4-2023

Regulating by new technology: The impacts of the SEC data analytics on the SEC investigations

Tian DENG
Singapore Management University, tian.deng.2018@phdacc.smu.edu.sg

REGULATING BY NEW TECHNOLOGY: THE IMPACTS OF THE SEC DATA ANALYTICS ON THE SEC INVESTIGATIONS

TIAN DENG

SINGAPORE MANAGEMENT UNIVERSITY

2023

Regulating by New Technology: The Impacts of the SEC Data Analytics on the SEC Investigations

Tian Deng

Submitted to School of Accountancy in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Accounting

Dissertation Committee

Qiang Cheng (Chair)
Professor of Accounting
Singapore Management University

Sterling Huang
Associate Professor of Accounting
Singapore Management University

Liandong Zhang
Professor of Accounting
Singapore Management University

Fangjian Fu
Associate Professor of Finance
Singapore Management University

Singapore Management University
2023

I hereby declare that this PhD dissertation is my original work and it had been written by me in its entirety.

I have duly acknowledged all the sources of information which have been used in this dissertation.

This PhD dissertation has also not been submitted for any degree in any university previously.

Deng Tian

Tian Deng

19 Apr 2023

Regulating by New Technology: The Impacts of the SEC Data Analytics on the SEC Investigations

Tian Deng

ABSTRACT

Despite the Securities and Exchange Commission's (SEC) growing emphasis on data analytics in recent years, there is scant research about whether the investment in data analytics accomplishes its objective of enhancing enforcement efficiency. This study examines the effects of the SEC regional offices' use of data analytics on their investigation outcomes. The utilization of data analytics reduces information processing costs, thereby streamlining the enforcement process as a whole. I find that the SEC's use of data analytics is associated with a 12% increase in the SEC's investigation success rate. Such an improvement is greater for firms whose disclosure are more machine-friendly, those with a greater level of complexity, and those located further away from the SEC regional offices. Furthermore, I find that firms are less inclined to engage in fraud after the SEC's use of data analytics, probably due to a higher perceived detection likelihood. Additional tests suggest that the investigation time is shorter, and the detected fraud is more complex after the SEC's use of data analytics. Collectively, the results provide evidence that the SEC's use of data analytics increases its enforcement efficiency and deters firms' fraud behavior.

Table of Contents

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Qiang Cheng, for his continuous guidance, support, and encouragement throughout my PhD journey. He has shared his deep understanding of accounting research, encouraged me to purse novel ideas, and offered tremendous support whenever I encounter difficulties, without which this dissertation would not have been possible. His passion and commitment to research have left a substantial impact on me and will keep motivating me for the rest of my life. I am deeply honored to be his student.

I would also like to thank Professor Sterling Huang, Professor Liandong Zhang, and Professor Fangjian Fu, for their insightful comments and suggestions for my dissertation. I truly enjoy discussions with Sterling on research ideas, and his suggestions greatly help me develop my research skills and mindset. My appreciation also extends to Professor Yun Lou, Professor Rencheng Wang, Professor An-Ping Lin, and other faculty members at SMU. I am lucky to be in the first batch of PhD in Accounting at SMU, and I've learned so much from the interaction with them during classes, seminars, and meetings.

My sincere thanks also extend to my PhD friends. To Mengjie Yang, who is the best officemate I could ask for, and I enjoyed every walk and meal we had together. To Shuo Li, I will miss the time we spent together hiking, climbing, and playing squash. To Amanda Aw Yong, who shared the PhD journey with me throughout.

Finally, my most heartful appreciation goes to my parents for their tremendous support, endless care, and unconditional love. To Dr. Li, for accompanying me, through the ups and downs.

1. Introduction

There is a long-standing interest in understanding the effectiveness of the monitoring role of the Securities and Exchange Commission (SEC). Prior studies show that the effective enforcement of securities regulations plays an important role in financial reporting outcomes (Holthausen 2009; Kedia and Rajgopal 2011; Blackburne 2014). With advancements in new technologies, the SEC has increasingly leveraged data analytics to enhance enforcement efficiency (SEC 2018, 2019, 2020). Despite the growing emphasis on data analytics by the SEC, there is little research on whether the SEC's investment in data analytics achieves its stated objective of enhancing enforcement efficiency, and more importantly, whether it impacts firms' fraud behavior. In this paper, I fill in this gap by examining the effects of the SEC's use of data analytics on both SEC enforcement efficiency and firms' fraud decisions.

The use of data analytics has been emphasized by the SEC and is becoming increasingly important in recent years.[1] Among all federal agencies, the SEC is the second in applying analytics tools (Engstrom, Ho, Sharkey, and Cuellar 2020). The word "*Analytics*" is mentioned four times in the SEC's 2010 annual report but 45 times in its 2015 annual report. In 2020, the SEC appointed its first Chief Data Officer,[2] and the responsibility of this new position is to help "develop the SEC's data management strategy and priorities, enable data

[1] According to the National Institute of Standards and Technology (2015), "Data analytics" including two parts: data management and analytics. Data management consists of technologies and processes that deliver accurate and reliable data to support a range of functions and promotes integrity, completeness, and accuracy throughout the data life cycle, including acquisition, storage, maintenance, access, use, and disposal. Analytics refers to the discovery of meaningful patterns in data and is the process of synthesising knowledge from information. In my paper, I employ this broad definition of data analytics, which will include the use of machine learning and big data in assisting analytics work.

[2] *Austin Gerig Named as SEC's Chief Data Officer* (Jan 2020), available at https://www.sec.gov/news/press-release/2020-11.

analytics to support enforcement, examinations, and policymaking". Based on the budget information released by the SEC, in 2020, it allocates approximately $120 million towards data management and about $20 million on analytics, which in total accounts for 7% of its annual budget (SEC 2020). As suggested by the SEC, "Long-term investment and development in technology and analytical tools will be critical to the future success of the Commission's oversight responsibilities. These tools will provide the staff with a greater ability to monitor trends and emerging risks, ultimately enabling the staff to allocate SEC resources more effectively (SEC 2018, p.114)".

The primary driving force behind the SEC's investment in data analytics is to enhance its work efficiency in light of resource constraints. Over the past decade, the SEC officials contend that resource constraints have hindered their ability to efficiently investigate and prosecute all the committing misconduct (Thomsen 2009). Prior research has also provided evidence that resource constraints manifest as an important impediment to the regulatory process (Kedia and Rajpoal 2011; Bonsall, Holzman, and Miller 2019; Gunny and Hermis 2020; Ege, Glenn, and Robinson 2020; Hills, Kubic and Mayew 2021). Considering the SEC's prioritization of data analytics and its crucial role in upholding fair and orderly capital markets, it is imperative to examine the impacts of the SEC's use of data analytics.

I first examine whether the SEC's use of data analytics affects its enforcement efficiency. The use of data analytics has drastically reduced the costs of information processing, facilitating the SEC's detection of misbehavior despite its resource constraints. First, data analytics reduces information acquisition costs. Corporate disclosures are becoming increasingly lengthy and

complex over time (Cohen, Malloy, and Nguyen 2020), making the whole corpus of firm disclosures beyond the processing ability of human brains. Data analytics techniques, including machine learning algorithms that gather both quantitative and qualitative data from multiple sources, enable the SEC to reveal patterns that are otherwise difficult to detect. Second, data analytics reduces information integration costs (Blankespoor, deHaan, and Marinovic 2020). The more advanced data analytics methods allow the SEC to incorporate high-dimensional data and more sophisticated models into their analysis, thereby enhancing its ability to identify misbehavior. As a result, the decreased information processing costs facilitate the enforcement process by (1) discovering the most suspicious activities and selecting the cases that have a higher probability of involving fraud, and (2) collecting useful evidence more efficiently during the formal investigation process. Therefore, I predict that the SEC's use of data analytics is associated with higher enforcement efficiency.

Utilizing a dataset from Blackburne, Kepler, Quinn, and Taylor (2021) that contains all SEC investigations conducted by the SEC regional offices from 2008 to 2017, I operationalize SEC enforcement efficiency by using the investigation success rate. Specifically, I examine whether the SEC regional offices' use of data analytics is associated with a higher investigation success rate. A successful investigation is an investigation that concludes with an enforcement action. To capture the use of data analytics by the SEC, I use the regional offices' job postings that require data analytics skills, which provides me with both time-series and cross-sectional variations for firms under the

jurisdiction of different regional offices.[3] Consistent with my prediction, I find that the regional offices' use of data analytics is positively associated with the investigation success rate. The economic magnitude is also significant. Specifically, the regional offices are 12% more likely to conduct a successful investigation after the use of data analytics, which doubles the overall success rate of all investigations in my sample.

I next examine whether the SEC's use of data analytics affects firms' fraud decisions. The seminal work of Becker (1968) describes the deterrence effects of fraud detection. According to Becker (1968), the decision to commit fraud depends on its expected benefits and costs, and the probability of committing fraud is increasing in the expected benefits and decreasing in the expected costs of fraud. To the extent that firms are aware of the SEC's use of data analytics and thus expect a higher probability of detection, they are less inclined to engage in fraudulent activities. However, recent theory work shows that strengthening enforcement could have unintended consequences that lead to a higher probability of fraud. Specifically, Samuels, Taylor, and Verrecchia (2021) argue that on the one hand, high levels of public scrutiny facilitate monitoring, suggesting a negative relation between enforcement and fraud, but on the other hand, public scrutiny also increases the weight that investors place on earnings in valuing the firm, in turn increasing the benefits of fraud and suggesting a positive relation. Therefore, it is an empirical question whether the SEC's use of data analytics affects firms' fraud decisions.

[3] The SEC's enforcement activities are conducted by the Division of Enforcement, which has various regional offices across the country. Please see Section 2.1 for more details about the SEC's organization structure and enforcement process.

Empirical analysis of the effect of SEC scrutiny on fraud decisions faces the challenge of partial observability (Wang, Winton, and Yu 2010; Barton, Burnett, Gunny, and Miller 2022). Specifically, I do not observe all the fraud that has been committed, but only the ones that have been committed and subsequently detected. Using an ex post measure of detected fraud would not allow me to accurately quantify how the firms' propensity to commit fraud changes after observing the SEC's use of data analytics. To examine the effect of the SEC's use of data analytics on firms' fraud decisions, I measure the underlying probability of fraud using *F-score* – a measure of ex ante fraud probability (Dechow, Ge, Larson, and Sloan 2011; Berge and Lee 2022). The *F-score* is calculated using a prediction model based on various financial statement variables, and a higher *F-score* is associated with a higher probability of fraud. I find that the *F-score* is lower after the SEC regional offices' use of data analytics, suggesting that firms' fraud incentives are deterred by the SEC regional offices' use of data analytics.

While it is very difficult to precisely measure the usage of data analytics at the regional office, I conduct some additional tests to complement my main measure which uses job posting data. First, the main analyses use lagged one-year job posting to capture the actual hire of employees with data analytics skills. It's also possible that there is a longer gap between job advertising and actual hire. To examine this possibility, I construct another measure, *SEC data analytics_3year*, which identifies data analytics-related job postings in any of the previous three years, and find consistent results using this alternative measure. Second, a job posting does not necessarily lead to successful hiring, and it depends on both the demand and supply of talent. To complement the job

posting measure, I also measure SEC data analytics using a sample of SEC employees that have resume and skill information. I find that the resume-based measure is positively correlated with the job posting measure, suggesting that the skills listed in the job posting reasonably mirror the skills of individuals working at the SEC.[4]

I further explore the circumstances under which the SEC's use of data analytics plays a bigger role in improving enforcement efficiency. First, I find that the effect of data analytics on enforcement efficiency is stronger for firms whose disclosures are more machine-friendly, indicating that the SEC is indeed utilizing data analytics techniques in its enforcement activities. I measure disclosure machine-friendly by using the *Disclosure Scriptability* measure constructed by Allee, DeAngelis, and Moon (2018), which captures the ease with which a filing can be processed and parsed by an automated program. Second, I predict that the impact of data analytics on enforcement efficiency is stronger for complex firms because the reduction in information processing costs is greater for complex firms with the help of data analytics compared to other firms. Using the number of business segments to capture firm complexity, I find that the effect of data analytics on enforcement efficiency is mainly driven by complex firms. Third, I examine the effect of geographical proximity between firms' headquarters and the SEC regional offices on the relation between the use of data analytics and enforcement efficiency. Prior research finds that the SEC is more likely to investigate the proximate firms due to lower enforcement costs (Kedia and Rajgopal 2011). The use of data analytics enables

[4] Section 5.3.2 discuss the advantages and caveats of this resume-based measure in more detail.

the SEC to collect and analyze hard information more efficiently for geographically distant firms, which in turn results in lower enforcement costs and a higher investigation success rate for such firms. Using the geographical distance between firms' headquarters and the corresponding regional offices, I find results suggesting that the effect of data analytics on enforcement efficiency is stronger for distant firms than for proximate firms.

In the additional tests, I examine the impacts of the SEC's use of data analytics on other investigation outcomes. First, I find that the investigation time is shortened after the SEC's use of data analytics, consistent with my hypothesis that the use of data analytics facilitates the formal investigation process. Second, I investigate whether the nature of the detected fraud changes after the SEC's use of data analytics. The detected fraud is the outcome of two stages: fraud occurrence and fraud detection. For the fraud occurrence, as the expected costs of committing fraud are higher given the higher detection likelihood, the expected benefits should also be higher to induce the managers to engage in fraud, resulting in more severe and sophisticated committed fraud. For the fraud detection, with the use of data analytics, the decrease in information processing costs is larger for the more complex cases. As a result, the observed detected fraud is more severe and complex. To measure the severity and complexity of fraud, I use the number of defendants for each case and the number of alleged violations for each defendant following Kalmenovitz (2021) and Zheng (2021). Consistent with my prediction, I find that the detected fraud is more complex and severe after the SEC's use of data analytics.

To better understand the decision to use data analytics by the regional offices, I explore several characteristics of the regional offices that may be

related to their decision to use data analytics. I collect and analyze two sets of variables. The first category measures the size and capacity of the regional offices, including the budget, the number of employees, and whether there are leadership changes at the regional offices. The second category captures the local economic conditions, including GDP per capita, and the unemployment rate of the local state. The results suggest that regional office budget is positively associated with the use of data analytics, which is consistent with the idea that the regional offices need a sufficient budget to invest in data analytics. I also find that the number of employees at the regional offices is negatively associated with the use of data analytics, suggesting that a lack of human resources may affect the use of data analytics to increase work efficiency. I control for these potential determinants of the SEC's use of data analytics and find that the effects of the SEC's use of data analytics remain.

This paper contributes to two streams of literature. First, it contributes to the literature about the SEC enforcement process (Kubic and Rajgopal 2011; Correia 2014). There is a large literature studying the enforcement role of the SEC because it oversees the U.S. capital markets with the stated objective of ensuring fair and orderly capital markets. Given the limited resources that potentially constrain the SEC's ability to achieve its goal and the advancement of new technologies, it is crucial to assess whether the use of new technologies enhances enforcement efficiency, and to what extent data analytics contributes to the improvement of enforcement efficiency. To my knowledge, this is the first paper that examines the impacts of the SEC's use of data analytics. The findings of this study also have implications for the SEC's budget allocation and other regulatory agencies. The SEC's Director of Enforcement, Stephanie

Avakian, recently discussed the Division's focus on shortening investigations and stated that the SEC "will continue to look for ways to accelerate the pace of its investigations (SEC 2020, p.129)". The findings of this study suggest that the use of data analytics could be one avenue for the regulators to pursue to achieve these goals.

Second, this paper contributes to the fraud literature. The majority of the previous literature has focused on how executive compensation (Burns and Kedia 2006; Goldman and Slezak, 2006; Armstrong, Jagolinzer, and Larcker, 2010) or corporate governance (Dechow, Sloan, and Sweeny, 1996; Lennox and Pittman, 2010) is associated with corporate fraud, while there has been little focus on the monitoring role of the SEC (Kedia and Rajgopal, 2011; Blackburne, 2014). I provide evidence that regulatory oversight is an important mechanism that has a deterrence effect and influences firms' reporting decisions.

2. Background

2.1 SEC regional offices and the enforcement process

Created by the Securities Exchange Act of 1934, the Securities and Exchange Commission (SEC) is an independent federal agency that enforces federal securities law. Since 2007, the SEC has had 11 regional offices that located in Atlanta, Boston, Chicago, Denver, Fort Worth, Los Angeles, Miami, New York, Philadelphia, Salt Lake City, and San Francisco (SEC, 2007). Figure 1 illustrates the location of the 11 regional offices and their areas of jurisdiction. The SEC states that the primary aim of enforcement is to "protect investors and the markets by investigating potential violations of the federal securities laws and litigating the SEC's enforcement actions (SEC 2017)". The SEC utilizes the

11 regional offices to carry out this goal, as these regional offices handle most of the investigations of potential violations. Each regional office is headed by a regional director and is usually staffed with enforcement attorneys, accountants, investigators, and compliance examiners.[5] Table 1 shows the number of employees, the number of attorneys, and the annual budget for each office.[6] In terms of staffing, the New York regional office has the largest number of employees, followed by the Chicago regional office. The Salt Lake City regional office is the smallest in terms of personnel.

Figure 2 outlines the typical SEC enforcement process (SEC Enforcement Manual 2017; Blackburne and Quinn 2023; Blackburne, Kepler, Quinn, Taylor 2021). As shown in Figure 2, the enforcement process begins with a trigger event when the SEC enforcement staff receives a "lead" about possible violations of securities laws. These leads can come from a variety of sources, both internally and externally. Internally, the SEC conducts its own surveillance activities, such as reviewing corporate filings and trading data. Externally, the SEC receives tips or complaints from whistleblowers (e.g., investors, former or current employees), and referrals from other regulatory organizations such as securities exchanges and foreign regulatory authorities (O'Malley, Harnisch, and Umayam 2007). Once an office receives a lead, the assigned assistant regional director and staff members would perform an initial evaluation of the credibility and severity of the potential violation and recommend whether a lead is promising enough to become a matter-under-inquiry (MUI), and a regional director would then decide whether to open an

[5] Note that the widely studied comment letters are conducted at the Division of Corporate Finance, which is located at the SEC's headquarter, Washington D.C.
[6] I thank Joseph Kalmenovitz for sharing the SEC employee data.

MUI. Within 60 days of initiating a MUI, the SEC staffs collect additional information and consult with the associate director to determine whether to convert the MUI to a formal investigation.

Oftentimes the formal investigations will involve the issuance of a formal order by the SEC, in which the investigative staff handling the investigation acquire subpoena power, allowing the staff to require the production of documents and appearance of witness for testimony under oath.[7,8] The evidence is typically gathered through interviewing witnesses and examining company records and relevant data. If the firm does not submit an offer of settlement, the staff will summarize their findings and make an enforcement recommendation to the head of the Division of Enforcement. If the firm submits an offer of settlement, the staff will also make a recommendation regarding the settlement. Next, the head of the Division of Enforcement will take the recommendations to the five commissioners at the SEC headquarter, who will jointly decide the enforcement outcome in a closed meeting. If the commissioners believe there is sufficient useful evidence to litigate, then there will be an enforcement action via either civil action in federal courts or internally through administrative proceedings.

2.2 The use of data analytics at SEC

The SEC has been increasingly relying on data analytics to guide its operational activities and to make more informed and effective decisions.

[7] SEC investigations are confidential "to preserve the integrity of its investigative process as well as to protect persons against whom unfounded charges may be made or where the SEC determines the enforcement action is not necessary or appropriate" (SEC 2020), and law enforcement exemptions in the Freedom of Information Act (FOIA) allow the SEC to withhold information about them.

[8] For example, the SEC may send the target firm a Wells Notice, which outlines why it intends to purse an enforcement action, and the firm under investigation is given the opportunity to respond by rebutting violation charges. The firm's response is known as a "Well submission". If the staffs agree with the rebuttal, the matter is closed.

According to the budget information released by the SEC, it spent about $120 million on data management and about $20 million on analytics in 2020, which in total accounts for 7% of the annual budget (SEC 2020). In addition, the SEC's Strategic Plan for the year 2018-2022 continues to emphasize enhancing analytics of market and industry data to prevent, detect, and prosecute improper behavior (SEC 2018). Data analytics has been helpful in identifying promising leads, conducting investigations, and litigating cases. As suggested by the SEC chairs, Mary Jo White, and Jay Clayton, the use of data analytics is critical to maximizing the SEC's limited resources and developing more effective and efficient enforcement programs (Mary Jo White 2016; Jay Clayton 2019).

The SEC has undertaken several initiatives in the past decade to enhance its utilization of data analytics. At its headquarter in Washington D.C., the SEC created the Division of Economic and Risk Analysis (DERA) in 2009. The DERA employs economists, analysts, data scientists, computer engineers, and statisticians, with the aim to "integrate financial economics and rigorous data analytics into the core mission of the SEC." The DERA interacts with all other SEC divisions and offices by providing economic analyses, data, and insights from research to support the agency's policymaking and enforcement actions. It develops customized analytics tools and analyses to proactively detect risks that could indicate possible violations of federal securities laws. For example, developed by the DERA and launched in 2012, the Accounting Quality Model (AQM) is designed to provide a set of quantitative analytics to assess the degree to which the registrants' financial statements appear anomalous. Corporate filings are processed and assigned a risk score by AQM within 24 hours of filing with the SEC. The risk score can provide promising leads to the enforcement

staff and assist them in further investigations. The AQM is further developed into the Corporate Issuer Risk Assessment (CIRA) in 2015 which helps the enforcement staff identify trends and aberrational reporting by public companies.[9]

In addition to the DERA, the Division of Enforcement (DOE) has also developed its data analytic capabilities. For example, the Center for Risk and Quantitative Analytics was established in July 2013 to support and coordinate the DOE's data analytic activities, assist staff in conducting risk-based investigations, and develop methods of monitoring signs of possible wrongdoings. One analytics tool they developed is the *EPS initiative*, which uses data analytics to uncover potential accounting and disclosure violations caused by earnings management practices. Such a tool has helped the DOE in identifying fraud in many cases. The regional offices are also developing their data analytics capability in recent years. Appendix A shows a few examples of job postings from the regional offices that require data analytics skills. For example, the New York regional office is hiring a *Quantitative Research Analyst* to collect and analyze large volumes of structured and unstructured data and is required to have a strong background in machine learning, statistics, as well as experience in creating predictive analytics on noisy data. The recruitment of employees skilled in data analytics strengthens the regional offices' capabilities to conduct investigations in the rapidly evolving capital market.

[9] CIRA is a dashboard of 200 metrics that are used to detect anomalous patterns in financial reporting. For example, CIRA enables the staff to look at how inventory at a manufacturing company is moving relative to reported sales. The SEC staff who saw increased inventory and declining sales may flag the company as ripe for fraudulent accounting adjustments.

Some prominent cases are brought with the help of data analytics tools over the years. For example, On September 2020, the SEC announced a settlement action against *Interface*, a Georgia-based modular carpet manufacturer who reported its EPS improperly to inflate its revenue and stock price. The action was arising from an investigation generated by the *EPS initiative* that identified the accounting adjustments that were not compliant with GAAP. These adjustments were made when Interface's internal forecasts indicated that the company would likely fall short of analyst consensus EPS estimates. The SEC has also brought about significant trading-related cases that may not have been possible without its ability to analyze voluminous amounts of data, including trading data and communications metadata. One prime example is SEC v. *leremenko*, which the SEC filed in January 2019. In this case, the SEC filed charges against nine defendants for their alleged roles in a scheme to hack into the SEC's EDGAR system and extract nonpublic information for use in illegal trading. As DOE noted in its 2019 Annual Report, the case required: "…painstaking analysis of numerous events in which the defendants allegedly traded during the window between when the material nonpublic information was extracted and when it was disseminated to the public, and it showcased a number of [the SEC's] complex analytic tools and capabilities. Market and trading specialists, using proprietary systems, identified suspicious trading in advance of more than 150 announcements. Through statistical analyses, the staffs determined that the odds the defendants would have randomly chosen to trade in front of these disparate events ranged from less than 7 in 10 million to less than 1 in 1 trillion". These anecdotal cases suggest that

the use of data analytics has helped the SEC to identify potential violations and facilitate the investigation process.

3. Literature Review and Hypothesis

3.1 Literature review

This paper is related to a growing literature about the SEC oversight and enforcement. Prior research has shown that factors such as location, political connections, lawyers' career concerns, voters' interests, and case materiality can affect whether and how the SEC carries out enforcement actions (Kedia and Rajgopal 2011; Correia 2014; deHaan, Kedia, Koh, and Rajgopal 2015; Heese 2019; Bonsall, Holzman, and Miller 2019; Zheng 2021). For example, given the resource constraints at the SEC, Kedia and Rajgopal (2011) find that the SEC is more likely to investigate firms located closer to its regional offices because of (1) reduced travel time for proximate firms, (2) a greater familiarity and knowledge about firms that are closely located, and (3) a higher likelihood of receiving tips as employees of proximate firms are likely to be aware of the SEC. Bonsall, Holzman, and Miller (2019) suggest that a high office case backlog decreases the likelihood of opening an SEC investigation. Prior studies examining the comment letter process provide similar evidence about the resource constraints at the SEC (Ege, Glenn, and Robinson 2020; Gunny and Hermis 2020; Hills, Kubic, and Mayew 2021). A growing literature also studies how the internal organizational design of the SEC impacts the enforcement process. For example, Kalmenovitz (2021) finds that tournament incentives, as reflected in hierarchical pay gaps and promotion opportunities inside the SEC, affect enforcement activities and enforcement outcomes. Kubic (2021)

documents a positive association between comment letter review team size and the error detection rates, and this association is driven by the number of accountants in the review team. My study contributes to the literature by focusing on the use of data analytics by the SEC, which has been the main focus of the SEC in recent years and has not been examined by any prior studies.

This paper is also related to the broad fraud literature. The majority of previous studies has focused on how executive compensation (Burns and Kedia 2006; Goldman and Slezak, 2006; Armstrong, Jagolinzer, and Larcker, 2010) or corporate governance (Dechow, Sloan, and Sweeny, 1996; Lennox and Pittman, 2010) is associated with corporate fraud, while a few focus on the monitoring role of the SEC (Kedia and Rajgopal, 2011; Blackburne, 2014). I contribute to this literature by documenting the deterrence effect of the SEC's use of new technology that affects the firms' incentives to commit fraud.

3.2 Hypothesis development

As shown in Figure 3, I model the observed fraud as an outcome of two latent processes: firms' fraud occurrence and the SEC's fraud detection. I'm interested in the effects of the SEC's use of data analytics at both stages. The theoretical work on fraud has emphasized the strategic interdependence between a firm's fraud decision and a monitor's effort. For example, the theoretical work on corporate fraud, such as Bar-Gill and Bebchuk (2003), Goldman and Slezak (2006), Noe (2008), Povel, Singh, and Winton (2007), and Stein (1989), models the interdependence between firm managers and shareholders. The firm's fraud decision depends on its assessment of the likelihood of being caught and the monitor's decision to investigate a firm depends on the likelihood that the firm has committed fraud. In my paper, I

focus on the monitoring role of the SEC, and by considering the interdependent relationship between firms and the SEC, I examine whether and how the SEC's use of data analytics affects both the SEC enforcement process and firms' fraud occurrence likelihood.

I first examine whether and how the use of data analytics affects the SEC's enforcement efficiency, and I measure enforcement efficiency by using the investigation success rate. As illustrated in Section 2.1 and Figure 2, an investigation can conclude with or without an enforcement action, and I define a successful investigation as an investigation that concludes with an enforcement action. A successful investigation suggests that the SEC has chosen the right case to investigate and collected enough evidence during the investigation process so that an investigation is concluded with an enforcement action, but not closed without further actions. In this sense, a successful investigation means that the SEC has allocated its constrained resources efficiently.

The utilization of data analytics by the SEC has drastically reduced the costs of information processing, facilitating it to detect misbehavior despite its resource constraints. First, data analytics reduces information acquisition costs. Corporate disclosures are becoming increasingly lengthy and complex over time (Cohen, Malloy, and Nguyen 2020), making the whole corpus of firm disclosures beyond the processing ability of human brains. Data analytics techniques, including machine learning algorithms that gather both quantitative and qualitative data from multiple sources, enable the SEC to reveal irregularities patterns that are otherwise difficult to detect. Second, data analytics reduces information integration costs (Blankespoor, deHaan, and

Marinovic 2020). The more advanced data analytics methods allow the SEC to incorporate high-dimensional data and more sophisticated models into their analysis, thereby enhancing its ability to unravel complex patterns of fraudulent behavior. For example, Bao, Ke, Li, Yu, and Zhang (2020) show that a more advanced machine learning model can extract more useful information from raw financial data. As a result, the utilization of data analytics enables the SEC to identify the cases that have a higher probability of being detected and increase the work efficiency during the formal investigation process, eventually leading to a higher investigation success rate. Below I illustrate how the use of data analytics can improve the SEC's work efficiency throughout the enforcement process.

Before the formal investigation starts, the SEC regional offices need to carefully evaluate all the "leads" and choose the right case to further investigate. The use of data analytics can help the SEC discover the most suspicious activities and thus divert their resources to cases that have a higher probability of being detected.[10] Internally, the use of data analytics enables the SEC to conduct its own surveillance activities more efficiently. For example, the use of data analytics helps the SEC review corporate filings more efficiently and spot the potential misconduct. Externally, the SEC utilizes data analytics to review the tips, complaints, and referrals (TCR) received and determine whether they should be further investigated. Every year the SEC received more than 10,000 TCRs that needed to be reviewed. With the help of data analytics, the SEC can efficiently and thoroughly analyze the information from these external

[10] SEC suggests this point in their annual report that "the increasing use of sophisticated analytic tools that identify suspicious patterns and activities, allowing enforcement to more quickly identify and purse unlawful conduct" (SEC 2011).

resources and conduct further investigation if needed.[11] By utilizing data analytics to evaluate both internal and external leads, the SEC is able to select the case that has a higher probability of detection to conduct a formal investigation.

After selecting the cases to formally investigate, the SEC staffs continue to collect more evidence until they can conclude the investigation. The use of data analytics can improve the SEC's work efficiency during the formal investigation. Specifically, data analytics enable the SEC to electronically retrieve and organize an extraordinary volume of documents obtained during the investigation process, and the decreased information processing costs help the SEC analyze and collect more useful evidence. For example, the use of data analytics helps the SEC analyze the mass trading data and identify the illegal trading behavior. More useful evidence collected during the investigation process increases the probability of successful detection.

In sum, as a result of decreased information processing costs, the SEC's use of data analytics improves work efficiency both before and during the formal investigation process, which in turn leads to a higher probability of a successful investigation. Based on the above arguments, I state my first hypothesis as follows:

H1: The SEC's use of data analytics is positively associated with the investigation success rate.

[11] For example, the SEC is using natural language processing tools such as Latent Dirichlet Allocation (LDA) to analyze the information in the TCRs.

I then examine whether the SEC's use of data analytics affects firms' fraud decisions. The conventional wisdom is that more intense oversight or greater enforcement dampens firms' fraud incentives. The seminal work of Becker (1968) describes the deterrence effects from the regulators' fraud detection. According to Becker (1968), the decision to commit fraud depends on its expected benefits and cost, and the probability of committing fraud is increasing in the expected benefits of fraud and decreasing in the expected costs of fraud. The expected benefits can come from the equity incentives (Goldman and Slezak 2006) and the need for growth and external financing (Dechow, Ge, Larson, and Sloan 2011). The expected costs of fraud are determined by the probability of being detected and the penalty upon detection. If a factor can affect the probability of fraud detection and if its effect can be anticipated when the fraud decision is made, then this factor should affect the probability of fraud occurrence in the opposite direction. This is the deterrence effect of fraud detection. Theory also suggests that managers will adjust their behavior in response to changes in the SEC oversight if they can either anticipate it or observe it because the potential cost of fraud is changed (Fischer and Verrecchia 2000). The SEC's use of data analytics increases the detection likelihood conditional on the fraud occurrence. Managers and the general counsel can get the information about the SEC activities from various sources, including the SEC's website, the SEC's social media account, personal networks between the general counsel and the SEC staff, and other sources. For example, Lin (2021) finds that a majority of the executives at the S&P 1500 firms follow the SEC's Twitter account and get the relevant information. The firms observe the SEC's use of data analytics and thus perceived a higher probability of being detected

if they commit fraud, i.e., the cost of committing fraud is higher, and thus they are less likely to commit fraud. This is related to perceptual deterrence, which refers to the behavior adjustment of offenders after observing changes in policing (Apel 2013).

However, recent theory work shows that strengthening enforcement could have unintended consequences that may even lead to a lower reporting quality. Samuels, Taylor, and Verrecchia (2021) argue that on the one hand, high levels of public scrutiny facilitate monitoring, suggesting a negative relation between scrutiny and misreporting. On the other hand, public scrutiny also increases the weight that investors place on earnings in valuing the firm, in turn increasing the benefit of misreporting and suggesting a positive relation.

Based on the above arguments, the effect of the SEC's use of data analytics on firms' fraud occurrence likelihood is unclear ex ante. So I state my second hypothesis as follows in the null form:

H2: The SEC's use of data analytics is not associated with firms' fraud occurrence likelihood.

4. Research Design and Data

4.1 Research design

To examine the effect of SEC regional offices' use of data analytics on the investigation success rate, I estimate the following equation using ordinary least squares (OLS) regression with standard errors clustered by regional office:[12]

[12] I use OLS (i.e., a linear probability model) to estimate this equation to facilitate coefficient interpretation and the usage of fixed effects. The inferences are similar if I use logistic regression.

$$Detect_{i,j,t} = \beta_0 + \beta_1 SEC\ data\ analytics_{i,t-1} + Office\ Controls + Firm\ Controls + Year\ FE + Office\ FE + Industry\ FE + \varepsilon_{i,j,t} \quad (1)$$

where *i*, *j*, *t* denote regional office *i*, firm *j*, and year *t*, respectively. The dependent variable in Equation (1) is *Detect*, which is defined as one if an investigation opens in a year that leads to an enforcement action in later years, and zero otherwise. The variable of interest is *SEC data analytics*, which is defined as one if a regional office is using data analytics in a year, and zero otherwise. I measure the use of data analytics at the regional office level to capture cross-sectional variations for firms under the jurisdiction of different regional offices. Specifically, I use the job postings from the regional offices that require data analytics skills to measure the utilization of data analytics at the regional office level. Section 4.2.2 describes the details of the variable measurement. Consistent with the use of data analytics decreasing information processing costs and increasing enforcement efficiency, I expect β_1 to be significantly positive.

I control for several regional-office-level factors that potentially affect the investigation outcome. Specifically, I control for the annual budget (*Budget*), the number of employees (*N_employee*), regional director change (*Leadership change*), and legal expertise (*SEC legal expertise*) at the regional office level. I measure *Budget* by using the total salary of all employees at an office in a year, which essentially control for the available sources at the office. I use *N_employee* to control for the available human resources at a regional office. Both *Budget* and *N_employee* are important factors affecting the regional office

investigation process.[13] I also control for leadership change (*Leadership change*) at the regional office, which is defined as one if the regional director change in a year, and zero otherwise.[14] A new regional director may adopt a different enforcement focus and thus affect the investigation process. Finally, I control for the legal expertise at the regional office (*SEC legal expertise*) as the attorneys play an important role in the SEC investigation and litigation process. Specifically, I measure *SEC legal expertise* as the percentage of postings that require a legal-related skill by the regional office in a year.

I also include a vector of firm-level factors to control for the potential impacts of a firm's financial position and information environment on enforcement action (Kedia and Rajgopal 2011). Specifically, I control for a firm's size (*Size*), accounting performance (*ROA*), leverage ratio (*Leverage*), market-to-book ratio (*MTB*), return volatility (*Ret_Vol*), R&D expenditure (*R&D*), capital expenditure (*CAPEX*), institutional ownership (*IO*), and engagement of Big 4 auditors (*Big4*). I also control for a firm's geographical proximity to the regional office (*Proximate*) because Kedia and Rajgopal (2011) find that the SEC is more likely to investigate firms located closer to its offices .[15]

Finally, I include two macroeconomy factors that capture the economic conditions in the local area that potentially affect the enforcement activities.

[13] I obtained regional-office-level budget and employee data from Kalmenovitz (2021).
[14] I obtained regional director information from the SEC's website (https://www.sec.gov/news/pressreleases).
[15] I use the historical headquarter location to measure the distance between a firm and the corresponding office. The information about historical headquarter location is obtained from the Augmented 10-X Header Data provided by the Notre Dame Software Repository for Accounting and Finance (https://sraf.nd.edu/data/augmented-10-x-header-data/).

Specifically, I control for the gross domestic product per capita (*GDP*) and unemployment rate (*UR*) of the regional offices' state.

To examine whether the SEC's use of data analytics impact firms' fraud occurrence likelihood, I estimate the following equation using OLS regression with standard errors clustered by regional office:

$$\begin{aligned} Prob(Fraud)_{i,j,t} \\ = \alpha_0 + \alpha_1 SEC\ data\ analytics_{i,t-1} + Firm\ Controls \\ + Year\ FE + Office\ FE + Industry\ FE + \varepsilon\_(i,j,t) \end{aligned} \quad (2)$$

where *i, j, t* denote regional office *i*, firm *j*, and year *t*. Following Berger and Lee (2022), I use *F-score* to measure the ex ante fraud probability (Dechow et al. 2011). Specifically, I calculate the *F-score* using a prediction model based on financial statement variables capturing accrual quality, firm performance, and external financing measures. A higher *F-score* is associated with a higher probability of accounting fraud. The details of the measurement are in Appendix D. The vector of firm-level control variables for this equation is the same set of controls from Equation (1).

4.2 Variable measurement

4.2.1 Measurement of investigation success rate

To measure the SEC investigation success rate, I utilize a dataset that includes all SEC investigations between 2008 and 2017. As has been discussed in Section 2.1, the information about SEC investigations has historically been unavailable to researchers because of the confidential nature of the investigations. Blackburne, Kepler, Quinn, and Taylor (2021) recently acquired data that includes investigation targets, opening dates, and closing dates through

a series of FOIA requests. An investigation is closed without an enforcement action when there is insufficient evidence of wrongdoings, thus a successful detection is an investigation that concludes with an enforcement action. As such, the dependent variable *Detect* in equation (1) is an indicator variable that equals one if an investigation opened in a year that concludes with an enforcement action, and zero otherwise. I collect information about the SEC enforcement action from the SEC website.[16]

4.2.2 Measurement of SEC data analytics

I use the hiring of personnel with data analytics skills at the regional offices as a proxy for the utilization of data analytics by the SEC regional offices.[17] Specifically, *SEC data analytics* is equal to one for office-years that have a job posting that requires a data analytics skill, and zero otherwise. Instead of measuring *SEC data analytics* at the entire entity level, I measure it at the regional office level for the following reasons: (1) the regional offices are responsible for monitoring, investigating, and enforcing securities violations in its specific geographic area, and they have the discretion to the ways of conducting investigations;[18] (2) focusing on the regional office provides me with both time-series and cross-sectional variations in the utilization of data analytics.

Job posting data is obtained from Burning Glass, which provides real-time data on job postings and the skills demanded of prospective candidates.

[16] https://www.sec.gov/litigation/litreleases.htm.
[17] To the extent that the job postings may not capture all the current employees at the SEC, I construct another measure using SEC employees' resume data. The details of resume-based variable are in Section 5.5.3.
[18] While there is a centralized Division of Enforcement at the SEC headquarter located at the Washington D.C, the majority of (75%) the investigation cases are carried out by the regional offices.

According to Burning Glass, its algorithm crawls nearly 40,000 online job boards and company websites to scrape and code information on job postings. Its proprietary algorithms remove duplicate postings and convert them into a machine-readable format. Importantly, Burning Glass also standardizes the job-level characteristics such as employer name, job title, location of the position, salary, education requirements, and skill requirements. Recent labor economics studies have used the Burning Glass data to examine the changing landscape of the U.S. labor market (e.g., Deming and Kahn 2018; Hershbein and Kahn 2018).

To measure data analytics skills, I follow Acemoglu, Autor, Hazell, and Restrpo (2022), Chen and Srinivasan (2023), and Gao, Huang, and Wang (2021) to construct a list of data analytics skills. These data analytics skills are related to analytics, automation, artificial intelligence, big data, cloud, digitalization, and machine learning. Appendix B shows the keywords used to identify data analytics skills.

4.2.3 Measurement of firms' fraud occurrence likelihood

It is ideally to observe all underlying fraud and see whether it declines after the SEC's use of data analytics. However, fraud is unobservable until it is detected, and prior studies using detected fraud to measure the probability of fraud (Wilde 2017) suffer the challenge of partial observability (Wang, Winton, and Yu 2010; Wang 2013; Barton, Burnett, Gunny, and Miller 2022). Specifically, we do not observe all the fraud that has been committed but observe only the ones that have been committed and subsequently detected. Moreover, because the SEC's use of data analytics is expected to reduce fraud occurrence likelihood and increase fraud detection likelihood, changes in

observed detected fraud (which is the net of these two opposing effects) are uninformative about changes in underlying fraud.

Following Berge and Lee (2022), I rely on an imputed measure of fraud probability as a closer approximation to underlying fraud: *F-score* (Dechow, Ge, Larson, and Sloan 2011). Specifically, I calculate the *F-score* using a prediction model based on financial statement variables capturing accrual quality (noncash net operating assets, changes in receivables and inventory, and percentage of soft assets), firm performance (changes in cash sales and return on assets), and external financing measures (equity and debt issuance). A higher *F-score* is associated with a higher probability of fraud. The *F-score* can also capture earnings management within GAAP, as evidenced by a high *F-score* during the pre-misstatement period (Dechow, Ge, Larson, and Sloan 2011).[19] A detailed calculation is provided in Appendix D.[20]

4.3 Sample selection

Table 2 presents the sample selection procedures. To estimate Equation (1), I start with all SEC investigations opened from 2008 to 2017 using data obtained from Blackburne, Kepler, Quinn, and Taylor (2021). I start with investigations opened in 2008 because the SEC formalized the 11 regional offices in 2007 and I use one-year-lagged value of *SEC data analytics* to examine the effect of data analytics in the subsequent investigations. I then keep investigations that are related to public firms and their executives.[21] I keep

[19] Given the *F-score* is constructed from detected fraud, it may still proxy for the probability of detected, instead of existing, fraud. I utilize a bivariate probit model to further address the partial observability problem in Section 5.6.

[20] Beneish and Vorst (2022) evaluate the ability of *F-score*, *M-score*, current accruals, unexplained audit fees, and Benford's Law to predict financial statement fraud and find that *F-score* ranks first among the five models.

[21] The target of the investigations can be public firms, executives or auditors of the public firms, registered market participants (e.g., investment advisers), and several self-regulatory organizations (for example, the national securities exchanges and the registered clearing

investigations conducted at the regional office level.[22] Finally, I keep firm-years with non-missing control variables. The data about the control variables are from various database including Compustat, CRSP, Thomson Reuters and Audit Analytics. The final sample for estimating Equation (1) consists of 913 investigations and 742 unique firms. To estimate Equation (2), I start from firm-years in Compustat-CRSP database from 2008 to 2017. I then keep firm-years under the jurisdiction of regional offices based on the historical location of the company headquarters.[23] I also keep firm-years with non-missing control variables. The final sample for estimating Equation (2) consists of 25,664 firm-years and 4,260 unique firms.

4.4 Descriptive statistics

Table 3 presents the descriptive statistics for my sample. Panel A presents the descriptive statistics for the sample used in estimating Equation (1) about the effect of the SEC's use of data analytics on investigation success rate. As shown by the mean value of *Detect*, the SEC detects fraud for 12.4% of the opened investigations. 8.7% of the investigations are associated with offices with data analytics skills. The average budget of an office is 26.83 million and there are 182 employees at an office on average. 19.4% of the job postings from the regional office require at least a legal skill. Table 3 Panel B presents the descriptive statistics for the full sample used in estimating Equation (2) about the effect of the SEC's use of data analytics on firms' fraud occurrence

agencies). I focus on public firms and their executives to better control for factors that potentially affect the investigation outcome.

[22] Around one quarter of the investigations are conducted at the Headquarter located at Washington D.C.

[23] Historical headquarter location data is obtained form the Augmented 10-X Header Data provided by the Notre Dame Software Repository for Accounting and Finance ((https://sraf.nd.edu/data/augmented-10-x-header-data/).

likelihood. The distributions of the control variables generally follow those from prior studies. For instance, 73.5% of the firms have a Big 4 auditor (*Big4*), and 60.6% of firms' shares are owned by institutional investors (*IO*) on average. 65.5% of the firms are located within 100km to the SEC regional offices.

5. Results

5.1 SEC data analytics and investigation success rate

Table 4 presents the results for the estimation of Equation (1) about the effect of the SEC regional offices' use of data analytics on investigation success rate. Column (1) presents the results with year fixed effects and Columns (2) and (3) add regional office and industry fixed effects progressively. Consistent with my H1, the coefficient on *SEC data analytics* is positive and significant in all three columns (*t-statistics*=2.46, 4.80, and 2.87, respectively). These results indicate that the utilization of data analytics significantly improves the detection likelihood of SEC investigations. The magnitude of the economic significance is also meaningful. Focusing on the results of Column (3), I find that *SEC data analytics* increases the probability of fraud detection by 12.1%, which almost doubles the mean probability of detection (12.4%) of all the investigations in my sample.

The results for other control variables are also informative. Focusing on the results of Column (3), I find that the regional office budget (*Budget*) is positively associated with investigation success rate, consistent with the idea that regional office budgets affect the office's work efficiency. It also suggests that the SEC's use of data analytics has an incremental effect on detection likelihood besides the impact of office budget, providing implications for the

regional offices' future budget allocation. I also find that a leadership change at the regional office (*Leadership change*) is positively associated with the investigation detection likelihood, suggesting that a new regional director can affect the investigation process and increase the detection likelihood.

5.2 SEC data analytics and fraud occurrence likelihood

Table 5 presents the results for the estimation of Equation (2) about the effect of the SEC regional offices' use of data analytics on firms' fraud occurrence likelihood. Column (1) presents the results with year fixed effects and Columns (2) and (3) add regional office and industry fixed effects progressively. I find that the coefficients on *SEC data analytics* are all negative and significant in all three columns (*t-statistics*=-2.18, -4.03, and -3.98, respectively), suggesting that firms are less likely to commit fraud after observing the SEC's use of data analytics. The effect of the SEC data analytics in deterring fraud is also economic significant. Focusing on the results of Column (3), I find that *SEC data analytics* decreases the probability of fraud by 16.2%, which is around 14.5% of the sample mean of *F-score*. The results suggest that after the SEC regional offices employ the data analytics, firms adjust their expectations of being caught when committing fraud, and reduce their incentives to commit fraud (i.e., fraud is deterred). This finding of fraud deterrence resonates with the concept of perceptual deterrence in criminology, where offenders adjust their behaviour accordingly after observing a change in policing (Apel, 2013).

Several control variables also have predicted relations that are consistent with prior studies. For example, firms that are proximate to the SEC regional

offices (*Proximate*) are less likely to commit fraud, and firms with greater volatility (*Ret_Vol*) are more likely to commit fraud.

Combining the results of fraud detection and fraud occurrence likelihood together, on the one hand, the probability of fraud detection increases with the help of data analytics, and on the other hand, firms are less likely to commit fraud considering the higher detection risk. A new equilibrium is achieved where firms trade off the expected benefits and expected costs of committing fraud, and the underlying nature of the fraud changes.

5.3 Additional tests

In this section, I conduct some additional tests to collaborate my main findings, and complement the main measure of SEC data analytics.

5.3.1 Falsification test using comment letter

An alternative explanation for the higher probability of fraud detection is the SEC-wide changes in oversight, but not necessarily the use of data analytics. An SEC-wide increase in oversight would also impact the SEC's filing review process (and issuance of comment letter), and this process is executed by the Division of Corporate Finance at the SEC's headquarters in Washington D.C, whereas the work of the Division of Enforcement is conducted by various regional offices. Accordingly, I test whether the use of data analytics at the regional offices affect the issuance and quality of comment letter as a falsification test. Following Gunny and Hermis (2020), I use the issuance of a comment letter and the number of days to process to capture the efficiency of the comment letter process. The results in Table 6 Panel A suggest that the use of data analytics at the regional offices does not affect the comment letter process.

5.3.2 Alternative measures for SEC data analytics

In my main analysis I use lagged one-year *SEC data analytics* to test the effects of the SEC's use of data analytics. There are two potential problems of this measure. First, there could be more than one year lag between the job demand and the actual hiring, so the lagged one-year job posting may not accurately capture the current employees' profile. Second, it's also possible that the effect of the SEC's use of data analytics can persist in subsequent years. To alleviate these concerns, I construct another measure, *SEC data analytics_3year*, that equals to one if an office hires an employee with a data analytics skill in any of the previous 3 years, and zero otherwise. As shown in Table 6 Panel B, I find consistent results with my main findings using this alternative measure.

One assumption of using job posting to measure the use of data analytics is that the supply of relevant talents matches the demand, in other words, the SEC has successfully hired the employee with the desired skillset. Another caveat of the job posting data is that for each posting, it does not tell how many employees the SEC actually hires (i.e., one posting may have multiple hirings). To complement the job posting measure, I also use the SEC employees' resume data to construct the data analytics measure. Specifically, I obtained the SEC employees' resumes from a leading labor markets analytics provider, *Revelio Labs*. This data provider collects data from employees' online profiles and resumes from various websites and social media platforms such as LinkedIn (Li, Lourie, Nekrasov, and Shevlin 2022; Renschler, Ahn, Hoitash, and Hoitash 2023). The data contains individuals from public firms, private firms, small and medium-sized enterprises, non-profits, government entities, universities, etc.

More importantly, it contains information about individuals' self-identified skills.

Using *Revelio Labs*, I identify 8,294 individuals who have once worked at the SEC from 2008 to 2017. Among them, 4,659 have skills information. I construct a new variable, *SEC Data analytics_resume*, as the ratio of employees with a data analytics skill. Untabulated results indicate that *SEC Data analytics_resume* is positively correlated with my main variable *SEC Data analytics* (with a correlation coefficient of 0.277, significant at the 0.01 level), suggesting that skills listed in the job posting reasonably mirror the skills of individuals working at the SEC. Note that *SEC Data analytics_resume* also has its caveats: (1) I cannot see the exact time that the individual possesses the skills, and (2) skills are self-reported and individuals have incentives to inflate their profile.

5.4 Cross-sectional analyses

To shed light on the mechanisms through which data analytics affect the SEC's enforcement process, I further explore situations where data analytics is going to play a bigger role. In this section, I examine the effect of the following factors on the relation between data analytics and investigation success rate: (1) corporate disclosure scriptability (2) firm complexity, and (3) geographic proximity between firms and the SEC regional offices.

5.4.1 Disclosure Scriptability

Reviewing corporate fillings submitted to the SEC (10-Ks, 10-Qs, 8-Ks, DEA14As, etc) is one of the important surveillance activities that the agency conducts. Prior research has documented that corporate disclosures have been becoming increasingly complex over the years (Li 2008, Cohen, Malloy, and

Nguyen 2020). For example, Cohen, Malloy, and Nguyen (2020) documented that the length of 10-Ks has grown five times, and the number of textual changes (changes to the language and construction of the financial report) has grown twelve times from 1995 to 2017. Disclosures are constructed in different ways and can differ in their machine-readability, depending on their structure and format (Allee, DeAngelis, and Moon 2018; Cao, Jiang, Yang, Zhang 2023). By utilizing data analytics tools, including machine learning tools and natural language kits, the SEC is better able to analyze data and incorporate information from disclosures that are more machine friendly.

I measure disclosure machine-friendly by using the *Disclosure Scriptability* measure constructed by Allee, DeAngelis, and Moon (2018), which measures the ease with which a filing can be processed and parsed by an automated program. It contains two main aspects: the ease of identifying data of interest and the ease of processing that data into useful information. For ease of identifying data, it includes four disclosure characteristics: the ease with which a script can (1) separate tables from text, (2) decompose text into logical sections, (3) identify the content of logical sections based on the quality of headings, and (4) find the relevant content in the filing itself rather than following links to external documents. For ease of processing data into information, it also includes four disclosure characteristics: (1) the proportion of the filing that is machine-readable as text, (2) the portion of numeric information in the filing that is tabulated, (3) the ease of processing textual information, and (4) the ease of processing tabular information.[24] *Disclosure*

[24] I thank the authors of Allee, DeAngelis, and Moon (2018) for sharing the measurement from their paper.

Scriptability is measured by the average of these eight variables for all disclosures for a firm-year.

Table 7 Panel A presents the results of subsample regressions of Equation (1). Firms with high *Disclosure Scriptability* are those *Disclosure Scriptability* is higher than the sample mean. The Wald test result of examining the differences in the coefficient estimates in the different subsamples suggests that the effect of SEC data analytics is stronger for firms whose disclosure is more machine friendly (*p-value*=0.08). This test provides evidence for my argument that data analytics tools are indeed utilized by the SEC to analyze corporate filings and contribute to the enforcement process.

5.4.2 Firm complexity

I next examine the effect of firm complexity on the relation between the SEC data analytics and investigation success rate. I argue that data analytics increases the amount of data that the SEC can potentially use and decrease the information processing costs of analyzing all the relevant data, leading to a higher investigation success rate. Firms vary in the level of complexity, and more complex firms presumably have a larger amount of and more complex data for the SEC to analyze. In this case, the effects of the SEC data analytics on investigation success rate should be more consequential for more complex firms than for other firms.

I test this cross-sectional prediction by estimating subsample regressions of Equation (1). I use the number of business segments to capture firm complexity and I classify firm-years into high complexity if the number of business segments is larger than the sample mean. Columns (1) and (2) of Table 7 Panel B shows the results of subsample regressions for Equation (1). I find

that the effects of the SEC data analytics on investigation success rate are concentrated in more complex firms. The coefficients of interest are also significantly different between groups (*p-value*=0.08), suggesting that the effect of data analytics is stronger for more complex firms.

5.4.3 Geographical proximity

I next examine the moderating effect of geographical proximity on the relation between the SEC data analytics and fraud detection likelihood. Kedia and Rajgopal (2011) suggest that the investigation costs are lower for firms that are proximate to the regional offices because proximity facilitates interactions between the SEC officials and firms' executives that might inform the SEC about potential misconduct. With the use of data analytics, the SEC regional office is able to collect and analyze more hard information about firms that are distant from the SEC regional office. Thus I predict that the detection likelihood increases more for firms that are farther away from the SEC regional offices than for firms that are located proximate to the SEC regional offices.

I test this cross-sectional prediction by estimating subsample regressions for Equation (1). I form the subsamples based on the distance between the SEC regional offices and the firms' headquarters. Instead of using a continuous variable to partition the sample, I use a discrete number to classify the subsamples following Kedia and Rajgopal (2011). Specifically, I classify firm-years that have a short distance to the SEC regional offices as firms that are located within 100 km of the regional offices. Columns (1) and (2) of Table 7 Panel C shows the results of subsample regressions for Equation (1). I find that the increase in investigation success rate is larger for firms that are distant from

the SEC regional offices than for firms that are proximate to the SEC regional offices, and the difference is also significantly different (*p-value*=0.05).

5.5 Data analytics and other investigation outcomes

In my main analysis I focus on the effect of the SEC's use of data analytics on the investigation success rate, and it's also possible that the SEC's use of data analytics could impact the investigation process in other ways. In this section, I examine the effect of the SEC's use of data analytics on other investigation outcomes, including investigation time and the complexity of the detected fraud.

Investigation time is the number of days between the investigation open date and the investigation close date, and a shorter investigation time suggests a more efficient investigation process. The use of data analytics facilitates the whole investigation process and thus speeds up the investigation period, leading to a shorter investigation time.

I also predict the detected fraud is more severe and complex after the SEC's use of data analytics. The detected fraud is the outcome of two stages: fraud occurrence and fraud detection. For the fraud occurrence, as the expected costs of committing fraud are higher given the higher detection likelihood, the expected benefits should also be higher to induce the managers to engage in fraud, resulting more severe and sophisticated committed fraud. For the fraud detection, the decrease in information processing costs is larger for the more complex cases with the use of data analytics. As a result, the observed detected fraud is more severe and complex. Following Kalmenovitz (2021) and Zheng (2021), I measure case complexity using the number of defendants for each case and the number of alleged violations for each defendant. I collect available

information about defendants and the number of violations at the defendant level for each SEC enforcement case in my sample from the SEC website.[25] I am able to collect available information for 166 defendants for 92 cases with available information for control variables.[26]

Table 8 presents the results of the effect of the SEC data analytics on investigation time and complexity of detected fraud. I find that after the SEC's use of data analytics, the investigation time is shorter. In addition, the detected fraud has more defendants for each case and more alleged violations for each defendant, suggesting that the SEC detects more complex fraud cases after the use of data analytics.

5.6 Alternative model specification using a bivariate probit model

As discussed in Section 4, one empirical challenge in studying the fraud process is the partial observability problem. In this section, I utilize a bivariate probit model to better shed light on the effect of the SEC's use of data analytics on both fraud occurrence and fraud detection process. The bivariate probit model is a well-established technique increasingly employed in the economic and finance literature to overcome partial observability problems (Wang Winton, and Yu 2010; Wang 2011; Barton, Burnett, Gunny, and Miller 2022). Specifically, I specify two distinct but latent processes: fraud occurrence and fraud detection. The observed incidence of detected fraud depends on the outcomes of both processes. This model allows me to infer the probabilities of both fraud occurrence and fraud detection using observed data on detected

[25] From https://www.sec.gov/litigation/litreleases.htm.
[26] Note that the testing power is reduced due to the small sample size.

frauds by maximizing log-likelihood. The details about the specifications and derivations of the bivariate probit model are in Appendix E.

The regression results of the bivariate probit model are presented in Table 9. Column (1) presents the results of the effect of the SEC data analytics on firms' fraud occurrence likelihood and Column (2) presents the results of the effect of the SEC data analytics on the SEC fraud detection likelihood. The results provide consistent evidence with my main results that the SEC's use of data analytics increases the fraud detection likelihood, and the higher detection risk further deters firms' incentives to commit fraud.

5.7 Determinants of SEC data analytics

To better understand the decision to use data analytics by the regional office, I examine several characteristics of the regional offices that may be related to their decision to use data analytics. I collect and analyze two sets of variables. The first category measures the size and capacity of the regional offices, including the budget, the number of employees, and whether there are leadership changes at the regional offices. The second category captures the local economic conditions, including GDP per capita, and the unemployment rate of the local state. To control for changes that occur at the level of the SEC headquarter, I include year fixed effects in my model. Table 10 shows the results for the determinants of the regional offices' use of data analytics. The results suggest that regional office budget (*Budget*) is positively associated with the use of data analytics, which is consistent with the idea that the regional offices need a sufficient budget to invest in data analytics. I also find that the number of employees at the regional offices (*N_employee*) is negatively associated with the use of data analytics, suggesting that a lack of human resources may affect

the use of data analytics to increase work efficiency. I control for these potential determinants of the SEC's use of data analytics and find that the effects of the SEC' use of data analytics remain. Nevertheless, I acknowledge that I cannot completely eliminate all endogeneity concerns because many other regional office characteristics remain unobservable.

6. Conclusions

The use of data analytics has long been emphasized by the SEC and is becoming increasingly important in recently years, while little is known about whether the SEC's investment in data analytics has achieved its stated goal of improving enforcement efficiency. Using a sample of all investigations conducted by the SEC regional offices from 2008 to 2017, I find that the use of data analytics increases the SEC investigation success rate. I further examine the impacts of the SEC data analytics on firms' fraud occurrence likelihood. Anticipating a higher detection risk after the SEC's use of data analytics, firms are less likely to engage in fraud. In the cross-sectional analysis, I find that the effect of data analytics is greater for firms whose disclosure is more machine-friendly, those have a higher level of complexity, and those are geographically distant from the SEC regional offices. I also find that the investigation time is shorter, and the detected fraud is more complex after the SEC's use of data analytics. My results are robust across different measures of SEC data analytics and different specifications.

The findings in this paper contribute to the literature about the SEC enforcement process and have implications for the SEC's future budget

allocation. This paper also contributes to the corporate fraud literature by documenting the deterrence effect on firms' fraud incentives from the regulators.

Appendix A
SEC Data Analytics Job Posting

This Appendix includes a few examples of job postings for employees with data analytics skills, with the most relevant passage underlined.

Posting A:
JOB TITLE: Quantitative Research Analyst

ORGANIZATION: Securities and Exchange Commission

JOB LOCATION: New York, NY

JOB DESCRIPTION:
The U.S. Securities and Exchange Commission is looking for the best and brightest to join our team. Our mission includes advocating for investors who seek to secure a future for their family, providing guidance and regulations for the nation's securities industry in an increasingly global market, and taking action with an eye toward promoting the capital formation necessary to sustain economic growth. Typical Duties Include:

- Serve as a quantitative research analyst working with SEC staff in building sophisticated models, determining proper empirical methodology, organizing data collection, writing unique programs, preparing written reports, and summarizing the studies in formal and informal presentations.
- Provide senior level technical expertise for the design and conduct of comprehensive, complicated financial data studies, surveys, reviews, and research projects where the boundaries are extremely broad and difficult to determine in advance.
- Conduct research in areas such as the analysis of new financial instruments and strategies, options, and derivates which involves the application of financial engineering methodologies and employing financial theory and applied mathematics, as well as computation and the practice of programming.
- Develop state-of-the-art software tools to collect and analyze large volumes of structured and/or unstructured data.
- Work with large volumes of financial data from different sources for back-testing and validation of models, algorithms, and strategies.

QUALIFICATIONS:

- Knowledge of financial engineering to develop, maintain and/or validate models used for forecasting, valuation, instrument and strategy selection, portfolio construction, and risk management covering a wide range of financial instruments, including equities, fixed income, currencies, futures, commodities, and/or derivatives.
- Strong background in machine learning, statistics, or probability at the graduate school level or higher, as well as experience creating predictive analytics on noisy data.

- Proficiency in computer processes, methods, and languages such as Java, C/C++, Java, C#, Matlab, R, SQL, VBA, Perl, Python, Haskell, Clojure, Racket, Lisp, F#, Julia; familiarity with UNIX, shell scripting, distributed/parallel computing, a scripting language such as Python or Perl, fluency with regular expressions; or similar languages and the state-of-the-art database techniques. Demonstrated proficiency in a wide range of programmer's tools (e.g. sed, awk, xargs, google, stack overflow, etc)

Posting B:
JOB TITLE: Supervisory Financial Economist

ORGANIZATION: Securities and Exchange Commission

JOB LOCATION: New York City, NY

JOB DESCRIPTION:

- Provides financial and risk modeling expertise and support to other offices and divisions with corporate issuer, broker/dealer, and investment adviser risk assessment and oversight activity. It will also support the SEC's staff with examination planning, including providing guidance on the collection and analysis of data to help promote risk-based examination programs.
- Exercising the full range of supervisory and personnel management responsibilities pertinent to work performed by subordinate staff, assuring the fulfillment of quality work products to meet changing requirements and contingencies as they develop.
- Serving as liaison for agency staff to the field of predictive analytics and advising SEC senior management on economic issues related to risks in securities markets and financial system stability, addressing a wide range of complex and potentially controversial matters.
- Directing the creation and implementation of quantitative methods and models to provide data-driven analytical support for Commission supervisory, surveillance, and investigative programs as they relate to corporate issuers, broker/dealers, investment advisors, and exchanges and trading platforms.
- Providing executive leadership and management of the staff and work products by maintaining and exhibiting knowledge, insight, and understanding of state of the art risk assessment tools and techniques; identifying and developing new databases necessary to advance risk assessment programs; anticipating trends and practices in the markets; fostering productive work relationships with national and international agencies.

QUALIFICATIONS:

- Supervisory Program Management managerial and leadership skills necessary to effectively plan, schedule, and carry out major projects and

studies surrounding high profile issues for the SEC by providing leadership and direction through subordinate staff and project teams while exercising the full range of supervisory and personnel management responsibilities.

- Financial Economics: Knowledge and ability to critically analyze economic principles, theories, concepts, methods, and techniques.
- Technical Competence: An understanding of predictive analytics, statistical methods, and modeling techniques relevant to the risk assessment of financial market entities; knowledge of how to collect, manage, and financial market/entity data necessary to implement such methods and techniques.
- Written Communication: Advanced demonstrated skill in presenting concise and clear written information and opinions on topics of financial economics to support critical decisions at the SEC.
- Oral Communication: Ability to expresses clear and convincing ideas and facts to individuals and in group settings to effectively represent the SEC.

Posting C:
JOB TITLE: Case Management Specialist

ORGANIZATION: Securities and Exchange Commission

JOB LOCATION: Los Angeles, CA

JOB DESCRIPTION:
The U.S. Securities and Exchange Commission is looking for the best and brightest to join our team. Our mission includes advocating for investors who seek to secure a future for their family, providing guidance and regulations for the nation's securities industry in an increasingly global market, and taking action with an eye toward promoting the capital formation necessary to sustain economic growth. Typical Duties Include:

- Support the Division of Enforcement's Case Management Systems and Reporting (CMSR) Group by coordinating and executing the quarterly review of information to be provided to the Office of Financial Management (OFM).
- Develop training guidance and assist in the coordination and execution of the training of the Division's Case Management Specialists.
- Develop exception reports and perform analysis of data generated by the Division's case management and tracking system (the Hub).
- Support the CMSR Group by serving as a technical advisor for operational policies and procedures relevant to the Division's case management and tracking system.
- Perform multi-office and national data quality assurance and oversight related to the CMSR Group's case management and tracking system.

QUALIFICATIONS:

- Knowledge of principles, concepts, and methods of legal research and reference sources sufficient to locate legal decisions and court orders.
- Skill in communicating, preparing reports, drawing conclusions and recommending courses of action.
- Ability to perform data analysis utilizing various analytic tools such as Excel and/or Webi.
- Skill in accessing computerized legal research services (PACER; LEXIS-NEXIS; and Westlaw) in order to maintain Division's case management system (the HUB), and to manage and monitor the accuracy of data and generate reports.
- Ability to train others and provide continuing monitoring and guidance in accessing computerized legal research services.

Appendix B
Key Words for Data Analytics Skills

This appendix lists the keywords used to identify SEC employees' data analytics skills.

acl, ai chatbot, ai related, ai tech, amazon web services, analytics, apache, apache drill, apache flink, apache hbase, apache hdfs, apache hive, apache pig, apache presto, apache samza, apache spark, apache storm, apache zookeeper, artificial intelligence, audit command language, augmented reality, automation solutions, autonomous tech, big data, biometric, business intelligence, caffe, caseware analytics, chatbot, cloud based, cloud computing, cloud deployment, cloud enablement, cloud platform, cntk, cognitive computing, computer vision, conversational ai, customer intelligence, data lake, data mining, data scien, data visualization, deep learning, devops, digital marketing, digital revolution, digital strateg, digital transformation, digital twin, digiti, eclipse deeplearning4j, edge computing, evolutionary ai, evolutionary computing, facial recognition, gradient boost, hadoop, hybrid cloud, idea data analysis, image processing, image recognition, intelligent automation, intelligent system, keras, kernel method, kylin, latent dirichlet allocation, latent semantic analysis, libsvm, machine learning, machine translation, machine vision, mahout, mapreduce, marketing automation, microsoft powerbi, microsoft visio, mongodb, mxnet, mysql, natural language processing, neural network, nosql, object recognition, opencv, operating intelligence, opinion mining, pattern recognition, predictive model, process automation, proprietary algorithm, python, pytorch, qlikview, random forest, recommender system, robotic process automation, sas, scala, scikit-learn, scipy, sentiment analysis, sentiment classifi, smart data, spark mllib, speech recognition, spss, sql, structured query language, supervised learning, support vector machine, tableau, tensorflow, text mining, theano, unsupervised learning, vba, virtual agent, virtual assistant, virtual machine, virtual realit, visual basic for application, visualization, word2vec, xgboost

Appendix C
Variable Definitions

Variable	Definition
Detect	An indicator variable that equals one for an investigation that leads to an enforcement action, and zero otherwise.
F-score	The predicted value for a firm of earnings misstatement developed by Dechow et al. (2011). The detailed computation is shown in Appendix D.
SEC Data analytics	An indicator variable that equals one if a regional office has a job posting that requires a data analytics skill, and zero otherwise. For public companies, their regional offices are assigned based on their headquarters states. If a firm's headquarters is under the jurisdiction of a regional office, I assign the firm to that office.
SEC Data analytics_3year	An indicator variable that equals to one if a regional office has a job posting that require a data analytics skill in any of the previous three years, and zero otherwise.
Budget	The total salary of all employees at a regional office in a year.
N_employee	The number of employees at a regional office in a year.
Leadership change	An indicator variable that equals one if the regional office director changes in a year, and zero otherwise.
SEC legal expertise	The number of job postings that require a legal skill divided by the total number of postings at a regional office in a year. Legal skills are skills containing keywords "litigation", "legal" and "law".
Size	The natural logarithm of a firm's total assets.
ROA	Return on assets, defined as net income before extraordinary items divided by total assets.
Leverage	Total debt is divided by total assets.
MTB	The market value of equity divided by book value of equity.
Ret_Vol	The standard deviation of daily return over a fiscal year.
R&D	R&D expenditure divided by total assets in a year.
CAPEX	Capital expenditure divided by total assets in a year.
IO	The average percentage of shares held by institutional investors in a year.
Big4	An indicator variable equals to one if a firm's auditor is from Deloitte & Touche, Ernst & Young, PricewaterhouseCoopers, or KPMG in a year, and zero otherwise.

Proximate	An indicator variable that equals one if the distance from a firm's headquarters to the SEC regional office is less than 100 km, and zero otherwise.
GDP	GDP per capita for a state in a year, measured in thousands.
Unemployment rate	Unemployment rate for a state in a year.
Complexity	The number of business segments for a firm in a year.
No.Violations	The total number of violations against a defendant.
No.defendants	The total number of defendants in a case.
Comment letter	An indicator variable that equals one if a firm receives a comment letter in a year, and zero otherwise.
Days to process	The number of days between the first comment letter issuance date and the 10-K filling date.
Abnormal ROA	Residual from regression $ROA_1 = \alpha_0 + \alpha_1 ROA_0 + \alpha_2 ROA_{-1} + \delta$
Abnormal return volatility	The demeaned standard deviation of monthly stock returns in a year.
Abnormal stock turnover	The demeaned average monthly turnover in a year.

Appendix D
Calculation of the *F-score*

I use model 1 in Dechow et al. (2011) to calculate the *F-score*. They form a prediction model of fraud using financial statement variables capturing accrual quality (noncash net operating assets, changes in receivables and inventory, and percentage of soft assets), firm performance (changes in cash sales and return on assets), and a market-related measure (equity and debt issuance). Dechow et al. (2011) perform backward elimination in the estimation of logistic models for the various determinants of misstatements. They then regress an indicator variable that is equal to one for firm-years involving a AAER during 1982-2005 on the selected sets of predictors to estimate the coefficient on each component of the *F-score*, and compute the predicted value as follows:

$$
\begin{aligned}
\boldsymbol{Predicted\ Value} &= -\mathbf{7.893} \\
&+ \mathbf{0.79} \times \boldsymbol{Changes\ in\ noncash\ operating\ assets} \\
&+ \mathbf{2.518} \times \boldsymbol{Changes\ in\ receivables} \\
&+ \mathbf{1.191} \times \boldsymbol{Changes\ in\ inventory} + \mathbf{1.979} \times \%\boldsymbol{Non} \\
&- \boldsymbol{cash\ and\ Non} - \boldsymbol{PP\&E} \\
&+ \mathbf{0.171} \times \boldsymbol{Changes\ in\ cash\ sales} \\
&+ (-\mathbf{0.932}) \times \boldsymbol{Changes\ in\ ROA} \\
&+ \mathbf{1.029} \times \boldsymbol{Equity\ or\ debt\ issuance}
\end{aligned}
$$

$$
Probability = \frac{e^{Predicted\ value}}{1 + e^{Predicted\ value}}
$$

After calculating the probability of misstatement from the predicted value above, Dechow et al. (2011) compute the *F-score* by dividing the probability by the unconditional probability of misstatement. The unconditional probability, 0.0037, is the ratio of the number of misstatement firms over the total number of firms in their sample. Therefore, the *F-score* provides the likelihood that a firm is engaging in accounting misstatement relative to the unconditional expectation. A higher *F-score* is associated with a higher probability of misstatement.

Appendix E
Bivariate Probit Model

I follow Wang et al. (2010) and Wang (2013) to employ the following bivariate probit model. For each firm *i*, I denote $Fraud_{it}^{*}$ and $Detect_{it}^{*}$ as the latent variables determining firm *i*'s likelihood of committing a fraud in year *t* and the possibility of detecting it as follows:

$$Fraud_{it}^{*} = \beta X_{F,it} + \mu_{it}$$

$$Detect_{it}^{*} = \eta X_{D,it} + v_{it}$$

$X_{F,it}$ is a vector of variables explaining firm *i*'s likelihood of committing a fraud in year *t*, and $X_{D,it}$ contains variables explaining the firm's likelihood of being detected. μ_{it} and v_{it} are zero-mean disturbances with a bivariate normal distribution. The correlation between μ_{it} and v_{it} is ρ.[27]

I define $Fraud_{it}$ = 1 if $Fraud_{it}^{*} > 0$,[28] and $Fraud_{it}$ =0, otherwise; and $Detect_{it}$ =1 if $Detect_{it}^{*} > 0$, and $Detect_{it}$ =0, otherwise. The realizations of $Fraud_{it}$ and $Detect_{it}$ are not directly observed. Instead, we observe

$$Observe_{it} = Fraud_{it} \times Detect_{it},$$

where $Observe_{it}$ = 1 if firm *i* has committed fraud and has been detected, and $Observe_{it}$ = 0 if firm *i* has not committed a fraud or has committed fraud but has not been detected.

Let Φ denote the bivariate standard normal cumulative distribution function. The empirical model for $Observe_{it}$ is:

$$P\ (Observe_{it} = 1) = P\ (Fraud_{it} \times Detect_{it} = 1) = \Phi\ (\beta X_{F,it}, \eta X_{D,it}, \rho)$$

$$P\ (Observe_{it} = 0) = P\ (Fraud_{it} \times Detect_{it} = 0) = 1 - \Phi\ (\beta X_{F,it}, \eta X_{D,it}, \rho)$$

[27] If the estimated ρ is significantly non-zero, it suggests that the two processes are interdependent and should be estimated together.

[28] This means the expected benefits exceed the expected costs of fraud, and the firm decide to commit fraud.

Therefore, the log-likelihood function for the model, estimated using maximum likelihood is:

$$L(\beta, \eta, \rho) = \Sigma \log(P(Observe_{it}=1)) + \log(P(Observe_{it}=0))$$

According to Piorier (1980) and Feinstein (1990), one important condition for the full identification of the model parameters is that $X_{F,it}$ and $X_{D,it}$ do not contain exactly the same variables. I follow Wang et al. (2010) and Wang (2011) to include factors that affect a firm's ex post likelihood of being detected but not the firm's ex ante incentive to commit fraud. In particular, I use unexpected firm performance shocks (*Abnormal ROA*), abnormal stock return volatility (*Abnormal return volatility*), and abnormal turnover (*Abnormal stock turnover*).[29]

My baseline specification for the latent fraud commission equation is as follows:

$$Fraud^{*}_{it} = \alpha_F + \beta_F X_{F,it-1} + \mu_{it}$$

The vector $X_{F,it-1}$ is the set of variables that capture the cost and benefit of committing fraud. The firm characteristics are chosen to describe the condition or state underlying the decision of whether or not to commit fraud, and hence all the firm characteristics are measured one year prior to the violation period, i.e., in year *t-1* for the violation year *t*. The factors include *Size*, *Leverage*, *MTB*, *R&D* and *CAPEX*.

My baseline specification for the latent fraud detection equation is as follows:

$$Detect^{*}_{it} = \alpha_{D+} \delta_D X_{D,it-1} + \lambda_D X_{D,it+1} + \nu_{it}$$

[29] The model assumes no false detection. False detection refers to the situation in which no fraud has occurred but the SEC (inappropriately) detected fraud. No false detection is a reasonable assumption for the SEC fraud detection enforcement cases examined in the paper because false detections are arguably rare in the SEC cases (Wang 2013).

The vector $X_{D,it-1}$ is the set of *ex ante* factors whose effects on the probability of detection can be anticipated at the time that the decision to commit fraud is made (the same set of variables that are included in the fraud commission equation). The vector $X_{D1,it+1}$ is the set of *ex post* factors whose effects on the probability of detection cannot be anticipated at the time the fraud is committed. These variables are measured at the one year after the violation year *t*, i.e., year *t+1*, because fraud detection occurs *after* fraud is committed. These factors that are unpredictable when the fraud decision is made can influence the probability of detection ex post. These ex-post determinants of fraud detection are important in my analysis because they provide a natural set of variables for identification between the fraud commission equation and the fraud detection equation. Following Wang (2013) and Wang et al. (2010), these variables include abnormal accounting performance (*Abnormal ROA*), abnormal stock return volatility (*Abnormal return volatility*), and abnormal turnover (*Abnormal stock turnover*), all of which are measured as of one year after fraud begins.

I identify the detected fraud from the SEC enforcement actions. For each firm-years that are associated with fraud, I match with 50 control firm-years that are not associated with enforcement actions by industry and size.[30] The final sample consists of 12,882 firm-years. The regression results of the bivariate probit model are shown in Table 9.

[30] I choose 50 to have a reasonable sample size so that the bivariate probit model can be estimated (Barton et al. 2022). The inferences remain the same if I match each fraud firm-year with 30 control firm-years by industry and size.

Figure 1. SEC Regional Office Locations and Jurisdictions

The SEC has its headquarter in Washington, DC, and 11 regional offices located in various states. This figure illustrates the areas of jurisdiction of each regional office. Each regional office is in charge of the investigations against firms who have potentially violated the securities laws under its jurisdiction. *Source: SEC Agency Financial Report Fiscal Year 2021*

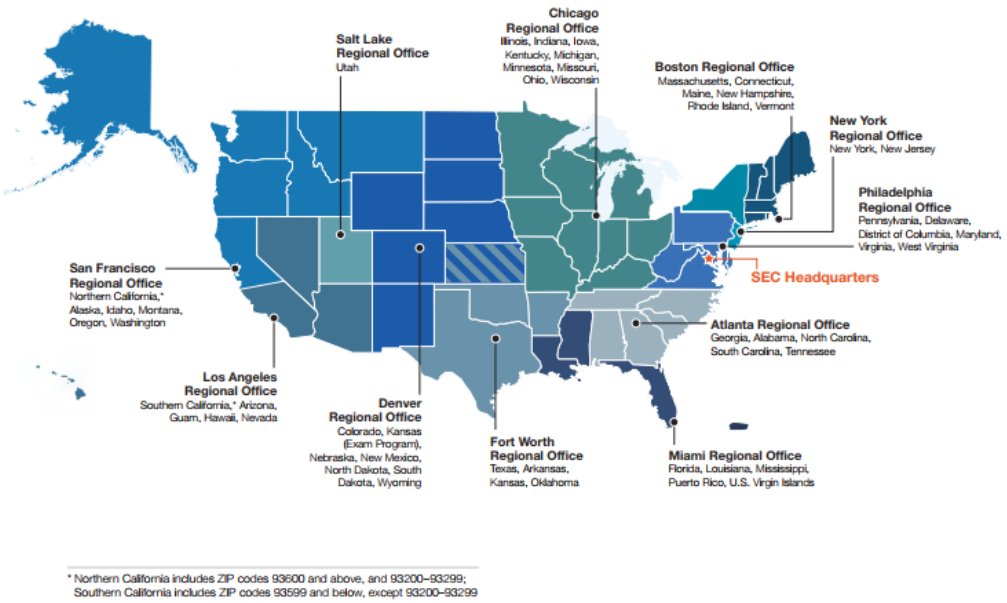

Figure 2. SEC Enforcement Process

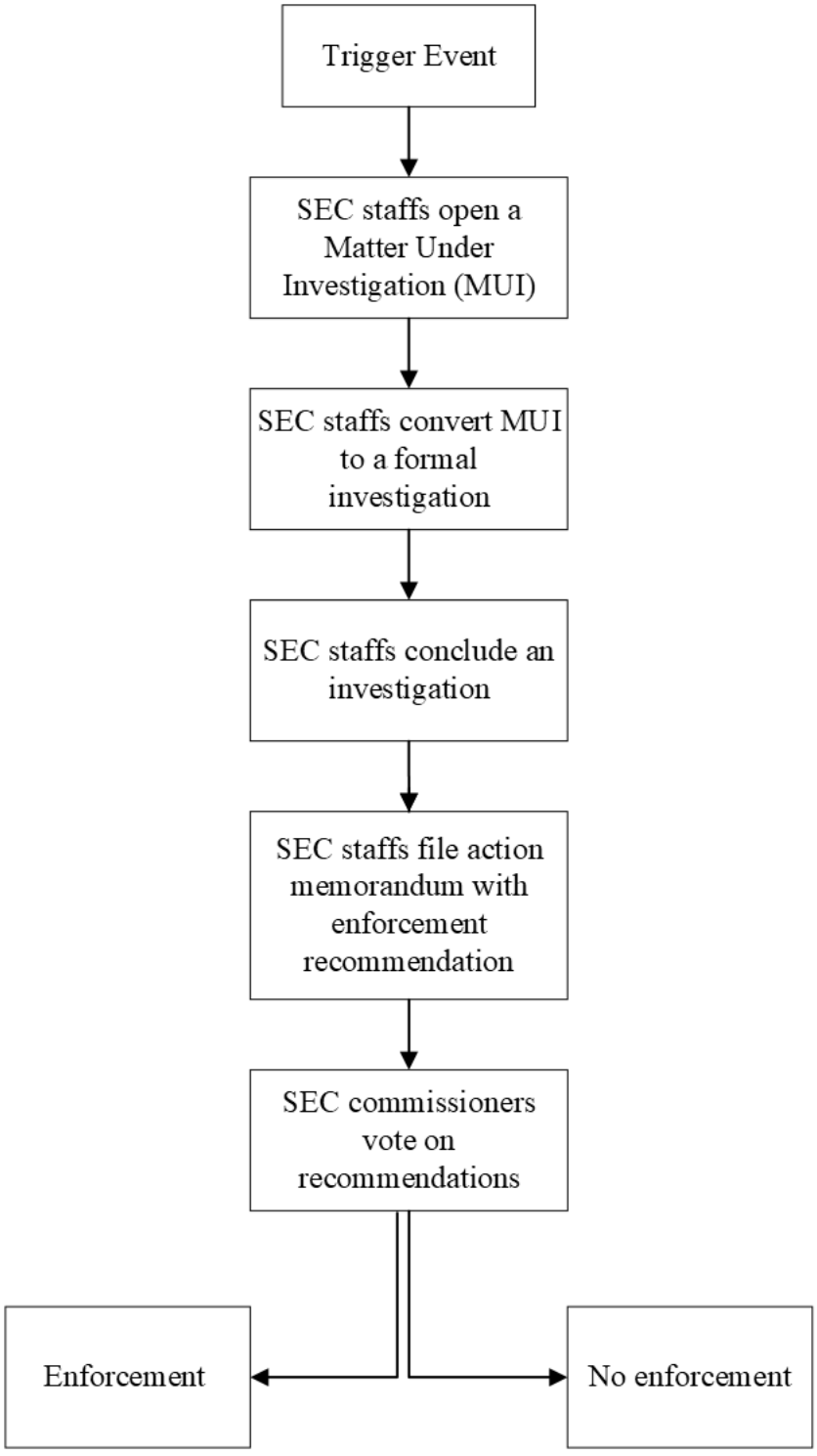

Figure 3. Two-stage Model of Corporate Fraud

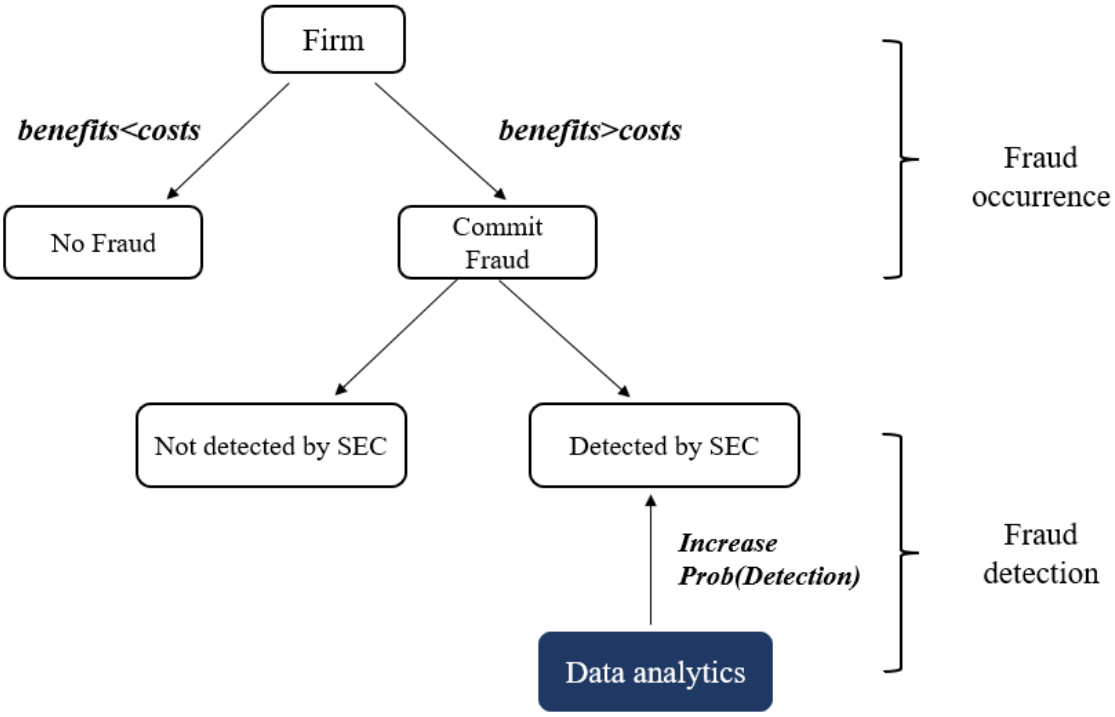

Table 1 Characteristics of SEC Regional Offices

This table presents the characteristics of the SEC regional offices. The number of employees is the average number of employees at an office from 2008 to 2017. The number of attorneys is the average number of attorneys at an office from 2008 to 2017. The amount of budget is the average amount of budget of an office from 2008 to 2017, and the budget is measured as the total salary of all employees at the office.

Regional Office	Number of Employees	Number of Attorneys	Annual Budget (in millions)
Atlanta	101	43	13
Boston	140	58	20
Chicago	246	96	33
Denver	104	55	14
Fort Worth	116	55	15
Los Angeles	169	87	23
Miami	110	58	14
New York	401	136	59
Philadelphia	128	47	17
Salt Lake City	24	13	3
San Francisco	110	46	17

Table 2 Sample Selection

This table presents the sample selection procedures.

Sample for Equation (1)		
	No. investigations	**No. firms**
Investigations conducted from 2008 to 2017	5,455	
Keep:		
Investigations related to public firms	1,506	1,179
Investigations conducted at the regional offices	1,147	942
Firm-years with non-missing control variables	913	742

Sample for Equation (2)	**No. firm-years**	**No. firms**
Firm-years in Compustat-CRSP from 2008 to 2017	41,111	6,585
Keep:		
Firm-years under the jurisdiction of regional offices	38,384	6,149
Firm-years with non-missing control variables	25,664	4,260

Table 3 Descriptive Statistics

This table presents the descriptive statistics of the variables used in the analyses. Please see Appendix C for the variable definitions.

Panel A: Full Sample Descriptive Statistics for Investigation Success Rate Test

Variable	N	Mean	SD	p25	p50	p75
Detect	913	0.124	0.329	0.000	0.000	0.000
SEC data analytics	913	0.087	0.281	0.000	0.000	0.000
Budget	913	26.830	17.780	14.720	19.330	30.320
N_employee	913	182	110	107	134	237
Leadership change	913	0.220	0.415	0.000	0.000	0.000
SEC legal expertise	913	0.194	0.299	0.000	0.000	0.333
Size	913	7.550	2.314	5.898	7.420	9.189
ROA	913	0.003	0.255	0.001	0.049	0.108
Leverage	913	0.197	0.212	0.011	0.145	0.298
MTB	913	2.884	10.730	1.105	1.955	3.713
Ret_Vol	913	0.035	0.020	0.020	0.030	0.044
R&D	913	0.059	0.135	0.000	0.000	0.060
CAPEX	913	0.048	0.068	0.007	0.024	0.059
IO	913	0.607	0.322	0.374	0.682	0.866
Big4	913	0.821	0.383	1.000	1.000	1.000
Proximate	913	0.212	0.409	0.000	0.000	0.000
GDP	913	57.310	9.109	51.140	57.280	62.940
Unemployment rate	913	7.759	2.413	5.900	7.900	9.100

Panel B: Full Sample Descriptive Statistics for Fraud likelihood Test

F-score	25,664	1.118	3.651	0.497	0.829	1.262
SEC data analytics	25,664	0.096	0.295	0.000	0.000	0.000
Leadership change	25,664	0.185	0.389	0.000	0.000	0.000
SEC legal expertise	25,664	0.199	0.302	0.000	0.000	0.400
Size	25,664	6.450	2.057	5.029	6.458	7.837
ROA	25,664	0.012	0.274	-0.004	0.065	0.116
Leverage	25,664	0.191	0.220	0.000	0.136	0.304
MTB	25,664	3.551	16.790	1.236	2.090	3.736
Ret_Vol	25,664	0.033	0.019	0.020	0.028	0.040
R&D	25,664	0.060	0.140	0.000	0.001	0.063
CAPEX	25,664	0.050	0.062	0.015	0.031	0.060
IO	25,664	0.606	0.327	0.335	0.686	0.878
Big4	25,664	0.735	0.442	0.000	1.000	1.000
Proximate	25,664	0.655	0.476	0.000	1.000	1.000
GDP	25,664	58.880	9.536	52.300	57.830	63.980
Unemployment rate	25,664	7.221	2.439	4.900	7.200	9.000

Table 4 SEC Data Analytics and Investigation Success Rate

This table reports the results of Equation (1) about the effect of SEC data analytics on investigation success rate. The dependent variable *Detect*, is an indicator variable that equals one for an investigation that leads to an enforcement action, and zero otherwise. The main variable of interest, *SEC data analytics*, is an indicator variable that equals one if a regional office has a job posting that requires a data analytics skill. Please see Appendix C for the detailed variable definitions. Standard errors are clustered at the regional office level. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.10 levels, respectively, based on two-tailed tests.

VARIABLES	(1) *Detect*	(2) *Detect*	(3) *Detect*
SEC data analytics$_{t-1}$	**0.091****	**0.128*****	**0.121****
	(2.46)	**(4.80)**	**(2.87)**
Budget$_{t-1}$	0.008	0.028**	0.024***
	(0.95)	(2.96)	(3.69)
N_employee$_{t-1}$	-0.001	-0.002	-0.002
	(-0.77)	(-1.38)	(-1.12)
Leadership change$_{t-1}$	0.013	0.039	0.044**
	(0.50)	(1.80)	(2.25)
SEC legal expertise$_{t-1}$	0.061	0.037	0.055
	(1.10)	(0.72)	(1.06)
Size$_{t-1}$	0.034***	0.034***	0.033**
	(3.60)	(3.64)	(3.13)
ROA$_{t-1}$	-0.052	-0.053	-0.070
	(-1.15)	(-0.93)	(-0.95)
Leverage$_{t-1}$	-0.066	-0.069	-0.033
	(-1.67)	(-1.62)	(-0.61)
MTB$_{t-1}$	-0.000	-0.000	0.000
	(-0.73)	(-0.67)	(0.74)
Ret_Vol$_{t-1}$	-0.537	-0.381	-0.418
	(-0.74)	(-0.46)	(-0.57)
R&D$_{t-1}$	0.076	0.102	0.044
	(1.11)	(1.28)	(0.34)
CAPEX$_{t-1}$	-0.218	-0.309	-0.056
	(-1.37)	(-1.65)	(-0.32)
IO$_{t-1}$	-0.031	-0.036	-0.051
	(-0.90)	(-0.96)	(-1.22)
Big4$_{t-1}$	-0.016	-0.014	0.004
	(-0.65)	(-0.49)	(0.14)
Proximate$_{t-1}$	0.003	0.005	0.005
	(0.11)	(0.20)	(0.11)

GDP_{t-1}	-0.005*	-0.010	-0.004
	(-2.22)	(-1.44)	(-0.71)
$Unemployment\ rate_{t-1}$	-0.006	-0.020	-0.000
	(-0.63)	(-0.93)	(-0.02)
Observations	913	913	913
R-squared	0.07	0.09	0.21
Year FE	Yes	Yes	Yes
Office FE	No	Yes	Yes
Industry FE	No	No	Yes

Table 5 SEC Data Analytics and Fraud Likelihood

This table reports the results of Equation (2) about the effect of SEC data analytics on firms' fraud occurrence likelihood. The dependent variable *F-score* is the predicted value for a firm of accounting fraud developed by Dechow et al. (2011). The main variable of interest, *SEC data analytics*, is an indicator variable that equals one if a regional office has a job posting that requires a data analytics skill. Please see Appendix C for the detailed variable definitions. Standard errors are clustered at the regional office level. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.10 levels, respectively, based on two-tailed tests.

VARIABLES	(1) *F-score*	(2) *F-score*	(3) *F-score*
SEC data analytics$_{t-1}$	**-0.096***	**-0.161*****	**-0.162*****
	(-2.18)	**(-4.03)**	**(-3.98)**
Leadership change$_{t-1}$	-0.01	-0.042	-0.047
	(-0.18)	(-0.74)	(-0.86)
SEC legal expertise$_{t-1}$	0.274**	0.142	0.135
	-2.41	-1.34	-1.26
Size$_{t-1}$	0.060**	0.061**	0.073***
	-2.92	-3.09	-4.33
ROA$_{t-1}$	-0.503*	-0.485	-0.598**
	(-1.91)	(-1.81)	(-2.29)
Leverage$_{t-1}$	-0.162	-0.192	-0.159
	(-1.03)	(-1.22)	(-0.93)
MTB$_{t-1}$	0.005	0.005	0.005
	-1.57	-1.59	-1.47
Ret_Vol$_{t-1}$	7.700***	7.925***	7.455***
	-4.34	-4.36	-3.92
R&D$_{t-1}$	1.145*	1.267**	0.855
	-2.1	-2.24	-1.48
CAPEX$_{t-1}$	-2.904***	-2.884***	-1.796**
	(-4.94)	(-4.92)	(-2.59)
IO$_{t-1}$	0.095	0.094	0.071
	-1.02	-1.02	-0.76
Big4$_{t-1}$	-0.09	-0.077	-0.069
	(-1.41)	(-1.18)	(-1.05)
Proximate$_{t-1}$	-0.084	-0.084	-0.129*
	(-1.36)	(-1.36)	(-2.10)
GDP$_{t-1}$	0.003	-0.005	-0.006
	-1.05	(-0.35)	(-0.45)
Unemployment rate$_{t-1}$	-0.018	0.062	0.067
	(-0.75)	-1.51	-1.65

Observations	25,664	25,664	25,664
R-squared	0.02	0.02	0.03
Year FE	Yes	Yes	Yes
Office FE	No	Yes	Yes
Industry	No	No	Yes

Table 6 Additional Tests

This table reports the results of additional tests about the effects of SEC data analytics. *Comment letter* is an indicator variable that equals to one if a firm receives a comment letter in a year, and zero otherwise. *Days to process* is the number of days between the first comment letter issuance date and the 10-K filling date. *SEC data analytics_3year* is an indicator variable that equals one if a regional office has a job posting that require a data analytics skill in any of the previous three years, and zero otherwise. Please see Appendix C for the variable definitions. Standard errors are clustered at the regional office level. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.10 levels, respectively, based on two-tailed tests.

Panel A: Falsification test using comment letter

	(1)	(2)
VARIABLES	*Comment letter*	*Days to process*
SEC data analytics$_{t-1}$	**-0.003**	**1.783**
	(-0.47)	**(0.40)**
Size$_{t-1}$	0.060***	-5.616***
	(28.89)	(-7.43)
ROA$_{t-1}$	-0.022*	17.271
	(-2.10)	(1.64)
Leverage$_{t-1}$	0.104***	-17.417***
	(5.49)	(-3.89)
MTB$_{t-1}$	-0.002***	0.197
	(-4.12)	(1.71)
Ret_Vol$_{t-1}$	1.362***	95.501
	(7.68)	(1.34)
R&D$_{t-1}$	-0.046***	43.005***
	(-4.22)	(3.95)
CAPEX$_{t-1}$	-0.010	-3.583
	(-0.11)	(-0.22)
IO$_{t-1}$	0.059***	9.893***
	(5.50)	(5.11)
Big4$_{t-1}$	-0.021**	8.549**
	(-2.93)	(2.44)
Proximate$_{t-1}$	-0.000	0.421
	(-0.00)	(0.14)
Observations	29,240	9,776
R-squared	0.11	0.06
Year FE	Yes	Yes
Office FE	Yes	Yes
Industry	Yes	Yes

Panel B: Alternative measure for SEC data analytics

VARIABLES	(1) *Detect*	(2) *F-score*
SEC data analytics_3year$_{t-1}$	**0.076*** **(2.03)**	**-0.149**** **(-2.79)**
Observations	913	25,664
R-squared	0.09	0.02
Control variables	Yes	Yes
Year FE	Yes	Yes
Office FE	Yes	Yes

Table 7 Cross-sectional Tests

This table reports the results of subsample tests for Equation (1). Firms with high disclosure scriptability are firms whose disclosure scriptability is higher than the sample mean. Firms with high complexity are firms whose number of business segments are larger than the sample mean. Firms that are distant from the SEC regional offices are firms that are located more than 100 km from the regional offices. Please see Appendix C for the variable definitions. Standard errors are clustered at the regional office level. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.10 levels, respectively, based on two-tailed tests.

Panel A: Cross-sectional tests for disclosure scriptability

Dependent Variable =	*Detect*	
	(1)	(2)
Sample Partitions	Firms with high Disclosure Scriptability	Firms with low Disclosure Scriptability
SEC data analytics$_{t-1}$	**0.153*** **	0.092*
	(3.17)	(1.95)
p-value for the tests of the difference	**0.08**	
N	503	410
R-squared	0.11	0.14
Control Variables	Yes	Yes
Year FE	Yes	Yes
Office FE	Yes	Yes

Panel B: Cross-sectional tests for firm complexity

Dependent Variable =	*Detect*	
	(1)	(2)
Sample Partitions	Firms with high Complexity	Firms with low Complexity
SEC data analytics$_{t-1}$	**0.135** **	0.102
	(3.16)	(1.25)
p-value for the tests of the difference	**0.08**	
N	527	385
R-squared	0.17	0.09
Control Variables	Yes	Yes
Year FE	Yes	Yes
Office FE	Yes	Yes

Panel C: Cross-sectional tests for firm proximity

Dependent Variable =	*Detect*	
	(1)	(2)
Sample Partitions	Firms with long distance to SEC	Firms with short distance to SEC
SEC Data analytics$_{t-1}$	**0.177***** **(4.20)**	0.066 (0.72)
p-value for the tests of the difference	**0.05**	
N	719	193
Adj. R^2	0.10	0.11
Control Variables	Yes	Yes
Year FE	Yes	Yes
Office FE	Yes	Yes

Table 8 Data Analytics and Other Investigation Outcomes

This table reports the results of the effect of the SEC data analytics on other investigation outcomes (*Investigation time, No.Defendatns,* and *No.Violations*). *Investigation time* is the natural logarithm of the number of days between investigation begin date and end date. *No.Defendatns* is the total number of defendants in a year. *No.Violations* is the total number of violations against a defendant. Please see Appendix C for the variable definitions. Standard errors are clustered at the regional office level. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.10 levels, respectively, based on two-tailed tests.

	(1)	(2)	(3)
VARIABLES	*Investigation time*	*No. Defendants*	*No. Violations*
SEC data analytics$_{t-1}$	**-0.397***	**1.039****	**2.612***
	(-2.04)	**(2.25)**	**(1.82)**
Size$_{t-1}$	0.012	-0.032	-0.772***
	(0.67)	(-0.36)	(-4.08)
ROA$_{t-1}$	-0.086	-0.416	-1.506
	(-0.55)	(-0.35)	(-0.56)
Leverage$_{t-1}$	-0.165	1.086	0.864
	(-1.38)	(1.30)	(0.51)
MTB$_{t-1}$	0.003*	0.012	0.043
	(2.15)	(0.89)	(1.51)
Ret_Vol$_{t-1}$	2.580	-8.820	-4.123
	(0.98)	(-0.84)	(-0.18)
R&D$_{t-1}$	-0.318	-1.611	-8.522*
	(-1.33)	(-0.76)	(-1.77)
CAPEX$_{t-1}$	-0.233	-1.228	13.786*
	(-0.63)	(-0.38)	(1.88)
IO$_{t-1}$	-0.101	0.434	3.567**
	(-1.32)	(0.60)	(2.26)
Big4$_{t-1}$	-0.057	-0.058	-1.589
	(-0.77)	(-0.11)	(-1.34)
Proximate$_{t-1}$	-0.100	0.364	0.014
	(-0.81)	(1.01)	(0.02)
Observations	913	92	166
R-squared	0.06	0.26	0.35
Year FE	No	Yes	Yes

Table 9 Bivaraite Probit Model

This table reports the results of the effect of the SEC data analytics on firms' fraud occurrence likelihood and the SEC fraud detection likelihood using the bivariate probit model. Please see Appendix C for the variable definitions. Standard errors are clustered at the regional office level. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.10 levels, respectively, based on two-tailed tests.

Variable	***P(Fraud)***	***P(Detection|Fraud)***
	(1)	(2)
SEC data analytics$_{t-1}$	**-0.545***	**0.489***
	(-1.904)	**(1.666)**
Size$_{t-1}$	0.019	-0.008
	(0.484)	(-0.153)
Leverage$_{t-1}$	2.072***	-1.989***
	(5.835)	(-4.415)
MTB$_{t-1}$	0.012***	-0.010**
	(2.828)	(-2.337)
R&D$_{t-1}$	-7.704***	8.043***
	(-4.536)	(5.715)
CAPEX$_{t-1}$	-3.210	3.116
	(-0.899)	(0.877)
Abnormal ROA$_{t+1}$		0.424*
		(1.748)
Abnormal return volatility$_{t+1}$		2.095
		(1.354)
Abnormal stock turnover$_{t+1}$		0.002
		(0.971)
	0.06	
Log Likelihood	-1286.6522	
Observations	12,882	

Table 10 Determinants of SEC Data Analytics

This table reports the results about the determinants of the SEC data analytics. *SEC data analytics*, is an indicator variable that equals one if a regional office has a job posting that requires a data analytics skill. Please see Appendix C for the variable definitions. ***, **, and * indicate statistical significance at the 0.01, 0.05, and 0.10 levels, respectively, based on two-tailed tests.

VARIABLES	(1) ***SEC data analytics***	(2) ***SEC data analytics***
$Budget_t$	0.056***	0.059***
	(3.71)	(2.82)
$N\_employee_t$	-0.008***	-0.009***
	(-3.43)	(-2.66)
$Leadership\ change_t$	-0.078	-0.081
	(-0.92)	(-0.96)
GDP_t	-0.006	-0.007
	(-1.16)	(-1.13)
$Unemployement\ rate_t$	-0.002	0.014
	(-0.10)	(0.55)
Observations	98	98
R-squared	0.20	0.35
Year FE	No	Yes

References

Acemoglu, D., D. Autor, J. Hazell, and P. Restrepo. 2022. Artificial Intelligence and Jobs: Evidence from Online Vacancies. *Journal of Labor Economics* 40 (S1): 293–340.

Allee, K. D., DeAngelis, M. D., & Moon Jr, J. R. 2018. Disclosure "scriptability". *Journal of Accounting Research* 56 (2): 363-430.

Apel, R. 2013. Sanctions, perceptions, and crime: Implications for criminal deterrence. *Journal of Quantitative Criminology* 29 (1): 67-101.

Armstrong, C. S., A. D. Jagolinzer, D. F. Larcker. 2010. Chief executive officer equity incentives and accounting irregularities. *Journal of Accounting Research* 48 (2): 225-271.

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. 2020. Detecting accounting fraud in publicly traded US firms using a machine learning approach. *Journal of Accounting Research* 58 (1): 199-235.

Bar-Gill, O., and Bebchuk, L.A. 2003, Misreporting corporoate performance, Working Paper. Available at SSRN: https://ssrn.com/abstract=354141.

Barton, J., Burnett, B., Gunny, K., & Miller, B. P. 2022. The Importance of Separating the Probability of Committing and Detecting Misstatements in the Restatement Setting. Forthcoming at *Management Science*.

Becker, G. S. 1968. Crime and punishment: An economic approach. *Journal of Political Economy* 76 (2): 13-68.

Beneish, M. D., & Vorst, P. 2022. The cost of fraud prediction errors. *The Accounting Review* 97 (6): 91-121.

Berger, P. G., & Lee, H. 2022. Did the Dodd–Frank Whistleblower Provision Deter Accounting Fraud? *Journal of Accounting Research* 60 (4): 1337-1378.

Blackburne, T. 2014. Regulatory oversight and reporting incentives: evidence from SEC budget allocations. Working paper. Available at SSRN: https://ssrn.com/abstract=4286862.

Blackburne, T. P., & Quinn, P. J. 2023. Disclosure speed: Evidence from nonpublic SEC investigations. *The Accounting Review* 98 (1): 55-82.

Blackburne, T., Kepler, J. D., Quinn, P. J., & Taylor, D. 2021. Undisclosed SEC investigations. *Management Science* 67 (6): 3403-3418.

Blankespoor, E., deHaan, E., & Marinovic, I. 2020. Disclosure processing costs, investors' information choice, and equity market outcomes: A review. *Journal of Accounting and Economics* 70 (2-3): 101344.

Bonsall, S., Holzman, E., & Miller, B. 2019. Wearing out the watchdog: SEC case backlog and investigation likelihood. Working paper. Available at https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/00ad069d-b853-4777-98d6-1d1e7a07b518/content.

Burns, N., S. Kedia. 2006. The impact of performance-based compensation on misreporting. *Journal of Financial Economics* 79 (1): 35-67.

Cao, S., Jiang, W., Yang, B., & Zhang, A. L. 2023. How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI. Forthcoming at *The Review of Financial Studies*.

Chen, W., & Srinivasan, S. 2023. Going digital: Implications for firm value and performance. *Review of Accounting Studies*: 1-47.

Cohen, L., Malloy, C., & Nguyen, Q. 2020. Lazy prices. *The Journal of Finance* 75 (3): 1371-1415.

Correia, M. M. 2014. Political connections and SEC enforcement. *Journal of Accounting and Economics* 57 (2-3): 241-262.

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. 2011. Predicting material accounting misstatements. *Contemporary Accounting Research* 28 (1): 17-82.

Dechow, P. M., R. G. Sloan, A. P. Sweeney. 1996. Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the SEC. *Contemporary Accounting Research* 13 (1): 1-36.

DeHaan, E., Kedia, S., Koh, K., & Rajgopal, S. 2015. The revolving door and the SEC's enforcement outcomes: Initial evidence from civil litigation. *Journal of Accounting and Economics* 60 (2-3): 65-96.

Deming, D., & Kahn, L. B. 2018. Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics* 36 (S1): S337-S369.

Donelson, D. C., Hopkins, J. J., & Yust, C. G. 2018. The cost of disclosure regulation: evidence from D&O insurance and nonmeritorious securities litigation. *Review of Accounting Studies* 23 (2): 528-588.

Ege, M., Glenn, J. L., & Robinson, J. R. 2020. Unexpected SEC resource constraints and comment letter quality. *Contemporary Accounting Research* 37 (1): 33-67.

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M. F. 2020. Government by algorithm: Artificial intelligence in federal administrative agencies. *NYU School of Law, Public Law Research Paper*, 20-54.

Gao, R., S. Huang, and R. Wang. 2021. Data Analytics and Audit Quality. Working paper. Available at https://ssrn.com/abstract=3928355.

Goldman, E., & Slezak, S. L. 2006. An equilibrium model of incentive contracts in the presence of information manipulation. *Journal of Financial Economics* 80 (3): 603-626.

Gunny, K. A., & Hermis, J. M. 2020. How busyness influences SEC compliance activities: Evidence from the filing review process and comment letters. *Contemporary Accounting Research* 37 (1): 7-32.

Heese, J. 2019. The political influence of voters' interests on SEC enforcement. *Contemporary Accounting Research* 36 (2): 869-903.

Hershbein, B., & Kahn, L. B. 2018. Do recessions accelerate routine-biased technological change? Evidence from vacancy postings. *American Economic Review* 108 (7): 1737-1772.

Hills, R., Kubic, M., & Mayew, W. J. 2021. State sponsors of terrorism disclosure and SEC financial reporting oversight. *Journal of Accounting and Economics* 72 (1): 101407.

Holthausen, R. W. 2009. Accounting standards, financial reporting outcomes, and enforcement. *Journal of Accounting Research* 47 (2): 447-458.

Hutton, A. P., Marcus, A. J., & Tehranian, H. 2009. Opaque financial reports, R2, and crash risk. *Journal of Financial Economics* 94 (1): 67-86.

Jackson, H. E., & Roe, M. J. 2009. Public and private enforcement of securities laws: Resource-based evidence. *Journal of Financial Economics* 93 (2): 207-238.

Jay Clayton, Keynote Remarks at the Mid-Atlantic Regional Conference. June 4, 2019. Available at https://www.sec.gov/news/speech/clayton-keynote-mid-atlanticregional-conference-2019.

Kalmenovitz, J. 2021. Incentivizing financial regulators. *The Review of Financial Studies* 34 (10): 4745-4784.

Kedia, S., & Rajgopal, S. 2011. Do the SEC's enforcement preferences affect corporate misconduct? *Journal of Accounting and Economics* 51 (3): 259-278.

Kubic, M. 2021. Examining the examiners: SEC error detection rates and human capital allocation. *The Accounting Review* 96 (3): 313-341.

Lennox, C., J. A. Pittman. 2010. Big Five audits and accounting fraud. *Contemporary Accounting Research* 27 (1): 209-247.

Li, F. 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics* 45 (2-3): 221-247.

Li, Q., Lourie, B., Nekrasov, A., & Shevlin, T. 2022. Employee turnover and firm performance: Large-sample archival evidence. *Management Science* 68 (8): 5667-5683.

Lin, J. 2021. Regulating via social media: Deterrence effects of the SEC's ese of Twitter. Working paper. *Available at* https://ssrn.com/abstrac=3952904.

Mary Jo White, Speech at the New York University School of Law Program on Corporate Compliance and Enforcement. Nov 18, 2016. Available at https://www.sec.gov/news/speech/chair-white-speech-new-yorkuniversity-111816.html.

National Institute of Standards and Technology Special Publication 1500-1, September 2015. NIST Big Data Interoperability Framework: Volume 1, Definitions.

Noe, Thomas H. 2008. Tunnel-proofing the Executive Suite: Temptation, and the Design of Executive Compensation, Working paper, University of Oxford

O'Malley, T., Harnisch, K., Umayam, M. 2007. An overview of the SEC enforcement

process. MFA Reporter, August/September 2007. Retrieved from https://www.friedfrank.com/siteFiles/Publications/D3C432EBDCA2BB28994CEFE2076F12FC.pdf.

Poirier, D. J. 1980. Partial observability in bivariate probit models. *Journal of Econometrics* 12 (2): 209-217.

Povel, Paul, Rajdeep Singh, and Andrew Winton. 2007. Booms, Busts, and Fraud, *Review of Financial Studies* 20 (4): 1219-1254.

Renschler, M., Ahn, J., Hoitash, R., & Hoitash, U. 2023. Internal Audit Competency and Financial Reporting Quality: Evidence from LinkedIn Human Capital Data. *Forthcoming: Auditing: A Journal of Practice and Theory*.

Samuels, D., Taylor, D. J., & Verrecchia, R. E. 2021. The economics of misreporting and the role of public scrutiny. *Journal of Accounting and Economics* 71 (1): 101340.

SEC, 2007, Enforcement Mannual. Available at https://www.sec.gov/divisions/enforce/enforcementmanual.pdf.

SEC, 2007. Press Release: SEC Elevates District Offices to Regional Level; 2007-59; March 30, 2007.

Securities and Exchange Commission. Agency Financial Report. Fiscal Year 2018. Available at https://www.sec.gov/files/sec-2018-agency-financial-report.pdf.

Securities and Exchange Commission. Agency Financial Report. Fiscal Year 2019. Available at https://www.sec.gov/files/sec-2019-agency-financial-report.pdf#mission.

Securities and Exchange Commission. Agency Financial Report. Fiscal Year 2020. Available at https://www.sec.gov/files/sec-2020-agency-financial-report_1.pdf#chairmessage.

Securities and Exchange Commission. Strategic Plan. Fiscal Year 2018-2022. Available at https://www.sec.gov/files/SEC_Strategic_Plan_FY18-FY22_FINAL.pdf.

Stein, Jeremy C. 1989. Efficient Capital Markets, Inefficient Firms: A Model of Myopic Corporate Behavior. *Quarterly Journal of Economics* 104 (4): 655-669

Thomsen, L. 2009. Testimony of Linda Chatman Thomsen before the United States Senate Committee on Banking, Housing and Urban Affairs Concerning Investigations and Examinations by the Securities and

Exchange Commission and Issues Raised by the Bernard L. Madoff Investment Securities Matter Tuesday, January 27, 2009.

Wang, T. Y. 2013. Corporate securities fraud: Insights from a new empirical framework. *The Journal of Law, Economics, & Organization* 29 (3): 535-568.

Wang, T. Y., Winton, A., & Yu, X. 2010. Corporate fraud and business conditions: Evidence from IPOs. *The Journal of Finance* 65 (6): 2255-2292.

Wilde, J. H. 2017. The deterrent effect of employee whistleblowing on firms' financial misreporting and tax aggressiveness. *The Accounting Review* 92 (5): 247-280.

Zheng, X. 2021. A Tale of Two Enforcement Venues: Determinants and Consequences of the SEC's Choice of Enforcement Venue After the Dodd-Frank Act. *The Accounting Review* 96 (6): 451-476.