THREE ESSAYS ON THE IMPACTS OF EMERGING TECHNOLOGIES ON ACCOUNTING AND AUDITING

By

CHANYUAN ZHANG

A dissertation submitted to the

Graduate School-Newark

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Management

Written under the direction of

Miklos Vasarhelyi

And approved by

Newark, New Jersey

May 2022

ABSTRACT OF THE DISSERTATION

THREE ESSAYS ON THE IMPACTS OF EMERGING TECHNOLOGIES ON ACCOUNTING AND AUDITING

by CHANYUAN ZHANG

Dissertation Director:

Miklos A. Vasarhelyi

This dissertation examines the adoption and impacts of emerging technologies on accounting and auditing. The first essay (Chapter 2) responds to an increasing need for research on Robotic Process Automation (RPA) in external auditing, especially research that concerns auditors' roles in an RPA-enabled audit workflow. Since more than half of the audit tasks require certain levels of auditors' judgment and cannot be fully automated, audit automation should include attended automation, in which auditors work alongside and interact with automation routines. This chapter adopts the Design Science Research (DSR) approach and proposes an Attended Process Automation (APA) framework that guides the implementation of attended automation in audits. This paper also demonstrates the APA framework by applying it to the planning process for Single Audits, a government-required external audit for beneficiaries of funding. The APA framework emphasizes auditors' vital role in an automated audit workflow in providing professional judgments that are currently irreplaceable by automation.

The second essay (Chapter 3) adopts Machine Learning (ML) to examine predictive factors of audit failure. Researchers have identified a broad set of explanatory variables for audit quality. However, little is known about the effectiveness of these

audit-related variables (ARV) in predicting audit failure (i.e., low-quality audits) and which are the most predictive. Understanding the predictive power of ARV can help researchers, regulators, and practitioners evaluate whether these provide practical value in red-flagging audit failure. Using machine learning techniques, we find that ARV have acceptable predictive power and that they outperform benchmark financial variables in predicting audit failure. The most predictive ARV reflect auditor competence, independence, effort, incentive, and the quality of the audited financial reports. We synthesize predictive ARV into a score that can significantly outperform existing academic measures in incrementally associating with audit failures. Our study informs researchers, regulators, practitioners, and investors about the usefulness of ARV in predicting audit failure and provides them with a list of predictive audit features and a score that can be used to predict the likelihood of an audit failure.

The third essay (Chapter 4) explores whether the implementation of artificial intelligence (AI) in firms' operations is associated with improved accuracy of management earnings forecasts. We identify non-technology firms that have implemented AI in their operations from 2014 to 2018. We find that AI is associated with more accurate management earnings forecasts after its implementation and that it improves management forecast accuracy indirectly through improving the performance of firms' operations. However, this indirect effect is small compared to AI's direct effect. We also find evidence that AI more profoundly improves management forecast accuracy when the forecast horizon is longer, and that ML is the primary AI technology that contributes to the improvements in management forecast accuracy. In contrast, we find no evidence that AI is associated with changes to the precision, frequency, or bias of management forecasts. We contribute to the literature by providing initial archival evidence about the association between AI implementation and improvements to the accuracy of management earnings forecasts.

Overall, this dissertation adopts a diverse set of research mythologies to examine the prevailing issues regarding the usage of RPA, ML, and AI in accounting practice and research.

ACKNOWLEDGEMENTS

My deepest appreciation goes to my academic advisor, Dr. Miklos A. Vasarhelyi. Without Dr. Vasarhelyi's support, guidance, and encouragement, I wouldn't have gone so far in my pursuit of enhancing human knowledge about accounting and information systems. I would like to thank my dissertation committee members: Dr. Aleksandr Kogan, Dr. Helen Brown-Liburd, Dr. Michael Alles, and Dr. David Wood. Dr. Kogan is my mentor of machine learning and design science research. He showed me the value of rigor, logic, critical thinking. Dr. Brown-Liburd guides me to connect my research with needs of auditing practice and advises me on how to conduct behavioral research. From Dr. Brown-Liburd, I learned to consider the big picture of my research and its practical implications. Dr. Alles always has great insights and gives me constructive feedback to improve my work. He taught me the importance of thinking through the problem. I knew Dr. Wood since the midyear AAA meeting of 2018. Dr. Wood's encouragement for me to pursue innovative research has motivated me since then. He sets a great example for accounting scholars to conduct practice-relevant and rigorous research. I deeply appreciate Dr. Wood's comments and suggestions for my work.

Next, I would like to thank my other academic advisors. Dr. Chanta Thomas advised Chapter 2 of my dissertation together with Dr. Vasarhelyi. Dr. Thomas is always articulate and insightful. I appreciate her patience in guiding me on how to craft a high-quality academic paper and how to communicate with readers effectively. By the time of this dissertation defense, Chapter 2 had been accepted for publication at the Journal of Information Systems (Zhang et al. 2022). Dr. Soohyun Cho advised Chapter 3 of my dissertation together with Dr. Vasarhelyi. Dr. Cho is an excellent scholar in accounting information systems. Under her guidance, I learned to be resilient under tough times and to cope with challenges properly. I worked with Dr. Rozario on Chapter 4 of my dissertation. Dr. Rozario not only guided me in academic studies but also in life. Over the years, Dr. Rozario has been not only my colleague and coauthor but also my role model and my dear friend.

I would also like to thank my cohorts during the Ph.D. program: Dr. Arion Cheong, Dr. Sheng-Feng Philip Hsieh, Dr. Eid Alotaibi, and Dr. Nuriddin Tojiboyev. I will always remember the days and nights when we tried to solve coding challenges, the train rides where five of us were reading academic papers, the coffee aroma in our cubicles, the long discussions, and the laughter we had together.

My appreciation also goes to Ms. Barbara Jensen from CARLab of Rutgers Business School. Without her help, my experience as an international student in the United States wouldn't have been so smooth. I want to thank Dr. Dan Palmon, the department chair of the accounting department of Rutgers Business School. Without his trust and support, I wouldn't have built valuable teaching experience via independently teaching a diverse set of courses. I also appreciate Dr. Ann Medinets. She is my mentor in teaching and academic writing. I learned a great deal from her on how to be an effective teacher and writer.

Furthermore, I sincerely thank my parents. Without my parents' understanding and support, I wouldn't be in the United States pursuing my passion. I deeply appreciate my husband, Greg Parker, for his continuous love and his strong support for me. Lastly, my appreciation also goes to those who have helped me during my PhD program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

The development of accounting and auditing is non-separable from Information Technology (IT) advancements. IT is primarily concerned with the acquisition, storage, organization, processing, output, and dissemination of data (Rajaraman 2018). Before the 1950s, accountants used basic tools like the abacus and adding machines to conduct calculations (Pennington 1955 as cited in the CPA Journal 2017). Between the 1950s to the 1990s, with the invention of computers, accountants and auditors started using personal computers and spreadsheet programs (e.g., Microsoft Excel) (e.g., Lennox 1965). Since the 1990s, software applications have emerged for accounting and auditing tasks (e.g., Sayana 2003; Tommie 2006). For example, representative accounting software includes QuickBook, SAP, and Oracles, and major auditing software includes CaseWare, IDEA, and ACL.

In recent years, the advancement in computer science, the availability of Big Data, and the exponential increase in computing power have catalyzed a series of emerging technologies: technologies that have radical novelty, fast growth, coherence, prominent impact, and uncertainty and ambiguity (Rotolo, Hicks, and Martin 2015; Brundage et al. 2018; Agrawal, Gans, and Goldfarb 2018; Deloitte 2019). Such emerging technologies include but are not limited to Artificial Intelligence (AI), Robotic Process Automation (RPA), Machine Learning (ML), and Blockchain. Anecdotal evidence suggests that accountants and auditors are actively adopting or considering adopting these emerging technologies in their practice (e.g., AICPA 2020).

Although technology adoption is not a new phenomenon in the history of accounting and auditing, today's emerging technologies pose special considerations in accounting and auditing applications. Examples of such concerns are:

- How feasible is applying emerging technologies (e.g., RPA, ML, and AI) in accounting and auditing?
- How will the adoption of emerging technologies affect the qualities of accounting and auditing practices?
- How can regulators utilize emerging technologies to supervise the accounting and auditing profession better, thus enhancing the healthy functioning of the capital market?

The above considerations are becoming increasingly important given the following features of emerging technologies:

- Technologies are made more and more user friendly and easy to implement
- There is a call for novel business measurement and assurance with emerging technologies given that existing practices are becoming less relevant (e.g., Cho, Vasarhelyi, and Zhang 2019)
- There is still little and explicit regulatory guidance on the proper usage of such technologies despite their growing applications in accounting and auditing.
- Stakeholders (e.g., investors and regulators) increasingly call for research to understand the impacts of emerging technologies on accounting and auditing.

This dissertation builds on the background of increasing adoption of emerging technologies in accounting and auditing and an urgent need better to understand the adoption and impacts of such technologies. There are three major chapters in this dissertation: Chapter 2 - Attended Process Automation in Audit: A Framework and A Demonstration; Chapter 3 - Identifying Informative Audit Quality Indicators (IAQI) Using Machine Learning; Chapter 4 - The Effects of Artificial Intelligence on Management Earnings Forecast Accuracy.

Chapter 2 examines the adoption of RPA in external auditing by following Design Science Research (DSR) to propose an Attended Process Automation (APA) framework that guides the implementation of attended automation in audits. The APA framework was demonstrated in the planning process for Single Audits, a government-required external audit for beneficiaries of funding. Chapter 2 emphasizes auditor-machine cooperation and auditors' vital role in an automated audit workflow in providing professional judgments that are currently irreplaceable by automation.

Chapter 3 focuses on how regulators and academics can utilize ML to understand better factors that are predictive of audit failure (Informative Audit Quality Indicators, or IAQI). The results from Chapter 3 can potentially enhance the supervision of proper auditing conduct, thus enhancing the healthy functioning of the capital market. This essay utilizes various machine learning algorithms and identifies 13 audit-related variables as IAQI with their predictive power validated. These IAQI reflect auditor competence, independence, effort, incentive, and the quality of the audited financial reports.

Chapter 4 explores whether the implementation of AI in firms' operations is associated with improvements in the accuracy of management earnings forecasts. This research identifies non-technology firms that have implemented AI in their operations from 2014 to 2018. The findings show that AI is associated with more accurate management earnings forecasts after its implementation. AI can more profoundly improve the accuracy of the first management earnings forecast compared to the last. Finally, this research shows that ML is the primary AI technology that contributes to the accuracy of management earnings forecasts. This chapter contributes to the literature by providing initial archival evidence about the association between AI implementation and management earnings forecast accuracy.

Overall, this dissertation showcases a diverse set of research methodologies to examine the general issues regarding RPA, ML, and AI in accounting practice and research.

CHAPTER 2: ATTENDED PROCESS AUTOMATION IN AUDIT: A FRAMEWORK AND A DEMONSTRATION

2.1 Introduction

Robotic Process Automation (RPA) is a software technology that allows a knowledge worker to automate repetitive, standardized, structured, and rule-based tasks on one or multiple software platforms (Willcocks, Lacity, and Craig 2015; IEEE 2017). RPA has been determined to be especially useful in task-specific areas of business, such as taxation (Cooper, Holderness, Sorensen, and Wood 2019), accounting (Cooper et al. 2019), and auditing (Moffitt, Rozario, and Vasarhelyi 2018; Huang and Vasarhelyi 2019; Zhang 2019; Eulerich, Pawlowski, Waddoups, and Wood 2021).

In this study, we focus on the implementation of RPA in the external audit profession. The nature of external auditing, in general, is very task-specific. Auditors must achieve the objectives set out by the regulatory agencies that establish the Generally Accepted Auditing Standards (GAAS). To achieve those objectives, auditors conduct a series of tasks, including data collection, entry, sorting, and analysis. Since these tasks are deterministic and repetitive, audit professionals can be assisted by RPA (Moffitt et al. 2018; Rozario and Vasarhelyi, 2018). Regulatory agencies, like the Public Company Accounting Oversight Board (PCAOB) and the American Institute of Certified Public Accountants (AICPA), have acknowledged audit firms' increasing effort in adopting emerging technologies in their audit methodologies (PCAOB 2019). They are specifically calling for research in understanding how emerging technologies like RPA will affect audit procedures and potentially impact audit quality.[1, 2] In addition, the recent COVID-

[1] https://pcaobus.org/Standards/research-standard-setting-projects/Pages/data-technology.aspx
[2] https://www.thecaq.org/rab-request-for-proposals-topics-of-interest-in-2020/

19 pandemic has caused the audit profession to explore ways that emerging technologies, specifically RPA, could increase remote work productivity and assist auditors during long periods of working remotely (Appelbaum, Budnik, and Vasarhelyi 2020; Cohn 2020).

In general, the adoption and use of RPA in audits is still at a nascent stage (Cooper et al., 2019). Since there is a lack of guidance and illustration on how to implement RPA and specifically attended RPA in auditing, more research is needed in this area (Moffit, Richardson, Snow, Weisner, and Wood 2016; Moffit et al. 2018; Christ, Eulerich, Krane, and Wood 2020a; Syed et al. 2020; Eulerich et al. 2021). Thus, this study provides necessary insight and guidance, for both researchers and auditing practitioners, in implementing attended automation in auditing.

We provide a methodological framework, the Attended Process Automation (APA) framework, that can guide the implementation of attended automation in audits for public accounting professionals with at least basic RPA knowledge. This APA framework is in line with the Technological Process Reframing (TPR) theory, which proposes that professionals should reconsider methods and processes adopted in their area with the appearance of disruptive technologies (Issa, Sun, and Vasarhelyi 2016).

Existing RPA research in audit, including studies that propose implementation frameworks (Moffitt et al. 2018; Huang and Vasarhelyi 2019; Eulerich et al. 2021) and case studies (Cohen, Rozario, and Zhang 2019; Huang and Vasarhelyi 2019; Rozario, Vasarhelyi, and Zhang 2019), implicitly focus on unattended automation, in which an automation routine can complete a defined process independently with little auditor intervention (Mullakara and Asokan 2020). However, since most audit tasks in a typical financial statement audit require certain levels of auditor judgment (Abdolmohammadi 1999; Krieger and Drews 2018; Mactavish, McCracken, and Schmidt 2018; Boland, Daugherty, and Dickins 2019), they cannot easily be automated in an unattended manner. Therefore, it is necessary to consider attended automation, in which humans provide real-time inputs to an automation routine and intensively interact with the bot while it is running (Mullakara and Asokan 2020).

This study also answers the call to develop more practice-relevant research in the accounting field by demonstrating how attended automation can be used in an actual audit setting (Kaplan 2011; Basu 2012; Wood 2016; Burton, Summers, Wilks, and Wood 2020; Christ, Emett, Summers, and Wood 2020b; Rajgopal 2020; Burton, Summers, Wilks, and Wood 2021; Eulerich et al. 2021). We explore attended automation in the planning phase of an audit, which can be generalized to other phases and types of audit. Moreover, we add to the accounting studies that adopt Design Science Research (DSR) methodology to produce practice-relevant accounting knowledge (Arnold, Collier, Leech, Sutton, and Vincent 2013; Jans, Alles, and Vasarhelyi 2014; Lombardi and Dull 2016; Huang and Vasarhelyi 2019; Yan and Moffitt 2019; Christ et al. 2020b; Eulerich et al. 2021). We use DSR to propose a framework that can guide the implementation of attended automation in audits and demonstrate it based on an actual audit setting. In DSR, IT artifacts are created to solve identified organizational problems (Hevner, March, Park, and Ram 2004; Hevner and Chatterjee 2010). The problem identified in this study is that unattended automation has limited scope in an audit process because most audit tasks are not structured and cannot be automated without auditor-bot interaction (Abdolmohammadi 1999; Krieger and Drews 2018; Mactavish et al. 2018; Boland et al., 2019). Adoption of attended automation will address this problem by keeping auditors highly engaged in the automated workflow.

Another area in which this study informs academic theory and research is the audit task structure (Abdolmohammadi 1999; Krieger and Drews 2018; Mactavish et al. 2018; Boland et al. 2019). This study extends RPA application research by exploring the task structure of the auditing tasks being automated. Using the APA framework to guide attended automation will allow the auditor to meaningfully interact with the bot and shift them away from repetitive and mundane tasks to those that require higher-level analytical, critical thinking, and professional skepticism skills. Therefore, it also provides researchers with a baseline to explore the implications of attended automation on auditor judgment quality.

Lastly, this study speaks to a broader audience in that it informs those interested in RPA implementation. Namely, it advances the understanding of how attended automation can be implemented into an established auditing process using fundamental RPA skills, advances the understanding of the role of RPA in auditing, and illustrates how auditors can work alongside automation to achieve human-machine synergy, which is especially needed as professional auditing judgments are still valuable and cannot be entirely replaced by automation (Sutton, Arnold, and Holt 2018; Zhang 2019). In addition, with a better understanding of RPA's impact on audits, regulators could also consider adjusting auditing standards to acknowledge audits assisted with emerging technologies and create incentives for innovations balanced with proper quality controls.

2.2 Background Literature

2.2.1 Audit Automation and RPA

Information Technology (IT) development has driven digitization and automation in accounting and auditing (Keenoy 1958; Vasarhelyi 1984). Audit automation is defined as

> *"The use of computers in the management, planning, performance, and completion of audits to eliminate or reduce time spent on computational or clerical tasks, to improve the quality of audit judgments, and to ensure consistent audit quality"* (ICAEW 1993, as cited in Manson, McCartney, Sherer, and Wallace 1998)

RPA can facilitate audit automation (Moffitt et al. 2018; Huang and Vasarhelyi, 2019). Broadly speaking, RPA refers to software that can automate tasks that have structured data, rule-based processes, and a single correct outcome in one or more unrelated software systems (IEEE 2017; Lacity and Willcocks 2017; Cho, Vasarhelyi, and Zhang 2019; Kokina and Blanchette 2019; Zhang 2019; Kokina, Gilleran Blanchette, and Stoddard 2021). However, distinguished from other automation technologies, RPA can run applications in the same way that a person works with software, by interacting with applications at the User Interface (UI) level (Lacity and Willcocks 2017; Gartner 2019), and it is user-friendly (Lacity and Willcocks 2017). While using software to automate work is not a new idea, RPA makes the automation of knowledge workers' jobs more accessible, cheaper, and quicker (Hofmann, Samp, and Urbach 2020).

RPA has been adopted by some in the public accounting industry "to automate the input, processing, and output of data across computer applications to streamline repetitive and mundane business processes" (Cooper et al. 2019). Moffitt et al. (2018) and Huang and Vasarhelyi (2019) proposed using RPA to facilitate audit automation by automating audit tasks that are deterministic and highly repetitive, such as reconciliations, confirmations, internal control testing, and detail testing. Furthermore, Huang and Vasarhelyi (2019) also established a four-stage framework to implement unattended RPA in auditing and created an RPA prototype based on the framework to automate the audit confirmation procedure. Cohen et al. (2019) also used a case study to explore using RPA in the substantive testing of the employee benefit plan audit.

2.2.2 Unattended and Attended RPA

There are two types of RPA: attended and unattended RPA (Mullakara and Asokan 2020). The key difference between attended and unattended automation is that the former requires real-time human intervention, whereas the latter does not. Attended RPA acts like a digital assistant that a human user can call upon on-demand to help complete specific tasks (Ostdick 2017; Mullakara and Asokan 2020). The objective of adopting attended automation is to augment day-to-day knowledge work (Mullakara and Asokan 2020). Therefore, it is suitable for tasks that require real-time human judgment input. For example, a human worker can first trigger a bot that logs in and collects data from multiple systems. Then the bot presents results for the human to review. If the human approves the results, the bot will be authorized to generate a report and update the system. In contrast, unattended automation is suitable for tasks that require little to no human intervention or judgments (Ostdick 2017; Mullakara and Asokan 2020).

Attended automation emphasizes the collaboration between humans and machines. The disruption caused by this type of emerging technology to the accounting profession has raised concerns about accountants and auditors being replaced by automation (Smith 2016; Palmer 2017; Sheedy 2017; Coopers, Holderness, Sorensen, and Wood 2021). Early evidence suggests that RPA enables high-skilled accounting tasks rather than replacing them (Zhang, Issa, Rozario, and Soegaard 2021). Attended automation echoes the concept of "the missing middle," in which humans complement machines by training, interpreting, and maintaining them (Daugherty and Wilson 2018). Machines contribute to humans by amplifying their work and judgment (Daugherty and Wilson 2018). Given the importance of auditors' professional judgment in audits (Nelson and Tan 2005), this paper conjectures that the future of audit automation should focus on attended automation in which auditors and automation cooperate to achieve human-machine synergy.

Compared to unattended RPA, attended RPA has a broader application scope. Unattended RPA bots are applied to individual audit tasks that are structured and can be automated without real-time human intervention. In contrast, attended RPA proposed in this paper is applied to an audit process that is composed of a series of audit tasks that can be either structured or unstructured, with real-time human judgment to moderate and intervene (the concept of task and process is introduced in the "process understanding" step of the APA framework). However, the level of complexity can be adapted to what is needed for the audit solution, and firms can make automation as complex or simple as necessary.

2.2.3 Technological Process Reframing (TPR)

Issa et al. (2016) define TPR as "the reconsideration of methods and processes on an area of endeavor consequent of the advent of a disruptive technology," and they discuss the potential changes to audit procedures brought by the disruptive technology. In addition, other researchers have used the TPR theory to discuss how RPA can be adopted in the audit practice, specifically how auditing processes, audit standards, and audit assurance may be impacted by RPA (Moffit et al. 2018).

While previous research using TPR may have focused on the sustaining and disruptive aspects of emerging technologies (Christensen 2013; Yan and Moffit 2019), this paper argues that the emergence of RPA should make the audit profession rethink how it can better assist the audit process. Moreover, this study proposes that the attended process automation, which features real-time human-machine interaction, can be a potential solution that makes auditors and RPA bots complement each other.

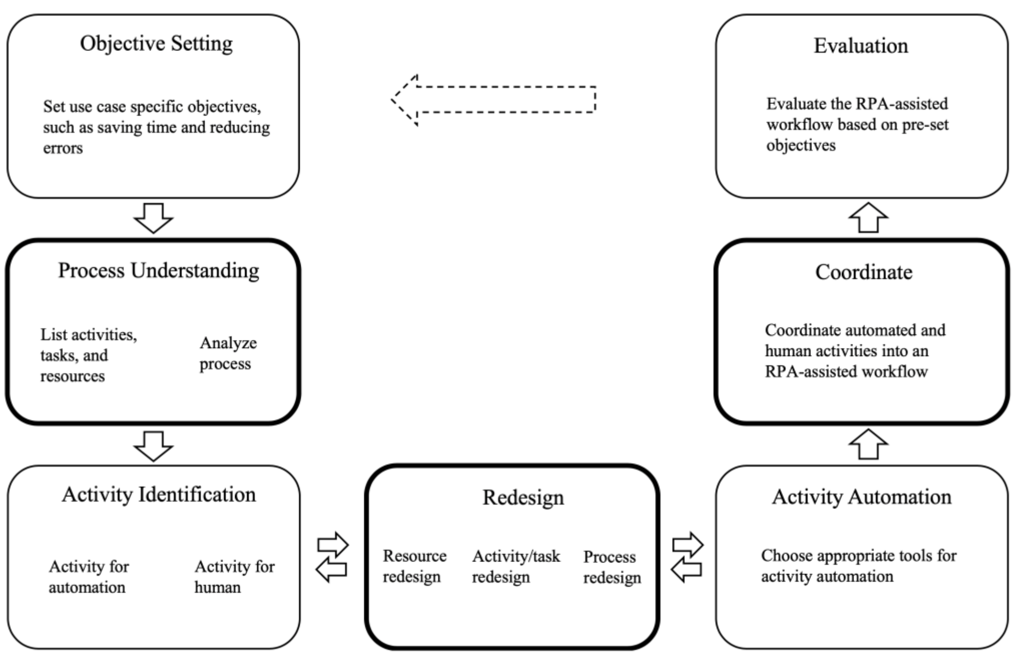
2.2.4 Design Science Research (DSR)

To explore how to implement attended automation and achieve auditor-bot interaction, we follow the DSR methodology proposed by Peffers, Tuunanen, Rothenberger, and Chatterjee (2007), which provides the principles, practices, and procedures required to conduct DSR. This study adopts the six steps in the DSR methodology: problem identification and motivation, the definition of the objectives for a solution, design and development, demonstration, evaluation, and communication (Peffers et al. 2007). The *problem* identified is that unattended automation has relatively limited scope in an audit process because more than half of the audit tasks are not structured and cannot be fully automated (e.g., Abdolmohammadi 1999). Since there is a lack of research studying this level of human-bot interaction in audits, the *objective* throughout this study is to provide such a methodology that can guide the implementation of attended automation in audit, which is more valuable and appropriate for the majority of audit tasks. The *artifact* created in this study is a methodological framework called the Attended Process Automation (APA) framework that can guide the implementation of attended automation in audits. The use of this methodology is *demonstrated* by conducting a case study on a specific audit procedure in a CPA firm. Next, the usefulness of this methodology is *evaluated* via the CPA firm's decision to use the framework along with the prototype. Finally, the problem identified and the artifact created are *communicated* via this academic paper.

2.3 Attended Process Automation (APA) Framework

We developed an APA Framework that can guide the implementation of attended process automation in auditing. Public accounting professionals with at least basic RPA knowledge can use this framework by incorporating the auditor into the process via an attended audit workflow. The framework consists of seven steps: objective setting, process understanding, activity identification, redesign, activity automation, coordination, and evaluation. Figure 1 provides an overview of the APA framework.

Figure 1: Audit Process Automation (APA) Framework



The APA framework distinguishes itself from previous RPA frameworks (e.g., Moffitt et al. 2018 and Huang and Vasarhelyi 2019) in several ways. First, in the "Process Understanding" step, Step 2, we differentiate between the concepts of process, activity, task, and resource based on the workflow literature, and we perform process understanding at each level, respectively. Discerning these concepts is vital because understanding different process elements is fundamental to process automation and redesign projects (Davenport and Short 1990; Davenport and Stoddard 1994; Earl 1994). Next, since the APA framework focuses on attended automation, it contains an "Activity Identification" step, Step 3, where tasks for automation and tasks for humans are classified. The purpose of this step is for "Coordination," Step 6, where tasks that can be automated and tasks that cannot be automated are coordinated together. Another distinguishing factor is that in Step 4, the APA framework focuses on process redesign's role in attended automation. In achieving human-machine synergy, automated tasks and human tasks need to be reorganized. This process redesign is essential for process improvement (Davenport and Short 1990; Davenport and Stoddard 1994; Earl 1994). These steps are discussed further below.

Objective Setting (Step 1)

The APA framework starts with setting specific objectives for audit automation. Examples of such objectives include but are not limited to timesaving, error minimization, and process improvement (Rozario 2019). The objectives set at the beginning stage can also evaluate the automation project at the end of the framework.

Process Understanding (Step 2)

The next step is process understanding. This stage aims to understand the process of interest, including the activities, tasks, and resources in the process, so that later the auditor can classify activities into those that can be automated and those that cannot. We follow the workflow and business process redesign literature and define a process as a sequence of activities performed to achieve a defined outcome (Davenport and Short 1990; Earl 1994; Stohr and Zhao 2001), and an activity as “a discrete step performed either by a machine or a human agent” (Stohr and Zhao 2001). Tasks are granular steps of an activity (Stohr and Zhao 2001), and workflow is an automated process (Stohr and Zhao 2001). Resources refer to the infrastructure, software, files, and data used in an activity or task.

Changes were made to the “process understanding” used in prior literature (Rozario 2019) to make this step more usable in audit settings. First, we put “Process Understanding” before “Activity Identification” because one can better identify which activities can or cannot be automated after sufficient understanding of a process. Second, we propose understanding the process at different levels: process, activities, tasks, and resources. For this to occur, the auditor needs to list each activity involved in the process of interest, each task contained in an activity, and the resources required for each activity (e.g., files and software). Then, it will be possible to analyze the process as a whole to identify places for improvement. Details of operationalizing process understanding are illustrated in the section “Demonstration of the APA Framework.”

Activity Identification (Step 3)

This step identifies audit activities suitable for automation and those that need to remain manual to humans. Following prior literature, we consider an activity suitable for automation if it is rule-based, repetitive, has structured data, has standardized data, and has machine-readable inputs (Huang and Vasarhelyi 2019; Eulerich et al. 2021). Otherwise, the activity will remain manual to the auditor. For this study, an audit activity is considered rule-based if it follows predefined rules and does not require high-level human judgment. An activity is repetitive if it is performed across many different auditees. In addition, structured data is defined as data stored in columns and rows, and standardized data are those with identical data fields and data types across different auditees. Machine-readable inputs refer to digitized data that machines can process, such as data from Excel spreadsheets, word documents, and PDF files. If the original activity does not meet all of the five criteria but does after redesigning (discussion provided in the next section), it is still considered suitable for automation.

Redesign (Step 4)

Huang and Vasarhelyi (2019) included "Procedure Modification" as part of their framework to consider necessary redesigns on the audit procedure and audit data to fit RPA-based audits. This paper expands Huang and Vasarhelyi (2019) and considers redesign at three levels: resource redesign, activity/task redesign, and process redesign. In resource redesign, auditors consider modifying the file type, file layout, or file format to make them suitable for automation. Resource redesign is consistent with the concept of "Audit Data Standardization"[3] (Moffitt et al. 2018; Rozario 2019; Huang and Vasarhelyi

[3] https://www.aicpa.org/interestareas/frc/assuranceadvisoryservices/auditdatastandardworkinggroup.html

2019), and the purpose is to make audit-relevant data suitable for automation. The non-standard or unstructured format of audit data is a common challenge in external audits. AICPA has issued Audit Data Standards (ADS) to homogenize audit data formats to facilitate audit analytics and automation.[4] However, since it is voluntary for auditees and accounting information software providers to follow ADS, and due to the competition among accounting information software vendors, the actual implementation of ADS is not ideal. Nevertheless, standardization of audit data is foundational to audit automation (Alles, Kogan, and Vasarhelyi 2008; Cohen et al. 2019; Huang and Vasarhelyi 2019).

In activity/task redesign, auditors transform initial human-executed processes into the robot-automated process by adopting programming logic (Rutschi and Dibbern 2020). In the process redesign step, auditors consider adjusting the sequence of how each activity/task is executed to improve the workflow of RPA-based audits. The step of redesign and that of activity identification can be iterative. Details of how to operationalize redesigning are illustrated in the section “Demonstration of the APA Framework.”

Activity Automation (Step 5)

Similar to the “Implementation” stage in Huang and Vasarhelyi (2019) and “Audit App Prototyping” in Rozario (2019), the “Activity Automation” step is to choose appropriate tools to automate the activities identified for automation. This step and the previous step of redesigning can also be iterative.

[4] See: https://www.aicpa.org/interestareas/frc/assuranceadvisoryservices/auditdatastandards.html

Coordination (Step 6)

This step is to bridge manual and automated activities into an attended audit workflow based on the redesigned process after automating automatable activities. RPA can coordinate human and automation routines using interactive tools like the message box, email, and event triggers.[5] For example, after the bot generates a report, it will send the report to auditors via an email informing them to review and sign it. Next, auditors review and save the signed report in a designated folder, triggering the bot to perform the following activities. Details of how to operationalize coordination are illustrated in the section "Demonstration of the APA Framework."

Evaluation (Step 7)

Similar to Moffitt et al. (2018), Huang and Vasarhelyi (2019), and Rozario (2019), this stage evaluates the RPA-assisted audit workflow based on the objectives set at the beginning of the framework.

Skills Needed to Use the APA Framework

Public accounting professionals with basic RPA skills can use the APA Framework. To gain such basic RPA knowledge, public accounting professionals can register for free courses offered by RPA software vendors.[6] Accounting professionals without a programming background can configure simple RPA bots after a few hours spent following tutorials. However, to create effective bots that can scale across audit

---

[5] Event triggers are what starts the automation. For example, a bot can be configured to be triggered by a newly saved file in a designated folder, a specific incoming email, and status changes of a system.
[6] For example, UiPath Academy: https://academy.uipath.com/search-filter/MQ%3D%3D?contenttype=learningpath and Automation Anywhere university: https://university.automationanywhere.com/training/rpa-courses/

engagements, public accountants usually need to partner with experienced bot developers or IT personnel to complement each other in knowledge and skills.

2.4 Demonstration of The APA Framework

This section demonstrates how to apply the APA framework to a real-life audit automation task. This demonstration of the proposed methodological framework illustrates the practicality and feasibility of this framework via 'proof by construction' (Nunamaker, Chen, and Purdin 1990; Hevner et al. 2004; Peffers, Rothenberger, Tuunanen, and Vaezi 2012; Venable, Pries-Heje, and Baskerville 2012). To that end, we worked with a medium-sized CPA firm in the United States to apply the APA framework to the audit planning of the firm's Single Audit practice.[7] Single Audits are organization-wide audits of non-federal entities (e.g., states, local governments, universities, and not-for-profit entities) that expend $750,000 or more of Federal funds in one year.[8] They are performed annually to assure the U.S. federal government of the management and use of Federal funds. At the time of this study, this firm was considered one of the major providers of Single Audit services.

In order to help improve the efficiency and effectiveness of the CPA firm's Single Audit process, the firm provided access to their Single Audit work papers. The process involved in conducting Single Audits lends itself to the study of attended automation. Single Audits present an optimal way to demonstrate the use of attended automation in audits because the primary tasks performed in this type of audit are generalizable to all

[7] The authors, not the auditors, configured the demonstration automation project because the purpose here is to illustrate how to apply the APA framework to a real-life situation.
[8] https://www.aicpa.org/content/dam/aicpa/interestareas/governmentalauditquality/resources/singleaudit/downloadabledocuments/fundamental-series/safund2018part1.pdf

audits.[9] Additionally, the small scale of Single Audits also makes them suitable for building a prototype and demonstrating the APA framework. Thus, lessons learned from implementing the APA framework in Single Audits should apply to other types of audits.

In addition, we focused on the audit planning phase for the following reasons. First, audit planning is one of the main phases of an audit engagement, and it has an essential role in the effectiveness of an audit. In audit planning, auditors establish the overall audit strategy for the engagement and develop an audit plan that includes the planned risk assessment procedures and planned responses (including the nature, timing, and extent of further audit procedures) to material misstatement risks (Louwers, Ramsay, Sinason, Strawser, and Thibodeau 2015; PCAOB AS 2101). Second, audit planning is, in general, a time-consuming phase of an audit (Aobdia, Choudhary, and Newberger 2019) and, therefore, would benefit from automation. Third, audit planning contains many tasks that require auditors' professional judgment (e.g., risk assessments), which cannot currently be automated in an unattended manner (Abdolmohammadi 1999; Louwers et al. 2015; Mactavish et al. 2018). At the same time, the audit planning process also contains many deterministic and highly repetitive activities, such as documentation, file referencing, data processing, and calculation (Abdolmohammadi 1999; Louwers et al. 2015; Krieger and Drews 2018). These structured tasks can potentially be automated. Therefore, audit planning is an ideal phase to illustrate attended automation in which humans and automation routines are coordinated to work together. Detailed

---

[9] See: https://www.aicpa.org/content/dam/aicpa/interestareas/governmentalauditquality/resources/singleaudit/downloadabledocuments/fundamental-series/safund2018part1.pdf

demonstration of applying each step of the APA framework to the planning process of Single Audits is provided as follows.

Objective Setting (Step 1)

The partner and senior managers of this CPA firm set the main objective for this pilot project to reduce the time auditors spend on repetitive and rule-based tasks so that auditors will be left with more time to handle tasks that require professional judgments.

Process Understanding (Step 2)

The CPA firm provided the authors with a list of activities involved in Single Audit planning (see Appendix).[10] For each activity, the firm indicated whether they expect the activity to be automated or remain manual. Out of 11 activities provided by the firm, nine were expected to be fully automated, and the other two were expected to be partially automated with some human intervention. A partner and three senior auditors accompanied the observation process to ensure an accurate comprehension of the audit planning procedure. The authors documented the details of the Single Audit planning, which were reviewed by the partner and senior auditors of the CPA firm for completeness and accuracy.

List Activities, Tasks, and Resources

Before each Single Audit starts,[11] the auditees will send the auditors an Excel form named the Schedule of Expenditures of Federal Awards (SEFA) via email. In SEFA, auditees are required to list:[12]

1. The name of the Federal grantor agency or organization;
2. The official program title of the Federal award;

---

[10] The authors also observed the auditors performing the planning tasks from the list of activities.
[11] Assuming the auditee has already been accepted by the audit firm.
[12] https://researchadmin.asu.edu/financial-accountability/sefa

3. The applicable Catalog of Federal Domestic Assistance (CFDA) number[13] for each award;
4. The contract or grant numbers assigned by Federal or state agencies, in addition to the CFDA number;
5. Current year expenditures;
6. Footnote disclosures;
7. The cluster groups[14] of the programs, if they belong to any.

An example of SEFA is provided in Figure 2. Although SEFA requires the same type of information from different auditees, the form completed by different auditees usually has different layouts. Upon the receipt of SEFA, auditors then audit the information on it. Therefore, SEFA is fundamental to Single Audits.

Figure 2: Example of SEFA

Auditee Name
Schedule of Expenditures of Federal Awards
For the Year Ended XXX

Federal Award Source	Federal Program Name	CFDA Number	Additional Award Identification	Expenditures: From Pass-Through Awards	Expenditures: From Direct Awards	Expenditures: Total	Passed through to Subrecipients
STATE OF NEW JERSEY DEPT. OF EDUCATION	TITLE I GRANTS TO LOCAL EDUCATIONAL AGENCIES	84.01		$521,801	-	$521,801	0
STATE OF NEW JERSEY DEPT. OF EDUCATION	ENGLISH LANGUAGE ACQUISITION STATE GRANTS	84.365		$13,415	-	$13,415	0
STATE OF NEW JERSEY DEPT. OF EDUCATION	IMPROVING TEACHER QUALITY STATE GRANTS	84.367	PART A	$52,796	-	$52,796	0
STATE OF NEW JERSEY DEPT. OF EDUCATION	SAFE & DRUG FREE	84.186		$25,892	-	$25,892	0
SPECIAL EDUCATION CLUSTER (IDEA)							
STATE OF NEW JERSEY DEPT. OF EDUCATION	SPECIAL EDUCATION_GRANTS TO STATES	84.027		$132,143	-	$132,143	0
STATE OF NEW JERSEY DEPT. OF EDUCATION	SPECIAL EDUCATION_PRESCHOOL GRANTS	84.173		$2,177	-	$2,177	0
	Total SPECIAL EDUCATION CLUSTER (IDEA):			**$134,320.00**	**-**	**$134,320.00**	
STATE OF NEW JERSEY DEPT. OF EDUCATION	CHARTER SCHOOLS	84.282		$168,768	-	$168,768	0
CHILD NUTRITION CLUSTER							
STATE OF NEW JERSEY DEPT. OF EDUCATION	SCHOOL BREAKFAST PROGRAM	10.553		$73,613	-	$73,613	0
STATE OF NEW JERSEY DEPT. OF EDUCATION	NATIONAL SCHOOL LUNCH PROGRAM	10.555		$322,468	-	$322,468	0
	Total CHILD NUTRITION CLUSTER:			**$396,081**	**-**	**$396,081**	
STATE OF NEW JERSEY DEPT. OF EDUCATION	CHILD AND ADULT CARE FOOD PROGRAM	10.558		$24,121	-	$24,121	0
	Total Federal Awards Expended:			**$1,337,194**	**-**	**$1,337,194**	

[13] Catalog of Federal Domestic Assistance (CFDA) is a list of all federal financial assistance and nonfinancial assistance programs available to a variety of applicants. CFDA numbers uniquely identify the federal programs. Each CFDA number contains five digits and appears in the following format: ##.### (e.g., 10.001 or 98.102). https://blog.grants.gov/2018/06/04/what-is-a-cfda-number-2/
[14] The federal programs are grouped into different clusters defined by the compliance supplement. The list of clusters is updated periodically.

Based on the firm's practice, we grouped the Single Audit planning process into five phases: Pre-SEFA Testing, SEFA Testing, Planning Step 1, Planning Step 2, and Planning Step 3. The purpose of each phase is provided in Table 1. The resources (i.e., files and software) required for each activity are presented in Table 2. The activities and tasks in each phase are listed in Table 3.

Table 1: Phases of Single Audit Planning

Phase	Purpose
Pre-SEFA Testing	Ensure the integrity of the data provided by the clients and prepare data ready to be audited
SEFA Testing	Examine whether the current year's total expenditure and total payments materially deviate from the previous year's
Planning Step 1	Determine if a single audit or program-specific audit is necessary
Planning Step 2	Determine if the auditee should be considered a low-risk auditee
Planning Step 3	Identify Type A and Type B programs and perform related risk assessments

Table 2: Process Understanding: Activities and Resources

Phase	Activity	Files Involved	File format	Software involved
Pre-SEFA Testing	Compare Current Year SEFA (hereafter, CY SEFA) with supporting documents to check data integrity on CY SEFA.	CY SEFA	Excel spreadsheet	Excel
		Client-specific supporting documents	PDF	Adobe
	Compare program information from CY SEFA with that from Previous Year SEFA (hereafter, PY SEFA) to check the integrity of data on CY SEFA.	CY SEFA	Excel spreadsheet	Excel
		PY SEFA	Excel spreadsheet	Excel
	Group programs on CY SEFA that are in the same cluster.	CY SEFA	Excel spreadsheet	Excel
		Part 5-Clusters of programs	PDF	Adobe
	Calculate the total expenditure for each cluster on CY SEFA.	CY SEFA	Excel spreadsheet	Excel
SEFA Testing	Input total expenditure of current and previous year in SEFA Materiality worksheet.	CY SEFA	Excel spreadsheet	Excel
		PY SEFA	Excel spreadsheet	Excel

		SEFA Materiality worksheet	Excel spreadsheet	Excel
	In SEFA Materiality worksheet, calculate the percentage change of the total expenditure from CY SEFA to PY SEFA and compare the change with the pre-set materiality.	SEFA Materiality worksheet	Excel spreadsheet	Excel
Planning Step 1	On Federal Program Planning Worksheet (hereafter, FPPW), answer the question "Do the total federal awards expended equal or exceed $750,000?". If the total expenditure from CY SEFA is above $750,000, choose "Yes" as the answer. Otherwise, choose "No".	CY SEFA	Excel spreadsheet	Excel
		FPPW	CaseView document[15]	CaseWare Working Papers[16]
	On FPPW, if the answer to the previous question is "Yes", choose "No" for the second question to indicate that a single audit is required. Otherwise, select "Yes".	FPPW	CaseView document	CaseWare Working Papers
Planning Step 2	Determine if the client is considered a low-risk auditee by going through questions in Planning Step 2 on FPPW with reference to the previous year working papers.	PY SEFA	Excel spreadsheet	Excel
		PY Single Audit Reports	Word document	Word
		FPPW	CaseView document	CaseWare Working Papers
Planning Step 3	Input the total expenditure into FPPW.	CY SEFA	Excel spreadsheet	Excel
		FPPW	CaseView document	CaseWare Working Papers
	On FPPW, choose the range of amount to which the total expenditure belongs.	FPPW	CaseView document	CaseWare Working Papers
	On CY SEFA, identify Type A and Type B programs.	CY SEFA	Excel spreadsheet	Excel

[15] CaseView is a file type special to CaseWare Working Paper.
[16] CaseWare Working Paper is a commonly used software that public accounting firms in the United States use to manage audit workpapers. See: https://www.caseware.com/us/products/audit

	Update Federal Type A worksheet.	CY SEFA	Excel spreadsheet	Excel
		Federal Type A Worksheet	Excel spreadsheet	Excel
	Update Federal Type B worksheet.	CY SEFA	Excel spreadsheet	Excel
		Federal Type B Worksheet	Excel spreadsheet	Excel
	Assess the risk level of each Type A program on Federal Type A Worksheet.	Federal Type A Worksheet	Excel spreadsheet	Excel
	Assess risk level of each Type B program on the Federal Type B Worksheet.	Federal Type B Worksheet	Excel spreadsheet	Excel
	If there is at least one Type A program, choose “Yes” for the question “Is there at least one Type A program” on FPPW. Otherwise, choose “No”.	Federal Type A Worksheet	Excel spreadsheet	Excel
		FPPW	CaseView document	CaseWare Working Papers
	If there is at least one Type A program that is not-low-risk, choose “Yes” for the question “Is there one or more Type A programs that may not be considered a low-risk program” on FPPW. Otherwise, choose “No”.	Federal Type A Worksheet	Excel spreadsheet	Excel
		FPPW	CaseView document	CaseWare Working Papers
	Input the threshold that classifies Type A and Type B programs into FPPW.	CY SEFA	Excel spreadsheet	Excel
		FPPW	CaseView document	CaseWare Working Papers
	If there is at least one Type B program, choose “Yes” for the question “Are there any Type B programs?” on FPPW. Otherwise, choose “No”.	Federal Type B Worksheet	Excel spreadsheet	Excel
		FPPW	CaseView document	CaseWare Working Papers
	If there is at least one high-risk Type B program, choose “Yes” for the question “Are any of the Type B programs considered high-risk?” on FPPW. Otherwise, choose “No”.	Federal Type B Worksheet	Excel spreadsheet	Excel
		FPPW	Excel spreadsheet	Excel
	Update the Historical Summary of Major Programs.	Federal Type A Worksheet	Excel spreadsheet	Excel
		Federal Type B Worksheet	Excel spreadsheet	Excel

		Historical Summary of Major Programs	Excel spreadsheet	Excel
	Add extra major programs	Historical Summary of Major Programs	Excel spreadsheet	Excel
	Calculate the total expenditures of all major programs.	Historical Summary of Major Programs	Excel spreadsheet	Excel
	Input the value of the total expenditure of all major programs into FPPW.	Historical Summary of Major Programs	Excel spreadsheet	Excel
		FPPW	CaseView document	CaseWare Working Papers
	On FPPW, if the total expenditure of all major programs is above a pre-set percentage of the total expenditure of all programs, choose “Yes” for the question “Has the percentage of coverage requirement been met?”. Otherwise, choose “No”.	FPPW	CaseView document	CaseWare Working Papers

Table 3: Process Understanding: Activities and Tasks

Phase	No.	Activity	Manual Tasks
Pre-SEFA Testing	1	Compare CY SEFA with supporting documents to check the integrity of data on CY SEFA.	1) Select the programs based on judgments to review. 2) Read client-specific supporting documents and look for information for the selected programs. 3) Compare information from supporting documents with that in CY SEFA to check whether CY SEFA is appropriately reported in terms of total amount and other details.
	2	Compare program information from CY SEFA with that from PY SEFA to check the integrity of data on CY SEFA.	1) Select the programs based on judgments to review. 2) Read PY SEFA and look for information for the selected programs. 3) Compare information from PY SEFA with that in CY SEFA to check whether CY SEFA is appropriately reported.
	3	Group programs on CY SEFA that are in the same cluster.	1) Open Part 5-Cluster of programs. 2) Go through each program in CY SEFA and make a note of the cluster to which this program belongs. 3) On CY SEFA, group programs that are in the same cluster
	4	Calculate the total expenditure for each cluster on CY SEFA.	For each cluster, input Excel function to calculate the total expenditure of that cluster[17].
SEFA Testing	5	Input total expenditure of current and previous year	Copy the total expenditure from CY SEFA and paste it to SEFA Materiality worksheet. Copy the

[17] Program expenditures that do not belong to any cluster and cluster expenditures are used as the units of testing in later procedures.

		in SEFA Materiality worksheet.	total expenditure from PY SEFA and paste it to SEFA Materiality worksheet.
	6	In SEFA Materiality worksheet, calculate the percentage change of the total expenditure from CY SEFA to PY SEFA and compare the change with the pre-set materiality.	In SEFA Materiality worksheet, input Excel function to calculate the percentage change.
Planning Step 1	7	On FPPW, answer the question “Do the total federal awards expended equal or exceed $750,000?”. If the total expenditure from CY SEFA is above $750,000, choose "Yes" as the answer. Otherwise, choose “No”.	Auditor read the total expenditure from CY SEFA and choose “Yes” for this question if it is above $750,000. Otherwise, the auditor chooses “No”.
	8	On FPPW, if the answer to the previous question is "Yes", choose "No" for the second question to indicate that a single audit is required. Otherwise, select “Yes”.	On FPPW, if the answer to the previous question is "Yes", choose "No" for the second question to indicate that a single audit is required. Otherwise, select “Yes”.
Planning Step 2	9	Determine if the client is considered a low-risk auditee by going through questions in Planning Step 2 on FPPW with reference to previous year working papers.	Read previous year Single Audit reports and answer each question accordingly.
Planning Step 3	10	Input the total expenditure into FPPW.	Copy the total expenditure from CY SEFA and paste it to FPPW.
	11	On FPPW, choose the range of amount to which the total expenditure belongs.	On FPPW, choose the range of amount to which the total expenditure belongs.
	12	On CY SEFA, identify Type A and Type B programs.	On CY SEFA, 1) add two columns in and name them "Type A" and "Type B" respectively, and 2) manually compare the cluster expenditure (or program expenditure) with predefined thresholds, and then 3) mark in either "Type A" or "Type B" column to indicate which type does the cluster or program belong to.
	13	Update Federal Type A worksheet.	Copy the information of Type A programs from CY SEFA and paste it to Federal Type A worksheet.
	14	Update Federal Type B worksheet.	Copy the information of Type B programs from CY SEFA and paste it to Federal Type B worksheet.
	15	Assess risk level of each Type A program on	Open Federal Type A worksheet of previous two years, answer questions on current year Federal Type A worksheet such as “Has it been audited in

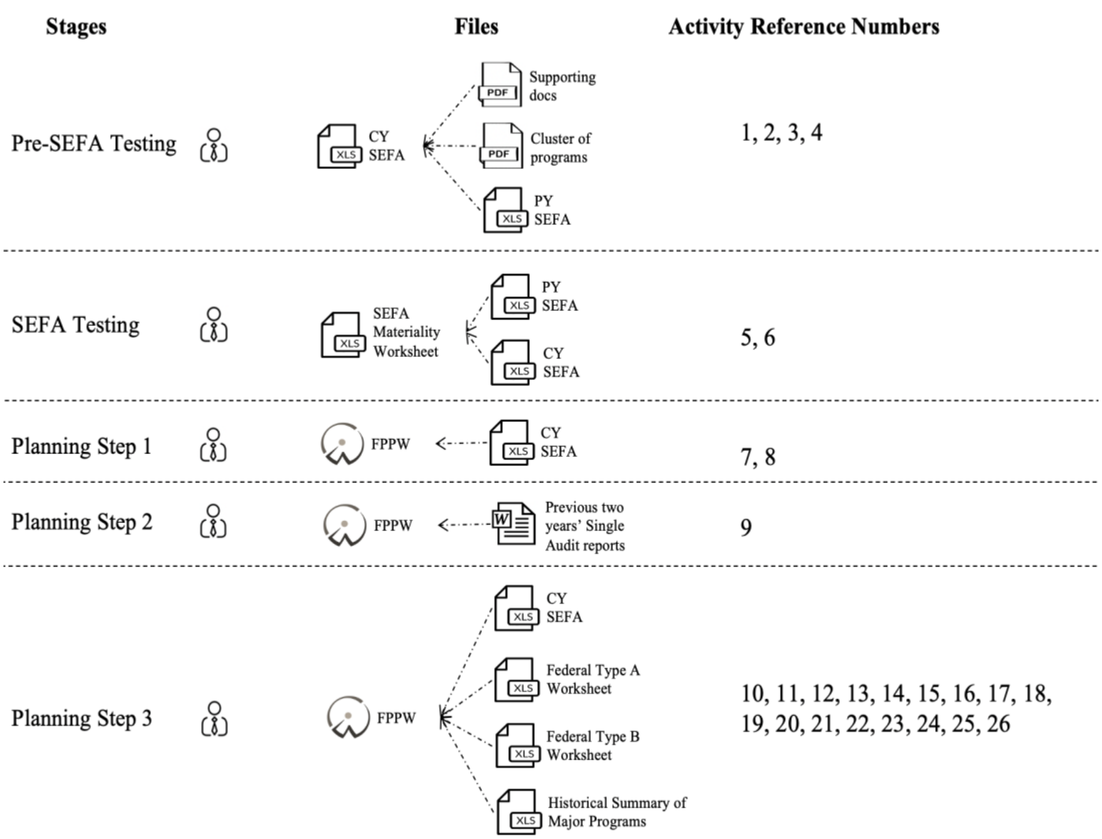
	Federal Type A Worksheet.	the past 2 years? If 'yes' which year was it audited" and "Is this program low-risk?"
16	Assess risk level of each Type B program on Federal Type B Worksheet.	Open Federal Type B worksheet of previous two years, answer questions on current year Federal Type B worksheet such as "Has it been audited before? If 'yes' which year was it audited" and "Is this program low-risk?"
17	If there is at least one Type A program, choose "Yes" for the question "Is there at least one Type A program" on FPPW. Otherwise, choose "No".	On Federal Type A Worksheet, count the number of Type A programs. If there is at least one Type A program, choose "Yes" for the question "Is there at least one Type A program" on FPPW. Otherwise, choose "No".
18	If there is at least one Type A program that is not-low-risk, choose "Yes" for the question "Is there one or more Type A programs that may not be considered a low-risk program" on FPPW. Otherwise, choose "No".	On Federal Type A Worksheet, count the number of Type A programs that are not-low-risk. If there is at least one Type A program that is not-low-risk, choose "Yes" for the question "Is there one or more Type A programs that may not be considered a low-risk program" on FPPW. Otherwise, choose "No".
19	Input the threshold that classifies Type A and Type B programs into FPPW.	From CY SEFA, copy the threshold that classifies Type A and Type B programs and paste it to FPPW.
20	If there is at least one Type B program, choose "Yes" for the question "Are there any Type B programs?" on FPPW. Otherwise, choose "No".	On Federal Type B Worksheet, count the number of Type B programs. If there is at least one Type B program, choose "Yes" for the question "Are there any Type B programs?" on FPPW. Otherwise, choose "No".
21	If there is at least one high-risk Type B program, choose "Yes" for the question "Are any of the Type B programs considered high-risk?" on FPPW. Otherwise, choose "No".	On Federal Type B Worksheet, count the number of Type B programs that are high-risk. If there is at least one high-risk Type B program, choose "Yes" for the question "Are any of the Type B programs considered high-risk?" on FPPW. Otherwise, choose "No".
22	Update the Historical Summary of Major Programs.	1) From Federal Type A worksheet, copy the information of Type A programs that are not-low-risk and paste it to the Historical Summary of Major Programs. 2) From Federal Type B worksheet, copy information of Type B programs that are high-risk and paste it to the Historical Summary of Major Programs.
23	Add extra major programs	In the Historical Summary of Major Programs, add extra major programs besides not-low-risk Type A and high-risk Type B programs.
24	Calculate the total expenditures of all major programs.	In the Historical Summary of Major Programs, input Excel function to calculate the sum of expenditures of all major programs.
25	Input the value of the total expenditure of all major programs into FPPW.	From the Historical Summary of Major Programs, copy the total expenditure of all major programs and paste it to FPPW.

26	On FPPW, if the total expenditure of all major programs is above a pre-set percentage of the total expenditure of all programs, choose "Yes" for the question "Has the percentage of coverage requirement been met?". Otherwise, choose "No".	On FPPW, if the total expenditure of all major programs is above a pre-set percentage of the total expenditure of all programs, choose "Yes" for the question "Has the percentage of coverage requirement been met?". Otherwise, choose "No".

Analyze the Process

A holistic view of the original planning process of the Single Audit is illustrated in Figure 3.

Figure 3: Original Process Overview



Note:
- Activity reference numbers are from Table 3
- Human icon indicates manual activities

Based on the observation of the Single Audit planning process and discussions with the CPA firm's senior manager and auditors, four main characteristics of this process were identified: 1) manual, 2) time-consuming, 3) error-prone, and 4) non-standard.

1) Manual

Even though the audit work papers were digitized using CaseWare and Excel, intensive manual procedures were still prevalent, especially in locating and opening different files, copying and pasting information, performing rule-based data analysis on Excel, and completing the Federal Program Planning Worksheet (hereafter, FPPW).

2) Time-consuming

Since the planning process involved intensive manual steps, it was time-consuming. According to the senior auditor, auditors generally spend at least three hours to complete the rule-based tasks in the audit planning, counting a substantial portion of the entire audit planning phase.

3) Error-prone

The original planning procedures were prone to errors caused by manual referencing, calculation, and classification.[18]

4) Non-standard

Though audit work papers have standardized templates, auditors tended to annotate them in different ways. Besides, for the same task, such as categorizing Type A and Type B programs on Excel (Activity 12 in Table 3), different auditors usually

[18] The auditors from the CPA firm stated that common errors came from data entry, copying and pasting, classifying, and annotating program type, and inputting Excel calculation functions. These activities were a time-consuming part of the Single Audit planning process.

approached it in various ways. These not only resulted in unstructured work papers but also added complexity to the audit reviewing process.

Since the original audit planning processes of Single Audits were manual, time-consuming, error-prone, and non-standard, there was a need to optimize this process by using appropriate automation tools to reduce the intensive manual procedures, save time, improve the accuracy of the deterministic tasks, enhance the standardization of audit working papers, and importantly, include auditors in the automated workflow to provide their judgments.

Activity Identification (Step 3)

Table 4 presents whether each activity meets different criteria and whether it is chosen for automation or left for auditors. Five out of 26 activities (Activity 1, 9, 15, 16, and 23) were not chosen for automation but were left for auditors. Activity 1 and 9 are not rule-based and do not have structured and standardized data. Activity 15 and 16 are risk assessments that require auditors' professional judgment, so they do not follow predefined rules. Besides, the data required for activities 15 and 16 are not structured and usually are not in machine-readable formats. Activity 23 does not meet any of the criteria because it does not follow prescribed rules, is ad hoc, and does not have suitable data for automation.

Table 4: Activities Selected for Automation

Phase	No.	Activity	Rule-based	Repetitive	Structured data	Standardized data	Machine-readable data	Choose for automation
Pre-SEFA Testing	1	Compare CY SEFA with supporting documents to check the	**No**	Yes	**No**	**No**	Yes	**No**

		integrity of data on CY SEFA.						
	2	Compare program information from CY SEFA with that from PY SEFA to check the integrity of data on CY SEFA.	Yes	Yes	Yes	Yes	Yes	Yes
	3	Group programs on CY SEFA that are in the same cluster.	Yes	Yes	Yes	Yes	Yes	Yes
	4	Calculate the total expenditure for each cluster on CY SEFA.	Yes	Yes	Yes	Yes	Yes	Yes
SEFA Testing	5	Input total expenditure of current and previous year in SEFA Materiality worksheet.	Yes	Yes	Yes	Yes	Yes	Yes
	6	In SEFA Materiality worksheet, calculate the percentage change of the total expenditure from CY SEFA to PY SEFA and compare the change with the pre-set materiality.	Yes	Yes	Yes	Yes	Yes	Yes
Planning Step 1	7	On FPPW, answer the question “Do the total federal awards expended equal or exceed $750,000?”. If the total expenditure from CY SEFA is above $750,000, choose "Yes" as	Yes	Yes	Yes	Yes	Yes	Yes

		the answer. Otherwise, choose “No”.						
	8	On FPPW, if the answer to the previous question is "Yes", choose "No" for the second question to indicate that a single audit is required. Otherwise, select “Yes”.	Yes	Yes	Yes	Yes	Yes	Yes
Planning Step 2	9	Determine if the client is considered a low-risk auditee by going through questions in Planning Step 2 on FPPW with reference to previous year working papers.	**No**	Yes	**No**	**No**	Yes	**No**
Planning Step 3	10	Input the total expenditure into FPPW.	Yes	Yes	Yes	Yes	Yes	Yes
	11	On FPPW, choose the range of amount to which the total expenditure belongs.	Yes	Yes	Yes	Yes	Yes	Yes
	12	On CY SEFA, identify Type A and Type B programs.	Yes	Yes	Yes	Yes	Yes	Yes
	13	Update Federal Type A worksheet.	Yes	Yes	Yes	Yes	Yes	Yes
	14	Update Federal Type B worksheet.	Yes	Yes	Yes	Yes	Yes	Yes
	15	Assess risk level of each Type A program on Federal Type A Worksheet.	**No**	Yes	**No**	**No**	**No**	**No**
	16	Assess the risk level of each	**No**	Yes	**No**	**No**	**No**	**No**

	Type B program on Federal Type B Worksheet.						
17	If there is at least one Type A program, choose “Yes” for the question “Is there at least one Type A program” on FPPW. Otherwise, choose “No”.	Yes	Yes	Yes	Yes	Yes	Yes
18	If there is at least one Type A program that is not low risk, choose “Yes” for the question “Is there one or more Type A programs that may not be considered a low-risk program” on FPPW. Otherwise, choose “No”.	Yes	Yes	Yes	Yes	Yes	Yes
19	Input the threshold that classifies Type A and Type B programs into FPPW.	Yes	Yes	Yes	Yes	Yes	Yes
20	If there is at least one Type B program, choose “Yes” for the question “Are there any Type B programs?” on FPPW. Otherwise, choose “No.”	Yes	Yes	Yes	Yes	Yes	Yes
21	If there is at least one high-risk Type B program, choose “Yes” for the question “Are any of the Type B	Yes	Yes	Yes	Yes	Yes	Yes

	programs considered high-risk?” on FPPW. Otherwise, choose “No”.						
22	Update the Historical Summary of Major Programs.	Yes	Yes	Yes	Yes	Yes	Yes
23	Add extra major programs	**No**	**No**	**No**	**No**	**No**	**No**
24	Calculate the total expenditures of all major programs.	Yes	Yes	Yes	Yes	Yes	Yes
25	Input the value of the total expenditure of all major programs into FPPW.	Yes	Yes	Yes	Yes	Yes	Yes
26	On FPPW, if the total expenditure of all major programs is above a pre-set percentage of the total expenditure of all programs, choose “Yes” for the question “Has the percentage of coverage requirement been met?”. Otherwise, choose “No”.	Yes	Yes	Yes	Yes	Yes	Yes

Redesign (Step 4)

Before automation, auditors need to consider whether an existing audit procedure should be redesigned to make automation easier (Alles et al. 2008; Rozario and

Vasarhelyi 2018; Huang and Vasarhelyi 2019). We consider redesign from three aspects following the APA framework: resource, activity/task, and process.

Resource Redesign

In resource redesign, auditors consider modifying the resources (e.g., software, files, data) involved in each activity to adapt to automation, such as transforming data into machine-readable formats,[19] structuring data in columns and rows, and standardizing data fields and data type. In this demonstration case, we redesigned the Current Year (CY) SEFA and other Excel audit working papers to make them structured. The unaudited CY SEFA provided by different clients usually has different layouts, which is not ideal for automation that requires standardized and structured inputs (Cohen et al., 2019; Huang and Vasarhelyi, 2019). To illustrate, in the SEFA example provided in Figure 2, different federal programs are grouped into their corresponding clusters. The total expenditure of that cluster is calculated at the end of each cluster. However, different auditees have different clusters, and the number of federal programs in each cluster varies among different auditees. Therefore, no “rules” can instruct an automation program to locate the federal programs, clusters, and total expenditures precisely.[20]

To structure and standardize SEFA, we adopted the SEFA template from the auditors' data collection form with the Federal Audit Clearinghouse (FAC). The auditee or auditors use this template at the end of the Single Audit to send the auditor’s report to FAC. Since the FAC uses the template to store Single Audit information into a database,

[19] For example, it is better to use Excel spreadsheet than PDF file, and it is better to use PDF file that is directly converted from Word Document than scanned PDF.

[20] Specifically, for different auditees, their program funds to be audited could be located in different columns with non-standardized column titles. The usage of merged cells further increases the challenge to precisely locate a cell across auditees.

it is structured and standardized. We suggested that the auditors require auditees to submit the unaudited SEFA using this template. The same information in Figure 2 is presented in the SEFA template in Figure 4.

Figure 4: Standardized SEFA Template from Federal Audit Clearinghouse

1. Federal Awards Expended During Fiscal Period

Schedule of Expenditures of Federal Awards

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
	CFDA #								Loan Programs		Federal Award Source			Passed Through	
Row Number (auto-generated)	Federal Awarding Agency Prefix 1	CFDA Three-Digit Extension 2	Additional Award Identificat	Federal Program Name	Amount Expended ($)	Cluster Name	Federal Program Total 4 (auto-generated) ($)	Cluster Total 5 (auto-generated) ($)	Loan/Loan Guarantee (Loan) Y/N	If Loan, the End of the Audit Period Outstanding Loan Balance 6 ($)	Direct Award (Direct) Y/N	If not Direct, list Name of Pass-through Entity	If not Direct, list Identifying Number Assigned by the Pass-through Entity, if assigned 7	Federal Award Passed Through to Subrecipients Y/N	If Passed Through, provide Amount Passed Through ($)
1	84	010		TITLE I GRANTS TO LOCAL EDUCATIONAL AGENCIES	$521,801	N/A	$521,801		N		N	STATE OF NEW JERSEY DEPT. OF EDUCATION	N/A	N	
2	84	365		ENGLISH LANGUAGE ACQUISITION STATE GRANTS	$13,415	N/A	$13,415		N		N	STATE OF NEW JERSEY DEPT. OF EDUCATION	N/A	N	
3	84	367	PART A	IMPROVING TEACHER QUALITY STATE GRANTS	$52,796	N/A	$52,796		N		N	STATE OF NEW JERSEY DEPT. OF EDUCATION	N/A	N	
4	84	186		SAFE & DRUG FREE	$25,892	N/A	$25,892		N		N	STATE OF NEW JERSEY DEPT. OF EDUCATION	N/A	N	
5	84	027		SPECIAL EDUCATION_GRANTS TO STATES	$132,143	SPECIAL EDUCATION CLUSTER (IDEA)	$132,143	$134,320	N		N	STATE OF NEW JERSEY DEPT. OF EDUCATION	N/A	N	
6	84	173		SPECIAL EDUCATION_PRESCHOOL GRANTS	$2,177	SPECIAL EDUCATION CLUSTER (IDEA)	$2,177	$134,320	N		N	STATE OF NEW JERSEY DEPT. OF EDUCATION	N/A	N	
7	84	282		CHARTER SCHOOLS	$168,768	N/A	$168,768		N		N	STATE OF NEW JERSEY DEPT. OF EDUCATION	N/A	N	
8	10	553		SCHOOL BREAKFAST PROGRAM	$73,613	CHILD NUTRITION CLUSTER	$73,613	$396,081	N		N	STATE OF NEW JERSEY DEPT. OF EDUCATION	N/A	N	
9	10	555		NATIONAL SCHOOL LUNCH PROGRAM	$322,468	CHILD NUTRITION CLUSTER	$322,468	$396,081	N		N	STATE OF NEW JERSEY DEPT. OF EDUCATION	N/A	N	
10	10	558		CHILD AND ADULT CARE FOOD PROGRAM	$24,121	N/A	$24,121		N		N	STATE OF NEW JERSEY DEPT. OF EDUCATION	N/A	N	
				Total Federal Awards Expended =	$1,337,194										

Similarly, we also adjusted the formats of other relevant working papers (i.e., SEFA Materiality Worksheet, Federal Type A Worksheet, Federal Type B Worksheet, and Historical Summary of major Programs Worksheet) to make them structured and standardized.

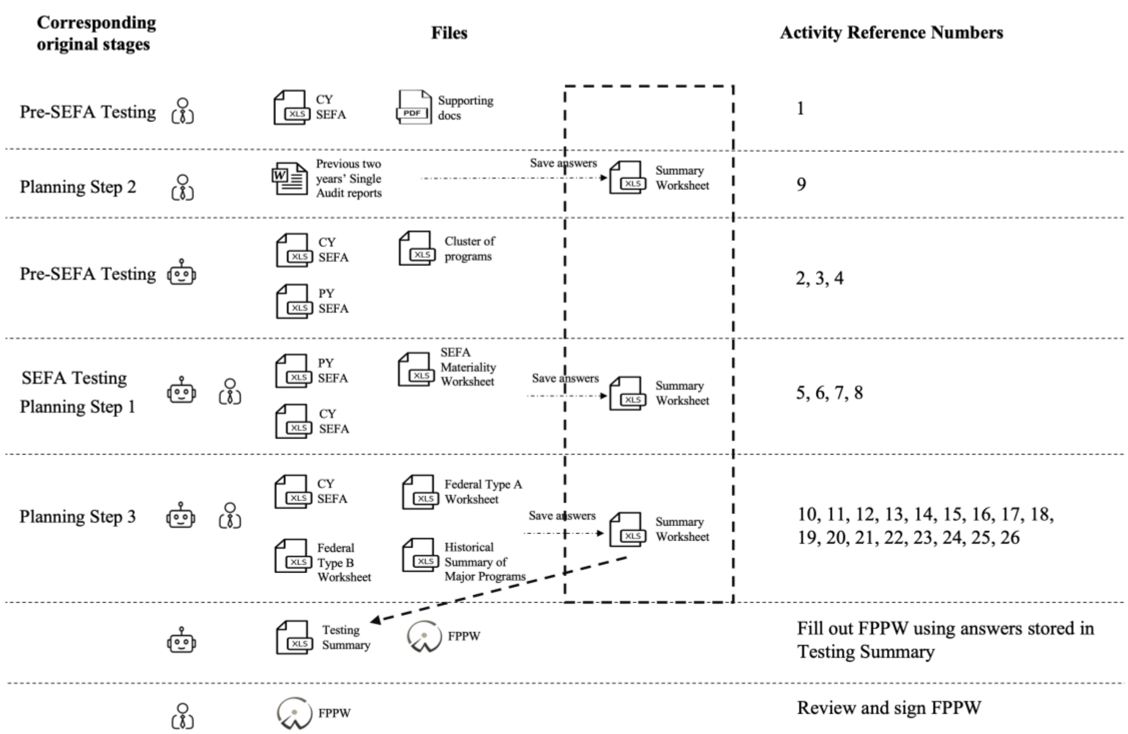
Activity/Task Redesign

Here the auditors consider improving how an activity/task is performed to enhance efficiency. In Single Audit Planning, the activity/task redesign is mainly about transforming the prior human-executed process into a robot-automated process by adopting programming logic (Rutschi and Dibbern 2020). For example, in activity 3 (Group programs on CY SEFA in the same cluster), instead of automating the tasks by mimicking how an auditor performs each task, we used programming logic, such as database queries, to simplify and automate this activity.[21]

Process Redesign

Next, we considered adjusting the sequence of how each activity is executed to adapt to the attended automation workflow. The redesigned process is presented in Figure 5. First, we moved manual activities 1 and 9 upfront in the workflow because they can be standalone activities (i.e., their execution does not depend on results from other activities). Manual activities 15, 16, and 23 depend on results from their previous activities, so they remained in their original sequence.

Figure 5: The Redesigned Process Overview



Note:

- Activity reference numbers are from Table 3
- Human icon indicates manual activities
- Robot icon indicates automated activities

[21] Such as the "groupby()" function of the pandas Python library and the SQL queries.

The process was next modified to smooth the workflow. An Excel spreadsheet named "Summary Worksheet" was created to store the answers required to complete the FPPW. In Planning Steps 1-3 in the original process (Figure 3), auditors could open FPPW in CaseWare and complete it by *simultaneously* referring to different Excel audit working papers, performing rule-based data analysis, and executing professional judgment when needed. While this sequence of activity execution is intuitive in a manual process, it is inefficient for an automated workflow because process automation usually follows a structured flow of logic rather than an unstructured and ad-hoc sequence of activities. The Summary Worksheet is presented in Figure 6. In the redesigned process (Figure 5), an auditor will first manually perform Planning Step 2 to provide answers to questions "Step2.1", "Step2.2", and "Step2.3" on the Summary Worksheet. Other activities (except for activities 7, 8, 17, and 18) will be automatically executed to provide answers to other questions on the Summary Worksheet. At the end of Planning Step 3, the results stored in the Summary Worksheet will be directly entered into the FPPW. Lastly, the auditor will review and sign the automatically filled FPPW.

Figure 6: The Summary Worksheet

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Step1.1	Step1.2	Step 2.1	Step 2.2	Step 2.3	Step 3.1	Step 3.3	Step 3.4	Step 3.5	Step 3.6	Step 3.7.b	Step 3.7.c	Step 3.8	Step 3.9	Step 3.10
2	Yes	No	Yes	Yes	No	$35,914,675	b	Yes	Yes		$1,077,440.25	Yes	Yes	$29,294,532	Yes
3															
4															
5															
6															

Activity Automation (Step 5)

We chose the appropriate automation tools to automate the redesigned audit process/activity/task at this stage.[22] For the redesigned activity in which the FPPW is

[22] When automating activities, we obtained remote access to the firm's selected audit working papers with the auditees' information anonymized. The configuration environment was close to auditors' real working environment.

completed based on results saved in the Summary Worksheet, we automated it using UiPath RPA software by mimicking auditors' actions of clicking checkboxes and typing numbers. RPA is suitable to automate this activity because, without an application programming interface (API), it is relatively easy to mimic human behavior from the user interface than hard coding to interact with CaseWare.

Other automatable activities we identified were related to Excel automation. Available tools for Excel automation include but are not limited to the Excel automation packages provided by RPA software tools, Visual Basic for Applications (VBA), and the Python pandas package. If the activities to be automated in Excel are simple, such as copying and pasting and inputting functions, one can consider using the Excel automation packages provided by RPA software. However, when the activities involve many steps or require complicated data analysis, other tools like VBA or Python are often more efficient. In this case, since there were many steps involved in the audit testing, and some testing required complex logic (especially grouping federal programs by clusters), Python is adopted.[23] [24]

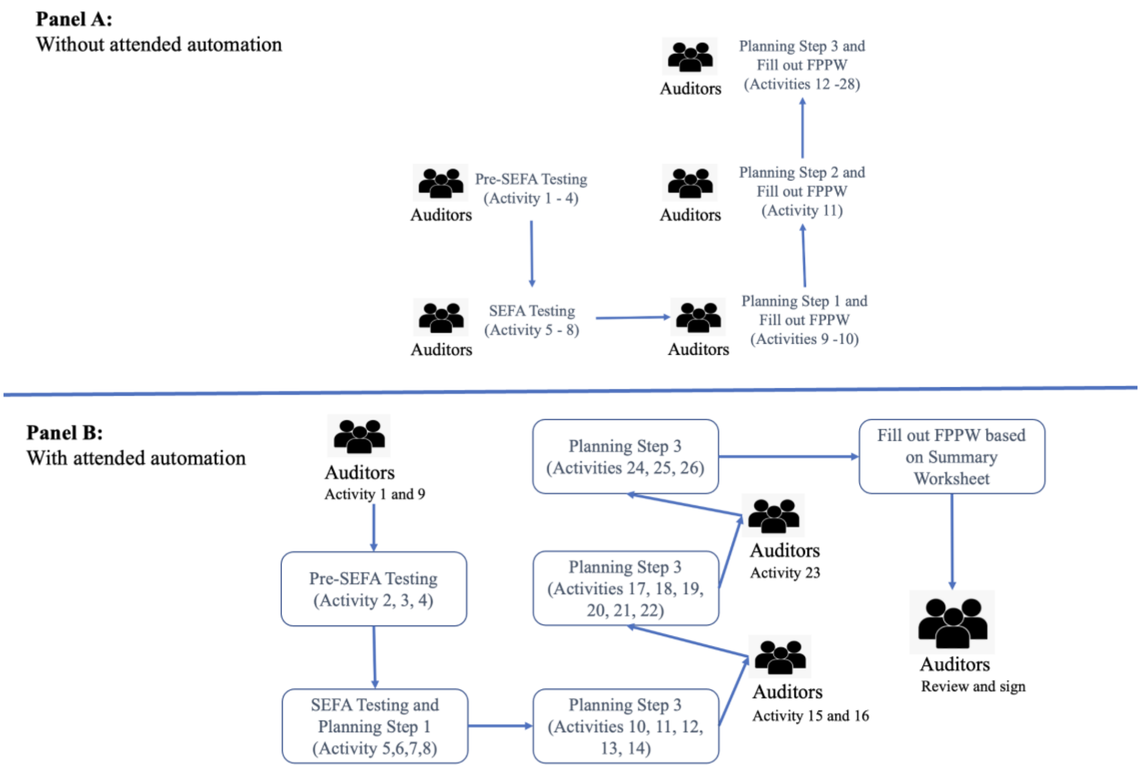
Coordination (Step 6)

Then, the automated and manual activities together were coordinated into an attended automation workflow. The flowchart of the attended audit planning workflow is provided in Panel B, Figure 7. The rectangles in Figure 7 represent automated tasks, and the human icons indicate tasks performed by auditors. RPA software was used to

[23] Other tools like SQL can achieve the same objective.
[24] The use of Python programming *alone* is not RPA. Instead, the use of programming like Python should be used in conjunction with or layered with RPA. In this particular demonstration case study, Python is embedded within the overarching RPA infrastructure to help better perform certain complicated data analysis tasks that are costly to be directly configured using RPA.

coordinate automated and human activities. To better compare the process with and without attended process automation, we visualize the original process in Panel A, Figure 7.

Figure 7: Audit Planning of Single Audits with and Without Attended Automation



Note:

- Activity reference numbers are from Table 3
- The human icon indicates manual activities
- Activities enclosed in rectangles are automated

When auditors finish activities 1 and 9, the automated workflow will be triggered to complete pre-SEFA testing, SETA-Testing, Planning Step 1, and the first part of Planning Step 3. At the end of activity 14 in Planning Step 3, the automated workflow will notify the auditors that they can perform activities 15 and 16 (provide risk assessments to Type A and Type B programs) based on prior activities. After auditors finish the risk assessments and save the file,[25] the workflow will be triggered and continue from activity 17. A similar interaction between the machine and the auditor happens at activity 23. After the FPPW is automatically populated, the automated workflow will notify auditors to review and sign the FPPW.

Evaluation (Step 7)

This step evaluates the prototype constructed from the above steps.[26] Following the objective setting, auditors focus on the efficiency improvements of the *mundane* audit activities. With the RPA prototype built for the firm, the time spent on the *deterministic part* of the audit planning was reduced from approximately three hours to no more than five minutes, leaving auditors more time to focus more on tasks that require higher levels of professional judgment.[27] This CPA firm performs at least 500 Single Audits each year. Thus, the time estimated to be saved from automating deterministic audit planning tasks can be around 1500 hours (3*500) each year.

Furthermore, auditors from this CPA firm expressed concerns about the audit quality of their original Single Audit planning process. Their specific concerns included spending time correcting data-entry errors, obtaining data from different sources, and manually converting data into useable formats. Implementing attended automation helps to mitigate these specific issues by 1) further standardizing the structured audit tasks and audit working papers; 2) improving the accuracy of data processing and data entry, and 3)

---

[25] In this prototype, auditors need to perform activity 15 and 16 in real time. However, in production, the bot can be configured in a way that leaves auditors a long period of time to perform activity 15 and 16.
[26] In section 2.5, we discuss the validation of the APA framework.
[27] The five minutes do not contain time to re-format the working papers. Re-formating or redesigning is a one-time task and once completed, the reformatted templates will be used firm-wide.

allowing auditors to allocate more time to complicated tasks that require higher levels of cognitive effort and professional judgments.

Applying the APA framework in this real-world context demonstrates how the framework can be applied to automate audit tasks and illustrates the framework's feasibility. In addition, the CPA firm that participated in this study chose to adopt the framework methodology and prototype. Currently, they are continuing to adapt both to fit their needs as a firm better. [28]

2.5 Validation of The APA Framework

This study aims to provide a methodological framework that guides the implementation of attended automation in audits. Therefore, it was essential to hear and incorporate feedback directly from those most likely to use the framework. Input from RPA experts, public accounting professionals, and accounting students with basic RPA knowledge was used to inform the construction of the framework like that used in prior research (No, Lee, Huang, and Li 2019). This input was necessary for developing the framework to ensure that it would successfully guide human-bot cooperation during the attended audit workflow. The CPA firm’s decision to use the framework and the prototype validates this framework's useability in practice. It also demonstrates the framework’s viability to other parts of the audit process. Like the CPA firm in this study, other CPA firms can adapt the framework and prototype to fit their specific needs.

---

[28] A general timeline of developing this demonstration case study is as follows. Objective setting and process understanding: about 7 weeks, including visiting the audit firm, meetings with auditors, and analyzing documents provided by auditors to understand the Single Audit planning process. Activity identification and process redesign: about 2 weeks, including identifying tasks suitable for automation and human, and redesigning the process. Bot configuration and coordination: about 3 weeks, including environment setup, bot development, troubleshooting, and debugging. Most of the time spent in this step is in troubleshooting and debugging. Since the audit firm provided us with an opinion-based evaluation until this point, we do not provide a time estimate for the evaluation step.

2.6 Discussion

2.6.1 Implications of Attended Process Automation to the Audit Profession

Attended automation in audit is consistent with the concept of "auditors-in-the-loop," which emphasizes the importance of auditors in an automated audit workflow in providing professional judgment and human intervention that cannot be replaced by automation (Zhang 2019). It also underscores that the role of RPA or automation in general in the audit should be assisting auditors in handling mundane and highly repetitive tasks rather than replacing auditors. In a time-constrained audit engagement, if automation can take over the tedious and routine tasks, not only can the accuracy and efficiency of these tasks be improved, auditors will also be left with more time to exercise professional judgment (Alles et al. 2008; Moffitt et al. 2018; Huang and Vasarhelyi 2019). Since attended automation acts as a digital assistant that a human user can call upon on-demand to help complete specific tasks (Ostdick 2017; Mullakara and Asokan 2020), auditors' job satisfaction is expected to improve because they do not have to struggle with repetitive, rule-based, and tedious tasks (Cooper et al. 2021). However, since auditors will have more time available for challenging and high value-add tasks, higher standards will be set for auditors' professional skepticism and analytical skills (Zhang, Dai, and Vasarhelyi 2018). Furthermore, auditors may take on new roles to audit the automation tools used in the audit procedure (Appelbaum and Kozlowski 2020).

2.6.2 Limitations

Although attended RPA is expected to deliver benefits such as enabling auditors to focus on more challenging tasks and increasing the overall audit efficiency, it may lead to unintended consequences. For example, an error or a bot malfunction may cause mistakes in the automating engagements. To reduce the chance of this type of mistake, the automation engagement, specifically the bot, should be strictly tested for errors and malfunctions before deployment and be regularly reviewed and monitored after deployment. In addition, as a part of the attended automation process and consistent with typical auditing methodology, auditors need to review the results generated by the automated activities for reliability and reasonableness. Detailed bot documentation is crucial because it can explain the bot's nature, function, and governance to satisfy quality and reliability requirements from management and regulators. It is noted that a “meta-bot” can be configured to assist the auditor with these types of review activities by automatically monitoring and auditing the bots that are assisting the auditor.

Another limitation to this RPA study is that by the time this study was completed, the CPA firm that supported the demonstration case study had not completely rolled out attended automation in their audit practice. They needed more time to research applicable auditing standards and regulations associated with using automation in auditing.

2.7 Concluding Remarks

This paper proposed a methodological framework to guide public accounting professionals with at least basic RPA knowledge in implementing attended automation in audits. A demonstration of how to apply the APA framework was provided using the planning process of Single Audits. This framework for attended automation emphasized the auditor’s role in an automated workflow while still utilizing their professional judgment. This study also provided another example of the use of DSR to produce practice-relevant knowledge in auditing. Since this study demonstrated the APA framework using only the audit planning of Single Audits, future research can apply the

APA framework to other settings of external audits and highlight the specific circumstances in which the framework would be the most useful in auditing. Therefore, there is room for further research in this area that could empirically address the usefulness of this framework and its impact on auditor judgment and decision making in an experimental setting. Other topics to consider for future research include examining the challenges and potential risks (e.g., unintended consequences from bot errors) of RPA adoption in audits, the governance and deployment models of RPA in audits, and the cost and benefit analysis of RPA adoption in different audit settings is also of interest.

CHAPTER 3: IDENTIFYING INFORMATIVE AUDIT QUALITY INDICATORS (IAQI) USING MACHINE LEARNING

3.1 Introduction

Which aspects of an audit can best predict audit failure? Investors, auditors, academics, and regulators around the world are increasingly searching for answers to this question, especially after the recent accounting scandals of Wells Fargo (Wall Street Journal 2016) and Wirecard (Financial Times 2020). Following prior literature (e.g., Francis 2004; Francis and Michas 2013; Li, Qi, Tian, and Zhang 2017), we define an audit failure, or a low-quality audit, as the failure to issue a modified or qualified audit report when there is a material misstatement in the audit client's financial statements.

Although extant research has tested a wide array of theories and developed a broad set of explanatory variables for audit quality (hereafter, audit-related variables, or ARV), little is known about the effectiveness of these variables in predicting audit failure out-of-sample and what variables are the most predictive. "Out-of-sample" refers to using holdout samples to measure model performance that is not used during model construction (Shmueli 2010; Bao, Ke, Li, Yu, and Zhang 2020). Examining out-of-sample predictive power can assess a model's performance in unfamiliar situations, thus reflecting the model's practical usefulness (Shmueli 2010; Shmueli and Koppius 2011). Assessing the out-of-sample predictive power of ARV can help researchers, regulators, and practitioners evaluate whether ARV that are commonly used in academic research provide practical value in alerting audit failure in novel scenarios. Furthermore, identifying the most predictive ARV can inform researchers, regulators, and practitioners which audit features are most relevant. This research aims to explore the following two questions:

1. How effectively can ARV, identified from and supported by prior literature, predict audit failure out-of-sample?

2. Which ARV are the most predictive of audit failure?

To address these questions, we use machine learning, which is a computational method that identifies hidden and non-linear patterns from large and high-dimensional datasets and selects a subset of variables that make the best out-of-sample predictions (Alpaydin 2014; Cecchini, Aytug, Koehler, and Pathak 2010; Bertomeu 2020). With a long list of theory-driven ARV and a sizable amount of historical data, machine learning is suitable for uncovering the predictive power of ARV and which ARV are the most predictive of audit failure.

In this machine learning research design, the outcome variable is audit failure, and the predictors are ARV. As a proxy for audit failure, we use material restatements of annual financial reports, generated due to violations of Generally Accepted Accounting Principles (GAAP) or fraud (hereafter, material annual restatements or MAR), guided by earlier examples of auditing literature (e.g., Lobo and Zhao 2013; Kinney, Palmrose, and Scholz 2004; Newton, Wang, and Wilkins 2013). MAR indicate that auditors signed off on materially misstated financial statements (Defond and Zhang 2014; Lobo and Zhao 2013; Center of Audit Quality 2013; Tan and Young 2015; Aobdia 2019; Audit Analytics 2020) because auditors are responsible for expressing opinions on whether financial statements are presented in conformity with GAAP and obtaining reasonable assurance about whether the financial statements are free of material misstatement, whether caused by error or fraud (AS 1001; Kinney et al. 2004; Stanley and DeZoort 2007; Newton et al. 2013; Francis, Michas, and Yu 2013; Eshleman and Guo 2014). Thirty-one publicly available ARV were gleaned from related literature for use as predictors and categorized as either audit inputs, audit processes, or audit outputs (Francis 2011).

We compile a dataset of U.S. public firms spanning the years 2005 to 2017 with 26,339 firm-year observations. Then, we use ARV to predict audit failure using a cost-sensitive learning and rolling-window prediction design with six commonly used machine learning algorithms (Logistic Regression or LR, Random Forest or RF, Support Vector Machine or SVM, AdaBoost or AB, Gradient Boosting or GB, and Extreme Gradient Boosting or XGB) (Perols 2011; Perols, Bowen, Zimmermann, and Samba 2017; Carmona et al. 2019; Bao et al. 2020; Brown, Crowley, and Elliott 2020; Ding, Lev, Peng, Sun, and Vasarhelyi 2020; Chen, Cho, Dou, and Lev 2020). We follow prior literature and measure each model's overall predictive ability using the area under the curve or AUC based on holdout testing samples (e.g., Bao et al. 2020; Brown et al. 2020).

We find that when all 31 ARV are used as predictors, the ensemble machine learning model, Gradient Boosting, has the highest average AUC of 0.727 during the hold-out test years 2015-2017. This level of AUC is considered acceptable in the machine learning literature (Mandrekar 2010; Hosmer, Lemeshow, and Sturdivant 2013). To put the predictive power of ARV in perspective, we identify from prior literature two groups of benchmark variables, most of which are financial variables. Statistical tests show that ARV have significantly higher predictive power measured in AUC than those benchmark variables, by a margin of roughly 8-14%. Interestingly, ARV also significantly outperform a combined model using both ARV and benchmark variables by about 2-4%. Taken together, we find that ARV's predictive power is acceptable and that they outperform benchmark variables in predicting audit failure.

To identify the most predictive ARV, which we term Informative Audit Quality Indicators (IAQI), we use each of the six machine learning algorithms and apply feature subset selection to select a subset of the ARV that have the most predictive power. Using a voting mechanism, we identify IAQI as the ARV that are determined to be the most predictive variables by most of the machine learning algorithms examined. This process generates 13 out of 31 variables to be regarded as IAQI: auditor tenure, auditor resignation, auditors' report of clients' internal control weakness, absolute value of total accruals, absolute value of total accruals scaled by cash flow from operation, audit report lag, audit office size, integrated audit, discretionary accruals, auditor city-level industry specialization, auditor competition, audit fees, and client influence. Table 5 summarizes the IAQI. We find that IAQI, which are a selected subset of ARV, can significantly outperform ARV by about 3%, as measured by AUC. In contrast, the least predictive ARV are auditor industry specialization at the national firm level, non-timely issuance of 10K due to audit issues, prior ROA meet, small profit, abnormal audit fee, non-audit fee ratio, auditor workload compression, and audit-related fees. Identifying the most and least predictive audit features serves as a reality check on which ARV are the most useful in predicting audit failure out-of-sample, and can inform researchers, regulators, and practitioners which variables to focus on when predicting audit failure.

Table 5: Informative Audit Quality Indicators (IAQI)

Category	**Sub-category**	**Aspects Captured**	**Variable**	**Measurement**
Audit input	Auditor characteristics	Competence, Resource, Independence	Office Size	Natural logarithm of one plus total annual audit fees of an audit office (Aobdia 2019)
		Competence, Information advantage	Industry Specialization_MSA	Auditor's annual market share based on audit fees within a two-digit SIC category for a particular Metropolitan

				Statistical Area (MSA) (Reichelt and Wang 2010)
Audit process	Task characteristics	Informational advantage, Independence	Tenure	Number of years that the company is audited by the same audit firm (Bell et al. 2015)
		Audit effort, Audit efficiency	Audit Report Lag	Natural logarithm of the number of days between fiscal year-end and the signature date of audit opinion (Lobo and Zhao 2013)
		Audit effort, Workload	Integrated Audit	Indicator variable equal to 1 when the audit engagement is an integrated audit of financial statements and internal controls, and 0 otherwise (Aobdia 2019)
	Auditor-client contracting features	Audit effort	Audit Fee	Natural logarithm of 1 plus the audit fees charged to the auditee (Aobdia 2019)
		Incentive	Auditor Resignation	Indicator variable equal to 1 if the current auditor will resign (instead of being dismissed by the company) from the next fiscal year, 0 otherwise (Krishnan and Krishnan 1997)
		Independence	Influence	The ratio of a company's total fees (i.e., audit fees plus non-audit fees) relative to the aggregate annual total fees generated by the local office that audits the company (Lopez and Peters 2012)
	Environmental characteristics	Incentive	Auditor Competition_MSA	MSA-level auditor concentration based on Herfindahl index. Details are provided in (Newton et al. 2013)
Audit output	Auditor communications	Competence, Independence	Internal Control Weakness	Indicator variable equal to 1 if a material weakness is reported for the year, 0 otherwise (Aobdia 2019)
	Quality of the audited financial statements	Within-GAAP manipulation	Disc. Accruals	Residual from the cross-sectional modified Jones model in Aobdia (2019)
			Abs (Accruals)	The absolute value of accruals deflated by beginning assets (Aobdia 2019)
			Abs (Accruals/CFO)	The absolute value of accruals deflated by cash flow from operations (Aobdia 2019)

To explore how researchers, regulators, and practitioners can more efficiently utilize IAQI, we synthesize IAQI into a predictive score by taking the standardized probability prediction from one of the best-performing machine learning models. We term this predictive score the P Score and compare it with the F Score from Dechow, Ge, Larson, and Sloan (2011) and the M Score from Beneish (1997, 1999) in a head-to-head comparison (Price, Sharp, and Wood 2011) of their incremental association with audit failure. A head-to-head comparison holds the sample, control variables, and dependent variables constant and only changes the testing variables (Price et al., 2011). We find that P Scores carry a significantly higher incremental association with audit failure than both F and M scores. P Score's advantage over F or M Scores could come from advanced machine learning models' ability to capture hidden and non-linear relationships among data.

Additionally, based on Bao et al. (2020), who use raw financial variables to predict accounting misconduct, we also explore the predictive powers of audit variables that are readily available in their raw or close-to-raw form from the Audit Analytics database. In total, we collect 91 such variables. We find that these raw or close-to-raw audit variables carry significantly higher predictive powers in terms of AUC than ARV by about 6% and IAQI by about 3%, consistent with Bao et al. (2020)'s finding that raw financial variables can outperform highly engineered financial ratios in accounting misconduct prediction.

To further examine whether the level of predictive power that can be achieved using audit related variables and the state-of-the-art machine learning model is considered useful by practitioners, we conducted a survey among a selected group of practitioners including audit partners, audit managers, senior auditors, and auditing standard setters. Nine out of 12 participants responded to our survey with high confidence (see Section VIII for more details), and they expressed an expected average AUC of 0.832 for an audit failure prediction model to be useful in practice. Since the average AUC that can be achieved by audit variables input into advanced machine learning models is within the range of 0.70-0.80, it appears that these models must be further refined to be considered practically useful.

Bertomeu, Cheynel, Floyd, and Pan (2020), in a near analog to the present study, use machine learning to predict material misstatements using a broad set of variables from accounting, capital markets, governance, and auditing. Their objective is to effectively detect material misstatement, regardless of sources of predictor variables. In contrast, the present study aims to inform academics, regulators, and practitioners on the practical usefulness of commonly used and theory-supported audit-related variables by examining their out-of-sample predictive power of audit failure. Assessing the predictive power of ARV can also help researchers, regulators, and practitioners assess the distance between theory and practice (Shmueli 2010; Shmueli and Koppius 2011) and identify the most relevant audit features. Furthermore, this study includes a more complete list of ARV (31 variables) than Bertomeu et al. (2020) (8 variables). Appendix presents a detailed comparison between this study and relevant literature.

This study makes several important contributions. First, it adds to the growing body of accounting research that uses machine learning to predict accounting outcomes that are difficult to model using traditional methods (e.g., Cecchini et al. 2010; Perols 2011; Perols et al. 2017; Bao et al. 2020; Brown et al. 2020; Ding et al. 2020; Bertomeu et al. 2020). We examine if ARV that are commonly used by academics have practical value in red-flagging audit failure. In a domain like audit quality, in which researchers have tested a wide array of theories and accumulated a broad knowledge of explanatory factors, it is vital to understand the out-of-sample predictive power of ARV that are operationalized based on theoretical constructs (Shmueli 2010; Shmueli and Koppius 2011). Although ARV can significantly outperform benchmark variables, and their predictive power is acceptable by academic standards, the predictive power of ARV has yet to exceed expectations from practitioners, suggesting a gap between theory and practice and, therefore, room for improvement.

Second, this study contributes to the stream of auditing literature that tries to validate audit quality measures (e.g., Aobdia 2019; Rajgopal, Srinivasan, and Zheng 2021). Specifically, it focuses on the perspective of the out-of-sample predictive power of audit quality measures. Although Aobdia (2019) investigates the degree of agreement between fifteen measures of audit quality used in academia and two proprietary measures of audit process quality, he focuses on the within-sample correlations rather than the out-of-sample predictive power of audit-related variables. Even though Rajgopal et al. (2021) evaluate how well existing audit quality proxies predict specific allegations related to audit deficiencies, they also construct within-sample evaluation metrics. This study furthers the understanding of the validity of audit-related variables by examining their out-of-sample predictive power. Understanding that predictive power can spur comparisons of competing theories, different operationalizations of constructs, and different measuring instruments (Shmueli 2010; Shmueli and Koppius 2011). The predictive audit-related variables (i.e., IAQI) identified in this study based on out-of-sample predictive power can guide future research in using these most relevant audit features in predicting audit failure. This study also suggests that some commonly used ARV do not have a strong out-of-sample predictive power of audit failure, such as non-audit fee ratio, abnormal audit fee, and auditors' firm-level industry specialization, suggesting future researchers should exercise caution in using these variables if the objective is to predict audit failure.

Third, this study provides tools that can be used in future research, including a predictive score (i.e., P Score) based on IAQI and advanced machine learning models that can capture non-linear relationships among data. P Score outperforms commonly used academic measures, F Score and M Score, in providing incremental association with audit failure. In future research, P Scores can be used as an alternative measure by academics and practitioners when inferring audit failure.

Fourth, findings from this research have implications for future audit research. The survey of audit partners, audit managers, senior auditors, and auditing standard setters suggests a desired average AUC of 83.2% in audit failure prediction. This threshold can be used as a benchmark for future studies that predict audit failure or material misstatement. Furthermore, we find that ARV have higher predictive power than benchmark financial variables, suggesting that future research should examine whether the external audit process or the clients' innate features are the driving factor for the observed audit failure. Moreover, since we find that raw audit variables have higher predictive power than either IAQI or ARV, future research can examine the source of incremental predictive power from raw audit variables.

Fifth, as a response to PCAOB's call for research on audit quality indicators using public source information (PCAOB 2015), this study has practical implications for regulators and other stakeholders, including audit committees, audit firms, and investors. When interpreted within specific contexts, IAQI and P Score can assist regulators and stakeholders in assessing audit quality and can facilitate relevant decision-making during the risk-based inspection process (PCAOB 2015; Eutsler 2020). In addition, these measures can aid audit firms in risk assessment and management while helping investors assess firms' reporting risks more effectively (PCAOB 2015). Furthermore, this study provides regulators and other stakeholders methodologies to be used to identify predictive variables and establish predictive scores using proprietary data (PCAOB 2019).

3.2 Machine Learning

Machine learning is a computational method that can identify hidden patterns from data and make predictions (Alpaydin 2014). Compared to traditional data analysis approaches, machine learning works better with large or high-dimensional datasets, requires fewer underlying assumptions, and can model complex and hidden patterns among variables (Alpaydin 2014; Cecchini et al. 2010; Bertomeu 2020). It is used in predictive modeling to predict new or future observations (Shmueli 2010). Predictive modeling is evaluated by out-of-sample measures that are derived from holdout samples, such as AUC (Shmueli 2010; Bao et al. 2020). Out-of-sample metrics can assess a model's performance on unseen situations, reflecting the model's predictive power (Shmueli 2010; Shmueli and Koppius 2011).

An important subset of machine learning is supervised learning, in which an algorithm learns from available examples or experiences with known positive or negative "labels" and then makes predictions about future instances (Alpaydin 2014). For example, after being provided with examples of known fraudulent and legitimate transactions, a supervised learning algorithm can be trained to extract identifiable patterns. The trained algorithm will then be able to predict whether a new transaction is fraudulent. Since labeled data is available (i.e., with known MAR or not), supervised learning is adopted for this study. Popular supervised learning algorithms used in accounting research include Logistic Regressions (LR), Random Forest (RF), Support Vector Machine (SVM), and AdaBoost (AB), Gradient Boosting (GB), and Extreme Gradient Boosting (XGB) (Alpaydin 2014; Cecchini et al. 2010; Perols 2011; Perols et al. 2017; Bao et al. 2020; Brown et al. 2020; Chen et al. 2020). Detailed descriptions for each of the popular supervised learning algorithm can be found in Alpaydin (2014) and Hastie et al. (2017b). Among the six algorithms, RF, AB, GB, XGB belong to a category called ensemble method that works by combining predictions from two or more models (Hastie et al. 2017b; Bao et al. 2020). Ensemble methods generally provide superior performance than single models (Hastie et al. 2017b; Bao et al. 2020).

Accounting researchers have explored the use of machine learning to predict fraud (Perols 2011; Perols et al. 2017; Cecchini et al. 2010; Bao et al. 2020), bankruptcy (Gentry, Shaw, Tessmer, and Whitford 2002), restatement/misstatement (Dutta, I., Dutta, S., and Raahemi 2017; Bertomeu et al. 2020; Hunt, E., Hunt, J., Richardson, and Rosser 2022), and accounting estimates (Ding et al. 2020). They also use machine learning as a tool to make automatic classifications or to extract or generate variables that can be used in explanatory-modeling studies (e.g., Li 2010; Sun and Sales 2018; Sun 2018; Hayes and Boritz 2019; Brown et al. 2020). This study adds to the accounting and auditing literature by using machine learning to assess the level of predictive power ARV have in predicting audit failure and identify audit-related variables that are the most predictive of audit failure.

3.3 Material Annual Restatements and Audit-Related Variables

3.3.1 Material Annual Restatements (MAR)

Audit quality is either unobservable or can only be observed when there are known errors or deficiencies (DeAngelo 1981; PCAOB 2015; Causholli and Knechel 2012). Consequently, researchers, regulators, and professionals often describe what high audit quality is not (i.e., in terms of errors or deficiencies that reduce audit quality) rather than defining audit quality for what it is (Knechel, Krishnan, Pevzner, Shefchik, and Velury 2013). In academic research, various proxies are used to measure audit quality. There is no consensus on what the best measure of audit quality is because different measures capture discrete aspects of an audit (Defond and Zhang 2014). Overall, compared to less direct measures of audit quality (e.g., accruals quality), more direct measures (e.g., restatements) can capture egregious audit failures and have higher consensus on measurement, but they are less able to capture the continuous nature of audit quality (Defond and Zhang 2014).

In this study, material annual restatements (MAR) of reports due to GAAP violations or frauds is used as the measure of audit failure based on the following reasons. First, we capture the aspect of materiality in audits (Christensen, Glover, Omer, and Shelley 2016) by focusing on the material restatements (a.k.a., “Big R” or re-issuance) that are disclosed in Item 4.02 of Form 8-K (Center of Audit Quality 2013; Audit Analytics 2020). Restatements announced in 8-Ks address a material error that requires re-issuance of past financial statements, and these "Big Rs" are the primary type of restatements to garner concern (Center of Audit Quality 2013; Audit Analytics 2020). In contrast, restatements disclosed in periodic reports (e.g., 10-Ks, 10-K/As. 10-Qs) are immaterial changes that are considered ongoing adjustments made in the ordinary course of business (Audit Analytics 2020). Therefore, material restatements are a tacit admission that auditors signed off on materially misstated financial statements (Defond and Zhang 2014; Lobo and Zhao 2013; Center of Audit Quality 2013; Tan and Young 2015; Aobdia 2019; Audit Analytics 2020).

Second, to reflect the expected responsibilities of auditors, we limit the reasons for material restatement to GAAP violations or frauds because auditors are held responsible for expressing opinions on whether the financial statements are presented in conformity with GAAP and obtaining reasonable assurance about whether the financial statements are free of material misstatement, whether caused by error or fraud (AS 1001; Kinney et al. 2004; Stanley and DeZoort 2007; Newton et al. 2013; Francis et al. 2013; Eshleman and Guo 2014). Furthermore, we restrict material restatements to those of annual reports because only the annual financial reports of public firms are required to be audited (Kinney et al. 2004; Stanley and DeZoort 2007; Cao, Myers, and Omer 2012; Lobo and Zhao 2013; Bills, Cunningham, and Myers 2016).

Third, there is a consensus in the literature that restatements can capture actual audit failure with little measurement error and that they are a relatively direct output-based measure of audit failure as compared to other proxies (e.g., Romanus et al. 2008;

Defond and Zhang 2014; Francis et al. 2013; Knechel and Sharma 2012; Newton et al. 2013; Ettredge et al. 2014; Eshleman and Guo 2014; Bills et al. 2016; Lennox 2016; Aobdia 2018; Bhaskar et al. 2019; Cunningham et al. 2019; Ahn et al. 2020; Rajgopal et al. 2021).

Fourth, both audit professionals and investors also identify financial statement restatements as the most readily available and outcome-based signal of audit failure since the existence of a restatement indicates that an improved audit process could have identified the error (Christensen et al. 2016; Gaynor et al. 2016).

Fifth, material restatements have been found to be strongly associated with the audit process quality measured by PCAOB's Part I Findings and with the internal assessments of audit quality from audit firms (Aobdia 2019).

Limitations of Using MAR as a Proxy for Audit Failure

While this paper uses MAR as a proxy of audit failure, it also acknowledges the limitations of using this proxy. First, not all poor audit quality incidences produce a material restatement. A material annual restatement is a joint outcome of three events: a material misstatement happened, auditors failed to identify the misstatement, and the misstatement was eventually unveiled and disclosed (Gaynor et al. 2016). Therefore, an absence of MAR does not necessarily indicate high audit quality. Second, MAR cannot capture the procedure-related characteristics of audit quality, such as the extent and appropriateness of evidence supporting the auditor's opinion, and the degree of correspondence between the auditor's procedures and auditing standards (Bell, Causholli, and Knechel 2015). Third, evidence from lawsuits data of whether auditors are held accountable for restatements is mixed.

Despite its limitations, MAR remains the most readily accessible indicator of audit failure from public source information (Christensen et al. 2016) and it is commonly used in audit quality research (Romanus et al. 2008; Defond and Zhang 2014; Francis et al. 2013; Knechel and Sharma 2012; Newton et al. 2013; Ettredge et al. 2014; Eshleman and Guo 2014; Bhaskar et al. 2019).

3.3.2 Audit-Related Variables

ARV are variables that fall into the framework of audit input, audit process, and audit output, as specified in Francis (2011) and Gaynor et al. (2016). ARV were collected by obtaining a list of audit quality literature from Web of Science (WoS), which is a publisher-independent global citation database. The results were limited to articles published since 2003 (post-SOX) and until 2020 (year of research), yielding 1116 articles. To capture the most influential and widely used ARV, the articles were ranked by the number of times they were cited so that the 200 most cited papers could be manually reviewed, along with papers they referenced and those that cited them. Then, researchers extracted ARV by looking for variables that could be categorized as audit inputs, audit processes, and audit outputs (Francis 2011; Gaynor et al. 2016). ARV often appear measures of audit quality, audit-related features used as testing variables, or control factors of audit quality. This process yielded 31 ARV that are publicly available and supported by theoretical constructs.

Following the frameworks of Gaynor et al. (2016), Francis (2011), and PCAOB (2013), the 31 ARV are further classified into sub-categories, including auditor characteristics, task characteristics, environmental characteristics, auditor-client contracting features, auditor communication, and the quality of the audited financial statements. The aspects of an audit each ARV captures are summarized, following relevant literature. Table 6 provides the list of ARV and their classifications.

Table 6: Audit-Related Variables (ARV)

No.	Variable	Description	Main source	Category	Sub-Category	Aspects Captured
1	Industry Specialization_National	Auditor's annual market share based on audit fees within a two-digit SIC category (Aobdia 2019)	Aobida (2019); Balsam et al. (2003); Reichelt and Wang (2010); Romanus et al. (2008); Francis and Yu 2009; Asthana and Boone 2012	Audit input	Auditor characteristics	Competence, Informational advantage
2	Industry Specialization_MSA	Auditor's annual market share based on audit fees within a two-digit SIC category for a particular Metropolitan Statistical Area (MSA)[29](Reichelt and Wang 2010)	Reichelt and Wang (2010); Ferguson et al. (2003); Ahn et al. (2020); Francis and Yu 2009; Asthana and Boone 2012	Audit input	Auditor characteristics	Competence, Informational advantage
3	Office Size	Natural logarithm of one plus total annual audit fees of an audit office (Aobdia 2019)	Aobdia (2019); Choi et al. (2010); Francis and Yu (2009); Asthana and Boone 2012	Audit input	Auditor characteristics	Competence, Resource, Independence
4	Big 4	Indicator variable equal to one if the audit firm is a Big 4, and 0 otherwise (Aobdia 2019)	Lobo and Zhao (2013); Newton et al. (2013); Eshleman and Guo (2014) Francis and Yu 2009; Reichelt and Wang 2010	Audit input	Auditor characteristics	Competence, Resource, Independence
5	New Client (or Auditor Change)	Indicator variable equal to 1 if the auditor-client relationship is in its first year, and 0 otherwise (Aobdia 2019)	Aobida (2019); Francis et al. (2013); Cohen et al. (2014); Ettredge et al. 2014; Schroeder (2016)	Audit process	Task characteristic	Informational advantage, Independence

[29] According to Reichelt and Wang (2010), the geographical city available from Audit Analytics is not the MSA. MSA information is available from the U.S. Census Bureau.

6	Tenure	Number of years that the company is audited by the same audit firm (Bell et al. 2015)	Myers et al. (2003); Johnson et al. (2002); Knechel and Vanstraelen (2007); Lim et al. (2010); Bell et al. (2015); Stanley and DeZoort (2007); Francis and Yu (2009); Lobo and Zhao (2013); Reichelt and Wang 2010; Choi et al. 2010; Boone et al. 2010	Audit process	Task characteristic	Informational advantage, Independence
7	Local Auditor_ MSA	Indicator variable equal to 1 if the audit engagement office is located in the same MSA where audit clients are headquartered, and 0 otherwise (Choi et al. 2012)	Choi et al. (2012); Francis et al. (2013)	Audit process	Task characteristic	Informational advantage
8	Integrate d Audit	Indicator variable equal to 1 when the audit is an integrated audit of financial statements and internal controls, and 0 otherwise (Aobdia 2019)	Aobdia (2019); Bhaskar et al. (2019)	Audit process	Task characteristic	Audit effort, Workload
9	Accelerat ed Filer	Indicator variable equal to 1 for firms that are accelerated filers, and 0 otherwise (Newton et al. 2013)	Lambert et al. (2017); Ettredge et al. 2014; Newton et al. (2013)	Audit process	Task characteristic	Time pressure, Workload
10	Busy	Indicator variable equal to 1 if a company has a fiscal year-end date of December, and 0 otherwise (Lopez and Peters 2012)	Lopez and Peters (2012); Lobo and Zhao (2013)	Audit process	Environmental characteristics	Time pressure, Workload

11	Workload Compression	The relative level of workload compression of an auditor office during the fiscal year end month of the auditee[30] (Lopez and Peters 2012)	Lopez and Peters (2012)	Audit process	Environmental characteristics	Workload
12	Auditor Competition_MSA	MSA-level auditor concentration based on Herfindahl index. Details are provided in (Newton et al. 2013)	Newton et al. (2013); Francis et al. 2013; Ettredge et al. 2014	Audit process	Environmental characteristics	Incentive
13	Auditor Resignation	Indicator variable equal to 1 if the current auditor will resign (instead of being dismissed by the company) from the next fiscal year, 0 otherwise (Krishnan and Krishnan 1997)	Krishnan and Krishnan (1997); Huang and Scholz (2012)	Audit process	Auditor-client contracting features	Incentive
14	Audit Fees	Natural logarithm of 1 plus the audit fees charged to the auditee (Aobdia 2019)	Aobida (2019); Paterson and Valencia (2011); Cao et al. (2012); Francis et al. (2013); Lobo and Zhao (2013); Newton et al. (2013); Eshleman and Guo (2014); Cohen et al. (2014); Choi et al. 2010; Srinidhi and Gul 2007; Ettredge et al. 2014	Audit process	Auditor-client contracting features	Audit effort
15	Tax Fee	Natural logarithm of 1 plus the total tax fees charged to the auditee	Kinney et al. (2004); Paterson and Valencia (2011)	Audit process	Auditor-client contracting features	Independence, Informational advantage

[30] "For each month, we add the audit fees charged to clients with the same fiscal year-end month in each local office; we then divide each monthly sum by the total audit fees collected by the local office for the year." (Lopez and Peters 2012)

16	Audit-Related Fee	Natural logarithm of 1 plus the audit-related fees charged to the auditee	Kinney et al. (2004); Paterson and Valencia (2011)	Audit process	Auditor-client contracting features	Independence, Informational advantage
17	Other Fees	Natural logarithm of 1 plus the other fees charged to the auditee	Kinney et al. (2004); Paterson and Valencia (2011)	Audit process	Auditor-client contracting features	Independence, Informational advantage
18	Non-Audit Fee Ratio	Non-audit fees deflated by total fees paid (audit plus non-audit fees). Non-audit fee equals to the sum of benefit fee, IT fee, Tax fee, audit related fee, and other fees. (Ruddock et al. 2006)	Lim and Tan (2008); Shinidhi and Gul (2007); Ruddock et al. (2006); Cao et al. (2012); Newton et al. (2013); Cohen et al. (2014); Boone et al. 2010; Ettredge et al. 2014	Audit process	Auditor-client contracting features	Independence, Informational advantage
19	Influence	The ratio of a company's total fees (i.e., audit fees plus non-audit fees) relative to the aggregate annual total fees generated by the local office that audits the company (Lopez and Peters 2012)	Lopez and Peters (2012); Francis and Yu (2009); Eshleman and Guo 2014; Ettredge et al. 2014	Audit process	Auditor-client contracting features	Independence
20	Abnormal Audit Fee	The unscaled residual from the audit fee model used in Blankley et al. (2012)[31]	Blankley et al. (2012); Asthana and Boone (2012); Lobo and Zhao (2013); Schroeder (2016)	Audit process	Auditor-client contracting features	Abnormal audit effort
21	Audit Report Lag	Natural logarithm of the number of days between fiscal year-end and the signature date of audit opinion (Lobo and Zhao 2013)	Knechel and Sharma (2012); Lobo and Zhao (2013); Ettredge et al. 2014; Asthana and Boone 2012; Lambert et al. 2017	Audit process	Task characteristic	Audit effort, Audit efficiency

[31] We base the model on Blankley et al. (2012) but use the definition of ICWeak from Newton et al. (2013).

22	Non-timely Issuance of 10K_Due to Audit	Indicator variable equal to 1 if the company filed 10-K late and the lateness is due to audit, 0 otherwise	Cao et al. (2016); Lambert et al. (2017); Wang et al. (2013)	Audit process	Task characteristic	Audit effort, Audit efficiency
23	Going Concern	Indicator variable equal to 1 if auditor gave a going concern opinion (Aobdia 2019)	Aobida (2019); Carey and Simnett (2006); Lobo and Zhao (2013); Ettredge et al. (2014); Lennox (2016); Reichelt and Wang 2010; Francis and Yu 2009; Lim and Tan 2008; Boone et al. 2010; Ettredge et al. 2014	Audit output	Auditor communication	Independence, Competence
24	Internal Control Weaknes s	Indicator variable equal to 1 if a material weakness is reported for the year, 0 otherwise (Aobdia 2019)[32]	Aobdia (2019); Anantharaman and Wans (2019); Francis and Yu 2009; Ettredge et al. 2014	Audit output	Auditor communication	Independence, Competence
25	Disc. Accruals	Residual from the cross-sectional modified Jones model in Aobdia (2019)	Aobida (2019); Dechow et al. (1995); Kothari et al. (2005); Reichelt and Wang (2010)	Audit output	Quality of the audited financial statements	Within-GAAP manipulation
26	Abs (Disc. Accruals)	Absolute value of Disc. Accruals (Aobdia 2019)	Aobida (2019); Dechow et al. (1995); Kothari et al. (2005); Reichelt and Wang (2010); Bills et al. (2016); Krishnan, Krishnan, and Song (2017)	Audit output	Quality of the audited financial statements	Within-GAAP manipulation
27	DD Residual	Residual from the Dechow and Dichev model in Aobdia (2019)	Aobida (2019); Dechow and Dichev (2002); McNichols (2002)	Audit output	Quality of the audited financial statements	Within-GAAP manipulation
28	Abs (Accruals)	The absolute value of accruals	Aobida (2019); Leuz et al. (2003)	Audit output	Quality of the audited	Within-GAAP manipulation

[32] Here, a value of 1 indicates that the audit engagement is an integrated audit and the auditor expressed weakness in ICFR.

		deflated by beginning assets (Aobdia 2019)			financial statements	
29	Abs (Accruals /CFO)	The absolute value of accruals deflated by cash flow from operations (Aobdia 2019)	Aobida (2019); Leuz et al. (2003)	Audit output	Quality of the audited financial statements	Within-GAAP manipulation
30	Small Profit	Indicator variable equal to 1 if ROA is between 0% and 3%, 0 otherwise (Aobdia 2019)	Aobida (2019); Francis and Yu (2009)	Audit output	Quality of the audited financial statements	Within-GAAP manipulation
31	Prior ROA Meet	Indicator variable equal to 1 if the year-on-year change in ROA is between 0% and 1%, 0 otherwise (Aobdia 2019)	Aobida (2019)	Audit output	Quality of the audited financial statements	Within-GAAP manipulation

3.4 Data and Research Design

3.4.1 Data and Sample

We started with the COMPUSTAT population from 2000 to 2019.[33] After removing firm-year observations that do not have Central Index Key (CIK) numbers, that do not file 10-K, and that are duplicates, we matched the remaining data with the Audit Analytics dataset by fiscal year-end.[34] After removing observations with missing values and keeping observations from 2005 to 2017, we had 26,339 firm-year observations. The details of sample derivation are provided in Table 7. We chose the starting year to be 2005 because the Form 8-K disclosure requirement came into effect in August of 2004.[35]

[33] We download the data from 2000 so that we can calculate tenure (i.e., the length of auditor-client relationship) accurately.
[34] COMPUSTAT and Audit Analytics have different ways of deciding fiscal year. Therefore, we use the field "Data Date" in COMPUSTAT and "Fiscal year ended" in Audit Analytics for matching.
[35] See https://www.sec.gov/rules/final/33-8400.htm

The data ended in 2017 because there is an average of 2-year lag between the restatement filing date (the most recent filing date before this study was 2019) and the restated date (Lobo and Zhao 2013; Eshleman and Guo 2014).

Table 7: Sample Determination

	Number of firm-year observations
COMPUSTAT population from 2000 to 2019	224,047
Less: observations without CIK	-25,279
Less: observations that do not file 10K	-16,615
Less: duplicated CIK and Fiscal Year-End	-23,761
Match with Audit Analytics by fiscal year-end	
Less: missing audit fee records	-48,604
Less: audit fees reported in foreign currencies	-4,723
Less: observations without going concern opinion records	-4,105
Less: observations with foreign auditors or business	-10,933
Less: observations without SIC code	-657
Less: observations without MSA information	-9,497
Less: observations without abnormal audit fee	-31,696
Less: observations without audit fee lag	-59
Less: missing auditor resigned data	-520
Less: observations with infinite values of Industry Specialization_MSA	-3
Less: observations with infinite values of workload compression	-3
Less: observations without discretionary accruals	-5,167
Less: observations with missing DD residual variable	-6,609
Less: observations outside 2005 to 2017	-8,975
Less: observations that are not first-year restatements in serial restatements	-502
Final sample size	**26,339**

We obtained restatement data from the Audit Analytics database, which houses a complete population of restatements records originated from Form 8-Ks or periodic reports (Karpoff, Koester, Lee, and Martin 2017; Lobo and Zhao 2013; Audit Analytics 2020).[36] To derive MAR, we first consulted the “Non-Reliance Restatements” section of the Audit Analytics database and downloaded restatement data from 2000 to 2019. We

[36] Other databases such as the one associated with the Center for Financial Reporting and Management (CFRM) mainly provide misstatements disclosed on AAERs that were generated due to SEC investigations for accounting or auditing misconduct or that led to lawsuits (Karpoff et al. 2017). Accordingly, such databases are more often used in research related to fraud/misconduct prediction.

selected only 8-K sources of disclosure (Center of Audit Quality 2013; Audit Analytics 2020),[37] then further restricted the selection to material restatements that could be ascribed to GAAP violations or fraud.[38] Finally, we excluded interim restatements and kept the annual restatements.[39] The resulting restatement instances are MAR. For the restated firm-year observations in the sample, the "year" denotes the fiscal year in which the firm's annual report ultimately received material restatement and not the year the restatement was announced.

Additionally, some MAR in our sample occur across consecutive years. For example, a firm may have materially restated its 10-K in 2011, 2012, and 2013. According to the data provider, most consecutive instances occur because the material misstatement from the starting year carries over to subsequent years; therefore, a MAR's starting year is usually the year when the material misstatement originated. To capture audit failure more accurately, we retain only the starting year of MAR instances in our sample (Stanley and Dezoort 2007). [40] It also is necessary to remove repeated MAR because serial MAR that span both the training and testing periods could overstate the performance of ensemble learning algorithms (e.g., AdaBoost and Random Forest) (Bao et al. 2020).

---

[37] In the "Non-Reliance Restatements" section of the Audit Analytics database, there is one data field that provides information about the source document in which the restatement has been announced.

[38] In the "Non-Reliance Restatements" section of the Audit Analytics database, there are data fields indicating whether a restatement is related to GAAP violations, fraud (financial fraud, irregularities and misrepresentations), or clerical errors.

[39] The regular module of Audit Analytics only provides a time range of the restated financial statements, which comprise both interim restatements and annual restatements. We use a Python script to derive the annual restatements and will provide the Python codes by request. This way of identifying annual restatements is consistent with Audit Analytics 2020.

[40] For example, if a firm restated its 10-K in consecutive years from 2011 to 2013, we only keep the observation for 2011 and delete those for 2012 and 2013.

Our sample contains 446 starting-year material annual restatement observations, which account for around 1.69% of the entire population. Table 8 presents the starting-year distribution of MAR by fiscal year. The distribution of our sample is consistent with Audit Analytics' 2020 restatement report (Audit Analytics 2020). Table 9 provides the descriptive statistics of the 31 ARV, most of which are comparable to previous studies. The pairwise correlations of ARV are provided in Appendix.

Table 8: Sample Distribution by Years

Fiscal Year	Number of Firms	Number of Starting year MAR	Percentage
2005	2188	69	3.15%
2006	2066	48	2.32%
2007	1864	31	1.66%
2008	2054	44	2.14%
2009	1894	27	1.43%
2010	1981	45	2.27%
2011	1946	35	1.80%
2012	1963	35	1.78%
2013	2076	31	1.49%
2014	2064	26	1.26%
2015	2049	25	1.22%
2016	2087	11	0.53%
2017	2107	19	0.90%
Total	**26339**	**446**	**1.69%**

*Note: MAR is material annual restatement due to GAAP violations or fraud.

Table 9: Descriptive Statistics*

Variable	Mean	Std. Dev.	Min	Max	Comparable research, if any
Industry Specialization_National	0.17	0.16	0.00	1.00	Aobdia (2019)
Industry Specialization_MSA	0.44	0.37	0.00	1.00	Choi et al. (2012); Francis et al. (2013); Lopez and Peters (2012)
Office Size	16.19	2.14	8.01	20.22	Aobdia (2019); Newton et al. (2013); Lopez and Peters (2012)
Big 4	0.63	0.48	0.00	1.00	Aobdia (2019); Newton et al. (2013); Choi et al. (2012)
New Client	0.11	0.31	0.00	1.00	Aobdia (2019)
Tenure	6.19	4.26	1.00	18.00	Lobo and Zhao (2013); Bell et al. (2015)
Local Auditor_MSA	0.68	0.47	0.00	1.00	Choi et al. (2012); Francis et al. (2013)
Integrated Audit	0.63	0.48	0.00	1.00	Aobdia (2019); Zhao et al. (2017)

Accelerated Filer	0.63	0.48	0.00	1.00	Newton et al. (2013)
Busy	0.75	0.43	0.00	1.00	Aobdia (2019); Lopez and Peters (2012); Blankley et al. (2012)
Workload Compression	0.67	0.33	0.00	1.00	Lopez and Peters (2012)
Auditor Competition_MSA	0.28	0.14	0.09	1.00	Competition level higher than Newton et al. (2013), maybe because of differences in the sample period.
Auditor Resignation	0.02	0.12	0.00	1.00	Huang and Scholz (2012)
Audit Fees	13.43	1.54	0.00	18.23	Aobdia (2019); Newton et al. (2013); Francis et al. (2013)
Tax Fee	7.17	5.67	0.00	16.86	Paterson and Valencia (2011)
Audit-Related Fee	5.91	5.73	0.00	17.91	Paterson and Valencia (2011)
Other Fees	2.73	4.39	0.00	16.97	Paterson and Valencia (2011)
Non-Audit Fee Ratio	0.13	0.14	0.00	1.00	Lower non-audit fee ratio compared to Newton et al. (2013) maybe because of differences in the sample period.
Influence	0.16	0.24	0.00	1.00	Asthana and Boone (2012); Lopez and Peters (2012)
Abnormal Audit Fee	0.01	0.62	-15.86	2.76	Asthana and Boone (2012); Blankley et al. (2012); Lobo and Zhao (2013)
Audit Report Lag	4.23	0.30	0.00	7.34	Asthana and Boone (2012); Lopez and Peters (2012)
Non-timely Issuance of 10K_Due to Audit	0.01	0.10	0.00	1.00	Lower value compared to Cao et al. (2016) due to our revised definition of the variable
Going Concern	0.11	0.32	0.00	1.00	Minutti-Meza (2013)
Internal Control Weakness	0.03	0.16	0.00	1.00	Aobdia (2019); Newton et al. (2013)
Disc. Accruals	0.00	0.87	-12.62	12.22	Aobdia (2019); Reichelt and Wang (2009); Francis and Yu (2009)
Abs (Disc. Accruals)	0.29	0.83	0.00	12.62	Francis and Yu (2009)
Abs (Accruals)	0.34	1.35	0.00	10.99	Lower compared to Aobdia (2019) maybe because of sample composition
Abs (Accruals/CFO)	1.78	3.72	0.00	24.39	Aobdia (2019)
DD Residual	0.06	0.12	0.00	1.41	Aobdia (2019)
Small Profit	0.07	0.25	0.00	1.00	Lower compared to Aobdia (2019) maybe because of sample composition
Prior ROA Meet	0.01	0.11	0.00	1.00	Lower compared to Aobdia (2019) maybe because of sample composition

*Note: the number of observations for each variable is 26339.

3.4.2 Research Design

We examine the predictive power of ARV by inputting them into multiple machine learning algorithms to predict MAR via a cost-sensitive learning and rolling-window prediction mechanism with hyperparameter tuning. To identify the most predictive ARV, which we term Informative Audit Quality Indicator or IAQI, we perform feature subset selection using six popular machine learning algorithms to select a subset of ARV that can best predict MAR. Then, we check the predictive power of IAQI using the same procedure of cost-sensitive learning and rolling-window prediction with hyperparameter tuning.

Examining the Predictive Power of ARV

Cost-Sensitive Learning

In predicting audit failure, a machine learning algorithm can make two types of mistakes: false-positive errors (i.e., Type 1 errors) and false-negative errors (i.e., Type 2 errors). In this study, “positive” means that observations have known MAR and “negative” otherwise. A false negative error happens when an algorithm mistakenly classifies a positive instance as negative, while a false positive error happens when the algorithm classifies a negative instance as positive. In the context of this research, a false negative error is a more severe mistake than a false positive error because the investigation costs incurred from false positives are usually much lower than the financial costs (e.g., financial losses for investors and litigation costs for the issuer and audit firm) arising from audit failure (Beneish and Vorst 2020). Similar cost imbalances also occur in other domains like fraud detection and loan default prediction (e.g., Perols 2011; Perols et al. 2017; Beneish and Vorst 2020). We define a *misclassification cost* as the cost ratio between false negatives and false positives (Perols 2011). For example, a misclassification cost of 20 indicates that a false negative is 20 times as costly as a false positive. Since the actual misclassification cost of audit quality prediction is unknown, we test different misclassification costs ranging between 1 and 100 (i.e., 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100) (Perols 2011).

To sufficiently consider the cost imbalance issue in our research setting, we adopt a cost-sensitive learning mechanism that can adjust the ratio between the number of positive and negative instances in the training dataset (Elkan 2001). We follow Perols et al. (2017) and adopt the multi-subset Observation Undersampling (OU) method (Chan and Stolfo 1998) to implement cost-sensitive learning. Details of implementing cost-sensitive learning via the multi-subset OU method are provided in Appendix.

Rolling-Window Prediction with Hyperparameter Tuning

To mimic the decision-making process of learning from the past and predicting the future, and to ensure that each prediction model adapts to changing environments through time, we adopt a rolling-window prediction in which each machine learning model is trained with five years of historical data and then the trained models are used to predict outcomes in the following two years (Bao et al. 2020; Brown et al. 2020).[41] We use seven sets of training and testing data in the rolling-window prediction, with the testing years spanning from 2011 to 2017.[42] Our predictions target periods that are two years into the future because the lag between financial report filing and misstatement identification is, on average two years (Lobo and Zhao 2013; Eshleman and Guo 2014; Bao et al. 2020). In this way, we can ensure that most of the past MAR have already been revealed at the time the prediction is made, reducing look-ahead biases (Brown et al. 2020; Jiang, Vasarhelyi, and Zhang 2022).

---

[41] The main findings hold when we also set out a validation sample to tune the hyper-parameters of the model.

[42] These seven sets are: 1) training using 2005 to 2009 data in order to predict the outcome in 2011; 2) training using 2006 to 2010 data in order to predict the outcome in 2012; 3) training using 2007 to 2011 data in order to predict the outcome in 2013; 4) training using 2008 to 2012 data in order to predict the outcome in 2014; 5) training using 2009 to 2013 data in order to predict the outcome in 2015; 6) training using 2010 to 2014 data in order to predict the outcome in 2016; and 7) training using 2011 to 2015 data in order to predict the outcome in 2017.

Hyperparameters are parameters whose values are set manually instead of being "learned" from training data. One example of a hyperparameter is the maximum depth of decision trees used in AdaBoost. Appendix provides the grid of hyperparameters tested for each algorithm. In our ML experiment design, for each misclassification cost and test year combination, we further split the training set into training for tuning and validation for tuning. The latter is used to collect hold-out performance and select the best combination of hyperparameters for an algorithm. Figure 8 illustrates the design if rolling window prediction with hyperparameter tuning.

Figure 8. Illustration of Rolling Window Prediction with Hyperparameter Tuning

	Subset 1		Subset 2		Subset 3		Subset 4		Subset 5		Subset 6		Subset 7	
Year	Tune Hyperparameters	Experiment with Selected Model	Tune Hyperparameters	Experiment with Selected Model	Tune Hyperparameters	Experiment with Selected Model	Tune Hyperparameters	Experiment with Selected Model	Tune Hyperparameters	Experiment with Selected Model	Tune Hyperparameters	Experiment with Selected Model	Tune Hyperparameters	Experiment with Selected Model
2005	Tune_Train	Train												
2006	Tune_Train	Train	Tune_Train	Train										
2007	Tune_Train	Train	Tune_Train	Train	Tune_Train	Train								
2008		Train	Tune_Train	Train	Tune_Train	Train	Tune_Train	Train						
2009	Tune_Vali	Train		Train	Tune_Train	Train	Tune_Train	Train	Tune_Train	Train				
2010			Tune_Vali	Train		Train	Tune_Train	Train	Tune_Train	Train	Tune_Train	Train		
2011		Test			Tune_Vali	Train		Train	Tune_Train	Train	Tune_Train	Train	Tune_Train	Train
2012				Test			Tune_Vali	Train		Train	Tune_Train	Train	Tune_Train	Train
2013						Test			Tune_Vali	Train		Train	Tune_Train	Train
2014								Test			Tune_Vali	Train		Train
2015										Test			Tune_Vali	Train
2016												Test		
2017														Test

Performance Evaluation Metric

Following prior accounting literature (e.g., Dechow et al. 2011; Bao et al. 2020; Brown et al. 2020), we use area under the Receiver Operating Characteristic (ROC) curve, or AUC, to evaluate the overall predictive ability of a machine learning algorithm. ROC curve is a plot of the true positive rate (i.e., the percentage of audit failure accurately classified as audit failure) on the y-axis against the false positive rate (i.e., the percentage of non-audit-failure firms incorrectly classified as audit failure) on the x-axis for different possible classification thresholds (Bradley 1997). AUC ranges from 0 to 1: AUC of 0.5 means random prediction; AUC below 0.5 means the prediction is worse than a random guess; AUC above 0.5 means the prediction is better than a random guess; and AUC of 1 means perfect prediction (Bradley 1997).

Previous studies have also used the estimated relative cost of misclassification (ERC) or expected cost of misclassification (ECM) to evaluate algorithm performance (Perols 2011; Perols et al. 2017). However, ERC and ECM are only informative when the prior probability of positive instances and the costs of misclassification can be reasonably estimated (Perols 2011; Perols et al. 2017). In contrast, AUC does not require such estimates, representing an average comparison of classifiers in domains with class and cost imbalances (Perols 2011). Since there is scant research relating to estimates of the prior probability of an audit's deficiency or the range of misclassification costs for "Big R" predictions, we consider the use of AUC to measure the overall performance of the machine learning algorithms in audit quality prediction an appropriate application. In a sensitivity analysis, an alternative measure is used to evaluate the algorithm performance.

Identifying IAQI Using Feature Subset Selection

After examining the predictive power of using all ARV to predict audit failure, we then examine if a selective subset of ARV can achieve higher predictive power than ARV. In real-world situations, the most predictive features for a target outcome are often unknown a priori (Dash and Liu 1997; Tang, Alelyani, and Liu 2014). To identify those predictive features, we start with a list of candidate features that are identified from domain knowledge (Dash and Liu 1997; Tang, Alelyani, and Liu 2014). In this study, the candidate features are the 31 ARV identified from prior audit quality literature. In many applications, including all candidate features to predict the target outcome does not necessarily generate better performance than including only a selected subset of the candidate features (Dash and Liu 1997; Hocking and Leslie 1967; Perols 2011; Hastie, Tibshirani, and Tibshirani 2017a; Hastie et al. 2017b; Bao et al. 2020; Bertomeu et al. 2020). Some candidate features may be redundant or irrelevant in predicting the target outcome, which causes model overfitting (Dash and Liu 1997; Tang et al. 2014; Hastie et al. 2017a; Hastie et al. 2017b). Removing those irrelevant/redundant features leaves behind a parsimonious model with enhanced algorithm performance and reduced computational complexity (Hall and Smith 1998; Tang et al. 2014; Bao et al. 2020).

Feature subset selection is a technique to select a subset of features that can maximize the performance of a learning algorithm (Dash and Liu 1997; Tang et al. 2014). In this study, we designed a sequential backward feature selection process that identifies the best number and combination of features to be retained. The pseudo-code of this sequential backward feature selection is provided in Appendix. Backward stepwise selection starts with all candidate features, then it iteratively removes the feature that has the least impact on the pre-defined performance metrics and stops when there is no significant improvement to the performance (Hastie et al. 2017b). We use backward stepwise selection in this study because it not only makes feature selection computationally feasible (Hastie et al. 2017a; Hastie et al. 2017b), but also avoids potential omission of predictive candidate features (Guyon and Elisseeff 2003).

Depending on the learning mechanism, the best subset of features will differ across discrete learning algorithms (Perols 2011). To reduce any potential bias arising from the choice of a particular machine learning algorithm, this study refers to Perols (2011) and adopts a “voting mechanism” in which ARV are considered IAQI if they are selected as the best subset of features by most of the machine learning algorithms adopted in this research.

3.5 Results

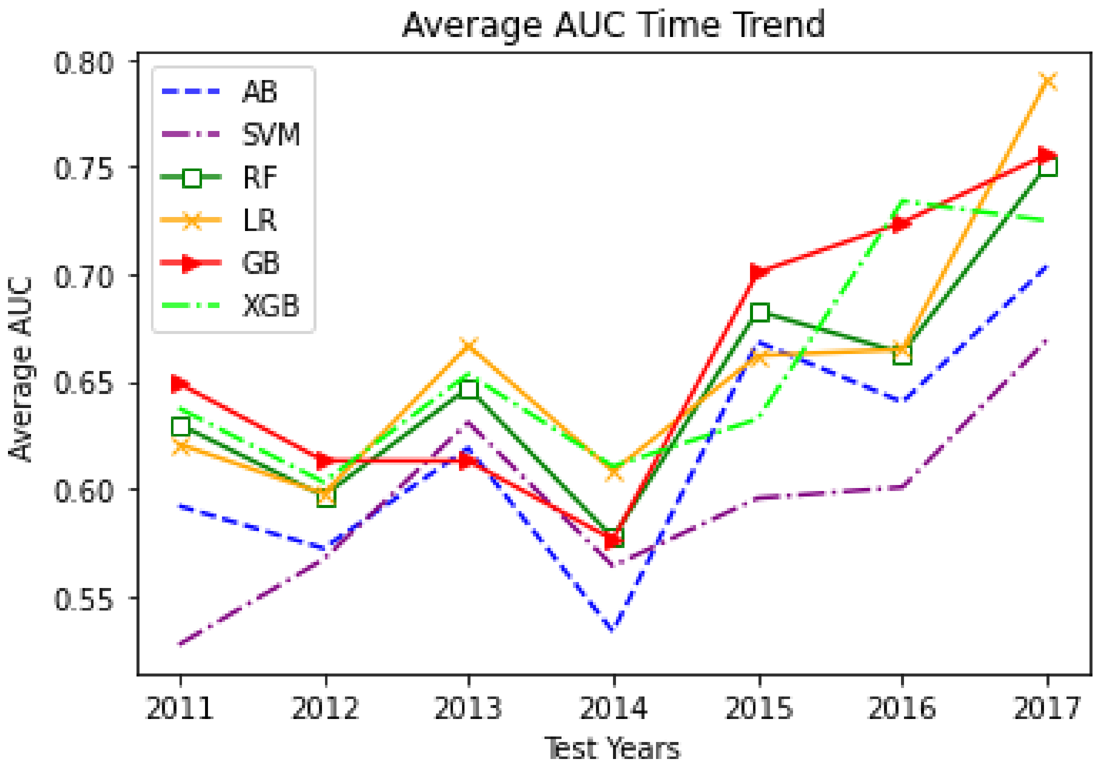
3.5.1 Predictive Power of ARV

The models with tuned hyperparameters for each algorithm under different classification costs for different test years are available in our online appendix.[43] These models are trained on the training set and applied to the hold-out testing set.

Figure 9 provides a time trend plot of the average AUC across misclassification costs for each model. In general, AUC levels are higher in test years 2015 – 2017 than in the period 2011 – 2014. Panel A, Table 10 shows that such differences are statistically significant, potentially because the training set contains data from 2008, a financially distressing year, for the testing period 2011 – 2014. To reduce the potential impact of a year of financial distress on the trained model, and eventually, the testing results, we utilize subsets with test years from 2015 – 2017 (i.e., Subsets 5, 6, 7 shown in Figure 1) from here on. Panel B, Table 10 provides summary statistics of the AUC of the six machine learning algorithms using ARV as predictors in test years 2015 – 2017. Overall, the average predictive power of ARV in terms of AUC across the six machine learning algorithms we examined is 0.687. Among the six algorithms, GB, an ensemble machine learning model, produces the highest average AUC of 0.727 with the lowest variation (standard deviation of 0.026), whereas SVM produces the lowest average AUC of 0.622 with the highest variation (standard deviation of 0.098).

[43] Online Appendix is available via https://github.com/IAQI/IAQI_online_appendix/tree/Tuned-Models.

Figure 9. Time Trend of ARV's Predictive Power



Note:
AB is AdaBoost, GB is Gradient Boosting, LR is Logistic Regression, RF is Random Forest, SVM is Support Vector Machine, and XGB is Extreme Gradient Boosting.

Table 10

Panel A. Comparison of AUC of ARV in Test Years 2011–2014 and 2015-2017

Algorithm	**Mean1 (2011-2014)**	**Mean2 (2015-2017)**	**Diff (Mean2 – Mean1)**	**t stat**
AB	0.579	0.671	0.092***	19.00
GB	0.613	0.727	0.114***	36.87
LR	0.623	0.706	0.082***	15.31
RF	0.613	0.699	0.086***	21.01
SVM	0.572	0.622	0.050***	5.49
XGB	0.626	0.697	0.071***	15.37

Note: *** two-sided p value < 0.01; ** two-sided p value < 0.05; * two-sided p value < 0.10. AB is AdaBoost, GB is Gradient Boosting, LR is Logistic Regression, RF is Random Forest, SVM is Support Vector Machine, and XGB is Extreme Gradient Boosting. ARV include 31 theory-driven and publicly available audit-related variables.

Panel B. Summary Statistics of AUC of ARV in Test Years 2015 - 2017

Algorithm	Mean	Max	Min	Std
AB	0.671	0.776	0.546	0.047
GB	0.727	0.770	0.675	0.026
LR	0.706	0.806	0.654	0.060
RF	0.699	0.771	0.592	0.042
SVM	0.622	0.810	0.428	0.098
XGB	0.697	0.773	0.593	0.051
Average	0.687	0.784	0.581	0.054

Note: AB is AdaBoost, GB is Gradient Boosting, LR is Logistic Regression, RF is Random Forest, SVM is Support Vector Machine, and XGB is Extreme Gradient Boosting. ARV include 31 theory-driven and publicly available audit-related variables

Thus far, our results show that the highest predictive power that ARV can achieve in terms of out-of-sample AUC is 0.727 via the ensemble learning model GB. By academic standards, an AUC of 0.5-0.6 indicates a “random guess”; 0.6-0.7 is “poor” classification; 0.7 to 0.8 is considered “acceptable” performance; 0.8 to 0.9 is “good” prediction; and 0.9 to 1 is “excellent” performance (Mandrekar 2010; Hosmer et al. 2013). Therefore, using the ensemble machine learning model, GB, ARV can predict audit failure out-of-sample with an acceptable performance by academic standards.

To further put ARV’s predictive power in perspective, we compare the AUC that can be achieved by ARV with that of variables commonly used in prior literature to predict misstatement or accounting misconduct. Our first set of benchmark variables (BMK1) are the 35 predictive variables of material misstatement identified by Bertomeu et al. (2020).[44] Our second set of benchmark variables (BMK2) are 35 raw financial variables identified from research that predicts material misstatements, accounting misconduct, or fraud, including Beneish (1997, 1999), Dechow et al. (2011), Cecchini et al. (2011), Perols (2011), Perols et al. (2017) and Bao et al. (2020). We use raw financial

[44] The 35 variables come from Table 7 in Bertomeu et al. (2020). We did not include the variable “% Outsiders appointed” due to lack of access to the database Equilar as indicated in Larcker et al. (2017).

variables as a second group of benchmark variables because Bao et al. (2020) show that they produce higher predictive power than financial ratios when predicting accounting fraud. The details of the two sets of benchmark variables are provided in the Appendix.

Table 11 summarizes the comparison in predictive powers between ARV, benchmark variables (BMK1 and BMK2), and ARV combined with benchmark variables (ARVBMK1 and ARVBMK2). Overall, we observe that ARV have significantly higher predictive power in terms of AUC than BMK1 (BMK2) by an average of 14.1% (8.2%), and that ARV alone outperforms ARVBMK1 (ARVBMK2) by an average of 4.2% (2.1%). We further perform a Tukey's Honest Significance Difference (i.e., Tukey's HSD test; Perols 2011) to compare the predictive powers of the five groups of variables (i.e., ARV, BMK1, BMK2, ARVBMK1, and ARVBMK2) regardless of algorithm types and misclassification costs. Figure 10 visualizes the result of Tukey's HSD test. The horizontal lines in Figure 10 indicate the 95% confidence internals of the predictive powers of each variable group in terms of AUC. We observe that the horizontal line of ARV falls outside of those of other variable groups, suggesting that the predictive power of ARV is significantly greater than that of other variable groups.

Table 11

Panel A. Predictive Power Comparison between ARV and Predictive Variables from Bertomeu et al. (2020)

	BMK1	**ARV**	**ARV-BMK1**			**ARVBMK1**	**ARV- ARVMBK1**		
Algorithm	**Mean**	**Mean**	**Diff**	**t stat**	**Perc.**	**Mean**	**Diff**	**t stat**	**Perc.**
AB	0.595	0.671	0.076***	11.149	12.8%	0.653	0.018***	2.771	2.8%
GB	0.616	0.727	0.111***	17.188	18.0%	0.652	0.075***	14.707	11.5%
LR	0.611	0.706	0.095***	6.275	15.5%	0.706	-0.000	-0.118	0.0%
RF	0.607	0.699	0.092***	9.304	15.2%	0.659	0.040***	5.160	6.1%
SVM	0.561	0.622	0.061***	5.131	10.9%	0.633	-0.011	-0.801	-1.7%
XGB	0.595	0.697	0.102***	6.568	17.1%	0.677	0.020**	2.239	3.0%
Average	0.602	0.687	0.085***	19.809	14.1%	0.659	0.028***	7.699	4.2%

Note:
*** two-sided p value < 0.01; ** two-sided p value < 0.05; * two-sided p value < 0.10. AB is AdaBoost, GB is Gradient Boosting, LR is Logistic Regression, RF is Random Forest, SVM is Support Vector

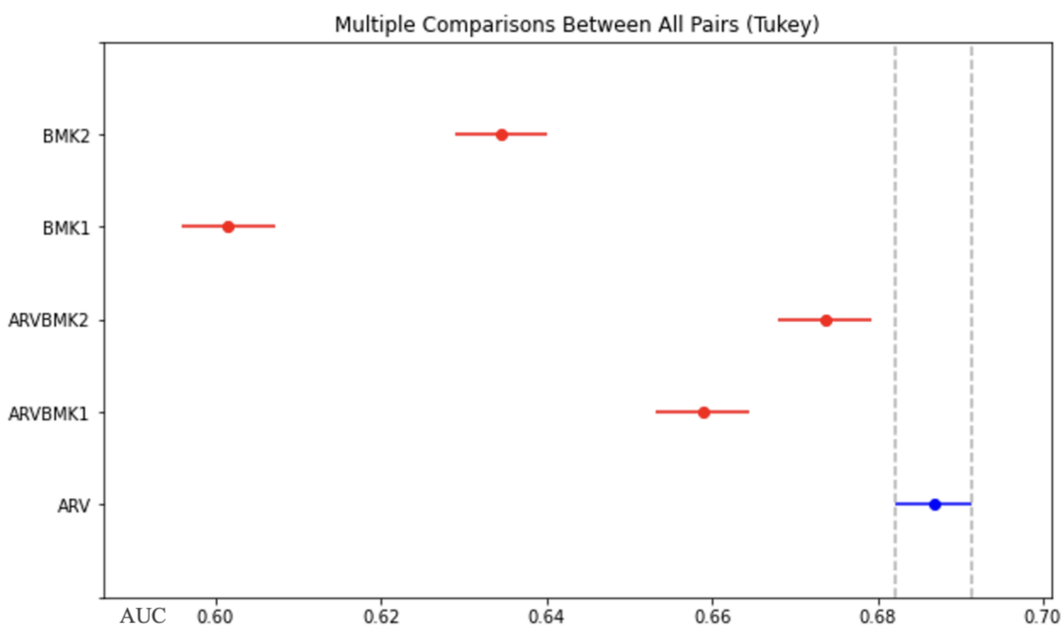
Machine, and XGB is Extreme Gradient Boosting. ARV include 31 theory-driven and publicly available audit-related variables. BMK1 include 35 predictive variables for material misstatement identified from Bertomeu et al. (2020). ARVBMK1 include both the 31 ARV and 35 BMK1.

Panel B. Predictive Power Comparison between ARV and raw financial variables

	BMK2	**ARV**	**ARV-BMK2**			**ARVBMK2**	**ARV- ARVMBK2**		
Algorithm	**Mean**	**Mean**	**Diff**	**t stat**	**Perc.**	**Mean**	**Diff**	**t stat**	**Perc.**
AB	0.664	0.671	0.007	1.216	1.1%	0.681	-0.010*	-1.956	-1.5%
GB	0.664	0.727	0.062***	16.592	9.4%	0.696	0.031***	5.025	4.5%
LR	0.545	0.706	0.161***	18.852	29.5%	0.644	0.061***	1.810	9.5%
RF	0.643	0.699	0.056***	10.639	8.7%	0.673	0.026***	9.946	3.8%
SVM	0.506	0.622	0.116***	11.250	22.8%	0.595	0.027***	5.511	4.6%
XGB	0.650	0.697	0.047***	6.056	7.2%	0.696	0.001	0.183	0.2%
Average	0.635	0.687	0.052***	14.387	8.2%	0.673	0.014***	4.065	2.1%

Note:
*** two-sided p value < 0.01; ** two-sided p value < 0.05; * two-sided p value < 0.10. AB is AdaBoost, GB is Gradient Boosting, LR is Logistic Regression, RF is Random Forest, SVM is Support Vector Machine, and XGB is Extreme Gradient Boosting. ARV include 31 theory-driven and publicly available audit-related variables. BMK2 include 35 raw financial variables from Dechow et al. (2011), Cecchini et al. (2011), Perols (2011), Perols et al. (2017), and Bao et al. (2020). ARVBMK2 include both the 31 ARV and 35 BMK2.

Figure 10. Comparison of Predictive Power of ARV and Benchmark Variables Based on Tukey HSD Test



Note:
ARV include 31 theory-driven and publicly available audit-related variables. BMK1 include 35 predictive variables for material misstatement identified from Bertomeu et al. (2020). ARVBMK1 include both the 31 ARV and 35 BMK1. ARVMBK1 include both ARV and MBK1. BMK2 include 33 raw financial variables from Dechow et al. (2011), Cecchini et al. (2011), Perols (2011), Perols et al. (2017), and Bao et al. (2020). ARVBMK2 include both the 31 ARV and 35 BMK2. This Tukey HSD test is performed on results collected from all tested algorithms: AdaBoost, Gradient Boosting, Logistic Regression, Random Forest,

Support Vector Machine, and Extreme Gradient Boosting. The dotted lines indicate the 95% confidence internals. Horizontal lines that fall outside (inside) of the dotted lines indicate that their values are (not) significantly different from those represented by the horizontal line enclosed in the dotted line.

Overall, we find that ARV have acceptable out-of-sample predictive power and that they outperform benchmark variables in predicting audit failure.

3.5.2 Informative Audit Quality Indicators

Table 12 displays the results of feature subset selection for each machine learning algorithm. For each row in Table 12, the value "1" indicates that an algorithm selects this variable from the sequential backward feature selection process described in the previous section and Appendix. ARV that are deemed as significant features by the majority (i.e., more than 50%) of the tested algorithms include auditor tenure, auditor resignation, auditors' report of clients' internal control weakness, absolute value of total accruals, absolute value to total accruals scaled by cash flow from operation, audit report lag, audit office size, integrated audit, discretionary accruals, auditor city-level industry specialization, auditor competition, audit fees, and client influence. We consider these 13 most predictive ARV as IAQI. In contrast, the least predictive ARV (less than 50% vote) are auditor industry specialization at the national firm level, non-timely issuance of 10K due to audit issues, prior ROA meet, small profit, abnormal audit fee, non-audit fee ratio, auditor workload compression, and audit-related fees.

Table 12. Feature Subset Selection Results

Variables	AB	SVM	RF	LR	GB	XGB	Total Vote	Percentage
Tenure	1	1	1	1	1	1	6	100.00%
Auditor Resignation	1	0	1	1	1	1	5	83.33%
Internal Control Weakness	1	0	1	1	1	1	5	83.33%
Abs (Accruals)	1	1	0	1	1	1	5	83.33%
Abs (Accruals/CFO)	1	0	1	1	1	1	5	83.33%
Audit Report Lag	1	0	1	1	1	1	5	83.33%
Office Size	0	1	1	1	0	1	4	66.67%
Integrated Audit	1	0	1	1	1	0	4	66.67%
Disc. Accruals	1	0	1	0	1	1	4	66.67%
Industry Specialization_MSA	0	1	1	0	1	1	4	66.67%

Auditor Competition_MSA	0	1	0	1	1	1	4	66.67%
Audit Fees	1	0	1	1	1	0	4	66.67%
Influence	1	0	1	0	1	1	4	66.67%
Accelerated Filer	0	1	0	0	1	1	3	50.00%
Busy	0	0	1	0	1	1	3	50.00%
Going Concern	0	0	0	1	1	1	3	50.00%
Abs (Disc. Accruals)	1	0	0	1	0	1	3	50.00%
DD Residual	0	0	1	0	1	1	3	50.00%
Big 4	0	0	1	0	1	1	3	50.00%
New Client	0	0	1	1	1	0	3	50.00%
Local Auditor_MSA	0	0	1	0	1	1	3	50.00%
Tax Fee	1	0	0	0	1	1	3	50.00%
Other Fees	0	0	1	0	1	1	3	50.00%
Audit-Related Fee	0	0	0	0	1	1	2	33.33%
Workload Compression	1	0	0	0	0	1	2	33.33%
Non-Audit Fee Ratio	0	0	1	1	0	0	2	33.33%
Abnormal Audit Fee	0	0	1	0	0	1	2	33.33%
Small Profit	0	0	0	0	1	1	2	33.33%
Prior ROA Meet	0	0	1	0	1	0	2	33.33%
Non-timely Issuance of 10K_Due to Audit	0	0	0	0	1	0	1	16.67%
Industry Specialization_National	0	0	1	0	0	0	1	16.67%

Note:
AB is AdaBoost, GB is Gradient Boosting, LR is Logistic Regression, RF is Random Forest, SVM is Support Vector Machine, and XGB is Extreme Gradient Boosting. “1” indicates that an algorithm chooses this variable as the most predictive variable from the feature selection process.

The predictive modeling process serves as validation of the predictive power of the audit-related variables examined in previous studies. Among the 31 theory-driven and publicly accessible audit-related variables, IAQI are the most relevant in predicting audit failure. However, variables that are not selected as IAQI are not irrelevant to understanding audit quality. Instead, the implication is that when the objective is to *predict* MAR, IAQI are more predictive than other variables. Based on our categorization of ARV in Table 6, we present the 13 IAQI together with their categories, sub-categories, and aspects of audit captured in Table 5. The 13 IAQI are composed of variables from the entire cycle of an audit engagement: audit input, audit process, and audit output.

We also examine the predictive power of IAQI by following the same process applied to ARV. The tuned models for each algorithm under different classification costs for different test years are available in the online appendix. [45] Table 13 compares the predictive powers between ARV and IAQI in terms of the average AUC for each algorithm. Furthermore, we observe that when IAQI are used as predictors, AB generates the highest AUC of 0.749, followed by XGB (AUC .735), and GB (AUC 0.734), all of which are ensemble machine learning models. Similar to what we observe in ARV's predictive power, IAQI achieves the lowest predictive power under SVM (0.628). Overall, the average predictive power of IAQI in terms of AUC across the six examined algorithms is 0.710, greater than that of ARV (0.687) by about 3.3%. Based on the results previously reported in Table 7, IAQI outperform the benchmark financial variables BMK1 (average AUC 0.602) by 17.9% and MBK2 (average AUC 0.635) by 11.8%.

Table 13. Predictive Power Comparison Between IAQI and ARV

	IAQI	**ARV**	**IAQI-ARV**		
Algorithm	**Mean**	**Mean**	**Diff**	**t stat**	**Perc.**
AB	0.749	0.671	0.0780***	13.054	11.6%
GB	0.734	0.727	0.008**	2.306	1.0%
LR	0.692	0.706	-0.014**	-2.230	-2.0%
RF	0.720	0.699	0.021***	4.875	3.0%
SVM	0.628	0.622	0.006	0.667	1.0%
XGB	0.735	0.697	0.038***	5.710	5.5%
Average	0.710	0.687	0.023***	7.260	3.3%

Note:
*** two-sided p value < 0.01; ** two-sided p value < 0.05; * two-sided p value < 0.10. AB is AdaBoost, GB is Gradient Boosting, LR is Logistic Regression, RF is Random Forest, SVM is Support Vector Machine, and XGB is Extreme Gradient Boosting. ARV include 31 theory-driven and publicly available audit-related variables. IAQI is a subset of ARV that include 13 predictive audit-related variables.

In summation, we have identified IAQI, a selective subset of ARV that are deemed the most predictive by most of the machine learning algorithms examined, and we show that IAQI can outperform ARV and benchmark financial variables. Nevertheless, even though IAQI outperforms ARV and benchmark variables, the highest

[45] Online Appendix is available via https://github.com/IAQI/IAQI_online_appendix/tree/Tuned-Models.

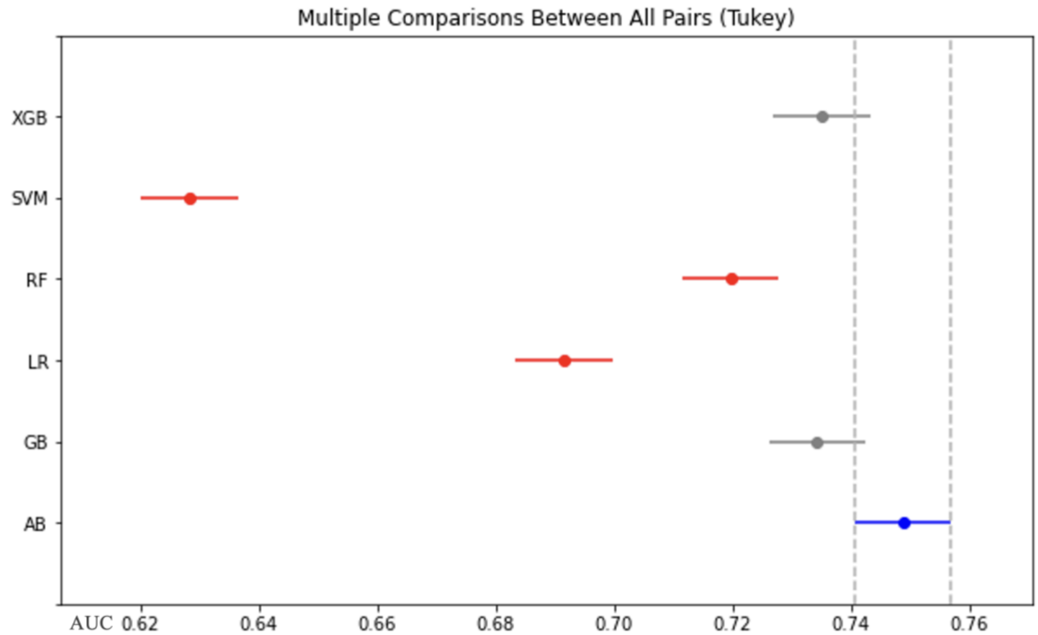
average AUC that can be achieved by AB models (average AUC of 0.749) is still considered acceptable but not excellent by academic standards.

3.6 Predictive Score of Audit Failure

Having identified IAQI, a set of literature-based, predictive, audit-related variables of audit failure that can predict earnings management and misstatement, we establish IAQI-based predictive scores (P Scores) of audit failure. Like the F score in Dechow et al. (2011), the P Score is a scaled probability of audit failure for each firm-year. However, unlike the F score, which is calculated based on probability from linear regression, P Score is the standardized probability output from advanced machine learning models that can model non-linear relationships.

To decide which machine learning algorithm to use in developing P Score, we apply Tukey’s HSD test to examine the relative performance of different machine learning models using IAQI as predictors. Figure 11 presents the Tukey HSD test result. As seen in Table 13, AB generates the highest predictive power in terms of AUC when IAQI are used as inputs. However, since the 95% confidence internals of AB overlap with those of GB and XGB, these three algorithms have statistically equivalent performance. In contrast, the 95% confidence intervals of SVM, LR, and RF fall outside of AB’s confidence interval, suggesting that they have significantly inferior performance compared to AB.

Figure 11. IAQI Predictive Power Comparison by Algorithms



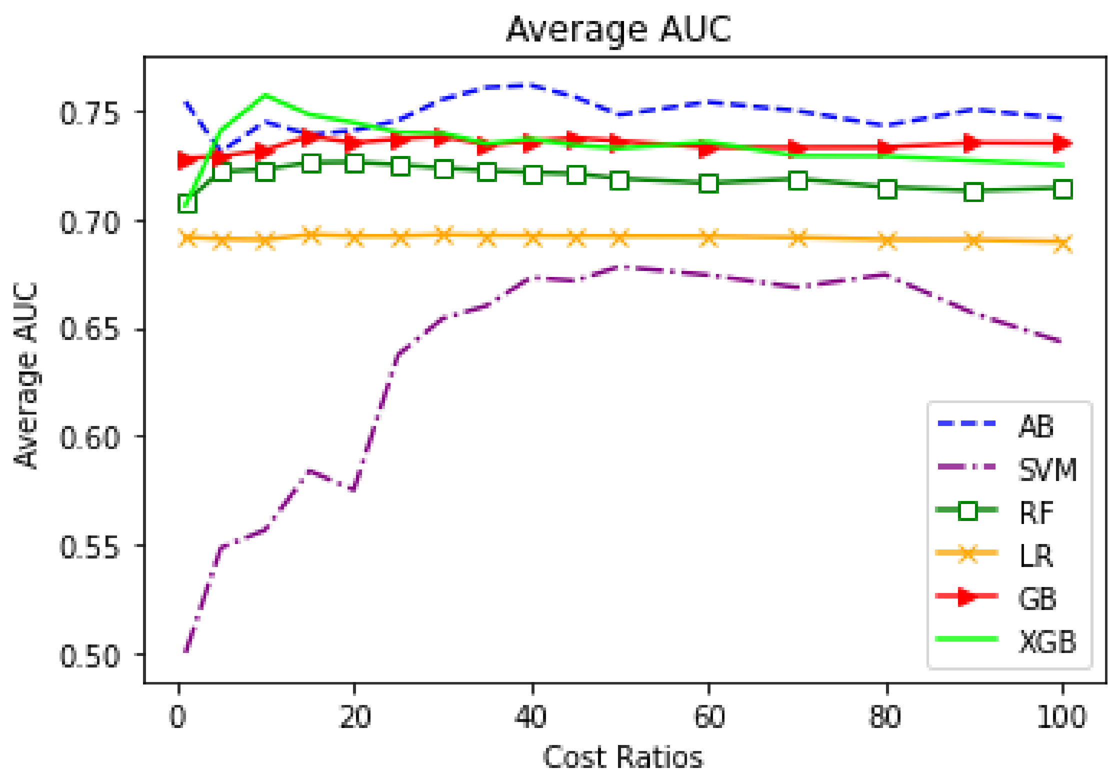
Note:
This Tukey HSD test is performed on results collected from all tested algorithms when IAQI are used as predictors. AB is AdaBoost, GB is Gradient Boosting, LR is Logistic Regression, RF is Random Forest, SVM is Support Vector Machine, and XGB is Extreme Gradient Boosting. The dotted lines indicate the 95% confidence internals. Horizontal lines that fall outside (inside) of the dotted lines indicate that the values they represent are (not) significantly different from the model represented by the horizontal line enclosed in the dotted line.

Since AB, GB, and XGB have equivalent performance using IAQI, we generate three alternative P Scores (i.e., PSCORE_AB, PSCORE_GB, and PSCORE_XGB). To generate P Scores based on each algorithm, we obtain the probability predictions from the tuned models of each algorithm for testing years 2015 – 2017 using IAQI.[46] To decide which level of misclassification cost to set in this process, we plot IAQI's predictive power in terms of AUC across testing years 2015-2017 by misclassification costs for each algorithm in Figure 12. The performance of AB, GB, and XGB are not sensitive to misclassification costs, so we set misclassification cost at 40, as it is the middle point of

[46] Available at: https://github.com/IAQI/IAQI_online_appendix/tree/Tuned-Models.

the range of misclassification costs that we tested (i.e., 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100). In un-tabulated analysis, P Scores generated from alternative misclassification costs show similar results. To obtain each P Score, we standardize the probability prediction output from the tuned models of each algorithm.[47]

Figure 12. IAQI's Predictive Power by Misclassification Costs Averaging from 2015 to 2017



Note:
AB is AdaBoost, GB is Gradient Boosting, LR is Logistic Regression, RF is Random Forest, SVM is Support Vector Machine, and XGB is Extreme Gradient Boosting.

Table 14 provides summary statistics for the P Scores (PSCORE_AB, PSCORE_GB, and PSCORE_XGB). For comparison, we also included two commonly

---

[47] Standardization transforms the original distribution of the predicted probability into one with a mean of 0 and a standard deviation of 1. For example, if an observation in the 2016 test set has a final probability prediction of 0.45, the mean and standard deviation of the final probability prediction for all observations in the 2016 test set is 0.32 and 0.12 respectively, then the Predictive MAR Score for this observation is 1.08 ((0.45-0.32)/0.12). We adopt standardization as a rescaling method because it is relatively easy to interpret scores in a standardized distribution and to identify extreme values of the scores.

used academic measures for detecting accounting misstatements and earnings manipulation: F score (FSCORE) from Dechow et al. (2011) and M score (MSCORE) from Beneish (1997, 1999). We calculate the F and M scores following Price et al. (2011). Table 14 shows that observations in the hold-out test sets spanning from 2015 - 2017 that have actual MAR feature P Scores that are significantly higher than those without actual MAR. However, there is no such difference for F Score or M Score.

Table 14. Univariate Analysis of P Score, F Score, and M Score

Score	Actual MAR		Actual Non-MAR		Difference	t stat
	Mean	Count	Mean	Count		
PSCORE_AB	0.579	55	-0.005	6188	0.584***	5.474
PSCORE_GB	0.760	55	-0.007	6188	0.767***	6.148
PSCORE_XGB	0.743	55	-0.007	6188	0.750***	7.179
FSCORE	-6.411	50	-6.021	5209	-0.390	-0.503
MSCORE	-1.837	46	-2.452	4019	0.615	0.587

Note:
*** two-sided p value < 0.01; ** two-sided p value < 0.05; * two-sided p value < 0.10. PSCORE is the P Score developed in "Section VI. Predictive Audit Failure Score." F Score is developed by Dechow et al. (2011). M Score is developed by Beneish (1999).

3.6.1 P Score's Incremental Association with Audit Failure

To compare P Score, F Score, and M Score's incremental association with MAR, we establish the following model, Model (1):

$$MAR_{it} = \beta_0 + \beta_1 * SCORE_{it} + \boldsymbol{Controls_{it}} * \boldsymbol{\beta} + \varepsilon \qquad (1)$$

MAR_{it} equals one if the financial statement for firm i in fiscal year t has a material restatement due to GAAP violations or fraud.[48] $SCORE_{it}$ takes the values of P Score (PSCORE_AB, PSCORE_GB, and PSCORE_XGB), FSCORE, and MSCORE for firm i in fiscal year t. Meanwhile, we obtain control variables from related literature (See the footnotes of Table 11). If $SCORE_{it}$ can provide incremental association with an actual

[48] Consistent with the machine learning experiment, only starting-year MAR is used.

MAR, we expect parameter β_1 to be significantly positive. We also control for year and industry fixed effects and calculate the standard errors using firm clusters. All tests are estimated using logistic regressions. The results are documented in Panel A, Table 15. Overall, we observe that all else being held equal, PSCORE_AB, PSCORE_GB, and PSCORE_XGB are significantly and positively associated with actual MAR, while FSCORE and MSCORE are not. *Ceteris paribus*, a 1-point increase in the PSCORE_AB, PSCORE_GB, and PSCORE_XGB increases the odds of having an actual MAR by 2.01, 1.61, and 1.83 times, respectively.[49]

We also follow Price et al. (2011) and perform head-to-head comparisons of each of the three P Scores with FSCORE or MSCORE, generating six comparison groups (i.e., 3 by 2). We adopted Model (1) for each comparison group and made sure that the sample has values for both the scores to be compared. We use seemingly unrelated estimation (SUEST) to test coefficients across models and determine which score has a higher coefficient. The SUEST test results are presented in Panel B, Table 15. SUEST test results show that all P Scores (PSCORE_AB, PSCORE_GB, and PSCORE_XGB) are significantly greater than either FSCORE or MSCORE, suggesting that P Scores have greater detection ability of audit failure than FSCORE or MSCORE. Panel B, Table 15, also shows that Pseudo R squared is the highest in the model with PSCORE_AB and the lowest in the model with MSCORE. The PSCORE_AB model also has the highest in-sample AUC of 0.864, while MSCORE has the lowest in-sample AUC of 0.826.[50] The

[49] The coefficient for PSCORE_AB is 0.699, so the odds ratio is 2.01 (e^0.699); the coefficient for PSCORE_GB is 0.478, so the odds ratio is 1.61 (e^0.478); the coefficient for PSCORE_XGB is 0.607, so the odds ratio is 1.83 (e^0.607).

[50] Different from the out-of-sample AUC that we provide in the rest of the paper that is calculated from hold-out sample unseen by the model, the in-sample AUC is measured using the sample that constructs the model. Thus, in-sample AUC is generally higher than out-of-sample AUC.

comparative advantages of P Score might come from the ability of complex machine learning algorithms to capture non-linear relationships in data.

Table 15

Panel A. Comparison of P Score, F Score, and M Score

	(1)	(2)	(3)	(4)	(5)
	MAR	MAR	MAR	MAR	MAR
	Coef. (z score)	Coef. (z score)	Coef. (z score)	Coef. (z score)	Coef. (z score)
PSCORE_AB	0.699***				
	(2.96)				
PSCORE_GB		0.478**			
		(2.27)			
PSCORE_XGB			0.607***		
			(2.79)		
FSCORE				0.001	
				(0.01)	
MSCORE					-0.004
					(-0.60)
CONTROLS	Yes	Yes	Yes	Yes	Yes
Industry Fixed Effects	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes
Pseudo R2	0.224	0.215	0.219	0.201	0.181
In-Sample AUC	0.864	0.859	0.864	0.844	0.826
Observations	3124	3124	3124	2907	2388

Panel B. Head-to-Head Comparison

	PSCORE_AB	**FSCORE**	**PSCORE_AB**	**MSCORE**
Coef.	0.6948***	0.0006	0.5570***	-0.0043
Z score	3.01	0.01	2.93	-0.60
CONTROLS	Yes	Yes	Yes	Yes
Industry and Year Fixed Effects	Yes	Yes	Yes	Yes
Significant Difference from SUEST	Two-sided P value = 0.0048		Two-sided P value = 0.0033	
Pseudo R2	0.224	0.201	0.198	0.181
In-Sample AUC	0.864	0.844	0.843	0.826
N	2907		2388	
	PSCORE_GB	**FSCORE**	**PSCORE_GB**	**MSCORE**
Coef.	0.4912***	0.0006	0.4494**	-0.0043
Z score	2.34	0.01	2.08	-0.60
CONTROLS	Yes	Yes	Yes	Yes
Industry and Year Fixed Effects	Yes	Yes	Yes	Yes
Significant Difference from SUEST	Two-sided P value = 0.0310		Two-sided P value = 0.0359	
Pseudo R2	0.216	0.201	0.194	0.181

In-Sample AUC	0.859	0.844	0.847	0.826
N	2907		2388	
	PSCORE_XGB	**FSCORE**	**PSCORE_XGB**	**MSCORE**
Coef.	0.6395***	0.0006	0.6518***	-0.0043
Z score	2.90	0.01	2.85	-0.60
CONTROLS	Yes	Yes	Yes	Yes
Industry and Year Fixed Effects	Yes	Yes	Yes	Yes
Significant Difference from SUEST	Two-sided P value = 0.0060		Two-sided P value = 0.0042	
Pseudo R2	0.220	0.201	0.202	0.181
In-Sample AUC	0.864	0.844	0.854	0.826
N	2907		2388	

Note: ***, **, and * indicate significance level at 1%, 5%, and 10% respectively (two-sided p-value). All standard errors are estimated by clustering firms. MAR is material annual restatement due to GAAP violations or financial fraud. Predictive Audit Failure Score. F Score is developed by Dechow et al. (2011). M Score is developed by Beneish (1999). Big 4 is 1 if the auditor is one of the Big 4 audit firms. 0 otherwise. Control variables include the following variables. Delta_Rec = change in accounts receivable scaled by total assets: $RECT_t/AT_t$ - $RECT_{t-1}/AT_{t-1}$ (Lobo and Zhao 2013). Delta_INV = change in inventory scaled by total assets: $INVT_t/AT_t$ - $INVT_{t-1}/AT_{t-1}$ (Lobo and Zhao 2013) Soft asset = soft assets as a percentage of total assets: $(AT_t - PPENT_t - CHE_t)/AT_t$ (Lobo and Zhao 2013). LNASSET = log of total assets: log (AT_{t-1}) (Eshleman and Guo 2014). ATRUN = total sales divided by lagged total assets: $REVT_t/AT_{t-1}$ (Eshleman and Guo 2014). ROA = income before extraordinary items divided by average total assets: $IB_t/((AT_t + AT_{t-1})/2)$ (Eshleman and Guo 2014). Leverage = financial leverage, defined as long-term debt plus debt in current liabilities, all scaled by total assets: $(DLTT_t+DLC_t)/AT_t$ (Eshleman and Guo 2014). CURR = the current ratio, defined as current assets divided by current liabilities: ACT_t/LCT_t (Eshleman and Guo 2014). MKBK = the market-to-book ratio, defined as market value at fiscal year-end scaled by book equity: $(PRCC\_F_t * CSHO_t)/CEQ_t$ (Eshleman and Guo 2014). EPS = the earnings-to-price ratio, defined as income before extraordinary items, scaled by market value at fiscal year-end: $IB_t/(PRCC\_F_t * CSHO_t))$ (Eshleman and Guo 2014). EPSGrowth = the growth rate of EPS: $(EPS_t – EPS_{t-1})/EPS_{t-1}$ (Eshleman and Guo 2014). TotoalAccruals = change in noncash assets (noncash total assets minus total liabilities and preferred stocks) from year t— 1 to year t scaled by average total assets: $\{((AT_t – CHE_t) - (LT_t + PSTK_t)) - ((AT_{t-1} – CHE_{t-1}) - (LT_{t-1} + PSTK_{t-1}))\}/((AT_t + AT_{t-1})/2)$ (Lobo and Zhao 2013). SalesGrowth = change in sales from the prior year to the current year: $(SALE_t – SALE_{t-1})/SALE_{t-1}$ (Lobo and Zhao 2013). MA = 1 if involved in merger activity (Stanley and DeZoort 2007). Restructure = 1 if the firm has restructuring changes during the year (Newton et al. 2013). FirmAge= The natural logarithm of the number of years the firm has been listed on COMPUSTAT (Eshleman and Guo 2014): ln(fiscal year-IPODATE). GoingConcern= 1 if the company received a going concern modified opinion in year t, zero otherwise (Ettredge Emeigh Fuerherm Li 2014). AuditorChange= 1 if an audit engagement occurs within the first year of an auditor change, and 0 otherwise (Francis et al. 2013). This is equivalent to the variable NewClient defined in Table1. Influence = is the ratio of a specific client's total fees (audit fees plus nonaudit fees) relative to annual fees of SEC registrants generated by the practice office in a given year (Francis et al. 2013), Eshleman and Guo 2014). FREEC = Demand for external financing, defined as operating cash flows (OANCF) less capital expenditures (CAPX), all scaled by lagged assets (Eshleman and Guo 2014). Abnormal Audit Fee = The unscaled residual from the audit fee model used in Blankley et al. (2012). Busy = 1 if a company has a fiscal year-end date of December, and 0 otherwise (Lopez and Peters 2012). Auditor Competition_MSA = MSA-level auditor concentration based on Herfindahl index. Details are provided in (Netown et al. 2013). Industry Specialization_MSA = auditor's annual market share of audit fees within a two-digit SIC category for a particular city. A city is defined as a Metropolitan Statistical Area (MSA) (Reichelt and Wang 2010). Non-Audit Fee Ratio = Non-audit fees deflated by total fees paid (audit plus non-audit fees). Non-audit fee equals to the sum of benefit fee, IT fee, Tax fee, audit related fee, and other fees. (Ruddock et al. 2006) Local Auditor_MSA = 1 if the audit engagement office is located in the same MSA where audit clients are headquartered, 0 otherwise (Choi, Kim, Qiu, and Zhang 2012). We did not include qualified opinion from Eshleman and Guo (2014) because observations with qualified opinion were removed due to missing other variables. We did not include Leases from Lobo and Zhao (2013) because the remaining observations all have future operating lease obligations that are greater than 0 (MRCT>0).

3.6.2 Can P Scores Reflect Audit Quality?

To assess whether P Scores can reflect audit quality, we establish the following model based on a commonly recognized conclusion that Big 4 auditors provide higher audit quality (Eshleman and Guo 2014; DeFond, Erkens, and Zhang 2017; Jiang, Wang, I., and Wang, K. 2019).[51] If P Scores can indicate audit quality, the parameter γ_1 should be significantly negative in Model (2) below.

$$SCORE_{it} = \gamma_0 + \gamma_1 * Big4_{it} + \boldsymbol{Controls_{it}} * \boldsymbol{\gamma} + \varepsilon \qquad (2)$$

The value of $Big4_{it}$ is one if firm i in fiscal year t was audited by a Big 4 auditor. Similar control variables are used in this model (See the footnotes of Table 16). We use OLS (Ordinary Least Squares) to perform this test while also controlling for year and industry fixed effects and calculating the standard errors using firm clusters. The results in Table 16 show that holding other factors constant, financial reports audited by Big 4 auditors have significantly lower PSCORE_AB, PSCORE_GB, and PSCORE_XGB compared to other engagements, confirming our expectations. In contrast, we do not observe such significant association between Big4 and FSCORE or MSCORE.

Table 16. Relationship Between Big4 and P Score, F Score, and M Score

	(1)	(2)	(3)	(4)	(5)
	PSCORE_AB	PSCORE_GB	PSCORE_XGB	FSCORE	MSCORE
	Coef. (t stats)	Coef. (t stats)	Coef. (t stats)	Coef. (t stats)	Coef. (t stats)
Big4	-0.329***	-0.524***	-0.602***	0.217	0.457
	(-6.28)	(-9.15)	(-10.26)	(1.13)	(0.89)
CONTROLS	Yes	Yes	Yes	Yes	Yes
Industry Fixed Effects	Yes	Yes	Yes	Yes	Yes

[51] Although Big 4 auditor status is not chosen as a *predictive* factor of MAR, it is an important *explanatory* factor of audit quality. Additional details on these differences can be found in Shmueli (2010) and Shmueli and Koppius (2011).

Year Fixed Effects	Yes	Yes	Yes	Yes	Yes
Adj. R2	0.287	0.308	0.334	0.081	0.053
Observations	3982	3982	3982	3710	3126

Note: ***, **, and * indicate significance level at 1%, 5%, and 10% respectively (two-sided p-value). All standard errors are estimated by clustering firms. MAR is material annual restatement due to GAAP violations or financial fraud. Predictive Audit Failure Score. F Score is developed by Dechow et al. (2011). M Score is developed by Beneish (1999). Big 4 is 1 if the auditor is one of the Big 4 audit firms. 0 otherwise. Control variables include the following variables. Delta_Rec = change in accounts receivable scaled by total assets: $RECT_t/AT_t$ - $RECT_{t-1}/AT_{t-1}$ (Lobo and Zhao 2013). Delta_INV = change in inventory scaled by total assets: $INVT_t/AT_t$ - $INVT_{t-1}/AT_{t-1}$ (Lobo and Zhao 2013) Soft asset = soft assets as a percentage of total assets: $(AT_t - PPENT_t - CHE_t)/AT_t$ (Lobo and Zhao 2013). LNASSET = log of total assets: log (AT_{t-1}) (Eshleman and Guo 2014). ATRUN = total sales divided by lagged total assets: $REVT_t/AT_{t-1}$ (Eshleman and Guo 2014). ROA = income before extraordinary items divided by average total assets: $IB_t/((AT_t + AT_{t-1})/2)$ (Eshleman and Guo 2014). Leverage = financial leverage, defined as long-term debt plus debt in current liabilities, all scaled by total assets: $(DLTT_t+DLC_t)/AT_t$ (Eshleman and Guo 2014). CURR = the current ratio, defined as current assets divided by current liabilities: ACT_t/LCT_t (Eshleman and Guo 2014). MKBK = the market-to-book ratio, defined as market value at fiscal year-end scaled by book equity: $(PRCC\_F_t * CSHO_t)/CEQ_t$ (Eshleman and Guo 2014). EPS = the earnings-to-price ratio, defined as income before extraordinary items, scaled by market value at fiscal year-end: $IB_t/(PRCC\_F_t * CSHO_t))$ (Eshleman and Guo 2014). EPSGrowth = the growth rate of EPS: $(EPS_t – EPS_{t-1})/EPS_{t-1}$ (Eshleman and Guo 2014). TotoalAccruals = change in noncash assets (noncash total assets minus total liabilities and preferred stocks) from year t— 1 to year t scaled by average total assets: $\{((AT_t – CHE_t) - (LT_t + PSTK_t)) - ((AT_{t-1} – CHE_{t-1}) - (LT_{t-1} + PSTK_{t-1}))\}/((AT_t + AT_{t-1})/2)$ (Lobo and Zhao 2013). SalesGrowth = change in sales from the prior year to the current year: $(SALE_t – SALE_{t-1})/SALE_{t-1}$ (Lobo and Zhao 2013). MA = 1 if involved in merger activity (Stanley and DeZoort 2007). Restructure = 1 if the firm has restructuring changes during the year (Newton et al. 2013). FirmAge= The natural logarithm of the number of years the firm has been listed on COMPUSTAT (Eshleman and Guo 2014): ln(fiscal year-IPODATE). GoingConcern= 1 if the company received a going concern modified opinion in year t, zero otherwise (Ettredge Emeigh Fuerherm Li 2014). AuditorChange= 1 if an audit engagement occurs within the first year of an auditor change, and 0 otherwise (Francis et al. 2013). This is equivalent to the variable NewClient defined in Table1. Influence = is the ratio of a specific client's total fees (audit fees plus nonaudit fees) relative to annual fees of SEC registrants generated by the practice office in a given year (Francis et al. 2013), Eshleman and Guo 2014). FREEC = Demand for external financing, defined as operating cash flows (OANCF) less capital expenditures (CAPX), all scaled by lagged assets (Eshleman and Guo 2014). Abnormal Audit Fee = The unscaled residual from the audit fee model used in Blankley et al. (2012). Busy = 1 if a company has a fiscal year-end date of December, and 0 otherwise (Lopez and Peters 2012). Auditor Competition_MSA = MSA-level auditor concentration based on Herfindahl index. Details are provided in (Netown et al. 2013). Industry Specialization_MSA = auditor's annual market share of audit fees within a two-digit SIC category for a particular city. A city is defined as a Metropolitan Statistical Area (MSA) (Reichelt and Wang 2010). Non-Audit Fee Ratio = Non-audit fees deflated by total fees paid (audit plus non-audit fees). Non-audit fee equals to the sum of benefit fee, IT fee, Tax fee, audit related fee, and other fees. (Ruddock et al. 2006) Local Auditor_MSA = 1 if the audit engagement office is located in the same MSA where audit clients are headquartered, 0 otherwise (Choi, Kim, Qiu, and Zhang 2012). We did not include qualified opinion from Eshleman and Guo (2014) because observations with qualified opinion were removed due to missing other variables. We did not include Leases from Lobo and Zhao (2013) because the remaining observations all have future operating lease obligations that are greater than 0 (MRCT>0).

3.7 Using Raw Audit Variables

Bao et al. (2020) show that raw financial variables outperform financial ratios in predicting fraud. In the same vein, we explore whether using raw audit variables can outperform highly engineered ARV in predicting audit failure. To do so, we expand the selection of audit variables to variables that are readily accessible in their raw or close-to-raw forms from the Audit Analytics database. In total, we gathered 91 such variables, including 74 raw variables and 17 close-to-raw variables that require minimal calculation from Audit Analytics. For ease of reference, we term these 91 raw or close-to-raw audit related variables RawAudit. Among RawAudit, 67 are binary variables, 23 are continuous variables, and 3 are categorical variables. The Appendix lists the 91 RawAudit variables.

Using RawAudit, we adopted the same cost-sensitive learning and rolling window calculation with hyperparameter tuning design.[52] Table 17 presents a comparison of RawAudit's predictive power with that of IAQI and ARV. When RawAudit are used as predictors, GB models achieve the highest average AUC of 0.781. On average, RawAudit carry significantly higher AUC predictive powers than ARV (roughly 6%) and IAQI (roughly 3%), consistent with Bao et al. (2020)'s conclusion that raw financial variables can significantly outperform financial ratios.

Table 17. Predictive Power Comparison Between RawAudit, IAQI, and ARV

	RawAudit	**IAQI**	**RawAudit - IAQI**			**ARV**	**RawAudit - ARV**		
Algorithm	**Mean**	**Mean**	**diff.**	**t stat**	**Perc.**	**Mean**	**diff.**	**t stat**	**Perc.**
AB	0.726	0.749	-0.023*	-1.943	-3.1%	0.671	0.055***	4.674	8.2%
GB	0.781	0.734	0.047***	6.625	6.4%	0.727	0.055***	7.802	7.5%
LR	0.731	0.692	0.040***	3.762	5.7%	0.706	0.026**	2.317	3.6%
RF	0.723	0.720	0.004	0.271	0.5%	0.699	0.024*	1.780	3.5%
SVM	0.681	0.628	0.053***	3.851	8.4%	0.622	0.059***	3.926	9.5%
XGB	0.732	0.735	-0.003	-0.373	-0.4%	0.697	0.035***	4.405	5.0%

[52] Tuned models are available at: https://github.com/IAQI/IAQI_online_appendix/tree/Tuned-Models

Average	0.729	0.710	0.019***	3.883	2.7%	0.687	0.042***	8.331	6.1%

Note:
*** two-sided p value < 0.01; ** two-sided p value < 0.05; * two-sided p value < 0.10. AB is AdaBoost, GB is Gradient Boosting, LR is Logistic Regression, RF is Random Forest, SVM is Support Vector Machine, and XGB is Extreme Gradient Boosting. RawAudit include 91 audit variables that are readily available in raw or close-to-raw form from Audit Analytics database. ARV include 31 theory-driven and publicly available audit-related variables. IAQI is a subset of ARV that include 13 predictive audit-related variables.

Like how P Scores were calculated for IAQI, we generated P Scores using the probability prediction output from tuned GB models with RawAudit as model inputs, as GB models generate the highest average AUC with RawAudit as input. Untabulated results show that, all else being held equal, a 1-point increase in the PSCORE_RAW_GB increases the odds of having an actual MAR by 2.9 times and that engagements audited by Big4 auditors have significantly lower PSCORE_RAW_GB.[53]

3.8 Is the Predictive Power Good Enough?

We find that the highest average AUC that can be achieved by ARV, IAQI, and RawAudit are 0.727, 0.749, and 0.781, respectively. Although an AUC between 0.7-0.8 is considered acceptable by academic standards, and ARV, IAQI, and RawAudit have significantly higher AUC predictive power than benchmark financial variables, it is not clear whether the level of predictive power that can be achieved using these methods is perceived to be useful by practitioners. To ascertain practitioners' threshold for usability, we conducted a survey among senior auditing practitioners and standard setters.[54] Since auditing practitioners, in general, lack technology literacy (e.g., Jackson, Michelson, and Munir 2020) and evaluating predictive power requires such knowledge, we targeted

[53] In untabulated results, the coefficient for PSCORE_RAW_GB is 1.065, so the odds ratio is 2.01 (e^0.699); the coefficient for PSCORE_GB is 0.478, so the odds ratio is 1.61 (e^0.478); the coefficient for PSCORE_XGB is 0.607, so the odds ratio is 1.83 (e^0.607).
[54] This survey was approved by Institutional Review Boards (IRB).

practitioners who are actively involved in audit innovation to obtain reliable survey results.

Before the survey was sent to potential participants, we pilot tested it with a retired audit partner of a Big4 firm and a senior manager of audit analytics in audit innovation of a Big4 firm to ensure that the survey uses wordings understandable to practitioners. To approximate expected average AUC, the question asked, “How many times out of 100 do you think an auditing software should correctly predict material misstatement for this software to be useful?” [55] Our pilot testing participants indicated that this way of question framing can be understood by most auditing practitioners.

We identified potential participants who are actively involved in audit innovation via the authors’ professional connections and sent the survey to 12 potential practitioners (including 1 participant affiliated with standard setters). The survey was voluntary, and all responses were anonymous. In total, we got 10 responses. Table 18 summarizes the responses.

Table 18. Survey Results Summary

Demographic Information			**Survey Questions**		
Position	**Years of auditing experience**	**Knowledge about machine learning (1 to 7 indicating completely unfamiliar to very familiar)**	**How many times out of 100 do you think this auditing software should correctly predict material misstatement for this software to be useful? (Please provide an integer**	**How confident are you in the judgments you made in the previous question? (1 to 7 indicating completely unsure to completely confident)**	**Is your organization using machine learning in audit engagements?**

[55] AUC shows the model’s ability to discriminate between positive examples and negative examples (Faraggi and Reiser 2002). For example, an AUC of 0.5 means that 50% of the time, the model will correctly assign a higher absolute risk to a randomly selected positive instance than to a randomly selected negative instance. Thus, the question “How many times out of 100 do you think [an] auditing software should correctly predict material misstatement for this software to be useful?” can approximately capture the level of expected AUC from a prediction model.

			between 0 and 100)		
Head of Audit	33	3	30	2 (removed from analysis)	Yes
Audit manager	18	5	90	5	No
Audit partner	38	6	90	7	No
Audit manager	10	6	50	6	No
Audit manager	5	3	99	5	No
Senior auditor	6	3	90	5	Maybe
Audit partner	40	5	80	5	No
Audit partner	28	6	70	6	Maybe
Audit partner	11	6	90	5	No
Asst. Chief Auditor (Standard Setter)	10	3	90	6	(No answer provided)

Among the respondents, there were 4 audit partners, 3 audit managers, 1 head of audit, 1 senior auditor, and 1 standard setter. On average, the respondents have 20 years of audit experience. When answering the question, “How many times out of 100 do you think an auditing software should correctly predict material misstatement for this software to be useful?”, respondents provided an integer between 0 and 100. To assess how confident respondents were in providing answers, we asked the quality-check question, “How confident are you in the judgments you made above?”, where respondents could choose from 1 to 7, ranging from completely unsure to completely confident. Since one respondent, the head of audit, was very unsure about his/her response (confidence value of 2 out of 7), we excluded his/her response from calculations.

The average value of responses with high confidence (confidence value of at least 5 out of 7) to the question "How many times out of 100 do you think an auditing software should correctly predict material misstatement for this software to be useful?" is 83.2, suggesting that practitioners expect an average AUC to be at least 83.2% for a prediction model of audit failure to be useful in practice. Thus, there is still a 5%-10% shortfall in predictive power measured by AUC for advanced machine learning models using audit variables as inputs to be perceived as useful in practice.[56] This threshold can be used as an alternative benchmark for researchers to assess if the predictive power of their models meets practitioners' expectations. For audit failure prediction model to achieve practitioners' expectations, future research may explore whether adding variables collected from unconventional sources, such as social media, can increase the predictive power of audit failure prediction models. Furthermore, with the fast progress of machine learning, future research can explore whether novel machine learning algorithms can better model and predict audit failure.

3.9 Sensitivity Analysis

3.9.1 Alternative Feature Selection

In the main analysis, we adopted a sequential backward feature selection process to identify predictive variables for each algorithm. In this section, we explore two alternative feature selection methods: an algorithm-independent method called univariate feature selection, and permutation feature importance based on random forest.[57]

[56] The highest average AUC that can be achieved by ARV, IAQI, and RawAudit are 0.727, 0.749, and 0.781, suggesting a gap of 0.105, 0.083, and 0.051, respectively.

[57] Instead of using feature selection methods, one could also use feature construction techniques to aggregate features in novel ways and to reduce feature dimension (Sondhi 2009). For example, one of the most common feature construction methods is Principal Component Analysis or PCA (Sondhi 2009). In un-tabulated analysis, we found that PCA does not generate superior performance compared to using the

Univariate feature selection involves running a one-way ANOVA test between a feature and an outcome variable to see if there is a significant association between them (Pedregosa et al. 2011). We used 5-fold stratified cross-validation to implement univariate feature selection. For each fold, we applied univariate feature selection to select half (i.e., 15) of the most significant features out of the 31 ARV, ranked by their F statistic. This process generated 12 features that were chosen as significant features in each integration, and seven of them overlap with IAQI.[58]

Permutation feature importance identifies significant features from an established prediction model by highlighting features whose permutated values will increase the most significant prediction error (Pedregosa et al. 2011; Molnar 2021). We implemented permutation feature selection based on 5-fold stratified cross-validation, ranking features based on their average importance values. Seven out of the 11 most important features identified by permutation feature importance overlap with IAQI.[59]

---

original predictors. For example, our analysis shows that the average AUC in the hold-out test years 2015-2017 is 0.56, 0.52, 0.56, 0.59 when we construct 2, 5, 10, and 15 factors from audit-related variables, respectively. Furthermore, another challenge about using feature construction in auditing research is that the constructed features often lack interpretation. However, being able to interpret the model input and output is important for practitioners and researchers (Zhang et al. 2022). Since one of the main objectives of this research is to examine which audit-related variables that have been widely used in the literature have the most predictive power, we need to be able to clearly identify which feature has the most predictive power. Thus, due to the non-superior predictive power of using feature construction and the challenge of interpretability, we do not report feature construction results in the paper.

[58] The 12 most significant features from univariate feature selection are: Industry Specialization_National, Office Size, Big4, New Client, Tenure, Integrated Audit, Accelerated Filer, Auditor Resignation, Audit Report Lag, Internal Control Weakness, Abs (Disc. Accruals), and Abs (Accruals/CFO). The seven overlapping variables with IAQI are: Office Size, Tenure, Integrated Audit, Auditor Resignation, Audit Report Lag, Internal Control Weakness, and Abs (Accruals/CFO).

[59] The 11 most important features from permutation feature importance are: Audit Report Lag, Abs (Accruals/CFO), Tenure, Abnormal Audit Fee, Audit Fee, Industry Specialization_National, Influence, DD Residual, Office Size, Abs (Accruals), and Disc. Accruals. The seven overlapping variables with IAQI are: Audit Report Lag, Audit Fees, Abs (Accruals/CFO), Tenure, Office Size, Abs (Accruals), and Disc. Accruals.

3.9.2 Alternative Measure of Audit Failure

In this section, we adopt an alternative measure of audit failure: MAR that also receive Accounting and Auditing Enforcement Releases (AAER) from the SEC. AAER are results of SEC investigations for securities law violations (SEC 2017). Prior literature adopts AAER instances to identify severe frauds or accounting misconducts (e.g., Perols 2011; Perols et al. 2017; Cecchini et al. 2010; Brown et al. 2020; Bao et al. 2020). The alternative measure of audit failure used here is a subset of MAR that are also investigated by SEC and eventually received AAER, indicating that these material misstatements involve severe frauds or accounting misconducts. We use this alternative measure of audit failure and perform the feature selection stated in the main analysis, obtaining 10 most predictive features, nine out of which overlap with IAQI.[60]

3.9.3 Alternative Measure of Algorithm Performance

As an alternative to AUC, or area under the ROC curve, we adopt an alternative evaluation metric: area under the Precision-Recall Curve (PR-AUC). The Precision-Recall curve is similar to the ROC curve. A ROC curve plots false positive rate on one axis and true positive rate on the other. A Precision-Recall curve replaces the false positive rate axis with precision (Jeni, Cohn, and De La Torre 2013; Saito and Rehmsmeier 2015). PR-AUC balances precision (i.e., fraction of true audit failures among the predicted audit failures) and recall (i.e., the fraction of true audit failures predicted), and it is useful when data are imbalanced (Jeni et al. 2013; Saito and Rehmsmeier 2015). Some prior studies use F-score to measure algorithm performance

---

[60] The nine overlapping features are office size, tenure, audit fee, audit report lag, influence, auditor competition, Disc. Accruals, Abs(Accruals), and Abs(Accruals/CFO). The only different predictive feature is abnormal audit fee.

(e.g., Bertomeu et al. 2020). Like PR-AUC, F-score also weighs precision and recall (Rijsbergen and Joost 2004). However, unlike F-scores, whose value depends on specific classification thresholds, PR-AUC is an aggregated value under different possible classification thresholds, and better reflects overall predictive power (Jeni et al. 2013; Saito and Rehmsmeier 2015). We use PR-AUC as an alternative measure of algorithm performance and perform the same feature selection as stated before. Under this alternative algorithm performance measure, nine out of the 13 IAQI from the main analysis are also found to be the most predictive variables. [61]

Overall, across different measures of audit failure and algorithm performance, tenure, audit fee, audit report lag, influence, and discretionary accruals were consistent predictors of audit failure.

3.10 Discussion and Conclusion

We identify a fundamental question to audit quality research: how well can audit-related variables (ARV) identified from extant literature predict audit failure out-of-sample, and which are the most predictive? Over decades, auditing research has identified a broad set of theory-based variables that are associated with audit quality. However, little is known about how effectively these ARV forecast audit failure out-of-sample. Learning about the predictive power of ARV can help assess the distance between theory and practice (Shmueli 2010; Shmueli and Koppius 2011) and narrow down a long list of ARV to those are the most predictive of audit failure.

We use material restatement of annual reports due to GAAP violations or frauds (or MAR) as the proxy for audit failure. We collect 31 publicly available ARV as

[61] They are auditor's industry specialization at the city level, internal control weakness, tenure, auditor resignation, audit fee, client influence, audit report lag, integrated audit, and discretionary accruals.

predictors and compiled a dataset of U.S. public firms spanning the period between 2005 and 2017. Using machine learning, we find that ARV have acceptable predictive power by academic standards, and that ARV significantly outperform benchmark variables. Based on a feature subset selection process, we identified 13 ARV to be the most predictive and we term them informative audit quality indicators or IAQI. As a selected subset of ARV, IAQI carry significantly higher predictive power than ARV in forecasting audit failure. The 13 selected IAQI variables represent auditor characteristics, audit task characteristics, auditor-client contracting features, auditor communication, and the quality of the audited financial statements. We further synthesize IAQI into a predictive score (P Score) and show that P Score significantly outperform existing academic measures in incrementally associating with audit failure, providing an alternative measure for practitioners and researchers in inferring audit failure. Although ARV, and especially IAQI, have significantly higher predictive power than benchmark variables in predicting audit failure, their predictive power has yet to exceed practitioners' expectations. Consequently, predictions generated from ARV, IAQI, or P Scores should not be treated as absolute and definitive predictions of audit failure. Instead, predictions should be interpreted within specific contexts to *assist* stakeholders in gauging the likelihood of audit failure.

This study adds to the growing body of accounting research that makes use of machine learning by comparing different machine learning methods to examine whether ARV that are commonly used by academics are effective in predicting audit failure out-of-sample. Our finding that ARV's predictive power is acceptable by academic standards but does not exceed expectations from practitioners suggests a gap between theory and practice and, therefore, room for improvement. We also add to the stream of auditing literature that validates audit quality measures by focusing on out-of-sample predictive power. Our results inform practitioners and researchers by proffering IAQI that are most likely to predict audit failure while pointing out which ARV have low predictive power. Furthermore, the P Score developed in this research can be used as an alternative measure by academics and practitioners when inferring audit failure. Moreover, we provide an accuracy benchmark for real-world implementation (average AUC of 83.2%) according to practitioner opinions, which may be useful in future studies that predict audit failure. Overall, this study has both academic and practical implications in audit failure prediction.

A limitation for this paper is the usage of MAR as a proxy for audit failure. Because not all audit failures produce a material restatement (Gaynor et al. 2016; Suresh and Guttag 2019), the use of MAR may have introduced measurement error. However, we tried to reduce the potential impact of measurement error by using a relatively large sample, as large samples makes the parameters in a machine learning model converge towards the correct value (Suresh and Guttag 2019). Furthermore, we cross-validated the main results by using alternative measures of audit failure and algorithm performance and found similar results.

Future research may seek to adopt other proxies of audit quality, such as those outlined in the Part 1 Findings of the PCAOB inspections and proposed in audit firms' internal assessments of audit quality. However, these proxies are inherently limited by their small sample size and limited representativeness. Researchers must balance the trade-offs of using alternative proxies of audit quality. Future research can also explore developing innovative measures of audit failure and utilize audit-related variables generated from alternative sources, such as social media and online forums.

CHAPTER 4: THE EFFECTS OF ARTIFICIAL INTELLIGENCE ON THE ACCURACY OF MANAGEMENT EARNINGS FORECASTS

4.1 Introduction

In recent years, an increasing number of firms have made, or are planning to make, substantial investments in artificial intelligence (AI) technologies to optimize business operations, which in turn, may relate to improvements in managerial decision-making (Brynjolfsson et al. 2018a; Deloitte 2019; Colson 2019). AI is the theory and development of computer systems that can perform tasks that typically require human intelligence (IEEE-USA 2017). AI technologies include machine learning (ML), natural language processing (NLP), computer vision, and robotics (Russell and Norvig 2002). Perrault et al. (2019) report that 58% of large firms have adopted AI and that global investment in AI is growing steadily. While prior research has focused on providing descriptive discussions about the potential impact of AI on accounting or designing AI systems to predict accounting balances and firm events (e.g., Issa et al. 2016; Kokina and Davenport 2017; Perols 2011; Perols et al. 2017; Bao et al. 2020; Ding et al. 2020), little is known about the relation between AI and an important managerial outcome we expect AI to affect, the accuracy of management earnings forecasts.

We focus on the association between AI adoption and the accuracy of management guidance for several reasons. First, management guidance is a key voluntary disclosure mechanism, and accurate management guidance reduces information asymmetry, lowers litigation costs, and improves a firm's reputation (Williams 1996; Skinner 1997; Coller and Yohn 1997; Graham et al. 2005). Second, the accuracy of management earnings forecasts generally depends on the quality of financial and non-financial internal information, and AI arguably improves the quality of such information.

Since firms use AI to improve forecast accuracy for sales and costs, predict fraud, and assist with the accounting and financial reporting process (Perols 2011; Dechow et al. 2011; Perols et al. 2017; Mckinsey 2021; Bao et al. 2020), it may affect the accuracy of management guidance by improving the quality of management reports that are used as input to produce forecasts. Third, the notion that AI could improve the accuracy of management guidance is conceptually consistent with the “management coefficients theory” (Bowman 1963), which proposes that decision-support systems can help managers make reliable decisions. Thus, understanding the impact of AI on the accuracy of management guidance – which is an observable output – may present implications for other less directly observable managerial outcomes that depend upon the quality of internal management reports.

We posit that the association between AI and the accuracy of management guidance occurs through two channels. First, AI technologies have frequently been shown to improve *predictions* by enhancing the analysis of large amounts of data across business operations and supply chains (Sullivan and Zutavern 2017; Deloitte 2019; Agrawal et al. 2018; Cho et al. 2019). In turn, more accurate predictions of sales and costs generated by AI should lead to more accurate earnings forecasts by improving the quality of managers’ internal reports used to develop earnings forecasts. Second, AI technologies have been shown to increase the *quality of data* by automating procedures for data collection, processing, and validation (Lacity et al. 2015; Willcocks et al. 2015a; Willcocks et al. 2015b; Moffitt et al. 2018; Cho et al. 2019; Zhaokai and Moffitt 2019). AI’s improvement to the quality of data overall also enhances the quality of the internal information used as inputs to earnings forecasts.

Despite the purported benefits of AI to the accuracy of earnings guidance, there are also concerns about its potential impact. For example, issues with potential bias in the data that feed into the AI models and managers' overreliance on AI could lead to suboptimal inputs to earnings forecasts (Munoko et al. 2020). Moreover, AI implementation requires a substantial amount of time, money, and resources because complementary organizational capabilities, such as IT expertise, take time to develop (Cockburn et al. 2018; Bergstein 2019). Therefore, the purported benefits of AI to earnings forecast accuracy may not yet be realized after only a few years' implementations.

Hence, it is an empirical question as to whether AI improves the accuracy of management guidance. Considering the expected benefits of AI in this area, we hypothesize that, after a firm has implemented AI, managers will be more likely to issue more accurate guidance compared to firms that do not implement AI.

Although the link may be direct between AI and the quality of internal information that produces accurate management guidance, AI may indirectly impact the accuracy of management earnings forecasts through improvements in operational performance and financial reporting quality. As a result, we further explore the mediating factors through which AI may impact the accuracy of management earnings forecasts.

To study the main research question, we identify 117 non-technology firms that have implemented AI from 2014 to 2018 from three sources: a proprietary dataset, firms' public disclosures, and news articles. [62, 63] Using a difference-in-differences research design, we find evidence of a positive relation between AI and the accuracy of management earnings forecasts. Economically, we find that AI implementers have 21.3% more accurate management guidance than a population-based control group. Because the linkage between AI and the accuracy of earnings guidance may be indirect, we employ a structural equation modeling (i.e., regression path analysis) to explore whether firms' operational performance and financial reporting quality are the mediating factors through which AI may impact forecast accuracy. We find that AI has an indirect effect on the accuracy of earnings guidance through firms' operational performance. However, this indirect effect is small (15.6% of the total effect) compared to AI's direct effect on earnings forecasts accuracy (84.3% of the total effect). We do not find evidence that financial reporting quality mediates the relation between AI and forecast accuracy. Furthermore, we find that the impact of AI on the accuracy of guidance is more pronounced when the forecast horizon is longer, suggesting that AI reduces information uncertainty in developing earnings guidance. Finally, since AI encompasses different technologies, we examine the type of AI technology that significantly contributes to improvements in the accuracy of management guidance and find that ML contributes the most compared to others, namely robotic process automation (RPA), natural language

[62] We target non-technology firms because they adopt AI in business operations, which generate data (e.g., sales and cost of goods sold) that flows into management guidance. Although technology firms can also use AI in business operations, they often sell AI as products or services, which could introduce noise to the sample since it will not be clear whether the observed impact of AI on the accuracy of management earnings forecasts is from AI's ability to influence the quality of information generated in business operations or from selling AI as a product or service. For our sample firms, we manually searched AI announcements of technology firms and observed that technology firms such as SAP and Cisco disclose sales of their AI products and services, rather than the use of AI to improve firm operations.

[63] Our starting point for identifying AI adopters is 2014 as this is the first year that the proprietary dataset provider identifies AI adoption by firms. Additionally, a trend analysis of firms' public disclosure of AI adoption shows that there is an increase in the number of AI disclosures starting in 2014.

processing (NLP), and computer vision. We conduct sensitivity tests related to the measurement of forecast accuracy, endogeneity originating from AI adoption and management guidance disclosure, alternative control groups, and possible omitted variables to investigate the robustness of our main results and find that they hold.

Although the accuracy of management guidance is our primary focus due to its natural linkage with the functionalities of AI to produce accurate forecasts, we additionally explore the relationship between AI adoption and other aspects of management guidance, including precision, frequency, and bias. We do not find evidence of a relation between AI adoption and either precision or frequency of management guidance, consistent with the notion that precision and frequency are primarily affected by firms' policies prescribing the types and frequency of guidance issuance (Graham et al. 2005; Einhorn and Ziv 2008; Tang 2013) and are, therefore, less likely to be affected by AI adoption. The null finding on precision is also consistent with the trade-off between precision and accuracy since a forecast with a high level of precision is more likely to be wrong (Hayward and Fitza 2017). Finally, we find no evidence that AI adopters produce directionally biased guidance.

Considering that our sample size is small, the results of this study should be interpreted with caution. Nevertheless, our study contributes to the accounting and accounting information systems literature in several ways. First, our study adds to the interdisciplinary literature of AI and accounting by examining whether AI implementation impacts the accuracy of management guidance, an important managerial outcome we expect AI to affect, given that AI may better inform managers' decisions. Unlike existing research that provides descriptive discussions about the potential impact of AI on accounting (Issa et al. 2016; Kokina and Davenport 2017) and that designs AI systems to predict accounting balances and firm events (e.g., Perols 2011; Perols et al. 2017; Bao et al. 2020; Ding et al. 2020), our study provides archival evidence that the *actual* use of AI in firms is associated with improvements to the accuracy of earnings guidance.

Second, our study contributes to the management forecast disclosure literature by providing insights into the earnings forecast properties that AI might affect. We find evidence that AI translates into more accurate forecasts, consistent with expectations of the core functions of AI to improve data quality and predictions (Perols 2011; Perols et al. 2017; Bao et al. 2020; Ding et al. 2020). In contrast, we find no evidence that AI is associated with management earnings forecasts' precision, frequency, or bias. Further, we find that AI can more profoundly improve the accuracy of management earnings forecasts when the horizon is longer, which further illustrates its influence on forecast uncertainty. Finally, we find that ML is the main contributor to the improved accuracy of management guidance. Collectively, our findings suggest that the adoption of AI can contribute to improvements in the accuracy of management guidance.

Third, our study contributes to the literature on the impact of the information environment quality on decision making by identifying an IT-based determinant that might influence firms' internal information. Studies document that internal control quality and enterprise system implementation impact information quality. For example, studies show that enterprise system implementation affects the quality of internal information (Brazel and Dang 2008; Dorantes et al. 2013; Paredes and Whitley 2018) and that lower quality of internal controls decreases the quality of information (Feng et al.

2009). We extend this vein of literature by focusing on AI, an emerging technology that, unlike enterprise systems, processes and analyzes large amounts of information to aid the managerial decision-making process.

Finally, our study should be of interest to managers as they consider the adoption of AI to improve their business operations. Regulators overseeing public firms might be interested in the implementation of AI that could improve certain aspects of the quality of firms' internal information and, more specifically, the quality of voluntary disclosures. For instance, the Securities and Exchange Commission (SEC) is currently considering the impact of AI on accounting quality and asserts that firms are increasingly using it to conduct comprehensive analyses on financial information (SEC 2018; SEC 2019; SEC 2021).

4.2 Related Literature and Hypothesis Development

4.2.1 Artificial Intelligence

AI is the field of making computers achieve human-like intelligence, including thinking and acting humanly and/or rationally (IEEE-USA 2017). It accomplishes this by utilizing capabilities of cognitive automation, machine learning, reasoning, hypotheses generation and analysis, and knowledge representation (Russell and Norvig 2002; IEEE-USA 2017). Although the term "Artificial Intelligence" was coined in 1956 (Russell and Norvig 2002) and certain forms of AI technologies like expert systems have existed for several decades, it is not until recently that there have been substantial improvements to it to make it more useful, more accessible, and more affordable than before (Deloitte 2019). These improvements are attributable to the increase in computing power, decrease in storage and processing costs, and the abundance of data. AI is also becoming more accessible and affordable than before (Brundage et al. 2018; Deloitte 2019; Agrawal et al. 2018).

AI can be disaggregated into a spectrum where on one end is the least intelligent form of AI, such as RPA and expert systems to execute structured activities, and on the other end is the most intelligent form of AI, such as artificial general intelligence (AGI), which requires minimal human intervention to conduct complex cognitive activities (Davenport and Ronanki 2018). The in-between sections within the spectrum would include technologies like computer vision, robotics, and machine learning as these technologies are capable of interpreting images, text, and adapting to their environment to learn from it, respectively (Brynjolfsson et al. 2018). Collectively, AI can be referred to as an umbrella of different types of technologies that range from RPA to AGI.[64]

AI can improve business operations, which, in turn, could influence IIQ. In addition, AI complements the firms' existing large-scale enterprise systems. While enterprise systems facilitates the timely dissemination of internal information across business functions, AI leverages this information and produces recommendations that are crucial to managerial decision-making. Prior research has shown that enterprise systems can improve the accuracy of management earnings forecasts (e.g., Brazel and Dang 2008; Dorantes et al. 2013). While enterprise systems are one type of information technology that can improve the accuracy of management earnings forecasts, our focus in this paper

[64] This definition and disaggregation of AI is also consistent with McKinsey's firm survey description of AI. McKinsey (2020) defines AI as: "the ability of a machine to perform cognitive functions that we associate with human minds (such as perceiving, reasoning, learning, and problem solving) and to perform physical tasks using cognitive functions (for example, physical robotics, autonomous driving, and manufacturing work)." Accordingly, the questions in the survey relate to RPA, computer vision, NLP, robotics, machine learning, and AGI.

is on the effect of AI, a different type of technology, on the accuracy of management earnings forecasts.[65]

4.2.2 Hypothesis Development

Management guidance represents a key voluntary disclosure mechanism that provides valuable information to capital market participants. The issuance of management guidance is beneficial to a firm as it can reduce information asymmetry, litigation costs, and improve its reputation for credible and transparent reporting (Williams 1996; Skinner 1997; Coller and Yohn 1997; Graham et al. 2005). For example, managers may be motivated to issue earnings forecasts because it gives market participants a signal about their managerial ability to forecast future business outcomes (Trueman 1986). In addition, litigation costs are lower when disclosure is timely (Skinner 1997). Coller and Yohn (1997) suggest that managers are more likely to issue earnings forecasts when there is information asymmetry between managers and investors. Furthermore, Clement et al. (2003) suggest that forecasts that confirm the market's beliefs about earnings reduce uncertainty about future earnings. Beyer and Dye (2012) argue that firms enhance their reputation by being forthcoming about unfavorable news in their forecasts. Moreover, the literature finds that managers' disclosure activities are sensitive to stock price incentives and may strategically walk down analysts to meet or beat their earnings estimates. For example, managers sell more of their companies' shares

[65] Although we do not control for enterprise system adoption in the main analysis because we do not have access to this proprietary information, we include industry fixed effects, which reduces the possibility of omitted variable bias. In the robustness tests section, we performed a manual search of enterprise system adoption for our treatment firms and include enterprise system adoption as a control variable. The main results hold after controlling for enterprise system adoption for treatment firms.

after good news is released than when bad news is released (Noe 1999). Moreover, managers meet or beat analysts' expectations to avoid negative earnings surprises (Matsumoto 2002). In sum, several incentives may affect the properties of management guidance (i.e., whether, when, and how to issue forecasts).

Regardless of managers' incentives or abilities to produce earnings guidance, guidance issued with poor-quality information will likely be less accurate. Feng et al. (2009) document that lower quality of financial information decreases managers' propensity to issue guidance and that the issued guidance is more likely to be less accurate and specific. Jennings et al. (2013) suggest that managers must access high-quality internal information to provide accurate management forecasts. Cassar and Gibson (2008) survey private firms and show that firms with internal accounting reporting and budgeting processes make more accurate sales forecasts. Using data on enterprise system implementations, Dorantes et al. (2013) find that managers are more likely to issue forecasts and that these forecasts are more accurate after a firm has implemented an enterprise system. Cheng et al. (2018) examine the impact of SFAS 142 on the information environment and find that firms that are impacted by this regulation experience an improvement in the accuracy of management earnings forecasts. In summary, these studies indicate that higher-quality internal data leads to more accurate management guidance. In this study, we focus on the accuracy of management guidance not only because it signals guidance quality (e.g., Ajinkya et al. 2004; Feng et al. 2009; Hilary et al. 2014), but also because it corresponds to the core function of AI, which is to generate accurate predictions (Alpaydin 2014; Agrawal et al. 2018).

The Channels Through Which AI Might Improve the Accuracy of Management Earnings Forecasts

The "management coefficients theory" proposes that decision-support systems can aid managers in making reliable decisions (Bowman 1963). Consistent with this theory, it is plausible that when AI is used to enhance business operations, its implementation is associated with more accurate management guidance from the following two main channels.

First, when AI is used to enhance business operations, it can generate relevant and accurate *predictions* by analyzing large amounts of data across business operations and supply chains (Sullivan 2017; Agrawal et al. 2018; Deloitte 2019; Cho et al. 2019). Better predictions could be reflected as higher quality information, potentially translating into more informed managerial decisions (Agrawal et al. 2018). For example, Ding et al. (2020) explore using machine learning to make accounting estimates related to insurance claims, and they conclude that it can significantly improve their accuracy. Applications of AI for fraud prediction show the superior performance of machine learning compared to conventional methods (Cecchini et al. 2010; Perols 2011; Perols et al. 2017; Bao et al. 2020). Research has also proposed AI applications to predict stock prices (Akita et al. 2016), asset returns (Feng et al. 2018), inventory levels (Kartal et al. 2016), and sales (Syam and Sharma 2018). While this stream of research advances knowledge by designing AI systems that predict accounting balances or firm events, they are limited in that they do not examine whether the actual use of these systems in firms might contribute to an important outcome that AI may affect, the accuracy of management guidance.

Second, AI technologies can increase the *quality of data* (e.g., accuracy, timeliness, completeness, and validity) that are the inputs in prediction generation by automating the procedures for data collection, processing, and validation (Lacity et al. 2015; Willcocks et al. 2015a; Willcocks et al. 2015b; Moffitt et al. 2018; Cho et al. 2019; Zhaokai and Moffitt 2019). For example, RPA can optimize and streamline business processes and automate routine accounting and finance tasks to increase the accuracy, timeliness, and validity of accounting data (Lacity et al. 2015; Willcocks et al. 2015a; Willcocks et al. 2015b; Lacity and Willcocks 2018; Moffitt et al. 2018; Cooper et al. 2021). Models for textual analysis, including deep learning and natural language processing, can be trained to automatically extract the sentiment features from business documents, such as financial statements, contracts, and even social media, to potentially deliver relevant and reliable information for managers and auditors. This information might be used to predict internal control weaknesses (Sun 2018; Zhaokai and Moffitt 2019), procurement irregularities (Sun and Sales 2018), and financial misstatements (Bartov et al. 2018; Zhaokai and Moffitt 2019; Rozario et al. 2022). Similar to the AI for accounting prediction literature, these studies do not examine the actual use of AI systems and their link to the accuracy of management guidance.

Consider the following announcements:

Chipotle Mexican Grill Inc. – Conference call transcript for Q2 2018:

> *"New tool will have best-in-class sales forecasting component that will leverage machine learning to remove guess work of determining sales and labor needs for business."*

Johnson & Johnson – Article from the Institute of Chartered Accountants in England and Wales:

> *"Implementing Robotic Process Automation (RPA) in Johnson & Johnson's finance department has resulted in standardization, improvements in workflow and greater accuracy."*

In the first example, the firm's use of AI for sales forecasting will provide managers with a recommendation for expected sales, and sales forecast will likely be one input that management incorporates as they estimate earnings guidance. In the second example, the firm uses AI to improve the quality of internal management reports in the finance department. As a result, managers may use higher-quality information from finance to estimate earnings more accurately. These examples illustrate that AI might improve the accuracy of guidance through two channels: the accuracy of forecasts and the quality of data from which forecasts may be generated.

Hurdles That May Delay or Prevent AI From Benefiting the Accuracy of Management Earnings Forecasts

Although AI is expected to deliver benefits regarding more accurate management guidance via the channels mentioned above, several hurdles may delay or prevent AI from improving the accuracy of management earnings forecasts. First, AI is still in the early stages of adoption for most non-technology firms (Agrawal et al. 2018; Brundage et al. 2018; Deloitte 2019). Since AI technologies, especially tools based on machine learning, usually take time to learn and improve (Agrawal et al. 2018), their expected benefits to the accuracy of management earnings forecasts may not yet be realized after only a few years' implementations. Second, similar to other IT technology adoption that requires developing complementary organizational capabilities (Bresnahan et al. 1996), AI requires large amounts of machine-readable data, employees who have data expertise, and new organizational structures that emphasize knowledge sharing (Cockburn et al.

2018), all of which take time to develop. Third, there may be bias in the data that feed into the AI models, or managers may over-rely on AI, potentially leading to suboptimal decisions that hamper the accuracy of management earnings forecasts (Munoko et al. 2020).

To sum up, various challenges may delay or prevent AI from improving the accuracy of management guidance. As a result, it is an empirical question as to whether AI might improve the accuracy of management earnings forecasts. Since AI can potentially improve the accuracy of the information that is generated and consumed within the firm, we expect that firms that have implemented this technology will be likely to produce more accurate guidance (Baginski and Hassell 1997; Ajinkya et al. 2005; Feng et al. 2009; Dorantes et al. 2013; Ittner and Michels 2017; Cheng et al. 2018). Accordingly, we formulate the following main hypothesis:

> HYPOTHESIS: *Ceteris Paribus, AI implementation is associated with more accurate management guidance after its implementation.*

To test this hypothesis, we first examine the association between AI adoption and the accuracy of management earnings forecasts using a Difference-in-Differences (DID) design. We next decompose the observed association into direct and indirect effects. We further examine the condition in which AI's improvement to the accuracy of management earnings forecasts is amplified. Finally, we decompose AI into ML versus the other AI, i.e., RPA, NLP, and Computer Vision, with the expectation that ML is the driving factor for AI's improvement to the accuracy of management earnings forecasts.

4.3 Research Design

4.3.1 Data and Sample

We identified AI usage by non-technology public firms from 2014 to 2018 from three sources. First, we used a proprietary dataset from Global Software Leads (GSL). GSL is a global technology marketing company that provides software vendors with information about when and how firms implement certain technologies (e.g., enterprise systems and artificial intelligence) to help them with marketing campaigns. Therefore, GSL has a strong incentive to maintain an accurate list of technology implementations (Murthy et al. 2020). GSL created the AI usage dataset in 2014 for firms that began to adopt AI starting from that time, and it has been updating the dataset every year via telephone surveys and web searches.

Nonetheless, because the GSL dataset may not include all firms that adopt AI, we extended our search of AI adopters to include firms' public disclosures of AI adoption in financial reports, conference calls, and press releases. Furthermore, since not all firms have the incentive to disclose the adoption of AI, we also extended our search to firms that have all necessary variables from Compustat and IBES Guidance and that are not matched to either GSL or firms' public disclosures. We searched news articles from sources such as Lexis Nexis and Factiva for these firms to identify AI adoption. The use of the three sources to identify the treatment firms, i.e., the AI adopters, should produce a more complete and accurate dataset than a single source.

We adopt the following keywords in our search of firms' public disclosures and news articles: "artificial intelligence," "machine learning," "natural language processing," "natural language generation," "robotic process automation," "intelligent system," "intelligent decision aid," "intelligent decision support system," "intelligent agent," "audit support system," "expert system," and "knowledge-based system" (Sutton et al. 2016; Davenport and Ronaki 2018; Brynjolfsson et al. 2018a; Mckinsey 2021).[66] Next, we follow prior literature that identifies the implementation year for enterprise systems (Nicolaou 2004; Masli et al. 2010; Dorantes et al. 2013; Paredes and Wheatley 2018) as a guide to obtaining the implementation year for AI. Based on this, we extract the first year of mention of the AI adoption.

We manually search for the context scenario of AI use and the type of AI (e.g., machine learning or robotics) primarily implemented by the treatment firms identified from the three sources. We verify that our treatment firms have implemented AI to improve operations rather than selling AI as direct products/services.[67] Operation improvement is the context in which AI has been shown to improve internal management reports' quality via better prediction and better quality of data. Furthermore, the procedure for the manual search helps to ensure that the implementation year we collect is for the actual AI implementation, rather than the intention to implement or high-level comments about AI, such as "AI can improve firm value" (Nicolaou 2004; Cheng et al. 2019).

The GSL dataset contains 1423 public firms that implemented AI between 2014 and 2018 and are covered by COMPUSTAT. From the 1423 firms, we removed 918 public firms that are not covered by the IBES Guidance database since 2014. Further, we removed 407 public firms whose AI implementation year cannot be determined, do not

---

[66] We adopt these keywords from the academic literature as they encompass a comprehensive list of AI technologies. Moreover, these keywords are consistent with those in the McKinsey survey on AI adoption.
[67] For example, most technology and telecommunication firms like Google, Microsoft, and AT&T that provide AI products and services to other firms were excluded.

use AI to improve operations, and belong to technology industries. [68] We then removed 20 firms that do not have all the necessary dependent and control variables. The total subsample from GSL consists of 78 firm-year observations.

Next, we identified 734 public firms covered by COMPUSTAT that mention AI keywords in their public disclosures (i.e., financial reports, conference calls, and press releases). We removed 163 firms that are duplicated from the GSL database. Next, we removed 544 firms not covered by the IBES Guidance database, whose AI implementation year cannot be determined, that do not use AI to improve operations, and that belongs to technology industries. We then removed nine firms that do not have all the necessary variables, resulting in a subsample of 18 additional firm-year observations.

We found 21 additional firm-year observations after searching for AI adoption for the rest of the public firms (i.e., firms that are not covered by the GSL dataset nor disclose AI adoption in public disclosures) in news articles from Lexis Nexis and Factiva. The final sample consists of 117 firm-year observations (78, 18, and 21 observations, respectively). The sample construction process for the treatment firms is illustrated in Panel A of Table 19.

The types of AI technologies adopted in our treatment firms are provided in Panel B, Table 19. More than half of the treatment firms explicitly disclosed using ML as a specific AI technology in the firm's operations, consistent with the description of the AI technologies adopted by firms as described in the McKinsey survey (2021). Moreover, our treatment firms adopt AI technologies with the objectives to predict operational

[68] SIC codes that belong to "Office of Technology" were excluded: 3510, 3523, 3524, 3530, 3531, 3532, 3533, 3537, 3540, 3541, 3550, 3555, 3559, 3560, 3561, 3562, 3564, 3567, 3569, 3570, 3571, 3572, 3575, 3576, 3577, 3578, 3579, 3580, 3585, 3590, 4812, 4813, 4822, 4832, 4833, 4841, 4899, 7370, 7371, 7372, 7373, 7374. See: https://www.sec.gov/info/edgar/siccodes.htm.

performance, financial information, supply chain movements, and to automate business processes, all of which are to improve firms' internal operations (Panel C, Table 19).

Most of the treatment firms in our sample implemented AI in firm operations in 2017 and 2018 (Panel D, Table 19). Most firms have started to implement or disclose AI in recent years, presumably because substantial improvements have been made to it and because it is becoming more affordable in recent years compared to decades ago (Agrawal et al., 2018; Brundage et al., 2018; Deloitte 2019). Moreover, the majority of our treatment firms are in trade and services (34%) and manufacturing (21%) industries (Panel E, Table 19), comparable with the survey results from McKinsey (2021).

Table 19. Sample Construction

Panel A: Sample Determination

GSL AI adoption firm list from 2014-2018 covered by COMPUSTAT (number of unique firms)	1423	
Less: not covered by IBES Guidance database from 2014-2018	-918	
Less: firms whose AI implementation year cannot be determined, or those who do not use AI to improve business operations	-379	
Less: firms in the technology industry	-28	
Less: missing dependent or control variables	-20	
GSL sample		**78**
Public disclosure of AI from financial reports/press release/conference calls from 2014-2018 covered by COMPUSTAT (number of unique firms)	734	
Less: duplicated firms from GSL	-163	
Less: not covered by IBES Guidance database from 2014-2018	-463	
Less: firms whose AI implementation year cannot be determined, or those who do not use AI to improve operations	-14	
Less: firms in the technology industry	-67	
Less: missing dependent or control variables	-9	
Financial reports/press release/conference calls sample		**18**
Firms with all financial data from COMPUSTAT and IBES Guidance but not matched to GSL or public disclosure of AI	436	
Less: firms in the technology industry	-71	
Less: firms whose AI implementation cannot be identified	-344	
Additional search from news articles		**21**
Total Treatment sample		**117**

Panel B: Types of AI Technologies Implemented by Treatment Firms

Types of AI	**Count**	**Example**
Machine Learning	77	(The Hershey Company) " …use **machine learning** to analyze previous purchases to identify customers with a propensity to purchase particular products and services. The company wanted to gain valuable, revenue-generating insight faster than a traditional analytics implementation could deliver."
Intelligent Automation/Robotics	22	(Aflac Inc.) "In addition to delivering better customer experience, the company wanted to cut costs by reducing call center volume and seeing better overall accuracy in responses; The company uses **RPA** to mimic and repeat activities humans typically perform. Some 20 software bots perform activities such as paying claims and processing invoices."
Natural Language Processing/Virtual Assistant/Chatbots	9	(Macy's Inc.) "Macy’s is testing a mobile tool using **artificial intelligence** that lets shoppers get answers customized to the store they’re in… [The tool] uses **natural language** and offers feedback in seconds… [The goal is to] boost sales while freeing up employees to focus on more complicated customer requests."
Image Recognition/Optical AI	3	(Avangrid Renewables) "This new system blends **artificial intelligence with high-precision optical technology** to identify protected avian species such as golden eagles, evaluate their flight paths and send an alert to shut down specific turbines if a collision risk is detected. Avangrid Renewables is evaluating the technology’s effectiveness and will potentially use the results to support continued development of effective avian risk management strategies."
Non-specified AI	20	(AES Corp.) "With this big operation, the company is moving towards **artificial intelligence** in energy and utilities to explore how it can improve the efficiency and maintenance of its electric grid systems... The company is creating a suite of data science tools to take advantage of a large amount of data they accumulate on a daily basis. The data comes from the power plants they manage, including solar energy farms, batteries, and gas plants. AI is required to make sense of all the data being acquired, and help the company improve the daily operations of the energy farms, either for more efficiency, produce more power or to lower the cost of operations."
Total	133*	

*Note: Some firms disclosed implementing more than one AI technology. Therefore, the total is higher than 117, the number of treatment firms.

Panel C: Objectives of AI Technologies Implemented by Treatment Firms

Objectives of AI	**Count**	**Example**
Predict internal operational performance	85	(Walmart, Inc.) "Walmart crafted a compelling business need for AI …Commodity processes generally include **accounts payable, accounts receivable general ledger and a bit of compensation and benefits**. The company can apply data science to ensure that an invoice is 100% accurate, paying the correct amount to the right invoice."
Automate business process	37	(Johnson and Johnson, Inc.) "Implementing Robotic Process Automation (RPA) in Johnson & Johnson's **finance department has resulted in standardization, improvements in workflow and greater accuracy**...Johnson & Johnson started with a pilot focused on automating aspects of intercompany requests, invoice creation and postings."
Predict financial information	12	(Rite Aid Corp) "...uses machine learning and AI to design intelligent tests. It incorporates specific brand inputs to adapt to **the pricing parameters, long purchase cycles and stock-up products** that can make promoting in drug stores challenging."
Predict supply chain movement	5	(BorgWarner, Inc.) "BorgWarner's Supplier Performance Monitor (SPM) system uses artificial intelligence and sophisticated mathematical theories to **analyze current supplier performance and forecast future trends**. The easy-to-install software can be transferred to new locations and **improves supply chain response times significantly.**"
Total	139*	

*Note: Some firms disclosed applying AI technologies to more than one use case. Therefore, the total is higher than 117, the number of treatment firms.

Panel D: Distribution of the Treatment Firms by Implementation Year

Implementation Year	**Count**	**Percentage**
2014	6	5.0%
2015	6	5.0%
2016	12	10.1%
2017	50	42.9%
2018	43	37.0%
Total	117	

Panel E: Distribution of the Treatment Firms by Industry

Industry	**Count**	**Percentage**
Trade and Services	40	34.19%
Manufacturing	25	21.37%
Life Sciences	21	17.95%
Energy and Transportation	15	12.82%
Finance	12	10.26%
Real Estate and Construction	4	3.42%
Grand Total	117	100.0%

We use population-based control firms as the baseline control group. Population controls are non-technology public firms that are not included in the GSL database, do not mention AI adoption in public disclosures, are not reported to adopt AI in news articles, and have all necessary dependent and control variables.

4.3.2 Difference-in-Differences (DID) Model Specification

We adopt a DID model to isolate the effect of AI adoption on the accuracy of management earnings forecasts:

$$\begin{aligned} ACCURACY\_AVE &= b_0 + b_1 TREATMENT * POST + b_2 TREATMENT + b_3 POST + b_4 HORIZON \\ &+ b_5 LNAT + b_6 ABSCHGROA + b_7 NEWEQUITY + b_8 ANALYST + b_9 LOSS \\ &+ b_{10} NEWS + b_{11} ORGCHG + b_{12} COMPLEXITY + b_{13} BIG4 + b_{14} EARNVOL \\ &+ b_{15} BETA + Industry\ Fixed\ Effects + Year\ Fixed\ Effecs \\ &+ \varepsilon \qquad (1) \end{aligned}$$

In model (1), TREATMENT equals 1 for treatment firms and 0 for control firms. POST equals 1 (or 0) if the firm-year observation is one year after (or before) AI implementation. The variable of interest is the interaction term, TREATMENT*POST. ACCURACY_AVE is the average accuracy of management guidance on annual earnings per share issued in one year (e.g., Ittner and Michels 2017; Heizman and Huang 2019). ACCURACY_AVE equals -1 multiplied by the average absolute value of the forecast errors scaled by prior year stock price, where the forecast error is the difference between the actual earnings per share and management forecast (Ajinkya et al. 2005; Dorantes et al. 2013).[69, 70] The greater the ACCURACY_AVE, the more accurate the management guidance. Therefore, we expect TREATMENT*POST to be positively associated with ACCURACY_AVE. We include control variables that can potentially affect management earnings forecasts identified in previous literature. Their descriptions and expected impact on ACCURACY_AVE are provided in the Appendix.

4.4 Results

4.4.1 Descriptive Statistics and Univariate Analysis

Table 20 presents the descriptive statistics for both treatment (TREATMENT=1) and control (TREATMENT=0) firms before (POST=0) and after (POST=1) AI adoption. Before AI adoption, although treatment firms issue slightly more accurate guidance than control firms, such difference becomes more economically and statistically significant after AI adoption. In terms of control variables, treatment firms are significantly larger in size, followed by more analysts, incur fewer losses, are more likely to be audited by Big 4 auditors, and are less volatile in terms of the market beta than control firms.

[69] When management forecast is a range estimation, we use the midpoint of the upper and lower values of the forecast range.

[70] About 95% of our sample firms issue more than 1 management earnings forecast in a given year and 81% of them issue management earnings forecast on a quarterly and more frequent basis.

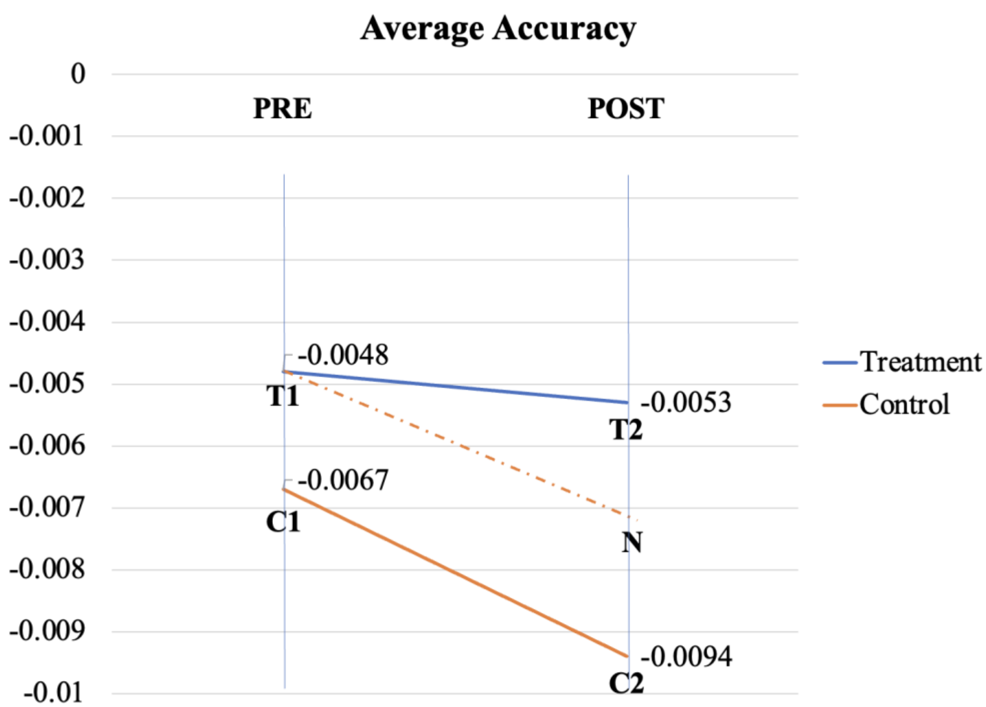
Table 20. Descriptive Statistics and Univariate Analysis

	POST=0					POST=1				
	TREATMENT=1		TREATMENT=0		Diff	TREATMENT=1		TREATMENT=0		Diff
	mean	std	mean	std	(Treatment - Control)	mean	std	mean	std	(Treatment - Control)
ACCURACY_AVE	-0.0048	0.0093	-0.0067	0.0118	0.0019*	-0.0053	0.0076	-0.0094	0.0173	0.0041***
HORIZON	5.1463	0.273	5.1391	0.2701	0.0072	5.0965	0.3186	5.0826	0.3222	0.0139
LNAT	9.3557	1.677	7.8721	1.4571	1.4836***	9.3038	1.7825	7.9147	1.4576	1.3891***
ABSCHGROA	0.0267	0.0435	0.0301	0.0471	-0.0034	0.0311	0.0401	0.0368	0.054	-0.0057
NEW_EQUITY	0.7788	0.417	0.7874	0.4094	-0.0086	0.7265	0.4477	0.7194	0.4495	0.0071
ANALYST	2.7066	0.4879	2.1015	0.7441	0.6051***	2.6066	0.6037	2.0113	0.7536	0.5953***
LOSS	0.0769	0.2678	0.0905	0.2871	-0.0136	0.0598	0.2382	0.1403	0.3475	-0.0805***
NEWS	0.6154	0.4889	0.5775	0.4943	0.0379	0.6068	0.4906	0.5552	0.4972	0.0516
ORGCHG	-0.1143	1.0659	-0.0973	1.2551	-0.017	0.1662	1.3008	-0.0318	1.3483	0.198
COMPLEXITY	0.2659	1.3676	0.1505	1.2725	0.1154	0.0078	1.3447	-0.0307	1.3206	0.0385
BIG4	0.9712	0.1682	0.8532	0.3541	0.118***	0.9658	0.1825	0.8478	0.3594	0.118***
EARNVOL	0.0121	0.016	0.0126	0.0137	-0.0005	0.0135	0.0155	0.0147	0.0153	-0.0012
BETA	0.9296	0.3212	1.0149	0.3442	-0.0853**	0.8847	0.3559	0.9829	0.3579	-0.0982***
Observations	104		729		833	117		1005		1122

Note: ***p < 0.01, **p < 0.05, *p < 0.10 (two-tailed tests). Variable descriptions are provided in the Appendix.

We plot a parallel trend of the accuracy of management earnings forecasts for treatment and control firms in Figure 13. Points T1 and C1 (T2 and C2) represent the average accuracy of management earnings guidance for treatment and control firms one year before (after) AI adoption, respectively. The dashed line is parallel to the line from C1 to C2, and it represents the counterfactual (hypothetical) trend of average accuracy of management earnings guidance had the treatment firms not adopted AI. Therefore, the distance between points N and C2 represents the inherent difference between treatment and control firms in the average accuracy of management earnings forecasts, and the treatment effect from AI adoption is represented as the difference between points T2 and N.

Figure 13. Parallel Trend Plot of Management Earnings Forecast Accuracy Before and After AI Adoption



4.4.2 Main Test of the Hypothesis

Table 21 presents the DID analysis results for the main hypothesis. The regression is estimated using ordinary least squares (OLS). We find that the adoption of AI is associated with more accurate management guidance subsequent to its adoption compared to the population-based control firms, providing support for the main hypothesis. The coefficient of TREATMENT*POST is positive and significant at the 5% level. The coefficient value 0.002 is equivalent to an improvement of 21.3% on the accuracy of management earnings forecasts for treatment firms compared to control firms to AI adoption.[71]

---

[71] For ACCURACY_AVE, the coefficient of TREATMENT*POST is 0.002. This represents a 21.3% percent improvement for treatment firms compared to the control group after the adoption of AI, considering that the mean ACCURACY_AVE for the control group is -0.0094 in the post-adoption period.

For our control variables, we find that firms that issue forecasts earlier (longer HORIZON), incur a loss (LOSS), and are more volatile (larger EARNVOL) tend to issue less accurate management guidance. In contrast, firms that are followed by more analysts (larger ANALYSTS) and issued new equity or new debt in the following year (NEWEQUITY) tend to issue more accurate earnings guidance. Collectively, the findings support the main hypothesis as AI implementation is associated with more accurate management guidance subsequent to its implementation.

Table 21. Main Test of the Hypothesis

		ACCURACY_AVE
	Expected sign	Coef. (t stats)
TREATMENT*POST	+	0.002**
		(1.72)
TREATMENT	?	0.000
		(0.30)
POST	?	-0.001***
		(-2.72)
HORIZON	-	-0.007***
		(-3.34)
LNAT	+	0.000
		(-0.27)
ABSCHGROA	-	-0.007
		(-0.56)
NEWEQUITY	+	0.002**
		(1.74)
ANALYST	+	0.003**
		(2.58)
LOSS	-	-0.012***
		(-4.24)
NEWS	+	0.001
		(0.81)
ORGCHG	-	0.001
		(1.42)
COMPLEXITY	-	0.000
		(0.35)
BIG4	+	0.000
		(0.17)
EARNVOL	-	-0.118**
		(-2.30)
BETA	-	0.001
		(0.46)
Industry Fixed Effects		Yes
Year Fixed Effects		Yes

Adj.R2		0.25
N		1955

Note: ***p < 0.01, **p < 0.05, *p < 0.10 (one-tailed tests when the coefficients have the predicted signs, two-tailed tests otherwise). We do not place any asterisk for one-tailed tests when the estimated sign is opposite to the predicted sign). T-statistics are based on standard errors adjusted for firm clustering effects. Variable descriptions are provided in the Appendix.

4.4.3 Direct and Indirect Impacts of AI on the Accuracy of Management Earnings Forecast

To explore whether the observed association between AI and more accurate management guidance is driven by AI's direct or indirect impact, we conduct a structural equation modeling (i.e., regression path analysis) (Huang et al., 2016; Laplante et al., 2021).

When AI is used in business operations, it is likely to improve operational performance, which in turn improves the accuracy of management earnings forecasts. As a result, AI may indirectly affect the accuracy of management earnings forecasts through firms' operational performance. Similarly, another possible mechanism through which AI may indirectly affect the accuracy of management earnings forecasts is financial reporting quality. For example, we would expect AI to improve the quality of accounting information, which then improves the accuracy of management earnings forecasts. Firms' operational performance is measured by return on assets (ROA), calculated as earnings before extraordinary items scaled by total assets (Brown and Caylor 2009). The higher the ROA, the greater the firm performance. Financial reporting quality (FRQ) is measured by -1 multiplied by the absolute value of discretionary accruals which are the residuals from the industry and year-specific regression models for the performance-adjusted discretionary accruals (Hope et al., 2013). The larger the FRQ, the higher the financial reporting quality. We expect the indirect effects to be positive for both ROA and FRQ.

Table 22 presents the path analysis results.[72] The rows of interest in Table 22 are the direct effect, which represents AI's direct impact on the accuracy of management earnings forecast, and the indirect effect, which represents the indirect effect of AI through ROA or FRQ on the accuracy of management earnings forecasts. Panel A, Table 22 shows that AI's direct improvement to the accuracy of management earnings forecasts is significant at the 1% level and that the indirect effect via ROA is significant at the 5% level. The indirect effect through ROA makes up 15.6% of AI's total effect on the accuracy of management earnings forecasts, while AI's direct effect takes up the majority of 84.3%.[73] Panel B, Table 22 shows no statistically and economically significant indirect impact of AI on the accuracy of management earnings forecast via FRQ. In sum, our findings suggest that while AI can indirectly improve the accuracy of management earnings guidance via enhancing ROA, such indirect effect is small compared to AI's direct improvement to the accuracy of management earnings forecasts.

Table 22

Panel A. AI Adoption's Effect Decomposition with ROA as a Mediator

	ACCURACY_AVE		
Effect	**Expected sign**	**Coeff.**	**z score**
Total effect	+	0.0032***	3.03
Direct effect (AI adoption -> Management Earnings Forecast Accuracy)	+	0.0027***	2.64
Indirect effect (AI adoption -> ROA -> Management Earnings Forecast Accuracy)	+	0.0005**	1.97

[72] In the path analysis, AI is measured by the variable TREATMENT. The path analysis results are similar if we use TREATMENT*POST instead of TREATMENT.
[73] For ACCURACY_AVE, the indirect effect is 0.0005, equivalent to 15.6% of the 0.0032 total effect; the direct effect is 0.0027, equivalent to 84.3% of the 0.0032 total effect.

AI adoption -> ROA	+	0.0109**	2.00
ROA -> Management Earnings Forecast Accuracy	+	0.0442***	10.38
Observations		1950	

Note: ***p < 0.01, **p < 0.05, *p < 0.10 based on one-tailed tests. The "Direct effect" represents AI's direct effect on management earnings forecast accuracy after controlling for ROA (coefficient 0.0027). The "Indirect effect" is AI's impact on management earnings forecast accuracy through ROA (coefficient 0.0005), which is a product of AI's impact on ROA (coefficient 0.0109) and ROA's impact on management earnings forecast accuracy (coefficient 0.0442) (i.e., 0.0005=0.0109× 0.0442). The "Total effect" includes both AI's direct and indirect effects on management earnings forecast accuracy (i.e., 0.0032 = 0.0027 + 0.0005). In this path analysis, AI adoption is measured by the variable TREATMENT. Results are similar if TREATMENT*POST is used. Variable descriptions are provided in the Appendix.

Panel B. AI Adoption's Effect Decomposition with FRQ as a Mediator

	ACCURACY_AVE		
Effect	**Expected sign**	**Coeff.**	**z score**
Total effect	+	0.0039***	2.65
Direct effect (AI adoption -> Management Earnings Forecast Accuracy)	+	0.0039***	2.64
Indirect effect (AI adoption -> FRQ -> Management Earnings Forecast Accuracy)	+	0.0000	0.34
AI adoption -> FRQ	+	0.0080	0.82
FRQ -> Management Earnings Forecast Accuracy	+	0.0016	0.38
Observations		1125	

***p < 0.01, **p < 0.05, *p < 0.10 based on one-tailed tests. The "Direct effect" represents AI's direct effect on management earnings forecast accuracy after controlling for FRQ (coefficient 0.0039). The "Indirect effect" is AI's impact on management earnings forecast accuracy through FRQ (coefficient 0.0000), which is a product of AI's impact on FRQ (coefficient 0.0080) and FRQ's impact on management earnings forecast accuracy (coefficient 0.0016) (i.e., 0.0000=0.0080× 0.0016). The "Total effect" includes both AI's direct and indirect effects on management earnings forecast accuracy (i.e., 0.0039 = 0.0039 + 0.0000). In this path analysis, AI adoption is measured by the variable TREATMENT. Results are similar if TREATMENT*POST is used. Variable descriptions are provided in the Appendix.

4.4.4 Moderation Effects: Information Availability and Firm's Earnings Uncertainty

To add another perspective to the main test of the hypothesis, we examine whether AI is more useful in situations where it is more difficult to predict. As such, we consider the moderating effects of information availability and firm earnings uncertainty.

AI models have the ability to learn from large volumes of historical data and identify hidden patterns, producing more accurate forecasts than conventional methods (Alpaydin 2014; Agrawal et al. 2018). In the context of producing earnings guidance, information uncertainty decreases managers' ability to issue more accurate earnings predictions (Rogers and Stocken 2005). Information uncertainty can be present when managers have limited information sets, such as when forecasts are issued with a longer horizon, i.e., when the time between the forecast and the realization of the forecast is longer. The length of the horizon captures the availability of earnings information of the target fiscal year. That is to say, the longer the horizon, the less earnings information of the target fiscal year that is available, and thus, the higher the uncertainty to generate the earnings forecast. For ease of reference, we refer to this source of uncertainty as "information availability uncertainty." Since AI is able to incorporate more historical data points and better identify hidden patterns than conventional methods to generate predictions (e.g., Shim and Pourhomayoun 2017; Syam and Sharma 2018; Pavlyshenko 2019),[74] it may reduce the information availability uncertainty associated with issuing forecasts with longer horizons. Thus, we expect the effect of AI on the accuracy of management earnings forecasts to be more pronounced for guidance with a longer horizon than that with a shorter horizon. We use model (2) to test the moderating effect of information availability uncertainty measured by horizon on the accuracy of management earnings forecast.

[74] For example, AI has the capability to process large amounts of historical data whereas conventional heuristic methods leverage less data, such as prior year, or few years of data to produce forecasts.

$$
\begin{aligned}
ACCURACY = \; & b_0 + b_1 TREATMENT * POST * HORIZON + b_2 TREATMENT + b_3 POST \\
& + b_4 HORIZON + b_5 LNAT + b_6 ABSCHGROA + b_7 NEWEQUITY \\
& + b_8 ANALYST + b_9 LOSS + b_{10} NEWS + b_{11} ORGCHG + b_{12} COMPLEXITY \\
& + b_{13} BIG4 + b_{14} EARNVOL + b_{15} BETA + Industry\ Fixed\ Effects \\
& + Year\ Fixed\ Effecs + \varepsilon \qquad (2)
\end{aligned}
$$

In model (2), the dependent variable is the accuracy of management earnings forecasts with the longest and the shortest horizons issued in a year to demonstrate AI's ability to reduce information uncertainty. The testing variable is an interaction term of TREATMENT*POST*HORIZON.[75] HORIZON is the natural logarithm of the number of days between the date when the guidance was made and the fiscal year-end of the actual annual EPS. The larger the HORIZON, the longer the time lag between the guidance date and the fiscal year-end. We expect TREATMENT*POST*HORIZON to be positively associated with ACCURACY. Panel A of Table 23 presents the results. Consistent with our expectation, we find that AI significantly improves the accuracy of guidance that has longer HORIZON than that with shorter HORIZON, as reflected in the positive association between TREATMENT*POST*HORIZON and ACCURACY. The coefficient for TREATMENT*POST*HORIZON is 0.0004 and significant at 5% level, suggesting that a one-unit increase in HORIZON can lead to an additional 4.3% improvement of the accuracy of management earnings forecasts for treatment firms compared to control firms after the adoption of AI.[76, 77]

---

[75] In model (2), we do not include interaction terms TREATMENT*POST, TREATMENT*HORIZON, and POST*HORIZON to avoid multicollinearity since they are highly correlated (higher than 90%) with TREATMENT*POST*HORIZON.

[76] For ACCURACY_AVE, the coefficient of TREATMENT*POST*HORIZON is 0.0004. This represents a 4.3% percent improvement for treatment firms compared to control group after the adoption of AI, considering that the mean ACCURACY_AVE for the control group is 0.0094 in the post-adoption period.

[77] The results are similar if we replace HORIZON with a dummy variable indicating whether the forecast has the longest horizon.

Besides information availability uncertainty, we further explore another type of uncertainty originated from the volatility of firms' earnings (hereafter, "earnings uncertainty") and examine if AI can maintain its advantage in improving the accuracy of management earnings forecasts for firms with high earnings uncertainty using Model (3)

$$
\begin{aligned}
ACCURACY\_AVE
&= b_0 + b_1 TREATMENT * POST * VOL + b_2 TREATMENT * POST \\
&+ b_3 TREATMENT * VOL + b_4 POST * VOL + b_5 TREATMENT + b_6 POST \\
&+ b_7 VOL + b_8 HORIZON + b_9 LNAT + b_{10} ABSCHGROA + b_{11} NEWEQUITY \\
&+ b_{12} ANALYST + b_{13} LOSS + b_{14} NEWS + b_{15} ORGCHG + b_{16} COMPLEXITY \\
&+ b_{17} BIG4 + b_{18} BETA + Industry\ Fixed\ Effects + Year\ Fixed\ Effecs \\
&+ \varepsilon \qquad\qquad (3)
\end{aligned}
$$

We measure firms' earnings uncertainty by VOL, which is the decile of the standard deviation of quarterly earnings over 12 quarters ending in the current fiscal year, divided by the median asset value for the period. The higher the VOL, the more volatile the firms' earnings (i.e., higher uncertainty in relation to future earnings).[78] Similar to our expectation of how information uncertainty would moderate AI's impact on the accuracy of earnings guidance, we expect AI to improve the accuracy of management earnings forecasts more profoundly for firms with high earnings uncertainty compared to those with low earnings uncertainty. The results in Panel B of Table 23 provide weak evidence that AI improves guidance accuracy when earnings volatility is higher. The coefficient for TREATMENT*POST*VOL is 0.001 and significant at 10% level, suggesting that a one-unit increase in VOL can lead to an additional 10.6% improvement of the accuracy

[78] Alternatively, a dummy variable that indicates whether earnings volatility over 12 quarters ending in the current year is above the 10% percentile generates similar results.

of management earnings forecast for treatment firms compared to control firms after the adoption of AI.[79, 80]

Table 23

Panel A. The Moderation Effect of Information Uncertainty Measured by HORIZON

		ACCURACY
	Expected sign	Coef. (t stats)
TREATMENT*POST*HORIZON	+	0.0004**
		(1.81)
TREATMENT	?	0.000
		(0.09)
POST	?	-0.001***
		(-2.76)
HORIZON	-	-0.004***
		(-8.69)
CONTROLS		Yes
Industry Fixed Effects		Yes
Year Fixed Effects		Yes
Adj.R2		0.238
N		3782

Note: ***p < 0.01, **p < 0.05, *p < 0.10 (one-tailed tests when the coefficients have the predicted signs, two-tailed tests otherwise). The number of observations almost doubled compared to other analyses because instead of using the average management earnings forecast accuracy as the dependent variable, we use the accuracy of forecasts that have the longest and the shortest HORIZON so that we can better demonstrate AI's ability to reduce information uncertainty. Variable descriptions are provided in the Appendix.

Panel B. The Moderation Effect of Earnings Uncertainty Measured by VOL

		ACCURACY_AVE
	Expected sign	Coef. (t stats)
TREATMENT*POST*VOL	+	0.001*
		(1.62)
TREATMENT*POST	?	-0.003
		(-1.23)
TREATMENT*VOL	?	0.000
		(-0.78)
POST*VOL	?	-0.000**
		(-2.29)
TREATMENT	?	0.002
		(1.08)
VOL	?	0.000
		(-0.63)

[79] For ACCURACY_AVE, the coefficient of TREATMENT*POST*VOL is 0.001. This represents a 10.6% percent improvement for treatment firms compared to control group after the adoption of AI, considering that the mean ACCURACY_AVE for the control group is 0.0094 in the post-adoption period.
[80] The results are similar if we replace VOL with a dummy variable indicating whether the earnings volatility is above the 90th percentile.

POST	?	0.001
		(1.16)
CONTROLS		Yes
Industry Fixed Effects		Yes
Firm Fixed Effects		Yes
r2_a		0.248
N		1950

Note: ***p < 0.01, **p < 0.05, *p < 0.10 (one-tailed tests when the coefficients have the predicted signs, two-tailed tests otherwise). Variable descriptions are provided in the Appendix.

Together, the findings in this section suggest that AI's improvement to the accuracy of management earnings forecast is amplified when there is higher information availability uncertainty.

4.4.5 Types of AI Technologies and Management Earnings Forecast Accuracy

To further disaggregate the impact of AI on earnings guidance, we examine the specific type of AI technology (or technologies) that is (are) driving the observed benefits of AI to the accuracy of management earnings forecasts. As such, as an extension to the main test of the hypothesis, we partition AI technologies by types with a focus on comparing ML versus other types of AI. This way of partitioning is consistent with the channels via which AI may improve the accuracy of management earnings forecasts: better prediction and better data quality. In particular, better prediction is mainly achieved by ML since, by design, ML is developed to produce more accurate predictions. Better data quality is mainly achieved from other types of AI (i.e., RPA, NLP, computer vision) via task automation. Moreover, specific types of AI technologies are used for different objectives. ML technologies are often used for the objective of predicting internal operational performance, financial information, and supply chain movement. In contrast, other AI technologies (i.e., RPA, computer vision, and NLP) are often used to automate business processes. We expect ML to improve the accuracy of management earnings forecasts more profoundly than other AI technologies because ML's inherent function to generate accurate predictions for operations, financial information, and supply chain has a natural linkage with the accuracy of management earnings forecasts. We use Model (4) to examine the impact of ML and other AI technologies on the accuracy of management earnings forecasts.

$$
\begin{aligned}
ACCURACY\_AVE \\
&= b_0 + b_1 ML * POST + b_2 OTHERAI * POST + b_3 ML + b_4 OTHERAI \\
&+ b_5 POST + b_6 HORIZON + b_7 LNAT + b_8 ABSCHGROA + b_9 NEWEQUITY \\
&+ b_{10} ANALYST + b_{11} LOSS + b_{12} NEWS + b_{13} ORGCHG + b_{14} COMPLEXITY \\
&+ b_{15} BIG4 + b_{16} EARNVOL + b_{17} BETA + Industry\ Fixed\ Effects \\
&+ Year\ Fixed\ Effecs + \varepsilon \qquad (4)
\end{aligned}
$$

ML is a binary variable that equals 1 if the treatment firm adopts ML. OTHERAI is a binary variable that equals 1 if the treatment firms adopt any other AI technologies.[81] Table 24 presents the results for Model (4). Consistent with our expectation, ML is the driving factor of AI's improvement to the accuracy of management earnings forecasts, suggesting that the main channel through which AI improves the accuracy of management earnings forecast is via better prediction than better data quality.

[81] ML and OTHERAI are not mutually exclusive: firms can adopt both ML and other types of AI.

Table 24. Types of AI Technologies and Management Earnings Forecast Accuracy

		ACCURACY_AVE
	Expected sign	Coef. (t stats)
ML*POST	+	0.003**
		(1.66)
OTHERAI*POST	+	0.001
		(0.81)
ML	?	0.000
		(0.09)
OTHERAI	?	0.001
		(0.58)
POST	?	-0.001***
		(-2.72)
CONTROLS		Yes
Industry Fixed Effects		Yes
Year Fixed Effects		Yes
Adj. R2		0.249
N		1950

Note: ***p < 0.01, **p < 0.05, *p < 0.10 (one-tailed tests when the coefficients have the predicted signs, two-tailed tests otherwise). Variable descriptions are provided in the Appendix.

Overall, our main test of the hypothesis shows that AI adopters generate more accurate earnings guidance than control firms after the adoption of AI. We further find that the majority of this observed association is from the direct impact of AI on the accuracy of management earnings forecasts rather than the indirect effect through firms' improvements to operational performance. Moreover, we find that the improvement of AI on the accuracy of guidance is more pronounced when the forecast horizon is longer, suggesting that AI can reduce information uncertainty in producing earnings guidance. Finally, we show that ML is the main contributor to AI's improvement in management earnings forecasts' accuracy.

4.5 Other Aspects of Management Guidance

In our main hypothesis, we examine the association between AI and the accuracy of management guidance due to the natural linkage between AI's functionalities to produce accurate forecasts. However, it is plausible that AI affects other aspects of management guidance, such as precision, frequency, and bias.

Precision is another important property of management guidance (e.g., Cheng et al. 2013). Management guidance might be issued as point forecasts or range forecasts. Prior studies find that management forecasts issued as point estimates or as range forecasts that are narrow in width capture managers' beliefs about the future and, in turn, impact the beliefs of investors and analysts about managers' certainty about the future (Baginski and Hassell 1997). Feng et al. (2009) find that managers who have access to lower quality internal information are more likely to issue less specific forecasts. More recently, Ittner and Michels (2017) find that firms with more advanced risk-based forecasting and planning practices issue narrower forecasts. Similar to Feng et al. (2009), Chen et al. (2018) study a setting of internal information asymmetry and found that firms with lower information quality are more likely to issue less specific forecasts. As a result, it is plausible that higher-quality management reports produced by AI systems would affect the precision of earnings guidance.

Besides guidance precision, prior work also suggests that guidance frequency reflects disclosure transparency (Bhojraj et al. 2011; Wang and Tan 2012). Managers who act in the firm's best interest are expected to issue more frequent forecasts (Ajinkya et al. 2005). Furthermore, to issue more frequent earnings guidance, managers need to reconcile a greater amount of information on a more timely basis (Wang and Tan 2012). Since AI can more efficiently process and generate relevant information than conventional methods, we expect AI adoption to increase the frequency of guidance issuance.

Another property of earnings guidance is the extent to which the guidance is optimistically biased. Optimistically or upward biased earnings guidance reflects managers' optimistic views to the investors about firms' earnings performance (Hurwitz 2018). However, such an optimistic view can sometimes be due to managers' overconfidence or overestimation of earnings performance (e.g., Hribar and Yang 2016). Thus, prior literature considers high-quality earnings guidance to have less optimistically biased earnings guidance. For example, Ajinkya et al. (2005) find that firms with more outside directors and greater institutional ownership issue less optimistically biased earnings forecasts. Since we expect AI to improve the quality of information that flows into earnings guidance, AI adopters will likely issue less optimistically biased earnings guidance.

Following prior literature, we measure precision (PRECISION) using a binary variable that equals 1 if the forecast is a point forecast and 0 if the forecast is a range forecast (Dorantes et al. 2013). When a firm issues multiple forecasts in a given year, we use the average of point estimates throughout the year (Dorantes et al., 2013). The greater the PRECISION, the more precise the management earnings guidance.[82] The greater the PRECISION, the more precise the management earnings forecast. FREQUENCY is the total number of forecasts issued by a firm in one year (Ajinkya et al. 2005). BIAS is the difference between guidance value and the realized EPS value scaled by the prior year's

[82] We also adopted an alternative measure of precision, PRECISION_ALT, which equals -1 multiplied by the average of the absolute value of the difference between the upper and lower values of the management earnings forecast scaled by prior year stock price (Ittner and Michels 2017). The greater the PRECISION_ALT, the more precise the management earnings guidance. The results are similar when PRECISION_ALT is used instead of PRECISION.

stock price (Ajinkya et al. 2005). A positive value of BIAS indicates optimistically biased management guidance (Ajinkya et al. 2005).

Table 25 presents the results of the impact of AI on PRECISION, FREQUENCY, and BIAS. Overall, we do not find evidence that AI significantly impacts management guidance's precision, frequency, or bias significantly. One possible explanation for the null finding for precision and frequency is that these forecast properties are primarily affected by firms' earnings guidance policies prescribing the types (point vs. range) and frequency of guidance issuance. Prior literature finds that such earnings guidance policy is usually "sticky" and is rarely changed over a long period since altering the guidance policy is costly (Graham et al. 2005; Einhorn and Ziv 2008; Tang 2013). Another potential explanation for the null finding on precision is that there might be a trade-off between precision and accuracy since a forecast with a high level of precision is more likely to be wrong, and, thus, erroneous (Hayward and Fitza 2017). Lastly, the null result of AI's impact on the direction of guidance (bias) suggests that AI does not directionally bias earnings guidance.

Table 25. AI Adoption and Other Properties of Management Earnings Forecast

	PRECISION		FREQUENCY		BIAS	
	Expected sign	Coef.	Expected sign	Coef.	Expected sign	Coef.
		(t stat)		(t stat)		(t stat)
TREATMENT*POST	+	-0.007	+	0.120	-	-0.011
		(-0.23)		(0.48)		(-0.56)
TREATMENT	?	0.006	?	-0.172	?	0.008
		(0.21)		(-0.66)		(0.82)
POST	?	0.004	?	-0.057	?	0.001
		(0.54)		(-0.94)		(0.75)
CONTROLS		Yes		Yes		Yes
Industry Fixed Effects		Yes		Yes		Yes
Firm Fixed Effects		Yes		Yes		Yes
Adj. R2		0.093		0.212		0.047
N		1955		1955		1950

Note: ***p < 0.01, **p < 0.05, *p < 0.10 (one-tailed tests when the coefficients have the predicted signs, two-tailed tests otherwise). Variable descriptions are provided in the Appendix.

4.6 Robustness Tests

To examine the robustness of our main conclusion that AI is associated with more accurate management guidance, we conduct various sensitivity tests, including the use of an alternative measure of management earnings forecast accuracy, adoption of selection models to reduce selectivity bias from AI adoption, and guidance issuance, construction of propensity score-matched control firms, entropy balancing, examination of the impact of AI on the change of management earnings forecast accuracy, controlling for a possible omitted variable (enterprise system adoption), and using a firm fixed effect model. Overall, our main conclusion is robust against these sensitivity tests.

4.6.1 Alternative Measure of the accuracy of management earnings forecast

We adopt an alternative measure of the accuracy of management earnings forecasts, ACCURACY_AVE_ALT. Unlike ACCURACY_AVE, which is scaled by prior year stock price, ACCURACY_AVE_ALT is scaled by asset per share (Feng et al. 2009). Table 26 presents the results for the main hypothesis. We find that AI leads to more accurate guidance after its adoption using this alternative measure. In Table 26, the coefficient of TREATMENT*POST is positive and significant at the 1% level, and the economic improvement is about 36.1%.

Table 26. Alternative Measure of Management Earnings Forecast Accuracy

		ACCURACY_AVE_ALT
	Expected sign	Coef. (t stats)
TREATMENT*POST	+	0.003***
		(2.69)
TREATMENT	?	-0.003***
		(-2.68)
POST	?	-0.001**

		(-2.43)
CONTROLS		Yes
Industry Fixed Effects		Yes
Year Fixed Effects		Yes
Adj.R2		0.332
N		1850

Note: ***p < 0.01, **p < 0.05, *p < 0.10 (one-tailed tests when the coefficients have the predicted signs, two-tailed tests otherwise). Variable descriptions are provided in the Appendix.

4.6.2 Selection Models

There are two sources of selectivity bias in our main analysis: firms' choice to adopt AI and firms' decision to disclose earnings guidance. Below we describe the application of selection models to reduce such selectivity bias.

Firms' Choice to Adopt AI

Firms that choose to adopt AI might be inherently superior to other firms in unobserved factors like firm culture and manager motivation, which may enable the treatment firms to generate more accurate earnings guidance. Thus, the observed association between AI adoption and more accurate earnings guidance might be positively biased from these omitted unobservable factors associated with AI adoption and management earnings forecast accuracy.

We identify an instrument variable that can directly impact AI adoption but has little direct impact on management earnings forecast accuracy. This instrument variable is the ranking of AI skill penetration (AISKILLRANK) for US regions based on LinkedIn data aggregated from 2015 to 2018, and it was developed by the AI Index Report Steering Committee from Stanford University (Perrault et al. 2019). AI skill penetration measures the concentration of AI skills among top skills in each city added by LinkedIn members.[83] The details of how the AI skill penetration is measured can be found in Perrault et al. (2019).

The smaller the ranking number for a region, the higher the penetration rate of AI skills in that region.[84] We assume that the higher the AI skill penetration rate (i.e., the smaller the ranking number) for a region where a firm is headquartered, the more likely this firm will adopt AI due to the abundance of AI talent supply. We model firms' choice to adopt AI using the following first-stage selection model.

$$\begin{aligned} TREATMENT_{it} = {} & b_0 + b_1 AISKILLRANK_i + b_2 CONCENTRATION_{it} + b_3 LOGSALE_{it} \\ & + b_4 UNCERTAINTY_{it} + b_5 ABSCHGROA_{it} + b_6 NEWEQUITY_{it} \\ & + b_7 ANALYST_{it} + b_8 LOSS_{it} + b_9 NEWS_{it} + b_{10} ORGCHG_{it} \\ & + b_{11} COMPLEXITY_{it} + b_{12} BIG4_{it} + b_{13} EARNVOL_{it} + b_{14} BETA_{it} \\ & + Industry\ Fixed\ Effects + Year\ Fixed\ Effects \\ & + \varepsilon \qquad (5) \end{aligned}$$

We include control variables, CONCENTRATION, LOGSALE, and UNCERTAINTY, identified by previous literature to impact IT implementation (Kobelsky et al. 2008; Pincus et al. 2017). The definitions of these variables are provided in Appendix. Following the guidance from Lennox et al. (2012), we also include control variables that we will include in the second-stage model (ABSCHGROA, NEWEQUITY, ANALYST, LOSS, NEWS, ORGCHG, COMPLEXITY, BIG4, EARNVOL, and BETA). The definitions of these variables are also presented in Table A1 of the Appendix. Table A4 in Appendix presents the results for the first-stage selection model of AI implementation.

---

[83] The AI skill ranking is based on LinkedIn data aggregated from 2015 to 2018. Therefore, in our analysis, we treat this ranking as a time-invariant variable during our sample period of 2014-2018. In untabulated analysis, our main findings still hold if we use a sample period of 2015-2018.

[84] For example, Bryan College Station in Texas is ranked 1, and it has the highest relative AI skill penetration in the country, followed by San Francisco Bay Area, Lafayette, Indiana, Binghamton, New York, and Urbana-Champaign, Illinois.

As expected, AISKILLRANK is significantly and negatively associated with AI implementation after controlling for other factors. Below we list reasons that support AISKILLRANK being an appropriate instrument variable for AI adoption, following Larcker and Rusticus (2010) and Lennox et al. (2012). First, AISKILLRANK is an exogenous factor as it reflects the supply of AI talents in a US region, and it is not likely to be affected by firms' endogenous choices. Second, our first-stage model (model (5) as shown in Table A4) provides evidence that AISKILLRANK is a significant determinant of AI implementation after controlling for other factors. In addition, AISKILLRANK is not significantly associated with management forecast accuracy. In un-tabulated analysis, the association between AISKILLRANK and management forecast accuracy is insignificant, with a two-sided p-value above 0.8.

The pseudo R squared of the model (5) is 0.286. Besides AISKILLRANK having a significant and expected association with AI adoption, we also observe that higher sales (LOGSALE), issuing new equity or debt (NEWEQUITY), more analysts following (ANALYSTS), a greater current year EPS than the previous year's (NEWS), and higher levels of firm complexity (COMPLEXITY) are associated with a higher propensity of AI adoption. Based on this selection model of AI implementation, we calculate the Inverse Mill's Ratio for AI implementation (IMR_AI) and include it in the following second-stage model.[85]

[85] In estimating the second-stage model, population controls were used with entropy balancing to reduce inherent differences between control and treatment firms. The results are similar without including entropy weights.

$$
\begin{aligned}
ACCURACY\_AVE_{it+1} \\
= \ & b_0 + b_1 TREATMENT_{it} + b_2 HORIZON_{it+1} + b_3 LNAT_{it+1} \\
& + b_4 ABSCHGROA_{it+1} + b_5 NEWEQUITY_{it+1} + b_6 ANALYST_{it+1} + b_7 LOSS_{it+1} \\
& + b_8 NEWS_{it+1} + b_9 ORGCHG_{it+1} + b_{10} COMPLEXITY_{it+1} + b_{11} BIG4_{it+1} \\
& + b_{12} EARNVOL_{it+1} + b_{13} BETA_{it+1} + Industry\ Fixed\ Effects \\
& + Year\ Fixed\ Effects + IMR\_AI_{it} + \varepsilon \qquad (6)
\end{aligned}
$$

Firms' Choice to Disclose Guidance

To address the concern about the selection bias from firms' decisions to issue earnings guidance, we model firms' choice to issue management earnings forecasts following the prior literature (e.g., Ajinkya et al. 2005; Feng et al. 2009; and Li et al. 2012).

$$
\begin{aligned}
DISCLOSE_{it} = \ & b_0 + b_1 LOGSALE_{it} + b_2 ABSCHGROA_{it} + b_3 NEWEQUITY_{it} + b_4 ANALYST_{it} \\
& + b_5 LOSS_{it} + b_6 NEWS_{it} + b_7 ORGCHG_{it} + b_8 COMPLEXITY_{it} + b_9 BIG4_{it} \\
& + b_{10} EARNVOL_{it} + b_{11} BETA_{it} + b_{12} STD\_AF_{it} + b_{13} FINCHA_{it} \\
& + Industry\ Fixed\ Effects + Year\ Fixed\ Effects + \varepsilon \qquad (7)
\end{aligned}
$$

The instrument variable we use for management earnings guidance disclosure is the natural logarithm of sales (LOGSALE).[86] Table A5 in Appendix presents the selection model of management earnings disclosure. LOGSALE is an appropriate instrument for management earnings guidance disclosure because it is a significant determinant of management earnings guidance disclosure, as shown in Table A5 in Appendix. Second, in un-tabulated results, we find that, after controlling for other factors, it is not significantly associated with the accuracy of management earnings forecast (two-sided p-value of above 0.9).

The pseudo R squared of model (7) is 0.289. Model (7) shows that firms that have smaller changes in ROA from the previous year to the current year (ABSCHGROA),

[86] Unlike prior literature (e.g., Feng et al. 2009), we do use not the natural logarithm of the number of analysts following (ANALYST) as an instrument variable since ANALYST is significantly associated with management guidance in our sample and it is used as a control variable.

more analysts following (ANALYSTS), incur a loss (LOSS), have less dispersion in analysts forecasts (STD_AF), and less financial changes (FINCHA) are associated with a higher likelihood of issuing management earnings forecasts, consistent with findings from prior literature (e.g., Ajinkya et al. 2005; Feng et al. 2009; Dorantes et al. 2013). Based on this selection model of management earnings forecast disclosure, we calculate the Inverse Mill's Ratio for management earnings forecast disclosure (IMR_DISCLOSE) and include it in the following second-stage model.

$$\begin{aligned} ACCURACY\_AVE_{it+1} &= b_0 + b_1 TREATMENT_{it} + b_2 HORIZON_{it+1} + b_3 LNAT_{it+1} \\ &+ b_4 ABSCHGROA_{it+1} + b_5 NEWEQUITY_{it+1} + b_6 ANALYST_{it+1} + b_7 LOSS_{it+1} \\ &+ b_8 NEWS_{it+1} + b_9 ORGCHG_{it+1} + b_{10} COMPLEXITY_{it+1} + b_{11} BIG4_{it+1} \\ &+ b_{12} EARNVOL_{it+1} + b_{13} BETA_{it+1} + Industry\ Fixed\ Effects \\ &+ Year\ Fixed\ Effects + IMR\_DISCLOSE_{it+1} \\ &+ \varepsilon \qquad (8) \end{aligned}$$

Table 27 presents the results for the second-stage models that control for firms' choice to adopt AI and issue guidance. The first column in Table 27 shows the results after controlling for IMR_AI. We observe that treatment firms experienced an increase in management earnings forecast accuracy compared to control firms after controlling for other factors and IMR_AI. We observe similar results in the second column in Table 9 where IMR_DISCLOSE is controlled. The third column of Table 27 presents the results controlling for IMR_AI and IMR_DISCLOSE. Overall, we find consistent evidence that AI adoption is significantly associated with management earnings forecast accuracy after controlling for the selection bias of AI adoption and management earnings forecast disclosure.[87, 88]

Table 27. Second-stage Model Results after Controlling for Firms' Choice of AI Adoption and Management Earnings Forecast Disclosure

		DV=ACCURACY_AVE		
		(1)	(2)	(3)
	Expected sign	Coeff. (t stats)	Coeff. (t stats)	Coeff. (t stats)
TREATMENT	+	0.006***	0.003***	0.004***
		(3.26)	(2.43)	(2.87)
IMR_AI	?	0.001		0.003
		(0.17)		(0.51)
IMR_DISCLOSE	?		-0.035**	-0.036***
			(-2.46)	(-2.79)
CONTROLS		Yes	Yes	Yes
Industry Fixed Effects		Yes	Yes	Yes
Year Fixed Effects		Yes	Yes	Yes
Adj.R2		0.437	0.403	0.509
N		853	1094	828

Note: ***p < 0.01, **p < 0.05, *p < 0.10 (one-tailed tests when the coefficients have the predicted signs, two-tailed tests otherwise). Variable descriptions are provided in the Appendix.

4.6.3 Propensity Score Matching

The descriptive statistics table (Table 20) shows that treatment and control firms are inherently different in some control variables (i.e., LNAT, ANALYSTS, LOSS, BIG4, and BETA). To further reduce the concern that the inherent differences between treatment and control firms are biasing our main result, we construct an alternative control group based on PSM (propensity score matching). We generate propensity scores for AI implementation using a probit regression and then match control and treatment firms with propensity scores within 5% of the treatment firm, without replacement. The

[87] Another advantage of adopting a selection model is that it can reduce the potential bias from measurement errors of our test variable, TREATMENT. TREATMENT represents firms that adopted AI, and it captures both, firms that adopted AI and disclosed AI adoption. As a result, our selection model for AI adoption also captures disclosure of AI adoption (Angrist and Pischke 2008).
[88] As another sentitivity test, we incorporate the IMR_AI and IMR_DISCLOSE variables in the main DID model and find similar results.

appendix provides the detailed procedures of the PSM approach. Table 28 presents the results for the main hypothesis based on PSM controls with the DID specification. As shown in Table 28, we continue to find consistent evidence supporting the main hypothesis that AI is associated with improvements to the accuracy of earnings guidance. We also constructed PSM control firms based on alternative matching thresholds (one-to-one, 1%, and 3%), with or without replacement, and the results are similar (see Table A7 Appendix).

Table 28. Adopt Propensity Score Matched Controls

		ACCURACY_AVE
	Expected sign	Coef. (t stats)
TREATMENT*POST	+	0.002**
		(1.74)
TREATMENT	?	0.001
		(1.49)
POST	?	-0.001
		(-1.11)
CONTROLS		Yes
Industry Fixed Effects		Yes
Year Fixed Effects		Yes
Adj.R2		0.318
N		883

Note: ***p < 0.01, **p < 0.05, *p < 0.10 (one-tailed tests when the coefficients have the predicted signs, two-tailed tests otherwise). Variable descriptions are provided in the Appendix. We matched control and treatment firms that have propensity scores within 5% of the treatment firm, without replacement. Appendix presents results under alternative PSM design choices and the results are consistent.

4.6.4 Entropy Balancing

Another approach we adopt to reduce the potential bias from the inherent differences between treatment and control firms is entropy balancing. Entropy balancing is a multivariate method that preserves the full sample while reweighting each observation in the control group so that the means and variances of the covariates for the control sample and the treatment sample are balanced, i.e., not statistically different

(Hainmueller 2012; Hainmueller and Xu 2013). Unlike other matching methods like PSM, entropy balancing achieves a high degree of covariate balance, which helps to reduce the concern that potential imbalance in covariates across the control and treatment groups drives the results (Hainmueller 2012; Hainmueller and Xu 2013; Glendening et al. 2019). Besides, entropy balancing allows unit weights to vary smoothly across units compared to other matching methods like Propensity Score Matching (PSM), which assigns a binary “weight” to each control firm (i.e., match=1, unmatch=0).

We weigh each population control firm from the main analysis using the entropy balancing method based on data prior to AI adoption. After reweighting the observations via entropy balancing, the means and variances for the control variables are nearly identical between the treatment and control groups prior to AI adoption, with no significant differences across the groups. In untabulated results, we apply entropy balancing in conjunction with the DID design adopted in the main analysis and find that TREATMENT*POST is positive and significant at the 5% level with a coefficient value of 0.003.

4.6.5 AI Adoption and Change in Management Earnings Forecast Accuracy

The summary statistics in Table 20 and the parallel trend plot in Figure 13 show a decreasing trend of management earnings forecast accuracy. Another way to examine the effect of AI on management earnings forecast accuracy is to test whether AI adoption is associated with a smaller decrease in management earnings accuracy. Following Masli et al. (2010), we use model (9) to perform this analysis.[89]

[89] We use entropy balancing to reduce the inherent differences in control variables between treatment and control firms.

$$
\begin{aligned}
\Delta ACCURACY\_AVE \\
&= b_0 + b_1 TREATMENT + b_2 \Delta HORIZON + b_3 \Delta LNAT + b_4 \Delta ABSCHGROA \\
&+ b_5 \Delta NEWEQUITY + b_6 \Delta ANALYST + b_7 \Delta LOSS + b_8 \Delta NEWS \\
&+ b_9 \Delta ORGCHG + b_{10} \Delta COMPLEXITY + b_{11} \Delta BIG4 + b_{12} \Delta EARNVOL \\
&+ b_{13} \Delta BETA + Industry\ Fixed\ Effects + Year\ Fixed\ Effects \\
&+ \varepsilon \qquad (9)
\end{aligned}
$$

ΔACCURACY_AVE represents the decrease in management earnings forecast accuracy after AI adoption. It equals -1 multiplied by the percentage change of ACCURACY_AVE from the year of adoption to one year after AI adoption.[90] Since we expect to observe AI's benefits to management earnings forecast accuracy, we expect the parameter sign for TREATMENT to be negative, indicating that AI adoption is associated with a smaller decrease in management earnings forecast accuracy. Table 29 presents the results for model (7). Consistent with our expectations, we find evidence that AI adoption is associated with a smaller decrease in management earnings forecast accuracy.

Table 29. AI Adoption and Change in Management Earnings Forecast Accuracy

		ΔACCURACY_AVE
		Coeff.
		(t stat)
TREATMENT	-	-0.893***
		(-2.45)
ΔControls		Yes
Industry Fixed Effects		Yes
Year Fixed Effects		Yes
Adj. R2		0.231
N		945

Note: ***p < 0.01, **p < 0.05, *p < 0.10 (one-tailed tests when the coefficients have the predicted signs, two-tailed tests otherwise. Variable descriptions are provided in the Appendix.

[90] For continuous variables, the change term is -1 multiply by the difference between t+1 and t values scaled by the t value. For indicator variables, the change value is -1 multiply by the direct difference between t+1 and t values.

4.6.6 Control for Enterprise System Adoption

A potential correlated omitted variable that we can identify is firms' enterprise system adoption. First, enterprise system adoption could be positively correlated with AI adoption since AI complements existing large-scale enterprise systems (Gray 2018). As a result, firms might implement new AI when implementing a new enterprise system. Second, enterprise system adoption is associated with the accuracy of management earnings forecasts (e.g., Brazel and Dang 2008; Dorantes et al. 2013). Therefore, to control for the potential impact of enterprise adoption in our main conclusion, we performed a manual search of enterprise system adoption for our treatment firms and included enterprise system adoption as a control variable. We use the terms "enterprise system," "SAP," "Oracle," "information systems," and "relational databases" as keywords, followed by the words "implement" and "adopt" to identify the adoption of an enterprise system in conjunction with AI. We identified 16 firms that adopt enterprise systems and AI from this search. In untabulated analysis, we find that our main conclusion, that AI leads to more accurate management earnings forecasts, still holds.

4.6.7 Firm Fixed Effects

As another way to control for omitted variables that are time-invariant and unobservable, we adopt a firm fixed effects model. Untabulated analysis shows that while the coefficient values for TREATMENT*POST remain similar to the one reported in the main analysis, the inclusion of firm fixed effects reduces the statistical significance of the impact of AI on the accuracy of management guidance compared to the main test results. However, this result should be interpreted with caution due to the limitations of fixed effects groupings that do not have within fixed effects variation in the variable of interest

(Li and Prabhala 2007; Minutti-Meza 2013; deHaan 2021). Specifically, from our review of the sample, there is no within-firm variation concerning AI adoption (i.e., the variable TREATMENT). As a result, controlling for industry fixed effects is more appropriate than controlling for firm fixed effects in this research setting.

4.6.8 Placebo Analysis

To further mitigate the concern that the observed improvement of management earnings forecast accuracy from AI adoption comes from potential model misspecification, we perform a placebo analysis (Lim et al. 2020) where the AI implementation is artificially reversed three years before its actual implementation. In untabulated analysis, a dummy variable TREATMENT_PLACEBO represents the falsified AI adoption for the treatment firms three years before their actual implementation of AI (Lim et al. 2020). [91] If the actual AI adoption drives the improvement of management earnings forecast accuracy, we should not observe the significant effect of TREATMENT_PLACEBO*POST on management earnings forecast accuracy. Thus, any significant impact of TREATMENT_PLACEBO*POST on management's earnings forecast accuracy will undermine the findings from the main analysis. In un-tabulated results, we do not observe significant impacts of TREATMENT_PLACEBO*POST on the accuracy of management earnings forecast, lending support to our main findings that the improvement of management earnings forecast accuracy comes from the actual implementation of AI.

[91] In untabulated analysis, we used two years before AI implementation and find similar results.

4.7 Conclusion

We examine whether AI contributes to improvements in the accuracy of management earnings guidance, an important managerial outcome we expect AI to affect. We find evidence that earnings guidance is more accurate for AI adopters than non-AI adopters, consistent with the premise that AI can aid in the development of managerial outcomes. Further, we find that AI has an indirect effect on the accuracy of earnings guidance through firms' operational performance but that this effect is small compared to the direct effect of AI on accuracy. We also find that AI improves the accuracy of guidance that has a longer horizon, suggesting that AI reduces information uncertainty in producing earnings guidance, and that ML contributes the most to improving the accuracy of guidance compared to RPA, NLP, and computer vision. Last, we do not find evidence that AI affects precision, frequency, or bias in earnings forecasts. We subject our main findings to a series of sensitivity tests and find that they hold. Collectively, our findings provide early evidence that AI adoption is likely to improve the quality of earnings guidance.

Our study contributes to the accounting and accounting information system literature that designs AI systems to predict accounting balances and firm events (e.g., Ding et al. 2020), discussing the potential benefits of AI to accounting (e.g., Kokina and Davenport 2017). We do so by documenting initial archival evidence about the impact of the *actual* use of AI on the firm's information environment. We also contribute to the management forecast disclosure literature by providing insights into the earnings forecast properties that AI might affect since we find that AI influences the accuracy of earnings forecasts. Finally, our study contributes to the literature on the impact of the information environment quality on decision making by identifying a novel information technology-based determinant that might influence firms' internal information.

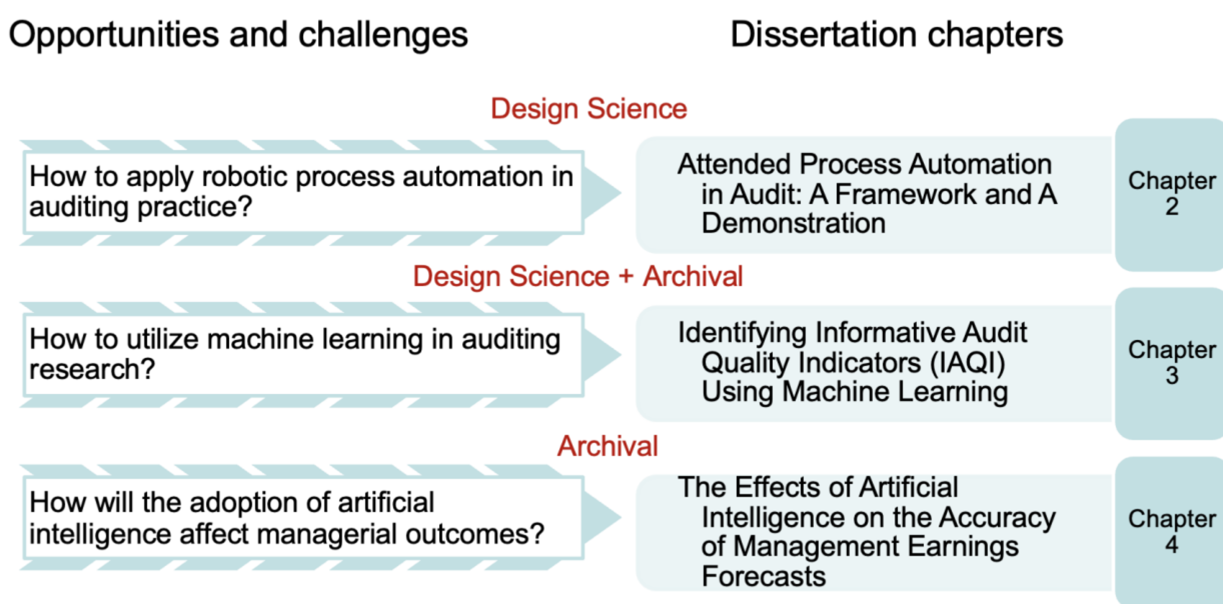
Our study is subject to the following limitations. First, the construction of our sample is based on firms' AI implementation data collected from a data provider, firms' public disclosures, and news mentions of using AI to improve firms' operations. The identification of AI adopters from three sources should provide a more comprehensive, complete, and accurate dataset. However, there may be additional AI adopters that do not disclose the adoption of AI. Future research could expand the sample size of AI firms. Finally, we find evidence regarding the benefits of AI to the quality (accuracy) of management earnings forecasts, but there are also unintended consequences that warrant examination. For example, data that is used to produce predictions can be biased. Future research is needed to examine the impact of such unintended consequences on management guidance.

CHAPTER 5: CONCLUSION

5.1 Summary

This dissertation adopts a diverse set of research mythologies to examine the prevailing issues regarding the usage of emerging technologies in accounting practice and research. Figure 14 summarizes how each dissertation chapter responds to specific opportunities and challenges in accounting research.

Figure 14. Summary of This Dissertation



Chapter 2 answers to the question "how to apply robotic process automation in auditing practice?" by proposing a framework that guides the implementation of attended RPA in audits. The proposed framework emphasizes auditors' vital role in an automated audit workflow in providing professional judgments that are currently irreplaceable by automation. Chapter 3 explores adopting ML in auditing research by using it to examine the out-of-sample predictive power of audit-related variables (i.e., ARV) for audit failure and identify which factors are the most predictive of audit failure. This study finds that

ARV have acceptable predictive power and that they outperform benchmark financial variables in predicting audit failure. The most predictive ARV reflect auditor competence, independence, effort, incentive, and the quality of the audited financial reports. Chapter 4 explores whether the implementation of AI in firms' operations is associated with improved accuracy of management earnings forecasts, an important managerial outcome. This study finds that AI is associated with more accurate management earnings forecasts after its implementation, that AI more profoundly improves management forecast accuracy when the forecast horizon is longer, and that ML is the primary AI technology that contributes to the improvements in management forecast accuracy.

5.2 Contributions

This dissertation extends the AIS literature by designing a framework that guides auditors to implement RPA in audits, utilizing machine learning to examine the out-of-sample predictive power of audit-related variables, and empirically examining the impacts of AI adoption on management earnings forecast accuracy.

Chapter 2 answers the call to develop more practice-relevant research in the accounting field by demonstrating how attended automation can be used in an actual audit setting. Furthermore, Chapter 2 extends RPA application research by exploring the task structure of the auditing tasks being automated. The proposed framework in Chapter 2 also provides researchers with a baseline to explore the implications of attended automation on auditor judgment quality. Lastly, Chapter 2 advances the understanding of the role of RPA in auditing and illustrates how auditors can work alongside automation to achieve human-machine synergy, which is especially needed as professional auditing judgments are still valuable and cannot be entirely replaced by automation (Sutton, Arnold, and Holt 2018; Zhang 2019).

Chapter 3 adds to the growing body of accounting research that adopts machine learning by examining whether ARV that are commonly used by academics have practical value in flagging audit failure. The finding that the predictive power of ARV has yet to exceed expectations from practitioners suggests a gap between theory and practice and, therefore, room for improvement. Chapter 3 also contributes to the stream of auditing literature that tries to validate audit quality measures by focusing on the perspective of the out-of-sample predictive power of audit quality measures. Furthermore, Chapter 3 provides tools that can be used in future research, including a predictive score (i.e., P Score) based on IAQI and advanced machine learning models that can capture non-linear relationships among data. Additionally, Chapter 3 shows that audit partners, audit managers, senior auditors, and auditing standard setters expect a model to have an average AUC of 83.2% for it to be useful in audit failure prediction. This threshold can be used as a benchmark for future studies that predict audit failure or material misstatement. Lastly, Chapter 3 responds to PCAOB's call for research on audit quality indicators using public source information (PCAOB 2015).

Chapter 4 adds to the interdisciplinary literature of AI and accounting by examining whether AI implementation impacts the accuracy of management guidance, an important managerial outcome. Second, Chapter 4 contributes to the management forecast disclosure literature by providing insights into the earnings forecast properties that AI might affect. Third, Chapter 4 contributes to the literature on the impact of the information environment quality on decision making by identifying an IT-based determinant that might influence firms' internal information. Finally, Chapter 4 should be of interest to managers as they consider the adoption of AI to improve their business operations.

5.3 Limitations

This dissertation is subject to several limitations. A limitation for Chapter 2 is that by the time this study was completed, the CPA firm that supported the demonstration case study had not completely rolled out attended automation in their audit practice. They needed more time to research applicable auditing standards and regulations associated with using automation in auditing. A limitation for Chapter 3 is the usage of MAR as a proxy for audit failure. The use of MAR may introduce measurement error because not all audit failures produce a material restatement (Gaynor et al. 2016; Suresh and Guttag 2019). However, this potential impact of measurement error is reduced by using a relatively large sample, as large samples make the parameters in a machine learning model converge towards the correct value (Suresh and Guttag 2019). A limitation for Chapter 4 is that the construction of our sample is based on firms' AI implementation data collected from a data provider, firms' public disclosures, and news mentions of using AI to improve firms' operations. Although AI adopters are identified from three complementary sources, there may be additional AI adopters that do not disclose the adoption of AI in public sources.

5.4 Future Research

Since Chapter 2 demonstrated the APA framework using only the audit planning of Single Audits, future research can apply the APA framework to other settings of external audits and highlight the specific circumstances in which the framework would be the most useful in auditing. Therefore, there is room for further research in this area that could empirically address the usefulness of this framework and its impact on auditor judgment and decision making in an experimental setting. Other topics to consider for future research include examining the challenges and potential risks (e.g., unintended consequences from bot errors) of RPA adoption in audits, the governance and deployment models of RPA in audits, and the cost and benefit analysis of RPA adoption in different audit settings is also of interest.

Chapter 3 finds that ARV have higher predictive power than benchmark financial variables, suggesting that future research should examine whether the external audit process or the clients' innate features are the driving factor for the observed audit failure. Moreover, since we find that raw audit variables have higher predictive power than either IAQI or ARV, future research can examine the source of incremental predictive power from raw audit variables. Chapter 3 also finds that the predictive power of audit-related variables has not exceeded practitioners' expectations. For audit failure prediction model to achieve practitioners' expectations, future research may explore whether adding variables collected from unconventional sources, such as social media, can increase the predictive power of audit failure prediction models. Furthermore, with the fast progress of machine learning, future research can explore whether novel machine learning algorithms can better model and predict audit failure. Chapter 3 uses material restatements as a proxy for audit failure. Future research may seek to adopt other proxies of audit quality, such as those outlined in the Part 1 Findings of the PCAOB inspections and proposed in audit firms' internal assessments of audit quality. Future research can also explore developing innovative measures of audit failure and utilize audit-related variables generated from alternative sources, such as social media and online forums.

Chapter 4 shows evidence regarding the benefits of AI to the accuracy of management earnings forecasts, but there could also be unintended consequences from AI adoption that warrant examination. Future research is encouraged to examine the impact of such unintended consequences on managerial outcomes.

APPENDIX

Appendix for Chapter 2

Table A1: Activity List of Single Audit Planning Provided by the CPA Firm

Expected manual (M) or automated (A)	Activity Description	File Type
A	Answer Step I Planning Worksheet questions based on the data in CY SEFA	Caseware doc
A	Answer Step II Planning Worksheet questions based on information available in other locations	Caseware doc
A	Populate Step III Planning Worksheet questions with current year federal award information available in CY SEFA	Caseware doc
A and M	Populate Row 12 of Federal Type A Risk Assessment Worksheet with current year Type A programs identified in previous step	Excel
A and M	Populate Row 9 of Federal Type B Risk Assessment Worksheet with current year Type B programs identified in previous step	Excel
A	Produce a report of major Type A and B programs, determine if coverage is sufficient	Caseware doc
A	Finish Planning Worksheet based on steps above	Caseware doc
A	Add a quality check that total of all sections on the form Historical Summary of Major Programs exceed the 20% (low risk) or 40% (not low risk) of total expenditure.	Caseware doc
A	Populate materiality worksheet based on major programs identified on the form Historical Summary of Major Programs	Excel
A	Populate historical summary of major programs based on major programs identified on the form Historical Summary of Major Programs	Excel
A	Include ability to repeat these steps as changes are made to DCF and produce an exception report with changes for the auditor to review.	Excel

Appendix for Chapter 3

Table A2. Comparison Between This Paper and the Most Related Literature

Literature	Objectives	Target Variable (Dependent Variable)	Predictors (Independent Variable)	Methodology	Evaluation Method
Cecchini et al. (2010)	1. Predict management fraud using basic financial data 2. Develop a kernel specific to the domain of finance	Fraud investigations enforced by the SEC and disclosed in AAER	Financial variables that have been used in fraud prediction	Predictive modeling (Support Vector Machine)	Out-of-sample evaluation (Recall rate, AUC)
DeChow et al. (2011)	1. Predict material misstatements 2. Develop a comprehensive database of financial misstatements	Material misstatements enforced by the SEC and disclosed on the AAER	Accrual quality Financial performance Nonfinancial measures Off-balance-sheet activities	Explanatory modeling (Regression)	In-sample evaluation (Recall rate, Specificity rate, AUC)
Perols (2011)	1. Identify which classification algorithms provide the most utility in predicting fraud 2. Identify the best prior fraud probability and misclassification cost when training classifiers 3. Identify useful predictors for classification algorithms	Fraud investigations enforced by the SEC and disclosed in AAER	Mostly financial variables that have been identified to be significant in fraud research	Predictive modeling (J48, SMO, Multilayer Perceptron, Logistics, stacking, and bagging)	Out-of-sample evaluation (Estimated Relative Costs, or ERC, of misclassification)
Perols (2016)	Introduce and evaluate three data analytics preprocessing methods to address challenges related to (1) the rarity of fraud observations, (2) the relative abundance of	Fraud investigations enforced by the SEC and disclosed in AAER	Mostly financial variables that have been identified to be significant in fraud research	Predictive modeling (Support Vector Machine)	Out-of-sample evaluation (Estimated Relative Costs, ERC, of misclassification)

	explanatory variables identified in the prior literature, and (3) the broad underlying definition of fraud.				
Dutta et al. (2017)	Predict restatements	All types of restatements from Audit Analytics database (fraudulent or erroneous, disclosed in all sources)	Financial variables that are related to fraud/restatement	Predictive modeling (Decision Tree, Artificial Neural Network, Naïve Bayes, Support Vector Machine, and Bayesian Belief Network)	Out-of-sample evaluation (Recall rate Specificity rate AUC)
Aobdia (2019)	Investigate the degree of concordance between fifteen measures of audit quality used in academia and two measures of audit process quality determined either by audit firms' internal inspections or by PCAOB inspections of individual engagements.	Part 1 Findings and internal inspection ratings from PCAOB	15 measures of audit quality used in academia	Explanatory modeling (Regression)	Statistical significance
Bao et al. (2020)	Predict accounting fraud using machine learning using financial data	Accounting frauds from SEC's AAER in CFRM database	28 raw financial variables from Dechow et al. (2011) and Cecchini et al. (2010)	Predictive modeling (Ensemble learning)	Out-of-sample evaluation (AUC and NDCG@k)
Brown et al. (2020)	Use a machine learning technique to assess whether the thematic content of financial statement	Accounting frauds from SEC's AAER in CFRM database Fraud-related restatements	F-score from Dechow et al. (2011) Thematic content variables (topic and style)	Use Bayesian topic modeling algorithm to determine and quantify the topic content of	Out-of-sample evaluation (AUC)

	disclosures is incrementally informative in predicting intentional misreporting.	from Audit Analytics database Fraud-related restatements from 10K/A		a large collection of 10-K narratives Use predictive modeling (logistic regression) to examine the incremental predictive power of thematic contents	
Bertomeu et al. (2020)	Use machine learning to predict material misstatements.	Material restatements disclosed in Form 8-K from Audit Analytics database	Variables from accounting, capital markets, governance, and auditing datasets	Predictive modeling (Gradient Boosted Regression Tree and other common algorithms)	Out-of-sample evaluation (AUC)
Rajgopal et al. (2021)	Provide detailed descriptive analyses of how poor audits are perceived in both public and private litigation settings. And evaluate how well existing audit quality proxies predict detailed allegations related to how auditors actually performed in specific engagements.	Audit deficiencies in specific engagements alleged by the SEC or private law firms	14 frequently used audit quality proxies	Explanatory modeling (Logistic Regression)	In-sample evaluation (Statistical significance, AUC)
This paper	Identify IAQI, which are theory-driven audit-related variables that are the most predictive of audit failure.	Material annual restatements due to GAAP violations or financial fraud disclosed in Form 8-K from Audit	Theory-driven audit-related variables	Predictive modeling (Random Forest, Artificial Neural Network, Support Vector Machine,	Out-of-sample evaluation (AUC)

		Analytics database		Logistic Regression, AdaBoost)	

Table A3. Pairwise Correlation of Audit-Related Variables

	Industry Specialization_ National	Industry Specialization_ MSA	Office Size	Big 4	New Client	Tenure	Local Auditor _MSA	Integrated Audit	Accelerated Filer	Busy
Industry Specialization_ National	1									
Industry Specialization_ MSA	0.4511*	1								
Office Size	0.6419*	0.2237*	1							
Big 4	0.7786*	0.4264*	0.7870*	1						
New Client	-0.1903*	-0.1085*	-0.2037*	-0.2267*	1					
Tenure	0.3581*	0.2320*	0.3397*	0.4066*	-0.4276*	1				
Local Auditor _MSA	0.1627*	0.0171*	0.2323*	0.2065*	-0.0556*	0.1105*	1			
Integrated Audit	0.4175*	0.2433*	0.5155*	0.5242*	-0.1733*	0.3735*	0.1582*	1		
Accelerated Filer	0.3927*	0.2194*	0.5196*	0.5189*	-0.1671*	0.3423*	0.1639*	0.9090*	1	
Busy	0.1148*	0.0694*	0.1307*	0.1295*	-0.0399*	0.0558*	0.0069	0.0613*	0.0620*	1
Workload Compression	0.0980*	0.1090*	-0.0155*	0.0891*	-0.0241*	0.0692*	0.0077	0.0387*	0.0348*	0.7846*
Auditor Competition_MSA	0.0889*	0.4197*	-0.1413*	0.0920*	-0.0190*	0.0797*	-0.1014*	0.0554*	0.0419*	0.0064
Auditor Resignation	-0.1060*	-0.0365*	-0.1420*	-0.1283*	0.0243*	-0.0506*	-0.0387*	-0.1023*	-0.0968*	-0.0149*
Audit Fees	0.6095*	0.3624*	0.7317*	0.6949*	-0.1983*	0.3999*	0.2263*	0.6324*	0.6131*	0.1037*
Tax Fee	0.3116*	0.2068*	0.3310*	0.3704*	-0.1456*	0.2333*	0.1199*	0.3133*	0.3092*	0.0023
Audit-Related Fee	0.3037*	0.2037*	0.3112*	0.3290*	-0.1078*	0.2046*	0.0955*	0.3035*	0.2811*	0.0371*
Other Fees	0.2101*	0.1061*	0.2235*	0.2356*	-0.0745*	0.1582*	0.0843*	0.1835*	0.1834*	0.0307*
Non-Audit Fee Ratio	0.0885*	0.0567*	0.0854*	0.1138*	-0.0340*	0.0673*	0.0417*	0.0791*	0.0763*	-0.0150*
Influence	-0.2498*	0.0433*	-0.6217*	-0.3299*	0.0874*	-0.0844*	-0.0995*	-0.1410*	-0.1458*	-0.0795*
Abnormal Audit Fee	0.1310*	0.0438*	0.2311*	0.1546*	-0.0722*	0.0299*	0.0554*	-0.0379*	-0.0166*	-0.0352*
Audit Report Lag	-0.3680*	-0.2241*	-0.4189*	-0.4222*	0.1761*	-0.3145*	-0.1435*	-0.4632*	-0.4522*	-0.0606*
Non-timely Issuance of 10K_Du	-0.0800*	-0.0448*	-0.0933*	-0.0947*	0.0771*	-0.0669*	-0.0564*	-0.0770*	-0.0768*	-0.0186*

e to Audit										
Going Concern	-0.3049*	-0.1879*	-0.4025*	-0.3654*	0.1174*	-0.1973*	-0.1668*	-0.3864*	-0.3824*	-0.0223*
Internal Control Weakness	0.0061	-0.0045	0.0346*	0.0177*	0.0337*	-0.0325*	0.0129*	0.1236*	0.1125*	-0.0077
Disc. Accruals	0.0130*	0.0173*	0.0144*	0.0176*	-0.0262*	0.0176*	0.0180*	0.0138*	0.0143*	0.0133*
Abs (Disc. Accruals)	-0.2044*	-0.1443*	-0.2674*	-0.2408*	0.0934*	-0.1478*	-0.1220*	-0.2425*	-0.2369*	-0.0217*
Abs (Accruals)	-0.1973*	-0.1250*	-0.2750*	-0.2391*	0.0827*	-0.1318*	-0.1387*	-0.2362*	-0.2306*	-0.0328*
Abs (Accruals/CFO)	-0.1153*	-0.0713*	-0.1340*	-0.1321*	0.0559*	-0.1013*	-0.0588*	-0.1340*	-0.1341*	-0.0009
DD Residual	-0.2407*	-0.1615*	-0.2948*	-0.2794*	0.0902*	-0.1654*	-0.1198*	-0.2777*	-0.2705*	-0.0272*
Prior ROA Meet	0.0353*	0.0338*	0.0349*	0.0362*	-0.0151*	0.0385*	0.0167*	0.0236*	0.0236*	0.0072
Small Profit	0.0036	0.0049	0.0252*	-0.0002	0.0142*	-0.008	0.0066	0.0194*	0.0141*	-0.0208*

(Continued)

	Workload Compression	Auditor Competition_MSA	Auditor Resignation	Audit Fees	Tax Fee	Audit-Related Fee	Other Fees	Non-Audit Fee Ratio	Influence	Abnormal Audit Fee
Workload Compression	1									
Auditor Competition_MSA	0.0768*	1								
Auditor Resignation	0.0046	0.0172*	1							
Audit Fees	0.1062*	0.0305*	-0.1283*	1						
Tax Fee	0.0036	0.0703*	-0.0597*	0.4464*	1					
Audit-Related Fee	0.0510*	0.0608*	-0.0588*	0.4725*	0.3328*	1				
Other Fees	0.0400*	0.0094	-0.0237*	0.2749*	0.2053*	0.1418*	1			
Non-Audit Fee Ratio	-0.0119	0.0402*	-0.0201*	0.1043*	0.5048*	0.4658*	0.2475*	1		
Influence	0.2409*	0.2477*	0.0770*	-0.1278*	-0.0246*	-0.0026	-0.0155*	0.0536*	1	
Abnormal Audit Fee	-0.0449*	-0.0893*	-0.0249*	0.3705*	0.0998*	0.0145*	0.0449*	-0.1673*	-0.0364*	1

Audit Report Lag	-0.0502*	-0.0723*	0.0866*	-0.4962*	-0.2609*	-0.2796*	-0.1731*	-0.1151*	0.1308*	0.0743*
Non-timely Issuance of 10K_Due to Audit	-0.0158*	-0.0240*	0.0131*	-0.0869*	-0.0522*	-0.0467*	-0.0261*	-0.0289*	0.0361*	0.0036
Going Concern	-0.0242*	-0.0529*	0.0793*	-0.4633*	-0.2224*	-0.2130*	-0.1261*	-0.0991*	0.1033*	0.0191*
Internal Control Weakness	-0.0089	-0.0137*	0.011	0.0776*	0.0085	0.0021	0.0149*	-0.0231*	0.0211*	0.0549*
Disc. Accruals	0.0212*	0.0245*	-0.0096	0.0128*	0.0073	-0.0082	0.0053	-0.0112	0.0043	0.0112
Abs (Disc. Accruals)	-0.0341*	-0.0467*	0.0506*	-0.3238*	-0.1518*	-0.1413*	-0.0855*	-0.0508*	0.0452*	-0.0220*
Abs (Accruals)	-0.0478*	-0.0477*	0.0574*	-0.3341*	-0.1612*	-0.1411*	-0.0929*	-0.0718*	0.0425*	-0.0076
Abs (Accruals/CFO)	-0.0137*	-0.0336*	0.0506*	-0.1410*	-0.0841*	-0.0844*	-0.0548*	-0.0424*	0.0321*	-0.0027
DD Residual	-0.0372*	-0.0537*	0.0561*	-0.3631*	-0.1643*	-0.1721*	-0.0980*	-0.0757*	0.0549*	0.0187*
Prior ROA Meet	0.0104	0.0068	-0.0049	0.0466*	0.0266*	0.0230*	0.0187*	0.0097	-0.0034	-0.0056
Small Profit	-0.0136*	-0.0083	-0.0123*	0.0426*	-0.0044	0.012	0.0029	-0.0084	0.0026	-0.0008

(Continued)

	Audit Report Lag	Non-timely Issuance of 10K_Due to Audit	Going Concern	Internal Control Weakness	Disc. Accruals	Abs (Disc. Accruals)	Abs (Accruals)	Abs (Accruals/CFO)	DD Residual	Prior ROA Meet
Audit Report Lag	1									
Non-timely Issuance of 10K_Due to Audit	0.1480*	1								
Going Concern	0.3952*	0.1180*	1							
Internal Control Weakness	0.1125*	0.0642*	-0.0228*	1						
Disc. Accruals	-0.0359*	-0.0061	-0.0514*	-0.0066	1					
Abs (Disc.	0.2518*	0.0561*	0.4126*	-0.0167*	-0.3005*	1				

Accruals)										
Abs (Accruals)	0.3011*	0.0903*	0.4767*	-0.0148*	-0.3808*	0.4947*	1			
Abs (Accruals/CFO)	0.1885*	0.0459*	0.1555*	0.0329*	-0.1568*	0.1644*	0.2516*	1		
DD Residual	0.2953*	0.0919*	0.4543*	-0.0145*	-0.0593*	0.4381*	0.5111*	0.1353*	1	
Prior ROA Meet	-0.0492*	-0.0032	-0.0207*	-0.0083	-0.0054	-0.0121*	-0.0081	-0.0201*	-0.0225*	1
Small Profit	0.0067	-0.0095	-0.0757*	0.0261*	0.0118	-0.0500*	-0.0512*	0.0164*	-0.0543*	-0.0222*

Note: * indicates significant at 5% level.

Cost-Sensitive Learning (CSL) and Multi-Subset Observation Under-sampling (OU) Method

CSL works mainly by adjusting the ratio between the number of positive and negative instances in the training dataset (Elkan 2001). Specifically, this adjustment allows us to keep all positive instances and a proportion (i.e., 1/misclassification cost) of the negative instances in the training dataset (Thai-Nghe, Gantner, and Schmidt-Thieme 2010). In other words, rendering an algorithm cost-sensitive is equivalent to under-sampling the negative instances in the training dataset (Thai-Nghe et al. 2010). However, merely under-sampling the negative examples can result in the discarding of potentially useful information for the classification task (Perols et al. 2017). To avoid information loss, we follow the example of Perols et al. (2017) and adopt the Multi-Subset Observation Undersampling (OU) method (Chan and Stolfo 1998), which creates n (n equals misclassification cost) training datasets, each of which contains all positive observations as well as a different subsample of negative observations. Perols et al. (2017) document a detailed description of the OU method.

Table A4. Grid of Hyperparameters Tested for Each Algorithm

Algorithm	Hyperparameter tested
LR	penalty: ['l2','none'] C: [0.1, 1.0, 10] solver: ['sag', 'saga','lbfgs']
SVM	C: [0.1, 1.0, 10] kernel: ['linear', 'poly', 'rbf'] gamma: ['scale','auto']
AB	base_estimator: [DecisionTreeClassifier(max_depth=1), DecisionTreeClassifier(max_depth=3), DecisionTreeClassifier(max_depth=5)] n_estimators: [30, 50, 70] learning_rate: [0.5, 1, 1.5]
RF	n_estimators: [50, 100, 200] max_features: ['auto', 'sqrt'] max_depth: [3, 5,10] min_samples_split:[2, 5, 10]
GB	min_samples_split: [2, 4, 6] min_samples_leaf: [1, 3] min_weight_fraction_leaf: [0.0, 0.3, 0.5] max_depth: [3, 5] max_leaf_nodes: [None] max_features: ['auto', 'sqrt', None]
XGB	eta: [0.01, 0.2, 0.3], min_child_weight: [1, 5], max_depth: [3, 6, 10], subsample: [0.5, 1], colsample_bytree: [0.5, 1]

Sequential Backward Feature Subset Selection

We designed a sequential backward feature subset selection process that performs a greedy search of the best subset of features for all possible total number of features. In this process, for each algorithm, the model with default parameter was adopted. The pseudo code of this process is as follows.

1. Start with the original feature size of 31 (there are 31 ARV)

2. For each possible subset feature size n in {30, 29, 28, … 1}

3. For i in {1, 2, …, 31-n}:

3.1 Generate all possible feature subsets of size 31-i

3.2 Split the entire dataset into five stratified randomized folds

3.3 For each candidate feature subset:

3.3.1 For each fold:

3.3.1.1 Obtain the AUC of the candidate feature subset on the hold out set

3.3.2 Average the AUC across five folds

3.5 Keep the candidate feature subset that generates the highest AUC (i.e., we remove the feature whose absence results in highest AUC gain)

4. Select the subset feature that has the highest AUC from the best subset feature with all possible sizes n in {30, 29, 28, … 1}

Benchmark Variables

The first set of benchmark variables (BMK1) are the 35 predictive variables for material misstatement identified from Bertomeu et al. (2020). Table 7 of Bertomeu et al. (2020) lists 36 variables including financial, audit, and corporate governance variables that counts 76.17% of cumulative importance in detecting material misstatements. While the variable “% Outsiders appointed” was inaccessible for this study, as it is from a commercial database called Equilar (Larcker et al. 2017), we included the other 35 variables. These 35 variables are listed below.

Table A5. BMK1: Predictive Variables of Material Misstatement (Adapted from Bertomeu et al. 2020)

Predictor	**Calculation**
% Soft assets	(Total Assets-PP&E-Cash and Cash Equivalent)/Total Assets
Bid ask spread	Past 252 days average bid-ask spread
Non-audit fee / total fee	Ratio of non-audit fees to total fees
Qualified opinion (internal control)	An indicator variable coded 1 if the firm gets a qualified opinion in term of internal control from auditor
Change in operating lease activity	The change in the present value of future noncancelable operating lease obligations deflated by average total assets
Short interest	Past 12 months average short interest percentage from COMPUSTAT
Stock return volatility	Past 252 days stock return volatility
Log of non-audit fee	Log value of non-audit fees
Percentile rank of audit fee by auditor	Percentile rank of audit fees by auditor
Leverage	Long-term Debt/Total Assets
Level of finance raised	Financing Activities Net Cash Flow /Average total assets
Abnormal change in employees	Percentage change in the number of employees-percentage change in assets
WC accruals	[(Change in Current Assets- Change in Cash and Short-term Investments)-(Change in Current Liabilities- Change in Debt in Current Liabilities- Change in Taxes Payable)]/ Average Total Assets
% Outsiders own	The fraction of outstanding shares held by the average outside director
Book-to-market	Equity/Market Value
Change in inventory	Change in Inventory/Average Total Assets
Lag one year return	Previous year's annual buy-and-hold return inclusive of delisting returns minus the annual buy-and-hold value-weighted market return
Earnings-to-price	Earnings/Market Value
Return	Annual buy-and-hold return inclusive of delisting returns minus the annual buy-and-hold value-weighted market return
Lag mean-adjusted absolute value of DD residuals	The following regression is estimated for each two-digit SIC industry: Change in WC = b0+ b1 Change in CFOt−1+b2 Change in CFOt +b3 Change in CFOt+1 +%. The mean absolute value of the residual is calculated for each industry and is then subtracted from the absolute value of each firm's observed residual.
% Outsiders appointed	The fraction of outside directors who were appointed by existing insiders
Analyst forecast errors	Absolute value of difference between actual earnings and analyst forecast consensus, deflated by stock price at the fiscal year-end
Change in cash margin	Percentage change in cash margin where cash margin is measure as 1- [(Cost of Good-Change in Inventory + Change

	in Accounts Payable)/ (Sales - Change in Accounts Receivable)]
RSST accruals	(Change in WC+ Change in NCO+ Change in FIN)/Average Total Assets, Where WC = (Current Assets- Cash and Short-term Investments) - (Current Liabilities-Debt in Current Liabilities); NCO = (Total Assets - Current Assets- Investments and Advances) - (Total Liabilities - Current Liabilities- Long-term Debt; FIN = (Short-term Investments + Long-term Investments) - (Long-term Debt + Debt in Current Liabilities + Preferred Stock)
Firm age	Firm age since listing
Change in cash sales	Percentage change in cash sales (Sales - Change in Accounts Receivable)
Deferred tax expense	Deferred tax expense for year t/ total assets for year t-1
Lag studentized DD residuals	Scales each residual by its standard error from the industry-level regression
Percentile rank of non-audit fee by auditor	Percentile rank of non-audit fees by auditor
Percentile rank of total fee by auditor	Percentile rank of total fees by auditor
Log of total fee	Log value of total audit fees
Log of audit fee	Log value of audit fees
Change in receivables	Change in Accounts Receivable/Average Total Assets
Change in return on assets	(Earnings t /Average Total Assets t)- (Earningst−1/Average Total Assetst−1)
Performance matched discretionary accruals	The difference between the modified Jones discretionary accruals for firm i in year t and the modified Jones discretionary accruals for the matched firm in year t, following Kothari et al. (2005); each firm-year observation is matched with another firm from the same two-digit SIC code and year with the closest return on assets.
Modified Jones model discretionary accruals	The modified Jones model discretionary accruals are estimated cross-sectionally each year using all firm-year observations in the same two-digit SIC code: WC Accruals =b0 + b1 (1/Beginning assets) + b2 (Change in Sales- Change in Rec) /Beginning assets+b3 Change in PPE /Beginning assets + e. The residuals are used as the modified Jones model discretionary accruals.

The second set of benchmark variables (BMK2) are 33 raw financial variables, including 28 raw financial variables that Bao et al. (2020) identified from Dechow et al. (2011) and Cecchini et al. (2011) and five extra raw financial variables used in Perols (2011) and Perols et al. (2017). The 33 raw financial variables are listed below, and they are readily accessible from COMPUSTAT.

Table A6. BMK2: Raw Financial Variables

Variable	Main Source
Cash and Short-Term Investments	Beneish (1997, 1999), Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011) and Perols et al. (2017)
Receivables - Total	Beneish (1997, 1999), Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011) and Perols et al. (2017)
Inventories - Total	Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011) and Perols et al. (2017)
Short-Term Investments - Total	Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011) and Perols et al. (2017)
Current Assets - Total	Beneish (1997, 1999), Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011) and Perols et al. (2017)
Property, Plant and Equipment - Total (Gross)	Beneish (1997, 1999), Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011) and Perols et al. (2017)
Investment and Advances - Other	Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011) and Perols et al. (2017)
Assets - Total	Beneish (1997, 1999), Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011) and Perols et al. (2017)
Accounts Payable - Trade	Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011) and Perols et al. (2017)
Debt in Current Liabilities - Total	Beneish (1997, 1999), Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011) and Perols et al. (2017)
Income Taxes Payable	Beneish (1997, 1999), Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011) and Perols et al. (2017)
Current Liabilities - Total	Beneish (1997, 1999), Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011) and Perols et al. (2017)
Long-Term Debt - Total	Beneish (1997, 1999), Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011) and Perols et al. (2017)
Liabilities - Total	Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011), and Perols et al. (2017)
Common/Ordinary Equity - Total	Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011), and Perols et al. (2017)
Preferred/Preference Stock (Capital) - Total	Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011), and Perols et al. (2017)
Retained Earnings	Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011), and Perols et al. (2017)
Sales/Turnover (Net)	Beneish (1997, 1999), Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011) and Perols et al. (2017)

Cost of Goods Sold	Beneish (1997, 1999), Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011) and Perols et al. (2017)
Depreciation and Amortization	Beneish (1997, 1999), Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011) and Perols et al. (2017)
Interest and Related Expense - Total	Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011), and Perols et al. (2017)
Income Taxes - Total	Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011), and Perols et al. (2017)
Income Before Extraordinary Items	Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011), and Perols et al. (2017)
Net Income (Loss)	Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011), and Perols et al. (2017)
Long-Term Debt - Issuance	Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011), and Perols et al. (2017)
Sale of Common and Preferred Stock	Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011), and Perols et al. (2017)
Price Close - Annual - Calendar	Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011), and Perols et al. (2017)
Common Shares Outstanding	Bao et al. (2020), Cecchini et al. (2010), Dechow et al. (2011), Perols (2011), and Perols et al. (2017)
Common Shares Issued	Perols (2011) and Perols et al. (2017)
Operating Activities - Net Cash Flow	Perols (2011) and Perols et al. (2017)
Operating Income Before Depreciation	Perols (2011) and Perols et al. (2017)
Property, Plant and Equipment - Total (Net)	Perols (2011) and Perols et al. (2017)
Working Capital (Balance Sheet)	Perols (2011) and Perols et al. (2017)
Selling, General and Administrative Expense	Beneish (1997, 1999)
Amortization of Intangibles	Beneish (1997, 1999)

Table A7. Raw Audit Variables

Variable Name	Description (Excerpted from Audit Analytics Data Dictionary)	Source Module from Audit Analytics (unless indicated otherwise)
GOING_CONCERN	Indicates the auditor's opinion contains an explanatory paragraph regarding the going concern assumption.	AUDIT OPINIONS
OP_AUD_PCAOB	Identifies current PCAOB registration status of the auditor - 0=Not,1=registered	AUDIT OPINIONS
INTEGRATED_AUDIT	An "integrated audit" is defined as occurring when in addition to the auditor expressing an opinion on the financial statements, they also issued an opinion on the effectiveness of internal controls over financial reporting. (Per PCAOB standard #5)	REVISED AUDIT OPINION
IS_NTH_ADD_OP	Identifies multiple opinions for the same period	REVISED AUDIT OPINION
AUDIT_FEES	Consists of all fees necessary to perform the audit or review in accordance with GAAS. This category also may include services that generally only the independent accountant reasonably can provide, such as comfort letters, statutory audits, attest services, consents and assistance with and review of documents filed with the SEC	AUDIT FEES
NON_AUDIT_FEES	The sum of Audit Related Fees, Benefit Plan Related Fees, FISDI Fees, Tax Related Fees and Other/Misc Fees	AUDIT FEES
TOTAL_FEES	The sum of Audit Fees and Total Non-Audit Fees. Rows in which the total fees are zero for a particular year were due to the registrant disclosing an auditor as having been engaged as their independent accountants for the year, yet not disclosing the corresponding auditor fees	AUDIT FEES
BENEFITS_FEES	In general, these fees compose part of the total audit related fee number. In cases where the registrant itemizes their audit related fees and discloses the fees associated with benefit plan audits, the benefit plan fees are subtracted from the total audit	AUDIT FEES

	related fees and entered under this field	
TAX_FEES	Typically, this category would include fees for tax compliance, tax planning, and tax advice. Tax compliance generally involves preparation of original and amended tax returns, claims for refund and tax payment-planning services. Tax planning and tax advice encompass a diverse range of services, including assistance with tax audits and appeals, tax advice related to mergers and acquisitions, employee benefit plans and requests for rulings or technical advice from taxing authorities. This category would not capture those services related to the audit	AUDIT FEES
TAX_FEES_COMPLIANCE	Tax compliance generally involves preparation of original and amended tax returns, claims for refund and tax payment-planning services. Tax Related Fees = (Tax Related Fees - Compliance) + (Tax Related Fees - Non-Compliance) Please note that this field was added in December 2011. It will be populated on a going-forward basis. For a select number of companies some historical data will be added.	AUDIT FEES
TAX_FEES_NON_COMPLIANCE	Tax planning and tax advice encompass a diverse range of services, including assistance with tax audits and appeals, tax advice related to mergers and acquisitions, employee benefit plans and requests for rulings or technical advice from taxing authorities. Tax Related Fees = (Tax Related Fees - Compliance) + (Tax Related Fees - Non-Compliance) Please note that this field was added in December 2011. It will be populated on a going-forward basis. For a select number of companies some historical data will be added.	AUDIT FEES
AUDIT_RELATED_FEES	In general, these are assurance and related services (e.g., due diligence services) that traditionally are performed by the independent accountant. More	AUDIT FEES

	specifically, these services would include, among others: employee benefit plan audits, due diligence related to mergers and acquisitions, accounting consultations and audits in connection with acquisitions, internal control reviews, attest services that are not required by statute or regulation and consultation concerning financial accounting and reporting standards.	
OTHER_FEES	All other auditor fees. Note that prior to the implementation of SEC Rule 33-8183 this category included tax related fees and audit related fees.	AUDIT FEES
FEES_PCAOB_REG	Flag indicates whether auditor is currently registered with the PCAOB	AUDIT FEES
DISMISSED_GC	Flagged if the filing indicated that the previous auditor issued a going concern opinion	AUDITOR CHANGES
DISMISSED_DISAGREE	Indicates that the company and the auditor are or have been in disagreement on a matter of accounting principles or practices, financial statement disclosure, or auditing scope or procedure. This will be checked even when the filing indicates the disagreement has been resolved.	AUDITOR CHANGES
AUDITOR_RESIGNED	Indicates whether the departed auditor resigned from the engagement or was dismissed from the engagement by the registrant. 1=outgoing auditor resigned; 0=outgoing auditor was dismissed by company	AUDITOR CHANGES
IS_BENEFIT_PLAN	Indicates whether or not the "Subsidiary - Plan" field is a benefit plan.	AUDITOR CHANGES
MERGER	Flagged when the ITEM 4.01 AUDITOR CHANGES was due to a merger or acquisition	AUDITOR CHANGES
DISMISS_PCAOB_REG	Flag indicates whether outgoing old auditor is currently registered with the PCAOB	AUDITOR CHANGES
ENGAGED_AUDITOR_PCAOB	Flag indicates whether incoming new auditor is currently registered with the PCAOB.	AUDITOR CHANGES
ISS_INTERNAL_CONTROL	Indicates the registrant specifically identifies an internal control issue. This does not mean	AUDITOR CHANGES

	that a lack of these controls, whether corrected or not, were the cause of the auditor change. Rather, it indicates simply that they were mentioned. In circumstances where reasons are given for resignations or dismissals that seem like internal control deficiencies, but the expression "internal controls" is not used, the field will not be flagged. In circumstances where "reportable conditions" are identified but the expression, "internal controls" is not used, the field will not be flagged.	
ISS_ACCOUNTING	Indicates issues related to accounting treatments and/or disagreements about accounting principles were disclosed. Flagged even if the company states that they have made the necessary changes for compliance or if they say that there is no longer any disagreement between the registrant and the auditor.	AUDITOR CHANGES
ISS_SCOPE	Indicates that the registrant disclosed that the auditor identified a scope limitation issue. Example of a Scope Limitation disclosure: "Because of the absence of significant accounting records of the Partnership for 2001, it was not practicable to extend our auditing procedures to enable us to form an opinion on the accompanying 2002 consolidated statements... the scope of our work was not sufficient to enable us to express an opinion on the results of operations, changes in partners' capital and..."	AUDITOR CHANGES
ISS_RESTATE_FINS	Indicates that the registrant reported that a restatement of the financials either occurred or will occur.	AUDITOR CHANGES
ISS_AUDIT_OPINION	Indicates that the registrant disclosed that there are questions regarding the veracity or applicability of previous or upcoming audit opinions. The field covers such areas as companies issuing unauthorized opinions to concerns being raised	AUDITOR CHANGES

	about the veracity of opinions that have been issued.	
ISS_MNGMT_REP	Indicates that the registrant disclosed that there are questions regarding the reliability of information that has or had been reported by the former and/or current management of the registrant.	AUDITOR CHANGES
ISS_ILLEGAL_ACTS	Indicates that the registrant disclosed the identification of an illegal or allegedly illegal act. This could range from embezzlement by company personnel to violations of the Foreign Corrupt Practices Act. These acts may be corrected, rectified or related to personnel no longer at the company.	AUDITOR CHANGES
ISS_REDUCE_FEES	Indicates that the registrant has identified the desire to reduce audit fees as a reason for changing auditors.	AUDITOR CHANGES
ISS_SEC_INVESTIGATE	Indicates an SEC investigation is mentioned in the auditor change disclosure.	AUDITOR CHANGES
ISS_EXIT_AUDITS	Indicates that the departing auditor will be discontinuing their public company audit practice.	AUDITOR CHANGES
ISS_LACK_INDEPENDENCE	Indicates that the registrant disclosed that there are questions regarding the departing auditor's independence. These issues may include circumstances such as an audit firm losing independence due to an employee having been employed by the registrant. Also includes such circumstances as the auditor and registrant being in dispute, as such disputes impairs the auditor's independence.	AUDITOR CHANGES
ISS_AUD_MERGE	Indicates that the auditor change has occurred due to a merger, acquisition or reorganization between two audit firms.	AUDITOR CHANGES
ISS_PCAOB_NOT_REG	Indicates auditor not registered with PCAOB issue	AUDITOR CHANGES
ISS_REAUDIT_PREVIOUS	Indicates incoming auditors are going to re-audit previous periods	AUDITOR CHANGES
ISS_IC_REPORTABLE	Indicates that the registrant disclosed a reportable condition exists either as referenced to SEC regulations or professional standards (GAAS/GAAP). In circumstances where reasons are	AUDITOR CHANGES

	given for resignations or dismissals that seem like reportable conditions, but the expression "reportable condition" is not used, the field will not be flagged.	
ISS_SEC_INQUIRE_CO	Indicates the registrant disclosed that they have received an inquiry from the SEC concerning accounting and related treatments on the registrant's filings.	AUDITOR CHANGES
ISS_SEC_INQUIRE_AUD	Indicates the registrant disclosed that there have been SEC inquiries regarding the registrant's auditor related to the treatment of certain accounting and reporting matters.	AUDITOR CHANGES
ISS_FEE	Indicates that the registrant disclosed that there was a fee dispute between the registrant and the departing auditor.	AUDITOR CHANGES
ISS_BANKRUPT	Indicates the registrant disclosed that it will have to file, is concerned about or is emerging from some sort of bankruptcy, receivership or insolvency.	AUDITOR CHANGES
ISS_BAN_AUD	Indicates an auditor has resigned or has been dismissed because they have been banned by the SEC for performing audit work.	AUDITOR CHANGES
ISS_OTHER	Indicates other issue(s) disclosed which is not part of this taxonomy.	AUDITOR CHANGES
BOARD_APP_OUT	Indicates that the audit committee and/or the board of the registrant approved the dismissal or resignation of the departing auditors.	AUDITOR CHANGES
BOARD_APP_IN	Indicates that the audit committee and/or the board of the registrant approved the engagement of the new auditor.	AUDITOR CHANGES
CONSULT_INCOMING	Indicates that the registrant consulted previously with the new auditor prior to the engagement.	AUDITOR CHANGES
AUD_LETTER_DISAGREE	Indicates whether the auditor's letter agrees or disagrees (in any part) with the registrant's disclosure of the auditor change.	AUDITOR CHANGES
AUD_LETTER_NO_COMM	Indicates whether the auditor's letter agrees or disagrees (in any part) with the registrant's disclosure of the auditor change.	AUDITOR CHANGES
AUD_LETTER_AGREE	Indicates whether the auditor's response letter agrees or disagrees	AUDITOR CHANGES

	(in any part) with the registrant's disclosure of the auditor change.	
AUD_CO_DISAGREE	Indicates that the company and the auditor are in disagreement about a significant issue affecting the company. The disagreement could be concerning a range of issues, including, but not limited to, an accounting issue and its impact on the financial statements, whether the financial statements should be restated, and whether management has engaged in illegal acts.	AUDITOR CHANGES
IS_EFFECTIVE	According to assessment of disclosure controls: Y (1) =yes, disclosure controls were found to be effective; N (0) =no, disclosure controls were not found to be effective; ND (2) =not disclosed	SOX 302 DISCLOSURE CO
MATERIAL_WEAKNESS	Indicates whether the management indicated a material weakness existed in the registrant's disclosure controls. 0 means NO, 1 means YES	SOX 302 DISCLOSURE CO
SIG_DEFICIENCY	Indicates whether the management indicated other deficiencies or disclosures regarding the effectiveness of their disclosure controls. 0 means NO, 1 means YES	SOX 302 DISCLOSURE CO
NOTEFF_ACC_RULE_x	Indicates that the assessment of disclosure controls identified accounting rule application failures. The particular reasons for the failures noted are listed.	SOX 302 DISCLOSURE CO
NOTEFF_FIN_FRAUD_x	Indicates that the assessment of disclosure controls identified financial fraud, irregularities and misrepresentations. The particular reasons for the failures are listed.	SOX 302 DISCLOSURE CO
NOTEFFERRORS_x	Errors in accounting and clerical applications noted in assessment of disclosure controls	SOX 302 DISCLOSURE CO
NOTEFF_OTHER_x	Material weakness identified in assessment of disclosure controls	SOX 302 DISCLOSURE CO
AUDITOR_AGREES	Auditor agrees with management's assessment of internal controls, valid data for auditor assessment. 0 = No; 1 = Yes; 2 = Not Disclosed.	SOX 404 INTERNAL CONT
COMBINED_IC_OP	Set = 1 if the auditor's financial statement report is integrated with their assessment of internal	SOX 404 INTERNAL CONT

	controls, valid data for auditor assessment (A) only	
IC_IS_EFFECTIVE	According to assessment of internal controls: Y=yes, internal controls were found to be effective; N=no, internal controls were not found to be effective; ND=not disclosed	SOX 404 INTERNAL CONT
COUNT_WEAK	Count of material weaknesses identified in assessment of internal controls	SOX 404 INTERNAL CONT
NOTEFF_ACC_RULE_y	Indicates that the assessment of disclosure controls identified accounting rule application failures. The particular reasons for the failures noted are listed.	SOX 404 INTERNAL CONT
EXEMPTION	Indicates exemptions to the assessment of internal controls over financial reporting were identified. The particular exemptions are listed. (2=yes, 0=no)	SOX 404 INTERNAL CONT
NOTEFF_FIN_FRAUD_y	Indicates that the assessment of disclosure controls identified financial fraud, irregularities and misrepresentations. The particular reasons for the failures are listed.	SOX 404 INTERNAL CONT
NOTEFFERRORS_y	Errors in accounting and clerical applications noted in assessment of disclosure controls	SOX 404 INTERNAL CONT
NOTEFF_OTHER_y	Material weakness identified in assessment of disclosure controls	SOX 404 INTERNAL CONT
HST_SEASON_ISSUER	Indicates how the registrant identified their "Well known seasoned issuer" status. (1= Yes; 2 = Did not disclose; 0 = No)	ACCELERATED FILER
HST_IS_SHELL_CO	Indicates how the registrant identified their shell company status. (1= Yes; 2 = Did not disclose; 0 = No)	ACCELERATED FILER
HST_IS_ACCEL_FILER	Indicates how the registrant identified their accelerated filer status. (1= Yes; 2 = Did not disclose; 0 = No)	ACCELERATED FILER
HST_IS_LARGE_ACCEL	Indicates how the registrant identified their large accelerated filer status. (1= Yes; 2 = Did not disclose; 0 = No)	ACCELERATED FILER
HST_IS_VOLUNTARY_FILER	Indicates whether the registrant filed as a voluntary filer. (1= Yes; 2 = Did not disclose; 0 = No)	ACCELERATED FILER
HST_IS_SMALL_REPORT	Indicates whether the registrant filed as a smaller reporting company. (1= Yes; 2 = Did not disclose; 0 = No)	ACCELERATED FILER

DID_NOT_DIS	Indicates if the company did not disclose their accelerated filer status. (1= Yes; 0 = No)	ACCELERATED FILER
Auditor Opinion (AUOP)	This item contains the code that indicates the auditor's opinion on a company's financial statements.0 indicates that the financial statements were not audited. 1 indicates that the financial statements are presented fairly. 2 indicates that the financial statements are presented fairly, but the auditing firm is concerned about either limitation on the scope of the examination or unsatisfactory financial statement presentations. 3 indicates that the auditing firm does not express an opinion on the financial statements. 4 indicates that the auditing firm's opinion is unqualified, but explanatory language has been added to the standard report. 5 indicates that the financial statements are not presented fairly.	COMPUSTAT (not from Au
Auditor Opinion - Internal Control (AUOPIC)	This item represents the auditor's opinion of the effectiveness of the company's internal control over financial reporting in conjunction with auditing a company's financial statements. 0 indicates no Auditor's report. 1 indicates effective (No Material Weakness). 2 indicates adverse (Material Weakness Exists). 3 indicates disclaimer (Unable to Express Opinion). 4 indicates delayed filing.	COMPUSTAT (not from Au
Fee_Res	(self-created) Indicates if there is a restatement of audit fees	AUDIT FEES RESTATED
Adj_AUDIT_FEES	(self-created) Difference between the AUDIT_FEES in the AUDIT FEES RESTATED module and that in the AUDIT FEES module	AUDIT FEES RESTATED a
Adj_NON_AUDIT_FEES	(self-created) Difference between the NON_AUDIT_FEES in the AUDIT FEES RESTATED module and that in the AUDIT FEES module	AUDIT FEES RESTATED a
Adj_TOTAL_FEES	(self-created) Difference between the TOTAL_FEES in the AUDIT FEES RESTATED module and that in the AUDIT FEES module	AUDIT FEES RESTATED a

Adj_BENEFITS_FEES	(self-created) Difference between the BENEFITS_FEES in the AUDIT FEES RESTATED module and that in the AUDIT FEES module	AUDIT FEES RESTATED a
Adj_TAX_FEES	(self-created) Difference between the TAX_FEES in the AUDIT FEES RESTATED module and that in the AUDIT FEES module	AUDIT FEES RESTATED a
Adj_TAX_FEES_COMPLIANCE	(self-created) Difference between the TAX_FEES_COMPLIANCE in the AUDIT FEES RESTATED module and that in the AUDIT FEES module	AUDIT FEES RESTATED a
Adj_TAX_FEES_NON_COMPLIA NCE	(self-created) Difference between the TAX_FEES_NON_COMPLIAN CE in the AUDIT FEES RESTATED module and that in the AUDIT FEES module	AUDIT FEES RESTATED a
Adj_AUDIT_RELATED_FEES	(self-created) Difference between the AUDIT_RELATED_FEES in the AUDIT FEES RESTATED module and that in the AUDIT FEES module	AUDIT FEES RESTATED a
Adj_OTHER_FEES	(self-created) Difference between the OTHER_FEES in the AUDIT FEES RESTATED module and that in the AUDIT FEES module	AUDIT FEES RESTATED a
NT	(self-created) Indicates if there is a non-timely filing notice	NT
NT_AUDIT	(self-created) Indicates if the non-timely filing notice is related to audit issues	NT
SOX302report	(self-created) Indicates if there is a SOX 302 disclosure	SOX 302 DISCLOSURE CO
SOX404_Auditor	(self-created) Indicates if there is a SOX 404 disclosure	SOX 404 INTERNAL CONT
Big4	(self-created) Indicates if the auditor is one of the Big4	AUDIT OPINIONS
Tenure	(self-created) The length of auditor-client relationship	AUDITOR CHANGES
New Client	(self-created) Indicates if this is the first year of the auditor-client relationship	AUDITOR CHANGES

Appendix for Chapter 4

Control Variables

We control for the timeliness of the management forecast disclosure by using HORIZON, the natural logarithm of the number of days between the forecast date and the fiscal period end date (Ajinkya et al. 2005; Dorantes et al. 2013). The larger the HORIZON, the earlier the forecast is provided. We expect HORIZON to be negatively associated with management earnings forecast accuracy. We use LNAT, the natural logarithm of total assets and the end of year t, to control firm size. We expect LNAT to be positively associated with management earnings forecast accuracy (e.g., Dorantes et al. 2013; Cheng et al. 2018; Huang et al. 2018). We use ABSCHGROA, the absolute value of the change in ROA (earnings before extraordinary items scaled by total assets) from year t-1 to year t, to control for the expected difficulty of earnings predictions (Dorantes et al. 2013). We expect ABSCHGROA to be negatively associated with management earnings forecast accuracy (more errors). To control for the incentive to provide better quality forecasts, we use NEWEQUITY, an indicator that equals 1 if a firm issues new equity or new debt in the following year, and 0 otherwise (Frankel, McNichols, and Wilson 1995; Dorantes et al. 2013). We expect NEWEQUITY to be positively associated with management earnings forecast accuracy. ANALYST is used to control for the private information production, and it is the natural logarithm of the number of analysts following at the end of year t (Ajinkya et al. 2005; Dorantes et al. 2013). We expect ANALYST to be positively associated with management earnings forecast accuracy. LOSS, a binary variable that equals 1 if the firm reported a loss in the current period and 0 otherwise, is used to control for profitability (Ajinkya et al. 2005;

Dorantes et al. 2013). We expect LOSS to be negatively associated with management earnings forecast accuracy. NEWS is a binary variable that equals 1 if the current-period EPS is greater than or equal to the previous-period EPS, and 0 otherwise, and it is to control for the intention to prevent unfavorable litigation (Ajinkya et al. 2005). We expect NEWS to be positively associated with management earnings forecast accuracy. To control for firms' organizational changes, we include ORGCHG, which is a factor variable composed of asset growth, sales growth, leverage, and merger and acquisition activity (Dorantes et al. 2013). We expect ORGCHG to be negatively associated with management earnings forecast accuracy. We include COMPLEXITY that is a factor variable composed of the number of segments, the existence of foreign transactions and the existence of a restructuring to control for firm complexity (Dorantes et al. 2013). We expect COMPLEXITY to be negatively associated with management earnings forecast accuracy. BIG4 is used to control for auditor's reputation, and we expect BIG4 to be positively associated with management earnings forecast accuracy. The volatility of earnings is controlled by using EARNVOL, the standard deviation of quarterly earnings over 12 quarters ending in the current fiscal year, divided by median asset value for the period. We expect EARNVOL to be negatively associated with management earnings forecast accuracy. BETA is used to control for the sensitivity of the asset returns to the market returns, and it is the equity beta for the last fiscal period. We also expect BETA to be negatively associated with management earnings forecast accuracy. As presented in the following sections of Appendix, the correlations between the control and dependent variables are, in general, as expected.

Table A8. Variable Descriptions

Variables	Definition
ACCURACY	Accuracy of management earnings (annual earnings per share) forecast. It equals -1 multiplied by the absolute value of the forecast error scaled by prior year stock price, where the forecast error is the difference between the actual earnings per share and management forecast (Ajinkya et al. 2005; Dorantes et al. 2013). When management forecast is a range estimation, we use the midpoint of the upper and lower values of the forecast range (Ajinkya et al. 2005; Feng et al. 2009; Dorantes et al. 2013). The larger the value, the more accurate the management earnings forecast.
ACCURACY_AVE	The average of ACCURACY throughout the year. The larger the ACCURACY_AVE, the more accurate the management earnings forecasts (Dorantes et al. 2013).
ACCURACY_AVE_ALT	An alternative measure of the average management earnings forecast accuracy throughout the year. Compared to ACCURACY_AVE, the only difference is that ACCURACY_AVE_ALT uses asset per share as the scaler instead of prior year stock price (Feng et al. 2009).
PRECISION	1 if the forecast is a point forecast, and 0 if the forecast is a range. If a firm issues multiple forecasts in a given year, we use the average of point estimates throughout the year (Dorantes et al. 2013). The greater the PRECISION, the more precise the management earnings forecast.
FREQUENCY	The number of annual earnings forecasts issued by a firm in one year (Ajinkya et al. 2005)
BIAS	The difference between guidance value and the realized earnings per share value scaled by prior year stock price (Ajinkya et al. 2005)
TREATMENT	A binary variable that equals to 1 if the firm adopts AI (treatment firms), 0 otherwise
POST	A binary variable that equals to 1 if the firm-year observation is in the post-AI adoption period.
ML	A binary variable that equals to 1 if the treatment firm adopts ML, 0 otherwise
OTHERAI	A binary variable that equals to 1 if the treatment firm adopts other AI technologies than ML – RPA, NLP, Computer Vision, 0 otherwise
LNAT	Natural logarithm of total assets at the end of year t
ABSCHGROA	The absolute value of the change in ROA (earnings before extraordinary items scaled by total assets) from year t-1 to year t
NEWEQUITY	1 if a firm issues new equity or new debt in the following year, and 0 otherwise
ANALYST	The natural logarithm of the number of analysts following at the end of year t
HORIZON	The natural logarithm of the number of days between the forecast's date and the fiscal period end date. When the dependent variable is ACCURACY_AVE, we use the average of HORIZON.

LOSS	1 if the firm reported losses in the current period, and 0 otherwise
NEWS	1 if the current-period EPS is greater than or equal to the previous-period EPS, and 0 otherwise
ORGCHG	A factor comprised of asset growth, sales growth, leverage, and merger and acquisition activity. See details in Appendix C.
COMPLEXITY	A factor comprised of the number of segments, the existence of foreign transactions and the existence of a restructuring. See details in Appendix C.
BIG4	1 if the company is audited by one of the Big 4 auditors, and 0 otherwise
EARNVOL	The standard deviation of quarterly earnings over 12 quarters ending in the current fiscal year, divided by median asset value for the period.
BETA	The equity beta for the last fiscal period*
VOL	The decile of the standard deviation of quarterly earnings over 12 quarters ending in year t, divided by median asset value for the period.
ROA	Earnings before extraordinary items scaled by total assets (Brown and Caylor 2009).
FRQ	-1 multiplied by the absolute value of discretionary accruals, which are the residuals from the industry and year-specific regression models for the performance-adjusted discretionary accruals (Hope et al. 2013).
AISKILLRANK	The natural logarithm of the state-level AI skill penetration ranking compiled by Perrault et al. (2019)
LOGSALE	Natural logarithm of sales (COMPUSTAT item SALE) (Pincus et al. 2017)
UNCERTAINTY	Standard deviation of firm i's net income for the previous 5 years, scaled by sales (COMPUSTAT items IB and SALE) (Pincus et al. 2017)
CONCENTRATION	Four-firm industry concentration at the four-digit SIC level (COMPUSTAT item SALE) (Pincus et al. 2017)
STD_AF	The standard deviation of the individual analyst forecasts in year t (Feng et al. 2009)
FINCHA	A factor comprised of ROA, losses, research and development, and special items (Feng et al. 2009). ROA is income before extraordinary items scaled by prior year total assets; Losses is a binary variable that equals to 1 if the sum of income before extraordinary items of current and prior is less than zero; research and development is the research and development expense scaled by prior year total assets; Special items is the absolute value of special items scaled by prior year total assets (Feng et al. 2009).
IMR_AI	The Inverse Mills Ratio for AI adoption based on model (a) in Appendix E.
IMR_DISCLOSE	The Inverse Mills Ratio for management earnings forecast disclosure based on model (b) in Appendix E.

*Note: We imputed missing values of BETA using the year and industry average to avoid sample loss. We conducted the analysis without the imputed value and the results remain similar.

Factor Formation

Following Feng et al. (2009) and Dorantes et al. (2013), we use principal component analysis to aggregate seven variables into two factors, representing organizational changes (ORGCHG) and firm complexity (COMPLEXITY) (see Table A1). To present a parsimonious model, we report results using the two factors. Our results are not sensitive to including the two factors versus the 7 individual controls.

Table A9

Factors	Component Loadings
ORGCHG	
ASSETGROWTH	0.6267
LEVERAGE	0.5110
SALESGROWTH	0.5105
MA	0.2924
COMPLEXITY	
RESTRUCTURING	0.5034
FOREIGN	0.6239
LNSEGMENT	0.5978

Note:
SALESGROWTH equals sales growth from prior year to current year (Feng et al. 2009; Dorantes et al. 2013; Huang et al. 2018); ASSETGROWTH is asset growth from prior year to current year (Feng et al. 2009; Dorantes et al. 2013)**;** LEVERAGE is total liabilities /lagged total assets (Ajinkya et al. 2005; Feng et al. 2009; Dorantes et al. 2013; Cheng et al. 2018; Huang et al. 2018); MA is 1 if the firm undertook a large merger or acquisition in year t, and 0 otherwise (Feng et al. 2009; Dorantes et al. 2013; Huang et al. 2018); RESTRUCTURING is 1 if the firm recognized restructuring charges in current year, and zero otherwise (Feng et al. 2009; Dorantes et al. 2013; Huang et al. 2018); FOREIGN is 1 if the firm has foreign transactions in current year, and zero otherwise (Feng et al. 2009; Dorantes et al. 2013; Huang et al. 2018); LNSEGMENT is the natural logarithm of the total number of geographic and operating segments (Feng et al. 2009; Dorantes et al. 2013; Cheng et al. 2018; Huang et al. 2018). We imputed missing values of LNSEGMENT with 0 to avoid sample loss. The results are not sensitive to imputation.

Table A10. Correlation Among Variables

	ACCURACY_AVE	HORIZON	LNAT	ABSCHGROA	NEWEQUITY	ANALYST	LOSS	NEWS	ORGCHG	COMPLEXITY	BIG4	EARNVOL	BETA
ACCURACY_AVE	1												
HORIZON	-0.1157*	1											
LNAT	0.1230*	0.0737*	1										
ABSCHGROA	-0.1983*	0.0486	-0.2203*	1									
NEWEQUITY	0.1032*	0.0457	-0.1100*	0.0282	1								
ANALYST	0.1767*	0.0655*	0.6202*	-0.0677*	0.0984*	1							
LOSS	-0.3548*	0.0191	-0.1095*	0.3383*	-0.0253	-0.0969*	1						
NEWS	0.1169*	-0.0571	-0.0585*	-0.1035*	0.0770*	0.0537	-0.2335*	1					
ORGCHG	0.0285	0.0196	0.0525	0.1495*	0.0610*	0.0006	0.0067	-0.004	1				
COMPLEXITY	0.0221	0.0567	0.0172	0.0773*	0.0463	0.1424*	0.0466	0.0009	-0.0626*	1			
BIG4	0.0721*	-0.0092	0.4267*	-0.1041*	-0.0407	0.3124*	0.0037	-0.0109	-0.0289	0.0609*	1		
EARNVOL	-0.2177*	-0.0002	-0.2718*	0.7038*	0.0346	-0.0957*	0.3485*	-0.0341	0.0960*	0.0868*	-0.0941*	1	
BETA	-0.1522*	0.0966*	-0.1802*	0.1471*	-0.0155	-0.0617*	0.0900*	0.0286	0.0211	0.2272*	-0.0358	0.1647*	1

Note: *p value < 0.01 (two-tailed tests). Variable descriptions are provided in Table A8.

Table A11. Selection Model for AI Implementation

DV=TREATMENT		
	Expected sign	Coef. (Z score)
AISKILLRANK	-	-0.205***
		(-2.85)
UNCERTAINTY	-	-0.008***
		(-2.52)
CONCENTRATION	+	-0.069
		(-0.33)
LOGSALE	+	0.106***
		(3.55)
ABSCHGROA	?	0.02
		(0.40)
NEWEQUITY	?	0.214***
		(2.92)
ANALYST	?	0.323***
		(6.11)
LOSS	?	0.294***
		(3.99)
NEWS	?	0.127**
		(2.15)
ORGCHG	?	0.044
		(0.60)
COMPLEXITY	?	0.081***
		(3.05)
BIG4	?	-0.063
		(-0.72)
EARNVOL	?	-0.129
		(-0.77)
BETA	?	-0.098
		(-1.23)
Industry Fixed Effects		Yes
Year Fixed Effects		Yes
Pseudo R2		0.286
N		7319

Note: ***p < 0.01, **p < 0.05, *p < 0.10 (one-tailed tests when the coefficients have the predicted signs, two-tailed tests otherwise). We do not place any asterisk for one-tailed tests when the estimated sign is opposite to the predicted sign). Z scores are based on standard errors adjusted for firm clustering effects. Variable descriptions are provided in Table A8 in the Appendix.

Table A12. Selection Model for Management Earnings Forecast Disclosure

DV=DISCLOSE		
	Expected sign	Coef. (Z score)
LOGSALE	+	0.032**
		(1.67)
ABSCHGROA	-	-0.569**
		(-2.11)
NEWEQUITY	+	0.013
		(0.19)
ANALYST	+	0.490***
		(12.14)
LOSS	-	-0.289***
		(-4.14)
NEWS	-	0.008
		(0.23)
ORGCHG	-	-0.016
		(-0.37)
COMPLEXITY	-	0.197
		(7.14)
BIG4	+	-0.126
		(-1.30)
EARNVOL	-	-0.978
		(-1.03)
BETA	-	-0.034
		(-0.45)
STD_AF	-	-0.347***
		(-3.26)
FINCHA	-	-0.264***
		(-3.40)
Industry Fixed Effects		Yes
Year Fixed Effects		Yes
Pseudo R2		0.289
N		10551

Note: ***p < 0.01, **p < 0.05, *p < 0.10 (one-tailed tests when the coefficients have the predicted signs, two-tailed tests otherwise). We do not place any asterisk for one-tailed tests when the estimated sign is opposite to the predicted sign). Z scores are based on standard errors adjusted for firm clustering effects. The loadings for the factor variable FINCHA are -0.63 from ROA, 0.34 from losses, 0.52 from special items, and 0.47 from research and development. Variable descriptions are provided in Table A8.

Propensity Score Matching

We establish the following model to generate propensity scores of firms choosing to implement AI in base year t:[92]

$$
\begin{aligned}
TREATMENT_{it} = {} & b_0 + b_1 LNAT_{it} + b_2 HORIZON_{it} + b_3 ANALYST_{it} + b_4 ABSCHGROA_{it} \\
& + b_5 NEWEQUITY_{it} + b_6 LOSS_{it} + b_7 NEWS_{it} + b_8 BIG4_{it} + b_9 ORGCHG_{it} \\
& + b_{10} COMPLEXITY_{it} + b_{11} EARNVOL_{it} + b_{12} BETA_{it} + Year\ Fixed\ Effects + \varepsilon
\end{aligned}
$$

where *i* and *t* index firm and year, respectively, and the definitions of the predictor variables can be found in Table A1. Our PSM and main analysis models are similar in variable choice, following the guidance in Shipman et al. (2017) that explanatory variables in the main model be included in the PSM model. Table A6 presents the model that generates propensity scores.

Table A13. Propensity Score Matching Model

DV=TREATMENT		
	Expected sign	Coef. (Z score)
LNAT	+	0.046
		(1.13)
HORIZON	?	-0.369***
		(-2.75)
ABSCHGROA	?	0.395
		(0.50)
NEWEQUITY	+	0.126
		(1.13)
ANALYST	+	0.438***
		(4.88)
LOSS	-	0.327**
		(2.40)
NEWS	+	0.222**
		(2.37)
ORGCHG	+	-0.059
		(-0.35)
COMPLEXITY	+	0.041
		(1.27)
BIG4	+	0.322*
		(1.79)
EARNVOL	-	0.023
		(0.01)

[92] We did not include industry fixed effects in the PSM model because doing so would significantly reduce the number of treatment firms (reducing by about one third). However, we controlled for industry fixed effects in the difference-in-difference models with PSM control firms.

BETA	-	-0.07
		(-0.53)
Year Fixed Effects		Yes
Pseudo R2		0.145
N		1441

Note: ***p < 0.01, **p < 0.05, *p < 0.10 (two-tailed tests). Z-scores are standard errors adjusted for firm clustering effects. Control variable descriptions are provided in Table A8.

In the “Propensity Score Matching” section in “Robustness Tests”, we matched control and treatment firms that have propensity scores within 5% of the treatment firm, without replacement. Below we present the sensitivity of results under alternative PSM design choices. Overall, our conclusion that AI adoption is associated with more accurate management earnings forecast holds under different matching thresholds and with or without replacement.

Table A14. Results Under Alternative PSM Design Choices

	With replacement				**Without replacement**			
	One to One	**1%**	**3%**	**5%**	**One to One**	**1%**	**3%**	**5%**
TREATMENT*POST	0.004**	0.003**	0.003**	0.003**	0.003**	0.003*	0.002	0.002**
	(2.10)	(1.74)	(1.94)	(2.08)	(1.76)	(1.63)	(1.35)	(1.74)
TREATMENT	0.002	0.002	0.002	0.001	0.003*	0.002	0.002**	0.001
	(1.13)	(1.36)	(1.55)	(1.31)	(1.78)	(1.44)	(2.06)	(1.49)
POST	-0.002	-0.001	0.000	-0.001	-0.001	0.000	-0.001	-0.001
	(-0.84)	(-0.49)	(-0.56)	(-0.96)	(-0.56)	(-0.14)	(-0.97)	(-1.11)
CONTROLS	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adj. R2	0.369	0.376	0.485	0.479	0.351	0.298	0.332	0.318
N	423	553	1447	2234	381	421	750	883

BIBLIOGRAPHY

Agrawal, A., J. Gans, and A. Goldfarb. (2018). Prediction machines: the simple economics of artificial intelligence. Harvard Business Press.

Abdolmohammadi, M. J. (1999). A comprehensive taxonomy of audit task structure, professional rank and decision aids for behavioral research. Behavioral Research in Accounting, 11: 51-92.

Ahn, J., Hoitash, R., and Hoitash, U. (2020). Auditor task-specific expertise: The case of fair value accounting. The Accounting Review, 95(3), 1-32.

Ajinkya, B., S. Bhojraj, and P. Sengupta. 2005. The association between outside directors, institutional investors and the properties of management earnings forecasts. Journal of accounting research, 43(3), 343-376.

Akita, R., A. Yoshihara, T. Matsubara, and K. Uehara. (2016). Deep learning for stock prediction using numerical and textual information. In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS) (pp. 1-6). IEEE.

Al-Surmi, A., M. Bashiri, and I. Koliousis, (2021). AI based decision making: combining strategies to improve operational performance. International Journal of Production Research, 1-23.

Alles, M., A. Kogan, and M. Vasarhelyi. (2008). Audit automation for implementing continuous auditing: Principles and problems. Available at: https://www.researchgate.net/publication/228530458_Audit_Automation_for_Implementing_Continuous_Auditing_Principles_and_Problems

Alpaydin, E. (2014). Introduction to Machine Learning (3rd ed.). The MIT Press.

Anantharaman, D., and Wans, N. (2019). Audit office experience with SOX 404 (b) filers and SOX 404 audit quality. The Accounting Review, 94(4), 1-43.

Angrist, J. D., and Pischke, J. S. (2008). Mostly harmless econometrics. Princeton university press.

Aobdia, D. (2018). The impact of the PCAOB individual engagement inspection process—Preliminary evidence. The Accounting Review, 93(4), 53-80.

Aobdia, D. (2019). Do practitioner assessments agree with academic proxies for audit quality? Evidence from PCAOB and internal inspections. Journal of Accounting and Economics, 67(1), 144-174.

Aobdia, D., P. Choudhary, and N. Newberger. (2019). The economics of audit production: What matters for audit quality?. Available at: https://ssrn.com/abstract=3300277 or http://dx.doi.org/10.2139/ssrn.3300277

Appelbaum, D., and S. Kozlowski. (2020). The inherent risk assessment of an RPA-enabled accounting information system. Working paper, Montclair State University.

Appelbaum, D., S. Budnik, and M. Vasarhelyi. (2020). Auditing and accounting during and after the COVID-19 crisis. The CPA Journal 906: 14-19.

Arnold, V., P. A. Collier, S. A. Leech, S. G. Sutton, and A. Vincent. (2013). Incase: Simulating experience to accelerate expertise development by knowledge workers. Intelligent Systems in Accounting, Finance and Management 20 (1): 1–21.

AS 1001: Responsibilities and Functions of the Independent Auditor. Available at: https://pcaobus.org/Standards/Auditing/Pages/AS1001.aspx

Asthana, S. C., and Boone, J. P. (2012). Abnormal audit fee and audit quality. Auditing: A Journal of Practice and Theory, 31(3), 1-22.

Audit Analytics. (2020). 2019 Financial Restatements: A Nineteen Year Comparison. Available at: http://auditanalytics-2019restatementreport-j.pagedemo.co/

Baginski, S. P., and J. M. Hassell. (1997). Determinants of management forecast precision. Accounting Review, 303-312.

Balsam, S., Krishnan, J., and Yang, J. S. (2003). Auditor industry specialization and earnings quality. Auditing: A journal of practice and Theory, 22(2), 71-97.

Bao, Y., Ke, B., Li, B., Yu, Y. J., and Zhang, J. (2020). Detecting Accounting Fraud in Publicly Traded US Firms Using a Machine Learning Approach. Journal of Accounting Research.

Bartov, E., L. Faurel, and P. S. Mohanram. (2018). Can Twitter help predict firm-level earnings and stock returns?. The Accounting Review, 93(3), 25-57.

Basu, S. (2012). How can accounting researchers become more innovative? Accounting Horizons, 26 (4): 851–870.

Bell, T. B., Causholli, M., and Knechel, W. R. (2015). Audit firm tenure, non-audit services, and internal assessments of audit quality. Journal of Accounting Research, 53(3), 461-509.

Beneish, M. D. (1997). Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance. Journal of Accounting and Public Policy 16: 271–309.

Beneish, M. D. (1999). The detection of earnings manipulation. Financial Analysts Journal (September/October): 24–36.

Beneish, M. D. and Vorst, P. (2020). The Cost of Fraud Prediction Errors. Kelley School of Business Research Paper No. 2020-55, Available at SSRN: https://ssrn.com/abstract=3529662 or http://dx.doi.org/10.2139/ssrn.3529662

Bergstein, B. (2019). This is why AI has yet to reshape most businesses. MIT Technology Review.

Bertomeu, J. (2020). Machine learning improves accounting: discussion, implementation and research opportunities. Rev Account Stud. https://doi.org/10.1007/s11142-020-09554-9

Bertomeu, J., Cheynel, E., Floyd, E., and Pan, W. (2020). Using Machine Learning to Detect Misstatements. Review of Accounting Studies, Forthcoming, Available at SSRN: https://ssrn.com/abstract=3496297

Beyer, A. and R. A. Dye. (2012). Reputation management and the disclosure of earnings forecasts. Review of Accounting Studies, 17(4), 877-912.

Bhaskar, L. S., Schroeder, J. H., and Shepardson, M. L. (2019). Integration of internal control and financial statement audits: Are two audits better than one? The Accounting Review, 94(2), 53-81.

Bhojraj, S., Libby, R., and Yang, H. (2011). Guidance frequency and guidance properties: The effect of reputation-building and learning-by-doing. Johnson School Research Paper Series, (12-2010).

Bills, K. L., Cunningham, L. M., and Myers, L. A. (2016). Small audit firm membership in associations, networks, and alliances: Implications for audit quality and audit fees. The Accounting Review, 91(3), 767-792.

Blankley, A. I., Hurtt, D. N., and MacGregor, J. E. (2012). Abnormal Audit Fees and Restatements. Auditing: A Journal of Practice and Theory, 31(1), 79–96.

Boland, C. M., B. E. Daugherty, and D. Dickins. (2019). Evidence of the relationship between PCAOB inspection outcomes and the use of structured audit technologies. Auditing: A Journal of Practice and Theory 382: 57-77.

Boone, J. P., Khurana, I. K., and Raman, K. K. (2010). Do the Big 4 and the second-tier firms provide audits of similar quality?. Journal of accounting and public policy, 29(4), 330-352.

Bowman, E. H. (1963). Consistency and optimality in managerial decision making. Management Science, 9(2), 310-321.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7), 1145-1159.

Brazel, J. F. and L. Dang. (2008). The effect of ERP system implementations on the management of earnings and earnings release dates. Journal of information systems, 22(2), 1-21.

Bresnahan, T., S. Greenstein, D. Brownstone, and K. Flamm. (1996). Technical progress and co-invention in computing and in the uses of computers. Brookings Papers on Economic Activity. Microeconomics, 1996, 1-83.

Brown, L. D., and Caylor, M. L. (2009). Corporate governance and firm operating performance. Review of quantitative finance and accounting, 32(2), 129-144.

Brown, N. C., Crowley, R. M., and Elliott, W. B. (2020). What are you saying? Using topic to detect financial misreporting. Journal of Accounting Research, 58(1), 237-291.

Brundage, M., S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel,... and D. Amodei. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.

Brynjolfsson, E., D. Rock, and C. Syverson. (2018a). Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. In The economics of artificial intelligence: An agenda (pp. 23-57). University of Chicago Press.

Brynjolfsson, E., T. Mitchell, and D. Rock. (2018b). What can machines learn, and what does it mean for occupations and the economy? In AEA Papers and Proceedings (Vol. 108, pp. 43-47).

Burton, F. G., S. L. Summers, T. J. Wilks, and D. A. Wood. (2020). Creating relevance of accounting research ROAR scores to evaluate the relevance of accounting research to practice. Accounting Horizons, Forthcoming.

Burton, F. G., S. L. Summers, T. J. Wilks, and D. A. Wood. (2021). Do we matter? Attention the general public, policymakers, and academics give to accounting research. Issues in Accounting Education 36 (1), 1-22.

Cao, J., Chen, F., and Higgs, J. L. (2016). Late for a very important date: financial reporting and audit implications of late 10-K filings. Review of Accounting Studies, 21(2), 633-671.

Cao, Y., Myers, L. A., and Omer, T. C. (2012). Does company reputation matter for financial reporting quality? Evidence from restatements. Contemporary Accounting Research, 29(3), 956-990.

Carey, P., and Simnett, R. (2006). Audit partner tenure and audit quality. The accounting review, 81(3), 653-676.

Carmona, P., Climent, F., and Momparler, A. (2019). Predicting failure in the US banking sector: An extreme gradient boosting approach. International Review of Economics and Finance, 61, 304-323.

Cassar, G. and B. Gibson. (2008). Budgets, internal reports, and manager forecast accuracy. Contemporary Accounting Research, 25(3), 707-738.

Causholli, M., and Knechel, W. R. (2012). An examination of the credence attributes of an audit. Accounting Horizons, 26(4), 631-656.

Cecchini, M., Aytug, H., Koehler, G. J., and Pathak, P. (2010). Detecting Management Fraud in Public Companies. Management Science, 56(7), 1146–1160.

Center of Audit Quality. (2013). Financial Restatement Trends in the United States: 2003-2012. Available at: https://www.thecaq.org/wp-content/uploads/2019/03/financial-restatement-trends-in-the-united-states-2003-2012.pdf

Chan, P. K., and Stolfo, S. J. (1998). Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. In KDD (Vol. 1998, pp. 164-168).

Chen, C., X. Martin, S. Roychowdhury, X. Wang, and M. T. Billett. (2018). Clarity begins at home: Internal information asymmetry and external communication quality. The Accounting Review, 93(1), 71-101.

Chen, X., Cho, Y. H., Dou, Y., and Lev, B. (2020). Fundamental Analysis of XBRL Data: A Machine Learning Approach. Available at SSRN.

Cheng, Q., T. Luo, and H. Yue. (2013). Managerial incentives and management forecast precision. The Accounting Review, 88(5), 1575-1602.

Cheng, Q., Y. J. Cho, and H. Yang. (2018). Financial reporting changes and the internal information environment: Evidence from SFAS 142. Review of Accounting Studies, 23(1), 347-383.

Cheng, S. F., G. De Franco, H. Jiang, and P. Lin. (2019). Riding the blockchain mania: public firms' speculative 8-K disclosures. Management Science, 65(12), 5901-5913.

Cho, S., M. A. Vasarhelyi, and C. Zhang. (2019). The forthcoming data ecosystem for business measurement and assurance. Journal of Emerging Technologies in Accounting.16 (2): 1–21

Choi, J. H., Kim, C., Kim, J. B., and Zang, Y. (2010). Audit office size, audit quality, and audit pricing. Auditing: A Journal of practice and theory, 29(1), 73-97.

Christ, M. H., M. Eulerich, R. Krane, and D. A. Wood. (2020a). New frontiers for internal audit research. (June 8). Available at SSRN: https://ssrn.com/abstract=3622148 or http://dx.doi.org/10.2139/ssrn.3622148

Christ, M. H., S. A. Emett, S. L. Summers, and D. A. Wood. (2020b). Prepare for takeoff: Improving audit efficiency and effectiveness with drone-enabled inventory audit procedures. Review of Accounting Studies.

Christensen, B. E., Glover, S. M., Omer, T. C., and Shelley, M. K. (2016). Understanding audit quality: Insights from audit professionals and investors. Contemporary Accounting Research, 33(4), 1648-1684.

Christensen, C. M. (2013). The innovator's dilemma: When new technologies cause great firms to fail. Harvard Business Review Press.

Clement, M., R. Frankel, and J. Miller. (2003). Confirming management earnings forecasts, earnings uncertainty, and stock returns. Journal of Accounting Research, 41(4), 653-679.

Cockburn, I. M., R. Henderson, and S. Stern. (2018). The impact of artificial intelligence on innovation (No. w24449). National bureau of economic research.

Cohen, J. R., Hoitash, U., Krishnamoorthy, G., and Wright, A. M. (2014). The effect of audit committee industry expertise on monitoring the financial reporting process. The Accounting Review, 89(1), 243-273.

Cohen, M., A. Rozario, and C. Zhang. (2019). Exploring the use of robotic process automation RPA in substantive audit procedures. The CPA Journal 897: 49-53.

Cohn, M. (2020). Accountants leverage RPA amid coronavirus. Available at: https://www.accountingtoday.com/news/accountants-leverage-rpa-amid-coronavirus?utm_medium=socialandutm_source=twitterandutm_campaign=accountingtoday-twandutm_content=socialflow

Coller, M., and T. L. Yohn. (1997). Management forecasts and information asymmetry: An examination of bid-ask spreads. Journal of accounting research, 35(2), 181-191.

Colson, E. (2019). What AI-driven decision making looks like. Harvard Business Review. Available at: https://hbr.org/2019/07/what-ai-driven-decision-making-looks-like

Cooper, L. A., D. K. Holderness Jr, T. L. Sorensen, and D. A. Wood. (2019). Robotic process automation in public accounting. Accounting Horizons 33 (4): 15-35.

Cooper, L., K. Holderness, T. Sorensen, and D. A. Wood. (2021). Perceptions of robotic process automation in public accounting. Journal of Emerging Technologies in Accounting. Forthcoming.

Cunningham, L. M., Li, C., Stein, S. E., and Wright, N. S. (2019). What's in a name? Initial evidence of US audit partner identification using difference-in-differences analyses. The Accounting Review, 94(5), 139-163.

Dash, M., and Liu, H. (1997). Feature selection for classification. Intelligent data analysis, 1(1-4), 131-156.

Daugherty, P. R., and H. J. Wilson. (2018). Human+ machine: Reimagining work in the age of AI. Harvard Business Press.

Davenport, T. H. and R. Ronanki. (2018). Artificial intelligence for the real world. Harvard business review, 96(1), 108-116.

Davenport, T. H., and D. B. Stoddard. (1994). Reengineering: business change of mythic proportions?. MIS Quarterly: 121-127.

Davenport, T. H., and J. E. Short. (1990). The new industrial engineering: information technology and business process redesign. MIT Sloan Management Review (July 15). Available at : https://sloanreview.mit.edu/article/the-new-industrial-engineering-information-technology-and-business-process-redesign/

DeAngelo, L. E. (1981). Auditor size and audit quality. Journal of accounting and economics, 3(3), 183-199.
Dechow, P. M., and Dichev, I. D. (2002). The quality of accruals and earnings: The role of accrual estimation errors. The accounting review, 77(s-1), 35-59.
Dechow, P. M., Ge, W., Larson, C. R., and Sloan, R. G. (2011). Predicting material accounting misstatements. Contemporary accounting research, 28(1), 17-82.
Dechow, P. M., Sloan, R. G., and Sweeney, A. P. (1995). Detecting earnings management. Accounting review, 193-225.
DeFond, M., and Zhang, J. (2014). A Review of Archival Auditing Research. Journal of Accounting and Economics, 58(2–3), 275–326.
DeFond, M., Erkens, D. H., and Zhang, J. (2017). Do client characteristics really drive the Big N audit quality effect? New evidence from propensity score matching. Management Science, 63(11), 3628-3649.
deHaan, E., (2021). Using and interpreting fixed effects models. Available at SSRN 3699777.
Deloitte. (2019). State of AI in the Enterprise, 2nd Edition. Available at: https://www2.deloitte.com/content/dam/insights/us/articles/4780_State-of-AI-in-the-enterprise/DI_State-of-AI-in-the-enterprise-2nd-ed.pdf
Ding, K., Lev, B., Peng, X., Sun, T., and Vasarhelyi, M. A. (2020). Machine learning improves accounting estimates: evidence from insurance payments. Review of Accounting Studies, 1-37.
Dorantes, C. A., C. Li, G. F. Peters, and V. J. Richardson. (2013). The effect of enterprise systems implementation on the firm information environment. Contemporary Accounting Research, 30(4), 1427-1461.
Dutta, I., Dutta, S., and Raahemi, B. (2017). Detecting financial restatements using data mining techniques. Expert Systems with Applications, 90, 374–393.
Earl, M. J. (1994). The new and the old of business process redesign. The Journal of Strategic Information Systems 31: 5-22.
Einhorn, E., and Ziv, A. (2008). Intertemporal dynamics of corporate voluntary disclosures. Journal of Accounting Research, 46(3), 567-589.
Elkan, C. (2001). The foundations of cost-sensitive learning. In International joint conference on artificial intelligence (Vol. 17, No. 1, pp. 973-978). Lawrence Erlbaum Associates Ltd.
Eshleman, J. D., and Guo, P. (2014). Do Big 4 Auditors Provide Higher Audit Quality After Controlling for the Endogenous Choice of Auditor? Auditing: A Journal of Practice and Theory, 33(4), 197–220.
Ettredge, M., Fuerherm, E. E., and Li, C. (2014). Fee pressure and audit quality. Accounting, Organizations and Society, 39(4), 247-263.
Eulerich, M., J. Pawlowski, N. Waddoups, and D. A. Wood. (2021). Using robotic process automation in the internal audit function: Use cases and framework guidance for evaluation. Contemporary Accounting Research (forthcoming).
Eutsler, J. (2020). Risky Business: Assessing the Generalizability of the PCAOB's Risk-Based Inspection Reports for Annually Inspected Audit Firms. AUDITING: A Journal of Practice and Theory, 39(2), 117-138.
Faraggi, D., and Reiser, B. (2002). Estimation of the area under the ROC curve. Statistics in medicine, 21(20), 3093-3106.

Feng, G., J. He, and N. G. Polson. (2018). Deep learning for predicting asset returns. arXiv preprint arXiv:1804.09314.

Feng, M., C. Li, and S. McVay. (2009). Internal control and management guidance. Journal of accounting and economics, 48(2-3), 190-209.

Ferguson, A., Francis, J. R., and Stokes, D. J. (2003). The effects of firm-wide and office-level industry expertise on audit pricing. The accounting review, 78(2), 429-448.

Financial Times (2020). After Wirecard: is it time to audit the auditors? Available at: https://www.ft.com/content/b220719a-edca-4ebf-b6bc-5f7a67078745

Francis, J. R. (2004). What do we know about audit quality? The British accounting review, 36(4), 345-368.

Francis, J. R. (2011). A framework for understanding and researching audit quality. Auditing: A journal of practice and theory, 30(2), 125-152.

Francis, J. R., and Michas, P. N. (2013). The contagion effect of low-quality audits. The Accounting Review, 88(2), 521-552.

Francis, J. R., and Yu, M. D. (2009). Big 4 Office Size and Audit Quality. The Accounting Review, 84(5), 1521–1552.

Francis, J. R., Michas, P. N., and Yu, M. D. (2013). Office size of Big 4 auditors and client restatements. Contemporary Accounting Research, 30(4), 1626-1661.

Frankel, R., M. McNichols, and G. P. Wilson. (1995). Discretionary disclosure and external financing. Accounting Review, 135-150.

Gartner. (2019). Magic Quadrant for Robotic Process Automation Software.

Gaynor, L. M., Kelton, A. S., Mercer, M., and Yohn, T. L. (2016). Understanding the relation between financial reporting quality and audit quality. Auditing: A Journal of Practice and Theory, 35(4), 1-22.

Gentry, J. A., Shaw, M. J., Tessmer, A. C., and Whitford, D. T. (2002). Using inductive learning to predict bankruptcy. Journal of Organizational Computing and Electronic Commerce, 12(1), 39-57.

Glendening, M., E. G. Mauldin, and K. W. Shaw. (2019). Determinants and consequences of quantitative critical accounting estimate disclosures. The Accounting Review, 94(5), 189-218.

Graham, J. R., C. R. Harvey, and S. Rajgopal. (2005). The economic implications of corporate financial reporting. Journal of accounting and economics, 40(1-3), 3-73.

Gray, A. M. (2018). SAP Leonardo: Making Machine Learning and Artificial Intelligence. Retrieved from: https://www.asug.com/insights/sap-leonardo-making-machine-learning-and-artificial-intelligence-real

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political analysis, 25-46.

Hainmueller, J., and Y. Xu. (2013). Ebalance: A Stata package for entropy balancing. Journal of Statistical Software, 54(7).

Hall, M. A., and Smith, L. A. (1998). Practical feature subset selection for machine learning.

Hastie, T., Tibshirani, R., and Friedman, J. (2017b). The elements of statistical learning. New York: Springer series in statistics.

Hastie, T., Tibshirani, R., and Tibshirani, R. J. (2017a). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv preprint arXiv:1707.08692.

Hayes, B. L. and Boritz, E. (2019). Classifying Restatements: An Application of Machine Learning and Textual Analytics. Working paper. University of Guelph. Available at SSRN: https://ssrn.com/abstract=2716166 or http://dx.doi.org/10.2139/ssrn.2716166

Hayward, M. L., and M. A. Fitza. (2017). Pseudo-precision? Precise forecasts and impression management in managerial earnings forecasts. Academy of Management Journal, 60(3), 1094-1116.

Heitzman, S., and Huang, M. (2019). Internal information quality and the sensitivity of investment to market prices and accounting profits. Contemporary Accounting Research.

Hevner, A., and S. Chatterjee. (2010). Design science research in information systems. In Design research in information systems : 9-22. Boston, MA: Springer.

Hevner, A., S. T. March, J. Park, and S. Ram. (2004). Design science research in information systems. MIS Quarterly 281: 75-105.

Hilary, G., C. Hsu, and R. Wang. (2014). Management forecast consistency. Journal of Accounting Research, 52(1), 163-191.

Hocking, R. R., and Leslie, R. N. (1967). Selection of the best subset in regression analysis. Technometrics, 9(4), 531-540.

Hofmann, P., C. Samp, and N. Urbach. (2020). Robotic process automation. Electronic Markets 301: 99-106.

Honigsberg, C., Rajgopal, S., and Srinivasan, S. (2019). The Changing Landscape of Auditor Liability. Journal of Law and Economics, Forthcoming.

Hope, O. K., W. B. Thomas, and D. Vyas. (2013). Financial reporting quality of US private and public firms. The Accounting Review, 88(5), 1715-1742.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley and Sons.

Hribar, P., and Yang, H. (2016). CEO overconfidence and management forecasting. Contemporary accounting research, 33(1), 204-227.

Huang, F., and M. A. Vasarhelyi. (2019). Applying robotic process automation RPA in auditing: A framework. International Journal of Accounting Information Systems 35: 100433.

Huang, T. C., Chang, H., and Chiou, J. R. (2016). Audit market concentration, audit fees, and audit quality: Evidence from China. Auditing: A Journal of Practice & Theory, 35(2), 121-145.

Huang, Y., and Scholz, S. (2012). Evidence on the association between financial restatements and auditor resignations. Accounting Horizons, 26(3), 439-464.

Hunt, E., Hunt, J. O., Richardson, V. J., and Rosser, D. (2022). Auditor Response to Estimated Misstatement Risk: A Machine Learning ApproachAuditor Response to Estimated Misstatement Risk. Accounting Horizons.

Hurwitz, H. (2018). Investor sentiment and management earnings forecast bias. Journal of Business Finance & Accounting, 45(1-2), 166-183.

IEEE-USA. (2017). Artificial Intelligence Research, Development and Regulation. Available at: http://globalpolicy.ieee.org/wp-content/uploads/2017/10/IEEE17003.pdf

Institute of Chartered Accountants in England and Wales (ICAEW). (1993). Audit Automation, IT Briefing Number 4. Chartech Books.

Institute of Electrical and Electronics Engineers (IEEE). (2017). IEEE guide for terms and concepts in intelligent process automation. Available at: https://ieeexplore.ieee.org/document/8070671

Issa, H., T. Sun, and M. A. Vasarhelyi. (2016). Research ideas for artificial intelligence in auditing: The formalization of audit and workforce supplementation. Journal of Emerging Technologies in Accounting, 13(2), 1-20.

Ittner, C. D., and J. Michels. (2017). Risk-based forecasting and planning and management earnings forecasts. Review of Accounting Studies, 22(3), 1005-1047.

Jackson, D. E. N. I. S. E., Michelson, G. R. A. N. T., and Munir, R. A. H. A. T. (2020). The impact of technology on the desired skills of early career accountants. CPA Australia. Retrieved July, 15, 2021.

Jans, M., M. G. Alles, and M. A. Vasarhelyi. (2014). A field study on the use of process mining of event logs as an analytical procedure in auditing. The Accounting Review 895:1751-1773.

Jeni, L. A., Cohn, J. F., and De La Torre, F. (2013). Facing imbalanced data--recommendations for the use of performance metrics. In 2013 Humaine association conference on affective computing and intelligent interaction (pp. 245-251). IEEE.

Jennings, J. N., H. Seo, and L. Tanlu. (2013). The effect of organizational complexity on earnings forecasting behavior. AAA 2013 Management Accounting Section (MAS) Meeting Paper.

Jiang, J., Wang, I. Y., and Wang, K. P. (2019). Big N auditors and audit quality: New evidence from quasi-experiments. The Accounting Review, 94(1), 205-227.

Jiang, L., Vasarhelyi, M., and Zhang, C. A. (2022). Using Semi-Supervised Learning to Detect and Predict Unlabeled Restatements. Available at SSRN 4003621.

Johnson, V. E., Khurana, I. K., and Reynolds, J. K. (2002). Audit-firm tenure and the quality of financial reports. Contemporary accounting research, 19(4), 637-660.

Kaplan, R. S. (2011). Accounting scholarship that advances professional knowledge and practice. The Accounting Review 86 (2): 367-383.

Karpoff, J. M., Koester, A., Lee, D. S., and Martin, G. S. (2017). Proxies and databases in financial misconduct research. The Accounting Review, 92(6), 129-163.

Kartal, H., A. Oztekin, A. Gunasekaran, and F. Cebi. (2016). An integrated decision analytic framework of machine learning with multi-criteria decision making for multi-attribute inventory classification. Computers and Industrial Engineering, 101, 599-613.

Keenoy, C. L. (1958). The impact of automation on the field of accounting. The Accounting Review: 230-236.

Kinney, W. R., Palmrose, Z.-V., and Scholz, S. (2004). Auditor Independence, Non-Audit Services, and Restatements: Was the U. S. Government Right? Journal of Accounting Research, 42(3), 561–588.

Knechel, W. R., and Sharma, D. S. (2012). Auditor-provided nonaudit services and audit effectiveness and efficiency: Evidence from pre-and post-SOX audit report lags. Auditing: A Journal of Practice and Theory, 31(4), 85-114.

Knechel, W. R., and Vanstraelen, A. (2007). The relationship between auditor tenure and audit quality implied by going concern opinions. AUDITING: A journal of practice and theory, 26(1), 113-131.

Knechel, W. R., Krishnan, G. V., Pevzner, M., Shefchik, L. B., and Velury, U. K. (2013). Audit quality: Insights from the academic literature. Auditing: A Journal of Practice and Theory, 32(sp1), 385-421.

Kobelsky, K. W., V. J. Richardson, R. E. Smith, and R. W. Zmud. (2008). Determinants and consequences of firm information technology budgets. The Accounting Review, 83(4), 957-995.

Kokina, J., and S. Blanchette. (2019). Early evidence of digital labor in accounting: Innovation with robotic process automation. International Journal of Accounting Information Systems 35: 100431.

Kokina, J., and T. H. Davenport. (2017). The emergence of artificial intelligence: How automation is changing auditing. Journal of emerging technologies in accounting, 14(1), 115-122.

Kokina, J., R. Gilleran, S. Blanchette, and D. Stoddard. (2021). Accountant as digital innovator: Roles and competencies in the age of automation. Accounting Horizons 351: 153-184.

Kothari, S. P., Leone, A. J., and Wasley, C. E. (2005). Performance matched discretionary accrual measures. Journal of accounting and economics, 39(1), 163-197.

Krieger, F., and P. Drews. (2018). Leveraging big data and analytics for auditing: Towards a taxonomy. Available at: https://www.researchgate.net/profile/Paul-Drews/publication/328902212_Leveraging_Big_Data_and_Analytics_for_Auditing_Towards_a_Taxonomy/links/5beaa034a6fdcc3a8dd2122d/Leveraging-Big-Data-and-Analytics-for-Auditing-Towards-a-Taxonomy.pdf

Krishnan, J., and Krishnan, J. (1997). Litigation risk and auditor resignations. Accounting Review, 539-560.

Krishnan, J., Krishnan, J., and Song, H. (2017). PCAOB international inspections and audit quality. The Accounting Review, 92(5), 143-166.

Lacity, M. C., and L. P. Willcocks. (2017). Robotic process automation and risk mitigation. United Kingdom: SB Publishing.

Lacity, M., and L. P. Willcocks. (2018). Robotic process and cognitive automation: The next phase. SB Publishing.

Lacity, M., L. P. Willcocks, and A. Craig. (2015). Robotic process automation at Telefonica O2.

Lambert, T. A., Jones, K. L., Brazel, J. F., and Showalter, D. S. (2017). Audit time pressure and earnings quality: An examination of accelerated filings. Accounting, Organizations and Society, 58, 50-66.

Laplante, S. K., Lynch, D. P., and Vernon, M. E. (2021). Internal Information Quality and State Tax Planning. Contemporary Accounting Research, 38(4), 2589-2621.

Larcker, D. F., and Rusticus, T. O. (2010). On the use of instrumental variables in accounting research. Journal of accounting and economics, 49(3), 186-205.

Larcker, D. F., Richardson, S. A., and Tuna, I. R. (2007). Corporate governance, accounting outcomes, and organizational performance. The accounting review, 82(4), 963-1008.

Lennox, C. S. (2016). Did the PCAOB's restrictions on auditors' tax services improve audit quality? The Accounting Review, 91(5), 1493-1512.

Lennox, C. S., J. R. Francis, and Z. Wang. (2012). Selection models in accounting research. The accounting review, 87(2), 589-616.

Lennox, C., and Li, B. (2020). When Are Audit Firms Sued for Financial Reporting Failures and What Are the Lawsuit Outcomes? Contemporary Accounting Research.

Leuz, C., Nanda, D., and Wysocki, P. D. (2003). Earnings management and investor protection: an international comparison. Journal of financial economics, 69(3), 505-527.

Li, C., Peters, G. F., Richardson, V. J., and Watson, M. W. (2012). The consequences of information technology control weaknesses on management information systems: The case of Sarbanes-Oxley internal control reports. Mis Quarterly, 179-203.

Li, F. (2010). The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. Journal of Accounting Research, 48(5), 1049-1102.

Li, K., and Prabhala, N. R. (2007). Self-selection models in corporate finance. Handbook of empirical corporate finance, 37-86.

Li, L., Qi, B., Tian, G., and Zhang, G. (2017). The contagion effect of low-quality audits at the level of individual auditors. The Accounting Review, 92(1), 137-163.

Lim, C. Y., and Tan, H. T. (2008). Non-audit service fees and audit quality: The impact of auditor specialization. Journal of accounting research, 46(1), 199-246.

Lim, C. Y., Tan, H. T., and Cheng, Q. (2010). Does Auditor Tenure Improve Audit Quality? Moderating Effects of Industry Specialization and Fee Dependence. Contemporary Accounting Research, 27(3), 923.

Lim, J.H., J. Park, G.F., Peters, V.J. Richardson. (2020). Examining the Potential Benefits of Audit Data Analytics. University of Hawaii. Working paper.

Lobo, G. J., and Zhao, Y. (2013). Relation between audit effort and financial report misstatements: Evidence from quarterly and annual restatements. The Accounting Review, 88(4), 1385-1412

Lombardi, D.R. and R.B. Dull. (2016). The development of AudEx: An audit data assessment system. Journal of Emerging Technologies in Accounting 13 (1): 37–52.

López, D. M., and Peters, G. F. (2012). The effect of workload compression on audit quality. Auditing: A Journal of Practice and Theory, 31(4), 139-165.

Louwers, T. J., R. J. Ramsay, D.Sinason, and J.Strawser, Eds.. (2015). Auditing and assurance services. McGraw-Hill Education.

Mactavish, C., S. McCracken, and R. N. Schmidt. (2018). External auditors' judgment and decision making: An audit process task analysis. Accounting Perspectives 173: 387-426.

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. Journal of Thoracic Oncology, 5(9), 1315-1316.

Manson, S., S. McCartney, M. Sherer, and W. A. Wallace. (1998). Audit automation in the UK and the US: A comparative study. International Journal of Auditing 23: 233-246.

Masli, A., G. F. Peters, V. J. Richardson, and J. M. Sanchez. (2010). Examining the potential benefits of internal control monitoring technology. The Accounting Review, 85(3), 1001-1034.

Matsumoto, D.A., (2002). Management's incentives to avoid negative earnings surprises. The accounting review, 77(3), pp.483-514.

McKinsey. (2021). Global AI Survey: The state of AI in 2021, 2021. Available at: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2021

McNichols, M. F. (2002). Discussion of the quality of accruals and earnings: The role of accrual estimation errors. The accounting review, 77(s-1), 61-69.

Minutti-Meza, M. (2013). Does auditor industry specialization improve audit quality?. Journal of Accounting Research, 51(4), 779-817.

Moffitt, K. C., A. M. Rozario, and M. A. Vasarhelyi. (2018). Robotic process automation for auditing. Journal of Emerging Technologies in Accounting, 15(1), 1-10.

Moffitt, K. C., V. J. Richardson, N. M. Snow, M. M. Weisner, and D. A. Wood. (2016). Perspectives on past and future AIS research as the Journal of Information Systems turns thirty. Journal of Information Systems 303: 157-171.

Molnar, C. (2020). Interpretable machine learning. Lulu. com.

Mullakara, N., and A. K. Asokan. (2020). Robotic process automation projects: Build real-world RPA solutions using UiPath and Automation Anywhere. Packt Publishing. ISBN 978-1-83921-735-7.

Munoko, I., H. L. Brown-Liburd, and M. Vasarhelyi. (2020). The ethical implications of using artificial intelligence in auditing. Journal of Business Ethics, 1-26.

Murthy, U. S., T. J. Smith, J. D. Whitworth, and Y. Zhang. (2020). The effects of information systems compatibility on firm performance following mergers and acquisitions. Journal of Information Systems, 34(2), 211-233.

Myers, J. N., Myers, L. A., and Omer, T. C. (2003). Exploring the term of the auditor-client relationship and the quality of earnings: A case for mandatory auditor rotation? The accounting review, 78(3), 779-799.

Nelson, M., and H. T. Tan. (2005). Judgment and decision making research in auditing: A task, person, and interpersonal interaction perspective. Auditing: A Journal of Practice and Theory 24 (s-1): 41-71.

Newton, N. J., Wang, D., and Wilkins, M. S. (2013). Does a lack of choice lead to lower quality? Evidence from auditor competition and client restatements. Auditing: A Journal of Practice and Theory, 32(3), 31-67.

Nicolaou, A. I. (2004). Firm performance effects in relation to the implementation and use of enterprise resource planning systems. Journal of information systems, 18(2), 79-105.

No, W. G., K. Lee, F. Huang, and Q. Li. (2019). Multidimensional audit data selection MADS: A framework for using data analytics in the audit data selection process. Accounting Horizons 333: 127-140.

Noe, C.F., (1999). Voluntary disclosures and insider transactions. Journal of Accounting and Economics, 27(3), pp.305-326.

Nunamaker Jr, J. F., M. Chen, and T. D. Purdin. (1990). Systems development in information systems research. Journal of Management Information Systems 73: 89-106.

Ostdick, N. (2017). Attended Or unattended RPA? Advantages for both solutions. Available at: https://www.uipath.com/blog/unattended-attended-automation

Palmer, S. (2017). The 5 jobs robots will take first. Available at: https://www.cnbc.com/2017/02/28/the-5-jobs-robots-will-take-first.html

Paredes, A. A. P., and C. M. Wheatley. (2018). Do enterprise resource planning systems (ERPs) constrain real earnings management?. Journal of information systems, 32(3), 65-89.

Paterson, J. S., and Valencia, A. (2011). The effects of recurring and nonrecurring tax, audit-related, and other nonaudit services on auditor independence. Contemporary Accounting Research, 28(5), 1510-1536.

Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. Data, 4(1), 15.

PCAOB. (2013). Standing Advisory Group Meeting, Discussion-Audit Quality Indicators. Available at: https://pcaobus.org/news/events/documents/05152013_sagmeeting/audit_quality_indicators.pdf

PCAOB. (2015). Concept Release on Audit Quality Indicators. Available at: https://pcaobus.org/Rulemaking/Docket%20041/Release_2015_005.pdf

PCAOB. (2019). Strategic Plan 2019-2023. Available at: https://pcaobus.org/About/Administration/Documents/Strategic%20Plans/Strategic%20Plan-2019-2023.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Peffers, K., M. Rothenberger, T. Tuunanen, and R. Vaezi. (2012). Design science research evaluation. In International Conference on Design Science Research in Information Systems (May): 398-410. Berlin, Heidelberg: Springer.

Peffers, K., T. Tuunanen, M. A. Rothenberger, and S. Chatterjee. (2007). A design science research methodology for information systems research. Journal of Management Information Systems 243: 45-77.

Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. Auditing: A Journal of Practice and Theory, 30(2), 19-50.

Perols, J. L., R. M. Bowen, C. Zimmermann, and B. Samba. (2017). Finding needles in a haystack: Using data analytics to improve fraud prediction. The Accounting Review, 92(2), 221-245.

Perrault, R., Y. Shoham, E. Brynjolfsson, J. Clark, J. Etchemendy, B. Grosz, T. Lyons, J. Manyika, S. Mishra, and J. C. Niebles. (2019). Artificial Intelligence Index Report 2019. Available at: https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf

Pincus, M., F. Tian, P. Wellmeyer, and S. X. Xu. (2017). Do clients' enterprise systems affect audit quality and efficiency? Contemporary Accounting Research, 34(4), 1975-2021.

Price III, R. A., Sharp, N. Y., and Wood, D. A. (2011). Detecting and predicting accounting irregularities: A comparison of commercial and academic risk measures. Accounting Horizons, 25(4), 755-780.

Public Company Accounting Oversight Board (PCAOB) AS 2101. Auditing planning. Available at: https://pcaobus.org/Standards/Auditing/Pages/AS2101.aspx

Public Company Accounting Oversight Board (PCAOB). (2019). Staff Preview of 2018 Inspection Observations. Available at: https://pcaobus.org/Inspections/Documents/Staff-Preview-2018-Inspection-Observations.pdf

Rajgopal, S. (2020). Integrating practice into accounting research. Management Science 67 (9).

Rajgopal, S., Srinivasan, S. and Zheng, X. (2021). Measuring audit quality. Rev Account Stud. https://doi.org/10.1007/s11142-020-09570-9

Reichelt, K. J., and Wang, D. (2010). National and office-specific measures of auditor industry expertise and effects on audit quality. Journal of Accounting Research, 48(3), 647-686.

Rijsbergen, V. and Joost, C. (2004). The geometry of information retrieval. Cambridge University Press.

Rogers, J. L., and P. C. Stocken. (2005). Credibility of management forecasts. The Accounting Review, 80(4), 1233-1260.

Romanus, R. N., Maher, J. J., and Fleming, D. M. (2008). Auditor Industry Specialization, Auditor Changes, and Accounting Restatements. Accounting Horizons, 22(4), 389–413.

Rorzario, A.M., A.M. Vasarhelyi, and T.Wang. (2022). On the Use of Consumer Tweets to Assess Revenue Risk. Stevens Institute of Technology. Working Paper.

Rozario, A. M. (2019). Three essays on audit innovation: using social media information and disruptive technologies to enhance audit quality. Doctoral dissertation, Rutgers University-Graduate School-Newark.

Rozario, A. M., and M. A. Vasarhelyi. (2018). How robotic process automation is transforming accounting and auditing. The CPA Journal 886: 46-49.

Rozario, A. M., M. A.Vasarhelyi, and C. Zhang. (2019). Examining automation in audit. International Federation of Accountants. Available at: https://www.ifac.org/knowledge-gateway/preparing-future-ready-professionals/discussion/examining-automation-audit

Ruddock, C., Taylor, S. J., and Taylor, S. L. (2006). Nonaudit services and earnings conservatism: Is auditor independence impaired? Contemporary Accounting Research, 23(3), 701-746.

Russell, S., and P. Norvig. (2002). Artificial intelligence: a modern approach.

Rutschi, C., and J. Dibbern. (2020). Towards a framework of implementing software robots: Transforming human-executed routines into machines. ACM SIGMIS Database: the DATABASE for Advances in Information Systems 511: 104-128.

Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one, 10(3), e0118432.

Schroeder, J. H. (2016). The impact of audit completeness and quality on earnings announcement GAAP disclosures. The Accounting Review, 91(2), 677-705.

SEC (Securities and Exchange Commission). (2018). From the Data Rus to the Data Wars: A Data Revolution in Financial Markets. Available at: https://www.sec.gov/news/speech/speech-stein-092718

SEC (Securities and Exchange Commission). (2019).Policy and Challenges and Research Opportunities in the Era of Big Data. Available at: https://www.sec.gov/news/speech/policy-challenges-research-opportunities-era-big-

SEC (Securities and Exchange Commission). (2021). Finhub. Strategic Hub for Innovation and Financial Technology. Available at: https://www.sec.gov/finhub

SEC. (2017). How Investigations Work. Available at: https://www.sec.gov/enforce/how-investigations-work.html

Sheedy, C. (2017). What CPAs need to do to survive the automation revolution. The Journal of Accountancy.

Shim, S., and M. Pourhomayoun. (2017). Predicting movie market revenue using social media data. In 2017 IEEE International Conference on Information Reuse and Integration (IRI) (pp. 478-484). IEEE.

Shipman, J. E., Q. T. Swanquist, and R. L. Whited. (2017). Propensity score matching in accounting research. The Accounting Review, 92(1), 213-244.

Shmueli, G. (2010). To explain or to predict? Statistical science, 25(3), 289-310.

Shmueli, G., and Koppius, O. R. (2011). Predictive analytics in information systems research. MIS quarterly, 553-572.

Skinner, D. J. (1997). Earnings disclosures and stockholder lawsuits. Journal of accounting and economics, 23(3), 249-282.

Smith, A. (2016). Public predictions for the future of workforce automation. Pew Research Center 10.

Sondhi, P. (2009). Feature construction methods: a survey. sifaka. cs. uiuc. edu, 69, 70-71.

Srinidhi, B., and Gul, F. A. (2007). The differential effects of auditors' non-audit and audit fees on accrual quality. Contemporary Accounting Research.

Stanley, J. D., and DeZoort, F. T. (2007). Audit firm tenure and financial restatements: An analysis of industry specialization and fee effects. Journal of Accounting and Public Policy, 26(2), 131-159.

Stohr, E. A., and J. L. Zhao. (2001). Workflow automation: Overview and research issues. Information Systems Frontiers 33: 281-296.

Sullivan, J., and A. Zutavern. (2017). The mathematical corporation: Where machine intelligence and human ingenuity achieve the impossible. Hachette UK.

Sun, T. (2018). The incremental informativeness of the sentiment of conference calls for internal control material weaknesses. Journal of Emerging Technologies in Accounting 15 (1): 11–27. https://doi.org/10.2308/jeta-51969

Sun, T., and L. J. Sales. (2018). Predicting public procurement irregularity: An application of neural networks. Journal of Emerging Technologies in Accounting 15 (1): 141–154. https://doi.org/10.2308/jeta-52086

Suresh, H., and Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:1901.10002.

Sutton, S. G., M. Holt, and V. Arnold. (2016). "The reports of my death are greatly exaggerated"—Artificial intelligence research in accounting. International Journal of Accounting Information Systems, 22, 60-73.

Sutton, S. G., V. Arnold, and M. Holt. (2018). How much automation is too much? Keeping the human relevant in knowledge work. Journal of Emerging Technologies in Accounting 152: 15-25.

Syam, N., and A. Sharma. (2018). Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. Industrial Marketing Management, 69, 135-146.

Syed, R., S. Suriadi, M. Adams, W. Bandara, S. J. Leemans, C. Ouyang, and H. A. Reijers. (2020). Robotic process automation: Contemporary themes and challenges. Computers in Industry 115: 103162.

Tan, C. E., and Young, S. M. (2015). An analysis of "Little r" restatements. Accounting Horizons, 29(3), 667-693.

Tang, J., Alelyani, S., and Liu, H. (2014). Feature selection for classification: A review. Data classification: Algorithms and applications, 37.

Tang, M. M. (2013). What guides the guidance? An empirical examination of the dynamic disclosure theory. University of Rochester.

Thai-Nghe, N., Gantner, Z., and Schmidt-Thieme, L. (2010). Cost-sensitive learning methods for imbalanced data. In The 2010 International joint conference on neural networks (IJCNN) (pp. 1-8). IEEE.

Trueman, B. (1986). Why do managers voluntarily release earnings forecasts?. Journal of accounting and economics, 8(1), 53-71.

Vasarhelyi, M. A. (1984). Automation and changes in the audit process. Auditing: A Journal of Practice and Theory 41: 100–106.

Venable, J., J. Pries-Heje, and R. Baskerville. (2012). A comprehensive framework for evaluation in design science research. International Conference on Design Science Research in Information Systems (May): 423-438.

Wall Street Journal (2016). Wells Fargo: Where Was the Auditor? Available at: https://www.wsj.com/articles/wells-fargo-where-was-the-auditor-1478007838

Wang, E. and Tan, H. T. (2013). The effects of guidance frequency and guidance goal on managerial decisions. Journal of Accounting Research, 51(3), 673-700.

Willcocks, L. P., M. Lacity, and A. Craig. (2015a). The IT function and robotic process automation.

Willcocks, L. P., M. Lacity, and A. Craig. (2015b). Robotic process automation at Xchanging.

Williams, P. A. (1996). The relation between a prior earnings forecast by management and analyst response to a current management forecast. Accounting Review, 103-115.

Wood. D. A. (2016). Comparing the publication process in accounting, economics, finance, management, marketing, psychology, and the natural sciences. Accounting Horizons 303: 341–61.

Yan, Z., and K. C. Moffitt. (2019). Contract analytics in auditing. Accounting Horizons 333: 111-126.

Zhang, C. (2019). Intelligent process automation in audit. Journal of Emerging Technologies in Accounting.

Zhang, C. A., Cho, S., and Vasarhelyi, M. (2021). Explainable Artificial Intelligence (XAI) in Auditing: A Framework and Research Needs. Working paper.

Zhang, C., H. Issa, A. Rozario, and J. S. Soegaard. (2021). Robotic process automation RPA implementation case studies in accounting: A holistic perspective. Working paper. Rutgers University.

Zhang, C. A., Thomas, C., and Vasarhelyi (2022), M. Attended Process Automation in Audit: A Framework and A DemonstrationAttended Automation in Audit. Journal of Information Systems.

Zhang, C., J. Dai, and M. A. Vasarhelyi. (2018). The impact of disruptive technologies on accounting and auditing education: How should the profession adapt? The CPA Journal 889: 20-26.

Zhaokai, Y., and K. C. Moffitt. (2019). Contract analytics in auditing. Accounting Horizons, 33(3), 111-126.