ECON JOURNAL WATCH
Scholarly Comments on
Academic Economics

ECON JOURNAL WATCH 18(1)
March 2021: 71–78

A Response to "Critique of an Article on Machine Learning in the Detection of Accounting Fraud"

Yang Bao[1], Bin Ke[2], Bin Li[3], Y. Julia Yu[4], and Jie Zhang[5]

LINK TO ABSTRACT

This is a response to Stephen Walker's (2021) critique of our article, "Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach" (Bao, Ke, Li, Yu, and Zhang 2020). We received the final version of Walker's critique on February 24, 2021. Walker (2021) raises two empirical issues about our paper. The first one is about our treatment of missing values for the raw financial statement variables. The second one is about our treatment of serial fraud. We find no evidence that these two issues alter our paper's inferences. Walker (2021) also implies that we were not transparent in disclosure. To the contrary, we demonstrated our transparency by releasing not only our program codes but also our full data set, so that people can freely replicate and extend our findings.

Missing values

Following prior research, Bao et al. (2020) started with 28 raw financial variables for all fraud prediction models. Some observations of the 28 raw financial

[1] Shanghai Jiao Tong University, Shanghai, China 200030.
[2] National University of Singapore, Singapore 119245.
[3] Wuhan University, Wuhan, China 430072.
[4] University of Virginia, Charlottesville, VA 22903.
[5] Nanyang Technological University, Singapore 639798.

variables contain missing values. Walker (2021, 67) questions why we recoded the missing values to zero for *txp*, *ivao*, *ivst*, and *pstk* but not for the other raw financial variables. Our treatment of missing values is consistent with the existing literature, and we applied the same treatment consistently for all fraud prediction models. Specifically, we follow Mark Cecchini et al. (2010) and Patricia Dechow et al. (2011) by dropping the firm-year observations with missing values. The only exceptions are the raw variables *ivao* (Investments and Advances), *ivst* (Short-term Investment), *pstk* (Preferred/Preference Stock), and *txp* (Taxes Payable), which are set to zero if missing. These four raw data items are used to calculate the financial ratios "Changes in RSST accruals" and "Changes in working capital accruals," used in the logit model based on the 14 financial ratios in our published paper. Recoding missing values of *ivao*, *ivst*, *pstk*, and *txp* to zero follows the common practice in the prior accounting literature (e.g., Richardson et al. 2005; Dechow et al. 2011; Allen, Larson, and Sloan 2013). Scott Richardson et al. (2005, 451 n.8) explain in their study that these variables "represent a balance sheet item that may not be relevant for many companies (e.g., preferred stock), so we set them to zero rather than needlessly discarding observations."

Walker's approach to dealing with serial fraud

Walker (2021) also suggests that our treatment of serial fraud observations leads to incorrect inferences. Ours is one of the first studies to directly examine the influence of serial fraud on fraud prediction. To mitigate the influence of serial fraud cases that span both the training and testing samples on our inferences, we adopted a conservative approach in the published paper by recoding the serial fraud observations to zero (i.e., non-fraud) in the training sample if the serial fraud spans both the training and test periods. Walker questioned our implementation of this conservative approach because we required serial fraud to be *consecutive in our final sample*. Walker proposes an alternative approach by recoding the serial fraud observations to zero in the training sample if the serial fraud observations spanning both the training and test periods share the same primary AAER ID (i.e., *p_AAER* in Dechow et al.'s fraud database).[6] Walker claims that our best machine learning-based fraud prediction model, RUSBoost, underperforms in terms of precision the logit model based on the 14 financial ratios when his approach is adopted.

We believe that Walker's criticism is flawed for several reasons. First, he did not recalibrate the most important parameter of RUSBoost, number of trees,

[6] Note that *p_AAER* is the primary AAER number used by Dechow et al. (2011) to identify all the fraud years for the same firm that are part of the same fraud incident.

after changing the fraud training samples using his approach. As a result, he has understated the true performance of RUSBoost (see below for evidence). Second, Walker focuses only on the results in Table 3 of our published paper (using 2003–2008 as the test sample) while ignoring the results in Table 5 of our published paper (using 2003–2005, 2003–2011 and 2003–2014 as three alternative test samples), which is equally important for our inferences due to the underreporting of fraud in the latter part of our test years (i.e., 2006–2014). As explained in Sections 3.1 and 7.1 of our paper (Bao et al. 2020, 207–209, 224–225), many accounting frauds could remain undetected due to reduced regulatory enforcement of accounting fraud that approximately coincided with the 2008 financial crisis. In addition, Alexander Dyck et al. (2010) show that it takes approximately 24 months, on average, for the initial disclosure of a fraud, implying that the detected accounting fraud labels could be understated for the test years as early as 2006/2007. Therefore, the test years 2003–2005 should represent the best test period to assess the performance of different fraud prediction models. Third, there is no clear consensus on how to handle serial fraud in model building (see section 4 for more elaboration).

To address Walker's concern head on, here we replicate the results for our Tables 3 and 5 using Walker's approach. Walker does not replicate all of those results (Bao et al. 2020, 219 Table 3, 226 Table 5) using his approach. As Walker's approach would require the recoding of a significant number of fraud observations into non-fraud in the training sample (see footnote 8 below), we re-optimized an important model parameter of RUSBoost in model training (i.e., number of trees). The results are reported in the following Table 1.[7] We refer the reader to Bao et al. (2020) for the detailed definitions of the variables included in this table.

Several important findings emerge from Table 1. First, the performance of RUSBoost relative to the logit model based on the 14 financial ratios using both AUC and NDCG@k is the strongest for the test years 2003–2005, which are argued to be the best test period as noted above. Second, even though serial fraud is the most severe in 2003–2005, RUSBoost continues to outperform the logit model based on the 14 ratios after adjusting the serial fraud issue using Walker's approach. This evidence suggests that serial fraud is not driving the superior results of RUSBoost.[8] Third, contrary to Walker's claim that the precision of RUSBoost

[7] The results in Table 1 are generated using Matlab R2020b on Windows 10.

[8] Following Walker's approach, the percentage of recoded fraud firm years due to this serial fraud issue is close to 20 percent for each of the test years 2003–2005 but this percentage drops monotonically for the test years after 2005 (e.g., 6.12 percent for test year 2006, 3.74 percent for test year 2007, and so on). It is clear that the serial fraud issue is the most severe for the test years 2003–2005. Therefore, if serial fraud is a concern, it should negatively affect our RUSBoost model's relative performance the most for the test years 2003–2005. Our results in Table 1 do not support this prediction.

TABLE 1. Results using Walker's approach, corresponding to Tables 3 and 5 in Bao et al. (2020)

Panel A. Performance metrics averaged over the test period 2003–2005

	Model			
Metric	RUSBoost using 28 raw financial data items	Logit using 28 raw financial data items	Logit using 14 financial ratios	SVM-FK using 28 raw financial data items
AUC	0.7428	0.6736	0.6483	0.6255
NDCG@k	0.0394	0.0084	0.0109	0.0172
Sensitivity	0.0453	0.0097	0.0137	0.0265
Precision	0.0502	0.0113	0.0128	0.0164
# of True Fraud Obs Identified	9	2	2	4

Panel B. Performance metrics averaged over the test period 2003–2008

	Model			
Metric	RUSBoost using 28 raw financial data items	Logit using 28 raw financial data items	Logit using 14 financial ratios	SVM-FK using 28 raw financial data items
AUC	0.7228	0.6842	0.6711	0.6195
NDCG@k	0.0237	0.0042	0.0273	0.0199
Sensitivity	0.0291	0.0048	0.0399	0.0263
Precision	0.0281	0.0057	0.0262	0.0193
# of True Fraud Obs Identified	10	2	8	6

Panel C. Performance metrics averaged over the test period 2003–2011

	Model			
Metric	RUSBoost using 28 raw financial data items	Logit using 28 raw financial data items	Logit using 14 financial ratios	SVM-FK using 28 raw financial data items
AUC	0.7171	0.6981	0.6711	0.6297
NDCG@k	0.0243	0.0105	0.0235	0.0184
Sensitivity	0.0325	0.0171	0.0349	0.0300
Precision	0.0249	0.0100	0.0222	0.0168
# of True Fraud Obs Identified	13	5	10	9

Panel D. Performance metrics averaged over the test period 2003–2014

	Model			
Metric	RUSBoost using 28 raw financial data items	Logit using 28 raw financial data items	Logit using 14 financial ratios	SVM-FK using 28 raw financial data items
AUC	0.7196	0.7061	0.7015	0.6519
NDCG@k	0.0182	0.0101	0.0223	0.0261
Sensitivity	0.0244	0.0172	0.0345	0.0346
Precision	0.0187	0.0090	0.0185	0.0170
# of True Fraud Obs Identified	13	6	11	9

underperforms the precision of the logit model based on the 14 ratios for the test period 2003–2008, we find that the precision of RUSBoost continues to outperform the precision of the logit model based on the 14 ratios (2.81 percent vs. 2.62 percent). Specifically, the RUSBoost catches 10 accounting fraud firm years while the logit model catches 8 accounting fraud firm years in the test period 2003–2008. Fourth, using the AUC as a performance metric (a common performance evaluation metric in the fraud prediction literature), RUSBoost always outperforms the logit model based on the 14 ratios for all test periods in Table 1. Finally, similar to the reported results in our published paper, a counterintuitive finding in Table 1 is that the AUC of the logit model based on the 14 ratios increases over time for the test samples from 2003–2005 to 2003–2014, even though the problem of undetected fraud grows over time as noted above and therefore the reported fraud frequencies are not an accurate measure of the true fraud frequencies. Overall, we conclude that adopting Walker's approach does not alter our inferences.

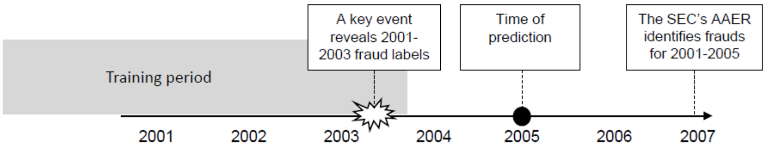
What is the optimal approach for dealing with serial fraud?

Walker's critique raises the broader question of whether there is an optimal approach for dealing with serial fraud. Ex ante, there is no clear answer to this question. As noted by Bao et al. (2020), most prior studies do not deal with the serial fraud issue at all and instead treat each fraud firm-year as an independent observation. If we follow the approach of these prior studies, as shown in Bao et al. (2020, 227 Table 6), the performance of RUSBoost would far exceed that of the logit model based on the 14 ratios.

It is also unclear whether Walker's approach of using *p_AAER* to define serial fraud is the most appropriate. This is because the AAERs issued by the SEC often come at the last stage of the fraud revelation timeline (Karpoff et al. 2017). Many other fraud revelation events (e.g., restatement, litigation, analyst report, whistleblower, etc.) can provide more timely fraud label information before the announcement of the SEC's AAERs. Hence, Walker's approach could be inappropriate and represent overkill.

To illustrate our reasoning, let's consider the following hypothetical scenario (see Figure 1 for the timeline): a firm has consecutive fraud labels from 2001 to 2005 identified by the SEC's AAER issued in 2007. These fraud labels are summarized in Dechow et al.'s database using the same *p_AAER*. The fraud labels for 2001–2003 were known to investors in 2004 due to a fraud revelation event in the middle of 2004, but the fraud labels for 2004–2005 were not known to investors until the date of AAER release at the end of 2007.

Figure 1. A serial fraud example with a key fraud revelation event during training period


A key event reveals 2001-2003 fraud labels
Time of prediction
The SEC's AAER identifies frauds for 2001-2005
Training period

	2001	2002	2003	2004	2005	2006	2007
Fraud labels identified by the same p_AAER:	1	1	1	1	1		
Fraud labels known at the time of prediction:	1	1	1				
Fraud labels recoded in training sample per Walker's approach:	0	0	0				

Now let's assume that, right after the release of fiscal year 2005's 10-K (i.e., time of prediction in Figure 1), we wish to predict whether the firm commits fraud or not in its 2005 annual report. As the fraud labels for 2001–2003 were already known to investors in 2004, we can include the available fraud labels in 2001–2003 to train a fraud prediction model for the test year 2005. On the other hand, Walker's approach would recode the fraud labels in 2001–2003 to zero in model training for test year 2005, which seems inappropriate because the fraud labels in 2001–2003 were already known to the public in 2005 and therefore should be included in model training.[9] We reviewed a few serial fraud cases and found that cases similar to our example above existed in our sample (e.g., AAER #2591 and AAER #2819), but it is beyond the scope of our study to identify all such cases.

In conclusion, Walker's approach of relying solely on *p_AAER* ID to define serial fraud could be inappropriate, given that the reported frequency of detected fraud in our sample is very low to start with (less than 1 percent, on average, as shown in Bao et al. 2020, 211 Table 1) and a portion of the serial fraud observations, which represents the majority of the fraud observations in our sample, could have been known to the public at the time of model training.

Appendix

Code related to this research is available from the journal website (link).

[9] Even if the fraud labels for 2001–2003 were known to investors in 2006 only (i.e., *after* the time of prediction in Figure 1), we can still use the fraud labels for the years 2001–2003 to retrain a prediction model for the test year 2005. As we did not know whether year 2005 was a fraud or not until 2007, such a retrained prediction model based on more updated fraud labels from the training years 2001–2003 should be still valuable to investors at the time of prediction 2005. Walker's approach would preclude the inclusion of these fraud labels in model training.

References

Allen, Eric J., Chad R. Larson, and Richard G. Sloan. 2013. Accrual Reversals, Earnings and Stock Returns. *Journal of Accounting and Economics* 56(1): 113–129.

Bao, Yang, Bin Ke, Bin Li, Y. Julia Yu, and Jie Zhang. 2020. Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach. *Journal of Accounting Research* 58(1): 199–235.

Cecchini, Mark, Haldun Aytug, Gary J. Koehler, and Praveen Pathak. 2010. Detecting Management Fraud in Public Companies. *Management Science* 56(7): 1146–1160.

Dechow, Patricia M., Weili Ge, Chad R. Larson, and Richard G. Sloan. 2011. Predicting Material Accounting Misstatements. *Contemporary Accounting Research* 28(1): 17–82.

Dyck, Alexander, Adair Morse, and Luigi Zingales. 2010. Who Blows the Whistle on Corporate Fraud? *Journal of Finance* 65(6): 2213–2253.

Karpoff, Jonathan M., Allison Koester, D. Scott Lee, and Gerald S. Martin. 2017. Proxies and Databases in Financial Misconduct Research. *Accounting Review* 92(6): 129–163.

Richardson, Scott A., Richard G. Sloan, Mark T. Soliman, and İrem Tuna. 2005. Accrual Reliability, Earnings Persistence, and Stock Prices. *Journal of Accounting and Economics* 39(3): 437–485.

Walker, Stephen. 2021. Critique of an Article on Machine Learning in the Detection of Accounting Fraud. *Econ Journal Watch* 18(1): 61–70. **Link**

About the Authors

Yang Bao is an assistant professor at Shanghai Jiao Tong University. His email address is baoyang@sjtu.edu.cn.

Bin Ke is a professor at National University of Singapore. His email address is bizk@nus.edu.sg.

Bin Li is a professor at Wuhan University. His email address is binli.whu@whu.edu.cn.

Y. Julia Yu is an assistant professor at University of Virginia. Her email address is julia.yu@virginia.edu.

Jie Zhang is an associate professor at Nanyang Technological University. His email address is zhangj@ntu.edu.sg.

Go to archive of Comments section
Go to March 2021 issue

Discuss this article at Journaltalk:
https://journaltalk.net/articles/6028/