun (2021). After alternative online sales data became available in December 2018, we compare the reported sales growth rate between the period from January 2017 to December 2018 and the period from January 2019 to December 2020, by plotting distribution of *GROWTH*, the growth rate of sales revenue disclosed in the company's annual report. In Figure 2, we find the distribution of *GROWTH* after 2018 has shifted left and is closer to zero. It suggests that firms reported lower sales growth when alternative data became available to the public.

4 | Empirical Results

The main objective of our study is to examine the effect of changes in the information environment on financial statement fraud. In Section 4.1, we first examine the association between financial statement fraud and the difference between growth of online sales and growth of sales from financial statements before alternative data becomes available. In Section 4.2, we reexamine this association after the introduction of online sales data from December 2018. In Section 4.3, we explore the possible external governance mechanism to explain the relationship

TABLE 2 | Summary statistics.

Variable	*N*	Unique firm	Mean	Median	P25	P75	SD
FRAUD	647	163	0.189	0	0	0	0.391
DIFF	647	163	−0.223	−0.189	−0.451	0.023	0.415
POST	647	163	0.502	1	0	1	0.500
GROWTH	647	163	0.107	0.085	−0.012	0.200	0.233
E_GROWTH	647	163	0.316	0.277	0.056	0.568	0.375
SIZE	647	163	22.617	22.435	21.896	23.095	1.152
LEV	647	163	0.397	0.365	0.260	0.512	0.176
MTB	647	163	3.324	2.383	1.517	3.689	3.130
LOSS	647	163	0.104	0	0	0	0.305
Non-BIG10	647	163	0.400	0	0	1	0.490
VOL	647	163	0.024	0.024	0.019	0.028	0.006
ANAL	647	163	1.966	2.079	0.693	3.296	1.381
MAO	647	163	0.046	0.000	0.000	0.000	0.210

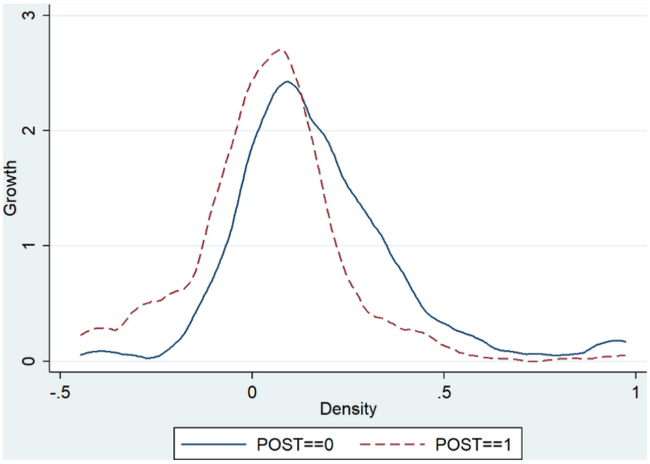

FIGURE 2 | Density plot of sales growth rate. This figure plots the distribution of reported sales growth rate. The solid line represents the density plot of the reported sales growth prior to the availability of alternative data. The dotted line represents the density plot of the reported sales growth after alternative data are available. [Colour figure can be viewed at wileyonlinelibrary.com]

between the availability of alternative data and financial statement fraud. In Section 4.4, we examine the market reaction to the introduction of alternative data, providing indirect supporting evidence that investors value the information from this data source. In Sections 4.5 to 4.7, we provide additional robustness checks to validate our main results.

4.1 | Univariate and Multivariate Tests

First, we investigate the association between corporate financial statement fraud before alternative data became available. Following Allee, Baik, and Roh (2021), we estimate the coefficients of the following OLS regression model:

$$FRAUD_{i,t} = \gamma_0 + \gamma_1 DIFF_{i,t} + \sum \gamma_i CONTROLS_{i,t} + YEAR\ FE + IND\ FE + \varepsilon \quad (1)$$

where the dependent variable $FRAUD_{i,t}$ represents whether the firm i commits financial fraud in year t. $DIFF_{i,t}$ represents the difference between firm i's sales growth in financial statements and the online sales growth in year t. $CONTROL_{i,t}$ is a set of

TABLE 3 | Pearson correlation matrix.

Panel A: 2017–2018

	FRAUD	DIFF	SIZE	LEV	MTB	LOSS	Non-BIG10	VOL	MAO
FRAUD	1								
DIFF	0.094**	1							
SIZE	−0.123**	0.151***	1						
LEV	0.152***	0.178***	0.377***	1					
MTB	−0.043	−0.032	−0.009	−0.078	1				
LOSS	0.305***	0.043	−0.191***	0.175***	−0.096*	1			
Non-BIG10	0.046	−0.090	−0.051	−0.010	−0.008	0.002	1		
VOL	0.097*	0.003	−0.142**	0.110**	0.059	0.232***	0.062	1	
ANAL	−0.256***	−0.030	0.457***	−0.068	0.362***	−0.298***	−0.048	−0.030	
MAO	0.347***	0.009	−0.197***	0.241***	−0.014	0.282***	−0.098*	0.079	−0.222***

Panel B: 2019–2020

	FRAUD	DIFF	SIZE	LEV	MTB	LOSS	Non-BIG10	VOL	ANAL
FRAUD	1								
DIFF	−0.046	1							
SIZE	−0.137**	0.013	1						
LEV	0.171***	−0.035	0.253***	1					
MTB	−0.039	0.009	0.073	0.051	1				
LOSS	0.204***	−0.115**	−0.189***	0.201***	−0.071	1			
Non-BIG10	−0.013	−0.058	−0.062	0.011	0.004	0.120**	1		
VOL	0.178***	−0.050	−0.233***	0.194***	0.110**	0.178***	−0.013	1	
ANAL	−0.170***	0	0.513***	−0.080	0.413***	−0.303***	−0.095*	−0.210***	
MAO	0.196***	0.003	−0.135**	0.109*	−0.044	0.351***	−0.013	0.212***	−0.174***

Note: The superscripts ***, **, and * denote statistical significance at 1%, 5% and 10%, respectively.

TABLE 4 | Growth difference and financial fraud prior to the availability of alternative data.

	(1)	(2)	(3)
	FRAUD	**FRAUD**	**FRAUD**
DIFF	**0.029*****	**0.031*****	**0.026*****
	(7.97)	**(9.70)**	**(7.61)**
SIZE			−0.009
			(−0.62)
LEV			0.232
			(1.05)
MTB			0.008
			(0.90)
LOSS			0.272***
			(21.03)
Non-BIG10			0.048***
			(3.84)
VOL			3.500
			(0.75)
ANAL			−0.058***
			(−6.73)
MAO			0.171
			(1.12)
Constant	0.224***	0.225***	0.366
	(23.49)	(143.26)	(1.07)
Ind and Year FE	No	Yes	Yes
Observations	322	322	322
R^2	0.009	0.023	0.168

Note: This table presents regression results investigating the relationship between the growth difference and financial statement fraud prior to the availability of alternative data (2017–2018). *FRAUD*, indicator variable equals one if a firm committed a fraud related to financial reporting in a given year, and zero otherwise. *DIFF*, the difference between a firm's sales growth in financial statements and the online sales growth. Detailed definitions of all variables are provided in Appendix A. The *t*-statistics in parentheses are based on standard errors clustered at the industry-year level. The superscripts ***, **, and * denote statistical significance at 1%, 5% and 10%, respectively.

control variables that affect a firm's fraud probability. Industry and year fixed effects are included to control for calendar-year-specific and industry-specific effects.

Following prior literature (e.g., Chen et al. 2016; Wu and Wang 2018; Allee, Baik, and Roh 2021), we control for firm size (*SIZE*), financial leverage (*LEV*), and market-to-book value (*MTB*) that are known to be associated with financial statement fraud. And we further include the loss indicator (*LOSS*), stock return volatility (*VOL*), non-big10 external auditing indicator (*Non-BIG10*), and analyst coverage (*ANAL*) in our model because a firm's financial health and external environment are important determinants of financial statement fraud (Lennox and Pittman 2010; Allee, Baik, and Roh 2021). To control for auditor's opinion, we constructed modified audit opinion (*MAO*), which is an indicator variable equal to one if the company receives a modified audit opinion, and zero otherwise.

Table 3 presents the correlation matrix for our main variable. Consistent with our prediction, we find that the Pearson correlation between *FRAUD* and *DIFF* is 0.094 and significant prior to the availability of alternative data (Panel A). However, the correlation becomes insignificant when alternative data become public (Panel B).

Across different specifications in Table 4, we find that the coefficients on *DIFF* are positive and significant before 2019 with or without industry fixed effects and/or control variables. *DIFF*, earnings manipulation proxy using alternative online sales data, one standard deviation increase in *DIFF* is associated with a 1.08% increase in the financial statement fraud rate, which is economically important considering that the base rate of financial statement fraud is just 18.9% (Column 3). This result suggests that before the alternative data of online sales are available to the public, earnings manipulation proxy leveraging the online sales data is both economically and statistically significant in association with corporate fraud. The results imply that alternative data can be used to detect financial statement fraud prior to the publicity of the data.

4.2 | Introduction of Alternative Data

Next, we re-examine this association between corporate financial statement fraud and sales growth difference after the alternative data becomes available. Based on Allee, Baik, and Roh (2021), we estimate the coefficients of the following OLS regression model:

$$FRAUD_{i,t} = \gamma_0 + \gamma_1 DIFF_{i,t} + \gamma_2 DIFF_{i,t} \times POST_t + \sum \gamma_i CONTROLS_{i,t} + YEAR\ FE + IND\ FE + \varepsilon \quad (2)$$

where the dependent variable $FRAUD_{i,t}$ represents whether the firm i commits financial fraud in year t. If company i releases fraudulent data on its sales for 2019 in 2020, and this fraud is discovered in 2021, the specification is $FRAUD_{i,2019}$. $DIFF_{i,t}$ represents the difference between firm i's sales growth in financial statements and the online sales growth in year t. $POST_t$ is a dummy variable, equals one if year t is 2019 and 2020, and zero otherwise. $CONTROL_{i,t}$ is a set of control variables that affect a firm's fraud probability. Industry and year fixed effects are included to control for calendar-year-specific and industry-specific effects.

In Table 5 Panel A, we use the full sample from 2017 to 2020, and further examine whether the availability of alternative data weakens the association between financial fraud and sales growth difference by using model (2). In Columns (1) and (2) without any control variables, the coefficients on *DIFF* are positive and significant, and the coefficients on the interaction of *DIFF* and *POST* are negative and significant. In Column (3), we find the coefficient on the interaction term remains significantly negative after adding control variables. The results imply that the association between corporate fraud and sales growth difference becomes weaker after the publicity of alternative data. Consistent with previous literature (e.g., Chen, Harford, and Lin 2015), we find that firms with higher leverage ratio, negative earnings,

TABLE 5 | Growth difference and financial fraud after alternative data are available.

Panel A

	(1)	(2)	(3)
	FRAUD	**FRAUD**	**FRAUD**
DIFF	0.029***	0.031***	0.028***
	(8.13)	(9.97)	(7.02)
POST	−0.065**		
	(−2.18)		
DIFF × POST	**−0.051****	**−0.056****	**−0.044*****
	(−2.30)	**(−2.43)**	**(−3.39)**
SIZE			−0.009
			(−1.39)
LEV			0.170*
			(1.71)
MTB			0.002
			(0.59)
LOSS			0.148**
			(2.23)
Non-BIG10			0.027***
			(3.53)
VOL			6.260**
			(2.15)
ANAL			−0.038***
			(−2.87)
MAO			0.341***
			(3.59)
Constant	0.224***	0.192***	0.235
	(23.98)	(39.86)	(1.38)
DIFF + DIFF × POST = 0	p-value = 0.330	p-value = 0.293	p-value = 0.245
Ind and Year FE	No	Yes	Yes
Observations	647	647	647
R^2	0.009	0.033	0.170

Panel B

	(1)	(2)	(3)
	FRAUD	**FRAUD**	**FRAUD**
DIFF	**−0.022**	**−0.024**	**−0.020**
	(−0.97)	**(−1.02)**	**(−1.05)**
SIZE			−0.007
			(−1.21)

(Continues)

TABLE 5 | (Continued)

Panel B	(1)	(2)	(3)
	FRAUD	FRAUD	FRAUD
LEV			0.088***
			(3.47)
MTB			−0.002
			(−1.01)
LOSS			0.045
			(0.67)
Non-BIG10			0.026***
			(11.25)
VOL			10.268***
			(4.81)
ANAL			−0.015
			(−1.58)
MAO			0.472***
			(5.28)
Constant	0.160***	0.159***	0.027
	(5.50)	(20.13)	(0.22)
Ind and Year FE	No	Yes	Yes
Observations	325	325	325
R^2	0.002	0.062	0.228

Note: Panel A presents regression results investigating whether the availability of alternative data weakens the relationship between the growth difference and financial statement fraud. Panel B presents regression results investigating the relationship between the growth difference and financial statement fraud after alternative data are available (2019–2020). *FRAUD*, indicator variable equals one if a firm committed a fraud related to financial reporting in a given year, and zero otherwise. *DIFF*, the difference between a firm's sales growth in financial statements and the online sales growth. Detailed definitions of all variables are provided in Appendix A. The *t*-statistics in parentheses are based on standard errors clustered at the industry-year level. The superscripts ***, **, and * denote statistical significance at 1%, 5% and 10%, respectively.

return volatility, and fewer analyst coverage are more likely to commit fraud. The results suggest that the difference between sales growth from online sales data and financial statements cannot detect financial statement fraud when the online sales data becomes available to the public. We suspect that the managers understand the increased cost of such manipulation of sales data on financial statements and therefore modify their behavior once aware the online sales data would be available to the public.

In Table 5, Panel B, we use model (1) to re-examine the previous association after 2018. In contrast to prior findings in Table 4, we find that across different specifications, the coefficient on *DIFF* is insignificant after 2018. The results are consistent with what we documented previously in Table 5, Panel A.

4.3 | External Governance Channels

Previous studies show that the role of information intermediaries in corporate governance is crucial. Information intermediaries can act as an external regulatory agency and play a significant role in studying fraud detection, insider trading, executive compensation, and other corporate governance fields. The alternative data provider can act as a new type of information intermediary and the alternative data itself can play an important role in providing external governance (Zhu 2019), therefore mitigating financial statement fraud or fraud. In this section, we explore the possible external governance mechanism to explain the relationship between the availability of alternative data and financial statement fraud.

Prior research documents that high-quality external auditors can perform a governance role and discourage financial statement fraud (Fan and Wong 2005; Lennox and Pittman 2010). If the alternative data can play an external governance role, we expect that firms audited by low-quality auditors are more likely to benefit from the availability of alternative data. To test this hypothesis, we first construct the variable *Non-BIG10*, a dummy variable equals 1 if the company is not audited by the top 10 auditors in China, and then interact *Non-BIG10*, *DIFF* and *POST* in model (2). Consistent with our prediction, in Table 6, we find that the coefficients on interaction term are

TABLE 6 | Cross-sectional tests: external auditors.

	(1) FRAUD	(2) FRAUD	(3) FRAUD
DIFF	0.030***	0.033***	0.029***
	(4.34)	(5.29)	(3.79)
POST	−0.124***		
	(−3.36)		
DIFF × POST	−0.120***	−0.122***	−0.078***
	(−5.50)	(−5.61)	(−3.70)
Non-BIG10	0.045***	0.059***	0.053***
	(3.18)	(8.58)	(10.06)
DIFF × Non-BIG10	−0.002	−0.002	0.001
	(−0.29)	(−0.25)	(0.11)
POST × Non-BIG10	−0.092***	−0.084***	−0.067***
	(−3.59)	(−6.01)	(−3.11)
DIFF × POST × Non-BIG10	**−0.107***	**−0.103***	**−0.053**
	(−4.40)	**(−4.61)**	**(−2.18)**
SIZE			−0.009
			(−1.43)
LEV			0.167
			(1.63)
MTB			0.002
			(0.58)
LOSS			0.148**
			(2.10)
VOL			6.302**
			(2.09)
ANAL			−0.037**
			(−2.60)
MAO			0.338***
			(3.46)
Constant	0.252***	0.201***	0.231
	(15.07)	(30.97)	(1.36)
Ind and Year FE	No	Yes	Yes
Observations	647	647	647
R^2	0.017	0.041	0.173

Note: This table presents regression results investigating the influence of external auditors on the relationship between the availability of alternative data and financial statement fraud. *FRAUD*, indicator variable equals one if a firm committed a fraud related to financial reporting in a given year, and zero otherwise. *DIFF*, the difference between a firm's sales growth in financial statements and the online sales growth. *POST*, indicator variable equals one if the year is 2019 and 2020 when alternative data are available. *Non-BIG10*, equals 1 if the client firm is audited by a non-big10 auditing firm, and 0 otherwise. Detailed definitions of all variables are provided in Appendix A. The *t*-statistics in parentheses are based on standard errors clustered at the industry-year level. The superscripts ***, **, and * denote statistical significance at 1%, 5% and 10%, respectively.

TABLE 7 | Cross-sectional tests: financial analysts.

	(1) FRAUD	(2) FRAUD	(3) FRAUD
DIFF	0.019***	0.021***	0.018***
	(5.59)	(6.92)	(6.11)
POST	−0.029**		
	(−2.57)		
DIFF × POST	−0.016**	−0.022***	−0.028***
	(−1.99)	(−5.03)	(−4.29)
Low_ANAL	0.221***	0.223***	0.166***
	(10.48)	(9.93)	(5.05)
DIFF × Low_ANAL	0.027***	0.026***	0.028***
	(6.70)	(6.10)	(4.87)
POST × Low_ANAL	−0.118**	−0.116**	−0.129**
	(−2.49)	(−2.44)	(−2.12)
DIFF × POST × Low_ANAL	**−0.070***	**−0.069***	**−0.035***
	(−2.67)	**(−3.22)**	**(−1.85)**
SIZE			−0.012*
			(−1.85)
LEV			0.174
			(1.59)
MTB			−0.001
			(−0.56)
LOSS			0.159**
			(2.27)
Non-BIG10			0.026***
			(3.19)
VOL			6.756**
			(2.50)
MAO			0.171
			(1.24)
Constant	0.132***	0.115***	0.018***
	(20.34)	(10.89)	(6.11)
Ind and Year FE	No	Yes	Yes
Observations	647	647	647
R^2	0.057	0.081	0.178

Note: This table presents regression results investigating the influence of financial analysts on the relationship between the availability of alternative data and financial statement fraud. *FRAUD*, indicator variable equals one if a firm committed a fraud related to financial reporting in a given year, and zero otherwise. *DIFF*, the difference between a firm's sales growth in financial statements and the online sales growth. *POST*, indicator variable equals one if the year is 2019 and 2020 when alternative data are available. *Low-ANAL*, equals 1 if the industry-adjusted analyst coverage is negative, and 0 otherwise. Detailed definitions of all variables are provided in Appendix A. The *t*-statistics in parentheses are based on standard errors clustered at the industry-year level. The superscripts ***, **, and * denote statistical significance at 1%, 5% and 10%, respectively.

negative and significant in all columns. The results suggest that the impact from alternative data availability on financial statement fraud is greater among firms with low-quality auditors.

Second, Chen, Harford, and Lin (2015) and Chen et al. (2016) find that financial analysts play an important external governance role in scrutinizing management behavior and deterring corporate fraud. We expect that firms with less analyst coverage are suffering from severer agency problems and the effects of alternative data will be more pronounced. To test this hypothesis, we first construct variable *Low_ANAL*, which equals one if the industry-adjusted analyst coverage is negative, and then we add the interaction term of *Low_ANAL*, *DIFF*, and *POST* in model (2). In Table 7, we show that the coefficients on interaction term are significantly negative, implying that the effects of alternative data availability on financial statement fraud is concentrated among firms with less analyst coverage.

4.4 | Market Reaction

If the alternative data provide additional information for investors to detect corporate financial statement fraud, we expect a positive market reaction when alternative data are introduced for all the firms covered by the WIND online sales database.

We use the cumulative abnormal return around the announcement date of the introduction of online sales data to investigate the market reaction to the availability of alternative data.[4] We follow a standard event study procedure (Fracassi and Tate 2012; Kim and Klein 2017) to estimate the stock prices changes around the earliest public announcement of the introduction of the online sales database. The market model is applied to an estimation period of 250 to 20 trading days prior to the announcement date of the alternative data introduction. We ensure that a firm has at least 60 trading days of data to calculate benchmark returns. The abnormal return (AR) for firm i on day t is calculated as follows:

$$AR_{it} = R_{it} - \hat{\alpha} - \hat{\beta} R_{mt} \tag{3}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the ordinary least squares estimates of the intercept and slope coefficient derived from the market model. R_{it} represents the daily stock return, and R_{mt} represents the market return using the value-weighted index to measure.

Return data are extracted from CSMAR for all companies covered by the WIND database. After obtaining abnormal return of each period of the stock, the cumulative abnormal returns (CAR) are calculated by aggregating daily ARs over time starting from before the day of alternative data availability announcement to after the announcement date. Cumulative abnormal returns for an event period from t_1 to t_2 can compute as follows:

$$\mathrm{CAR}(t_1, t_2) = \sum_{t_1}^{t_2} AR_t \tag{4}$$

where CAR_j is the cumulative abnormal return of firm j over the interval. Abnormal returns are computed during the event period from one trading day, two trading days, and three trading days before and after the event date.

Event study results in Table 8 shows the cumulative abnormal returns for all firms covered by the WIND online sales database. We present results for the 1-day event Window (−1, +1), 2-day event Window (−2, +2), 3-day event Window (−3, +3), 5-day event Window (−5, +5), and 10-day event Window (−10, +10). Averaged over all firms in the sample, a firm's stock covered by the WIND online sales database leads to an increase in market valuation of 1.32% (significant at the 1% level) between day −1 and +1. We get consistent findings when using longer event Windows.[5] These findings are broadly consistent with our expectation that the market has responded positively to those stock that are covered by the alternative data from WIND. These results provide indirect supporting evidence that alternative data can play an external governance role as investors do pay attention to this source of information. Unfortunately, our dataset lacks the necessary identifiers to distinguish between different types of investors. Consequently, our analysis captures aggregate investor behavior without the ability to segment by specific investor categories, which may limit the granularity of our findings.

4.5 | Difference-In-Differences Analysis

The previous identification strategy used a pre-post analysis, which supported our findings about the diminished ability to detect financial statement fraud after the publicity of alternative data. To enhance the robustness of our analysis, we now employ a formal difference-in-differences (DID) approach. Table 9 (Panel A) presents the regression results using a DID model during the same sample period, and we find consistent results. Treated group is the firms that are covered by the WIND online sales database. To construct the matched control group, we first identify the sample of companies not covered by WIND. We then follow a standard matching process by estimating the probability that a firm is treated using a logit regression with data from the pre-publicity period. This logit regression includes the control variables from the main regression, as well as year and industry fixed effects. We create a matched sample using the 1:1 nearest neighbor matching technique with a caliper set at 0.03, following Shipman, Swanquist, and Whited (2017).

TABLE 8 | Market reaction to the availability of alternative data news.

Window period	(−1, 1)	(−2, 2)	(−3, 3)	(−5, 5)	(−10, 10)
CAR	1.32%***	1.71%***	2.11%***	1.06%**	1.14%**
p-value	< 0.001	0.001	0.007	0.021	0.034

Note: This table contains the average cumulative abnormal returns (CAR) around the earliest alternative data availability news date of December 21, 2018, the date when WIND started to promote the online sales database on their official website. Daily abnormal returns are calculated as a firm's raw return minus the expected returns from a market model with parameters estimated over the −250 to −20 trading-day Window relative to the event date. The superscripts ***, **, and * denote statistical significance at 1%, 5% and 10%, respectively.

TABLE 9 | Difference-in-difference analysis.

Panel A: PSM + DID

	(1)	(2)	(3)
	FRAUD	**FRAUD**	**FRAUD**
TREAT	0.026		
	(0.53)		
POST	0.014		
	(0.44)		
TREAT × POST	**−0.096****	**−0.102****	**−0.102****
	(−2.27)	**(−2.42)**	**(−2.38)**
SIZE			0.148***
			(3.21)
LEV			0.262
			(1.42)
MTB			0.003
			(0.52)
LOSS			0.045
			(0.93)
Non-BIG10			0.018
			(0.51)
VOL			−2.958
			(−1.19)
ANAL			0.014
			(0.79)
MAO			0.210***
			(3.53)
Constant	0.206***	0.229***	−3.201***
	(5.72)	(19.82)	(−3.11)
Firm and Year FE	No	Yes	Yes
Observations	902	902	902
R^2	0.007	0.671	0.688

Panel B: Falsification analysis

	(1)	(2)	(3)
	FRAUD	**FRAUD**	**FRAUD**
TREAT	−0.037		
	(−0.88)		
POST2016	0.037		
	(0.99)		
TREAT × POST2016	**0.037**	**0.050**	**0.048**
	(0.72)	**(0.96)**	**(0.91)**
SIZE			−0.078
			(−0.99)
LEV			0.330
			(1.40)
MTB			−0.009
			(−1.08)
LOSS			0.034
			(0.40)
Non-BIG10			−0.000
			(−0.00)
VOL			1.584
			(0.50)
ANAL			0.018
			(0.61)
MAO			0.111
			(0.74)
Constant	0.195***	0.190***	1.747
	(6.05)	(13.03)	(1.03)
Firm and Year FE	No	Yes	Yes
Observations	874	874	874
R^2	0.006	0.515	0.522

(Continues)

TABLE 9 | (Continued)

Note: Panel A presents regression results using a difference-in-difference (DID) model. The dependent variable, *FRAUD*, is an indicator variable that equals one if a firm committed financial reporting fraud in a given year, and zero otherwise. The variable *TREAT* is an indicator that equals one if the firm is covered by the WIND online sales database, and *POST* is an indicator that equals one for the years 2019 and 2020, when alternative data are available. To improve comparability, the sample includes observations from industries (3-digit) with more than five unique treated firms. To construct the matched control group, we first estimate the probability that a firm is treated using a logit regression with data from the pre-publicity period. The logit regression includes the control variables used in the main regression, as well as year and industry fixed effects. The estimated coefficients for the main variables from the first stage are (bold coefficients indicate significance at the 1% level; constant, year and industry dummies suppressed): TREAT = 0.34**SIZE**-1.07LEV + 0.00MTB + 0.43LOSS-0.37Non-BIG10–35.12VOL + 0.38**ANAL** + 0.51MAO. We then create a matched sample using the 1:1 nearest neighbor matching technique with a caliper set at 0.03, following Shipman, Swanquist, and Whited (2017). Panel B reports the falsification test results using a difference-in-difference model. We use a placebo coverage year of 2016 and use the sample from the pre-period, 2015 to 2018. *FRAUD*, indicator variable equals one if a firm committed a fraud related to financial reporting in a given year, and zero otherwise. *TREAT*, indicator variable equals one if the firm is covered by the WIND online sales database. *POST2016*, indicator variable equals one if the year is 2017 and 2018. We construct matched control group same as Panel A. Detailed definitions of all variables are provided in the Appendix A. The *t*-statistics in parentheses are based on standard errors clustered at the firm level. The superscripts ***, **, and * denote statistical significance at 1%, 5% and 10%, respectively.

In addition, Table 9 (Panel B) presents the falsification test results using a DID model. We use a placebo coverage year of 2016 and a sample from the pre-period, 2015 to 2018. The variable *POST2016* is an indicator that equals one for the years 2017 and 2018. We construct the matched control group using the same method as in Panel A. We do not find any significant results across different specifications.

The paper primarily relies on the constructed *DIFF* indicator to identify the effect of the availability of alternative data on the prediction of corporate fraud. While the DID approach cannot construct the *DIFF* indicator, it provides great robustness support for our results.

4.6 | Alternative Dependent Variable

The current dependent variable is binary and may not capture the intensity of financial statement fraud. To address this, we have constructed a new variable, *PUNISH*, to capture the intensity of financial statement fraud based on the announcement documents issued by the regulatory authorities. This variable is

TABLE 10 | The severity of financial statement fraud.

	(1)	(2)	(3)
	Ln (PUNISH+1)	**Ln (PUNISH+1)**	**Ln (PUNISH+1)**
DIFF	0.021***	0.023***	0.020***
	(6.94)	(8.17)	(5.35)
POST	−0.039*		
	(−1.88)		
DIFF × POST	**−0.035****	**−0.037****	**−0.028*****
	(−2.16)	**(−2.28)**	**(−2.94)**
SIZE			−0.011
			(−1.49)
LEV			0.208**
			(2.71)
MTB			−0.003
			(−1.06)
LOSS			0.094
			(1.50)
Non-BIG10			0.055***
			(4.73)
VOL			4.843**
			(2.02)
ANAL			−0.027**
			(−2.33)
MAO			0.340***
			(4.03)
Constant	0.159***	0.140***	0.251
	(17.59)	(35.73)	(1.46)
Ind and Year FE	No	Yes	Yes
Observations	633	633	633
R^2	0.006	0.030	0.182

Note: This panel presents regression results using an alternative variable that captures the severity of the financial statement fraud. *PUNISH*, a variable equals one if the firm is ordered to rectify by the regulatory authority, equals two if the executives of the firm are criticized or warned by the regulatory authority, equals three if the firm and executives is publicly reprimanded by the regulatory authority, equals four if the firm is fined by the regulatory authority, and zero otherwise. *DIFF*, the difference between a firm's sales growth in financial statements and the online sales growth. *POST*, indicator variable equals one if the year is 2019 and 2020 when the alternative data are available. Detailed definitions of all variables are provided in the Appendix A. The *t*-statistics in parentheses are based on standard errors clustered at the industry-year level. The superscripts ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

TABLE 11 | Robustness tests: firm fixed effects.

	(1)	(2)
	FRAUD	FRAUD
DIFF	0.023**	0.013**
	(2.51)	(2.41)
DIFF × POST	**−0.035*** **	**−0.028*** **
	(−3.29)	**(−2.70)**
SIZE		0.265***
		(4.88)
LEV		0.207
		(1.32)
MTB		0.005
		(0.60)
LOSS		0.124**
		(2.41)
Non-BIG10		0.000
		(0.02)
VOL		−3.235*
		(−1.91)
ANAL		−0.000
		(−0.01)
MAO		0.229***
		(3.02)
Constant	0.192***	−5.839***
	(44.03)	(−4.77)
Firm and Year FE	Yes	Yes
Observations	647	647
R^2	0.625	0.648

Note: This table presents regression results using firm fixed effects instead of industry fixed effects. *FRAUD*, indicator variable equals one if a firm committed a fraud related to financial reporting in a given year, and zero otherwise. *DIFF*, the difference between a firm's sales growth in financial statements and the online sales growth. *POST*, indicator variable equals one if the year is 2019 and 2020 when the alternative data are available, and 0 otherwise. Detailed definitions of all variables are provided in the Appendix A. The *t*-statistics in parentheses are based on standard errors clustered at the industry-year level. The superscripts ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

TABLE 12 | Robustness tests: logit model.

	(1)	(2)	(3)
	FRAUD	FRAUD	FRAUD
DIFF	0.324***	0.370***	0.376***
	(3.54)	(3.47)	(3.21)
POST	−0.443**		
	(−2.11)		
DIFF × POST	**−0.457*** **	**−0.518*** **	**−0.492*** **
	(−2.89)	**(−2.96)**	**(−3.29)**
SIZE			−0.096*
			(−1.88)
LEV			1.225*
			(1.79)
MTB			−0.004
			(−0.14)
LOSS			0.751**
			(2.27)
Non-BIG10			0.172***
			(3.60)
VOL			44.404**
			(2.26)
ANAL			−0.280***
			(−3.62)
MAO			1.640***
			(3.30)
Constant	−1.214***	−1.149***	0.285
	(−19.98)	(−28.18)	(0.24)
Ind and Year FE	No	Yes	Yes
Observations	647	621	621
Pseudo R^2	0.011	0.019	0.149

Note: This table presents main regression results using the *logit* model instead of linear probability model. *FRAUD*, indicator variable equals one if a firm committed a fraud related to financial reporting in a given year, and zero otherwise. *DIFF*, the difference between a firm's sales growth in financial statements and the online sales growth. *POST*, indicator variable equals one if the year is 2019 and 2020 when the alternative data are available. Detailed definitions of all variables are provided in Appendix A. The *z*-statistics in parentheses are based on standard errors clustered at the industry-year level. The superscripts ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

defined as follows: one if the firm is ordered to rectify by the regulatory authority, two if the executives of the firm are criticized or warned by the regulatory authority, three if the firm and its executives are publicly reprimanded by the regulatory authority, four if the firm is fined by the regulatory authority, and zero otherwise. In China, directives for rectification or warnings from regulatory authorities are relatively mild, whereas public reprimands or fines signify more serious violations. Table 10 shows that the results are robust and consistent with our main findings.

4.7 | Robustness Check

Although we have panel data, the period for examining growth difference changes spans only 4 years. Following Breuer and Dehaan (2024), we use industry fixed effects instead of firm fixed effects to ensure sufficient variation in our analysis. However, we include firm fixed effects in our main regressions, and the results remain consistent and robust, as shown in Table 11. In models (1) and (2), we use OLS regression instead of logit model to facilitate the interpretation of the regression coefficients. In Table 12, we report the results using logit model and find consistent results. The results suggest that our main findings are not sensitive to regression models.

The COVID-19 pandemic has significantly impacted businesses worldwide, with many experiencing a sharp decline in sales (Rababah et al. 2020). Notably, there has been a shift towards online sales, particularly during 2020 (Roggeveen and Sethuraman 2020). To ensure our results are not influenced by this factor, we exclude data from 2020 and re-evaluate our findings. In Table 13, we observe that our results remain consistent with our previous findings, suggesting that they are not solely driven by the effects of COVID-19.

TABLE 13 | Robustness tests: Covid-19 influence.

	(1)	(2)	(3)
	FRAUD	**FRAUD**	**FRAUD**
DIFF	0.027***	0.031***	0.028***
	(6.48)	(9.98)	(6.77)
POST	−0.032**		
	(−2.63)		
DIFF×POST	**−0.025*** **	**−0.025*** **	**−0.025*** **
	(−5.29)	**(−6.44)**	**(−3.74)**
SIZE			−0.012
			(−1.32)
LEV			0.177
			(1.22)
MTB			0.001
			(0.13)
LOSS			0.169
			(1.69)
Non-BIG10			0.038***
			(4.28)
VOL			6.193
			(1.43)
ANAL			−0.049***
			(−4.20)
MAO			0.250***
			(2.90)
Constant	0.224***	0.214***	0.361*
	(23.76)	(69.35)	(1.71)
Ind and Year FE	No	Yes	Yes
Observations	484	484	484
R^2	0.007	0.019	0.144

Note: This table presents main regression results using the sample period from 2017 to 2019 to avoid the influence of Covid-19 in 2020. *FRAUD*, indicator variable equals one if a firm committed a fraud related to financial reporting in a given year, and zero otherwise. *DIFF*, the difference between a firm's sales growth in financial statements and the online sales growth. *POST*, indicator variable equals one if the year is 2019 when alternative data are available. Detailed definitions of all variables are provided in Appendix A. The *t*-statistics in parentheses are based on standard errors clustered at the industry-year level. The superscripts ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

5 | Conclusion

In this study, we first examine the association between financial statement fraud and the availability of alternative data derived from online consumer transactions. We show that online sales data independently collected by third-party vendors can detect financial statement fraud in the e-commerce industry before 2018 when the alternative data are not available to the public. One standard deviation increase of the gap between the alternative data from online sales and actual revenue from financial statements is associated with a 1.08% increase in the probability of corporate financial statement fraud, which is around 5.71% increase considering that the base rate of corporate fraud is 18.9%. Second, we leverage an exogenous event to the companies in 2018, when the WIND company made alternative online sales data available to third parties for purchase. The results imply that the association between financial statement fraud and sales growth difference becomes weaker after the publicity of alternative data. Third, we examine a possible external governance mechanism to explain the relationship between the availability of alternative data and corporate financial statement fraud. We show that the impact from alternative data availability on financial statement fraud is greater among firms with less analyst coverage, as well as lower quality auditors. At the end, we provide indirect supporting evidence that investors value the information from this alternative data source as the market shows positive reaction to the introduction of alternative data.

Based on the above, our study proposes a new information intermediary, alternative data provider, and illustrates the positive role of alternative data in detecting corporate financial statement fraud while providing supporting evidence that alternative data can act as an external governance mechanism in mitigating financial statement fraud in the firms with relatively less external governance systems. Second, we contribute to the growing literature on the impact of technology on financial statement fraud and corporate governance. In recent papers, multiple technological innovations have been documented in the context of the capital market. We provide new insights and supporting evidence that the introduction of alternative data from online sales plays an important role in mitigating financial statement fraud. Third, our findings contribute to research in forensic economics in general and in nonfinancial measures (NFMs) in particular. Our study highlights the importance of alternative data from a separate source, such as online sales, for financial statement fraud and fraud detection. While we conduct a variety of tests to enhance the validity of our results, we still use a single data source to measure online sales activity. Considering the emergence of big data and alternative data, we anticipate that future researchers will seek to examine whether other alternative data and availability of those alternative data are useful in detecting financial statement fraud and have an impact on corporate governance.

Acknowledgments

The authors make equal contributions and are listed in alphabetical order. Le is grateful for financial support from the National Natural Science Foundation of China (Grant No. 72402105).

Ethics Statement

The authors have nothing to report.

Consent

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from CSMAR and WIND. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the authors with the permission of CSMAR and WIND.

Endnotes

[1] Available to the public:WIND subscribers can browse, download, and use online sales data for free. This data was not available to the public before December 2018.

[2] https://www2.deloitte.com/content/dam/Deloitte/cn/Documents/consumer-business/deloitte-cn-cb-consumer-goods-and-retail-industry-overview-zh-210922.pdf.

[3] However, unlike upward earnings manipulation, such as inflating revenue, downward manipulation is generally more acceptable to external auditors and regulatory authorities in China (Lennox, Wu, and Zhang 2016). As a result, firms engaging in downward performance manipulation face a lower likelihood of penalties (Kinney and Martin 1994).

[4] We carefully review Kim and Klein (2017) and believe that our setting adheres to the four foundational assumptions of the event study methodology.

[5] Untabulated results show a significant increase for companies that committed accounting manipulation in the past (2.19% vs. 1.26%). However, this increase was not statistically significant for companies with fewer analysts or less reputable auditing firms. A possible explanation for this is that investors may prioritize concerns about a company's history of fraud over factors like analyst coverage and auditor reputation (Brazel et al. 2015).

References

Agarwal, S., W. Qian, and X. Zou. 2021. "Disaggregated Sales and Stock Returns." *Management Science* 67: 7167–7183.

Allee, K. D., B. Baik, and Y. Roh. 2021. "Detecting Financial Misreporting With Real Production Activity: Evidence From an Electricity Consumption Analysis." *Contemporary Accounting Research* 38: 1581–1615.

Allen, F., J. Qian, and M. J. Qian. 2005. "Law, Finance, and Economic Growth in China." *Journal of Financial Economics* 77: 57–116.

Amiram, D., Z. Bozanic, J. D. Cox, Q. Dupont, J. M. Karpoff, and R. Sloan. 2018. "Financial Reporting Fraud and Other Forms of Misconduct: A Multidisciplinary Review of the Literature." *Review of Accounting Studies* 23: 732–783.

Beyer, A., D. A. Cohen, T. Z. Lys, and B. R. Walther. 2010. "The Financial Reporting Environment: Review of the Recent Literature." *Journal of Accounting and Economics* 50, no. 2–3: 296–343.

Blankespoor, E., E. deHaan, and C. Zhu. 2018. "Capital Market Effects of Media Synthesis and Dissemination: Evidence From Robo-Journalism." *Review of Accounting Studies* 23, no. 1: 1–36.

Blankespoor, E., B. E. Hendricks, J. Piotroski, and C. Synn. 2022. "Real-Time Revenue and Firm Disclosure." *Review of Accounting Studies* 27: 1079–1116.

Brazel, J. F., K. L. Jones, J. Thayer, and R. C. Warne. 2015. "Understanding Investor Perceptions of Financial Statement Fraud and Their Use of Red Flags: Evidence From the Field." *Review of Accounting Studies* 20: 1373–1406.

Brazel, J. F., K. L. Jones, and M. F. Zimbelman. 2009. "Using Nonfinancial Measures to Assess Fraud Risk." *Journal of Accounting Research* 47: 1135–1166.

Breuer, M., and E. D. Dehaan. 2024. "Using and Interpreting Fixed Effects Models." *Journal of Accounting Research* 62: 1183–1226.

Chen, J., D. Cumming, W. Hou, and E. Lee. 2016. "Does the External Monitoring Effect of Financial Analysts Deter Corporate Fraud in China?" *Journal of Business Ethics* 134, no. 4: 727–742.

Chen, Q., X. Chen, K. Schipper, Y. X. Xu, and J. Xue. 2012. "The Sensitivity of Corporate Cash Holdings to Corporate Governance." *Review of Financial Studies* 25: 3610–3644.

Chen, T., J. Harford, and C. Lin. 2015. "Do Analysts Matter for Governance? Evidence From Natural Experiments." *Journal of Financial Economics* 115: 383–410.

Chiu, P.-C., S. H. Teoh, Y. Zhang, and X. Huang. 2023. "Using Google Searches of Firm Products to Detect Revenue Management." *Accounting, Organizations and Society* 109: 101457.

Croux, C., J. Jagtiani, T. Korivi, and M. Vulanovic. 2020. "Important Factors Determining Fintech Loan Default: Evidence From a Lending Club Consumer Platform." *Journal of Economic Behavior & Organization* 173: 270–296.

Dai, L., J. T. Parwada, and B. Zhang. 2015. "The Governance Effect of the Media's News Dissemination Role: Evidence From Insider Trading." *Journal of Accounting Research* 53, no. 2: 331–366.

Dang, C., S. Foerster, Z. F. Li, et al. 2021. "Analyst Talent, Information, and Insider Trading." *Journal of Corporate Finance* 67: 101803.

Dichev, I. D., and J. Qian. 2022. "The Benefits of Transaction-Level Data: The Case of NielsenIQ Scanner Data." *Journal of Accounting and Economics* 74: 101495.

Dimmock, S. G., W. C. Gerken, and N. P. Graham. 2018. "Is Fraud Contagious? Coworker Influence on Misconduct by Financial Advisors." *Journal of Finance* 73: 1417–1450.

Dyck, A., A. Morse, and L. Zingales. 2010. "Who Blows the Whistle on Corporate Fraud?" *Journal of Finance* 65: 2213–2253.

Dyck, A., N. Volchkova, and L. Zingales. 2008. "The Corporate Governance Role of the Media: Evidence From Russia." *Journal of Finance* 63, no. 3: 1093–1135.

Fan, J., T. Wong, and T. Zhang. 2007. "Politically Connected CEOs, Corporate Governance, and Post-IPO Performance of China's Newly Partially Privatized Firms." *Journal of Financial Economics* 84: 330–357.

Fan, J. P. H., and T. J. Wong. 2005. "Do External Auditors Perform a Corporate Governance Role in Emerging Markets? Evidence From East Asia." *Journal of Accounting Research* 43: 35–72.

Fracassi, C., and G. Tate. 2012. "External Networking and Internal Firm Governance." *Journal of Finance* 67: 153–194.

Froot, K., N. Kang, G. Ozik, and R. Sadka. 2017. "What Do Measures of Real-Time Corporate Sales Say About Earnings Surprises and Post-Announcement Returns?" *Journal of Financial Economics* 125: 143–162.

Goldstein, I., J. Jagtiani, and A. Klein. 2019. Fintech and the New Financial Landscape. Bank Policy Institute (BPI): Banking Perspectives, 1st Quarter, March.

Hendershott, T., C. M. Jones, and A. J. Menkveld. 2011. "Does Algorithmic Trading Improve Liquidity?" *Journal of Finance* 66, no. 1: 1–33.

Jagtiani, J., and C. Lemieux. 2019. "The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence From the Lending Club Consumer Platform." *Financial Management* 48, no. 4: 1009–1029.

Jensen, M. C., and W. H. Meckling. 1976. "Theory of the Firm: Managerial Behaviour, Agency Costs and Ownership Structure." *Journal of Financial Economics* 3: 305–360.

Joe, J. R., H. Louis, and D. Robinson. 2009. "Managers' and Investors' Responses to Media Exposure of Board Ineffectiveness." *Journal of Financial and Quantitative Analysis* 44, no. 3: 579–605.

Kang, J. K., L. Stice-Lawrence, and Y. T. F. Wong. 2021. "The Firm Next Door: Using Satellite Images to Study Local Information Advantage." *Journal of Accounting Research* 59: 713–750.

Kim, S., and A. Klein. 2017. "Did the 1999 NYSE and NASDAQ Listing Standard Changes on Audit Committee Composition Benefit Investors?" *Accounting Review* 92, no. 6: 187–212.

Kinney, W. R., and R. D. Martin. 1994. "Does Auditing Reduce Bias in Financial Reporting? A Review of Audit-Related Adjustment Studies." *Auditing: A Journal of Practice & Theory* 13: 149–156.

Kolanovic, M., and R. Krishnamachari. 2017. Big Data and AI Strategies: Machine Learning and Alternative Data Approach to Investing. Report, J.P. Morgan, May 18.

Lennox, C., and J. A. Pittman. 2010. "Big Five Audits and Accounting Fraud." *Contemporary Accounting Research* 27: 209–247.

Lennox, C., X. Wu, and T. Y. Zhang. 2016. "The Effect of Audit Adjustments on Earnings Quality: Evidence From China." *Journal of Accounting and Economics* 61: 545–562.

Liao, L., X. Cui, and Q. Sun. 2021. "The Information Content of Alternative Data: Evidence From E-Commerce Sales." *Journal of Management World* 37, no. 9: 90–103 (in Chinese).

Monk, A., M. Prins, and D. Rook. 2019. "Rethinking Alternative Data in Institutional Investment." *Journal of Financial Data Science* 1: 14–31.

Murphy, P. R., and M. T. Dacin. 2011. "Psychological Pathways to Fraud: Understanding and Preventing Fraud in Organizations." *Journal of Business Ethics* 101: 601–618.

Rababah, A., L. Al-Haddad, M. S. Sial, Z. Chunmei, and J. Cherian. 2020. "Analyzing the Effects of COVID-19 Pandemic on the Financial Performance of Chinese Listed Companies." *Journal of Public Affairs* 20, no. 4: e2440.

Ren, L., X. Zhong, and L. Wan. 2021. "Missing Analyst Forecasts and Corporate Fraud: Evidence From China." *Journal of Business Ethics* 181: 1–24.

Roggeveen, A. L., and R. Sethuraman. 2020. "How the COVID-19 Pandemic May Change the World of Retailing." *Journal of Retailing* 96, no. 2: 169–171.

Shipman, J., Q. Swanquist, and R. Whited. 2017. "Propensity Score Matching in Accounting Research." *Accounting Review* 92, no. 1: 213–244.

Wu, X., and J. Wang. 2018. "Management's Materiality Criteria of Internal Control Weaknesses and Corporate Fraud: Evidence From China." *International Journal of Accounting* 53, no. 1: 1–19.

Zhu, C. 2019. "Big Data as a Governance Mechanism." *Review of Financial Studies* 32: 2021–2061.

Zingales, L. 2000. "In Search of New Foundations." *Journal of Finance* 55, no. 4: 1623–1653.

Appendix A

Variable Definitions

FRAUD	Financial fraud, indicator variable equals one if a firm committed a fraud related to financial reporting (e.g., inflated profits and revenues) in a given year, and zero otherwise.
DIFF	Growth difference, the difference between a firm's sales growth in financial statements (*GROWTH*) and the online sales growth (*E_GROWTH*).
POST	Indicator variable equals one if the year is 2019 and 2020.
SIZE	The natural logarithm of total assets.
LEV	Financial leverage, calculated as total liabilities divided by total assets.
MTB	Market-to-book, calculated as the market value of equity divided by the book value of equity.
LOSS	Indicator variable equals one if the firm reports a negative net income.
Non-BIG10	Indicator variable equals one if the client firm is audited by a non-big10 auditing firm, and 0 otherwise.
VOL	Annualized standard deviation of daily stock returns.
ANAL	The natural logarithm of the number of analysts following the firm plus 1.
MAO	An indicator variable equal to one if the company is issued a modified audit opinion, and zero otherwise.
PUNISH	Fraud severity, equals one if the firm is ordered to rectify by the regulatory authority, equals two if the executives of the firm are criticized or warned by the regulatory authority, equals three if the firm and executives is publicly reprimanded by the regulatory authority, equals four if the firm is fined by the regulatory authority, and zero otherwise.
GROWTH	The percentage change in total sales reported in financial statements.
E_GROWTH	The percentage change in online sales.

Appendix B

Online Sales Growth and Reported Sales Growth

	(1)	(2)
	GROWTH	GROWTH
E_GROWTH	**0.022*** **	**0.018*** **
	(2.82)	**(2.75)**
SIZE		−0.012
		(−1.02)
LEV		0.155**
		(2.36)
MTB		0.004
		(1.42)
LOSS		−0.172***
		(−6.01)
Non-BIG10		−0.024
		(−1.37)
VOL		2.610*
		(1.71)
ANAL		0.032***
		(3.78)
MAO		−0.096***
		(−3.07)
Constant	0.095***	0.183
	(8.82)	(0.68)
Ind and Year FE	Yes	Yes
Observations	647	647
R^2	0.151	0.291

Note: This table presents the results investigating the relationship between the reported sales growth (*GROWTH*) in financial reports and the online sales growth (*E_GROWTH*). *GROWTH*, the percentage change in total sales reported in financial statements. *E_GROWTH*, the percentage change in online sales. Detailed definitions of all variables are provided in Appendix A. The *t*-statistics in parentheses are based on standard errors clustered at the industry-year level. The superscripts ***, **, and * denote statistical significance at 1%, 5% and 10%, respectively.