Should Bao et al. (2020) be retracted?

Ian D. Gow

2022-10-11

Walker (2022) calls "for an investigation at the *Journal of Accounting Research* (JAR) into academic research misconduct" related to Bao et al. (2020). This paper examines whether this call is justified and considers possible responses JAR might take. As the title of this piece suggests, I reframe the call of Walker (2022) as a question likely more applicable to JAR: Should Bao et al. (2020) be retracted? I argue that some form of retraction is required, even if only to provide (through a "retract and republish" approach) a research record that is clear and free from known error.

However, I suggest that there are a number of factors that the editors of JAR might consider in deciding whether to provide the authors of Bao et al. (2020) the opportunity to republish.

1 Dramatis personae

Given the dramatic title, it seems appropriate to start with a listing of the characters appearing in this story. In order of appearance (in the world, not in this paper) we have:

- The Original (BKLYZ0, Bao et al. 2015): Version submitted to JAR in 2015.
- The Publication (BKLYZ1, Bao et al. 2020): Version published in JAR in 2020.
- The Critique (W1, Walker 2021a)
- The Reply (BKLYZ2, Bao et al. 2021): Initial reply to W1.
- The Rejoinder (W2, Walker 2021b)
- The Erratum (BKLYZ3, Bao et al. 2022): Erratum published in JAR in 2022.
- The Call for an Investigation (W3, Walker 2022): Response to BKLYZ3.

2 BKLYZ1

What exactly is the contribution of BKLYZ1? Bao et al. (2020, 199) suggest that it is "a state-of-the-art fraud prediction model" where the fraud that is predicted is accounting fraud

resulting in an Accounting and Auditing Enforcement Release (AAER) by the SEC.[1]

BKLYZ1 do indeed provide such a model. An analyst—whether an academic or a practitioner—can go to the GitHub page associated with BKLYZ1, download the code and data, open Matlab, and generate the model.[2]

But, even on the terms of BKLYZ1, this model has limitations. First, it doesn't really detect accounting fraud in a general sense, so much as AAERs. Accounting fraud might not result in AAERs, either because it is never detected, or because it is detected but does not rise to the level that leads to an AAER, or even because the fraud is so profound that an AAER is somewhat irrelevant.

With regard to the last category, it is not even clear that Enron, the public company at the heart of one of most notorious cases of accounting fraud this century, was the subject of an AAER. While the CEO (Jeffrey Skilling) and CFO (Andrew Fastow) of Enron ended up serving time in prison, there is no AAER related either Skilling or Fastow (many AAERs relate to individuals). There is no AAER directed specifically at Enron, perhaps because it entered bankruptcy shortly after fraud was detected.[3]

The BKLYZ1 sample "ends in 2008 because the regulators reduced the enforcement of accounting fraud starting from around 2009, increasing the possibility that many accounting fraud cases remain undetected for the post-2008 period" (Bao et al. 2020, 203–4). In other words, the specific outcome (AAERs) that the BKLYZ1 model is designed to predict becomes simply too difficult to predict after 2008. This means that the model is only useful for "predicting" AAERs before 2009. Obviously no practitioner would find such a model useful and even academics—who appear to have an unhealthy appetite for using outputs of prediction models in regression analyses—probably have little use for a model whose utility ended with financial crisis of 2008.

But even if the BKLYZ1 model itself is not useful for any identifiable purpose, perhaps the contribution of BKLYZ1 is in showing that a model based on ensemble learning ("one of the most powerful machine learning methods") "outperforms … by a large margin" approaches commonly used in prior accounting research, including models based on logistic regression. Of course, this would not be a contribution to the vast literature in statistical learning, as

[1] See the SEC website for details.

[2] In some ways, the requirement for Matlab is unfortunate, as Matlab is proprietary software and offers less transparency than an implementation using one of the open-source alternatives, such as Python or R, would. Unfortunately there is no implementation of RUSBoost in the popular Python library, `scikit-learn`, though this library does have an implementation of AdaBoost. There is also no well-documented implementation of RUSBoost in R, though there are several implementations of the AdaBoost in R. Fortunately, it is possible to implement RUSBoost in R, and I include an implementation in the R package `farr` that I created as a complement to the course book found here. A chapter using RUSBoost will be forthcoming in the near future.

[3] The one AAER in the Bao et al. (2020) sample connected to Enron actually covers the order for Citigroup to pay a amount in a settlement arising because "Citigroup assisted [Enron and Dynegy] in enhancing artificially their financial presentations through a series of complex structured transactions … to allow those companies to report proceeds of financings as cash from operating activities".

any practitioner in that field is unlikely to be surprised by a general result covered in an introductory textbook.[4]

Nor would we view the contribution of BKLYZ1 to be in demonstrating the superiority of the specific ensemble method used in the paper ("RUSBoost") over alternative approaches. First, BKLYZ1 do not evaluate RUSBoost relative to other ensemble methods, such as the AdaBoost approach on which it is based. Second, earlier research has already provided evidence on this point (Seiffert et al. 2008).

Rather the contribution of BKLYZ1 seems more specific to the setting of accounting fraud. Bao et al. (2020, 204) summarize their results:

> The average AUC and the average NDCG@k for the ensemble learning model are 0.725 and 0.049, respectively, representing a performance increase of 7.9% and 75%, respectively, relative to the performance of the better benchmark model, the Dechow et al. model [based on logistic regression]. These performance differences are also economically significant: Using the NDCG@k approach (where k = 1%), our best model, the ensemble learning model, identified a total of 16 fraud cases in the test period 2003–08 [versus] 9 for the Dechow et al. model.

Assuming that the results of BKLYZ1 generalize from the prediction of AAERs in 2003–2008 to prediction of accounting fraud, we might conclude that BKLYZ1 provides the helpful result that a model based on RUSBoost may provide superior performance in predicting accounting fraud than a model based on logistic regression. Presumably this was the central contribution of BKLYZ1 that led to its publication in an accounting outlet as prestigious as the *Journal of Accounting Research*.

3 BKLYZ3

Unfortunately, we need to update the central finding of BKLYZ1 in light of BKLYZ3. BKLYZ3 corrects a "coding error" in BKLYZ1 identified by W1. The updated results are found in Panel B of Table 1 of BKLYZ3. There we see that the quoted summary of results should instead read.

> The average AUC and the average NDCG@k for the ensemble learning model are 0.7228 and 0.0237, respectively, representing a performance increase of 7.7% and *performance decrease of 13.2%*, respectively, relative to the performance of the better benchmark model, the Dechow et al. model [based on logistic regression]. These performance differences are also economically significant: Using the NDCG@k approach (where k = 1%), our best model, the ensemble learning model, identified

[4] The general result here being that statistical learning methods such as ensemble learning usually improve out-of-sample prediction performance relative to models—such as logistic regression that tend to *overfit* the data used to train them.

> a total of 10 fraud cases in the test period 2003–08 [versus] 8 for the Dechow et al. model.

So, based on the criteria used in BKLYZ1, correction of the "coding error" in that paper overturns the central result of BKLYZ1. So what seems to be the sole basis for publishing BKLYZ1 is no longer true.

Indeed, W3 (Walker 2022, 192) suggests that the second sentence may need an additional correction to read:

> These performance differences are also economically significant: Using the NDCG@k approach (where k = 1%), our best model, the ensemble learning model, identified a total of 8 fraud cases in the test period 2003–08 [versus] 8 for the Dechow et al. model.

In my view, given the way the erratum is written, a reader who stumbled upon BKLYZ1 and then checked BKLYZ3 might be forgiven for missing how the primary result of BKLYZ1 is undermined by the correction in BKLYZ3. BKLYZ3 arguably obfuscates the main result of BKLYZ1 by emphasizing alternative test periods, placing renewed emphasis on AUC (the metric with weaker results in BKLYZ1), introducing novel results using NDCG@k at cut-offs not considered in BKLYZ1, and providing a rather beside-the-point discussion of an "alternative approach to coding serial fraud" that really does not provide an alternative approach.[5]

4 Are their grounds for retraction?

The Committee on Publication Ethics provides guidance "intended to advise editors and publishers on expected practices when considering whether a retraction is appropriate, and how to issue a retraction."[6] In light of BKLYZ3, "editors should consider retracting a publication if … they have clear evidence that the findings are unreliable … as a result of major error (eg, miscalculation or experimental error)." Given that BKLYZ3 provides evidence of an error and, based on our discussion above, this error goes to the central result of the BKLYZ1, retraction seems to be an appropriate response.

But note that the COPE guidelines [p. 8] provide the option of "retract and republish": "journals may wish to work with authors to concurrently retract an article that was found to be fundamentally flawed while simultaneously publishing a linked and corrected version of the work. This strategy … may provide an opportunity for journals and authors to transparently correct the literature when a simple correction cannot sufficiently address the flaws of the original article."

[5] My own analysis suggests that AAERs are never released prior to the last affected period, so AAERs that affect test years are always in the "future" relative to that test year, and should never be coded as anything other than zero in the training sample.

[6] See p.2 of Retraction Guidelines.

Note that while many grounds for retraction involve evidence of academic misconduct, there is no requirement that misconduct be shown for a retraction to occur. Indeed, it seems that the "retract and republish" option that has merit in the present cases is predicated on the *absence* of misconduct.

5 Is there evidence of misconduct?

W3 claims to "make the case that there is evidence of academic misconduct and make the recommendation that the *Journal of Accounting Research* launch a full and independent investigation into the matter." I can imagine that the editors of the *Journal of Accounting Research* would be reluctant to get into questions of whether there is academic misconduct, given the implications of any such allegation.

While I argue that a better presentation of the research record requires some kind of retraction, the COPE guidelines cited above appear to afford some latitude as to whether a journal will "in some instances, … may wish to work with authors to concurrently retract an article that was found to be fundamentally flawed while simultaneously publishing a linked and corrected version of the work." In other words, the *Journal of Accounting Research* arguably enjoys wide discretion over whether to "republish" a corrected version of BKLYZ1.

The remainder of this note collects some information that I conjecture the *Journal of Accounting Research* may consider in reaching its decision on the best course of action.

6 Some observations

6.1 Test period

One thing that BKLYZ1 is very clear on is that the test period is 2003–2008. The main results ("performance increase of 7.9% and 75%") of that paper are all based on this test period.

Yet the erratum [BKLYZ3, p. 1636] mysteriously seems to emphasize 2003–2005 as the test period: "Using NDCG@k as a performance measure RUSBoost … continues the dominate the performance of the other models for the test period 2003–2005." The published erratum [p. 1636] even suggests that BLKYZ1 "argued that this test period was the cleanest". Yet this is simply an impossible reading of BKLYZ1, where 2003–2008 is clearly the test period. The one common theme appears to be that the best test period is the test period that delivers the most favorable "out-of-sample" performance for RUSBoost, which would be a thoroughly dishonest approach to use.

At this point, a careful reader might point out that 2003–2005 was actually the test period used in BKLYZ0. Hopefully, BKLYZ would not be the ones to point this out, because this choice of test period was justified in BKLYZ0 on the following two items. First, BKLYZ0 state

[p. 4] that "the SEC's Accounting and Auditing Enforcement Releases (AAERs) available to us end in September 2010". Second, "there is an average of five-year gap between a fraud occurrence and the AAER publication date."

As the sample period in BKLYZ1 includes AAERs that extend to 2014, the first item seems to suggest that (on the logic of BKLYZ0 itself) the test period could be updated to 2003–2009 (i.e., adding four years) with the new sample.

But it is important to note that BKLYZ0 included in their training sample frauds that were also in the test periods, even though the AAER publication dates would have been after the test periods in question.[7]

If the second statement were maintained and the issues of "serial fraud" were addressed, then by the logic of BKLYZ1, the "gap" of two years used in BKLYZ1 would have to be five years, and thus the feasible test sample could not *begin* until 2006. BKLYZ1 [p. 209] "require a gap of 24 months between the financial results announcement of the last training year and the results announcement of a test year … because Dyck, Morse, and Zingales (2010) find that it takes approximately 24 months, on average, for the initial disclosure of the fraud."[8]

6.2 The "coding error"

BKLYZ3 states [p. 1635] that [W1 and W2] "identified an error in the program codes [sic] of BKLYZ1 posted on Github that led to an overstatement of model performance metrics. This erratum corrects this error …." It is difficult to disagree with W3's claim that this statement is false. There is nothing in the code posted on GitHub that created this error, instead the "error" was in data used by that code.

W3 claims that "to this date, the authors have offered no explanation as to why they did what they did" in recoding certain frauds to have different identifiers. This seems correct. In BKLYZ2, BKLYZ provide what may be best described as a non-explanation for what they did.

The example in Figure 1 illustrates the "coding error" made in BKLYZ1 if there are missing items for the affected firm in 2004. If such missing items exist, firm-years in 2001 through 2003 would be given a different fraud ID from the 2005 firm-year (in such cases BKLYZ appended the character "1" to the fraud ID for 2001 through 2003, and the character "2" to the fraud ID for 2005). Because the fraud IDs for 2001–2003 no longer appear in the test year (2005), they are not recoded as zero. In BKLYZ2, the practice of *not* recoding these frauds in this way is described as "Walker's approach" even though (as W2 points out), this is not so much Walker's approach as the approach described in BKLYZ1.

[7] That this is the case is implicit in footnote 10 to BKLYZ1 and the discussion under "serial fraud" in BKLYZ1 that is not found in BKLYZ0.

[8] While this shorter period is definitely convenient for BKLYZ1, it seems less convenient that there is no evidence of the claim in Dyck, Morse, and Zingales (2010) itself. Perhaps the BKLYZ1 authors obtained underlying data from the authors of Dyck, Morse, and Zingales (2010).

That this approach to recoding frauds is not legitimate is evidenced by the fact that doing it that way has been relabelled a "coding error" in BKLYZ3.

But relabelling "serial frauds" also does not make sense in that it is recoding a fraud as 1 in 2003 so that a "prediction" of the same underlying fraud can be made in 2005 even though, by the terms of the example itself (Figure 1 of BKLYZ), the SEC does not release an AAER until 2007.

That training models using data on frauds that are released until after the test period is problematic seems obvious. And it seems clear from BKLYZ1's discussion of "serial fraud" that the authors were well aware of these issues. In fact, they seem aware of the underlying issue even in BKLYZ0. As discussed above, the BKLYZ0 "sample ends in 2005 because … a significant portion of the accounting frauds that occurred over 2006–2010 are likely still unreported by the AAERs as of the end of 2010."

One response the authors might have is that there might be a "key event [that] reveals 2001–2003 fraud labels" before 2004 (and before the AAER publication date in 2007). But this implies a completely different prediction problem from that studied in BKLYZ1 (or BKLYZ0), which identifies frauds as AAER events. The only reliable "key event" that identifies an AAER is the release of an AAER. As discussed above, disclosed accounting fraud might not result in AAERs for a number of reasons.

Saying that sometimes information is released that suggests a high likelihood of a future AAER event transforms the prediction problem from one about predicting confirmed AAERs using financial statement features into one about predicting confirmed AAERs using financial statement features and also some *unidentified* information about possible future AAERs.[9]

Even if we expand the information set to include the "maybe-future-AAERs" as is done in the "not Walker's approach" depicted in Figure 1 of BKLYZ2, there is no rationale provided as to why missing values for some items on Compustat in 2004 should be assumed to precipitate a fraud revelation event before 2004. In fact, it is hard to conceive of one.

And this "coding error" is not something that happened by mistake. The strident defence of their approach provided in BKLYZ2 suggests that the authors did this consciously (only in BKLYZ3 do they suggest it was a "coding error"). And the "coding error" would require some tricky coding. In replicating BKLYZ1 for a chapter on prediction for a book on accounting research, I needed 16 lines of code to create the "new" fraud ID used in BKLYZ1 and, based on other code posted by the authors, I conjecture that either they needed significantly more than 16 lines of code, or perhaps they just recoded the fraud IDs manually.

At the very least, I think the authors should be required to provide more information of how the "coding error" was implemented so that the editors can assess the likelihood that it was indeed a "coding error" (as it is framed in BKLYZ3) and not a deliberate research design choice

[9] This a completely different prediction problem is perhaps the one embedded in BKLYZ0, which appears to use information about serial frauds that will be revealed after the test period in developing a training model using those very same frauds.

(as it is framed in BKLYZ2). If it seems that it was a deliberate research design choice, then I think the authors need to explain the rationale for it and also the rationale for dissembling its existence in the code posted to GitHub upon publication of BKLYZ1.

One possible response by the BKLYZ team might be that they have more than complied with the JAR data policy in effect when they submitted their paper and therefore do not need to account for the "coding error" beyond the code they have already provided. I do not think this kind of response would be helpful to their case.

There appears to be no data policy at *The Accounting Review* (see discussion here), but this did not prevent the retraction of Bird and Karolyi (2017), where the journal stated "the authors were unable to provide the original data and code requested by the publisher" to support an assertion made in the paper.

Given the "coding error" of BKLYZ3 appears to have been a conscious decision (see BKLYZ2), I think that the onus is on the authors of BKLYZ1 to show that the "coding error" was made in good faith.

6.3 Meta-parameters

In BKLYZ2, the authors suggested that W1 was flawed because he "did not recalibrate the most important parameter of RUSBoost, number of trees, after changing the fraud training samples using his approach [i.e., the approach described in BKLYZ1]." This is somewhat understandable, as BKLYZ1 does not describe the process of calibrating these meta-parameters in the first place. Indeed, there are several meta-parameters used in BKLYZ1: number of trees, `MinLeafSize`, `LearnRate`, and `RatioToSmallest`. It is critical that such meta-parameters be fixed using the training and validation data prior to evaluating model performance against test data, lest these parameters be selected based on test performance, thus overstating the predictive value of the model.

In this regard, it is somewhat concerning that the number of trees parameter that is selected for the BKLYZ1 specification is 1,000 according to W3, not the 3,000 used in BKLYZ1.[10] This creates the unfortunate impression that the idea of selecting meta-parameters in a transparent fashion using only validation data emerged only after observing a decline in test performance upon switching to "Walker's approach" in preparing the erratum. Some credit is due to BKLYZ and JAR for facilitating this sharing of data and code in the first place.[11]

[10] In my own analysis, I found that 900 trees maximized performance in the validation sample, which is very close to the value found in W3.

[11] We should not infer that the situation is better at journals without a data-sharing policy. If anything, we might expect them to be worse.

6.4 Other data issues

While it is possible to explain changes in the code between BKLYZ1 and BKLYZ3 using GitHub, some concerns remain. For example, as detailed here, the two data files provided are not consistent. While the PDF-rendered SAS code suggests that one data file depends on the other, there are observations on AAERs not found in the former.

6.5 How unusual are the issues in BKLYZ1?

Some readers may be surprised to learn that the core results of a published paper can disappear when a "coding error" is detected and corrected. I am not surprised. Having replicated many papers, I conjecture that the kinds of issues observed with BKLYZ1 are commonplace. Papers that say one thing, but do another (most papers using "regression discontinuity designs" fit here). Papers with genuine coding errors. Papers with results that are very sensitive to "design choices" that are difficult to rationalize (see here and here). If such issues abound, then the merit of singling out BKLYZ1 seems to be low.

Another factor that seems relevant is JAR's data policy. While not perfect, arguably JAR's policy was critical in helping to unearth the issues not only in BKLYZ1, but in the replications I refer to above. If BKLYZ1 had been accompanied by a very perfunctory effort to comply with the data policy (e.g., "here is the list of CIKs for the fraud firms in our sample"), then it would not have been possible to detect the issue raised by W1 and corrected in BKLYZ3.

References

Bao, Yang, Bin Ke, Bin Li, Y. Julia Yu, and Jie Zhang. 2015. "Detecting Accounting Frauds in Publicly Traded U.S. Firms: New Perspective and New Method." https://www.rieb.kobe-u.ac.jp/tjar/conference/6th/CA3_YingriJuliaYU.pdf.

———. 2020. "Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach." *Journal of Accounting Research* 58 (1): 199–235. https://doi.org/10.1111/1475-679X.12292.

———. 2021. "A Response to 'Critique of an Article on Machine Learning in the Detection of Accounting Fraud'." *Econ Journal Watch* 18 (1): 71–78. https://econjwatch.org/articles/a-response-to-critique-of-an-article-on-machine-learning-in-the-detection-of-accounting-fraud.

———. 2022. "Erratum." *Journal of Accounting Research* 60 (4): 1635–46. https://doi.org/10.1111/1475-679X.12454.

Bird, Andrew, and Stephen A. Karolyi. 2017. "Governance and Taxes: Evidence from Regression Discontinuity." *The Accounting Review* 92 (1): 29–50. https://doi.org/10.2308/1558-7967-92.1.000.

Dyck, Alexander, Adair Morse, and Luigi Zingales. 2010. "Who Blows the Whistle on Corporate Fraud?" *The Journal of Finance* 65 (6): 2213–53. https://doi.org/10.1111/j.1540-6261.2010.01614.x.

Seiffert, Chris, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. 2008. "RUSBoost: Improving Classification Performance When Training Data Is Skewed." In *2008 19th International Conference on Pattern Recognition*, 1–4.

Walker, Stephen. 2021a. "Critique of an Article on Machine Learning in the Detection of Accounting Fraud." *Econ Journal Watch* 18 (2): 61. https://econjwatch.org/articles/critique-of-an-article-on-machine-learning-in-the-detection-of-accounting-fraud.

———. 2021b. "Rejoinder to the Critique of an Article on Machine Learning in the Detection of Accounting Fraud." *Econ Journal Watch* 18 (2): 230. https://econjwatch.org/articles/rejoinder-to-the-critique-of-an-article-on-machine-learning-in-the-detection-of-accounting-fraud.

———. 2022. "Erroneous Erratum to Accounting Fraud Article." *Econ Journal Watch* 19 (2): 190–203. https://econjwatch.org/articles/erroneous-erratum-to-accounting-fraud-article.